Theoretical Improvements in Privacy Auditing with Single-Run Methods

Michael Jung
michael.jung@emory.edu
Emory University
, Atlanta , Georgia , USA

ABSTRACT

This paper explores potential theoretical improvements to the existing framework of privacy auditing using a single training run, as proposed by Steinke et al. (2023). While their method reduces computational overhead by auditing privacy with a single run, challenges remain related to the precision of privacy loss estimation and applicability across various differential privacy (DP) models. We propose optimizations to the auditing process by integrating more advanced attack methods, such as Likelihood Ratio Attack (LiRA) and Transfer Learning Attack (TLA) which aim to enhance the precision of Membership Inference Attacks (MIA). Additionally, we explore the use of Rényi Differential Privacy (RDP) to tighten the bounds of privacy loss estimation. Our investigation introduces practical extensions that improve audit accuracy while maintaining computational efficiency. While theoretical, these improvements provide new insights for future empirical work in privacy-preserving machine learning frameworks.

ACM Reference Format:
Michael Jung. 2024. Theoretical Improvements in Privacy Auditing with Single-Run Methods. In . ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In today's data-driven world, privacy has become a critical concern, especially in machine learning systems that rely on vast amounts of personal data. As organizations use increasingly sensitive information to train models, ensuring that individual privacy is preserved has never been more important. Differential Privacy (DP) has emerged as a widely adopted standard for providing rigorous privacy guarantees while still allowing models to maintain their utility. The core principle of DP ensures that the presence or absence of a single individual's data in a dataset has a negligible effect on the output of the model, thus protecting user privacy.

Despite the robustness of DP, auditing its guarantees remains computationally expensive. Privacy audits typically require multiple training runs to estimate privacy loss parameters like epsilon (ϵ), which directly measure the level of privacy exposure. These repetitive training cycles are time-consuming and resource-intensive, creating a demand for more efficient auditing mechanisms.

A recent framework introduced by Steinke et al. offers a solution to this challenge through their one-training-run privacy audit method. This approach significantly reduces the computational burden by estimating privacy parameters after a single training run. By leveraging multiple independent data points and examining their inclusion or exclusion from the dataset, their method simplifies the auditing process without compromising the integrity of the model. However, while the one-training-run method has achieved promising results, there are discrepancies between the empirical privacy loss estimates and the theoretical privacy guarantees, leaving room for further refinement.

In this paper, we explore potential improvements to the one-training-run privacy audit framework. We aim to enhance the precision of privacy loss estimation while maintaining computational efficiency. Specifically, we investigate how advanced attack methods, such as Likelihood Ratio Attack (LiRA) and Transfer Learning Attack (TLA), and alternative DP models, such as Rényi Differential Privacy (RDP), can optimize the audit process. By bridging the gap between empirical results and theoretical bounds, we hope to contribute to more accurate and broadly applicable privacy auditing methods for machine learning systems.

2 BACKGROUND

2.1 Differential Privacy (DP) and Its Role in Privacy Auditing:

Differential Privacy (DP) is a mathematical framework used to ensure that the inclusion or exclusion of a single data point does not significantly affect the outcome of an analysis, thus protecting individual privacy. It provides a formal privacy guarantee, commonly measured by a parameter ϵ, which quantifies the risk of data leakage. In the context of machine learning, DP mechanisms are employed to ensure that models trained on sensitive datasets maintain privacy guarantees while still being useful.

2.2 Steinke et al.'s One-Training-Run Privacy Audit:

Steinke et al. (2023) introduced a novel privacy auditing method designed to reduce the computational cost of traditional privacy audits, which often require multiple training runs. Their approach focuses on auditing privacy by performing a single model training run, while analyzing the inclusion or exclusion of multiple data points. This method helps estimate privacy parameters, such as ϵ, with fewer resources. Although the method is efficient, it faces

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

challenges in precisely estimating privacy loss, and empirical results sometimes diverge from theoretical upper bounds.

2.3 Privacy Auditing Attack Methods

2.3.1 Membership Inference Attack (MIA). Membership Inference Attacks aim to determine whether a particular data point was part of the training set by analyzing how a model reacts to different inputs. Two notable techniques are used under MIA: Gradient Space Attacks, where attackers evaluate how gradients change depending on data inclusion, and Input Space Attacks, which analyze how input variations affect predictions to infer data membership.

