

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

- Split the action-value function estimate **two components**:
 - The expected reward $\tilde{R}(x, u)$
 - The expected next state value function $\tilde{Q}(x, u)$
- Use **different learning rates** for the two components
- 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

$$\begin{aligned} \tilde{R}(x, u) &= \mathbb{E}[r(x, u, x') \mid x' \sim \mathcal{P}(x' \mid x, u)] & \tilde{Q}(x, u) &= \mathbb{E}\left[\max_{u'} Q^*(x', u') \mid x' \sim \mathcal{P}(x' \mid x, u)\right] \\ Q^*(x, u) &= \tilde{R}(x, u) + \gamma \tilde{Q}(x, u) & \tilde{R}_{t+1}(x, u) &\leftarrow \tilde{R}_t(x, u) + \alpha_t (R(x, u, x') - \tilde{R}_t(x, u)) \\ & & \tilde{Q}_{t+1}(x, u) &\leftarrow \tilde{Q}_t(x, u) + \beta_t (\max_{u'} Q_t(x', u') - \tilde{Q}_t(x, u)) \end{aligned}$$

Q-Learning: $\beta_t = \alpha_t$

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

RQ_δ-Learning: $\beta_t = \alpha_t \delta_t$

Exploit the **variance of estimation** to set the learning rate

- Estimate the variance of the estimator \tilde{Q} , using the sample variance of the target:

$$\text{Var}[\tilde{Q}] \approx S_t^2 \omega_t$$

$$\omega_{t+1} = (1 - \beta_t)^2 \omega_t + \beta_t^2$$

- Compute the learning rate according to the **precision of the estimate**:

- Inversely** proportional β :

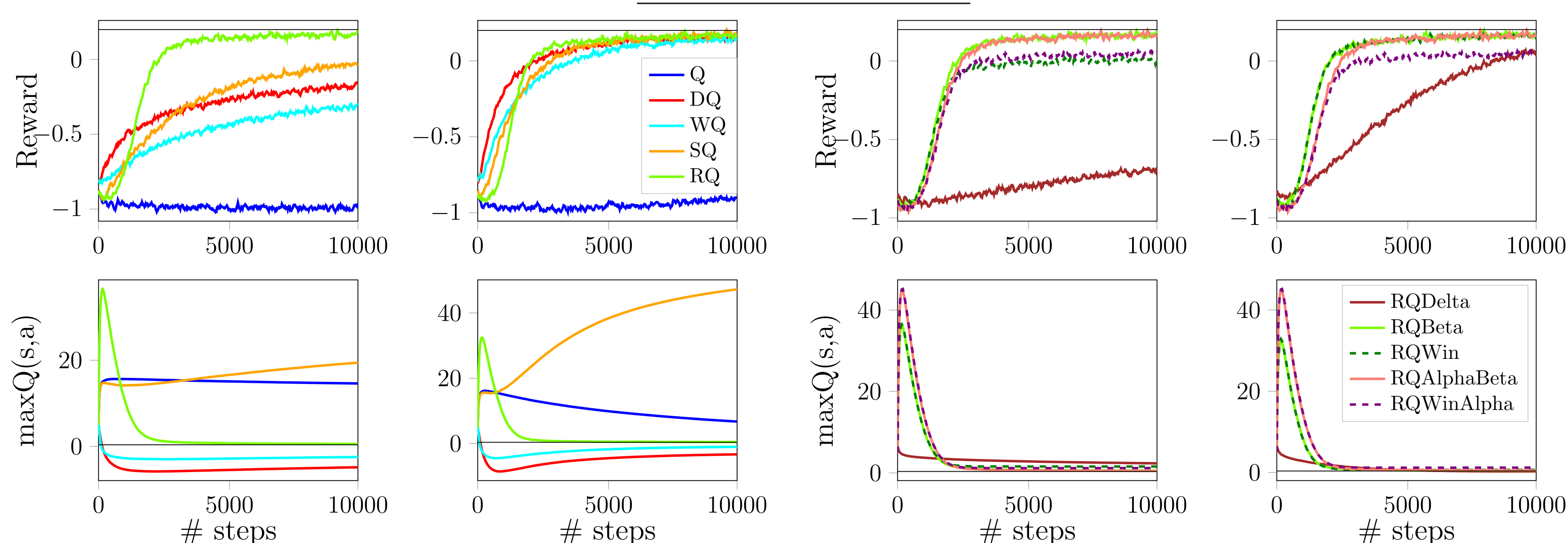
$$\beta_t = \frac{\sigma_e^2(t)}{\sigma_e^2(t) + \eta}$$

- Directly** proportional δ :

