Machine Learning with Membership Privacy via Knowledge Transfer

Virat Shejwalkar
University of Massachusetts Amherst
vshejwalkar@cs.umass.edu

Amir Houmansadr
University of Massachusetts Amherst
amir@cs.umass.edu

Abstract

Machine learning models are prone to membership inference attacks, which aim to infer whether the target sample is a member of the target model's training dataset. The serious privacy concerns due to the membership inference have motivated multiple defenses against membership inference attacks, e.g., differential privacy and adversarial regularization. Unfortunately, these defenses produce machine learning models with unacceptably low utility, e.g., classification accuracy. We propose a new defense based on knowledge distillation, called *Distillation for Membership Privacy* (DMP), against membership inference attacks that preserves the utility of the resulting models significantly better than prior defenses. We provide a novel criterion to tune the data used for knowledge transfer in DMP in order to adjust the tradeoffs between utility and privacy of resulting models. Our evaluations clearly demonstrate the state-of-the-art membership privacy-utility tradeoffs of DMP.[1]

1 Introduction

Machine learning (ML) models trained using privacy sensitive data can leak private information about their data owners. This has been demonstrated through various inference attacks [8, 10, 5], and most notably the *membership inference attack* (MIA) [24] which is the focus of our work. An MIA adversary with a blackbox or whitebox access to a target model aims to determine if a given target sample belonged to the private training data of the target model or not. MIAs are able to distinguish the members from non-members by *learning* the behavior of the target model on member versus non-member inputs.

Recent literature has investigated several defenses against membership inference attacks based on differential Privacy (DP), e.g., DP-SGD [2] and PATE [15], and based on regularization, e.g., adversarial regularization [13] and L2-regularization [24]. DP based defenses tend to add large amounts of noise during learning or inference phase and significantly reduce model accuracies. Furthermore, as we show, adversarial regularization and other state-of-the-art regularizations, e.g., label smoothing [27] and dropout [26], also fail to provide acceptable membership privacy-utility tradeoffs (simply called 'tradeoffs' here onward). In summary, existing defenses against MIAs offer poor tradeoffs between model utility and membership privacy.

Motivated by the poor tradeoffs, we propose a defense against MIAs that significantly improves the tradeoffs compared to prior defenses. That is, for a given degree of membership privacy (i.e., MIA resistance), our defense produces models with significantly higher classification accuracies compared to prior defenses. Our defense, called ***D**istillation for **M**embership **P**rivacy* (DMP), leverages *knowledge distillation* [9] (more generally called *knowledge transfer*), which transfers the knowledge of large models to smaller models, and is primarily used for model compression. Intuitively, DMP protects membership privacy by thwarting the access of the resulting models to the private

[1] The full version of this work [23] is accepted at AAAI, 2021

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

training data. Similar to adversarial regularization, DMP assumes access to a possibly sensitive and "unlabeled" *reference data* drawn from the same distribution as the "labeled" private training data, and uses such reference data to train its final models. We provide a novel criterion to select/generate reference data to improve membership privacy due to DMP. While some prior work [15] combined knowledge transfer and DP to protect data privacy, our work is *the first* to study the promise of knowledge transfer as the sole technique to train membership privacy-preserving models.

2 Preliminaries

Knowledge distillation. [4, 3] proposed knowledge distillation, which uses the outputs of a large teacher model to train a smaller student model, in order to *compress* large models to smaller models. The outputs used for distillation can vary, e.g., Hinton et al. [9] use class probabilities generated by the teacher as the outputs, while Romero et al. [19] use the intermediate activations along with class probabilities of the teacher. It is well established that *knowledge distillation produces students with accuracies similar to their teachers* [6, 29]. This also allows DMP to produce highly accurate target models. Note that, although we use term "distillation", DMP uses teacher and student models of the same sizes, because DMP is not concerned with the size of the resulting model.

