Bogdan Kulynych, Mohammad Yaghini, Giovanni Cherubin, Michael Veale, and Carmela Troncoso

Disparate Vulnerability to Membership Inference Attacks

Abstract: A membership inference attack (MIA) against a machine-learning model enables an attacker to determine whether a given data record was part of the model's training data or not. In this paper, we provide an in-depth study of the phenomenon of *disparate vulnerability* against MIAs: unequal success rate of MIAs against different population subgroups. We first establish necessary and sufficient conditions for MIAs to be prevented, both on average and for population subgroups, using a notion of distributional generalization. Second, we derive connections of disparate vulnerability to algorithmic fairness and to differential privacy. We show that fairness can only prevent disparate vulnerability against limited classes of adversaries. Differential privacy bounds disparate vulnerability but can significantly reduce the accuracy of the model. We show that estimating disparate vulnerability by naïvely applying existing attacks can lead to overestimation. We then establish which attacks are suitable for estimating disparate vulnerability, and provide a statistical framework for doing so reliably. We conduct experiments on synthetic and real-world data finding significant evidence of disparate vulnerability in realistic settings.

Keywords: membership inference attacks, machine learning, fairness

DOI 10.2478/popets-2022-0023
Received 2021-05-31; revised 2021-09-15; accepted 2021-09-16.

1 Introduction

Membership Inference Attacks (MIAs), in which an adversary aims to determine whether an example is part of the training set, are one of the main privacy attacks against machine-learning (ML) models. Since they were first described [39], many works have studied the potential of these attacks under diverse circumstances [22, 24, 29, 30, 32, 33]; and the causes and limits of these attacks [16, 26, 43]. In both empirical and theoretical approaches researchers focus on the *average* MIA success across the records. However, there is empirical evidence that the vulnerability to MIAs is not always evenly distributed: it can differ across target classes [39], it can be more effective against some individuals [29], and it can vary across subgroups [6]. These results imply that average-based studies can overestimate the privacy for some individuals [15].

In this paper, we provide the first theoretical analysis of the *disparate vulnerability* to MIA across populations subgroups. Our contributions are the following:

✓ We introduce a novel characterization of the vulnerability to MIAs, which provides a *necessary and sufficient* condition for these attacks to succeed: lack of *distributional generalization*. Vulnerability to MIA arises when the *distribution* of a model's property (e.g., loss, or outputs) is different for samples in and out of the training dataset. This result complements previous studies that demonstrated the lack of standard generalization (i.e., overfitting) to be a sufficient but not necessary condition for vulnerability to MIAs [29, 43].

✓ We introduce the first formal analysis of disparate vulnerability and extend our results on necessary and sufficient conditions for preventing MIAs to subgroup vulnerability and disparate vulnerability.

✓ We show that estimating the magnitude of the disparate vulnerability is non-trivial when subgroups are small. We provide a statistical framework and methods to estimate disparate vulnerability and its significance. We show that not all vulnerability estimation mechanisms used in prior work are adequate for subgroups. We discuss the implications of these difficulties for regulation compliance.

✓ We prove that satisfying algorithmic-fairness constraints can decrease disparate vulnerability to limited classes of attackers. We also show that training with differential privacy bounds the magnitude of the disparate vulnerability.

✓ We empirically evaluate disparate vulnerability both on synthetic and on real-world datasets,

Bogdan Kulynych, Carmela Troncoso: EPFL
Mohammad Yaghini: University of Toronto, Vector Institute
Giovanni Cherubin: Alan Turing Institute
Michael Veale: University College London

demonstrating that disparate vulnerability exists in realistic models, with high statistical significance.

✓ We discuss the importance of disaggregating privacy measurements when evaluating the legal implications of privacy attacks. In particular, the importance of studying the consequences of privacy attacks for subgroups when analyzing the privacy risks of a deployment, as opposed to studying individual privacy risks [29] that can be dismissed as residual and acceptable.

2 Related Work

Theory studies on MIA. Yeom et al. studied the relation of MIAs to overfitting [43]; in their work, they formalize MIA as an indistinguishability game, which we adapt to construct our theoretical framework. Farokhi et al. analyzed the dependence of MIA's success on the amount of information the model memorizes [16], and Jayaraman et al. investigated their dependence on the prior probability that the example given to the adversary is a member or non-member of the training set [22]. Yeom et al. [43], and Cherubin et al. [9] showed that MIAs success is bounded by DP. Humphries et al. [21] showed these bounds only apply so long as the training data are i.i.d.-sampled. All these analyses, however, are only meaningful for the *average*-case MIA. A classifier thought to be secure according to these analyses may provide weaker protection to certain individuals or subpopulations. Our work complements these studies and generalizes the notion of MIA risk to *subgroups* of the population, enabling study of vulnerability for subsets of the records' labels, individuals, and subpopulations.

Disparity and machine learning. The work on disparity in machine learning is centered on understanding and mitigating disparate impact of algorithmic decisions on subpopulations [2, 10, 28]. Bagdasaryan et al. [1] and Pujol et al. [35] study disparity in accuracy under differential privacy (DP), and show that training with DP can increase disparate impact. In this work, we develop a theory that supports the empirical evidence that disparate impact would also cause disparity in vulnerability to MIAs [6, 29, 39].

3 Membership Inference Attacks

Let Ω be a *population* of examples, where each example represents an individual: $x \in \Omega$. We assume that the population is partitioned in disjoint *subgroups*. Each subgroup $G_z \subset \Omega$ is formed by examples that share one or several attributes (e.g., race or gender in the way they are commonly represented in data), such that $\bigcup_{z=1}^{t} G_z = \Omega$. We consider a *data-generating distribution* $\mathcal{D}$ over Ω.

We indicate with $A(\cdot)$ the training algorithm that produces a model A_S given training data $S \subset \Omega$. The learning task for this model is to infer the value of the *label* $y = y(x)$ associated with an individual x. We assume that the model can be either a regressor (y takes values in a set with total order, e.g. $\mathbb{R}$) or a classifier (y takes values in a finite set).

The goal of a membership inference attack (MIA) is to predict whether an example $x \in \Omega$ is a *member* or a *non-member* of the training set S. We assume a threat model where a MIA adversary observes the target model's behavior that relates to x, and has information about the data distribution $\mathcal{D}$, training-data sampling, and the training algorithm. We formalize MIAs using the indistinguishability game by Yeom et al. [43]:

```
MIA($\mathcal{A}, A, n, \mathcal{D}$)
1 : $S \leftarrow \mathcal{D}^n; A_S = A(S)$
2 : $m \xleftarrow{\$} \{0, 1\}$
3 : if $m = 1$ then
4 :     $x \xleftarrow{\$} S$
5 : else
6 :     $x \leftarrow \mathcal{D}$
7 : endif
8 : $\hat{m} \leftarrow \mathcal{A}(x, A_S, n, \mathcal{D})$
9 : return $m = \hat{m}$
```

In this game, the challenger samples S from the population, and trains a model A_S using training algorithm A (line 1). The challenger then randomly draws a secret m (line 2) whose value denotes x's *m*embership in S: $m = 1$ if the *challenge example* x is sampled from the training set S (line 4), and $m = 0$ if it is sampled from the data distribution $\mathcal{D}$ (line 6). As Yeom et al. [43], we assume that the population is large enough that the chance of sampling a member $x \in S$ from $\mathcal{D}$ is negligible. Given the challenge example x, the target model A_S

and its training algorithm $A(\cdot)$, the sampling parameter n, and the distribution of the training data $\mathcal{D}$, the MIA adversary $\mathcal{A}(\cdot)$ makes a guess $\hat{m}$ about the example's membership in S (line 8). We use this formalization as it is the most common, although there are other ways to formalize MIAs [21].

The MIA game defines a joint probability distribution over training datasets S, membership "coins" m, and challenge examples x. We denote by M the random variable taking the value of the membership coin (line 2), by X the challenge example, by $Y = y(X)$ the label associated with the challenge example x, by Z the subgroup of the population z to which the x belongs, and by $\hat{Y} = A_S(X)$ the output the model A_S at x.

3.1 Attack Strategy

As described in the MIA game, the adversary's knowledge is limited to $(x, A_S, n, \mathcal{D})$, and their goal is to guess the membership of x. For brevity, we use A_S to indicate both the access to trained models A_S and their training algorithm $A(\cdot)$.

We define a general strategy to perform a membership attack that encompasses several instances of MIA, e.g., [32, 39, 43]. This strategy consists of two phases.

First, the adversary prepares an attack algorithm $\texttt{Att}_{A,n,\mathcal{D}}(\cdot)$ which depends on the target training algorithm $A(\cdot)$, and data-sampling parameters n and $\mathcal{D}$, e.g., by training a shadow-model attack classifier [39]. We drop the subscripts in $\texttt{Att}_{A,n,\mathcal{D}}$ where the setting is clear from the context.

In the second phase, the adversary extracts *features*, $w \leftarrow \phi(A_S, x)$, describing the target model and the example, and applies the attack algorithm to the extracted features to obtain the membership guess, $\hat{m} \leftarrow \texttt{Att}_{A,n,\mathcal{D}}(w)$. Thus, the guess $\hat{m}$ is obtained by applying the attack algorithm to the extracted features:

$$\mathcal{A}(x, A_S, n, \mathcal{D}) \triangleq \texttt{Att}_{A,n,\mathcal{D}} \circ \phi(A_S, x)$$

This formalization is flexible: it captures both white-box and black-box adversarial models. For example, the features could be the outputs of the model and the example's label $w = (A_S(x), y(x))$ [39], the model's loss for the challenge example, $w = \ell(A_S(x), y(x))$ [43], or the model's gradients as in some white-box attacks [32].

We use random variable W to indicate the extracted features w across instances of the MIA game. For example, if the attacker uses the model's output and the label as features [39], we denote them as $W = (\hat{Y}, Y)$. With a slight abuse of notation, we use $\phi_W : (A_S, x) \mapsto w$ to indicate the procedure that extracts features w that are realizations of the W random variable. Furthermore, we denote by $\mathcal{A}_W$ an adversary that uses features W.

We distinguish two kinds of adversaries depending on the features they use: *regular* adversaries that do not use subgroup information ($Z \notin W$), and *subgroup-aware* adversaries that do use this information ($Z \in W$). We assume that the latter adversary can obtain the subgroup z from the examples x themselves, encoded in an example (e.g., gender, race). That is the case for our experiments on real-world data in Section 7. However, in practical scenarios, this knowledge could be encoded in the label $y(x)$, or come from external sources. Prior work has mainly considered regular adversaries.

3.2 Vulnerability

We introduce the concept of *vulnerability* of an ML model to membership inference attacks (MIAs). Vulnerability measures the success of an adversary against the model. We also introduce worst-case (Bayes) vulnerability, i.e., vulnerability against an information-theoretically optimal adversary.

Vulnerability to MIAs is the normalized advantage [43] of adversary $\mathcal{A}$ over random guessing:

Definition 1. We define *vulnerability* to adversary $\mathcal{A}$ as:

$$V(\mathcal{A}) \triangleq 2\Pr[\mathsf{MIA}(\mathcal{A}, A, n, \mathcal{D}) = 1] - 1 \qquad (1)$$

We also extend the definition to subgroups:

Definition 2. Let z be a subgroup of the population. We define *subgroup vulnerability* to adversary $\mathcal{A}$ as:

$$V_z(\mathcal{A}) \triangleq 2\Pr[\mathsf{MIA}(\mathcal{A}, A, n, \mathcal{D}) = 1 \mid Z = z] - 1.$$

which captures the normalized advantage of a MIA adversary $\mathcal{A}$ for challenge examples coming from a given subgroup z.

