Post-doctorat : Inférence bayésienne adossée à un simulateur stochastique (ABySS)

Contexte

Dans de nombreuses disciplines scientifiques, de la physique des particules à la cosmologie, en passant par la biologie moléculaire et l'épidémiologie, il est aujourd'hui usuel de développer des outils de simulations complexes afin de décrire des phénomènes d'intérêt. Ces modèles basés sur la simulation sont souvent stochastiques et comportent de multiples paramètres d'entrée. Alors que l'objet premier de la simulation stochastique est de pouvoir générer des données à partir d'une configuration de paramètres (simulation *forward*), son intérêt réside fréquemment dans le problème inverse : déterminer une configuration de paramètres du modèle permettant de générer des données suffisamment proches de celles observées dans la Nature. Or, la résolution de ce problème indirect et non linéaire est en général une tâche ardue. Notre objectif est d'élaborer un cadre rigoureux d'inférence statistique pour l'estimation des paramètres d'entrée d'un simulateur stochastique. En particulier, nous proposons d'adopter le paradigme bayésien pour la résolution du problème inverse afin de caractériser l'ensemble des solutions via leur distribution a posteriori. Cependant, cet objectif se heurte ici à une difficulté fondamentale : on ne dispose pas de l'expression analytique de la vraisemblance dans le contexte de la simulation stochastique. Or, les schémas conventionnels d'exploration de lois *a posteriori* basés sur les techniques Monte Carlo par Chaînes de Markov (MCMC) requièrent cette connaissance analytique.

Inférence sans vraisemblance analytique

Ce défi méthodologique a inspiré une partie de la communauté statistique ces dernières années. C'est d'abord dans les domaines de la génétique des populations et de l'épidémiologie que sont apparues les méthodes *Approximated Bayesian Computations* (ABC) pour des espaces d'observables discrets. L'ABC procède à la génération aléatoire d'un paramètre et ne retient celui-ci que si les observables alors simulées sont égales aux observations expérimentales. Si ce simple schéma peut fonctionner sur des espaces discrets, on comprend qu'il soit gourmand en nombre de simulations *forward* et ne puisse s'appliquer comme tel pour des espaces d'observations continus (physique, cosmologie, économie, *etc*.). Les versions continues de l'ABC introduisent alors une distance permettant de comparer observables simulées et expérimentales et ne retiennent les propositions de paramètres que si les observables simulées sont jugées suffisamment proches des expérimentales. Ceci requiert de fixer un seuil d'acceptation/rejet. Par ailleurs, la métrique a le plus souvent recours à des résumés statistiques des données simulées. De nombreuses variantes et évolutions comme les MCMC-ABC, SMC-ABC ou encore l'ajustement *post hoc*, ont été développées afin d'améliorer l'efficacité de l'algorithme. Cependant, ces techniques n'ont pas comme priorité de réduire drastiquement le nombre de générations *forward*. L'amortissement des calculs est en revanche un de nos objectifs vu la complexité algorithmique de la simulation stochastique. Pour cela, des approches modélisant la vraisemblance ont été proposées. L'ABC séminal peut être relié à une approximation non paramétrique par noyau (uniforme) de la vraisemblance. Des modèles paramétriques plus informatifs (vraisemblance synthétique) considèrent une gaussienne multivariée - ou un mélange de gaussiennes - comme loi des résumés statistiques des observables. Ce modèle peut être plus économique en termes d'évaluations directes mais s'avère restrictif car il nécessite un choix judicieux de résumés et de s'assurer que ceux-ci sont convenablement approximés par le modèle gaussien. Or, ce manque de flexibilité peut se révéler incompatible avec la complexité de la distribution des observables issues de la simulation stochastique (ex.: la distribution en énergie de particules où chaque dépôt observable est le résultat d'un transport stochastique). Nous proposons par conséquent d'abandonner le passage par des résumés et d'envisager directement l'estimation de la distribution conditionnelle des observables. L'approximation de la vraisemblance est ainsi placée dans un cadre d'estimation de densité conditionnelle.

Densité conditionnelle et *machine learning*

La modélisation de densités de probabilité complexes est aujourd'hui un enjeu majeur dans le domaine du *deep learning*. Elle est par exemple au coeur des modèles génératifs profonds dont les *Variational Auto-Encoder* et les *Generative Advsersarial Network*s sont parmi les plus célèbres et étudiés.

Ces modèles n'ont pourtant pas été conçus pour résoudre les problèmes inverses contrairement à d'autres modèles de *machine learning* ayant des architectures inversibles par construction. Ces modèles génératifs inversibles permettent à la fois de calculer explicitement la densité normalisée ainsi que de générer efficacement des échantillons. Ils peuvent par exemple prendre la forme de *Normalizing Flows, Mixture Density Networks, Autoregressive Models, etc*.
Dans ce cadre, la recherche de nouvelles architectures inversibles profondes et des algorithmes d'apprentissage associés pourrait permettre de respecter la topologie des données pour assurer que la physique des modèles de simulation ait été correctement capturée. Cet apprentissage du lien entre la distribution des données et les configurations des paramètres permettra alors d'accéder à une nouvelle représentation des modèles. L'optimisation de l'apprentissage des réseaux en grande dimension sera également un défi primordial.
D'autre part, par rapport à un maillage prédéfini de l'espace des paramètres, l'apprentissage actif constitue un élément déterminant dans la réduction du nombre de simulations Monte Carlo. Celui-ci vise à orienter progressivement les propositions aléatoires des paramètres vers les régions de plus fortes probabilités de façon à maximiser l'efficacité de l'échantillonnage et contribue à l'amortissement des calculs.

Programme de travail

Dans le cadre de ce post-doctorat, nous proposons ainsi d'évaluer la faisabilité de l'inférence bayésienne pour la simulation stochastique, en combinant un modèle génératif profond de la vraisemblance implicite et les concepts d'apprentissage actif. La simulation Monte Carlo et l'inférence bayésienne par échantillonnage étant connues toutes deux pour leurs coûts calculatoires, nous serons guidés par la modération du nombre des simulations *forward* à opérer.

L’application vise à estimer les paramètres d’une source de rayons X de tomodensitométrie à partir d’un spectre expérimental de photons X acquis dans l’axe du dispositif. En particulier, nous rechercherons l’angle entre le faisceau d’électrons et la cible de conversion ainsi que des épaisseurs et formes de filtrations (moins d’une dizaine de paramètres). Nous considérerons comme observable stochastique chaque dépôt d’énergie détecté. Dans une première phase de preuve de concept, les données, considérées comme expérimentales, seront en fait elles-mêmes simulées. La modélisation du générateur X de tomodensitométrie sera effectuée avec le code PHOEBE (portage du simulateur *Penelope* en C++, développé au CEA/LIST/LM2S). L’implémentation de l’estimation de la *surrogate* du simulateur par *deep learning* et le développement de l’algorithme d’inférence bayésienne (échantillonnage paramètres, apprentissage actif) seront réalisés dans un environnement de programmation probabiliste (ex. Tensorflow-Probability).

Profil et encadrement

Le sujet s’inscrivant dans un programme de recherche transverse, le post-doctorant devra travailler aux interfaces de plusieurs domaines complémentaires. Pour cela, il bénéficiera au CEA d'un encadrement opéré par plusieurs chercheurs. Le candidat pourra présenter un profil en sciences de l’informatique ou en physique théorique et posséder de solides notions en statistiques mathématiques et modélisation bayésienne.

Contacts

eric.barat@cea.fr
eiji.kawasaki@cea.fr
guillaume.damblin@cea.fr