A Practical Guide to DP Deep Learning

…Using the Pseudo Posterior Mechanism

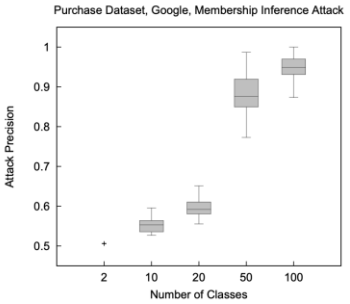
Rob Chew[1], Sandy Preiss[1], Amanda Konet[1], Matt Williams[1], Elan Segarra[2], David Oh[2], Erin Boon[2], Terrance Savitsky[2]

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

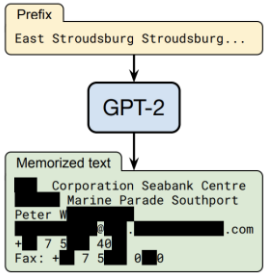
RTI INTERNATIONAL

Motivation

- Statistical agencies are increasingly interested in releasing internally developed autocoding models.
- However, releasing deep learning (DL) models can pose serious privacy risks.
- Differential privacy (DP) can help, but existing DP DL methods are difficult to apply in practice.
- Guidance for newer DP DL paradigms is largely absent.
- This talk presents practical guidance for diagnosing SWAG-PPM, a DP DL method developed with autocoding as a primary use case.



Shokri et al., 2017



Carlini et al., 2021

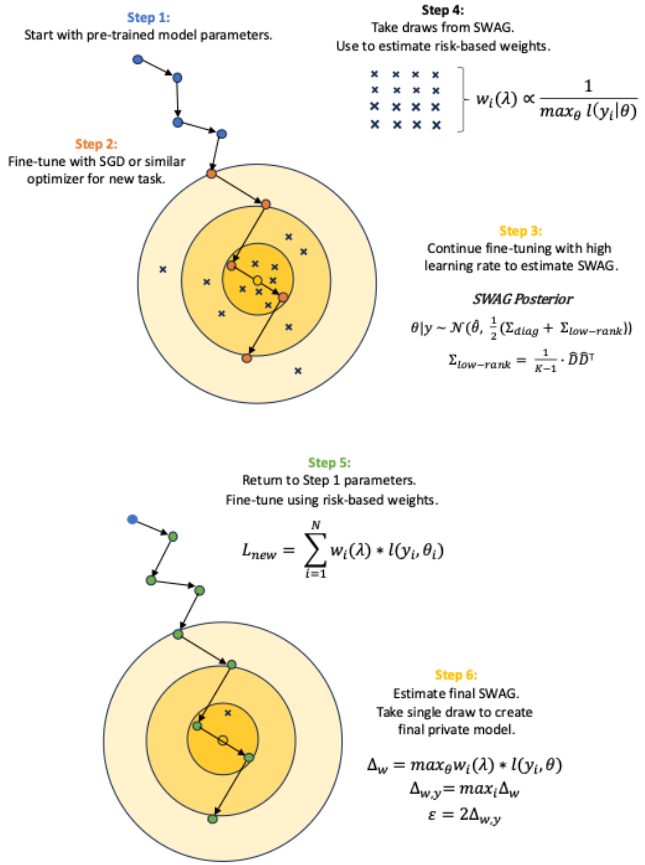

Contents

Ponomareva et al., 2023

SWAG-PPM (Chew et al., 2025)

- **PPM** (Pseudo Posterior Mechanism)
  - Provides DP by randomly drawing from the posterior of model parameters
  - Downweighs riskier records during training, limiting how much any single record can influence the model
- **SWAG** (Stochastic Weight Averaging – Gaussian)
  - Fits a multivariate Gaussian from additional training updates around a local minima
  - This Gaussian acts as an approximate (pseudo) posterior that can then be sampled from for a model release

Article Link →



Key Hyperparameters

- **Fine-tuning Epochs** (“*FT Epochs*”)
  - # of full passes over the training data
    - **Too few:** leave parameters off-mode, biasing SWAG
    - **Too many:** overfitting, model memorize training data
- **Disclosure Risk Weight Parameter** (*“c”*)
  - How quickly downweighing increases with increased disclosure risk
    - **High c:** high risk records downweighed more relative to low risk
    - **Low c:** flattens decline, lowering max contribution of all records
- **Hyperparameter Interactions**
  - *FT epochs* & *c* are not independent
  - Changing c alters both curvature and location of posterior, requiring re-running fine tuning

