On the Privacy Risks of Algorithmic Fairness

Hongyan Chang, Reza Shokri
School of Computing
National University of Singapore
{hongyan, reza}@comp.nus.edu.sg

***Abstract*—Algorithmic fairness and privacy are essential elements of trustworthy machine learning for critical decision making processes. Fair machine learning algorithms are developed to minimize discrimination against protected groups in machine learning. This is achieved, for example, by imposing a constraint on the model to equalize its behavior across different groups. This can significantly increase the *influence* of some training data points on the fair model. We study how this change in influence can change the information leakage of the model about its training data. We analyze the privacy risks of statistical notions of fairness (i.e., equalized odds) through the lens of *membership inference attacks*: inferring whether a data point was used for training a model. We show that fairness comes at the cost of privacy. However, this privacy cost is not distributed equally: the information leakage of fair models increases significantly on the unprivileged subgroups, which suffer from the discrimination in regular models. Furthermore, the more biased the underlying data is, the higher the privacy cost of achieving fairness for the unprivileged subgroups is. We demonstrate this effect on multiple datasets and explain how fairness-aware learning impacts privacy.**

***Index Terms*—Algorithmic Fairness, Data Privacy, Machine Learning, Membership Inference Attacks**

1. Introduction

The rapid growth of applying machine learning algorithms for automated decision making in sensitive domains has raised concerns that machine learning models might reflect and amplify existing human bias, especially in applications of significant individual-level consequences (e.g., healthcare or bail assignment) [1, 2]. An example of discrimination in machine learning is shown evidently by the Pro-Publica investigation of COMPAS (an algorithm for scoring criminal defendant's likelihood of reoffending) [1], the disparity on accuracy in computer vision systems [3], and the gender bias in word vectors [4].

In order to reduce the effect of human bias and minimize discrimination in machine learning, there has been a flurry of research work recently, which includes different mathematical notions of fairness and various fair learning algorithms [5–14]. The statistical notions of fairness (i.e., equalized odds) [8] suggest equalizing the model's predictive power across groups that are identified based on a protected attribute (e.g., race or gender). A prevalent approach is to add fairness constraints that encode the criteria that a classifier must satisfy for the learning problem to achieve fairness.

At the same time, the growing rate of applying machine learning algorithms on personal data has raised a prominent concern that the individuals' information might be leaked through models which are trained on them. Recent work shows that machine learning models are highly susceptible to leaking information about the members of their training data. *Membership inference attacks* achieve a high accuracy against machine learning models in a wide range of settings, where the adversary infers if a data point was used for training a model [15–21]. The risk of such attacks is particularly striking when the membership can reveal an individual's sensitive attributes.

Privacy and fairness, as two societal concerns about machine learning, do not exist in isolation. For instance, in the recidivism prediction application, demographic groups (e.g., black defendants and white defendants) should experience similar treatments, namely similar accuracy. Simultaneously, participation in the training data means that an individual once committed a crime. Therefore, fairness and membership privacy are both needed for ethical use of machine learning. It is, therefore, imperative to understand the interactions between them.

Despite the importance of this matter, there has been little effort to analyze the information leakage of fairness-awareness learning. Models can be trained with differential privacy and fairness constraints [22–24]. It is shown that differentially private models (i.e., DP-SGD [25]) have a larger accuracy reduction on "underrepresented" subgroups [26]. The privacy-preserving learning algorithms worsen the "unfairness" of the learned models. In other words, privacy comes at the cost of fairness. In this paper, we ask the complementary yet related question: *Is there a privacy cost for achieving fairness?* In other words, does imposing fairness constraints on the learning algorithm impact privacy risk of the training data?

Towards answering this question, we first formalize data privacy risk as the success of membership inference attacks. This reflects the information leakage of the model about the *individual* data points in its training set. The adversary observes the model's predictions and aims to find a "distinguisher" that identifies members and non-members of the training dataset.

Finding a single global "distinguisher" for all data points (e.g., a threshold on the loss of the model on its inputs) is the prevalent method in the existing membership inference attacks [17, 19, 20]. However, this approach is evidently sub-optimal. In practice, data can come from distinct distributions. For instance, samples from different groups can have slightly different underlying distributions. Thus, the machine learning models might learn different

patterns on each group, for the same task. This can lead to the performance disparity of the learned models, which motivates using fair algorithms. Therefore, the way that the model's predictions (loss) for training versus test data differs, can be distinct from one group to another. This can be exploited by the membership inference attacks.

We, thus, propose an effective attack strategy where the adversary finds a "distinguisher" per sub-group (e.g., qualified male applicants in the loan approval application). We empirically show that this simple modification of existing membership inference attacks results in a higher attack accuracy, thus leads to a more accurate estimation of the privacy risk. We focus on the information leakage of models through their output: *black box* setting, in which the adversary cannot observe the internal state of the model (e.g., its parameters and gradients).

Based on our attack strategy, we empirically show that the fairness-aware learning raises the privacy risk of the unprivileged subgroup and have disparate impacts on the privacy risk of subgroups (which is determined by the labels and protected attributes). We show that fairness comes at the cost of privacy, and the privacy cost is not equal across subgroups. Furthermore, there is a trade-off between fairness and privacy. When the underlying data (and corresponding unconstrained model) is more "unfair", the trained fair models leak more information about the unprivileged subgroup. Additionally, the more fair a model on its training dataset is, the higher the privacy risk of the model on unprivileged subgroups is.

We use synthetic datasets to study how, when, and why models trained using the fair learning algorithm leak more information about the training data, in a wide range of settings. Intuitively, fairness-aware learning imposes fairness constraints to force models to equally fit the unprivileged subgroup. Yet, when the size of the unprivileged subgroup is small, or the data in the unprivileged subgroup is hard to model, fair models memorize the training data from the unprivileged subgroups (instead of learning a general pattern on them). This memorization gives rise to high privacy risk as it is easier for the adversary to infer the membership. Hence, the unprivileged subgroups have a higher privacy risk on fair models.

We also conduct experiments on multiple real-world datasets, including the Law School dataset [27], Bank Marketing dataset [28], and COMPAS datasets [29]. After imposing fairness constraints, the privacy risk increases from 64.5% to 70.0% for the smallest subgroup on the Bank Marketing dataset. On the contrary, for the largest subgroup, the privacy risk only increases by 0.5%. In addition, tightening the fairness constraint on fair models consistently increases the privacy risk of subgroups. With a tighter fairness constraint on the Law School dataset, the privacy risk of a subgroup on fair models increases by 4.1%.

2. Related Work

2.1. Fairness

Various algorithmic fairness definitions are studied in the literature, including metric equality across sensitive groups [6, 8], individual fairness [7], causality [10]. For training models that satisfy those fairness definitions, many techniques are proposed in recent years, such as pre-processing methods [11, 14], in-processing methods [5, 9, 12, 13, 30], and post-processing methods [8]. In pre-processing methods, the goal is to find a new representation of data to retain information of input features about the learning task without the information that can lead to bias. In-processing methods enforce fairness during the training process by incorporating the fairness principle into the objective function or reducing the constrained optimization problems to a sequence of cost-sensitive classification problems. Instead, given a trained model, post-processing methods correct its predictions to satisfy fairness criteria. As mentioned previously, fairness and privacy concerns do not exist in isolation. We analyze the interaction between fairness and privacy by investigating the information leakage of fair machine learning algorithms. In this work, we focus on Equalized odds [8] (a statistical notion of fairness) and use in-processing approaches [5] to train fair models and leave the analysis of privacy risk for other fair algorithms and other notions of fairness to future work.

2.2. Membership inference attacks

In the machine learning context, the membership inference attacks aim to determine whether a given data point was used to train the model or not [17, 19, 21, 31]. In prior works, membership inference attacks are used to measure the information leakage of different machine learning algorithms, including deep learning algorithms [19], adversarial robust learning algorithms [18], learning algorithms for explanations models [16], learning algorithms for embedding models [15] and reinforcement learning algorithms [32]. It is worth emphasizing that our main goal is to analyze the membership information leakage of fair learning algorithms by taking advantage of the membership inference attacks instead of designing membership inference attack algorithms.

Yaghini et al. [33] demonstrates the information leakage of standard machine learning algorithms (without fairness consideration) varies across subgroups (e.g., female versus male). Imposing fairness constraints during the training does not eliminate the disparity of vulnerability. The results are in coincide with the findings we have on unconstrained models. It worth highlighting that our focus is on the effect of fairness constraints on the privacy risk of subgroups and individuals. In other words, we aim to analyze the privacy cost of fairness instead of the disparity of the vulnerability. The follow-up work [34] propose to evaluate the privacy risk of individuals and not only the privacy risk in aggregate and show that there are highly vulnerable records even the average attack accuracy is low. We analyze how fairness constraints affect individual privacy risk and show that fairness constraints can raise the privacy risk of individuals. In addition, the privacy risk of the most vulnerable points increases after imposing fairness constraints.

