EasyChair Preprint

№ 8027

Two-Stage High Precision Membership Inference Attack

Shi Chen and Yubin Zhong

May 22, 2022

Two-stage High Precision Membership Inference Attack

Shi Chen and Yubin Zhong(✉)

Guangzhou University, Guangzhou 510006, China
Zhong_yb@163.com

Abstract. Most membership inference attacks (MIA) identify training set records by observing the particular behavior of machine learning models on training data, but these methods based only on overfitting are difficult to achieve high precision. Even though recent difficulty calibration techniques have alleviated this problem, calibrated attacks can still only identify a smaller number of memberships with high precision. In this work, we rethink the value of overfitting for MIA and we argue that overfitting can provide clear signals of non-membership to the adversary. In scenarios where the cost of an attack is high, such signals can prevent the adversary from launching unnecessary attacks. We propose a simple and efficient two-stage high-precision MIA that uses an overfitting-based attack to perform "membership exclusion" before performing the MIA. We show that this two-stage attack can significantly increase the number of identified members while guaranteeing high precision.

Keywords: Machine Learning · Membership Inference · Overfitting.

1 Introduction

Machine learning methods rely on large amounts of data for training, and when training data contains sensitive information, one concern is whether the model will reveal private information about the training data. Unfortunately, recent research[3,18,19,21,22] has shown that attackers can gain access to models to steal sensitive information from datasets. One of the more representative of existing privacy attacks is the membership inference attack (MIA)[19], where the adversary aims to infer whether a record exists in the training set of the target model.

However, Rezaei et al.[16] discovered that previous MIAs tend to predict non-member samples as member samples and suffer from a high false positive rate (FPR). The previous attacks [18,19,20]are not suitable in attack scenarios where the false positive cost is high[13,14]. Some recent work [1,17,22]has considered the use of difficulty calibration to mitigate the high FPR problem, and this approach has also been shown to achieve high precision attacks on models that generalize well[13,14]. However, identifying more members with high precision requirements is still a difficult task. For example, for CNNs that achieve 98.61% accuracy at MNIST, the C-Conf attack proposed by Watson et al.[22] is able

to identify only 21 out of 10000 members with 100% precision. Although it is already a big improvement compared to the attack without considering the difficulty calibration, we still want to know how to identify more members with a high precision.

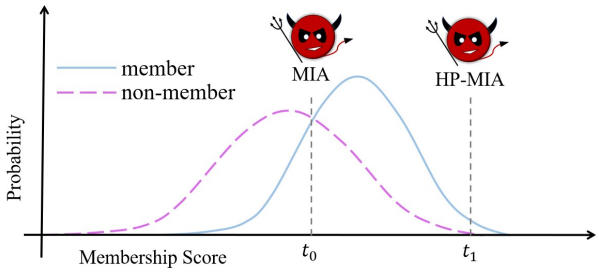

Fig. 1. The difference between HP-MIA and MIA.The aim of MIA is to find a suitable threshold that achieves the highest accuracy in distinguishing members from non-members, while the goal of HP-MIA is to find the threshold that identifies members with a high prediction rate.

Overfitting-based MIAs, such as the Loss attack by Yeom et al.[24] and the confidence attack by Salem et al.[18] suffer from serious high FPR problems. Although the relationship between overfitting and member privacy leakage has received extensive attention[5,12,13,24], these works do not discuss how to better exploit the overfitting tendency of machine learning models for black-box MIA design.

Deep learning models have strong learning ability, even if the samples in the training set are contaminated, their Loss tend to be not too high[25]. Suppose the target model is overfitting so severely that the Loss of all members in the training set is less than a small real number ϵ, then we just need to mark all records with a Loss greater than ϵ as non-members and we can achieve a "membership exclusiontechnique" with 100% precision. However, easy-to-predict non-members may have very low Loss[22], and Hintersdorf et al.[6] also found that neural networks are prone to overconfidence in records outside the training data.Therefore, we argue that **overfitting may not provide a valid basis for membership inference directly, but can provide an explicit non-membership signal to the adversary.**

In this paper, we focus on how to design a high precision MIA that can identify more members, and we name this type of attack as High Precision MIA (HP-MIA). As shown in the figure, the goal of the previous MIA was to obtain an optimal threshold for distinguishing members from non-members, and this threshold corresponded to a high accuracy rate. In contrast, the goal of HP-MIA is to determine a threshold such that most of the identified samples are members. We treat the construction of a high prediction rate MIA as an optimization problem with constraints, setting the number of members identified

as the optimization objective and the high precision as the constraint. Instead of using overfitting-based attacks for direct membership inference, we use them as a "membership exclusiontechnique", specifically, we propose a simple and effective two-stage HP-MIA. We first exclude non-membership samples from the target dataset by traditional overfitting-based attack[24], and then use calibrated attack[22] to identify the true members.