2.3.2 Transfer Learning Attack (TLA). Transfer Learning Attack (TLA), as discussed by Adrian et al. in "Transfer Learning for Security: Challenges and Future Directions," involves leveraging pre-trained models from one domain and fine-tuning them on a related task to enhance the attacker's inference capabilities. In the context of membership inference attacks, the attacker utilizes knowledge from the source model to make more informed guesses about whether specific data points were part of the training set of the target model. By focusing on shared representations learned during pre-training, TLA can exploit vulnerabilities in the model transfer process, making it a powerful tool for attacking models with less training data or highly sensitive datasets.

2.3.3 Likelihood Ratio Attack (LiRA). Nicholas et al., in "Membership Inference Attacks From First Principles," describe Likelihood Ratio Attack (LiRA) as an attack method that significantly improves the inference accuracy by applying a likelihood-based approach. LiRA calculates the likelihood of whether a data point was in the training set by comparing the model's confidence scores for that point when trained on a dataset including the point versus excluding it. Unlike other attacks that rely solely on direct confidence scores, LiRA improves robustness by using statistical tests to quantify the uncertainty and reduce the chances of overfitting to the attack model's assumptions, making it a more accurate tool for privacy auditing.

2.4 Rényi Differential Privacy (RDP):

Rényi Differential Privacy (RDP) extends the standard DP definition by using Rényi divergence to provide a tighter and more flexible privacy guarantee. RDP enables finer control over the trade-off between privacy and utility, especially for mechanisms that involve repeated subsampling or multiple queries. Steinke et al.'s framework could potentially benefit from adopting RDP to achieve tighter privacy bounds and more accurate auditing results, as it provides a more nuanced view of privacy loss compared to traditional DP.

3 METHODS

3.1 Differentially Private Stochastic Gradient Descent (DP-SGD)

Input:

- Dataset $X = \{x_1, x_2, ..., x_n\}$ with n data points.
- Model parameters $w \in \mathbb{R}^d$, where d is the number of model parameters.
- Loss function $f : \mathbb{R}^d \times X \to \mathbb{R}$.
- Hyperparameters: Number of iterations t, clipping threshold c, noise multiplier σ, sampling probability q, learning rate η.

Output:

- Trained model parameters w_t after t iterations.

Objective: Minimize the loss function $f(w, x)$ while ensuring that the model satisfies differential privacy by adding noise to the gradients during training and clipping them to a threshold to control the influence of any single data point.

Main Related Equation:

$$w_t = w_{t-1} - \eta \left(\sum_{i \in S^t} \text{clip}(g_i^t, c) + \mathcal{N}(0, \sigma^2 c^2 I) \right)$$

where:

- η is the learning rate.
- g_i^t is the gradient at time t for data point x_i.
- $\text{clip}(g_i^t, c)$ limits the norm of the gradient to a threshold c.
- $\mathcal{N}(0, \sigma^2 c^2 I)$ is the Gaussian noise added for privacy preservation.

3.2 Transfer Learning Attack (TLA)

Input:

- Pre-trained source model M_{source} (e.g., trained on CIFAR-100).
- Fine-tuned target model M_{target} (e.g., fine-tuned on CIFAR-10).
- Target dataset $D = \{x_1, x_2, ..., x_n\}$ for which membership inference needs to be performed.

Output:

- Enhanced membership inference attack performance based on the transferred knowledge.

Objective: Leverage pre-trained models to improve the performance of membership inference attacks on small target datasets by transferring features learned in a related source domain.

Main Related Equation:

$$L_{\text{TLA}} = L_{\text{source}} + \alpha L_{\text{target}}$$

where L_{source} is the loss on the source dataset, L_{target} is the loss on the target dataset, and α is a balancing parameter that controls the contribution of the source model to the target model's fine-tuning.

3.3 Likelihood Ratio Attack (LiRA)

Input:

- Model outputs (such as confidence scores or probability distributions) on in-sample and out-of-sample data points.
- Target dataset D and its associated model.

Output:

- A likelihood ratio score for each data point, used to infer whether the point was part of the training set.

Objective: Infer data point membership in the training set by calculating the likelihood ratios, thereby assessing the vulnerability of the model to membership inference attacks.

Main Related Function:

$$\Lambda = \frac{P(\text{conf}_{\text{obs}} | \mathcal{N}(\mu_{\text{in}}, \sigma_{\text{in}}^2))}{P(\text{conf}_{\text{obs}} | \mathcal{N}(\mu_{\text{out}}, \sigma_{\text{out}}^2))}$$

This function compares the likelihood of the observed confidence score conf_{obs} given the Gaussian distributions of the confidence scores for the "in" and "out" datasets. The ratio helps infer whether the data point was likely included in the training set D or not.

