Finite Markov Decision Processes

Exercises from Chapter 3

Reference book: Reinforcement Learning, an introduction - 2nd Edition

Book club February 26, 2021

Exercise 3.1

Devise three example tasks of your own that fit into the MDP framework, identifying for each its states, actions, and rewards. [...]

Exercise 3.1

Devise three example tasks of your own that fit into the MDP framework, identifying for each its states, actions, and rewards. [...]

We give three examples:

- a robot escaping a maze;
- a robot picking up trash with a bionic arm;
- a robot heating an oven to a selected temperature;

Exercise 3.2

Is the MDP framework adequate to usefully represent all goal-directed learning tasks? Can you think of any clear exceptions?

Exercise 3.2

Is the MDP framework adequate to usefully represent all goal-directed learning tasks? Can you think of any clear exceptions?

Some conditions could be violated:

- we could not be able to compute all the possible states (*dimensionality problem*);
- the Markovian property could be violated and past has no observable in the present;

Exercise 3.3

Consider the problem of driving. [...] What is the right level, the right place to draw the line between agent and environment? [...] Is there any fundamental reason for preferring one location over another, or is it a free choice?

Exercise 3.3

Consider the problem of driving. [...] What is the right level, the right place to draw the line between agent and environment? [...] Is there any fundamental reason for preferring one location over another, or is it a free choice?

For what I understood the line between the environment and the agent should be drawn in a way that:

- actions from the agent have a direct effect and response from the environment;
- these responses are related to the goal in some ways;

Exercise 3.4

Give a table analogous to that in Example 3.3, but for $p(s', r|s, a)$. It should have columns for s, a, s', r, and $p(s', r|s, a)$, and a row for every 4-tuple for which $p(s', r|s, a) > 0$.

Exercise 3.4

Give a table analogous to that in Example 3.3, but for $p(s', r|s, a)$. It should have columns for s, a, s', r, and $p(s', r|s, a)$, and a row for every 4-tuple for which $p(s', r|s, a) > 0$.

s	a	s'	r	$p(s', r\|s, a)$
high	search	high	r_{search}	α
high	search	low	r_{search}	$1 - \alpha$
high	wait	high	r_{wait}	1
high	wait	low	-	0
low	search	high	-3	$1 - \beta$
low	search	low	r_{search}	β
low	wait	high	-	0
low	wait	low	r_{wait}	1
low	recharge	high	0	1
low	recharge	low	-	0

Exercise 3.5

The equations in Section 3.1 are for the continuing case and need to be modified to apply to episodic tasks. Show that you know the modifications needed by giving the modified version of (3.3).

Exercise 3.5

The equations in Section 3.1 are for the continuing case and need to be modified to apply to episodic tasks. Show that you know the modifications needed by giving the modified version of (3.3).

Episodic tasks have a special state (the terminal state) with transition probability that is:

- 1 to itself;
- 0 to all other states;

We consider that state only as a possible arriving state, so that $s' \in \mathbf{S}^+$ but $s \in \mathbf{S}$:

$$\sum_{r_j \in \mathbb{R}} \sum_{s' \in \mathbb{S}^+} p(s', r|s, a) = 1 \qquad \forall s \in \mathbb{S}, a \in \mathbb{A}$$

Exercise 3.6

Suppose you treated pole-balancing as an episodic task but also used discounting, with all rewards zero except for -1 upon failure. What then would the return be at each time? How does this return differ from that in the discounted, continuing formulation?

Exercise 3.6

Suppose you treated pole-balancing as an episodic task but also used discounting, with all rewards zero except for -1 upon failure. What then would the return be at each time? How does this return differ from that in the discounted, continuing formulation?

- for episodic discounted tasks we have a terminal state and a terminal reward which here is the only non-zero but -1 reward:

$$G_t = \sum_{k=0}^{T-t-1} \gamma^k R_{k+t+1} = \gamma^{T-t-1} R_T = -\gamma^{T-t-1};$$

- for continuing tasks:

$$G_t = -\gamma^{-K}.$$

  where K is the time step before the failure;

Exercise 3.7

Imagine that you are designing a robot to run a maze. [...] After running the learning agent for a while, you find that it is showing no improvement in escaping from the maze. What is going wrong? Have you effectively communicated to the agent what you want it to achieve?

