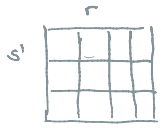
Exercises

Monday, June 1, 2020 3:58 PM

Chapter 3

Exercise 3.11 If the current state is S_t, and actions are selected according to a stochastic policy π, then what is the expectation of R_{t+1} in terms of π and the four-argument function p (3.2)? □

$p(s',r|s,a) \longrightarrow$ 

$\sum_{s,r} p(s',r|s,a) = 1$

$E[R_t | S_t = s] = \sum_a \pi(a|s_t) \cdot \sum_{s',r} p(s',r|s,a) \cdot r$?

Exercise 3.12 Give an equation for v_π in terms of q_π and π.

$v_\pi(s) = E_\pi[G_t | S_t = s] = E[\sum_{k=0}^{\infty} \gamma^k \cdot R_{t+k+1} | S_t = s]$

$= \sum_a \pi(a|s) \cdot E_\pi[G_t | S_t = s, A_t = a]$

Iteration over the whole action space, then multiply by the probability

and:

$q_\pi(s,a) = E_\pi[G_t | S_t = s, A_t = a]$

then:

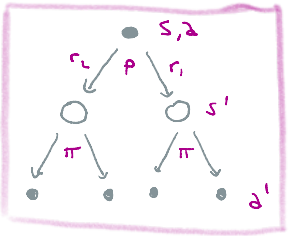
$v_\pi = \sum_a \pi(a|s) \cdot q_\pi(s,a)$

$v(s)$ is equal to the expectation of the $q(s,a)$ over all the actions.

Exercise 3.13 Give an equation for q_π in terms of v_π and the four-argument p. □

$v(s) = E_\pi[G_t | S_t = s]$

$q_\pi(s,a) = \underbrace{E_\pi[R_{t+1} | S_t = s, A_t = a]}_{①} + \gamma \cdot \underbrace{E_\pi[G_{t+1} | S_t = s, A_t = a]}_{②}$



① $E_\pi[R_{t+1} | S_t = s, A_t = a] = \underbrace{\sum_{s'} p(s' | S_t = s, A_t = a)}_{\sum_{s',r} p(s',r|s,a)} \cdot \underbrace{E_\pi[R_{t+1} | S_t = s, A_t = a, S'_{t+1} = s']}_{r(s,a,s')}$

pag. 49, 3.4 — pag. 49, 3.6

② $E_\pi[G_{t+1} | S_t = s, A_t = a] = \sum_{s'} \underbrace{E_\pi[\sum_{k=0}^{\infty} \gamma^k \cdot R_{t+k+2} | S_t = s, A_t = a, S_{t+1} = s']}_{v(s')} \cdot \underbrace{p(s' | S_t = s, A_t = a)}_{\sum_r p(s',r|s,a)}$

$S_t = s$, $A_t = a$ can be discarded since they don't give any information in a Markovian process.

Then:

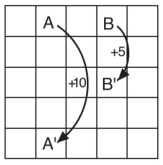
$q_\pi(s,a) = \sum_{s',r} p(s',r|s,a) \cdot r + \gamma \cdot \sum_{s',r} v(s) \cdot p(s',r|s,a) =$

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

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

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. (These numbers are accurate only to one decimal place.) □

$V(s) = 0{,}575 + 0{,}1 - 0{,}1 + 0{,}175 = 0.7$ $\quad \pi(a|s) = 0{,}25$



3.3	8.8	4.4	5.3	1.5
1.5	3.0	2.3	1.9	0.5
0.1	0.7	0.7	0.4	-0.4
-1.0	-0.4	-0.4	-0.6	-1.2
-1.9	-1.3	-1.2	-1.4	-2.0

Exercise 3.15 In the gridworld example, rewards are positive for goals, negative for running into the edge of the world, and zero the rest of the time. Are the signs of these rewards important, or only the intervals between them? 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. What is v_c in terms of c and γ? □

Equation 3.8 regulates the return at time t:

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

$$V_t = E_\pi\left[G_t \,|\, S_t = s\right]$$

Adding a c to all the rewards:

$$G_t = \sum_{k=0}^{\infty} \gamma^k \left(R_{k+t+1} + c\right)$$

$$V_\pi = E_\pi\left[\sum_{k=0}^{\infty} \gamma^k \cdot \left(R_{k+t+1} + c\right)\right] = E_\pi\left[\sum_{k=0}^{\infty} \gamma^k \cdot R_{k+t+1} + \underbrace{\sum_{k=0}^{\infty} \gamma^k \cdot c}_{\text{constant} = C}\right] = C + E_\pi\left[G_t \,|\, S_t = s\right]$$

Linearity of expectation

Esercizio 10. Prove that $\sum_{n=0}^{\infty} \gamma^k = \frac{1}{1-\gamma}$

$$\sum_{k=0}^{\infty} \gamma^k = \frac{1}{1-\gamma} \quad \text{, if } \gamma < 1$$

Proof:

$\sum_{k=0}^{\infty} \gamma^k =$ this is the sum of an infinite geometric series.

Let's start from the finite geometric series:

$$S_n = \sum_{k=0}^{n} a \cdot r^k = a + ar^1 + \ldots + ar^n$$

Let's multiply this by r:

$$r \cdot S = ar + ar^1 + \ldots + ar^{n+1}$$

$S_n - rS_n = a - a \cdot r^{n+1} = a(1 + r^{n+1})$

$S_n(1-r) = a(1-r^{n+1}) \rightarrow S_n = \frac{a(1-r^{n+1})}{1-r}$

An infinite sum can be seen as:

$$\lim_{n\to\infty} \sum_{k=0}^{n} a \cdot r^k = \lim_{n\to\infty} \frac{a \cdot (1-r^{n+1})}{1-r}$$

if $r < 1 \rightarrow \lim_{n\to\infty} \frac{a(1-r^{n+1})}{1-r} = \frac{a}{1-r}$ ($r^{n+1} \to 0$)

In our case $a = 1$ and $r = \gamma$:

then

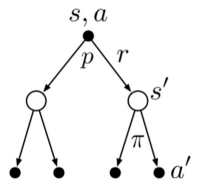
$$\sum_{k=0}^{\infty} \gamma^k = \frac{1}{1-\gamma}$$

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. □

$G_t = R_{t+1} + R_{t+2} + \dots + R_T = R_{t+1} + G_{t+1}$

It would change, because if the reward sometimes is slightly negative and we add a constant, we might lose the information of the penalty reward.

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). Hint: the backup diagram to the right corresponds to this equation. Show the sequence of equations analogous to (3.14), but for action values. □



q_π backup diagram

$$Q_\pi(s,a) = E_\pi\left[G_t \mid S_t = s, A_t = a\right] = E_\pi\left[\sum_{k=0}^{\infty} \gamma^k \cdot R_{k+t+1} \mid S_t = s, A_t = a\right] =$$

$$= E_\pi\left[R_{t+1} + \gamma \cdot R_{t+2} + \gamma^2 \cdot R_{t+3} + \dots \mid S_t = s, A_t = a\right] = E_\pi\left[R_{t+1} + \gamma\left(R_{t+2} + \gamma \cdot R_{t+3} + \dots\right) \mid S_t = s, A_t = a\right]$$

$$= E_\pi\left[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a\right] = \quad (G_{t+1})$$

$$\sum_{s',r} p(s',r \mid s,a) \cdot \left[r + \gamma E_\pi\left[G_{t+1} \mid S_t = s, A_t = a, S_{t+1} = s'\right]\right]$$

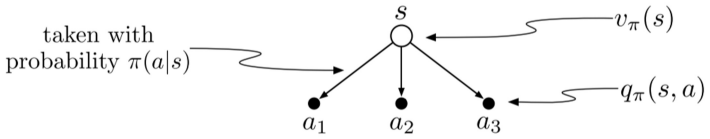
No additional info since a MDP is only determined by the current state, that for G_{t+1} is s'.

