Investigating Membership Inference Attacks under Data Dependencies

Thomas Humphries
University of Waterloo
Waterloo, Canada
thomas.humphries@uwaterloo.ca

Simon Oya
University of Waterloo
Waterloo, Canada
simon.oya@uwaterloo.ca

Lindsey Tulloch
University of Waterloo
Waterloo, Canada
lindsey.tulloch@uwaterloo.ca

Matthew Rafuse
University of Waterloo
Waterloo, Canada
matthew.rafuse@uwaterloo.ca

Ian Goldberg
University of Waterloo
Waterloo, Canada
iang@uwaterloo.ca

Urs Hengartner
University of Waterloo
Waterloo, Canada
urs.hengartner@uwaterloo.ca

Florian Kerschbaum
University of Waterloo
Waterloo, Canada
florian.kerschbaum@uwaterloo.ca

***Abstract*—Training machine learning models on privacy-sensitive data has become a popular practice, driving innovation in ever-expanding fields. This has opened the door to new attacks that can have serious privacy implications. One such attack, the Membership Inference Attack (MIA), exposes whether or not a particular data point was used to train a model. A growing body of literature uses Differentially Private (DP) training algorithms as a defence against such attacks. However, these works evaluate the defence under the restrictive assumption that all members of the training set, as well as non-members, are independent and identically distributed. This assumption does not hold for many real-world use cases in the literature. Motivated by this, we evaluate membership inference with statistical dependencies among samples and explain why DP does not provide meaningful protection (the privacy parameter ϵ scales with the training set size n) in this more general case. We conduct a series of empirical evaluations with off-the-shelf MIAs using training sets built from real-world data showing different types of dependencies among samples. Our results reveal that training set dependencies can severely increase the performance of MIAs, and therefore assuming that data samples are statistically independent can significantly underestimate the performance of MIAs.**

***Index Terms*—Membership Inference Attacks, Differential Privacy**

I. Introduction

Machine learning (ML) is increasingly used to make predictions on privacy-sensitive data. In recent years, large tech companies such as Google and Amazon have begun to offer machine learning as a service to the general public through their cloud platforms. Although these systems can yield new and interesting insights in a variety of fields, machine learning models trained on sensitive data also present a lucrative attack surface for adversaries. In a Membership Inference Attack (MIA) [1], [2], an adversary attempts to identify the data that was used to train an ML model (i.e., its members). Successful MIAs violate the privacy of individuals and pose a significant threat to unprotected ML models [1].

Differentially private (DP) mechanisms can be applied to an ML model during the learning process [1]–[6] to limit the effect that a single data point can have on the model's output. Training ML models with DP mechanisms is becoming more common with popular libraries such as TensorFlow [7] and PyTorch [8] offering several DP training algorithms.

Previous work has provided experimental and theoretical evidence of DP training as a defence against MIAs [1], [2], [9], [10]. Yeom et al. [2] prove that DP training algorithms ensure a *theoretical upper bound* on the privacy leakage caused by MIAs. This bound, later improved by Erlingsson et al. [11], offers strong protection for high DP levels (corresponding to small values of the DP parameter ϵ), but quickly weakens for large ϵ. Despite this, empirical evaluations of MIAs against DP-trained models [10], [11] suggest that large values of ϵ, theoretically regarded as low-privacy settings, provide sufficient protection against MIAs in the wild. Motivated by these findings, Murakonda and Shokri [12] developed a tool for tuning ϵ based solely on the empirical performance of MIAs, thus increasing the model's utility at the expense of theoretical privacy guarantees. In short, this body of literature suggests that applying DP mechanisms to ML models, even in severely weakened privacy regimes, provides sufficient protection against known MIAs.

These works on membership inference all follow a similar evaluation methodology where members and non-members are sampled independently from the same distribution. However, this underlying assumption is not always representative of real data [13]–[19]. Furthermore, prior work has shown that data dependencies can lead to an overestimation of the privacy protection provided by DP [9], [20]–[23]. Motivated by this, we conduct the first study of MIA performance on ML models trained with DP-SGD [4] *under data dependencies*.

We begin by studying the methodology from previous work on MIAs and identify the assumptions that underpin their theoretical findings [2], [10], [24]. Since there are no publicly known instances of MIAs against ML models in the wild, researchers are limited to simulating attacks using theoretical games between the model owner and attacker called *membership experiments* [2]. We show that previous membership experiments assume independence between training samples, which might not represent many membership inference cases in practice. To solve this, we propose a new membership experiment that loosens the restrictive independence assumption. We demonstrate that previous DP-based bounds on the adversary's success do not hold in this more general setting. We emphasize that this does not mean that DP is broken or flawed. Rather, the guarantees of DP mechanisms no longer apply when defining membership inference with data dependencies. Intuitively, this is because DP mechanisms limit the effect that a *single* sample has on the output. However, when training samples have dependencies, related samples can leak additional information about this single sample.

Prior work has shown the importance of studying MIA performance both theoretically and empirically [10], [11]. Thus, we also study the effect of data dependencies on the *empirical performance* of MIAs under different instantiations of our membership experiment. Using six real-world datasets, we evaluate the performance of off-the-shelf attacks. These attacks have black-box query access to the target model and are given no additional background information about the data dependencies we consider. Using these attacks, we investigate three different types of data dependencies. First, to study the extent to which data dependencies can affect MIAs, we artificially split a dataset into members and non-members using a clustering algorithm. Our results show that data dependencies can lead to perfect membership inference even in the presence of DP training. Second, we investigate the effects of an explicit bias in an attribute of the training set. In particular, we create a gender and an education bias in the training set, and see that MIA performance typically increases as the training set bias becomes more pronounced. Finally, we consider MIA performance on various real-world examples of data dependencies that occur during data collection. For example, we study the effect of all members being from a specific health region (or hospital) and non-members from all other regions. We also study the performance of MIAs when members and non-members come from the same source (the US Census) but are curated by different researchers.

In all of our evaluations with statistical dependencies among training samples we find a significant increase in MIA performance over the related work where all samples are Independent and Identically Distributed (IID) [1], [2], [10], [11], [24], [25]. This shows that the assumption of data independence among training samples underestimates the attack performance. Our empirical evaluation yields attack performance greater than the bounds of DP, which confirms that the currently known DP bounds do not apply once we consider dependent data, and thus DP mechanisms are not a cure-all solution for membership inference. This highlights the importance of considering data dependencies in future MIA evaluations and urges new research to develop better defences, and more realistic theoretical bounds on MIA performance.

II. Preliminaries

In this section, we summarize the concepts related to membership inference attacks and differentially private machine learning that are most relevant to our work. For reference, our notation is summarized in Table I.

TABLE I: Notation

Notation	**Description**
$z = (x, y)$	Data point with feature vector x and label y
$\mathcal{Z}$	Space of all possible data points
S	Training set of size n; $S \in \mathcal{Z}^n$
$\mathcal{D}$	Distribution of a data point (over $\mathcal{Z}$)
$\boldsymbol{\mathcal{D}}$	Joint distribution of n data points (over $\mathcal{Z}^n$)
$[K]$	Set of integers from 1 to K
$\mathcal{A}$	Space of all possible machine learning models
A	Learning algorithm $A : \mathcal{Z}^n \rightarrow \mathcal{A}$
a	Instance of a trained model ($a \in \mathcal{A}$)
`Att`	Membership inference attack, outputs a bit
`Adv`	Membership advantage, `Adv` $\in [0, 1]$

We use $z = (x, y)$ to denote an element or data sample, where x is its feature vector and y is its class or label. Let $\mathcal{Z}$ be the element space, i.e., $z \in \mathcal{Z}$. We use $S \in \mathcal{Z}^n$ to denote a training set that contains n elements $z \in \mathcal{Z}$. Let $\mathcal{D}$ be a probability distribution over $\mathcal{Z}$; $z \sim \mathcal{D}$ means that z is randomly sampled from $\mathcal{D}$, and $S \sim \mathcal{D}^n$ means that S consists of n independent samples from $\mathcal{D}$. We use $\boldsymbol{\mathcal{D}}$ to denote a probability distribution over $\mathcal{Z}^n$, i.e., a *joint* distribution for all samples in a dataset; $S \sim \boldsymbol{\mathcal{D}}$ means the dataset S is sampled from $\boldsymbol{\mathcal{D}}$. We use $[K]$ to denote the set of integers from 1 to K. A training algorithm A is a (possibly randomized) function that takes a training set $S \in \mathcal{Z}^n$ and outputs a trained model $a \in \mathcal{A}$, where $\mathcal{A}$ is the space of trained models. We use $a = A(S)$ to denote that a is the model that results from applying the training algorithm A to the training set S. The goal of the trained model a is to solve a classification task; i.e., assign a label y to a feature vector x. In this work, we focus on neural networks trained with DP-SGD [4], which are a popular model choice for solving classification problems in machine learning. However, our theoretical findings are generic and apply to other models as well.

A. Membership Inference Attacks

Though useful for solving classification tasks, machine learning models are subject to various privacy attacks. In this work we focus on the Membership Inference Attack (MIA), whose goal is to determine whether or not a specific data point z was included in the training set of a target model a. This attack is particularly dangerous when the model is trained on sensitive data, where an individual's inclusion or exclusion in the dataset could reveal sensitive or compromising information to an attacker. MIAs are related to property inference attacks [26]–[28], but are fundamentally different. In a Property Inference Attack (PIA), the goal of the attacker is to infer some property of the training set that the model producer did not intend to share (i.e., a property common to all training set samples), whereas the goal of an MIA is to infer whether or not a particular sample was in the training set. We discuss the relationship to PIAs further in Section V.

We use `Att` to denote a membership inference attack. Typically, an MIA receives the target model a and a data point z, and knows the training mechanism A and some statistical information about the training data (e.g., $\mathcal{D}$ or $\boldsymbol{\mathcal{D}}$). The attack outputs a bit, which is $\texttt{Att} = 0$ when it decides that z is a member, and $\texttt{Att} = 1$ when it decides it is a non-member. Two of the most well-known membership inference attacks are the proposals by Shokri et al. [1] and Yeom et al. [2]. The *shadow model attack* by Shokri et al. [1] uses public data to train a set of shadow models, designed to mimic the target model's functionality. Then, it trains an additional attack model to identify the membership status of a sample using outputs from the shadow models. The *threshold attack* by Yeom et al. [2] assumes that the adversary has access to the loss function of the target model as well as the distribution of the loss on the private training data. Given a data sample z and the model a, the attack queries the model to obtain the loss of z, and decides that z is a member if its loss is below a threshold computed from the distribution information. A weaker variant of this attack assumes the adversary only knows the expected loss of the training set, and uses this value as decision threshold.

A variety of metrics can be used to measure the success of an MIA. The True Positive Rate (TPR) is the probability that the attack correctly identifies a member as such, and the False Positive Rate (FPR) is the probability of incorrectly guessing that a non-member is a member. A popular metric to measure the success of an MIA, which we use in this work, is the *membership advantage*, which Yeom et al. [2] define as $\texttt{Adv} \doteq \text{TPR} - \text{FPR}$. This metric is 0 when the attack randomly decides the membership of z, and is 1 when the attack always guesses the membership of z correctly.

