First-passage bias of learning in stochastic environments

M. A. Larchenko, V. V. Palyulin

Skolkovo Institute of Science and Technology, Moscow, Russia

Skoltech
Skolkovo Institute of Science and Technology

Abstract

Reinforcement Learning presents a promising way to control active particles [1]. In this work we study the impact of stochasticity on a learning process.

We use Q-learning with table Q-function to find the fastest path to a target. Environments are 1D and 2D gridworlds having regions with different noise intensity T. The noise is modelled as T random actions, performed after every agent's action.

Stable bias is observed in our systems. On realistic timescales, the bias leads to selection of suboptimal strategies, increasing presence of agents in regions with high noise.

We assume that stochastic dynamic allows a poorly trained agent to reach a target earlier. Consequently, first-passage properties of media affect learned strategy.

Effect of learning rate

The update rule of Q-learning is governed by learning rate α, exploration-exploitation parameter ε and discount rate γ [3]

$$Q(s_t, a_t) := Q(s_t, a_t) + \alpha\big(R_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)\big)$$

The value $Q(s_t, a_t)$ is renewed in a cycle [2]

$$Q_{n+1} = Q_n + \alpha\big(R_n - Q_n\big) = (1-\alpha)^n Q_1 + \alpha \sum_{i=1}^{n} (1-\alpha)^{n-i} R_i$$

The higher α, the sooner an agent overwrites its previous experience. We expect that in stochastic dynamics there could be restriction on α, under which convergence to certain policies is possible.

2D simulation

An agent has 4 actions and moves on 10x10 grid, r_{step} = -1 and r_{target} = 100. Step penalty is not applied to additional random moves.

Noise level T of the red region can be varied. High learning rates prevent the algorithm from exploring this random area at $T > 0$, Figure 1. Low α results in opposite behaviour.

Presence of noise worsens the mean first passage time: 9.7% for $T = 3$. Learning time improves results slowly: 30k episodes give MFPT = 25.0 and 94% agents in random area, 300k – 19.5 and 80% respectively (T = 3, $\alpha = 0.1$).

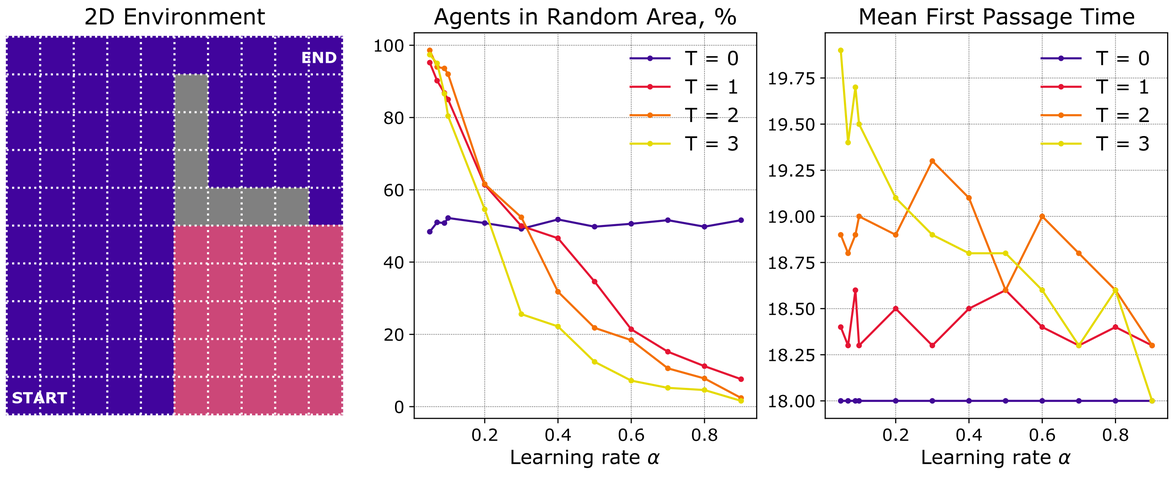
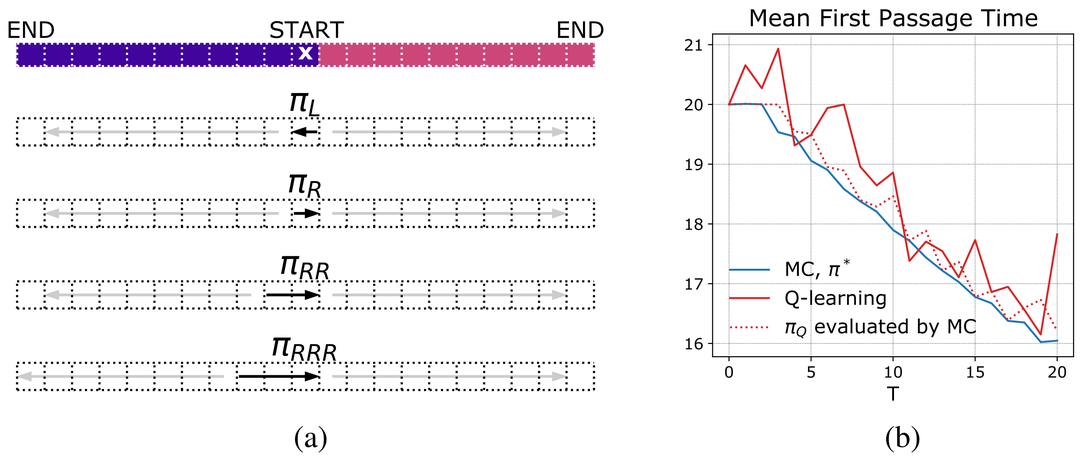


