Reinforcement Learning with Markov Risk Measures and Multipattern Risk Approximation

Andrzej Ruszczyński[1] and Tiangang Zhang[2]

[1,2]Rutgers University, Piscataway, NJ 08854, USA

Abstract

For a risk-averse finite-horizon Markov Decision Problem, we introduce a special class of Markov coherent risk measures, called mini-batch measures. We also define the class of multi-pattern risk-averse problems that generalizes the class of linear systems. We use both concepts in a feature-based Q-learning method with multi-pattern Q-factor approximation and we prove a high-probability regret bound of $\mathcal{O}(H^2N^H\sqrt{K})$, where H is the horizon, N is the mini-batch size, and K is the number of episodes. We also propose an economical version of the Q-learning method that streamlines the policy evaluation (backward) step. The theoretical results are illustrated on a stochastic assignment problem and a short-horizon multi-armed bandit problem.