Introduction To Overparameterized ML
Ben Gurion University
Semester B 2024
Computer Science Department

Final Project

Gal Alon | Yuval Sharon

1 Paper Summary - *A Blessing of Dimensionality in Membership Inference through Regularization [Tan et al., 2023]*

Jasper Tan, Daniel LeJeune, Blake Mason, Hamid Javadi, Richard G. Baraniuk

1.1 Introduction and Motivation

The paper main research question is - *Is overparameterization a privacy liability?*. In this paper the authors studies the connection between the parameterization level of a model and its vulnerability to Membership inference (MI) attack. The main contributions of the paper are:

- The authors show that increasing the number of parameters or decreasing regularization in a classification model can reduce privacy.
- The authors identify situations where wider neural networks (NNs) exhibit an improved privacy–utility trade-off due to regularization compared to narrower ones, while increased generalization performance from overparameterization aligns with privacy when controlled by regularization.
- The authors theoretically analyze high-dimensional logistic regression in the asymptotic regime, and derive fundamental MI vulnerability in overparameterized logistic regression models.

In this paper, the authors focused on the *black box* setting - where the attackers don't have an access to the model itself, and only to its predictions (and therefore we can also compute the loss).

1.2 Membership Inference Setting

Membership inference in this paper involves determining whether a specific data point was part of a model's training set. An MI adversary A predicts if a given data point (x, y) was in the training set by analyzing the model's output $\hat{f}(x)$ and the associated loss. The focus is on single-query black-box adversaries, which only have access to the data point and the model's output, not its internal workings. The attack used in this paper is a sample-specific loss threshold method, where the adversary compares the loss $\ell(y, f(x))$ against a learned threshold $\tau(x, y)$ to determine membership.

1.3 The Privacy-Utility Trade off

The core contribution of this paper lies in demonstrating how regularization can balance the trade-off between privacy and utility. The authors show that overparameterized models, when properly regularized, can achieve lower test errors while maintaining or even enhancing privacy. This phenomenon is illustrated in Figure 3, where the relationship between model width and MI advantage is plotted, show casing how increasing the models parameterization level generally increases MI advantage on the model even as it decreases its test error.

1.3.1 Theoretical Analysis

The analysis is conducted within the setting of logistic regression models, where they explore how varying the number of parameters and regularization affects model behavior. Central to their findings is Theorem 4, which highlights two key insights:

- As models become more overparameterized, their vulnerability to MI attacks increases, with the MI advantage converging to 1.
- Despite this increased risk, overparameterization can still enhance generalization under certain conditions, illustrating the trade-off between accuracy and privacy.

The authors utilized logistic regression models with a focus on the interplay between parameter count, regularization, and their effects on MI advantage and generalization performance. For a detailed mathematical analysis and proofs, refer to the original paper.

1.3.2 Experimental Analysis

The trade-off between privacy and utility was theoretically proven for logistic regression models. In this section, the authors extend this analysis to several neural network architectures to empirically demonstrate this phenomenon (3). Beyond merely exploring this trend, the authors also present a counter intuitive finding: when a model is jointly tuned for both its parameterization level and the amount of regularization, the privacy-utility trade-off can be eliminated. In this study, the focus is on early stopping as the regularization mechanism.
The key observations from these experiments are:

- The decrease in the model's generalization error and the increase in the adversary's MI advantage occur at different rates during training for networks with varying widths [4].
- Wider networks generally achieve better privacy-utility trade-offs compared to narrower networks.
- Proper regularization, specifically through early stopping, can effectively eliminate the privacy-utility trade-off in overparameterized models [5]

1.4 Discussion

This paper shows that while overparameterization can increase privacy risks, it doesn't always have to. With proper regularization, overparameterized models can achieve greater privacy. However, this result may not apply universally, as it depends on the model type, MI methods, training processes, and the choice of regularization.

