Off-Policy and Off-Actor Actor-Critic with Bootstrapped Dual Policy Iteration

D. Steckelmacher | H. Plisnier | D. M. Roijers | A. Nowé

dsteckel@vub.be | helenis@vub.be | dmo@vub.be | annew@vub.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 1/8.

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 $Q^\pi$, the actor quickly converges to the optimal policy:

$$n = E_{Q \sim Q^\pi}(G(Q)) = E_{Q \sim Q^\pi}(G(Q^\pi)) \rightarrow n^*$$

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

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

Experience Buffer (20 000)

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, $Q^\pi$ and $Q^\rho$.

$$Q_{k+1}(s_k, a_k) \leftarrow Q_k(s_k, a_k) + \alpha(n_k + V(V(s_{k+1}) - Q_k(s_k, a_k)))$$

$$V_{k+1}(s_k) = \min_{a_k, Q(k)^\pi(a_k)} Q^\pi(s_{k+1}, a_k) \text{argmax}_Q Q^\pi(s_{k+1}, a_k')$$

$$Q^\rho, Q^\pi \leftarrow Q_{k+1}, Q^\rho$$

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 $Q^\rho$ on $Q_{k+1}$ with the MSE loss, for 20 epochs
    - Swap $Q^\rho$ and $Q^\pi$

- Train the actor on the greedy policy to the critic

References


Acknowledgments

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