Amortized Robustness in Federated Learning

Abstract. Federated learning (FL) enables collaborative model training without centralizing user data, addressing privacy and computational challenges associated with data centralization. However, it remains highly vulnerable to adversarial participants, a.k.a., Byzantine clients, injecting poisoned updates that leads to model corruption. Existing defenses such as robust aggregation often significantly harms the accuracy and efficiency of the final model. This project aims to develop novel methods that enhance robustness against Byzantine clients while minimizing accuracy loss and performance trade-offs. By revisiting existing threat models and adopting more realistic adversarial scenarios, we seek to refine security guarantees and start adapting these techniques to next-generation AI models, such as large language models.

I. Introduction to Federated Learning & Robustness

Machine Learning (ML) is at the core of many data-oriented high-stake applications such as medicine, finance or recommendation algorithms. The recent success of ML essentially relies on the advent of a new generation of models that are trained using an ever increasing amount of data and computational resources. The dominant paradigm in ML consists in collecting user-generated data and centralizing them on a single machine (or a data-center) to train the model. While this approach is very useful for training highly accurate models, it raises serious privacy concerns regarding the utilization of the users' data. Accordingly, this kind of methodology cannot be used for critical applications where data is too sensitive to be shared, e.g. medical data collected by several hospitals. Besides, collecting data and learning the model centrally incurs overwhelming computational costs, especially since the size of the models has dramatically increased in recent years (now often exceeding 10^{12} parameters). To circumvent these limitations, federated learning (FL) aims to have multiple users collaboratively train a model while ensuring that their data remains on their personal devices [6]. This scheme distributes the computational burden to the local devices, while allowing users to retain control over their respective data. Even though FL is a very promising paradigm, it comes with its own challenges and vulnerabilities. In this project, we focus on one of the most pressing issues in FL, namely preserving *robustness* to malicious devices.

Robustness in Federated Learning

In FL, robustness issues can take various forms. In this project, we focus on inaccuracies in model predictions caused by corrupted (or misbehaving) clients. When deploying FL in real world environments, it is almost inevitable to encounter users that deviate from the prescribed learning algorithm. These deviations may stem from software or hardware bugs, poisoned or biased data, or even malicious adversaries attempting to manipulate the learning process by sending erroneous information. Such misbehaving devices are often referred to as *Byzantine*, by analogy with the similar problem in distributed computing [8]. Unfortunately, standard FL solutions fail to guarantee delivery of accurate models in such an adversarial setting. Indeed, if left unprotected, the learning procedure can be critically influenced by a handful of Byzantine devices [5]. This vulnerability may lead to severe societal repercussions if the resulting model is used in a high-stake application. For example, the enormous influence of today's recommendation algorithms could attract malicious political and private entities willing to bias the recommendations made to billions of users on a daily basis. To safely harness the potential of FL in real-world applications, it is therefore imperative to ensure the robustness of these solutions against such Byzantine actors.

Specifically, suppose the system comprises n clients represented by a set $[n] = \{1, \dots, n\}$, where for any given model θ each client i incurs a loss given by $\mathcal{L}_i(\theta)$. The local loss of each client is measured on its local dataset. In the classic FL setting the server aims to compute a model θ^* that minimizes the global average of all the clients' loss functions, i.e., $\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_i(\theta)$. However, in robust FL we assume that up to f clients are

adversarial and may send arbitrarily incorrect information (about their local loss functions) to the server. Let us denote the set of non-adversarial agents by $\mathcal{H} \subset [n]$ with $|\mathcal{H}| \geq n - f$. Thus, the objective of the server in robust FL reduces to minimizing the average of the honest clients' loss functions, i.e., $\frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \mathcal{L}_i(\theta)$. The main challenge in robust FL arises from the fact that the identity of $\mathcal{H}$ is apriori unknown to the server.

II. Prior Work on Byzantine-Robust FL

Numerous solutions have been proposed to address the aforementioned issue of Byzantine-robustness in FL (see, e.g., [5]). Most prior work has, until recently, assumed that the data held by machines is homogeneous (an assumption that rarely holds in practical settings). In fact, FL is inherently heterogeneous, as clients/users retain their data locally. Then, each client holds only a portion of the dataset that may not accurately represent the entire population. This disparity between theoretical assumptions and real-world conditions can be partially attributed to the fact that data homogeneity simplifies the problem of Byzantine-robustness. When users have homogeneous data, their local computational patterns are similar, making it easier to detect harmful erroneous information from Byzantine users, thereby limiting their disruptive potential. In contrast, when datasets are heterogeneous, Byzantine devices can exploit this variability to "hide in the crowd" and interfere with the learning process more effectively. This challenge has led recent research to establish theoretical limits on Byzantine robustness in FL [2, 3, 7], demonstrating that robustness is inherently constrained by data heterogeneity. In short, **achieving robustness often comes at the cost of very reduced model accuracy.**

Key Limitations

Prior work assumes a threat model wherein misbehaving clients can behave arbitrarily (i.e., can send malicious information to the server) and fully exploit the data heterogeneity across honest clients. While this modeling has proved useful in deriving foundational theorems on robustness in FL (e.g., see [5]), it leads to overly pessimistic conclusions about the impact of corrupted clients on learning accuracies. These negative results are mainly artifacts of the definition of data heterogeneity, combined with an unrealistically strong adversarial model. We believe that in practice weaker threat models, wherein the behavior of corrupted clients is constrained, are more reasonable. For instance, in FL for medical applications, it is natural to assume that clients (such as hospitals, clinics, or pharmacies) are honest by intention. In such cases, misbehavior is more likely to stem from inadvertent errors, such as mislabeled data, missing data, or biases within datasets, rather than from worst-case arbitrary actions. Moreover, the likelihood of misbehavior need not be uniform across all the clients. Some clients may indeed be more prone to corruption than others, depending upon their sources of data, ethical values and computational infrastructure.

