Model-Based Value Expansion for Efficient Model-Free Reinforcement Learning

Niharika Shrivastava
School of Computing
National University of Singapore
Singapore, 119077
niharika@comp.nus.edu.sg

Abstract

The authors of [1] propose to use a dynamics model to simulate the short-term horizon (near-future Model-Based (MB) component) and Q-learning to estimate the long-term value beyond the simulation horizon (distant-future Model-Free (MF) component) in order to reduce the sample complexity of learning.

1 Model-Based Value Expansion

The issue of model bias with MB approaches results in overfitting in exactly those low-data regions where they are most needed. The authors propose model-based value expansion (MVE) that controls for uncertainty in the model by using the learnt dynamics model only up to a fixed depth H. MVE estimate for the value of a given state is defined as a culmination of a component predicted by the dynamics model (on-policy) and its value at horizon H.

1.1 Implementation

It relies on an approximate fixed point construction from an empirical distribution of transitions β (replay buffer) that enables accurate simulation of H states from the dynamics model. These states constitute targets that are used to train the value function on the entire training distribution, instead of just the buffer. Authors also use a TD-k trick to skirt the distribution mismatch problem such that the training distribution is an approximate fixed point which helps in training Q values appropriately.

Results showcased that incorporating synthetic experience from a dynamics model via short horizon greatly improved the performance of model-free RL by being sample-efficient. However, key design decisions are necessary for its success such as the usage of TD-k trick and a trustworthy horizon value H.

1.2 Comparison with other approaches

MA-DDPG performs worse than MVE-DDPG due to the possible staleness of imaginary states and overfitting the actor using imaginary data. MVE can be extended onto ME-TRPO to achieve good performance. MVE enables faster learning by use of off-policy data as compared to other n-step return methods.

References

[1] Feinberg, V., Wan, A., Stoica, I., Jordan, M. I., Gonzalez, J. E., & Levine, S. (2018). Model-Based Value Estimation for Efficient Model-Free Reinforcement Learning. ArXiv. https://doi.org/10.48550/arXiv.1803.00101

Preprint. Under review.