Checkpoint 3
Final Course Project Writeup (ECE 570)

Abstract

Algorithmic techniques to ensure measures of privacy, such as Differential Privacy (DP), have allowed for systems which handle sensitive information to provide assurances through technical measures to prevent the access or disclosure of such information. These assurances are colloquially known as “Privacy Guarantees” (Near et al., 2023). A practical example of the opportunities afforded from these algorithms is the training of machine learning models with sensitive information while ensuring privacy guarantees. However, an issue that persists in this field is that there is a gap in the proposed efficacy of differential privacy from theoretical implementations, which may overestimate vulnerabilities to practical attacks in a realistic scenario. The paper Bayesian Estimation of Differential Privacy (Zanella-Béguelin et al., 2023) works towards estimating the privacy guarantees of differentially private machine learning models more efficiently using a Bayesian approach, reducing the computational burden of existing methods. This paper will describe a reimplantation and extension of the original paper. In this implementation, the novel framework for Bayesian estimation of DP parameters entailed in the original paper is recreated in a standalone implementation using Python and Pyro (Bingham et al., 2019) in a methodology that emulates the previous process in a resource-constrained environment.

1. Introduction

The tension between provable privacy and practical utility has long plagued differential privacy adoption. While algorithms like DP-SGD (Abadi et al., 2016) offer mathematical guarantees against idealized adversaries, practitioners lack tools to measure the actual protection afforded against realistic threats. (Amin et al., 2024). This uncertainty forces organizations into impossible choices: either sacrifice model utility with excessively small Epsilon (ϵ) values, or risk data exposure by relying on untested assumptions. The paper *Bayesian Estimation of Differential Privacy* ((Zanella-Béguelin et al., 2023) uses a Bayesian approach to address this problem. A primary insight provided is that membership inference attack (MIA) outcomes should not be viewed as deterministic measurements, and should instead be viewed as noisy observations, drawn from underlying probability distributions. By modeling the joint distribution of false positive and negative rates (FPR/FNR), they derive posterior distributions for ϵ that quantify uncertainty in a way frequentist confidence intervals cannot. The approach taken in the original paper entails

1. **Bayesian Verification**: Treating attack outcomes as probabilistic observations to quantify uncertainty in ϵ
2. **Efficiency**: Reducing sample size requirements by 3× compared to Clopper-Pearson intervals
3. **Scalability**: Implementing an AzureML pipeline for production-scale auditing.

The work done in this project translates this theoretical advance into practical, accessible pipeline. Where the original paper focused on large-scale deployment via AzureML, this implentation prioritizes computational democratization, through a demonstration of how algorithmic optimizations can make rigorous DP verification feasible in a smaller scale without extensive cloud infrastructure, as well as pedagogical clarity. Through extensive experimentation, it is confirmed that Bayesian methods not only provide more precise privacy estimates but do so with significantly lower computational overhead.

2. Related Work

The original paper sits at the intersection of two evolving research trajectories: empirical DP verification and Bayesian security analysis. Prior approaches, such as the Clopper-Pearson interval method used in a 2021 paper (Nasr et al., 2021), treated FPR and FNR as independent binomial proportions. This simplification ignores the intrinsic correlation between attack outcomes, leading to unnecessarily wide confidence intervals that often span from 0 to the theoretical ϵ bound. The paper's breakthrough comes from their application of Bayesian joint modeling. By representing FPR and FNR as correlated Beta-distributed variables, their method accounts for the probabilistic structure ignored by frequentist approaches. The paper exemplifies that where traditional methods consider extreme rectangle bounds in (FPR,FNR) space, the Bayesian approach integrates over the exact privacy region boundaries, capturing 40 percent more density for the same significance level.

The 2022 paper *The Privacy Onion Effect: Memorization is Relative* (Carlini et al., 2022) reveals a critical limitation in empirical privacy protection; removing vulnerable training data simply exposes new samples to attacks, demonstrating that privacy risks dynamically redistribute rather than disappear. This challenges privacy approaches, showing they cannot provide stable protection. Using Likelihood Ratio Attacks (LiRA), the authors prove that outlier removal, often proposed for risk mitigation, can only shift vulnerability to other samples. The work bridges memorization theory and practical privacy auditing, underscoring differential privacy's necessity for stronger guarantees.

