NUS CS-CS5562: Trustworthy Machine Learning	October 30, 2023

Assignment 4

Lecturer: Reza Shokri *Student: Niharika Shrivastava A0245355A*

2 Private Training via DP-SGD

2.1 DP-SGD

Model	Train Accuracy	Test Accuracy
Non-Private	83.51	82.52
Private	84.22	82.39

The test accuracy for the private model decreased compared to the non-private model, i.e., the utility for the private model decreased at the cost of increased privacy.

2.2 Computing Privacy Parameters of DP-SGD using Moments Accountant

Given,

Parameter	σ	B	E	N	δ	q	T
Value	0.05	1	20	10^4	10^{-5}	$B/N = 10^{-4}$	$E \times N = 2 \times 10^5$

Table 1: Caption

By Theorem 2.2 (Tail bound) in [1],

$$\delta = \min_{\lambda} e^{[\alpha(\lambda) - \lambda\epsilon]}$$

$$\frac{\partial \delta}{\partial \lambda} = e^{[\alpha(\lambda) - \lambda\epsilon]}[\frac{\partial \alpha(\lambda)}{\partial \lambda} - \epsilon] = 0$$

$$\Rightarrow \frac{\partial \alpha(\lambda)}{\partial \lambda} = \epsilon$$

By Theorem 2.1 and Lemma 3 in [1], the log moment of DP-SGD can be bounded as follows:

$$\alpha(\lambda) \leq Tq^2\lambda^2/\sigma^2$$

$$\frac{\partial \alpha(\lambda)}{\partial \lambda} \leq \frac{2Tq^2\lambda}{\sigma^2} \Rightarrow \epsilon \leq \frac{2Tq^2\lambda}{\sigma^2}$$

By Theorem 2 in [1], to guarantee DP-SGD to be (ϵ, δ)-differentially private, it suffices that:

$$\epsilon \geq \frac{2Tq^2\lambda}{\sigma^2}$$

$$\Rightarrow \epsilon = \frac{2Tq^2\lambda}{\sigma^2}, \alpha(\lambda) = Eq^2\lambda^2/\sigma^2$$

By substituting the value of ϵ and $\alpha(\lambda)$,

$$\lambda = \sqrt{\frac{-\sigma^2 \log \delta}{Tq^2}} = 3.7935$$

$$\epsilon = 6.0967$$

2.3 Effect of Clipping Norm on Accuracy

1. The l_2 norm trajectory of the privatized gradients is much noisier than the true gradients since we add Gaussian noise in each epoch. Thus, the gradient descent is not smooth.

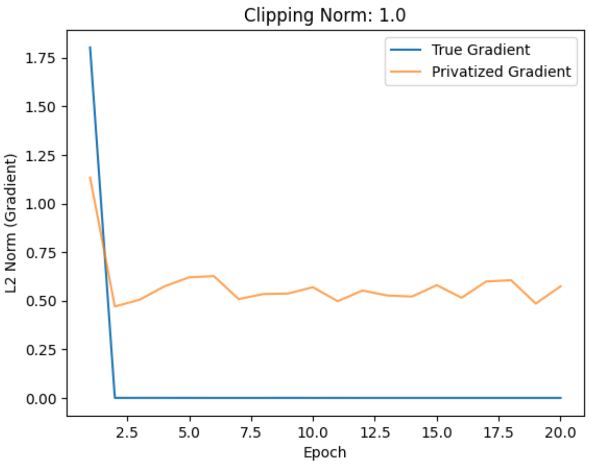
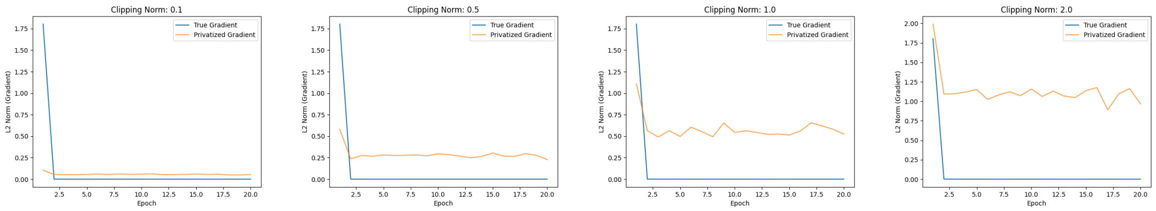


Figure 1: Question 2.3.1

2. As the clipping norm increases, the magnitude of the private gradients increases and they become increasingly jagged. This is because the standard deviation of the noise added to the gradients is linearly proportional to C.