Contribution We summarize our contributions and key finding as follows :

- We propose a new perspective on designing MIAs based on overfitting. Specifically, we find that overfitting provides the adversary with a far more reliable non-membership signal than the membership signal. In scenarios where the cost of attack is high, such signals can help the adversary avoid unnecessary losses. We propose to use the overfitting-based membership exclusion technique to assist the adversary in achieving high precision inference.
- Based on our new understanding of overfitting, we propose a Two-stage High Precision MIA(HP-MIA). we deploy our attack on various datasets and on various models, and our evaluation results show that Two-stage HP-MIA is able to identify more memberships than other attacks while guaranteeing high precision. For a victim model that achieves 98.59% on the MNIST dataset, Two-stage HP-MIA perfectly identified 258 members, which is more than 10 times the number of other methods, even though these methods already use difficulty calibration techniques.

2 Background

In this section, we give the definition of membership inference attack (MIA) and introduce the threshold-based MIA in Section 2.1. Then, we describe difficulty calibration techniques used to mitigate the high FPR problem of MIA in Section 2.2.

2.1 Membership Inference Attacks

Definition 1 (Membership Inference Attacks[19]) *Given a machine learning model h that has completed training on the training set $D \sim Q^n$, and a target sample $z = (x, y) \sim Q$, where Q denotes the probability distribution of the data points. The membership inference attack can be formalized as a binary classifier:*

$$\mathcal{A} : Z \times H \longrightarrow \{0, 1\} \tag{1}$$

where 0 means z does not belong to the training set D, otherwise it is 1. Z denotes the set of all samples $z \sim Q$ and H denotes the set of all classifiers trained on examples from a data distribution Q.

Most of the previous works[14,18,19,20,22] assumed that the adversary has only black-box access to the target model and infer membership information

through the posterior probability vector. In addition to this, the adversary uses shadow training techniques to train a shadow model to mimic the behavior of the target model. Shokri et al.[19] use neural networks to construct the attack model and train it based on the inputs and outputs of the shadow model.

A common binary classification in membership inference problems is the threshold model, which distinguishes members from non-members by computing a particular score $s(h, z)$ and setting a threshold t.

$$\mathcal{A}_{score}(h, z, s, t) = I\left[s(h, z) > t\right] \tag{2}$$

where the indicator function $I\left[x\right]$ equals to 1 if x is true and 0 otherwise.

2.2 Calibrated Membership Inference Attacks

For non-member examples with low prediction difficulty, the target model may exhibit a high degree of confidence. Most of the early MIA attacks[18,19,24] implicitly assumed that the prediction difficulty of members and non-members is the same, which is believed to be the main reason for the high FPR problem. To address this problem, Watson et al.[22] proposed the difficulty calibration technique.

Specifically, we assume that the adversary has a reference dataset D_{ref} with the same distribution as the training set of the target model. He trains some reference models on the D_{ref} before performing the attack, then, he determines whether the example is a membership by comparing the membership score of the target example on the target model and reference models. Formally, we define the calibrated score as:

$$s_{cal}(h, z) = s(h, z) - E_{g \leftarrow T(D_{ref})}\left[s(g, z)\right] \tag{3}$$

where T denotes the randomized training algorithm. The calibrated attack is performed by setting a threshold on the calibrated score.

The goal of this calibration technique is to eliminate the interference of the example's own characteristics with the MIA, similar approaches have been used in the work of Sablayrolles et al.[17] and Carlini et al.[3]

3 Methodology

3.1 Threat Model

In contrast to the definition in Section 2.1, for HPMIA, the adversary is more interested in the precision of the attack, specifically, we consider the following HP-MIA game:

Definition 2 (HP-MIA game $G(\mathcal{Q}, \mathcal{A}, T, n)$) *Let $\mathcal{Q}$ be a distribution over data points, $\mathcal{A}$ be an attack, T be a randomized training algorithm, n be a positive integer. The game proceeds as follows:*

1. *The challenger chooses a secret bit* $b \leftarrow \{0,1\}$ *uniformly at random, and samples a training dataset* $D \sim \mathcal{Q}^n$.
2. *If* $b = 1$, *the challenger randomly selects a record* z *in the training set* D. *Otherwise, the challenger samples a record* z *from the distribution* $\mathcal{Q}$.
3. *The challenger trains a model* $h \leftarrow T(D)$ *on* D *and sends* h *and* z *to the adversary.*
4. *The adversary tries to infer the secret bit as* b', *and performs an attack only if* $b' = 1$.
5. $$G(\mathcal{Q}, \mathcal{A}, T, n) = \begin{cases} 1 & b' = 1 \text{ and } b = 1 \\ 0 & b' = 1 \text{ and } b = 0 \end{cases}$$

In this paper, we assume that the adversary has only black-box access to the target model and has a shadow dataset D_{shadow} sampled from the same distribution as the training set of the target model, through which the adversary can train the shadow model to mimic the behavior of the target model. Given a target dataset $D_{target}, D_{shadow} \bigcap D_{target} = \emptyset$,the adversary aims to identify members of the target model from the target dataset D_{target} with as high a precision as possible. Our assumptions about the adversary's knowledge are similar to most prior work[14,18,19,20].

