Differential Privacy in Machine Learning

Vraj Thakkar

May 2025

Table of Contents

Table of Contents

Privacy Attacks

- Privacy Attack on Netflix(2008)
  - Netflix announced a prize of 1,000,000 USD for anyone who could create a better recommendation system than Netflix.
  - Data:- (anonymized) user ID, movie ID, rating, and date.
  - Narayanan and Shmatikov did a linkage attack on the dataset with the publicly available IMDb dataset.

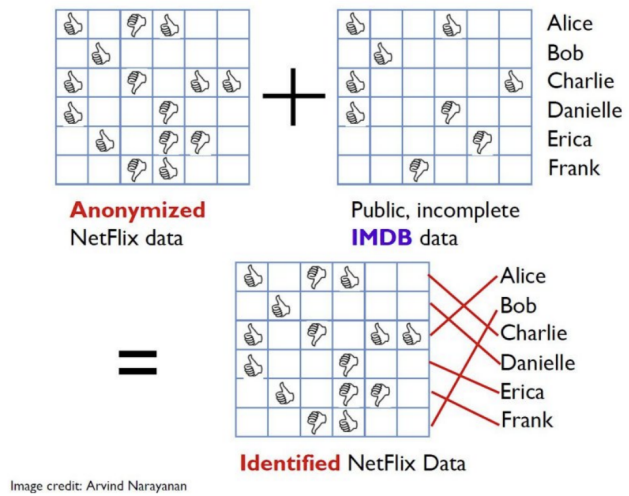


Figure: Linkage attack Demonstrated in [1]

[1] Arvind Narayanan and Vitaly Shmatikov, Robust de-anonymization of large sparse datasets, 2008

Privacy Attacks

- The Secret Sharer(2018)
  - Neural Networks memorize the training data even when not overfitting.
  - Increase in the number of times a phrase like "Rohit's credit card number is 091-119-202" appears in training data increases memorization.
  - These phrases can be extracted from the model.

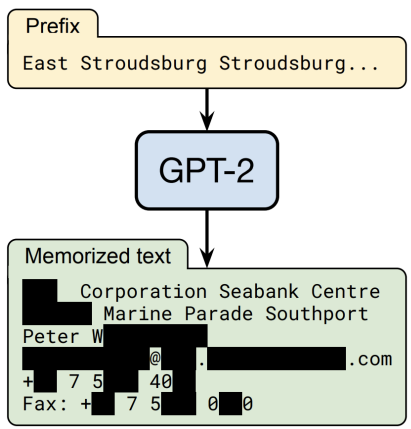


Figure 1: **Our extraction attack.** Given query access to a neural network language model, we extract an individual person's name, email address, phone number, fax number, and physical address. The example in this figure shows information that is all accurate so we redact it to protect privacy.

Figure: GPT-2[1]

[1] Extracting data from LLMs by Carilini et al., 2020.

Collecting Data

- Most recommendation systems/ surveys rely on collecting private data and training ML models or publishing statistics.
- Examples: Google ads, Facebook prevents harmful URL sharing, Collecting data for smoking
- Adversaries can use this to get training data, and this calls for privacy-preserving techniques.

Table of Contents

What is Privacy?

- Privacy for individuals in the context of data analysis is about not getting identified.
- If the results of a data analysis can revert to any individual, revealing his/her sensitive information about disease/ drug habits, etc., then it is a privacy violation.
- Privacy comes from Plausible deniability!

Disease or not?

- Toss a coin, if heads write true value, if tails toss again $\Longrightarrow$ heads then yes, tails then no.
- Plausible deniability!
- If p = true proportion of diseased, q = observed proportion of disease
- $q = p/2 + 1/4$ which implies $p = 2(q + 1/4)$.

