A PRACTICAL GUIDE TO DIFFERENTIALLY PRIVATE DEEP LEARNING USING THE PSEUDO POSTERIOR MECHANISM

Alexander J. Preiss[1], *Amanda Konet*[1], *Robert Chew*[1], *Matthew R. Williams*[1], *Elan A. Segarra*[2], *David H. Oh*[2], *Erin Boon*[2], *Terrance D. Savitsky*[2]

[1]RTI International
[2]U.S. Bureau of Labor Statistics

Balancing predictive performance with rigorous privacy protection is a critical challenge when training neural networks on sensitive individual-level data. Yet, practitioners often have little guidance for navigating the competing statistical, computational, and privacy demands of privacy-preserving machine learning. We present a procedural framework for the Stochastic Weight Averaging–Gaussian Pseudo Posterior Mechanism (SWAG-PPM), a differentially private deep learning method that combines Gaussian posterior approximation via parameter snapshotting with risk-based downweighting of high-disclosure-risk records. Our framework focuses on the joint tuning of two key hyperparameters (the disclosure-risk slope parameter (c) and the number of fine-tuning epochs) whose interaction governs model convergence behavior and the privacy–utility trade-off. We introduce diagnostic tools, including a novel max-delta plot, to evaluate convergence and guide hyperparameter adjustments. Using a transformer model for occupational injury classification, we demonstrate that diagnostic-guided tuning within SWAG-PPM can achieve strong privacy protection with minimal loss in accuracy. While our case study uses a specific dataset and model architecture, all methodological steps can apply to other settings where privacy risk is heterogeneously distributed.

Organizations handling sensitive individual-level data, such as government statistical agencies and technology firms, face increasing demands to produce public data products while rigorously protecting respondent privacy. Model-based approaches that privatize statistical models (rather than raw data) have gained prominence for their ability to generate synthetic datasets under strict privacy controls [1]. One such method is the Pseudo Posterior Mechanism (PPM) [7], which adjusts the likelihood contribution of each record in proportion to its disclosure risk. Parallel to these developments, deep learning models, particularly transformer-based neural networks [8], have become state-of-the-art tools for regression and classification tasks [4].

Recent work has extended the use of PPM to modern neural network architectures by using Stochastic Weight Averaging–Gaussian (SWAG) [6] to construct a Gaussian approximation of the posterior distribution [3]. This approach, Stochastic Weight Averaging–Gaussian Pseudo Posterior Mechanism (SWAG-PPM), demonstrated strong privacy guarantees with minimal utility loss, even under relatively strict privacy settings. However, implementing a Bayesian privacy mechanism like the PPM within an SGD-trained neural network introduces important challenges. Achieving high-quality parameter estimation requires tuning the number of fine-tuning epochs so that the model parameters evolve near the global mode. Simultaneously, the PPM risk-based weights must be calibrated to meet a target privacy guarantee, which in turn alters the shape of the model distribution. This introduces a non-trivial interaction: tuning the PPM affects the model's parameter trajectory, and vice versa.

In this paper, we develop a detailed procedural framework to coordinate these dual tuning processes, enabling robust application of the SWAG-PPM methodology to neural network models (Figure 1). Our approach ensures that data curators can achieve their desired privacy guarantees while maintaining the predictive performance necessary for practical deployment. To demonstrate this process in action, we use the publicly available Severe Injury Reports dataset [2] collected by the U.S. Occupational Safety and Health Administration (OSHA) from January 2015 through September 2023 under regulation 29 CFR 1904.39. Each record includes a "Final Narrative" free-text description of the incident and is coded using version 2.01 of the Occupational Injury and Illness Classification System (OIICS) into fields such as "Nature of Injury," which serves as the outcome variable for modeling. In its raw form, the dataset contains 86,210 observations spanning 199 distinct OIICS codes, exhibiting a pronounced class imbalance typical of occupational injury data. Fine-tuning was performed starting from the weights of a pre-trained DistilRoBERTa transformer [5].

