Systematic Evaluation of Privacy Risks of Machine Learning Models

Liwei Song
liweis@princeton.edu
Princeton University

Prateek Mittal
pmittal@princeton.edu
Princeton University

Abstract

Machine learning models are prone to memorizing sensitive data, making them vulnerable to membership inference attacks in which an adversary aims to guess if an input sample was used to train the model. In this paper, we show that prior work on membership inference attacks may severely underestimate the privacy risks by relying solely on training custom neural network classifiers to perform attacks and focusing only on the aggregate results over data samples, such as the attack accuracy. To overcome these limitations, we first propose to benchmark membership inference privacy risks by improving existing non-neural network based inference attacks and proposing a new inference attack method based on a modification of prediction entropy. We propose to supplement existing neural network based attacks with our proposed benchmark attacks to effectively measure the privacy risks. We also propose benchmarks for defense mechanisms by accounting for adaptive adversaries with knowledge of the defense and also accounting for the trade-off between model accuracy and privacy risks. Using our benchmark attacks, we demonstrate that existing defense approaches against membership inference attacks are not as effective as previously reported.

Next, we introduce a new approach for fine-grained privacy analysis by formulating and deriving a new metric called the privacy risk score. Our privacy risk score metric measures an individual sample's likelihood of being a training member, which allows an adversary to identify samples with high privacy risks and perform membership inference attacks with high confidence. We propose to combine both existing aggregate privacy analysis and our proposed fine-grained privacy analysis for systematically measuring privacy risks. We experimentally validate the effectiveness of the privacy risk score metric and demonstrate that the distribution of privacy risk scores across individual samples is heterogeneous. Finally, we perform an in-depth investigation to understand why certain samples have high privacy risk scores, including correlations with model properties such as model sensitivity, generalization error, and feature embeddings. Our work emphasizes the importance of a systematic and rigorous evaluation of privacy risks of machine learning models. We publicly release our code at https://github.com/inspire-group/membership-inference-evaluation and our evaluation mechanisms have also been integrated in Google's TensorFlow Privacy library.

1 Introduction

A recent thread of research has shown that machine learning (ML) models memorize sensitive information of training data, indicating serious privacy risks [4,11,12,17,37,41,43]. In this paper, we focus on the membership inference attack, where the adversary aims to guess whether an input sample was used to train the target machine learning model or not [41,48]. It poses a severe privacy risk as the membership can reveal an individual's sensitive information [3,35]. For example, participation in a hospital's health analytic training set means that an individual was once a patient in that hospital. As membership inference attacks expose the privacy risks of an individual user participating in the training data, they serve as a valuable tool to quantify the privacy provided by differential privacy implementations [19] and to help to guide the selection of privacy parameters in the broader class of statistical privacy frameworks [25]. Shokri et al. [41] conducted membership inference attacks against machine learning classifiers in the black-box manner, where the adversary only observes prediction outputs of the target model. They formalize the attack as a classification problem and train dedicated neural network (NN) classifiers to distinguish between training members and non-members. The research community has since extended the idea of membership inference attacks to generative models [7,13,16,46], to differentially private models [19,36], to decentralized settings where the models are trained across multiple users without sharing their data [30,32], and to white-box settings where the adversary also has the access to the target model's architecture and weights [32].

To mitigate such privacy risks, several defenses against membership inference attacks have been proposed. Nasr et al. [31] propose to include membership inference attacks

during the training process: they train the target model to simultaneously achieve correct predictions and low membership inference attack accuracy by adding the inference attack as an adversarial regularization term. Jia et al. [20] propose a defense method called MemGuard which does not require retraining the model: the model prediction outputs are obfuscated with noisy perturbations such that the adversary cannot distinguish between members and non-members based on the perturbed outputs. Both papers show that their defenses greatly mitigate membership inference privacy risks, resulting in attack performance that is close to random guessing.

In this paper, we critically examine how previous work [20, 31, 32, 38, 41] has evaluated the membership inference privacy risks of machine learning models, and demonstrate two key limitations that lead to a severe underestimation of privacy risks. First, many prior papers, particularly those proposing defense methods [20, 31], solely rely on training custom NN classifiers to perform membership inference attacks. These NN attack classifiers may underestimate privacy risks due to inappropriate settings of hyperparameters such as number of hidden layers and learning rate. Second, existing evaluations only focus on aggregate notions of privacy risks faced by all data samples, lacking a fine-grained understanding of privacy risks faced by individual samples.

Table 1: Benchmarking the effectiveness of existing defenses [20, 31] against membership inference attacks. Both Nasr et al. [31] and Jia et al. [20] report that for their defended models, custom NN classifiers achieve attack accuracy close to 50%, which is the accuracy of random guessing. By using a suite of non-NN based attacks as our benchmark, we find that the attack accuracy is significantly larger than previous estimates, ranging from an increase of 6% to 239%.

defense methods	dataset	reported attack acc	our benchmark attack acc
adversarial regularization [31]	Purchase100	51.6%	59.5%
	Texas100	51.0%	58.6%
MemGuard [20]	Location30	50.1%	69.1%
	Texas100	50.3%	74.2%

To overcome the limitation of reliance on NN-based attacks, we propose to use a suite of alternative existing and novel non-NN based attack methods to benchmark the membership inference privacy risks. These benchmark attack methods make inference decisions based on computing custom metrics on the predictions of the target model. Compared to NN-based attacks, our proposed benchmark attacks are easy to implement without hyperparameter tuning. We only need to set the threshold values using the shadow-training technique [41]. We also show that rigorously benchmarking defense mechanisms requires a careful consideration of strategic adversaries aware of the defense mechanism, as well as alternative baselines that trade-off accuracy of the target machine learning model with privacy risks. With our proposed benchmark attacks, we indeed find that that existing membership inference defense methods [20, 31] are not as effective as previously reported. As shown in Table 1, the adversary can still perform membership inference attacks on models defended by adversarial regularization [31] and MemGuard [20] with an accuracy ranging from 58.6% to 74.2%, instead of the reported accuracy around 50%, which is the accuracy of random guessing. Therefore, we argue that these non-NN based attacks should supplement existing NN based attacks to effectively measure the privacy risks.

Figure 1: Cumulative distribution of privacy risk scores for undefended models trained on Purchase100, Location30, and CIFAR100 datasets.

To overcome the limitation of a lack of understanding of fine-grained privacy risks in existing works, we propose a new metric called the privacy risk score, that represents an individual sample's probability of being a member in the target model's training set. Figure 1 shows the cumulative distributions of privacy risk scores on target undefended models trained on Purchase100, Location30, and CIFAR100 datasets respectively. We can see that the privacy risk faced by individual training samples is heterogeneous. By utilizing the privacy risk score, an adversary can perform membership inference attacks with high confidence an input sample is inferred as a member if and only if its privacy risk score is higher than a certain threshold value. Overall, we recommend that our per-sample privacy risk analysis should be used in conjunction with existing aggregate privacy analysis for an in-depth understanding of privacy risks of machine learning models. Conventional aggregate analysis provides an average perspective of privacy risks incurred by all samples, while privacy risk score provides a perspective on privacy risk from the viewpoint of an individual sample. The former provides an aggregate estimation of privacy risks, while the latter allows us to understand the heterogeneous distribution of privacy risks faced by individual samples and identify samples with high privacy risks. We summarize our contributions as follows:

1. We propose a suite of non-NN based attacks to benchmark target models' privacy risks by improving existing attacks with class-specific threshold settings and designing a new inference attack based on a modified prediction entropy estimation in a manner that incorporates the ground truth class label. Furthermore, to rigorously evaluate the performance of membership inference defenses, we make recommendations for comparison with early stopping baseline and considering adaptive attackers with knowledge of defense mechanisms.

2. With our benchmark attacks, we find that two state-of-the-art defense approaches [20, 31] are not as effective as previously reported. Furthermore, we observe that the defense performance of adversarial regularization [31] is no better than early stopping, and the evaluation of MemGuard [20] lacks a consideration of adaptive adversaries. We also find that the existing white-box attacks [32] have limited advantages over our benchmark attacks, which only need black-box access to the target model. We also show that our attacks with class-specific threshold settings strictly outperform attacks with class-independent thresholds, and our new inference attack based on modified prediction entropy strictly outperforms conventional prediction entropy based attack.

3. We propose to analyze privacy risks of machine learning models in a fine-grained manner by focusing on individual samples. We define a new metric called the privacy risk score, that estimates an individual sample's probability of being in the target model's training set.

