Off-Policy and Off-Actor Actor-Critic with **B**ootstrapped **D**ual **P**olicy **I**teration

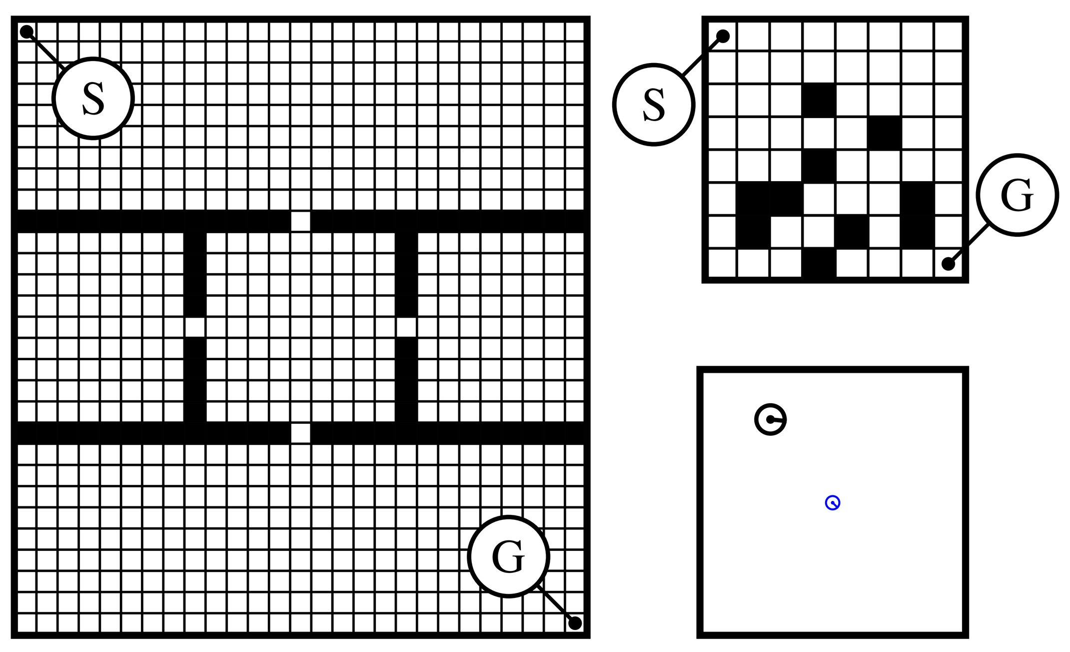
VRIJE UNIVERSITEIT BRUSSEL

D. Steckelmacher dsteckel@ai.vub.ac.be | H. Plisnier hplisnie@vub.ac.be | D. M. Roijers droijers@ai.vub.ac.be | A. Nowé anowe@vub.ac.be

Environment

Markov Decision Processes with **continuous states** and **discrete actions**. We focus in **high sample-efficiency** and **exploration quality**.

Table is a continuous-state environment where a simulated robot has to dock onto its charging station. Both are on a 1-by-1 square table. The charging station is at (0.5, 0.5), and the robot starts at (0.1, 0.1). Actions allow the robot to turn left/right 0.1 radians, or move forward 0.005 units. The robot docks (+100) when it is on the charging station ± tolerance. A reward of -50 is given if the robot falls off the table.
Five Rooms is a 47-by-49 gridworld with walls. The agent receives -1 per time-step, +100 when reaching the goal.
Frozen Lake (8x8) is a highly-stochastic gridworld from the OpenAI Gym. The agent receives a reward of +1 when reaching the goal, -1 when falling in one of the fatal pits. Actions allow the agent to move up, down, left or right, but cause a random move with a probability of ⅔.



action $a_t \sim \pi(s_t)$

state s_t

Actor

The critics produce greedy policies, that the actor progressively imitates:

$$\pi_{k+1} \leftarrow (1 - \lambda)\pi_k + \lambda G(Q_k)$$

Due to the moving average, the actor estimates the **expected greedy policy of the critics**. Because the *off-policy* and *off-actor* critics approximate **Q^* instead of Q^π**, the actor quickly converges to the optimal policy:

$$\begin{aligned}\pi &= E_{Q\sim P(Q=Q^*)}[G(Q)] \\ &= E_{Q\sim P(Q=Q^*)}[G(Q^*)] \\ &\rightarrow \pi^*\end{aligned}$$