Optimal adversaries. We base our analysis on information-theoretically optimal adversaries. The worst-case vulnerability to any adversary that leverages features W is:

$$\max_{\texttt{Att}_W : \mathbb{W} \mapsto \{0,1\}} V(\texttt{Att}_W \circ \phi_W), \qquad (2)$$

where $\mathbb{W}$ is the domain of W. The maximum is achieved by a *Bayes adversary* which uses the following strategy for the attack [9, 36]:

$$\texttt{Att}^*_W(w) \triangleq \arg\max_{m \in \{0,1\}} \Pr[M = m \mid W = w], \qquad (3)$$

We denote the Bayes adversary as $\mathcal{A}^*_W \triangleq \mathtt{Att}^*_W \circ \phi_W$, and drop the subscripts where no ambiguity arises.

Subgroup-aware Bayes adversary. We assume the adversary can know the subgroup z to which each example x belongs. Recall that we refer to this adversary as subgroup-aware. As the vulnerability to the Bayes adversary grows if the adversary has more information about the examples, the worst-case vulnerability to a subgroup-aware adversary is equal or higher compared to a regular adversary:

Proposition 1. $V(\mathcal{A}^*_{W,Z}) \geq V(\mathcal{A}^*_W)$.

We defer the proof to Appendix A.

In our experimental evaluations, we only consider subgroup-aware adversaries as they are guaranteed to attain higher advantage in the worst case.

4 Distributional Generalization and Vulnerability to MIAs

An ML model is said to *overfit*, or poorly *generalize*, when its average loss on the training set differs from its loss on new samples from the population. Previous work showed that, while overfitting is an important factor for evaluating MIA [39], it is not necessary for MIA vulnerability [29, 43].

Fig. 1 illustrates with an example why the absence of standard overfitting does not, in general, prevent MIAs. The figure shows a model's loss values on its training and test data. The standard, average-based definition of overfitting cannot distinguish between the two distributions; but an adversary potentially can, and the model can be vulnerable to MIAs.

4.1 Distributional Generalization

To establish the necessary and sufficient conditions for models to be vulnerable to MIAs, we introduce an extended notion of generalization that goes beyond comparing the average loss on train and test data. It covers the difference in the distributions of any given property of a model on the training data and outside. A *property* is any function that takes as input a model and an example: $\pi(A_S, x)$, and returns a numeric vector. A property function can be, for instance, a loss function, the gradient, or the prediction from the model.

We are interested in the distributions of properties on the examples x coming from the training dataset and from outside of the training dataset. For any set T from the range of π, we define the corresponding probability measures as:

$$\mu^\pi_1(T) \triangleq \Pr_{\substack{S\sim\mathcal{D}^n\\ x\sim S}}[\pi(A_S, x) \in T],$$

$$\mu^\pi_0(T) \triangleq \Pr_{\substack{S\sim\mathcal{D}^n\\ x\sim \mathcal{D}}}[\pi(A_S, x) \in T].$$

Definition 3. For any property function $\pi(A_S, x)$, we define *the distributional-generalization gap* as follows:

$$R(\pi, d) \triangleq d\big(\mu^\pi_1,\ \mu^\pi_0\big),$$

where $d(\mu, \mu')$ is a measure of dissimilarity between probability distributions.

This generic notion subsumes the standard notion of generalization. Standard generalization can be measured using the average-dataset generalization gap (see, e.g., in Yeom et al. [43]), the difference between the expected loss on the training dataset and the expected loss on the distribution:

$$R \triangleq \mathop{\mathbb{E}}_{\substack{S\sim\mathcal{D}^n\\ x\sim S}}[\ell(A_S, x)] - \mathop{\mathbb{E}}_{\substack{S\sim\mathcal{D}^n\\ x\sim \mathcal{D}}}[\ell(A_S, x)],$$

where $\ell(A_S, x)$ is a loss function. We can recover this standard notion of a generalization gap as $R(\ell, d_{\mathrm{MD}})$, using the loss function as the property function and the *mean discrepancy* $d_{\mathrm{MD}}(\mu, \mu')$ as a dissimilarity measure:

$$d_{\mathrm{MD}}(\mu, \mu') \triangleq \int \omega \,\mathrm{d}\mu(\omega) - \int \omega \,\mathrm{d}\mu'(\omega),$$

Whereas standard generalization quantifies how much the training algorithm tends to memorize the training dataset through the lens of its performance (loss), distributional generalization can do so (1) through the lens of other properties beyond losses, and (2) considering distributional information instead of only the difference between the means.

Evaluating distributional generalization enables us to assess the generalization of an ML model on the entire population, rather than on average. In Fig. 1 it is clear that the model's actual loss across the entire population is concentrated on a few individuals. Distributional generalization enables us to capture this discrepancy, whereas standard generalization does not.

Concurrently, Nakkiran and Bansal [31] have also proposed a similar notion of distributional generalization. Our proposal allows for more general distances between distributions, whereas Nakkiran and Bansal, when translated to our terms, define the gap using mean discrepancy, which is not sufficient for our analysis.

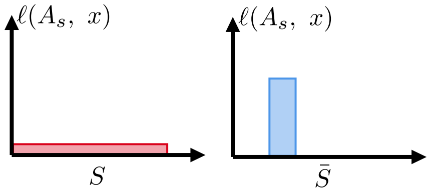

Fig. 1. Loss values of a model A_S on train data S (left) and test data $\bar{S}$ (right). According to standard notion of generalization, this model does not overfit: average loss (area) on training and test data is identical. Some population individuals, however, are more penalized on the test data. This discrepancy is captured by *distributional* generalization.

4.2 Relation Between Worst-case Vulnerability and Distributional Generalization

The ability of any classifier to successfully distinguish between observations of two classes can be characterized by the total variation between the class-conditional distributions of observations. By applying this fact to the worst-case MIA attackers, we can characterize vulnerability in terms of distributional generalization:

Proposition 2. *The worst-case vulnerability to MIAs with adversary's features W is equal to the distributional-generalization gap under total-variation distance:*

$$V(\mathcal{A}^*_W) = R(\phi_W, d_{\mathrm{TV}}),$$

where the total-variation distance is defined as:

$$d_{\mathrm{TV}}(\mu, \mu') \triangleq \sup_{T \subseteq \mathbb{W}} |\mu(T) - \mu'(T)|$$

According to Proposition 2, when the property function π is the adversary's feature extraction mechanism ϕ_W, the distributional-generalization gap is equal to the worst-case vulnerability to adversaries that use features $W = \phi_W(A_S, X)$.

Proof. Let us define the *Bayes error* L^*, the 0-1 classification error of the Bayes classifier. In the case of $\mathtt{Att}^*$:

$$L^* \triangleq \Pr[\mathtt{Att}^*(W) \neq M]$$

Recall that vulnerability is defined through the success probability of an adversary:

$$V(\mathcal{A}_W) \triangleq 2\Pr[\mathtt{Att}(W) = M] - 1$$

Thus, for a Bayes adversary, $V(\mathcal{A}^*_W)$ uses the complement of the Bayes error L^*:

$$V(\mathcal{A}^*_W) = 2(1 - \Pr[\mathtt{Att}^*(W) \neq M]) - 1 = 1 - 2L^*.$$

It is well-known that the the Bayes error of the binary classifier under uniform prior is equal to:

$$\begin{aligned} L^* &= \frac{1}{2} - \frac{1}{2} d_{\mathrm{TV}}\left(\Pr[W \mid M = 1],\ \Pr[W \mid M = 0]\right) \\ &= \frac{1}{2} - \frac{1}{2} d_{\mathrm{TV}}\left(\Pr_{\substack{S \sim \mathcal{D}^n \\ x \sim S}}[\phi_W(A_S, x)],\ \Pr_{\substack{S \sim \mathcal{D}^n \\ x \sim \mathcal{D}}}[\phi_W(A_S, x)]\right) \\ &= \frac{1}{2} - \frac{1}{2} d_{\mathrm{TV}}\left(\mu_1^{\phi_W},\ \mu_0^{\phi_W}\right), \end{aligned}$$

See, e.g., Devroye et al. [12, Chapter 3.9]. This implies the sought form.

□

This form is a straightforward consequence of our Bayes-optimal approach to vulnerability and is an application of a well-known result in statistical theory. It provides us with an intuitive interpretation of the worst-case vulnerability to MIAs—as it is equal to the distributional-generalization gap—thus with a guideline on how to prevent MIAs. The result holds for both white-box and black-box adversary models.

Let us visually illustrate distributional generalization and worst-case vulnerability. Consider adversarial features $W = \hat{Y}$. For the continuous property function $\phi_{\hat{Y}}$, the distributional-generalization gap becomes:

$$\begin{aligned} R(\phi_{\hat{Y}}, d_{\mathrm{TV}}) &= d_{\mathrm{TV}}\left(\mu_1^{\phi_{\hat{Y}}}, \mu_0^{\phi_{\hat{Y}}}\right) \\ &= \frac{1}{2} \int \left| f_1(\hat{y}) - f_0(\hat{y}) \right| \mathrm{d}\hat{y}, \end{aligned}$$

where f_1 and f_0 are probability density functions associated with measures μ_1 and μ_0, respectively. See Fig. 2 for a visualization. The worst-case vulnerability to adversaries using features $W = \hat{Y}$ is the area between the densities of the "in" and "out" output distributions.

Note that the distance used in Proposition 2 is *average-dataset*. That is, when computing the features $\phi(A_S, X)$, the model A_S is a random variable over the randomness of $A(\cdot)$ and $S \sim \mathcal{D}^n$. To train models with minimal vulnerability to MIAs, Li et al. [27] used a similar yet different notion of distance, the distance between outputs of a *fixed* model on its training dataset and a validation dataset. Although conceptually similar, such distance cannot be directly used to evaluate the worst-case vulnerability using Proposition 2.

Overfitting and worst-case vulnerability. The absence of overfitting in the standard sense does not necessarily preclude MIAs [29, 43]. But, a straightforward implication of Proposition 2 shows there is a case when the standard generalization gap does bound the worst-case vulnerability:

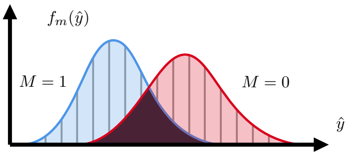

Fig. 2. Distributional-generalization gap for models' outputs $\hat{y}$. The curves represent the probability density functions of models' outputs on the training datasets ($M = 1$) and outside ($M = 0$). The striped area shows the distributional-generalization gap: total variation between distributions of model's outputs on training and outside. Proposition 2 shows that the the size of the striped area exactly equals to the worst-case vulnerability to any adversary that uses model outputs $\hat{y}$ as features for distinguishing members from non-members.

Corollary 1. *Let $\ell(A_S, x) = \mathbb{1}[A_S(x) \neq y(x)]$ be the 0-1 loss, and the adversary's features be the loss values $W = \ell(A_S, X)$. Then, the standard generalization gap equals worst-case vulnerability:*

$$V(\mathcal{A}^*_{\ell(A_S,X)}) = |R(\ell, d_{\mathrm{MD}})| \tag{4}$$

Proof. As loss is binary-valued, $R(\ell, d_{\mathrm{TV}})$ simplifies to:

$$\begin{aligned} R(\ell, d_{\mathrm{TV}}) &= |\Pr[\ell(A_S, X) = 1 \mid M = 1] \\ &\quad - \Pr[\ell(A_S, X) = 1 \mid M = 0]| \\ &= |\mathbb{E}[\ell(A_S, X) \mid M = 1] \\ &\quad - \mathbb{E}[\ell(A_S, X) \mid M = 0)]| \\ &= |R(\ell, d_{\mathrm{MD}})|. \end{aligned}$$

□

Therefore, if a MIA adversary only observes whether a queried example has a correct or incorrect prediction by the target model, the upper bound on the success of any such attack has a direct relationship to standard overfitting $R(\ell, d_{\mathrm{MD}})$. Thus, for such an adversarial model, no overfitting *does* imply no vulnerability to MIAs.

