FAKULTÄT FÜR WIRTSCHAFTSINFORMATIK
UND WIRTSCHAFTSMATHEMATIK

UNIVERSITÄT MANNHEIM

Prof. Dr. Leif Döring
André Ferdinand, Sara Klein

Reinforcement Learning

5. Solution 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

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

Solution:

$$\begin{aligned}V &= T^\pi V \\ &= \Big(\sum_{a\in\mathcal{A}} \pi(a\,;\,s)\Big(r(s,a) + \gamma \sum_{s'\in\mathcal{S}} p(s';\,s,a)V(s')\Big)\Big)_{s\in\mathcal{S}} \\ &= \Big(\sum_{a\in\mathcal{A}} \pi(a\,;\,s)r(s,a)\Big)_{s\in\mathcal{S}} + \gamma\Big(\sum_{s'\in\mathcal{S}}\sum_{a\in\mathcal{A}} \pi(a\,;\,s)p(s';\,s,a)V(s')\Big)_{s\in\mathcal{S}} \\ &= r_\pi + \gamma\Big(\sum_{s'\in\mathcal{S}} P_\pi(s,s')V(s')\Big)_{s\in\mathcal{S}} \\ &= r_\pi + \gamma P_\pi V.\end{aligned}$$

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 received 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 receive a catalog. If a costumer is in class H at the subsequent quarter and received 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 receives 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 receives 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.
Solution:

- $\mathcal{S} = \{L, H\}$ *is the state space,*
- $\mathcal{A} = \{C, NC\}$ *is the action space, where* C *stands for catalog and* NC *for no catalog,*
- $\mathcal{R} = \{5, 10, 35, 25\}$ *is the reward space,*
- *the transition function* $p(\{s', r'\}; s, a)$ *is given by*

$(s,a)\backslash(s',r')$	$L,5$	$L,10$	$H,35$	$H,25$	$L,35$	$L,25$	$H,5$	$H,10$
L,C	0.3	0	0.7	0	0	0	0	0
L,NC	0	0.5	0	0.5	0	0	0	0
H,C	0.2	0	0.8	0	0	0	0	0
H,NC	0	0.6	0	0.4	0	0	0	0

For example if costumer L *receives a catalog* $((s,a) = (L,C))$ *then he stays in class* L *with probability* 0.3 *and thus has an expected purchase of* $25 - 15 = 5$ $((s',r') = (L,5))$ *or he changes to class* H *with probability* 0.7 *and then has an expected purchase of* $50 - 15 = 35$ $((s',r') = (H,35))$*. All other combinations of* (s',r') *have zero probability.*

b) What is the expected one-step reward $r(s,a)$ for every state-action pair? Define the stationary policy which has greatest one-step reward.
Solution:
The one step reward is defined by $r(s,a) = \sum_{s',r'} r' p(\{s',r'\}; s, a)$*. So we have*

$$r(L,C) = 0.3 \cdot 5 + 0.7 \cdot 35 = 26, \quad r(L,NC) = 0.5 \cdot 10 + 0.5 \cdot 25 = 17.5$$
$$r(H,C) = 0.2 \cdot 5 + 0.8 \cdot 35 = 29, \quad r(H,NC) = 0.6 \cdot 10 + 0.4 \cdot 25 = 16$$

Thus, the stationary policy

$$\pi(C; L) = 1, \quad \pi(NC; L) = 0$$
$$\pi(C; H) = 1, \quad \pi(NC; H) = 0,$$

maximizes the 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).
Solution:
We choose $\pi^0 = \pi$ *from b) and first have to calculate* $V^{\pi^0} = (I - \gamma P_{\pi^0})^{-1} r_{\pi^0}$ *By the definition of* π^0 *we see that* $r_{\pi^0} = \left(\sum_{a \in \mathcal{A}} \pi^0(a\,;\, s) r(s,a) \right)_{s \in \mathcal{S}}$ *is given by*

$$r_{\pi^0} = \begin{pmatrix} 26 \\ 29 \end{pmatrix}.$$

For $P_{\pi^0}(s, s') = \sum_{a \in \mathcal{A}} \pi^0(a\,;\, s) p(s'\,;\, s, a)$ *we have*

$$P_{\pi^0} = \begin{pmatrix} 0.3 & 0.7 \\ 0.8 & 0.2 \end{pmatrix}.$$

Thus,

$$\begin{aligned}(I - \gamma P_{\pi^0})^{-1} &= \begin{pmatrix} 1 - 0.9 \cdot 0.3 & -0.9 \cdot 0.7 \\ -0.9 \cdot 0.8 & 1 - 0.9 \cdot 0.2 \end{pmatrix}^{-1} \\ &= \frac{1}{0.73 \cdot 0.82 - (-0.27) \cdot (-0.72)} \begin{pmatrix} 0.82 & 0.27 \\ 0.72 & 0.73 \end{pmatrix} \\ &= \begin{pmatrix} 2.03 & 0.67 \\ 1.78 & 1.81 \end{pmatrix}.\end{aligned}$$

Now we can caluclate V^{π^0} *by*

$$V^{\pi^0} = \begin{pmatrix} 2.03 & 0.67 \\ 1.78 & 1.81 \end{pmatrix} \begin{pmatrix} 26 \\ 29 \end{pmatrix} = \begin{pmatrix} 72.21 \\ 98.77 \end{pmatrix}.$$

We continue with the next step in the algorithm and calculate $Q^{\pi^0}(s, a) = r(s, a) + \gamma \sum_{s'} p(s'; s, a) V^{\pi^0}(s')$*:*

$$\begin{aligned} Q^{\pi^0}(L, C) &= 26 + 0.9(0.3 \cdot 72.21 + 0.7 \cdot 98.77) = 107.72 \\ Q^{\pi^0}(L, NC) &= 17.5 + 0.9(0.5 \cdot 72.21 + 0.5 \cdot 98.77) = 94.44 \\ Q^{\pi^0}(H, C) &= 29 + 0.9(0.2 \cdot 72.21 + 0.8 \cdot 98.77) = 113.11 \\ Q^{\pi^0}(H, NC) &= 16 + 0.9(0.6 \cdot 72.21 + 0.4 \cdot 98.77) = 90.55 \end{aligned}$$

We follow the policy

$$\begin{aligned} \pi^1(C; L) = 1, \quad \pi^1(NC; L) = 0 \\ \pi^1(C; H) = 1, \quad \pi^1(NC; H) = 0, \end{aligned}$$

and see that $\pi^0 = \pi^1$*. Hence the algorithm is terminated and the optimal policy is* $\pi^0 = \pi^1$*.*