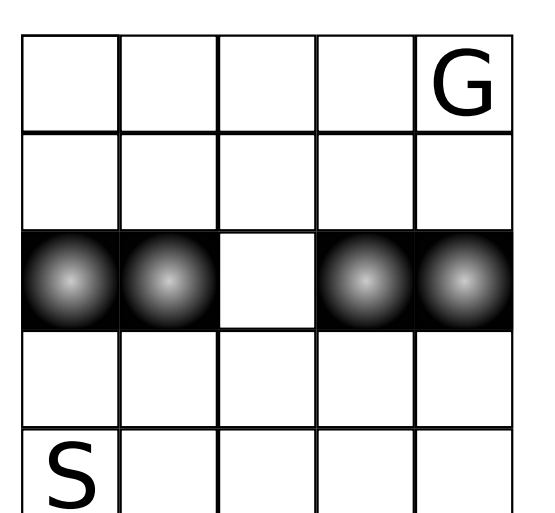
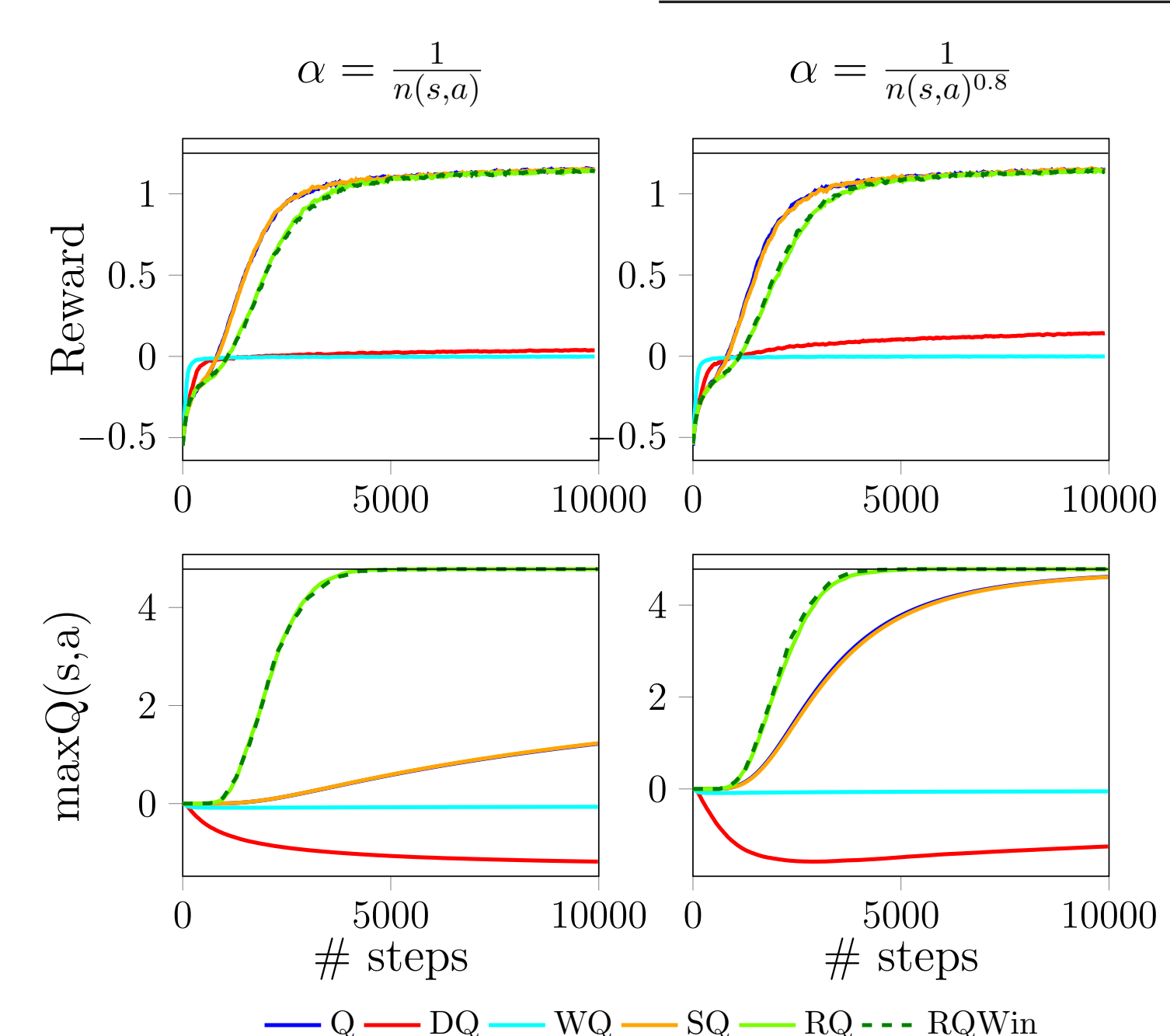
$$\delta_t = e^{\frac{\sigma_e^2}{\eta} \log \frac{1}{2}}$$

EMPIRICAL RESULTS

NOISY GRIDWORLD



GRIDWORLD WITH HOLES



DOUBLE CHAIN