4.3 Disparate Vulnerability

In this section, we provide a theoretical analysis of vulnerability to MIAs disaggregated by subgroups.

We introduce a subgroup-specific version of distributional generalization, in which the distributions of the property π are computed on examples that belong to a given subgroup. For any set T from the range of π, we define subgroup-specific measures:

$$\mu^{\pi}_{1,z}(T) \triangleq \Pr_{\substack{S\sim\mathcal{D}^n \\ x\sim(S|z)}} [\pi(A_S, x) \in T],$$

$$\mu^{\pi}_{0,z}(T) \triangleq \Pr_{\substack{S\sim\mathcal{D}^n \\ x\sim(\mathcal{D}|z)}} [\pi(A_S, x) \in T],$$

where $x \sim (\cdot \mid z)$ denotes sampling conditioned on the subgroup z.

Definition 4. For a property function $\pi(A_S, x)$, the *subgroup-specific distributional-generalization gap* is:

$$R_z(\pi, d) \triangleq d\big(\mu^{\pi}_{1,z},\ \mu^{\pi}_{0,z}\big),$$

where $d(\mu, \mu')$ is a measure of dissimilarity between probability distributions.

Subgroup vulnerability from distributional generalization. To extend the worst-case analysis to subgroups, we use the worst-case subgroup vulnerability under adversary's features W to the corresponding Bayes adversary: $V_z(\mathcal{A}^*_W)$. We show that this subgroup vulnerability is also related to distributional generalization:

Proposition 3. *The worst-case vulnerability of a subgroup z is bounded:*

$$V_z(\mathcal{A}^*_W) \leq R_z(\phi_W, d_{\mathrm{TV}}) \tag{5}$$

Moreover, for subgroup-aware adversaries the bound becomes an equality:

$$V_z(\mathcal{A}^*_{W,Z}) = R_z(\phi_W, d_{\mathrm{TV}}) \tag{6}$$

We defer the proof to Appendix A.

Formalizing disparate vulnerability. Finally, having discussed subgroup vulnerability, we can analyze disparate vulnerability. We define *disparity in vulnerability*:

Definition 5. Disparity in vulnerability (or disparity for short) between two subgroups z and z' is the difference in vulnerability of these subgroups:

$$\Delta V_{z,z'}(\mathcal{A}^*_W) \triangleq V_z(\mathcal{A}^*_W) - V^*_{z'}(\mathcal{A}^*_W)\,.$$

The previous results on the connection between subgroup vulnerability and distributional generalization enable us to relate disparity to degrees of distributional generalization across different population subgroups. From Proposition 3, we can see that the magnitude of disparity can be trivially bounded using distributional-generalization gaps of the involved subgroups:

Corollary 2. *Magnitude of disparity between subgroup z and z' is upper bounded:*

$$\left|\Delta V_{z,z'}(\mathcal{A}^*_W)\right| \leq \max\{R_z(\phi_W, d_{\mathrm{TV}}), R_{z'}(\phi_W, d_{\mathrm{TV}})\} \tag{7}$$

Moreover, disparity has an exact closed form for subgroup-aware adversaries:

Corollary 3. *Suppose that a subgroup-aware adversary uses features (W, Z). Then, disparity between subgroups z and z' is the difference between distributional generalization gaps of these subgroups:*

$$\Delta V_{z,z'}(\mathcal{A}^*_{W,Z}) = R_z(\phi_W, d_{\text{TV}}) - R_{z'}(\phi_W, d_{\text{TV}}) \,. \tag{8}$$

4.4 Takeaways

Necessary and sufficient condition for MIA existence. Without making any parametric assumptions, we have showed that the vulnerability to MIAs can be characterized using an extended notion of generalization, and that disparity is bounded by the difference in levels of distributional generalization across population subgroups. This interpretation of a standard result in statistical theory generalizes and complements the theoretical findings of Yeom et al. [43] and Sablayrolles et al. [36]. It also confirms that the presence of standard overfitting is not a necessary condition for MIAs to succeed [29, 43].

Hardness of defending against MIAs. The interpretation of worst-case vulnerability through distributional generalization has important consequences for practical defences against MIA that do not rely on differential privacy.

In order to reduce the vulnerability against adversaries that use features W, the distribution of W for examples that are outside of the training set has to be close to that for the training set examples. This means that, to avoid vulnerability, a target model has to—either implicitly or explicitly—learn the distribution of W [23] which is a stronger requirement than what is typically necessary for its main task (i.e. generalization in terms of accuracy, or average error).

Moreover, adversaries are not limited to one set of features W; thus, the distribution has to be learned for a multitude of possible configurations of adversarial features W. Additionally, to prevent disparity in vulnerability, the distribution of W has to be learned across population subgroups—an even more challenging task.

5 Detecting and Measuring Disparate Vulnerability

We showed in Section 4 that vulnerability to MIAs appears when a model lacks in distributional generalization. The degree to which records are vulnerable can vary across subgroups in the data, potentially resulting in disparate vulnerability. In this section, we provide mechanisms to reliably estimate subgroup vulnerability and its disparity in practice.

To empirically estimate MIA vulnerability, we simulate the MIA game with a real attack. If we could play the game infinite times, then estimating the success probability of the adversary would be trivial. In practice, however, we can only run the game a finite amount of times, which provides us with a finite number of challenge examples x. We group these examples into two sets of datasets of n elements: a set of r datasets $\{S_i\}_{i=1..r}$ composed of n “in” examples (i.e., sampled as in line 4 of the MIA game, used for training), and r datasets $\{\bar{S}_i\}_{i=1..r}$ composed of n “out” examples (i.e., sampled as in line 6 of the MIA game, not used for training). Each pair of datasets S_i and $\bar{S}_i$ can be seen as the train and test datasets of one model.

We define the estimate of vulnerability as:

$$\hat{V}(\mathcal{A}) \triangleq \frac{1}{r} \sum_{i=1}^{r} v_i \tag{9}$$

where v_i is the *model-specific estimate of vulnerability*: the advantage of the adversary against a single target model. We compute v_i for a pair of datasets S_i and $\bar{S}_i$ as:

$$\begin{aligned} v_i \triangleq 2 \cdot \frac{1}{2n} \Bigg(&\sum_{j=1}^{n} \mathbb{1}[\mathcal{A}(S_i^{(j)}, A_{S_i}, n, \mathcal{D}) = 1] \\ &+ \sum_{j=1}^{n} \mathbb{1}[\mathcal{A}(\bar{S}_i^{(j)}, A_{S_i}, n, \mathcal{D}) = 0] \Bigg) - 1, \end{aligned} \tag{10}$$

As r increases, $\hat{V}(\mathcal{A})$ approximates the value of the true vulnerability V.

We use the same approach to estimate subgroup vulnerability $V_z(\mathcal{A})$, but we only use examples that belong to the subgroup of interest z when computing the model-specific estimate of subgroup vulnerability $v_{i,z}$. We omit $\mathcal{A}$ when it is clear from context.

5.1 Statistical Detection of Disparity

When evaluating subgroup vulnerability, we have to rely on subsets of $(S_i, \bar{S}_i)$ formed by subgroup examples. These subsets are possibly of size much smaller than n. Due to the variance of the empirical averages in the Eq. (10), an estimate of subgroup vulnerability is in general less statistically reliable than the estimate of overall vulnerability that uses datasets $(S_i, \bar{S}_i)$ in their entirety.

As a result, when estimating disparate vulnerability using the estimates of subgroup vulnerability, we need to statistically ensure that, if found, disparity is not due to random chance.

More formally, given estimates $\{v_{i,z}\}_{i=1..r}$ across different subgroups, we want to find statistical evidence that the actual subgroup vulnerabilities differ:

$$V_{z_1} \overset{?}{\neq} V_{z_2} \overset{?}{\neq} \ldots \overset{?}{\neq} V_{z_t} \tag{11}$$

Multiple subgroups. This problem is an instance of a standard within-subjects experimental design: We have multiple measurements (model-specific vulnerability estimates for different subgroups $v_{i,z_1}, v_{i,z_2}, \ldots, v_{i,z_t}$) for the same subject (model A_{S_i}). We want to know whether the means of vulnerability values differ across subgroups. Therefore, we can determine whether the training algorithm exhibits disparate vulnerability using the repeated-measures one-way ANOVA model (see, e.g., Seltman [38, Chapter 14]). This approach enables us to use the ANOVA F-test to establish whether there is evidence of disparate vulnerability. Following the standard protocol, if the F-test is positive, we perform *post-hoc followup tests* to determine which particular pairs of subgroups exhibit disparity. For the post-hoc tests, we use pairwise dependent t-tests with correction for multiple comparisons. As the correction method, we use the standard Benjamini-Hochberg procedure for controlling the false detection rate.

Two subgroups. When comparing only two subgroups, z and z', the procedure naturally simplifies to running one dependent t-test that checks if the difference between means of two groups is significant.

5.2 The Bias Problem

Some attacks in the literature assume that the adversary has *additional knowledge* beyond the tuple $(x, A_S, n, \mathcal{D})$. This knowledge can result in the vulnerability estimation being positively biased: indicating higher vulnerability than the actual worst case within the knowledge model of $(x, A_S, n, \mathcal{D})$. Overestimating vulnerability is not necessarily an issue, as pessimistic estimates incentivize caution in deployment. However, if the positive bias is correlated with the parameters of a subgroup (e.g., higher bias for smaller subgroups), it leads to incorrect conclusions about *disparate* vulnerability.

In this section, we check whether estimates of vulnerability using attacks proposed in the literature are biased. We evaluate three attacks:

- **Shadow-model attack [39]**. An adversary trains a number of shadow models using the target training algorithm $A(\cdot)$ on datasets sampled from $\mathcal{D}^n$. The adversary uses these shadow models to train a machine-learning classifier to estimate the probability $\Pr[M \mid W]$. In our evaluation, we use 30 shadows and Gradient Boosting Trees as the attack classifier.
- **Average-threshold attack [43]**. An adversary has additional knowledge: the average loss on the training dataset τ and the loss function ℓ used to compute this average, $\big(\tau,\ \ell(\cdot,\cdot)\big)$, where $\tau \triangleq \sum_{x\in S} \ell(A_S, x)$. When attacking, the adversary uses τ as threshold to decide whether the challenge example was "in" (the example's loss less than threshold) or "out" (greater than threshold).
- **Optimal-threshold attack [6, 40]**. An adversary has additional knowledge: the loss function ℓ and the optimal loss threshold τ^* that separates the losses in the best way, $\big(\tau^*,\ \ell(\cdot,\cdot)\big)$, where

$$\begin{aligned}\tau^* \triangleq \arg\max_{\tau} \frac{1}{n} \sum_{x\in S} \mathbb{1}[\ell(A_S, x) \leq \tau] \\ + \mathop{\mathbb{E}}_{x\sim\mathcal{D}} \big[\mathbb{1}[\ell(A_S, x) > \tau]\big]\end{aligned}$$

The attack proceeds as the average-threshold one.

We deviate slightly from the attacks' original formulations. The threshold attacks use $W = \ell(A_S, X)$ as features, where the loss function is cross-entropy, whereas the original shadow-model attack used $W = (\hat{Y}, Y)$. For fairness, we make all adversaries use the threshold attacks' features.

As we want to evaluate subgroup-aware adversaries, we use features $W = \big(\ell(A_S, X), Z\big)$ for all attacks, with cross-entropy as loss function. We make the attacks subgroup-aware as follows. For the shadow-model attack, the adversary trains separate attack classifiers for each subgroup, and then applies the appropriate classifier to each challenge example. For the threshold attacks, we assume the adversary has different thresholds for each subgroup [6, 41], i.e., average loss, respectively optimal threshold, per subgroup.

