2026-2027
Sujet de stage M1/M2
Titre: Apprentissage fédéré

Contexte

L'apprentissage fédéré consiste à entraîner un modèle partagé sur de nombreux clients sans centraliser leurs données. Pourtant, le choix du mode de fédération (règle d'agrégation, échantillonnage des clients, nombre de pas locaux, schéma de pondération) repose encore largement sur des heuristiques, et l'on ne sait pas clairement quel choix donne la convergence la plus rapide dans un contexte donné [3, 4, 1, 2, 5]. Lorsque les données des clients sont hétérogènes, déséquilibrées, bruitées, voire corrompues, certains clients aident bien davantage le modèle global que d'autres, et le serveur n'a aucun moyen direct de les identifier.

Ce stage abordera deux questions liées. Premièrement, comment le serveur peut-il estimer la qualité et l'utilité de la contribution de chaque client (à partir des mises à jour, des gradients ou du comportement de la perte, sans accès aux données brutes) et utiliser cette estimation pour sélectionner ou pondérer les clients ? Deuxièmement, quelle est la signification mathématique des modes d'agrégation usuels (FedAvg et ses variantes pondérées, FedProx, optimiseurs adaptatifs côté serveur, agrégateurs robustes comme la médiane ou la moyenne tronquée) ? En particulier, se comportent-ils comme l'optimisation que l'on mènerait si toutes les données étaient réunies, ou introduisent-ils des biais tels que la dérive des clients (« client drift ») lorsque les distributions de données ne sont pas identiques ?

L'analyse mathématique constitue la partie la plus importante de ce stage : les expériences serviront principalement à illustrer, tester et motiver les résultats théoriques, plutôt qu'à constituer une fin en soi.

Objectifs

Le ou la stagiaire commencera par une revue de la littérature sur l'optimisation fédérée, la sélection de clients et les analyses de convergence. Il ou elle formalisera ensuite la comparaison entre l'agrégation fédérée et l'entraînement centralisé, en identifiant les conditions sous lesquelles les deux coïncident (par exemple, des distributions de données identiques ou un unique pas local) et en quantifiant l'écart dans les autres cas.

Sur le plan expérimental, le ou la stagiaire construira un cadre de simulation à hétérogénéité contrôlée (déséquilibre des classes, déséquilibre des quantités, clients bruités ou adversaires) afin de comparer les stratégies d'agrégation et de sélection de clients en termes de vitesse de convergence, de précision finale et de robustesse, et de tester des critères proposés pour détecter les clients précieux. Les résultats attendus sont un banc d'essai reproductible, une vision théorique et empirique claire de ce que chaque mode d'agrégation optimise, ainsi que des recommandations pratiques ou une nouvelle stratégie de sélection, pouvant donner lieu à une publication. Le profil recherché est celui d'un ou d'une étudiante de M2 (ou de M1 avancé) en apprentissage automatique, mathématiques appliquées ou informatique, avec de solides bases en optimisation et en probabilités et à l'aise avec Python/PyTorch.

Les travaux du/de la stagiaire pourront donner lieu à des publications dans des conférences et revues nationales et internationales, ainsi qu'à des présentations lors des séminaires de l'équipe. La soutenance

de stage aura lieu à l'institution d'origine de l'étudiant(e), avec une présence des encadrant(e)s de stage en visioconférence. Les encadrant(e)s américain(e)s délivreront un certificat de réussite de stage détaillé, ainsi que des lettres de recommandation pour la suite des études du/de la stagiaire. Les stages M2 peuvent donner lieu à des propositions de continuation en thèse, sous condition d'adéquation des compétences, des financements disponibles et des règles d'admission de l'École doctorale locale.

Etablissement d'accueil

L'Université du Texas à El Paso (UTEP) est une université publique américaine classée **R1** (à activité de recherche de haute catégorie) par la classification Carnegie, qui distingue les établissements ayant une très forte activité de recherche. Située au Texas, à la frontière entre les États-Unis et le Mexique, elle accueille plus de 26000 étudiants et propose 169 formations de licence, de master et de doctorat. Elle se distingue par son engagement en faveur de l'égalité des chances et de l'accès à un enseignement supérieur de qualité. L'UTEP figure parmi les 5% d'universités américaines les mieux classées en matière de recherche. Son département d'informatique, rattaché au *Miguel A. Loya College of Engineering*, propose des formations en informatique du niveau licence au doctorat, et collabore avec des partenaires académiques et industriels. Ses principaux axes de recherche couvrent le génie logiciel, la cybersécurité, l'intelligence artificielle et les fondements théoriques de l'informatique.