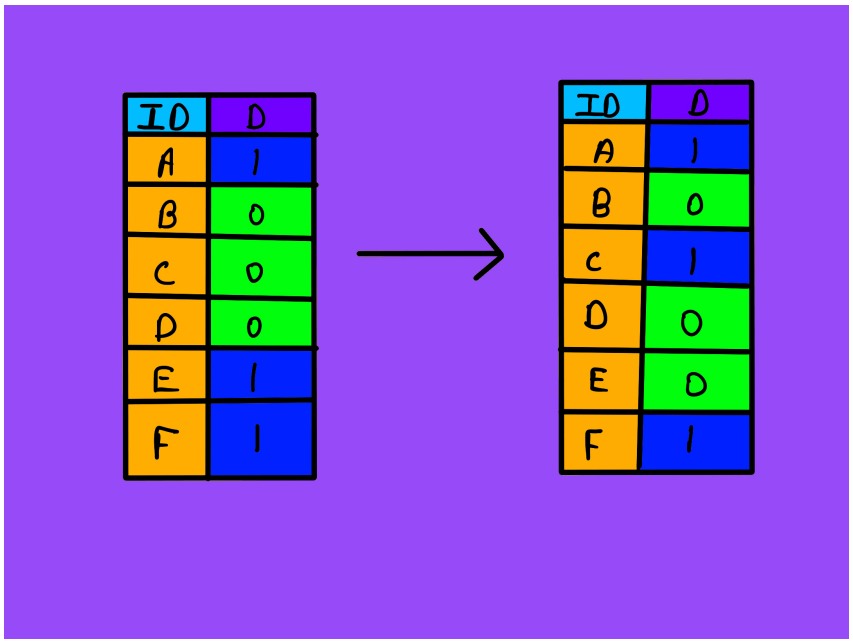


Figure: Original data and DP-data

Differential Privacy

Definition

A mechanism $M : \mathcal{X}^n \rightarrow \mathcal{Y}$ is ϵ-indistinguishable if for all pairs of neighbouring datasets X and X' and all $T \subseteq \mathcal{Y}$ we have

$$Pr(M(X) \in T) \leq e^{\epsilon} Pr(M(X') \in T)$$

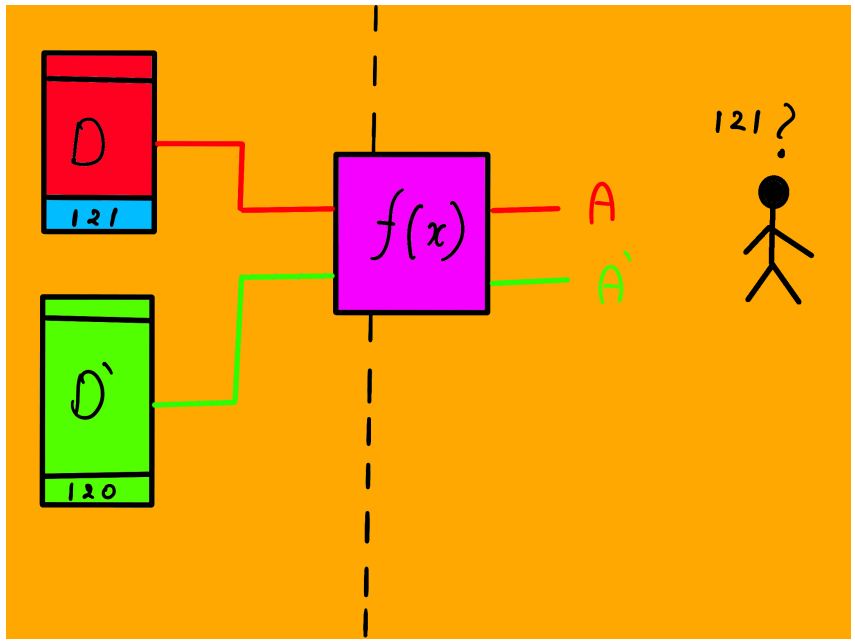


Figure: Differential Privacy Definition

Properties of Differential Privacy

- Post-Processing: If I add the guarantees of differential privacy in the pipeline in any function, then all the later functions enjoy the guarantee of Differential Privacy.
- Add DP to the data $\implies$ All data analysis has DP guarantees.
- This also means that no additional information will be able to violate privacy.