2 Connection to Course Material: Overparameterized Machine Learning

The paper's exploration of overparameterized models and their privacy-utility trade-offs directly relates to key concepts covered in our course on overparameterized machine learning. Here are the main connections:

2.1 Generalization in Overparameterized Regimes

As we studied in class machine learning models can still generalize well in the overparameterized regime (where p>n) and thus have similar or better test accuracy in those areas (of course, under certain conditions). This paper extends this concept by examining how overparameterization affects not only generalization but also privacy risks.

2.2 Trade-off Between Parameterization and Privacy

The paper builds on our course concepts by highlighting how overparameterized models are more vulnerable to Membership Inference (MI) attacks, a topic we've discussed in the context of model memorization risks. Specifically, the increased susceptibility of overparameterized models to privacy attacks, such as MI attacks, was established in [Tan et al., 2022], where a trade-off between parameterization and privacy was proven for linear models. Building on this work, [Tan et al., 2023] suggests that proper regularization techniques, such as early stopping, can mitigate these risks, striking a balance between model capacity and privacy without compromising utility.

2.3 Role of Regularization

In our course, we examined how regularization techniques can effectively manage the complexities of overparameterized models, particularly in mitigating test error peaks (as we saw in Figure 1 from [Nakkiran et al., 2021]). This paper extends that understanding by demonstrating how regularization, especially through early stopping, plays a crucial role in balancing the privacy-utility trade-off. The findings emphasize that with the right regularization, overparameterized models can retain their generalization benefits while significantly reducing MI vulnerabilities.

2.4 Empirical Analysis of Neural Networks

The paper's experimental analysis on various neural network architectures aligns with our course's (and especially its second part) focus on overparameterized deep learning models, through the following papers: [Nakkiran et al., 2019] [Somepalli et al., 2022]

3 Extension of the Experimental Results

Building on the analysis and findings of the paper, we conducted a series of additional experiments aimed at extending the original work. These experiments were designed to explore different aspects of the privacy-utility trade-off in overparameterized models by varying key factors that influence Membership Inference Attack (MIA) vulnerability. The objectives of our extension include:

- Repeating the original experiments with alternative MIA methods to assess the robustness of the findings across different attack strategies.
- Investigating the impact of removing data augmentation on the privacy-utility trade-off, motivated by the understanding that data augmentation inherently increases the regularization level of the model.
- Evaluating the effect of weight decay regularization, varying the levels of weight decay to better understand its role in controlling MIA vulnerability.

These extensions are intended to deepen our understanding of the conditions under which overparameterization affects model privacy and to explore how different regularization techniques and optimization strategies can mitigate these risks. Our code is available here.

3.1 Motivation

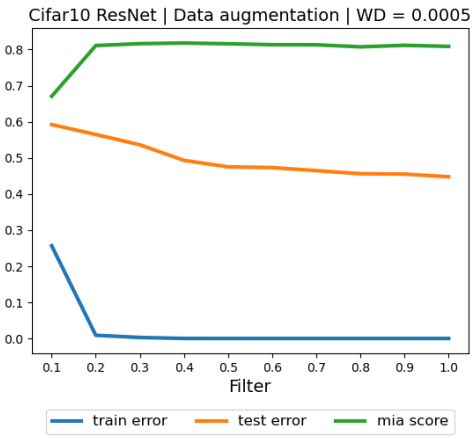
The original study's findings were based primarily on a single Membership Inference Attack method. To strengthen the generality of their conclusions, we incorporated another commonly used technique from [Kurmanji et al., 2023]. This extension demonstrates that the observed privacy-utility trade-offs are not dependent on a specific MI attack but hold across different adversarial strategies, providing stronger evidence for the paper's central claims.

A significant assertion of the paper is that regularization can mitigate MI risks in overparameterized models, primarily focusing on early stopping as the regularization method for Deep Neural Networks (DNNs). To broaden the scope of this claim, we explored additional regularization techniques, such as weight decay and data augmentation, to see if similar benefits are observed. This allows us to assess whether the proposed privacy-utility balance extends beyond early stopping to other common regularization methods.

We also examined the specific accuracy-privacy trade-offs introduced by each regularization method. Understanding how different strategies impact this balance could offer valuable insights for practical applications, guiding the choice of techniques depending on the need for privacy or accuracy in overparameterized models.

