Prof. Dr. Leif Döring
André Ferdinand, Sara Klein

Reinforcement Learning

9. Solution Sheet

1. Proofs for T-step MDPs

Prove the following claims from the lecture by comparing with the discounted counterpart.

a) Proposition 3.4.4: Given a Markovian policy $\pi = (\pi_t)_{t\in D}$ and a T-step Markov decision problem. Then the following relation between the state and state-action value function hold

$$V_t^\pi(s) = \sum_{a\in\mathcal{A}_s} \pi_t(a\,;\,s) Q_t^\pi(s,a),$$
$$Q_t^\pi(s,a) = r(s,a) + \sum_{s'\in\mathcal{S}} p(s'\,;\,s,a) V_{t+1}^\pi(s')$$

for all $t < T$. In particular (plugging-in), the Bellman expectation equations

$$V_t^\pi(s) = \sum_{a\in\mathcal{A}} \pi_t(a\,;\,s)\Big[r(s,a) + \sum_{s'\in\mathcal{S}} p(s'\,;\,s,a) V_{t+1}^\pi(s')\Big],$$
$$Q_t^\pi(s,a) = r(s,a) + \sum_{s'\in\mathcal{S}}\sum_{a'\in\mathcal{A}_s} p(s'\,;\,s,a)\pi_t(a\,;\,s') Q_{t+1}^\pi(s',a')$$

hold.

Solution:

By the definition of the time-state value function we have

$$V_t^\pi(s) = \mathbb{E}_s^{\hat\pi}\Big[\sum_{t'=0}^{T-t-1} R_{t'+1}\Big] = \sum_{a\in\mathcal{A}} \pi_t(a;s)\mathbb{E}_s^{\hat\pi}\Big[\sum_{t'=0}^{T-t-1} R_{t'+1} | A_0 = a\Big]$$
$$= \sum_{a\in\mathcal{A}} \pi_t(a;s)\mathbb{E}_s^{\hat\pi_a}\Big[\sum_{t'=0}^{T-t-1} R_{t'+1}\Big] = \sum_{a\in\mathcal{A}} \pi_t(a;s) Q_t^\pi(s,a),$$

where $\hat\pi$ is π shifted by t, i.e. $\hat\pi_0 = \pi_t, \dots, \hat\pi_{T-t-1} = \pi_{T-1}$.

For the time-state-action value function we have that

$$
\begin{aligned}
Q_t^{\pi}(s,a) &= \mathbb{E}_s^{\hat{\pi}_a}\Big[\sum_{t'=0}^{T-t-1} R_{t'+1}\Big] \\
&= r(s,a) + \mathbb{E}_s^{\hat{\pi}_a}\Big[\sum_{t'=1}^{T-t-1} R_{t'+1}\Big] \\
&= r(s,a) + \sum_{s'\in\mathcal{S}} \mathbb{P}_s^{\hat{\pi}_a}(S_1=s')\mathbb{E}_s^{\hat{\pi}_a}\Big[\sum_{t'=1}^{T-t-1} R_{t'+1}|S_1=s'\Big] \\
&= r(s,a) + \sum_{s'\in\mathcal{S}} p(s'\,;\,s,a)\mathbb{E}_{s'}^{\tilde{\pi}}\Big[\sum_{t'=0}^{T-(t+1)-1} R_{t'+1}\Big] \\
&= r(s,a) + \sum_{s'\in\mathcal{S}} p(s'\,;\,s,a)V_{t+1}^{\pi}(s'),
\end{aligned}
$$

where $\hat{\pi}$ is defined as above and $\tilde{\pi}$ is π shifted by $t+1$.

b) Lemma 3.4.6: The following holds for the optimal time-state value function and the optimal time-state-action value function for any $s \in \mathcal{S}$:

(i) $V_t^*(s) = \max_{a\in\mathcal{A}_s} Q_t^*(s,a)$ for all $t \le T-1$.
Solution:
Similar to the discounted infinite time MDP we have

$$
\begin{aligned}
\max_{a\in\mathcal{A}} Q_t^*(s,a) &= \max_{a\in\mathcal{A}} \sup_{\pi\in\Pi_t^{T-1}} Q_t^{\pi}(s,a) \\
&= \sup_{\pi\in\Pi_t^{T-1}} \max_{a\in\mathcal{A}} Q_t^{\pi}(s,a) \\
&= \sup_{\pi\in\Pi_t^{T-1}} \max_{a\in\mathcal{A}} \mathbb{E}_s^{\pi^a}[\sum_{t'=0}^{T-t-1} R_{t'+1}] \\
&= \sup_{\pi\in\Pi_t^{T-1}} \sup_{\tilde{\pi}\in\Pi} \mathbb{E}_s^{(\tilde{\pi},\pi_{t+1},\dots,\pi_{T-1})}[\sum_{t'=0}^{T-t-1} R_{t'+1}] \\
&= \sup_{\pi\in\Pi_t^{T-1}} \mathbb{E}_s^{\pi}[\sum_{t'=0}^{T-t-1} R_{t'+1}] \\
&= \sup_{\pi\in\Pi_t^{T-1}} V_t^{\pi}(s).
\end{aligned}
$$

We can replace $\max_{a\in\mathcal{A}}$ by $\sup_{\tilde{\pi}\in\Pi}$ in the forth equation by the same reason as in the infinite time case:
'$\le$': is always true ($\max \le \sup$), because all deterministic policies are included in Π.

'≥': we have that

$$\begin{aligned}\mathbb{E}_s^{(\tilde{\pi},\pi)}[\sum_{t'=0}^{T-t-1} R_{t'+1}] &= \sum_{a\in\mathcal{A}} \tilde{\pi}(a|s)\mathbb{E}_s^{(\pi^a)}[\sum_{t'=0}^{T-t-1} R_{t'+1}] \\ &\leq \max_{a\in\mathcal{A}} \mathbb{E}_s^{(\pi^a)}[\sum_{t'=0}^{T-t-1} R_{t'+1}]\sum_{a\in\mathcal{A}} \tilde{\pi}(a|s) \\ &= \max_{a\in\mathcal{A}} \mathbb{E}_s^{(\pi^a)}[\sum_{t'=0}^{T-t-1} R_{t'+1}].\end{aligned}$$

And therefore

$$\sup_{\tilde{\pi}\in\Pi} \mathbb{E}_s^{(\tilde{\pi},\pi)}[\sum_{t'=0}^{T-t-1} R_{t'+1}] \leq \max_{a\in\mathcal{A}} \mathbb{E}_s^{(\pi^a)}[\sum_{t'=0}^{T-t-1} R_{t'+1}].$$

(ii) $Q_t^*(s,a) = r(s,a) + \sum_{s'\in\mathcal{S}} p(s'\,;\,s,a)V_{t+1}^*(s')$ for all $t < T-1$.
Solution:
Using a) this follows directly by

$$\begin{aligned}Q_t^*(s,a) &= \sup_{\pi\in\Pi_t^T} Q_t^\pi(s,a) \\ &= \sup_{\pi\in\Pi_t^T} (r(s,a) + \sum_{s'\in\mathcal{S}} p(s';s,a)V_{t+1}^\pi(s')) \\ &= \sup_{\pi\in\Pi_{t+1}^T} (r(s,a) + \sum_{s'\in\mathcal{S}} p(s';s,a)V_{t+1}^\pi(s')) \\ &= r(s,a) + \sum_{s'\in\mathcal{S}} p(s';s,a) \sup_{\pi\in\Pi_{t+1}^T} V_{t+1}^\pi(s') \\ &= r(s,a) + \sum_{s'\in\mathcal{S}} p(s';s,a)V_{t+1}^*(s').\end{aligned}$$

2. Example: T-step MDPs

Recall the Ice Vendor example from the lecture. Assume the maximal amount of ice cream is $m = 3$ and the damand distribution is given by $\mathbb{P}(D_t = d) = p_d$ with $p_0 = p_2 = \frac{1}{4}, p_1 = \frac{1}{2}$. Suppose the revenue function f, ordering cost function o and storage cost function h are given by

$$\begin{aligned}&f : \mathbb{N}_0 \to \mathbb{R},\, x \mapsto 9x, \\ &o : \mathbb{N}_0 \to \mathbb{R},\, x \mapsto 2x, \\ &h : \mathbb{N}_0 \to \mathbb{R},\, x \mapsto 2 + x.\end{aligned}$$

a) Set up the transition matrix $p(s_{t+1}; s_t, a_t)$ in a table, such that every $s_t + a_t$ maps to the probability to land in s_{t+1}, and the reward function $r(s_t, a_t, s_{t+1})$ for this example.