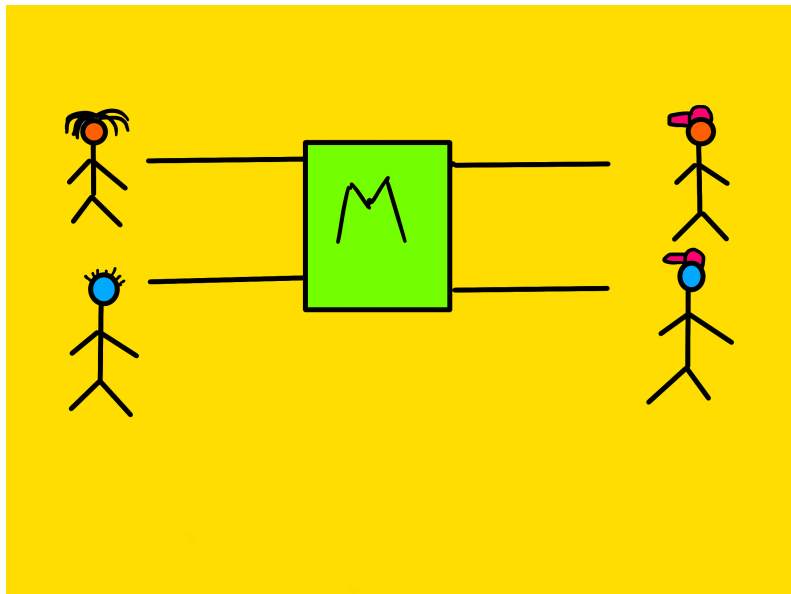


Figure: Differential Privacy ensures that sensitive user data is not extracted

Properties of Differential Privacy

- Composition: If we have mechanism M where $M = (M_1, M_2 \ldots\ldots M_k)$ where each M_i is ϵ-DP then M is $k\epsilon$-DP.
- With more iterations, the randomization adds up and could lead to converging values.

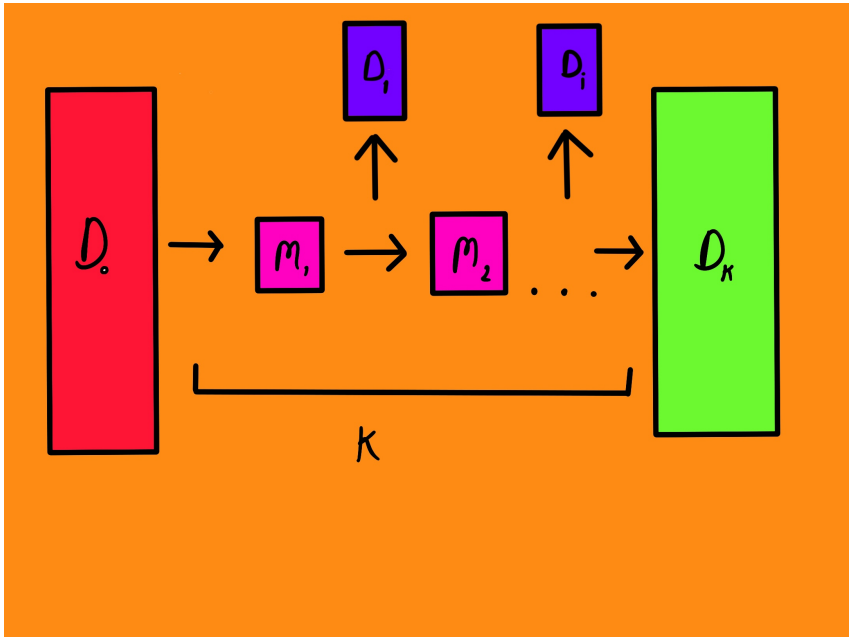


Figure: Composition of k-mechanisms

Approximate Differential Privacy

- Difference in definition for (ϵ, δ)

$$\Pr[M(X) \in W] \leq e^{\epsilon} \Pr\left[M\left(X'\right) \in W\right] + \delta$$

- This is a relaxation of the definition of pure differential privacy.
- There are two main benefits of this guarantee.
  - We need to add much less noise than for pure DP, and hence we get much greater utility
  - Also because of the lesser noise we get better bounds for composition, of the order of $\sqrt{k}$ instead of k

Mechanisms for Differential Privacy

Definition

Let $f : \mathcal{X}^n \to \mathbb{R}^k$. The ℓ_p-sensitivity of f is

$$\Delta_p^{(f)} = \max_{X,X'} \left\| f(X) - f\left(X'\right) \right\|_p$$

- Here, X and X' are neighbouring datasets and p depends on the specific mechanism and privacy gaurantee.
- $f(X) = \frac{1}{n} \sum_i^n x_i$ where $x_i \in \{0, 1\}$ then $\Delta_1^f = 1/n$.