3.2 The Proposed MI Attack

We employed a Membership Inference Attack (MIA) inspired by the method in [Kurmanji et al., 2023]. This black-box approach only utilizes the model's inputs and outputs, without access to internal parameters.



(a)

Figure 1: MIA score and train/test errors for ResNet18 trained on CIFAR=100. here we used the MI method from 3.2 . Note that filter is equivelent to the width parameter of ResNet18, where filter=1 is the standart model, and filter=0.1 is ResNet18 with width parameter 1.

Originally described by [Shokri et al., 2017], this method uses shadow models to simulate the target model's behavior. These shadow models, trained on datasets with known membership status, enable the attacker to learn a function $h_\theta(f(x))$ that predicts whether a data point x was part of the training set D_{train} based on the model's output.
In our implementation, the attacker is a binary classifier trained to distinguish between members and non-members of the training set based on cross-entropy loss values. The loss $l(x, y)$ is calculated for each example, and the classifier predicts membership status using loss values from both the training set D_{train} and a held-out test set D_{test}. The classifier's performance is then evaluated on a separate, disjoint set.

3.3 Alternative MI Attack

3.3.1 Introduction

This was our initial experiment, aimed at verifying that the main claim of the paper holds true for this method before extending it to other experiments.

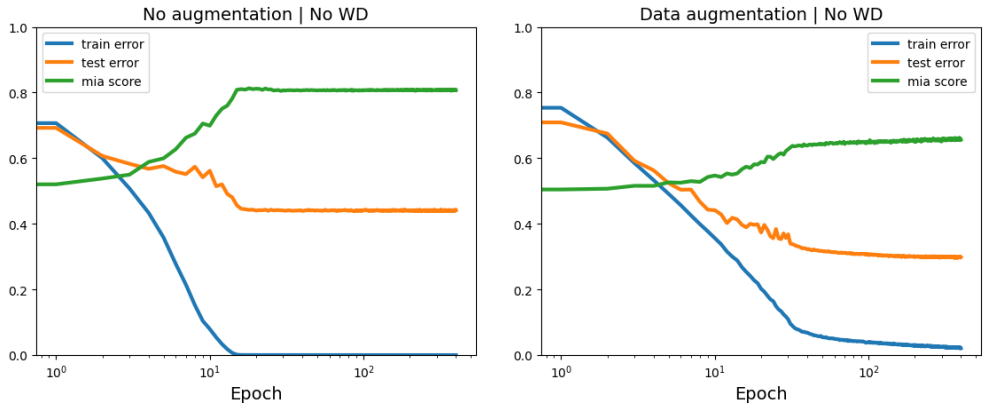
3.3.2 Results

Our results are in Figure 1, we can see that same trend still stand - over parameterized models are more volunerable to MI attacks.

3.4 The Effect of Removing Data Augmentation

3.4.1 Introduction

In the original paper, the authors used early stopping as the primary regularization method. Although they employed data augmentation, it wasn't explicitly discussed as a source of regularization. Given that data augmentation inherently introduces a level of implicit regularization, we aimed to explore its impact on the Membership Inference Attack (MIA) score by removing it from the training process.



(a) (b)

Figure 2: The effect of data augmentation on the test and train error, and on the MIA score. The model here is ResNet-18 trained of CIFAR-100.

3.4.2 Results

Figure 2 illustrates the test/train error and MIA score as a function of the epoch number. As observed, without data augmentation, the model exhibits higher test error and increased vulnerability to MI attacks. Conversely, applying standard data augmentation results in better generalization (lower test error) and reduced susceptibility to MI attacks. It is also noteworthy that with data augmentation, the increased data complexity leads to a longer convergence time (more epochs) for the model.

3.5 The Effect of Weight Decay as The Regularization Method

3.5.1 Introduction

To further support the claim that regularization can mitigate the risk of MI attacks in overparameterized models, we conducted experiments by training a model with varying levels of weight decay regularization, both with and without data augmentation (details in the next subsection). Our objective was to investigate the trade-off between the weight decay level and both the test/train error and the MIA score, aiming to understand how different levels of regularization influence this balance between model performance and vulnerability.