Method. It is hard to tell exactly if an estimate is higher than the worst-case vulnerability, as in practice the worst case is unknowable. We propose a simple test for bias within our adversarial model: run the estimation method against *data-independent models*. A target model can be independent of its training data, e.g., if it is completely random, constant, or trained with differential privacy parameter $\varepsilon \approx 0$ (see Section 6.2). If the model is independent of the data, we expect the esti-

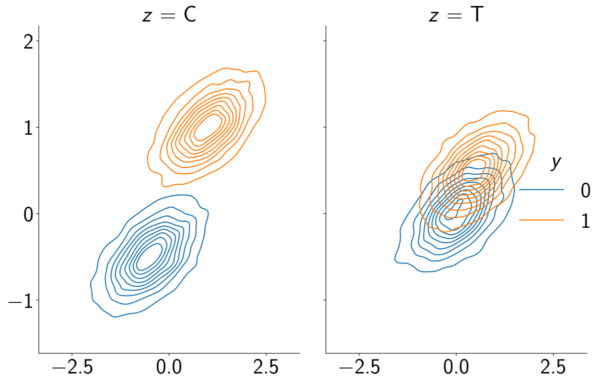

Fig. 3. Distribution of values in our synthetic data. *x-axis:* value of the 1-st dimension of the synthetic data, *y-axis:* value of the 2-nd dimension. We use 100-dimensional data for our experiments.

mates of overall and subgroup vulnerabilities, as well as disparity, to all be zero in expectation. We refer to any violation of this property as *null-model bias.* We are not only interested in whether a method exhibits such bias, but in whether this bias is correlated with subgroups.

Dataset. To have control over the distributions of subgroups and their representation, we create a synthetic dataset. We assume that the examples have binary class labels $y \in \{0, 1\}$, and belong to one of two subgroups $z \in \{C, T\}$. We generate the examples from the multivariate normal distributions:

$$\begin{aligned}\Pr(x \mid y = 0, z = C) &\sim \mathcal{N}(-1/2 \cdot \mathbf{1}^d, \Sigma)\\ \Pr(x \mid y = 1, z = C) &\sim \mathcal{N}(1 \cdot \mathbf{1}^d, \Sigma)\\ \Pr(x \mid y = 0, z = T) &\sim \mathcal{N}(0 \cdot \mathbf{1}^d, \Sigma)\\ \Pr(x \mid y = 1, z = T) &\sim \mathcal{N}(1/2 \cdot \mathbf{1}^d, \Sigma),\end{aligned}$$

where $\mathbf{1}^d$ is a d-vector of all ones, and the covariance matrix Σ is generated such that $||\Sigma||_{\max} \leq 1$. We use $d = 100$ dimensions, and set $\Pr[y = 1] = 1/2$. See Fig. 3 for an illustration.

To reflect that some subgroups can be harder to learn than others, the distributions are designed in such a way that the subgroup $z = C$ is more separable and hence more easily learnable than the subgroup $z = T$. In our experiments we use the subgroup $z = C$ as the control (or *majority*) subgroup with fixed number of representatives in the data, and $z = T$ as the treatment (or *minority*) subgroup whose size we vary.

Setup. To see if the potential null-model bias depends on the sizes of subgroups, we generate multiple synthetic datasets such that each contains data belonging to two subgroups: control and treatment. The control subgroup has 1000 representatives in each dataset; the size of the treatment subgroup varies between 25 and 1000, with 8 distinct values. We run 8 experiments with different subgroup proportions. Within each experiment, we train 200 target models on freshly generated datasets. We set the target training algorithm to output the same classifier for any input training dataset. Recall that because the models are independent of the input, we expect all vulnerability estimates to be zero on average. We estimate disparity using three attacks described above, and run t-tests to see if the estimates are statistically significant as explained in Section 5.1.

Results on our synthetic dataset. In Fig. 4, we can see that the estimates of disparity produced with the shadow-model attack and the average-threshold attack are centered around zero, with the statistical tests confirming no significant difference from zero. The estimates coming from the optimal-threshold attack, however, are highly biased compared to the other attacks, as the estimates are consistently and significantly ($p <$ 0.001) different from zero. The bias is always positive — overestimates disparity — and gets higher as the size of the treatment subgroup decreases. As the target models are independent of their training data and thus cannot have disparate vulnerability, we conclude that the use of the optimal-threshold attack results in significant null-model bias that grows as the subgroup size gets smaller.

Results on the dataset by Chang and Shokri [6]. To verify that our results are not artifacts of our specific synthetic data setup, we also reproduce the data setup used by Chang and Shokri to evaluate their subgroup-aware optimal-threshold attack. In their setup, they have one fixed dataset containing four subgroups that we denote as “0-0”, “0-1”, “1-0”, “1-1”, where the first number indicates simulated demographic group and the second number the class y (we refer to the original work [6] for details). The subgroups have 50, 450, 1000, and 1000 examples, respectively, with the total dataset size of 2500 examples. Following Chang and Shokri, we randomly subsample training datasets of size 1250 from the full dataset, and train one model on each. As before, we “train” a data-independent model. In this experiment, we only use threshold attacks due to the small size of the dataset (see Section 7 for more details). We use the ANOVA F-test as described in Section 5.1 to determine whether any of the subgroups have differing subgroup vulnerabilities.

Fig. 5 shows that significant null-model bias of the optimal-threshold attack also appears on this dataset (F-test $p <$ 0.001). In particular, the subgroup vulnera-

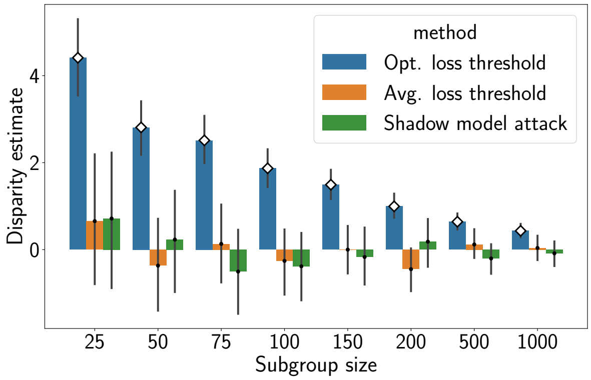

Fig. 4. Null-model bias of methods to estimate disparate vulnerability. Disparity in percentage points (*y-axis*) vs. size of the treatment subgroup in the training data (*x-axis*). Computed on synthetic datasets with fixed control subgroup (1000 examples) .The target training algorithm is data-independent: actual MIA vulnerability, subgroup vulnerabilities, and disparity in vulnerability are all zero. The error bars represent the variation across 200 model-specific estimates. The diamond marker (◊) means that an estimate significantly differs from zero with $p < 0.001$.

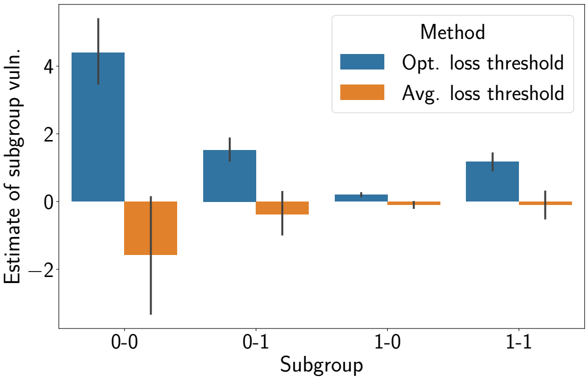

Fig. 5. Null-model bias on the synthetic data setup from Chang and Shokri [6]. Estimate of disparity in percentage points (*y-axis*) vs. subgroup (*x-axis*). The target training algorithm is data-independent, thus actual MIA vulnerability, subgroup vulnerabilities, and disparity in vulnerability are all zero.

bility for the smallest subgroup “0-0” with 50 examples appears as 4%. At the same time, the estimates from the average-threshold attack are centered around 0 and do not significantly differ (F-test $p \approx 0.1$), suggesting no null-model bias.

This bias, however, should not affect the conclusions by Chang and Shokri [6]. Rather than directly using the estimates of subgroup vulnerability, their analysis used *differences* in estimates of subgroup vulnerability between two models (a “fair” and a “regular” model). In their particular scenario, the bias introduced by the estimation should be cancelled out in the final difference. Although the conclusions of Chang and Shokri should not be affected by the bias, estimation methods such as the optimal-threshold attack should be avoided when evaluating disparate vulnerability in general.

Biased estimator in a prior version. A pre-print version of our work[1] used a vulnerability estimation method that, like the optimal-threshold attack, leveraged information about the training dataset of the target model. This estimator was therefore biased, and so were the numerical results of that version.

Takeaways. Biased estimators of vulnerability can result in consistent overestimation of disparity if the bias correlates with subgroup parameters. The shadow-model attack does not have such bias as it does not have access to any information about a specific target. Interestingly, the average-threshold attack, despite using an additional piece of knowledge that goes beyond our adversarial model, also does not exhibit such bias. On the contrary, the optimal-threshold attack produces significantly biased estimates for small groups.

Our results show the need to evaluate bias of the estimation method when measuring disparate vulnerability. To this end, we proposed to measure null-model bias, which detects bias when the worst-case vulnerability is zero. This test does not preclude a method from having bias if the worst-case vulnerability is larger. However, in practice MIA vulnerability has been shown to be relatively low.

5.3 Does Disparate Vulnerability Exist in ML Models?

Having established suitable methods for measuring disparate vulnerability, we apply them in a synthetic setup, and show that disparate vulnerability arise in practice.

Setup. To capture the effect of subgroup size in the training data, we create several experiments with different subgroup proportions. Within each experiment, we sample 200 dataset pairs S_i and $\bar{S}_i$ from our data distribution. In each dataset, the size of the control subgroup is fixed at 2500, and we vary the size of the treatment subgroup between experiments: 100, 500, 1000, and 2500. We estimate subgroup vulnerabilities using the subgroup-aware shadow-model attack (see Sec-

[1] https://arxiv.org/abs/1906.00389v2

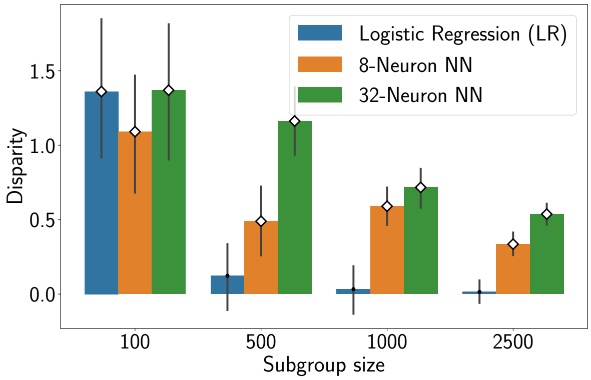

Fig. 6. Disparate vulnerability vs. subgroup representation in a training dataset. The *y-axis* represents disparity in vulnerability between the treatment group z and control group z' whose size is fixed to 2500, in percentage points. The error bars represent the variation across 200 model-specific estimates. Statistical significance markers: $p < 0.001$ ($\Diamond$), $p < 0.01$ ($\circ$), $p \geq 0.01$ ($\cdot$).

tion 5.2), because this attack is guaranteed to not have null-model bias. As before, we use $W = (\ell(A_S, X), Z)$ as adversary's features. To train shadow models, we independently sample 30 fresh datasets from our data distribution. We use t-tests to determine whether measured disparity is statistically significant (see Section 5.1).

Targets. We evaluate the following model families: logistic regression, and two ReLU neural networks with one hidden layer containing 8 and 32 neurons, respectively. We use the *scikit-learn* library [34] to train these models. All our models attain close to 100% test accuracy in our synthetic data setup.