2.3. Privacy and Fairness

We center on the privacy cost of fairness. By no means we are the first to consider the interaction be-

tween fairness and privacy. Early work [7] explores the relationship between fairness in machine learning and differential privacy. The authors point out that the tools from differential privacy can be adapted for satisfying fairness constraints. Later on, the position paper [35] raises questions about understanding how fairness and privacy interact. Our results provide answers to one of the questions in the paper that fairness enhancement schemes could diminish the privacy of its subjects.

Kuppam et al. [36] shows that resource allocation based on differentially private statistics can affect some subgroups disproportionately. In the machine learning context, Bagdasaryan et al. [26] study the impact of differential privacy on the accuracy drops on subgroups and demonstrate that if the original model is unfair, the unfairness becomes worse once privacy-preserving algorithms (i.e., DP-SGD [25]) is applied. This implies that privacy comes at the price of fairness. In our work, we show that there is a privacy cost of fairness.

Several works propose algorithms to learn a model that satisfies differential privacy and fairness (demographic parity [22, 23], equality of opportunity [24]) simultaneously. Recently, Tran et al. [37] introduce a differential privacy framework to train deep learning models that satisfy several group fairness notions, including equalized odds, accuracy parity, and demographic parity. Instead, we are interested in analyzing how the privacy risk of individuals and subgroups changes after imposing fairness constraints. In addition, it should be pointed out that the works [37, 38] propose differentially private fair learning algorithms for equalized odds (the focus of our paper) guaranteeing the privacy of protected attributes (e.g., race, gender) instead of the membership information. Thus, their privacy concerns are different from ours.

3. Background

We study membership privacy risks related to machine learning algorithms and fair machine learning algorithms. We start with introducing machine learning and fairness, followed by notations we use throughout the paper.

3.1. Machine Learning

A machine learning model can be viewed as a function mapping inputs (features) to an output. To be more specific, the input is typically a vector of feature values, and the output is a label for classification and a real number for regression. We consider classification tasks and denote a machine learning model as $M : \mathcal{X} \to \mathcal{Y}$, mapping from the input (feature) space $\mathcal{X}$ to the output space $\mathcal{Y}$. We let M_S denote the model obtained by applying a machine learning algorithm on a *training set* S of size n, where each data point in S is sampled i.i.d. from a data distribution $\mathcal{D}$. In the typical machine learning setting, the machine learning algorithm outputs a model that minimizes a loss function ℓ on the training set S. We refer to the models obtained by the standard learning algorithm (without fairness consideration) as *unconstrained models*.

3.2. Fairness

The central problem of fair machine learning is to enhance machine learning algorithms with fairness principles to ensure that the value of a protected feature (e.g., race, gender) does not 'unfairly' influence a learning algorithm's outcome. Each data points can be represent by $z = (x, g, y) \in \mathcal{X} \times \mathcal{G} \times \mathcal{Y}$, where $x \in \mathcal{X}$ represents a set of features which is the input to the machine learning model, $g \in \mathcal{G}$ represents the *protected feature* and $y \in \mathcal{Y}$ represents a label. We consider the binary classification setting, where $y \in \mathcal{Y} = \{-, +\}$. We use X, Y, and G denote the random variables associated with the feature vector, the label, and the protected attribute, respectively. The input vector X can either contain G as one of its features or contain other features correlated with G. For instance, each data point corresponds to an applicant in the loan approval application, where X represents the demographics, income level, and loan amount, and G represents race. X can contain G or other features such as zip code that is often correlated with race.

Assume all the data points are split into groups based on the protected feature, *group fairness* requires that different protected groups experience the same treatment in an average sense. Throughout the paper, we use *equzlied odds*, which is a widely-used definition for group fairness [8]. A model is fair with respect to equalized odds if the accuracy for different groups is the same. In other words, given the true label for a data point, a fair model's prediction on a data point and its protected attribute are conditionally independent. Following previous works [5, 39], we use a relaxed notion of equalized odds allowing a small violation of fairness, formally defined as follows:

Definition 1 (δ - Equalized Odds Fairness). A classifier M satisfies δ-Equalized Odds conditions with respective to the protected attribute G, if for all $g, g' \in \mathcal{G}$, the false positive rate and false negative rate of the classifier in the group $\{G = g\}$ and $\{G = g'\}$ are within δ of one another. In other words,

$$\Delta(M, \mathcal{D}) \triangleq \max_{\substack{y \in \{-,+\} \\ g, g' \in \mathcal{G}}} \Big| \Pr_{\mathcal{D}}[M(X) \neq y | S = g, Y = y] - \Pr_{\mathcal{D}}[M(X) \neq y | S = g', Y = y] \Big| \leq \delta, \tag{1}$$

where the probabilities are computed over the data distribution $\mathcal{D}$. We refer to Δ as the model's **fairness gap** under equalized odds. A model satisfies exact fairness under equalized odds when $\delta = 0$.

In practice, the data distribution $\mathcal{D}$ is unknown. A fair model is learned by ensuring δ-fairness empirically on the training set S, e.g., through minimizing the model's empirical loss under δ-fairness as a constraint or post-processing [8], where the probability in (1) is computed over the training set S. We denote the δ-fair model trained on S as M_S^δ and refer to δ as *enforced fairness level*.

3.3. Notations

We let $\mathcal{D}_g^y$ denote the distribute of the data from the with protected attribute $G = g$ and label $Y = y$. In addition to that, we use G_g to represent the group $\{(X, G, Y) | G = g\}$ and let G_g^y represent the subgroup

TABLE 1: List of Notations

Symbol	Description	Section
z	Data point.	3.1
x	Feature vector.	3.1
g	Protected feature.	3.1
y	A label	3.1
$\mathcal{X}$	Feature space.	3.1
$\mathcal{Y}$	Label space	3.1
$\mathcal{G}$	Protected feature space.	3.1
$\mathcal{D}$	Data distribution.	3.1
S	Training set.	3.1
X	Random variable associated with feature vector.	3.1
Y	Random variable associated with label.	3.1
G	Random variable associated with protected attribute.	3.1
M	Machine learning model.	3.1
M_S	Machine learning model trained on S.	3.1
ℓ	Loss function.	3.1
Δ	Fairness gap.	3.2
δ	Enforced fairness level.	3.2
M_S^δ	δ-fair model trained on S.	3.2
δ	Enforced fairness level.	3.2
$\mathcal{D}_g^y$	Distribution of the data with protected attribute g and label y.	3.3
G_g	Group with protected attribute g.	3.3
G_g^y	Subgroup with protected attribute g and label y.	3.3
$\mathcal{A}$	Adversary.	4.1
$\mathcal{K}$	Knowledge about the model that is known to the adversary.	4.1
A	Learning algorithm.	4.1
A_S	Machine learning model obtained by applying the learning algorithm A on S.	4.1
$\mathcal{A}_\ell$	Adversary based on loss threshold.	4.2
$\mathcal{N}$	Gaussian distribution.	5

with $\{(X, G, Y)|G = g, Y = y\}$. Table 1 lists all the notations used in the paper.

4. Problem Statement

Membership privacy is about the information, whether a participant *opts* in or *out* of the input dataset of an algorithm. In the machine learning context, the algorithm is the learning algorithm, the input dataset is the training dataset, and the output is the learned model. The membership privacy concern is evident when the training dataset contains sensitive information about individuals. For instance, in the recidivism prediction application, the training dataset is collected from individuals who have committed crimes once. In this setting, membership reveals the criminal records of individuals, which is sensitive information. As demonstrated by the *differential privacy* [40], a formal definition of privacy, we say that privacy is preserved if the output distributions are indistinguishable when a participant in or out of the input dataset.

At the same time, in automated decision-making applications, fairness is also a serious social concern. It requires that different groups (e.g., male versus female, black people versus white people) experience the same treatment. In other words, the learned model should not discriminate on the protected attribute (e.g., race, gender). For example, in the recidivism prediction application, the model should have a similar true positive rate and true negative rate on white defendants and black defendants. Multiple algorithms are proposed to build fair classifiers by imposing fairness constraints during training or after training. More precisely, the learning algorithm produces models that have a minimal loss on the training dataset and satisfy δ-fairness on the training dataset.

Motivated by the recidivism prediction application, we find the fact that in decision-making applications where fairness is a pressing need, the training dataset typically contains sensitive information about individuals (e.g., recidivism prediction application, loan approval application, health condition assessment). The immediate question is whether or not fair machine learning leaks more information about the training data. More specifically, we ask whether, how, why, and when models trained by fair machine learning algorithms leak more membership information about the private data than models obtained by standard learning algorithms (without fairness consideration).

Towards answering our questions, we first formalize the privacy risk and present a method for analyzing the privacy risk. In Section 5, we show our experimental findings to answer our questions.