3.4 Rényi Differential Privacy (RDP)

Input:

- A dataset $D = \{x_1, x_2, \ldots, x_n\}$ and a trained model $\mathcal{M}$
- Privacy parameter α, the order of the Rényi divergence.

Output:

- A privacy loss estimate $\epsilon(\alpha)$, giving a tighter bound on privacy leakage compared to traditional DP.

Objective: Provide a more flexible privacy accounting mechanism by using Rényi divergence, which generalizes traditional differential privacy by allowing the use of different privacy parameters (orders).

Main Related Function:

$$D_\alpha(P||Q) \triangleq \frac{1}{\alpha - 1} \log \mathbb{E}_{x \sim Q} \left(\frac{P(x)}{Q(x)} \right)^\alpha$$

This measures the Rényi divergence between two probability distributions P and Q, which quantifies the privacy loss under the RDP framework. The parameter α controls the sensitivity of the divergence measurement and its generalization over various divergence metrics.

4 RELATED WORK

In recent years, differential privacy has gained prominence as a robust framework for ensuring privacy in machine learning. Various methods have been proposed and studied to improve privacy guarantees, balance privacy-utility trade-offs, and mitigate privacy attacks.

4.1 Differentially Private Stochastic Gradient Descent (DP-SGD)

DP-SGD is one of the most widely used methods for ensuring privacy in machine learning. It introduces noise into the gradient updates during training and clips the gradients to prevent any single data point from having a disproportionate effect on the model's parameters. Abadi et al. (2016) first proposed DP-SGD, demonstrating its effectiveness for training deep neural networks under differential privacy constraints. Cynthia Dwork, one of the pioneers in differential privacy, introduced the foundational concept of differential privacy in her seminal works, which laid the foundation for subsequent methods, including DP-SGD. Further enhancements to DP-SGD, including optimal noise addition strategies and more efficient gradient clipping methods, have been explored.

4.2 Membership Inference Attacks (MIA)

Membership Inference Attacks (MIA) aim to determine whether a specific data point was included in the training dataset, posing a significant privacy risk. Shokri et al. (2017) first introduced MIAs in the context of deep learning, highlighting how adversaries could exploit model outputs to infer membership. Recent work by Nicholas et al. (2020) extended this concept by analyzing membership inference attacks from first principles, providing deeper insights into how these attacks can be executed and prevented.

4.3 Transfer Learning Attack (TLA)

Transfer Learning Attack (TLA) is an emerging technique to enhance the performance of membership inference attacks by leveraging a model pre-trained on a related dataset. Adrian et al. (2020) highlight the challenges and future directions for transfer learning in security, including its application in privacy attacks. This technique allows attackers to exploit shared features between datasets, potentially increasing the attack's effectiveness. Understanding the interaction between transfer learning and differential privacy mechanisms is an area of ongoing research.

4.4 Rényi Differential Privacy (RDP) and Other Advanced DP Methods

While traditional differential privacy methods like DP-SGD are effective, more advanced techniques, such as Rényi Differential Privacy (RDP), Approximate Differential Privacy (ADP), and Local Differential Privacy (LDP), are being actively explored to provide stronger privacy guarantees. RDP, introduced by Mironov (2017), offers a more flexible privacy accounting framework and allows for tighter bounds on privacy loss. ADP and LDP extend traditional DP by considering approximate privacy guarantees and decentralized data collection scenarios, respectively. These methods provide new avenues for improving privacy-preserving machine learning, particularly in settings with high-dimensional data or multiple data sources.

5 EXPERIMENTAL SETTINGS

5.1 Datasets

CIFAR-10 dataset is used for training the models and evaluating the impact of privacy mechanisms. CIFAR-10 consists of 60,000 32x32 color images across 10 classes, with 6,000 images per class. Additionally, **CIFAR-100** will be used for the Transfer Learning Attack (TLA), where the model is pre-trained on CIFAR-100 and fine-tuned on CIFAR-10 to simulate an adversarial transfer learning scenario. The CIFAR-10 and CIFAR-100 datasets are commonly used in image classification tasks and are well-suited for evaluating model performance and privacy mechanisms.

5.2 Model Architecture

ResNet-18, a convolutional neural network (CNN) known for its residual learning capabilities, which helps in training deeper models by mitigating the vanishing gradient problem will be used. This model architecture is selected due to its efficiency in image classification tasks while being computationally feasible for our privacy experiments.

