UNIVERSITY OF LIVERPOOL

Membership Inference Attacks and Differential Privacy

Outline

- Introduction on Membership Inference Attacks

(1) Typical Approach: Shadow model
(2) Threshold-based Membership Inference Attacks

- Differential Privacy

(1) Definition and Properties
(2) DP-SGD
(2) PATE

- Visual Prompting

Membership Inference Attacks

- Attack goal: determine whether an individual data example is inside the training dataset of the target model or not

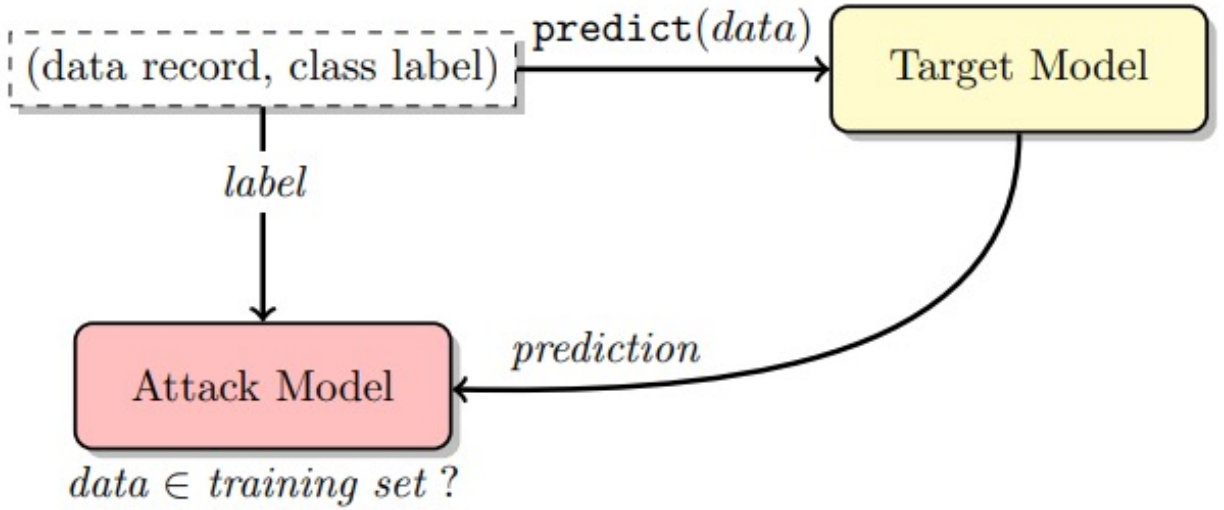
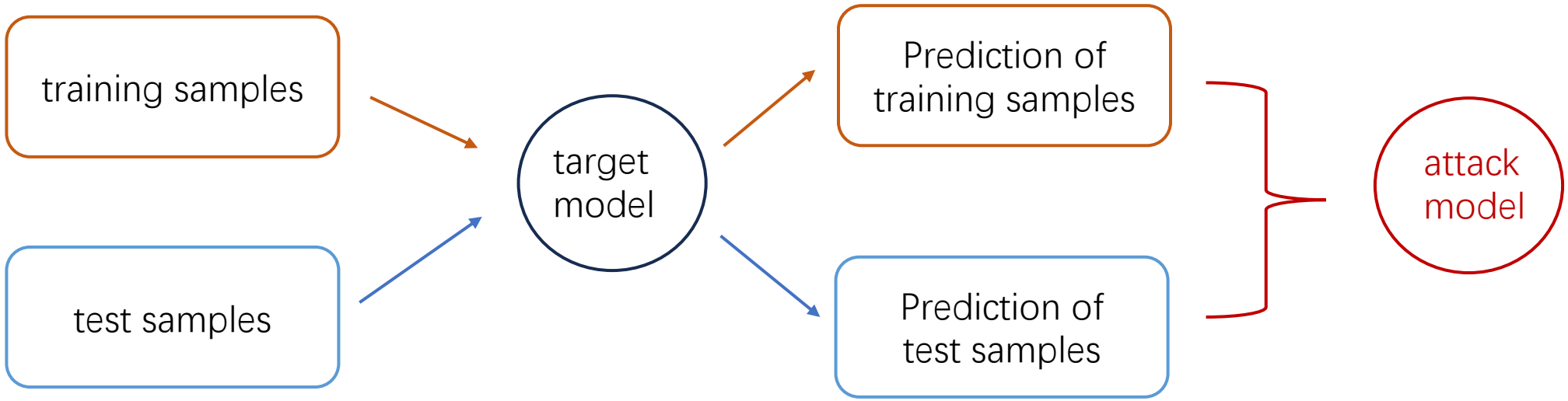


Fig. 1: Membership inference attack in the black-box setting. The attacker queries the target model with a data record and obtains the model's prediction on that record. The prediction is a vector of probabilities, one per class, that the record belongs to a certain class. This prediction vector, along with the label of the target record, is passed to the attack model, which infers whether the record was *in* or *out* of the target model's training dataset.

Shokri R, Stronati M, Song C, et al. Membership inference attacks against machine learning models[C]//2017 IEEE symposium on security and privacy (SP). IEEE, 2017: 3-18.

Membership Inference Attacks

- under the General Data Protection Regulation (GDPR), MIAs can increase the risks that private personal information can be inferred from publicly accessible ML models
- First MIA: *Shokri et al.* proposed the first MI attack for classification models in the context of ML, which utilized all features of multiple shadow models to train a binary classifier-based attack model in a black-box scenario

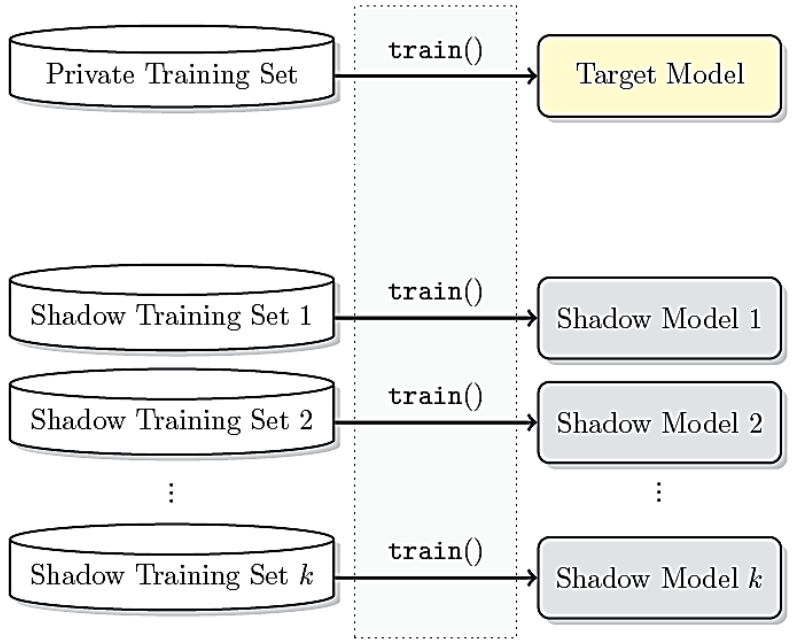


Shadow Training Attack

- Threat model:
  - The adversary has back-box query access to the target model
  - The goal is to infer whether input samples were part of its private training set
- ***Shadow training*** approach:
  - Create several shadow models to substitute the target model
  - Each shadow model is trained on a dataset that has a similar distribution as the private training dataset of the target model

E.g., if the target model performs celebrity face recognition, the attacker can collect images of celebrities from the Internet

Then, query the target model with images of Brad Pitt, and if the confidence of the target model is high, then probably the private training set contains images of Brad Pitt: use those images for the shadow training sets



Shokri R, Stronati M, Song C, et al. Membership inference attacks against machine learning models[C]//2017 IEEE symposium on security and privacy (SP). IEEE, 2017: 3-18.