Results. The results in Fig. 6 show that ML models can exhibit disparate vulnerability, even on a simple dataset. For all treatment sizes and targets, our estimates of disparity are significant ($p < 0.001$), with the exception of the logistic regression when the treatment subgroup is relatively well-represented ($500 - 2500$ examples). We also see that the sample size of the subgroup plays an important role in disparate vulnerability: *the less represented is a group in the training data, the higher the disparate vulnerability as compared to a better represented group*. Even though the sample size seems to be the dominant effect, we observe small but significant disparate vulnerability even when the subgroups are equally represented in training.

6 Mitigating Disparate Vulnerability

We study whether existing methods for addressing privacy and fairness in ML prevent disparate vulnerability.

6.1 Fairness Constraints

Due to the dependency of disparate vulnerability on the disparate *behavior* of the model across subgroups, minimizing the between-subgroup discrepancy in any given property, such as model's outputs or loss [11], intuitively could decrease disparate vulnerability.

Formally, let us denote by gap^π the total-variation distance between distributions of some property of a model $\pi(A_S, x)$ on examples coming from two subgroups z and z':

$$\mathsf{gap}^\pi \triangleq d_{\mathrm{TV}}\left(\Pr_{\substack{S\sim\mathcal{D}^n\\ x\sim(\mathcal{D}|z)}}[\pi(A_S,x)],\ \Pr_{\substack{S\sim\mathcal{D}^n\\ x\sim(\mathcal{D}|z')}}[\pi(A_S,x)]\right)$$

Certain notions of algorithmic fairness upper bound, or are equivalent to, the above gap given an appropriate choice of the property function: if we choose the model property to be its outputs, then for $\pi(A_S, x) = A_S(x)$, we obtain *demographic parity* [14]. Similarly, for the 0-1 loss property of the model, choosing $\pi(A_S, x) = \mathbb{1}[A_S(x) = y(x)]$ gives us *accuracy equality* [4].

In practice, a notion of fairness is satisfied on the training dataset rather than the whole data distribution. To capture this, we define an in-training gap as follows:

$$\mathsf{gap}^\pi_S \triangleq d_{\mathrm{TV}}\left(\Pr_{\substack{S\sim\mathcal{D}^n\\ x\sim(S|z)}}[\pi(A_S,x)],\ \Pr_{\substack{S\sim\mathcal{D}^n\\ x\sim(S|z')}}[\pi(A_S,x)]\right)$$

The following proposition establishes that, if the in-training gap is bounded and the model generalizes its fairness condition well, then vulnerability disparity is bounded to adversaries that use the property addressed by the fairness notion:

Proposition 4. *Suppose a subgroup-aware adversary uses features (W, Z), and the following two conditions are satisfied:*

1. *Fairness on the training data:* $\mathsf{gap}^{\phi_W}_S \leq \gamma$
2. *Fairness generalization:* $|\mathsf{gap}^{\phi_W} - \mathsf{gap}^{\phi_W}_S| \leq \delta$

Then, the magnitude of disparity in worst-case vulnerability is bounded as follows:

$$|\Delta V_{z,z'}(\mathcal{A}^*_{W,Z})| \leq 2\gamma + \delta$$

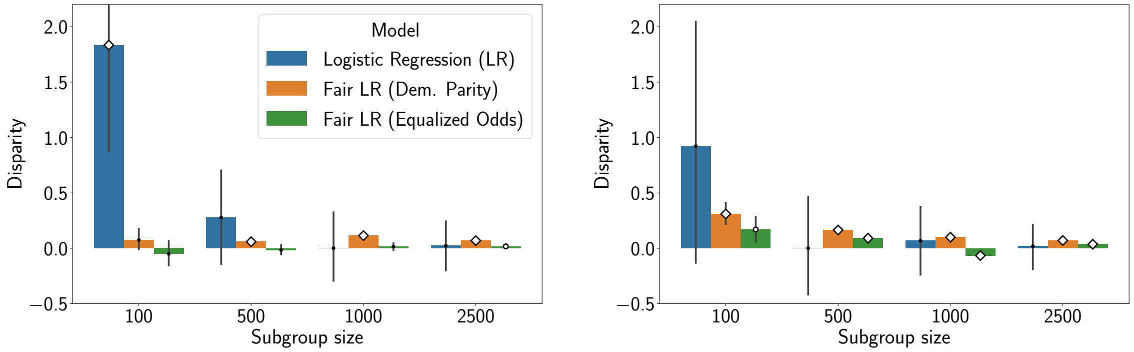

Fig. 7. Effect of algorithmic-fairness constraints on disparate vulnerability. The vulnerability is estimated with subgroup-aware attacks that use models' outputs as the feature *(left)*, and the models' loss *(right)*. The results for logistic regression are provided for reference (its values here are not comparable with the results of other experiments as the data dimensionality is different). See Fig. 6 caption for details.

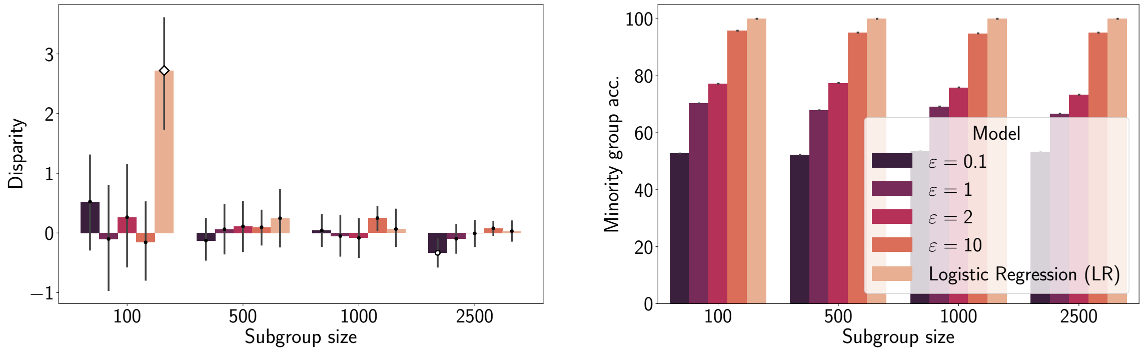

Fig. 8. Effect of differentially private training on disparate vulnerability *(left)*, and test accuracy *(right)*. The results for logistic regression are provided for reference. See Fig. 6 caption for details.

We defer the proof to Appendix A.

We note that these guarantees only apply to adversaries targeting the features addressed by implemented the fairness notion. In other words, just as in algorithmic-fairness literature where no single fairness measure is appropriate in a general context [17], no one fairness measure can provide guarantees for bounding disparate vulnerability for any adversary.

6.1.1 Empirical Evaluation

Fairness notions. To validate the theoretical results, we estimate vulnerability of models that satisfy two algorithmic-fairness notions: First, *demographic parity* [14] which ensures that distributions of model outputs between demographic subgroups are close: $\mathsf{gap}^{\phi_{\hat{Y}}} \approx 0$. Second, *equalized odds*, which ensures that true-positive rates and false-positive rates between the subgroups are close [19]. We choose these notions as they are common in the literature, and there exist efficient algorithms and tooling for producing classifiers that satisfy them. To train the classifiers, we use the threshold post-processing approach [19] from the *fairlearn* library [5], applied to a logistic regression classifier.

Setup. Within the setup of Section 5.3, we run the following two experiments:

E1 We fulfill the requirements of Proposition 4. For this, we estimate vulnerability using features equalized by demographic parity: $W = (\hat{Y}, Z)$. By Proposition 4, we expect low disparity in vulnerability *for both classifiers* as long as they generalize their fairness property well. In Appendix A, we show that in our data setup equalized odds implies demographic parity, thus the theoretical guarantee also applies for equality of odds.

E2 We estimate vulnerability using adversary's features $W = (\ell(A_S, X), Z)$ which do *not* match what the fairness property does, so the requirements of Proposition 4 are not fulfilled.

We find that with 100 dimensions in our setup, the threshold-optimization algorithm produces models that classify the data with 100% accuracy and no vulnerability. To demonstrate a setting where disparate vulnerability arises, we deviate from the parameters of Section 5.3 and use the synthetic dataset with 10 dimensions.

Results. We present the results in Fig. 7. For E1, we see that demographic parity decreases disparate vulnerability compared to standard logistic regression. This empirically confirms Proposition 4. For E2, as expected, both equalized odds and demographic parity do not completely prevent disparate vulnerability. Yet, they do decrease its magnitude by 3× compared to the standard logistic regression.

In our particular setup, the constrained models do not perform worse than the unconstrained models. In general, however, fairness notions can be inherently at odds with accuracy [44].

6.2 Differentially Private Training

In this section, we look at how learning with differential privacy [13] relates to disparity in vulnerability. We use the basic notion of differential privacy:

Definition 6. Training algorithm A satisfies ε-differential privacy (DP) if for any two datasets S, S' differing by the records of one individual, for any set of models T:

$$\Pr[A_S \in T] \leq \exp(\varepsilon) \Pr[A_{S'} \in T]$$

DP training limits the contribution of any individual in the dataset to the model training. Thus, DP should decrease vulnerability to MIAs. In particular, Yeom et al. [43], Chatzikokolakis et al. [7] and Humphries et al. [21], showed the advantage of a MIA adversary is bounded by DP in the setting of the MIA game. For example:

Proposition 5 (Adapted from Yeom et al. [43]). *If the training algorithm satisfies ε-DP, the worst-case vulnerability with any adversary's features W is bounded:*

$$V(\mathcal{A}_W^*) \leq \exp(\varepsilon) - 1 \tag{12}$$

These guarantees extend to disparate vulnerability:

Proposition 6. *If the training algorithm satisfies ε-DP, the worst-case subgroup vulnerability of any z, as well as magnitude of vulnerability disparity between any subgroups z and z', are uniformly bounded for any adversary's features W:*

$$V_z(\mathcal{A}_W^*) \leq \exp(\varepsilon) - 1, \quad \left|\Delta V_{z,z'}(\mathcal{A}_W^*)\right| \leq \exp(\varepsilon) - 1 \tag{13}$$

We defer the proof to Appendix A.

6.2.1 Empirical Evaluation

To study how DP affects disparate vulnerability we train DP models with different privacy levels. As target models, we use DP logistic regression with private empirical risk minimization [8], trained using the *diffprivlib* [20] implementation. We use a min-max scaler, and provide a maximum row norm estimated on a separate sample from the data distribution. We use privacy levels $\varepsilon = 0.1, 1, 2, 10$.

We see in Fig. 8 that, for all evaluated values of ε, DP training considerably reduces disparity compared to the non-private logistic regression, with statistical tests not detecting significant deviations from 0.

On the downside, unlike training with fairness constraints, DP training results in a significant decrease in accuracy of the models: from 45 p.p. to 5 p.p. drop depending on the value of ε.

6.3 Takeaways

Fairness only bounds disparate vulnerability in certain scenarios. Even when the classifier's fairness property generalizes beyond the training set, the bound is restricted to the adversarial strategy covered by the chosen fairness notion. Covering one adversarial strategy, however, is a weak security guarantee: the model could be (disparately) vulnerable to other strategies. Moreover, it is known that different fairness constraints are at odds with each other [17]. Hence, a model protected by one fairness notion may be inherently insecure against adversaries exploiting non-protected features.

Differential privacy bounds disparate vulnerability. We show that DP provides an upper bound on the vulnerability of all individuals, subgroups, and therefore on disparate vulnerability too. On the flip side, because DP guarantees are often at odds with accuracy, in practical applications ε is usually set high, allowing for a lot of variation within the upper bound of Proposition 6.

