1 Exercise

Consider the following system.

State variables: $X = \{b, c\}$
Initial state : $b \vee \neg c$
Transition relation: $(b@0 \leftrightarrow c@1) \wedge (c@0 \leftrightarrow b@1)$

Derive a formula that represents those states that are reachable by two steps with the transition relation, by using the logic-based image operation.

Solution

We are computing the image *of $b \vee \neg c$ w.r.t. the transition relation formula, and then again the image of the resulting formula with the same transition relation formula.*
The first step is forming the relevant conjunction of the formula for the set of current states *and the formula for the* transition relation*.*

$$(b@0 \vee \neg c@0) \wedge (b@0 \leftrightarrow c@1) \wedge (c@0 \leftrightarrow b@1)$$

Then we existentially abstract away $b@0$ and $c@0$. We do this by first abstracting $c@0$, and then $b@0$ (or we could do it the other way round, the result will be logically the same.)

$$\begin{aligned}
&\exists b@0 \exists c@0\, ((b@0 \vee \neg c@0) \wedge (b@0 \leftrightarrow c@1) \wedge (c@0 \leftrightarrow b@1)) \\
&= \exists b@0(((b@0 \vee \neg\bot) \wedge (b@0 \leftrightarrow c@1) \wedge (\bot \leftrightarrow b@1)) \vee ((b@0 \vee \neg\top) \wedge (b@0 \leftrightarrow c@1) \wedge (\top \leftrightarrow b@1))) \\
&= \exists b@0(((b@0 \leftrightarrow c@1) \wedge \neg b@1) \vee (b@0 \wedge (b@0 \leftrightarrow c@1) \wedge b@1)) \\
&= (((\bot \leftrightarrow c@1) \wedge \neg b@1) \vee (\bot \wedge (\bot \leftrightarrow c@1) \wedge b@1)) \vee (((\top \leftrightarrow c@1) \wedge \neg b@1) \vee (\top \wedge (\top \leftrightarrow c@1) \wedge b@1)) \\
&= (\neg c@1 \wedge \neg b@1) \vee (c@1 \wedge \neg b@1) \vee (c@1 \wedge b@1) \\
&= (c@1 \wedge \neg b@1)
\end{aligned}$$

To complete the first image computation, we rename all the variables by replacing $@1$ by $@0$, thereby obtaining.

$$(c@0 \wedge \neg b@0)$$

The second image computation proceeds analogously. Obviously, the transition relation can be viewed as exchanging c and b, and therefore the result will be

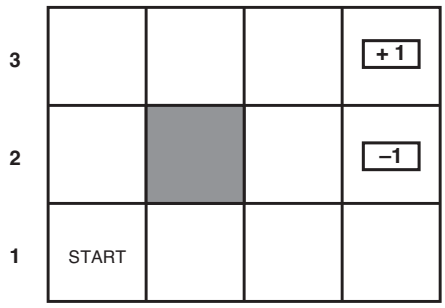
$$(b@0 \wedge \neg c@0),$$

and the set of states reachable by two steps of the transition relation is the one represented by

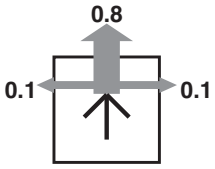
$$b \wedge \neg c.$$

2 Exercise

For the 4×3 world shown in figure below, calculate which squares can be reached from START by the action sequence [Up, Up, Right, Right, Right] and with what probabilities. The intended outcome occurs with probability 0.8, but with probability 0.2 the agent moves at right angles to the intended direction. Collision with a wall (including the walls of the gray cell) results in no movement.



(a)



(b)

Solution

The difficult way of solving is to identify all possible paths and calculate their probabilities. A much easier, dynamic-programming style solution is to tabulate the probability of "state (x, y) is occupied at time t".

At time 0, only state (1,1) is possible, and its probability is 1.0. For time 1, calculate probabilities of all states, based on the probabilities of their predecessors. And after that, same for time 2 and all states, and so on. This results in the following table.

	last action and time					
		↑	↑	→	→	→
	0	*1*	*2*	*3*	*4*	*5*
(1,1)	*1.0*	*0.1*	*0.02*	*?*		
(2,1)		*0.1*	*0.09*	*?*		
(3,1)			*0.01*	*?*		
(4,1)				*?*		
(1,2)		*0.8*	*0.24*	*?*		
(2,2)	□	□	□	□	□	□
(3,2)				*?*		
(4,2)						
(1,3)			*0.64*	*?*		
(2,3)				*?*		
(3,3)						
(4,3)						

This can be filled in by hand, but if the table was a bit bigger, filling it in by a small program would be much easier.

3 Exercise

Sometimes Markov decision processes (MDPs) are formulated with a reward function $R(s)$, $R(s, a)$, or $R(s, a, s')$.

1. Write (and simplify) Bellman equations for these formulations.
2. Show how an MDP with reward function $R(s, a, s')$ can be transformed into a different MDP with reward function $R(s, a)$, such that optimal policies in the new MDP correspond exactly to optimal policies in the original MDP.

Solution

1. *The Bellman equation is*

$$v(s) = \max_{a \in A} \sum_{s' \in S} P(s, a, s')[R(s, a, s') + \gamma v(s')].$$

If $R(s, a, s') = R(s, a)$ for some 2-place function R, then there is no dependency of the reward from the successor state s' in $\sum_{s' \in S}$, allowing the reward term to be moved outside the sum.

$$v(s) = \max_{a \in A} \sum_{s' \in S} P(s, a, s')[R(s, a) + \gamma v(s')] \tag{1}$$

$$= \max_{a \in A} \left(R(s, a) + \sum_{s' \in S} P(s, a, s')\gamma v(s') \right). \tag{2}$$

Here important is of course that for any given s and a, $\sum_{s' \in S} P(s, a, s') = 1$.

If $R(s,a,s') = R(s)$ for some 1-place function R, then the immediate reward does not depend on the action in $\max_{a \in A}$ either, and can be moved outside the maximization too.

$$v(s) = \max_{a \in A} \left(R(s) + \sum_{s' \in S} P(s,a,s') \gamma v(s') \right) \tag{3}$$

$$= R(s) + \max_{a \in A} \left(\sum_{s' \in S} P(s,a,s') [\gamma v(s')] \right). \tag{4}$$

Here we use the algebraic property $\max_{a \in A}(g + f(a)) = g + \max_{a \in A} f(a)$ that holds whenever g is independent of a.

2. *We construct an MDP that has exactly the same optimal policies, but, the sequence of rewards on a particular execution of the respective optimal policies may differ.*
We define the new reward function as

$$R(s,a) = \sum_{s' \in S} P(s,a,s') R(s,a,s')$$

This means that the expected value of the reward we always get when taking action a in state s is the same as the expected value of the possible *different rewards when taking action a in state s and* ending up in different successor states*. Since the expected value of the immediate reward for a given state and action is the same as in the original MDP, the new MDP has exactly the same optimal policies with exactly the same values for all states.*
In the original MDP the expected immediate reward *(ignoring the term $\gamma v(s')$ in the Bellman equation) in state s with action a is*

$$\sum_{s' \in S} P(s,a,s') R(s,a,s').$$

In the new MDP the expected immediate reward is by definition

$$\begin{aligned} &\sum_{s' \in S} P(s,a,s') R(s,a) \\ &= R(s,a) + \sum_{s' \in S} P(s,a,s') \cdot 0 \\ &= \sum_{s' \in S} P(s,a,s') R(s,a,s') + 0 \end{aligned}$$

These are obviously the same.