Improving Robustness to Model Inversion Attacks via Mutual Information Regularization

Tianhao Wang[1*], **Yuheng Zhang**[2*], **Ruoxi Jia**[3]
[1] Harvard University, [2] University of Illinois Urbana-Champaign, [3] Virginia Tech
tianhaowang@fas.harvard.edu, yuhengz2@illinois.edu, ruoxijia@vt.edu

Abstract

This paper studies defense mechanisms against model inversion (MI) attacks – a type of privacy attacks aimed at inferring information about the training data distribution given the access to a target machine learning model. Existing defense mechanisms rely on model-specific heuristics or noise injection. While being able to mitigate attacks, existing methods significantly hinder model performance. There remains a question of how to design a defense mechanism that is applicable to a variety of models and achieves better utility-privacy tradeoff.

In this paper, we propose the **M**utual **I**nformation Regularization based **D**efense (MID) against MI attacks. The key idea is to limit the information about the model input contained in the prediction, thereby limiting the ability of an adversary to infer the private training attributes from the model prediction. Our defense principle is model-agnostic and we present tractable approximations to the regularizer for linear regression, decision trees, and neural networks, which have been successfully attacked by prior work if not attached with any defenses. We present a formal study of MI attacks by devising a rigorous game-based definition and quantifying the associated information leakage. Our theoretical analysis sheds light on the inefficacy of DP in defending against MI attacks, which has been empirically observed in several prior works. Our experiments demonstrate that MID leads to state-of-the-art performance for a variety of MI attacks, target models and datasets.

Introduction

Machine learning (ML) techniques have revolutionized many fields, such as computer vision, natural language processing, and robotics. However, the application of ML to domains involving sensitive and proprietary datasets is currently hindered by potential privacy threats. Recent studies have demonstrated various privacy attacks, which can expose the information about private training data through the access to a target model. The access could either be whitebox or blackbox. In the whitebox setting, the adversary has complete knowledge of the target model, whereas in the blackbox setting, the adversary is just allowed to make prediction queries against the model. Both settings can find concrete examples in today's ML-as-a-service platforms, such as Tensorflow Hub which offers a library of models that users can download and Microsoft Azure Cognitive Services which produce predictions for user-input data.

This paper focuses on MI attacks, a type of privacy attacks aimed at reconstructing the input associated with any given model output. MI attacks have been used to recover images of any person from a face recognition model given just their name (Zhang et al. 2019) and infer the genetic markers of individuals based on the medicine dosage prediction and some demographic information (Fredrikson et al. 2014).

Existing defense mechanisms against MI attacks can be categorized into two threads. One thread studies model-specific heuristics to mitigate attacks. For example, for decision trees, Fredrikson, Jha, and Ristenpart (2015) studied the relationship between a sensitive feature's depth in the tree and the model's susceptibility to MI attacks and provided guidance for placing the sensitive feature in the tree. Although simple and efficient to implement, these heuristics cannot be easily generalized to a broader class of models and often only provide very limited protection against attacks. Another thread studies generic defense strategies that can be applied to any models. An example of such strategies is to train differentially private ML models. However, prior work (Fredrikson et al. 2014) has empirically shown that DP can mitigate the MI attack only when the injected noise is large enough and as a side effect, the model suffers significant performance degradation. There still remains a question of how to design a defense mechanism that is applicable to a variety of models and achieves better utility-privacy tradeoff.

In this paper, we present a defense mechanism against MI attacks, called MID, which can achieve the state-of-the art performance for a variety of target models, datasets, and attack algorithms in both blackbox and whitebox settings. Since MI adversaries exploit the correlation between the model input and output for successful attacks, our idea for the defense is to limit the redundant information about the model input contained in the prediction. Algorithmically, we introduce a regularizer into the training loss, which penalizes the mutual information between the model input and the prediction. We present tractable approximations of the regularizer for all the models that have been successfully attacked before, which include linear regression, decision trees and neural networks.

In addition, we provide a formal study of MI attacks and defenses. Particularly, existing theoretical framework for MI

[*]Equal Contribution

attacks focuses only on the privacy implications to training data while neglecting the fact that the attack also breaches privacy of other data from the same distribution. We take a first step toward formalizing the distributional privacy risk of MI attacks. Our theoretical analysis provides insights into the phenomenon of the inefficacy of DP in defending MI attacks observed in previous works. We evaluate our defense mechanism for various target models, datasets, and attack algorithms, and demonstrate the superiority of our defense to the existing methods in terms of utility-privacy tradeoff.

Related Work

Attack Algorithms. The first MI attack was demonstrated in (Fredrikson et al. 2014), where the authors presented a general algorithm to recover training data associated with an arbitrary model output given the model and some auxiliary information about the training set. The general idea of the algorithm is to formulate the MI attack as an optimization problem seeking for the sensitive input that achieves the maximum likelihood or posterior probability for the given model output and auxiliary information. Fredrikson et al. (2014) applied the algorithm to recover genetic markers given the linear regression model that uses them as part of input features. They also found that MI attacks are able to recover sensitive attributes for not only training data but also test data drawn independently from domain distribution. Fredrikson, Jha, and Ristenpart (2015) discussed the application of the general MI attack algorithm to more complex models including decision trees and some shallow neural networks. Zhang et al. (2019) proposed a whitebox attack algorithm that can distill generic knowledge from public data and leverage it to improve the realism of reconstructed images for deep neural networks. Yang, Chang, and Liang (2019) focused on the blackbox setting and proposed to train a separate model that swaps the input and output of the target model to perform MI attacks. Salem et al. (2019) studied the blackbox MI attacks for online learning, where the attacker has access to the versions of the target model before and after an online update and the goal is to recover the training data used to perform the update. They proposed to train a shadow model to mimic the target model, and then trains a separate model that transforms the change of the target model output in two online learning iterations into the private attributes.

