FAKULTÄT FÜR WIRTSCHAFTSINFORMATIK
UND WIRTSCHAFTSMATHEMATIK

UNIVERSITÄT
MANNHEIM

Prof. Dr. Leif Döring
Sara Klein, Benedikt Wille

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. Markov Decision Process 1

A rental car agency serves two cities with one office in each. The agency has M cars in total. At the start of each day, the manager must decide how many cars to move from one office to the other to balance stock. Let $f_i(q)$ for $i = 1,2$ denote the probability that the daily demand for cars to be picked up at city i and returned to city i equals q. Let $g_i(q)$, $i = 1,2$, denote the probability that the daily demand for "one-way" rentals from city i to the other equals q. Assume that all rentals are for one day only, that it takes one day to move cars from city to city to balance stock, and, if demand exceeds availability, customers go elsewhere. The economic parameters include a cost of K per car for moving a car from city to city to balance stock, a revenue of R per car for rentals returned to the rental location and R per car for one-way rentals.

Formulate this problem as discounted infinite-horizon Markov decision model.

Solution:

Since we know that there are a total of M cars we only need to describe the number of cars that are at one of the cities, say city 1, in every time-step and the number of cars at the other city will follow. So we define the state space as $\mathcal{S} := \{0, \dots, M\}$. Similarly, we only need to know the number of cars moved from one city (or transported to said city), say city 1, to (resp. from) the other one. We cannot move more cars from city 1 to city 2 than there are currently in city 1 in each period. Similarly at city 1 we cannot receive more cars from city 2 than there are currently in city 2 in each period. Therefore for each state $s \in \mathcal{S}$ we define $\mathcal{A}_s := \{(M-s), \dots, s\}$. The whole action-space is therefore $\mathcal{A} := \{-M, \dots, M\}$. For notational simplicity we further define the demands $(D_{1O}, D_{1R}, D_{2O}, D_{2O}) \sim g_1 \otimes f_1 \otimes g_2 \otimes f_2$, where O stands for "one-way" and R for "returned". We recognize that since it takes the whole day to move cars from city 1 to city 2, the number of cars available for rent at city 1 in some state $(s,a) \in \mathcal{S} \times \mathcal{A}_s$ is $s-a$ whenever $a > 0$ (i.e. if we decide to transfer a cars from city 1 to city 2) and it is s whenever $a \leq 0$ (i.e. if we decide to transfer $|a|$ cars from city 2 to city 1). For the same reason the number of available cars for rental in city 2 for the same state-action pair is $(M-s)+(a \vee 0)$. Furthermore we can define the number of rentals by category in a given period in dependence of the current state-action pair as $(X_{1O}, X_{1R}, X_{2O}, X_{2O})(s,a)$. Next, we will compute the distribution of this probability vector. We assume that the demands are given at the beginning of each period[*] *and that we choose to first accommodate "one-way" rental demand*[†]*. Since the number of rentals in each city only depends on the state-action pair and the demand in the city and all demands are independent we can follow that (X_{1O}, X_{1R}) is independent of (X_{2O}, X_{2R}). Let's first look at (X_{1O}, X_{1R}). There are three different cases of which values can be obtained for the pair. The first case is when the demand D_{1O} is bigger or equal than the cars available in city 1 in that period. In this case all cars will be rented for "one-way" rentals and there are* 0 *rentals that will be returned at city 1. If the demand D_{10} is smaller than the cars available in city 1 there are 2 sub cases. One in which the demand D_{1R} is bigger than the rest of the available cars, in which case the rest will be rented out, and one in which D_{1R} is smaller, in which case the number of rented cars exactly matches the demand. Therefore*

$$(X_{1O}, X_{1R})(s,a) = \begin{cases} (s-(a \wedge 0), 0) & , D_{1O} \geq s-(a \wedge 0) \\ (D_{1O}, s-(a \wedge 0) - D_{1O}) & , D_{1O} < s-(a \wedge 0),\ D_{1R} \geq s-(a \wedge 0) - D_{1O} \\ (D_{1O}, D_{1R}) & , D_{1O} < s-(a \wedge 0),\ D_{1R} < s-(a \wedge 0) - D_{1O} \end{cases}$$

[*] If this is not given we have no chance of knowing the order in which costumers come, but if we accept costumers as long as we have stock, the order highly influences the next state.