4. We experimentally demonstrate the effectiveness of our new metric in being able to capture the likelihood of an individual sample being a training member. We also show how an adversary can exploit our metric to launch membership inference attacks on individual samples with high confidence. Finally we perform an in-depth investigation of our privacy risk score metric, and its correlations with model sensitivity, generalization error, and feature embeddings.

Our code is publicly available at https://github.com/inspire-group/membership-inference-evaluation for the purpose of reproducible research. Furthermore, our evaluation mechanisms have also been integrated in Google's TensorFlow Privacy library.

2 Background and Related Work

In this section, we first briefly introduce machine learning basics and notations. Next, we present existing membership inference attacks, including black-box attacks and white-box attacks. Finally, we discuss two state-of-the-art defense methods: adversarial regularization [31] and MemGuard [20].

2.1 Machine learning basics and notations

Let $F_q : R^d \to R^k$ be a machine learning model with d input features and k output classes, parameterized by weights q. For an example $z = (x, y)$ with the input feature x and the ground truth label y, the model outputs a prediction vector over all class labels $F_q(x)$ with $\sum_{i=0}^{k-1} F_q(x)_i = 1$, and the final classification result will be the label with the largest prediction probability $\hat{y} = \arg\max_i F_q(x)_i$.

Given a training set D_{tr}, the model weights are optimized by minimizing the prediction loss over all training examples.

$$\min_q \frac{1}{|D_{tr}|} \sum_{z \in D_{tr}} \ell(F_q, z); \quad (1)$$

where $|D_{tr}|$ denotes the size of training set, and ℓ computes the prediction loss, such as cross-entropy loss. In this paper, we skip the model parameter q for simplicity and use F to denote the machine learning model.

2.2 Membership inference attacks

For a target machine learning model, membership inference attacks aim to determine whether a given data point was used to train the model or not [26, 38, 41, 48]. The attack poses a serious privacy risk to the individuals whose data is used for model training, for example in the setting of health analytics.

2.2.1 Black-box membership inference attacks

Shokri et al. [41] investigated the membership inference attacks against machine learning models in the black-box setting. For an input sample $z = (x, y)$ to the target model F, the adversary only observes the prediction output $F(x)$ and infers if z belongs to the model's training set D_{tr}. To distinguish between target model's predictions on members and non-members, the adversary learns an attack model using the shadow training technique: (1) the adversary first trains multiple shadow models to simulate the behavior of the target model; (2) based on shadow models' outputs on their own training and test examples, the adversary obtains a labeled (member vs non-member) dataset, and (3) finally trains multiple neural network (NN) classifiers, one for each class label, to perform inference attacks against the target model.

Salem et al. [38] show that even with only a single shadow model, membership inference attacks are still quite successful. Furthermore, in the case where the adversary knows a subset of target model's training set and test set, the attack classifier can be directly trained with target model's predictions on those known samples, and then tested on unknown training and test sample [31, 32]. Nasr et al. [31] redesign the attack by using one-hot encoded class labels as part of input features and training a single NN attack classifier for all class labels.

Besides membership inference attacks that rely on training NN classifiers, there are non-NN based attack methods that

make inference decisions based on computing custom metrics on the predictions of the target model. Leino et al. [24] suggest using the metric of prediction correctness as a sign of being a member or not. Yeom et al. [48] and Song et al. [44] find that the metric of prediction confidence of correct label $F(x)_y$ can be compared with a certain threshold value to achieve similar attack performance as NN-based attacks. Shokri et al. [41] show a large divergence between prediction entropy distributions over training data and test data, although this metric was not explicitly used for attacks.

Despite the existence of such non-NN based attacks, many research papers [20,31,32] still only train NN attack classifiers to evaluate target models' privacy risks. We find that this can lead to severe underestimation of privacy risks by re-evaluating the same target models with non-NN based attacks. Furthermore, we improve existing non-NN based attacks by setting different threshold values for different class labels, building upon the motivation of separated attack classifiers for each class label by Shokri et al. [41]. We also propose a new inference attack method by considering ground truth label when evaluating prediction uncertainty.

2.2.2 White-box membership inference attacks

Nasr et al. [32] analyze membership inference attacks in the white-box setting, where the adversary has the full access to the target machine learning model and knows the model architecture and model parameters. They find that simply combining target model's final predictions and its intermediate computations to learn the attack classifier results in attack accuracy no better than that of the corresponding black-box attacks. Instead, by using the gradient of prediction loss with regard to model parameters $\frac{\nabla_\theta \ell(F_\theta, z)}{\nabla \theta}$ as additional features, the white-box membership inference attacks obtain higher attack accuracy than the black-box attacks. We show that the gap between white-box attack accuracy and black-box attack accuracy is much smaller than previous estimates in this paper.

2.3 Defenses against membership inference attacks

To mitigate the risks of membership inference attacks, several defense ideas have been proposed. L_2 norm regularization [23] and dropout [45] are standard techniques for reducing over fitting in machine learning. They are also shown to decrease privacy risks to some degree [38, 41]. However, target models can still be quite vulnerable after applying these techniques. Differential privacy [9, 10] can also be applied to ML models for provable risk mitigation [1, 29, 33, 40], however, it induces significant accuracy drop for desired values of the privacy parameter [19]. Two dedicated defenses, adversarial regularization [32] and MemGuard [20], were recently proposed against membership inference attacks. Both defenses are reported to have the ability of decreasing the attack accuracy to around 50%, which is the performance of random guessing. We explain their details below.

2.3.1 Adversarial regularization [31]

Nasr et al. [31] propose to include the membership inference adversary with the NN-based attack into the training process itself to mitigate privacy risks. At each training step, the attack classifier is first updated to distinguish between training data (members) and validation data (non-members), and then the target classifier is updated to simultaneously minimize the prediction loss and mislead the attack classifier.

More specifically, to train the classifier F with parameters θ in a manner that is resilient against membership inference attacks, Nasr et al. [31] use another classifier I with parameters ω to perform membership inference attacks. The attack classifier I takes the target model's prediction $F(x)$ and the input label y as input features and generate one single output $I(F(x), y)$, which is in the range [0, 1]. It infers the input sample as a member if the output is larger than 0.5, a non-member otherwise. At each training step, they first update the attack classifier I by maximizing the membership inference gain over the training set D_{tr} and the validation set D_{val}.

$$\arg\max_{\omega} \frac{\sum_{z \in D_{tr}} \log(I(F(x), y))}{|D_{tr}|} + \frac{\sum_{z \in D_{val}} \log(1 - I(F(x), y))}{|D_{val}|} \tag{2}$$

They further train the target classifier by minimizing both model prediction loss and membership inference gain over the training set D_{tr}.

$$\arg\min_{\theta} \frac{1}{|D_{tr}|} \sum_{z \in D_{tr}} \ell(F(x), y) + \lambda \log(I(F(x), y)); \tag{3}$$

where λ is a penalty parameter for the privacy risk. In this way, the target model F is trained with an additional regularization term to defend against membership inference attacks.

2.3.2 MemGuard [20]

Jia et al. [20] propose MemGuard as a defense method against membership inference attacks, which, different from Nasr et al. [31], does not need to modify the training process. Instead, given a pre-trained target model F, they obfuscate its predictions with well-designed noises to confuse the membership inference classifier I, without changing classification results.

The attack classifier I is trained following the shadow-training technique [41], which takes the model prediction $F(x)$ with the sample label y, and outputs a score $I(F(x), y)$ in the range [0 ,1] for membership inference: if the output is larger than 0.5, the data sample is inferred as a member, and vice versa. The key question of how to add noise to $F(x)$

can be formulated as the following optimization problem:

$$\begin{aligned} &\min_{n} d(F(x)+n, F(x)); \\ &\text{subject to: } \operatorname{argmax}_i(F(x)_i + n_i) = \operatorname{argmax}_i F(x)_i; \\ &I(F(x)+n) = 0.5; \\ &F(x)_i + n_i \geq 0, \forall i \\ &\sum_i n_i = 0; \end{aligned} \tag{4}$$

where the objective is to minimize the distance between original predictions and noisy predictions. The first constraint ensures the classification result does not change after adding noise, the second constraint ensures the attack classifier cannot determine whether the sample is a member or a non-member with the noisy predictions, and last two constraints ensure the noisy predictions are valid.

When evaluating the defense performance, both Nasr et al. [31] and Jia et al. [20] train NN classifiers for inference attacks. As shown in the following section, we find that their evaluations underestimate privacy risks. With our benchmark attacks, the adversary achieves significantly higher attack accuracy on defended models than previous estimates. We further find that the performance of adversarial regularization [31] is no better than early stopping, and the evaluation of MemGuard [20] lacks consideration of strategic adversaries.