Formalization of MI attacks. Several recent works started to formalize MI attacks and study the factors that affect a model's vulnerability from a theoretical viewpoint. Wu et al. (2016) characterized model invertibility for Boolean functions using the concept of influence from Boolean analysis; Yeom et al. (2018) formalized the risk that the model poses specifically to individuals in the training data and showed that the risk increases with the degree of model overfitting. Both works define the privacy loss of MI attacks as the extent of information leakage of training data exceeds that of test data. On the contrary, our paper recognizes the privacy loss suffered by the test data drawn from the same distribution as training data and presents a formalism of MI attacks that allows for the analysis of distributional privacy leakage.

Defenses. There are very few studies about defenses against MI attacks. One idea is to use DP (Fredrikson et al. 2014). DP provides a theoretical guarantee for training data privacy, but is not meant to protect the entire distribution. Zhang et al. (2019) and Fredrikson et al. (2014) observed through empirical studies that DP cannot provide protection against MI attacks with reasonable model utility. Our paper presents a theoretical analysis that can explain the inefficacy of DP. Other defense ideas are model-specific. For instance, Fredrikson, Jha, and Ristenpart (2015) proposed to place sensitive features at a particular depth to improve the robustness of decision trees. There also exists defenses designed specifically for blackbox attacks on neural networks, such as injecting uniform noise to confidence scores (Salem et al. 2019), reducing their precision (Fredrikson, Jha, and Ristenpart 2015) or dispersion (Yang et al. 2020). Our defense differs from existing methods in that our approach is model-agnostic and applicable to both whitebox and blackbox settings. Also, we will show that our approach can achieve better utility-privacy tradeoff than the existing methods.

Defense via Mutual Information Regularization

Problem Setup

In MI attacks, an adversary, given the access to a target model f trained to predict specific labels Y, uses it to infer the information about the training data distribution X. We denote the output of the target model as $\hat{Y}$, i.e., $\hat{Y} = f(X)$. In addition, the adversary may also have access to some auxiliary knowledge T that facilitates the inference. Consider the MI attacks against a face recognition model that labels an image containing a face with an identity corresponding to the individual depicted in the image. The attack goal is to recover a representative image for some target identity y based on the access to the target model. Possible auxiliary knowledge could be corrupted or blurred face images (Yang, Chang, and Liang 2019; Zhang et al. 2019).

For the face recognition example, both the recovery of a training image for the target identity in and that of a test image (i.e., the image that does not appear in the training set but drawn from the same distribution) would incur privacy loss to the individual. Hence, it is important to design defenses to ensure the privacy of the entire training distribution, rather than just the members in training set. The goal of our defense is to design an algorithm to train the target model f on the data distribution (X, Y) such that the access to the resulting model does not allow an adversary to infer the information about $P(X|\hat{Y} = y)$.

A naive defense is to produce a classifier where $\hat{Y}$ is independent of X. In this case, the adversary cannot learn anything about the input data distribution for a given label. However, this classifier would clearly be useless in practice. Hence, there is a tradeoff between privacy and model utility and we want a defense that can achieve the best tradeoff between the two.

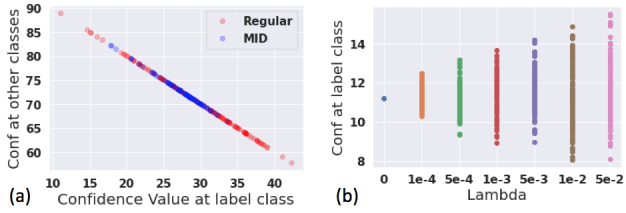

Figure 1: Illustration of the effect of penalizing $\mathcal{I}(X;\hat{Y})$ on the model output.

Algorithm

Intuitively, we will need to limit the dependency between X and $\hat{Y}$ to prevent the adversary from inferring the training data distribution associated with a specific label. Our idea is to quantify the dependence between X and $\hat{Y}$ using their mutual information $\mathcal{I}(X;\hat{Y})$ and incorporate it into the training objective as a regularizer. Specifically, our defense, which we call MID, trains the target model via the following loss function:

$$\min_{f\in\mathcal{H}} E_{(x,y)\sim p_{X,Y}(x,y)}[\mathcal{L}(y,f(x))] + \lambda\mathcal{I}(X,\hat{Y}) \quad (1)$$

where $\mathcal{I}(X,\hat{Y}) = \int_{\mathcal{X}}\int_{\mathcal{Y}} p_{X,Y}(x,y)\log(\frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)})dydx$, $\mathcal{L}(y,f(x))$ is the loss function for the main prediction task, and λ is the weight coefficient that controls the tradeoff between privacy and utility on the main prediction task.