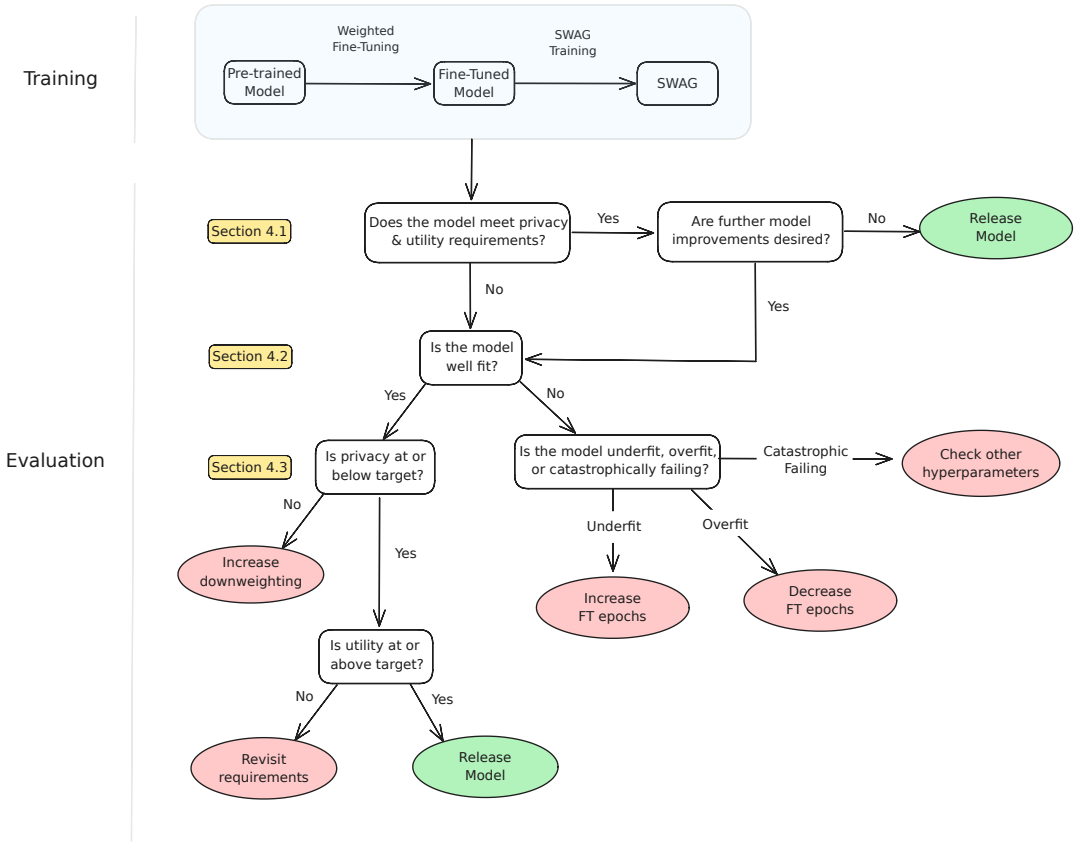

Figure 1. SWAG-PPM Evaluation Flowchart

References

[1] Martin Abadi et al. "Deep learning with differential privacy". In: *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. 2016, pp. 308–318.

[2] Rob Chew. *OSHA Severe Injury Reports: Jan 2015 - Sep 2023*. DOI: 10.6084/m9.figshare.28669604.v1. Mar. 2025. DOI: 10.6084/m9.figshare.28669604.v1. URL: https://figshare.com/articles/dataset/OSHA_Severe_Injury_Reports_Jan_2015_-_Sep_2023/28669604.

[3] Robert Chew et al. "Bayesian Pseudo Posterior Mechanism for Differentially Private Machine Learning". In: *arXiv preprint arXiv:2503.21528* (2025).

[4] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of NAACL-HLT* (2019).

[5] Zhuang Liu et al. "A robustly optimized BERT pre-training approach with post-training". In: *China national conference on Chinese computational linguistics*. Springer. 2021, pp. 471–484.

[6] Wesley J Maddox et al. "A simple baseline for bayesian uncertainty in deep learning". In: *Advances in neural information processing systems* 32 (2019).

[7] Terrance D Savitsky, Matthew R Williams, and Jingchen Hu. "Bayesian pseudo posterior mechanism under asymptotic differential privacy". In: *Journal of Machine Learning Research* 23.55 (2022), pp. 1–37.

[8] Ashish Vaswani et al. "Attention Is All You Need". In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017.