FAKULTÄT FÜR WIRTSCHAFTSINFORMATIK
UND WIRTSCHAFTSMATHEMATIK

UNIVERSITÄT MANNHEIM

Prof. Dr. Leif Döring
Sara Klein, Benedikt Wille

Reinforcement Learning

6. Solution Sheet

1. Policy evaluation

Consider Algorithm 7 from the lecture. In Theorem 3.3.2 we proved convergence for this algorithm if $\gamma < 1$. Now assume $\gamma = 1$ and set $\Delta = 2\epsilon$ in the initialisation and choose termination condition $\Delta < \epsilon$. Give an example such that Algorithm 8 does not converge using $\gamma = 1$. You are allowed to initialise the value function V arbitrarily.

Solution:

For example define $\mathcal{S} = \{0, 1\}$ *and* $\mathcal{A} = \{A, B\}$. *Furthermore, we assume that the reward* R_t *is deterministic given* S_t, A_t *and given by the function* $R(s, a)$ *with values*

$$R(0, A) = 1, \quad R(0, B) = 0$$
$$R(1, A) = 0, \quad R(1, B) = 1.$$

The transition probabilities are independent of the reward given in the following table

$p(s', s, a)$	0	1
$0, A$	1	0
$0, B$	0	1
$1, A$	1	0
$1, B$	0	1

We define the policy π *by*

$$\pi(A; 0) = 0, \quad \pi(B; 0) = 1,$$
$$\pi(A; 1) = 1, \quad \pi(B; 1) = 0$$

and initialise the value functions $V = V_{new}$ *by* $V(0) = 1$, $V(1) = 0$. *Furthermore, we choose* $\epsilon < 1$.

We start with the first loop and calculate

$$V_{new}(s) = \sum_{a \in \mathcal{A}} \pi(a; s) \Big(r(s, a) + \sum_{s' \in \mathcal{S}} p(s'; s, a) V(s)),$$

i.e we get

$$V_{new}(0) = \pi(B; 0)(R(0, B) + p(1; 0, B)V(1)) = 1(0 + 1 \cdot 0) = 0$$
$$V_{new}(1) = \pi(A; 1)(R(1, A) + p(0; 1, A)V(0)) = 1(0 + 1 \cdot 1) = 1.$$

We see that $\Delta = 1$ and thus we set $V = V_{new}$ and continue with net next loop:

$$V_{new}(0) = \pi(B;0)(R(0,B) + p(1;0,B)V(1)) = 1(0 + 1 \cdot 1) = 1$$
$$V_{new}(1) = \pi(A;1)(R(1,A) + p(0;1,A)V(0)) = 1(0 + 1 \cdot 0) = 0,$$

which results in our initial value function from the beginning. Hence, we see directly that the value iteration algorithm alternates between these two value functions and never converges.

2. Convergence of the in-place policy evaluation algorithm

Recall Algorithm 8 from the lecture. We aim to prove convergence of the algorithm (without termination) to V^π. Therefore, label the state space $\mathcal{S}$ by $s_1, \dots, s_K$ and define

$$T_s^\pi V(s') = \begin{cases} T^\pi V(s) & : s = s' \\ V(s) & : s \neq s' \end{cases}$$

Define the composition $\overline{T}^\pi : U \to U$, $\overline{T}^\pi(v) := \Big(T_{s_K}^\pi \circ \cdots \circ T_{s_1}^\pi\Big)(v)$ on the space of all functions $U = \{u : \mathcal{S} \to \mathbb{R}\}$ equipped with the supremum-norm.

a) Argue why $\overline{T}^\pi$ is different from the Bellman operator T^π.
Solution:
Applying the Bellman operator T^π updates the function v in every state s using the fixed values $v(s)$ for all s. More precisely, we store all values $v(s)$ for all $s \in \mathcal{S}$ and calculate

$$v_{new}(s) = \sum_{a \in \mathcal{A}} \pi(a;s)\Big(r(s,a) + \sum_{s' \in \mathcal{S}} p(s';s,a)v(s')\Big),$$