Mechanisms for Differentail Privacy

Definition

Let $f : \mathcal{X}^n \to \mathbb{R}^k$. The Laplace mechanism is defined as

$$M(X) = f(X) + (Y_1, \ldots, Y_k)$$

where the Y_i are independent Laplace (Δ_1/ϵ) random variables.

- **Example**
- $f'(X) = f(X) + Laplace(1/n\epsilon)$ where $Laplace(b) = \frac{1}{2b} \exp(-|x|/b)$
- f' is ϵ-DP
- We can also give the utility of the algorithm using Markov's inequality.

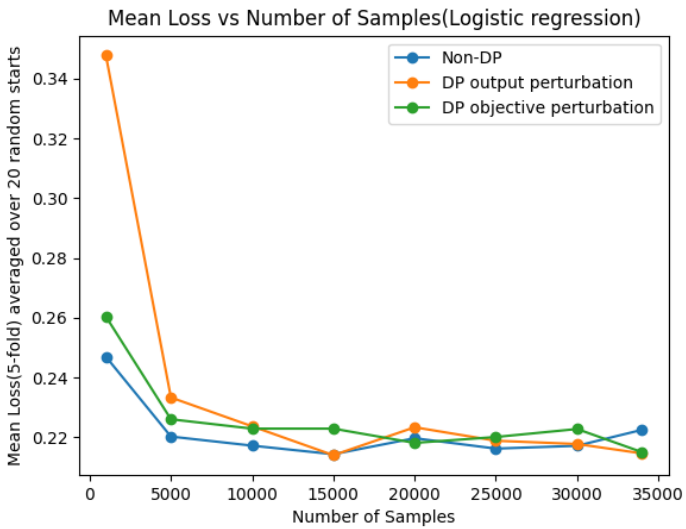
Table of Contents

Understanding the pipeline

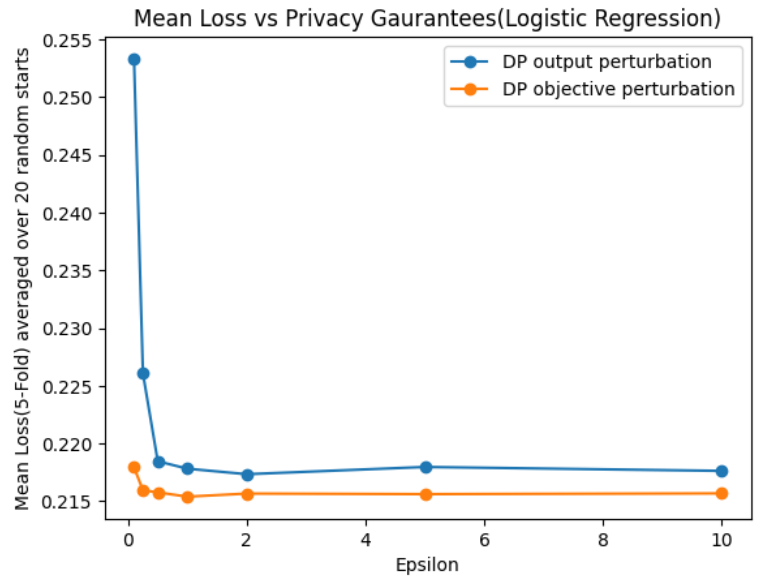
- How we define DP also depends on how we plan to use the model.
- For example, if we are to give the weights of the model, then we need to provide DP guarantees either in the data or during training.
- Broadly, there are three places to add DP in the Machine Learning models.
  - Noising Weights
  - Objective Modification
  - Gradient noising

Algorithms for Logistic Regression

- Algorithm 1: Do normal logistic regression and then add noise to the final weight parameters.
- Algorithm 2: Add noise to the objective function and do gradient descent with added noise.[2]



