FAKULTÄT FÜR WIRTSCHAFTSINFORMATIK
UND WIRTSCHAFTSMATHEMATIK

UNIVERSITÄT MANNHEIM

Prof. Dr. Leif Döring
André Ferdinand, Sara Klein

Reinforcement Learning

4. Solution Sheet

1. Proof of Theorem 3.1.3

Complete the Proof of Theorem 3.1.3. from the lecture:

a) Show that defining $\mathbb{P}_T$ on the singletons

$$\begin{aligned}&\mathbb{P}_T(\{(s_0, r_0, a_0, \dots, s_T, r_T, a_T)\})\\ &:= \mu(\{s_0\}) \cdot \delta_0(r_0) \cdot \pi_0(\{a_0\}\,;\, s_0) \cdot \prod_{i=1}^{T} p(\{(s_i, r_i)\}\,;\, s_{i-1}, a_{i-1}) \cdot \pi_i(\{a_i\}\,;\, s_0, a_0, \dots, a_{i-1}, s_i).\end{aligned}$$

yields a probability measure.

Solution:

We denote with Ω_T the space of all trajectories until timepoint T,

$$\Omega_T = ((s_0, r_0, a_0, \dots, s_T, r_T, a_T) : s_i \in \mathcal{S}, a_i \in \mathcal{A}_{s_i}, r_0 \in \mathcal{R}, \forall i \in \{0, \dots, T\}).$$

We show the claim by induction:

Induction beginning: *Assume $T = 1$, then*

$$\begin{aligned}\mathbb{P}_1(\Omega_1) &= \sum_{s_0}\sum_{r_0}\sum_{a_0}\sum_{s_1}\sum_{r_1}\sum_{a_1} \mathbb{P}_T(\{(s_0, r_0, a_0, s_1, r_1, a_1)\})\\
&= \sum_{s_0}\sum_{r_0}\sum_{a_0}\sum_{s_1}\sum_{r_1}\sum_{a_1} \mu(\{s_0\}) \cdot \delta_0(r_0) \cdot \pi_0(\{a_0\}\,;\, s_0) \cdot\\
&\quad p(\{(s_1, r_1)\}\,;\, s_0, a_0) \cdot \pi_i(\{a_1\}\,;\, s_0, a_0, s_1)\\
&= \sum_{s_0} \mu(\{s_0\}) \cdot \sum_{r_0} \delta_0(r_0) \cdot \sum_{a_0} \pi_0(\{a_0\}\,;\, s_0) \cdot\\
&\quad \sum_{s_1}\sum_{r_1} p(\{(s_1, r_1)\}\,;\, s_0, a_0) \cdot \underbrace{\sum_{a_1} \pi_i(\{a_1\}\,;\, s_0, a_0, s_1)}_{=1}\\
&= \sum_{s_0} \mu(\{s_0\}) \cdot \sum_{r_0} \delta_0(r_0) \cdot \sum_{a_0} \pi_0(\{a_0\}\,;\, s_0) \cdot \underbrace{\sum_{s_1}\sum_{r_1} p(\{(s_1, r_1)\}\,;\, s_0, a_0)}_{=1}\\
&= \sum_{s_0} \mu(\{s_0\}) \cdot \sum_{r_0} \delta_0(r_0) \cdot \underbrace{\sum_{a_0} \pi_0(\{a_0\}\,;\, s_0)}_{=1}\\
&= \sum_{s_0} \mu(\{s_0\}) \cdot \underbrace{\sum_{r_0} \delta_0(r_0)}_{=1}\\
&= \underbrace{\sum_{s_0} \mu(\{s_0\})}_{=1}\\
&= 1\end{aligned}$$

Induction assumtion: *Assume the claim holds for* $\mathbb{P}_{T-1}(\Omega_{T-1}) = 1$.
Induction conclusion: *Then for* T,

