Markov Decision Processes II

Csaba Szepesvári

Friday 17th February, 2023

1 Definitions

As usual, for a set U, $\mathcal{M}_1(U)$ stands for the set of all probability measures (or, probability distributions) over U. Let U be finite. Then any such probability measure $P \in \mathcal{M}_1(U)$ is just a map from all possible subsets of U to $[0,1]$. In this case P is also uniquely identified with the probabilities it assigns to singletons of U and we, in fact, let p denote the corresponding map, which does map U to $[0,1]$. The connection is that for any $u \in U$, $p(u) = P(\{u\})$. We call p the probability mass function (pmf) underlying P. Since pmfs and probability measures have a one-to-one correspondence, oftentimes we use pmfs instead of probability measures and will abuse the notation by writing $p \in \mathcal{M}_1(U)$, where p is a pmf. If we have a map $f : U \to \mathcal{M}_1(V)$ for finite sets U and V, we further abuse the notation by writing $f(v|u)$ instead of $(f(u))(v)$. (Most notably, we do this for the transition dynamics function p and the policy π.)

Definition 1.1 (Finite MDP)**.** A finite MDP is given by the tuple $M = (\mathcal{S}, \mathcal{A}, \mathcal{R}, p)$, where $\mathcal{S}$ is the finite state space, $\mathcal{A}$ is the finite action space, $\mathcal{R}$ is the finite set of possible rewards, and $p : \mathcal{S} \times \mathcal{A} \to \mathcal{M}_1(\mathcal{R} \times \mathcal{S})$ is the transition dynamics function. In particular, $p(r', s'|s, a)$ is the probability of seeing reward $r' \in \mathcal{R}$ and next state $s' \in \mathcal{S}$ given that action $a \in \mathcal{A}$ is taken in state $s \in \mathcal{S}$.[1]

Definition 1.2 (Histories and policies)**.** For $t \geq 0$, the set of t-step histories is defined recursively as follows: $\mathcal{H}_0 = \mathcal{S}$ and for $t \geq 1$, $\mathcal{H}_t = \mathcal{H}_{t-1} \times \mathcal{A} \times \mathcal{R} \times \mathcal{S}$. A policy of a finite MDP $M = (\mathcal{S}, \mathcal{A}, \mathcal{R}, p)$ is $\pi = (\pi_t)_{t \geq 0}$, where $\pi_t : \mathcal{H}_t \to \mathcal{M}_1(\mathcal{A})$ is the map used at time t. In words, for $h \in \mathcal{H}_t$, $a \in \mathcal{A}$, $\pi_t(a|h)$ is the probability that action a is taken when the history is h at time t. For $t \geq 1$ we let $\mathcal{H}_t^{-\mathcal{S}} = \mathcal{H}_{t-1} \times \mathcal{A} \times \mathcal{R}$ ("missing the last state"), and for $t \geq 0$ we let $\mathcal{H}_t^{+\mathcal{A}} = \mathcal{H}_t \times \mathcal{A}$ ("appending an action").

Definition 1.3 (Memoryless policy)**.** If π is a policy such that for all $t \geq 0$, π_t only depends on the last state in the history, then π is called a *memoryless* policy. For such policies, one only needs to specify $\pi_0 : \mathcal{S} \to \mathcal{M}_1(\mathcal{A})$, as opposed to the sequence $(\pi_t)_{t \geq 0}$. And so by abusing language, any map from states to distributions over actions will be treated as a memoryless policy.

Definition 1.4 (Probability measures induced by using a policy in an MDP)**.** Fix a finite MDP $M = (\mathcal{S}, \mathcal{A}, \mathcal{R}, p)$, a policy π of this MDP, and a distribution $\mu \in \mathcal{M}_1(\mathcal{S})$. Then, the probability measure—induced by using π in M and μ as the initial state distribution—is a probability distribution $\mathbb{P}$ over some sample space Ω, such that there are random variables $S_0, S_1, \ldots : \Omega \to \mathcal{S}$, $A_0, A_1, \ldots : \Omega \to \mathcal{A}$, and $R_1, R_2, \ldots : \Omega \to \mathcal{R}$, for which the following holds for every $t \geq 0$ and every history

$h_t = (s_0, a_0, r_1, s_1, a_1, \ldots, a_{t-1}, r_t, s_t) \in \mathcal{H}_t$:

$$\begin{aligned}\mathbb{P}(H_t = h_t) = \mu(s_0) \cdot \pi_0(a_0|s_0) \cdot p(r_1, s_1|s_0, a_0) \\ \cdot \pi_1(a_1|s_0, a_0, r_1, s_1) \cdot p(r_2, s_2|s_1, a_1) \\ \vdots \\ \cdot \pi_{t-1}(a_{t-1}|s_0, a_0, \ldots, s_{t-1}) \cdot p(r_t, s_t|s_{t-1}, a_{t-1}) \,.\end{aligned}$$

When the dependence of $\mathbb{P}$ on μ and π is important, we write $\mathbb{P}_{\mu,\pi}$ to signify this dependence. To simplify notation, we also use $\mathbb{P}$, $\mathbb{P}_\mu$ and $\mathbb{P}_\pi$, when one, or both of these objects are clear from the context.

A reasonable question to ask is whether the above definition is even a correct one? Does a distribution $\mathbb{P}$ (and a corresponding sample space Ω) with the above properties exist? The answer is yes and it is the Kolmogorov's extension theorem which guarantees this. The next question is whether $\mathbb{P}$ and Ω are uniquely defined? Or are there distinct probability measures, $\mathbb{P}$ and $\mathbb{P}'$ (and corresponding sample spaces), that satisfy the above definition? The answer is that $\mathbb{P}$ and Ω are not uniquely defined. Once we have a pair $(\Omega, \mathbb{P})$ that satisfies the above definition, we can always construct some $\mathbb{P}'$ and Ω' such that $\mathbb{P}'$ and Ω' will also satisfy the definition with appropriate sequences of random variables, $S'_0, S'_1, \ldots, A'_0, A'_1, \ldots$ and $R'_1, R'_2, \ldots$, over Ω' (for instance, consider $\Omega' = \Omega \times \{1\}$). However, that $(\Omega, \mathbb{P})$ are not uniquely defined does not matter, as in every calculation involving $\mathbb{P}$ (e.g., definition of value functions), only the properties of $\mathbb{P}$ mentioned in Definition 1.4 will be used.