3 Systematically Evaluating Membership Inference Privacy Risks

In this section, we first present a suite of non-NN based attacks to benchmark privacy risks, which only need to observe target model's output predictions (i.e., black-box setting). Next, we provide two recommendations, comparison with early stopping and considering adaptive attacks, to rigorously measure the effectiveness of defense approaches. Finally, we present experiment results by re-evaluating target models in prior work [20, 31, 32] with our proposed benchmark attacks.

3.1 Benchmarks of membership inference attacks

We propose to use a suite of non-NN based attack methods to benchmark membership inference privacy risks of machine learning models. We call these attack methods "metric-based attacks" as they first measure the performance metrics of target model's predictions, such as correctness, confidence, and entropy, and then compare those metrics with certain threshold values to infer whether the input sample is a member or a non-member [24, 44]. We improve existing metric-based attacks by setting different threshold values for different class labels of target models. Then we propose another new metric-based attack by considering ground truth label when evaluating prediction uncertainty. We denote the inference strategy as I, which codes members as 1, and non-members as 0. Overall, we propose that existing NN based attacks should be supplemented with our metric-based attacks for systematically and rigorously evaluating privacy risks of ML models.

3.1.1 Existing attacks

Inference attack based on prediction correctness. Sablayrolles et al. [24] observe that the membership inference attacks based on whether the input is classified correctly or not achieve comparable success as NN-based attack on target models with large generalization errors. The intuition is that the target model is trained to predict correctly on training data (members), which may not generalize well on test data (non-members). Thus, we can rely on the prediction correctness metric for membership inference. The adversary infers an input sample as a member if it is correctly predicted, a non-member otherwise.

$$I_{corr}(F,(x,y)) = \mathbb{1}\{\operatorname{argmax}_i F(x)_i = y\}; \tag{5}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function.

3.1.2 Improving existing attacks with class-dependent thresholds

Inference attack based on prediction confidence. Yeom et al. [48] and Song et al. [44] show that the attack strategy of using a threshold on the prediction confidence results in similar attack accuracy as NN-based attacks. The intuition is that the target model is trained by minimizing prediction loss over training data, which means the prediction confidence of a training sample $F(x)_y$ should be close to 1. On the other hand, the model is usually less confident in predictions on a test sample. Thus, we can rely on the metric of prediction confidence for membership inference. The adversary infers an input example as a member if its prediction confidence is larger than a preset threshold, a non-member otherwise.

$$I_{conf}(F,(x,y)) = \mathbb{1}\{F(x)_y \geq t_y\}. \tag{6}$$

Yeom et al. [48] and Song et al. [44] choose to use a single threshold for all class labels. We improve this method by setting different threshold values t_y for different class labels y. The reason is that the dataset may be unbalanced so that the target model indeed has different confidence levels for different class labels. Our experiments show that this class-dependent thresholding technique leads to better attack performance. The class-dependent threshold values t_y are learned with the shadow-training technique [41]: the adversary (1) first trains a shadow model to simulate the behavior of the target model; (2) then obtains the shadow model's prediction confidence values on both shadow training and shadow test data; (3) finally leverages knowledge of membership labels (member vs non-member) of the shadow data to select the

threshold value t_y which achieves the highest accuracy in distinguishing between shadow training data and shadow test data with the class label y based on Equation (6).

Inference attack based on prediction entropy Although there is no prior work using prediction entropy for membership inference attacks, Shokri et al. [41] indeed present the difference of prediction entropy distributions between training and test data to explain why privacy risks exist. Salem et al. [38] also mention the possibility of using prediction entropy for attacks. The intuition is that the target model is trained by minimizing the prediction loss over training data, which means the prediction output of a training sample should be close to a one-hot encoded vector and its prediction entropy should be close to 0. On the other hand, the target model usually has a larger prediction entropy on an unseen test sample. Therefore, we can rely on the metric of prediction entropy for membership inference. The adversary classifies an input example as a member if its prediction entropy is smaller than a preset threshold, a non-member otherwise.

$$I_{entr}(F;(x;y)) = 1\{ -\sum_i F(x)_i \log(F(x)_i) \quad t_y\}: \quad (7)$$

Similar to the confidence-based attack, we propose to use the threshold values t_y that are dependent on the class labels and are set with the shadow-training technique [41].

3.1.3 Our new inference attack based on modified prediction entropy

The attack based on prediction entropy has one serious issue: it does not contain any information about the ground truth label. In fact, both a correct classification with probability of 1 and a totally wrong classification with probability of 1 lead to zero prediction entropy values.

To resolve this issue, we design a new metric with following two properties to measure the model prediction uncertainty given the ground truth label: it should be (1) monotonically decreasing with the prediction probability of the correct label $F(x)_y$, and (2) monotonically increasing with the prediction probability of any incorrect label $F(x)_i; 8i \neq y$. Let $x \in [0;1]$ denote the prediction probability for a certain label, the function used in conventional entropy computation $-x\log x$ is not a monotonic function. As a contrast, $(1-x)\log x$ is a monotonically decreasing function, and $-x\log(1-x)$ is a monotonically increasing function. Therefore, we propose the modified prediction entropy metric, computed as follows.

$$\begin{aligned} Mentr(F(x);y) = \; & (1-F(x)_y)\log(F(x)_y) \\ & -\sum_{i\neq y} F(x)_i \log(1-F(x)_i): \end{aligned} \quad (8)$$

In this way, a correct classification with probability of 1 leads to modified entropy of 0, while a wrong classification with probability of 1 leads to modified entropy of infinity.

Now, with the new metric of modified prediction entropy, the adversary classifies an input example as a member if its modified prediction entropy is smaller than a preset threshold, a non-member otherwise.

$$I_{Mentr}(F;(x;y)) = 1\{Mentr(F(x);y) \quad t_y\}: \quad (9)$$

Similar to previous scenarios, we set different threshold values t_y for different class labels, which are learned with the shadow training technique [41]. Experiments show that the inference attack based on our modified prediction entropy is strictly superior to the inference attack based on prediction entropy.

3.2 Rigorously evaluating membership inference defenses

To evaluate the effectiveness of defenses against membership inference attacks, we make the following two recommendations, besides using our metric-based benchmark attacks.

3.2.1 Comparison with early stopping

During the training process, the target model's parameters are updated following gradient descent methods, so the training error and test error usually get reduced gradually with an increasing number of training epochs. However, as the number of training epochs increases, the target model also becomes more vulnerable to membership inference attacks, due to increased memorization. We thus propose to use early stopping [6, 34, 47] as a benchmark defense method, in which fewer training epochs are used in order to tradeoff a slight reduction in model accuracy with lower privacy risk.

Figure 2: Test accuracy at different training epochs for Purchase100 classifiers without defense and with adversarial regularization defense [31]. We should compare the final defended model to the model with early stopping.

We recommend that whenever a defense method is proposed in the literature that reduces the threat of membership inference attacks at the cost of degradation in model accuracy, the performance of the defense method should be benchmarked against our early stopping approach. This is indeed the

case for the defense method of adversarial regularization (AdvReg) [31]. As shown in Figure 2, the defended Purchase100 classifier should be compared to the undefended model with fewer training steps and similar accuracy.

3.2.2 Adaptive attacks

There always exists an arms race between privacy attacks and defenses for machine learning models. When evaluating the defense performance, it is critical to put the adversary into the last step, i.e., the adversary knows the defense mechanism and performs adaptive attacks against the defended models. A perfect defense performance with non-adaptive attacks does not mean that the defense approach is effective [2, 5, 15].

Specifically for defenses proposed against membership inference attacks, we should consider that the adversary knows the defense mechanism such that he or she can train shadow models following the defense method. From these defended shadow models, the adversary then learns an attack classifier or sets threshold values for metric-based attacks, and finally performs attacks on the defended target model.

3.3 Experiment results

We first re-evaluate the effectiveness of two membership inference defenses [20, 31], and then re-evaluate the white-box membership inference attacks proposed by Nasr et al. [32]. Following prior work [41, 44, 48], we sample the input (x, y) from either the target model's training set or test set with an equal 0.5 probability to maximize the uncertainty of membership inference attacks. Thus, the random guessing strategy results in a 50% membership inference attack accuracy.

3.3.1 Datasets

Purchase100 This dataset is based on Kaggle's Acquire Valued Shoppers Challenge[1], which contains shopping records of several thousand individuals. We obtain a simplified and preprocessed purchase dataset provided by Shokri et al. [41]. The dataset has 197,324 data samples with 600 binary features. Each feature corresponds to a product and represents whether the individual has purchased it or not. All data samples are clustered into 100 classes representing different purchase styles. The classification task is to predict the purchase style based on the 600 binary features. We follow Nasr et al. [31, 32] to use 10% data samples (19,732) to train a model.