Membership inference attacks. Below we give a general methodology of MIAs. Consider a target model θ and a sample $(\mathbf{x}, y)$. MIAs exploit the differences in the behavior of θ on members and non-members of the private D_{tr}. Therefore, MIAs train a binary attack model to classify target samples into members and non-members. Such attack models can be neural networks [24, 21] or simple thresholding functions where threshold is tuned for maximum attack performance [28, 25]. The adversary computes various features of θ for given $(\mathbf{x}, y)$, e.g., prediction $\theta(\mathbf{x}, y)$, θ's loss on $(\mathbf{x}, y)$, and the gradients of the loss. Let $F(\mathbf{x}, y, \theta)$ denote the combination of these features. The attack model h takes $F(\mathbf{x}, y, \theta)$ as its input and outputs the probability that $(\mathbf{x}, y)$ is a member of D_{tr}. Let $\Pr_{D_{\mathsf{tr}}}$ and $\Pr_{\setminus D_{\mathsf{tr}}}$ be the conditional probabilities of the members and non-members of D_{tr}, respectively. Hence, the expected gain of the attack model for the above setting is given by:

$$G^{\theta}(h) = \mathop{\mathbb{E}}_{\substack{(\mathbf{x},y)\\ \sim \Pr_{D_{\mathsf{tr}}}}} [\log(h(F))] + \mathop{\mathbb{E}}_{\substack{(\mathbf{x},y)\\ \sim \Pr_{\setminus D_{\mathsf{tr}}}}} [\log(1 - h(F))] \tag{1}$$

In practice, the adversary knows only a finite set of the members D and non-members D'^A required to train h, hence computes the above gain empirically as in (2), and solves for h^* that maximizes (2).

$$G^{\theta}_{D^A, D'^A}(h) = \sum_{\substack{(\mathbf{x},y)\\ \in D^A}} \frac{\log(h(F))}{|D^A|} + \sum_{\substack{(\mathbf{x},y)\\ \in D'^A}} \frac{\log(1 - h(F))}{|D'^A|} \tag{2}$$

3 Distillation for Membership Privacy (DMP)

DMP is a strong meta-regularizer and the main intuition behind DMP is based on the results by Sablayrolles et al. [20]. They assume a posterior distribution, $\mathbb{P}(\theta|D)$, of the parameters trained on data $D = \{z_1, .., z_n\}$ as given in (3). Consider a neighboring dataset $D' = \{z_1, .., z'_j, .., z_n\}$ of D, which is obtained by modifying at most one sample of D [7]. Sablayrolles et al. [20] show that, to provide membership privacy to z_j, the log of the ratio of probabilities of obtaining the same θ from D and D' should be bounded, i.e., (3) should be bounded.

$$\log\Big|\frac{\mathbb{P}(\theta|D)}{\mathbb{P}(\theta|D')}\Big| = |\ell(\theta, z_j) - \ell(\theta, z'_j)| \quad \ldots \mathbb{P}(\theta|D) \propto \exp(\sum_{z_i \in D} \ell(\theta, z_i)) \tag{3}$$

(3) implies that, if θ was indeed trained on z_j, then to provide membership privacy to z_j, the loss of θ on z_j should be same as on any non-member sample z'_j. DMP is built on this intuition and aims to train a model with statistically close losses on the members and non-members. DMP achieves this via knowledge transfer and restricts the direct access of its final models to the private data and significantly reduces the membership information leakage about the private data. DMP's final models have superior utility due to the well-established efficiency of knowledge transfer in producing student models with accuracies close to teacher models.

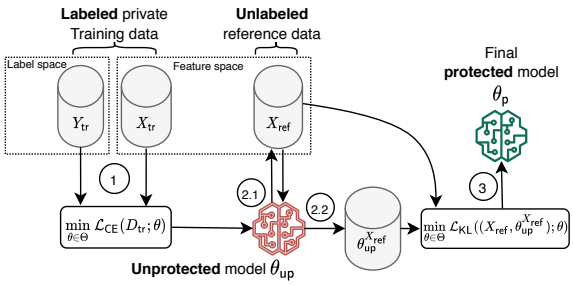

Figure 1: Three phases (described alongside) of our ***D**istillation for **M**embership **P**rivacy* (DMP) defense.

Details of the DMP technique. We now detail our DMP defense depicted in Figure 1. In *pre-distillation phase* (step-1 Figure 1), DMP trains an *unprotected model*, θ_{up}, on the private training data, D_{tr}, using standard SGD optimizer, e.g., Adam. Such unprotected θ_{up} is highly susceptible to MIA due to large generalization error, i.e., difference between train and test accuracies [24, 28].

