Chapter 3

- G_t is the cumulative reward following t

(3.1) $G_t = R_{t+1} + R_{t+2} + R_{t+3} + \dots + R_T$

↳ where T is the final step time

(3.2) $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots$

$= \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}, \quad 0 \leq \gamma \leq 1$

γ is the discount rate

- $V_\pi(s)$ is the state-value function for policy π. This is the expected return when starting in state s and following policy π thereafter.

- $q_\pi(s,a)$ is the action-value function for policy π. This is the expected return starting from state s, taking action a, and following policy π thereafter.

- π is the policy, decision making rule

- $\pi(s)$ returns the action taken in state s under determistic policy π.

- $\pi(a|s)$ is the probability of taking action a when in state s under stochastic policy π.

- $p(s', r | s, a)$ is the probability of transitioning to state s' with reward r from state s with action a.

(3.6) $$p(s', r | s, a) = \Pr\{S_{t+1} = s', R_{t+1} = r \mid S_t = s, A_t = a\}$$

- The state is whatever information is available to the agent. A state signal has the Markov property if it summarizes all past sensations compactly while retaining all relevant information. The environment's response at $t+1$ depends only on the state and action representations at t, independent of the path, or history, of signals that led up to it. Decisions and values are a function only of the current state.

A reinforcement learning task that satisfies the Markov property is called a Markov decision process. If the state and action spaces are finite, then it is called a finite Markov decision process (finite MDP).

(3.10) - $$v_\pi(s) = E_\pi\left[G_t \mid S_t = s \right]$$

$$= E_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \;\middle|\; S_t = s \right]$$

(3.11) - $$q_\pi(s,a) = E_\pi\left[G_t \mid S_t = s, A_t = a \right]$$

$$= E_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \;\middle|\; S_t = s, A_T = a \right]$$

- $E_\pi[\cdot]$ is the expected value of a random variable given that the agent follows policy π, and t is any time step.

- $$v_\pi(s') = E_\pi\left[G_{t+1} \mid S_{t+1} = s' \right]$$

$$= E_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{(t+1)+k+1} \;\middle|\; S_{t+1} = s' \right]$$

$$= E_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+2} \;\middle|\; S_{t+1} = s' \right]$$

$$(3.12) \quad v_\pi(s) = E_\pi\left[G_t \mid S_t = s \right]$$

$$= E_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s \right]$$

$$= E_\pi\left[R_{t+1} + \sum_{k=1}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s \right]$$

$$= E_\pi\left[R_{t+1} + \sum_{k=0}^{\infty} \gamma^{k+1} R_{t+k+2} \,\middle|\, S_t = s \right]$$

$$= E_\pi\left[R_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k R_{t+k+2} \,\middle|\, S_t = s \right]$$

$$= E_\pi\left[R_{t+1} + \gamma v_\pi(S_{t+1}) \,\middle|\, S_t = s \right]$$

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

Bellman equation for v_π

Chapter 3

Gridworld

actions = { North, South, East, West }

$A = (0,1)$
$A' = (4,1)$
$B = (0,3)$
$B' = (2,3)$

	0	1	2	3	4
0		A		B	
1					
2				B'	
3					
4		A'			

$R = 0$ normal case, move N, S, E, or W

$R = -1$ if move would take agent off the grid, position unchanged

$R = +10$ when taking any action from state A, move to A'

$R = +5$ when taking any action from state B, move to B'

$\gamma = 0.9$

- Optimal value functions

$\pi \geq \pi'$ iff $v_\pi(s) \geq v_{\pi'}(s) \quad \forall s \in S$

π_* is an optimal policy

(3.13) $v_*(s) = \max_\pi v_\pi(s) \quad \forall s \in S$

(3.14) $q_*(s,a) = \max_\pi q_\pi(s,a) \quad \forall s \in S, a \in A$

- q_* in terms of v_*

(3.15) $q_*(s,a) = E\left[R_{t+1} + \gamma v_*(S_{t+1}) \mid S_t = s, A_t = a\right]$

- $v_*(s) = \max_a q_*(s,a)$, where $a \in A(s)$

(3.16) $= \max_a E\left[R_{t+1} + \gamma v_*(S_{t+1}) \mid S_t = s, A_t = a\right]$

