

Objet / Thèse Méthodes d'apprentissage statistique rapides, sobres et asymptotiquement efficaces / Septembre 2024

L'utilisation des données massives (big-data) pour les collectivités et les industries est un enjeu majeur aujourd'hui pour améliorer leurs services et leurs performances dans un monde en transition avec ses nouvelles contraintes environnementales. Notamment, l'empreinte carbone du numérique et de l'intelligence artificielle que nous sommes amenés à utiliser plus largement est un enjeu majeur pour nos organisations. Par exemple, l'empreinte carbone de l'apprentissage de GPT-3 est estimée à environ 550 tonnes en 2021 [1]. Celle de l'apprentissage de plus petits modèles (pour l'analyse marketing des retours clients dans des sociétés de services) équivaut à l'empreinte carbone de 7 ans d'une personne (en moyenne dans le monde) [2]. Les progrès technologiques permettent de réduire l'impact CO₂, mais les modèles grossissent également en parallèle. Il faut donc pouvoir proposer des méthodologies d'apprentissage statistique sobres pour atteindre l'objectif de neutralité carbone.

Le projet « Méthodes d'apprentissage statistique rapides, sobres et asymptotiquement efficaces » aborde le concept d'efficacité asymptotique, le développement d'outils rapides et sobres d'aide à la décision basés sur ces recherches et leur transfert vers le monde socio-économique. Les méthodologies développées serviront à l'analyse des données énergétiques de Le Mans Métropole dans le cadre du projet POPSU Transitions et des données du programme de recherche « Efficience et Sobriété Numérique » avec la société EREN Groupe (<https://www.eren-groupe.com/fr>) qui développe une offre de services autour de la décarbonation (production d'énergie renouvelable, solution de stockage d'énergie, production d'hydrogène, ...) au travers de ses différentes sociétés.

Grâce à la notion d'efficacité asymptotique dans la théorie de la *normalité locale asymptotique*, il est possible de définir, pour un problème impliquant beaucoup de données, ce qu'est la « meilleure » procédure d'estimation des paramètres d'un modèle ou la « meilleure » décision statistique (tests d'hypothèses). Pour certaines expériences statistiques, les procédures efficaces ne sont pas connues. Pour d'autres expériences, ces procédures efficaces sont connues mais elles sont généralement complexes et très longues à calculer numériquement pour des données massives ou des données acquises à haute-fréquence. Des alternatives doivent donc être trouvées pour gérer des ensembles de données volumineux ou acquis à haute fréquence disponibles dans les différents domaines d'application. Le projet ANR EFFI 2022-2025 (www.ffi-stats.fr) vise à améliorer les connaissances sur l'efficacité asymptotique de plusieurs expériences statistiques et à fournir de nouveaux estimateurs et de nouvelles procédures de test efficaces et innovants pour des applications réelles. Ce projet est également soutenu par la région Pays de la Loire.

Le doctorant s'intéressera à ces développements théoriques sur l'efficacité asymptotique en statistique paramétrique. Notamment, il explorera les méthodes *one-step* (par exemple [3, 4, 5]) pour le machine learning (méthodes de descente de gradient stochastique et extensions) et les méthodes de deep-learning étendant deux travaux récents [6, 7].



Laboratoire Manceau de Mathématiques
<http://lmm.univ-lemans.fr>
contact-lmm@univ-lemans.fr

Le Mans Université
Avenue Olivier Messiaen
72085 - LE MANS cedex 9
univ-lemans.fr

Des compétences de programmation en R et/ou en tensorflow/PyTorch (ou équivalent) seraient particulièrement appréciées. Le doctorant participera également à l'analyse des données énergétiques étudiées dans le cadre de la coopération avec de Le Mans Métropole et la société de décarbonation EREN Groupe.

Laboratoire Manceau de Mathématiques Le Laboratoire Manceau de Mathématiques (LMM) développe à Le Mans Université une recherche en mathématiques théorique et appliquée, principalement à l'assurance et aux énergies. Cette recherche s'articule historiquement autour de deux axes thématiques : probabilités, finance et risques d'une part et statistique des processus et applications d'autre part. Plus récemment, des activités de recherche transverses en actuariat, risque et assurance se sont développées avec la fondation de l'Institut du Risque et de l'Assurance (IRA). Le LMM est membre de la Fédération de recherche CNRS Henri Lebesgue (<https://www.fpl.math.cnrs.fr>) et du LabEx Centre Henri Lebesgue (<https://www.lebesgue.fr/fr>).

Laboratoire d'Innovation Territorial Le Laboratoire d'Innovation Territorial est une démarche collective de réflexion, innovation, expérimentation et diffusion à destination des acteurs publics, économiques et de la recherche initiée par Le Mans Métropole. Il vise à croiser les savoirs scientifiques, expertises opérationnelles et expériences du territoire pour construire des savoirs et expérimenter des solutions en faveur de la neutralité carbone. Dans ce cadre, Le Mans Métropole soutient la recherche à travers son appel à projets Allocations doctorales Neutralité Carbone, et a rejoint avec Le Mans Université le programme national de recherche-action POPSU Transitions pour la période 2023-2026.

Contact : Alexandre.Brouste@univ-lemans.fr

- [1] D. Patterson and al. (2021) Carbon Emissions and Large Neural Network Training, arxiv:2104.10350.
- [2] E. Strubell and al. (2020) Energy and Policy Considerations for Modern Deep Learning Research, Proceedings of the AAAI Conference on Artificial Intelligence Energy and Policy Considerations for Modern Deep Learning Research.
- [3] A. Brouste, M. Soltane and E. Votsi (2020) *One-step estimation for the fractional Gaussian noise model at high-frequency*, ESAIM PS, 24, 827-841.
- [4] A. Brouste, C. Dutang and D. Noutsia Mieniedou (2021) *OneStep - Le Cam's one-step estimation procedure*, R Journal, 13(1), 366-377.
- [5] S. Ben-Hariz, A. Brouste, Y. Esstafa and M. Soltane (2023) *Fast calibration of weak FARIMA models*, ESAIM PS, 27, 156-173.
- [6] A. Brouste, N. Dugué and W. Meskini *Speeding up the training of neural networks with the one-step procedure*, su.
- [7] A. Bensoussan, A. Brouste, Y. Esstafa *Asymptotically efficient one-step corrected stochastic gradient descent for statistical estimation*, in progress.