Privacy, Utility and Fairness: Navigating Trade-offs in Differentially Private Machine Learning

Lea Demelius

Graz University of Technology
Know Center Research GmbH
ldemelius@know-center.at

Abstract

Developing *trustworthy AI* requires advancing methods that meet key requirements such as privacy or fairness while maintaining strong utility, as well as understanding the intricate interdependencies between these dimensions, which often manifest as trade-offs. My PhD research focuses on differential privacy, which is widely regarded as the state-of-the-art for protecting privacy in data analysis and machine learning. I investigate the relationships between differential privacy, utility and fairness, with the goal of advancing the adoption of differentially private machine learning in real-world settings.

Introduction

Nowadays, AI and machine learning significantly impact our daily lives, from chatbots like ChatGPT and recommender systems on platforms such as Netflix or Spotify, to decision support tools in healthcare, human resources, and finance. As the prevalence of algorithms increases, particularly in high-stakes settings, so does the demand for them to be *trustworthy*, including requirements such as transparency, robustness, fairness and privacy preservation. The need for privacy protection is particularly pressing as AI systems necessitate the collection and analysis of ever-larger amounts of (personal) data.

One concern are potential information leakages through the AI model. It has been demonstrated that AI models are vulnerable to *reconstruction attacks*, where an adversary can reconstruct (portions of) the training data from the model's outputs, and *membership inference attacks*, where attackers can determine whether a specific individual's data was included in the model's training set. It has also been shown that traditional anonymization techniques, such as k-anonymity (Sweeney 2002), cannot reliably guard against these vulnerabilities, which has led to the growing interest in differential privacy (DP) (Dwork and Roth 2014).

DP is a mathematical framework that enables the protection of individual data points (e.g. information about specific individuals) while still allowing queries about the whole dataset (e.g. population-level insights) by adding curated noise. In contrast to traditional anonymization techniques, DP provides a privacy guarantee by quantifying the maximum information loss. The most widely used method for achieving differential privacy in (non-convex) neural networks is *Differentially Private Stochastic Gradient Descent* (DPSGD) (Abadi et al. 2016), which incorporates gradient clipping and perturbation during the training process. The main challenge of employing DP lies in its impact on the algorithm, which can lead to unintended consequences for its properties. This often manifests as trade-offs, where improving privacy comes at the expense of performance, accuracy, usability or other critical measures like fairness.

The focus of my PhD is to investigate the intricate relationships between privacy, utility and fairness in machine learning. I am particularly interested in how the integration of DP requires adjustments in standard machine learning practices, such as hyperparameter selection, with the goal of enabling a broader adoption of methods like DPSGD.

Contributions

At the beginning of my PhD, I conducted an extensive analysis of the state-of-the-art literature of both DP in deep learning in particular and privacy in AI in general (with a focus on data protection during computations, which not only includes DP but also other privacy-enhancing technologies like homomorphic encryption, multi-party computation, and federated learning). This literature analysis concluded with the publication of a white paper on privacy and security in AI in collaboration with the company SGS (Demelius et al. 2023), a survey on trustworthy AI (Kowald et al. 2024), and a survey on recent advances of differential privacy in deep learning (Demelius, Kern, and Trügler 2023). I also contributed to a book chapter on modern data protection and trustworthy AI in health care (Demelius, Jantscher, and Trügler 2024).

In the course of my literature analysis, I came across an intriguing line of research that showed that DPSGD has a disparate impact on accuracy (Bagdasaryan, Poursaeed, and Shmatikov 2019; Farrand et al. 2020; Tran, Dinh, and Fioretto 2021; Xu, Du, and Wu 2021). They demonstrated that integrating DP not only leads to a decrease in overall accuracy - a common challenge associated with DP - but also disproportionately impacts certain sub-groups, raising a significant fairness concern. Interestingly, a follow-up study

(de Oliveira et al. 2023) showed that DPSGD does not necessarily have a negative impact on fairness, as long as the DP model's hyperparameters are optimized for performance. They infer that the disparate impact of DPSGD primarily occurs when hyperparameter settings that perform well for non-private models are re-used for DP models without further tuning. While this finding would greatly benefit practitioners aiming to train private and fair neural networks, a closer look revealed that there are still open questions regarding the general applicability and the impact of different performance and fairness metrics. I, therefore, conducted in-depth experiments answering the following questions:

1. How does DPSGD influence disparities across a variety of metrics, and do they necessarily co-occur?
2. How dependent are these disparities on the choice of hyperparameters, and how effective and reliable is hyperparameter tuning in developing private models with similar (or even better) performance and fairness than non-private models?
3. How does hyperparameter choice affect DPSGD-Global-Adapt, a variant of DPSGD specifically designed to mitigate the disparate impact of DP?

