FAKULTÄT FÜR WIRTSCHAFTSINFORMATIK
UND WIRTSCHAFTSMATHEMATIK

UNIVERSITÄT MANNHEIM

Prof. Dr. Leif Döring
André Ferdinand, Sara Klein

Reinforcement Learning

6. Solution Sheet

1. Policy evaluation

Consider Algorithm 8 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 $\gamma = 1$. Your 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+1} *is determinstic 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*

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

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 the value function we started with. We see directly that we iterate between these two value functions and never converge.

2. Convergence of the in-place policy evaluation algorithm

Recall Algorithm 9 from the lecture. We aim to prove convergence of the algorithm (without termination) to V^π. Therfore 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 $U = \{u : \mathcal{S} \to \mathbb{R}\}$ equiped with the supremums norm.

a) Argue why $\overline{T}^\pi$ is different from the Bellman operator T^π.
Solution:
Appling 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 fixpoint 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 fixpoint with respect to the Bellam 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

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

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

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