Review-36

Expected Eligibility Traces

Assigning credit to state-action pairs is a challenging problem in Reinforcement Learning (RL). Eligibility traces provide a mechanism for assigning credit to recent state-action pairs but do not take into account counterfactual sequences. The work introduces expected eligibility traces, which update all past state and action pairs which could have preceded the current state. More specifically, the work introduces ET(λ) and ET(λ, η) which present a generalization of eligibility traces and a machanism for interpolating between instantaneous and expected traces respectively. Experiments carried in the linear and nonlinear settings demonstrate suitability of expected traces.

One-step Temporal Difference (TD) methods assign credit slowly when compared to multi-step updates. On the other hand, multi-step methods such as Monte-Carlo (MC) learning present high variance. This allows one to couple multi-step methods with eligibility traces which keep track of past updates diminishing over temporal span. However, eligibility traces do not take into account counterfactual state-action pairs. To achieve this, the work introduces expected eligibility traces which provision single-step counterfactual updates using the expectation of a given trace under the agent's state and policy. The first of the two methods, ET(λ), is an extension of TD(λ) and brings the expected trace closer to the actual trace. ET(λ) updates have lower variance than TD($lambda$). The second method, ET(λ, η), is called the mixture trace which is a generalization of ET(λ) and TD(λ) and lies between MC and state-based estimate updates. ET(λ, η) allows one to interpolate between expected and instantaneous traces using a recursive relation. At $\eta = 1$, ET(λ, η) represents the instantaneous TD(λ) while at $\eta = 0$, ET(λ, η) denotes ET(λ). A theoretical analysis of ET(λ, η) reveals that the method converges in the linear case as TD(λ).

Experiments carried out on linear and nonlinear task settings cnsisting of gridworld, multi-chain and Atari environments demonstrate the suitability of expected traces. In the case of gridworld and multi-chain tasks, ET(λ) is found to assign values over longer temporal spans and reduced prediction errors as a result of counterfactual knowledge respectively. In the nonlinear setting of two Atari environments, ET(λ) is combined with Deep Q-Learning to demonstrate its suitable scalability in comparison to pre-existing Deep RL methods. While expected traces present a generalized method for counterfactual credit assignment, its theoretical and empirical study could be improved with regards to a few caveats. Firstly, the