5.3 Privacy Mechanisms

In this experiment, we evaluate four different privacy mechanisms:

- **Differentially Private Stochastic Gradient Descent (DP-SGD)**: DP-SGD adds noise to the gradients during the training process to ensure privacy by limiting the amount of

information any single data point can influence. We explore various values of the privacy parameter ϵ to assess the trade-off between privacy and model performance.

- **Transfer Learning Attack (TLA)**: The TLA is evaluated by pre-training the model on **CIFAR-100** and fine-tuning it on **CIFAR-10**. The model's vulnerability to membership inference attacks is tested by assessing the success of the transfer learning process.
- **Likelihood Ratio Attack (LiRA)**: LiRA attacks will utilize confidence scores produced by the model to determine the likelihood that a data point was part of the training set. The attack will rely on the likelihood ratio as described in the relevant literature.
- **Rényi Differential Privacy (RDP)**: RDP provides a more flexible privacy accounting mechanism by using Rényi divergence, which generalizes traditional differential privacy. We will compute the privacy loss for each model under the RDP framework using varying values of the Rényi divergence order α.

5.4 Attack Setup

- **TLA**: The attack will be performed by fine-tuning the model on **CIFAR-100** and using it to infer membership for **CIFAR-10** data points. The success of the attack will be evaluated based on how accurately the transfer-learned model identifies the membership of data points.
- **LiRA**: For LiRA, the likelihood ratio will be computed for each data point, comparing the confidence scores of the model for in-sample and out-of-sample data. The model will be trained with DP-SGD and the attack will assess its ability to distinguish between in-sample and out-of-sample data.

5.5 Evaluation Metrics

We will evaluate the performance of the models and attacks using the following metrics:

- **Accuracy**: The accuracy of the model on CIFAR-10, to assess its classification performance under different privacy mechanisms.
- **Attack Success Rate**: For both TLA and LiRA, the success rate will be measured, which indicates how accurately the attacks can infer the membership of a data point in the training set.
- **Privacy Parameter** ϵ: We will compare the impact of various values of ϵ in DP-SGD on model accuracy and privacy guarantees.
- **RDP Privacy Loss**: The privacy loss under the RDP framework will be computed to assess the strength of privacy guarantees provided by RDP.

6 LIMITATIONS AND FUTURE DIRECTIONS

6.1 Limitations

Despite the comprehensive theoretical framework and review of differential privacy mechanisms presented in this paper, several limitations should be acknowledged:

- **Technical Integration Challenges:** A key limitation of this work was the inability to successfully implement the `Opacus` library for integrating differential privacy into the model training process. Specifically, the `PrivacyEngine` encountered errors related to optimizer initialization and parameter mismatches, which prevented the completion of the planned experiments. This hindered the empirical validation of Differentially Private Stochastic Gradient Descent (DP-SGD), Transfer Learning Attacks (TLA), and Likelihood Ratio Attacks (LiRA).
- **Lack of Empirical Results:** As a consequence of the technical issues, this paper lacks an experimental evaluation section that quantitatively compares the performance of privacy-preserving models. The absence of experimental results limits the ability to demonstrate the practical trade-offs between privacy and model accuracy, which is a critical component for understanding the real-world impact of privacy mechanisms.
- **Limited Scope of Attacks Evaluated:** While this paper discussed several privacy attacks (TLA, LiRA, Membership Inference Attacks), the failure to implement these methods restricted the exploration of additional attack vectors and more robust attack-defense evaluations that would provide deeper insights into privacy vulnerabilities in machine learning models.

6.2 Future Directions

To address the limitations and further advance the field, several directions for future work are suggested:

- **Successful Implementation of Privacy Mechanisms:** Future work should prioritize overcoming the technical challenges associated with the `Opacus` library and `PrivacyEngine`. Correcting the errors and ensuring compatibility between the privacy mechanisms and the model architecture is essential to obtaining empirical results that support the theoretical findings. Additionally, exploring alternative libraries or updated versions of `Opacus` could help resolve the integration issues faced in this work.
- **Broader Attack Vectors:** Expanding the scope of privacy attacks to include adversarial examples, model inversion attacks, and more sophisticated membership inference techniques could enhance the understanding of privacy risks. Comparative evaluations between these methods and differential privacy mechanisms would provide a more comprehensive analysis of privacy threats.
- **Empirical Privacy-Utility Trade-offs:** Future research should aim to empirically test the privacy-utility trade-offs by implementing DP-SGD, TLA, and LiRA in working environments. This would allow researchers to measure the impact of varying privacy parameters (ϵ) on model accuracy and attack success rates, leading to more informed recommendations for privacy preservation in machine learning.
- **Transferability of Privacy Mechanisms:** Another promising direction is exploring the transferability of privacy mechanisms to other datasets and model architectures. For example, the impact of DP-SGD or LiRA on large-scale datasets

(e.g., ImageNet) and deeper networks (e.g., ResNet-50, ResNet-101) could yield valuable insights into the generalizability of privacy-preserving techniques.