Texas100 This dataset is based on the Hospital Discharge Data public use files with patients' information released by the Texas Department of State Health Services[2]. Each data record contains the external causes of injury (e.g., suicide, drug misuse), the diagnosis (e.g., schizophrenia), the procedures the patient underwent (e.g., surgery) and some generic information (e.g., gender, age, race). We obtain a simplified and preprocessed Texas dataset provided by Shokri et al. [41]. The classification task is to predict the patient's main procedure based on the patient's information. The dataset focuses on 100 most frequent procedures, resulting in 67,330 data samples with 6,170 binary features. Following previous papers [20, 31, 32], we use 10,000 data samples to train a model.

Location30 This dataset is based on Foursquare dataset[3], which contains location "check-in" records of several thousand individuals. We obtain a simplified and preprocessed Location dataset provided by Shokri et al. [41]. The dataset contains 5,010 data samples with with 446 binary features. Each feature corresponds to a certain region or location type and represents whether the individual has visited the region/location or not. All data samples are clustered into 30 classes representing different geosocial types. The classification task is to predict the geosocial type based on the 466 binary features. Following Jia et al. [20], we use 1,000 data samples to train a model.

CIFAR100 This is a major benchmark dataset for image classification [21]. It is composed of 32×32 color images in 100 classes, with 600 images per class. For each class label, 500 images are used as training samples, and the remaining 100 images are used as test samples.

We choose these datasets for fair comparison with prior work [20, 31, 32]. Since all datasets except CIFAR100 are binary datasets, we also provide attack results with more complex datasets in Appendix A, where our benchmark attacks achieve higher attack success than NN-based attacks.

3.3.2 Re-evaluating adversarial regularization [31]

We follow Nasr et al. [31] to train both defended and undefended classifiers on Purchase100 and Texas100 datasets. For both datasets, the model architecture is a fully connected neural network with 4 hidden layers. The numbers of neurons for hidden layers are 1024, 512, 256, and 128, respectively. All hidden layers use hyperbolic tangent (Tanh) as the activation function. We note that the defense method of adversarial regularization [31] incurs accuracy drop. After applying the defense, the test accuracy drops from 80.9% to 76.6% on the Purchase100 dataset, and from 52.3% to 46.4% on the Texas100 dataset. As we discuss in Section 3.2.1, to further evaluate the effectiveness of adversarial regularization [31], we also obtain models with early stopping by saving the undefended models in every training epoch and picking the saved epochs with similar accuracy performance as defended models. Table 2 presents the membership inference attack results.

From Table 2, we can see that the defended models are still vulnerable to membership inference attacks, indicat-

[1] https://www.kaggle.com/c/acquire-valued-shoppers-challenge

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

[3] https://sites.google.com/site/yangdingqi/home/foursquare-dataset

Table 2: Benchmarking the effectiveness of using adversary regularization [31] as defense against membership inference attacks. We can see that the defended models are still vulnerable to membership inference attacks.

dataset	Model Performance: using defense [31]?	training acc	test acc	Membership Inference Attacks: attack acc by [31]	attack acc (I_{corr})	attack acc (I_{conf})	attack acc (I_{entr})	attack acc (I_{Mentr})
Purchase100	no	99.8%	80.9%	67.6%	59.5%	67.1%	65.7%	67.1%
Purchase100	yes	92.7%	76.6%	51.6%	58.1%	59.4%	55.8%	59.5%
Purchase100	early stopping	92.9%	76.4%	N.A.	58.2%	59.2%	55.9%	59.1%
Texas100	no	81.0%	52.3%	63.0%	64.4%	67.8%	60.2%	67.7%
Texas100	yes	56.6%	46.4%	51.0%	55.1%	58.6%	53.5%	58.6%
Texas100	early stopping	59.3%	47.9%	N.A.	55.7%	59.4%	54.0%	59.5%

ing the necessity of our metric-based benchmark attacks. We achieve 59.5% and 58.6% attack accuracy on the defended Purchase100 classifier and the defended Texas100 classifier with our benchmark attacks, significantly larger than 51.6% and 51.0% as reported by Nasr et al. [31]. Furthermore, on all models except the undefended Purchase100 classifier, the largest attack accuracy achieved by benchmark attacks is larger than that of NN based attacks used in Nasr et al. [31]. Note that the defense method provides limited mitigation of privacy risks: it reduces attack accuracy from around 67% to around 59% on tested models. We also find that our new attack based on the modified entropy (I_{Mentr}) always outperforms the conventional entropy based attack (I_{entr}). It is also very competitive among all benchmark attacks.

From Table 2, we also surprisingly find that adversarial regularization [31] is no better than our early stopping benchmark method: with early stopping, the undefended Purchase100 classifier and the undefended Texas100 classifier have the attack accuracy 59.2% and 59.5%, which are quite close to those of defended models. Therefore, when evaluating the effectiveness of a future defense mechanism that trades lower model accuracy for lower membership inference risk, we argue to compare the defended model to the naturally trained model with early stopping for a fair comparison. We emphasize that our early stopping baseline can be calibrated to achieve similar model accuracy as the defended model. In contrast, the adversarial regularization approach may have a model accuracy which is different from the defended model under consideration, and will thus not represent a fair comparison.

To show the attack improvement yielded by our class-dependent thresholding technique, we compare with metric-based attacks when the same threshold is applied to all class labels. Table 3 shows the results on Texas100 classifiers without defense, with AdvReg [31], and with early stopping. We can see that with the class-dependent thresholding technique, we increase the attack accuracy by 1% – 4%.

Table 3: Comparing attack performance between conventional class-independent thresholding attacks and our class-dependent thresholding attacks.

attack methods	defense methods for Texas100 classifier: no defense	AdvReg [31]	early stopping
I_{conf} (class-independent)	64.7%	55.5%	55.8%
I_{conf} (class-dependent)	67.8%	58.6%	59.4%
I_{entr} (class-independent)	58.3%	52.9%	53.2%
I_{entr} (class-dependent)	60.2%	53.5%	54.0%
I_{Mentr} (class-independent)	64.8%	55.4%	55.9%
I_{Mentr} (class-dependent)	67.7%	58.6%	59.5%

3.3.3 Re-evaluating MemGuard [20]

We follow Jia et al. [20] to train classifiers on Location30 and Texas100 datasets. For both datasets, the model architecture is a fully connected neural network with 4 hidden layers. The numbers of neurons for hidden layers are 1024, 512, 256, and 128, respectively. All hidden layers use rectified linear unit (ReLU) as the activation function. MemGuard [20] does not change the accuracy performance, so the comparison with early stopping benchmark is not applicable. Table 4 lists the attack accuracy on both undefended and defended models, with attack methods in Jia et al. [20] and our metric-based benchmark attack methods. In fact, Jia et al. [20] use 6 different NN attack classifiers to measure the privacy risks, and we pick the highest attack accuracy among them.

From Table 4, we again emphasize the necessity of our benchmark attacks by showing that the defended models still have high membership inference accuracy: 69.1% on the defended Location30 classifier and 74.2% on the defended

Table 4: Benchmarking the effectiveness of using MemGuard [20] as defense against membership inference attacks. We can see that the defended models are still vulnerable to membership inference attacks.

dataset	Model Performance: using defense [20]?	training acc	test acc	Membership Inference Attacks: attack acc by [20]	attack acc (I_{corr})	attack acc (I_{conf})	attack acc (I_{entr})	attack acc (I_{Mentr})
Location30	no	100%	60.7%	81.1%	68.7%	76.3%	61.6%	78.1%
Location30	yes	100%	60.7%	50.1%	68.7%	69.1%	52.1%	68.8%
Texas100	no	99.95%	51.77%	74.0%	74.2%	79.0%	66.6%	79.4%
Texas100	yes	99.95%	51.77%	50.3%	74.2%	74.1%	54.6%	74.0%

Table 5: Benchmarking the effectiveness of white-box membership inference attacks proposed by Nasr et al. [32]. We can see that compared with our black-box benchmark attacks, the advantage of white-box attacks is limited.

dataset	Model Performance: training acc	test acc	Membership Inference Attacks: attack acc by [32] (white-box)	attack acc by [32] (black-box)	attack acc (I_{corr})	attack acc (I_{conf})	attack acc (I_{entr})	attack acc (I_{Mentr})
Purchase100	99.8%	80.9%	73.4%	67.6%	59.5%	67.1%	65.7%	67.1%
Texas100	81.0%	52.3%	68.3%	63.0%	64.4%	67.8%	60.2%	67.7%
CIFAR100	100%	83.00%	74.3%	67.7%	58.5%	73.7%	73.3%	73.6%

Texas100 classifier, much larger than 50.1% and 50.3% reported by Jia et al. [20]. We even achieve higher membership inference accuracy than attacks in Jia et al. [20] on all models, except the undefended Location30 classifier. Note that the defense still works but to a limited degree: it reduces the attack accuracy by 12% on the Location30 classifier and by 5% on the Texas100 classifier. Similar to Section 3.3.2, our proposed modified-entropy based attack always achieves higher attack accuracy than the entropy based attack, and is very competitive among all benchmark attacks.