3.2 High-Precision Membership Inference Attack

For the above attack setup, the two important metrics we focus on are the precision and recall of the attack. We give a formal definition of HP-MIA in terms of a constrained optimization problem:

Definition 3 (High-Precisionα Membership Inference Attack) *Given a target model* h, *a target dataset* D_{target}, $\alpha \in [0,1]$ *is a precision constraint value. We call* $\hat{\mathcal{A}}_{score}$ *is a High-Precision*α *Membership Inference Attack (HP*α*-MIA) if* $\hat{\mathcal{A}}_{score}$ *satisfies:*

$$\hat{\mathcal{A}}_{score} = \underset{\mathcal{A}_{score}}{argmax}\, R(\mathcal{A}_{score}, D_{target}) \qquad s.t. P(\mathcal{A}_{score}, D_{target}) \geqslant \alpha \tag{4}$$

In practice, we do not have access to the training set of the target model and thus cannot solve the optimization problem on the real dataset D_{target}. According to our assumptions on adversary knowledge in Section 3.1, we can construct the member dataset D^{in}_{shadow} and non-member dataset D^{out}_{shadow} of the shadow model for supervised training of the attack model. As for the score-based attacks, the process of constructing a construction is actually finding an optimal threshold,and we choose the appropriate threshold in the following set U:

$$U(h, s, D_{shadow}) = \left\{ u_i = \frac{s(h, z_i) + s(h, z_{i+1})}{2} : z_i \in D_{shadow} \right\} \tag{5}$$

where $s(h, z_i) \leqslant s(h, z_{i+1}), i = 1, 2, ..., m$,and m is the amount of members of the shadow data set. Specifically, we iterate through all elements in U and calculate

the attack precision and recall corresponding to each element, select the subset U' that satisfies precision $\geqslant \alpha$, and return the element in U' that corresponds to the largest recall. The construction process of the attack model is shown in Algorithm 1.

3.3 Two-stage HPMIA

We rethink the significance of overfitting for black-box membership inference attacks. Even though overfitting-based attacks cannot achieve high precision membership inference, they are powerful "membership exclusiontechnique". High-precision member inference is often used in scenarios where the cost of attack is high, so a reliable exclusion technique can help the adversary avoid unnecessary losses.

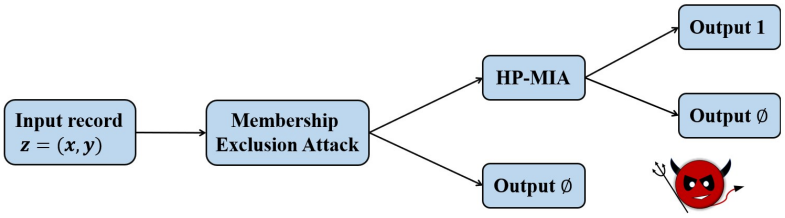

Fig. 2. Two-stage HP-MIA overview

We consider a simple two-stage attack. As shown in the Figure 2, we first exclude the non-membership samples from the target dataset D_{target} by high-precision membership exclusion attacks, and then use HP-MIA on the remaining data. We set two thresholds for the two phases of the attack, and we construct the optimal attack by adjusting the accuracy constraint value of membership exclusion attacks. The details are given in Algorithm 2. In this paper, we use Loss[24] as the membership score to construct membership exclusion attacks and calibrated Loss score[22] for the second stage of membership inference.

4 Experiments

In this section, we first show the experimental setup in Section 4.1, including dataset, target model architecture, training setup, and evaluation metrics of the attack model. Then, we evaluate our attack and compare it with the previous MIA in Section 4.2. Finally, we analyze the impact of some factors on the attack performance in Section 4.3.

Table 1. Accuracy of the target model

	MNIST	F-MNIST	CIFAR10	Purchase100	Texas100
Model	CNN	CNN	AlexNet	MLP	MLP
Train_Acc	100%	100%	100%	100%	100%
Test_Acc	98.59%	88.74%	70.61%	80.59%	43.89%

4.1 Experimental Setup

Dataset We conducted experiments on several baseline datasets of different complexity: MNIST[11], Fashion-MNIST (F-MNIST)[23], CIFAR10[9], Purchase100[1], and Texas100[2]. We randomly divide each of these datasets into six datasets, two of which are used as the training set D^{in}_{target} and test set D^{out}_{target} for the target model, two of which are used as the training set D^{in}_{shadow} and test set D^{out}_{shadow} for the shadow model, and the remaining two datasets are used as the reference training set for the training of the reference model.

For MNIST and F-MNIST, the target model test set has 10,000 images, and the remaining datasets have 12,000 images each. For CIFAR10, each dataset has 10,000 images. For Purchase100, each dataset has 20,000 records, and for Texas100, each dataset has 10,000 records.

Model architectures The target model for MNIST and Fashion-MNIST is a small CNN with two convolutional layers and a maximum pooling layer, two convolutional layers with 24 and 48 output channels, and a kernel size of 5, followed by a fully connected layer with 100 neurons as the classification head, and we use Tanh as the activation function. For CIFAR10, we use AlexNet[10] as the structure of the target model. For Purchase100 and Texas100, we refer to the work of Song et al.[20] and use a multilayer perceptron (MLP) as the target model with four hidden layers with the number of neurons of 1024, 512, 256, and 128, respectively, and use Tanh as the activation function.