Visa et mobilité

La réalisation du stage à l'Université du Texas à El Paso (UTEP) nécessite l'obtention d'un visa J-1. L'établissement pourra accompagner le/la candidat(e) dans ses démarches administratives, mais les frais liés à la demande de visa et aux formalités associées resteront à sa charge. L'accueil en stage est conditionné à l'obtention du visa requis.

Durée, rémunération et frais de scolarité

Le stage, d'une durée de quatre à six mois, n'est pas rémunéré. Aucun frais de scolarité (tuition fees) ne sera demandé par l'UTEP au stagiaire dans le cadre de ce stage, et une attestation de stage sera délivrée à l'issue de ce dernier (cf. ci-dessus). Les encadrant(e)s aideront le/la stagiaire pour toute démarche pour obtenir une bourse de mobilité internationale.

Encadrement

Le (la) stagiaire sera encadré(e) par :

- Martine Ceberio – Professeure en Informatique à l'UTEP (`https://expertise.utep.edu/profiles/mceberio`)
- Christoph Lauter – Maître de conférences en informatique (Associate Professor), HDR, à l'Université de l'UTEP (`https://www.christoph-lauter.org/`)
- Hanane Benmaghnia – Post-Doctorante en validation numérique au LIP6, Sorbonne Université – CNRS, future Chargée de recherche à l'UTEP, (`https://hanane-benmaghnia.github.io/`)

Profil recherché

Le stage s'adresse à des étudiant(e)s (M1 ou M2) en informatique, mathématiques appliquées, électronique ou dans un domaine connexe en premier lieu, mais est ouvert aussi aux étudiant(e)s d'autres disciplines

s'ils/elles peuvent démontrer une forte capacité de travail interdisciplinaire. Les compétences recherchées sont :

- Modélisation de problèmes découlant des différentes disciplines des science à l'aide d'outils mathématiques et informatiques.
- Compréhension générale des approches modernes des sciences des données et de l'Intelligence Artificielle, y compris pour les modèles moins repandus commes les modèles de Kolmogorov-Arnold.
- Curiosité pour aborder une problématique de recherche dans un contexte international, et ouverture d'esprit pour la découverte de la langue anglaise, utilisée dans un contexte universitaire de recherche.
- Compétences basique en programmation et en génie logiciel basique.

La curiosité scientifique, l'autonomie, la rigueur et la capacité à analyser des résultats expérimentaux seront particulièrement appréciées.

Modalités de candidature

Les candidatures doivent être envoyées à Martine Ceberio (mceberio@utep.edu), Christoph Lauter (cqlauter@utep.edu) et Hanane Benmaghnia (hanane.benmaghnia@lip6.fr). Elles doivent inclure :

- un curriculum vitae,
- une lettre de motivation,
- des liens vers des projets personnels, professionnels ou universitaires représentatifs, par exemple des dépôts Git (si disponibles) ou publications antérieures.
- une ou deux personnes référentes à contacter.

References

[1] Yae Jee Cho, Jianyu Wang, and Gauri Joshi. Client selection in federated learning: Convergence analysis and power-of-choice selection strategies. *arXiv preprint arXiv:2010.01243*, 2020.

[2] Tiansheng Huang, Weiwei Lin, Li Shen, Keqin Li, and Albert Y Zomaya. Stochastic client selection for federated learning with volatile clients. *IEEE Internet of Things Journal*, 9(20):20055–20070, 2022.

[3] Jian Li, Tongbao Chen, and Shaohua Teng. A comprehensive survey on client selection strategies in federated learning. *Computer Networks*, 251:110663, 2024.

[4] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189*, 2019.

[5] Yuxin Shi, Zelei Liu, Zhuan Shi, and Han Yu. Fairness-aware client selection for federated learning. In *2023 IEEE international conference on multimedia and expo (ICME)*, pages 324–329. IEEE, 2023.