PROPOSITION SUJET DE THÈSE
CONTRATS DOCTORAUX - 2023-2026

X Contrat doctoral fléché FR Agorantic
Contrat doctoral fléché EUR InterMEDIUS

Pour candidater sur ce sujet, les personnes intéressées doivent contacter le.la directeur.rice de thèse avant le **2 juin 2023.**

Les auditions des candidats retenus auront lieu début juillet.

Directeur.rice de thèse :	**Rachid Elazouzi**	**Mail :**	**rachid.elazouzi@univ-avignon.fr**
Laboratoire :	**LIA**	**Téléphone :**	**+33490843562**
Co-directeur.rice et/ou encadrant.e :	**Anna Melnykova et Pierre-Henri Morand**		
Laboratoires :	**LMA et LBNC**		

Titre en français : apprenTissage fédéRé poUr deS héTérogènes et sensibles (TRUST)

Titre en anglais : Federated learning for highly heterogeneous and sensitive data.

Abstract: This thesis deals with challenges of using data in machine learning algorithms in a way that respects the privacy and confidentiality of each user. One of the statistical learning approaches which respects the General Data Protection Regulation (GDPR) is a family of federated learning algorithms (FL), which trains the model on the user machines, without transmitting sensitive data to the server. However, FL is always confronted with a problem of statistical heterogeneity, which can result in less performant models. The goal of this PhD thesis is to study the impact of statistical heterogeneity on FL and will propose efficient solutions to improve its performance while maintaining data confidentiality. The open databases hosted in the FR Agorantic can be used as a field of application of the work of the thesis.

Keywords: Federated Learning, Statistical heterogeneity, GDPR, Data protection, Game Theory

1- Detailed description of the subject

There is a growing interest in a new distributed Machine Learning approach called Federated Learning (FL) [La17]. The main principle of FL is that clients compute their local gradients and communicate them to a central server. Then, this centralized server orchestrates learning cycles on large volumes of data, which are created and stored locally in a large number of clients. This learning procedure is repeated until a certain criterion is reached. This allows participating customers to protect their data and address data security and privacy issues mandated by law. Indeed, as several data protection authorities like GDPR have already suggested, FL has the potential to facilitate compliance with the principle of data privacy. This is especially important in health or political applications where the data are full of personal and

highly sensitive information, and where data analysis methods likely need to comply with regulatory guidelines. It is therefore a real challenge to develop highly accurate FL tools while complying with privacy regulations. Since the first proposal of FL by Google AI in 2017, many efforts have recently been devoted to the development of FL's algorithms to build machine learning models with different privacy preservation approaches. In addition, powerful GPUs are becoming more accessible, allowing larger models to be deployed, accelerating FL deployment.

This growing demand for FL technology will open new challenges in addition to those appearing in traditional ML. These include, for example, the extension of the learning period due to the correlation between customer databases. More generally, federated learning still faces the challenges of statistical heterogeneity. Statistical heterogeneity results from non-IID data generated by different clients, which have different characteristics or probability distributions of labels [Kari19, Li20, Wang20]. It is proven to have a negative impact on the convergence and accuracy of the model compared to homogeneous data (independent and identically distributed). For example, if the customer data is heavy-tailed and heterogeneous at the same time, the common problem becomes complicated and difficult because each customer may hold different tail classes [Li22].

In the thesis, we will study several avenues to achieve our objectives:

(i) quantitatively study the impact of statistical heterogeneity on federated learning,

(ii) define metrics (eg Kullback-Leibler, or Wasserstein distance) allowing to evaluate the statistical heterogeneity of the data while ensuring the confidentiality of the data,

(iii) develop a strategy to select customers during training using the metric that measures the statistical heterogeneity of each customer, which can effectively limit model divergence during federated learning.

On the other hand, we will always have the choice between the effectiveness of the proposed methods and the risk of confidentiality of the data. One of the approaches that will be explored in this thesis is the Bayesian neural network, introduced in FL to solve the problem of model overfitting by representing all network parameters in the global model with probability distributions [Chen21].

The subject of this thesis is by construction interdisciplinary. It requires triple expertise in computer science, mathematics and law. IT brings all the expertise on federated learning. Mathematics will provide a theoretical analysis on metrics to be explored to meet the challenges related to data heterogeneity. The law plays an important role on the aspect related to the constraints imposed by the EU GDPR. Finally, we plan to set up Federated Learning using several databases provided by Agorantic Research Federation to test our solutions under the constraints of the GDPR.

2- Candidate requirements

- The candidate will have a Master or Engineering Degree with a Major in Informatics, Applied Mathematics, Statistics or Data Science
- She or he will be interested in and capable of developing machine learning models, and be at ease with statistics and Data Analysis
- She or he will have an interest in interdisciplinary research and possess good programming skills (for example, Python, R or similar)
- Good English level (written and spoken) is required. French would be a plus.

3- International mobility opportunities for doctoral students

PhD candidate will work in LIA (Laboratoire Informatique d'Avignon) in a close collaboration with mathematics and LBNC labs. The candidate will be provided all the necessary technical equipment to conduct the experiments (laptop, screens access to computational server), as well as funding for conferences and research stays. In particular, within a framework of ongoing collaboration with Carnegie Mellon University CyLab, the candidate will be proposed to do a 4 month stay in Carnegie Mellon University.

4- Références bibliographiques

[La17] P. Kairouz, M. Bennis, et al. “Advances and Open Problems in Federated Learning,” https://arxiv.org/pdf/1912.04977.

[Chen21] Chen, H.-Y. and Chao, W.-L. FedBE: Making Bayesian model ensemble applicable to federated learning. In Inter-national Conference on Learning Representations, 2021.

[**Wang20**] Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. Advances in neural information processing systems, 33:7611–7623, 2020.

[Li20] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. Proceedings of Machine Learning and Systems, 2:429–450, 2020.

[Kari19] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J Reddi, Sebastian U Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for on-device federated learning. 2019.

[Li22] Qinbin Li, Yiqun Diao, Quan Chen, and Bingsheng He. Federated learning on non-iid data silos: An experimental study. In 2022 IEEE 38th International Conference on Data Engineering (ICDE), pages 965–978. IEEE, 2022.