Next, in *distillation phase* (step-2.1 Figure 1), DMP obtains the reference data, X_{ref}, required to transfer the knowledge of θ_{up} in the final *protected model*, θ_{p}. Note that, X_{ref} is *unlabeled* and cannot be used directly for any learning. Then, we compute soft labels of X_{ref}, i.e., $\theta_{\mathsf{up}}^{X_{\mathsf{ref}}} = \theta_{\mathsf{up}}(X_{\mathsf{ref}})$ (step-2.2 Figure 1).

Finally, in *Post-distillation phase* (step-3 Figure 1), DMP trains a protected model θ_{p} on $(X_{\mathsf{ref}}, \theta_{\mathsf{up}}^{X_{\mathsf{ref}}})$ using the Kullback-Leibler divergence loss based optimization in (4) In (4), $\overline{\mathbf{y}}$ is the target soft label.

$$\theta_{\mathsf{p}} = \underset{\theta}{\operatorname{argmin}} \frac{1}{|X_{\mathsf{ref}}|} \sum_{(\mathbf{x},\overline{\mathbf{y}})\in(X_{\mathsf{ref}},\theta_{\mathsf{up}}^{X_{\mathsf{ref}}})} \mathcal{L}_{\mathsf{KL}}(\mathbf{x},\overline{\mathbf{y}}) \quad \ldots \mathcal{L}_{\mathsf{KL}}(\mathbf{x},\overline{\mathbf{y}}) = \sum_{i=0}^{\mathbf{c}-1} \overline{\mathbf{y}}_i \log\Big(\frac{\overline{\mathbf{y}}_i}{\theta_{\mathsf{p}}(\mathbf{x})_i}\Big) \tag{4}$$

Due to KL-divergence loss in (4), the resulting model, θ_{p}, perfectly learns the behavior of θ_{up} on the X_{ref}. Furthermore, X_{ref} being a representative non-member data, i.e., test data, we expect that the test accuracies of the final DMP model, θ_{p}, and the unprotected model, θ_{up}, are close [3, 19].

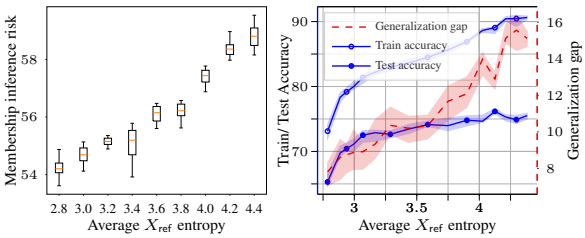

Figure 2: The lower the entropy of predictions of θ_{up} on X_{ref}, the higher the membership privacy.

Fine-tuning the DMP defense. Selecting the appropriate reference data, X_{ref}, is important to achieve the desired privacy-utility tradeoffs in DMP. To this end, we give an interesting result in Proposition 1.

Proposition 1. *Consider θ_{up} trained on a private D_{tr}. Then, the membership leakage about D_{tr} through predictions $\theta_{\mathsf{up}}(X_{\mathsf{ref}})$ can be reduced by selecting/generating X_{ref} that are far from D_{tr} in the input feature space and whose predictions, $\theta_{\mathsf{up}}(X_{\mathsf{ref}})$, have low entropies.*

Intuitively, such reference data are easy to classify and none of the members of the private D_{tr} significantly affects their predictions, and therefore, these predictions do not leak membership information of any particular member. Due to space limitations, we defer the proof to the full version of this work. In Figure 2, we empirically verify Proposition 1 using Purchase dataset [18]: Figure 2 (left) shows the increase in the MIA risk and Figure 2 (right) shows the increase in the classification performance of θ_{p} with the increase in average entropy of the predictions of unprotected model, θ_{up}, on the reference data, X_{ref}, used.

4 Evaluations

Experimental setup. We evaluate DMP on four datasets and corresponding model architectures that are consistent with the previous works [24, 14, 13]: Purchase dataset [18], Texas hospital dataset [1], CIFAR10 and CIFAR100 [12]. We measure the membership privacy due to various defenses as the accuracy of three state-of-the-art whitebox and blackbox attacks proposed in [14], and entropy-based blackbox attack proposed in [28]; we denote the attack accuracies by A_{wb}, A_{bb}, and A_{bl}, respectively. Additional experimental details are in Appendix A.1.

