Stochastic Calculus in Machine Learning: Optimization, Sampling, Simulation

Maxim Raginsky

Abstract

A great deal of recent research activity has focused on using continuous-time processes to analyze discrete-time algorithms and models. In particular, diffusion processes have been examined as a way towards a better understanding of first-order optimization methods, as they afford an analysis of behavior over non-convex landscapes. Gradient flows and diffusions have also found a role in the analysis of deep neural networks, where they are interpreted as describing the limiting case of infinitely many layers, each in effect infinitesimally thin.

In this tutorial, I will give an informal treatment of some of the recent applications of stochastic calculus of K. Ito to some problems at the intersection of optimization and machine learning. Specifically, I will cover the following topics:

I) Optimization — I will discuss non-convex learning using continuous-time Stochastic Gradient Langevin Dynamics (SGLD). I will first show that, under reasonable regularity assumptions on the objective function, SGLD finds an approximate global minimizer of the population risk in finite time (which, generally, be exponential in the problem dimension), and then discuss the metastability phenomenon of the Langevin dynamics at “intermediate” time scales. Here, by metastability I mean that, with high probability, the trajectory of the Langevin diffusion will either spend an arbitrarily long time in a small neighborhood of some local minimum or will quickly escape that neighborhood within a short recurrence time.

II) Sampling and simulation — I will show that diffusion processes with drift given by a sufficiently deep feedforward neural net provide a flexible and expressive class of probabilistic generative models. I will first show that sampling in such generative models can be phrased as a stochastic control problem (revisiting the classic results of Föllmer and Dai Pra) and then build on this formulation to quantify the expressive power of these models. Specifically, I will prove that one can efficiently sample from a wide class of terminal target distributions by choosing the drift of the latent diffusion from the class of multilayer feedforward neural nets, with the accuracy of sampling measured by the Kullback-Leibler divergence to the target distribution.