Exercise 3.7

Imagine that you are designing a robot to run a maze. [...] After running the learning agent for a while, you find that it is showing no improvement in escaping from the maze. What is going wrong? Have you effectively communicated to the agent what you want it to achieve?

- Yes escaping a maze could be seen as episodic but if I run in circles I could never escape the maze and learn nothing during an episode;

Exercise 3.7

Imagine that you are designing a robot to run a maze. [...] After running the learning agent for a while, you find that it is showing no improvement in escaping from the maze. What is going wrong? Have you effectively communicated to the agent what you want it to achieve?

- Yes escaping a maze could be seen as episodic but if I run in circles I could never escape the maze and learn nothing during an episode;
- if we add a discount or a punishment for each time step wasted the agent learn how to escape before the end of an episode;

Exercise 3.8

Suppose $\gamma = 0.5$ and the following sequence of rewards is received $R_1 = -1$, $R_2 = 2$, $R_3 = 6$, $R_4 = 3$, and $R_5 = 2$, with $T = 5$. What are G_0, G_1, G_2, G_3, G_4 and G_5? Hint: Work backwards.

Exercise 3.8

Suppose $\gamma = 0.5$ and the following sequence of rewards is received $R_1 = -1$, $R_2 = 2$, $R_3 = 6$, $R_4 = 3$, and $R_5 = 2$, with $T = 5$. What are G_0, G_1, G_2, G_3, G_4 and G_5? Hint: Work backwards.

We work backwards from 5 to 0:

- $G_T = G_5 = 0$;
- $G_4 = R_5 + \gamma G_5 = R_5 = 2$;
- $G_3 = R_4 + \gamma G_4 = R_4 + 2\gamma = 4$;
- $G_2 = R_3 + \gamma G_3 = R_3 + 4\gamma = 8$;
- $G_1 = R_2 + \gamma G_2 = R_2 + 8\gamma = 6$;
- $G_0 = R_1 + \gamma G_1 = R_1 + 6\gamma = 2$;

Exercise 3.9

Suppose $\gamma = 0.9$ and the reward sequence is $R_1 = 2$ followed by an infinite sequence of 7s. What are G_1 and G_0?

Exercise 3.9

Suppose $\gamma = 0.9$ and the reward sequence is $R_1 = 2$ followed by an infinite sequence of 7s. What are G_1 and G_0?

$$G_1 = \sum_{k=0}^{\infty} \gamma^k (R_{i>2} = 7) = 7 \cdot \frac{1}{1-\gamma} = 70;$$

$$G_0 = R_1 + \gamma G_1 = 2 + \gamma 70 = 65;$$

Exercise 3.10

Prove the second equality in (3.10) (geometric series).

Exercise 3.10

Prove the second equality in (3.10) (geometric series).

$$\sum_{k=0}^{\infty} \gamma^k = \lim_{n\to\infty} \sum_{k=0} \gamma^k = \lim_{n\to\infty} \sum_{k=0} \gamma^k \frac{(1-\gamma)}{(1-\gamma)} =$$

$$= \lim_{n\to\infty} \frac{(1-\gamma^{n+1})}{(1-\gamma)} = \frac{1}{(1-\gamma)}$$

as $\gamma^{n+1} \xrightarrow{n\to\infty} 0$ for $|\gamma < 1|$.
We used the progressive geometric series:

$$\sum_{k=0}^{n} \gamma^k \cdot (1-\gamma) = \left((1-\gamma) + (\gamma - \gamma^2) + \cdots + (\gamma^n - \gamma^{n+1})\right)$$

$$= \left(1 - \gamma^{n+1}\right)$$

Exercise 3.11

If the current state is S_t, and actions are selected according to stochastic policy π, then what is the expectation of R_{t+1} in terms of π and the four-argument function p (3.2)?

Exercise 3.11

If the current state is S_t, and actions are selected according to stochastic policy π, then what is the expectation of R_{t+1} in terms of π and the four-argument function p (3.2)?

$$\mathbb{E}\left[R_{t+1}|S_t = s\right] = \sum_a \pi(a|S_t) \sum_{s'} \sum_r r \cdot p(s', r|a, s)$$

Exercise 3.12

Give an equation for v_π in terms of q_π and π?