B. Differential Privacy in Machine Learning

Differential Privacy (DP), a privacy notion introduced by Dwork et al. [29], has become the gold standard in database privacy and is a popular privacy notion in machine learning:

Definition II.1 ((ϵ, δ)-DP)**.** A training algorithm A provides (ϵ, δ)-DP iff, for any two neighbouring datasets $S, S' \in \mathcal{Z}^n$ (i.e., S and S' differ by a single entry), and all possible subsets of the space of trained models $\mathcal{R} \subseteq \mathcal{A}$,

$$\Pr(A(S) \in \mathcal{R}) \leq \Pr(A(S') \in \mathcal{R}) \cdot e^{\epsilon} + \delta\,. \tag{1}$$

The parameter ϵ captures the degree of leakage of the mechanism A [29]. Small values of ϵ indicate that a model trained with S is indistinguishable from a model trained with S' which, intuitively, makes it difficult to infer whether or not an element z is in the training set. The parameter δ makes it easier to satisfy the DP constraint by allowing a small chance of failure in the privacy guarantee. It is typical to choose $\delta < 1/n$, where n is the number of elements in the dataset [30].

There are different approaches to developing a differentially private training algorithm A. In the case of neural networks, the differentially private stochastic gradient descent technique by Abadi et al. [4] is widely used. This technique involves clipping the gradients used for updating the network's weights during training time, and adding Gaussian noise to the average of the gradients.

III. Membership Experiments

Evaluating the performance of MIAs is crucial to understanding how dangerous these attacks are in practice and how to protect against them. However, since there are no publicly known cases of MIAs against ML models in practice, researchers typically rely on *membership experiments* to evaluate MIAs. A membership experiment is a theoretical game between the data owner and the adversary, where the adversary tries to guess the membership of a data sample. The game specifies how the data owner generates the training data and which information is available to the attacker. A membership experiment provides a *theoretical definition of membership inference*, specifying the rules one has to follow when evaluating MIAs empirically, and allowing researchers to prove DP-based theoretical guarantees that hold within this controlled environment.

In this section, we review two membership experiments that appear in related work [2], [24]. We argue that these experiments make unrealistic assumptions and thus might not be representative of what one could expect in a real-world attack. We then propose a new membership experiment that relaxes these assumptions. We call this experiment the Mixture Model (MM) membership experiment. We argue that the MM experiment better represents an MIA that could occur in practice, and thus it provides a more realistic definition of what membership inference is.

A. Strong-Adversary Membership Experiment

Differential privacy implies, by Definition II.1, that an adversary observing a model a cannot easily distinguish whether it has been trained with $S = \tilde{S} \cup \{z\}$ or $S' = \tilde{S} \cup \{z'\}$, where $\tilde{S} \in \mathcal{Z}^{n-1}$, and $z, z' \in \mathcal{Z}$. This naturally leads to the *strong-adversary membership experiment* that we describe in Algorithm 1 ($\texttt{Exp}_{\texttt{STR}}$). This experiment has five inputs: an attack $\texttt{Att}$, a training algorithm A, a set of $n-1$ samples $\tilde{S}$, and two data samples z and z'. The data owner flips a bit b to decide whether to train a model with $S = \tilde{S} \cup \{z\}$ ($b = 0$) or with $S' = \tilde{S} \cup \{z'\}$ ($b = 1$). The adversary receives the trained model a, the set $\tilde{S}$, and data samples z and z', and also knows the training algorithm A. The adversary succeeds ($\texttt{Exp}_{\texttt{STR}} = 1$) if it correctly learns whether z or z' was a member of the training set of a (i.e., bit b). Nasr et al. [24] use this membership experiment, formulated as a privacy game, to prove that the ϵ-DP guarantees of current training algorithms are tight.[1]

The membership advantage in $\texttt{Exp}_{\texttt{STR}}$ is

$$\texttt{Adv}(\texttt{Att}, A, \tilde{S}, z, z') = \Pr(\texttt{Att} = 0 | b = 0) - \Pr(\texttt{Att} = 0 | b = 1)\,.$$

Here, the positive summand is the TPR and the negative one is the FPR (if we consider that $b = 0$ is a positive). Note that these terms also depend on the input variables $(\texttt{Att}, A, \tilde{S}, z, z')$ but

[1] Nasr et al. [24] use the so-called *unbounded* DP notion, where one of the neighbouring datasets has one additional element (i.e., S vs. $S' = S \cup \{z\}$). We stick to the so-called *bounded* DP notion (S and S' have the same size) here since this is the privacy notion followed by Yeom et al. in their membership experiment [2]

Algorithm 1 Strong-Adversary Membership Experiment [24]

```
procedure $\texttt{Exp}_{\texttt{STR}}(\texttt{Att}, A, \tilde{S}, z, z')$
    Choose $b \sim \{0,1\}$ uniformly at random;
    if $b = 0$ then
        Train $a = A(\tilde{S} \cup \{z\})$;
    else
        Train $a = A(\tilde{S} \cup \{z'\})$;
    Return 1 if $\texttt{Att}(a, z, z', \tilde{S}, A) = b$; else 0.
```

we have omitted this dependence in the notation for simplicity. Next, we prove that an (ϵ, δ)-DP training algorithm provides the following upper bound on the membership advantage:

Theorem III.1 (New Membership Advantage Bound). *Let A be an (ϵ, δ)-DP training algorithm. Then, for all attacks $\texttt{Att}$, sets $\tilde{S}$, and data samples z and z', the membership advantage in $\texttt{Exp}_{\texttt{STR}}$ satisfies*

$$\textit{Adv}(\texttt{Att}, A, \tilde{S}, z, z') \leq (e^{\epsilon} - 1 + 2\delta)/(e^{\epsilon} + 1)\,. \quad (2)$$

Proof. Kairouz et al. [31] show that, for an adversary that wants to distinguish between two neighbouring inputs of a DP mechanism, the following bounds on the adversary's TPR and FPR hold:

$$\text{FPR} + e^{\epsilon} \cdot (1 - \text{TPR}) \geq 1 - \delta\,, \quad (1 - \text{TPR}) + e^{\epsilon} \cdot \text{FPR} \geq 1 - \delta\,.$$

These expressions follow from (1) and trivially apply to the TPR and FPR in $\texttt{Exp}_{\texttt{STR}}$. Adding these expressions together we get

$$(1 - \text{TPR} + \text{FPR})(1 + e^{\epsilon}) \geq 2(1 - \delta)\,.$$

Applying the definition of membership advantage ($\texttt{Adv} \doteq \text{TPR} - \text{FPR}$) and leaving $\texttt{Adv}$ in one side of the inequality yields the desired upper bound. □

For values of δ and ϵ close to zero, the bound above is close to 0. This means that DP is potentially a very strong privacy guarantee since it ensures that the advantage of an adversary that *knows all training set samples except one* (i.e., $\tilde{S}$), and has two possible candidates for the unknown sample (z and z') will always be below a certain value. However, as Nasr et al. [24] admit, this adversary is often unrealistically strong, and perhaps assessing the performance of MIAs that are given all training set samples but one is not very helpful towards understanding MIA in practical cases.

Another problem with this experiment is the *relation* between the membership bit b, the observed samples z and z', and the trained model a. In $\texttt{Exp}_{\texttt{STR}}$, the bit b determines the training set of a and thus they are causally *related* ($\Pr(a|b) \neq \Pr(a)$). We argue that there is no such dependence in an actual MIA scenario. In practice, the data owner trains their model once, with their (one and only) training set S. Then, the adversary receives the model and tries to determine whether a sample z is a member of its training set. Here, the bit b is *a property of* z representing its membership status, and not a property of the model a, i.e., $\Pr(a|b) = \Pr(a)$. Yeom et al.'s experiment [2], described next, captures this relation.

B. IID Membership Experiment

Yeom et al. [2] propose the membership experiment that we show in Algorithm 2 This experiment does not receive the training set S as an input, but a distribution $\mathcal{D}$ from which the samples of S are independently sampled (IID). The data owner trains a using S and then chooses a bit $b \in \{0,1\}$ uniformly at random. This bit determines whether the adversary receives a member (z chosen randomly from S) or a non-member (a fresh sample $z \sim \mathcal{D}$). The adversary receives the element z and the trained model a, and knows the training set size n, the training algorithm A, and the element distribution $\mathcal{D}$.With this information, the adversary carries out an attack $\texttt{Att}(z, a, n, A, \mathcal{D})$ that outputs a bit, indicating the predicted membership of z. The adversary succeeds (denoted $\texttt{Exp}_{\texttt{IID}} = 1$) if the attack correctly infers the membership status of z.

Algorithm 2 IID Membership Experiment [2]

```
procedure $\texttt{Exp}_{\texttt{IID}}(\texttt{Att}, A, n, \mathcal{D})$
    Sample $S \sim \mathcal{D}^n$;
    Train $a = A(S)$;
    Choose $b \sim \{0,1\}$ uniformly at random;
    if $b = 0$ then
        Draw $z \sim S$;
    else
        Draw $z \sim \mathcal{D}$;
    Return 1 if $\texttt{Att}(z, a, n, A, \mathcal{D}) = b$; else 0.
```

Yeom et al. [2] prove the following upper bound on the membership advantage in $\texttt{Exp}_{\texttt{IID}}$:

Theorem III.2 (Yeom et al.'s bound [2]). *Let A be an $(\epsilon, 0)$-DP learning algorithm. Then, for all attacks $\texttt{Att}$, training set sizes n, and data point distributions $\mathcal{D}$, the membership advantage in $\texttt{Exp}_{\texttt{IID}}$ satisfies*

$$\textit{Adv}(\texttt{Att}, A, n, \mathcal{D}) \leq e^{\epsilon} - 1\,. \quad (3)$$

Note that the advantage is also upper bounded by 1, so this bound is loose for $\epsilon > \ln 2$. More recently, Erlingsson et al. [32] provide a tighter bound using results from Hall et al. [33] for the more generic setting of (ϵ, δ)-DP:

Theorem III.3 (Erlingsson et al.'s bound [32]). *Let A be an (ϵ,δ)-DP learning algorithm. Then, for all attacks $\texttt{Att}$, training set sizes n, and data point distributions $\mathcal{D}$, the membership advantage in $\texttt{Exp}_{\texttt{IID}}$ satisfies*

$$\textit{Adv}(\texttt{Att}, A, n, \mathcal{D}) \leq 1 - e^{-\epsilon}(1 - \delta)\,. \quad (4)$$

We can see that this bound is less than or equal to 1.

In Appendix A, we prove that the advantage bound we derive in Theorem III.1 holds in the case where members and non-members are *statistically exchangeable*, i.e., when the joint distribution of a sequence of n members and one non-member does not depend on the order they appear in the sequence. We also explain the IID experiment is a particular case of statistical exchangeability, and thus our bound holds in this case:

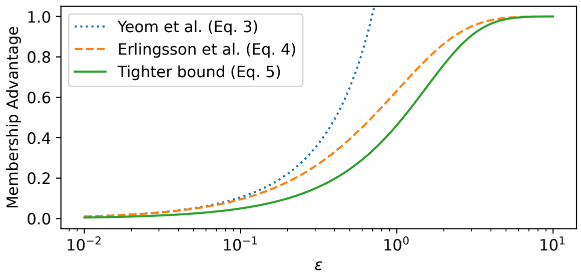

Fig. 1: Upper bounds on the membership advantage in the IID Membership Experiment.

Theorem III.4. *[New Membership Advantage Bound] Let A be an (ϵ,δ)-DP learning algorithm. Then, for all attacks $\texttt{Att}$, training set sizes n, and data point distributions $\mathcal{D}$, the membership advantage in $\texttt{Exp}_{\texttt{IID}}$ satisfies*

$$\mathit{Adv}(\texttt{Att}, A, n, \mathcal{D}) \leq (e^{\epsilon} - 1 + 2\delta)/(e^{\epsilon} + 1)\,. \qquad (5)$$

This bound improves over previous results [2], [32]; i.e., it is tighter than (3) and (4) for all $\epsilon \geq 0$ and $0 \leq \delta \leq 1$. We prove this in Appendix B. Figure 1 shows a comparison between the existing membership advantage upper bounds (Yeom et al. (3) and Erlingsson et al. (4)) and our bound (5). We used a value of $\delta = 10^{-5}$ for (4) and (5), since this is the value of δ we use in our experiments (described in Section IV-A) to ensure that δ is less than the inverse of the training set size. Our bound improves Erlingsson et al.'s by almost 0.2 in advantage for relevant privacy values $0.1 \leq \epsilon \leq 2$. These bounds suggest that high privacy settings $0.01 \leq \epsilon \leq 0.1$ completely thwart membership inference attacks (in the IID setting), values of $\epsilon \approx 1$ achieve intermediate privacy levels, and large values of $\epsilon \geq 10$ provide no worst-case privacy guarantee on the membership advantage.