To deconstruct the proposed regularizer, we re-write the mutual information as follows:

$$\mathcal{I}(X,\hat{Y}) = \mathcal{H}(\hat{Y}) - \mathcal{H}(\hat{Y}|X) \quad (2)$$

When f is a deterministic model, $\mathcal{H}(\hat{Y}|X) = 0$ and introducing the mutual information regularizer effectively reduces the entropy of the model output, i.e., $\mathcal{H}(\hat{Y})$. When f is stochastic, the regularizer will additionally promote the uncertainty of the model output for a fixed input, i.e., $\mathcal{H}(\hat{Y}|X)$. In practice, reducing $\mathcal{H}(\hat{Y})$ encourages the model output to be more concentrated and different inputs to be mapped into the same or very similar outputs; increasing $\mathcal{H}(\hat{Y}|X)$ makes the output to have a larger variance for a given x. Both terms will make X less likely to be determined from the $\hat{Y}$. Figure 1 illustrates this "two-pronged" privacy protection implied by our mutual information regularizer for a stochastic neural network consisting of a mean and a variance network trained on FaceScrub dataset. Figure 1 (a) plots the confidence value on the label class versus the sum of confidence on other classes for all test data with a particular label. It demonstrates that the regularizer will lead to more concentrated model outputs. Figure 1 (b) shows the simulated distribution of confidence values for a particular image through models trained with MID. We can see an increased trend of prediction variance for this input with increased λ, i.e. the strength of the MID regularizer.

Instantiations of MID

MID provides an appealing defense principle, as it defines what we mean by an effective defense, in terms of the fundamental tradeoff between reducing the input-output dependency and retaining good predictive power. However, computing mutual information is, in general, computational expensive. It requires modeling the joint distribution of model input and output and taking an integral over both domains, which is impracticable for most of the real-world prediction tasks. Here, we present efficient methods to implement the mutual information regularizer for different ML models that have been successfully attacked by previous work, including linear regression, decision trees, and neural networks. The unified idea underlying these methods is to find a tractable approximation to the mutual information regularizer.

Linear Regression. A linear regression model can be written as $\hat{y} = Ax$. Due to the deterministic nature of the model, the mutual information regularizer is reduced to $\mathcal{H}(\hat{Y})$. We proposed to approximate the distribution of $\hat{Y}$ by a Gaussian mixture:

$$p(\hat{y}) = \frac{1}{N}\sum_{i=1}^{N}\mathcal{N}(\hat{y}|Ax_i;\sigma^2) \quad (3)$$

where $\{x_i\}_{i=1}^N$ is the training set and σ is a free parameter. We utilizes a Taylor-expansion based approximation for the entropy of Gaussian mixtures described in Huber et al. (2008) and derive the following approximation to $\mathcal{I}(X,\hat{Y})$:

$$\tilde{\mathcal{I}}_{\text{lin}}(X,\hat{Y}) = \quad (4)$$
$$-\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2}(\frac{Ax_i - Ax_j}{\sigma})^2\right)\right)$$

Decision Trees. We modify the ID3 (Quinlan 1986), one of the most classic algorithms for training decision trees, to incorporate the mutual information regularizer. ID3 begins with the original training set as the root node. On each iteration of the algorithm, it iterates through every unused feature and calculates the splitting criterion (e.g. Information Gain or Gini Impurity), which measures the homogeneity of the labels within subsets. It then selects the feature achieving highest score based on the splitting criterion. The training set is then partitioned by the selected feature to produce subsets of the data. The algorithm continues to recurse on each subset and consider the features never selected before. In the inference phase, the decision tree is used to classify new test cases by traversing the tree using the features of the datum to arrive at a leaf node and label the datum with the most common class of the examples in the leaf node. Since decision trees trained with ID3 are deterministic, the mutual information regularizer again reduces to $\mathcal{H}(\hat{Y})$. To defend against the MI attacks, we add $-\mathcal{H}(\hat{Y})$ into the splitting criterion. Specifically, let A denote a feature and $\mathcal{C}(A)$ denote some homogeneity measure. We select the feature A that maximizes $\mathcal{C}(A) - \lambda\mathcal{H}(\hat{Y})$ to split on, where $\mathcal{H}(\hat{Y})$ is evaluated empirically with the training set.

Neural Networks. To come up with a tractable approximation to $\mathcal{I}(X;\hat{Y})$ for neural networks, we get inspiration

from the line of work on information bottleneck (Shwartz-Ziv and Tishby 2017; Alemi et al. 2016) and regard the neural network as a Markov chain $Y - X - Z - \hat{Y}$, where X is the feature, Y is the ground truth label, Z is a stochastic encoding of the input X at some intermediate layer and defined by $P(Z|X;\theta)$, and $\hat{Y}$ is the prediction. By the data processing inequality (Cover 1999), we have $\mathcal{I}(X,\hat{Y}) \leq \mathcal{I}(X,Z)$. Prior work has provided various efficient approximation to $\mathcal{I}(X,Z)$ (Alemi et al. 2016; Kolchinsky, Tracey, and Wolpert 2019). Thus, we replace $\mathcal{I}(X,\hat{Y})$ with its upper bound $\mathcal{I}(X,Z)$ in the training objective and train the neural network with the following loss function:

$$\min_{\theta} -\mathcal{I}(Z;Y) + \lambda \mathcal{I}(Z,X) \tag{5}$$

The first term encourages the learned encoding to be maximally informative about the label Y and measures the prediction performance of the model. The second term reduces the dependency between X and $\hat{Y}$ by minimizing its upper bound and improves the robustness against the MI attacks. Note that the training objective above boils down to the classic information bottleneck and we will employ the variational method (Alemi et al. 2016) to approximate the mutual information terms in the training objective.