[†] We can do this since the revenue in both cases is the same. If we assume the opposite, the distribution of the number of rentals vector will change and therefore the solution to the model may be different but the revenue should be the same.

for all state-action pairs $(s, a) \in \mathcal{S} \times \mathcal{A}_s$. *Analogously, it holds*

$$
\begin{aligned}
&(X_{2O}, X_{2R})(s, a) = \\
&\begin{cases}
(M - s + (a \vee 0), 0) & , D_{2O} \geq M - s + (a \vee 0) \\
(D_{2O}, M - s + (a \vee 0) - D_{2O}) & , D_{2O} < M - s + (a \vee 0),\ D_{2R} \geq M - s + (a \vee 0) - D_{2O} \\
(D_{2O}, D_{2R}) & , D_{2O} < M - s + (a \vee 0),\ D_{2R} < M - s + (a \vee 0) - D_{2O}
\end{cases}
\end{aligned}
$$

for all state-action pairs $(s, a) \in \mathcal{S} \times \mathcal{A}_s$. *Keeping in mind that the demands are independent and that in the cases where the demands are higher than the number of cars left in the city we can simply sum the probabilities we can write down the distributions of these random variables:*

$$
\begin{aligned}
&h_1(x_O, x_R)(s, a) = \\
&\begin{cases}
\sum_{k \geq s - (a \wedge 0)} g_1(k) & , x_O = s - (a \wedge 0),\ x_R = 0 \\
g_1(x_O) \sum_{k \geq s - (a \wedge 0) - x_O} f_1(k) & , 0 \leq x_O < s - (a \wedge 0),\ x_R = s - (a \wedge 0) - x_O \\
g_1(x_0) f_1(x_R) & , 0 \leq x_O < s - (a \wedge 0),\ 0 \leq x_R < s - (a \wedge 0) - x_O
\end{cases} \quad \text{, and} \\
&h_2(x_O, x_R)(s, a) = \\
&\begin{cases}
\sum_{k \geq M - s + (a \vee 0)} g_2(k) & , x_O = M - s + (a \vee 0),\ x_R = 0 \\
g_2(x_O) \sum_{k \geq M - s + (a \vee 0) - x_O} f_2(k) & , 0 \leq x_O < M - s + (a \vee 0),\ x_R = M - s + (a \vee 0) - x_O \\
g_2(x_0) f_2(x_R) & , 0 \leq x_O < M - s + (a \vee 0),\ 0 \leq x_R < M - s + (a \vee 0) - x_O
\end{cases}
\end{aligned}
$$

for all state-action pairs $(s, a) \in \mathcal{S} \times \mathcal{A}_s$. *The revenue generated in dependence of the state-action pair and on how many cars are rented per category is simply the number of cars times the revenue substracted by the costs of moving* $|a|$ *many cars:*

$$R(s, a)(x_{1O}, x_{1R}, x_{2O}, x_{2R}) = R(x_{1O} + x_{1R} + x_{2O} + x_{2R}) - K * |a|.$$

The number of cars in city one after a period, i.e. the next state, changes by the number of cars transferred to or received from city 2 and the number of "one-way" rentals from and to city 1:

$$S'(s, a)(x_{1O}, x_{1R}, x_{2O}, x_{2R}) = s - a - x_{1O} + x_{2O}.$$

It only remains to give the transition kernel and we are done:

$$
\begin{aligned}
p(\{s'\} \times \{r\}\ \ s, a) &= \sum_{x_{i,j}} \delta_{\{s' = S'(s,a)(x_{1O}, x_{1R}, x_{2O}, x_{2R})\}} \delta_{\{r = R(s,a)(x_{1O}, x_{1R}, x_{2O}, x_{2R})\}} \\
&\qquad \cdot \mathbb{P}((X_{1O}, X_{1R}, X_{2O}, X_{2O}) = (x_{1O}, x_{1R}, x_{2O}, x_{2R})) \\
&= \sum_{x_{i,j}} \delta_{\{s' = S'(s,a)(x_{1O}, x_{1R}, x_{2O}, x_{2R})\}} \delta_{\{r = R(s,a)(x_{1O}, x_{1R}, x_{2O}, x_{2R})\}} \\
&\qquad \cdot h_1(x_{1O}, x_{1R}) h_2(x_{2O}, x_{2R})(s, a),
\end{aligned}
$$

where we keep in mind that (X_{1O}, X_{1R}) *is independent of* (X_{2O}, X_{2R}).

3. Markov Decision Process 2

Two identical machines are used in a manufacturing process. From time to time, these machines require maintenance which takes three weeks. Maintenance on one machine costs c, per week while maintenance on both machines costs c_2 per week. Assume $c_2 > 2c$. The probability that a machine breaks down if it has been i periods since its last maintenance is p_i,with p_i non-decreasing in i. Maintenance begins at the start of the week immediately following breakdown, but preventive maintenance may be started at the beginning of any week. The decision maker must choose when to carry out preventive maintenance, if ever, and, if so, how many machines to repair. Assume the two machines fail independently.

a) What is the significance of the condition $c_2 > 2c$?
Solution:
If $c_2 \leq 2c$, then the maintenance of both machines together would be less expensive than a separate maintenance. Hence, the optimal policy would be to wait until the machines break down and only repair them then, since there is no disadvantage if both machines fail simultaneously.

b) Formulate this problem as discounted infinite-horizon Markov decision model.
Solution:
If the machine is not yet in maintenance, we need to know the time period since the last maintenance and if the machine is in maintenance, we need to know when maintenance started to guarantee that it lasts exactly three weeks. So we define the state space $\mathcal{S}_i = \{-2, -1, 0, 1, 2, 3, \dots\}$ for machine i, where the states $\{-2, -1, 0\}$ mean that machine i is in maintenance in the first, second, third week respectively. The other states represent the periods since the last maintenance.
The overall state space is then $\mathcal{S} = \mathcal{S}_1 \times \mathcal{S}_2$.
We have two possible actions for each machine, sending them into maintenance ($a_i = 1$) or not ($a_i = 0$). The action space is therefore $\mathcal{A} = \{0, 1\} \times \{0, 1\}$ also two-dimensional.
For the transition probabilities, we can treat each machine separately due to independence of the breakdown probabilities. Moreover, we will just define the transition probability into the next state, instead of the next state-reward pair. We define the transition function $p_1(\{s'\}; s, a) = p_2(\{s'\}; s, a)$ by

$(s,a)\backslash s'$	-2	-1	0	1	$i+1$
$(-2,a)$	0	1	0	0	0
$(-1,a)$	0	0	1	0	0
$(0,a)$	0	0	0	1	0
$(i,1)$	1	0	0	0	0
$(i,0)$	p_i	0	0	0	$1-p_i$

Where i in the table stands for any natural number $i \in \{1, 2, 3, \dots\}$.

The expected reward in a state-action pair is given by

$$r(s,a) = r((s_1,s_2),(a_1,a_2)) = r((s_1,s_2))$$
$$= \begin{cases} -c, & s_i \in \{-2,-1,0\}, s_j \in \{1,2,\dots\} \textit{ for } i \neq j \textit{ and } i,j \in \{1,2\} \\ -c_2, & s_1,s_2 \in \{-2,-1,0\} \\ 0, & s_1,s_2 \in \{1,2,\dots\}. \end{cases}$$

Note here, that the reward is independent of the action and the next state and only depends on the current state!
Given the transition probabilities p_i and the expected rewards we can define the transition kernel

$$p\left(\{(s_1',s_2')\} \times \{r\}\,;\,(s_1,s_2),(a_1,a_2)\right) = \delta_{\{r=r((s_1,s_2),(a_1,a_2))\}} \cdot p_1(s_1'\,;\,s_1,a_1)p_2(s_2'\,;\,s_2,a_2).$$

With this we have defined a (discounted) Markov decision model.

c) Provide an educated guess about the form of the optimal policy when the decision maker's objective is to minimize expected operating cost.
Solution:
Due to the independence of the breakdown probabilities, we expect that the optimal policy is symmetric with respect to the states of the two machines.