Model tranning We use Adam optimizer[8] to train the target model 200 epochs. For the target models of MNIST, F-MNIST, CIFAR10, Purchase100, and Texas100, we set the learning rates to 0.0005, 0.0005, 0.0001, 0.0002, and 0.0002, respectively, and the corresponding batch sizes to 100, 100, 50, 100, and 100, respectively. Table 1 shows the performance of several target models.

The shadow and reference models are trained using the same training algorithm and hyperparameters as the target model. For each dataset, we train a shadow model on the shadow dataset for attack model construction and 20 reference models on the reference dataset for calculating the calibrated score.

[1] https://www.kaggle.com/c/acquire-valued-shoppers-challenge/data
[2] https://www.dshs.texas.gov/THCIC/Hospitals/Download.shtm

Success metrics We consider an extremely cautious MIA adversary who wants to identify as many members as possible with high precision. We use the following metrics to evaluate the attack performance: number of correctly identified members (TP), recall (Recall) and precision (Pr). Recall and Pr are calculated as follows:

$$\text{Recall} = \frac{\text{TP}}{|D^{in}_{target}|} \qquad \text{Pr} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

where FP denotes the number of non-members identified as members by the attack model.

4.2 Attack Evaluation

Effectiveness of two-stage HP-MIA To highlight the effect of Two-stage attack, we use two calibrated attacks, C-Loss and C-Conf, to compare with our attacks. These two attacks use Loss and Confidence as membership scores, respectively, and use difficulty calibration[22] to remove the effect of example difficulty. It is worth noting that C-Loss can be viewed as a direct HP-MIA without using the membership exclusion technique.

Table 2. Evaluation of various data sets, model structures, and MIA methods, $\alpha = 90\%$

		MNIST	F-MNIST	CIFAR10	Purchase100	Texas100
C-Loss	TP	41	224	100	4	1872
	Recall	0.41%	2.24%	1.00%	0.02%	18.72%
	Pr	91.11%	**92.18%**	**95.24%**	**100%**	88.83%
C-Conf	TP	111	686	1427	46	4087
	Recall	1.11%	6.86%	14.27%	0.23%	40.87%
	Pr	92.50%	88.98%	87.17%	93.88%	89.98%
Two-stage	TP	**590**	**1406**	**3866**	**8283**	**5792**
	Recall	**5.90%**	**14.06%**	**38.66%**	**41.42%**	**57.92%**
	Pr	**92.91%**	89.21%	85.34%	89.32%	**90.02%**

$\alpha = 90\%$

We can only implement HP-MIA that satisfies the precision constraint on the shadow model, so the precision on the target model may be biased, and the bias size depends on how close the shadow model is to the target model. Most of the time, as the precision constraint value rises, the accuracy of the attack on the target model becomes higher. Table 2, Table 3 and Table 4 show the attack performance of the three attacks when α is set to 0.9, 0.94 and 0.98 respectively. The Two-stage attack consistently identifies the most members on all models at various precision constraint settings.

The precision of various attacks on the target model varies under the same precision constraint value setting. We compare the experimental results under

Table 3. Evaluation of various data sets, model structures, and MIA methods, $\alpha = 94\%$

		MNIST	F-MNIST	CIFAR10	Purchase100	Texas100
C-Loss	TP	21	133	75	4	1142
	Recall	0.21%	1.33%	0.75%	0.02%	11.42%
	Pr	**100%**	**96.38%**	**96.15%**	**100%**	91.65%
C-Conf	TP	81	511	796	18	2489
	Recall	0.81%	5.11%	7.96%	0.09%	24.89%
	Pr	93.10%	93.76%	92.67%	94.74%	**94.42%**
Two-stage	TP	**390**	**1104**	**2301**	**3465**	**4022**
	Recall	**3.90%**	**11.04%**	**23.01%**	**17.33%**	**40.22%**
	Pr	97.74%	93.16%	90.84%	93.27%	94.18%

$\alpha = 94\%$

Table 4. Evaluation of various data sets, model structures, and MIA methods, $\alpha = 98\%$

		MNIST	F-MNIST	CIFAR10	Purchase100	Texas100
C-Loss	TP	17	70	40	4	183
	Recall	0.17%	0.70%	0.40%	0.02%	1.83%
	Pr	100%	95.89%	93.02%	100%	**96.32%**
C-Conf	TP	49	195	353	8	678
	Recall	0.49%	1.95%	3.53%	0.04%	6.78%
	Pr	92.45%	97.01%	**96.45%**	**100%**	**99.41%**
Two-stage	TP	**258**	**461**	**1059**	**71**	**1543**
	Recall	**2.58%**	**4.61%**	**10.59%**	**0.36%**	**15.43%**
	Pr	**100%**	**97.26%**	96.27%	95.95%	97.84%

$\alpha = 98\%$

Table 5. Evaluation of various data sets, model structures, and MIA methods,$\alpha = 99\%$