4.1. Definition of Privacy Risk

In this subsection, we formalize the privacy risk of individuals and subgroups. Recall that, the privacy is preserved if the output distributions are indistinguishable when a participant in or out of the input dataset. In other words, the privacy of a participant is preserved if upon observing the output, an adversary could not be able to tell whether the record of the participant is in training dataset or not.

Based on this, it is natural to measure the privacy risk of an individual as the success of the adversary whose goal is to infer whether or not the individual's data record was used for training a model. Such attacks are called *membership inference attacks* which are used as a *tool* to measure information leakage of different machine learning algorithms, including deep learning algorithms [19], adversarial robust learning algorithms [18], learning algorithms for explanations models [16], learning algorithms for embedding models [15] and reinforcement learning algorithms [32]. In our paper, by leveraging the membership inference attacks, we analyze the information leakage of fair machine learning algorithms and the gap of the leakage between standard learning algorithms and fair learning algorithms.

Roughly speaking, in membership inference attacks, an adversary attempts to infer whether a specific data record (challenge data point) was included in the training dataset of a learned model or not. Suppose a training dataset S of size n and each data point in S is sampled i.i.d. from the distribution $\mathcal{D}$, a model A_S is learned by applying a learning algorithm A on S. The learning algorithm A can be a fair learning algorithm or a standard learning algorithm. Given knowledge about the trained model, which is denoted as $\mathcal{K}(A_S)$, and the learning algorithm A, the adversary $\mathcal{A}$ needs to infer whether a challenge point z is a member of S or not. It is important to emphasize that the adversary can only access the learned model via function $\mathcal{K}$. In *black-box setting*, the $\mathcal{K}$ is generally the loss function of the learned model on test data points[20]. In *white-box setting*, the $\mathcal{K}$ outputs the learned model [21]. In this paper, we are interested in the privacy risk of individuals through the output of the

model. Accordingly, we measure the privacy risk in the black-box setting where the adversary only observes the loss.

To precisely formulate the membership inference attack, we describe an attack game in Attack Game 1 played between two parties: the challenger and an adversary. The challenger first samples fresh dataset S' of size $n-1$ from $\mathcal{D}$ and then flips a coin b to decide whether to add the challenge point z or add a point sampled from $\mathcal{D}$ to the training set S. Given the knowledge describe before, the adversary $\mathcal{A}$ needs to guess the value of b. The game outputs 1 when the adversary wins the game, namely, the adversary infers the membership of z for training dataset S successfully.

Attack Game 1 (Membership Inference Attack AG$(z, \mathcal{A}, A, n, \mathcal{D}, \mathcal{K})$). Let z be the challenge point, $\mathcal{A}$ be an adversary, A be a learning algorithm, n be a positive integer, $\mathcal{D}$ be a distribution over $(\mathcal{X}, \mathcal{G}, \mathcal{Y})$ and $\mathcal{K}$ be the information about a model given to the adversary. The attack game proceeds as follows:

1) The challenger samples $S' \sim \mathcal{D}^{n-1}$ and chooses $b \leftarrow \{0,1\}$ uniformly at random.
2) If $b = 0$, the challenger draws $z' \sim \mathcal{D}$ and lets the training dataset $S = S' \cup \{z'\}$. If $b = 1$, the challenger lets the training dataset $S = S' \cup \{z\}$ if $b = 1$.
3) The challenger trains a model A_S by applying the learning algorithm A on S and sends $(z, \mathcal{K}(A_S), A)$ to the adversary.
4) The adversary outputs $\mathcal{A}(z, \mathcal{K}(A_S), A) \in \{0,1\}$.
5) The game outputs 1 if $\mathcal{A}(z, \mathcal{K}(A_S), A) = b$ and 0 otherwise.

Naturally, we define the privacy risk of z as the probability that the adversary wins the game.

Definition 2 (Individual Privacy Risk). Given the learning algorithm A, the data distribution $\mathcal{D}$, $\mathcal{K}$ and the adversary $\mathcal{A}$, the privacy risk of z is defined as

$$\text{PR}(z, \mathcal{A}, A, n, \mathcal{D}, \mathcal{K}) = \Pr[\text{AG}(z, \mathcal{A}, A, n, \mathcal{D}, \mathcal{K}) = 1],$$

where the probability is taken over the randomness of S', the randomness of b, the randomness (if any) in the learning algorithm A, and the randomness (if any) of the adversary. Equivalently, the right-hand side, which is the probability that the adversary wins the game, can be expressed as the average of $\mathcal{A}$'s true positive rate and true negative rate as

$$\frac{\Pr[\mathcal{A} = 1 | b = 1] + \Pr[\mathcal{A} = 1 | b = 0]}{2}. \tag{2}$$

We emphasize that the privacy risk in our framework is different from the sort of membership privacy risk in prior works [17], in which the privacy risk is measured as the average privacy risk of points sampled randomly from the distribution. We analyze the privacy risk on individuals, as privacy concern is usually about each individual's data. The learning algorithms are still considered as leaking significant membership information if the privacy risk of a single data is high but low on average.

Privacy Risk of Subgroups. In addition to the individual privacy risk, we are also interested in the average privacy risk for individuals from a particular subgroup (e.g., black people who do not recommit crimes) and the gap of the information leakage between unconstrained models and fair models on each subgroup. Similar to the definition of the individual privacy risk, we formalize the privacy risk of the subgroup G_g^y in Attack Game 2. The only difference is that the challenge point is sampled from G_g^y by the challenger.

Attack Game 2 (Membership Inference Attack AG$(G_g^y, \mathcal{A}, A, n, \mathcal{D}, \mathcal{K})$). Let G_g^y determine the subgroup the adversary is interested in, $\mathcal{A}$ be an adversary, A be a learning algorithm, n be a positive integer, $\mathcal{D}$ be a distribution over data points (x, g, y) and $\mathcal{K}$ be the information about a model given to the adversary. The attack game proceeds as follows:

1) The challenger samples a challenge point $z \sim \mathcal{D}_g^y$, samples $S' \sim \mathcal{D}^{n-1}$ and chooses $b \to \{0,1\}$ uniformly at random.
2) If $b = 0$, the challenger draws $z' \sim \mathcal{D}$ and lets the training dataset $S = S' \cup \{z'\}$. If $b = 1$, the challenger lets the training dataset $S = S' \cup \{z\}$.
3) The challenger trains a model A_S by applying the learning algorithm A on S and sends $(z, \mathcal{K}(A_S), A)$ to the adversary.
4) The adversary outputs $\mathcal{A}(z, \mathcal{K}(A_S), A) \in \{0,1\}$.
5) The game outputs 1 if $\mathcal{A}(z, \mathcal{K}(A_S), A) = b$ and 0 otherwise.

Definition 3 (Subgroup Privacy Risk). Given the learning algorithm A, the data distribution $\mathcal{D}$, $\mathcal{K}$ and the adversary $\mathcal{A}$, the privacy risk of G_0^y is defined as

$$\text{PR}(G_0^y, \mathcal{A}, A, n, \mathcal{D}, \mathcal{K}) = \Pr[\text{AG}(G_0^y, \mathcal{A}, A, n, \mathcal{D}, \mathcal{K}) = 1],$$

where the probability is taken over the randomness of S', the randomness of b, the randomness (if any) in the learning algorithm A, the randomness (if any) of the adversary, and also the randomness of z.

The privacy risk of G_g^y is also the expectation of the privacy risk of data point z which belongs to G_g^y.

$$\text{PR}(G_0^y, \mathcal{A}, A, n, \mathcal{D}, \mathcal{K}) = \mathbb{E}_{z \sim \mathcal{D}_g^y}[\text{PR}(z, \mathcal{A}, A, n, \mathcal{D}, \mathcal{K})] \tag{3}$$

Based on the formalizations, we investigate individual privacy risk and subgroups privacy risk for standard learning algorithms (without fairness consideration) and the fair learning algorithms. More precisely, we are interested in the difference in $\text{PR}(z, \mathcal{A}, A, n, \mathcal{D}, \mathcal{K})$ and $\text{PR}(G_g^y, \mathcal{A}, A, n, \mathcal{D}, \mathcal{K})$ between the standard learning algorithm and fair machine learning algorithm.

4.2. Measuring Privacy Risk

This section presents the attack algorithm for measuring individual privacy risk and subgroup privacy risk for standard learning algorithms and fair learning algorithms.

We first need to specify the function $\mathcal{K}$. Recall that the adversary can only access the learned model via function $\mathcal{K}$. Specifically, we focus on the privacy risk of the data point through the output of the learned model, where $\mathcal{K}$ is the loss function of the learned model (i.e., $\mathcal{K}(A_S) = \ell(A_S, \cdot)$). This is the black-box setting as the learned model's parameters are unknown to the adversary [17, 19, 20].