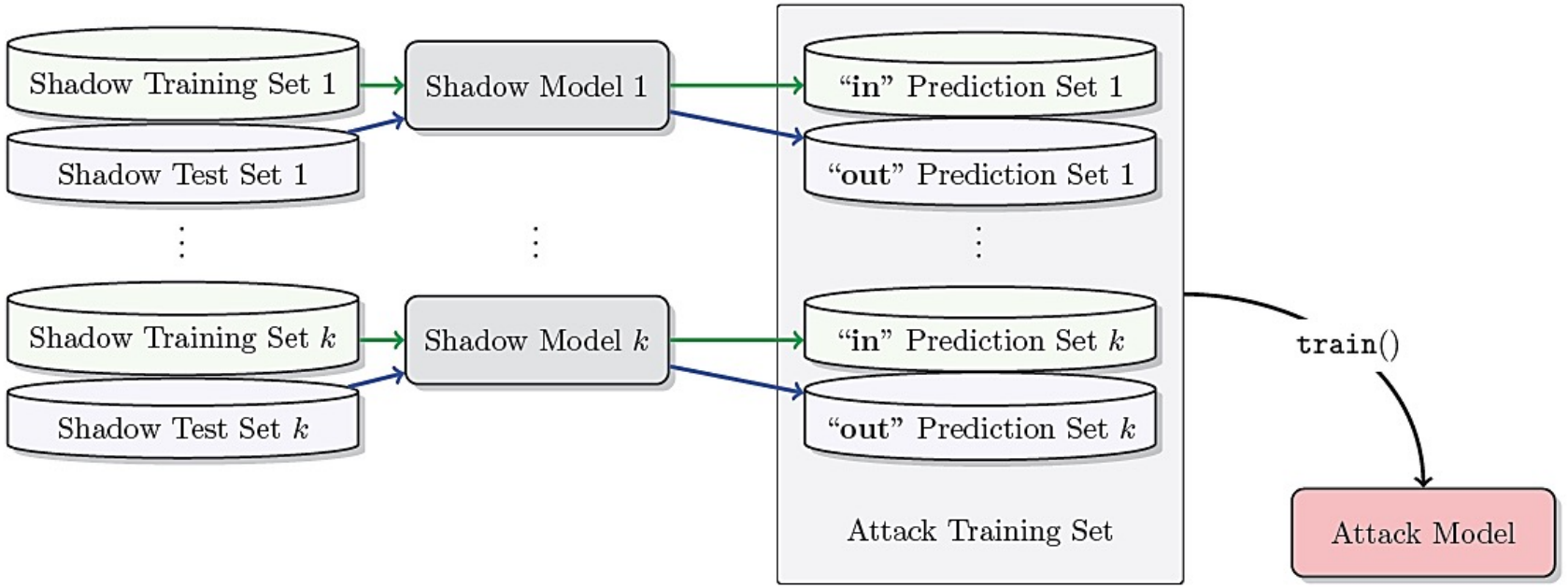
Shadow Training Attack

The output probability vectors from the shadow models are next used as inputs for training attack models (as binary classifiers) for each class

E.g., the probability vectors for all input images of Brad Pitt from all shadow training sets are labeled with 1 (meaning 'in' the training set)

The probability vectors for all input images of Brad Pitt from all shadow test sets are labeled with 0 (meaning 'out' or not in the training set)

An attack model is trained on these inputs to perform binary classification (in or out)



Shokri R, Stronati M, Song C, et al. Membership inference attacks against machine learning models[C]//2017 IEEE symposium on security and privacy (SP). IEEE, 2017: 3-18.

Shadow Training Attack

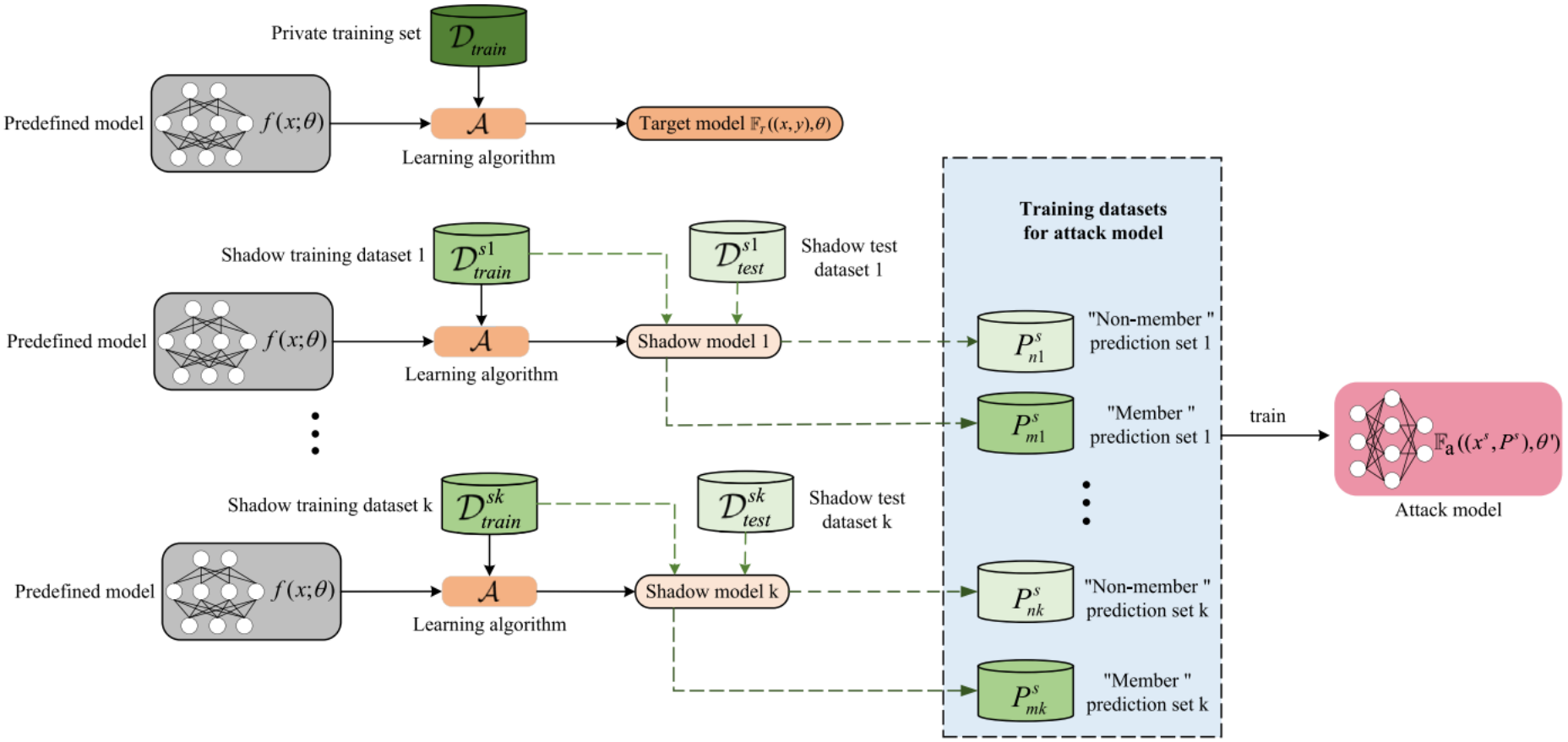


Fig. 7. Overview of the shadow training technique.

Niu J, Liu P, Zhu X, et al. A survey on membership inference attacks and defenses in Machine Learning[J]. Journal of Information and Intelligence, 2024.

Shadow Training Attack

The table shows the accuracy of a target model on training and testing sets, and the success of the attack for several models

One can note that the larger the overfitting (difference between the training and testing accuracy), the more successful the membership inference attack is

Conclusively, overfitting not only reduces the generalization of a model, but also makes the model more likely to leak sensitive information about the training data

In addition, the attack was more successful for training datasets that are more diverse and have larger number of classes (e.g., compare Purchase model with 100 classes to Purchase with 2 classes)

Dataset	*Training Accuracy*	*Testing Accuracy*	*Attack Precision*
Adult	0.848	0.842	0.503
MNIST	0.984	0.928	0.517
Location	1.000	0.673	0.678
Purchase (2)	0.999	0.984	0.505
Purchase (10)	0.999	0.866	0.550
Purchase (20)	1.000	0.781	0.590
Purchase (50)	1.000	0.693	0.860
Purchase (100)	0.999	0.659	0.935
TX hospital stays	0.668	0.517	0.657

TABLE II: Accuracy of the Google-trained models and the corresponding attack precision.

Shokri R, Stronati M, Song C, et al. Membership inference attacks against machine learning models[C]//2017 IEEE symposium on security and privacy (SP). IEEE, 2017: 3-18.

Threshold-based MIAs

Although NN-based shadow training attacks are a classic form of MIAs, they are less efficient, especially, we need more shadow models to get good attack performance. Threshold-based MIAs have been shown to achieve performance close to shadow training attacks and are much simpler.