Next, we discuss why Jia et al. [20] fail to achieve high membership inference accuracy for their defended models. We find that most of their attacks (4 out of 6) are non-adaptive attacks, where the adversary has no idea of the implemented defense, and thus the membership inference attacks are not successful. For the two adaptive attacks, Jia et al. [20] not put the adversary in the last step of the arms race between attacks and defenses. In their attacks, the adversary is aware that the model predictions will be perturbed with noises but does not know the exact algorithm of noise generation implemented by the defender. In their first adaptive attack, Jia et al. [20] round the model predictions to be one decimal during the attack classifier's inference to mitigate the effect of the perturbation. However, the attack performance is greatly degraded when the applied perturbation is large. In the second adaptive attack, Jia et al. [20] train the attack classifier using the state-of-the-art robust training algorithm by Madry et al. [28], with the hope that noisy perturbation will not change the classification. However, the robust training algorithm [28] has a very poor generalization property: the predictions on test points are still likely to be wrong after adding well-designed noises. For a thorough evaluation of the defense, we should consider that the attacker has the full knowledge of the defense mechanism, and he or she learns the attack model based on the defended shadow models.

3.3.4 Re-evaluating white-box membership inference attacks [32]

We have shown that previous work may underestimate the target models' privacy risks, and the metric-based attacks with only black-box access can result in higher attack accuracy than NN based attacks for most models. Recently Nasr et al. [32] demonstrated that a white-box membership inference adversary can perform stronger NN based attacks by using gradient with regard to model parameters. Next, we evaluate whether the advantage of white-box attacks still exists by using our metric-based black-box benchmark attacks.

We follow Nasr et al. [32] to obtain classifiers on Purchase100, Texas100 and CIFAR100 datasets. The Purchase100 classifier and the Texas100 classifier are same as undefended classifiers in Section 3.3.2. The CIFAR100 classifier is a publicly available pre-trained model[4] with the DenseNet architecture [18]. Table 5 lists all attack results.

[4] https://github.com/bearpaw/pytorch-classification

From Table 5, we can see that compared to the black-box metric-based attacks, the improvement of white-box membership inference attacks is limited. The attack accuracy of white-box membership inference adversary is only 0.5% and 0.6% higher than the attack accuracy achieved by our black-box benchmark attacks, on the Texas100 and the CIFAR100 classifiers. The white-box attack on the Purchase100 classifier still has 5.8% increase in attack accuracy compared to black-box attacks. As a validation of our observations, we note that Shejwalkar and Houmansadr also report close membership inference attack accuracy between white-box attacks and black-box attacks in their recent work [39].

4 Fine-Grained Analysis on Privacy Risks

Prior work [20, 31, 32, 41, 44] focuses on an aggregate evaluation of privacy risks by reporting overall attack accuracy or a precision-recall pair, which are averaged over all samples. However, the target machine learning model's performance is usually varied across samples, which denotes the heterogeneity of samples' privacy risks. Therefore, a fine-grained privacy risk analysis of individual samples is needed, with which we can understand the distribution of privacy risks over samples and identify which samples have high privacy risks.

In this section, we first define a metric called privacy risk score to quantitatively measure the privacy risks for each individual training member. Then we use this metric to experimentally measure fine-grained privacy risks of target models. Overall, we argue that existing aggregate privacy analysis of ML models should be supplemented with our fine-grained privacy analysis for a thorough evaluation of privacy risks.

4.1 Definition of privacy risk score

For membership inference attacks, the privacy risk of a training member arises due to the distinguishability of its model prediction behavior with non-members. This motivates our definition of the privacy risk score as following.

Definition 1 The privacy risk score of an input sample $z = (x, y)$ for the target machine learning model F is defined as the posterior probability that it is from the training set D_{tr} after observing the target model's behavior over that sample denoted as $O(F, z)$, i.e.,

$$r(z) = P(z \in D_{tr} | O(F, z)) \tag{10}$$

Based on Bayes' theorem, we further compute the privacy risk score as following.

$$\begin{aligned} r(z) &= \frac{P(z \in D_{tr}) \; P(O(F,z) | z \in D_{tr})}{P(O(F,z))} \\ &= \frac{P(z \in D_{tr}) \; P(O(F,z) | z \in D_{tr})}{P(z \in D_{tr}) \; P(O(F,z) | z \in D_{tr}) + P(z \in D_{te}) \; P(O(F,z) | z \in D_{te})}; \end{aligned} \tag{11}$$

where D_{te} stands for the test set. The observation $O(F, z)$ depends on the adversary's access to the target model: in the black-box membership inference attack [41], it is the model's final output, i.e., $O(F, z) = F(x)$; in the white-box membership inference attacks [32], it also includes the model's intermediate layers' outputs and gradient information at all layers. Our proposed benchmark attacks only need black-box access to the target model, and most existing attack methods [41, 44, 48] work in the black-box manner. Therefore, we focus on the black-box scenario for the computation of the privacy risk score in this paper and leave the discussion on white-box scenario as future work. In the black-box attack scenario, the privacy risk score can be expressed as

$$r(z) = \frac{P(z \in D_{tr}) \; P(F(x) | z \in D_{tr})}{P(z \in D_{tr}) \; P(F(x) | z \in D_{tr}) + P(z \in D_{te}) \; P(F(x) | z \in D_{te})} \tag{12}$$

From Equation (12), we can see that the risk score depends on both prior probabilities $P(z \in D_{tr})$, $P(z \in D_{te})$ and conditional distributions $P(F(x) | z \in D_{tr})$, $P(F(x) | z \in D_{te})$. For the prior probabilities, we follow previous work [41, 48] to assume that an example is sampled from either training set or test set with an equal 0.5 probability, where the uncertainty of membership inference attacks is maximized. Note that the privacy risk score is naturally applicable to any prior probability scenario, and we present the results with different prior probabilities in Appendix B. With the equal probability assumption, we have

$$r(z) = \frac{P(F(x) | z \in D_{tr})}{P(F(x) | z \in D_{tr}) + P(F(x) | z \in D_{te})} \tag{13}$$

For the conditional distributions $P(F(x) | z \in D_{tr})$, $P(F(x) | z \in D_{te})$, we empirically measure these values using shadow-training technique: (1) train a shadow model to simulate the behavior of the target model; (2) obtain the shadow model's prediction outputs on shadow training and shadow test data; (3) empirically compute the conditional distributions on shadow training and shadow test data. Furthermore, as the class-dependent thresholding technique is shown to improve the attack success in Table 3, we compute the distribution of model prediction over training data $P(F(x) | z \in D_{tr})$ in a class-dependent manner ($P(F(x) | z \in D_{te})$ is computed in the same way).

$$P(F(x) | z \in D_{tr}) = \begin{cases} P(F(x) | z \in D_{tr}; y = y_0); & \text{when } y = y_0 \\ P(F(x) | z \in D_{tr}; y = y_1); & \text{when } y = y_1 \\ \vdots & \\ P(F(x) | z \in D_{tr}; y = y_n); & \text{when } y = y_n \end{cases} \tag{14}$$

Since we empirically measure the conditional distributions using the shadow model's predictions over shadow data, the quality of measured distributions highly depends on the shadow model's similarity to the target model and the size of

shadow data. On the one hand, the size of shadow data is usually limited. Especially in our analysis where the distribution is computed in a class-dependent manner, for each class label y_n, we may not have enough samples[5] to adequately estimate the multi-dimension distribution $P(F(x)|z \in D_{tr}; y = y_n)$. On the other hand, in Section 3.3 we show that by only using the one-dimension prediction metric such as confidence and modified entropy, our proposed benchmark attacks in fact achieve comparable or even better success that NN-based attacks which leverage the whole prediction vector as features. Thus, we propose to further approximate the multi-dimension distribution in Equation (14) with the distribution of modified prediction entropy, since using modified entropy usually results in highest attack accuracy among all benchmark attacks.[6]

$$P(F(x)|z \in D_{tr}) \approx \begin{cases} P(Ment_r(F(x); y)|z \in D_{tr}; y = y_0); & \text{when } y = y_0 \\ P(Ment_r(F(x); y)|z \in D_{tr}; y = y_1); & \text{when } y = y_1 \\ \vdots & \\ P(Ment_r(F(x); y)|z \in D_{tr}; y = y_n); & \text{when } y = y_n \end{cases} \quad (15)$$

We also approximate $P(F(x)|z \in D_{te})$ in the same way. By plugging Equation (15) into Equation (13), we can get the privacy risk score for a certain sample.