(a) Loss vs Samples



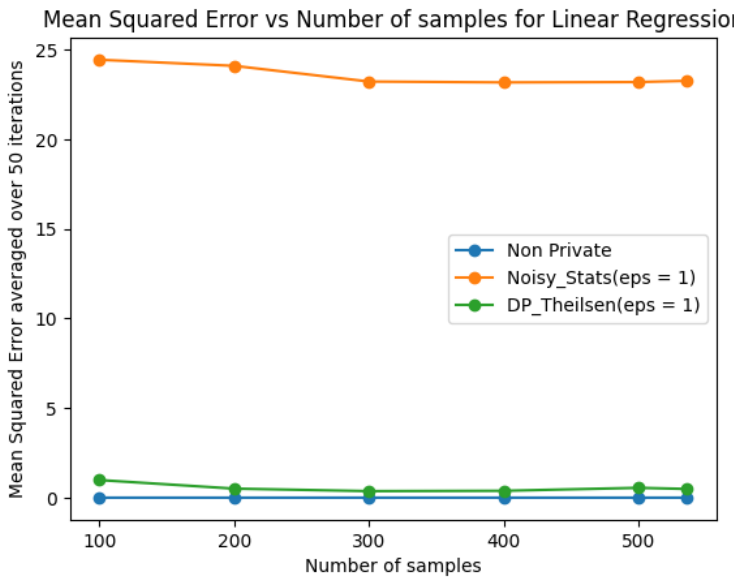
(b) Loss vs ϵ

Figure: Private Logistic Regression

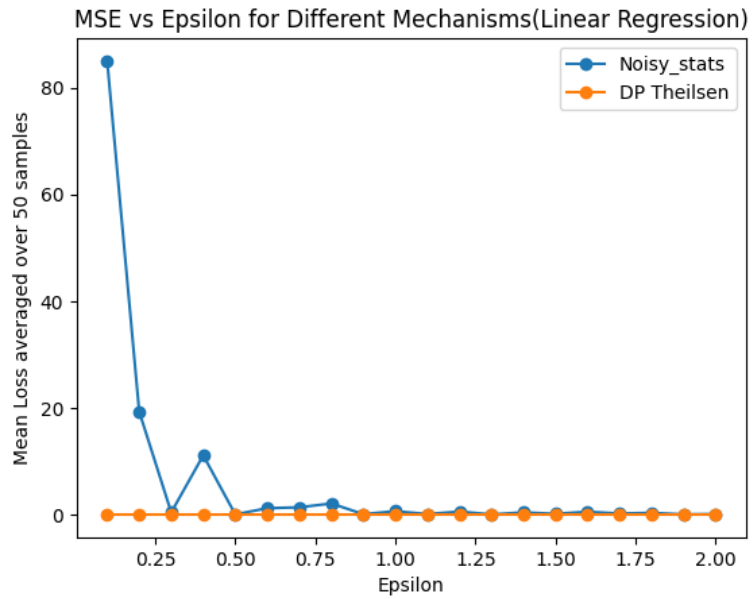
[2] Kamalika Chaudhuri, Claire Monteleoni, Privacy-preserving logistic regression, (NIPS, 2008).

Algorithms for Simple Linear Regression

- Algorithm 3: Add noise to the output parameters.
- Algorithm 4: Calculate the slope of all pairs of points and choose the median using differential privacy.[3]



(a) Loss vs Samples



(b) Loss vs ϵ

Figure: Private Linear Regression

[3] Daniel Alabi, Audra McMillan, Jayshree Sarathy, Adam Smith, and Salil Vadhan, Differentially private simple linear regression, 2020

Differential Privacy for Deep learning

Algorithm 5: DP-SGD[4]

1. Parameters: Noise rate(σ), Clipping norm(C), Batch size(L), Learning rate(ℓ), number of epochs(n).
2. Take each point with probability L/N where N is the total number of samples.
3. Compute the gradient for each point and clip it.
4. Add each gradient term and Gaussian noise according to σ.
5. Perform gradient step according to the learning rate.