Moreover, the actor selects actions in a way comparable to **Thompson sampling**:

$$\pi(s, a) = P[a = \text{argmax}_{a'} Q(s, a')]$$

$s_{512} \rightarrow G_{512}(Q)$
greedy policy

Off-Policy and Off-Actor Critics

Taking inspiration from Bootstrapped DQN [3], **16 critics** learn **Q*** from experiences sampled in the experience buffer. They use an **off-policy** version of Clipped DQN [1], that, like Double DQN [2], maintains two Q-functions per critic, Q^A and Q^B:

$$\begin{aligned}&Q_{k+1}(s_t, a_t) \leftarrow Q_k(s_t, a_t) + \alpha(r_t + \gamma V(s_{t+1}) - Q_k(s_t, a_t)) \\ &V(s_{t+1}) = \min_{A,B} Q^{A,B}(s_{t+1}, \text{argmax}_{a'} Q^A(s_{t+1}, a')) \\ &Q^A, Q^B \leftarrow Q_{k+1}, Q^A\end{aligned}$$

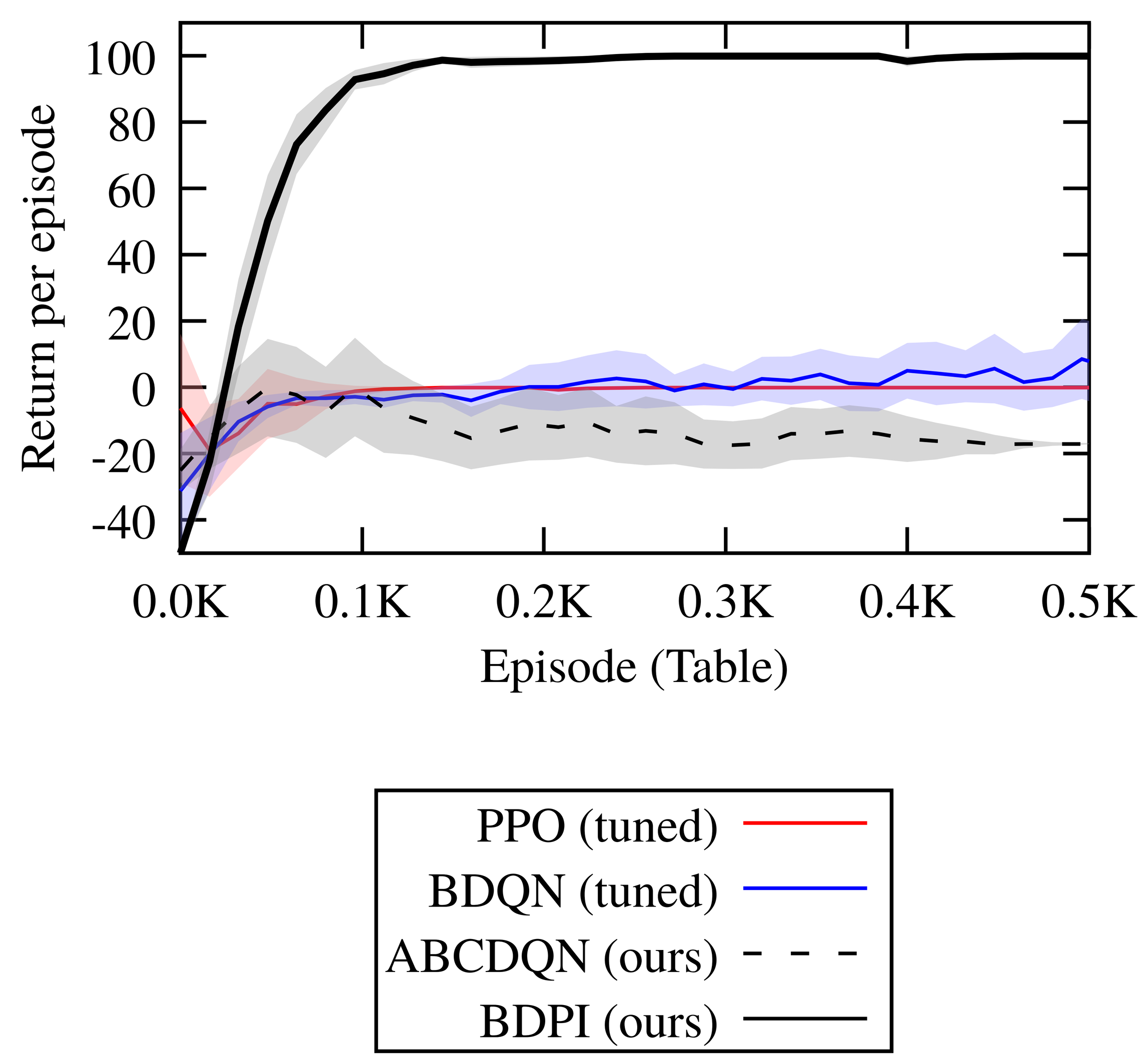
Our **Aggressive Bootstrapped Clipped DQN** (**ABCDQN**, the critic part of **BDPI**) algorithm goes several steps further:

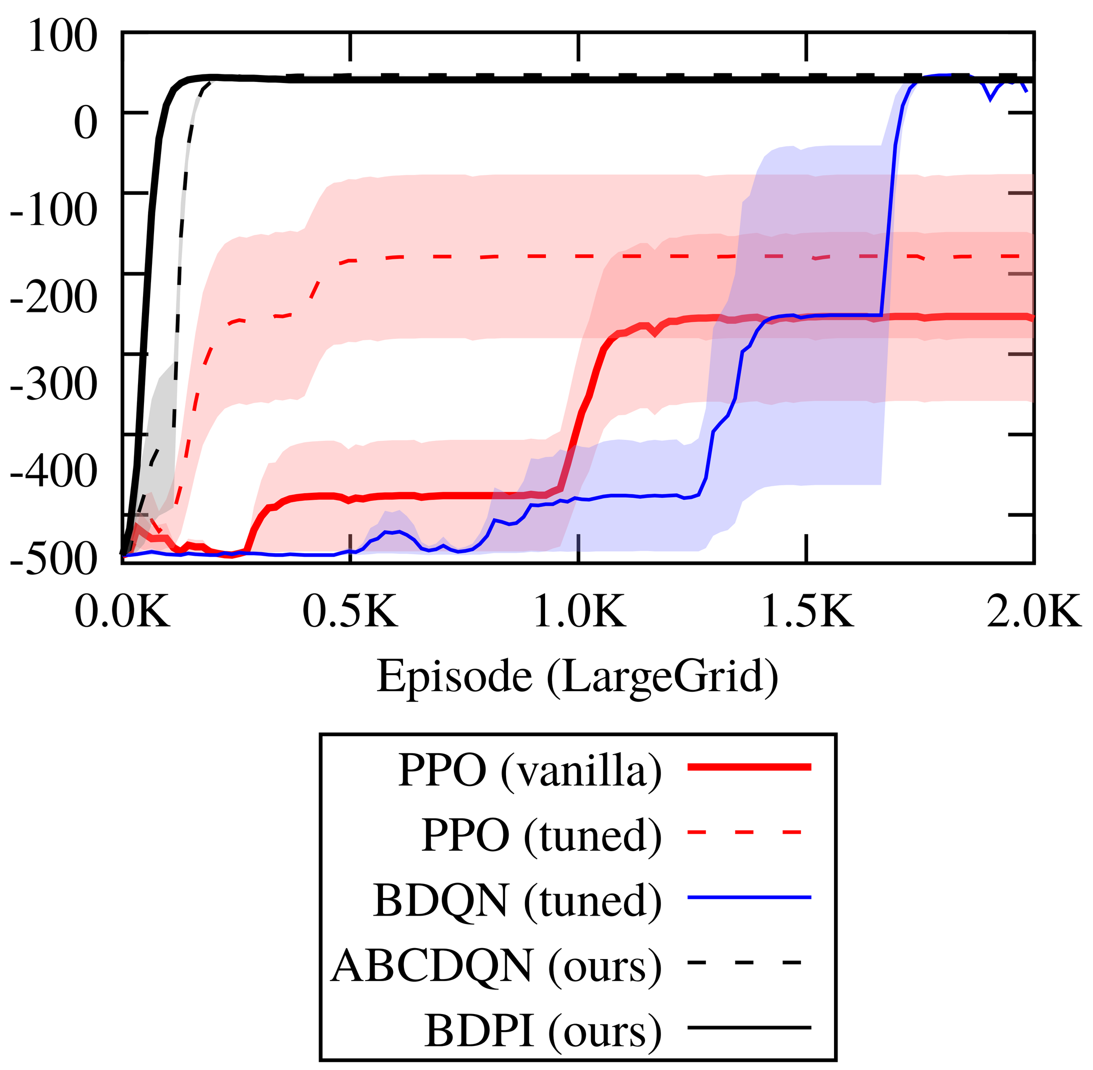
At every time-step:
- For each critic:
  - Sample **512** experiences
  - Repeat **4** times:
    - Compute Q_{k+1} from the experiences
    - Fit Q^A on Q_{k+1} with the MSE loss, for **20 epochs**
    - Swap Q^A and Q^B
  - Train the actor on the **greedy policy** to the critic