Given a training set Strain, a test set Stest, and a trained model hθ(·). Suppose a data point (x, y) comes from Strain or Stest with equal probabilities. Then, the membership inference attack accuracy with a threshold ζ is calculated as follows

$$Acc(\zeta) = \frac{1}{2} \times \left(\frac{\sum_{(x,y)\in S_{\text{train}}} \mathbf{1}[h_\theta(x)_y \geq \zeta]}{|S_{\text{train}}|} + \frac{\sum_{(x,y)\in S_{\text{test}}} \mathbf{1}[h_\theta(x)_y < \zeta]}{|S_{\text{test}}|} \right),$$

where hθ(x)y is the output confidence for label y and 1[·] is the indicator function. Therefore, the goal of the threshold-based attack model is to find an optimal threshold ζoptim that maximizes the attack accuracy,

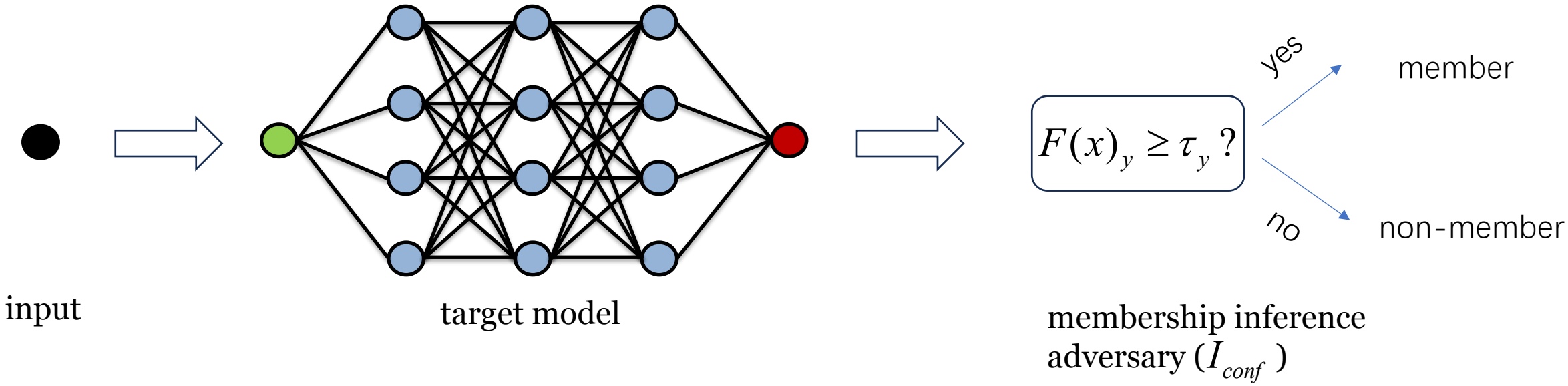
$$\zeta_{\text{optim}} = \arg\max_{\zeta} Acc(\zeta),$$

and this can be done by enumerating all possible threshold values ζ.

Class-Dependent Thresholds

We infers a sample as a member if the prediction confidence >= threshold, otherwise its a non-member

Class-dependent thresholds: setting different values of τ_y for different labels y

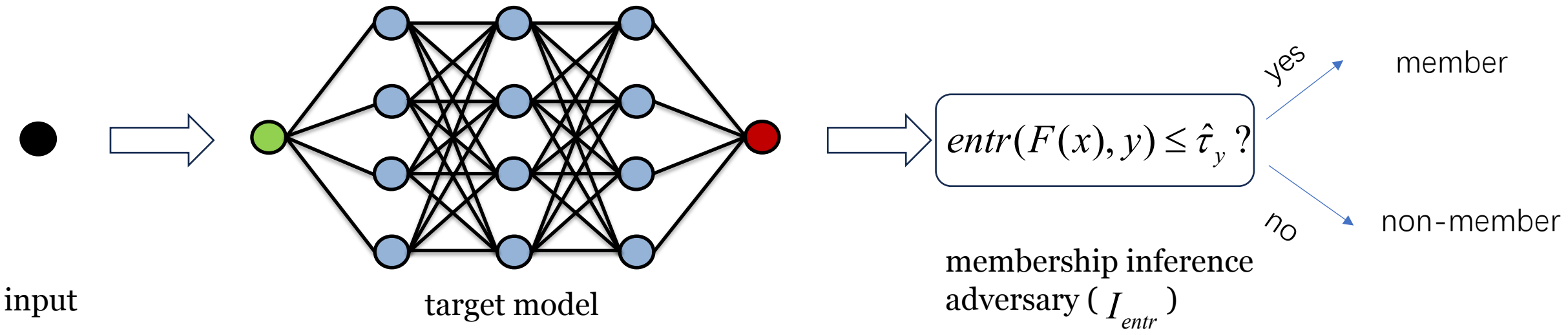


Song L, Shokri R, Mittal P. Privacy risks of securing machine learning models against adversarial examples[C]//Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. 2019: 241-257.

Prediction Entropy Thresholds

We infers a sample as a member if the prediction entropy <= threshold, otherwise its a non-member

$$I_{\text{entr}}(F,(\mathbf{x},y)) = \mathbb{1}\{-\sum_i F(\mathbf{x})_i \log(F(\mathbf{x})_i) \leq \hat{\tau}_y\}.$$



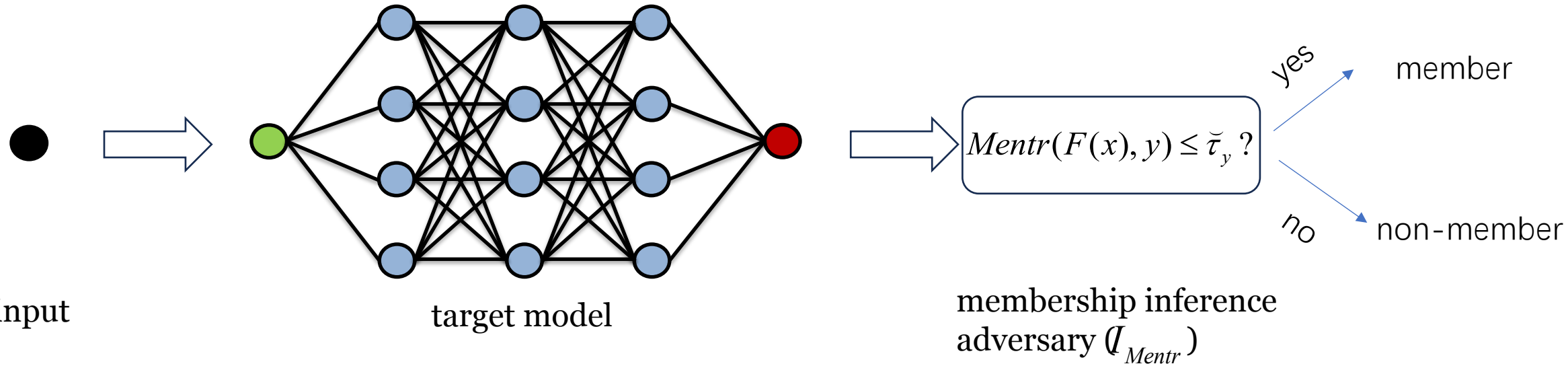
Song L, Shokri R, Mittal P. Privacy risks of securing machine learning models against adversarial examples[C]//Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. 2019: 241-257.

Modified Prediction Entropy

A new metric with following two properties given the ground truth label y:
(1) monotonically decreasing with the prediction probability of the correct label $F(x)_y$;
(2) monotonically increasing with the prediction probability of any incorrect label $F(x)_i, i \neq y$

$$\mathrm{Mentr}(F(\mathbf{x}), y) = -(1 - F(\mathbf{x})_y)\log(F(\mathbf{x})_y) - \sum_{i \neq y} F(\mathbf{x})_i \log(1 - F(\mathbf{x})_i).$$

We infers a sample as a member if the modified prediction entropy <= threshold, otherwise its a non-member



Song L, Shokri R, Mittal P. Privacy risks of securing machine learning models against adversarial examples[C]//Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. 2019: 241-257.