Proposition 1.5. *The stochastic process $S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, \ldots$ under the the probability distribution $\mathbb{P}_{\mu,\pi}$ satisfies the Markov property, regardless of the choice of μ and π.*

We will also use $\mathbb{E}_{\mu,\pi}$ to denote the expectation underlying $\mathbb{P}_{\mu,\pi}$. When μ is clear from the context, we also use just $\mathbb{E}_\pi$, etc.

Some important properties of $\mathbb{P}_{\mu,\pi}$ are given as follows.

Proposition 1.6. *The probability measure $\mathbb{P} := \mathbb{P}_{\mu,\pi}$ satisfies the following properties: For any $t \geq 0$, $h_t = (s_0, a_0, r_1, s_1, a_1, \ldots, r_t, s_t) \in \mathcal{H}_t$, $r' \in \mathcal{R}$, and $s' \in \mathcal{S}$, it holds that*

$$\mathbb{P}(S_0 = s_0) = \mu(s_0) \,, \tag{1}$$

$$\mathbb{P}(A_t = a_t | H_t = h_t) = \pi_t(a_t|h_t) \,, \tag{2}$$

$$\mathbb{P}(R_{t+1} = r', S_{t+1} = s' | H_t = h_t, A_t = a_t) = p(r', s'|s_t, a_t) \,. \tag{3}$$

Further, if some probability measure $\mathbb{P}$ satisfies the above properties then it also satisfies the properties used in Definition 1.4.

Note that Equation (3), given above, is what we recognize as the Markov property.

The following proposition will be useful in the future. This proposition will allow us to use the same sequence of random variables to denote states, actions, and rewards regardless of the initial distribution or the policy used.

Proposition 1.7 (Single sample space)**.** *There exist a sample space Ω and random variables $S_0, S_1, \ldots : \Omega \to \mathcal{S}$, $A_0, A_1, \ldots : \Omega \to \mathcal{A}$, and $R_1, R_2, \ldots : \Omega \to \mathcal{R}$ such that for any policy π and initial state distribution $\mu \in \mathcal{M}_1(\mathcal{S})$, there exists a probability measure $\mathbb{P}_{\mu,\pi} \in \mathcal{M}_1(\Omega)$ that satisfies Definition 1.4.*

2 Value functions

We will use the expected total discounted reward as the criterion to evaluate policies. Let $0 \leq \gamma < 1$ be the discount factor used.

Definition 2.1 (Value function of a policy)**.** Fix an arbitrary policy π. The value function $v_\pi : \mathcal{S} \to \mathbb{R}$ of π is defined by

$$v_\pi(s) = \mathbb{E}_{\delta_s,\pi}\left[\sum_{t\geq 0} \gamma^t R_{t+1}\right], \qquad s \in \mathcal{S}, \tag{4}$$

where $\delta_s \in \mathcal{M}_1(\mathcal{S})$ is the "Dirac" probability measure on $\mathcal{S}$ with a point-mass at s. In particular, if we take δ_s to be the pmf of this measure then

$$\delta_s(s') = \mathbb{I}\{s = s'\}, \quad \text{for all } s' \in \mathcal{S}.$$

In words, $v_\pi(s)$ is the expected total discounted reward incurred by the agent when it follows policy π from state s.

Is $v_\pi(s)$ well-defined? Since we are in a finite MDP, the infinite sum (also known as the return) $\sum_{t\geq 0} \gamma^t R_{t+1}$ is well-defined: Indeed, defining

$$f_n = \sum_{t=0}^{n} \gamma^t R_{t+1},$$

we can reason that f_n is a convergent sequence of functions; one can use the Cauchy criterion to show this. Thus, the sequence of functions $(f_n)_n$ converges to some function f, which we denote by $\sum_{t\geq 0} \gamma^t R_{t+1}$. That the expectation of f exists follows from Lebesgue's dominated convergence theorem. To use this theorem, we need to show that there exists a function $g : \Omega \to \mathbb{R}$ such that $|f_n| \leq g$ for all $n \geq 0$, and that $\mathbb{E}_{\delta_s,\pi}[g]$ exists. Indeed, $g = r_{\max}/(1-\gamma)$ can be shown to be such a function where

$$r_{\max} = \max\{|r| \,:\, r \in \mathcal{R}\}.$$

Therefore, Lebesgue's dominated convergence theorem guarantees that $\mathbb{E}_{\delta_s,\pi}[f]$ exists and

$$\lim_{n\to\infty} \mathbb{E}_{\delta_s,\pi}[f_n] \to \mathbb{E}_{\delta_s,\pi}[f].$$

Note that here

$$\mathbb{E}_{\delta_s,\pi}[f_n] = \mathbb{E}_{\delta_s,\pi}\left[\sum_{t=0}^{n} \gamma^t R_{t+1}\right] = \sum_{t=0}^{n} \gamma^t \,\mathbb{E}_{\delta_s,\pi}\left[R_{t+1}\right],$$

where the last equality follows from the linearity of expectation (which can be used because $\mathbb{E}_{\delta_s,\pi}\left[R_{t+1}\right]$ is well-defined). It then follows from the above argument that

$$v_\pi(s) = \lim_{n\to\infty} \sum_{t=0}^{n} \gamma^t \,\mathbb{E}_{\delta_s,\pi}\left[R_{t+1}\right].$$

Again, following the standard convention, the limit on the right-hand side is denoted by $\sum_{t=0}^{\infty} \gamma^t \, \mathbb{E}_{\delta_s,\pi} [R_{t+1}]$. Hence,

$$v_\pi(s) = \sum_{t=0}^{\infty} \gamma^t \, \mathbb{E}_{\delta_s,\pi} [R_{t+1}] \; . \tag{5}$$

Note that this could also be used as the definition of $v_\pi(s)$ (this expression differs from the definition of v_π in the sense that here the infinite sum is moved outside of the expectation). In fact, if we started with this definition, we would have spared the need for using Lebesgue's dominated convergence theorem to argue that v_π is well-defined (why?).

