Analysis on "Privacy Auditing with One (1) Training Run"

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

ABSTRACT

This paper presents a novel approach for auditing differentially private (DP) machine learning systems efficiently. By using the connection between DP and statistical generalization, the authors propose a single-training-run method that avoids the computational overhead of traditional group privacy techniques.

The proposed method involves including or excluding "canaries" into the training data and analyzing the algorithm's output to estimate privacy parameters. The authors demonstrate the effectiveness of their approach on DP-SGD training, achieving meaningful lower bounds on privacy while maintaining minimal impact on model accuracy.

ACM Reference Format:
Michael Jung. 2024. Analysis on "Privacy Auditing with One (1) Training Run". In . ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Differential Privacy (DP) has emerged as a key framework for protecting personal privacy in the era of data-driven machine learning. By ensuring that the output of an algorithm is insensitive to changes in a single individual's data, DP provides a rigorous mathematical guarantee against privacy breaches. This is particularly crucial in sensitive domains like healthcare, finance, and social sciences, where the misuse of personal data can lead to severe consequences.

While DP offers a theoretical guarantee, it is nonetheless difficult to confirm that a particular DP algorithm implementation actually protects privacy. This process, known as privacy auditing, is essential to ensure that the algorithm sticks to its claimed privacy guarantees and is not vulnerable to privacy attacks.

Traditional privacy auditing methods often involve multiple training runs of the algorithm with different subsets of data. This can be computationally expensive and time-consuming, restricting the practical application of privacy auditing in real-world scenarios. To address these limitations, the authors have focused on developing efficient auditing techniques that require only a single training run.

One(1) training run privacy auditing presents a viable way to overcome the computational challenges associated with traditional methods. By reducing the number of required training runs, single-run auditing techniques make privacy auditing more practical and scalable. This method makes use of cutting-edge approaches to estimate privacy parameters to minimize computing costs and estimate privacy parameters.

This paper explores the improvements in single-run privacy auditing techniques. We explore the theoretical foundations, practical implementations, and limitations of these approaches, providing a comprehensive overview. Our goal is to provide insight into the potential benefits and challenges of single-run auditing and to contribute to the development of more efficient and effective tools for ensuring privacy in machine learning.

2 BACKGROUND

Privacy auditing is a rapidly evolving field driven by the increasing importance of data privacy in the era of machine learning. As organizations deploy data-driven models in sensitive domains, the need for robust methods to verify the privacy guarantees of these models has become crucial.

The goal of privacy auditing is to estimate the actual privacy parameters (ϵ and δ) of a DP algorithm based on its behavior. This involves analyzing the algorithm's output and comparing it to the expected behavior under different privacy levels.

However, as mentioned in the paper, estimating these probabilities requires running the algorithm hundreds of times. This approach to privacy auditing is computationally expensive, which raises the question "Can we perform privacy auditing using a single run of the algorithm M?" which can be considered the main motivation of this survey.

3 PROBLEM FORMULATION

Key Concepts:

Privacy Parameter: A parameter that quantifies the level of privacy protection. A smaller ϵ value indicates stronger privacy.

Privacy Loss: The maximum amount of information an adversary can learn about an individual's data from the algorithm's output.

Noise Addition: DP algorithms typically add noise to their outputs to mask individual contributions and protect privacy.

Key Challenges:

Black-Box vs. White-Box Access: Researchers may have varying levels of access to the algorithm's internals. White-box access allows for direct inspection of the algorithm's code and intermediate computations, while black-box access is limited to observing the input and output.

Computational Efficiency: The auditing process should be computationally efficient to be practical for real-world applications.

Accuracy: The estimated privacy parameters should be accurate and reflect the true privacy protection offered by the algorithm.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

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

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

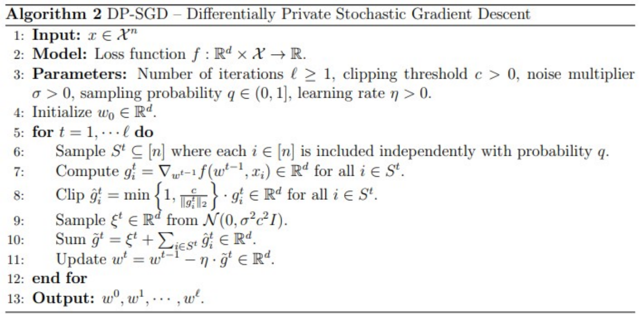
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

4 METHODS

White-Box Auditing: It analyzes the internal structures the used data structures, internal design, code structure, and the working of the software rather than just the functionality as in black box testing.[2]

Black-Box Auditing: A type of software testing in which the tester is not concerned with the software's internal knowledge or implementation details but rather focuses on validating the functionality based on the provided specifications or requirements. The tester only focuses on the input and output of the software.[2]