Figure 1: Q-learning, 500 agents, $\varepsilon = 0.1$, $\gamma = 0.9$, learning time = 300k episodes, performance was measured for greedy behaviour ($\varepsilon = 0.0$)

1D simulation

An agent starts in the center x_0 of the interval, consisting of 41 states, r_{step} = -1 and $r_{target} = 0$. Right-hand side $x > x_0$ has the noise level T.

It will be convenient to introduce some notation for policies, Figure 2a.



(a) (b)

Figure 2: (a) 1D environment and policies notation, (b) MC: 10^6 runs, Q-learning: 10^4 agents, $\alpha = 0.1$, $\varepsilon = 0.1$, $\gamma = 0.9$, learning time = 50k episodes

Policies of interest were evaluated by 10^6 Monte Carlo runs. The best found policy $\pi^\star$ depends on temperature T, π_Q represents the most common policy of 10^4 Q-learning agents (π_0 means that π_L and π_R have the same value)

$$\pi^\star = \begin{cases} \pi_0, & \text{for T} \in [0, 2] \\ \pi_R, & \text{for T} \in [3, 9] \\ \pi_{RR}, & \text{for T} = 10 \end{cases} \qquad \pi_Q = \begin{cases} \pi_0, & \text{for T} = 0 \\ \pi_R, & \text{for T} = 1, 2 \\ \pi_{RR}, & \text{for T} \in [3, 8] \\ \pi_{RRR}, & \text{for T} = 9, 10 \end{cases}$$

Nearly 100% of agents turn right one cell before it yields lower cost, Table 1. Evaluation of π_Q shows that it differs from $\pi^\star$ by 1.5%, Figure 2b.

In the second part of the simulation we have employed a drift. Its purpose is to gradually make π_R policy less profitable. The drift occurs only in the last quarter of the interval and is defined by a probability to make a left move.

Table 2 shows that $2/3$ of agents follow π_R despite 7.5% worse score, $1/4$ of them follow π_R when it is 12% worse than π_L (for T=3).

T	the most common, π_Q	R actions,%		
		x_0	$x_0 - 1$	$x_0 - 2$
0	π_0	50	0	0
1	π_R	99	0	0
2	π_R	97	0	0
3	π_{RR}	100	75	0
4	π_{RR}	100	91	0
5	π_{RR}	100	96	1
6	π_{RR}	100	100	1
7	π_{RR}	100	100	50
8	π_{RR}	100	100	41
9	π_{RRR}	100	100	95
10	π_{RRR}	100	100	80

Table 1: The most common policy derived from Q-learning values, 10^4 agents, α = 0.1, ε = 0.1, γ = 0.9, learning time = 50k episodes

T	drift	$\frac{\pi_R - \pi_L}{\pi_L}$, %	R actions in x_0, %
1	0.10	4.3	75
	0.15	6.7	31
	0.20	9.5	10
2	0.10	3.7	67
	0.15	5.9	36
	0.20	8.3	16
3	0.10	0.5	99
	0.15	2.2	96
	0.20	3.7	94
	0.25	5.9	86
	0.30	7.5	68
	0.35	9.2	42
	0.40	12.0	25

Table 2: Q-learning: 10^4 agents, α = 0.1, ε = 0.1, γ = 0.9, learning time = 50k episodes, MC: 10^4 runs.

Conclusions

- High learning rate prevents the algorithm from exploring stochastic media.
- For small enough α agents tend to go through noisy regions.
- Loss in performance can reach 7-10% for majority of agents and is more evident for higher dimensions (state-action set size).
- The performance of Q-learning is the highest among all tested algorithms (Double Q-learning, Expected SARSA and SARSA)
- Apparently, first-passage properties of media affect the policy selection.

References

[1] F. Cichos et al. "Machine learning for active matter". In: *Nature Machine Intelligence* 2.2 (2020), pp. 94–103.
[2] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.
[3] C. J. Watkins and P. Dayan. "Q-learning". In: *Machine learning* 8.3-4 (1992), pp. 279–292.