We note that we have considered the original experiment by Yeom et al. [2], where members and non-members are sampled independently from $\mathcal{D}$. In practice, members and non-members are typically different samples, so *sampling without replacement* is a more realistic setting. Sampling without replacement introduces dependencies among members and non-members, but the exchangeability property still holds in that case, and thus our bound still applies. For simplicity, in the remainder of the paper we use the term *IID sampling* to denote either *sampling with or without replacement*, since our findings apply to both settings. The IID membership experiment captures the evaluation approach of most related works on membership inference [1], [2], [10], [11], [24], [25]. These works build the member and non-member sets by randomly splitting a dataset, which simulates IID sampling. However, this procedure excludes the case where there are dependencies between training set samples (e.g., a training set bias). There is a large body of literature in ML that studies this scenario with the goal of improving the accuracy of models on out-of-distribution samples [13]–[15]. For example, in a health care setting such as MRI scanning, the images seen in the wild are often from a different distribution to the training data of the model [16]. Spurious correlations in ML datasets have also received a lot of attention in the ML literature as they can have a detrimental effect on model accuracy [17]–[19]. Despite the problems of non-IID data being known to the ML community, this case has been completely neglected in the MIA literature.

C. A Membership Experiment with Data Dependencies

We have argued that the strong-adversary membership experiment considers an unrealistically strong adversary and does not capture the idea that membership should be a property of a given data point, and not the model itself. Also, the IID experiment is limited to a scenario that, as suggested by related work in ML, does not hold in many use cases. Motivated by this, we propose a new membership experiment that extends $\texttt{Exp}_{\texttt{IID}}$ to cases where the training set samples may have dependencies. We consider the particular case where the joint dataset distribution, denoted $\boldsymbol{\mathcal{D}}$, is a *multivariate mixture model*. A mixture model is a probabilistic model that characterizes the distribution of an overall population by taking into account the presence of subpopulations within the overall population. Each of the subpopulations is characterized by a probability distribution called a *mixture component*. To draw one sample from a mixture model, one first selects one of the mixture components (subpopulations) at random, and then draws a sample from the mixture component. Note that the mixture components may have overlapping support, which implies that, given a point sampled from the mixture model (i.e., the overall population), it might not be possible to identify from which of the mixture components (i.e., subpopulations) it was sampled from. We use *multivariate* mixture components, i.e., each component models the distribution of all attributes of a data sample within a subpopulation. This particular case of $\boldsymbol{\mathcal{D}}$ is amenable to theoretical analyses (e.g., Sect. III-C1) and is general enough to accommodate many real-life examples, as we show in Section IV-D. We call this the Mixture Model (MM) membership experiment, and denote it by $\texttt{Exp}_{\texttt{MM}}$.

Algorithm 3 Mixture Model (MM) Membership Experiment

```
procedure $\texttt{Exp}_{\texttt{MM}}(\texttt{Att}, A, n, \boldsymbol{\mathcal{D}})$    $\triangleright\ \boldsymbol{\mathcal{D}} = \{\mathcal{D}_1, \dots, \mathcal{D}_K\}$
    Choose $k \sim [K]$, sample $S \sim \mathcal{D}_k^n$;
    Train $a = A(S)$;
    Choose $b \sim \{0, 1\}$ uniformly at random;
    if $b = 0$ then
        Draw $z \sim S$;
    else
        Choose $k' \sim [K] \setminus k$, draw $z \sim \mathcal{D}_{k'}$;
    Return 1 if $\texttt{Att}(z, a, n, A, \boldsymbol{\mathcal{D}}) = b$; else 0.
```

Algorithm 3 shows the MM membership experiment. The joint distribution $\boldsymbol{\mathcal{D}}$ is a multivariate mixture model with K mixture components. Each mixture component $k \in [K]$ models the training set samples independently following a distribution $\mathcal{D}_k$. We overload the notation of $\boldsymbol{\mathcal{D}}$ in this case for notational simplicity, and use it to denote the set of all mixture distributions $\boldsymbol{\mathcal{D}} = \{\mathcal{D}_1, \dots, \mathcal{D}_K\}$. In short, to sample $S \sim \boldsymbol{\mathcal{D}}$, we choose a mixture component uniformly at random ($k \sim [K]$) and then draw n samples *independently* from $\mathcal{D}_k$ (line 2). Note that, even though the samples are independently drawn from a mixture component $\mathcal{D}_k$, they are statistically dependent since they are drawn from the *same* (unknown) mixture component. When $b = 1$, the non-member z is sampled from a mixture component that is *different* from the training set's component; i.e., $z \sim \mathcal{D}_{k'}$, for $k' \neq k$ (line 8). This experiment can also encompass the IID case in $\texttt{Exp}_{\texttt{IID}}$ when all $\mathcal{D}_k$ (for $k \in [K]$) are identical.

The MM membership experiment accommodates many realistic settings. As an example, consider a case where there are K hospitals in a city. Let $\mathcal{D}_k$ be the distribution that models the samples from the kth hospital, for all $k \in [K]$. In this example, $\texttt{Exp}_{\texttt{MM}}$ represents the case where the target model a is trained with samples belonging to a particular hospital $k \in [K]$, and the samples from all the other hospitals $k' \in [K] \setminus k$ are non-members. If the distributions of samples in each hospital are slightly different, there will be certain dependencies among members and non-members. Previous theoretical experiments ($\texttt{Exp}_{\texttt{STR}}$ and $\texttt{Exp}_{\texttt{IID}}$) cannot account for these dependencies. In Section IV we study this and other realistic examples that follow $\texttt{Exp}_{\texttt{MM}}$.

We note that, following $\texttt{Exp}_{\texttt{MM}}$, the adversary has statistical knowledge about the possible dependencies between training set samples ($\boldsymbol{\mathcal{D}}$). However, since the choice of k is not revealed, the adversary cannot tell which mixture component distribution the members follow. Even though this statistical knowledge is powerful, it does not provide the adversary any a-priori membership advantage. Indeed, note that, when $\epsilon = 0$ (or, alternatively, if the adversary did not receive the target model a), the adversary cannot do better than randomly guessing the membership of z ($\texttt{Adv} = 0$). This is because, without the model a, the distribution of z is the same regardless of the membership bit b. However, when $\epsilon > 0$, the adversary can exploit the statistical dependencies among samples to improve the membership inference. We also note that, in our evaluation in Section IV, we evaluate off-the-shelf MIAs that *do not use the statistical dependencies* captured by $\boldsymbol{\mathcal{D}}$. However, we will see that these dependencies still cause these existing attacks to perform better than in the IID scenario.

The MM membership experiment captures the relation between b, z, and a, considers a realistic adversary that only has statistical knowledge about the dataset, and accommodates both IID training set distributions ($\mathcal{D}_k \equiv \mathcal{D}_{k'}, \forall k, k' \in [K]$) as well as those that incorporate dependencies among samples or bias ($\mathcal{D}_k \not\equiv \mathcal{D}_{k'}$). Therefore, we believe that this experiment provides a better template to define membership inference, and that evaluating MIAs under this experiment will provide results that are more representative than previous experiments of what one could expect in practice.

1) DP-based bounds on the generalized membership experiment: We note that existing DP-based bounds [2], [32] on the membership advantage $\texttt{Adv}$, as well as our new bound in Eq. (5), do not apply in the MM membership experiment. We note that this is not a flaw of DP: the training algorithms we evaluated provide differential privacy, but this is not enough to bound the leakage when defining membership inference as in $\texttt{Exp}_{\texttt{MM}}$. The reason for this is that DP mechanisms limit the effect that a *single* sample has on the output (see Def. II.1). However, when training samples have dependencies, limiting the effect of a single input is less effective since the remaining data points can also leak information about this input due to their co-dependencies [23].

We provide more intuition as to why the previous bounds do not hold with an example. Without loss of generality and for simplicity, in this analysis we assume that the space of samples $\mathcal{Z}$ is discrete. In both $\texttt{Exp}_{\texttt{IID}}$ and $\texttt{Exp}_{\texttt{MM}}$, the adversary observes the released model a and a sample z (along with general parameters n, A, $\mathcal{D}$, that we omit from the probability expressions here for notational simplicity). The joint distribution of a and z depends on whether z was a member of a's training set ($b = 0$), or a non-member ($b = 1$). Next, we write general expressions for the likelihood of a and z given the value of b. We use these expressions to compute the maximum likelihood ratio, which bounds the adversary advantage, and study how an $(\epsilon, 0)$-DP mechanism bounds this ratio in the IID and non-IID cases.

First, when z is a member ($b = 0$), and using $\tilde{S}$ to denote all other $n - 1$ members, we can write

$$\Pr(a, z|b = 0) = \Pr(z) \cdot \sum_{\tilde{S}} \Pr(\tilde{S}|z) \Pr(A(\tilde{S} \cup \{z\}) = a) \,.$$

Here, $\Pr(\tilde{S}|z)$ is the probability of $n - 1$ training set samples (all but z) conditioned on the fact that z is also a training set sample.

When z is a non-member (denoted $b = 1$), $\tilde{S}$ still denotes $n-1$ members, and we use z' to define the nth member. Then, using the law of total probability we can write the likelihood as

$$\Pr(a, z|b = 1) = \Pr(z) \sum_{z'} \Pr(z') \sum_{\tilde{S}} \Pr(\tilde{S}|z') \Pr(A(\tilde{S} \cup \{z'\}) = a).$$

(As above, $\Pr(\tilde{S}|z')$ is the probability of the $n - 1$ training set samples $\tilde{S}$ given the remaining training sample z'.)

The maximum likelihood ratio $L_{max} = \max_{a,z,\mathfrak{b}} \frac{\Pr(a,z|b=\mathfrak{b})}{\Pr(a,z|b=1-\mathfrak{b})}$ can be used to bound the adversary advantage.[2] We show why using an $(\epsilon, 0)$-DP training algorithm provides a strong bound on this ratio in the IID case ($\texttt{Exp}_{\texttt{IID}}$), but there are non-IID cases ($\texttt{Exp}_{\texttt{MM}}$) where the bound is not meaningful. First, recall that an $(\epsilon, 0)$-DP training algorithm ensures that, for two training sets S and S' that differ in *one* sample, and for all models $a \in \mathcal{A}$, $\Pr(A(S) = a) \leq e^{\epsilon} \cdot \Pr(A(S') = a)$. In the IID scenario ($\texttt{Exp}_{\texttt{IID}}$), samples are independent and therefore

[2] We omit the derivations for space issues; in short, L_{max} lower-bounds the adversary's probability of error by $1/(1 + L_{max})$, which in turn upper-bounds the advantage by $\texttt{Adv} \leq (L_{max} - 1)/(L_{max} + 1)$. Note this matches (5) for $\delta = 0$.

