

Apprentissage fédéré non-IID : méthodes d’optimisation pour une sélection efficace des clients

Direction : Sana Ben Hamida, Sonia Guehis, Inès Alaya
LAMSADE – Université Paris Dauphine PSL

Contexte

L’apprentissage fédéré (*Federated Learning*, FL) [1], [2] permet d’entraîner des modèles d’apprentissage automatique de manière distribuée sans centraliser les données, répondant ainsi aux contraintes croissantes de confidentialité et de souveraineté. Cette approche est aujourd’hui utilisée dans de nombreux domaines sensibles (santé, finance, IoT, edge computing).

Cependant, les déploiements réels à grande échelle font face à deux difficultés majeures :

- **Hétérogénéité statistique (non-IID)** : les clients possèdent des distributions de données différentes, ce qui dégrade la convergence et la qualité du modèle global [3], [4].
- **Hétérogénéité système** : disponibilité intermittente, latence variable et ressources limitées, augmentant fortement le temps d’entraînement réel.

Dans ces conditions, sélectionner efficacement quels clients participent à chaque ronde devient un levier clé pour réduire le *time-to-accuracy*. Des travaux récents montrent l’intérêt de stratégies guidées [5] ou d’échantillonnage optimal [6], mais la sélection reste largement heuristique et sans garanties théoriques sous non-IID sévère.

Objectif de la thèse

Cette thèse vise à développer un cadre d’optimisation combinatoire pour la sélection adaptative de clients en apprentissage fédéré non-IID, et à concevoir des métaheuristiques hybrides (algorithmes génétiques, colonies de fourmis, approches mémétiques) capables de produire des sélections efficaces à grande échelle.

L’idée centrale est de formuler la sélection de clients comme un problème multi-objectif combinant :

- qualité statistique des mises à jour (variance/biais des gradients),
- coûts système (latence, communication, énergie),
- réduction du temps d’entraînement global.

Les métaheuristiques seront utilisées comme méthodes d’approximation scalables pour explorer cet espace combinatoire de très grande dimension, éventuellement guidées par des modèles prédictifs appris à partir des historiques d’exécution.

Axes de recherche

La thèse s’articulera autour des axes suivants :

1. **Analyse du non-IID** : modélisation de l’hétérogénéité inter-clients et étude de son impact sur la convergence des algorithmes fédérés [7].

2. **Formulation optimisationnelle** : définition d'un problème de sélection multi-objectif sous contraintes de ressources.
3. **Conception de métaheuristiques** : développement d'algorithmes génétiques, colonies de fourmis et variantes hybrides dédiées au FL.
4. **Hybridation apprentissage–optimisation** : intégration de prédicteurs d'utilité/coût pour guider la recherche.
5. **Validation expérimentale à grande échelle** : évaluation sur des plateformes et benchmarks de référence (LEAF, FedScale) [8], [9].

Contributions attendues

- formalisation théorique du problème de sélection de clients sous non-IID,
- nouveaux algorithmes métaheuristiques adaptés au FL,
- réduction mesurable du *time-to-accuracy* par rapport aux stratégies existantes (sélection aléatoire, heuristiques guidées, importance sampling),
- publications dans des conférences/journaux en apprentissage automatique, optimisation et systèmes distribués,
- implémentations open-source reproductibles.

Environnement expérimental

Les développements s'appuieront sur des plateformes modernes de simulation et d'évaluation du FL telles que :

- LEAF (benchmarks fédérés réalistes) [8],
- FedScale (évaluation conjointe modèle/système à grande échelle) [9].

Profil recherché

Nous recherchons un(e) candidat(e) motivé(e) avec :

- solide formation en machine learning / optimisation / statistiques,
- bonnes bases en Python (PyTorch/TensorFlow),
- intérêt pour les systèmes distribués ou l'edge computing,
- goût pour la recherche expérimentale et/ou théorique.

Une expérience en optimisation combinatoire, métaheuristiques ou apprentissage fédéré est un plus.

1 Contact et candidature

Merci d'envoyer par e-mail :

- CV
- Lettre de motivation
- Relevés de notes
- Éventuellement une lettre de recommandation

sana.mrabet@dauphine.psl.eu

sonia.guehis@dauphine.fr

ines.alaya@parisnanterre.fr

Références clés

- [1] T. Z. SANA et al., “Advancing Federated Learning : A Systematic Literature Review of Methods, Challenges, and Applications,” *IEEE Access*, t. 13, p. 153 817-153 844, 2025. DOI : 10.1109/ACCESS.2025.3605165
- [2] B. YURDEM, M. KUZLU, M. K. GULLU, F. O. CATAK et M. TABASSUM, “Federated learning : Overview, strategies, applications, tools and future directions,” *Heliyon*, t. 10, n° 19, e38137, 2024, ISSN : 2405-8440. DOI : <https://doi.org/10.1016/j.heliyon.2024.e38137> adresse : <https://www.sciencedirect.com/science/article/pii/S2405844024141680>
- [3] K. HSIEH et al., “The Non-IID Data Quagmire of Decentralized Machine Learning,” 2020, arXiv :1910.00189. DOI : 10.48550/arXiv.1910.00189
- [4] S. P. KARIMIREDDY, S. KALE, M. MOHRI, S. REDDI, S. STICH et A. SURESH, “SCAFFOLD : Stochastic Controlled Averaging for Federated Learning,” 2021, arXiv :1910.06378. DOI : 10.48550/arXiv.1910.06378
- [5] F. LAI, Y. DAI, S. SINGAPURAM et al., “Oort : Efficient Federated Learning via Guided Participant Selection,” in *USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, 2021, p. 19-35.
- [6] W. CHEN, S. HORVATH et P. RICHTARIK, “Optimal Client Sampling for Federated Learning,” 2022, arXiv :2010.13723. DOI : 10.48550/arXiv.2010.13723
- [7] P. KAIROUZ, H. B. MCMAHAN et al., “Advances and Open Problems in Federated Learning,” *Foundations and Trends in Machine Learning*, 2021, arXiv :1912.04977. DOI : 10.48550/arXiv.1912.04977
- [8] S. CALDAS et al., “LEAF : A Benchmark for Federated Settings,” 2019, arXiv :1812.01097. DOI : 10.48550/arXiv.1812.01097
- [9] F. LAI et al., “FedScale : Benchmarking Model and System Performance of Federated Learning at Scale,” 2022, arXiv :2105.11367. DOI : 10.48550/arXiv.2105.11367