Differentially Private Linear Regression

Due: Dec 8, 2025, 10 AM PT

In this homework assignment, we implement DP training of a simple linear regression task and explore the role of different parameters.

1 Synthetic Data Generation

Implement a mechanism to generate synthetic data for a linear regression task. The data generation process should be flexible to accommodate different choices of:

- n: The number of data points.
- d: The dimensionality of the feature space (number of parameters).

The data should follow a linear model $Y = X\beta^* + \eta$, where $X \in \mathbb{R}^{n\times d}$ is the feature matrix, $\beta^* \in \mathbb{R}^d$ is the true parameter vector, and η represents additive Gaussian noise.

2 Implementation of Optimization Algorithms

Implement the following algorithms for training the linear regression model (minimizing the Mean Squared Error loss):

- **SGD:** Standard Stochastic Gradient Descent (this will serve as your non-private baseline).
- **DP-SGD:** Differentially Private Stochastic Gradient Descent.
- **DP-Adam:** Differentially Private Adam.

For the DP implementations (DP-SGD and DP-Adam), ensure correct implementation of the following key steps:

1. **Per-Sample Gradient Clipping:** Clipping the L2 norm of the gradients for each individual data point to a threshold C.
2. **Noise Addition:** Adding Gaussian noise to the aggregated clipped gradients, calibrated to the clipping threshold and the desired privacy budget.

3 Privacy Accounting

In all differentially private experiments:

- Set the privacy parameter $\delta = 1/n^{1.1}$.
- Use advanced composition theorem to compute the value of ϵ in each private training task.

4 Experiments and Analysis

Conduct the following experiments and provide a detailed analysis of the results. Use the Mean Squared Error (MSE) on a separate, non-private test set to evaluate model utility.

4.1 Exploring the (ϵ, δ)-DP Trade-off

- Fix n and d (e.g., $n = 10000, d = 100$).
- Train models using DP-SGD and DP-Adam for a range of target ϵ values (e.g., $\epsilon \in \{0.5, 1, 4, 10\}$). Adjust the noise multiplier accordingly while keeping δ fixed.
- Plot the utility (MSE) versus ϵ.
- Compare the performance of DP-SGD and DP-Adam under the same privacy budget.

4.2 Exploring the Role of Batch-Size

- Fix all parameters and just change the batch-size
- Perform an ablation study by varying the batch-size
- Discuss how the choice of batch-size impacts convergence and the final utility

4.3 Role of the Clipping Threshold (C)

- Fix n, d, and the target ϵ (or equivalently, the noise multiplier).
- Perform an ablation study by varying the clipping threshold C. You may want to explore values based on the percentiles of the observed gradient norms.
- Discuss how the choice of C impacts the convergence speed and the final utility of both DP-SGD and DP-Adam.
- Discuss the trade-off between the bias introduced by clipping and the variance introduced by the privacy noise.

4.4 Impact of Data Size (n) and Dimensionality (d)

Investigate how the characteristics of the dataset affect the utility of private training.

- **Impact of n:** Fix d and the target ϵ. Vary n (e.g., $n \in \{10^2, 10^3, 10^4\}$). Analyze how the "cost of privacy" (the utility gap between the private models and the non-private SGD baseline) changes as the dataset size increases.
- **Impact of d:** Fix n and the target ϵ. Vary d (e.g., $d \in \{10, 100, 500\}$). Analyze how the dimensionality of the problem affects the performance of the private models.

5 Deliverables

You should submit a single concise report (a single pdf file) summarizing your methodology, the implementation details of the algorithms, the experimental results (including clear plots and tables), and an analysis of the findings for each section.

Place your code in a well-documented Google Colab notebook that includes the data generation process, algorithm implementations, and all experiments. Add the link to your Google Colab at the top of your report.