$\Pr(\tilde{S}|z) = \Pr(\tilde{S})$. The likelihood ratio between hypotheses $b = 1$ and $b = 0$ is bounded by:

$$\begin{aligned}\frac{\Pr(a, z|b=1)}{\Pr(a, z|b=0)} &= \frac{\sum_{z'} \Pr(z') \sum_{\tilde{S}} \Pr(\tilde{S}) \Pr(A(\tilde{S} \cup \{z'\}) = a)}{\sum_{\tilde{S}} \Pr(\tilde{S}) \Pr(A(\tilde{S} \cup \{z\}) = a)} \\ &\leq \frac{\sum_{z'} \Pr(z') \sum_{\tilde{S}} \Pr(\tilde{S}) \Pr(A(\tilde{S} \cup \{z\}) = a) e^{\epsilon}}{\sum_{\tilde{S}} \Pr(\tilde{S}) \Pr(A(\tilde{S} \cup \{z\}) = a)} \\ &= \sum_{z'} \Pr(z') \cdot e^{\epsilon} = e^{\epsilon} \,.\end{aligned}$$

The same applies to the reciprocal $\Pr(a, z|b=0)/\Pr(a, z|b=1)$. In the MM membership experiment, if we are explicitly in a non-IID case, then $\Pr(\tilde{S}|z) \neq \Pr(\tilde{S})$ and the derivations above cannot be carried out. As an example, consider the pathological case where the mixture components in $\boldsymbol{\mathcal{D}}$ are different and deterministic, i.e., all the training set samples are identical. In that case, $\Pr(\tilde{S}|z)$ is only non-zero when $\tilde{S} = \{z\}^{n-1}$ (in that case, $\Pr(\{z\}^{n-1}|z) = 1$). Then,

$$\begin{aligned}\frac{\Pr(a, z|b=1)}{\Pr(a, z|b=0)} &= \frac{\sum_{z'} \Pr(z') \sum_{\tilde{S}} \Pr(\tilde{S}|z') \Pr(A(\tilde{S} \cup \{z'\}) = a)}{\sum_{\tilde{S}} \Pr(\tilde{S}|z) \Pr(A(\tilde{S} \cup \{z\}) = a)} \\ &= \sum_{z'} \Pr(z') \frac{\Pr(A(\{z'\}^n) = a)}{\Pr(A(\{z\}^n) = a)} \,.\end{aligned}$$

We can apply the DP definition n times and bound the expression above by $e^{n \cdot \epsilon}$, but this is certainly not useful since n is typically very large. More generally, bounding the expression above requires using DP to provide *group privacy*, which does not yield useful bounds for large n [30]. Even though this is an extreme example, it illustrates why the previous DP-based bounds (that are independent of dataset size) on the membership advantage do not apply to $\texttt{Exp}_{\text{MM}}$.

The pathological example above shows that an upper bound on $\texttt{Adv}$ can be very large in non-IID scenarios. However, there might be particular distributions $\boldsymbol{\mathcal{D}}$ for which one could prove an upper bound that lies closer to the IID bound (5). Also, empirical MIAs have been shown to achieve advantage levels $\texttt{Adv}$ far below DP-based upper bounds [10] in the IID scenario. Thus, it is important to empirically measure the performance of MIAs in the non-IID case, as performance could be similar to the IID case. We study this in the next section.

IV. Empirical Evaluation of Membership Inference with Data Dependencies

In this section, we empirically investigate the effects of non-IID training sets on the performance of off-the-shelf MIAs, following our MM membership experiment $\texttt{Exp}_{\text{MM}}$.

A. Evaluation Setup

We explain how we instantiate $\texttt{Exp}_{\text{MM}}$ with real-world datasets. In each of our experiments, we initialize K disjoint sets of samples $\{D_k\}_{k \in [K]}$ with real data (in most of our experiments, we use $K = 2$). Each of these sets characterizes one of the mixture components: drawing a sample from $\mathcal{D}_k$ is simply drawing a sample from the set D_k. We use sampling *without replacement* to ensure that every member and non-member is a distinct data sample. Even though this is an obvious assumption in practice (duplicate samples do not make sense in most problems), we note that all the theoretical experiments in Section III allow for identical members and non-members.

In each of the sections below, we follow a different approach to build the sets of samples $\{D_k\}_{k \in [K]}$, to study different types of data dependencies. In each experiment, given a training set size n, we take n samples from a mixture component D_k as the member set, and draw n samples at random from all other mixture components as non-members. We train the model on the members, and evaluate each of the attacks we consider on all members and non-members. We compute an attack's TPR by looking at the proportion of members correctly identified as such, and the FPR by computing the proportion of non-members wrongly classified as members by the attack. Then, the attack's membership advantage is simply $\texttt{Adv} =$ TPR – FPR. We label this advantage as "non-IID" in our plots. We also measure the classification accuracy of the model over members and non-members in this non-IID case, to give insight into the attacks' performance.

In order to study the effect of data dependencies vs. independent data sampling, we also measure the membership advantage if members and non-members were IID samples. To do this, for each non-IID experiment, we merge the member and non-member sets, and generate a new member and non-member set by resampling from this merged set (we keep the original member and non-member set sizes). The advantage in this case represents the attack's performance in the IID experiment, and we label it "IID" in our plots.

a) Datasets: We use three publicly available datasets considered in prior work [1], [34] and three additional datasets from the UCI machine learning repository [35]:

1) The `adult` dataset [36], which contains 48 842 data samples from the 1994 US census, each with 14 attributes such as age, gender, and education. The binary classification task is deciding whether the individual earns more than $50k a year.
2) The `compas` dataset [37], extracted from ProPublica's investigation into racial bias in ML, which contains 6 172 samples with 15 attributes such as age, sex, and number of prior offenses. The classification task is predicting whether or not an individual re-offended within 2 years.
3) The `texas` dataset [38] contains a series of records of inpatient stays in various hospitals published by the Texas Department of State Health Services. The classification task is predicting the procedure. We follow the approach by Shokri et al. [1] and compute the top-100 most popular procedures as the classification label (we discard any rows not in the top-100). The final dataset contains 350 280 samples each with 66 attributes such as length of stay, age, and total charges.

TABLE II: Evaluation Setup Summary

Evaluation (Section)	Dataset and experiment	Number of Features (after encoding)	Number of classes	Members	Non-members	Shadow model data
IV-B	`adult` cluster	104	2	10 000	10 000	15 132
	`compas` cluster	15	2	2 000	2 000	-
IV-C	`adult` education	87	2	10 000	10 000	10 772
	`adult` gender	103	2	10 000	10 000	11 192
IV-D	`heart`, Cleveland	22	2	303	617	-
	`heart`, Hungary	22	2	294	626	-
	`heart`, Switzerland	22	2	123	797	-
	`heart`, VA Long Beach	22	2	200	720	-
	`students`, Gabriel Pereira	42	2	423	226	-
	`texas`, region #3	269	100	10 000	10 000	20 000
	`census` and `adult`	100	2	10 000	10 000	20 000

4) The `heart` datasets [39] are four datasets collected from hospitals in Cleveland, Hungary, Switzerland, and the VA Long Beach. The datasets contain 303, 294, 123, and 200 data samples, respectively. We use the 14 attributes common to all datasets such as age, cholesterol, and fasting blood sugar with the classification task of predicting if the patient has a heart condition.
5) The `students` dataset [40] contains data samples from a study on student achievement in Portuguese schools. We use the Portuguese language dataset, which has 649 samples from two schools (423 and 226 samples each) containing 30 attributes such as age, study time, and number of absences. The classification task is predicting if the student passed or failed the course, based on the final grade. We remove the two intermediate grades from the attribute list, as they are highly correlated with the final grade.
6) The `census` dataset [41] contains data extracted from the 1994 and 1995 US Census. We take all 99 827 data samples from the 1994 dataset and extract the 10 attributes in common with the `adult` dataset. The classification task is to predict whether an individual earns more than $50k a year.

We replace missing values with the mean or mode value for that attribute and use a one-hot encoding for all categorical attributes. For example, if the categorical attribute has values red, green, and blue we add three new binary attributes, one for each colour. Since we only used a subset of the attributes for certain datasets, the rows were not always unique. To address this, we remove duplicate rows, keeping a single copy of each duplicate. Table II summarizes how we use these datasets in our experiments. In most experiments, we set a value for n ($n = 10\,000$ in most cases, and $n = 2\,000$ in `compas`), and sample n members and n non-members using the procedure described above. When the remaining data is large enough, we use it to evaluate the shadow model attack (when we do not have enough data, denoted "–" in Table II, we skip this attack). In experiments where the datasets are small (`heart` and `students`), we use an uneven number of members and non-members (we explain this in Section IV-D).

b) Model Architecture: Following the work of Jayaraman et al. [10], and others [1], [4], we use a ReLU network with 2 hidden layers, each with 256 neurons trained for 100 epochs as the target model in all evaluations. The models use the DP ADAM Gaussian Optimizer and RDP accountant from TensorFlow Privacy [42] with ℓ_2 regularization to avoid overfitting. The default hyper-parameters are 10^{-5} as the ℓ_2 regularization coefficient, 10^{-2} as the learning rate, and 200 as the batch size [10]. We vary ϵ in our evaluations and fix δ to equal 10^{-5}. This follows the general recommendation that δ should be less than the inverse of the training set size [30]. We note that $\epsilon = 0$ still leaks information, since $\delta > 0$.

c) Membership Inference Attacks: We consider two membership inference attacks evaluated in prior work [10] that we summarized in Section II-A. The *shadow model attack* [1] requires a large pool of publicly available data in order to train shadow models. Therefore we only run this attack on datasets that have leftover samples after building our member and non-member sets (see Table II). Following prior work [1], [10], we train five shadow models, each with the same architecture as the target model. The attack model uses the same architecture as the target model but with 64 neurons in the hidden layer. We consider two variants of the *threshold attack* [2]: one that knows the true loss distribution and therefore chooses the optimal decision threshold (optimal threshold attack) and one that uses the average training loss as decision threshold (average threshold attack). We remark that we do not modify these attacks. The attacks only get black box query access to the model a (with confidence scores), the training loss (log loss attack), public data following $\mathcal{D}$ (shadow model attack), and the sample z. That is, they do not use all the information that we allow the attacker in Algorithm 3. Specifically, we omit the set size n and the training algorithm A. The attacks were not designed to exploit dependencies between training samples and the adversary never explicitly learns these dependencies. Furthermore, in each experiment, we are careful not to include attributes that can trivially identify the mixture component that a sample was drawn from (we clarify this below).

d) Implementation: We used Python 3.8 for our implementation[3], building upon the code of Jayaraman et al. [43]. This code includes the implementation of the threshold attack and shadow model attack, where the latter is based on Shokri et al. [1]. To implement RDP, we use the RDP accountant in TensorFlow Privacy.

[3] Source code available at https://github.com/t3humphries/non-IID-MIA

B. Cluster-based Dependencies

We begin by studying to what extent data dependencies can affect MIAs. To do this, we artificially create two sets D_1 and D_2 that have samples whose features lie in different regions of the feature space. A similar idea was explored in previous work [17] by adding spurious perturbations to the training data that caused a large discrepancy between training and testing accuracy. Instead of modifying existing attributes, we take a real-world dataset and run a k-means clustering algorithm with $k = 2$ to find clusters of samples that lie in different regions of the feature space. We run the clustering algorithm independently for all the samples of each class, and add all samples in cluster i to the set D_i ($i \in \{1, 2\}$). Intuitively, D_1 and D_2 represent mixture components where the characteristics used to distinguish the classes are significantly different.

We use `adult` with $n = 10\,000$ and `compas` with $n = 2\,000$ in this experiment. We evaluate what happens when we use D_1 to sample the members and D_2 for the non-members, and vice-versa. We use leftover samples in `adult` (up to n per cluster) to train the shadow models for the shadow model attack (see Table II). We do not evaluate the shadow model attack for the `compas` dataset, since we do not have enough samples to provide the adversary.

We run the clustering algorithm once to define D_1 and D_2, and then repeat the creation of member and non-member sets, the training, and the attack evaluation 50 times. For each of these experiments, we also run the IID counterpart as explained above, i.e., by randomly mixing the member and non-member sets and repeating the training and attack evaluation.