Hence, the adversary needs to infer the membership of z given the loss of the trained model on the challenge point (i.e., $\ell(A_S, z)$). In the typical setting, the adversary pre-computes a loss threshold based on the knowledge about the population [17, 20] or shadow models [19] and outputs 1 ("member") if the loss is below the threshold or 0 ("non-member") if the loss is above the threshold [17, 19, 20]. Such adversary can be formalized as follows:

Adversary 1. Let ℓ_{pre} be an adversary's pre-computed loss threshold. On the input $z = (x, g, y), \mathcal{K}(A_s)$ and A, the membership adversary proceeds as follows:

1) Query the model to get $\ell(A_s, z)$.
2) Output 1 if $\ell(A_s, z) < \ell_{pre}$ and 0 otherwise.

However, there is an evident shortcoming of Adversary 1. In practice, the data can come from distinct distributions. For instance, in a binary classification task, it is impossible to distinguish positive samples and negative samples if the samples are from the same distribution. In addition, the model typically learns distinct patterns on the data from different distributions. Thus, the data from different distributions influence the learning procedure differently. As a consequence, a loss threshold, which results in the best attack accuracy for one data distribution (e.g., samples with positive labels), is usually a sub-optimal loss threshold for other distributions (e.g., samples with negative labels). Thus, we argue that, when the adversary knows which distribution the challenge data point belongs to, a better strategy is to use different loss thresholds for different distributions.

At the same time, the "unfairness" of the standard machine learning arises when the data from different groups are from different distributions. For example, in the loan approval application, qualified male applicants and qualified female applicants can come from slightly distinct distributions. Thus, in this case, the pattern learned on qualified male applicants can not be generalized to qualified female applicants and vice versa. Consequently, the unconstrained model may result in a much better accuracy on qualified male applicants than qualified female applicants and vice versa. This leads to the "unfairness" of the unconstrained models.

Based on this, we propose to use different loss thresholds for different subgroups, as described in Adversary 2. More specifically, the adversary pre-computes loss thresholds, one for each subgroup, and determines which one to use based on the protected attribute and the label of the challenge point. The adversary is formalized as follows:

Adversary 2. Let $\ell_{pre}^{(g,y)}$ be an adversary's pre-computed loss threshold for subgroup G_g^y. On the input $z = (x, g, y), \mathcal{K}(A_s)$ and A, the membership adversary proceeds as follows:

1) Query the model to get $\ell(A_s, z)$.
2) Output 1 if $\ell(A_s, z) < \ell_{pre}^{(g,y)}$ and 0 otherwise.

For each subgroup, we find a loss threshold that separates the members and non-members from the subgroup the best, and we measure the individual privacy risk and subgroup privacy risk based on Adversary 2.

In the following section, we present the experimental findings based on our framework. We show that our simple yet effective attack strategy would significantly improve the attack accuracy, which induces a better estimation of the privacy risk.

5. Experiments

In this section, we show our experimental findings on multiple datasets. We start by presenting our results on synthetic datasets, followed by our results on three real-world datasets.

Fair learning algorithm. We train fair models using the reduction approach proposed by [5], which incorporate specific definitions of fairness into existing machine learning methods. The reduction approach treats the underlying machine learning method as a "black box" while implementing a wrapper by reducing the constrained optimization problem to a sequence of cost-sensitive classification problems. It is important to note that the output of the reduction approach is a *randomized classifier*. To measure the classification performance, we use the expected accuracy, given by

$$\mathrm{Acc}(M_S, (x, g, y)) = 1 - |M_S(x) - y|, \qquad (4)$$

where $M_S(x)$ is the expected prediction of a randomized classifier M_S. For the unconstrained models, $M_S(x)$ is a deterministic prediction. We use the implementation provided in [5][1]. The underlying machine learning methods for fair models are the same as the machine learning methods for training unconstrained models.

Measurement. We repeat Attack Game 1 and Attack Game 2 to measure the individual privacy risk and subgroup privacy risk for different learning algorithms based on Adversary 2. The adversary can query the cross-entropy loss of a point on the unconstrained models and expected cross-entropy loss on fair models. We find a loss threshold for each subgroup that separates members and non-members from the subgroup the best in each attack game. We refer to the difference in the individual privacy risk and subgroup privacy risk between unconstrained models and fair models as *individual privacy cost* and *subgroup privacy cost*. A positive privacy cost implies that fair models leak information than unconstrained models.

We also compute the difference in the loss of a point between models trained on it and models not trained on it, referred to as *memorization*. If the model memorizes a point instead of learning a pattern to classify the point correctly, the memorization is large.

As a consequence, it is easier for the adversary to distinguish whether the data point is in the training dataset or not. Consequently, there is a higher privacy risk of the data point.

5.1. Experimental results on synthetic datasets

In this section, we present experimental results on synthetic datasets to show the effect of fairness constraints on the privacy risk to answer our questions in Section 4.

Dataset and models. We generate synthetic datasets with the same setting as in [41] in which the synthetic datasets are used to analyze the models' predictive behavior under the fairness constraint. Specifically, we draw 2,500 binary sensitive attributes from a Bernoulli distribution: $P_0 = \Pr[G = 0]$ and $P_1 = \Pr[G = 1]$ and draw binary labels per groups from a Bernoulli distribution: $P_g^- = \Pr[Y = -|G = g]$ and $P_g^+ = \Pr[Y = +|G = g]$.

1. https://github.com/fairlearn/fairlearn

TABLE 2: Accuracy and fairness gap of unconstrained models and fair models on the training dataset and the test dataset. The "Train Δ" and "Test Δ" columns show the results for the fairness gap on the training data and the test data which is defined in (1).

Model	Train acc	Test acc	Train Δ	Test Δ
Unconstrained	85.9%	85.6%	0.371	0.452
Fair($\delta = 0.001$)	83.0%	81.8%	0.001	0.265

Then we assign a 2-dimensional feature vector per subgroup by drawing samples from four different Gaussian distributions:

$$\begin{aligned}
&\text{For subgroup } G_0^- : \ X \sim \mathcal{N}([0,-1],[7,1;1,7]).\\
&\text{For subgroup } G_1^- : \ X \sim \mathcal{N}([-5,0],[5,1;1,5]).\\
&\text{For subgroup } G_0^+ : \ X \sim \mathcal{N}([1,2],[5,2;2,5]).\\
&\text{For subgroup } G_1^+ : \ X \sim \mathcal{N}([2,3],[10,1;1,4]). \qquad (5)
\end{aligned}$$

$\mathcal{N}(\mu, \Sigma)$ represents a Gaussian distribution with mean vector μ and covariance matrix Σ. We set $P_0 = 0.2$, $P_0^- = 0.1$ and $P_1^- = 0.5$. Accordingly, the P_1, P_0^+ and P_1^+ is 0.8, 0.9 and 0.5 respectively. The group G_0 has a smaller number of samples and has unbalanced positive samples and negative samples. The subgroup G_0^- is the smallest subgroup. The whole dataset is of size 2500, and we use 50% for the training dataset and 50% for the test dataset.

We train fully connected neural network models with hidden layer size {32, 16, 8} for unconstrained models and train fair models using the reduction approach. We use Adam optimizer and set the learning rate as 0.001.

The accuracy and fairness gap of unconstrained models and fair models are shown in Table 2. The test accuracy of unconstrained models is 39.4%, 84.6%, 84.6% 89.4% for the subgroup G_0^-, G_1^-, G_0^+ and G_1^+ respectively. The subgroup G_0^- is the *unprivileged subgroup* as it has much worse accuracy compared with the negative samples from G_1 (i.e., subgroup G_1^-). The fairness constraints are imposed to equalize the accuracy between G_0^- and G_1^-. We train fair models on the synthetic dataset with enforced fairness gap $\delta = 0.001$ (the upper bound of the fairness gap on the training dataset), which obtain average accuracy 50.3%, 76.8%, 85.2% and 86.9% of 20 runs for the subgroup G_0^-, G_1^-, G_0^+ and G_1^+ respectively. It shows that after imposing fairness constraints, the subgroup G_0^- *gains* a better accuracy.

Measuring privacy risk. To measure individual privacy cost, we compute each data point's privacy risk on 30 unconstrained models and 30 fair models. More precisely, we train 30 unconstrained models on 1250 data points chosen randomly among a synthetic dataset of size 2500. As a result, for each data point, on average, we have 15 models trained on it and 15 models not trained on it. On the same set of datasets, we train 30 fair models. We measure the individual privacy risk of a point as the average attack accuracy on the trained 30 models (for unconstrained models and fair models). We compute subgroup privacy risk as the average value of individual privacy risk from each subgroup.

Fairness comes at the cost of privacy. We measure the individual privacy cost and show the histograms for each subgroup in Figure 1. We observe that the subgroup G_0^-, unlike other subgroups, has a larger fraction of samples that have a positive privacy cost. The average value of the individual privacy cost is 6.9%, -1.47%, -2.04%, -1.97% for subgroup G_0^-, G_0^+, G_1^-, G_1^+ respectively. That is to say that the subgroup G_0^- has a higher privacy risk on fair models compared with unconstrained models. Fairness comes at a much higher privacy cost for subgroup G_0^-, compared with other subgroups.