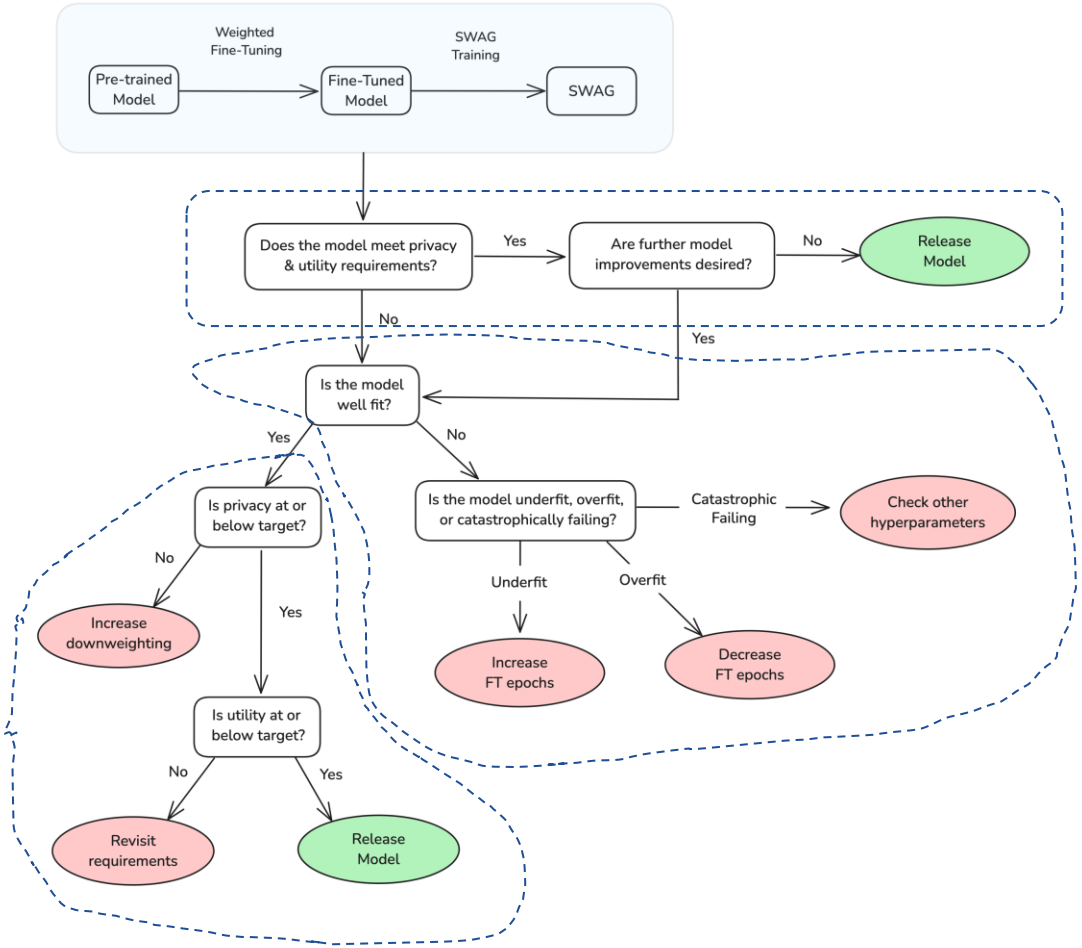
Training Flowchart

Three steps:

1. Initial privacy & utility assessment
2. Model convergence (*“FT epochs”*)
3. Privacy tuning (*“c”*)

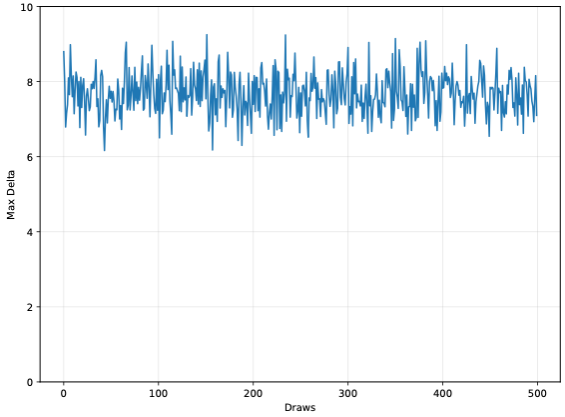
To illustrate these steps, we present a case study that trains an autocoder using the OSHA severe injury reports data.