3.5.2 Results

Weight Decay Level	Train Error	Test Error	MIA Score
0	0.01%	44%	80.7%
0.0005	0.01%	46.4%	88.6%
0.001	0.01%	45.3%	84.1%
0.01	1.2%	**43.2%**	77.6%
0.1	72.3%	82.7%	**59.8%**

Table 1: Results for different weight decay levels, without data augmentation.

Weight Decay Level	Train Error	Test Error	MIA Score
0	2%	29.8%	65.6%
0.0005	3.6%	28.6%	63.4%
0.001	5.7%	**27.8%**	60.8%
0.01	28.7%	36.7%	52.9%
0.1	82.1%	82%	**49.9%**

Table 2: Results for different weight decay levels, with data augmentation.

With data augmentation, the best balance between utility and privacy is achieved with a weight decay (wd) of 0.001, whereas without augmentation, the optimal result is obtained with wd of 0.01. This outcome can be expected, as data augmentation itself acts as a form of regularization. Overall, these findings further highlights that selecting the appropriate regularization level is crucial for effectively balancing the utility-privacy trade-off.
More visualization of these experiments is in the appendix 6, 7.

3.6 Significance of the Extension

In the experiments above, we empirically demonstrated that the main claims of the original paper hold true across different MIA methods and various regularization techniques. These results reinforce and extend the original findings, showing that it is indeed possible to control the privacy levels of overparameterized models through regularization.
The regularization methods we explored directly influence the model's weights, making them more comparable to other commonly used techniques (compared to early stopping). This suggests that similar effects on privacy might be achievable with other regularization methods as well.
Moreover, our findings regarding the use of data augmentation and weight decay regularization have significant practical implications. They suggest that it is possible to maintain a certain level of privacy without relying on a held-out validation set, offering more flexibility and efficiency in model training and deployment.

3.7 Suggestion For Future Work

The MIAs used in this study are common but not the most advanced. Future work could focus on using more sophisticated and accurate attacks to better understand model vulnerabilities and privacy risks in more realistic scenarios.
We suspect that the training procedure, including factors like different optimizers, learning rates, or extended training, significantly affects this trade-off. However, due to the scope of this work, we chose to focus on other aspects.

References

[Kurmanji et al., 2023] Kurmanji, M., Triantafillou, P., Hayes, J., and Triantafillou, E. (2023). Towards unbounded machine unlearning.

[Nakkiran et al., 2019] Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. (2019). Deep double descent: Where bigger models and more data hurt.

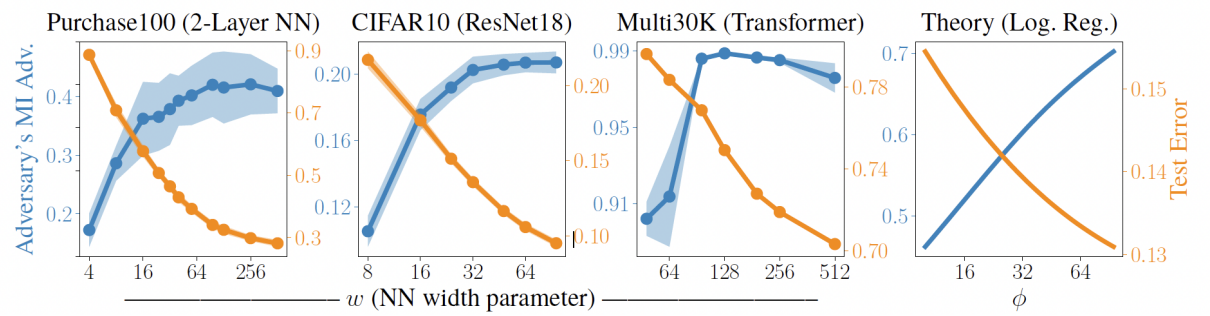
[Nakkiran et al., 2021] Nakkiran, P., Venkat, P., Kakade, S., and Ma, T. (2021). Optimal regularization can mitigate double descent.