***FOR ALL** $s \in \mathcal{S}$ **and afterwards we set** $T^\pi v = v_{new}$.*
If we apply the operator $\overline{T}^\pi$, we first apply the operator $T_{s_1}^\pi$: For s_1 we get a new value $\overline{v}_{new}(s_1) = \sum_{a \in \mathcal{A}} \pi(a;s_1)\Big(r(s_1,a) + \sum_{s' \in \mathcal{S}} p(s';s_1,a)v(s')\Big)$ and for all other states we change nothing. We set $\overline{v_1}(s_1) = v_{new}(s_1)$, $\overline{v_1}(s_k) = v(s_k)$ for all $k > 1$ and continue with $T_{s_2}^\pi$. The next operator $T_{s_2}^\pi$ applies the Bellman operator at state s_2 and leave all other variables untouched. The fundamental change is now that we apply the Bellman operator at state s_2 for the vector $\overline{v_1}$ and not for v, i.e. we have $\overline{v}_{new}(s_2) = \sum_{a \in \mathcal{A}} \pi(a;s_2)\Big(r(s_2,a) + \sum_{s' \in \mathcal{S}} p(s';s_2,a)\overline{v_1}(s')\Big)$! Hence, $v_{new}(s_2)$ from the Bellman operator is different from $\overline{v}_{new}(s_2)$ which we get from the operator $\overline{T}^\pi$. We set $\overline{v_2}(s_k) = \overline{v_1}(s_k)$ for all $k \neq 2$ and $\overline{v_2}(s_2) = \overline{v}_{new}(s_2)$. We continue after this scheme and see that the operators are different.

b) Show that V^π is a fixed point of the operator $\overline{T}^\pi$.
Solution:
We have that $T_{s_i}^\pi$ only changes the i-th coordinate of the vector $v \in \mathbb{R}^{|S|}$. By induction, we show that $(\overline{T}^\pi)(V^\pi) = V^\pi$, by proving that V^π is a fixed point in every coordinate $s \in \mathcal{S}$.

So for s_1 we have

$$\begin{aligned}(\overline{T}^\pi)(V^\pi)(s_1) &= (T^\pi_{s_K} \circ \cdots \circ T^\pi_{s_1})(V^\pi)(s_1)\\ &= T^\pi_{s_1}(V^\pi)(s_1)\\ &= T^\pi(V^\pi)(s_1)\\ &= V^\pi(s_1),\end{aligned}$$

because V^π is a fixed point with respect to the Bellman operator. We see also from this calculation, that $(\overline{T}^\pi)(V^\pi) = (T^\pi_{s_K} \circ \cdots \circ T^\pi_{s_1})(V^\pi) = (T^\pi_{s_K} \circ \cdots \circ T^\pi_{s_2})(V^\pi)$.
Now we assume that $(\overline{T}^\pi)(V^\pi) = (T^\pi_{s_K} \circ \cdots \circ T^\pi_{s_{i+1}})(V^\pi)$, and $(\overline{T}^\pi)(V^\pi)(s_i)V^\pi(s_i)$ for fixed $i < K$, then for $i+1$ we get

$$\begin{aligned}(\overline{T}^\pi)(V^\pi)(s_{i+1}) &= (T^\pi_{s_K} \circ \cdots \circ T^\pi_{s_1})(V^\pi)(s_{i+1})\\ &= (T^\pi_{s_K} \circ \cdots \circ T^\pi_{s_{i+1}})(V^\pi)(s_{i+1})\\ &= (T^\pi_{s_{i+1}})(V^\pi)(s_{i+1})\\ &= T^\pi(V^\pi)(s_{i+1})\\ &= V^\pi(s_{i+1}).\end{aligned}$$

This proves the claim.

c) Prove that $\overline{T}^\pi$ is a contraction on $(U, ||\cdot||_\infty)$.
Solution:
Consider u and v in U, then

$$\begin{aligned}&\|(T^\pi_{s_K} \circ \cdots \circ T^\pi_{s_1})(u) - (T^\pi_{s_K} \circ \cdots \circ T^\pi_{s_1})(v)\|_\infty\\ &= \max_{i=1,\ldots,K} \Big\{\Big|(T^\pi_{s_K} \circ \cdots \circ T^\pi_{s_1})(u)(s_i) - (T^\pi_{s_K} \circ \cdots \circ T^\pi_{s_1})(v)(s_i)\Big|\Big\}\\ &= \max_{i=1,\ldots,K} \Big\{\Big|(T^\pi_{s_i} \circ \cdots \circ T^\pi_{s_1})(u)(s_i) - (T^\pi_{s_i} \circ \cdots \circ T^\pi_{s_1})(v)(s_i)\Big|\Big\}\\ &= \max\Big\{|T^\pi_{s_1}(u)(s_1) - T^\pi_{s_1}(v)(s_1)|, |T^\pi_{s_2}(\tilde{u}^{(1)})(s_2) - T^\pi_{s_2}(\tilde{v}^{(1)})(s_2)|, \ldots,\\ &\quad |T^\pi_{s_K}(\tilde{u}^{(K-1)})(s_K) - T^\pi_{s_K}(\tilde{v}^{(K-1)})(s_K)|\Big\},\end{aligned}$$

where $\tilde{u}^{(i)} := (T^\pi_{s_i} \circ \cdots \circ T^\pi_{s_1})(u)$. Then we can continue

