Journal & Proceedings of the Royal Society of New South Wales, vol. 158, part 2, 2025, pp. 276–278. ISSN 0035-9173/25/020276-03

Thesis abstract

On the security and robustness of federated learning with application to smart grid infrastructures

Cody Lewis

Abstract of a thesis for a Doctorate of Philosophy submitted to
The University of Newcastle, Callaghan, Australia

In the past two decades, machine learning has fast emerged as a popular approach to solving many high-dimensional problems. It focuses on analysing and interpreting patterns and structures in data to enable learning, reasoning, and decision making. A major benefit of machine learning arises from its ability to produce a model or models that can be applied to solve a wide range of problems. For example, machine learning is commonly used in computer vision systems to detect various objects within images, such as traffic signs at a roadside, without having a developer write code that accounts for every possible way that object can be present within an image.

However, machine learning is known to be "data hungry" in that it requires vast datasets with significant amount of variation, to produce accurate models and results. Hence, sourcing of this data can have major implications, especially when it comes to private data pertaining to people. This led to the development of federated learning, which is a form of distributed machine learning across many clients who hold their own independent data that is not shared with the central machine learning model. The clients each train a copy of the machine learning model on their own dataset and upload the resulting trained model to a central server. The server aggregates the client models together to produce a new global model, which is sent back to the clients for the next round of training. The federated learning algorithm aims to maintain data privacy by replacing the requirement that distributed learning needs to have the clients share data, instead tasking them with sharing the model.

Despite its improvements to privacy, federated learning still has several challenges when it comes to security and robustness. In this thesis, we make several theoretical and analytical contributions to the challenges of robustness, privacy and fairness and their combined effects in the federated learning setting.

Robustness is concerned with the ability of the federated learning system to be resilient in the face of attacks or faults. A major class of attacks in federated learning arises due to poisoning, where the clients can intentionally train their model incorrectly (e.g. train a car vision system to recognise stop signs as 40 km/h) and upload the poisoned model to the server to corrupt the global model. We have considered opportunistic adversaries timing their poisoning attacks to undermine robustness mechanisms. We propose the "on-off" attack, where adversaries switch between submitting honest and poisoned updates to fool the stateful variant of robustness mechanisms that are

often found in federated learning defence systems (FLDS). We have also proposed good- and bad-mouthing poisoning attacks, where adversaries submit copies (or negated copies) of the target client's update. Such attacks can fool both stateful and stateless robust aggregation systems to either become biased towards the victim or away from the victim. We then propose a robust aggregation algorithm, which helps to mitigate the on-off attack by monitoring the consistency of client updates and punishing those who make sudden and vast changes. Our proposed robust aggregation algorithm simultaneously mitigates the good and bad-mouthing attacks by ensuring that each overly similar update is diminished in influence upon aggregation, to the point that their sum is effectively equivalent to a single update. We analyse our proposed attacks and mitigation strategy, first from a theoretical point of view, and then from an empirical perspective, showing their effectiveness.

The emergence of *gradient inversion attacks* has posed a significant issue to the *privacy* of federated learning. Gradient inversion attacks are often initiated by the server, which observes the updates provided by the clients, and, as a background process, performs a search for the minibatch dataset that produces the same update. If this task is completed successfully, then the server would have violated privacy by recreating the data that the client was not sharing. Recently, gradient inversion attacks have seen many improvements in their effectiveness, for instance, when inverting larger minibatch sizes in the training process and when clients perform a greater number of steps of training. Though some mitigation strategies exist for these attacks, they tend to have significant overheads in terms of network structure, performance trade-offs, and computation time. Moreover, these techniques fail to consider the more recent advancements to gradient inversion attacks. We propose a secure adaptation of the client-side Adam optimisation algorithm to mitigate gradient inversion attacks, with a particular focus on attacks that become more effective as the learning progresses. Our algorithm better mitigates inversion attacks by ensuring that there are always many overlapping samples representative of each class, regardless of the minibatch size, thus confusing recent gradient inversion attacks since they rely on single sample representations for each class. Simultaneously, our algorithm improves the model performance, instead of hindering, unlike many other privacy preserving mechanisms. We prove that this technique prevents inversion attacks and converges effectively, through theoretical and empirical analysis.

Fairness in federated learning tends to have a basis on the equal opportunity to contribute to the global model. We consider the issue of *fairness in the context of device heterogeneity*. Where, to be fair to the clients, the system would have to ensure that it has nearly equal performance with respect to devices with different computational capabilities. Though there have been some prior works addressing this issue, we found that they do not attempt to maximise fairness. For this reason, we propose a federated learning framework that maximises equal opportunity fairness through a game theoretical analysis of the synchronous federated learning setting where clients can train subsets of the global model. We provide theoretical analysis and proofs of our proposed scheme. Additionally, we demonstrate the effective-

ness of our proposed scheme using empirical studies. Furthermore, we address the issue of privacy in conjunction with the issue of fairness under device heterogeneity. Solutions to device heterogeneity often require server-side analysis of submitted updates, directly opposing the privacy requirements in federated learning. This motivated us to propose an algorithm that enables secure aggregation while providing fairness with device heterogeneity. This algorithm considers device heterogeneity fairness through model partitioning and then combines this with secure aggregation whereby the clients pad and encrypt their updates to be uploaded. These updates are made to be of the same size as the global model. The server aggregates the gradients forming the sum of gradients that is sent back to the clients. Note that the server is unable to correctly average the parameters under the private setting, as it is not aware of how many clients have updated each parameter. We demonstrate both empirically and theoretically that the proposed scheme has little to no negative impact on performance, and in some cases even improves the performance.

Finally, we apply the techniques that we have developed for robustness, privacy, and fairness to *smart grid infrastructures*. We propose a hierarchical federated learning-based framework for smart grid infrastructure and demonstrate how it improves client dropout and poisoning robustness, using relatively lightweight models suitable for devices with limited computational capability. We provide theoretical justification underlying our design and compare our framework and algorithms empirically with existing relevant works.

Dr. Cody Lewis
School of Information and Physical Sciences
The University of Newcastle
Callaghan NSW 2308
Australia

E-mail: hello@codymlewis.com

URL: https://hdl.handle.net/1959.13/1517607