4.2 Experiment results

In our experiments, we first validate that our proposed privacy risk score really captures the probability of being a member. Next, we compare the distributions of training samples' privacy risk scores for target models without defense and with defenses [20, 31]. We then demonstrate how to use privacy risk scores to perform membership inference attacks with high confidence. Finally, we perform an in-depth investigation of individual samples' privacy risk scores by correlating them with model sensitivity, generalization errors, and feature embeddings. To have enough diversity of data and models, and to further evaluate defense methods, we perform experiments on 3 Purchase100 classifiers (without defense, with AdvReg [31], and with early stopping) and 2 Texas100 classifiers (without defense, and with MemGuard [20]). Both Purchase100 classifiers and Texas classifiers use fully connected neural networks with 4 hidden layers, and the numbers of neurons for hidden layers are 1024, 512, 256, and 128, respectively. Purchase100 classifiers use Tanh as the activation function [31], and Texas100 classifiers use ReLU as the activation function [20].

[5] In our experiments, on average we have 197 samples per class for Purchase100 dataset; 100 samples per class for Texas100 dataset; 33 samples per class for Location30 dataset; and 500 per class for CIFAR100 dataset.

[6] In most cases, both modified entropy based attack and confidence based attack give best attack performance. However, for undefended Location30 and Texas100 classifiers in Table 4, the modified entropy based attack achieves significantly higher attack accuracy.

4.2.1 Validation of privacy risk score

Before presenting the detailed results for privacy risk score, we first validate its effectiveness here. For the target machine learning model, we first compute the privacy risk scores following the method in Section 4.1 for all training and test samples. Next we divide the entire range of privacy risk scores into multiple bins, and count the number of training points (n_{tr}) and the number of test points (n_{te}) in each bin. Then we compute the fraction of training points ($\frac{n_{tr}}{n_{tr}+n_{te}}$) in each bin, which indicates the real likelihood of a sample being a member (y axis of the last column in Figure 3a). If the privacy risk score truly corresponds to the probability that a sample is from a target model's training set, then we expect the actual values of privacy risk scores and fraction of training points in each bin to closely track with each other.

As a baseline to compare with, we also consider using NN based attacks to estimate privacy risks of individual samples. Prior papers suggest using the attack classifier's prediction to measure the input's privacy risk [20, 31]. The attack classifier has only one output, which is within [0, 1] and can serve as a proxy to estimate the probability of being a member. Following same steps as above, we compute the real probability of being a member and the average outputs of the attack classifier. Specifically, we follow Nasr et al. [31] to train the attack classifier by using the target model's predictions and one-hot encoded input labels as features.

Figure 3 shows the distribution of training samples' privacy risk scores (top row) and attack classifier's outputs on training data (bottom row) for Purchase100 classifiers without defense, with AdvReg [31], and with early stopping. We also compare the privacy risk score and attack classifier's output with the real probability of being a member, as shown in the last column of Figure 3 where the ideal case is used to check the effectiveness of metrics. We can see that our proposed privacy risk score closely aligns with the actual probability of being a member: the privacy risk score curves for all three models are quite close to the line of the ideal case. On the other hand, the attack classifiers' outputs fail to capture the membership probability. This is because the NN classifiers are trained to minimize the loss, i.e., the output of a member should be close to 1 while the output of a non-member should be close to 0. With this training goal, the obtained attack classifiers failed to capture the privacy risks for individual samples. We also quantitatively measure the root-mean-square error (RMSE) between estimated probability of member and real probability of member. On the three Purchase100 classifiers, the RMSE values of our privacy risk score are 0.05, 0.09, and 0.06; in contrast, the RMSE values of NN classifier's outputs are 0.26, 0.26, 0.25, respectively. We observe similar results on the undefended Texas100 classifier and the defended classifier by MemGuard [20], with details in Appendix C.

We also validate the effectiveness of privacy risk score across varied model architectures. For Purchase100 and

(a) The first three figures present the distributions of training samples' privacy risk scores on Purchase100 classifiers without defense, with AdvReg [31], and with early stoppping. The last figure shows that the privacy risk score can well represent the real probability of being a member, with root mean square error (RMSE) of 0.05, 0.09, and 0.06.

(b) The first three figures present the distributions of NN attack classifier's outputs over training samples on Purchase100 classifiers without defense, with AdvReg [31], and with early stoppping. The last figure shows that the NN classifier's output fails to represent the real probability of being a member, with RMSE values of 0.26, 0.26, and 0.25.

Figure 3: Estimate the probability of being a member with our proposed privacy risk score (Figure 3a), and with the NN attack classifier's output (Figure 3b).

Figure 4: Validation of privacy risk score with different model architectures on (undefended) Purchase100 (left), Texas100 (middle), and CIFAR100 (right) classifiers. For Purchase100 and Texas100 classifiers, the legend is expressed as (activation function, width, depth). The RSME values between privacy risk score (x-axis) and probability of being a member (y-axis) for all lines are smaller than 0.09.

Texas100 classifiers, we test two additional neural network depths by deleting the last hidden layer (depth=3) or adding one more hidden layer with 2048 neurons (depth=5); we test two additional neural network widths by halving the numbers of hidden neurons (width=0.5) or doubling the numbers of hidden neurons (width=2.0); we also test ReLU, Tanh, or Sigmoid as the activation functions. For CIFAR100 classifiers, besides DenseNet [18], we test other popular convolutional neural network architectures, including AlexNet [22], VGG [42], ResNet [14], and Wide ResNet [49]. As show in Figure 4, our proposed privacy risk score metric indeed well represents the likelihood of a sample being in the training set under different model architectures. On the Texas100 dataset, the classifier fails to learn meaningful features using the Sigmoid activation function, achieving an accuracy of only 4%, and is thus omitted from the figure. We provide validation results with defended classifiers in Appendix D.

4.2.2 Heterogeneity of members' privacy risk scores

After validating the effectiveness of the privacy risk score metric, we show the heterogeneity of training samples' privacy risks by plotting the cumulative distribution of their privacy risk scores. We also investigate the performance of mem-

bership inference defense methods [20, 31] with comparison between defended and undefended classifiers.

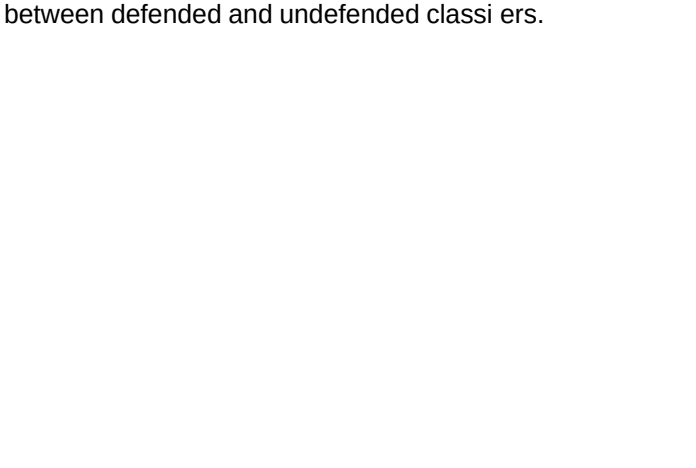

Figure 5: The cumulative distribution of privacy risk scores for Purchase100 classifiers in Nasr et al. [31].

Figure 5 presents the cumulative distributions of training points' privacy risk scores for Purchase100 classifiers. We can see that, compared with the undefended classifier, defended classifier with adversarial regularization [31] has smaller privacy risk scores on average. However, we can also see that the defended classifier has a small portion of training data with higher privacy risk scores than the undefended model: the undefended model has all members' privacy risk scores under 0.8, in contrast, the defended model has several training points with privacy risk scores higher than 0.8. Furthermore, the classifier with early stopping has a similar risk score distribution as the defended classifier.