Furthermore, we find the top 20 vulnerable points which have the highest privacy risk on fair models. In Figure 2, we show the privacy risk for these vulnerable data points before and after imposing fairness constraints. We notice that these points are mainly from the subgroup G_0^- and have a low privacy risk on unconstrained models. It indicates that fairness constraint increases the privacy risk for the data points from subgroup G_0^- significantly. As a result, these points could have a very high privacy risk on fair models. For instance, the fairness constraints increase the privacy risk of a data point from 63.3% to 93.3%.

Thus, imposing fairness constraints increases the subgroup privacy risk for subgroup G_0^- and also the individual privacy risk for points from G_0^- significantly. Hence, we conclude that fairness comes at the cost of privacy. In addition, the highest individual privacy risk on fair models is 93%, which is much larger than that on fair models (86.4%).

Fairness constraints increase the distinguishability between members and non-members. Figure 3 compares the loss distribution of the training data and the test data from subgroup G_0^- on a fair model and an unconstrained model. It is clear that the loss distributions on the training data and test data become more separable after imposing fairness constraints. The fair model achieves a low loss on the training data from G_0^- compared with the unconstrained model. As a result, the adversary achieves a lower false-negative rate on the fair model. The figure demonstrates that fairness constraint increases the separability of the loss distribution between training data and test data. Thus, the adversary achieves a better attack accuracy. In consequence, there is a higher privacy risk for subgroup G_0^-.

Fair models memorize the training data. Figure 4 (a) compares the average training accuracy of unconstrained models and fair models for all subgroups over 30 experiments (different random seeds). The unconstrained models have a low accuracy on subgroup G_0^- and are not fair with respect to equalized odds as the accuracy on subgroup G_1^- is much higher than that on G_0^-. After imposing fairness constraints, the training accuracy on the subgroup G_0^- increases from 50.1% to 78.8%.

However, as shown in Figure 4 (b), the memorization of fair models about the points from G_0^- is 0.58, which is two times larger than that of unconstrained models. It shows that imposing fairness constraints increases the training accuracy on the G_0^- but does not increase its test accuracy at the same magnitude. It means that fair models memorize the points from G_0^- instead of learning a general pattern on it. As a result, the output distributions are more distinguishable when a participant in or out of

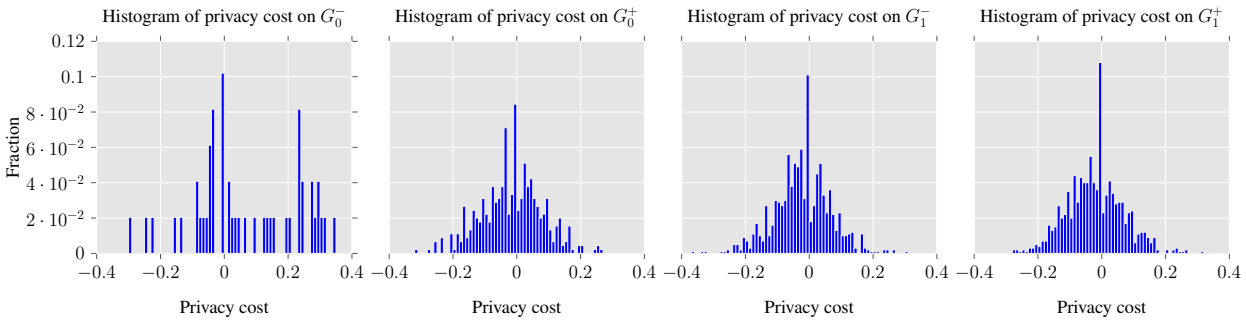

Figure 1: Histograms of individual privacy cost - Synthetic dataset. The x-axis is the privacy cost, which is the individual privacy risk difference on fair models and unconstrained models. A higher privacy cost of a data point reflects that the data point's privacy risk on fair models is much higher than that on constrained models.

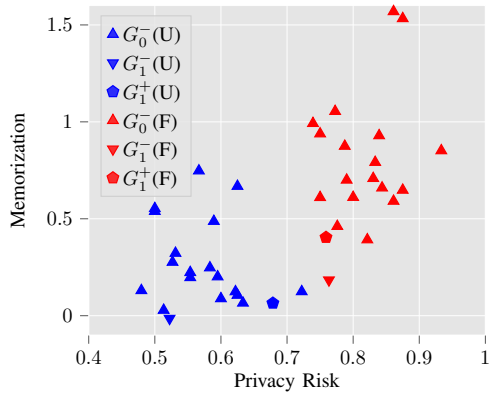

Figure 2: The most vulnerable points on fair models - Synthetic dataset. We find the top 20 vulnerable points that have the highest privacy risk on fair models. For each vulnerable point, we show its privacy risk and the memorization of models before (blue color) and after (red color) imposing fairness constraints. The shape of the marker shows which subgroup a point belongs to.

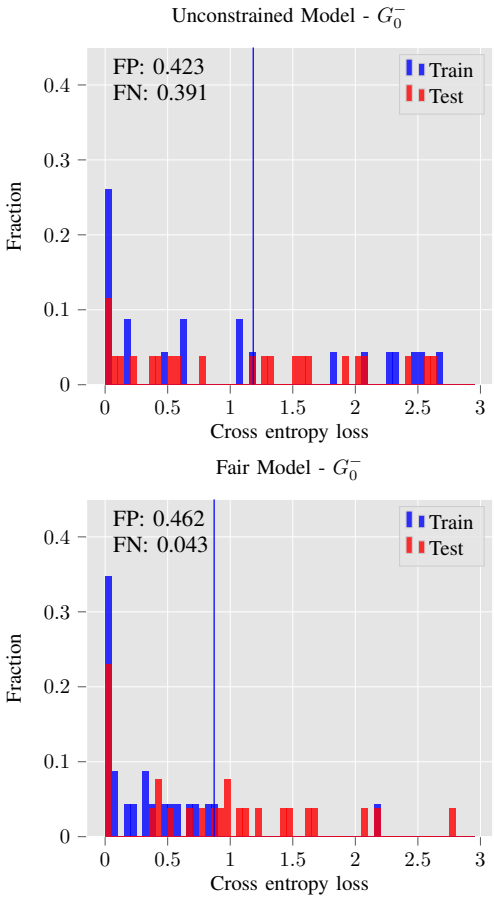

Figure 3: Loss distribution of an unconstrained model and a fair model on subgroup G_0^-. The line shows the loss threshold used by the adversary and FN and FP represent the corresponding false-negative rate and false-positive rate of the adversary.

the training datasets. Thus, the privacy risk of the points from G_0^- increases. This is also demonstrated in Figure 2 that the vulnerable points are mainly from G_0^- and fair models have larger memorizations on those points than unconstrained models.

On the contrary, fairness constraints only minorly change the privacy risk and the memorization for other subgroups. Accordingly, the privacy cost on other subgroups is low. This shows that fair constraints, in turn, incurs the "unfairness" of the model in terms of the privacy cost.

Fairness incurs a conflict between accuracy and privacy in the training dataset. In Figure 5, we show the effect of fairness constraints on the individual privacy cost and accuracy gain of the points from subgroup G_0^- in the training dataset. For each data point from subgroup G_0^- in the training dataset, we measure the training accuracy gain for each point induced by fairness constraints as the difference in the accuracy between fair models and unconstrained models. Thus, a positive accuracy gain implies that fairness constraints improve accuracy.

As we consider the privacy cost when the point is in the training dataset, we measure the true positive rate of the adversary on each point for fair models and unconstrained models and report the difference in it between fair models and unconstrained models. The true positive rate of the adversary reflects the probability of correctly predicting the membership of a point when the point is in the training dataset. A larger difference indicates that, with the same training dataset, fair models leak more information about the training dataset than unconstrained models. We report the average value of 30 runs.

We observe that fair models have better training accu-

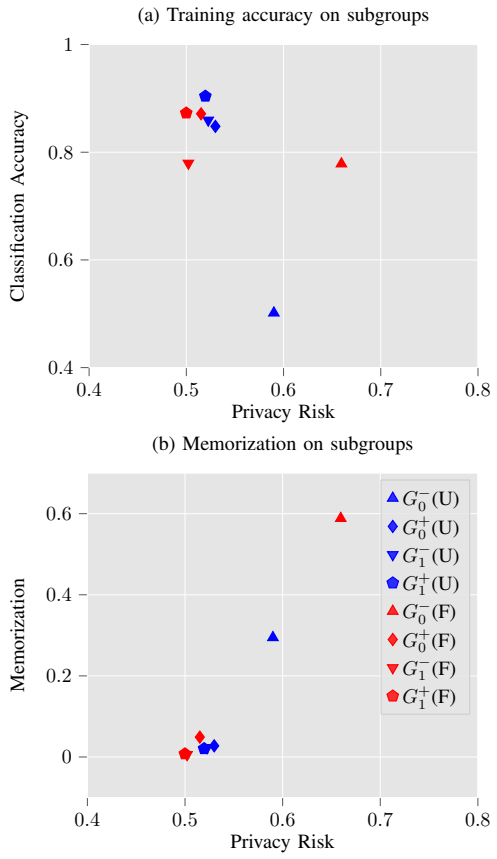

Figure 4: Memorization and training classification accuracy of fair models and unconstrained models on subgroups - Synthetic dataset. We use red color to show the results on the unconstrained models and blue for the fair models. The shape of the marker shows subgroup information.