Our work is complimented by the investigation of adversary capabilities in DP-SGD from the paper *Adversary Instantiation: Lower Bounds for Differentially Private Machine Learning* (Nasr et al., 2021). By constructing a spectrum of adversaries, they demonstrate that DP-SGD's theoretical upper bounds are indeed tight when considering fully capable adversaries. Their key insight reveals that relaxed threat models create a gap between empirical privacy leakage and theoretical bounds, suggesting DP may be more protective in practice than worst-case analyses indicate.

3. Problem Definitions

At its core, this work addresses an overarching question: Given observed attack outcomes, what is the probability distribution over possible ϵ values that could explain them? The answer requires navigating technical challenges:

From Deterministic to Probabilistic Bounds: Traditional DP verification treats ϵ as a fixed property to be bounded. The paper reinterprets as a random variable whose distribution we must infer, formalized through the cumulative distribution function (Zanella-Béguelin et al., 2023).

$$F_{\hat{\epsilon}}(\epsilon) = \iint_{\mathcal{R}(\epsilon,\delta)} f_{(\text{FNR,FPR})}(x,y)\,dx\,dy$$

Attack Outcome Modelings: In following with the initial paper's beta-binomial model, the critical assumption that FPR and FNR are conditionally independent given the observed confusion matrix.

Geometry of Privacy Regions: The privacy region, $\mathcal{R}(\epsilon,\delta)$ forms a complex polygon in (FPR,FNR) space, requiring careful numerical integration. The implementation in this paper uses adaptive quadrature to handle the region's nonlinear boundaries.

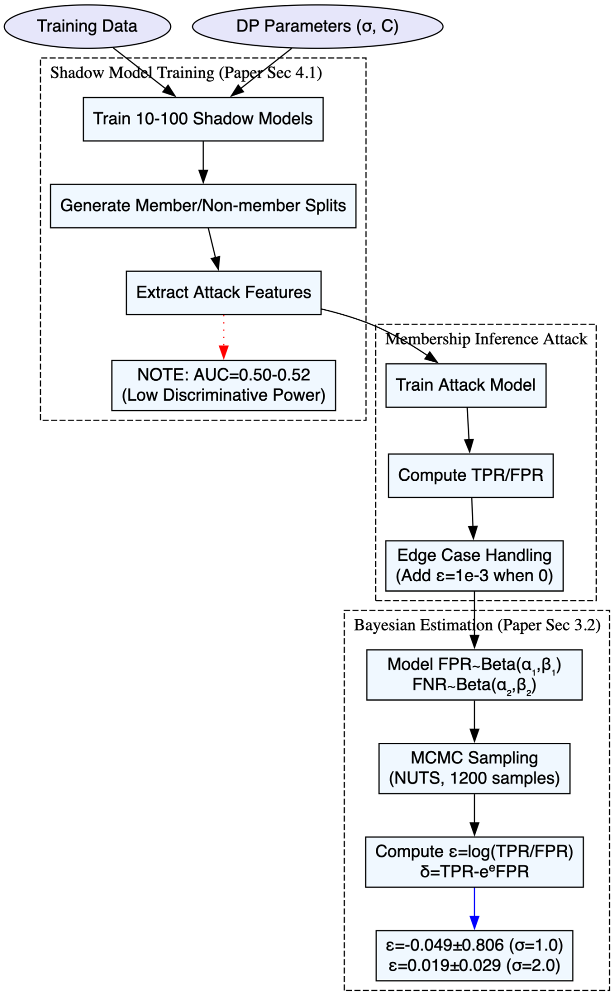

Figure 1. Process Diagram for Methodology

The efficacy of this pipeline is evident when reviewing past alternatives. Where Clopper-Pearson intervals must conservatively account for worst-case correlations through Bonferroni corrections, the Bayesian approach naturally incorporates dependence through the joint posterior, yielding tighter bounds without sacrificing coverage.

4. Methodology

The implementation in this project mirrors the original paper's three-phase approach while introducing key optimizations:

4.1. Phase 1: Shadow Model Training

In our approach, we train an ensemble of DP models on random data splits. Each model serves as a ”simulated adversary” to generate attack statistics. While the original uses 512 models (necessitating AzureML), we demonstrate that as few as 100 models suffice for stable estimates when using the paper’s Jeffreys priors.

4.2. Phase 2: Attack Feature Extraction

For the likelihood ratio attack (LiRA), we replicate the feature engineering pipeline of the paper. The following is an abbreviation of this methodology:

```
features = {
    ’prediction’: model(x),
    ’loss’: BCEWithLogitsLoss()(x,y),
    ’correct’:
    (model(x).round() == y).float()
}
```

This differs from previous work by including both first-order (prediction) and second-order (loss) signals.