Class-dependent threshold

Use the shadow-training technique to learn the threshold value τ_y , $\hat{\tau}_y$, or $\breve{\tau}_y$:

- (1) first trains a shadow model to simulate the behavior of the target model;
- (2) then obtains the shadow model's prediction confidence/(modified) entropy 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 which achieves the highest accuracy in distinguishing between shadow training data and shadow test data with the class label y based on the following equation for different thresholds

$$I_{\text{conf}}(F,(\mathbf{x},y)) = \mathbb{1}\{F(\mathbf{x})_y \geq \tau_y\}.$$

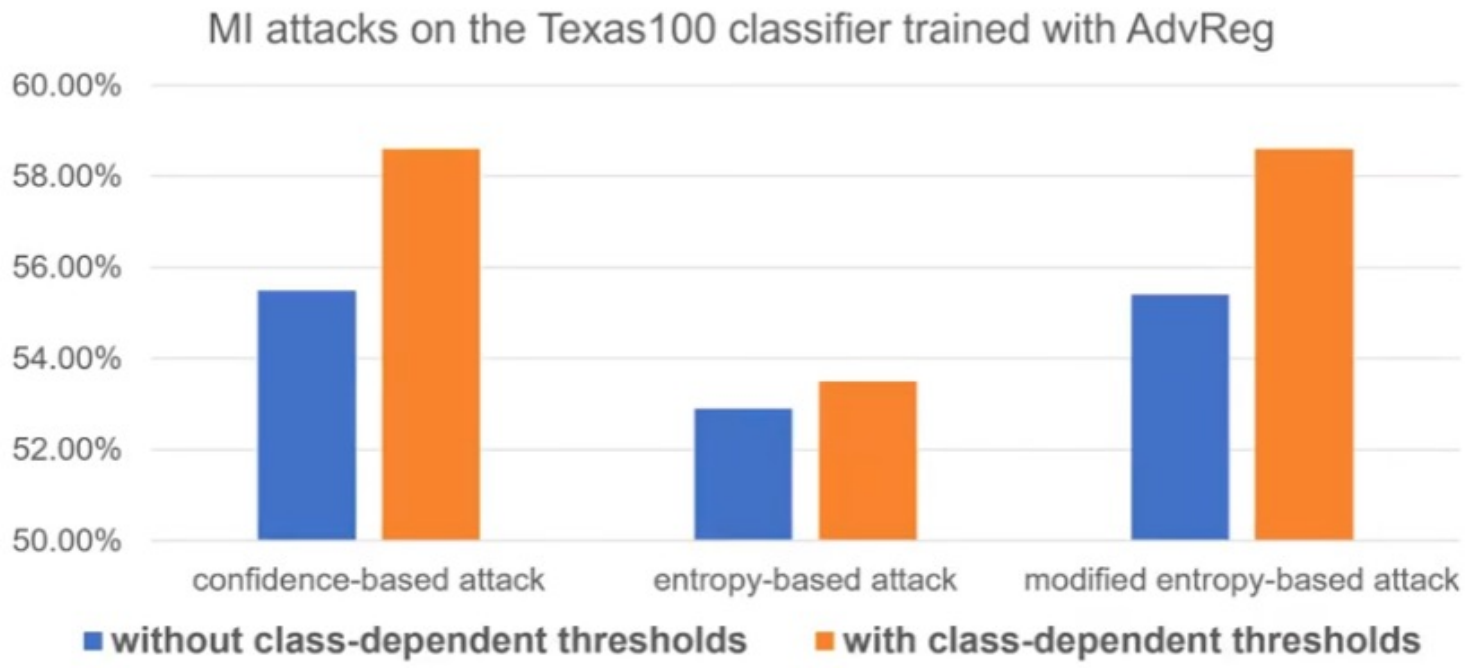
$$I_{\text{entr}}(F,(\mathbf{x},y)) = \mathbb{1}\{-\sum_i F(\mathbf{x})_i \log(F(\mathbf{x})_i) \leq \hat{\tau}_y\}.$$

$$I_{\text{Mentr}}(F,(\mathbf{x},y)) = \mathbb{1}\{\text{Mentr}(F(\mathbf{x}),y) \leq \breve{\tau}_y\}.$$

Comparison

Using the class-dependent thresholds, we can increase the MIA success by 1%-4%;

The attack based on modified entropy always outperforms the conventional entropy-based attack, results in highest attack success



Song L, Shokri R, Mittal P. Privacy risks of securing machine learning models against adversarial examples[C]//Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. 2019: 241-257.

Target Model's Sensitivity

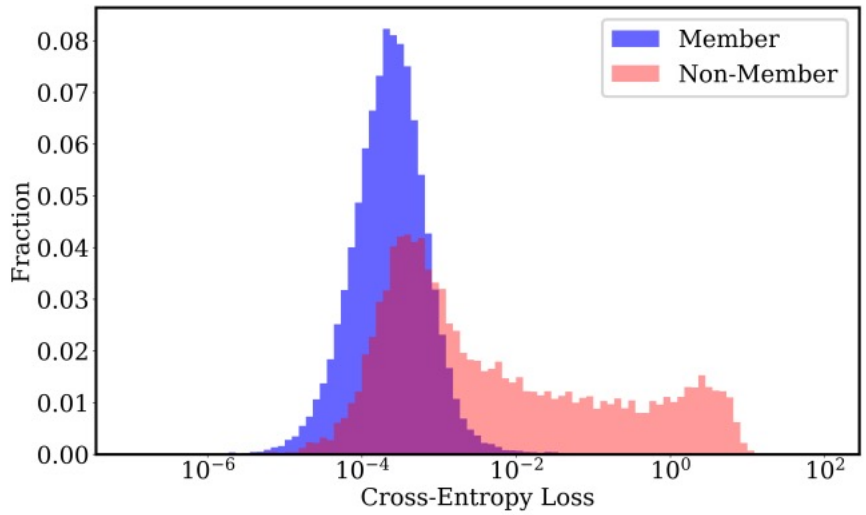
The performance of membership inference attack is related to the target model's sensitivity with regard to training data.

Definition of the sensitivity:
The sensitivity measure is the influence of one data point on the target model's performance by computing its prediction difference, when trained with and without this data point.

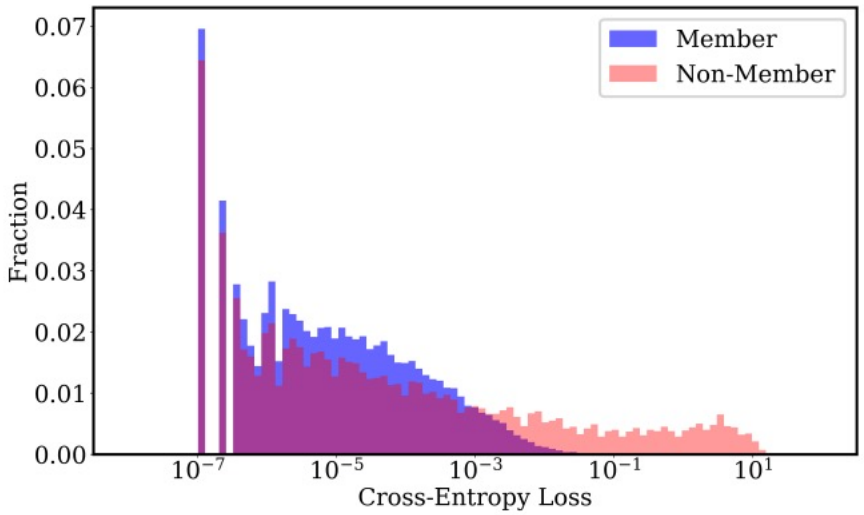
Relation to MIAs:
Intuitively, when a training point has a large influence on the target model (high sensitivity), its model prediction is likely to be different from the model prediction on a test point, and thus the adversary can distinguish its membership more easily

MIAs and Robustness

Conclusion: 1. the robust models might leak more membership information, due to exhibiting a larger generalization error, in both the benign or adversarial settings



(a) Adversarially robust model from Madry et al. [33], with 99% train accuracy and 87% test accuracy.



(b) Naturally undefended model, with 100% train accuracy and 95% test accuracy. Around 23% training and test examples have zero loss.

Figure 1: Histogram of CIFAR10 classifiers' loss values of training data (members) and test data (non-members). We can see the larger divergence between the loss distribution over members and non-members on the robust model as compared to the natural model. This shows the privacy risk of securing deep learning models against adversarial examples.

Song L, Shokri R, Mittal P. Privacy risks of securing machine learning models against adversarial examples[C]//Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. 2019: 241-257.