Definition 2.2 (Immediate reward function)**.** Given a finite MDP M, the immediate reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ of M is defined using[2]

$$r(s,a) = \sum_{r' \in \mathcal{R}, s' \in \mathcal{S}} r' \cdot p(r', s'|s, a) \, .$$

Sometimes the following is convenient:

Proposition 2.3. *For any policy π,*

$$\begin{aligned} v_\pi(s) &= \mathbb{E}_{\delta_s,\pi} \left[\sum_{t \geq 0} \gamma^t \, r(S_t, A_t) \right] \\ &= \sum_{t \geq 0} \gamma^t \, \mathbb{E}_{\delta_s,\pi} \big[r(S_t, A_t) \big] \, , \qquad s \in \mathcal{S} \, . \end{aligned}$$

Proof. From Equation (5), we have

$$v_\pi(s) = \sum_{t=0}^{\infty} \gamma^t \, \mathbb{E}_{\delta_s,\pi} [R_{t+1}] \; .$$

Now, by the property of conditional expectations,[3]

$$\mathbb{E}_{\delta_s,\pi} [R_{t+1}] = \sum_{s \in \mathcal{S}, a \in \mathcal{A}} \mathbb{E}_{\delta_s,\pi} [R_{t+1} \mid S_t = s, A_t = a] \, \mathbb{P}_{\delta_s,\pi}(S_t = s, A_t = a) \, .$$

Further, recall that for fixed $(s, a) \in \mathcal{S} \times \mathcal{A}$,

$$\mathbb{E}_{\delta_s,\pi} [R_{t+1} \mid S_t = s, A_t = a] = \sum_{r' \in \mathcal{R}} r' \cdot \mathbb{P}_{\delta_s,\pi}(R_{t+1} = r' \mid S_t = s, A_t = a).$$

Let (s, a) be such that $\mathbb{P}_{\delta_s,\pi}(S_t = s, A_t = a) > 0$. Hence, if we prove that

$$\mathbb{P}_{\delta_s,\pi}(R_{t+1} = r' \mid S_t = s, A_t = a) = \sum_{s' \in \mathcal{S}} p(r', s'|s, a) \, , \tag{6}$$

it follows that

$$\mathbb{E}_{\delta_s,\pi} [R_{t+1} \mid S_t = s, A_t = a] = r(s, a) \, ,$$

and thus,

$$\mathbb{E}_{\delta_s,\pi}\left[R_{t+1}\right] = \sum_{s,a} r(s,a) \cdot \mathbb{P}_{\delta_s,\pi}(S_t = s, A_t = a) = \mathbb{E}_{\delta_s,\pi}\left[r(S_t, A_t)\right] .$$

It remains to show that Equation (6) holds. For this, recalling that $\mathbb{P}_{\delta_s,\pi}(S_t = s, A_t = a) > 0$, we get

$$\begin{aligned}
&\mathbb{P}_{\delta_s,\pi}(R_{t+1} = r' \mid S_t = s, A_t = a) \\
&\quad = \frac{\sum_{h \in \mathcal{H}_t^{-\mathcal{S}}, s' \in \mathcal{S}} \mathbb{P}_{\delta_s,\pi}(R_{t+1} = r', S_{t+1} = s', S_t = s, A_t = a, H_t^{-S} = h)}{\mathbb{P}_{\delta_s,\pi}(S_t = s, A_t = a)} ,
\end{aligned}$$

where we define $H_t^{-S} = (S_0, A_0, R_1, S_1, A_1, R_2, S_2, \ldots, A_{t-1}, R_t)$. Let's calculate the numerator:

$$\begin{aligned}
&\sum_{h \in \mathcal{H}_t^{-\mathcal{S}}, s' \in \mathcal{S}} \mathbb{P}_{\delta_s,\pi}(R_{t+1} = r', S_{t+1} = s', S_t = s, A_t = a, H_t^{-S} = h) \\
&\quad = \sum_{h \in \mathcal{H}_t^{-\mathcal{S}}, s' \in \mathcal{S}} \Big(\mathbb{P}_{\delta_s,\pi}(R_{t+1} = r', S_{t+1} = s' | S_t = s, A_t = a, H_t^{-S} = h) \\
&\qquad\qquad \times \mathbb{P}_{\delta_s,\pi}(S_t = s, A_t = a, H_t^{-S} = h) \Big) \\
&\quad = \sum_{h \in \mathcal{H}_t^{-\mathcal{S}}, s' \in \mathcal{S}} p(r', s' | s, a) \cdot \mathbb{P}_{\delta_s,\pi}(S_t = s, A_t = a, H_t^{-S} = h) \qquad \text{(by Equation (3))} \\
&\quad = \left(\sum_{h \in \mathcal{H}_t^{-\mathcal{S}}} \mathbb{P}_{\delta_s,\pi}(S_t = s, A_t = a, H_t^{-S} = h) \right) \left(\sum_{s' \in \mathcal{S}} p(r', s' | s, a) \right) \\
&\quad = \mathbb{P}_{\delta_s,\pi}(S_t = s, A_t = a) \sum_{s' \in \mathcal{S}} p(r', s' | s, a) . \qquad \text{(by the law of total probability)}
\end{aligned}$$

