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 $\text{ET}(\lambda)$ and $\text{ET}(\lambda, \eta)$ which present a generalization of eligibility traces and a machanism for interpolating between instantaneous and expected traces respectively. Experiments carried out in linear and nonlinear settings demonstrate suitability of expected traces.

Multi-step methods such as Monte-Carlo (MC) learning or n-step returns 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, $\text{ET}(\lambda)$, is an extension of $\text{TD}(\lambda)$ and brings the expected trace closer to the actual trace. $\text{ET}(\lambda)$ updates have lower variance than $\text{TD}(\lambda)$. The second method, $\text{ET}(\lambda, \eta)$, is called the mixture trace which is a generalization of $\text{ET}(\lambda)$ and $\text{TD}(\lambda)$ and lies between MC and state-based estimate updates. $\text{ET}(\lambda, \eta)$ allows one to interpolate between expected and instantaneous traces using a recursive relation. At $\eta = 1$, $\text{ET}(\lambda, \eta)$ represents the instantaneous $\text{TD}(\lambda)$ while at $\eta = 0$, $\text{ET}(\lambda, \eta)$ denotes $\text{ET}(\lambda)$. A theoretical analysis of $\text{ET}(\lambda, \eta)$ reveals that the method converges in the linear case as $\text{TD}(\lambda)$.

Experiments carried out on linear and nonlinear task settings consisting of gridworld, multi-chain and Atari environments demonstrate the suitability of expected traces. In the case of gridworld and multi-chain tasks, $\text{ET}(\lambda)$ is found to assign values over longer temporal spans and reduce prediction errors as a result of counterfactual knowledge respectively. In Atari environments, $\text{ET}(\lambda)$ demonstrates 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, theoretical convergence of mixture trace is demonstrated for the linear case but the work does not throw light on its efficacy in the more general nonlinear setting. Secondly, empirical analysis in the nonlinear Atari domain presents a heavy dependence on learning rate. Moreover, $\text{Q}(\lambda)$ demonstrates consistent performance for $\alpha = 10^{-5}$ in comparison to $\text{QET}(\lambda)$. One might infer that learning improvements are originating from parameter tuning rather than trace estimate updates.

$\text{ET}(\lambda)$ and $\text{ET}(\lambda, \eta)$ present a suitable framework for credit assignment by utilizing trace updates in the form of state-based estimates. The work introduces, two new directions for future research. Firstly, linear expected traces present a close connection to predecessor features and can be theoretically improved in this regard. And secondly, expected traces can be extended to batch learning methods wherein the expected trace can be updated towards mixture trace and the resulting update can be thought of as forward-view update reversed in time.

Eligibility traces fail to take into account past counterfactual transitions during credit assignment. The work introduces expected eligibility traces which assign credit counterfactually based on all past transitions that could have occured. Expected trace updates are presented as $\text{ET}(\lambda)$ and $\text{ET}(\lambda, \eta)$ which is a generalization of expected and instantenous traces. Thereotical and empirical evaluation of expected traces demonstrate their suitability in credit assignment.