MIAs and Robustness

2. the robust training algorithms might make the model more susceptible to membership inference attacks, by increasing its sensitivity to its training data

We excluded 10 training points (one for each class label) and retrained the model;

We computed the sensitivity of each excluded point as the difference between its prediction confidence in the retrained model and the original model

We obtained the sensitivity metric for 60 training points by retraining the classifier 6 times

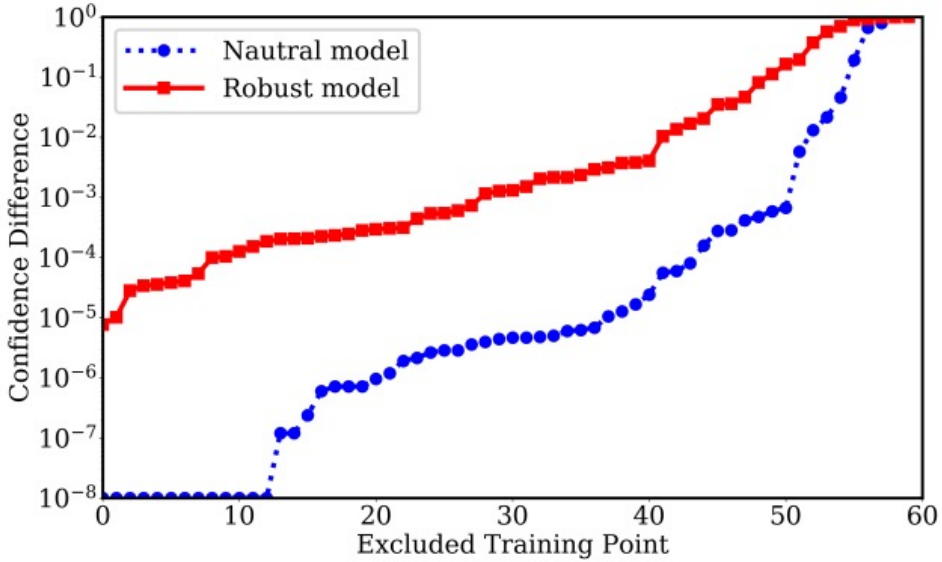
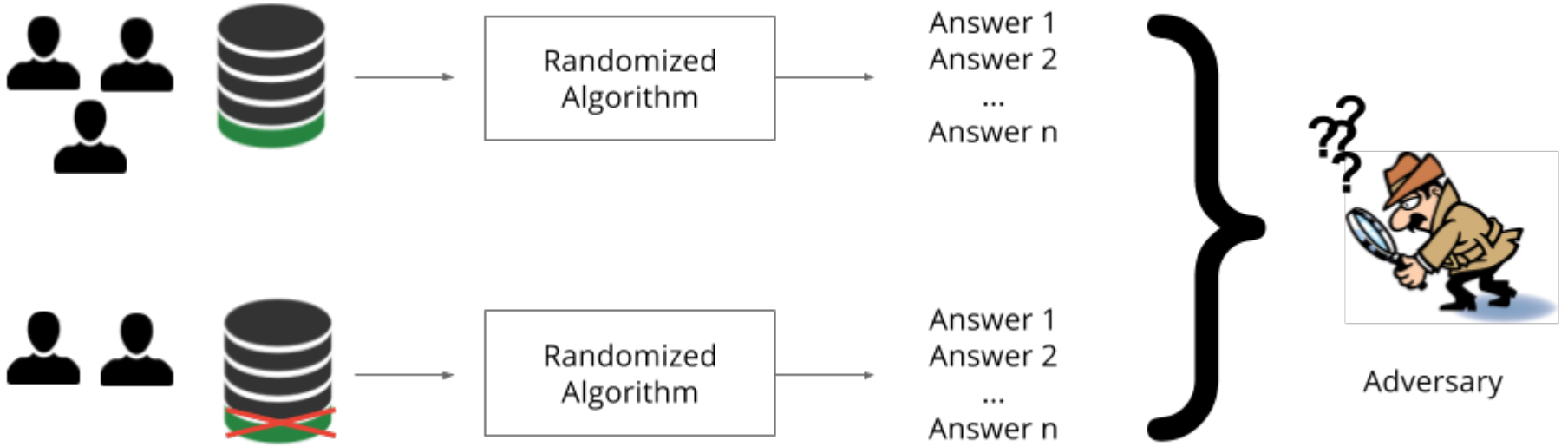


Figure 2: Sensitivity analysis of both robust [33] and natural CIFAR10 classifiers. x-axis denotes the excluded training point id number (sorted by sensitivity) during the retraining process, and y-axis denotes the difference in prediction confidence between the original model and the retrained model (measuring model sensitivity). The robust model is more sensitive to the training data compared to the natural model.

Song L, Shokri R, Mittal P. Privacy risks of securing machine learning models against adversarial examples[C]//Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. 2019: 241-257.

Differential Privacy

A basic introduction: Changing individual training sample in the training set, if the probability of learning any specific parameter remains roughly the same, this probability is referred to as privacy budget. A smaller privacy budget corresponds to stronger privacy protection. The intuition is the record of that individual sample will not be memorized, and its privacy will be respected.



http://www.cleverhans.io/privacy/2018/04/29/privacy-and-machine-learning.html

The attacker cannot distinguish the answer generated by a random algorithm of all three users and of two users, we have achieved differential privacy.

Formal Definition of Exact DP

For two datasets differ in exactly one record, A randomized mechanism M is $\mathcal{E}$ -DP, if it satisfies:

$$\Pr[\mathcal{M}(x) \in S] \leq e^{\varepsilon} Pr[\mathcal{M}(x') \in S]$$

- $\mathcal{E}$ **Privacy parameter/budget**. Controls the protection level

  Privacy-utility tradeoff: small $\mathcal{E}$ typically leads to lower utility

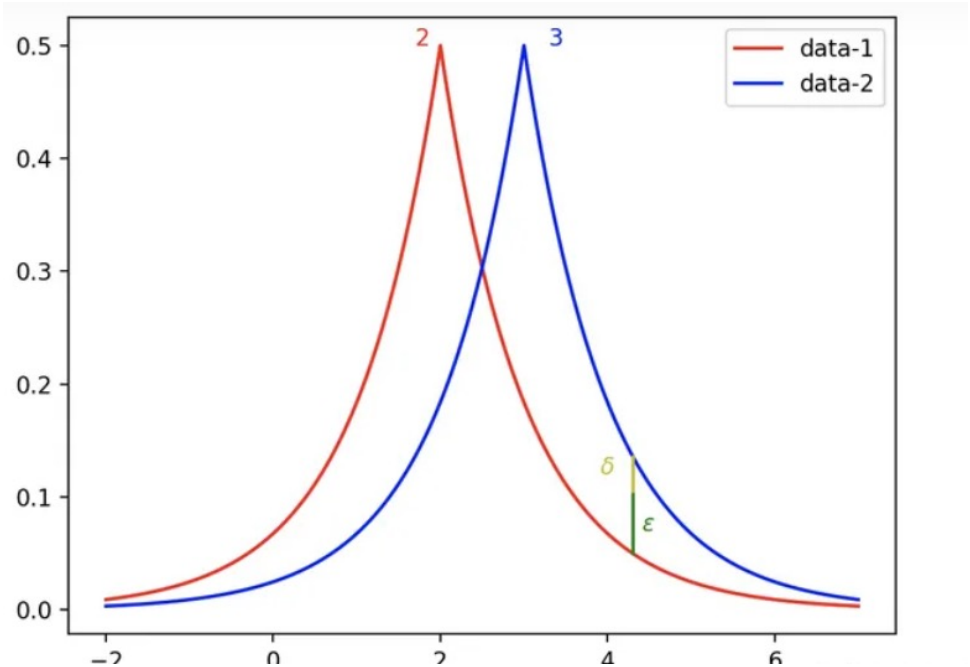
Approximate DP

A randomized mechanism $\mathcal{M}$ is (ε, δ)-DP , if it satisfies:

$$\Pr[\mathcal{M}(x) \in S] \leq e^{\varepsilon} Pr[\mathcal{M}(x') \in S] + \delta$$

δ : the probability of potential deviation from this privacy guarantee (relaxation)

$$D_{\infty}^{\delta}(Y \| Z) = \max_{S \subset Supp(Y); \Pr[Y \in S] \geq \delta} [\ln \frac{\Pr[Y \in S] - \delta}{\Pr[Z \in S]}] \leq \varepsilon$$



DP Properties: Sequential Composition

Sequential composition: Applying multiple DP mechanisms to the same dataset remains DP, while there are some degradation in the guarantees

$\mathcal{A}_1, \ldots, \mathcal{A}_t$ be a set of t mechanisms where the i-th mechanism satisfies $(\varepsilon_i, \delta_i)$-DP. Sequential composition states the joint output of the mechanisms, i.e., $(\mathcal{A}_1, \ldots, \mathcal{A}_t)$, is (ε', δ')-DP where $\varepsilon' := \sum_i \varepsilon_i$ and $\delta' := \sum_i \delta_i$

Ponomareva N, Hazimeh H, Kurakin A, et al. How to dp-fy ml: A practical guide to machine learning with differential privacy[J]. Journal of Artificial Intelligence Research, 2023, 77: 1113-1201.