Theoretical Analysis

In this section, we will formalize the MI attacks and quantify its *distributional* privacy loss. Then, we will try to provide a theoretical basis for the recent observation that DP – a canonical privacy notion nowadays – cannot provide protection against the MI attacks with reasonable model utility (Zhang et al. 2019; Fredrikson et al. 2014).

Formalizing MI Attacks

We present a methodology for formalizing the MI attacks. Unlike previous works that capture the privacy loss of *members in the training set* (Wu et al. 2016; Yeom et al. 2018), this is the first attempt of modeling the privacy loss of *members in the population*.

We have been mainly focused on the type of attacks that aim to recover all dimensions of the feature vector corresponding to a given label. Prior work on the MI attacks has also considered other types of attacks that only aim to reconstruct partial dimensions of the feature vector and have access to the remaining dimensions as auxiliary knowledge. For instance, Fredrikson et al. (2014) studied the MI attacks against a linear regression model that predict the medicine dosage based on the input of genotypes and some nonsensitive data like demographic information. To subsume this attack setting under our general formalization, we let the input random variable consists of the sensitive and nonsensitive part, i.e., $X = (X_s, X_{ns})$, and the attacker has access to X_{ns}.

The attacker could also be interested in some specific property of the feature vector instead of the entire dimensions. In the face recognition example, the attacker could be interested in learning about the properties of a face image, such as hair color and race, instead of reconstructing the entire face image (Zhang et al. 2019). To capture this common attack scenario, we introduce τ to denote the property function that maps the sensitive feature to the property of the interest to the attacker.

We abstract the MI attacks into the following semantic MI game played between the server and the adversary. The game is described by a tuple $(\mathcal{A}, \mathcal{M}, p, \tau)$, where $\mathcal{A}$ denotes the adversary which is a probabilistic machine with access to the learning algorithm $\mathcal{M}$ and p is the target data distribution. $\mathcal{M} : \mathcal{X}^n \to \mathcal{H}$ is a learning algorithm that takes a dataset as input and outputs a single, fixed classifier f. $\mathcal{H}$ represents the hypothesis class.

Experiment 1 (Semantic MI $\mathbf{Exp}^{\mathbf{SEM}}(\mathcal{A}, \mathcal{M}, p, \tau)$)

1. *Server draws a training set $S \sim p^n$, and trains a classifer $f \leftarrow \mathcal{M}(S)$.*
2. *Server draws $z = (x_{ns}, x_s, y) \sim p$. (f, x_{ns}, y) is presented to the adversary $\mathcal{A}$.*
3. *The adversary outputs $\mathcal{A}_{x_{ns},y}$. $\mathbf{Exp}^{\mathbf{SEM}}(\mathcal{A}, \mathcal{M}, p, \tau)$ is 1 if $\mathcal{A}_{x,y} = \tau(x_s)$, and 0 otherwise.*

The *gain* of the game is evaluated as

$$\begin{aligned} \texttt{gain}^{\mathbf{SEM}}(\mathcal{A}, \mathcal{M}, p, \tau) &= Pr[\mathbf{Exp}^{\mathbf{SEM}}(\mathcal{A}, \mathcal{M}, p, \tau) = 1] \\ &= Pr[\mathcal{A}_{x_{ns},y} = \tau(x_s)] \end{aligned} \tag{6}$$

where the probability is taken over the randomness of $S \sim p^n$, the randomness of $\mathcal{M}$, the randomness of $\mathcal{A}$ and the randomness of $(x_s, x_{ns}, y) \sim p$. This game directly formalizes the procedure of MI attack. Trivially, for this game, the best strategy for the adversary is always output the most probable $\tau(x_s)$ when $x_s \sim p_{X_s|x_{ns},y}$. Hence, the best possible gain for this game is

$$E_{(x,y)\sim p}[\max_{v} Pr_{x_s \sim p_{X_s|x_{ns},y}}[\tau(x_s) = v]] \tag{7}$$

To properly measure the adversary's advantage gained from the classifier f, we define an alternative underlying distribution $p_{X_s} \times p_{X_{ns},Y}$ such that X_s and (X_{ns}, Y) have no correlation with each other. That is, sampling $x_s \sim p_{X_s}$ independently from X_{ns} and Y. We define the advantage of $\mathcal{A}$ to be

$$\begin{aligned} \texttt{Adv}^{\mathbf{SEM}}(\mathcal{A}, \mathcal{M}, p, \tau) = \texttt{gain}^{\mathbf{SEM}}(\mathcal{A}, \mathcal{M}, p, \tau) \\ - \texttt{gain}^{\mathbf{SEM}}(\mathcal{A}, \mathcal{M}, p_{X_s} \times p_{X_{ns},Y}, \tau) \end{aligned} \tag{8}$$

i.e. $\texttt{gain}^{\mathbf{SEM}}(\mathcal{A}, \mathcal{M}, p_{X_s} \times p_{X_{ns},Y}, \tau)$ is the trivial baseline gain of the adversary when there are indeed no correlation between X_s and (X_{ns}, Y), and $\texttt{Adv}^{\mathbf{SEM}}$ measures the extra gain the adversary can get. Note that there are multiple other ways to define the adversary's advantage which capture different insights. We defer this discussion to the Appendix.

