Stochastic and learning algorithms for high energy physics

Motivation: High energy physics experiments generate large volume of data, among which only a fraction contains meaningful information. As an example, the Large Hadron Collider (LHC) experiment produces 1 Pb of data per year, a volume which will be increased by a factor of 30 in the ten years to come. Developing powerful automatic filtering and analysis methods are becoming a crucial challenge to address and an 4-fold improvement of the algorithmic performances is estimated necessary to scale with the future data volume and complexity.

Goal: This PhD thesis aims to improve data analysis in high energy physics experiments through innovations in stochastic algorithms, learning methods and quantitative uncertainty analysis related to the latter. Data representation generated by the non-supervised learning algorithms will also be studied and compared to the existing physics models. Goals are of both theoretical and numerical natures and aims to answer the new challenges set by the LHCb experiment, in particular by the triggering. They will also rely on public data set from CMS and ATLAS collaborations. Participations in big data challenges, as Kaggle ones, will also be considered.

Methodology: We propose three research axes: mock training dataset simulation, implementation of a Bayesian probabilist approach in supervised learning algorithms and the training and sampling of non- or semi-supervised learning methods. These three axes have in common a crucial need for efficient stochastic, so-called Monte Carlo, methods. Monte Carlo algorithms sample probability distributions via a Markov chain. The algorithm efficiency is directly linked to the exploration speed of its underlying chain. Recent advances were achieved in statistical physics by the development of non-reversible Markov chains and are now under a growing interest in Bayesian statistics. This thesis will revolve around these recent progresses by characterizing them further theoretically and applying them to Bayesian learning methods.

Environment: The PhD candidate will work at the Université Clermont-Auvergne and will be part of the LHCb collaboration (CERN, Geneva). This thesis lies in the interface between mathematics, computer science and physics and will be supervised by Arnaud Guillin and Manon Michel from Laboratoire de Mathématiques Blaise Pascal (UMR 6620) and by Stéphane Monteil from Laboratoire de Physique Corpusclaire de Clermont (UMR 6533) and member of the LHCb collaboration.

We encourage candidates with a background in mathematics, data science, statistical or particle physics to apply. The PhD thesis will start in 2019, at the latest in September. Interested candidates can contact Manon Michel (manon.michel@uca.fr) for more details. An internship taking place in the beginning of 2019 can precede the PhD.

Webpages:
Manon Michel's webpage: http://www.normalesup.org/~mmichel/
Arnaud Guillin's webpage: http://math.univ-bpclermont.fr/~guillin/
Webpage of the LPC team of the LHCb collaboration: http://lhcb-clermont.in2p3.fr/spip.php?rubrique7

References:
K. Albertsson et al (2018), *Machine Learning in High Energy Physics Community White Paper*: https://arxiv.org/abs/1807.02876
Michel et al (2018), *Forward Event-Chain Monte Carlo: Fast sampling by randomness control in irreversible Markov chains*: https://arxiv.org/abs/1702.08397
Durmus et al (2018), *Geometric ergodicity of the bouncy particle sampler*: https://arxiv.org/abs/1807.05401