Review-13

Model-Based Reinforcement Learning with Value-Targeted Regression

Parameteric models allow the scalability of Reinforcement Learning (RL) to large state and action spaces. The work proposes a novel algorithm for model parameter estimation. Th transition model is assumed to admit linear parameterization. Based on this formulation, the proposed algorithm carried out model parameter estimation by recursively solving a regression problem with target as the latest value estimate. Value-targeted regression yields an upper bound on the regret $\mathcal{O}(d\sqrt{H^3T})$. The regret bound is independent of the total number of states and actions and close to the proposed lower bound $\Omega(\sqrt{HdT})$.

Conventional model-based RL methods explicitly estimate transition probabilities and operate on raw states. To that end, the work aims to deviate from this notion and estimates model parameters by setting up a regression problem based on the value function. Targets in regression updates are latest value estimates. Value-targeted formulation yields one-dimensional targets which eliminates the need for multivariate tuning. Addtionally, the model parameters θ are updated through a simple recursive formula. Computation is carried out by constructing upper confidence estimates of Q-values which yield value estimates. Estimated value functions are used as targets in regression with $X_{h,k}^T.\theta$ being the predicted value consisting of approximate Monte-Carlo value estimates $X_{h,k} = \mathbb{E}[V_{h+1,k}(s)|s_h^k, a_h^k]$. The empirical loss function $(X_{h,k}^T.\theta - y_{h,k})^2$ consists of the taget $y_{h,k} = V_{h+1,k}(s_{h+1}^k)$ as the latest target estimates. Loss is updated using a ridge-regression setting with a regularization term. Recursive computations involve the utility of inner product $X_{h,k}.X_{h,k}^T$.

Linear parameterization of the model aids in recursive computation of value estimates and hence, provides an upper confidence bound with sub-linear regret. Additionally, Q-value estimates can be parameterized by d parameters and sufficiently balance exploration with epxloitation by making use of an exploration bonus. However, the work does not provide intuitive insights into the sufficiency of value-targeted regression. The formulation of value estimates as variable targets only leads to a theoretical result which does not provide strong guarantees on the empirical aspects of the bound. Additionally, target variables directly make use of latest value estimates as a result of which they may suffer from high stochasticity of the value function. This allows one to ask whether the bound trades off variance at the cost of regret? An intuitive perpsective of stochasticity in estimates would help explain the bound coherently.