# Prospectus

**The Effects of Clipping on Differentially Private Machine Learning**

(Technical Topic)

**Social Impact of RF London Trust giving unconsented patient data to GoogleAI DeepMind**

(STS Topic)

By
Kush Patel
October 1, 2020

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

Signed: _____

Approved:_____ Date _____
Ben Laugelli, Department of Engineering and Society

Approved:_____ Date _____
David Evans, Department of Computer Science

**Sociotechnical Problem**

Machine learning is an approach that is on the rise of user acceptance into society. Machine learning (ML) is a profound technique used to boost the computing and research powers of mankind. As this technology becomes increasingly common in consumer facing products, we need to examine the implications of such technology from a safety and security perspective. One issue is termed “privacy”, where such complications can occur in the form of data leakage of sensitive information or an accidental release of certain private and proprietary algorithms or neural net structures (Cristofaro et al., 2020). The level of privacy that a certain algorithm possesses can be seen as the degree of safety and security of that machine learning algorithm.

Differential privacy has become the standard method used to create privacy-preserving machine learning algorithms. While there are many examples of experiments that have been conducted to gain a better understanding of the implications of certain types of privacy tactics used, there is still little confidence within the research community on how the different variables or parameters involved need to be set to achieve acceptable privacy utility levels. Thus, current privacy-preserving machine learning implementations do not implement tested methods to achieve the desired results. My technical solution aims to identify trends in the form of relationships between certain parameters and the privacy of these machines.

There is a direct relationship between the rise of computing power and increase in data generated and this will become increasingly evident in the next coming decades (Anonymous, personal communication, 2020). As companies and organizations increase the sizes of their data lakes (pools of data that can be accumulated as a byproduct of humans using their technologies), consumers will be trading their privacy for personalization in exchange for the service. There are

technical and social aspects that need to be considered when handling this exchange. I will dive deeper into analyzing the event of the Royal Free London NHS Foundation Trust delivering information on over 1.6 million hospital patients to DeepMind, Google's AI, without the patients' consent. The general population needs to be educated about the dangers of artificial intelligence and what they are consenting to when using such technologies. From mild examples of Netflix and Google collecting user data to create personalized recommendations, to even the most extreme cases, such as healthcare and social security data begin used and stored, a simple technological fix through the means of patching data leaks from these models is not enough to fully address the social aspects of this issue.

To effectively bring light and reason to this pressing issue, both the technical and social aspects must be addressed. Below I outline my research in attempting to discover a fitting technical solution where I identify a trend in values of specific parameters to yield a privacy-preserving machine learning model. I also use the actor-network theory to analyze how we can make the proper strides to drive protection for the consumer using political strategies and regulations on the systems that use machine learning to harvest and analyze data. The ultimate goal is to ensure that we, as a society, can foster the growth of this powerful technology in a safe and inclusive manner.

**Technical Problem**

Clipping is a technique that is widely used in the frame of machine learning. Gradient clipping is an approach used to prevent exploding gradients (when large updates occur to the model in a single iteration which can lead to the model becoming unstable) in very deep networks. As learned in previous studies, the primary driver for successful experiments with

differential privacy methods applied in machine learning is the proper modification of machine learning algorithms to satisfy differential privacy (Chaudri et al., 2011). To convex optimization problems, such as empirical risk minimization (ERM) algorithms, we can add several methods to achieve differential privacy during the model's learning process.

Abadi described how every layer in a multi-layer neural network is considered separately (Abadi, et al., 2016). This yields the liberty of choosing different clipping thresholds and noise scales for each layer of the network, if desired. He also advised that the clipping and noise parameters may vary with the number of training steps. However, he deferred to only using constant clipping thresholds and noise scales for his experiments. This is explained by the contradicting effects of limiting the gradient. Clipping in general will destroy the unbiasedness of the gradient estimate (Krizhevsky et al., 2009). However, the average clipped gradient may point in a very different direction from the true gradient if the clipping parameter is set to be too small. On the contrary, if the clipping parameter is too large, we must add more noise to the gradients. In his experiment he proposes that the best way to choose such a clipping threshold is by taking the median of the norms of the unclipped gradients over the course of training, since noise is added based on the inverse relationship between the noise scale and clipping threshold. Thus, in Abadi's experiments he adhered to using only a fixed clipping threshold between the range 3 and 10 inclusive.

