Fakultät für Wirtschaftsinformatik
und Wirtschaftsmathematik

Universität
Mannheim

Prof. Dr. Leif Döring
André Ferdinand, Sara Klein

Reinforcement Learning

5. Excercise Sheet

1. Bellman expecation operator.

Recall the Bellman expectation operator for a stationary policy $\pi \in \Pi_s$:

$$\begin{aligned}(T^\pi u)(s) &= \sum_{a\in\mathcal{A}} r(s,a)\pi(a\,;\,s) + \gamma \sum_{s'\in\mathcal{S}} \mathbb{P}^\pi(S_1 = s'|S_0 = s)u(s') \\ &= \sum_{a\in\mathcal{A}} \pi(a\,;\,s)\Big(r(s,a) + \gamma \sum_{s'\in\mathcal{S}} p(s';\,s,a)u(s')\Big)\end{aligned}$$

Show that we can rewrite the fixed point equation in vector notation, i.e. check that indeed $T^\pi V = V$ is equivalent to $r_\pi + \gamma P_\pi V = V$, where

$$\begin{aligned}P_\pi &= \Big(\sum_{a\in\mathcal{A}} \pi(a\,;\,rts)p(s'\,;\,s,a)\Big)_{(s,s')\in\mathcal{S}\times\mathcal{S}} \\ r_\pi &= \Big(\sum_{a\in\mathcal{A}} \pi(a\,;\,s)r(s,a)\Big)_{s\in\mathcal{S}}.\end{aligned}$$

2. Exact policy iteration

Consider two types of costumers, L for low and H for high, shopping in a shopping center. Each quarter the manager devides all costumers into these classes based in their purchase behavior in the previous quarter. The manager wishes to determine to which classes of costumers he should send quarterly catalogs. Sending a catalog costs him \$15 per costumer. If a costumer recieved a catalog at the beginning of the quarter and is in class L at the subsequent quarter, then the expected purchase is \$20, and \$10 if he did not recieve a catalog. If a costumer is in class H at the subsequent quarter and recived a catalog, then the expected purchase is \$50, and \$25 if he did not. The decision wheather or not to send a catalog to a customer also affects the customer's classification in the subsequent quarter: If a costumer is in class L at the start of the present quarter, then the probability to stay in class L in the subsequent quarter is 0.3 if he recieves a catalog and 0.5 if he does not. For the class H costumer the probability to stay in H for the subsequent quarter is 0.8 if he recieves a catalog and 0.4 if he does not.
Assume that the discount rate is 0.9 and the manager wants to maximize the expected total discounted reward.

a) Formulate this problem as discounted infinte-horizion Markov decision model.

b) What is the expected one-step reward for every state-action pair? Define the stationary policy which has greatest one-step reward.

c) Find an optimal policy using the greedy exact policy iteration (algorithm 11) to find the optimal policy. Start with the stationary policy from b).

3. **Programming Task: Policy Iteration**

In this programming task we want to program a first game, as well as an agent that plays the game. In addition, we want to use the algorithms from the lecture to optimize the policy of the agent.

(a) Implement the Grid World game from the lecture (example 1.1.9). Create a class that can be used to play the game. This code might be helpful.

(b) Implement a class player that has a decision rule, such that the game from the previous task can be played.

(c) Use the generalised policy iteration paradigm (algorithm 12) to find the optimal policy for the game. For policy evaluation use Algorithm 8 and for policy improvement use Algorithm 10.