Plugging this back in, we get

$$\begin{aligned}
&\mathbb{P}_{\delta_s,\pi}(R_{t+1} = r' \mid S_t = s, A_t = a) \\
&\quad = \frac{\cancel{\mathbb{P}_{\delta_s,\pi}(S_t = s, A_t = a)} \sum_{s' \in \mathcal{S}} p(r', s' | s, a)}{\cancel{\mathbb{P}_{\delta_s,\pi}(S_t = s, A_t = a)}} = \sum_{s' \in \mathcal{S}} p(r', s' | s, a) .
\end{aligned}$$

□

3 Bellman equation for policy evaluation

Define

$$p(s' | s, a) = \sum_{r' \in \mathcal{R}} p(r', s' | s, a) .$$

Proposition 3.1 (Bellman equation for policy evaluation)**.** *Let $\pi : \mathcal{S} \to \mathcal{M}_1(\mathcal{A})$ be a memoryless policy. Then, the following holds:*

$$v_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) \left(r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) v_\pi(s') \right) . \tag{7}$$

Proof. Fix π, the sample space Ω, and the stochastic process $S_0, A_0, R_1, S_1, A_1, \ldots$ on it from Proposition 1.7. Further, for all $s \in \mathcal{S}$, let $\mathbb{P}_{\delta_s,\pi} \in \mathcal{M}_1(\Omega)$ be the probability measure from the same proposition.

Fix $t \geq 1$ and $(s_0, a_0, \ldots, s_t, a_t) \in (\mathcal{S} \times \mathcal{A})^{t+1}$. Assume that $\mathbb{P}_{\delta_{s_0},\pi}(A_0 = a_0, S_1 = s_1) > 0$. We claim that the following holds:

$$\begin{aligned}&\mathbb{P}_{\delta_{s_0},\pi}(S_1 = s_1, A_1 = a_1, S_2 = s_2, A_2 = a_2, \ldots, \quad S_t = s_t, \quad A_t = a_t \mid A_0 = a_0, S_1 = s_1)\\ &= \mathbb{P}_{\delta_{s_1},\pi}(S_0 = s_1, A_0 = a_1, S_1 = s_2, A_1 = a_2, \ldots, S_{t-1} = s_t, A_{t-1} = a_t)\,. \end{aligned} \tag{8}$$

In other words, under the memoryless policy π, "the state-action distribution from timestep $t = 1$ conditioned on starting from state s_0, choosing action a_0, and then arriving at state s_1" is the same as "the state-action distribution from timestep $t = 0$ starting from state s_1". We leave the proof of Equation (18) for later and continue with the proof of Equation (7).

From Proposition 2.3 we have

$$\begin{aligned} v_\pi(s_0) &= \sum_{t\geq 0} \gamma^t\, \mathbb{E}_{\delta_{s_0},\pi}\left[r(S_t, A_t)\right] \\ &= \mathbb{E}_{\delta_{s_0},\pi}[r(S_0, A_0)] + \gamma \sum_{t\geq 1} \gamma^{t-1} \mathbb{E}_{\delta_{s_0},\pi}\left[r(S_t, A_t)\right]\,. \end{aligned} \tag{9}$$

On one hand, a simple calculation shows (why?) that

$$\mathbb{E}_{\delta_{s_0},\pi}[r(S_0, A_0)] = \sum_{a_0 \in \mathcal{A}} \pi(a_0|s_0) \cdot r(s_0, a_0)\,. \tag{10}$$

On the other hand, for $t \geq 1$ we have

$$\mathbb{E}_{\delta_{s_0},\pi}\left[r(S_t, A_t)\right] = \sum_{s_t \in \mathcal{S}, a_t \in \mathcal{A}} \mathbb{P}_{\delta_{s_0},\pi}(S_t = s_t, A_t = a_t) \cdot r(s_t, a_t)\,. \tag{11}$$

Further,

$$\begin{aligned}&\mathbb{P}_{\delta_{s_0},\pi}(S_t = s_t, A_t = a_t)\\ &= \sum_{a_0 \in \mathcal{A}, s_1 \in \mathcal{S}} \mathbb{P}_{\delta_{s_0},\pi}(S_t = s_t, A_t = a_t \mid A_0 = a_0, S_1 = s_1) \cdot \mathbb{P}_{\delta_{s_0},\pi}(A_0 = a_0, S_1 = s_1)\\ &= \sum_{a_0 \in \mathcal{A}, s_1 \in \mathcal{S}} \mathbb{P}_{\delta_{s_0},\pi}(S_t = s_t, A_t = a_t \mid A_0 = a_0, S_1 = s_1) \cdot \pi(a_0|s_0) \cdot p(s_1|s_0, a_0)\,. \end{aligned} \tag{12}$$

From Equation (18), summing both sides over the variables $s_1, a_1, \ldots, s_{t-1}, a_{t-1}$, it follows that

$$\mathbb{P}_{\delta_{s_0},\pi}(S_t = s_t, A_t = a_t \mid A_0 = a_0, S_1 = s_1) = \mathbb{P}_{\delta_{s_1},\pi}(S_{t-1} = s_t, A_{t-1} = a_t)\,.$$

Combining with Equation (12), we have

$$\mathbb{P}_{\delta_{s_0},\pi}(S_t = s_t, A_t = a_t) = \sum_{a_0 \in \mathcal{A}, s_1 \in \mathcal{S}} \mathbb{P}_{\delta_{s_1},\pi}(S_{t-1} = s_t, A_{t-1} = a_t) \cdot \pi(a_0|s_0) \cdot p(s_1|s_0, a_0),$$

and from Equation (11) we get that

$$\begin{aligned}\mathbb{E}_{\delta_{s_0},\pi}\left[r(S_t,A_t)\right] &= \sum_{a_0\in\mathcal{A},s_1\in\mathcal{S}} \pi(a_0|s_0)\cdot p(s_1|s_0,a_0) \sum_{s_t\in\mathcal{S},a_t\in\mathcal{A}} \mathbb{P}_{\delta_{s_1},\pi}(S_{t-1}=s_t,A_{t-1}=a_t)\cdot r(s_t,a_t) \\ &= \sum_{a_0\in\mathcal{A},s_1\in\mathcal{S}} \pi(a_0|s_0)\cdot p(s_1|s_0,a_0)\cdot \mathbb{E}_{\delta_{s_1},\pi}[r(S_{t-1},A_{t-1})]\,.\end{aligned}$$