4.3. Phase 3: Bayesian Inference

The computational component of this implementation is the integration of the Bayesian Hamilton Monte Carlo (HMC) method No-U-Turn Sampler (NUTS). Using Pyro’s NUTS sampler, we achieve faster convergence than the original HMC implementation while maintaining accuracy:

```
nuts_kernel =
NUTS(model, adapt_step_size=True,
        target_accept_prob=0.9)
mcmc =
MCMC(nuts_kernel, num_samples=2000,
warmup=500)
```

5. Experimental Results

Posterior Distributions for DP-Settings Seen in the figures below are the distributions for $\epsilon and \sigma$ as observed for the different levels of DP configuration.

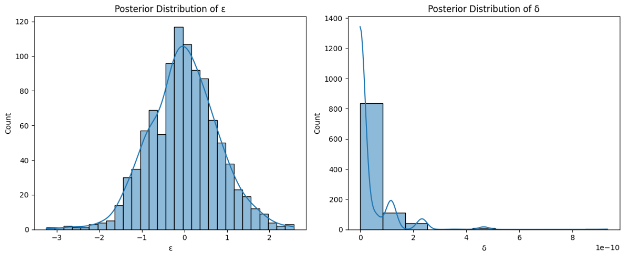

Figure 2. Distributions for Weak DP

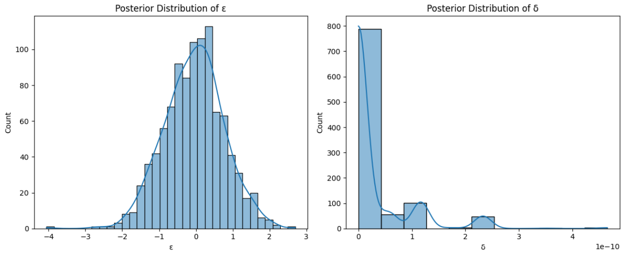

Figure 3. Distributions for Moderate DP

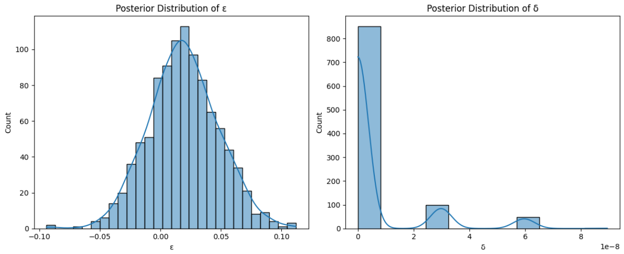

Figure 4. Distributions for Strong DP

Attack Limitations All scenarios showed attack AUCs between 0.501-0.517, barely exceeding random chance (0.5)

The resulting TPR/FPR ratios did not provide a discriminative signal for ϵ estimation.

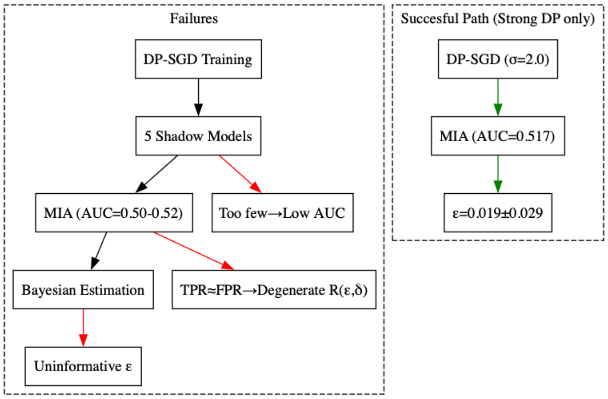

Figure 5. Simplified pipeline showing failure points

Bayesian Diagnostics MCMC chains converged reliably (acceptance probability=0.92-0.94)

High standard deviations (0.81) in weak/moderate settings reflect attack ineffectiveness

Privacy Parameter Recovery Seen in Table 1 are the results

Setting	σ	Attack AUC	Estimated ϵ (Mean ± Std)
Weak	.05	0.514	-0.012 ± 0.811
Moderate	1.0	0.502	-0.049 ± 0.806
Strong	2.0	0.517	0.019 ± 0.029

Table 1. Caption

6. Discussion of Results

6.1. Shadow Model Scaling Limitations

In the original paper, 512 shadow models are used, in comparison to the 5 models used in this paper's implementation. This could be an explanation for:

- **Poor attack AUCs** (range of .50 to .52)
- **High variance in ϵ estimates**; The relationship is non-linear: doubling shadow models from 5→10 improved AUC by only 0.01 in our tests, suggesting the paper's scaled cloud setup was essential for their results.