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

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

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

Bellman optimality equations

Exercise 3-8

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

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

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

$$= E_\pi\left[R_{t+1} + \sum_{k=0}^{\infty} \gamma^{k+1} R_{t+k+2} \mid S_t = s, A_t = a\right]$$

$$= E_\pi\left[R_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k R_{t+k+2} \mid S_t = s, A_t = a\right]$$

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

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

Bellman equation for q_π

- Exercise 3.9

$$V_\pi(s_{2,2}) = 0.25 \cdot \left[0 + 0.9 \cdot V_\pi(s_{1,2})\right]$$
$$+ 0.25 \left[0 + 0.9 \cdot V_\pi(s_{3,2})\right]$$
$$+ 0.25 \left[0 + 0.9 \cdot V_\pi(s_{1,2})\right]$$
$$+ 0.25 \left[0 + 0.9 \cdot V_\pi(s_{3,2})\right]$$

$$= 0.25 \cdot 0.9 (2.3)$$
$$+ 0.25 \cdot 0.9 (-0.4)$$
$$+ 0.25 \cdot 0.9 (0.7)$$
$$+ 0.25 \cdot 0.9 (0.4)$$

$$= 0.25 \cdot 0.9 \cdot (2.3 - 0.4 + 0.7 + 0.4)$$

$$= 0.25 \cdot 0.9 \cdot 3$$

$$= 0.675 \approx 0.7$$

Exercise 3.10

$$G_t = (R_{t+1} + c) + \gamma(R_{t+2} + c) + \gamma^2(R_{t+3} + c) + \ldots$$

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

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

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

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

where

$$V_c = \sum_{k=0}^{\infty} \gamma^k c$$

$$= c \cdot \sum_{k=0}^{\infty} \gamma^k$$

$$= c \cdot \frac{1}{1-\gamma}$$

$$= \frac{c}{1-\gamma}$$

Geometric series

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}$$

- Exercise 3.12

$$v_\pi(s) = E_\pi\left[q_\pi(s, A_t) \mid S_t = s \right]$$

$$= \pi(a_1|s)\, q_\pi(s,a_1) + \pi(a_2|s)\, q_\pi(s,a_2) + \pi(a_3|s)\, q_\pi(s,a_3)$$

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

- Exercise 3.13

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

$$= p(s_1', r_1 | s,a)\left[r_1 + \gamma v_\pi(s_1')\right] + p(s_2', r_2 | s,a)\left[r_2 + \gamma v_\pi(s_2')\right] + p(s_3', r_3 | s,a)\left[r_3 + \gamma v_\pi(s_3')\right]$$

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

- $q(s,a)$ is just the reward plus the value of the next state $v(s')$

Chapter 3

Exercise 3.17

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

$$V_*(S_{0,1}) = R_A + \gamma \cdot V_*(S_{4,1})$$

$$V_*(S_{4,1}) = 0 + \gamma \cdot V_*(S_{3,1})$$

$$V_*(S_{3,1}) = 0 + \gamma \cdot V_*(S_{2,1})$$

$$V_*(S_{2,1}) = 0 + \gamma \cdot V_*(S_{1,1})$$

$$V_*(S_{1,1}) = 0 + \gamma \cdot V_*(S_{0,1})$$

$$V_*(S_{1,1}) = \gamma V_*(S_{0,1})$$

$$V_*(S_{2,1}) = \gamma^2 V_*(S_{0,1})$$

$$V_*(S_{3,1}) = \gamma^3 V_*(S_{0,1})$$

$$V_*(S_{4,1}) = \gamma^4 V_*(S_{0,1})$$

$$V_*(S_{0,1}) = R_A + \gamma^5 V_*(S_{0,1})$$

$$V_*(S_{0,1}) = \frac{R_A}{1-\gamma^5} = \frac{10}{1-(0.9)^5} = 24.419$$

Exercise 3.18

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

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

Exercise 3.19

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

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

Exercise 3.20

$$\pi_*(s) = \operatorname{argmax}_a q_*(s,a)$$

Exercise 3.21

$$\pi_*(s) = \operatorname{argmax}_a \sum_{s',r} p(s',r|s,a)\left[r + \gamma v_*(s')\right]$$