Protection from Differential Privacy

One dominate privacy notion is DP, which carefully randomizes an algorithm so that its output does not to depend too much on any single individual in the dataset (Dwork, Roth et al. 2014).

Definition 1 (Differential Privacy)**.** *Let $\mathcal{M} : \mathcal{X}^n \to \mathcal{R}$ be a randomized mechanism. We say that $\mathcal{M}$ is (ϵ, δ)-differentially private if for every two adjacent datasets $S \sim S'$ and every subset $R \subseteq \mathcal{R}$,*

$$Pr[\mathcal{M}(S) \in R] \leq e^{\epsilon} Pr[\mathcal{M}(S') \in R] + \delta \tag{9}$$

We introduce an indistinguishability game to derive the robustness guarantee offered by DP. The game is described by a tuple $(\mathcal{A}, \mathcal{M}, p)$.

Experiment 2 (Indistinguishability $\mathbf{Exp}^{\mathbf{IND}}(\mathcal{A}, \mathcal{M}, p)$)

1. *Server chooses* $b \leftarrow \{0, 1\}$ *uniformly at random.*
2. *Server draws training set* $S \sim (p)^n$ *if* $b = 0$*;* $S \sim (p_{X_s} \times p_{X_{ns},Y})^n$ *if* $b = 1$.
3. *Server trains a classifier* $f \leftarrow \mathcal{M}(S)$ *and presents* f *to the adversary.*
4. $\mathcal{A}$ *outputs* $b' \in \{0, 1\}$. $\mathbf{Exp}^{\mathbf{IND}}(\mathcal{A}, \mathcal{M}, p)$ *is* 1 *if* $b = b'$*, and 0 otherwise.*

Let $\texttt{gain}^{\mathbf{IND}}(\mathcal{A}, \mathcal{M}, p) = Pr[\mathbf{Exp}^{\mathbf{IND}}(\mathcal{A}, \mathcal{M}, p) = 1]$. The following theorem shows that the gain of the semantic MI game can be upper bounded in terms of the best possible gain of this indistinguishability game.

Theorem 1. *For any attack strategy* $\mathcal{A}^*$*, learning algorithm* $\mathcal{M}$*, target distribution* p *and property function* τ*,*

$$Adv^{\mathbf{SEM}}(\mathcal{A}^*, \mathcal{M}, p, \tau) \leq 2 \max_{\mathcal{A}} gain^{\mathbf{IND}}(\mathcal{A}, \mathcal{M}, p) - 1 \tag{10}$$

Hence, to mitigate the threats of MI attack, we want $\max_{\mathcal{A}} \texttt{gain}^{\mathbf{IND}}$ to be not much greater than $1/2$.

We will provide a tight bound of $\texttt{gain}^{\mathbf{IND}}$ for any attack strategy $\mathcal{A}$ when the learning algorithm is differentially private. In the analysis, we assume that with high probability a training set drawn from p^n has no intersection with another set drawn from $(p_{X_s} \times p_{X_{ns},Y})^n$, i.e.,

$$Pr_{S \sim p^n, S' \sim (p_{X_s} \times p_{X_{ns},Y})^n}[S \sim_n S'] = 1 - \gamma \tag{11}$$

where $S \sim_n S'$ indicates that the two datasets $S, S' \in \mathcal{X}^n$ differ by n entries, and γ is a small value. This assumption is plausible for many practical scenarios where the feature vector has continuous domain or is high-dimensional.

Theorem 2. *If the learning algorithm* $\mathcal{M} : \mathcal{X}^n \rightarrow \mathcal{R}$ *is* (ϵ, δ)*-differentially private, then with probability at least* $1-\gamma$ *we have tight bound* $\max_{\mathcal{A}} gain^{IND}(\mathcal{A}, \mathcal{M}, p) \leq \frac{e^{n\epsilon}}{e^{n\epsilon}+1} + \frac{e^{n\epsilon}-1}{(e^{n\epsilon}+1)(e^{\epsilon}-1)}\delta$.

As we can see, to make the upper bound $\frac{e^{n\epsilon}}{e^{n\epsilon}+1} + \frac{e^{n\epsilon}-1}{(e^{n\epsilon}+1)(e^{\epsilon}-1)}\delta = \frac{1}{2} + \frac{e^{n\epsilon}-1}{2(e^{n\epsilon}+1)} + \frac{e^{n\epsilon}-1}{(e^{n\epsilon}+1)(e^{\epsilon}-1)}\delta \in \frac{1}{2} + o(1)$, the privacy budget ϵ needs to be set as $o(\frac{1}{n})$. However, this privacy budget is too small to allow any useful computation (see, e.g., Section 2.3.3 of Dwork, Roth et al. (2014), for a comprehensive review). Since the bound in Theorem 2 is tight, with high probability DP cannot mitigate the worst-case MI attacks with any reasonable model utility. We stress that large ϵ implies large attacker advantage only for worst-case attacks; for other attacks, their advantage might not monotonically increase with ϵ. However, in practice, a defender can hardly know the actual attack strategy; hence, defense methods that are effective against worst-case attacks are favored.

Experiments

In this section, we present the empirical evaluation of the efficacy of MID against different attacks and compare it with existing defense mechanisms.