Practically, the particular approach to DP training that we evaluated has mitigated disparity even with a high privacy level $\varepsilon = 10$ that results in vacuous theoretical bounds, but at significant accuracy costs.

7 Evaluation on Real-World Data

To investigate if we can detect disparate vulnerability in a realistic setting, we use the following two datasets as case studies:

- ADULT *dataset* [25]. The dataset contains 48,842 examples from the 1994 Census database[2]. The task is to determine if a yearly salary is over/under \$50K. It contains attributes such as age, sex, education, race, native country, etc. After one-hot encoding, the dataset contains 91 features. We use the race column as the subgroup attribute.
- TEXAS-50K *dataset.* We create this dataset based on 2013 Texas Hospital Discharge data[3]. As our evaluation setup is computationally expensive, to accommodate the same training algorithms as used in the synthetic data experiments, we randomly subsample 50,000 examples, and reduce the number of features for training. We use the following columns: type of admission, illness severity, mortality risk, principal diagnosis code (out of more than 6000 codes, we only keep the top 1000 and create one separate code for the rest), length of stay, and patient's demographic attributes: sex, race, ethnicity. After one-hot encoding, we have 1025 features. We use the race column as the subgroup attribute. As a task, analogously to the ADULT dataset, we use prediction of whether the total amount of charges is greater than a threshold (e.g., for health-insurance risk-scoring). As the threshold we pick the median total charges on the subsampled dataset.

Table 1 provides details about the subgroups.

Target models. We consider as target models logistic regression and neural networks with 8 and 32 neurons in the hidden layer (Section 5.3), logistic regression with fairness constraints (Section 6.1), and differentially private logistic regression with ε values 1, 2, and 10 (Section 6.2). All the models beat the random accuracy baseline on the tasks.

Table 1. Subgroup representation in the datasets.

Dataset	z	Size
ADULT	"White" (WH)	38,903
	"Black" (BL)	4,228
	"Asian-Pac-Islander" (AI)	1,303
	"Amer-Indian-Eskimo" (AE)	435
	"Other" (OT)	353
	All	48,842
TEXAS-50K	4	31,514
	5	10,883
	3	6,451
	2	1,019
	1	133
	All	50,000

Estimation method. As opposed to our synthetic data setup in which datasets to train shadow models can be directly sampled from the data-generating distribution, when real data is involved we can only sample data from the available finite dataset. We split the dataset in two parts: one for training of the shadow models, and one for evaluation of vulnerability [39]. As a result, the amount of available training data is greatly reduced, in particular, for minority subgroups that already have few representatives in the dataset. To avoid this problem, in this section we use the average-threshold attack for vulnerability estimation, which does not require training shadow models. Our evaluation in Section 5.2 showed that this attack is not null-model biased.

Setup. To train each target model, we randomly subsample 50% of the dataset to use for training (S_i), and hold out the remaining data ($\bar{S}_i$). We train 200 models for each model family on different splits of the dataset. For our statistical tests (see Section 5.1), we use $\alpha = 0.01$ as significance level.

Results. We summarize the results in Table 2. As in our synthetic experiments, we observe evidence of disparity in neural networks. Importantly, the results show that low vulnerability in absolute terms does not imply absence of disparity. On ADULT, the 8-neuron network shows relatively low 0.4% vulnerability but statistically significant disparity ($p < 10^{-4}$). Interestingly, on TEXAS-50K, we also see statistical evidence of disparate vulnerability for logistic regression with demographic-parity constraints, although its overall vulnerability of 1.46% is comparable to standard logistic regression.

[2] https://archive.ics.uci.edu/ml/datasets/adult

[3] https://www.dshs.texas.gov/THCIC/Hospitals/Download.shtm

Table 2. Summary of models performance and vulnerability on adult and texas-50K. Columns: *Disparity test:* p-value of the anova F-test that checks if any of the subgroups have differing subgroup vulnerabilities, *Test acc.:* Test accuracy of models, *Gen. gap:* Per-model difference between train accuracy and test accuracy, *Vuln.:* Aggregate vulnerability $V(\mathcal{A})$. Bold font indicates models that have statistically significant disparity ($p < 0.01$).

adult	Disparity test	Test acc.		Gen. gap		Vuln., %	
	p	avg	std	avg	std	avg	std
Model							
Logistic Regression (LR)	0.3230	0.8404	0.0018	0.0012	0.0034	0.0942	0.4093
8-Neuron NN	**0.0000**	0.8421	0.0018	0.0044	0.0033	0.4052	0.3927
32-Neuron NN	**0.0000**	0.8410	0.0019	0.0131	0.0033	1.1373	0.4178
DP LR, $\varepsilon = 1$	0.8534	0.7797	0.0135	0.0006	0.0040	0.0830	0.3478
DP LR, $\varepsilon = 2$	0.0500	0.8053	0.0076	0.0004	0.0036	0.0563	0.3360
DP LR, $\varepsilon = 10$	0.0419	0.8321	0.0023	0.0011	0.0032	0.0888	0.4100
Fair LR (Dem. Parity)	0.8945	0.8267	0.0018	0.0011	0.0035	0.0980	0.3331
Fair LR (Equalized Odds)	0.7089	0.7941	0.0095	0.0006	0.0038	0.0782	0.3521

texas-50K	Disparity test	Test acc.		Gen. gap		Vuln., %	
	p	avg	std	avg	std	avg	std
Model							
Logistic Regression (LR)	0.2666	0.7833	0.0021	0.0152	0.0036	1.3905	0.4374
8-Neuron NN	0.0112	0.8836	0.0068	0.0282	0.0055	2.2384	0.5916
32-Neuron NN	**0.0000**	0.8639	0.0060	0.0686	0.0060	6.6238	0.7212
DP LR, $\varepsilon = 1$	0.6192	0.6175	0.0191	0.0002	0.0045	0.0540	0.4317
DP LR, $\varepsilon = 2$	0.0522	0.6363	0.0136	0.0014	0.0040	0.2125	0.3916
DP LR, $\varepsilon = 10$	0.9737	0.7114	0.0146	0.0038	0.0041	0.5224	0.3245
Fair LR (Dem. Parity)	**0.0078**	0.7609	0.0028	0.0143	0.0039	1.2393	0.3444
Fair LR (Equalized Odds)	0.7174	0.7477	0.0180	0.0133	0.0038	1.4676	0.3983

For the models with F-test $p < 0.01$, we conduct follow-up post-hoc tests to see which particular pairs of subgroups have high disparity (we defer the detailed results of the post-hoc tests to Appendix B). On adult, consistently with our synthetic experiments, the smaller subgroups "Asian-Pac-Islander" (AI, 1,302 examples), and "Other" (OT, 353 examples), exhibit disparity between themselves and other more populous subgroups. On texas-50K, almost all subgroup pairs exhibit significant disparity for 32-neuron network.

The results for the logistic regression with fairness constraints are unlike the synthetic experiments. As opposed to a minority subgroup, as in the previous results, disparity appears between the most populous subgroup "4" (31,514 examples) and subgroups "2", "3" and "5". This disparity does not exist in the standard logistic regression. Thus, this result shows that fairness constraints can introduce disparity when the conditions of Proposition 4 are not met.

Discussion. We have used binary classification tasks for compatibility with the fairness definitions, but we expect disparity to be more pronounced in multi-class settings. As detailed in Section 4.4, disparate vulnerability is bound to happen whenever a model does not faithfully learn the distributional properties of the data for some subgroups. Prior research suggests it is likely to appear when the task has many features, or many classes in the case of classification [37].

We also only considered relatively small dataset sizes. Bigger datasets, on the one hand, enable better learning of the models thus decreasing vulnerability and disparate vulnerability, but on the other hand, they would enable the adversary to use shadow-model attacks that could provide better results than the average-threshold attack used in our experiments.

We leave investigations of the effect of th number of classes and dataset size on disparate vulnerability for future work.

8 Conclusions

We have provided the first formal analysis of the disparate vulnerability of population subgroups to membership inference attacks. Our analysis provides new insights into why and when vulnerability to MIAs arises and why and when these attacks have disparate impact.

Key takeaways. The first key learning of our study is that fully preventing MIAs, and thus preventing disparate vulnerability can only be done in two ways. Either by significantly increasing the complexity of the learning problem to ensure distributional generalization; or using a differentially-private training algorithm with the associated hit on performance.

The second learning surfaces a more general problem: the consequences of the unreliability of privacy estimation for demographic groups with a minority representation in the data. We show that for small subgroups it is easy to incorrectly estimate their protection indirectly via aggregate privacy measures, or directly when not considering biases adequately.

Why disparate vulnerability is important. Disparate vulnerability has crucial legal and policy significance. Companies moving data between organizations or across borders face frictions designed to protect fundamental rights established by the approximately 140 countries with largely conceptually and textually similar privacy regulation around the world [18]. For example, moving data from Europe into a country with significant state surveillance apparatus, such as the United States, is difficult after the European Court of Justice's judgement in *Schrems II*. Other countries, such as several in South Asia, have established specific personal data localization laws [3]. As a consequence, there is growing interest in attempting to replace a direct trade in personal data with various forms of trade in models trained on this data.

Yet vulnerability of models to MIAs or other attacks compromising confidentiality might in some situations qualify models themselves as personal data [42]. The accountability principle in European data protection law places the onus on data controllers to demonstrate that a model should not be classified this way, for example through privacy-estimation techniques. Our study indicates there is a real risk of "privacy-washing", laundering a model with aggregate statistics that mask vulnerabilities of subgroups. It is true that prior work has also indicated that aggregate analysis can hide MIA vulnerability to attacks focusing on structurally vulnerable records [29]. However, this appears easier to dismiss as an acceptable residual leakage risk compared to disparate risks concerning members of salient minority groups, as in a liberal democracy, a regulator is more accountable towards these than towards a socially arbitrary selection of persons.

Open challenges. Our results also uncover a new challenge. It is difficult for auditors or regulators to practically inspect disparate vulnerability, because they might lack a sufficient number of examples relating to a minority group. When the subgroup data is scarce, our methods could be underpowered to detect disparity; however, not using the statistical tests and unbiased estimation methods from Section 5 risks flagging disparity always whenever subgroup data differ, devaluing the meaning of the estimate.

This points to a need for theoretical results that can be used as foundation in practical regulatory contexts. Theoretical results may be able to help regulators better ascertain the limits of metrics presented to them, and the conditions under which a model is structurally likely to be vulnerable to different types of privacy attacks even without difficult-to-obtain empirical evidence. The initial results provided in this paper can already significantly contribute to discussions around the classification of machine learning systems in relation to their risk of data leakage as business practices of using models to transport information continue to evolve.

9 Acknowledgements

The authors would like to thank Maksym Andriushchenko and Simon Oya for the helpful feedback and discussions. We also thank Hongyan Chang and Reza Shokri for clarifying their use of risk estimates [6].

This work was partially funded by the Swiss National Science Foundation with grant 200021-188824.

References

[1] Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. In *Annual Conference on Neural Information Processing Systems, NeurIPS*, 2019.

[2] Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.*, 2016.

[3] Arindrajit Basu, Elonnai Hickok, and Aditya Singh Chawala. The Localisation Gambit: Unpacking Policy Measures for Sovereign Control of Data in India. *Centre for Internet and Society, India*, 2019.

[4] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 2018.

[5] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. Fairlearn: A toolkit for assessing and improving fairness in AI. Technical Report MSR-TR-2020-32, Microsoft, May 2020. URL https://

www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/.

[6] Hongyan Chang and Reza Shokri. On the privacy risks of algorithmic fairness. *IEEE European Symposium on Security and Privacy, EuroS&P*, 2021.

[7] Konstantinos Chatzikokolakis, Giovanni Cherubin, Catuscia Palamidessi, and Carmela Troncoso. The Bayes security measure. *arXiv preprint arXiv:2011.03396*, 2020.