Combining this with Equations (9) and (10), we get

$$\begin{aligned}v_\pi(s_0) &= \sum_{a_0\in\mathcal{A}} \pi(a_0|s_0) r(s_0,a_0) + \gamma \sum_{t\geq 1}\gamma^{t-1} \sum_{a_0\in\mathcal{A},s_1\in\mathcal{S}} \pi(a_0|s_0)p(s_1|s_0,a_0)\mathbb{E}_{\delta_{s_1},\pi}[r(S_{t-1},A_{t-1})] \\ &= \sum_{a_0\in\mathcal{A}} \pi(a_0|s_0)\left(r(s_0,a_0) + \gamma \sum_{s_1\in\mathcal{S}} p(s_1|s_0,a_0) \underbrace{\sum_{t\geq 1}\gamma^{t-1}\mathbb{E}_{\delta_{s_1},\pi}[r(S_{t-1},A_{t-1})]}_{=v_\pi(s_1)}\right),\end{aligned}$$

where the swapping of the sums is justified because there are finitely many states and actions (why does this justify swapping of the sums?).

Finally, it remains to prove Equation (18). To prove this, we just expand the definitions on the two sides and match the terms to notice that the equality does hold. In particular, the expression on the left hand side of Equation (18) is

$$\begin{aligned}&\mathbb{P}_{\delta_{s_0},\pi}(S_1=s_1,A_1=a_1,S_2=s_2,A_2=a_2,\ldots,S_t=s_t,A_t=a_t \mid A_0=a_0,S_1=s_1) \\ &\quad= \frac{\mathbb{P}_{\delta_{s_0},\pi}(A_0=a_0,S_1=s_1,A_1=a_1,S_2=s_2,A_2=a_2,\ldots,S_t=s_t,A_t=a_t)}{\mathbb{P}_{\delta_{s_0},\pi}(A_0=a_0,S_1=s_1)}\,.\end{aligned}$$

Further, one can show that the numerator of this expression is (why?)

$$\begin{aligned}&\mathbb{P}_{\delta_{s_0},\pi}(A_0=a_0,S_1=s_1,A_1=a_1,S_2=s_2,A_2=a_2,\ldots,S_t=s_t,A_t=a_t) \\ &\quad= \pi(a_0|s_0)p(s_1|s_0,a_0)\pi(a_1|s_1)p(s_2|s_1,a_1)\cdots p(s_t|s_{t-1},a_{t-1})\pi(a_t|s_t)\,,\end{aligned}$$

while

$$\mathbb{P}_{\delta_{s_0},\pi}(A_0=a_0,S_1=s_1) = \pi(a_0|s_0)p(s_1|s_0,a_0).$$

Thus,

$$\begin{aligned}&\mathbb{P}_{\delta_{s_0},\pi}(S_1=s_1,A_1=a_1,S_2=s_2,A_2=a_2,\ldots,S_t=s_t,A_t=a_t \mid A_0=a_0,S_1=s_1) \\ &\quad= \pi(a_1|s_1)p(s_2|s_1,a_1)\cdots p(s_t|s_{t-1},a_{t-1})\pi(a_t|s_t)\,.\end{aligned}$$

On the other hand, for the expression on the right hand side of Equation (18), we have

$$\begin{aligned}&\mathbb{P}_{\delta_{s_1},\pi}(S_0=s_1,A_0=a_1,S_1=s_2,A_1=a_2,\ldots,S_{t-1}=s_t,A_{t-1}=a_t) \\ &\quad= \pi(a_1|s_1)p(s_2|s_1,a_1)\cdots p(s_t|s_{t-1},a_{t-1})\pi(a_t|s_t)\,,\end{aligned}$$

which finishes the proof of Equation (18), and thus also the proof of the statement. □

4 Action-value functions

Definition 4.1 (Probability measure induced by using a policy in an MDP starting from a state-action distribution)**.** Fix a finite MDP $M = (\mathcal{S}, \mathcal{A}, \mathcal{R}, p)$, a policy π of this MDP, and a distribution $\nu \in \mathcal{M}_1(\mathcal{S} \times \mathcal{A})$. Then, the probability distribution $\mathbb{P}$—induced by using ν as the initial state-*action* distribution and then following π in M—is a probability distribution $\mathbb{P}$ over some sample space Ω such that there are random variables $S_0, S_1, \ldots : \Omega \to \mathcal{S}$, $A_0, A_1, \ldots : \Omega \to \mathcal{A}$, and $R_1, R_2, \ldots : \Omega \to \mathcal{R}$, for which the following holds for every $t \geq 0$ and $h_t = (s_0, a_0, r_1, s_1, a_1, \ldots, a_{t-1}, r_t, s_t) \in \mathcal{H}_t$:

$$\begin{aligned}\mathbb{P}(H_t = h_t) = \nu(s_0, a_0) \cdot p(r_1, s_1 | s_0, a_0) \\ \cdot \pi_1(a_1 | s_0, a_0, r_1, s_1) \cdot p(r_2, s_2 | s_1, a_1) \\ \vdots \\ \cdot \pi_{t-1}(a_{t-1} | s_0, a_0, \ldots, s_{t-1}) \cdot p(r_t, s_t | s_{t-1}, a_{t-1}) \,.\end{aligned}$$

When the dependence of $\mathbb{P}$ on ν and π is important, we use $\mathbb{P}_{\nu,\pi}$ to signify this dependence. To simplify notation, we also use $\mathbb{P}$, $\mathbb{P}_\nu$, and $\mathbb{P}_\pi$, when one, or both of these objects are clear from the context.