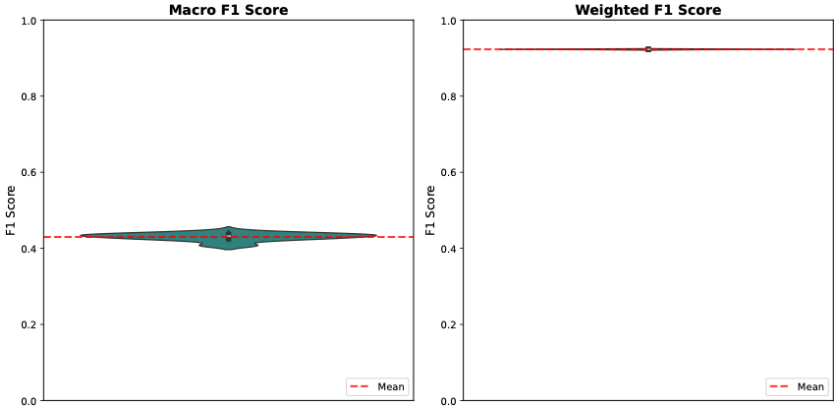
- https://www.osha.gov/severe-injury-reports



Step 1: Initial Assessment

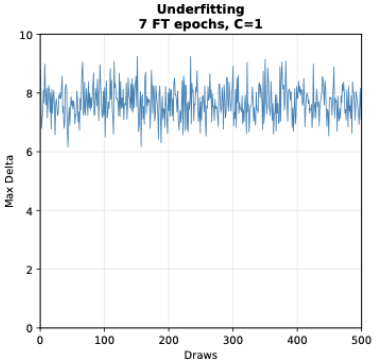
- **Targets**
  - **Privacy:** $\varepsilon \leq 10$
  - **Utility:** $\leq$ 25% degradation in F1 from non-private baseline
- **Privacy**
  - "Max Delta" plot
    - X-axis: Posterior draws
    - Y-axis: Max loss training data
  - $\varepsilon = (2 * Max\ Loss) = 18.5$
- **Utility**
  - "Utility" plot
    - F1 score (macro and weighted) across 30 posterior draws
- **Decision**
  - Does not meet privacy requirements

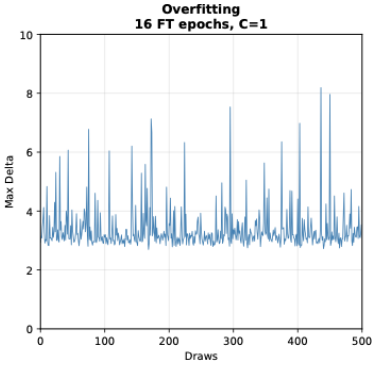


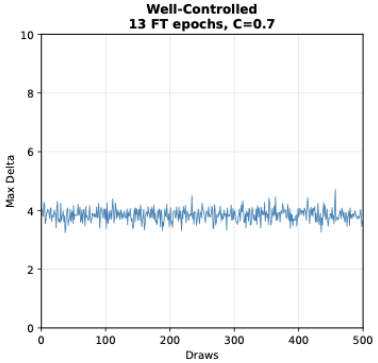


Step 2: Convergence

- Model convergence is important because both privacy and utility depend on the stability of the model's parameter estimates.
- Unlike mature Bayesian procedures, SWAG does not offer diagnostics to assess convergence
- We repurpose the max delta plots to help:
  - [**Top**] "Underfit" models produce high, persistent variability across draws
  - [**Bottom**] "Overfit" models produce low variability punctuated by sudden spikes
  - [**Middle**] "Well-fit" models exhibit low, stable max deltas across draws
- **Decision**
  - Vary the *FT epochs* until model is well-controlled