Figure 2 shows the average advantage of the average threshold attack, optimal threshold attack, and shadow model attack (shaded areas are the 95% confidence intervals for the mean) versus the privacy level ϵ. We only show the results for the case when members are sampled from D_1 and non-members from D_2 (the results for the opposite scenario were very similar). For the value $\epsilon = \infty$, we simply use non-private SGD. For reference, we also plot our membership advantage bound (5), which is only guaranteed to hold in the IID scenario. The rightmost plot includes the average classification accuracy of the target model over the member and non-member sets.

We see that these cluster-based dependencies significantly increase the adversary's performance compared to the IID case. In general, we observe that the attack performance increases as the privacy level decreases (higher ϵ). For $\epsilon \geq 1$ we see that the attack performs alarmingly well, close to perfect advantage. Our results also confirm that the advantage bound does not hold in non-IID cases.

The classification accuracy gives insight into the performance of the threshold attacks. Recall that these attacks classify all points with logloss *below* a threshold as members. The high discrepancy between the loss of members and non-member samples makes it easier for these attacks to identify their membership correctly. Another interesting observation is that, in Figure 2, the classification accuracy remains constant after $\epsilon > 1$ but the non-IID advantage of the threshold attack further increases when $\epsilon = \infty$. This corroborates the observation of Yeom et al. [2] that overfitting is not the only factor at play in the success of the attack.

In Figure 3, we show the results of the optimal threshold attack for the `compas` dataset. These results confirm our findings in the `adult` dataset, and even reach an almost-perfect advantage (≈ 1) for $\epsilon = 10$. To summarize, we have shown that, using off-the-shelf attacks and datasets, certain dependencies among members/non-members can significantly increase the performance of MIAs. The advantage we observe not only crosses the DP bound for the IID case, but can reach the maximum value of 1 in the worst case.

C. Dependencies Caused by an Attribute Bias

We have seen that the cluster-based dependencies can yield extremely high levels of membership advantage. However, a dependence this strong might not occur in practice. In this section, we consider a more realistic and interpretable dependency, based on an attribute in the dataset, that we call an *attribute* bias. We build two experiments using the `adult` dataset: one using the gender attribute (there are only two genders in this dataset, `'Male'` and `'Female'`), and another one using the education level (we consider that samples with at least `'some-college'` education are "high education", and all the others are "low education"). For each experiment, we use member and non-member sets with $n = 10\,000$ samples, and a bias value $p \in [0, 1]$ representing the proportion of the training set that has the particular attribute value.

We generate the sets D_1 and D_2 as follows. We use the gender attribute as an example, but the set generation works similarly for the education attribute. First, we generate D_1 by taking $\lceil p \cdot n \rceil$ samples at random from the dataset with the gender attribute `'Male'`, and the remaining samples have the `'Female'` attribute. We build D_2 by taking an even number of samples with each gender attribute value (i.e., $5\,000$ `'Male'` and $5\,000$ `'Female'`). We delete the gender attribute from all samples after generating the sets, so that the adversary cannot use the gender directly in identifying members and non-members. As before, we use remaining data samples to train the shadow models for the shadow model attack, ensuring we do not give the adversary more than n samples per attribute value. Since our goal here is to investigate training set bias, we train with the biased set (D_1) and vary the bias level p. The non-members (D_2) have an equal number of samples from each attribute value. We repeat the set generation, training, and evaluation process 50 times for each of the values of the bias p that we test.

Figure 4 shows the average advantage of the average threshold attack, optimal threshold attack, and shadow model attack for the gender attribute bias, versus the privacy level ϵ. Each colour represents the performance for a different training set bias, as shown in the legend. We show only some of the values of the bias p for the model's classification accuracy, for readability.

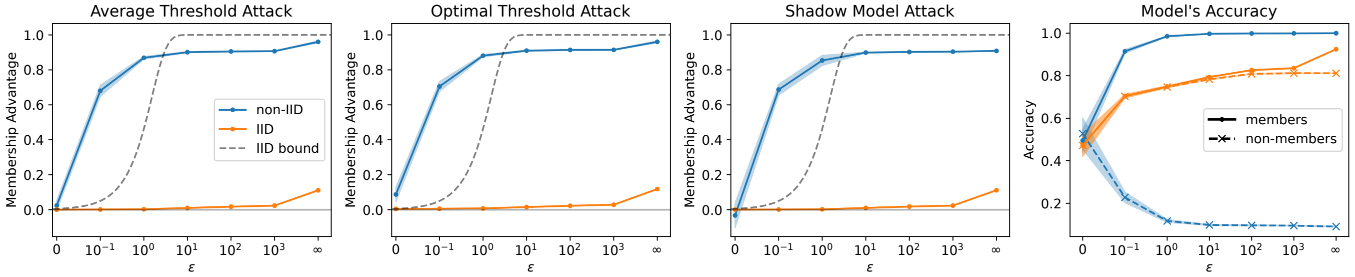

Fig. 2: Results in `adult` dataset with cluster-based dependencies.

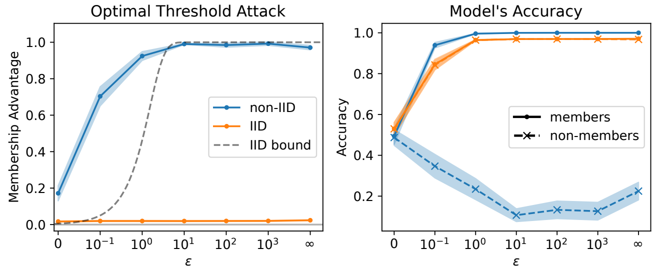

Fig. 3: Results in `compas` dataset with cluster-based dependencies.

We note that, when $p = 0.5$, both the member and non-member sets have the same number of samples from each attribute value, so this can be considered an instantiation of the IID experiment. Training bias ($p \neq 0.5$) can significantly increase the membership advantage, and can even have more impact than the privacy level ϵ. In fact, we see that the adversary achieves similar advantage in a model trained on unbiased data (50% males) without privacy ($\epsilon = \infty$), than in a model trained on highly biased data (0% males) with a high privacy level ($\epsilon = 0.1$). For bias levels $p \geq 0.75$, the membership advantage can be negative. The classification accuracy gives insight into this effect for the average threshold attack. When the training set consists of all male samples ($p = 1.00$), we see that the classification accuracy for non-members is higher than for members. This means that the classification task (salary prediction) is likely easier for predominately female samples than is on male samples (even when trained on males only). This misleads the threshold attack (which classifies all samples with low loss as members) causing it to perform poorly (negative `Adv`) when the training set is male-dominant.

In this experiment, we also consider the classification accuracy of a validation set that has the same bias as the training set. That is, we sample 2 500 data points that are disjoint from the training set with the same attribute bias as the training set. We see that the target models show very similar member and validation set accuracy, especially for low values of ϵ. This suggests that, to a data owner who only has access to biased data, the model appears to generalize well. Even if the data owner runs a tool such as ML Privacy Meter [12] to check their model, they might underestimate the vulnerability of their model.

We plot the results using the education split for the optimal threshold attack only in Figure 5. The results are qualitatively similar to the ones with the gender split. In this case, predicting the income of individuals with low education levels is generally easier than on individuals with high education. Even when training with high-education samples only ($p = 0$), the model's accuracy over non-members (samples with low education) is higher than in members, which causes the MIAs to perform poorly for low values of p. However, when p is large (majority of low-education samples in the training set), we see that the MIAs perform better, breaking the DP bound for the IID case.

D. *Inherent Dataset Dependencies*

We have seen that dependencies play a significant role in the success of membership inference attacks. A natural question to ask is whether such dependencies occur in practice. While there are no public examples of MIAs in the wild, we study naturally occurring dependencies found in off-the-shelf datasets. We do this by either finding different datasets for a similar classification problem or splitting existing datasets based on an attribute that would define a training set in practice, such as the location/institution where the samples were collected.

a) Hospital data: heart condition detection: For our first dataset-based dependency experiment, we use the `heart` datasets, which consist of four different datasets (Cleveland, Hungary, Switzerland, and the VA Long Beach), each containing the samples from one hospital/institution. The classification problem is detecting whether or not a patient has a heart condition (the classification problem is binary). We use all the samples from each database to define the $\{D_i\}_{i\in[4]}$ sets, and evaluate the performance of MIAs when we train with one database and all the samples from the other three databases are non-members.

Figure 6 shows the results when training with each of the hospitals. We see that there is certain property present within each hospital, since the membership advantage when we train with one hospital dataset (non-IID) is significantly higher than when we shuffle all datasets (IID). We note that the advantage in the IID case slightly breaks the bound. We posit that this is because the RDP accountant assumes batches are sampled without replacement and each batch is sampled

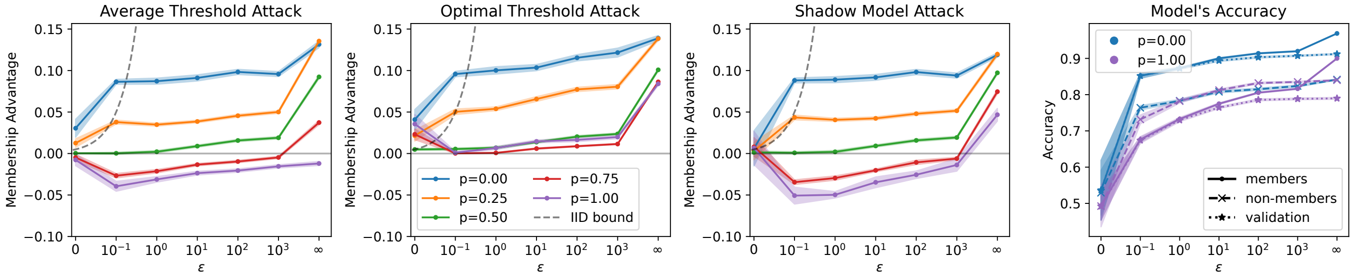

Fig. 4: Results in `adult` dataset with gender bias (p determines the proportion of males in the member set).

