Theory of stochastic simulation in machine learning

Supervisors: Henrik Hult and Jimmy Olsson

KTH Engineering Sciences

The aim of this project is to develop stochastic simulation methods in Machine Learning along with a corresponding rigorous efficiency analysis based on the theory of large deviations.

To facilitate learning of models for large and complex data the training mechanism needs to be sufficiently fast. The training of probabilistic models requires the evaluation of the likelihood, which, for latent variable models commonly used in Machine Learning, is given as an expectation of the conditional likelihood over the latent variables. In high-dimensional settings, sampling from the prior distribution typically gives very small likelihood, because it is unlikely to draw a latent variable that fits with the observed data. Importance sampling is a generic method for improving sampling efficiency of rare events, by sampling from a distribution under which the event of interest is no longer rare. In recent years the development of efficient importance sampling algorithms have taken significant steps forward through development of sequential Monte Carlo methods, Markov chain Monte Carlo methods (MCMC) and by considering asymptotic properties, based on large deviations theory, of the underlying stochastic models.

In the context of high-dimensional models encountered in machine learning and statistical physics standard implementation of MCMC may be slow because the energy landscape that defines the probability distribution is complex, with many local minima. Alternative designs, based on defining extended ensembles can facilitate faster convergence necessary for practical use in high-dimensional settings. Efficient designs can be mathematically analysed within large deviations theory, which aims at characterising the asymptotic properties of the resulting estimators. Recent progress of large deviations theory has lead to the development of new efficient algorithms for complex computations in chemical and statistical physics, and the development of such mathematical techniques is likely to have a significant future impact on the computational methods used in data science.

The sequential Monte Carlo (SMC) methods form another class of genetic-type algorithms for online approximation of sequences of probability distributions, with unknown partition function. The great generality and flexibility of the SMC methodology has together with the dramatic increase of computational power over recent years lead to a rapidly and steadily increasing interest in these methods. Today sequential Monte Carlo methods are successfully applied within a wide range of applications, including, e.g., computer vision, machine learning, automatic control, signal processing, optimization, robotics, econometrics, and finance. A future challenge lies in developing the sequential Monte Carlo methods to high-dimensional settings to cope with the problems and complex data arising in machine learning. We aim at addressing this challenge by combining the strength of SMC methods with deep recurrent neural networks.