$$= \sum_{s',r} p(s',r \mid s,a) \cdot \left[r + \gamma \cdot \sum_{a'} \pi(a' \mid s') \cdot E_\pi\left[G_{t+1} \mid S_{t+1} = s', A_{t+1} = a'\right]\right] =$$

look exercise 3.12 conditioning on $A_t = a'$, I add $\sum \pi(a' \mid s')$

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

Exercise 3.18 The value of a state depends on the values of the actions possible in that state and on how likely each action is to be taken under the current policy. We can think of this in terms of a small backup diagram rooted at the state and considering each possible action:

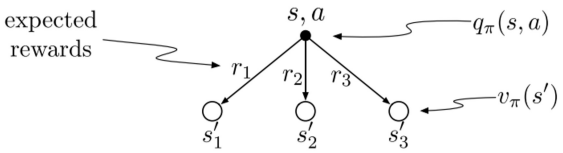


Give the equation corresponding to this intuition and diagram for the value at the root node, $v_\pi(s)$, in terms of the value at the expected leaf node, $q_\pi(s,a)$, given $S_t = s$. This equation should include an expectation conditioned on following the policy, π. Then give a second equation in which the expected value is written out explicitly in terms of $\pi(a|s)$ such that no expected value notation appears in the equation. □

$$v_\pi(s) = \mathbb{E}_\pi\left[Q_\pi(s,a) \mid S_t = s \right] = \sum_a \pi(a|s) \cdot Q_\pi(s,a)$$

value function

Exercise 3.19 The value of an action, $q_\pi(s,a)$, depends on the expected next reward and the expected sum of the remaining rewards. Again we can think of this in terms of a small backup diagram, this one rooted at an action (state–action pair) and branching to the possible next states:



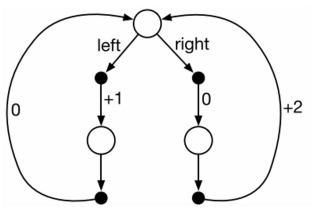
Give the equation corresponding to this intuition and diagram for the action value, $q_\pi(s,a)$, in terms of the expected next reward, R_{t+1}, and the expected next state value, $v_\pi(S_{t+1})$, given that $S_t = s$ and $A_t = a$. This equation should include an expectation but *not* one conditioned on following the policy. Then give a second equation, writing out the expected value explicitly in terms of $p(s', r|s, a)$ defined by (3.2), such that no expected value notation appears in the equation. □

$$Q_\pi(s,a) = \mathbb{E}_\pi\left[R_{t+1} + \gamma v_\pi(s') \mid S_t = s,\ A_t = a \right]$$

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

state-action function

Exercise 3.22 Consider the continuing MDP shown to the right. The only decision to be made is that in the top state, where two actions are available, left and right. The numbers show the rewards that are received deterministically after each action. There are exactly two deterministic policies, π_{left} and π_{right}. What policy is optimal if $\gamma = 0$? If $\gamma = 0.9$? If $\gamma = 0.5$? □



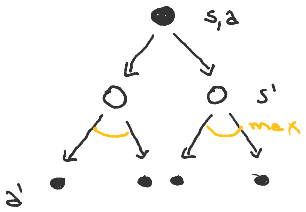
$$G_t = \sum_{k=0}^{\infty} \gamma^k \cdot R_{t+1+k}$$

TO DO . . .