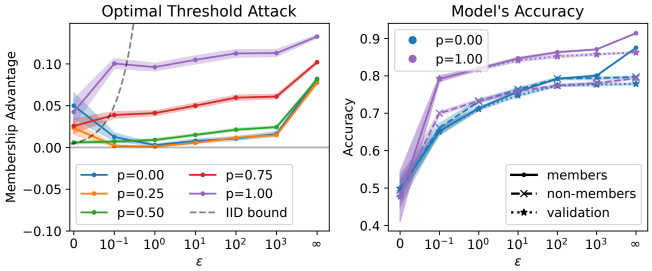

Fig. 5: Results in `adult` dataset with education bias (p determines the proportion of individuals with low education in the member set).

independently. In practice, and in our implementation, batches are non-overlapping. This difference in sampling would be particularly noticeable when the training set is small (as overlaps are more probable), which explains why the advantage in the IID case is slightly above the bound.

b) School data: students proficiency detection: Next, we consider the `students` dataset, which contains data about students from two Portuguese schools (Gabriel Pereira and Mousinho da Silveira), and the classification problem is to predict whether the student passes the course. We use all 423 samples from Gabriel Pereira school (D_1) as the training set, and all 226 samples from Moushinho da Silveira school (D_2) as non-members. Figure 7 shows the results. As before, we see that there is a common property in the students of each school that helps MIAs succeed with higher probability than in the case where we shuffle the data from both schools (IID).

c) Hospital data: large dataset with a multiclass problem: Next, we use the `texas` dataset, which is a larger dataset (350 280 samples) with the 100-class problem proposed by Shokri et al. [1]. Each sample contains an attribute specifying the health region where it was collected. We assign samples from health region number 3 to the set D_1, and assign all other samples to D_2. For each run of this experiment, we randomly sample $n = 10\,000$ samples from D_1 to use as members/training, and use the same number of samples from D_2 as non-members (we restrict the sizes for computational reasons). We give n leftover samples from each set to the adversary to train the shadow model attack. We note that such split could naturally occur in practice (e.g., the local authorities of health region 3 decide to train a machine learning model using data for all the hospitals in that region). Figure 8 shows the results. In this case, the difference between the non-IID case (training with health region 3) and IID case (training with samples from all health regions) is smaller than in previous experiments. We can see that the classification accuracy is only slightly higher for members in the non-IID than in the IID case (the opposite is true for non-members), which explains the more modest advantage increase due to the training set dependencies. However, the relative increase in advantage between the IID and non-IID experiments is significant, especially in the threshold attack when $\epsilon \in [1, 100]$.

d) Census data: dataset curation dependencies: Finally, we consider the dependencies that stem from techniques a researcher or institution follows when curating a dataset. We use both `adult` and `census` datasets, which have been extracted from *the 1994 US Census data*, but were curated into ML training data by different researchers. Specifically, each research group extracted a different number of records: the `adult` dataset consists of 48 842 records extracted using certain binary rules to filter out records, and the `census` dataset contains 299 285 records, but does not specify any filters. Both datasets consider the classification task of predicting if an individual's income is above 50 000 USD (binary classification). We take 10 attributes that both datasets have in common. The set D_1 is the data from `census`, and D_2 contains the samples from `adult` (we remove duplicates so that $D_1 \cap D_2 = \emptyset$). As in the previous experiment, we set $n = 10\,000$ and sample n data points from D_1 as members and n from D_2 as non-members.[4] We further draw n samples from each set to train the shadow model attack. We show the results in Figure 9.

We see that, even though the census data comes from the same source, the curation of these datasets introduces dependencies that can certainly help MIAs. The performance across all attacks is considerably higher for the non-IID setting, and DP-SGD does not seem to have a big effect at preventing this leakage. We note that the noise that DP-SGD adds during training decreases with n; this explains why, in this dataset,

[4] We remark that, since both of these datasets are extracted from the same source (the US Census), there a possibility that multiple unique records can describe the same individual. However, we consider membership inference at the data point level for this experiment.

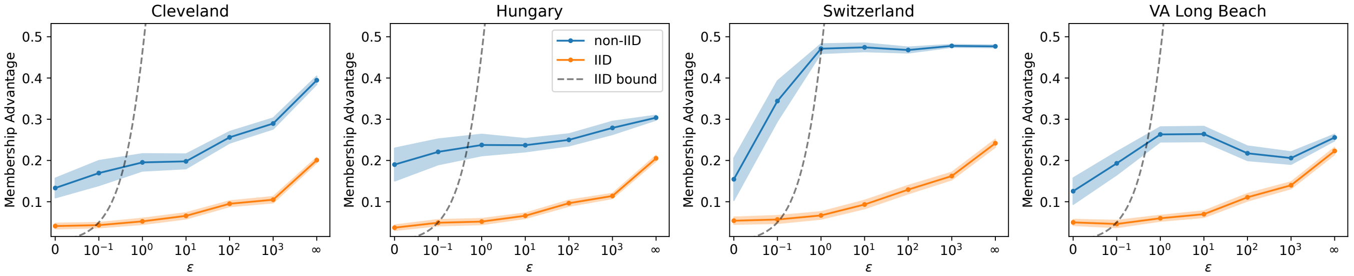

Fig. 6: Performance of the optimal threshold attack in the `heart` dataset when members belong to a particular database (hospital/institution), and non-members are taken from all other databases.

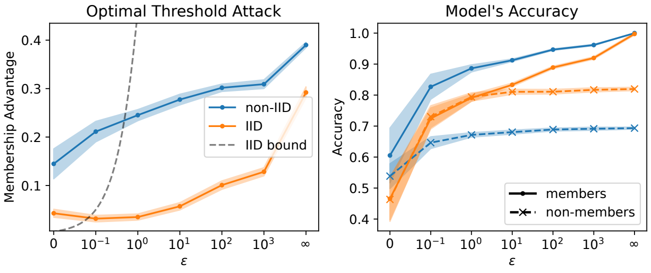

Fig. 7: Results in `students` dataset when members are from the Gabriel Pereira school, and non-members from Mousinho da Silveira school.

even small values of ϵ are not effective at stopping the MIAs in presence of data dependencies.

V. Discussion: Connection to Property Inference

In this paper, we investigate the effects of data dependencies on membership inference attacks. We show that *statistical differences* between members and non-members can improve the performance of existing off-the-shelf MIAs. One can interpret these data dependencies as a *property* that allows distinguishing members from non-members. This property can be tangible, such an existing attribute that already distinguishes members from non-members, or a non-interpretable and arbitrarily complex combination of attributes.

Property Inference Attacks (PIAs) [26]–[28] aim at revealing a property of the training set that the data owner might have wanted to keep secret. In contrast, recall that MIAs aim to infer whether an individual sample is a member of the training set. Given their definitions, MIAs and PIAs are very different attacks, with different goals altogether. However, by defining the data dependencies among members as a property, one could devise a *new MIA* that first runs a PIA to reveal a property of the training set and then checks the presence/absence of this property in the target sample to decide on its membership. This *MIA-via-PIA* draws connections between membership and property inference in the non-IID case: a successful PIA could lead to a successful MIA. In this section, we first discuss the feasibility of an MIA-via-PIA, then we discuss the privacy implications of this attack and, finally, explain the differences between an MIA-via-PIA and the MIAs we evaluated in Section IV.

a) Feasibility of MIA-via-PIAs: The first step in an MIA-via-PIA is finding a property that separates members from non-members. PIAs receive a description of the property to be inferred as an input, e.g., as two complementary hypotheses [26], [27] or as a binary function of the samples' attributes [28]. Generally speaking, finding the right property description that separates members from non-members is challenging. In the theoretical MM membership experiment (Algorithm 3), the adversary has access to $\boldsymbol{\mathcal{D}}$, a joint distribution made up of mixture components. Thus, in this setting, the adversary has a description of a set of properties $[K]$ representing the various mixture components. The adversary could train a meta-classifier [26] that receives a trained model as input, and infers which mixture component $\mathcal{D}_{\hat{k}}$ ($\hat{k} \in [K]$) the training samples followed. However, in practice, this property inference is likely to be unsuccessful for a number of reasons. First, the assumption that the adversary receives $\boldsymbol{\mathcal{D}}$ is very strong and unlikely to occur in practice. Second, the number of mixture components K is likely very large or might not even be a discrete set (e.g., in the case of hospital data, K could be the total number of "possible hospital distributions" that exist, which might not even be finite). This further increases the difficulty of the property inference (estimation of $\hat{k}$).

The second step in an MIA-via-PIA is to use the property $\hat{k}$ to attempt to infer the membership of a sample. For example, given the target sample z, the attacker could check its likelihood of being a member by checking if $\Pr(z|z \sim \mathcal{D}_{\hat{k}}) > \Pr(z|z \sim \mathcal{D}_{k'})$ for all $k' \neq \hat{k}$. Performing this step can be highly difficult for different reasons. First, we note that MIAs are a *statistical game*: even when the adversary correctly estimates the property, it is not always possible to correctly guess the membership of the target sample.[5] Second, the mixture components' support is very likely overlapping such that, given a particular data point z, there might be several mixture components from which the sample might have come from (i.e., $\Pr(z|z \sim \mathcal{D}_{\hat{k}}) \approx \Pr(z|z \sim \mathcal{D}_{k'})$ for at least one $k' \neq k$). In these cases, the MIA-via-PIA

[5] Note that 100% accuracy in membership inference is not possible, since the number of possible datasets typically exceeds the number of possible model parameters, and thus at least two different datasets may result in the same model, by the pigeonhole principle.

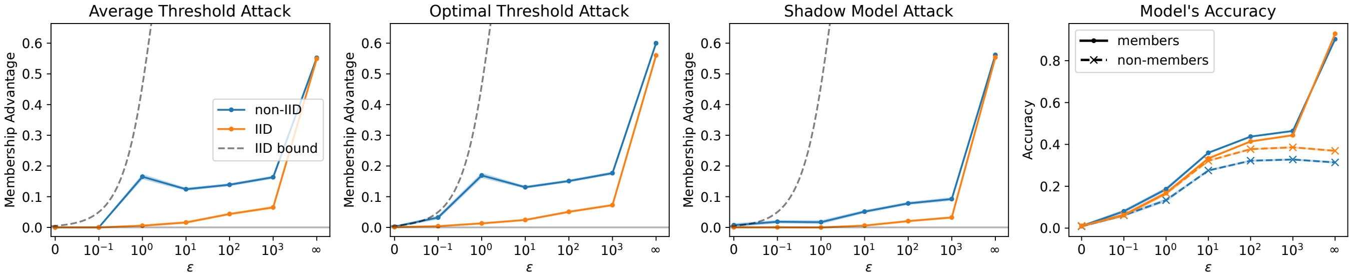

Fig. 8: Results in `texas` dataset when members are from hospitals in region 3, and non-members are from any other region.

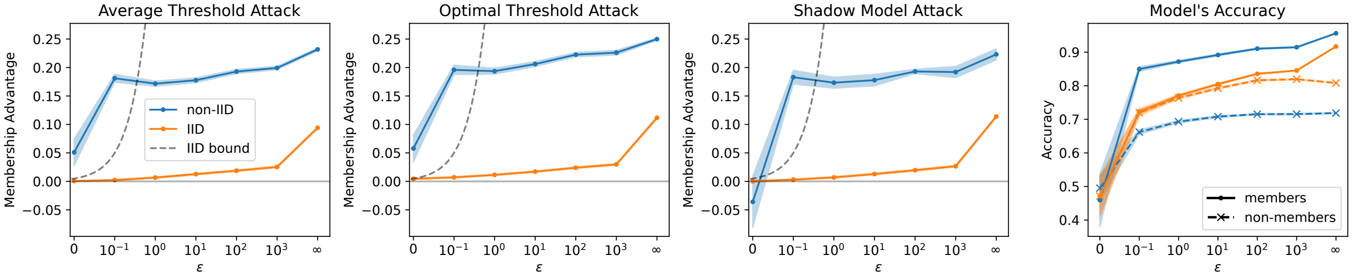

Fig. 9: Results when training a model with `census`, where non-members are from `adult` dataset.

will likely fail. Note that this will often be the case when K is large, which is likely to happen in practice, as we argue above.

b) Privacy Violations: Next, we discuss the meaning of privacy violations under an MIA-via-PIA. While PIAs violate the privacy of the training set as a whole, MIAs violate the privacy of individual members. This questions whether the privacy violation of a successful MIA-via-PIA is an individual or a group privacy violation. The MIA-via-PIA approach decides that any sample z that has the property is a member, regardless of whether or not that sample was truly in the training set. In that case, it would seem that correctly guessing membership is independent of the actual membership of the sample, and thus is not an individual privacy violation but a group privacy violation. However, we want to reiterate that MIA is a *statistical game*. In situations where there is a property that accurately distinguishes members from non-members, an MIA-via-PIA is a smart strategy from a Bayesian inference point of view, that could increase the adversary's odds of winning the statistical membership game. Leveraging the property *does not change the meaning* of the privacy violation from an individual to a group violation. We note that leveraging dependencies to improve the accuracy of statistical attacks is common in many privacy areas, such as location privacy (spatio-temporal correlations between a user's locations can help the adversary's estimation of these user's individual locations [44]), database privacy (correlations between client queries aid finding out the underlying secret keyword of each individual query [45]), etc. These privacy violations are generally considered as individual, and not group, violations.

c) MIAs under data dependencies: While we have argued that an MIA-via-PIA is a valid MIA that can be successful, a PIA is not necessary to build a successful MIA under statistical dependencies. Our theoretical experiments follow prior work [2] and consider an informed adversary with knowledge of $\mathcal{D}$. We have shown, theoretically, that DP bounds do not hold against such adversary (Section III-C1). However, in practice, assuming the adversary knows $\mathcal{D}$ might not be realistic. Our empirical evaluation considers a particular case of the theoretical membership experiment where the adversary simply does not use (and does not know) $\mathcal{D}$. We evaluate popular off-the-shelf attacks and see that these attacks achieve high accuracy under data dependencies, exceeding the DP bounds that hold for the IID scenario. This shows that knowledge of $\mathcal{D}$ (which could enable a MIA-via-PIA) is not necessary to exploit the data dependencies and break the IID DP bounds.