[8] Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. Differentially private empirical risk minimization. *J. Mach. Learn. Res.*, 2011.

[9] Giovanni Cherubin, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. F-BLEAU: Fast black-box leakage estimation. In *IEEE Symposium on Security and Privacy, S&P*, 2019.

[10] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 2017.

[11] Alexandra Chouldechova and Aaron Roth. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*, 2018.

[12] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

[13] Cynthia Dwork. *Differential Privacy*. Springer US, 2011.

[14] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science*, 2012.

[15] Michael D. Ekstrand, Rezvan Joshaghani, and Hoda Mehrpouyan. Privacy for all: Ensuring fair and equitable privacy protections. In *Conference on Fairness, Accountability and Transparency, FAT*, 2018.

[16] Farhad Farokhi and Mohamed Ali Kaafar. Modelling and quantifying membership information leakage in machine learning. *arXiv preprint arXiv:2001.10648*, 2020.

[17] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236*, 2016.

[18] Graham Greenleaf and Bertil Cottier. 2020 ends a decade of 62 new data privacy laws. *Privacy Laws & Business International Report*, 2020.

[19] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *NIPS*, 2016.

[20] Naoise Holohan, Stefano Braghin, Pól Mac Aonghusa, and Killian Levacher. Diffprivlib: The IBM differential privacy library. *arXiv preprint arXiv:1907.02444*, 2019.

[21] Thomas Humphries, Matthew Rafuse, Lindsey Tulloch, Simon Oya, Ian Goldberg, Urs Hengartner, and Florian Kerschbaum. Differentially private learning does not bound membership inference. *arXiv preprint arXiv:2010.12112*, 2020.

[22] Bargav Jayaraman, Lingxiao Wang, David Evans, and Quanquan Gu. Revisiting membership inference under realistic assumptions. *Proceedings on Privacy Enhancing Technologies*, 2021.

[23] Michael J. Kearns, Yishay Mansour, Dana Ron, Ronitt Rubinfeld, Robert E. Schapire, and Linda Sellie. On the learnability of discrete distributions. In *ACM Symposium on Theory of Computing*, 1994.

[24] Amir E Khandani, Adlar J Kim, and Andrew W Lo. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 2010.

[25] Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *International Conference on Knowledge Discovery and Data Mining, KDD*, 1996.

[26] Klas Leino and Matt Fredrikson. Stolen memories: Leveraging model memorization for calibrated white-box membership inference. In Srdjan Capkun and Franziska Roesner, editors, *USENIX Security Symposium*, 2020.

[27] Jiacheng Li, Ninghui Li, and Bruno Ribeiro. Membership inference attacks and defenses in classification models. In *CODASPY*, 2021.

[28] Zachary C. Lipton, Julian McAuley, and Alexandra Chouldechova. Does mitigating ML's impact disparity require treatment disparity? In *Annual Conference on Neural Information Processing Systems,NeurIPS*, 2018.

[29] Yunhui Long, Lei Wang, Diyue Bu, Vincent Bindschaedler, Xiaofeng Wang, Haixu Tang, Carl A Gunter, and Kai Chen. A pragmatic approach to membership inferences on machine learning models. In *IEEE European Symposium on Security and Privacy, EuroS&P*, 2020.

[30] Kristian Lum and William Isaac. To predict and serve? *Significance*, 2016.

[31] Preetum Nakkiran and Yamini Bansal. Distributional generalization: A new kind of generalization. *arXiv preprint arXiv:2009.08092*, 2020.

[32] Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Stand-alone and federated learning under passive and active white-box inference attacks. In *IEEE Symposium on Security and Privacy, S&P*, 2018.

[33] Ziad Obermeyer and Ezekiel J Emanuel. Predicting the future—big data, machine learning, and clinical medicine. *The New England journal of medicine*, 2016.

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

[35] David Pujol, Ryan McKenna, Satya Kuppam, Michael Hay, Ashwin Machanavajjhala, and Gerome Miklau. Fair decision making using privacy-protected data. In *Conference on Fairness, Accountability, and Transparency, FAT**, 2020.

[36] Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, and Hervé Jégou. White-box vs black-box: Bayes optimal strategies for membership inference. In *International Conference on Machine Learning, ICML*, 2019.

[37] Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. ML-leaks: Model and data independent membership inference attacks and defenses on machine learning models. In *26th Annual Network and Distributed System Security Symposium, NDSS*, 2019.

[38] Howard J Seltman. Experimental design and analysis. 2012.

[39] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy, S&P*, 2017.

[40] Reza Shokri, Martin Strobel, and Yair Zick. On the privacy risks of model explanations. *arXiv preprint arXiv:1907.00164*, 2019.

[41] Liwei Song and Prateek Mittal. Systematic evaluation of privacy risks of machine learning models. In *USENIX Security Symposium*, 2021.

[42] Michael Veale, Reuben Binns, and Lilian Edwards. Algorithms that remember: model inversion attacks and data protection law. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 2018.

[43] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *IEEE Computer Security Foundations Symposium, CSF*, 2018.

[44] Han Zhao and Geoffrey J. Gordon. Inherent tradeoffs in learning fair representations. In *Annual Conference on Neural Information Processing Systems, NeurIPS*, 2019.

A Proofs

In this section we provide the omitted proofs.

A.1 Regular vs. Subgroup-Aware Vulnerability

Proof of Proposition 1. Recall that the Bayes adversary uses a Bayes-optimal classifier that maximizes the success probability (i.e., vulnerability) among all the possible classifiers. That is, for the regular and subgroup-aware adversaries, we have respectively:

$$\begin{aligned} V(\mathcal{A}^*_W) &= \max_{g:\mathbb{W}\mapsto\{0,1\}} \Pr[g(\hat{Y}) = M] \\ V(\mathcal{A}^*_{W,Z}) &= \max_{g:\mathbb{W}\times\mathbb{Z}\mapsto\{0,1\}} \Pr[g(\hat{Y}, Z) = M]\,. \end{aligned}$$

Let $F = \{f \mid f = g \circ h, h(w,z) = w, g : \mathbb{W} \mapsto \{0,1\}\}$; that is, F is the set of functions $f : \mathbb{W} \times \mathbb{Z} \mapsto \{0,1\}$ that first reduce the tuple (w,z) to w and then apply a function g to the remaining input. Clearly, $F \subset \{g \mid g : \mathbb{W} \times \mathbb{Z} \mapsto \{0,1\}\}$.

Then, to prove this proposition it suffices to observe that the regular adversary is equivalent to a subgroup-aware one restricted to the set of functions F.

$$\begin{aligned} V(\mathcal{A}^*_{W,Z}) &= \max_{g:\mathbb{W}\times\mathbb{Z}\mapsto\{0,1\}} \Pr[g(w, Z) = M] \\ &\geq \max_{f\in F} \Pr[f(w, Z) = M] \\ &= \max_{g:\mathbb{W}\mapsto\{0,1\}} \Pr[g(w) = M] \\ &= V(\mathcal{A}^*_W)\,. \end{aligned}$$

□

A.2 Subgroup Vulnerability

To prove Proposition 3, we use the following statement:

Proposition 7. *For any two discrete probability measures μ and μ' the following holds:*

$$\sum_{x:\mu(x)>\mu'(x)} \big[\mu(x) - \mu'(x)\big] = \frac{1}{2}||\mu - \mu'||_1.$$

Proof. First, observe:

$$\begin{aligned} \frac{1}{2}||\mu - \mu'||_1 &= \frac{1}{2}\sum_x |\mu(x) - \mu'(x)| \\ &= \frac{1}{2}\sum_{\mu(x)>\mu'(x)} (\mu(x) - \mu'(x)) \\ &\quad - \frac{1}{2}\sum_{\mu(x)\leq\mu'(x)} (\mu(x) - \mu'(x)). \end{aligned}$$

Rearranging and grouping the terms, we get:

$$\begin{aligned} &= \frac{1}{2}\Big(\sum_{\mu(x)>\mu'(x)} \mu(x) - \sum_{\mu(x)\leq\mu(x)} \mu'(x) \\ &\quad - \sum_{\mu(x)>\mu'(x)} \mu'(x) + \sum_{\mu(x)\leq\mu'(x)} \mu'(x)\Big) \\ &= \sum_{x:\mu(x)>\mu'(x)} \big[\mu(x) - \mu'(x)\big] \end{aligned}$$

□

Proof of Proposition 3. We provide a proof for the case of discrete features W. The proof is analogous in the case of absolutely continuous W. Note that for discrete measures μ and μ', $d_{\mathrm{TV}}(\mu, \mu') = \frac{1}{2}||\mu - \mu'||_1$.

For convenience, let us define feature gaps as follows:

$$\begin{aligned} \mathsf{gap}(w) &\triangleq \mu_1(w) - \mu_0(w) \\ \mathsf{gap}_z(w) &\triangleq \mu_{1,z}(w) - \mu_{0,z}(w) \end{aligned}$$

Adversary's success for a subgroup has the following form that is useful for our proof:

$$\begin{aligned} &2\Pr[\mathtt{Att}^*(W) = M \mid Z = z] - 1 = \\ &= \Pr[\mathtt{Att}^*(W) = 1 \mid M = 1, Z = z] \\ &- \Pr[\mathtt{Att}^*(W) = 1 \mid M = 0, Z = z] \\ &= \sum_{w:\mathtt{Att}^*(w)=1} \mu_{1,z}(w) + \sum_{w:\mathtt{Att}^*(w)=1} \mu_{0,z}(w) \qquad (14) \\ &= \sum_{w:\mu_1(w)>\mu_0(w)} \big[\mu_{1,z}(w) - \mu_{0,z}(w)\big] \\ &= \sum_{w:\mathsf{gap}(w)>0} \mathsf{gap}_z(w) \end{aligned}$$

Table 3. Results of post-hoc tests on ADULT models. Columns: z *and* z'*:* identifiers of subgroups, t*:* value of the t statistic, p*:* uncorrected p-value, p*-corr.:* p-value after the correction for multiple comparisons.

NN-8	z	z'	t	p	p-corr.	NN-32	z'	z'	t	p	p-corr.
0	AE	AI	-4.4298	0.0000	**0.0001**	0	AE	AI	-11.3216	0.0000	**0.0000**
1	AE	BL	0.5143	0.6076	0.6751	1	AE	BL	0.9595	0.3385	0.3761
2	AE	OT	-1.7468	0.0822	0.1174	2	AE	OT	-4.1972	0.0000	**0.0001**
3	AE	WH	0.0498	0.9604	0.9604	3	AE	WH	0.5655	0.5724	0.5724
4	AI	BL	8.8677	0.0000	**0.0000**	4	AI	BL	24.1213	0.0000	**0.0000**
5	AI	OT	1.8976	0.0592	0.0987	5	AI	OT	6.1285	0.0000	**0.0000**
6	AI	WH	8.9236	0.0000	**0.0000**	6	AI	WH	25.4526	0.0000	**0.0000**
7	BL	OT	-2.6402	0.0089	0.0224	7	BL	OT	-6.4301	0.0000	**0.0000**
8	BL	WH	-1.3443	0.1804	0.2255	8	BL	WH	-1.2845	0.2005	0.2506
9	OT	WH	2.3290	0.0209	0.0417	9	OT	WH	6.1996	0.0000	**0.0000**

First, suppose that $Z \notin W$. Consider the following set:

$$C \triangleq \{w \mid \mathsf{gap}(w) > 0\} = \{w \mid \mu_1(w) > \mu_0(w)\}$$

For a given z, the set C is a union of two other disjoint sets A and B; $C = A \cup B$:

$$\begin{aligned} A &= \{w \mid \mu_{1,z}(w) \le \mu_{0,z}(w) \wedge \mu_1(w) > \mu_0(w)\} \\ B &= \{w \mid \mu_{1,z}(w) > \mu_{0,z}(w) \wedge \mu_1(w) > \mu_0(w)\} \end{aligned}$$

Thus, the sum in Eq. 14 can be decomposed into $\sum_A \mathsf{gap}_z(w) + \sum_B \mathsf{gap}_z(w)$, where

$$\sum_A \mathsf{gap}_z(w) = \sum_{\mathsf{gap}_z(w) \le 0 \wedge \cdots} \mathsf{gap}_z(w) \le 0$$

$$0 \le \sum_B \mathsf{gap}_z(w) \le \sum_{\mathsf{gap}_z(w) > 0} \mathsf{gap}_z(w) = \frac{1}{2}||\mu_{1,z} - \mu_{0,z}||_1$$

The last equality is by Proposition 7. Applying this bound to Eq. (14) we obtain the sought Eq. (5).