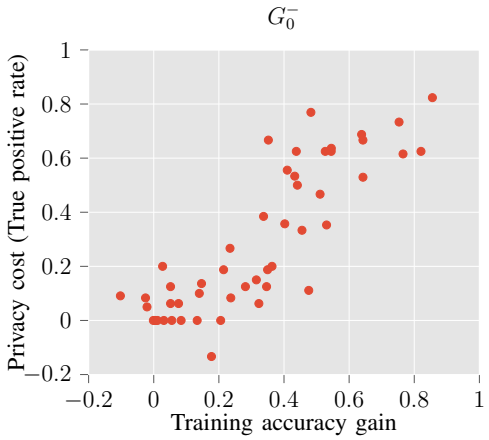

Figure 5: Accuracy gain and privacy cost of the points from subgroup G_0^- - Synthetic dataset. The x-axis is the training accuracy gain, which is the difference in the training accuracy between fair models and unconstrained models. A positive training accuracy gain implies that fair models fit better on the point than unconstrained models. The y-axis is the difference in the attack's true positive rate between fair models and unconstrained models, reflecting individuals' privacy cost in the training dataset. Each point in the plot represents a data point in the training dataset.

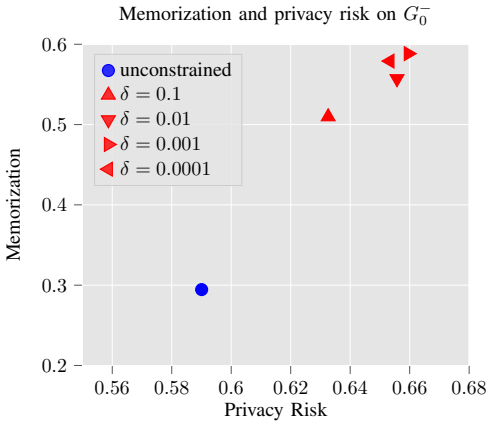

Figure 6: The effect of the enforced fairness gap δ on the privacy risk and memorization of the model - Synthetic dataset.

racy. However, a larger training accuracy gain of a point correlates with a larger increase in the adversary's true positive rate. Hence, there is a higher privacy cost for the point. In this setting, accuracy gain induced by fairness constraint hurts privacy.

Trade-offs between fairness and privacy. Figure 6 shows the effect of the enforced fairness gap δ on the privacy risk of fair models. Recall that the fairness gap of fair models on the training dataset is upper bounded by the enforced fairness gap δ. A decrease in the enforced fairness gap implies the fair models are less discriminatory on the training dataset. We observe that, when the fair model is less discriminatory on the training data, the memorization of fair models on subgroup G_0^- is larger, and the privacy risk is also larger. That is, a fairer model leaks more information about the unprivileged subgroup (G_0^- is our setting).

Additionally, we evaluate the correlation between the fairness gap on the unconstrained model and the subgroup privacy cost for G_0^-. We conduct experiments by varying the mean of the distribution for subgroup G_0^- (Change the mean for G_0^- in Equation 5) and show the corresponding training fairness gap of the unconstrained models $\Delta(S, M_S)$ and privacy cost of fair models in Figure 7 (a). For each experiment, we report the average of 20 runs.

Note that when the mean of the distribution for subgroup G_0^- varies, the separability between positive samples and negative samples changes. As a consequence, the difficulties of the classification task for subgroup G_0^- also differ. This leads to the accuracy changes of unconstrained models on G_0^-, which in turn influences the fairness gap. In this setting, a larger fairness gap means that subgroup G_0^- has a much worse accuracy on the unconstrained models compared with subgroup G_1^-. A clear observation is that the privacy cost increases when the fairness gap of the unconstrained model increases. It implies that, when the unconstrained model is more discriminatory, there is a higher privacy cost for the unprivileged subgroup. From a different point of view, it also shows that when data from the unprivileged subgroup is more complex (i.e., hard for unconstrained models to predict label correctly),

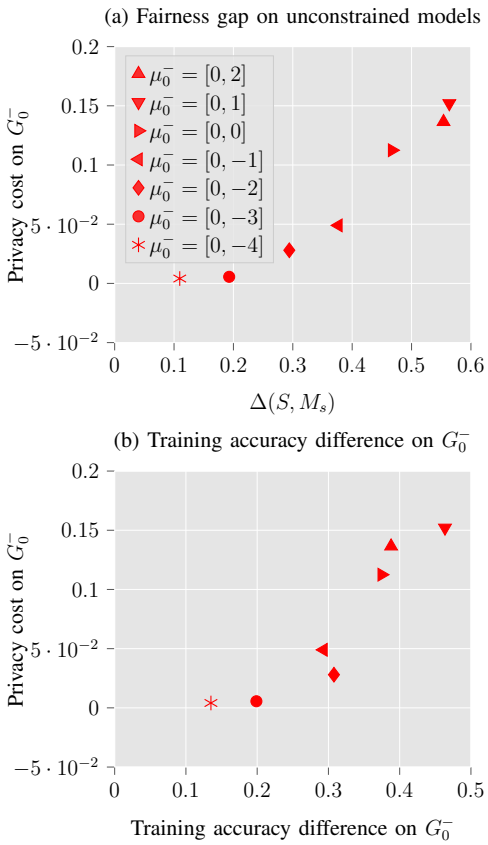

Figure 7: (a) The effect of the unconstrained model's fairness gap on the privacy cost of G_0^-. The x-axis is the fairness gap of unconstrained models on the training dataset. (b) The correlation between training accuracy gain and the privacy cost of G_0^-. The x-axis is the training accuracy difference between fair models and unconstrained models on subgroup G_0^-.

the privacy cost of the unprivileged subgroup is higher.

We also show the corresponding training accuracy gain on G_0^- in Figure 7 (b). We observe that when the fairness gap of the unconstrained model is larger, imposing fairness constraint will improve the training accuracy on the subgroup G_0^- more significantly. However, the privacy cost on G_0^- is also becoming higher when subgroup G_0^- has a larger accuracy gain from the fairness constraint (i.e., the increase of the training accuracy for the fair models is larger). It shows that the training accuracy gain for the unprivileged subgroup G_0^- comes at the cost of privacy.

A fewer number of samples implies a higher privacy risk. We evaluate the effect of the size of the subgroup on the privacy risk of fair models. In particular, we vary the fraction of samples from G_0 (i.e., $P_0 = \Pr[G = 0]$) and the fraction of samples from G_0^- (i.e., $P_0^- = \Pr[y = -|G = 0]$) and report the average subgroup privacy risk of subgroup G_0^- for the unconstrained models and fair models as shown in Figure 8 (a) and Figure 8 (b) respectively. For each experiment, we show the average of 20 runs. We observe that, when the size of the group G_0 becomes smaller, the privacy risk of fair models becomes much larger than that of unconstrained models. Other than that, when we fix the size of G_0 but decrease the number of samples with negative labels from the group G_0, the

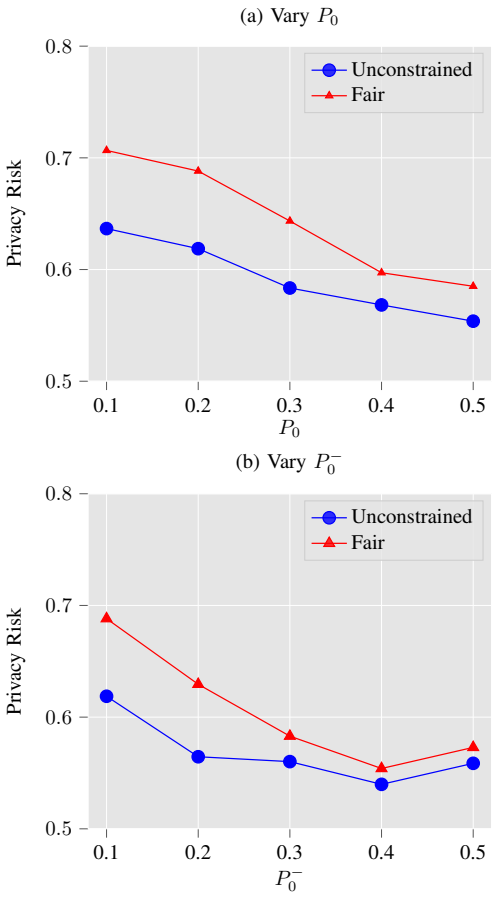

Figure 8: (a) The effect of the size of the group G_0 on the privacy risk of subgroup G_0^-. (b) The effect of the fraction of negative samples in group G_0 on the privacy risk of G_0^-.