Experiment Setting

Table 1 summarizes our experimental setting, including the attacks, models, datasets, metrics used in our evaluation as well as the other defenses that we compare with. We leave the description of datasets, attack hyper-parameters, model architectures as well as the details of the evaluation process and metrics to the Appendix.

Attack algorithms. We compare the efficacy of MID against the following MI attacks, which are the most effective ones presented in the literature thus far.

- MAP (Fredrikson et al. 2014; Fredrikson, Jha, and Ristenpart 2015) casts the MI attack as an optimization problem that seeks for the maximum a posteriori probability (MAP) estimate of the sensitive attribute under the target model. For decision trees, if the attacker also has access to the number of training examples for each path in the tree, it can further improve the attack performance by improving the MAP estimate (referred as "white-box with counts" in (Fredrikson, Jha, and Ristenpart 2015)). We dub these two attacks *Naive MAP* and *MAP with counts*, respectively.
- Knowledge Alignment (Yang, Chang, and Liang 2019) is a blackbox MI attack, in which the adversary trains an inversion model that swaps the input and output of the target model using the dataset drawn from a distribution similar to the private data distribution. The inversion model is then used to reconstruct the input feature for any given target model output.
- Update Leaks (Salem et al. 2019) is a blackbox MI attack designed for online ML models. It adopts a similar idea to (Yang, Chang, and Liang 2019) and trains an inversion model to reconstruct the private data. The difference from (Yang, Chang, and Liang 2019) is that the input to the inversion model now is the change of the target model output before and after an online update; moreover, the inversion model leverages the generative adversarial networks to improve the reconstruction accuracy.
- Generative MI (GMI) (Zhang et al. 2019) is a whitebox MI attack achieving the state-of-the-art performance against DNNs. GMI solves the MAP to recover the most possible private attribute via gradient descent. The key idea of GMI is to leverage public data to learn a generic prior for the private training data distribution and use it to regularize the optimization problem underlying the attack.

Defense baselines. We compare our defense with DP for all attack algorithms, as well as other existing defenses available for specific models or attacks. We implement the AdaSSP, differentially private ID3, and DPSGD in (Wang 2018; Friedman and Schuster 2010; Abadi et al. 2016) to construct differentially private linear regression models, decision trees, and neural networks, respectively. As for more specialized defenses, we compare with a defense proposed in (Fredrikson et al. 2014) for decision trees, which adjusts the depth of the sensitive features. We will use "Priority" to symbolize this defense. Previous work has also proposed to defend against blackbox attacks against neural networks by

	Attack	Model	Dataset	Defense Baseline	Utility Metrics	Attack Metrics
Blackbox	Naive MAP	Linear Regression	IPWC	DP-AdaSSP	MSE	Acc; AUROC
	Naive MAP	Decision Tree	FiveThirtyEight	DPID3; Priority	F1	F1
	Knowledge Alignment	Neural Networks	FaceScrub	DPSGD	Accuracy	Acc, L2, ECE
	Update-Leaks	Neural Networks	CIFAR10	DPSGD	Accuracy	Acc, ECE
Whitebox	MAP with counts	Decision Tree	FiveThirtyEight	DPID3; Priority	F1	F1
	GMI	Neural Networks	CelebA	DPSGD	Accuracy	Acc, L2

Table 1: Summary of experimental settings.

rounding the output confidence scores (Fredrikson, Jha, and Ristenpart 2015) or injecting uniform noise to the confidence vector (Salem et al. 2019). However, our experiments found that these two defenses are not effective to protect against the attacks considered in this paper regardless of the rounding precision or the noise magnitude, respectively. Hence, we do not exhibit the results for these two defenses.

Evaluation Protocol. We evaluate the performance of a defense mechanism in terms of the privacy-utility tradeoff. All the defenses considered in our paper have some hyperparameters which we could tune to vary the robustness and model performance. For the MID and DP, we can vary the weight parameter λ in Equation 1 and the privacy budget ϵ, respectively. For Priority, we can vary the depth of the tree where we place the sensitive features. In our experiment, we vary these hyperparameters and generate a utility-privacy tradeoff curve for each defense. We then use these curves to compare the performance of different defenses. When the target model is linear regression and a decision tree, we generate 100 models with different training and test data split and average the utility and privacy results over different models for each defense strategy and hyperparameter setting. When the target model is a neural network, we train 3 models and average the results. To illustrate the distributional privacy leakage, we demonstrate the attack performance on both training and test set for all attacks except GMI and Update-Leaks, since these two attacks aim at constructing a representative images for a given label instead of reconstruct a particular image in the training or test set. To compare the privacy risk of training and test data for GMI and Update-Leaks, we calculate the L_2 distances from deep feature representation of reconstructed images to that of training images and test images.