DP Properties: Parallel Composition

Parallel composition. Recall that in sequential composition all mechanisms were applied to the same dataset. In contrast, parallel composition assumes that the dataset is partitioned into mutually disjoint subsets, and each mechanism is applied to one unique subset. As before, we denote the set of mechanisms by $\mathcal{A}_1, \ldots, \mathcal{A}_t$, where the i-th mechanism satisfies $(\varepsilon_i, \delta_i)$-DP. Parallel composition guarantees that the combined mechanism, i.e., $(\mathcal{A}_1, \ldots, \mathcal{A}_t)$, is $(\max_i \varepsilon_i, \max_i \delta_i)$-DP. The guarantee here is stronger than that of sequential composition. Intuitively, this statement holds because in parallel composition the combined mechanism uses each record once, whereas in sequential composition each record is used multiple times.

Ponomareva N, Hazimeh H, Kurakin A, et al. How to dp-fy ml: A practical guide to machine learning with differential privacy[J]. Journal of Artificial Intelligence Research, 2023, 77: 1113-1201.

DP Properties: Post-processing

Invariance to post-processing. Applying any data-independent transformation to a DP mechanism is guaranteed to remain differentially private (with the same privacy parameters) Dwork & Roth (2014). This property has two important implications. First, it is impossible for an attacker to weaken the DP guarantee by post-processing the mechanism's output. Second, this property can be used to simplify the design and analysis of complex DP systems. For example, training a neural network with SGD is essentially a post-processing of gradients computed at successive iterations. Thus, based on the post-processing property, differentially private training of a neural network can be achieved by using differentially private gradients in each iteration; this method will be discussed in more detail in Section 4.2.

Ponomareva N, Hazimeh H, Kurakin A, et al. How to dp-fy ml: A practical guide to machine learning with differential privacy[J]. Journal of Artificial Intelligence Research, 2023, 77: 1113-1201.

Where to Introduce DP

Where to introduce DP

DP at input

DP synthetic data

if input is DP, ANY model trained on the data is DP

DP during training

Modify training process:
Gradient noise injection

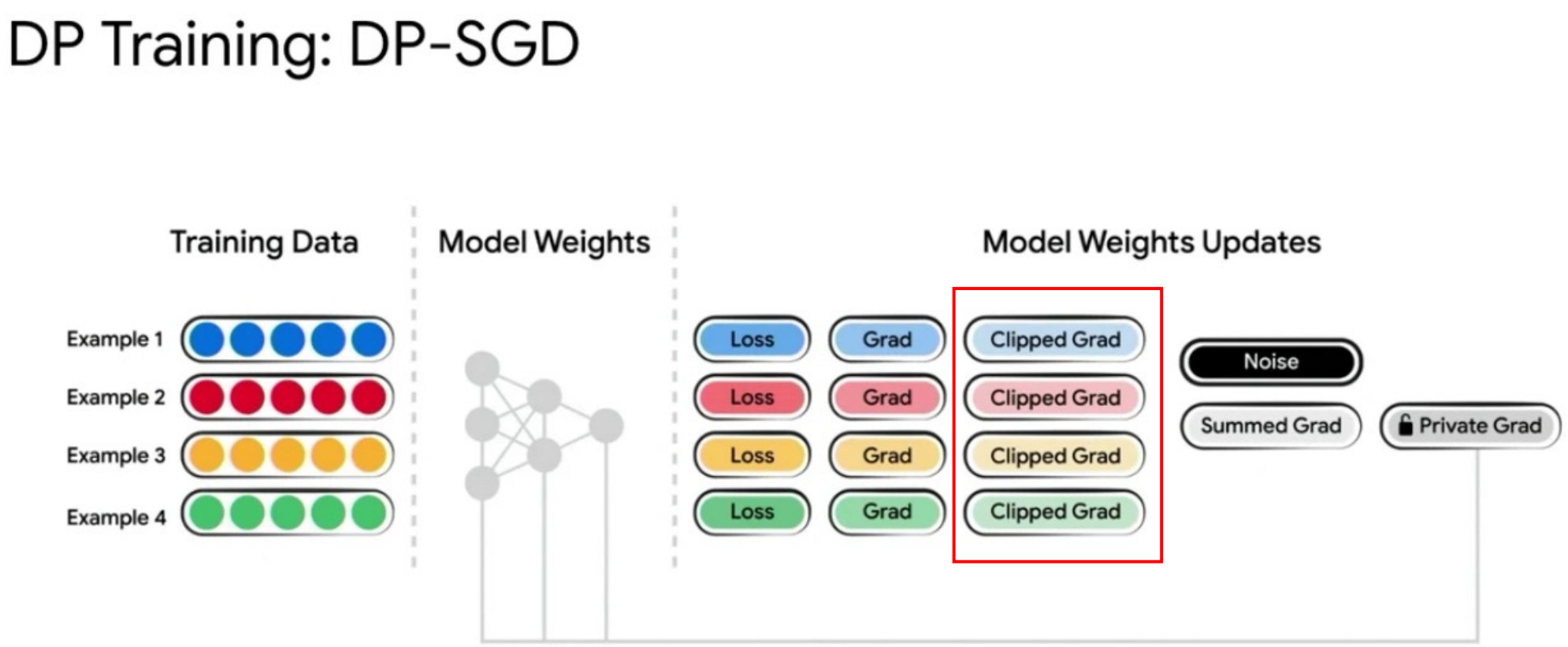
Only THIS model is DP

DP at prediction level

inject noise during inference

Only the predictions are DP

DP-SGD



Ponomareva N, Hazimeh H, Kurakin A, et al. How to dp-fy ml: A practical guide to machine learning with differential privacy[J]. Journal of Artificial Intelligence Research, 2023, 77: 1113-1201.

DP-SGD Algorithm

DP-SGD Algorithm

Input: Training data, consisting of features $X := \{x_1, x_2, ..., x_N\}$ and labels $Y := \{y_1, y_2, ..., y_N\}$.
$f(x; \theta)$ is the model applied to an input x and parameterized by θ.
$L(y, y')$ is the loss function for label y and prediction y'.
SGD hyperparameters: η learning rate, T number of iterations, B batch size.
DP hyperparameters: C clipping norm, σ noise level, δ (used only for privacy accounting).
Output: θ_T final model parameters
$\theta_0 \leftarrow$ randomly initialized values
for $t \leftarrow 1$ to T **do**
Randomly sample a batch B_t with sampling probability B/N for each data point.
Data are sampled with replacement for each batch.
for $i \in B_t$ **do**
$g_t(x_i) \leftarrow \nabla_{\theta_t} L(y_i, f(x_i; \theta_t))$ ▷ Compute per-example gradient wrt the weights
$g_t(x_i) \leftarrow g_t(x_i) / \max(1, \frac{||g_t(x_i)||_2}{C})$ ▷ Clip the per-example gradient
$\bar{g}_t \leftarrow \frac{1}{B}(\sum_i g_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 \mathbb{1}))$ ▷ Add noise
$\theta_{t+1} \leftarrow \theta_t - \eta \bar{g}_t$ ▷ Gradient descent step

Ponomareva N, Hazimeh H, Kurakin A, et al. How to dp-fy ml: A practical guide to machine learning with differential privacy[J]. Journal of Artificial Intelligence Research, 2023, 77: 1113-1201.

