Review-28

γ-Models: Generative Temporal Difference Learning for Infinite-Horizon Prediction

Compounding errrors in model predictions steer the agent away from the path towards optimal behavior. These errors are further alleviated by the choice of the horizon length which is suitably dealt with using a value function. To this end, the work presents γ-models which provide an infinite probabilistic horizon to the agent. γ-model is trained using a generative temporal difference learning analogous to successor representations. γ-models interpolate between a value function to store information about the future and predictive model to be independeent of task rewards. Models are instantiated as adversarial networks and normalizing flows depending on the presence of density evaluation.

γ-models facilitate infinite horizon prediction which serves constant-time prediction as generalized rollouts primary advantages. The γ-model constructs an off-policy framework for training generative models using temporal difference learning. The taget distribution is a weighted combination of single-step distribution and model bootstrap which represent the distribution at immediate next timestep and overall subsequent timesteps respectively. This mixture is reminiscient of a temporal diference target value. Training of the model is carried under two different scenarios, when the model permits sampling and when the model permits density evaluation. In the case of sampling, model is trained as a generative adversarial network minimizing the bootstraped f-divergence. In the case of density evaluation, the model is trained as normalizing flows which regresses to the constructed target density values. Training motivates the application of γ-models for carrying rollouts which depend on reweighting γ for appropriate model rollout lengths. Utilization γ-models for value expansion further leads to the γ-MVE estimator which trades away the necessity for finite step rollouts.

Suitability of γ-model trained using a generative scheme is validated upon observing the learned distributions from normalization flows on acrobot and pendulum tasks. Furthermore,m γ-models accurately represent the value map of tasks and yield improved performance on the evaluated benchmarks. While, the γ-model provides a theoretically rich connection to successor representations, the work empirical insights towards this theory. This leaves the role of γ-models as continuous successor features an open problem. Furthermore, the the γ-model is found suitable only for low dimensional tasks wherein the agent is not significantly impacted by compounding errors. It would be interesting to evaluate the model on high dimensional feature spaces consisting of pixel inputs and larger propioceptive state spaces.

Training of models using generative temporal difference learning provides a significanlty novel direction for future work. The model-based setup may be extended towards other generative techniques such as autoregressive flows which provide a temporal connection to sequential decision-making. Furthermore, the use of γ-MVE as an estimator can be explored in scenarios consistig of high dimensonal inputs and action spaces wherein compounding errors are a significant problem.