Exercise 3.12

Give an equation for v_π in terms of q_π and π?

$$v_\pi(s) = \sum_a q_\pi(s, a)\pi(a|s)$$

Exercise 3.13

Give an equation for q_π in terms of v_π and the four-argument p.

Exercise 3.13

Give an equation for q_π in terms of v_π and the four-argument p.

$$q_\pi(s,a) = \sum_{s'} \sum_{r} [r + \gamma v_\pi(s')]\, p(s', r|s, a)$$

Exercise 3.14

The Bellman equation (3.14) must hold for each state for the value function v_π shown in Figure 3.2 (right) of Example 3.5. Show numerically that this equation holds for the center state, valued at +0.7, with respect to its four neighboring states, valued at +2.3, +0.4, -0.4, and +0.7.

Exercise 3.14

The Bellman equation (3.14) must hold for each state for the value function v_π shown in Figure 3.2 (right) of Example 3.5. Show numerically that this equation holds for the center state, valued at +0.7, with respect to its four neighboring states, valued at +2.3, +0.4, -0.4, and +0.7.

Bellman equation can be written as:

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s'} \sum_r \left[r + \gamma v_{pi}(s') \right] p(s', r|s, a)$$

so that in this case:

$$v_{center} = \frac{0.9 \cdot (2.3 + 0.7 + 0.4 - 0.4)}{4} \approx 0.7$$

Exercise 3.15

[...] Prove, using (3.8), that adding a constant c to all the rewards adds a constant, v_c, to the values of all states, and thus does not affect the relative values of any states under any policies.[...]

Exercise 3.15

[...] Prove, using (3.8), that adding a constant c to all the rewards adds a constant, v_c, to the values of all states, and thus does not affect the relative values of any states under any policies.[...]

$$G_t \doteq \sum_{k=0}^{\infty} \gamma^k R_{k+t+1} \qquad \Longrightarrow$$

$$\begin{aligned}
\tilde{G}_t &= \sum_{k=0}^{\infty} \gamma^k \left(R_{k+t+1} + C\right) \\
&= G_t + C \sum_{k=0}^{\infty} \gamma^k \\
&= G_t + \frac{C}{1-\gamma}
\end{aligned}$$

Exercise 3.15

[...] Prove, using (3.8), that adding a constant c to all the rewards adds a constant, v_c, to the values of all states, and thus does not affect the relative values of any states under any policies.[...]

$$G_t \doteq \sum_{k=0}^{\infty} \gamma^k R_{k+t+1} \qquad \Longrightarrow \qquad v_\pi^* \doteq \mathbb{E}_\pi \left[G_t | S_t = s \right]$$

$$\begin{aligned} \tilde{G}_t &= \sum_{k=0}^{\infty} \gamma^k \left(R_{k+t+1} + C \right) \\ &= G_t + C \sum_{k=0}^{\infty} \gamma^k \\ &= G_t + \frac{C}{1-\gamma} \end{aligned}$$

$$\begin{aligned} \tilde{v}_\pi^* &= \mathbb{E}_\pi \left[\tilde{G}_t | S_t = s \right] \\ &= \mathbb{E}_\pi \left[G_t + \frac{C}{1-\gamma} \mid S_t = s \right] \\ &= v_\pi^* + \frac{C}{1-\gamma} \end{aligned}$$

Exercise 3.16

Now consider adding a constant c to all the rewards in an episodic task, such as maze running. Would this have any effect, or would it leave the task unchanged as in the continuing task above? Why or why not? Give an example.

Exercise 3.16

Now consider adding a constant c to all the rewards in an episodic task, such as maze running. Would this have any effect, or would it leave the task unchanged as in the continuing task above? Why or why not? Give an example.

- Adding a constant C to the reward change the value function only by a constant;
- The sign of this constant in an episodic task because negative rewards are used to accelerate the learning: goals must be reached before the end of the episode.

Exercise 3.17

What is the Bellman equation for action values, that is, for q_π? It must give the action value $q_\pi(s,a)$ in terms of the action values, $q_\pi(s', a')$, of possible successors to the state–action pair (s, a). [...]