Renyi DP and its Properties

Definition 2 (Rényi Differential Privacy). *A randomized mechanism $\mathcal{R}$ is (λ, ε)-RDP if it satisfies:*

$$\frac{1}{\lambda - 1} \log \mathbb{E}_{x \sim \mathcal{R}(D)} \left[\left(\frac{\Pr[\mathcal{R}(D) = x]}{\Pr\left[\mathcal{R}\left(D'\right) = x\right]} \right)^{\lambda - 1} \right] \leq \varepsilon, \tag{2}$$

Theorem 1. *[Composition] A randomized mechanism $\mathcal{R}$ consists of k sub mechanisms $\mathcal{R}_1, \mathcal{R}_2, ..., \mathcal{R}_k$, for $j \in [k]$, $\mathcal{R}_j$ satisfies (λ, ε_i)-RDP, then $\mathcal{R}$ satisfies $(\lambda, \sum_{i=1}^{k} \varepsilon_i)$-RDP with the same order λ. [27]*

Theorem 2. *[From RDP to DP] A randomized mechanism $\mathcal{R}$ satisfies (λ, ε)-RDP, equal to $\mathcal{R}$ satisfies $(\varepsilon + \frac{\log 1/\delta}{\lambda - 1}, \delta)$-DP for any $\delta \in (0, 1)$. [27]*

PATE

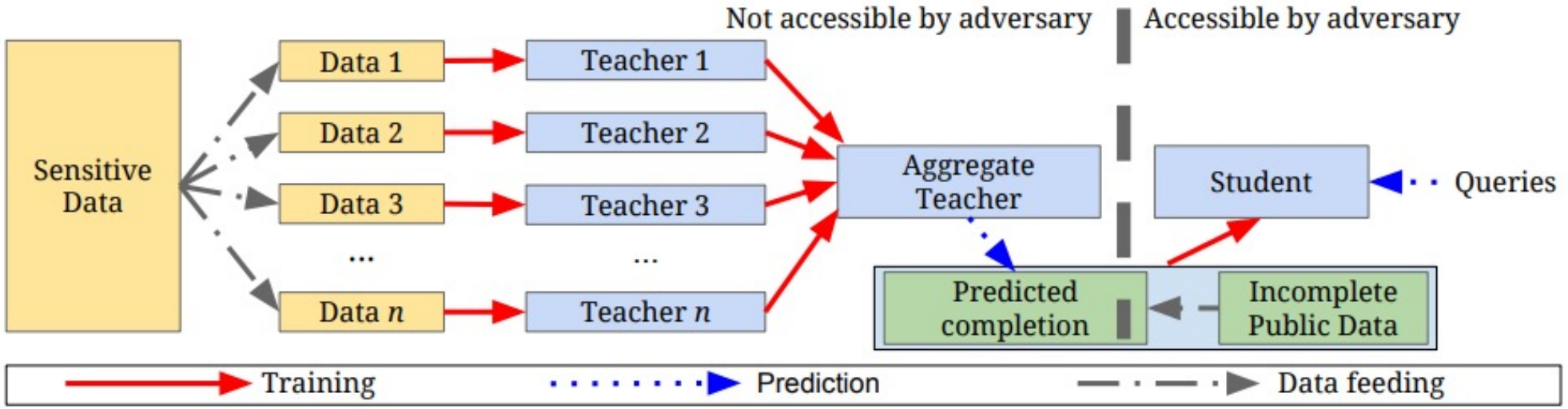


Figure 2: Overview of the approach: (1) an ensemble of teachers is trained on disjoint subsets of the sensitive data, (2) a student model is trained on public data labeled using the ensemble.

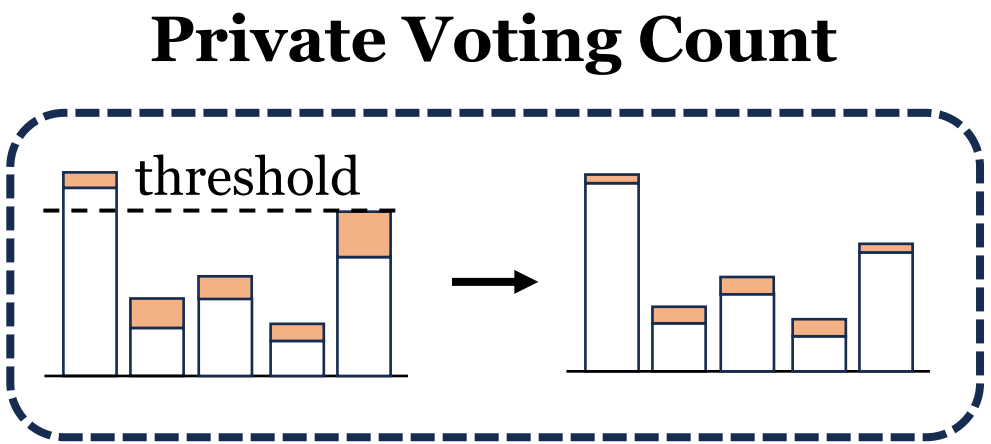
Papernot N, Song S, Mironov I, et al. Scalable private learning with pate[J]. arXiv preprint arXiv:1802.08908, 2018.

Private Voting Count (Teacher Aggregation)in PATE

Confident-GNMax mechanism: $\mathcal{M}_\sigma(\boldsymbol{x}') \triangleq \underset{k}{\operatorname{argmax}} \left\{ n_k(\boldsymbol{x}') + \mathcal{N}\left(0, \sigma^2\right) \right\}$ satisfies $(\lambda, \lambda/\sigma^2)$- RDP

(1) Sample Filtering. We set a threshold T, if the vote count for a particular label exceeds T (normally > 50% of the teacher classifiers) we retain such samples (queries) while discarding those with a vote count below the threshold T

(2) Assign Labels. For those samples pass stage one, we reapply GNMax with a smaller σ_2 to ensure the results from the majority of teacher classifier ensembles reflect the true labels to mitigate more potentially noisy labels



PATE Results

Dataset	Aggregator	Queries answered	Privacy bound ε	Accuracy	
				Student	Baseline
MNIST	LNMax (Papernot et al., 2017)	100	2.04	98.0%	99.2%
	LNMax (Papernot et al., 2017)	1,000	8.03	98.1%	
	Confident-GNMax (T=200, σ_1=150, σ_2=40)	286	**1.97**	**98.5%**	
SVHN	LNMax (Papernot et al., 2017)	500	5.04	82.7%	92.8%
	LNMax (Papernot et al., 2017)	1,000	8.19	90.7%	
	Confident-GNMax (T=300, σ_1=200, σ_2=40)	3,098	**4.96**	**91.6%**	
Adult	LNMax (Papernot et al., 2017)	500	2.66	83.0%	85.0%
	Confident-GNMax (T=300, σ_1=200, σ_2=40)	524	**1.90**	**83.7%**	
Glyph	LNMax	4,000	4.3	72.4%	82.2%
	Confident-GNMax (T=1000, σ_1=500, σ_2=100)	10,762	2.03	**75.5%**	
	Interactive-GNMax, two rounds	4,341	**0.837**	73.2%	

Table 1: **Utility and privacy of the students.** The Confident- and Interactive-GNMax aggregators introduced in Section 4 offer better tradeoffs between privacy (characterized by the value of the bound ε) and utility (the accuracy of the student compared to a non-private baseline) than the LNMax aggregator used by the original PATE proposal on all datasets we evaluated with. For MNIST, Adult, and SVHN, we use the labels of ensembles of 250 teachers published by Papernot et al. (2017) and set $\delta = 10^{-5}$ to compute values of ε (to the exception of SVHN where $\delta = 10^{-6}$). All Glyph results use an ensemble of 5000 teachers and ε is computed for $\delta = 10^{-8}$.

Papernot N, Song S, Mironov I, et al. Scalable private learning with pate[J]. arXiv preprint arXiv:1802.08908, 2018.