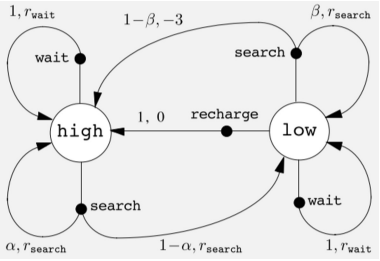
Exercise 3.23 Give the Bellman equation for q_* for the recycling robot. □

$$q_*(s,a) = \mathbb{E}\left[R_{t+1} + \max_{a'} q_*(s',a') \mid S_t = s, A_t = a \right]$$

$$\sum_{s',r} p(s',r \mid s,a) \cdot \left[r + \max_{a'} q(s',a') \right]$$



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



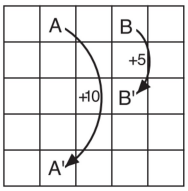
$s = \text{high}, \quad a = \text{wait}$

$$q_*(s,a) = r_{\text{wait}} + \gamma \max_{a'} q_*(\text{high}, a') = r_{\text{wait}} + \gamma \cdot \max_{a'} \{ q_*(\text{high}, \text{wait}), q_*(\text{high}, \text{search}) \}$$

$a' = \{\text{wait}, \text{search}\}$

and so on for all the states...

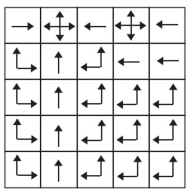
Exercise 3.24 Figure 3.5 gives the optimal value of the best state of the gridworld as 24.4, to one decimal place. Use your knowledge of the optimal policy and (3.8) to express this value symbolically, and then to compute it to three decimal places. □



Gridworld

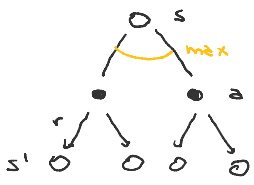
22.0	24.4	22.0	19.4	17.5
19.8	22.0	19.8	17.8	16.0
17.8	19.8	17.8	16.0	14.4
16.0	17.8	16.0	14.4	13.0
14.4	16.0	14.4	13.0	11.7

v_*



π_*

optimal value function



$$v_*(s) = \max_a q_*(s,a) =$$

$$= \max_a \mathbb{E}_{\pi_*}\left[G_t \mid S_t = s, A_t = a \right] =$$

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

$$= \max_a \mathbb{E}_{\pi_*}\left[R_{t+1} + \gamma \cdot v_*(S') \mid S_t = s, A_t = a \right] =$$

$$= \max_a \sum_{s',r} p(s',r \mid s,a) \cdot \left[r + \gamma \cdot v_*(s') \right]$$

$$= 1 \cdot \left[10 + \gamma \cdot 16 \right]$$

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. □

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

Exercise 3.27 Give an equation for π_* in terms of q_*. □

Exercise 3.28 Give an equation for π_* in terms of v_* and the four-argument p. □

Exercise 3.29 Rewrite the four Bellman equations for the four value functions (v_π, v_*, q_π, and q_*) in terms of the three argument function p (3.4) and the two-argument function r (3.5). □

Es. 3.27

$$\pi_* = \arg\max_a q_*(s,a)$$

Es. 3.28

$$\pi_* = \arg\max_a \sum_{s',r} p(s',r \mid s,a) \cdot [r + \gamma \cdot v_*(s')]$$

Not max. Imagine there are more actions that lead to an optimal $v_*(s)$, we want to calculate the average for all of them

Es. 3.29

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

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

$$q_\pi(s,a) = \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a] = \left[r(s,a) + \sum_{s',r} p(s',r \mid s,a) \cdot \sum_{a'} \gamma \cdot q_\pi(a',s') \cdot \pi(a' \mid s') \right]$$

$$q_{\pi_*}(s,a) = \mathbb{E}_{\pi_*}[G_t \mid S_t = s, A_t = a] = \left[r(s,a) + \gamma \sum_{s',r} p(s',r \mid s,a) \cdot \sum_a q_*(s',a') \cdot \pi_*(a' \mid s') \right]$$

Chapter 4 - Dynamic Programming

Exercise 4.1 In Example 4.1, if π is the equiprobable random policy, what is $q_\pi(11, \texttt{down})$? What is $q_\pi(7, \texttt{down})$? □

Example 4.1 Consider the 4×4 gridworld shown below.