Regardless of whether one uses an MIA-via-PIA in non-IID settings, it is a fact that data dependencies affect the performance of off-the-shelf MIAs (our empirical evaluation demonstrates this) and thus MIAs in the wild will inevitably (and perhaps unintentionally) also be affected by this. We believe that MIA advantage in the wild will be closer to the experiments that we have seen in this paper, and we maintain that future evaluations of MIAs should consider realistic member/non-member splits, since we have seen that random shuffling underestimates the privacy risks of membership inference. Finding techniques that help mitigate the leakage that stems from data dependencies, which has been ignored by past work, is an interesting topic for future work.

VI. Related Work

Chatzikokolakis et al. [46] provide a similar theoretical bound for the case of pure differential privacy. However, since DP-SGD uses the moments' accountant, which gives approximate DP, our bound is more useful in practice. We remark that the goal of this paper is not to be competitive with state-of-the-art attacks in MIA [11], [47]–[52]. Instead, we show the amplification effect of data dependencies on existing, well-known attacks by Shokri et al. [1] and Yeom et al. [2]. Showing that a single MIA breaks the upper bound for a single attack is enough to prove our point.

There have been several works that investigate the performance of MIAs under different assumptions. Nasr et al. [53] provide a comprehensive privacy analysis of MIAs under several white-box conditions and design an inference attack that targets vulnerabilities in the stochastic gradient descent algorithm. Recent work by Jagielski et al. [25] improves MIA success by developing a novel data poisoning attack. Nasr et al. [24] show experimentally that the strong adversary (see Alg. 1) can reach the theoretical DP upper bounds. These works all show stronger attack success than previously reported. However, they achieve this performance by giving the adversary increased (often unrealistic) capabilities while only considering the case of IID data. We instead focus on the weaker black box model of attacker capabilities and vary the data distribution (to be more realistic) to strengthen the attacks.

Kulynych et al. [34] demonstrate that data samples from underrepresented communities are more vulnerable to MIAs. Bagdasaryan et al. [54] find that DP-SGD results in a disproportionate amount of accuracy loss for vulnerable groups. These works complement our findings that MIAs are stronger when the training data has dependencies. Xiong et al. [55] study the risks of non-IID data in the federated learning setting. Hu et al. [56] design a new attack to uncover the source of data points (such as which hospital contributed the data point) in the federated setting. This work complements ours by showing an increase in attack performance when considering clients with different data distributions. Similarly, Gomrokchi et al. [57] study the problem of MIAs against reinforcement learning and find that temporal correlation between training trajectories plays a major role in attack success.

We recall that property inference attacks aim to discover hidden properties in a target classifier's training dataset, such as the proportion of members in the dataset that are students [27]. It is well known that DP does not protect against property inference attacks [27], [28]. As discussed in Section V, our goal is not to conduct a property inference attack or be competitive with state-of-the-art attacks in this field [26]–[28], [58], [59]. Instead, we show the effects of datasets that are non-IID on MIA attacks using properties as one way to introduce a dependency in the datasets.

While our work is the first to investigate the effects of dependent data on MIAs against Neural Networks trained with DP-SGD, other works have studied the limitations of DP under data dependencies, as summarized by Tschantz et al. [20]. Kifer and Machanavajjhala [21] claim that DP mechanisms do not always limit participation inference, and later propose the Pufferfish framework [22], which they use to show that data correlations can cause additional information leakage compared to the IID case. Li et al. [9] explain that DP implies membership privacy only when data is mutually independent. Liu et al. [23] empirically study the effect of data dependencies on inference attacks against the Laplace mechanism. They show that an adversary with knowledge of these dependencies can violate DP bounds that hold in the IID scenario. Although our conclusions are similar (besides the fact that we work in the ML domain), an important difference between this work and ours is that the attacks we evaluate do not use knowledge about these data dependencies. Instead, we show that existing attacks [1], [2], which were not designed to exploit data dependencies, also benefit from them.

VII. Conclusion

We show that current MIA experiments and evaluations hinge on the restrictive assumption that members and non-members are IID data samples. We studied the performance of MIAs when there are statistical dependencies among the training set samples, and found that they pose a far greater threat than previously reported. We show that current state-of-the-art MIAs can achieve near-optimal performance when we introduce artificial dataset dependencies, but even naturally occurring dependencies can increase MIA performance compared to the IID scenario. We show theoretically and empirically that, when members and non-members are not IID, the previous theoretical bounds of DP do not apply. Our work demonstrates that data dependencies should be taken into account when studying MIA performance, as they are a realistic assumption that, if ignored, can lead to a significant underestimation of the privacy risk that MIAs pose.

Acknowledgements

We gratefully acknowledge the support of NSERC for grants RGPIN-05849, RGPIN-03858, IRC-537591, and the Postgraduate Scholarship-Doctoral/ CGS-M programs. We further acknowledge the Royal Bank of Canada, and the Waterloo-Huawei Joint Innovation Laboratory for funding this research. This research was undertaken, in part, thanks to funding from the Canada Research Chairs program. This work was partially funded by an Ontario Graduate Scholarship (OGS). This work benefited from the use of the CrySP RIPPLE Facility at the University of Waterloo.

References

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

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

[3] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate, "Differentially private empirical risk minimization," *Journal of Machine Learning Research*, vol. 12, no. 3, 2011.

[4] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2016, pp. 308–318.

[5] B. K. Beaulieu-Jones, W. Yuan, S. G. Finlayson, and Z. S. Wu, "Privacy-preserving distributed deep learning for clinical data," https://arxiv.org/abs/1812.01484, 2018.

[6] L. Yu, L. Liu, C. Pu, M. E. Gursoy, and S. Truex, "Differentially private model publishing for deep learning," in *IEEE Symposium on Security and Privacy (SP)*, 2019, pp. 332–349.

[7] https://github.com/tensorflow/tensorflow.

[8] https://pytorch.org/.

[9] N. Li, W. Qardaji, D. Su, Y. Wu, and W. Yang, "Membership privacy: a unifying framework for privacy definitions," in *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2013, pp. 889–900.

[10] B. Jayaraman and D. Evans, "Evaluating differentially private machine learning in practice," in *USENIX Security Symposium*, 2019, pp. 1895–1912.

[11] B. Jayaraman, L. Wang, K. Knipmeyer, Q. Gu, and D. Evans, "Revisiting membership inference under realistic assumptions," https://arxiv.org/abs/2005.10881, 2020.

[12] S. K. Murakonda and R. Shokri, "Ml privacy meter: Aiding regulatory compliance by quantifying the privacy risks of machine learning," https://arxiv.org/abs/2007.09339, 2020.

[13] M. M. Kamani, S. Farhang, M. Mahdavi, and J. Z. Wang, "Targeted data-driven regularization for out-of-distribution generalization," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. New York, NY, USA: Association for Computing Machinery, 2020, p. 882–891.

[14] M. Arjovsky, "Out of distribution generalization in machine learning," https://arxiv.org/pdf/2103.02667.pdf, 2021.

[15] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do ImageNet classifiers generalize to ImageNet?" in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 5389–5400.

[16] G. Mårtensson, D. Ferreira, T. Granberg, L. Cavallin, K. Oppedal, A. Padovani, I. Rektorova, L. Bonanni, M. Pardini, M. G. Kramberger, J.-P. Taylor, J. Hort, J. Snædal, J. Kulisevsky, F. Blanc, A. Antonini, P. Mecocci, B. Vellas, M. Tsolaki, I. Kłoszewska, H. Soininen, S. Lovestone, A. Simmons, D. Aarsland, and E. Westman, "The reliability of a deep learning model in clinical out-of-distribution mri data: A multicohort study," *Medical Image Analysis*, vol. 66, p. 101714, 2020.

[17] V. Nagarajan, A. Andreassen, and B. Neyshabur, "Understanding the failure modes of out-of-distribution generalization," https://arxiv.org/pdf/2010.15775.pdf, 2021.

[18] S. Sagawa, A. Raghunathan, P. W. Koh, and P. Liang, "An investigation of why overparameterization exacerbates spurious correlations," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 13–18 Jul 2020, pp. 8346–8356.

[19] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang, "Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization," https://arxiv.org/pdf/1911.08731.pdf, 2020.

[20] M. C. Tschantz, S. Sen, and A. Datta, "SoK: Differential Privacy as a Causal Property," in *IEEE Symposium on Security and Privacy (SP)*, 2020, pp. 354–371.

[21] D. Kifer and A. Machanavajjhala, "No free lunch in data privacy," in *ACM SIGMOD International Conference on Management of Data (MOD)*, 2011, pp. 193–204.

[22] ——, "Pufferfish: A framework for mathematical privacy definitions," *ACM Trans. Database Syst.*, vol. 39, no. 1, Jan. 2014. [Online]. Available: https://doi.org/10.1145/2514689

[23] C. Liu, S. Chakraborty, and P. Mittal, "Dependence makes you vulnberable: Differential privacy under dependent tuples," in *Network and Distributed System Security Symposium (NDSS)*, vol. 16, 2016, pp. 21–24.

[24] M. Nasr, S. Song, A. Thakurta, N. Papernot, and N. Carlini, "Adversary instantiation: Lower bounds for differentially private machine learning," https://arxiv.org/abs/2101.04535, 2021.

[25] M. Jagielski, J. Ullman, and A. Oprea, "Auditing differentially private machine learning: How private is private sgd?" https://arxiv.org/abs/2006.07709, 2020.

[26] K. Ganju, Q. Wang, W. Yang, C. A. Gunter, and N. Borisov, "Property inference attacks on fully connected neural networks using permutation invariant representations," in *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2018, pp. 619–633.

[27] G. Ateniese, L. V. Mancini, A. Spognardi, A. Villani, D. Vitali, and G. Felici, "Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers," *International Journal of Security and Networks*, vol. 10, pp. 137–150, 2015.

[28] S. Mahloujifar, E. Ghosh, and M. Chase, "Property inference from poisoning," in *2022 IEEE Symposium on Security and Privacy (SP)*, 2022, pp. 1569–1569.

[29] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography Conference (TCC)*, 2006, pp. 265–284.

[30] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3-4, pp. 211–407, 2014.

[31] P. Kairouz, S. Oh, and P. Viswanath, "The composition theorem for differential privacy," *IEEE Transactions on Information Theory*, vol. 63, no. 6, pp. 4037–4049, 2017.

[32] Ú. Erlingsson, I. Mironov, A. Raghunathan, and S. Song, "That which we call private," https://arxiv.org/abs/1908.03566, 2019.

[33] R. Hall, A. Rinaldo, and L. Wasserman, "Differential privacy for functions and functional data," *Journal of Machine Learning Research*, vol. 14, no. Feb, pp. 703–727, 2013.

[34] B. Kulynych, M. Yaghini, G. Cherubin, M. Veale, and C. Troncoso, "Disparate vulnerability to membership inference attacks," in *Proceedings on Privacy Enhancing Technologies*, 2022, pp. 460–480.