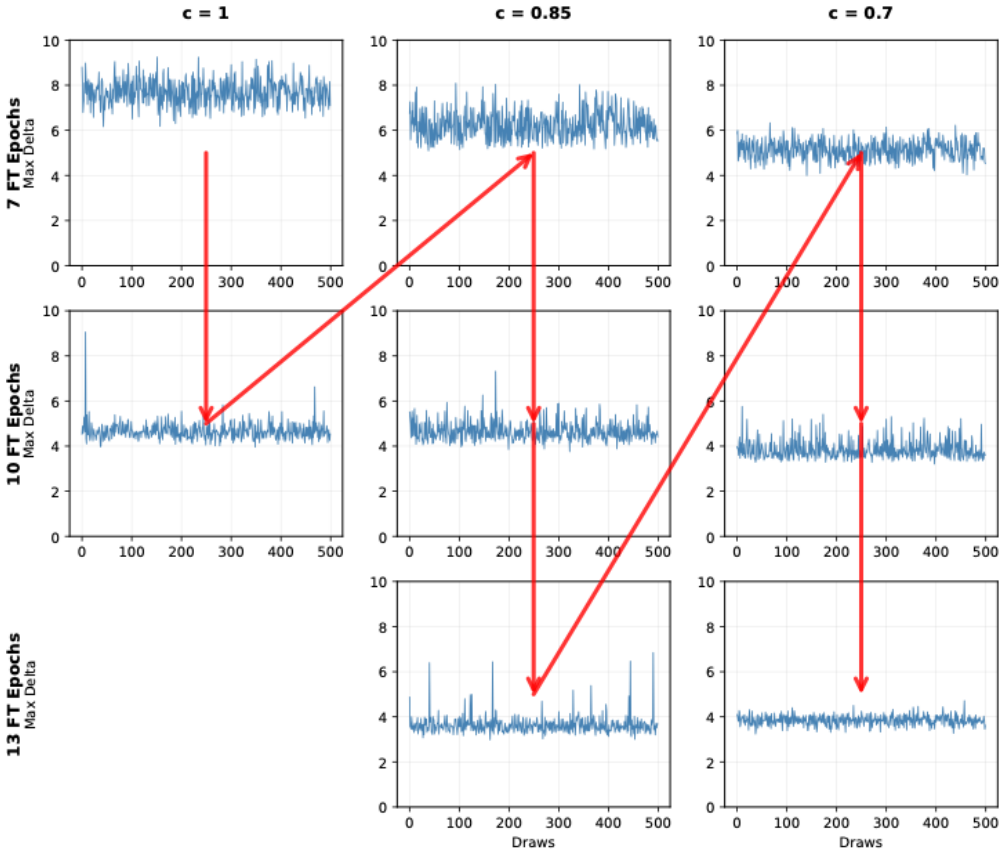






Step 3: Privacy Tuning

- If “converged” model does not meet privacy threshold, we can lower *c* to improve privacy at the expense of utility
- Start with high values of *c* and train until convergence, checking if privacy reqs are met along the way.
- If we see signs of overfitting, we reduce *c* and retrain.
- Since our privacy guarantee only relies on the final model release, intermediate tuning does not penalize privacy accounting
- **Decision**
  - Vary *FT epochs* and *c* until privacy and utility requirements are met.
  - If not possible, revisit requirements with stakeholders.



Discussion

- We presented a procedural framework for SWAG-PPM to help practitioners make decisions about training and privacy-utility trade-offs.
- **Future Work**
  - Adaptive hyperparameter optimization that jointly updates c and FT epochs during training.
  - Systematic evaluations of SWAG-PPM across different data distributions (class imbalance, text lengths, feature modalities, etc.) would help test robustness and potentially reveal domain-specific tuning patterns
  - Extend the theory of PPM to support finite-sample guaruntees to help strengthen confidence in practical deployments

Further Reading

Check upcoming GASP special issue in *Journal of Data Science*

Journal of Data Science 0 (0), 1–19
July 2025

DOI: 0000

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], and Terrance D. Savitsky[2]

[1]*Research Triangle Park, NC 27709, RTI International, USA*
[2]*Hillcrest Heights, MD 20746, U.S. Bureau of Labor Statistics, USA*

Abstract

Privacy-preserving machine learning methods seek to train useful models that do not disclose information about the data on which they were trained. Such methods are vital when organizations train neural networks on sensitive individual-level data and seek to release the models publicly. Their goal poses a trade-off between predictive performance (utility) and privacy protection. That trade-off makes privacy-preserving machine learning methods difficult to apply in practice, usually requiring extensive iteration and hyperparameter tuning. Yet, practitioners often have little guidance for navigating competing statistical, computational, and privacy demands. We present an implementation algorithm for the Stochastic Weight Averaging–Gaussian Pseudo Posterior Mechanism (SWAG-PPM), a Bayesian differentially private deep learning method. The implementation algorithm focuses on the joint tuning of two key hyperparameters whose interaction governs model convergence and the privacy–utility trade-off. We introduce novel diagnostic tools to evaluate convergence and guide hyperparameter adjustments. Using a transformer model for occupational injury classification, we demonstrate that diagnostic-guided tuning with SWAG-PPM can achieve strong privacy protection and utility. 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.

Keywords *Bayesian deep learning; Differential Privacy; Imbalanced learning; Official statistics; Pseudo posterior distribution*

Thank you

Rob Chew | rchew@rti.org