		MNIST	F-MNIST	CIFAR10	Purchase100	Texas100
C-Loss	TP	17	39	2	4	72
	Recall	0.17%	0.39%	0.02%	0.02%	0.72%
	Pr	100%	97.50%	**100%**	100%	96.00%
C-Conf	TP	46	160	207	8	452
	Recall	0.46%	1.60%	2.07%	0.04%	4.52%
	Pr	95.83%	96.97%	97.18%	**100%**	**99.34%**
Two-stage	TP	**223**	**318**	**864**	**51**	**728**
	Recall	**2.23%**	**3.18%**	**8.64%**	**0.26%**	**7.28%**
	Pr	**100%**	**97.78%**	97.08%	98.08%	98.91%

$\alpha = 99\%$

different precision constraint values. For MNIST, Two-stage can identify 258 memberships with 100% accuracy, while other methods fail to identify more than 120 members at various precision constraint settings. For F-MNIST, Two-stage identified 461 memberships with 97.26% when $\alpha = 0.98$, C-Conf identified 689 memberships with only 88.98% precision when $\alpha = 0.9$, and 511 memberships with only 93.76% precision when $\alpha = 0.94$. For CIFAR10, Two-stage identified 1059 memberships with 96.27%. Although C-Conf was able to identify 1427 memberships when $\alpha = 0.9$, the precision was only 87.17%. On the other hand, Two-stage was able to identify 2301 memberships with 90.84% precision when $\alpha = 0.94$. For Purchase100, the precision of Two-stage is lower than the other two attacks, but identifies far more memberships than them. For Texas100, Two-stage consistently identifies more memberships than the other two methods, but with lower precision at $\alpha = 0.94$ and $\alpha = 0.9$.

We believe that Two-stage is suitable for high-precision membership inference tasks that wish to identify a larger number of memberships.

Table 6. Performance of overfitting-based MIA under high-precision constraints

	MNIST	F-MNIST	CIFAR10	Purchase100	Texas100
Loss	56.56%	55.99%	68.00%	73.11%	74.32%
Conf	54.55%	55.99%	68.21%	73.11%	74.32%
Mentr	57.44%	57.22%	68.01%	74.50%	74.65%
$\alpha = 90\%$					

Failure of the direct Overfitting-based MIA MIA without the use of difficulty calibration fails under the requirement of high precision, so we did not compare these methods directly with our attacks in Section 4.1. We construct HP-MIA using three membership scores, Loss[24], Conf[18,19] and Mentr[20], respectively, and Table 6 shows the performance of these attacks on different datasets. Note that Algorithm1 will return a threshold that achieves the maximum accuracy when it finds that it cannot find a threshold that satisfies the accuracy requirement on the shadow model. We find that these attacks are completely unable to achieve the precision we require, even though we only set $\alpha = 0.9$.

Specifically, for a model generalized on the MNIST dataset, the overfitting-based direct attack can only achieve up to 57.44% precision. Even for the simple MLP with only 43.89% accuracy on Texas100, it can only achieve 74.65% precision. In scenarios where the cost of the attack is high, these attacks are completely inappropriate.

4.3 Ablation Experiments

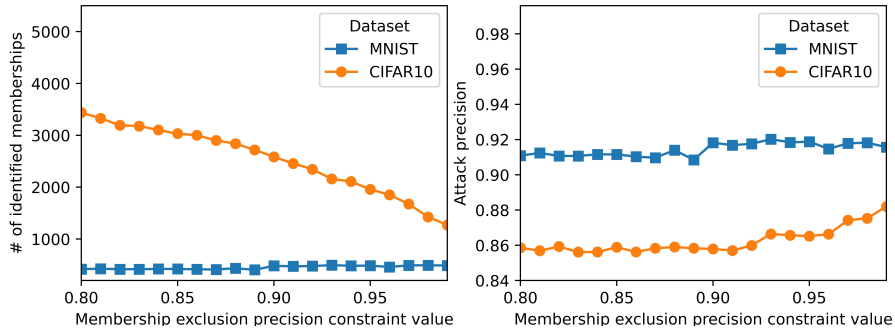

Fig. 3. Effect of precision constraint value for membership exclusion. We found that there is no more general precision constraint value for the membership exclusion technique, and the impact of different different precision constraint values on CIFAR10 and MNIST is not the same.

Attack performance vs. Membership exclusion precision constraint value Compared to other score-based attacks, Two-stage HP-MIA has two thresholds and thus takes more time in building the attack model. Some score-based MIAs provide an empirical threshold, for example, Waston et al.[22] point out that the empirical threshold for calibration attacks is a value only slightly greater than 0. We would like to explore whether we can reduce the computational effort with some experience. The adversary needs to adjust the precision constraint value β of the membership exclusion attack to achieve the most powerful attack when constructing Two-stage HP-MIA, and we would like to know whether β has an empirical value as a reference. We conduct experiments on the MNIST and CIFAR10 datasets to observe the performance of Two-stage HP-MIA when setting different β.

Unfortunately, we find that there is no relatively general precision constraint value for the membership exclusion attack. Figure 3 shows our experimental results, and we find that the value of β has different effects on the two datasets. the number of exposed examples on the MNIST dataset increases slowly with larger β, while the number of identified examples on CIFAR10 decreases rapidly with larger β. High-precision membership inference attacks require capturing more detailed model features and example characteristics, so it is difficult to have a general reference value. In order to construct more robust attacks, it is necessary to spend more time to optimize the thresholds.