Table 1: Models trained without any defenses have high test accuracies, A_{test}, but their high generalization errors, E_{gen} (i.e., $A_{\text{train}} - A_{\text{test}}$) facilitate high membership inference risks.

Dataset + model (Acronym)	No defense				
	E_{gen}	A_{test}	A_{wb}	A_{bb}	A_{bl}
Purchase + FC (P-FC)	24.0	76.0	77.1	76.8	63.1
Texas + FL (T-FC)	51.3	48.7	84.0	82.2	76.1
CIFAR100 + Alexnet (C100-A)	63.2	36.8	90.3	91.3	81.8
CIFAR100 + DenseNet-12 (C100-D12)	33.8	65.2	72.2	71.8	67.5
CIFAR100 + DenseNet-19 (C100-D19)	34.4	65.5	82.3	81.6	68.1
CIFAR100 + Alexnet (C10-A)	32.5	67.5	77.9	77.5	66.4

Table 2: Comparing test accuracy, A_{test}, and generalization error, E_{gen}, of DMP and Adversarial Regularization, for near-equal, low MIA risks (high membership privacy). A^{+}_{test} shows *the % increase* in A_{test} of DMP over Adversarial Regularization.

Dataset and model	Adversarial regularization (AdvReg)					DMP					
	E_{gen}	A_{test}	Attack accuracy			E_{gen}	A_{test}	A^{+}_{test}	Attack accuracy		
			A_{wb}	A_{bb}	A_{bl}				A_{wb}	A_{bb}	A_{bl}
P-FC	9.7	56.5	55.8	55.4	54.9	10.1	74.1	**+31.2%**	55.3	55.1	55.2
T-FC	6.1	33.5	58.2	57.9	54.1	7.1	48.6	**+45.1%**	55.3	55.4	53.6
C100-A	6.9	19.7	54.3	54.0	53.5	6.5	35.7	**+81.2%**	55.7	55.6	53.3
C100-D12	5.5	26.5	51.4	51.3	52.8	3.6	63.1	**+138.1%**	53.7	53.0	51.8
C100-D19	7.2	33.9	54.2	53.4	53.6	7.3	65.3	**+92.6%**	54.7	54.4	53.7
C10-A	4.2	53.4	51.9	51.2	52.1	3.1	65.0	**+21.7%**	51.3	50.6	51.6

4.1 Experimental results

Comparison with regularization techniques. Regularization improves the generalization of ML models, and hence, reduce the MIA risk [24]. Hence, we compare DMP with four regularization defenses, including the state-of-the-art MIA defense—adversarial regularization [13]. Table 1 shows accuracies of models trained without any defense. Note that, CIFAR models have lower than state-of-the-art accuracies due to smaller training datasets.

Comparisons with adversarial regularization (AdvReg). Table 2 compares A_{test} of DMP and AdvReg models, for similar MIA accuracies (i.e., membership privacy). As expected, these models also have similar E_{gen}'s. However, A_{test} of DMP models is significantly higher than AdvReg models; A^{+}_{test} column shows the % increase in A_{test} due to DMP over AdvReg: Accuracy improvements due to DMP over AdvReg are close to 100% for CIFAR-100, and 20% to 45% for other datasets. AdvReg uses accuracy of an MIA model to regularize and train its target models to fool the MIA model. However, AdvReg allows its target models to directly access D_{tr}. Hence, to effectively fool the MIA model, it puts relatively large weight on the regularization-loss term. This reduces the impact of the loss on main task and reduces the accuracy of AdvReg models. DMP uses appropriate reference data to transfer the knowledge of D_{tr} to its target models without allowing them direct access. Hence, DMP significantly outperforms AdvReg in terms of privacy-utility tradeoffs.

Comparisons with other regularizers. Next, we compare DMP with four state-of-the-art regularizers: weight decay (WD), dropout [26] (DR), label smoothing [27] (LS), and confidence penalty [17] (CP). Table 4 (Appendix A) shows the results, when MIA risks of regularized models is close that of DMP models (Table 2). We note that, in all the cases, A_{test} of DMP are significantly higher (up to 385% increase as A^{+}_{test} column specifies) than A_{test} of other regularizers. This is because, these regularizers aim to improve the test accuracies of target models, but are not designed to reduce MIA risk. Thus, to reduce MIA risk, these regularization techniques add large, suboptimal noises during training, and hurt the utility of resulting models.