Solution:

The transition matrix is given as follows

$(s+a)\backslash s'$	0	1	2	3
0	1	0	0	0
1	$\frac{3}{4}$	$\frac{1}{4}$	0	0
2	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	0
3	0	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

The reward function $R(s_t, a_t, s_{t+1}) = f(s_t + a_t - s_{t+1}) - o(a_t) - h(a_t + s_t)$ *is given by*

$$R(s_t, a_t, s_{t+1}) = 9(s_t + a_t - s_{t+1}) - 2a_t - 2 - (s_t + a_t) = 8s_t + 6a_t - 9s_{t+1} - 2.$$

b) Calculate the expected reward $r(s, a)$ for every state action pair. Can you guess an optimal strategy for a one time step MDP?

Solution:

The expected reward is given by

$$\begin{aligned} r(s,a) &= \sum_{r \in \mathcal{R}} r p(\mathcal{S} \times \{r\}\,;\, s, a) = \sum_{r \in \mathcal{R}} \sum_{s' \in \mathcal{S}} p(\{s'\} \times \{r\}\,;\, s, a) r \\ &= \sum_{s' \in \mathcal{S}} p(s'; s, a) R(s, a, s'), \end{aligned}$$

because the reward is deterministic for given s, a, s'. *The reward table is then*

$s\backslash a$	0	1	2	3
0	-2	$\frac{7}{4}$	1	-2
1	$\frac{15}{4}$	3	0	x
2	5	2	x	x
3	4	x	x	x

c) Suppose now you can play a 3-step MDP, hence you can order ice cream 3 times in $t = 0, 1, 2$. What is the optimal strategy for this finite time horizion MDP? Calculate the optimal state value, state-action value functions and the optimal policies using the greedy policy improvement algorithm from the lecture.

Hint: Use backward induction.

Solution:

We have as inition condition $V_3^* \equiv 0$ *and* $Q_2^* \equiv r$. *We follow from* Q_2^* *that the optimal policy is*

$$\pi_2^*(1; 0) = 1, \quad \pi_2^*(0; 1) = 1, \quad \pi_2^*(0; 2) = 1, \quad \pi_2^*(0; 3) = 1.$$

The value function $V_2^*(s) = \max_a Q_2^*(s, a)$, *are the red marked values in the reward tabel of b).*

It follows by

$$Q_1^*(s, a) = r(s, a) + \sum_{s' \in \mathcal{S}} p(s'; s, a) V_2^*(s')$$

that Q_1^* *is given by*

$s\backslash a$	0	1	2	3
0	$-\frac{1}{4}$	$\frac{61}{16}$	$\frac{67}{16}$	$\frac{9}{4}$
1	$\frac{93}{16}$	$\frac{99}{16}$	$\frac{17}{4}$	x
2	$\frac{131}{16}$	$\frac{25}{4}$	x	x
3	$\frac{33}{4}$	x	x	x

We follow from Q_1^* *that the optimal policy is*

$$\pi_1^*(2;0) = 1, \quad \pi_1^*(1;1) = 1, \quad \pi_1^*(0;2) = 1, \quad \pi_1^*(0;3) = 1.$$

The value function $V_1^*(s) = \max_a Q_1^*(s,a)$ *are the red numbers in the table. For the last timestep:*

$$Q_0^*(s,a) = r(s,a) + \sum_{s' \in \mathcal{S}} p(s';s,a)V_1^*(s')$$

that Q_0^* *is given by*

$s\backslash a$	0	1	2	3
0	$\frac{35}{26}$	$\frac{413}{64}$	$\frac{231}{32}$	$\frac{203}{32}$
1	$\frac{605}{64}$	$\frac{295}{32}$	$\frac{331}{32}$	x
2	$\frac{359}{32}$	$\frac{331}{32}$	x	x
3	$\frac{395}{32}$	x	x	x

We follow from Q_1^* *that the optimal policy is*

$$\pi_0^*(2;0) = 1, \quad \pi_0^*(2;1) = 1, \quad \pi_0^*(0;2) = 1, \quad \pi_0^*(0;3) = 1.$$

Finally we have that the red marked numbers in the last table are the optimal value function V_0^* *of this MDP.*