DP-SGD(Differentially Private Stochastic Gradient Descent): The main algorithm whose privacy the authors are most interested in auditing.



```
Algorithm 2 DP-SGD – Differentially Private Stochastic Gradient Descent
Input: $x \in \mathcal{X}^n$
Model: Loss function $f : \mathbb{R}^d \times \mathcal{X} \to \mathbb{R}$.
Parameters: Number of iterations $\ell \geq 1$, clipping threshold $c > 0$, noise multiplier
  $\sigma > 0$, sampling probability $q \in (0, 1]$, learning rate $\eta > 0$.
Initialize $w_0 \in \mathbb{R}^d$.
for $t = 1, \cdots \ell$ do
    Sample $S^t \subseteq [n]$ where each $i \in [n]$ is included independently with probability $q$.
    Compute $g_i^t = \nabla_{w^{t-1}} f(w^{t-1}, x_i) \in \mathbb{R}^d$ for all $i \in S^t$.
    Clip $\hat{g}_i^t = \min\left\{1, \frac{c}{\|g_i^t\|_2}\right\} \cdot g_i^t \in \mathbb{R}^d$ for all $i \in S^t$.
    Sample $\xi^t \in \mathbb{R}^d$ from $\mathcal{N}(0, \sigma^2 c^2 I)$.
    Sum $\tilde{g}^t = \xi^t + \sum_{i \in S^t} \hat{g}_i^t \in \mathbb{R}^d$.
    Update $w^t = w^{t-1} - \eta \cdot \tilde{g}^t \in \mathbb{R}^d$.
end for
Output: $w^0, w^1, \cdots, w^\ell$.
```

Figure 1: Algorithm of DP-SGD given in the paper

Detailed Discussion of Each Category

White-Box Auditing

Advantages: Can provide more accurate estimates of privacy parameters due to deeper insights into the algorithm's internal designs.

Disadvantages: Requires access to the algorithm's code or implementation details, which may not always be available.

Black-Box Auditing

Advantages: More practical in real-world scenarios where the algorithm's internals are not accessible.

Disadvantages: May be less accurate than white-box auditing due to limited visibility into the algorithm's behavior.

5 OUTCOMES

The development in single-run privacy auditing has demonstrated the feasibility of effectively assessing the privacy guarantees of DP algorithms without the need for multiple training runs. These techniques offer significant computational benefits and can be applied in a variety of real-world scenarios.

Possible groups earning the benefit of Single-Run Auditing may include:

Machine Learning Researchers: The paper provides valuable insights into privacy auditing techniques, enabling researchers to develop more secure and privacy-preserving machine learning models. Researchers can use these techniques to assess the privacy guarantees of their deployed models.

Organizations Handling Sensitive Data: Companies, institutions, and government agencies that handle sensitive data can use this paper to evaluate the privacy of their machine learning systems and identify potential vulnerabilities.

6 LIMITATIONS AND FUTURE DIRECTIONS

Despite the significant progress made in single-run privacy auditing, limitations and challenges remain:

Inherent problems: The authors state that the problem is due to a mismatch between realistic DP algorithms and pathological DP algorithms. This mismatch makes the lower bound much more sensitive to δ than predicted.

Future research directions mentioned in the paper include:

1. Improved Attacks:

Strengthening Existing Attacks: The authors acknowledge that their experimental evaluation uses existing attack methods. Enhancing these attacks with newly created attack methods could lead to more accurate and robust privacy auditing.

Addressing Example Hardness: The current attacks may not sufficiently account for the varying "hardness" of individual examples. Doing so could lead to improving the privacy audits.

2. Algorithm-Specific Analysis:

Leveraging Algorithm Structure: The paper's method approach is generic which can be applied to various DP algorithms. Exploring algorithm-specific analyses could potentially yield stronger results by exploiting the unique characteristics of certain algorithms, such as the iterative nature of DP-SGD.

3. Combining Single-Run and Multiple-Run Approaches:

Optimal Trade-offs: Investigating how to combine the benefits of single-run and multiple-run auditing to achieve tighter bounds with fewer runs.

4. Alternative Privacy Metrics:

Beyond Differential Privacy: Exploring other privacy definitions like Rényi DP to obtain more nuanced insights into the privacy guarantees of DP algorithms.

5. Beyond Lower Bounds:

Estimating True Privacy Loss: Developing techniques to estimate the "true" privacy loss, rather than just providing theoretical lower or upper bounds.

7 CONCLUSION

In conclusion, this analysis of "Privacy Auditing with One (1) Training Run" has demonstrated the significant potential of single-run privacy auditing techniques. By addressing the computational challenges associated with traditional auditing methods, this technique offers a promising approach for verifying the privacy guarantees of DP algorithms in real-world applications. As research in this area continues to evolve, we can expect further improvements in the accuracy, efficiency, and applicability of single-run auditing techniques.

REFERENCES

[1] Rahul Dev. Understanding data privacy and data privacy audit procedure, 2020.

[2] mahak$_j ain.Differencesbetweenblackboxtestingandwhiteboxtesting$, 2024.

[3] An Nguyen. Understanding differential privacy, 2019.

[4] Margaret Rouse and Natalie Medleva. Differential privacy, 2024.

[5] 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.

[2] [1] [3] [4] [5]