Attack performance vs. l_2 regularization l_2 regularization is a relatively simple defense technique for member inference attacks[7,14,19]. We assume that the adversary is unknown to the defense used by the victim, and both the shadow

Table 7. Two-stage experimental results on the target model using regularized defense,the datasets are F-MNIST and CIFAR10

Dataset	λ	Train_Acc	Test_Acc	TP	Pr
F-MNIST	0	100%	88.74%	1406	89.21%
	0.0001	100%	88.77%	283	84.99%
	0.0003	100%	88.16%	111	84.09%
	0.0005	99.96%	88.14%	83	81.37%
	0.0007	99.93%	88.34%	75	81.52%
	0.001	99.68%	87.87%	47	79.66%
	0.005	95.12%	88.50%	0	0
	0.01	89.68%	86.50%	0	0
CIFAR10	0	100%	70.61%	3866	85.34%
	0.0001	99.92%	67.88%	2044	92.57%
	0.0003	99.74%	69.13%	1781	90.77%
	0.0005	100%	68.85%	1531	92.79%
	0.0007	99.88%	70.23%	583	88.74%
	0.001	99.50%	66.59%	505	90.02%
	0.005	98.91%	69.26%	23	85.19%
	0.01	92.26%	62.26%	10	76.92%

model and the reference model are trained using the original algorithm. Table 7 shows the performance of the target model with the regularization technique and the inference effect of the Two-stage attack. As a common method to overcome overfitting, regularization can prevent the leakage of membership privacy to some extent.

In general, the number of memberships that can be inferred by Two-stage decreases significantly as λ grows. It is worth noting that lower levels of l_2 regularization may not reduce the attack precision as well. For CIFAR10, the attack precision at $\lambda < 0.001$ is instead higher than that without the l_2 regularization method.

Attack performance vs. Number of reference models Figure 4 shows the TP and Pr of Two-stage HP-MIA with different number of reference models. we use the datasets MNIST, F-MNIST and CIFAR10. in general, the attack precision receives little effect from the number of reference models, and the difference between the maximum and minimum precision on MNIST, F-MNIST and CIFAR10 are 0.97%, 0.89% and 0.43%. Besides, using fewer reference models may lead to a lower number of identified memberships. The TP of Two-stage HP-MIA using only one reference model is the least on the target model of three datasets.

However, we found that the increase in the number of reference models did not significantly improve the TP except for CIFAR10. For MNIST,the highest number of identified members was for the attack using 8 reference models, with 621, while the attack using 20 reference models identified 590 memberships.

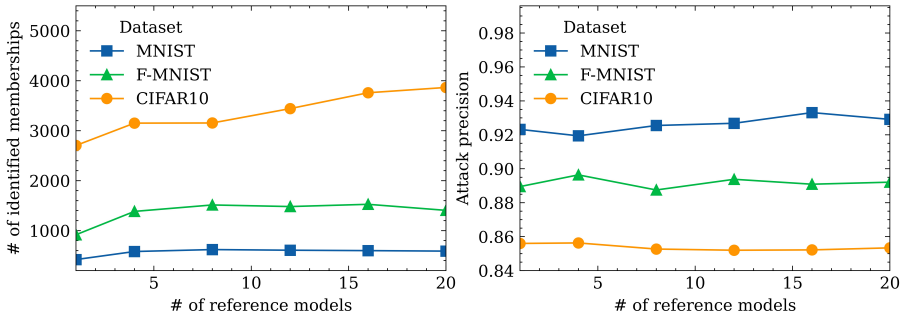

Fig. 4. The TP and Pr of Two-stage HP-MIA with different number of reference models. Attack precision receives little effect from the number of reference models. Using few reference models (e.g., one) may result in a low number of identified memberships.

For F-MNIST, the highest number of identified memberships was for the attack using 16 reference models, with 121 more members identified than when using 20 reference models. We do not recommend training too many reference models for calculating the calibrated score when not planning to spend too much time to deploy the attack.

5 Related Work

Black-box membership inference attacks against machine learning were first proposed by Shokri et al.[19] This attack uses the output predictions of the target model to distinguish between members and non-members of the training dataset. MIA can pose serious privacy risks to individuals when the participation status of members is considered sensitive. Besides, MIA can be used as a basis for more powerful attacks such as training data extraction attacks[2,3].

A more common approach in membership inference attacks is the threshold attack based on membership score. Yeom et al. [24] use Loss as membership score for the attack and discussed the relationship between MIA and generalization error. Salem et al.[18] found that maximum posterior probability and prediction entropy can also be used as membership score. Song et al. [20] proposed modified prediction entropy and set different thresholds for different classes to improve the attacks. Nasr et al. [15] extended MIA to white-box scenarios and found that the gradient norm is a strong signal to distinguish members from non-members. Choquette-Choo et al. [4] used the distance from the sample to the decision boundary as the membership score.This attack does not require access to the posterior probability vector, only the corresponding label.