Visual Prompt in PATE

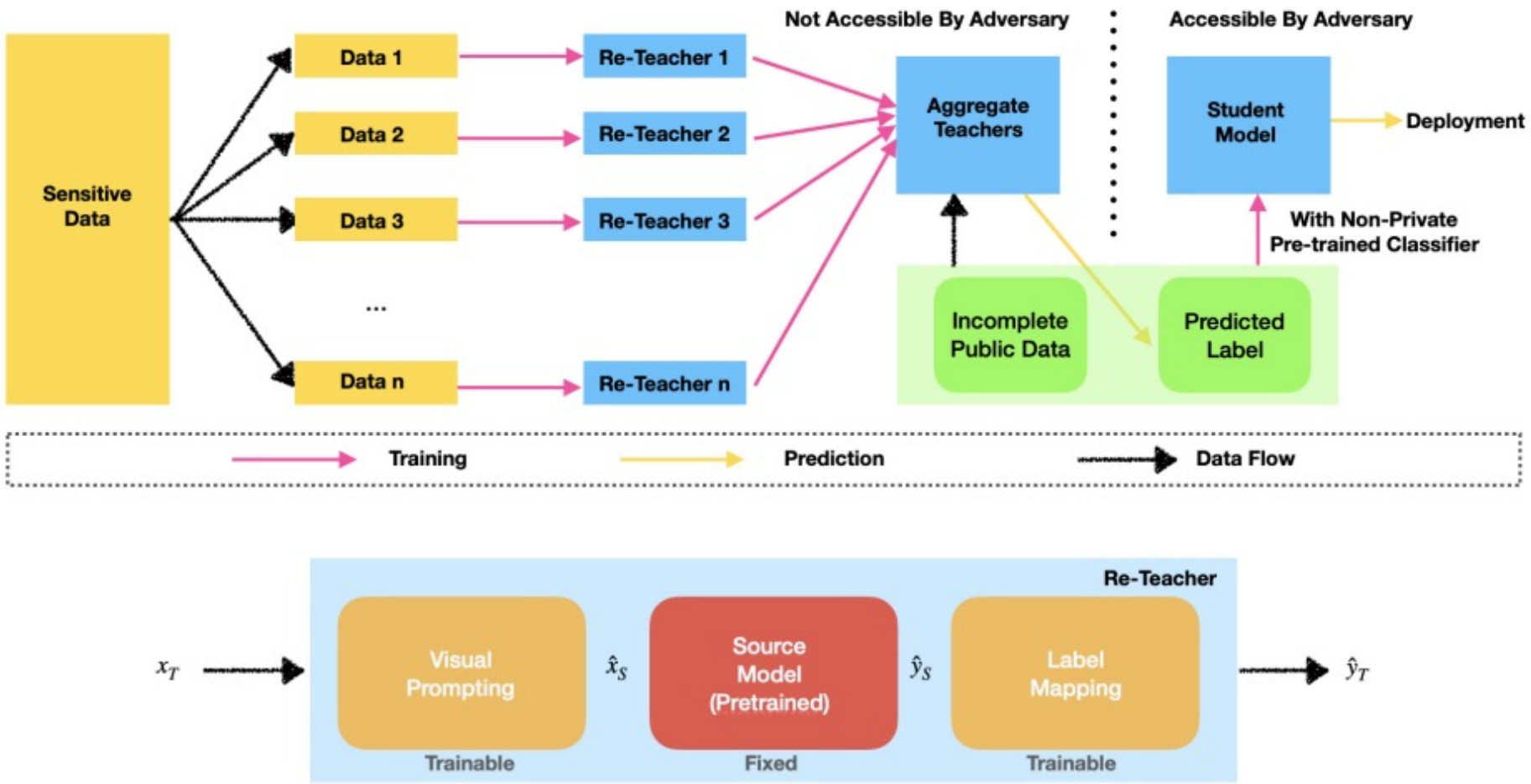


Figure 1. An overview of the proposed Prom-PATE framework.

Li Y, Tsai Y L, Yu C M, et al. Exploring the benefits of visual prompting in differential privacy[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 5158-5167.

Visual Prompt (Label Mapping) Training

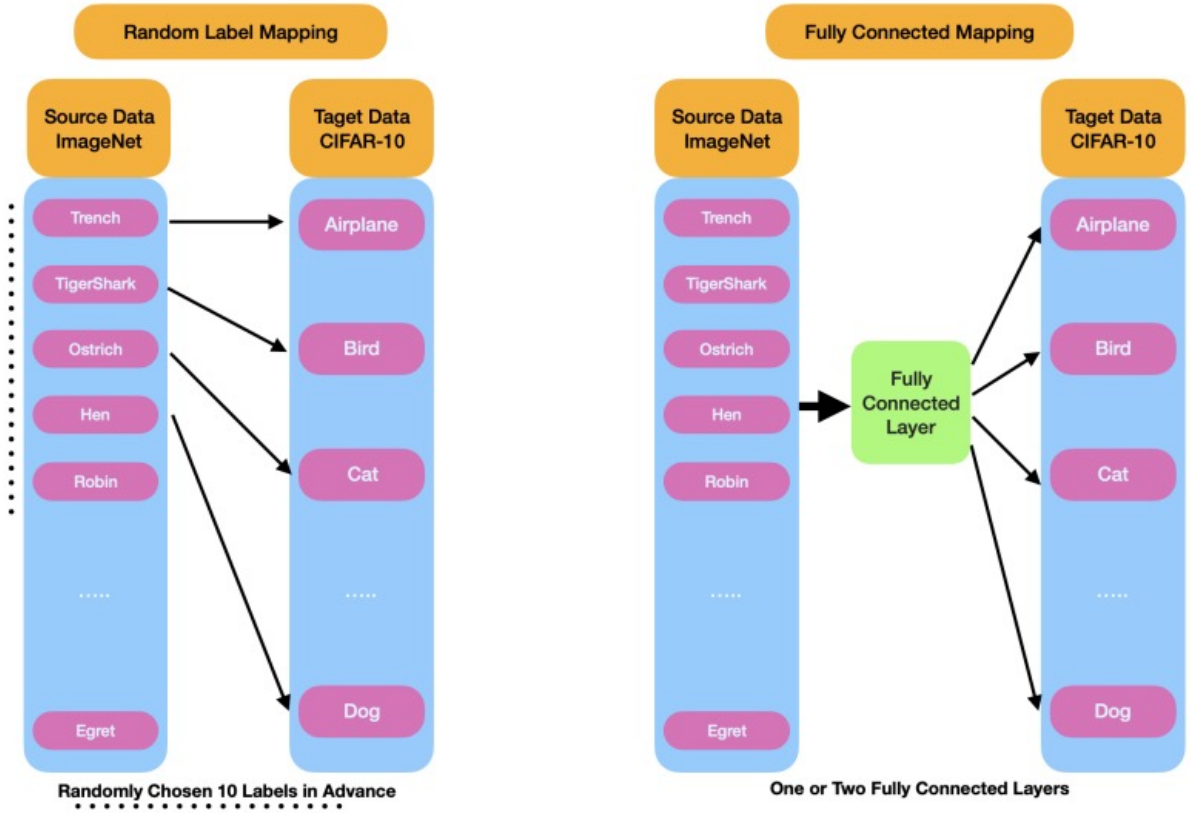


Figure 2. Illustration of different strategies for label mapping. *Left*: we follow the convention setting in VP [3] and apply randomly assigned label mapping that is pre-determined before training. *Right*: we simply apply a trainable fully-connected layer for the model to learn the appropriate mapping as proposed in [2]

Visual Prompts:

$$\hat{x}_S = M \odot \omega_1 + (I - M) \odot \text{ZeroPad}(x_T),$$

M: binary mask of the same dimension with the source data

Label Mapping:

$$\hat{y}_T = \text{softmax}(f_\ell(\omega_2; \hat{y}_S)).$$

$f_\ell(\omega_2; \cdot)$: label mapping function

Li Y, Tsai Y L, Yu C M, et al. Exploring the benefits of visual prompting in differential privacy[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 5158-5167.

Visual Prompt (Label Mapping) Training

Number of re-teachers	100	250	500	1000
ϵ	1.095	1.095	1.04	1.019
Queries	1000	1000	1000	1000
Answered Queries	18	46	90	684
Threshold T	430	500	650	500
σ_1	150	150	150	200
σ_2	50	100	100	50
Accuracy%) ± Std	59.20 ± 0	85.87 ± 0.55	96.53 ± 0.74	**97.07 ± 0.50**

Table 6. Effect on different numbers of re-teacher models.

Li Y, Tsai Y L, Yu C M, et al. Exploring the benefits of visual prompting in differential privacy[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 5158-5167.

Existing Challenges

- More Flexible Ways to Find the Optimal τ_y
- Trade-off between the Training Iteration and Privacy Budget in the Training-based DP Methods
- How to Create Dataset to Replace Public Dataset if No Public Dataset with Similar Distribution Exists
- Evaluation on Visual Prompting-Trained Models