Second, suppose that $Z \in W$. Let $w = (\cdots, z')$. If $z' \neq z$, then $\mathsf{gap}_z(w) = 0$, and so we only need to consider the case $z' = z$. In this case:

$$\begin{aligned} \mathbb{1}[\mathsf{gap}(w) > 0] &= \mathbb{1}[\mu_1(w) > \mu_0(w)] \\ &= \mathbb{1}\left[\mu_{1,z}(w) \cdot \Pr[z] > \mu_{0,z}(w) \cdot \Pr[z]\right] \\ &= \mathbb{1}[\mathsf{gap}_z(w) > 0]. \end{aligned}$$

After plugging this into Eq. (14), we obtain the equality in Eq. (6) by Proposition 7. □

A.3 Bounds on Disparity From Algorithmic Fairness

Proof of Proposition 4. First, observe that a combination of the two conditions implies:

$$\mathsf{gap}^{\phi_W} = d_{\mathrm{TV}}(\mu_{0,z}, \mu_{0,z'}) \le \gamma + \delta$$

By this implication and the triangle property of total variation we have that:

$$\begin{aligned} d_{\mathrm{TV}}(\mu_{0,z'}, \mu_{1,z'}) &\le \underline{d_{\mathrm{TV}}(\mu_{1,z'}, \mu_{0,z})} + d_{\mathrm{TV}}(\mu_{0,z}, \mu_{0,z'}) \\ &\le \underline{d_{\mathrm{TV}}(\mu_{1,z'}, \mu_{0,z})} + \gamma + \delta \end{aligned}$$

Applying the triangle inequality to the underlined term:

$$\begin{aligned} \underline{d_{\mathrm{TV}}(\mu_{1,z'}, \mu_{0,z})} &\le d_{\mathrm{TV}}(\mu_{0,z}, \mu_{1,z}) + d_{\mathrm{TV}}(\mu_{1,z}, \mu_{1,z'}) \\ &\le d_{\mathrm{TV}}(\mu_{0,z}, \mu_{1,z}) + \gamma \end{aligned}$$

Combining the two,

$$\begin{aligned} d_{\mathrm{TV}}(\mu_{0,z'}, \mu_{1,z'}) - \gamma - \delta &\le \underline{d_{\mathrm{TV}}(\mu_{1,z'}, \mu_{0,z})} \\ &\le d_{\mathrm{TV}}(\mu_{0,z}, \mu_{1,z}) + \gamma \end{aligned}$$

Implying:

$$R_{z'}(\phi_W, d_{\mathrm{TV}}) - R_z(\phi_W, d_{\mathrm{TV}}) \le 2\gamma + \delta$$

If we apply the previous steps analogously we can also obtain:

$$\begin{aligned} d_{\mathrm{TV}}(\mu_{0,z}, \mu_{1,z}) - \gamma - \delta &\le d_{\mathrm{TV}}(\mu_{1,z}, \mu_{0,z'}) \\ &\le d_{\mathrm{TV}}(\mu_{0,z'}, \mu_{1,z'}) + \gamma \end{aligned}$$

Thus,

$$R_z(\phi_W, d_{\mathrm{TV}}) - R_{z'}(\phi_W, d_{\mathrm{TV}}) \le 2\gamma + \delta$$

Combining the inequalities, we get:

$$|R_z(\phi_W, d_{\mathrm{TV}}) - R_{z'}(\phi_W, d_{\mathrm{TV}})| \le 2\gamma + \delta$$

By Corollary 3, we obtain the sought bound. □

A.4 Differential Privacy Bounds Subgroup Vulnerability and Disparity

Proof of Proposition 6. Observe that the following probability distributions are equivalent:

$$\begin{aligned}\Pr_{\substack{S'\sim\mathcal{D}^{n-1}\\ x\sim(\mathcal{D}|z)}}[\phi_W(A_{S'\cup\{x\}}, x)] &\equiv \Pr_{\substack{S\sim\mathcal{D}^{n}\\ x\sim(S|z)}}[\phi_W(A_S, x)]\\ \Pr_{\substack{S'\sim\mathcal{D}^{n-1}\\ x\sim(\mathcal{D}|z)\\ x'\sim\mathcal{D}}}[\phi_W(A_{S'\cup\{x'\}}, x)] &\equiv \Pr_{\substack{S\sim\mathcal{D}^{n}\\ x\sim(\mathcal{D}|z)}}[\phi_W(A_S, x)]\end{aligned} \tag{15}$$

Notice that datasets $S' \cup \{x\}$ and $S' \cup \{x'\}$ differ by the records of at most one individual. Therefore, for any fixed dataset S', the post-processing property of differential privacy applies:

$$\begin{aligned}&\Pr_{x\sim(\mathcal{D}|z)}[\phi_W(A_{S'\cup\{x\}}, x)] \leq\\ &\leq \exp(\varepsilon) \Pr_{\substack{x\sim(\mathcal{D}|z)\\ x'\sim\mathcal{D}}}[\phi_W(A_{S'\cup\{x'\}}, x)]\end{aligned}$$

Taking expectation over S' of both sides, we obtain:

$$\begin{aligned}&\Pr_{\substack{S'\sim\mathcal{D}^{n-1}\\ x\sim(\mathcal{D}|z)}}[\phi_W(A_{S'\cup\{x\}}, x)] \leq\\ &\leq \exp(\varepsilon) \Pr_{\substack{S'\sim\mathcal{D}^{n-1}\\ x\sim(\mathcal{D}|z)\\ x'\sim\mathcal{D}}}[\phi_W(A_{S'\cup\{x'\}}, x)]\end{aligned}$$

By equivalence in Eq. (15):

$$\Pr[W \mid M = 1, Z = z] \leq \exp(\varepsilon) \Pr[W \mid M = 0, Z = z]$$

To get the bound on subgroup vulnerability, recall that by Proposition 3 it is upper bounded by the total variation. Thus, for any set of feature values T:

$$\begin{aligned}V_z(\mathcal{A}_W^*) \leq &\sup_{T\subseteq\mathbb{W}} |\Pr[W \in T \mid M = 1, Z = z]\\ &- \Pr[W \in T \mid M = 0, Z = z]|\\ &\leq \exp(\varepsilon) - 1\end{aligned}$$

Applying Corollary 2, we also get the bound on disparity.

□

A.5 A Note on Equalized Odds vs. Demographic Parity

Let us define equalized odds (EO). With probabilities over the data distribution, a classifier satisfies EO if:

$$\Pr[\hat{Y} \mid Y, Z = z] = \Pr[\hat{Y} \mid Y, Z = z']$$

In these terms, demographic parity is defined as the following requirement for a classifier:

$$\Pr[\hat{Y} \mid Z = Z] = \Pr[\hat{Y} \mid Z = Z']$$

In general, these two notions are not equivalent. In our synthetic data setup (Section 5.2), however, it holds that (a) the distributions of classes are the same across subgroups: $\Pr[Y \mid Z = Z] = \Pr[Y \mid Z = Z']$, and (b) the two classes are balanced: $\Pr[Y = 1] = \Pr[Y = 0] = 1/2$. It is easy to see that in this case, EO implies demographic parity.

B Additional Tables

The rest of the appendix contains additional tables.

Table 4. Results of post-hoc tests on TEXAS-50K models. See Table 3 caption for details.

NN-32	z	z'	t	p	p-corr.
0	1	2	-3.4973	0.0006	**0.0007**
1	1	3	0.2056	0.8374	0.8374
2	1	4	4.2820	0.0000	**0.0000**
3	1	5	3.0576	0.0025	**0.0028**
4	2	3	10.0174	0.0000	**0.0000**
5	2	4	21.2727	0.0000	**0.0000**
6	2	5	17.4069	0.0000	**0.0000**
7	3	4	21.8804	0.0000	**0.0000**
8	3	5	13.2434	0.0000	**0.0000**
9	4	5	-8.1600	0.0000	**0.0000**

LR (Dem. Parity)	z'	z'	t	p	p-corr.
0	1	2	-1.2485	0.2133	0.3326
1	1	3	-1.1910	0.2351	0.3326
2	1	4	-2.4808	0.0139	0.0348
3	1	5	-0.9385	0.3491	0.3879
4	2	3	0.3151	0.7531	0.7531
5	2	4	-3.4931	0.0006	**0.0020**
6	2	5	1.1152	0.2661	0.3326
7	3	4	-8.8594	0.0000	**0.0000**
8	3	5	1.6787	0.0948	0.1896
9	4	5	12.8701	0.0000	**0.0000**

Table 5. Results on ADULT, disaggregated by subgroups, for models with disparity F-test $p < 0.01$.

		Test acc.		Gen. gap		Subgroup vuln.	
		avg	std	avg	std	avg	std
Model	z						
32-Neuron NN	Amer-Indian-Eskimo	0.9028	0.0139	0.0115	0.0253	1.1701	4.8259
	Asian-Pac-Islander	0.8165	0.0119	0.0693	0.0195	5.7713	2.6300
	Black	0.9043	0.0049	0.0138	0.0086	0.8200	1.6261
	Other	0.8881	0.0179	0.0492	0.0295	3.2550	5.1807
	White	0.8338	0.0021	0.0109	0.0035	0.9773	0.4496
8-Neuron NN	Amer-Indian-Eskimo	0.9042	0.0151	0.0041	0.0281	0.3701	4.7177
	Asian-Pac-Islander	0.8264	0.0119	0.0223	0.0214	2.1320	2.7965
	Black	0.9066	0.0047	0.0035	0.0093	0.1878	1.6152
	Other	0.8913	0.0165	0.0149	0.0309	1.2805	5.6344
	White	0.8345	0.0020	0.0039	0.0036	0.3535	0.4314

Table 6. Results on TEXAS-50K, disaggregated by subgroups, for models with disparity F-test $p < 0.01$.

		Test acc.		Gen. gap		Subgroup vuln.	
		avg	std	avg	std	avg	std
Model	z						
32-Neuron NN	1	0.8699	0.0380	0.0791	0.0451	8.5188	8.2829
	2	0.8644	0.0153	0.1013	0.0180	10.7429	3.0129
	3	0.8498	0.0085	0.0855	0.0106	8.3947	1.6121
	4	0.8644	0.0066	0.0637	0.0063	6.0331	0.8261
	5	0.8708	0.0063	0.0697	0.0074	6.7288	1.0840
Fair LR (Dem. Parity)	1	0.6932	0.0562	-0.0010	0.0839	0.0075	8.9200
	2	0.6934	0.0203	0.0095	0.0295	0.8381	2.9201
	3	0.7323	0.0084	0.0143	0.0099	0.7667	1.1361
	4	0.7771	0.0027	0.0155	0.0048	1.5751	0.4952
	5	0.7384	0.0068	0.0106	0.0088	0.5997	0.8448