In the past there have been two main methods of applying clipping: batch and per-instance. In Jayaraman's experiments both batch and per-instance clipping techniques were used (Jayaraman et al. 2019). He states that while batch clipping is the standard used in practicing deep learning, it is also more computationally efficient. However, from the perspective of differential privacy as it is used in machine learning, we can get more accurate models for a

given privacy budget, when per-instance clipping is used. This is the case because per-instance clipping uses the privacy budget more efficiently. Jayaraman fixed the clipping threshold at 1 and compared the accuracy loss of logistic regression models trained over the CIFAR-100 dataset with both batch clipping and per-instance clipping (Kaggle, 2014). He saw that per-instance clipping allows learning more accurate models for all values of epsilon and amplified the differences between the different mechanisms, making it certain that per-instance clippings seems like the clear choice when working with differential privacy in machine learning (Jayaraman, personal communication, 2020).

Recently, there have been many experiments conducted to analyze or work with deep learning using the lens of differential privacy. While the previous methods seem complex, they propose a simpler method for tackling this issue. This approach aims to add noise to the gradients, in the form of clipping, as a possible solution. In the light of understanding that there stand numerous unknowns on the true potential benefits or harms that can be caused by simply applying the clipping threshold, my experiment will elect to dive deeper into how variation in the application of clipping can impact the success of an attack. Through the methods of my experiment, I will visit the idea of looking at ranges of fixed and dynamic clipping thresholds to validate the design of my solution. Where dynamic clipping can be defined as a trained attack where the clipping threshold is changed for every epoch. The experiment will analyze the test accuracy and attack success to support my claim for the adoption of my solution.

**STS Problem**

As described previously, large corporations and organizations are forming a habit of creating large data lakes with consumer data. While it may not be in the best intention of these

entities to use this data for evil, hackers may try to exploit weaknesses in their machine learning algorithms to extract and use this data with negative intentions. On the flip side, many of the large giants that use these embedded machine learning algorithms in their products are left unregulated with the freedom to use data acquired without consent of the owners of the data points. The Royal Free London NHS Foundation Trust has recently been under scrutiny when they delivered information on over 1.6 million hospital patients to DeepMind, Google's AI, without the patients' consent (Wiggers, 2019). There are many social implications and factors that need to be unpacked to fully understand the severity of this event. Scholars currently understand the role of this act as unfaithful as its occurrence was not made public as the corporations quickly made deals with the media to hide them. Without proper regulation by the government, these large organizations can cause massive waves in the world of personal privacy and security. We must also consider the repercussions that may arise if hackers are able to steal information from these private datasets.

The current role of social actors is flawed since fingers can be pointed in both directions, where the companies can use and harvest data without consent from the consumers, while hackers can steal the data from the weak implementations of the machine learning algorithms (Roy, 2020). While this is true, current scholars and professionals are directing fault at the hands of the government as well in regards to loose regulations and policies for these companies. The players in this actor network include the consumers, companies, hackers, and government. The first take can be seen in the relationship between the government and the large entities, both of which are a crucial part of the actor network. The government needs to spend more time listening to the opinions of computer scientists on this matter and implement creative tacts to ensure that they close all loopholes that these companies are taking advantage of. Another actor at fault can

be the system that plans the education of the general population itself. As technology improves, it is readily available to the public but without any forewarning of the harm that it can bring. By bringing light to this, we may see a revelation occur among the users and drive a force in the actor network. By resolving this problem, society as a whole, the common population, will fully understand what the issue is and channel pressure towards the government to enact stricter regulation in essence changing society for the better. Finally, the last piece of the actor network to examine is the link between the corporations and the hackers. If the corporations are going to store the data and use them as material for their machine learning models then they need to operate within a certain degree of safety (Prabhu, 2019). I propose that each algorithm that is used needs to have a maximum data leakage threshold in place so that the hackers are not able to exploit the process and steal millions and millions of data points from these sensitive pools. My design has potential for success as it will trigger a chain of events within the defined actor network to create an ever repeating cycle that will self improve itself over time (Cressman, 2009).