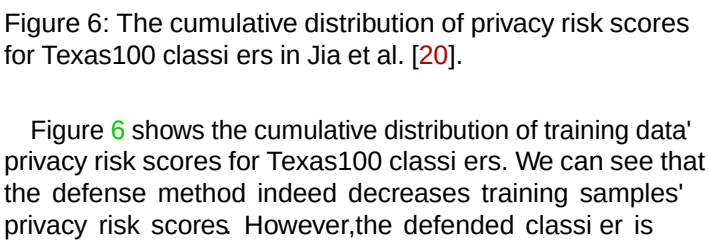

Figure 6: The cumulative distribution of privacy risk scores for Texas100 classifiers in Jia et al. [20].

Figure 6 shows the cumulative distribution of training data' privacy risk scores for Texas100 classifiers. We can see that the defense method indeed decreases training samples' privacy risk scores. However, the defended classifier is still quite vulnerable: 70% training samples have privacy risk scores higher than 0.6.

4.2.3 Usage of privacy risk score

From our definition and verification results in Section 4.2.1, we know that privacy risk score of a data point indicates its probability of being a member. Instead of pursuing high average attack accuracy, now the adversary can identify which samples have high privacy risks and perform attacks with high confidence: a sample is inferred as a member if and only if its privacy risk score is above a certain probability threshold.

We show the attack results with precision and recall values in Table 6 for target classifiers with varying threshold values on privacy risk scores. From Table 6, we can see that with larger threshold values on privacy risk scores, the adversary indeed has higher precision values for membership inference attacks. For MemGuard [20], when setting the same threshold value on privacy risk scores, both undefended and defended Texas100 classifiers have similar attack precision, but the defended classifier has a smaller recall value. However, the defended Texas100 classifier still has severe privacy risks: 70.5% training members can be inferred correctly with the precision of 71.3%, and 1.4% training members can be inferred correctly with the precision of 88.2%. Similarly, while adversarial regularization [31] can lower the average privacy risks, it increases the privacy risks for certain members: on the defended Purchase100 classifier, 0.2% training members can be inferred correctly with the precision of 83.3%. We urge designers of defense mechanisms to thus account for the full distribution of privacy risks in their analysis.

4.2.4 Impact of model properties on privacy risk score

We perform an in-depth investigation of privacy risk score by exploring its correlations with certain model properties, including sensitivity, generalization error, and feature embedding. We use the undefended Texas100 classifier from Jia et al. [20] for the following experiments.

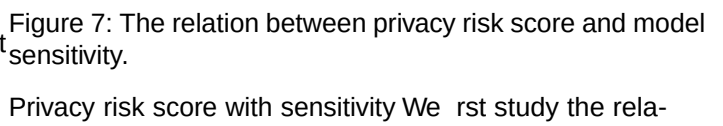

Figure 7: The relation between privacy risk score and model sensitivity.

Privacy risk score with sensitivity We first study the relationship between privacy risk scores and model sensitivity with regard to training samples. The sensitivity is defined as

Table 6: Membership inference attacks by setting different threshold values on privacy risk scores. For each threshold value, we report the (precision, recall) pair of membership inference attacks.

dataset	defense method	threshold values on privacy risk scores 1	0.9	0.8	0.7	0.6	0.5
Texas100	no defense	(85.4%, 21.2%)	(83.4%, 29.1%)	(81.2%, 45.3%)	(77.0%, 66.1%)	(72.8%, 85.4%)	(70.6%, 94.3%)
	MemGuard [20]	(88.2%, 1.4%)	(84.5%, 7.6%)	(82.6%, 18.7%)	(77.0%, 43.7%)	(71.3%, 70.5%)	(66.0%, 99.9%)
Purchase100	no defense	(N.A., 0%)	(N.A., 0%)	(N.A., 0%)	(63.4%, 7.8%)	(62.6%, 55.1%)	(61.6%, 90.9%)
	early stopping	(N.A., 0%)	(N.A., 0%)	(N.A., 0%)	(60.4%, 1.3%)	(57.1%, 20.2%)	(56.6%, 75.2%)
	AdvReg [31]	(83.3%, 0.2%)	(83.3%, 0.2%)	(65.9%, 0.5%)	(63.9%, 1.7%)	(58.9%, 24.6%)	(56.5%, 76.3%)

the influence of one training sample on the target model by computing the difference after removing that sample. Since the privacy risk score is obtained with the measured distributions of modified prediction entropy (Equation 15), we compute the model's sensitivity regard to a training point $z = (x; y)$ as the logarithm of $\frac{\text{Mentr}(F_z(x);y)}{\text{Mentr}(F(x);y)}$, where F_z means the retrained classifier after removing z from the training set.

Figure 7 shows the relation between privacy risk scores and the model sensitivity. For each privacy risk score, we show the first quartile, the average, and the third quartile of model sensitivities with regard to training data. We can see that, samples with higher privacy risk scores are likely to have a larger influence on the target model.

Privacy risk score with generalization error We observe that training samples with high risk scores are typically concentrated in a few class labels. Therefore, we further compare privacy risk scores among different class labels in this section.

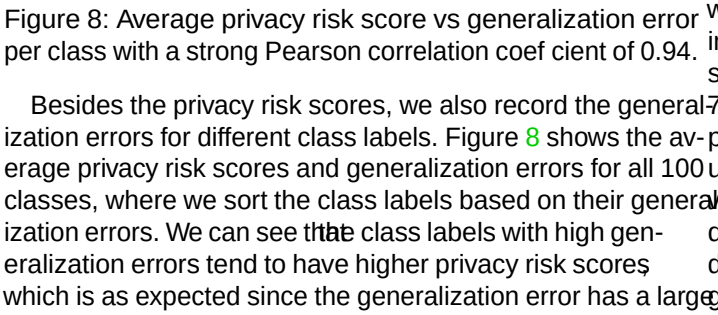

Figure 8: Average privacy risk score vs generalization error per class with a strong Pearson correlation coefficient of 0.94.

Besides the privacy risk scores, we also record the generalization errors for different class labels. Figure 8 shows the average privacy risk scores and generalization errors for all 100 classes, where we sort the class labels based on their generalization errors. We can see that the class labels with high generalization errors tend to have higher privacy risk scores, which is as expected since the generalization error has a large influence on the success of membership inference attacks [41]. The Pearson correlation coefficient between average privacy risk scores and generalization errors is as high as 0.94.

Privacy risk score with feature embeddings From the above experiment, we know that training samples from class labels with high generalization errors tend to have high privacy risk scores. Next, we investigate this further by looking into the feature representations of different class labels learned by the target classifier. We use the outputs of last hidden layer of the target classifier as the feature embedding of the input sample. We pick the top 5 class labels (30, 93, 97, 18, 98) with lowest average privacy risk scores (0.50, 0.52, 0.53, 0.54, 0.55) and at least 100 training samples, and the top 5 class labels (72, 49, 45, 51, 78) with highest average privacy risk scores (0.82, 0.83, 0.83, 0.85, 0.90) and at least 100 training samples. We record feature embeddings for both training and test examples from these 10 class labels. Finally, we adopt the t-Distributed Stochastic Neighbor Embedding (t-SNE) [27], a nonlinear dimensionality reduction technique, to visualize the feature embeddings.

Figure 9a and Figure 9b show the t-SNE plots of training samples and test samples, respectively. The training samples are separated clearly based on class labels since the target classifier has the training accuracy close to 100%. Test samples from class labels with low risk scores (classes 30, 93, 97, 18, 98) have quite similar feature embeddings as training samples and are still well separated. On the other hand, test samples from class labels with high risk scores (classes 72, 49, 45, 51, 78) exhibit differences in feature representations compared to corresponding training samples. From Figure 9, we also observe the heterogeneity of samples' privacy risks, in the granularities of both individual samples (e.g., different samples in class 78) and class labels (e.g., class 30 versus class 78). This further emphasizes the importance of fine-grained privacy risk analysis. It also validates our attack design of using class-dependent thresholds in Section 3.1. Our observations are also important for future defense work. A good defense approach should make training data and validation data have similar feature embeddings and consider the heterogeneity of samples' privacy risks.

(a) t-SNE plot for training samples in 10 class labels.

(b) t-SNE plot for test samples in 10 class labels

Figure 9: By using t-SNE [27], we visualize feature embeddings for both training samples and test samples. Training samples in the first 5 class labels have low privacy risk scores, and training samples in the last 5 class labels have high privacy risk scores.

