Prof. Dr. Leif Döring
André Ferdinand, Sara Klein

Reinforcement Learning

8. Excercise Sheet

1. Second version of Theorem 4.2.9 for SARSA

Show that the statement of Theorem 4.2.9 also holds if $\mathbb{E}[\varepsilon_n \mid \mathcal{F}_n] \neq 0$ but instead satisfies

$$\sum_{n=1}^{\infty} \alpha_i(n) \big| \mathbb{E}[\varepsilon_i(n) \mid \mathcal{F}_n] \big| < \infty \tag{1}$$

almost surely. It is enough to prove an improved version of Lemma 4.2.5 where the condition $\mathbb{E}[\varepsilon(t) \mid \mathcal{F}_t] = 0$ is replaced with

$$\sum_{n=1}^{\infty} \alpha(t) \big| \mathbb{E}[\varepsilon(t) \mid \mathcal{F}_t] \big| < \infty. \tag{2}$$

Apply the Robbins-Siegmund theorem to W^2 and use that $W \leq 1 + W^2$.

2. n-step TD

a) Write pseudocode for n-step TD algorithms for evaluation of V^π and Q^π in the non-terminating case and prove the convergence by checking the n-step Bellman expectation equations

$$T^\pi V(s) = \mathbb{E}_s^\pi \Big[R(s, A_0) + \sum_{t=1}^{n-1} \gamma^t R(S_t, A_t) + \gamma^n V(S_n) \Big]$$

and

$$T^\pi Q(s,a) = \mathbb{E}_s^{\pi^a} \Big[R(s, a) + \sum_{t=1}^{n-1} \gamma^t R(S_t, A_t) + \gamma^n Q(S_n, A_n) \Big]$$

and the conditions of Theorem 4.2.9 on the error term. Note that the algorithm only starts to update after the MDP ran for n steps. Can you also write down a version in the terminating case?

b) Write pseudocode for an n-step SARSA control algorithm in the non-terminating case. Try to prove convergence in the same way we did for 1-step SARSA in Theorem 4.3.6.

3. (Truncated, Clipped) Double Q-Learning

In this task we deal with Double Q-Learning, analyzing in particular the differences with Q-Learning.

(a) Implement algorithms 24 (Double Q-learning (with behavior policy)), 25 (Truncated Double Q-learning) and 26 (Clipped double Q-learning (with behavior policy)) of the lecture.

(b) Apply the algorithms to the Markov decision process from the last exercise (Constructed Max Bias).

(c) Compare the algorithms with the simple Q-learning algorithm. For example, apply similar metrics as in the lecture to the example (example 4.3.6).