III. Project Task: From Worst-Case to Expected Robustness

Consider the FL setting described above with n clients, where each client i has a local loss function $\mathcal{L}_i(\theta)$ for a global model θ. Since prior work assumes that any subset of clients $\mathcal{B} \subset [n]$ of size at most f can be corrupted, traditional robust FL algorithms aim to solve the following minimax optimization problem:

$$\underset{\theta}{\text{Minimize}} \max_{S \subset [n], |S| = n-f} \frac{1}{|S|} \sum_{i \in S} \mathcal{L}_i(\theta). \tag{1}$$

This particular learning objective aims to characterize the worst-case robustness, i.e., for any possible identity of misbehaving clients. However, it can also yield highly conservative robustness guarantees, especially in scenarios when the clients' behavior (being honest or adversarial) is correlated. In other words, honesty of any client is associated with the honesty of another. This is often true in settings wherein some clients may share common or similar data sources and ethical values. Moreover, the above minimax optimization formulation paints a pessimistic picture about robustness in cases when data heterogeneity across different subsets of clients need not be identical. By adapting the learning objective to these scenarios, this project aims to bridge the existing gap between theory and practice in robust FL.

Introduction to Expected Robustness

In this sub-task, we consider a new learning objective of minimizing the expected loss, defined over the probability distribution of honest subsets. Specifically, let $\mu_{\mathcal{H}}$ denote the probability distribution over the subsets of $[n]$ for being the set of honest clients. Then, the new robust FL goal is to solve the following problem:

$$\underset{\theta}{\text{Minimize}} \ \mathop{\mathbb{E}}_{S \sim \mu_{\mathcal{H}}} \left[\frac{1}{|S|} \sum_{i \in S} \mathcal{L}_i(\theta) \right], \tag{2}$$

While the optimal value of the above optimization problem is in general lower than that of the classic minimax optimization problem (1), the expected robustness guarantee in (2) can be significantly better for several existing robust FL algorithms.

Project Objectives

The objectives of this project can be summarized as follows:

1. Revisit the theoretical robustness guarantees of existing state-of-the-art (SOTA) robust FL algorithms, such as Robust DGD with nearest neighbor mixing (NNM) [2] or bucketing [7], through the lens of expected learning loss defined in (2). For this analysis, we will consider different ways of characterizing the probability distribution $\mu_{\mathcal{H}}$, e.g., modeling correlation between clients' behavior by means of an undirected graph where edges between clients indicate closeness in ethical values, geographical location and data population.
2. Empirically evaluate the tightness of the theoretical bounds for expected robustness by simulating various heterogeneity scenarios, e.g., invoking Dirichlet distribution, and different forms of adversarial behaviors, e.g., label-flipping [1] and a little is enough (ALIE) [4].
3. **(Ambitious)** If the SOTA robustness schemes are identified to be sub-optimal, then design a new robustness algorithm that yield optimal guarantees in this setting.

IV. Practical information

Organization: The internship will last from April to September 2026 at the LPSM (Laboratoire de Probabilités, Statistique et Modélisation) at Sorbonne University, funded by a grant from LPSM, with the possibility of continuing as a PhD student after the internship. The supervising team is constituted of -

- Rafael Pinot (Sorbonne University)
- Nirupam Gupta (University of Copenhagen)
- Aurélien Bellet (Inria Montpellier)

While the main location will remain at LPSM, visit to Copenhagen or Montpellier may be considered during the internship (and the following PhD).

Application: Strong M2 level trainee in statistics/machine learning/optimization with python programming skills are required. Applicants should send a CV and transcripts of the last two years to pinot@lpsm.paris, nigu@di.ku.dk, and aurelien.bellet@inria.fr

V. References related to the project

[1] Dan Alistarh et al. "Byzantine stochastic gradient descent". In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. 2018.

[2] Youssef Allouah et al. “Fixing by Mixing: A Recipe for Optimal Byzantine ML under Heterogeneity”. In: *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*. Vol. 206. Proceedings of Machine Learning Research. PMLR, 2023.

[3] Youssef Allouah et al. “Robust Distributed Learning: Tight Error Bounds and Breakdown Point under Data Heterogeneity”. In: *The 37th Conference on Neural Information Processing Systems [Spotlight]*. 2023.

[4] Moran Baruch et al. “A Little Is Enough: Circumventing Defenses For Distributed Learning”. In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems*. 2019.

[5] Rachid Guerraoui et al. *Robust Machine Learning*. Springer, 2024.

[6] Peter Kairouz et al. “Advances and Open Problems in Federated Learning”. In: *Foundations and Trends® in Machine Learning* 14.1–2 (2021).

[7] Sai Praneeth Karimireddy et al. “Byzantine-Robust Learning on Heterogeneous Datasets via Bucketing”. In: *International Conference on Learning Representations*. 2021.

[8] Leslie Lamport et al. “The Byzantine Generals Problem”. In: *ACM Trans. Program. Lang. Syst.* 4.3 (1982).