5 Conclusions

In this paper, we first argue that measuring membership inference privacy risks with neural network based attacks is insufficient. We propose to use a suite of metric-based attacks, including existing methods with our improved class-specific thresholds and a new proposed method based on modified prediction entropy, for benchmarking privacy risks of machine learning models. We also make recommendations of comparing with early stopping when benchmarking a defense that introduces a tradeoff between model accuracy and privacy risks, and considering adaptive attackers with knowledge of the defense to rigorously evaluate the performance of defense approaches. With these benchmark attacks, we show that (1) the defense approach of adversarial regularization, proposed by Nasr et al. [31], only reduces privacy risks to a limited degree and is no better than early stopping; (2) the defense performance of MemGuard, proposed by Jia et al. [20], is greatly degraded with adaptive attacks.

Next, we introduce a new metric called the privacy risk score for a fine-grained analysis of individual samples' privacy risks. We show the effectiveness of the privacy risk score in estimating the true likelihood of an individual sample being in the training set and observe the heterogeneity of samples' privacy risk scores with experimental results. Finally, we perform an in-depth investigation about the correlation between privacy risks and model properties, including sensitivity, generalization error, and feature embeddings. We hope that our work convinces the research community about the importance of systematically and rigorously evaluating privacy risks of machine learning models.

Acknowledgements

We are grateful to anonymous reviewers at USENIX Security for valuable feedback. We would also like to thank Google's TensorFlow Privacy team for integrating our methods. This work was supported in part by the National Science Foundation under grants CNS-1553437 and CNS-1704105, the ARL's Army Artificial Intelligence Innovation Institute (A2I2), the Office of Naval Research Young Investigator Award, the Army Research Office Young Investigator Prize, Faculty research award from Facebook, Schmidt DataX award, and by Princeton E-affiliates Award.

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *ACM Conference on Computer and Communications Security*, 2016.

[2] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning*, 2018.

[3] Michael Backes, Pascal Berrang, Mathias Humbert, and Praveen Manoharan. Membership privacy in microrna-based studies. In *ACM Conference on Computer and Communications Security*, pages 319–330, 2016.

[4] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *USENIX Security Symposium*, pages 267–284, 2019.

[5] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *ACM Workshop on Artificial Intelligence and Security*, pages 3–14. ACM, 2017.

[6] Rich Caruana, Steve Lawrence, and Lee Giles. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. In *Advances in neural information processing systems*, pages 402–408, 2001.

[7] Dingfan Chen, Ning Yu, Yang Zhang, and Mario Fritz. Gan-leaks: A taxonomy of membership inference attacks against gans. In NeurIPS workshop on privacy in machine learning, 2019.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255, 2009.

[9] Cynthia Dwork. Differential privacy. In 33rd International Colloquium on Automata, Languages and Programming, part II. Springer Verlag, July 2006.

[10] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In Shai Halevi and Tal Rabin, editors, Theory of Cryptography, pages 265–284, 2006.

[11] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit con dence information and basic countermeasures. In ACM Conference on Computer and Communications Security, 2015.

[12] Karan Ganju, Qi Wang, Wei Yang, Carl A Gunter, and Nikita Borisov. Property inference attacks on fully connected neural networks using permutation invariant representations. In ACM Conference on Computer and Communications Security, pages 619–633, 2018.

[13] Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. Logan: Membership inference attacks against generative models. In Proceedings on Privacy Enhancing Technologies, number 1, 2019.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[15] Warren He, James Wei, Xinyun Chen, Nicholas Carlini, and Dawn Song. Adversarial example defense: Ensembles of weak defenses are not strong. In USENIX Workshop on Offensive Technologies, 2017.

[16] Benjamin Hilprecht, Martin Härterich, and Daniel Bernau. Monte carlo and reconstruction membership inference attacks against generative models. In Proceedings on Privacy Enhancing Technologies, 2019.

[17] Briland Hitaj, Giuseppe Ateniese, and Fernando Perez-Cruz. Deep models under the GAN: information leakage from collaborative deep learning. In ACM Conference on Computer and Communications Security, 2017.

[18] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2017.

[19] Bargav Jayaraman and David Evans. Evaluating differentially private machine learning in practice. In USENIX Security Symposium, pages 1895–1912, 2019.

[20] Jinyuan Jia, Ahmed Salem, Michael Backes, Yang Zhang, and Neil Zhenqiang Gong. Memguard: Defending against black-box membership inference attacks via adversarial examples. In ACM Conference on Computer and Communications Security, 2019.

[21] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classi cation with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105, 2012.

[23] Anders Krogh and John A Hertz. A simple weight decay can improve generalization. In Advances in neural information processing systems, pages 950–957, 1992.

[24] Klas Leino and Matt Fredrikson. Stolen memories: Leveraging model memorization for calibrated white-box membership inference. arXiv preprint arXiv:1906.11798, 2019.

[25] Changchang Liu, Xi He, Thee Chanyaswad, Shiqiang Wang, and Prateek Mittal. Investigating statistical privacy frameworks from the perspective of hypothesis testing. In Proceedings on Privacy Enhancing Technologies, pages 233–254, 2019.

[26] 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 preprint arXiv:1802.04889, 2018.

[27] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. Journal of machine learning research, 9(Nov):2579–2605, 2008.

[28] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In International Conference on Learning Representations, 2018.

[29] H Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. In International Conference on Learning Representations, 2018.

[30] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting unintended feature leakage in collaborative learning. In IEEE Symposium on Security and Privacy, 2019.

[31] Milad Nasr, Reza Shokri, and Amir Houmansadr. Machine learning with membership privacy using adversarial regularization. In *ACM Conference on Computer and Communications Security*, 2018.

[32] Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *IEEE Symposium on Security and Privacy*, 2019.

[33] Nicolas Papernot, Martín Abadi, Ulfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. In *International Conference on Learning Representations*, 2017.

[34] Lutz Prechelt. Early stopping-but when? In *Neural Networks: Tricks of the trade*. Springer, 1998.

[35] Apostolos Pyrgelis, Carmela Troncoso, and Emiliano De Cristofaro. Knock knock, who's there? Membership inference on aggregate location data. In *Network and Distributed Systems Security Symposium*, 2018.

[36] Md Atiqur Rahman, Tanzila Rahman, Robert Laganière, Noman Mohammed, and Yang Wang. Membership inference attack against differentially private deep learning model. *Transactions on Data Privacy*, 2018.

[37] Ahmed Salem, Apratim Bhattacharya, Michael Backes, Mario Fritz, and Yang Zhang. Updates-leak: Data set inference and reconstruction attacks in online learning. In *USENIX Security Symposium*, 2020.

[38] Ahmed 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. In *Network and Distributed Systems Security Symposium*, 2019.

[39] Virat Shejwalkar and Amir Houmansadr. Reconciling utility and membership privacy via knowledge distillation. *arXiv preprint arXiv:1906.06589*, 2019.

[40] Reza Shokri and Vitaly Shmatikov. Privacy-preserving deep learning. In *ACM Conference on Computer and Communications Security*, 2015.

[41] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy*, pages 3–18, 2017.

[42] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[43] Congzheng Song, Thomas Ristenpart, and Vitaly Shmatikov. Machine learning models that remember too much. In *ACM Conference on Computer and Communications Security*, pages 587–601, 2017.

[44] Liwei Song, Reza Shokri, and Prateek Mittal. Privacy risks of securing machine learning models against adversarial examples. In *ACM Conference on Computer and Communications Security*, 2019.

[45] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[46] Bingzhe Wu, Shiwan Zhao, ChaoChao Chen, Haoyang Xu, Li Wang, Xiaolu Zhang, Guangyu Sun, and Jun Zhou. Generalization in generative adversarial networks: A novel perspective from privacy protection. In *Advances in Neural Information Processing Systems*, 2019.

[47] Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

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

[49] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference*, 2016.

A Membership inference attacks against other datasets

Here, we perform membership inference attacks on two more image datasets: CH-MNIST and Car196. The CH-MNIST dataset contains histology tiles from patients with colorectal cancer.[7] The dataset contains 64 64 black-and-white images from 8 different classes of tissue, 5,000 samples in total. We use 2,000 data samples to train a convolution neural network. The model contains 2 convolution blocks with the number of output channels equal to 32 and 64. The classifier achieves 99.0% training accuracy and 71.7% test accuracy.

The Car196 dataset contains colored images of 196 classes of cars.[8] The dataset is split into 8,144 training images and 8,041 testing images. To train a model with good accuracy, we use a public ResNet50 [14] classifier pretrained on ImageNet [8] and fine-tune it on the Car196 training set. The classifier achieves 99.3% training accuracy and 87.5% test accuracy.

[7] https://www.kaggle.com/kmader/colorectal-histology-mnist
[8] https://ai.stanford.edu/~jkrause/cars/car_dataset.html