Practical Differentially Private Deep Learning with SWAG-PPM

Alexander J. Preiss[1] (apreiss@rti.org)
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

Motivation

- Organizations seek to **release trained models** for reuse.
- Models can **leak data** via membership inference attacks.
- Common **differentially private training** methods (e.g., DP-SGD) often **degrade utility** severely.
- Disclosure **risk is heterogeneous**, concentrated in tails.

SWAG-PPM Methodology

- Combines the **Pseudo Posterior Mechanism (PPM)** with **Stochastic Weight Averaging-Gaussian (SWAG)**.
- High-risk records are downweighted, not noise-perturbed.
- Low-risk records retain near-full influence, preserving utility.
- SGD near the weighted mode yields an approximate Bayesian posterior of model parameters (SWAG).
- Sampling from this pseudo-posterior provides a formal privacy guarantee (PPM).
- Shifts DP from **global noise** to **local downweighting**.

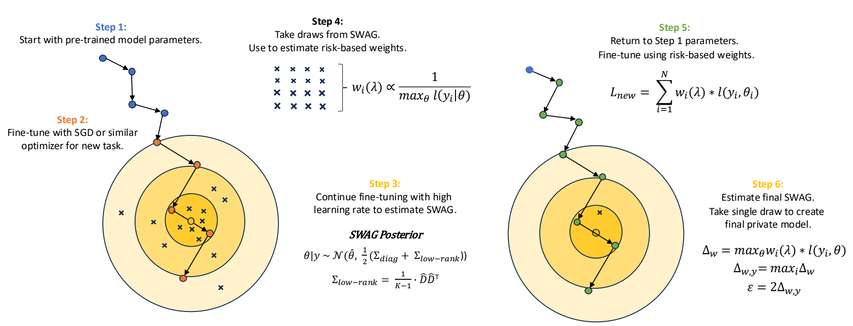

Figure 1: SWAG-PPM training, risk weighting, and posterior sampling workflow

Case Study: OSHA nature-of-injury classifier

- **Data (OSHA Severe Injury Reports):**
  - 86k labeled injury narratives
  - 199 highly imbalanced classes of *Nature of Injury* code
- **Task:** predict *Nature of Injury* code from free text
- **Models:** DistilRoBERTa fine-tuned with:
  - Non-private training
  - SWAG-PPM
  - Differentially private stochastic gradient descent (DP-SGD)

Table 1: Privacy and Utility Comparison by Model, OSHA Case Study

Model	Privacy		Utility	
	Epsilon	Delta	F1 Weighted	F1 Macro
Non-Private	-	-	0.76	0.49
SWAG-PPM	4.35	$O(n^{-1/2})$	0.75	0.44
DP-SGD	4	10^{-4}	0.08	0.03

Table 2: Macro F1 by Class Size Quartile, OSHA Case Study

Model	Top Quartile	Bottom Quartile
Non-Private	0.81	0.18
SWAG-PPM	0.81	0.08

- **Privacy guarantee on par with DP-SGD**
  - Epsilon is calculated from max record-level loss.
  - Delta is approximated; no finite sample estimate.
- **Utility on par with non-private training**
  - Downweighting mostly affects small classes.
- **No free lunch**
  - With default hyperparameters, SWAG-PPM training is roughly 40x slower than non-private.

Real-world Applications

- **Privacy-preserving SOII autocoder**
  - Training DP versions of Survey of Occupational Injuries and Illnesses autocoder model for public release on 3M records.
- **Implementation algorithm**
  - DP learning algorithms are notoriously finicky to tune.
  - Practitioner guidelines forthcoming in Journal of Data Science.

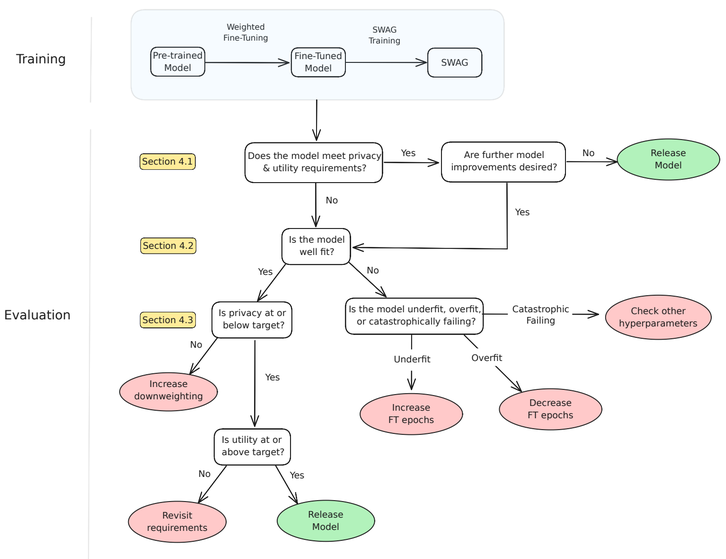

Figure 2: Diagnostic-driven implementation algorithm for SWAG-PPM

Ongoing Work

- Train and potentially release DP SOII autocoder.
- Python package to support DP training across agencies.
- Speedups: early stopping for SWAG training and posterior sampling.