[Shokri et al., 2017] Shokri, R., Stronati, M., Song, C., and Shmatikov, V. (2017). Membership inference attacks against machine learning models.

[Somepalli et al., 2022] Somepalli, G., Fowl, L., Bansal, A., Yeh-Chiang, P., Dar, Y., Baraniuk, R., Goldblum, M., and Goldstein, T. (2022). Can neural nets learn the same model twice? investigating reproducibility and double descent from the decision boundary perspective.

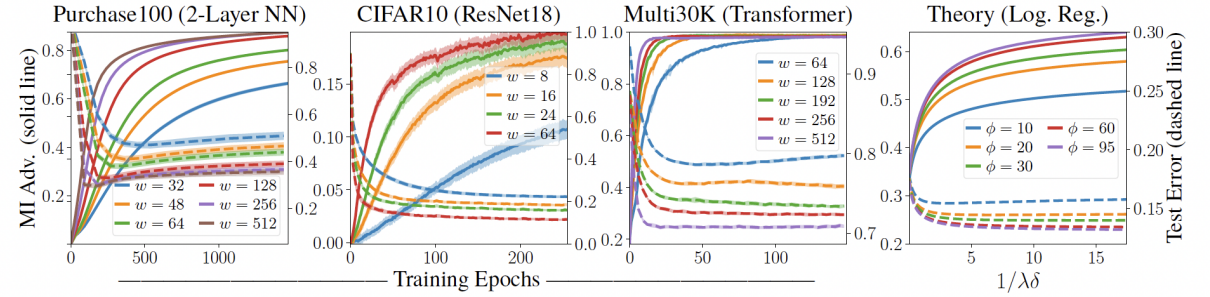
[Tan et al., 2023] Tan, J., LeJeune, D., Mason, B., Javadi, H., and Baraniuk, R. G. (2023). A blessing of dimensionality in membership inference through regularization. In Ruiz, F., Dy, J., and van de Meent, J.-W., editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 10968–10993. PMLR.

[Tan et al., 2022] Tan, J., Mason, B., Javadi, H., and Baraniuk, R. G. (2022). Parameters or privacy: A provable tradeoff between overparameterization and membership inference.



(a)

Figure 3: Figure 3 from [Tan et al., 2023], showing the effect of the parameterization level (NN width) on the test accuracy and on the MI Adv.



(a)

Figure 4: Figure 4 from [Tan et al., 2023], showing the effect of the early stopping regularization on the test accuracy and on the MI Adv for multiple net widths.

A Figures from original paper

Figure 3 is taken from [Tan et al., 2022].
Figure 4 is taken from [Tan et al., 2022].
Figure 5 is taken from [Tan et al., 2022].

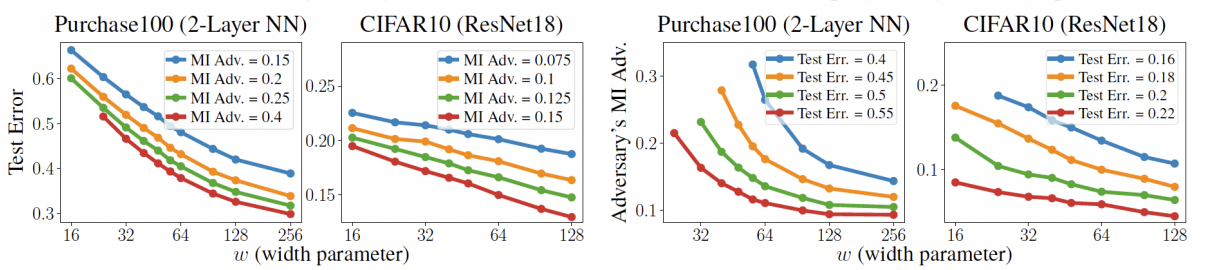
B Technical Details for 3

B.1 Data Set

For our experiments, we used the CIFAR-100 dataset, which is a widely recognized benchmark in the field of deep learning, particularly for testing and evaluating deep neural networks (DNNs). The CIFAR-100 dataset consists of 50,000 training images and 10,000 test images, all divided into 600 distinct classes. Each image in this dataset is a small 32x32 pixel color image. A subset of this dataset (CIFAR-10) was also used in the original paper to explore the privacy-utility trade-offs in overparameterized models. More details about the dataset can be found here.