Again, Kolmogorov's extension theorem guarantees the existence of $\mathbb{P}_{\nu,\pi}$.
The distribution also satisfies a proposition similar to Proposition 1.6:

Proposition 4.2. *The probability measure $\mathbb{P} := \mathbb{P}_{\nu,\pi}$ satisfies the following properties: For any $t \geq 0$, $h_t = (s_0, a_0, r_1, s_1, a_1, \ldots, r_t, s_t) \in \mathcal{H}_t$, $r' \in \mathcal{R}$, and $s' \in \mathcal{S}$, it holds that*

$$\mathbb{P}(S_0 = s_0, A_0 = a_0) = \nu(s_0, a_0)\,, \tag{13}$$

$$\mathbb{P}(A_t = a_t | H_t = h_t) = \pi_t(a_t | h_t), \quad \text{for } t \geq 1, \tag{14}$$

$$\mathbb{P}(R_{t+1} = r', S_{t+1} = s' | H_t = h_t, A_t = a_t) = p(r', s' | s_t, a_t)\,. \tag{15}$$

Further, if some probability measure $\mathbb{P}$ satisfies the above properties, then it also satisfies the properties used in Definition 4.1.

The next proposition follows from the definition.

Proposition 4.3. *For any $\mu \in \mathcal{M}_1(\mathcal{S})$ and policy π, if we define $\nu \in \mathcal{M}_1(\mathcal{S} \times \mathcal{A})$ to be*

$$\nu(s, a) = \mu(s) \cdot \pi(a | s)\,, \qquad (s, a) \in \mathcal{S} \times \mathcal{A}\,,$$

then we can choose a sample space Ω and the probability distributions $\mathbb{P}_{\nu,\pi}$ and $\mathbb{P}_{\mu,\pi}$ over Ω such that

$$\mathbb{P}_{\nu,\pi} = \mathbb{P}_{\mu,\pi}\,.$$

We also have the following strengthening of Proposition 1.7:

Proposition 4.4 (Single sample space)**.** *There exist a sample space Ω and random variables $S_0, S_1, \ldots : \Omega \to \mathcal{S}$, $A_0, A_1, \ldots : \Omega \to \mathcal{A}$, and $R_1, R_2, \ldots : \Omega \to \mathcal{R}$ such that the following hold:*

1. *for any policy π and initial state distribution $\mu \in \mathcal{M}_1(\mathcal{S})$, there exists a probability measure $\mathbb{P}_{\mu,\pi} \in \mathcal{M}_1(\Omega)$ that satisfies Definition 1.4; and*

2. *for any policy π and initial state distribution $\nu \in \mathcal{M}_1(\mathcal{S} \times \mathcal{A})$, there exists a probability measure $\mathbb{P}_{\nu,\pi} \in \mathcal{M}_1(\Omega)$ that satisfies Definition 4.1.*

The first condition in the above proposition is Proposition 1.7.

Similarly to δ_s, for $(s, a) \in \mathcal{S} \times \mathcal{A}$, we define $\delta_{s,a} \in \mathcal{M}_1(\mathcal{S} \times \mathcal{A})$ as follows:

$$\delta_{s,a}(s', a') = \mathbb{I}\{s = s', a = a'\}\,.$$

Definition 4.5 (Action-value function of a policy)**.** Fix an arbitrary policy π. The value function $q_\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ of π is defined by

$$q_\pi(s, a) = \mathbb{E}_{\delta_{s,a},\pi}\left[\sum_{t\geq 0} \gamma^t R_{t+1}\right], \qquad s \in \mathcal{S}\,, a \in \mathcal{A}\,.$$

In words, $q_\pi(s, a)$ is the expected total discounted reward incurred when we start at state s, use action a for the first time step, and in the subsequent timesteps follow policy π.

The next proposition gives the relationship between the two value functions v_π and q_π.

Proposition 4.6. *We have*

$$v_\pi(s) = \sum_{s,a} \pi(a|s) q_\pi(s, a)\,.$$

5 Why so complicated?

Most papers and books introduce value (and action-value) functions through conditioning. The formulae we see look as follows:

$$v_\pi(s) = \mathbb{E}[\sum_{t\geq 0} \gamma^t R_{t+1} | S_0 = s]\,, \qquad s \in \mathcal{S}\,.$$

In comparison to our definition (cf. Equation (4)) this looks nice, simple and elegant. However, careful inspection reveals that some things are lacking. First, the left-hand side depends on π, while the right-hand side does not show any dependence on π. What does π influence on the right-hand side? It must influence the distribution of the rewards, which comes into play through $\mathbb{E}$. Hence, at least $\mathbb{E}$ should be indexed by π, which is indeed what happens for example in our textbook. Then, the definition looks as follows:

$$v_\pi(s) = \mathbb{E}_\pi\left[\sum_{t\geq 0} \gamma^t R_{t+1} | S_0 = s\right], \qquad s \in \mathcal{S}\,. \tag{16}$$

Of course, this leaves the question of what exactly $\mathbb{E}_\pi$ stands for? An expectation operator must correspond to some probability measure. So there must be an underlying probability measure $\mathbb{P}_\pi$ defined over some event space with some properties. What are these? If we attempt a definition of $\mathbb{P}_\pi$ we discover that we want $\mathbb{P}_\pi$ to be defined over an event space that holds the state, action and reward variables and that it must satisfy certain properties. If we spell these out, we quickly arrive at something like Definition 1.4. And of course there must be an initial state distribution that $\mathbb{P}_\pi$ depends on (in fact, whatever $\mathbb{P}_\pi$ is, $\mu(s) = \mathbb{P}_\pi(S_0 = s)$, $s \in \mathcal{S}$, gives this distribution).