$$\begin{aligned}
\mathbb{P}_T(\Omega_T) &= \sum_{s_0}\sum_{r_0}\sum_{a_0}\cdots\sum_{s_T}\sum_{r_T}\sum_{a_T}\mathbb{P}_T(\{(s_0, r_0, a_0, \ldots, s_T, r_T, a_T)\}) \\
&= \sum_{s_0}\sum_{r_0}\sum_{a_0}\cdots\sum_{s_T}\sum_{r_T}\sum_{a_T}\mu(\{s_0\})\cdot\delta_0(r_0)\cdot\pi_0(\{a_0\}\,;\, s_0)\cdot \\
&\quad \prod_{i=1}^{T} p(\{(s_i, r_i)\}\,;\, s_{i-1}, a_{i-1})\cdot\pi_i(\{a_i\}\,;\, s_0, a_0, \ldots, a_{i-1}, s_i) \\
&= \sum_{s_0}\sum_{r_0}\sum_{a_0}\cdots\sum_{s_{T-1}}\sum_{r_{T-1}}\sum_{a_{T-1}}\mu(\{s_0\})\cdot\delta_0(r_0)\cdot\pi_0(\{a_0\}\,;\, s_0)\cdot \\
&\quad \prod_{i=1}^{T-1} p(\{(s_i, r_i)\}\,;\, s_{i-1}, a_{i-1})\cdot\pi_i(\{a_i\}\,;\, s_0, a_0, \ldots, a_{i-1}, s_i) \\
&\quad \sum_{s_T}\sum_{r_T} p(\{(s_T, r_T)\}\,;\, s_{T-1}, a_{T-1})\cdot \underbrace{\sum_{a_T}\pi_T(\{a_T\}\,;\, s_0, a_0, \ldots, a_{T-1}, s_T)}_{=1} \\
&= \sum_{s_0}\sum_{r_0}\sum_{a_0}\cdots\sum_{s_{T-1}}\sum_{r_{T-1}}\sum_{a_{T-1}}\mu(\{s_0\})\cdot\delta_0(r_0)\cdot\pi_0(\{a_0\}\,;\, s_0)\cdot \\
&\quad \prod_{i=1}^{T-1} p(\{(s_i, r_i)\}\,;\, s_{i-1}, a_{i-1})\cdot\pi_i(\{a_i\}\,;\, s_0, a_0, \ldots, a_{i-1}, s_i) \\
&\quad \underbrace{\sum_{s_T}\sum_{r_T} p(\{(s_T, r_T)\}\,;\, s_{T-1}, a_{T-1})}_{=1} \\
&= \sum_{s_0}\sum_{r_0}\sum_{a_0}\cdots\sum_{s_{T-1}}\sum_{r_{T-1}}\sum_{a_{T-1}}\mu(\{s_0\})\cdot\delta_0(r_0)\cdot\pi_0(\{a_0\}\,;\, s_0)\cdot \\
&\quad \prod_{i=1}^{T-1} p(\{(s_i, r_i)\}\,;\, s_{i-1}, a_{i-1})\cdot\pi_i(\{a_i\}\,;\, s_0, a_0, \ldots, a_{i-1}, s_i) \\
&= \mathbb{P}_{T-1}(\Omega_{T-1}) \\
&= 1
\end{aligned}$$

b) Check the claimed conditional probability identity

$$\mathbb{P}_\mu^\pi(S_{t+1} \in B, R_{t+1} \in D \,|\, S_t = s, A_t = a) = p(B \times D\,;\, s, a),$$

where we assume that $\mathbb{P}_\mu^\pi(S_t = s, A_t = a) > 0$.
Solution:

First we have,

$$
\begin{aligned}
&\mathbb{P}^{\pi}_{\mu}(S_{t+1} \in B, R_{t+1} \in D, S_t = s, A_t = a) \\
&= \mathbb{P}_{t+1}(S_0 \in \mathcal{S}, R_0 \in \mathcal{R}, \ldots, S_t = s, R_t \in \mathcal{R}, A_t = a, S_{t+1} \in B, R_{t+1} \in D, A_{t+1} \in \mathcal{A}) \\
&= \sum_{s_0}\sum_{r_0}\cdots\sum_{s_{t-1}}\sum_{r_{t-1}}\sum_{a_{t-1}}\sum_{r_t}\sum_{a_{t+1}} \mu(\{s_0\}) \cdot \pi_0(\{a_0\}\,;\, s_0)\cdot \\
&\quad \prod_{i=1}^{t-1} p(\{s_i, r_i\}\,;\, s_{i-1}, a_{i-1}) \cdot \pi_i(\{a_i\}\,;\, s_0, a_0, \ldots, a_{i-1}, s_i) \\
&\quad \times p(\{(s, r_t)\}\,;\, s_{t-1}, a_{t-1}) \cdot \pi_t(\{a\}; s_0, a_0, ..., s_{t-1}, a_{t-1}, s) \\
&\quad \times \sum_{s_{t+1}\in B}\sum_{r_{t+1}\in D} p(\{(s_{t+1}, r_{t+1})\}\,;\, s, a) \cdot \pi_t(\{a_{t+1}\}; s_0, a_0, ..., s, a, s_{t+1}) \\
&= \sum_{s_0}\sum_{r_0}\cdots\sum_{s_{t-1}}\sum_{r_{t-1}}\sum_{a_{t-1}}\sum_{r_t} \mu(\{s_0\}) \cdot \pi_0(\{a_0\}\,;\, s_0)\cdot \\
&\quad \prod_{i=1}^{t-1} p(\{s_i, r_i\}\,;\, s_{i-1}, a_{i-1}) \cdot \pi_i(\{a_i\}\,;\, s_0, a_0, \ldots, a_{i-1}, s_i) \\
&\quad \times p(\{(s, r_t)\}\,;\, s_{t-1}, a_{t-1}) \cdot \pi_t(\{a\}; s_0, a_0, ..., s_{t-1}, a_{t-1}, s) \\
&\quad \times \sum_{s_{t+1}\in B}\sum_{r_{t+1}\in D} p(\{(s_{t+1}, r_{t+1})\}\,;\, s, a) \cdot \sum_{a_{t+1}} \pi_t(\{a_{t+1}\}; s_0, a_0, ..., s, a, s_{t+1}) \\
&= \sum_{s_0}\sum_{r_0}\cdots\sum_{s_{t-1}}\sum_{r_{t-1}}\sum_{a_{t-1}}\sum_{r_t} \mu(\{s_0\}) \cdot \pi_0(\{a_0\}\,;\, s_0)\cdot \\
&\quad \prod_{i=1}^{t-1} p(\{s_i, r_i\}\,;\, s_{i-1}, a_{i-1}) \cdot \pi_i(\{a_i\}\,;\, s_0, a_0, \ldots, a_{i-1}, s_i) \\
&\quad \times p(\{(s, r_t)\}\,;\, s_{t-1}, a_{t-1}) \cdot \pi_t(\{a\}; s_0, a_0, ..., s_{t-1}, a_{t-1}, s) \\
&\quad \times \sum_{s_{t+1}\in B}\sum_{r_{t+1}\in D} p(\{(s_{t+1}, r_{t+1})\}\,;\, s, a)
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathbb{P}^{\pi}_{\mu}(S_t = s, A_t = a) \\
&= \sum_{s_0}\sum_{r_0}\cdots\sum_{s_{t-1}}\sum_{r_{t-1}}\sum_{a_{t-1}}\sum_{r_t} \mathbb{P}_t(S_0 = s_0, A_0 = a_0, ..., A_{t-1} = a_{t-1}, S_t = s, R_t = r_t, A_t = a) \\
&= \sum_{s_0}\sum_{r_0}\cdots\sum_{s_{t-1}}\sum_{r_{t-1}}\sum_{a_{t-1}}\sum_{r_t} \mu(\{s_0\}) \cdot \pi_0(\{a_0\}\,;\, s_0)\cdot \\
&\quad \prod_{i=1}^{t-1} p(\{s_i, r_i\}\,;\, s_{i-1}, a_{i-1}) \cdot \pi_i(\{a_i\}\,;\, s_0, a_0, \ldots, a_{i-1}, s_i) \\
&\quad \times p(\{(s, r_t)\}\,;\, s_{t-1}, a_{t-1}) \cdot \pi_t(\{a\}; s_0, a_0, ..., s_{t-1}, a_{t-1}, s)
\end{aligned}
$$

Hence, using the definition of conditional probability the fraction cancels to give

$$
\mathbb{P}^{\pi}_{\mu}(S_{t+1} \in B, R_{t+1} \in D \,|\, S_t = s, A_t = a) = \sum_{s_{t+1}\in B}\sum_{r_{t+1}\in D} p(\{(s_{t+1}, r_{t+1})\}\,;\, s, a).
$$

2. Probabilistic interpretation of MDPs

Use the formal definition of the stochastic process (S, A, R) to check that

$$p(s'; s, a) = \mathbb{P}^\pi_\mu(S_{t+1} = s' \,|\, S_t = s, A_t = a),$$
$$r(s, a) = \mathbb{E}^\pi_\mu[R_{t+1} \,|\, S_t = s, A_t = a],$$
$$r(s, a, s') = \mathbb{E}^\pi_\mu[R_{t+1} \,|\, S_t = s, A_t = a, S_{t+1} = s']$$

holds if the events in the condition have positive probability under $\mathbb{P}^\pi_\mu$, for any $t \in \mathbb{N}_0$ arbitrary start distribution μ and policy π.