Evaluation Metrics. The evaluation metrics for the model utility and attack performance are listed in Table 1. The utility metrics depend on the underlying prediction task and datasets. For regression tasks, we employ *mean squared error* (MSE); for classification tasks, we generally use *test accuracy* as the metric unless the dataset is highly imbalanced in which case we use the *F1 score*. For blackbox attacks on neural networks, there is a trivial defense to just release the prediction rather than the entire confidence vector and this defense achieves good utility in terms of test accuracy. However, this defense omits useful confidence information. To capture to what extent the defense worsen the confidence output, we leverage the *Expected Calibration Error* (ECE) (Guo et al. 2017), a metric commonly used to measure the miscalibration between confidence and accuracy. As for the attack performance, we follow the papers where the attacks were originally proposed and mostly use their metrics. Specifically, attack performance metrics for discrete attribute include *accuracy*, *AUROC*, and *F1-score*. The latter two are used when the sensitive attributes are highly imbalanced. To evaluate the attack performance for images, we build an evaluation classifier to see whether the image can be recognized as the target class and compute the *attack accuracy*. For all the metrics above, the high value they take, the better attack performance. We also calculate the L_2 distances from deep feature representation of reconstructed images to that of training images and test images in order to compare the privacy threats to training and test data.

Results

Figure 2 compares the performance of our defense against the existing baselines on various blackbox attacks, including Naive MAP (Figure 2 (a-b)), Knowledge Alignment (Figure 2 (c)), and Update-Leaks (Figure 2 (d)). For Naive MAP, we demonstrate the results on both linear regression and decision trees as these two models were successfully attacked by Naive MAP in prior work (Fredrikson et al. 2014; Fredrikson, Jha, and Ristenpart 2015). Knowledge Alignment and Update-Leaks are designed for attacking neural networks, so we only exhibit the results on neural networks for these two attacks. Figure 2 shows that MID can consistently achieve a better utility-privacy tradeoff than DP for protecting against all the attacks. For decision trees, there exists a model-specific defense strategy, named Priority, which adjusts the depth of the sensitive feature. Figure 2 (b) shows that MID is more robust than Priority in most cases except when model has a very high performance (i.e., F1 score > 0.57). Nevertheless, Priority has its own drawbacks: firstly, it is only applicable to decision trees; and second, it only protects a subset of attributes, or the sensitive attributes, which could be subject to additional privacy concern because one may leverage the correlation between nonsensitive attributes and sensitive attributes to recover the sensitive one. On the other hand, the principles of MID and DP are applicable to different types of models and protect all the attributes at once. A phenomenon consistently present in all attacks is that the more predictive power the model has, the more vulnerable it is to the attacks, regardless of the types of defenses attached to the model. This finding signifies the difficulty to defend against the MI attacks; yet, our defense can significantly improve the model robustness for any fixed model performance. We also evalu-

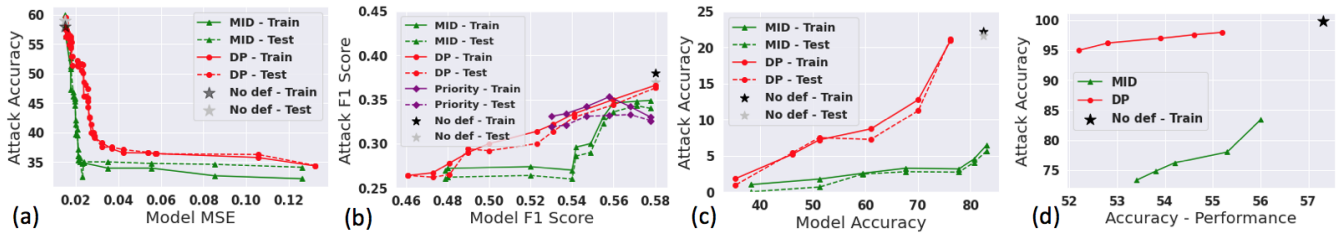

Figure 2: Defense result for blackbox MI attacks, including (a-b) Naive MAP on linear regression and decision trees, (c) Knowledge Alignment and (d) Update-Leaks on neural networks.

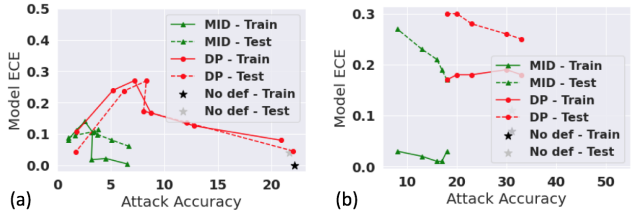

Figure 3: Confidence calibration of defenses evaluated on the (a) Knowledge Alignment and the (b) Update-Leaks attack.

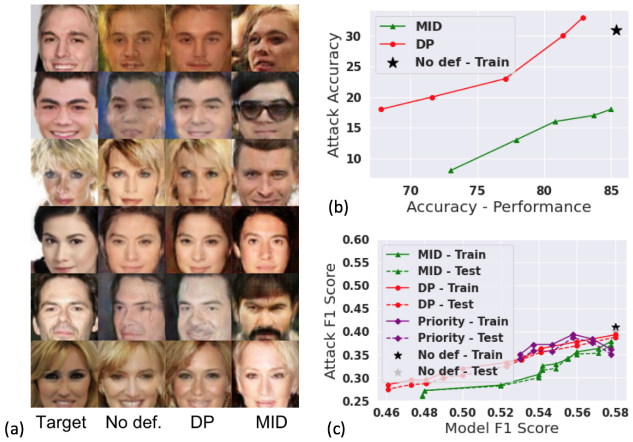

Figure 4: Defense results for whitebox MI attacks, including (a-b) GMI on neural networks and (c) MAP with counts on decision trees.

ate the attack performance in terms of other metrics shown in Table 1. As they give the similar results to the metrics exhibited in Figure 2, we will omit these results to the Appendix. Figure 3 illustrates the miscalibration between confidence and accuracy as the result of different defenses. The degree of miscalibration is measured by ECE. We find that MID mostly achieves significantly lower ECE error than DP for a given robustness level.