To support my claim, I will analyze evidence from these following similar pieces of corroborating evidence. For example, Google almost published data scans of chest X-rays, however they were held back since they would be violating the privacy rights of the patients. Finally, Microsoft recently pulled down a dataset that contained images of over 10 millions individuals when they realized that they had images of people without their consent (Wiggers, 2019).

## Conclusion

Machine learning is a research technique that is on the rise in terms of acceptance into society as a profound technique used to boost the computing and research powers of mankind. As this technology becomes increasingly common in consumer facing products, we need to examine the implications of such technology from a safety and security perspective. Such complications could occur in the form of data leakage of sensitive information or exposition of certain private and proprietary algorithms or neural net structures. As companies and organizations increase the sizes of their data lakes, the consumer will be trading their privacy for personalization in exchange for the service. There are technical and social aspects that need to be considered when handling this exchange. By experimenting with the liberty of choosing different clipping thresholds and noise scales for each layer of the network, we can use this insight to design privacy-preserving machine learning implementations without simply choosing arbitrary values to achieve the desired results. Finally, as seen in the case with the the Royal Free London NHS Foundation Trust delivering information on over 1.6 million hospital patients to DeepMind, Google's AI, without the patients' consent, my proposed solution of coupling research in government regulations with educating the general population about the dangers of artificial intelligence and what they are consenting to when they are using such technologies can be used a method to combat the social issue caused by the rise of machine learning.

## References


Anonymous. (2020). Personal Communication.

Abadi, M., Chu, A., Goodfellow, I., McMahan, H., Mironov, I., Talwar, k., & Zhang, L., (2016). *Deep learning with differential privacy* [Eds.]. Retrieved from https://arxiv.org/abs/1607.00133

Chaudhuri, K., Monteleoni, C., & Sarwate, A., (2011). *Differentially private Empirical Risk Minimization* [Eds.]. Retrieved from https://www.jmlr.org/papers/volume12/chaudhuri11a/chaudhuri11a.pdf

Cressman, D., (2009.) *A Brief Overview of Actor-Network Theory: Punctualization, Heterogeneous Engineering & Translation* [Article.]. Retrieved from https://summit.sfu.ca/item/13593

Cristofaro, E., (2020.) *An Overview of Privacy in Machine Learning* [Eds.]. Retrieved from https://arxiv.org/abs/2005.08679

Jayaraman, B., Evans, B., (2019). *Evaluating Differentially Private Machine Learning in Practice* [Eds.]. Retrieved from https://arxiv.org/pdf/1902.08874.pdf

Jayaraman, B., (2020). Personal Communication.

Kaggle, Inc. (2014). *Acquire Valued Shoppers Challenge* [Data]. Retrieved from https://kaggle.com/c/acquire-valued-shoppers-challenge/data

Krizhevsky, A., (2009). *Learning multiple layers of features from tiny images* [Eds.]. Retrieved from https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf

Prabhu, M., (2019.). *Security & Privacy considerations in Artificial Intelligence & Machine Learning — Part-6: Up close with Privacy* [Article.]. Retrieved from https://towardsdatascience.com/security-privacy-in-artificial-intelligence-machine-learning-part-6-up-close-with-privacy-3ae5334d4d4b

Roy, S., (2020.). *Why do we need Privacy-Preserving Machine Learning?* [Article.]. Retrieved from https://towardsdatascience.com/why-do-we-need-privacy-preserving-machine-learning-7480ddf9f114

Wiggers, K., (2019.) *AI has a privacy problem, but these techniques could fix it* [Article.]. Retrieved from https://venturebeat.com/2019/12/21/ai-has-a-privacy-problem-but-these-techniques-could-fix-it/