4. Banach fixed-point Theorem

Prove Banach fixed-point theorem (Theorem 3.1.17 in the lecture notes).
Banach Fixed-Point Theorem (Banach Space version): Let $(U, ||\cdot||)$ be a Banach space, and let $T : U \to U$ be a mapping such that for some $\lambda \in [0,1)$, we have

$$||T(x) - T(y)|| \leq \lambda ||x-y||, \quad \forall x,y \in U.$$

Then, T has a unique fixed point x^* in U, such that $Tx^* = x^*$. Moreover, for arbitrary $x_0 \in U$, the sequence $(x_n)_{n\in\mathbb{N}}$ defined by

$$x_{n+1} = Tx_n = T^{n+1}x_0$$

convergences in U to x^*.
Proof: *Consider the sequence (x_n) defined recursively by $x_0 \in U$ arbitrary, and $x_{n+1} = T(x_n)$ for $n \geq 0$.*
We first prove that (x_n) is a Cauchy sequence. For $n > 0$, we have:

$$\begin{aligned} ||x_{n+1} - x_n|| &= ||T(x_n) - T(x_{n-1})|| \\ &\leq \lambda ||x_n - x_{n-1}|| \\ &\leq \lambda^2 ||x_{n-1} - x_{n-2}|| \\ &\leq \dots \\ &\leq \lambda^n ||x_1 - x_0||. \end{aligned}$$

Now, for $n, m > 0$, we have that

$$\begin{aligned}
||x_{n+m} - x_n|| &\leq ||x_{n+m} - x_{n+(m-1)}|| + ||x_{n+(m-1)} - x_{n+(m-2)}|| + \cdots + ||x_{n+1} - x_n|| \\
&\leq \lambda^{m-1}||x_{n+1} - x_n|| + \lambda 1m - 2||x_{n+1} - x_n|| + \cdots + \lambda^0||x_{n+1} - x_n|| \\
&= \sum_{k=0}^{m-1} \lambda^k ||x_{n+1} - x_n|| \\
&\leq \lambda^n ||x_1 - x_0|| \sum_{k=0}^{m-1} \lambda^k \\
&\leq ||x_1 - x_0|| \lambda^n \sum_{k=0}^{\infty} \lambda^k \\
&< \frac{\lambda^n}{1-\lambda} ||x_1 - x_0||.
\end{aligned}$$

As $\lambda < 1$ we follow that

$$\frac{\lambda^n}{1-\lambda}||x_1 - x_0|| \to 0, \quad n \to \infty.$$

This shows that (x_n) is a Cauchy sequence.
Since U is a Banach space, there exists $x \in U$ such that $x_n \to x$ as $n \to \infty$. Now, as T is continuous, we have:

$$T(x) = T(\lim_{n\to\infty} x_n) = \lim_{n\to\infty} T(x_n) = \lim_{n\to\infty} x_{n+1} = x,$$

where the second equality follows from the continuity of T.
We have shown that there exists a fix point in U and that for every x_0 the sequence $(x_n)_{n\in\mathbb{N}}$ defined by

$$x_{n+1} = Tx_n = T^{n+1}x_0$$

convergences in U to some fixed point $x \in U$. It remains to show that the fixed point is unique.
Finally, suppose $y \in U$ is another fixed point of T. Then,

$$||x - y|| = ||T(x) - T(y)|| \leq \lambda||x - y||,$$

which implies $||x - y|| = 0$ since $\lambda < 1$. Therefore, $x = y$, proving the uniqueness of the fixed point.