3. As the clipping norm C increases, the accuracy of the model decreases. This is because the amount of noise added to the gradients increases with increasing clipping norm which lowers the accuracy.

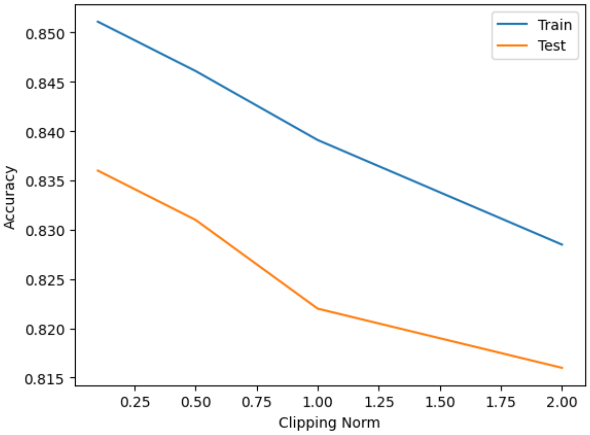


Figure 2: Question 2.3.2

3 Membership inference on DP models

3.1 Theorem Proof

From the definition of (ϵ, δ)-differential privacy,

$$Pr[Y = y|X = x] \leq e^{\epsilon} \cdot Pr[Y = y|X = x'] + \delta$$

where $x = D \bigcup z, x' = D$.

- $Pr[Y = y|X = x]$ means that the attacker correctly predicts that $z \in x$, i.e., the True Positive Rate (TPR).
- $Pr[Y = y|X = x']$ means that the attacker wrongly predicts that $z \in x'$, i.e., the False Positive Rate (FPR).

Thus, by rewriting the definition as follows:

$$TPR \leq e^{\epsilon} \cdot FPR + \delta$$
$$\Rightarrow e^{\epsilon} \cdot FPR + (1 - TPR) \geq 1 - \delta$$

Similarly, we can rewrite the definition of DP as follows:

$$Pr[Y \neq y|X = x'] \leq e^{\epsilon} \cdot Pr[Y \neq y|X = x] + \delta$$

- $Pr[Y \neq y | X = x']$ means that the attacker correctly predicts that $z \notin x'$, i.e., the True Negative Rate (TNR).
- $Pr[Y \neq y | X = x]$ means that the attacker wrongly predicts that $z \notin x$, i.e., the False Negative Rate (FNR).

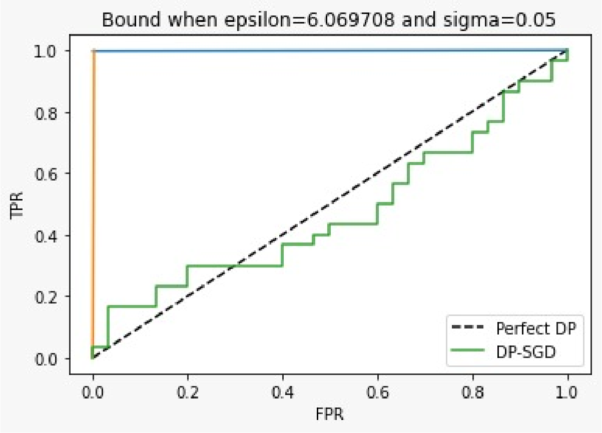
Thus, by rewriting the definition as follows:

$$
\begin{gathered}
TNR \leq e^{\epsilon} \cdot FNR + \delta \\
1 - FPR \leq e^{\epsilon} \cdot (1 - TPR) + \delta \\
\Rightarrow e^{\epsilon} \cdot (1 - TPR) + FPR \geq 1 - \delta
\end{gathered}
$$

Plotting

- The TPR, FPR values for distinguishing two histograms satisfy the inequalities in Theorem 1. Moreover, this membership inference attack is so poor that it is almost indistinguishable to infer membership even in the worst-case scenario.

  However, the bounds of our DP-SGD are too loose. The AUC is almost 1. This means that $\sigma = 0.05$ is not enough to make our algorithm a good differentially private algorithm.



- For a higher noise multiplier, ϵ would be lesser and the bounds would be much tighter, thereby making it a better defense algorithm.

References

[1] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. *In Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, page 308–318, 2016.