Comparisons with differentially private defenses. In Appendix A.2, we compare DMP with two state-of-the-art differentially private defenses, DP-SGD [2] and PATE [15], and demonstrate the superior membership privacy-utility tradeoffs of DMP over these defenses. Our comparisons with DP-based defenses emphasize the results of Jayaraman et al. [11], who study various DP-based defenses in depth and show that they fail to produce model with acceptable tradeoffs.

5 Conclusions

We proposed Distillation for Membership Privacy (DMP), a knowledge distillation based defense against membership inference attacks that significantly improves the membership privacy-model utility tradeoffs compared to state-of-the-art defenses. We provided a novel criterion to generate/select reference data in DMP and achieve the desired tradeoffs. Our extensive evaluation demonstrated the state-of-the-art privacy-utility tradeoffs of DMP. We believe our study highlights an important aspect of knowledge transfer: apart from its use as a sole membership inference defense, its simplicity can allow other defenses to incorporate knowledge transfer to improve their privacy-utility tradeoffs, which currently limits their use in practice.

References

[1] Texas hospital stays dataset. https://www.dshs.texas.gov/THCIC/Hospitals/Download.shtm. [Online; accessed 10-February-2020].

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

[3] Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? In *Advances in neural information processing systems*, pages 2654–2662, 2014.

[4] Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541, 2006.

[5] Nicholas Carlini, Chang Liu, Jernej Kos, Ulfar Erlingsson, and Dawn Song. The secret sharer: Measuring unintended neural network memorization and extracting secrets. *arXiv preprint arXiv:1802.08232*, 2018.

[6] Elliot J Crowley, Gavin Gray, and Amos J Storkey. Moonshine: Distilling with cheap convolutions. In *Advances in Neural Information Processing Systems*, pages 2888–2898, 2018.

[7] Zeyu Ding, Yuxin Wang, Guanhong Wang, Danfeng Zhang, and Daniel Kifer. Detecting violations of differential privacy. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pages 475–489. ACM, 2018.

[8] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2015.

[9] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *NIPS 2014 Deep Learning Workshop*, 2014.

[10] Briland Hitaj, Giuseppe Ateniese, and Fernando Pérez-Cruz. Deep models under the gan: information leakage from collaborative deep learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2017.

[11] Bargav Jayaraman and David Evans. Evaluating differentially private machine learning in practice. In *USENIX Security Symposium*, 2019.

[12] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009.

[13] Milad Nasr, Reza Shokri, and Amir Houmansadr. Machine learning with membership privacy using adversarial regularization. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pages 634–646. ACM, 2018.

[14] Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Stand-alone and federated learning under passive and active white-box inference attacks. *Security and Privacy (SP), 2019 IEEE Symposium on*, 2019.

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

[16] Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. Scalable private learning with pate. *arXiv preprint arXiv:1802.08908*, 2018.

[17] Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*, 2017.

[18] Purchase. Acquire Valued Shoppers Challenge. https://www.kaggle.com/c/acquire-valued-shoppers-challenge/data. [Online; accessed 11-September-2019].

[19] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

[20] Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, and Herve Jegou. White-box vs black-box: Bayes optimal strategies for membership inference. In *International Conference on Machine Learning*, pages 5558–5567, 2019.

[21] 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 NDSS*, 2019.

[22] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016.

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

[24] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *Security and Privacy (SP), 2017 IEEE Symposium on*, 2017.

[25] Liwei Song and Prateek Mittal. Systematic evaluation of privacy risks of machine learning models. *arXiv preprint arXiv:2003.10595*, 2020.

[26] 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.

[27] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[28] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pages 268–282. IEEE, 2018.

[29] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

A Appendix

A.1 Experimental Setup

A.1.1 Datasets and target model architectures

We use four datasets and corresponding model architectures that are consistent with the previous works (Shokri et al. [24]; Nasr et al. [14, 13]; Salem et al. [21]).