[4] Abadi et al., Deep Learning with Differential Privacy, 2016

DP-SGD

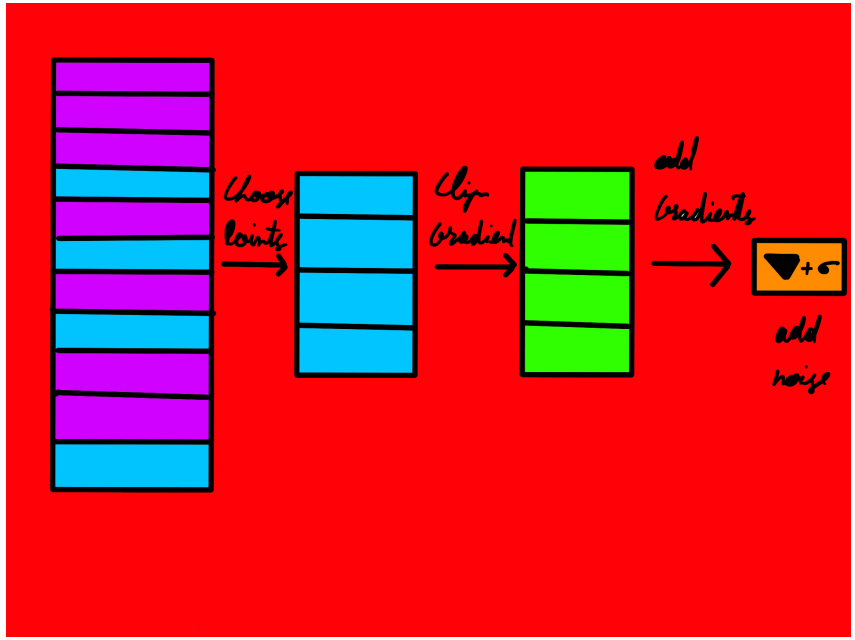


Figure: DP-SGD

DP-SGD results

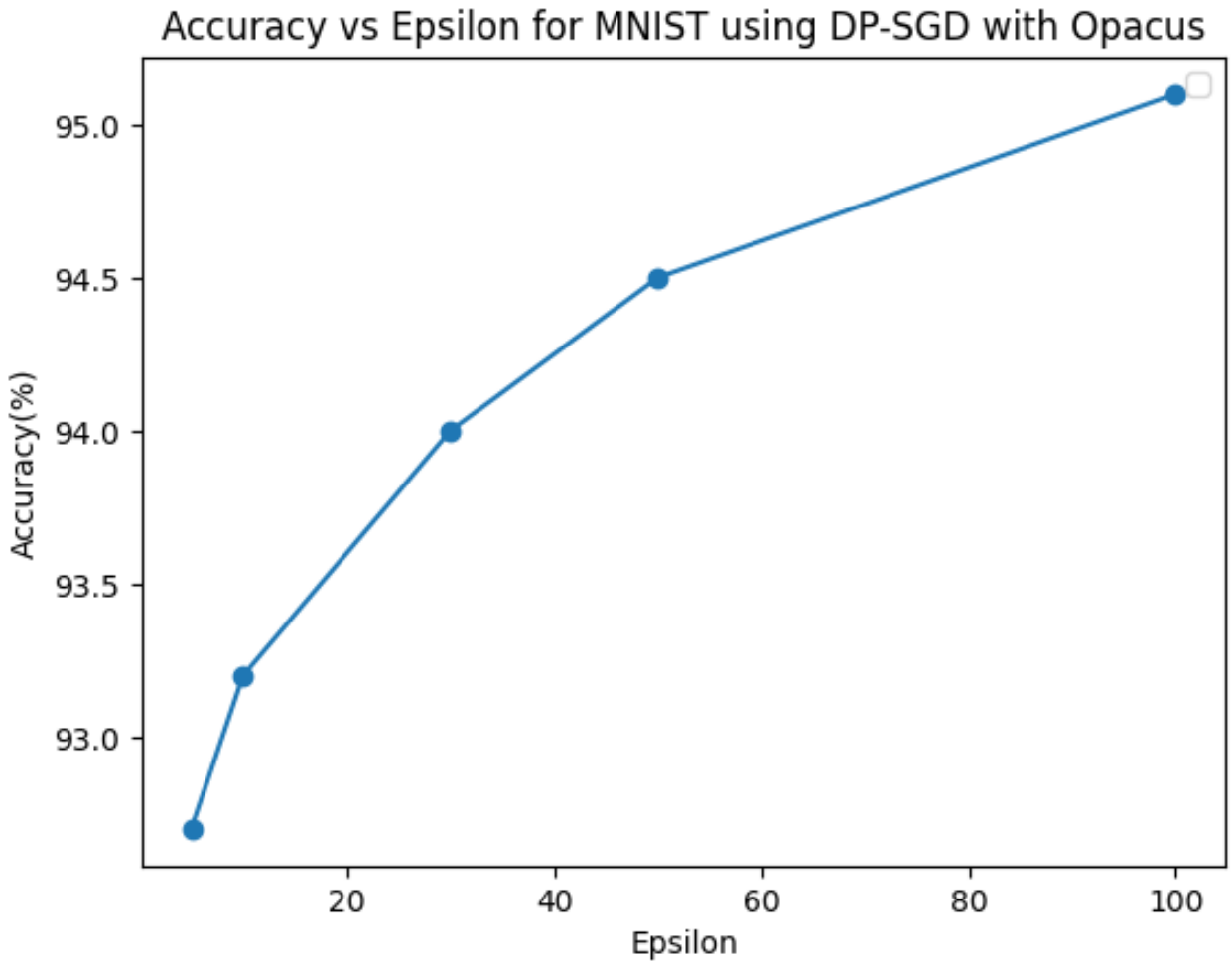