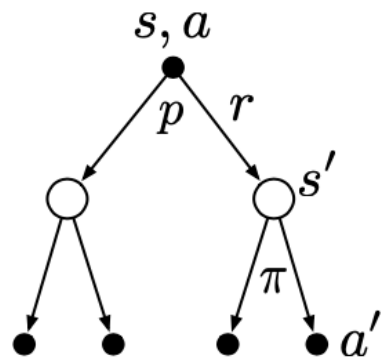
Exercise 3.17

What is the Bellman equation for action values, that is, for q_π? It must give the action value $q_\pi(s,a)$ in terms of the action values, $q_\pi(s', a')$, of possible successors to the state–action pair (s, a). [...]

$$q_\pi(s,a) \doteq \mathbb{E}\left[G_t \mid S_t = s, A_t = a\right]$$

$$= \mathbb{E}\left[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a\right]$$

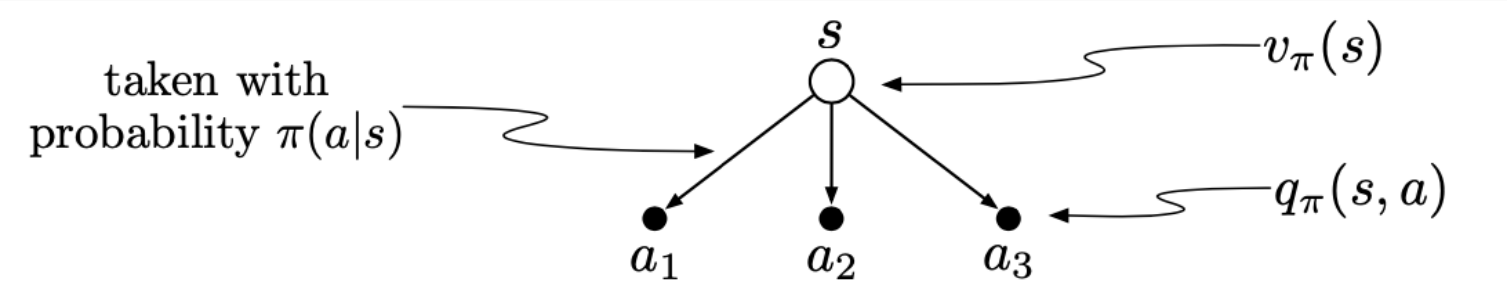
$$= \sum_{s'} \sum_{r} p(s', r|s, a) \left[r + \gamma \sum_{a'} \pi(a'|s') q_\pi(s', a') \right]$$



q_π backup diagram

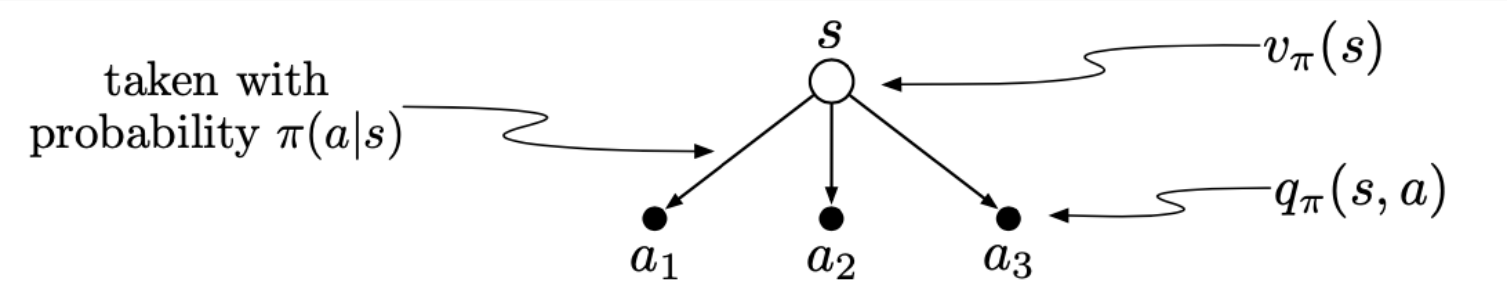
Exercise 3.18

[...] Give the equation corresponding to this intuition and diagram for the value at the root node, $v_\pi(s)$, in terms of the value at the expected leaf node, $q\pi(s, a)$, given $S_t = s$. [...] Then give a second equation in which the expected value is written out explicitly in terms of $\pi(a|s)$ [...].