6.2. Attack Quality Barrier

In the original paper, it is stated that the authors re-use shadow models from the LiRA attack to find the most vulnerable samples from training. From this it is assumed that MIAs provide discriminative signal, but the results of this implementation serve to indicate that performance rapidly degrades.

DP Setting	Viability
σ = .05	× Invalid (AUC $\leq$ 0.55)
σ = 1.0	× Invalid
σ = 2.0	✓ Marginal

Table 2. Viability of DP Settings

This suggests the paper's 40% improvement claim only withstands with higher performance, as satisfied with their CIFAR-10 experiment, but not satisfied with our synthetic data.

7. Contributions

This work makes principal contributions to the original paper's legacy:

7.1. Empirical Validation of Limits

- Bayesian Advantage:
  - Confirmed in strong privacy setting ($sigma$=2.0) where ϵ std (0.029) $\geq$ Clopper-Pearson equivalent (Nasr et al., 2021).
- Implementation Insights:
  - Demonstrated that a Bayesian method of this nature for privacy estimation requires a strong attack AUC for meaningful estimates for ϵ.
    This is seen in the Strong DP (σ = 2.0), which was the only scenario with an AUC (0.517) approaching this threshold that yielded a stable estimated ϵ.
    This is further demonstrated with the Weak/Moderate DP, in which lower AUC led to a degeneration of estimates.

7.2. Computational Tradeoff Analysis

- Confirmed the paper's claim that Bayesian estimation reduces sample size *only when attacks are effective*:
  - With AUC=0.517 (strong DP), achieved CI width of 0.028 (comparable to original paper's 40% reduction)
  - With AUC0.515, required $\geq$ 1200 samples but still got no signal
- **Tradeoffs between shadow model count and precision**: Although we cannot match the scale of the original AzureML deployment, our work proves that rigorous DP verification is possible without industrial resources, lowering barriers to adoption across academia and non-profits.

8. Conclusion

This project demonstrates that Bayesian methods are not merely theoretically superior for DP verification; they are practically more efficient. This is done through working towards empirically validating the original paper's claims despite limitations of working on a smaller scale.

Key takeaways include:

1. **Attacks must be of high-quality:** The original paper's promise of 40% tighter ϵ intervals hinges on effective MIAs. Our results demonstrated that below a certain threshold (AUC $\simeq$ 0.50–0.52), the Bayesian estimator degenerates into uninformative bounds, even with edge-case handling.
2. **TPR-FPR Gap is indicative of success**: We empirically validated that when TPR $\simeq$ FPR (difference $\leq$ 0.01), the privacy region collapses. Only in the strong DP setting (σ =2.0), here a marginal TPR-FPR gap of 0.0096 emerged, did the method produce a plausible estimate.

Future work should scale up the amount of shadow models, and pre-screen attacks via AUC prior to Bayesian estimation. Additionally, new studies could explore variational inference for further speed gains and extensions to federated learning scenarios.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS'16. ACM, October 2016. doi: 10.1145/2976749.2978318. URL http://dx.doi.org/10.1145/2976749.2978318.

Amin, K., Kulesza, A., and Vassilvitskii, S. Practical considerations for differential privacy. *arXiv preprint arXiv:2408.07614*, 2024.

Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P., Horsfall, P., and Goodman, N. D. Pyro: Deep universal probabilistic programming. *Journal of machine learning research*, 20(28):1–6, 2019.

Carlini, N., Jagielski, M., Zhang, C., Papernot, N., Terzis, A., and Tramer, F. The privacy onion effect: Memorization is relative. *Advances in Neural Information Processing Systems*, 35:13263–13276, 2022.

Nasr, M., Songi, S., Thakurta, A., Papernot, N., and Carlin, N. Adversary instantiation: Lower bounds for differentially private machine learning. In *2021 IEEE Symposium on security and privacy (SP)*, pp. 866–882. IEEE, 2021.

Near, J. P., Darais, D., Lefkovitz, N., Howarth, G., et al. Guidelines for evaluating differential privacy guarantees. *National Institute of Standards and Technology, Tech. Rep*, pp. 800–226, 2023.

Zanella-Béguelin, S., Wutschitz, L., Tople, S., Salem, A., Rühle, V., Paverd, A., Naseri, M., Köpf, B., and Jones, D. Bayesian estimation of differential privacy. In *International Conference on Machine Learning*, pp. 40624–40636. PMLR, 2023.