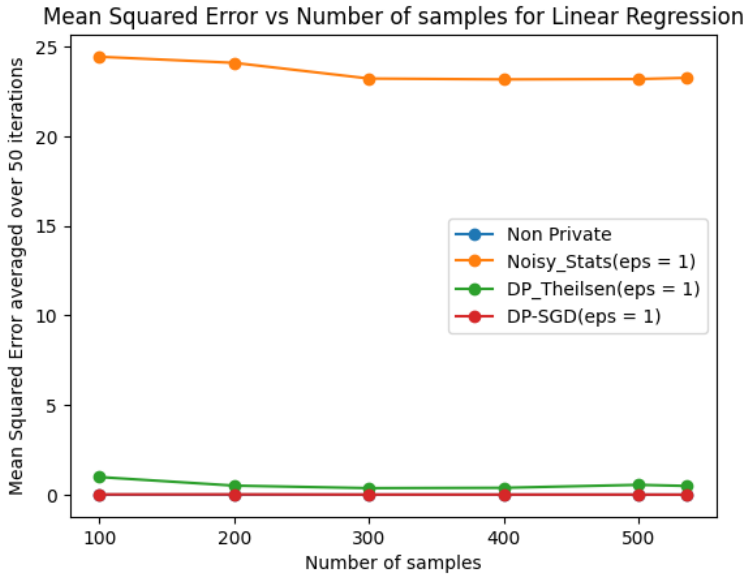
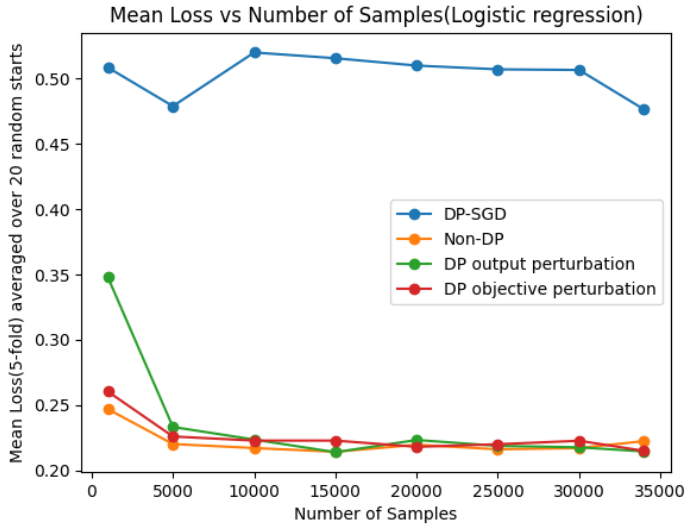


Figure: Mnist using DP-SGD with different ϵ

DP-SGD results



(a) Private Linear Regression(Loss vs Samples)