The previously mentioned attacks are all overfitting-based MIAs because they only exploit the fact that the models have different behaviors on members and non-members and do not consider the characteristics of individual samples. Rezaei et al. [16] found that these overfitting-based attacks cannot achieve high accuracy and low FPR at the same time.

Difficulty calibration is currently known to be a more effective technique to mitigate the high FPR problem, and this approach requires extra training of some reference models. It was originally proposed by Sablayrolles et al. [17] in their analysis of the Bayesian optimal strategy for MIA. They trained two sets of reference models, one set of models was trained using the target records while the other set was not. Then they calculated the mean of the membership score of the target records on the two sets of models separately, and then the mean of the two values was found as the Per-example hardness threshold.Waston et al. [22] considered a simpler approach which uses only a set of reference models that were not trained using the target records to compute the Per-example hardness threshold. Carlini et al. [1] proposed the LiRA attack, which uses a Gaussian function to fit the output of the reference models. Although this attack requires training a large number of reference models, it greatly improves the performance of MIA at low FPR.

6 Conclusion

In this work, we rethink the relationship between overfitting and membership inference attacks and demonstrate that using an overfitting-based approach for membership exclusion can effectively improve the performance of HP-MIA. Our evaluation results show that our attack is able to identify more members while guaranteeing high accuracy compared to other attacks.

We believe that our work in understanding membership privacy is preliminary and the relationship between example characteristics and privacy leakage needs to be further explored. In particular, we would like to know how adversaries should perform effective attacks on easy examples, and how victims and defenders have to work on the defense of hard examples.

Acknowledgments

This work was supported by the Guangzhou University Provincial College Student Innovation Training Program(No. S202111078114).

References

1. Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas F. Terzis, and Florian Tramèr. Membership inference attacks from first principles. *ArXiv*, abs/2112.03570, 2021.
2. Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 267–284, 2019.
3. Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Xiaodong Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *USENIX Security Symposium*, 2021.

4. Christopher A. Choquette-Choo, Florian Tramèr, Nicholas Carlini, and Nicolas Papernot. Label-only membership inference attacks. In *ICML*, 2021.
5. Ganesh Del Grosso, Georg Pichler, Catuscia Palamidessi, and Pablo Piantanida. Bounding information leakage in machine learning. *ArXiv*, abs/2105.03875, 2021.
6. Dominik Hintersdorf, Lukas Struppek, and Kristian Kersting. Do not trust prediction scores for membership inference attacks. *ArXiv*, abs/2111.09076, 2021.
7. Yigitcan Kaya, Sanghyun Hong, and Tudor Dumitras. On the effectiveness of regularization against membership inference attacks. *ArXiv*, abs/2006.05336, 2020.
8. Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.
9. Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.
10. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60:84 – 90, 2012.
11. Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86:2278–2324, 1998.
12. Klas Leino and Matt Fredrikson. Stolen memories: Leveraging model memorization for calibrated white-box membership inference. In *USENIX Security Symposium*, 2020.
13. Yunhui Long, Vincent Bindschaedler, Lei Wang, Diyue Bu, Xiaofeng Wang, Haixu Tang, Carl A. Gunter, and Kai Chen. Understanding membership inferences on well-generalized learning models. *ArXiv*, abs/1802.04889, 2018.
14. Yunhui Long, Lei Wang, Diyue Bu, Vincent Bindschaedler, Xiaofeng Wang, Haixu Tang, Carl A. Gunter, and Kai Chen. A pragmatic approach to membership inferences on machine learning models. *2020 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 521–534, 2020.
15. Milad Nasr, R. Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Stand-alone and federated learning under passive and active white-box inference attacks. *ArXiv*, abs/1812.00910, 2018.
16. Shahbaz Rezaei and Xin Liu. On the difficulty of membership inference attacks. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7888–7896, 2021.
17. Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, and Hervé Jégou. White-box vs black-box: Bayes optimal strategies for membership inference. In *ICML*, 2019.
18. A. Salem, Yang Zhang, Mathias Humbert, Mario Fritz, and Michael Backes. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. *ArXiv*, abs/1806.01246, 2019.
19. R. Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18, 2017.
20. Liwei Song and Prateek Mittal. Systematic evaluation of privacy risks of machine learning models. In *USENIX Security Symposium*, 2021.
21. Wei-Cheng Tseng, Wei-Tsung Kao, and Hung yi Lee. Membership inference attacks against self-supervised speech models. *ArXiv*, abs/2111.05113, 2021.
22. Lauren Watson, Chuan Guo, Graham Cormode, and Alexandre Sablayrolles. On the importance of difficulty calibration in membership inference attacks. In *International Conference on Learning Representations*, 2022.
23. Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv: Learning*, 2017.

24. Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pages 268–282, 2018.
25. Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *ArXiv*, abs/1611.03530, 2017.