Purchase [18] is a 100 class classification task with 197,324 binary feature vectors of length 600; each dimension corresponds to a product and its value states if corresponding customer purchased the product; the corresponding label represents the shopping habit of the customer.

Texas is dataset of patient records. It is a 100 class classification task with 67,300 binary feature vectors of length 6,170; each dimension corresponds to symptoms and its value states if corresponding patient has the symptom or not; the label represents the treatment given to the patient. For Purchase and Texas we use fully connected (FC) networks.

CIFAR10 and CIFAR100 [12] are popular image classification datasets, each has size 50k and 32×32 color images. We use Alexnet, DenseNet-12 (with 0.77M parameters), and DenseNet-19 (with 25.6M parameters) models for CIFAR100, and Alexnet for CIFAR10. Following previous works, *we measure the test accuracy of the target models as their utility*.

Sizes of dataset splits. The dataset splits are given in Table 3. For Purchase and Texas tasks, we use D_{ref} of size 10k and *select* X_{ref} of size 10k from the remaining data using our entropy-based criterion. For CIFAR datasets, we use D_{ref} of size 25k and due to small sizes of these datasets, use the entire remaining 25k data as X_{ref}. The 'Attack training' (described shortly) column shows the MIA adversary's knowledge of members and non-members of D_{tr}. Following all the previous works, we assume that the adversary knows 50% of D_{tr}. Further experimental details are provided in Appendix.

Table 3: All the dataset splits are disjoint. D, D' data are the members and non-members of D_{tr} known to MIA adversary.

Dataset	DMP training		Attack training	
	$\lvert D_{\mathsf{tr}}\rvert$	$\lvert X_{\mathsf{ref}}\rvert$	$\lvert D\rvert$	$\lvert D'\rvert$
Purchase (P)	10000	10000	5000	5000
Texas (T)	10000	10000	5000	5000
CIFAR100 (C100)	25000	25000	12500	8000
CIFAR10 (C10)	25000	25000	12500	8000

A.1.2 Membership inference attacks

We briefly review the four MIAs we use for evaluations. Following previous works, *we use the accuracy of MIAs on target models as a measure of their membership privacy*.

Bounded loss (BL) attack [28] decides membership using a threshold on the target model's loss on the target sample. When 0-1 loss is used, the attack accuracy is simply the difference in training and test accuracy of target model. We denote BL attack accuracy by A_{bl}.

NSH attacks (Nasr et al. [14]) are similar to NN attacks. They concatenate various whitebox (e.g., model gradients) and/or blackbox (e.g., model loss, predictions) features of target model, while NN attack uses only the target model predictions. We denote whitebox and blackbox NSH attack accuracies by A_{wb} and A_{bb}, respectively. For NN and NSH attacks, we use the same attack models the original works.

A.2 Comparison with differentially private defenses

A.2.1 Comparison with DP-SGD.

Following the methodology of Jayaraman et al. [11], we compare DMP and DP-SGD [2] using the empirically observed tradeoffs between membership privacy (MIA resistance) and A_{test} of models. We use only CIFAR10 for these experiments, as the DP-SGD achieves prohibitively low accuracies on difficult tasks such as Texas and CIFAR100. We evaluate MIA risk using the whitebox NSH attack. Table 5 shows the results of Alexnet trained on CIFAR10 using DMP and DP-SGD with different privacy budgets ϵ's; -ve E_{gen} implies A_{train} is lower than A_{test}. DP-SGD incurs significant (35%) loss in A_{test} at lower ϵ (12.5) to provide strong membership privacy. At higher ϵ, A_{test} of DP-SGD increases, but at the cost of very high generalization error, which facilitates stronger MIAs. Note that, further increase in privacy budget, ϵ, does not improve tradeoff of DP-SGD. More importantly, for low MIA risk of $\sim 51.3\%$, DMP models have 12.8% higher A_{test} (i.e., 24.5% improvement) than DP-SGD models, which shows the superior tradeoffs due to DMP.

Table 4: Evaluating three state-of-the-art regularizers, with similar, low MIA risks (high membership privacy) as DMP. A^{+}_{test} shows *the % increase* in A_{test} due to DMP over the corresponding regularizers.