Solution:

For the first claim we have by Excercise 1 b) that

$$\begin{aligned}
p(s'; s, a) &:= p(\{s'\} \times \mathcal{R}; s, a) \\
&= \mathbb{P}^\pi_\mu(S_{t+1} \in \{s'\}, R_{t+1} \in \mathcal{R} \,|\, S_t = s, A_t = a) \\
&= \frac{\mathbb{P}^\pi_\mu(S_{t+1} = s', R_{t+1} \in \mathcal{R}, S_t = s, A_t = a)}{\mathbb{P}^\pi_\mu(S_t = s, A_t = a)} \\
&= \frac{\mathbb{P}^\pi_\mu(S_{t+1} = s', S_t = s, A_t = a)}{\mathbb{P}^\pi_\mu(S_t = s, A_t = a)} \\
&= \mathbb{P}^\pi_\mu(S_{t+1} = s' \,|\, S_t = s, A_t = a).
\end{aligned}$$

For the next equation we have similar to above

$$\begin{aligned}
r(s, a) &:= \sum_r r \, p(\mathcal{S} \times \{r\}; s, a) \\
&= \sum_r r \, \mathbb{P}^\pi_\mu(S_{t+1} \in \mathcal{S}, R_{t+1} \in \{r\} \,|\, S_t = s, A_t = a) \\
&= \sum_r r \, \mathbb{P}^\pi_\mu(R_t = r \,|\, S_t = s, A_t = a) \\
&= \mathbb{E}^\pi_\mu[R_{t+1} \,|\, S_t = s, A_t = a],
\end{aligned}$$

where we have used the definition of expectation for discrete random variables in the last equation.

The last claim follows similar

$$
\begin{aligned}
r(s,a,s') &:= \sum_r r \, \frac{p(\{(s',r)\}; s, a)}{p((s',); s, a)} \\
&= \sum_r r \, \frac{\mathbb{P}^\pi_\mu(S_{t+1} = s', R_{t+1} = r \,|\, S_t = s, A_t = a)}{\mathbb{P}^\pi_\mu(S_{t+1} = s' \,|\, S_t = s, A_t = a)} \\
&= \sum_r r \, \frac{\frac{\mathbb{P}^\pi_\mu(S_{t+1}=s', R_{t+1}=r, S_t=s, A_t=a)}{\mathbb{P}^\pi_\mu(S_t=s, A_t=a)}}{\frac{\mathbb{P}^\pi_\mu(S_{t+1}=s', S_t=s, A_t=a)}{\mathbb{P}^\pi_\mu(S_t=s, A_t=a)}} \\
&= \sum_r r \, \frac{\mathbb{P}^\pi_\mu(S_{t+1} = s', R_{t+1} = r, S_t = s, A_t = a)}{\mathbb{P}^\pi_\mu(S_{t+1} = s', S_t = s, A_t = a)} \\
&= \sum_r r \, \mathbb{P}^\pi_\mu(R_{t+1} = r \,|\, S_{t+1} = s', S_t = s, A_t = a) \\
&= \mathbb{E}^\pi_\mu[R_{t+1} \,|\, S_t = s, A_t = a, S_{t+1} = s'].
\end{aligned}
$$

3. Bellman expectation equation for Q

Derive the Bellman expectation equation for the Q-function of a policy π:

$$
Q^\pi(s,a) = r(s,a) + \gamma \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} p(s' \,;\, s, a) \pi(a' \,;\, s) Q^\pi(s', a').
$$

Solution:

We use Lemma 3.1.17 and get directly

$$
\begin{aligned}
Q^\pi(s,a) &= r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s' \,;\, s, a) V^\pi(s') \\
&= r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s' \,;\, s, a) \sum_{a' \in \mathcal{A}} \pi(a' \,;\, s) Q^\pi(s', a') \\
&= r(s,a) + \gamma \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} p(s' \,;\, s, a) \pi(a' \,;\, s) Q^\pi(s', a').
\end{aligned}
$$