Appendix

A Additional experiments

Table 8. Evaluation of various data sets, model structures, and MIA methods,$\alpha = 92\%$

		MNIST	F-MNIST	CIFAR10	Purchase100	Texas100
C-Loss	TP	37	173	83	4	1444
	Recall	0.37%	1.73%	0.83%	0.02%	14.44%
	Pr	94.87%	**92.51%**	**94.32%**	**100%**	90.02%
C-Conf	TP	85	589	1154	18	3323
	Recall	0.85%	5.89%	11.54%	0.09%	33.23%
	Pr	93.41%	91.21%	89.67%	94.73%	92.43%
Two-stage	TP	**516**	**1316**	**3170**	**5959**	**5118**
	Recall	**5.16%**	**13.16%**	**31.70%**	**29.80%**	**51.18%**
	Pr	**94.87%**	90.70%	87.79%	91.42%	**92.27%**

$\alpha = 92\%$

Table 9. Evaluation of various data sets, model structures, and MIA methods,$\alpha = 96\%$

		MNIST	F-MNIST	CIFAR10	Purchase100	Texas100
C-Loss	TP	17	94	46	4	662
	Recall	0.17%	0.94%	0.46%	0.02%	6.62%
	Pr	**100%**	**96.91%**	**93.88%**	**100%**	94.44%
C-Conf	TP	64	320	660	8	1621
	Recall	0.64%	3.20%	6.60%	0.04%	16.21%
	Pr	94.12%	96.68%	93.48%	100%	**96.89%**
Two-stage	TP	**340**	**806**	**1644**	**1419**	**3014**
	Recall	**3.40%**	**8.06%**	**16.44%**	**7.10%**	**30.14%**
	Pr	98.84%	94.82%	93.46%	96.14%	95.99%

$\alpha = 96\%$

B Algorithm details

Algorithm 1: High-Precision Membership Inference Attack

function ***Membership Inference-threshold*:**
  Input: shadow dataset D_{shadow}, target model h, membership score s and precision constraint value α. Construct the set $U_{in}(h, s, D^{in}_{shadow})$ and $U_{out}(h, s, D^{out}_{shadow})$according to Equation 5
  $m \leftarrow |U_{in}| + |U_{out}|$
  $U' \leftarrow \emptyset$
  $TPset \leftarrow \emptyset$
  for $i \leftarrow 1 \ : \ m$ **do**
    $TP \leftarrow \sum_{u_j \in U_{in}} I[u_j > u_i]$
    $FP \leftarrow \sum_{u_j \in U_{out}} I[u_j > u_i]$
    $precision \leftarrow \frac{TP}{TP+FP}$
    if $precision \geqslant \alpha$ **then**
      $U' \leftarrow U' \cup u_i$
      $TPset \leftarrow TPset \cup TP$
  if $U' == \emptyset$ **then**
    return *fail*
  $TP, k \leftarrow \max_{k=1,2,\ldots,n} TPset$
  $t \leftarrow U'_k$
  return t, TP
function ***HP-mia*:**
  Input: target model h, target record z, membership score s and threshold t.
  if $s(h, z) \geqslant t$ **then**
    return 1
  else
    return $\emptyset$

Algorithm 2: Two-stage High-Precision Membership Inference Attack

function *Membership Exclusion-threshold*:
  Input: shadow dataset D_{shadow}, target model h, membership score s and precision constraint value β.
  Construct the set $U_{in}(h, s, D^{in}_{shadow})$ and $U_{out}(h, s, D^{out}_{shadow})$according to Equation 5
  $m \leftarrow |U_{in}| + |U_{out}|$
  $U' \leftarrow \emptyset$
  $TPset \leftarrow \emptyset$
  for $i \leftarrow 1 \ : \ m$ **do**
    Attack the shadow model with u_i as the threshold
    Calculate the number of correctly identified non-members (TP) and the precision (Pr)
    if $Pr \geqslant \beta$ **then**
      $U' \leftarrow U' \cup u_i$
      $TPset \leftarrow TPset \cup TP$
  if $U' == \emptyset$ **then**
    return *fail*
  $k \leftarrow \max_{k=1,2,\ldots,n} TPset$
  $t \leftarrow U'_k$
  return t, TP
function *Two-stage threshold*:
  Input: shadow dataset D_{shadow}, target model h, membership score used s_0 in the first stage, membership score s_1 used in the second stage and precision constraint value α.
  $TP^{opt}, t_0^{opt}, t_1^{opt}$ initialized to 0
  for $\beta \leftarrow 0, 1; step = 0.001$ **do**
    $t_0 \leftarrow$ Membership Exclusion-threshold$(D_{shadow}, h, s_0, \beta)$
    $D_{remaining} \leftarrow \{z_i : z_i \in D_{shadow}, s(h, z_i) < t_0\}$
    $t_1, TP \leftarrow$ Membership Inference-threshold$(D_{remaining}, h, s_1, \alpha)$
    if $TP > TP^{opt}$ **then**
      $TP^{opt} \leftarrow TP \quad t_0^{opt} \leftarrow t_0$
      $t_1^{opt} \leftarrow t_1$
  return t_0^{opt}, t_1^{opt}
function *Two-stage attack*:
  Input: target model h, target record z, membership score used s_0 in the first stage and its threshold t_0, membership score s_1 used in the second stage and its threshold t_1
  if $s_0(h, z) < t_0$ **then**
    return $\emptyset$
  else if $s_1(h, z) < t_1$ **then**
    return $\emptyset$
  else
    return 1