Exploiting Structure and Uncertainty of Bellman Updates in Markov Decision Processes

D. Tateo, C. D'Eramo, A. Nuara, M. Restelli, A. Bonarini
{davide.tateo, carlo.deramo, alessandro.nuara, marcello.restelli, andrea.bonarini}@polimi.it

Problem

- Learning is difficult in highly stochastic environments
- Uncertainty in action-value function estimates propagates
- Some algorithms face this problem focusing on the bias of the estimate

Contributions

1. Split the estimate in two components:
   - The expected reward $\widetilde{R}(x, u)$
   - The expected next state value function $\widetilde{Q}(x, u)$
2. Use different learning rates for the two components
3. We provide empirical results showing the effectiveness of our approach

RQ-Learning Algorithm

Idea

Split the action-value function in two components:

- $\widetilde{R}(x, u) = \mathop{\mathbb{E}}_{x' \sim \mathcal{P}(x'|x,u)} [r(x, u, x')]$
- $\widetilde{Q}(x, u) = \mathop{\mathbb{E}}_{x' \sim \mathcal{P}(x'|x,u)} \left[\max_{u'} Q^*(x', u') \right]$
- $Q^*(x, u) = \widetilde{R}(x, u) + \gamma \widetilde{Q}(x, u)$

Empirical Results