The resulting paper was recently submitted.

Future Work

Next, I plan to address a research gap regarding privacy leakage through hyperparameter tuning in differentially private machine learning that we identified during this work. Hyperparameter tuning can leak information even when the single training runs are differentially private (Papernot and Steinke 2021). While one could simply use composition to compute the actual privacy budget (by adding up the privacy budgets from all training runs), this would result in an unacceptable loose privacy guarantee. While alternative methods have been proposed to obtain improved privacy guarantees, for example, private random search (Papernot and Steinke 2021) or private grid search (Ding and Wu 2022), there are still open questions - particularly in scenarios involving machine learning practices such as cross-validation and early stopping. I am also interested in looking into practical attacks that exploit the additional information leakage through non-private hyperparameter tuning.

With the completion of my PhD, I will contribute both theoretical and practical insights into the complex relationships between privacy, utility and fairness to facilitate the adoption of differentially private machine learning in real-world settings. Participation in the AAAI-25 doctoral consortium will provide invaluable support for achieving this goal and will facilitate my future endeavors in both academia and industry-focused research.

Acknowledgments

This research is supported by the project "PRO'k'RESS" managed by the Austrian Research Promotion Agency (FFG) and the project "Z-T-G 004 (AI Privacy)" funded by the Zukunftsfonds Steiermark.

References

Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H. B.; Mironov, I.; Talwar, K.; and Zhang, L. 2016. Deep Learning with Differential Privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 308–318. Vienna Austria: ACM. ISBN 978-1-4503-4139-4.

Bagdasaryan, E.; Poursaeed, O.; and Shmatikov, V. 2019. Differential privacy has disparate impact on model accuracy. *Advances in neural information processing systems*, 32.

de Oliveira, A. S.; Kaplan, C.; Mallat, K.; and Chakraborty, T. 2023. An Empirical Analysis of Fairness Notions under Differential Privacy. In *The Fourth AAAI Workshop on Privacy-Preserving Artificial Intelligence*.

Demelius, L.; Jantscher, M.; and Trügler, A. 2024. Modern data protection and trustworthy AI [Moderner Datenschutz und vertrauenswürdige KI]. In Klein, A.; Dennerlein, S.; and Ritschl, H., eds., *Health Care und Künstliche Intelligenz. Ethische Aspekte verstehen – Entwicklungen gestalten*, 217–234. Narr Francke Attempto Verlag.

Demelius, L.; Kern, R.; and Trügler, A. 2023. Recent Advances of Differential Privacy in Centralized Deep Learning: A Systematic Survey. *arXiv preprint arXiv:2309.16398*.

Demelius, L.; Trügler, A.; Kopeinik, S.; Scher, S.; Nad, T.; and Kowald, D. 2023. Privacy and Security in AI. White paper, SGS.

Ding, Y.; and Wu, X. 2022. Revisiting Hyperparameter Tuning with Differential Privacy. In *NeurIPS ML Safety Workshop*.

Dwork, C.; and Roth, A. 2014. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4): 211–407.

Farrand, T.; Mireshghallah, F.; Singh, S.; and Trask, A. 2020. Neither private nor fair: Impact of data imbalance on utility and fairness in differential privacy. In *Proceedings of the 2020 workshop on privacy-preserving machine learning in practice*, 15–19.

Kowald, D.; Scher, S.; Pammer-Schindler, V.; Müllner, P.; Waxnegger, K.; Demelius, L.; Fessl, A.; Toller, M.; Mendoza Estrada, I. G.; Simic, I.; et al. 2024. Establishing and Evaluating Trustworthy AI: Overview and Research Challenges. *Frontiers in Big Data*, 7: 1467222.

Papernot, N.; and Steinke, T. 2021. Hyperparameter Tuning with Renyi Differential Privacy. In *International Conference on Learning Representations*.

Sweeney, L. 2002. k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems*, 10(05): 557–570.

Tran, C.; Dinh, M.; and Fioretto, F. 2021. Differentially private empirical risk minimization under the fairness lens. *Advances in Neural Information Processing Systems*, 34: 27555–27565.

Xu, D.; Du, W.; and Wu, X. 2021. Removing disparate impact on model accuracy in differentially private stochastic gradient descent. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 1924–1932.