

# 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



**MILANO 1863** 

### 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
- Despite empirical evidence, there is **no proof** that focusing on the bias is the solution

#### CONTRIBUTIONS

- 1. Split the estimate in **two components**:
  - The expected reward  $\widetilde{R}(x, u)$
  - The expected next state value function 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**

Improve accuracy of the estimate exploiting:

Structure of the Bellman update

Uncertainty of the estimation

APPROACH

Split the action-value function in two components

Compute the update as follows

$$\widetilde{R}(x,u) = \mathbb{E}\left[r(x,u,x')\right] \qquad \widetilde{Q}(x,u) = \mathbb{E}\left[\max_{x' \sim \mathcal{P}(x'|x,u)} Q^*(x',u')\right] \qquad \widetilde{R}_{t+1}(x,u) \leftarrow \widetilde{R}_t(x,u) + \alpha_t(R(x,u,x') - \widetilde{R}_t(x,u))$$

$$R_{t+1}(x,u) \leftarrow R_t(x,u) + \alpha_t(R(x,u,x) - R_t(x,u))$$

$$Q^*(x,u) = \widetilde{R}(x,u) + \gamma \widetilde{Q}(x,u)$$

$$\widetilde{Q}_{t+1}(x,u) \leftarrow \widetilde{Q}_t(x,u) + \beta_t(\max_{u'} Q_t(x',u') - \widetilde{Q}_t(x,u))$$

Exploit the variance of estimation to set the learning rate

1. Estimate the variance of the estimator  $\tilde{Q}$ , using the sample variance of the target:

$$\operatorname{Var}\left[\widetilde{Q}\right] \approx S_t^2 \omega_t$$

$$\operatorname{Var}\left[\widetilde{Q}\right] \approx S_t^2 \omega_t \qquad \qquad \omega_{t+1} = (1 - \beta_t)^2 \omega_t + \beta_t^2$$

2. Compute the learning rate according to the precision of the estimate:

Noisy Gridworld

 $\max Q(s, a)$ 

• Inversely proportional  $\beta$ :

-0.5

10000

$$\beta_t = \frac{\sigma_e^2(t)}{\sigma_e^2(t) + \eta}$$

Markethatelphatelphateleastra

5000

# steps

• Directly proportional  $\delta$ :

$$\delta_t = e^{\frac{\sigma_e^2}{\eta} \log \frac{1}{2}}$$

**Q-Learning:**  $\alpha_t = \beta_t$ 

5000

20000

10000

# steps

**RQ-Learning:**  $\beta_t \neq \alpha_t$ 

-0.5

10000

10000

5000

— RQDelta

--- RQWin

# steps

- RQBeta

— RQAlphaBeta

--- RQWinAlpha

10000

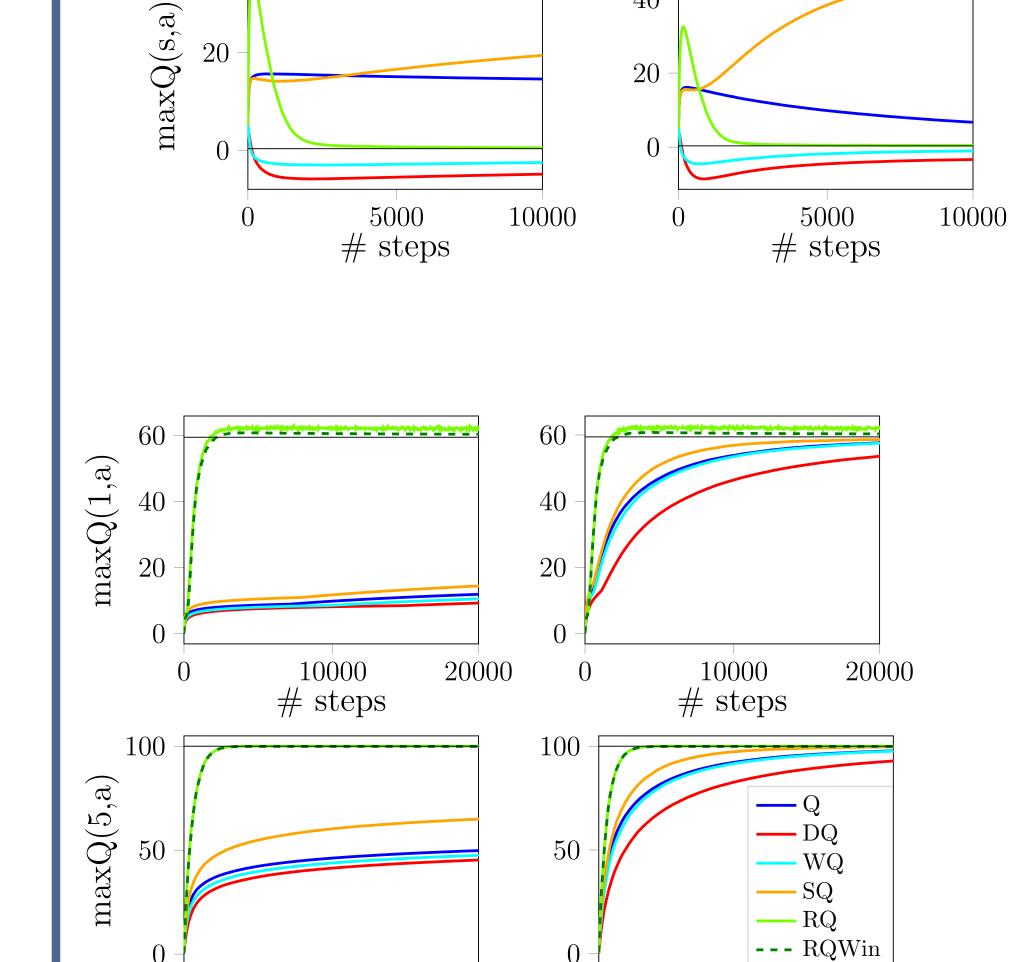
10000

**RQ** $_{\delta}$ -Learning:  $\beta_t = \alpha_t \delta_t$ 

## EMPIRICAL RESULTS

5000

Reward ...



20000

10000

# steps