$$\begin{aligned}&\|(T^\pi_{s_K} \circ \cdots \circ T^\pi_{s_1})(u) - (T^\pi_{s_K} \circ \cdots \circ T^\pi_{s_1})(v)\|_\infty\\ &= \max\Big\{|T^\pi(u)(s_1) - T^\pi(v)(s_1)|, |T^\pi(\tilde{u}^{(1)})(s_2) - T^\pi(\tilde{v}^{(1)})(s_2)|, \ldots,\\ &\quad |T^\pi(\tilde{u}^{(K-1)})(s_K) - T^\pi(\tilde{v}^{(K-1)})(s_K)|\Big\}\\ &\leq \max\Big\{\gamma\|u - v\|_\infty, \gamma\|\tilde{u}^{(1)} - \tilde{v}^{(1)}\|_\infty, \ldots, \gamma\|\tilde{u}^{(K-1)} - \tilde{v}^{(K-1)}\|_\infty\Big\}.\end{aligned}$$

By induction we will show that $\|\tilde{u}^{(i)} - \tilde{v}^{(i)}\|_\infty \leq \|u - v\|_\infty$ for all $i = 1, \ldots, K-1$.

First for $i = 1$ we have

$$\begin{aligned}
\|\tilde{u}^{(1)} - \tilde{v}^{(1)}\|_\infty &= \|T^\pi_{s_1}(u) - T^\pi_{s_1}(v)\|_\infty \\
&= \max_{i=1,\dots,K}\Big\{|T^\pi_{s_1}(u)(s_i) - T^\pi_{s_1}(v)(s_i)|\Big\} \\
&= \max\Big\{|T^\pi(u)(s_1) - T^\pi(v)(s_1)|, |u(s_2) - v(s_2)|, \dots, |u(s_K) - v(s_K)|\Big\} \\
&\leq \max\Big\{\gamma\|u - v\|_\infty, \|u - v\|_\infty, \dots, \|u - v\|_\infty|\Big\} \\
&\leq \|u - v\|_\infty.
\end{aligned}$$

Now we assume that $\|\tilde{u}^{(i)} - \tilde{v}^{(i)}\|_\infty \leq \|u - v\|_\infty$ for all $i < k \leq K - 1$. For k we follow then

$$\begin{aligned}
\|\tilde{u}^{(k)} - \tilde{v}^{(k)}\|_\infty &= \|(T^\pi_{s_k} \circ \cdots \circ T^\pi_{s_1})(u) - (T^\pi_{s_k} \circ \cdots \circ T^\pi_{s_1})(v)\|_\infty \\
&= \|(T^\pi_{s_k})(\tilde{u}^{(k-1)}) - (T^\pi_{s_k})(\tilde{v}^{(k-1)})\|_\infty \\
&= \max_{i=1,\dots,K}\Big\{|(T^\pi_{s_k})(\tilde{u}^{(k-1)})(s_i) - (T^\pi_{s_k})(\tilde{v}^{(k-1)})(s_i)|\Big\} \\
&= \max\Big\{\max_{i\neq k}\Big\{|(\tilde{u}^{(k-1)})(s_i) - (\tilde{v}^{(k-1)})(s_i)|\Big\}, |T^\pi(\tilde{u}^{(k-1)}) - T^\pi(\tilde{v}^{(k-1)})|\Big\} \\
&\leq \max\Big\{\|\tilde{u}^{(k-1)} - \tilde{v}^{(k-1)}\|_\infty, \gamma\|\tilde{u}^{(k-1)} - \tilde{v}^{(k-1)}\|_\infty\Big\} \\
&= \max\Big\{\|u - v\|_\infty, \gamma\|u - v\|_\infty\Big\} \\
&\leq \|u - v\|_\infty.
\end{aligned}$$

Overall we follow that

$$\begin{aligned}
&\|(T^\pi_{s_K} \circ \cdots \circ T^\pi_{s_1})(u) - (T^\pi_{s_K} \circ \cdots \circ T^\pi_{s_1})(v)\|_\infty \\
&\leq \max\Big\{\gamma\|u - v\|_\infty, \gamma\|\tilde{u}^{(1)} - \tilde{v}^{(1)}\|_\infty, \dots, \gamma\|\tilde{u}^{(K-1)} - \tilde{v}^{(K-1)}\|_\infty\Big\} \\
&\leq \gamma\|u - v\|_\infty.
\end{aligned}$$

3. Exact policy iteration

Consider two types of costumers, L for low and H for high, shopping in a shopping center. Each quarter the manager divides 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 weather 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-horizon 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* $\pi^0 = \pi^1$*.*