Purchase + FC (DMP's A_{test} = 74.1)						
Regularizer	E_{gen}	A_{test}	A^{+}_{test}	A_{wb}	A_{bb}	A_{bl}
WD	10.3	42.5	**+74.4%**	54.9	55.4	55.2
WD + DR	9.1	42.1	**+76.0%**	56.4	56.8	54.6
WD + LS	12.3	42.0	**+76.4%**	57.2	57.0	56.2
Texas + FC (DMP's A_{test} = 48.6)						
Regularizer	E_{gen}	A_{test}	A^{+}_{test}	A_{wb}	A_{bb}	A_{bl}
WD	5.0	22.5	**+116%**	58.3	57.7	52.5
WD + DR	6.1	14.2	**+242%**	63.1	62.6	53.1
WD + LS	8.3	37.3	**+30%**	61.7	61.0	54.2
CIFAR100 + DenseNet-12 (DMP's A_{test} = 63.1)						
Regularizer	E_{gen}	A_{test}	A^{+}_{test}	A_{wb}	A_{bb}	A_{bl}
WD	4.0	26.3	**+140%**	49.9	49.7	52.0
WD + DR	3.7	32.3	**+95.4%**	51.2	51.0	51.9
WD + LS	2.7	13.0	**+385%**	51.0	51.4	51.4
CIFAR10 + Alexnet (DMP's A_{test} = 65.0)						
Regularizer	E_{gen}	A_{test}	A^{+}_{test}	A_{wb}	A_{bb}	A_{bl}
WD	4.1	45.9	**+41.6%**	52.4	52.5	52.1
WD + DR	3.2	44.7	**+45.4%**	51.9	51.7	51.6
WD + LS	4.8	53.2	**+22.2%**	53.8	53.0	52.4

Table 5: DP-SGD versus DMP for CIFAR10 and Alexnet. For low MIA risk of $\sim 51.3\%$, DMP achieves 24.5% higher A_{test} than of DP-SGD (12.8% absolute increase in A_{test}).

Defense	Privacy budget (ϵ)	E_{gen}	A_{test}	A_{wb}
No defense	–	32.5	67.5	77.9
DMP	–	3.10	65.0	51.3
DP-SGD	198.5	3.60	52.2	51.7
	50.2	1.30	36.9	50.2
	12.5	0.30	31.7	50.0
	6.8	-1.60	29.4	49.9

A.2.2 Comparison with PATE.

PATE [15], a semi-supervized learning technique, requires a compatible pair of generator and disciminator to achieve acceptable performances. Hence, we use CIFAR10 dataset and, instead of Alexnet, use the generator-discriminator pair from [22], which has state-of-the-art performances. PATE trains a set of teachers, computes hard labels of each teacher on some X_{ref}, aggregates the labels for each $\mathbf{x} \in X_{\text{ref}}$ using majority voting, adds DP noise to the aggregate, and finally trains its target model on the noisy aggregate.

We train ensembles of 5, 10, and 25 teachers using D_{tr} of sise 25k. We use the optimized confident-GNMax (GNMax) aggregation scheme of [16] to label X_{ref} We present the results in Table 6. At low ϵ's ($<$10), GNMax hardly produces any labels, hence, the final target model has very low A_{test}, but at higher ϵ's ($>$1000), PATE target model has acceptable A_{test}. However, PATE cannot achieve performances even close to DMP, as it divides D_{tr} among its teachers. Such teachers have low accuracies and their ensemble cannot achieve the accuracy close to that of the unprotected model of DMP, which is trained on the entire D_{tr}. Hence, the quality of knowledge transferred in DMP is always higher than that in PATE.

Table 6: Comparing PATE with DMP. DMP has E_{gen}, A_{test}, and A_{wb} of 1.19%, 76.79%, and 50.8%, respectively. PATE has low accuracy even at high privacy budgets, as it divides data among teachers and produces low accuracy ensembles.

# of Teachers	Queries answered	Privacy budget (ϵ)	Target model E_{gen}	Target model A_{test}	A_{wb}
5	49	195.9	31.4	33.9	49.1
	1163	11684	65.4	68.1	49.0
10	23	42.9	39.1	38.3	50.1
	1527	6535	63.9	65.2	49.8
25	108	183.5	53.8	55.7	49.0
	4933	1794.1	57.8	60.3	48.6