Exercise 3.18

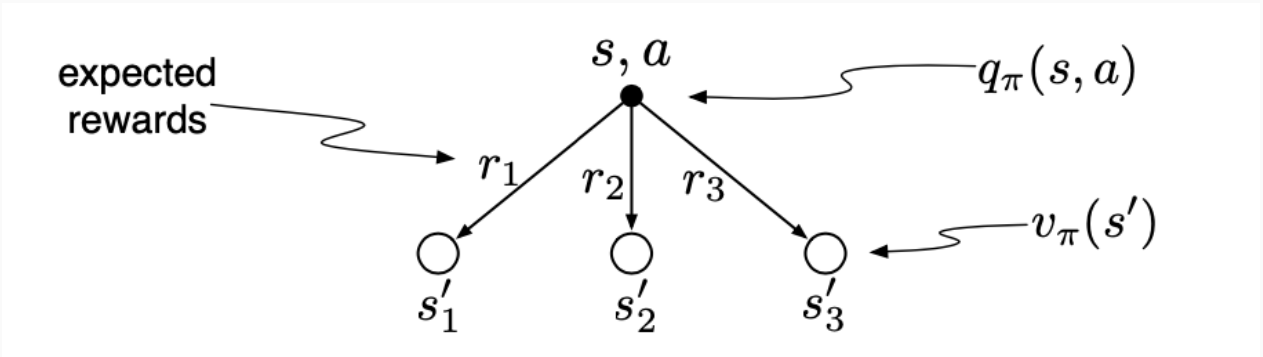
[...] Give the equation corresponding to this intuition and diagram for the value at the root node, $v_\pi(s)$, in terms of the value at the expected leaf node, $q\pi(s, a)$, given $S_t = s$. [...] Then give a second equation in which the expected value is written out explicitly in terms of $\pi(a|s)$ [...].



$$\begin{aligned} v_\pi(s) &\doteq \mathbb{E}_\pi\left[q_\pi(s, a)\right] \\ &= \sum_a \pi(a|s) q_\pi(s, a) \end{aligned}$$

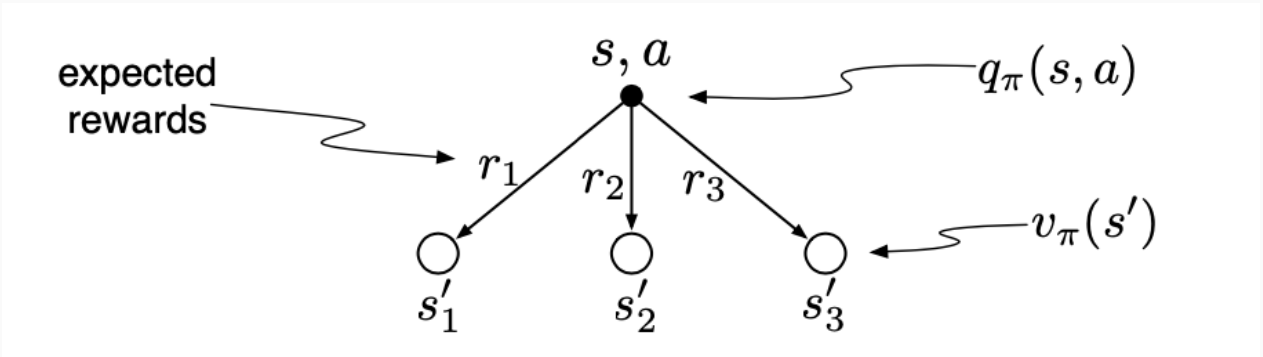
Exercise 3.19

[...] Give the equation corresponding to this intuition and diagram for the action value, $q_pi(s, a)$, in terms of the expected next reward, R_{t+1}, and the expected next state value, $v_\pi(S_{t+1})$, given that $S_t = s$ and $A_t = a$. [...] Then give a second equation, writing out the expected value explicitly in terms of $p(s', r|s, a)$ [...].



