FAKULTÄT FÜR WIRTSCHAFTSINFORMATIK
UND WIRTSCHAFTSMATHEMATIK

UNIVERSITÄT MANNHEIM

Prof. Dr. Leif Döring
Sara Klein, Benedikt Wille

Reinforcement Learning

4. Exercise Sheet

1. Markov Chains

Suppose that $(S_t)_{t\in\mathbb{N}}$ is a Markov Chain with values in some finite set $\mathcal{S}$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and denote by P the transition matrix. Assume further that $\mathbb{P}(S_n = s') > 0$ for some $n \in \mathbb{N}$ and $s' \in \mathcal{S}$ and define the probability measure $\tilde{\mathbb{P}}(\cdot \mid S_n = s')$. Prove that the shifted process $(\tilde{S}_t)_{t\in\mathbb{N}} = (S_{t+n})_{t\in\mathbb{N}}$ is again a Markov chain with transition matrix P.

2. Proof of Theorem 3.1.3

Complete the Proof of Theorem 3.1.3 in the lecture:

a) Show that defining $\mathbb{P}_T$ on the singletons

$$\mathbb{P}_T(\{(s_0, r_0, a_0, \dots, s_T, r_T, a_T)\}) := \mu(\{s_0\}) \cdot \pi_0(\{a_0\}\,;\, s_0) \cdot \prod_{i=1}^{T} p(\{(s_i, r_{i-1})\}\,;\, s_{i-1}, a_{i-1})$$
$$\cdot\, \pi_i(\{a_i\}\,;\, s_0, a_0, \dots, a_{i-1}, s_i) \cdot p(\mathcal{S} \times \{r_T\}; s_T, a_T).$$

yields a probability measure.

b) Check the claimed conditional probability identity

$$\mathbb{P}^\pi_\mu(S_{t+1} = s_{t+1}, R_t = r_t \,|\, S_t = s, A_t = a) = p(s_{t+1}, r_t\,;\, s, a).$$

where we assume that $\mathbb{P}^\pi_\mu(S_t = s, A_t = a) > 0$.

3. Probabilistic interpretation of MDPs

Use the formal definition of the stochastic process (S, A, R) to check that

$$\begin{aligned} p(s'; s, a) &= \mathbb{P}(S_{t+1} = s' \,|\, S_t = s, A_t = a), \\ p(r; s, a) &= \mathbb{P}(R_t = r | S_t = s, A_t = a), \\ r(s, a) &= \mathbb{E}[R_t \,|\, S_t = s, A_t = a], \\ r(s, a, s') &= \mathbb{E}[R_t \,|\, S_t = s, A_t = a, S_{t+1} = s'] \end{aligned}$$

holds if the events in the condition have positive probability, for any $t \in \mathbb{N}_0$.

4. Proof of Proposition 3.1.13

Prove that if $\pi \in \Pi_S$ then $(S_t, A_t, R_t)_{t\in\mathbb{N}}$ satisfies the Markov reward process property

$$\begin{aligned}\mathbb{P}((S_{t+1}, A_{t+1}, R_{t+1}) = (s_{t+1}, a_{t+1}, r_{t+1}) \mid (S_t, A_t) = (s_t, a_t), \ldots, (S_0, A_0) = (s_0, a_0)) \\ = \mathbb{P}((S_{t+1}, A_{t+1}, R_{t+1}) = (s_{t+1}, a_{t+1}, r_{t+1}) \mid (S_t, A_t) = (s_t, a_t))\end{aligned}$$

with time-homogeneous state/reward transition probabilities

$$p_{(s.a),(S',a',r')} = p(s', r; s, a)\pi(a'; s')$$

5. Programming task, Gumbel trick extended

On the previous exercise sheet, we looked at the Boltzmann Exploration and the relationship between the Softmax distribution and the Gumbel distribution. In this exercise you should try other random variables instead of the Gumbel distribution for the algorithmus and think about what a good or bad choice of random variables can be.
The following approach might be useful.

(a) Implement the algorithm 1 more generally by allowing them to pass any random variable instead of the Gumbel distribution. This code might be useful.

(b) Look in scipy.stats to see which random variables are implemented.

(c) Try the following random variables: Cauchy, Beta, Betaprime, Chi

Input : Initialization $\hat{Q}_a(0), a \in \mathcal{A}$, number of total timesteps n, number of arms $k \in \mathbb{N}$ and parameter $C \in \mathbb{R}$

Output: Trajectory of Rewards $(X_t)_{t\in\{1,\ldots,n\}}$ and actions $(A_t)_{t\in\{1,\ldots,n\}}$

begin

  for $t \leftarrow 1$ **to** n **do**

    Simulate $z_a, a \in \mathcal{A}$ independently identically standard Gumbel;

    Set $a_t = \arg\max_{a\in\mathcal{A}} \{\hat{Q}_a(t) + \sqrt{\frac{C^2}{N_a}} z_a\}$;

    Obtain reward x_t by playing arm a_t;

    Set $N_{a_t} = N_{a_t} + 1$;

    Update the estimated action value functions $\left(\hat{Q}_a(t)\right)$;

  end

  return $(X_t)_{t\in\{1,\ldots,n\}}, (A_t)_{t\in\{1,\ldots,n\}}$;

end

Algorithm 1: Boltzmann exploration algorithm modified