Policy Learning via Fully Probabilistic Design

Siavash Fakhimi Derakhshan, Tatiana Valentine Guy

Adaptive System Department, Institute of Information Theory and Automation, Czech Academy of Sciences, Prague 18200, Czech Republic

e-mail: Fakhimi@utia.cas.cz

Key words: Fully probabilistic design, imitation learning, Kullback-Liebler divergence, learning from demonstration, optimal policy.

The discovery of new molecules and materials with desired properties, known as inverse design problems, helps expand the horizons of novel and innovative real-life applications. There are a variety of inverse problems in chemistry encompassing various subfields like drug discovery, retrosynthesis, structure identification, etc. The domain size of the state variables on the chemical space makes it infeasible to search through all possible molecules. Recent developments in modern machine learning methods have shown great promise in tackling problems of this kind. Inverse reinforcement learning is one of the effective methods that has been used frequently in this domain.

Choosing the proper reward function is the challenging part of the problem and has a great impact on the performance of the learning mechanism in the inverse reinforcement learning technique, especially when we are dealing with systems that have high uncertainty. On the other hand, although the decision design algorithm based on fully probabilistic design (FPD) performs well in dealing with systems that are stochastic representation, it requires a high computational and processing cost.

Applying formalism of fully probabilistic design, we propose a new general data driven approach for finding a stochastic policy from demonstrations. The approach infers a policy directly from data without interaction with the expert or using any reinforcement signal. The expert's actions generally need not to be optimal. The proposed approach learns an optimal policy by minimising Kullback-Liebler divergence between probabilistic description of the actual agent-environment behaviour and the distribution describing the targeted behaviour of the optimised closed loop. We demonstrate our approach on simulated examples and show that the learned policy: i) converges to the optimised policy obtained by FPD; ii) achieves better performance than the optimal FPD policy whenever a mismodelling is present.