TABLE 3: Accuracy of the attack with a fixed loss threshold and multiple thresholds - Synthetic dataset.

Model	Attack	G_0^-	G_1^-	G_0^+	G_1^+
Unconstrained	Single	52.9%	51.2%	51.8%	51.2%
	Multiple	61.8%	52.8%	52.4%	52.2%
Fair ($\delta = 0.001$)	Single	60.8%	51.9%	51.6%	50.8%
	Multiple	69.2%	53.4%	52.5%	51.6%

privacy cost also increases. Hence, we conclude that when the unprivileged subgroup has fewer samples, the privacy cost becomes higher.

Effectiveness of multiple loss threshold. As discussed in section 4.2, fixing a loss threshold to infer the membership of all the data points will result in a suboptimal attack. Table 3 shows the accuracy of the adversary who uses a single loss threshold and the adversary who uses one threshold for each subgroup. For each experiment, we report the average of 30 runs. We observe that using one threshold for one subgroup effectively increases the attack accuracy, which results in a better estimation of the privacy risk.

5.2. Experimental results on real datasets

In this section, we show the experimental results on the real datasets.

TABLE 4: The distributions of the real-world datasets.

Name	G_0^-	G_1^-	G_0^+	G_1^+
Bank	2.2%	85.2%	0.7%	12.0%
COMPAS (race)	28.3%	24.4%	31.7%	15.6%
COMPAS (gender)	12.8%	39.9%	7.3%	40.0%
Law (race)	2.3%	2.7%	13.5%	81.5%
Law (gender)	2.5%	2.5%	41.4%	53.6%

TABLE 5: Accuracy and fairness gap of unconstrained models and fair models with $\delta = 0.001$ on the training dataset and test dataset – Decision tree model with max depth 10.

Dataset	Model	Train acc	Test acc	Train Δ	Test Δ
Bank	Unconstrained	94.7%	89.2%	0.056	0.06
	Fair	94.5%	89.3%	0.001	0.048
COMPAS (race)	Unconstrained	78.9%	64.5%	0.128	0.178
	Fair	78.7%	63.8%	0.001	0.065
COMPAS (gender)	Unconstrained	78.7%	64.8%	0.095	0.107
	Fair	78.4%	64.2%	0.001	0.072
Law (race)	Unconstrained	97.5%	93.7%	0.267	0.178
	Fair	97.5%	93.5%	0.001	0.115
Law (gender)	Unconstrained	97.5%	93.7%	0.031	0.027
	Fair	97.6%	93.6%	0	0.029

Datasets and models. We conduct experiments on the Law School dataset (Law) [2] [27], Bank Marketing dataset (Bank) [28], and COMPAS datasets [29], which have 16,672, 24,391, and 4,302 examples, respectively. The number of features is 19, 58, 11 for Law, Bank, and COMPAS respectively. We use the same preprocessing as in IBM's AI Fairness 360 [42]. For COMPAS and Law School datasets, we consider two versions for each dataset, one where the protected attribute G is race and the other where G is gender. For the Bank Marketing dataset, the protected attribute G is age. The distributions of the data points for real-world datasets are presented in Table 4.

We train decision tree models, which are commonly used in the fairness literature, and train corresponding fair models using the reduction approach [5]. For each experiment, we report the average of 20 runs. The performance of the fair models and unconstrained models are shown in Table 5. We also use fully connected neural network models with hidden layer size {32, 16, 8} on COMPAS dataset, {1024, 512, 256, 128, 64} for Bank Marketing dataset and {1024, 512, 256, 128, 64} for Law School dataset. For each experiment, we report the average of 10 runs.

Privacy cost of fairness. Table 6 compares the privacy risk on fair models and unconstrained models for subgroups. We observe that, on the unconstrained models, the privacy risk varies across subgroups. This implies that the information leakage of the standard learning algorithm about each subgroup is different. After imposing fairness constraints, the gap of the privacy risk across subgroups becomes larger. For instance, after imposing fairness constraint, the gap of the subgroup privacy risk between subgroup G_1^- and G_0^+ increases from 13.1% to 18.6% on Bank dataset.

In addition, the subgroup privacy cost on Law (race) dataset is high. Especially, the privacy risk of subgroup G_1^- increases by 10% after imposing fairness constraints. We notice that the training accuracy of subgroup G_1^- is 43.4% on unconstrained models and 70.0% on fair models. That is to say, the G_1^- is the unprivileged subgroup and gains accuracy due to the fairness constraints. In addition, the fairness gap of the unconstrained model on the Law (race) dataset is 0.267, which is much larger than that on other datasets, as shown in Table 5. Furthermore, there is only 2.7% fraction of the data points belong to G_1^-. Hence, based on our analysis, the fair models leak more information about subgroup G_1^- as the unprivileged gains large accuracy due to fairness and also has a small number of samples.

Compared with COMPAS (race) dataset, the privacy cost of fairness on the COMPAS (gender) dataset is relatively small. This is because the unconstrained model's fairness gap is 0.095, which is much smaller than that on the COMPAS (race) dataset. Hence, imposing fairness constraints has less impact on the behavior of the model. As a consequence, the privacy cost is also small. In contrast, on the Law (gender) dataset, even the fairness gap of the unconstrained model is small (0.031), we observe a high subgroup privacy cost on subgroup G_1^- (6.4%). The reason is that the size of the subgroup G_1^- is small, as shown in Table 4. As a result, it is harder for fair models to learn a general pattern based on a small number of samples. Accordingly, the fair models memorize the training points from G_1^-. Hence, the privacy cost for G_1^- is high. In short, the privacy cost of fairness is higher for unprivileged subgroups when it has less number of samples or is more complex to classify correctly for the unconstrained model.

In addition, in Table 7, we show the subgroup privacy

TABLE 6: Privacy risk of fair unconstrained models and fair models with $\delta = 0.001$ on multiple datasets – Decision tree models with max depth 10.

Dataset	Model	G_0^-	G_1^-	G_0^+	G_1^+
Bank	Unconstrained	54.5%	51.6%	64.5%	61.1%
	Fair	57.4%	52.1%	70.7%	64.4%
COMPAS (race)	Unconstrained	58.2%	56.5%	57.9%	61.1%
	Fair	59.9%	58.9%	60.1%	64.8%
COMPAS (gender)	Unconstrained	58.3%	56.9%	64.3%	57.6%
	Fair	57.2%	59.1%	64.3%	59.9%
Law (race)	Unconstrained	72.6%	71.1%	54.1%	51.0%
	Fair	74.5%	81.8%	55.5%	51.5%
Law (gender)	Unconstrained	72.1%	72.4%	51.4%	51.4%
	Fair	77.4%	78.8%	52.1%	51.9%

TABLE 7: Privacy risk of unconstrained models and fair models with different enforced fairness gap δ on decision tree models with max depth 15 – Law (race) dataset.

Model	G_0^-	G_1^-	G_0^+	G_1^+
Unconstrained	72.6%	71.1%	54.1%	51.0%
Fair ($\delta = 0.1$)	74.4%	77.7%	55.0%	51.3%
Fair($\delta = 0.01$)	74.3%	81.0%	55.3%	51.4%
Fair($\delta = 0.001$)	74.5%	81.8%	55.5%	51.5%

2. Downloaded from https://github.com/jjgold012/lab-project-fairness (Bechavod and Ligett, 2017)

TABLE 8: Privacy risk of unconstrained models and fair models for decision tree models – Law (race) dataset. The "DT-5" row shows the results on decision tree models with max depth 5.

Model type	Model	G_0^-	G_1^-	G_0^+	G_1^+
DT-5	Unconstrained	58.5%	56.1%	51.5%	50.3%
	Fair	57.4%	58.3%	51.7%	50.4%
DT-10	Unconstrained	72.6%	71.1%	54.1%	51.0%
	Fair	74.5%	81.8%	55.5%	51.5%
DT -15	Unconstrained	81.5%	87.4%	55.7%	51.6%
	Fair	87.9%	95.5%	58.9%	52.7%

TABLE 9: Privacy risk of fair unconstrained models and fair models with $\delta = 0.001$ on multiple datasets – Neural Network models.

Dataset	Model	G_0^-	G_1^-	G_0^+	G_1^+
Bank	Unconstrained	52.3%	50.5%	54.8%	51.3%
	Fair	52.3%	50.4%	54.8%	51.3%
COMPAS (race)	Unconstrained	57.9%	56.8%	57.3%	60.8%
	Fair	58.6%	57.8%	58.2%	62.5%
COMPAS (gender)	Unconstrained	57.8%	57.5%	63.2%	56.8%
	Fair	57.5%	58.6%	64.1%	58.2%
Law (race)	Unconstrained	61.0%	58.9%	52.4%	50.7%
	Fair	60.4%	62.8%	52.5%	50.7%
Law (gender)	Unconstrained	61.3%	60.0%	51.1%	50.9%
	Fair	63.1%	64.2%	51.2%	51.1%

risk on unconstrained models and fair models with different enforced fairness gap (i.e., δ) on the Law (race) dataset. There is a clear increase in the privacy risk on subgroup G_0^+ when we tighten the fairness constraint. Thus, there is a trade-off between fairness and privacy.