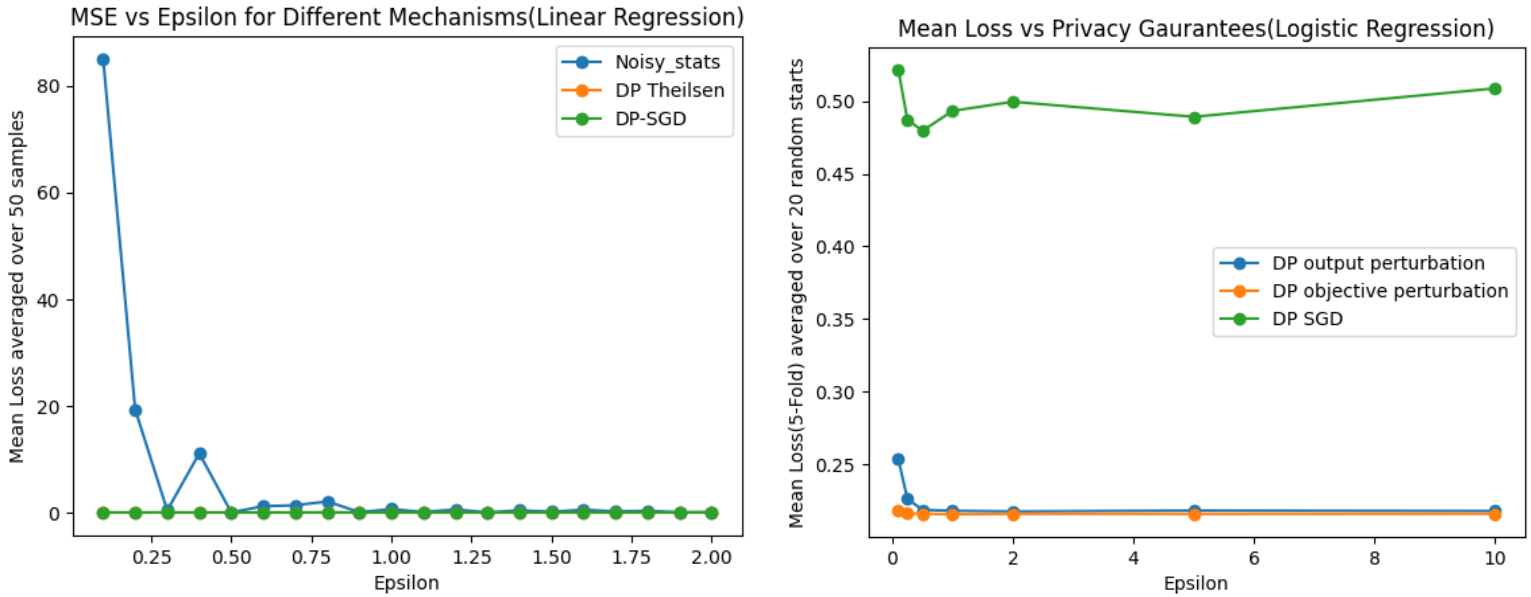


(b) Private Logistic Regression(Loss vs Samples)

Figure: Loss vs Samples for different tasks

- There are chances that for strong privacy guarantees($\epsilon \leq 1$) we have to add a lot of noise in the gradient and loose significant utility.

DP-SGD results



(a) Private Linear Regression(Loss vs ϵ)

(b) Private Logistic Regression(Loss vs ϵ)

Figure: Loss vs ϵ for different tasks

Discussion about DP-SGD

- Slow: GPU can't be utilized because of the need for the gradient of each data point.
- Low utility: Even SOTA implementations give about 69 % accuracy on CIFAR datasets whereas about 99.7 % accuracy can be obtained in non-private case.
- Moments accountant: Stronger composition guarantees than advanced composition.
- In LLMs, we usually train the model on public data and tune the model on private data using DP-SGD.

Real World Deployments

- Google Trends uses differential privacy to select which queries to proactively show on the website.
- Differential privacy was used to identify the movement of individuals in COVID-19 by Google[5] and Facebook.
- All the G-board prediction models have DP guarantees now.
- US census 2020 realses data using DP to protect identification of people in different demographics.
- DP allows companies to collect more data and it's use is only going to increase.

[5] Aktay et al., Google COVID-19 Community Mobility Reports: Anonymization Process Description, 2020

Future Work

- Federated Learning: Federated learning is always paired with DP. Examples: G-board, Self-driving cars, ads, etc.
- DP-auditing: How to form attacks?
- DP and fairness: How do fairness and DP interplay? Dhruv Shah(202103017)

References

- Dwork, C., Roth, A. (2014). The Algorithmic Foundations of Differential Privacy. Foundations and Trends in Theoretical Computer Science, 9(1-2), 1-180.
- Natalia Ponomareva, Hussein Hazimeh, Alex Kurakin, Zheng Xu, Carson Denison, H. Brendan McMahan, Sergei Vassilvitskii, Steve Chien, Abhradeep Thakurta. How to DP-fy ML: A Practical Guide to Machine Learning with Differential Privacy, JAIR 2023.
- Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In Proceedings of the 2016 ACM Conference on Computer and Communications Security, CCS '16, pages 308–318