Now, back to the definition of value functions with the help of conditioning: Recall that a conditional expectation $\mathbb{E}[\cdot|A]$ for an event $A \subset \Omega$ is simply the expectation with respect to the probability measure $\mathbb{P}_A(\cdot)$ defined by $\mathbb{P}_A(B) = \mathbb{P}(A \cap B)/\mathbb{P}(A)$ provided that $\mathbb{P}(A) > 0$. If, for example, X is a random variable over Ω that takes values in the discrete set $\mathcal{X} \subset \mathbb{R}$,

$$\mathbb{E}[X|A] = \sum_{x \in \mathcal{X}} x \; \mathbb{P}_A(X = x) \qquad \left(= \sum_{x \in \mathcal{X}} x \; \mathbb{P}(X = x|A) \right) .$$

However, notice that the definition does not say anything about how $\mathbb{E}[X|A]$ should be defined when $\mathbb{P}(A) = 0$. In fact, in this case, the value of $\mathbb{E}[X|A]$ is, by convention, *arbitrarily* determined. That is, if one wishes, one can use $\mathbb{E}[X|A] = 2$ or $\mathbb{E}[X|A] = -2$.

Hence, the problem with Equation (16): For any state $s \in \mathcal{S}$ such that $\mathbb{P}_\pi(S_0 = s) = 0$, the value on the right-hand side has an arbitrary value. This calls for trouble as we will quickly arrive at contradictions with various arbitrarily assigned values! Of course, the astute reader may not that we should therefore be careful so that the initial state distribution assigns a positive probability to any possible state. This will indeed do it for this definition for finite state spaces (and something similar can be made to work even if uncountably infinite state spaces provided we find a probability distribution that has full support over the whole state space, which is not that trivial actually).

There are two downsides (at least) to relying on this approach of defining value functions through conditioning: First, one needs to be still specific about the initial state distribution and second, extending this idea to the definition of action-value functions is just a no-go. To see what the problem with this is consider the following "definition" of the action-value function of a policy π:

$$q_\pi(s, a) = \mathbb{E}_\pi \left[\sum_{t \geq 0} \gamma^t R_{t+1} | S_0 = s, A_0 = a \right] , \qquad s \in \mathcal{S} \, , a \in \mathcal{A} \, , \tag{17}$$

which we find even in our textbook (Eq. (3.13), page 58, choosing $t = 0$ there). The problem is immediate once we write this down: This is only a good definition if $\mathbb{P}_\pi(S_0 = s, A_0 = a) = 0$. But can we guarantee this still? The answer is, in general, no. Take a very simple case when π is a deterministic memoryless policy. Then, whatever the state $s \in \mathcal{S}$, if $a \neq \pi(s)$, then $\mathbb{P}_\pi(S_0 = s, A_0 = a) = \mathbb{P}_\pi(S_0 = s)\mathbb{P}_\pi(A_0 = a|S_0 = s) = \mathbb{P}_\pi(S_0 = s) \cdot 0$. Thus, Equation (17) just cannot be used as a definition and we are forced to introduce $\mathbb{P}_{\delta_{(s,a)},\pi}$, or at least to use $\mathbb{P}_{\nu,\pi}$ with a probability measure ν over $\mathcal{S} \times \mathcal{A}$ which is positive over all state-action pairs. With this, Equation (17) becomes a valid definition.

However, then a new problem arises. We have at least two versions of $\mathbb{P}_\pi$, one used in Equation (16) (this is $\mathbb{P}_{\mu,\pi}$ with $\mu \in \mathcal{M}_1(\mathcal{S})$) and another one used in Equation (17) (this is $\mathbb{P}_{\nu,\pi}$ with $\nu \in \mathcal{M}_1(\mathcal{S} \times \mathcal{A})$). How do we know which one to use when? This is confusing to say the least. Can we perhaps just to use $\mathbb{P}_{\nu,\pi}$ in Equation (16)? This again does not work. For that definition, it is important that the distribution of actions in $t = 0$ follow the distributions coming from the policy π.

Another compounding issue with just using $\mathbb{P}_{\mu,\pi}$ is that when we derive the Bellman equation for v_π, one needs to show that the identity

$$\mathbb{E}_{\mu,\pi}[G_1|S_1 = s'] = \mathbb{E}_{\mu,\pi}[G_0|S_0 = s'] \tag{18}$$

holds for any $s' \in \mathcal{S}$ (cf. (3.14) in the book, again choosing $t = 0$). In fact, this identity will only make sense if $\mathbb{P}_{\mu,\pi}(S_1 = s') > 0$, but a careful inspection of the formulae shows that we do not need the identity to hold in the opposite case (lucky). Still, under $\mathbb{P}_{\mu,\pi}(S_1 = s') > 0$, Equation (18) calls for a proof (which is what we have also shown in these notes). In fact, this "stationarity" property is at the heart of Markov Decision Processes and is thus a good practice to work out a formal proof for this.

In conclusion, we choose to avoid using conditioning as the basis of definitions as these definitions, when properly carried out, require exactly as much preparation and notation of what we used: There is no sparing of the discussion of the role of initial distributions in these definitions. And then, there is no advantage of using conditioning as the definition compared to indexing, which makes this dependence clear.

One final detail is that our textbook chose a definition of v_π that emphasizes the stationarity of the return process under a memoryless policy. Going back to the definition in the book, we see there

$$v_\pi(s) = \mathbb{E}_\pi[G_t | S_t = s]\,, \qquad s \in \mathcal{S}\,. \tag{19}$$

Here, index $t \geq 0$ on the right-hand side is free. The intended meaning here is probably that the right-hand sides have the same value regardless of $t \geq 0$, which is the value that we assign to the left-hand side. Of course, that this is a correct definition (ie that the values on the right-hand side match) also calls for a proof (in fact, Equation (18) is equivalent to this). Note that here choosing μ cannot in general help to make $\mathbb{P}_\pi(S_t = s)$ positive. And then, with the standard math textbook definition, which, as mentioned before, defines a conditional expectation as an arbitrary value when the conditioning even has zero probability, we definitely run into trouble.