Effect of model complexity. We show the effect of model's complexity on privacy cost in Table 8. When the model has lower capacity, namely the max depth of the decision tree model is low, the privacy cost is small. This is because that the fair models can not achieve satisfactory accuracy on all the subgroups, even on the training dataset. In other words, there is a high accuracy cost of fairness. For the decision tree models with max depth 5, after imposing fairness constraints, the test accuracy drops from 33.4% to 13.3% on subgroup G_0^- and only increases from 9.2% to 13.1% for G_1^-. The accuracy of the fair models on these two subgroups is even much worse than the accuracy of random guessing (uniformly predicting 1 or 0). On the contrary, for the decision tree models with max depth 15, the training accuracy drops from 70.1% to 54.5% for G_0^- and increases from 43.4% to 53.3% for G_1^- and the corresponding privacy cost is 6.4%. Therefore, when the model has a lower capacity, after imposing fairness constraints, the model can not fit the data well. In other words, when the accuracy cost of privacy is high, the privacy cost is relatively low.

Similar results can be observed in Table 9, where we use neural network models. We observe that there is a slight increase in the privacy risk of subgroups and we also notice that there is a large accuracy reduction after imposing fairness constraints. On Bank dataset, the test accuracy of unconstrained models is 87.1%, 91.8%, 51.1% and 47.8% for subgroup G_0^-, G_1^-, G_0^+, G_1^+ respectively. However, after imposing fairness constraint, the test accuracy is 99.5%, 99.5%, 5.4% and 6.5% for subgroup G_0^-, G_1^-, G_0^+, G_1^+ respectively. It implies that, after imposing fairness constraint, the model tends to predict "+" for all the data points. In other words, the fair models do not learn the general pattern for all the subgroups. On the contrary, for the decision tree models with max depth 10, the test accuracy drops no more than 5% for subgroups after imposing fairness constraints. This demonstrates that, if the models do not have enough capacity, adding fairness constraints will significantly reduce the learned model's predictive power. Under this circumstance, the privacy cost of fairness is relatively small.

6. Conclusions and Future Work

Fairness-aware learning imposes fairness constraints during the training to enforce the model to fit the data from the unprivileged subgroups. However, the fair models tend to memorize the training data from unprivileged subgroups instead of learning a general pattern, especially when the size of the unprivileged subgroups is small or data from unprivileged subgroups is complex. Consequently, it is easier to infer whether data was used to train a model or not by measuring the performance of the model on the data point. This memorization leads to a high privacy risk for the data of the unprivileged subgroup. Hence, for the unprivileged subgroup, fairness is achieved at the cost of privacy. In our work, we also propose a simple yet effective attack strategy. On the datasets used in fairness literature, we empirically demonstrate that our strategy is much effective for attacking fair models and unconstrained models. Our work did not investigate theoretical bounds on the information leakage of fair machine learning, which would remain a topic of future research.

References

[1] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias," *ProPublica*, May, 23, 2016.

[2] S. Barocas and A. D. Selbst, "Big data's disparate impact," *Calif. L. Rev.*, vol. 104, p. 671, 2016.

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

[4] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," in *Advances in neural information processing systems*, 2016, pp. 4349–4357.

[5] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach, "A reductions approach to fair classification," *arXiv preprint arXiv:1803.02453*, 2018.

[6] T. Calders, F. Kamiran, and M. Pechenizkiy, "Building classifiers with independency constraints," in *2009 IEEE International Conference on Data Mining Workshops*. IEEE, 2009, pp. 13–18.

[7] 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*, 2012, pp. 214–226.

[8] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in neural information processing systems*, 2016, pp. 3315–3323.

[9] T. Kamishima, S. Akaho, and J. Sakuma, "Fairness-aware learning through regularization approach," in *2011 IEEE 11th International Conference on Data Mining Workshops*. IEEE, 2011, pp. 643–650.

[10] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," in *Advances in neural information processing systems*, 2017, pp. 4066–4076.

[11] D. Madras, E. Creager, T. Pitassi, and R. Zemel, "Learning adversarially fair and transferable representations," *arXiv preprint arXiv:1802.06309*, 2018.

[12] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi, "Fairness constraints: Mechanisms for fair classification," *arXiv preprint arXiv:1507.05259*, 2015.

[13] M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi, "Fairness constraints: Mechanisms for fair classification," in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 962–970.

[14] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *International Conference on Machine Learning*, 2013, pp. 325–333.

[15] C. Song and A. Raghunathan, "Information leakage in embedding models," *arXiv preprint arXiv:2004.00053*, 2020.

[16] R. Shokri, M. Strobel, and Y. Zick, "Privacy risks of explaining machine learning models," *arXiv preprint arXiv:1907.00164*, 2019.

[17] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, "Privacy risk in machine learning: Analyzing the connection to overfitting," in *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*. IEEE, 2018, pp. 268–282.

[18] L. Song, R. Shokri, and P. Mittal, "Privacy risks of securing machine learning models against adversarial examples," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 2019, pp. 241–257.

[19] 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)*. IEEE, 2017, pp. 3–18.

[20] A. Sablayrolles, M. Douze, C. Schmid, Y. Ollivier, and H. Jégou, "White-box vs black-box: Bayes optimal strategies for membership inference," in *International Conference on Machine Learning*, 2019, pp. 5558–5567.

[21] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning," in *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2019, pp. 739–753.

[22] J. Ding, X. Zhang, X. Li, J. Wang, R. Yu, and M. Pan, "Differentially private and fair classification via calibrated functional mechanism," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 01, 2020, pp. 622–629.

[23] D. Xu, S. Yuan, and X. Wu, "Achieving differential privacy and fairness in logistic regression," in *Companion Proceedings of The 2019 World Wide Web Conference*, 2019, pp. 594–599.

[24] 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*, 2019, pp. 309–315.

[25] 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*, 2016, pp. 308–318.

[26] E. Bagdasaryan, O. Poursaeed, and V. Shmatikov, "Differential privacy has disparate impact on model accuracy," in *Advances in Neural Information Processing Systems*, 2019, pp. 15 479–15 488.

[27] L. F. Wightman and H. Ramsey, "Law school admission council." 1998.

[28] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[29] J. Larson, S. Mattu, L. Kirchner, and J. Angwin, "COMPAS dataset," https://github.com/propublica/compas-analysis, 2017, [COMPAS dataset (2017)].

[30] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 335–340.

[31] A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz, and M. Backes, "Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models," *arXiv preprint arXiv:1806.01246*, 2018.

[32] X. Pan, W. Wang, X. Zhang, B. Li, J. Yi, and D. Song, "How you act tells a lot: Privacy-leaking attack on deep reinforcement learning," in *Proceedings*

of the 18th International Conference on Autonomous Agents and MultiAgent Systems, 2019, pp. 368–376.

[33] M. Yaghini, B. Kulynych, and C. Troncoso, "Disparate vulnerability: On the unfairness of privacy attacks against machine learning," *arXiv preprint arXiv:1906.00389*, 2019.

[34] Y. Long, L. Wang, D. Bu, V. Bindschaedler, X. Wang, H. Tang, C. A. Gunter, and K. Chen, "A pragmatic approach to membership inferences on machine learning models."

[35] M. D. Ekstrand, R. Joshaghani, and H. Mehrpouyan, "Privacy for all: Ensuring fair and equitable privacy protections," in *Conference on Fairness, Accountability and Transparency*, 2018, pp. 35–47.

[36] S. Kuppam, R. McKenna, D. Pujol, M. Hay, A. Machanavajjhala, and G. Miklau, "Fair decision making using privacy-protected data," *arXiv preprint arXiv:1905.12744*, 2019.

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

[38] M. Jagielski, M. Kearns, J. Mao, A. Oprea, A. Roth, S. Sharifi-Malvajerdi, and J. Ullman, "Differentially private fair learning," in *International Conference on Machine Learning*. PMLR, 2019, pp. 3000–3008.

[39] M. Donini, L. Oneto, S. Ben-David, J. S. Shawe-Taylor, and M. Pontil, "Empirical risk minimization under fairness constraints," in *Advances in Neural Information Processing Systems*, 2018, pp. 2791–2801.

[40] I. Dinur and K. Nissim, "Revealing information while preserving privacy," in *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 2003, pp. 202–210.

[41] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi, "Fairness constraints: A flexible approach for fair classification." *J. Mach. Learn. Res.*, vol. 20, no. 75, pp. 1–42, 2019.

[42] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic *et al.*, "Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias," *arXiv preprint arXiv:1810.01943*, 2018.