Replication of *A Family of Robust Stochastic Operators for Reinforcement Learning*

Faculty Sponsor: Michael Littman

Q-learning is a major technique in Reinforcement Learning, which functions by updating the agent's estimates of the value of state-action pairs according to some operator. While Q-learning using Bellman operator based updates can be shown to exactly solve for the optimal policy in the absence of approximation or estimation errors, this may not necessarily be the case for more complex problems. In addition, when the values of different actions at a particular state are close, these approximation errors further make learning the true optimal action difficult.

Recent efforts to mitigate these effects involve constructing Q-learning operators that are *optimality* preserving and *action-gap increasing*. Optimality preserving operators guarantee convergence to the optimal policy while gap-increasing operators improves Q-learning's rate of convergence. Most recently proposed operators with the above properties have been deterministic, which carry with them an inherent trade-off between optimality preservation and action-gap increasing. More specifically, operators which increase the action gap too much may violate optimality.

In their paper *A Family of Robust Stochastic Operators for Reinforcement Learning,* Lu, Squillante, and Wu propose adding randomness with certain properties to the standard Bellman operator to circumvent this trade-off. The authors provide rigorous justification that these stochastic operators are both optimality-preserving and gap increasing in a stochastic sense. They then provide experimental evidence of these claims as well as directly compare the performance of their stochastic operator, the standard Bellman operator and Consistent Bellman Operator, a type of deterministic gap increasing and optimality preserving operator, on OpenAI gym games.

In our replication study, we will briefly explain relevant notation and the relevance of the stochastically action-gap increasing and optimality-preserving. We compare the results found and our replicated results as well as analyze the performance of RSO under different exploration schemes, such as the ε-greedy and softmax methods. Finally, we compare the performance of different distributions for the perturbation term in the stochastic operator.