Figure 4 illustrates the defense results for whitebox attacks. Specifically, we consider GMI, the state-of-the-art attack against neural networks, and MAP with counts, the only known attack against decision trees; the corresponding privacy-utility tradeoff curves are shown in Figure 4 (b) and (c), respectively. We find that MID can outperform DP by a large margin to defend against whitebox MI attacks. Similar to the observation in blackbox experiment, model-specific defense strategies could achieve better robustness than MID occasionally; but overall, MID could be more preferable due to its broad applicability, thorough protection for all attributes, and reliable performance. Figure 4 (a) compares the reconstructed images of MID and DP. We tune the hyperparameters of both defenses so that the resulting target models have comparable performance. The test accuracies for the models trained with MID and DP are 80.8 and 81.4, respectively. We can see that MID can block the attack much better than DP. For instance, the reconstructions for DP can still retain many facial features of the target individual while the reconstructions for MID are almost completely different from the target individual.

One common observation from all defense results is that the attack performance is very close on training and test set. This implies that MI attacks pose similar privacy threats regardless of whether an individual's data is selected for training or not. Although some prior works (Yeom et al. 2018; Dwork et al. 2015) report that the attack performance on the training set is much higher than the test set, we conjecture that it is because the model overfits the training data in their evaluations. In our experiments, since both MID and DP can help mitigate overfitting, the privacy threats of MI attacks is presented to the population instead of the training set only.

Conclusion

We propose a defense against MI attacks based on regularizing the mutual information between the model input and prediction and further present tractable approximations to the regularizer for linear regression, decision trees, and neural networks. We provide theoretical basis for a common empirical observation that DP cannot defend against MI attacks with reasonable model utility. We perform experiments to compare the utility-privacy tradeoff of our defense against existing baselines on different models and datasets, and demonstrate that our proposed defense can achieve the state-of-the-art performance to protect against various attacks in both whitebox and blackbox settings. One limitation of our defense is that it does not prevent the attack but instead aim for better utility-privacy tradeoff. For future work, we would like to further improve our defense from the perspective of computational security, i.e., by assuming a polynomial time adversary.

References

Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H. B.; Mironov, I.; Talwar, K.; and Zhang, L. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 308–318.

Alemi, A. A.; Fischer, I.; Dillon, J. V.; and Murphy, K. 2016. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410* .

Cover, T. M. 1999. *Elements of information theory*. John Wiley & Sons.

Dwork, C.; Feldman, V.; Hardt, M.; Pitassi, T.; Reingold, O.; and Roth, A. 2015. Generalization in adaptive data analysis and holdout reuse. In *Advances in Neural Information Processing Systems*, 2350–2358.

Dwork, C.; Roth, A.; et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science* 9(3–4): 211–407.

Fredrikson, M.; Jha, S.; and Ristenpart, T. 2015. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, 1322–1333.

Fredrikson, M.; Lantz, E.; Jha, S.; Lin, S.; Page, D.; and Ristenpart, T. 2014. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In *23rd {USENIX} Security Symposium ({USENIX} Security 14)*, 17–32.

Friedman, A.; and Schuster, A. 2010. Data mining with differential privacy. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 493–502.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. *arXiv preprint arXiv:1706.04599* .

Huber, M. F.; Bailey, T.; Durrant-Whyte, H.; and Hanebeck, U. D. 2008. On entropy approximation for Gaussian mixture random vectors. In *2008 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, 181–188. IEEE.

Kolchinsky, A.; Tracey, B. D.; and Wolpert, D. H. 2019. Nonlinear information bottleneck. *Entropy* 21(12): 1181.

Quinlan, J. R. 1986. Induction of decision trees. *Machine learning* 1(1): 81–106.

Salem, A.; Bhattacharya, A.; Backes, M.; Fritz, M.; and Zhang, Y. 2019. Updates-leak: Data set inference and reconstruction attacks in online learning. *arXiv preprint arXiv:1904.01067* .

Shwartz-Ziv, R.; and Tishby, N. 2017. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810* .

Wang, Y.-X. 2018. Revisiting differentially private linear regression: optimal and adaptive prediction & estimation in unbounded domain. *arXiv preprint arXiv:1803.02596* .

Wu, X.; Fredrikson, M.; Jha, S.; and Naughton, J. F. 2016. A methodology for formalizing model-inversion attacks. In *2016 IEEE 29th Computer Security Foundations Symposium (CSF)*, 355–370. IEEE.

Yang, Z.; Chang, E.-C.; and Liang, Z. 2019. Adversarial neural network inversion via auxiliary knowledge alignment. *arXiv preprint arXiv:1902.08552* .

Yang, Z.; Shao, B.; Xuan, B.; Chang, E.-C.; and Zhang, F. 2020. Defending Model Inversion and Membership Inference Attacks via Prediction Purification. *arXiv preprint arXiv:2005.03915* .

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

Zhang, Y.; Jia, R.; Pei, H.; Wang, W.; Li, B.; and Song, D. 2019. The Secret Revealer: Generative Model-Inversion Attacks Against Deep Neural Networks. *arXiv preprint arXiv:1911.07135* .