B.2 Model

We used the ResNet18 model, as discussed in class, with the ability to control its width parameters, allowing us to adjust the level of parameterization. The model implementation



(a) Test error vs. network width for fixed MI adv.

(b) MI adv. vs. network width for fixed test error.

(a)

Figure 5: Figure 6 from [Tan et al., 2023], showing how proper regularization, specifically through early stopping, can effectively eliminate the privacy-utility trade-off in overparameterized models

can be found here.

B.3 Technical Details

We trained the model using the cross-entropy loss function and the SGD optimizer with a learning rate of 0.01 and a momentum of 0.9. The training was conducted with a batch size of 128 over 400 epochs.

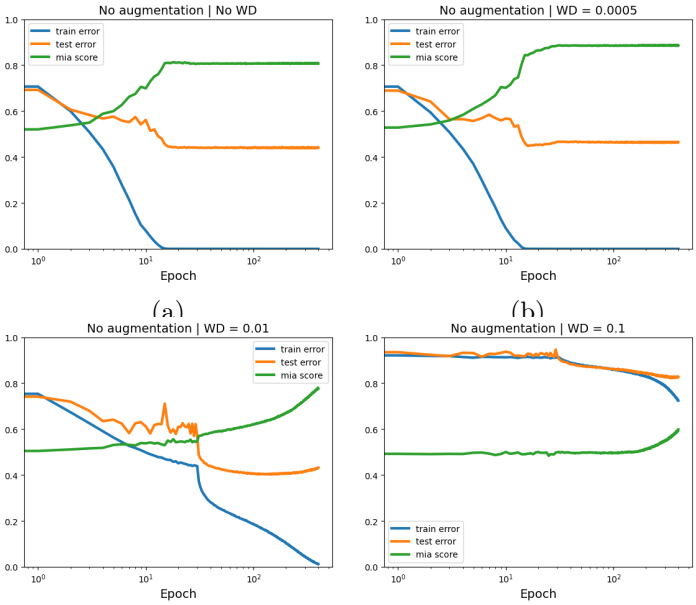


Figure 6: The effect weight decay on the test and train error, and on the MIA score. The model here is ResNet-18 trained of CIFAR-100.

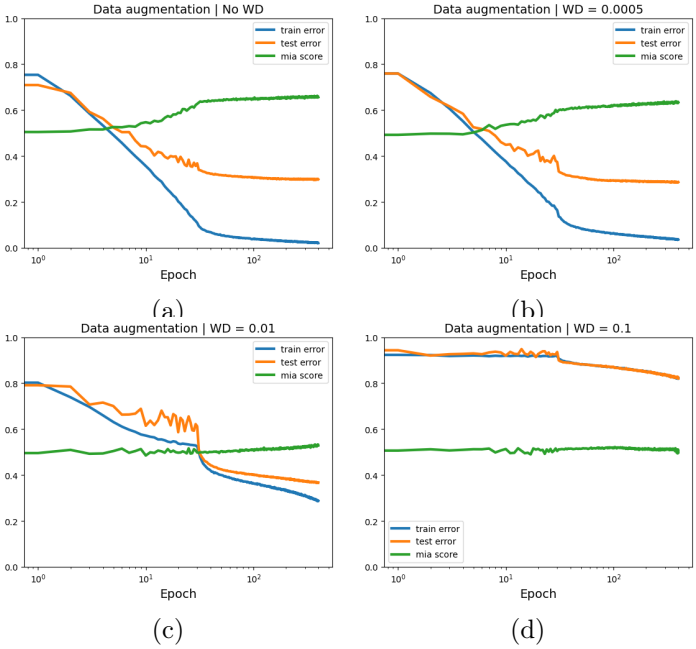


Figure 7: The effect of wieght deacy on on the test and train error, and on the MIA score. The model here is ResNet-18 trained of CIFAR-100.