[35] D. Dua and C. Graff, "Uci machine learning repository," http://archive.ics.uci.edu/ml, 2021.

[36] R. Kohavi, "Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid," in *KDD*, vol. 96, 1996, pp. 202–207.

[37] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias," https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, 2016.

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

[39] https://archive.ics.uci.edu/ml/datasets/Heart+Disease.

[40] P. Cortez and A. Silva, "Using data mining to predict secondary school student performance," in *European Concurrent Engineering and Future Business Technology Conference*, 2008.

[41] https://archive.ics.uci.edu/ml/datasets/Census-Income+(KDD).

[42] https://github.com/tensorflow/privacy.

[43] https://github.com/bargavj/EvaluatingDPML.

[44] R. Shokri, G. Theodorakopoulos, and C. Troncoso, "Privacy games along location traces: A game-theoretic framework for optimizing location privacy," *ACM Transactions on Privacy and Security (TOPS)*, vol. 19, no. 4, pp. 1–31, 2016.

[45] S. Oya and F. Kerschbaum, "IHOP: Improved statistical query recovery against searchable symmetric encryption through quadratic optimization," in *USENIX Security Symposium*, 2022, pp. 2407–2424.

[46] K. Chatzikokolakis, G. Cherubin, C. Palamidessi, and C. Troncoso, "The bayes security measure," https://arxiv.org/abs/2011.03396, 2020.

[47] L. Song and P. Mittal, "Systematic evaluation of privacy risks of machine learning models," in *USENIX Security Symposium*, 2021, pp. 2615–2632.

[48] 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.

[49] 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," in *2020 IEEE European Symposium on Security and Privacy (EuroS&P)*, 2020, pp. 521–534.

[50] L. Watson, C. Guo, G. Cormode, and A. Sablayrolles, "On the importance of difficulty calibration in membership inference attacks," https://arxiv.org/abs/2111.08440, 2021.

[51] J. Ye, A. Maddi, S. K. Murakonda, V. Bindschaedler, and R. Shokri, "Enhanced membership inference attacks against machine learning models," https://arxiv.org/abs/2111.09679, 2021.

[52] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramer, "Membership inference attacks from first principles," in *2022 IEEE Symposium on Security and Privacy (SP)*, 2022, pp. 1897–1914.

[53] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning," in *IEEE Symposium on Security and Privacy (SP)*, 2019, pp. 739–753.

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

[55] Z. Xiong, Z. Cai, D. Takabi, and W. Li, "Privacy threat and defense for federated learning with non-i.i.d. data in aiot," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 2, pp. 1310–1321, 2022.

[56] H. Hu, Z. Salcic, L. Sun, G. Dobbie, and X. Zhang, "Source inference attacks in federated learning," https://arxiv.org/abs/2109.05659, 2021.

[57] M. Gomrokchi, S. Amin, H. Aboutalebi, A. Wong, and D. Precup, "Where did you learn that from? surprising effectiveness of membership inference attacks against temporally correlated data in deep reinforcement learning," https://arxiv.org/abs/2109.03975, 2021.

[58] W. Zhang, S. Tople, and O. Ohrimenko, "Leakage of dataset properties in {Multi-Party} machine learning," in *USENIX Security Symposium*, 2021, pp. 2687–2704.

[59] M. Parisot, D. Spagnuelo *et al.*, "Property inference attacks on convolutional neural networks: Influence and implications of target model's complexity," in *18th International Conference on Security and Cryptography, SECRYPT 2021*, 2021, pp. 715–721.

Appendix

A. Proof of Theorem III.4

We prove that the bound in Theorem III.1 holds when members and non-members are *statistically exchangeable*. Then, we use this to show that this bound holds in Yeom et al.'s [2] membership experiment ($\texttt{Exp}_{\texttt{IID}}$), even when members and non-members are sampled without replacement.

Definition A.1 (Statistical Exchangeability in MIAs)**.** A joint distribution $\boldsymbol{\mathcal{D}}$ for n member samples $\{z_1, \ldots, z_n\}$ and a non-member sample z_{n+1} meets statistical exchangeability if the joint distribution of the samples $\{z_1, \ldots, z_{n+1}\} \sim \boldsymbol{\mathcal{D}}$ does not depend on the sample order in the sequence. Formally, for any permutation $\sigma : [n+1] \to [n+1]$,

$$\Pr(z_1, \ldots, z_{n+1}) = \Pr(z_{\sigma(1)}, \ldots, z_{\sigma(n+1)}). \tag{6}$$

We define a new membership experiment that generalizes $\texttt{Exp}_{\texttt{IID}}$ to the case where members and non-members are statistically exchangeable:

Algorithm 4 Exchangeable Membership Experiment

```
1: procedure $\texttt{Exp}_{\texttt{EX}}(\texttt{Att}, A, n, \boldsymbol{\mathcal{D}})$        ▷ $\boldsymbol{\mathcal{D}}$ exchangeable.
2:     Sample $\{z_1, z_2, \ldots, z_n, z'\} \sim \boldsymbol{\mathcal{D}}$;
3:     Choose $i \sim [n]$ uniformly at random;
4:     Set $z = z_i$ and $\tilde{S} = \{z_k\}_{k \neq i}$;
5:     Choose $b \sim \{0, 1\}$ uniformly at random;
6:     if $b = 0$ then
7:         Train $a = A(\tilde{S} \cup \{z\})$;
8:     else
9:         Train $a = A(\tilde{S} \cup \{z'\})$;
10:    Return 1 if $\texttt{Att}(z, a, n, A, \boldsymbol{\mathcal{D}}) = b$; else 0.
```

We first prove the following result:

Theorem A.1 (New Membership Advantage Bound)**.** *Let A be an (ϵ,δ)-DP learning algorithm. Then, for all attacks $\texttt{Att}$, training set sizes n, and statistically exchangeable joint distributions $\boldsymbol{\mathcal{D}}$, the membership advantage in $\texttt{Exp}_{\texttt{EX}}$ satisfies*

$$\mathit{Adv}(\texttt{Att}, A, n, \boldsymbol{\mathcal{D}}) \leq (e^\epsilon - 1 + 2\delta)/(e^\epsilon + 1)\,. \tag{7}$$

Proof. First, consider a variation of $\texttt{Exp}_{\texttt{EX}}$ that replaces the call to the attack $\texttt{Att}(z, a, n, A, \mathcal{D})$ in $\texttt{Exp}_{\texttt{EX}}$ (Alg. 4, line 10) with a more informed attack that receives both z and z', as well as $\tilde{S} \doteq S \setminus \{z\}$. We call this new experiment $\texttt{Exp}'_{\texttt{EX}}$. We write the call to the attack as $\texttt{Att}(a, z, z', \tilde{S}, A)$: we disregard n as an input, since it can be inferred from $\tilde{S}$, and $\boldsymbol{\mathcal{D}}$, since the adversary receives all variables sampled from it and, due to the statistical exchangeability property, $\Pr(\tilde{S}, z, z') = \Pr(\tilde{S}, z', z)$, and thus $\boldsymbol{\mathcal{D}}$ does not provide any information that helps distinguish a member from a non-member. Let $\texttt{Adv}_{\texttt{EX}}$ (resp. $\texttt{Adv}'_{\texttt{EX}}$) be the maximum advantage of any attack in $\texttt{Exp}_{\texttt{EX}}$ (resp. $\texttt{Exp}'_{\texttt{EX}}$). Notice that, since the attack in $\texttt{Exp}'_{\texttt{EX}}$ receives strictly more information than the attack in $\texttt{Exp}_{\texttt{EX}}$, then $\texttt{Adv}_{\texttt{EX}} \leq \texttt{Adv}'_{\texttt{EX}}$.

Finally, notice that the process between lines 5 and 10 in Algorithm 4 (resp. $\texttt{Exp}'_{\texttt{EX}}$) is exactly the same as $\texttt{Exp}_{\texttt{STR}}$ in Algorithm 1. In other words, we can rewrite $\texttt{Exp}'_{\texttt{EX}}$ as a new experiment, $\texttt{Exp}''_{\texttt{EX}}$, shown in Algorithm 5. We have proven that the membership advantage of $\texttt{Exp}_{\texttt{STR}}$ is upper-bounded by (5), regardless of how z, z', and $\tilde{S}$ were generated. Therefore, the same bound holds for $\texttt{Exp}''_{\texttt{EX}}$. Since $\texttt{Exp}''_{\texttt{EX}}$ is equivalent to $\texttt{Exp}'_{\texttt{EX}}$ and $\texttt{Adv}_{\texttt{EX}} \leq \texttt{Adv}'_{\texttt{EX}}$, it holds that our bound in (5) also holds for $\texttt{Exp}_{\texttt{EX}}$. □

We note that $\texttt{Exp}_{\texttt{IID}}$ can be written as $\texttt{Exp}_{\texttt{EX}}$ for the particular case where $\boldsymbol{\mathcal{D}} = \mathcal{D}^n$. This proves Theorem III.4. The variation of $\texttt{Exp}_{\texttt{IID}}$ that samples without replacement also results in a joint distribution $\boldsymbol{\mathcal{D}}$ that satisfies statistical exchangeability, and thus the bound also holds in this case. Finally, we note that the joint distribution $\boldsymbol{\mathcal{D}}$ in the mixture model experiment $\texttt{Exp}_{\texttt{MM}}$ does *not* satisfy the statistical exchangeability property, and thus we cannot prove the bound holds following this approach (in Section III-C we show why such a bound does not hold generally in the non-IID case).

Algorithm 5 Exchangeable Membership Experiment (v2)

```
procedure $\texttt{Exp}''_{\texttt{EX}}(\texttt{Att}, A, n, \boldsymbol{\mathcal{D}})$        ▷ $\boldsymbol{\mathcal{D}}$ exchangeable.
    Sample $\{z_1, z_2, \ldots, z_n, z'\} \sim \boldsymbol{\mathcal{D}}$;
    Choose $i \sim [n]$ uniformly at random;
    Set $z = z_i$ and $\tilde{S} = \{z_k\}_{k \neq i}$;
    Return $\texttt{Exp}_{\texttt{STR}}(\texttt{Att}^*, A, \tilde{S}, z, z')$.
```

B. Tightest Bound

We prove that our new bound ($(e^\epsilon - 1 + 2\delta)/(e^\epsilon + 1)$) is tighter than the bound by Yeom et al. [2] ($e^\epsilon - 1$) and Erlingsson et al. [32] ($1 - e^{-\epsilon}(1 - \delta)$). First, we note that Yeom et al's bound assumes $\delta = 0$. In that case, Erlingsson et al.'s bound is clearly tighter: $1 - e^{-\epsilon} = (e^\epsilon - 1)/(e^\epsilon) \leq e^\epsilon - 1$.

Thus, to prove our bound is the tightest of the three, it suffices to prove it is tighter than Erlingsson et al.'s bound, i.e.,

$$\frac{e^\epsilon - 1 + 2\delta}{e^\epsilon + 1} \leq 1 - e^{-\epsilon}(1 - \delta)$$

To do this, we simply compute the difference between the left and right terms of these inequalities, and perform basic arithmetic operations to show the difference cannot be positive:

$$\frac{e^\epsilon - 1 + 2\delta}{e^\epsilon + 1} - \left(1 - e^{-\epsilon}(1 - \delta)\right) = \frac{(\delta - 1)(e^{-\epsilon} + 1)}{e^\epsilon + 1} \leq 0$$

The last inequality comes from the fact that the terms $(e^{-\epsilon} + 1)$ and $(e^\epsilon + 1)$ are strictly positive, while $(\delta - 1)$ is non-positive (since $\delta \leq 1$). □