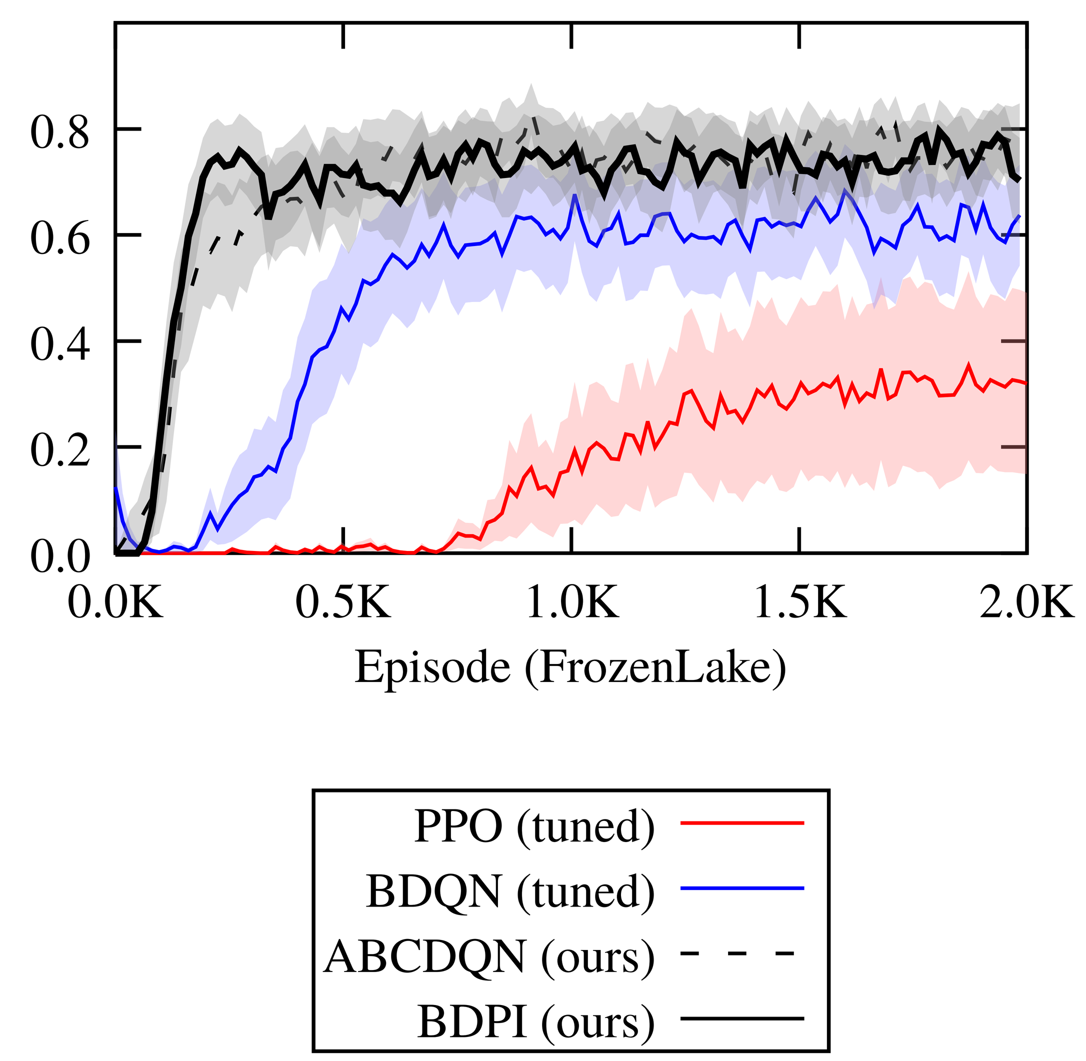
(s_t, a_t, r_t, s_{t+1})
experience

Experience Buffer (20 000)

$(s_t, a_t, r_t, s_{t+1})_{512}$
batch







References

[1]: Fujimoto, Van Hoof and Meger, *Addressing Function Approximation Error in Actor-Critic Methods*, ICML, 2018
[2]: Van Hasselt, *Double Q-Learning*, NIPS, 2010
[3]: Osband, Blundell, Pritzel and Van Roy, *Deep Exploration via Bootstrapped DQN*, NIPS, 2016
[4]: Pirotta, Restelli, Pecorino and Calandriello, *Safe Policy Iteration*, ICML, 2013

Acknowledgments

The first author is "Aspirant" at the Research Foundation - Flanders (FWO), grant number 1129319N.