Exercise 3.19

[...] Give the equation corresponding to this intuition and diagram for the action value, $q_pi(s, a)$, in terms of the expected next reward, R_{t+1}, and the expected next state value, $v_\pi(S_{t+1})$, given that $S_t = s$ and $A_t = a$. [...] Then give a second equation, writing out the expected value explicitly in terms of $p(s', r|s, a)$ [...].



$$
\begin{aligned}
q_\pi(s, a) &\doteq \mathbb{E}_\pi \left[R_{t+1} + \gamma v_\pi(S_{t+1}) | S_t = s, A_t = a \right] \\
&= \sum_{s'} \sum_{r} p(s', r|s, a) \left[r + \gamma v_\pi(s') \right]
\end{aligned}
$$

Exercise 3.22

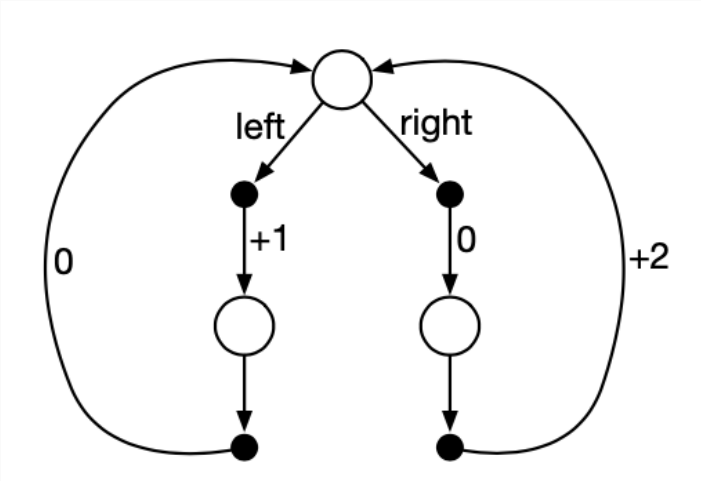
Consider the continuing MDP shown on to the right. The only decision to be made is that in the top state [...]. There are exactly two deterministic policies, π_{left} and π_{right}. What policy is optimal if $\gamma = 0$? If $\gamma = 0.9$? If $\gamma = 0.5$?

Exercise 3.22

Consider the continuing MDP shown on to the right. The only decision to be made is that in the top state [...]. There are exactly two deterministic policies, π_{left} and π_{right}. What policy is optimal if $\gamma = 0$? If $\gamma = 0.9$? If $\gamma = 0.5$?

$$G_{\pi left} = \sum_{k=0}^{\infty} \gamma^{2k} = \frac{1}{1-\gamma^2}$$

$$G_{\pi right} = \sum_{k=0}^{\infty} 2 \cdot \gamma^{2k+1} = \frac{2\gamma}{1-\gamma^2}$$



so:

- $\gamma > 0.5$ π_{right} is optimal;
- $\gamma < 0.5$ π_{left} is optimal;
- $\gamma = 0.5$ both optimal;

Exercise 3.23

Give the Bellman equation for q^* for the recycling robot.

Exercise 3.23

Give the Bellman equation for q^* for the recycling robot.

$$
\begin{aligned}
q^*(s,a) &= \mathbb{E}_{\pi^*}\left[G_t | S_t = s, A_t = a\right] \\
&= \sum_{s'} \sum_{r} p(s', r | s, a) \left[r + \gamma \max_{a'} q^*(s', a') \right]
\end{aligned}
$$

Exercise 3.23

Give the Bellman equation for q^* for the recycling robot.

$$
\begin{aligned}
q^*(s,a) &= \mathbb{E}_{\pi^*}\left[G_t | S_t = s, A_t = a\right] \\
&= \sum_{s'}\sum_{r} p(s', r|s, a)\left[r + \gamma \max_{a'} q^*(s', a')\right]
\end{aligned}
$$

So, in example, if $state = high = h$ and $action = wait = w$:

$$
\begin{aligned}
q^*(h,w) &= \sum_{s'}\sum_{r} p(s', r|s, a)\left[r + \gamma \max_{a'} q^*(s', a')\right] \\
&= \sum_{s'=high}\sum_{r=r_{wait}} 1 \cdot [r_{wait} + \gamma \max(q^*(s', w), q^*(s', search))] \\
&= r_{wait} + \gamma \max(q^*(high, wait), q^*(high, search))
\end{aligned}
$$

Exercise 3.24

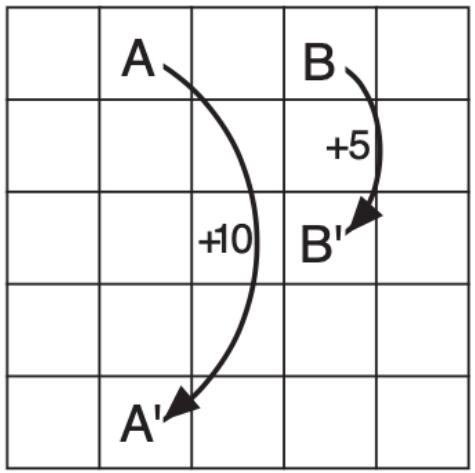
Figure 3.5 gives the optimal value of the best state of the gridworld as 24.4, to one decimal place. Use your knowledge of the optimal policy and (3.8) to express this value symbolically, and then to compute it to three decimal places.

