Prof. Dr. Leif Döring
André Ferdinand, Sara Klein

Reinforcement Learning

9. Excercise Sheet

1. Proofs for T-step MDPs

Prove the following claims from the lecture by comparing with the discounted counterpart.

a) Proposition 3.4.4: Given a Markovian policy $\pi = (\pi_t)_{t \in D}$ and a T-step Markov decision problem. Then the following relation between the state and state-action value function hold

$$V_t^\pi(s) = \sum_{a \in \mathcal{A}_s} \pi_t(a\,;\,s) Q_t^\pi(s,a),$$
$$Q_t^\pi(s,a) = r(s,a) + \sum_{s' \in \mathcal{S}} p(s'\,;\,s,a) V_{t+1}^\pi(s')$$

for all $t < T$. In particular (plugging-in), the Bellman expectation equations

$$V_t^\pi(s) = \sum_{a \in \mathcal{A}} \pi_t(a\,;\,s) \Big[r(s,a) + \sum_{s' \in \mathcal{S}} p(s'\,;\,s,a) V_{t+1}^\pi(s') \Big],$$
$$Q_t^\pi(s,a) = r(s,a) + \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}_s} p(s'\,;\,s,a) \pi_t(a\,;\,s') Q_{t+1}^\pi(s',a')$$

hold.

b) Lemma 3.4.6: The following holds for the optimal time-state value function and the optimal time-state-action value function for any $s \in \mathcal{S}$:

(i) $V_t^*(s) = \max_{a \in \mathcal{A}_s} Q_t^*(s,a)$ for all $t \leq T-1$.

(ii) $Q_t^*(s,a) = r(s,a) + \sum_{s' \in \mathcal{S}} p(s'\,;\,s,a) V_{t+1}^*(s')$ for all $t < T-1$.

2. Example: T-step MDPs

Recall the Ice Vendor example from the lecture. Assume the maximal amount of ice cream is $m = 3$ and the damand distribution is given by $\mathbb{P}(D_t = d) = p_d$ with $p_0 = p_2 = \frac{1}{4}, p_1 = \frac{1}{2}$. Suppose the revenue function f, ordering cost function o and storage cost function h are given by

$$f : \mathbb{N}_0 \to \mathbb{R},\ x \mapsto 9x,$$
$$o : \mathbb{N}_0 \to \mathbb{R},\ x \mapsto 2x,$$
$$h : \mathbb{N}_0 \to \mathbb{R},\ x \mapsto 2 + x.$$

a) Set up the transition matrix $p(s_{t+1}; s_t, a_t)$ in a table, such that every $s_t + a_t$ maps to the probability to land in s_{t+1}, and the reward function $r(s_t, a_t, s_{t+1})$ for this example.

b) Calculate the expected reward $r(s, a)$ for every state action pair. Can you guess an optimal strategy for a one time step MDP?

c) Suppose now you can play a 3-step MDP, hence you can order ice cream 4 times in $t = 0, 1, 2$. What is the optimal strategy for this finite time horizion MDP? Calculate the optimal state value, state-action value functions and the optimal policies using the greedy policy improvement algorithm from the lecture.
Hint: Use backward induction.

3. Multi Step Approximate Dynamic Programming

a) Implement Algorithm 26 of the lecture (First visit Monte Carlo for non-terminating MDPs).

b) Implement Algorithm 27 (First visit λ-return algorithm) of the lecture.

c) Implement Algorithm 28 (Offline TD(λ) policy evaluation with first-visit updates) of the lecture.