- **Tool and Framework Improvements:** Given the technical limitations encountered, future work may also involve contributing to the development of privacy libraries, addressing issues related to their ease of use, and improving their compatibility with different optimizers and model architectures. This could ensure more seamless adoption of differential privacy methods in the broader machine learning community.
- **Exploring Other Differential Privacy Methods:** A promising avenue for future research is the exploration of alternative or complementary differential privacy methods beyond the standard DP-SGD, such as Rényi Differential Privacy (RDP), Approximate Differential Privacy (ADP), and Local Differential Privacy (LDP). These methods offer different privacy-utility trade-offs and could provide enhanced privacy guarantees. RDP, for instance, allows for more flexible privacy accounting and might lead to stronger privacy guarantees for some use cases. ADP provides approximate privacy with lower computational overhead, making it useful for applications requiring high performance, while LDP allows for privacy preservation at the individual data level, even when data is shared across untrusted devices or parties. Research into these methods and their combination could help further refine the trade-offs between privacy and utility in machine learning systems.

7 CONCLUSION

This paper investigates various privacy-preserving techniques for machine learning models, focusing on Differentially Private Stochastic Gradient Descent (DP-SGD), Transfer Learning Attacks (TLA), Likelihood Ratio Attacks (LiRA), and Rényi Differential Privacy (RDP). While the theoretical analysis underscores the ability of DP-SGD to safeguard privacy by adding noise to gradients and clipping them, the challenges encountered during the implementation of Opacus limited the ability to empirically validate these methods. Thus, concrete experimental results could not be presented in this work.

Despite these implementation hurdles, the paper highlights the privacy-accuracy trade-offs inherent in these privacy methods. DP-SGD shows promise in protecting individual data points but introduces a loss in model accuracy as the privacy parameter (ϵ) increases. Transfer Learning Attacks (TLA) and LiRA pose significant risks to privacy, where TLA utilizes knowledge from pre-trained models, and LiRA exploits model confidence scores to infer membership in the training set. These attacks illustrate the need for more robust and advanced privacy mechanisms in machine learning systems.

The inclusion of Rényi Differential Privacy (RDP) in the theoretical analysis offers a more flexible approach to privacy accounting compared to traditional DP methods. By allowing for different privacy parameters (i.e. varying the order of the Rényi divergence), RDP may provide more precise privacy guarantees and further mitigate privacy risks in machine learning models.

In conclusion, while this study presents valuable insights into the theoretical underpinnings of privacy-preserving techniques, empirical validation and further exploration of alternative differential privacy methods are crucial for advancing the field. Future work should focus on overcoming implementation hurdles and empirically validating these methods in diverse environments, particularly on large-scale datasets and more complex models.

REFERENCES

[1] Martin 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*, CCS'16. ACM, October 2016.

[2] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles, 2022.

[3] Cynthia Dwork. Differential privacy. In *Proceedings of the 33rd International Conference on Automata, Languages and Programming - Volume Part II*, ICALP'06, page 1–12, Berlin, Heidelberg, 2006. Springer-Verlag.

[4] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images, 2009. https://www.cs.toronto.edu/~kriz/cifar.html.

[5] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images, 2009. https://www.cs.toronto.edu/~kriz/cifar.html.

[6] Adrian Shuai Li, Arun Iyengar, Ashish Kundu, and Elisa Bertino. Transfer learning for security: Challenges and future directions, 2024.

[7] Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, page 263–275. IEEE, August 2017.

[8] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models, 2017.

[9] Thomas Steinke, Milad Nasr, and Matthew Jagielski. Privacy auditing with one (1) training run. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 49268–49280. Curran Associates, Inc., 2023.

[10] Jiachen T. Wang, Saeed Mahloujifar, Shouda Wang, Ruoxi Jia, and Prateek Mittal. Renyi differential privacy of propose-test-release and applications to private and robust machine learning, 2022.

[9] [3] [2] [6] [7] [10] [1] [8] [4] [5]