Group Assignment- Project for SD6123

Spring 2025

Work in groups of 4-5. Choose ONE topic from the following two options.

1 Sum Estimation under Differential Privacy

Sum estimation, a fundamental analytical task, has been extensively studied under differential privacy (DP). Given a dataset $D = \{x_1, x_2, \ldots, x_n\}$, where each x_i is an integer from the domain $\{0, 1, 2, \ldots, U\}$, our goal is to compute $\mathrm{Sum}(D) = \sum_i x_i$. The standard approach to achieve DP is the Laplace Mechanism, which adds random noise with a scale of U/ε to the query result, yielding an error of $O(U/\varepsilon)$. However, the challenge lies in the choice of U, which must be large enough to cover all possible datasets. For example, when summing salaries, U must be set to the salary of the world's richest person, even though such extreme values rarely appear in typical datasets. To address this issue, a more desirable approach is to achieve an instance-specific error—one that scales with the maximum value in the dataset, i.e., $\mathrm{Max}(D) = \mathrm{Max}_i x_i$, rather than an overly conservative global bound. So far, these have been several DP protocols designed to achieve such an error.

What you should do in this project:

1. **Review existing solutions:** Select and review one paper from the five listed below.

   [A] https://arxiv.org/pdf/2403.10116

   [B] https://www.cse.ust.hk/ yike/DPMean.pdf

   [C] https://arxiv.org/abs/2111.02598

   [D] https://www.cse.ust.hk/ yike/R2T.pdf

   [E] https://www.cse.ust.hk/ yike/ShiftedInverse.pdf

2. **Theoretical analysis:** (a). Design a sum estimation algorithm based on the techniques presented in your chosen paper. Please note that these five papers may study more advanced problems but all include sum estimation as a special case. You should present the sum estimation algorithm. Give complete DP proof. (b). Analyze their error bounds (including constant factors) and (asymptotic) computational complexity.

3. **Evaluation:** Use at least three privacy-sensitive datasets from Kaggle to compare the errors of your selected techniques against the Laplace Mechanism. Discuss how the experimental results match your theoretical analysis.

2 Federated Learning

Federated Learning (FL) is an emerging approach that allows machine learning models to be trained across decentralized data sources without transferring raw data to a central location. This method is particularly useful for privacy-sensitive applications such as healthcare, finance, and personalized recommendations. However, FL also presents challenges, including communication overhead, potential security vulnerabilities, and trade-offs in model accuracy.

What you should do in this project:

1. **Dataset selection:** Choose any privacy-sensitive dataset (e.g., from Kaggle or Huggingface). You can use any framework of your choice (e.g., Flower).

2. **Baselines:** Develop a centralized learning baseline. Then, develop baselines using at least three different FL algorithms. You may refer to frameworks (e.g., Flower) or repositories, such as ours `https://ntu-zjy.github.io/DomainFL/`, for some implementation of common baselines.

3. **Privacy risks and protections:** Assess privacy leakage risks (i.e., attacks) and defences, such as differentially-private FL.

4. **Evaluation:** How do the different methods perform? Also assess the computational efficiency and communication overhead of the different methods.

The Deliverables

Here are the assignment requirements:

1. **Group Report:** Prepare a ten-page report to present your group project. The format should adhere to the AAAI formatting guidelines (same as the individual report). References do not count into the page limit. You MUST adhere strictly to the page limit; do not include "appendix". Your report should include: member names/matriculation number, introduction/motivation of your work, a review of related works and methods, discussion of methods chosen, an evaluation of experiments and/or results. Note: There is no need for abstract.

2. **Code:** Upload your codes on Github. Share a link that we can access for marking.

3. **How to get a good grade:** (i) your project shows a good understanding of the topic, (ii) your execution is well-documented and replicable, (iii) your results are presented clearly and evaluated with rigour, (iv) your thoughts are presented logically.

4. **Submission:** The report is due on **27 April, 2359.** Upload to NTULearn. Penalty of 20% every 12 hours late.