Finally, note that a definition like Equation (19) can only work for memoryless policies. Hence, if we want to get an answer to the question of whether memoryless policies are all one needs, we will be out of luck with a definition like the above.

6 Avoiding trouble

One simple way of avoiding some of the challenges we faced in the proofs is to avoid defining the return G_t and rather go by defining value functions directly via the total expected discounted reward (rather than via the expected total discounted reward):

$$v_\pi(s) = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{\delta_s,\pi}[R_{t+1}]\,.$$

Now, there is no need to reason about the existence of G_0, the existence of the expectation of G_0, or whether the infinite sum over the time steps can be moved outside of the expectation as we started with a definition where the infinite sum is already outside of the expectation. The definition of q_π can similarly be just

$$q_\pi(s,a) = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{\delta_{s,a},\pi}[R_{t+1}]\,.$$

Of course, these definitions are equivalent to the earlier ones when the earlier definitions "work". The definitions with the sum outside of the expectation is perhaps less intuitive. It is a good challenge to figure

out whether there are cases when the definition with the sum outside works, while the definition with the sum inside the expectation does not work (the expectation does not exist).

The new definitions also invite a different approach to proving the Bellman equation for memoryless policies:

Proposition 6.1. *Let π be a memoryless policy and for $t \geq 0$, let $r_\pi^{(t)}(s) = \mathbb{E}_{\delta_s,\pi}[R_{t+1}]$, $s \in \mathcal{S}$. Then,*

$$r_\pi^{(t)} = P_\pi^t r_\pi \,,$$

where we identify a function $f : \mathcal{S} \to \mathbb{R}$ with the vector $(f(s_1), \ldots, f(s_S))^\top$ and $P_\pi \in \mathbb{R}^{S\times S}$ is the matrix whose (i,j)th element is $(P_\pi)_{i,j} = \sum_{a\in\mathcal{A}} \pi(a|s_i)p(s_j|s_i,a)$ and $r_\pi(s) = \sum_{a\in\mathcal{A}} \pi(a|s)r(s,a)$, where $\mathcal{S} = \{s_1, \ldots, s_S\}$ and $|\mathcal{S}| = S$.

In the proposition P_π^t means the t-fold product of P_π with itself. In words, $r_\pi^{(t)}$ gives the vector of expected rewards after following π in the MDP for t transitions for the various initial states (and in particular, $r_\pi^{(0)} = r_\pi$). The proposition can be proved by induction on $t = 0, 1, 2, \ldots$.

Based on this proposition, the proof of the Bellman equation for π is immediate. First, note that v_π is well-defined. For $\|r\| = \max_{s\in\mathcal{S}} |r(s)|$, note that $\|r_\pi^{(t)}\| \leq r_{\max}$. By the triangle inequality, for $s \geq 0$,

$$\|\sum_{t\geq s} \gamma^t r_\pi^{(t)}\| \leq \sum_{t\geq s} \gamma^t \|r_\pi^{(t)}\| \leq r_{\max}\gamma^s/(1-\gamma) \to 0 \quad \text{as } s \to \infty \,,$$

hence, $\sum_{t\geq 0} \gamma^t r_\pi^{(t)}$ is convergent. It then follows that

$$v_\pi = \sum_{t\geq 0} \gamma^t r_\pi^{(t)} = \sum_{t\geq 0} \gamma^t P_\pi^t r_\pi = r_\pi + \gamma P_\pi \sum_{t\geq 0} \gamma^t P_\pi^t r_\pi = r_\pi + \gamma P_\pi \sum_{t\geq 0} \gamma^t r_\pi^{(t)} = r_\pi + \gamma P_\pi v_\pi \,.$$

Calculating with vectors and matrices are not only useful for very clean proofs, but they are also useful for computation. In particular, from the above, we see that

$$(I - \gamma P_\pi) v_\pi = r_\pi \,,$$

where I is the $S \times S$ identity matrix. Thus, v_π satisfies a linear system of equations. Thus, v_π can be calculated by first calculating the matrix $I - \gamma P_\pi$ and then solving the above equation with a standard linear algebra method for v_π. (Note that we already concluded that a solution to this equation exist. That it has a single solution, which also means that $I - \gamma P_\pi$ is invertible, will be seen in the lecture on dynamic programming.)

Notes

1. Note that we abuse the notation by sometimes writing $p(r', s'|s, a)$, such as in this definition, whereas at other times, we might write $p(s', r'|s, a)$. We want to assert to the reader that both these expressions are equivalent in the sense that both return the probability of observing the next state s' and reward r' given the current state s and the action a, i.e. the ordering of the first two variables is immaterial.

2. Note that we overload the notation by using the same symbol r for both the immediate reward function $r(s, a)$, and the variable denoting a particular value of the reward obtained r (or r'). However, this should not cause any confusion, and the usage should be clear from context.

3. This is just the tower property of expectations, which states that for any random variables X, Y, $\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$, provided that the expectation of X exists. Indeed, for Y discrete, $\mathbb{E}[\mathbb{E}[X|Y]] = \sum_{y \in \mathcal{Y}} \mathbb{P}(Y = y)\mathbb{E}[X|Y = y]$, where $\mathcal{Y}$ is the set of values Y takes with positive probability. For the sake of simplifying calculations, in the above expression we can replace $\mathcal{Y}$ with a superset of it, $\mathcal{Y}' \supset \mathcal{Y}$ if for $y \in \mathcal{Y}' \setminus \mathcal{Y}$ we give some meaning to $\mathbb{E}[X|Y = y]$. Indeed, in what follows we will always do this. But what value to give to this expression? Since this value will always gets multiplied by $\mathbb{P}(Y = y) = 0$, it turns out that this choice makes no difference: All the calculations give the same result regardless of what value we assign to $\mathbb{E}[X|Y = y]$. In summary, in what follows we allow expressions of the form $\mathbb{E}[X|Y = y]$ even when $\mathbb{P}(Y = y)$, in which case we assign an arbitrary value to this expression. Note that this is a standard convention.