Exercise 3.24

Figure 3.5 gives the optimal value of the best state of the gridworld as 24.4, to one decimal place. Use your knowledge of the optimal policy and (3.8) to express this value symbolically, and then to compute it to three decimal places.

- The optimal path is going from $A \to A'$ and then step by step going back to A;
- the shortest path is then made of 5 cells so:

$$v^*(A) = \sum_{k=0}^{\infty} 10\gamma^{5k} = \frac{10}{1-\gamma^5} = 24.419$$



Gridworld

Exercise 3.25

Give an equation for v_* in terms of q_*.

Exercise 3.25

Give an equation for v^* in terms of q^*.

$$v^*(s) = \max_{a'} q^*(s, a')$$

Exercise 3.26

Give an equation for q^* in terms of v^* and the four-argument p.

Exercise 3.26

Give an equation for q^* in terms of v^* and the four-argument p.

$$q^*(s,a) = \sum_{s'} \sum_{r} p(s', r|s, a) \left[r + \gamma v^*(s') \right]$$

Exercise 3.27

Give an equation for π^* in terms of q^*.

Exercise 3.27

Give an equation for π^* in terms of q^*.

Optimal policies π^* are the ones mapping the actions a^* maximizing the $q^*(s, a)$ into their probabilities:

$$\begin{aligned} a^* &= \arg\max_a q^*(s, a) \\ &= \arg \pi^*(a^*|s) \end{aligned}$$

Exercise 3.28

Give an equation for π^* in terms of v^* and the four-argument p.

Exercise 3.28

Give an equation for π^* in terms of v^* and the four-argument p.

As the previous exercise:

$$\begin{aligned} a^* &= \arg\max_a q^*(s,a) \\ &= \arg\max_a \sum_{s'} \sum_r p(s', r|s, a) \left[r +^* (s')\right] \end{aligned}$$

Exercise 3.29

Rewrite the four Bellman equations for the four value functions in terms of the three argument p and the two-argument r.

Exercise 3.29

Rewrite the four Bellman equations for the four value functions in terms of the three argument p and the two-argument r.

$$v_\pi(s) \doteq \mathbb{E}_\pi\left[G_t | S_t = s\right] = \sum_a \pi(a|s) \left[r(s,a) + \gamma \sum_{s'} p(s'|s,a) v_\pi(s') \right]$$

$$v^*(s) \doteq \mathbb{E}_{\pi^*}\left[G_t | S_t = s\right] = \sum_a \pi^*(a|s) \left[r(s,a) + \gamma \sum_{s'} p(s'|s,a) v^*(s') \right]$$

Exercise 3.29

Rewrite the four Bellman equations for the four value functions in terms of the three argument p and the two-argument r.

$$v_\pi(s) \doteq \mathbb{E}_\pi\left[G_t|S_t = s\right] = \sum_a \pi(a|s)\left[r(s,a) + \gamma \sum_{s'} p(s'|s,a) v_\pi(s')\right]$$

$$v^*(s) \doteq \mathbb{E}_{\pi^*}\left[G_t|S_t = s\right] = \sum_a \pi^*(a|s)\left[r(s,a) + \gamma \sum_{s'} p(s'|s,a) v^*(s')\right]$$

$$q_\pi(s,a) \doteq \mathbb{E}_\pi\left[G_t|S_t = s, A_t = a\right] = \mathbb{E}_\pi\left[R_{t+1} + \gamma G_{t+1}|S_t = s, A_t = a\right] =$$
$$= r(s,a) + \gamma \sum_{s'} p(s'|s,a) \sum_{a'} \pi(a'|s') q_\pi(s',a')$$

$$q^*(s,a) = r(s,a) + \gamma \sum_{s'} p(s'|s,a) \sum_{a'} \pi^*(a'|s') q^*(s',a')$$