Internship: Auditing Practical Privacy Guarantees of Differentially Private Machine Learning Models

November 6, 2024

Team and Contact

- Inria team PreMeDICal, based in Montpellier
- Supervisors: Aurélien Bellet (aurelien.bellet@inria.fr), Christian Janos Lebeda (christian-janos.lebeda@inria.fr), Tudor Cebere (tudor.cebere@inria.fr)

Keywords

Differential Privacy, Machine Learning, Privacy Auditing, Subsampling Schemes

Context

Differential Privacy (DP) [Dwork et al., 2006] is a rigorous mathematical framework that provides privacy guarantees for individuals in a dataset. It ensures that the outcome of any analysis is not significantly affected by any individual's data, thereby protecting personal information from being inferred. In a machine learning (ML) context, models trained on personal data can inadvertently reveal private details or allow adversaries to reconstruct original data [Nasr et al., 2019]. DP addresses this by enabling the training of models on sensitive datasets while preserving user privacy, with DP-SGD [Abadi et al., 2016] serving as the standard algorithm for implementing differential privacy in model training. DP-SGD follows the same steps as standard SGD but clips the gradients' norm to a threshold before perturbing them with carefully calibrated Gaussian noise, providing differential privacy guarantees for each gradient step.

However, practical implementations of DP mechanisms face some challenges:

- Implementation artefacts, such as numerical errors due to floating-point arithmetic [Mironov, 2012] or theoretical mistakes in privacy accounting [Tramer et al., 2022], can lead to underestimating the actual privacy leakage. This means the system may be less private than theoretically guaranteed.
- Conservative privacy analyses might provide pessimistic upper bounds on the privacy loss, overestimating privacy leakage. This can result in an unnecessary loss in the utility of the model due to excessive noise being added.

To address these issues, privacy auditing [Jagielski et al., 2020, Nasr et al., 2023] has emerged as a method to empirically estimate lower bounds on the privacy loss. Privacy auditing involves

simulating attacks—such as membership inference attacks—that attempt to determine whether a specific individual's data was included in the training dataset. By assessing the success rate of these attacks, we can measure the practical privacy leakage of the model.

Objectives

The goal of this internship is to investigate the practical guarantees of differentially private machine learning models. The internship will focus on (a subset of) the following questions:

1. **Auditing differentially private machine learning algorithms**: Recent work [Cebere et al., 2024] indicates that, under certain technical assumptions, the theoretical privacy analysis of DP-SGD is tight—that is, the theoretical bounds closely match the actual privacy loss. This project aims to i) **Explore additional assumptions**: Investigate whether introducing new or stronger assumptions can improve the upper bound provided by privacy accounting, potentially leading to better privacy-utility trade-offs for DP-SGD. ii) **Relax existing assumptions**: Examine if the assumptions required for tight analysis can be weakened without losing tightness, resulting in more realistic adversaries.

2. **Analyzing the impact of subsampling schemes on privacy guarantees**: Privacy amplification by subsampling [Kasiviswanathan et al., 2011, Balle et al., 2018] is instrumental in the privacy analysis of DP-SGD, where only a random subset of the data (a mini-batch) is used in each iteration. Recent studies [Chua et al., 2024, Lebeda et al., 2024] have shown that the choice of subsampling method (e.g., sampling with or without replacement) can significantly affect the theoretical privacy guarantees. On the other hand, sampling schemes considered in formal privacy guarantees do not always align with the ones used in practical implementations. This misalignment can lead to incorrect or loose privacy guarantees. The objectives are to: i) **Audit current approaches**: Examine existing DP-SGD implementations to identify discrepancies between assumed and actual subsampling methods and ii) **Detect theoretical gaps in practice:** Use privacy auditing to empirically detect any gaps between theoretical privacy guarantees and actual privacy leakage due to subsampling schemes.

Skills Required

- Background in Probability and Statistics, Machine Learning, and/or Algorithms and Computer Science.
- Proficiency in Python programming.
- Familiarity with Differential Privacy is a plus.
- Familiarity with Floating Point Arithmetics is a plus.

References

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)*, 2016.

Borja Balle, Gilles Barthe, and Marco Gaboardi. Privacy amplification by subsampling: Tight analyses via couplings and divergences. In *NeurIPS*, 2018.

Tudor Cebere, Aurélien Bellet, and Nicolas Papernot. Tighter privacy auditing of dp-sgd in the hidden state threat model, 2024. URL https://arxiv.org/abs/2405.14457.

Lynn Chua, Badih Ghazi, Pritish Kamath, Ravi Kumar, Pasin Manurangsi, Amer Sinha, and Chiyuan Zhang. How private are DP-SGD implementations? In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 8904–8918. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/chua24a.html.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In Shai Halevi and Tal Rabin, editors, *Theory of Cryptography*, pages 265–284, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-32732-5.

Matthew Jagielski, Jonathan Ullman, and Alina Oprea. Auditing differentially private machine learning: How private is private sgd? *Advances in Neural Information Processing Systems*, 33: 22205–22216, 2020.

Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

Christian Janos Lebeda, Matthew Regehr, Gautam Kamath, and Thomas Steinke. Avoiding pitfalls for privacy accounting of subsampled mechanisms under composition. *CoRR*, abs/2405.20769, 2024. doi: 10.48550/ARXIV.2405.20769. URL https://doi.org/10.48550/arXiv.2405.20769.

Ilya Mironov. On significance of the least significant bits for differential privacy. In *Proceedings of the 2012 ACM Conference on Computer and Communications Security*, CCS '12, 2012.

Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE Symposium on Security and Privacy, SP 2019, San Francisco, CA, USA, May 19-23, 2019*, pages 739–753. IEEE, 2019. doi: 10.1109/SP.2019.00065. URL https://doi.org/10.1109/SP.2019.00065.

Milad Nasr, Jamie Hayes, Thomas Steinke, Borja Balle, Florian Tramèr, Matthew Jagielski, Nicholas Carlini, and Andreas Terzis. Tight auditing of differentially private machine learning. In *USENIX Security*, 2023.

Florian Tramer, Andreas Terzis, Thomas Steinke, Shuang Song, Matthew Jagielski, and Nicholas Carlini. Debugging differential privacy: A case study for privacy auditing, 2022. URL https://arxiv.org/abs/2202.12219.