Contents

Code for exercises can be found at github.com/brynhayder/reinforcement_learning_an_introduction

Note that equation numbers in questions will refer to the original text.

1 Introduction

Exercise 1.1: Self-Play

Q

Suppose, instead of playing against a random opponent, the reinforcement learning algorithm described above played against itself, with both sides learning. What do you think would happen in this case? Would it learn a different policy for selecting moves?

A

- Would learn a different policy than playing a fixed opponent since the opponent would also be changing in this case.
- May not be able to learn an optimal strategy as the opponent keeps changing also.
- Could get stuck in loops.
- Policy could remain static since on average they would draw each iteration.

Exercise 1.2: Symmetries

Q

Many tic-tac-toe positions appear different but are really the same because of symmetries. How might we amend the learning process described above to take advantage of this? In what ways would this change improve the learning process? Now think again. Suppose the opponent did not take advantage of symmetries. In that case, should we? Is it true, then, that symmetrically equivalent positions should necessarily have the same value?

A

- We could label the states as unique up to symmetries so that our search space is smaller, this way we will get a better estimate of optimal play.
- If we are playing an opponent who does not take symmetries into account when they are playing then we should not label the states as the same since the opponent is part of the environment and the environment is not the same in those states.

Exercise 1.3: Greedy Play

Q

Suppose the reinforcement learning player was greedy, that is, it always played the move that brought it to the position that it rated the best. Might it learn to play better, or worse, than a nongreedy player? What problems might occur

A

- The greedy player will not explore, so will in general perform worse than the non-greedy player
- If the greedy player had a perfect estimate of the value of states then this would be fine.

Exercise 1.4: Learning from Exploration

Q

Suppose learning updates occurred after all moves, including exploratory moves. If the step-size parameter is appropriately reduced over time (but not the tendency to explore), then the state values would converge to a set of probabilities. What are the two sets of probabilities computed when we do, and when we do not, learn from exploratory moves? Assuming that we do continue to make exploratory moves, which set of probabilities might be better to learn? Which would result in more wins?

A

I think that an estimate for the probability of the state producing a win should be based on the optimal moves from that state.

- The one in which we only record the optimal moves is the probability of our optimal agent winning. If we include exploration then this is the probability of the training agent winning.
- Better to learn the probability of winning with no exploration since this is how the agent will perform in real time play.
- Updating from optimal moves only will increase probability of winning.

Exercise 1.5: Other Improvements

Q

Can you think of other ways to improve the reinforcement learning player? Can you think of any better way to solve the tic-tac-toe problem as posed?

A

I'm not too sure here...

- We could rank the draws as better than the losses.
- We might like to try running multiple iterations of games before updating our weights as this might give a better estimate.

2 Multi-armed Bandits

Exercise 2.1

Q

In ε-greedy action selection, for the case of two actions and $\varepsilon = 0.5$, what is the probability that the greedy action is selected?

A

0.5.

Exercise 2.2: Bandit example

Q

Consider a k-armed bandit problem with $k = 4$ actions, denoted 1, 2, 3, and 4. Consider applying to this problem a bandit algorithm using ε-greedy action selection, sample-average action-value estimates, and initial estimates of $Q_1(a) = 0$, for all a. Suppose the initial sequence of actions and rewards is $A_1 = 1$, $R_1 = 1$, $A_2 = 2$, $R_2 = 1$, $A_3 = 2$, $R_3 = 2$, $A_4 = 2$, $R_4 = 2$, $A_5 = 3$, $R_5 = 0$. On some of these time steps the ε case may have occurred, causing an action to be selected at random. On which time steps did this definitely occur? On which time steps could this possibly have occurred?

A

A_2 and A_5 were definitely exploratory. Any of the others *could* have been exploratory.

Exercise 2.3

Q

In the comparison shown in Figure 2.2, which method will perform best in the long run in terms of cumulative reward and probability of selecting the best action? How much better will it be? Express your answer quantitatively.

A

The $\varepsilon = 0.01$ will perform better because in both cases as $t \to \infty$ we have $Q_t \to q_*$. The total reward and probability of choosing the optimal action will therefore be 10 times larger in this case than for $\varepsilon = 0.1$.

Exercise 2.4

Q

If the step-size parameters, α_n, are not constant, then the estimate Q_n is a weighted average of previously received rewards with a weighting different from that given by (2.6). What is the weighting on each prior reward for the general case, analogous to (2.6), in terms of the sequence of step-size parameters?

A

Let $\alpha_0 = 1$, then

$$Q_{n+1} = \left(\prod_{i=1}^{n}(1-\alpha_i)\right) Q_1 + \sum_{i=1}^{n} \alpha_i R_i \prod_{k=i+1}^{n} (1-\alpha_k). \tag{1}$$

Where $\prod_{i=x}^{y} f(i) \doteq 1$ if $x > y$.

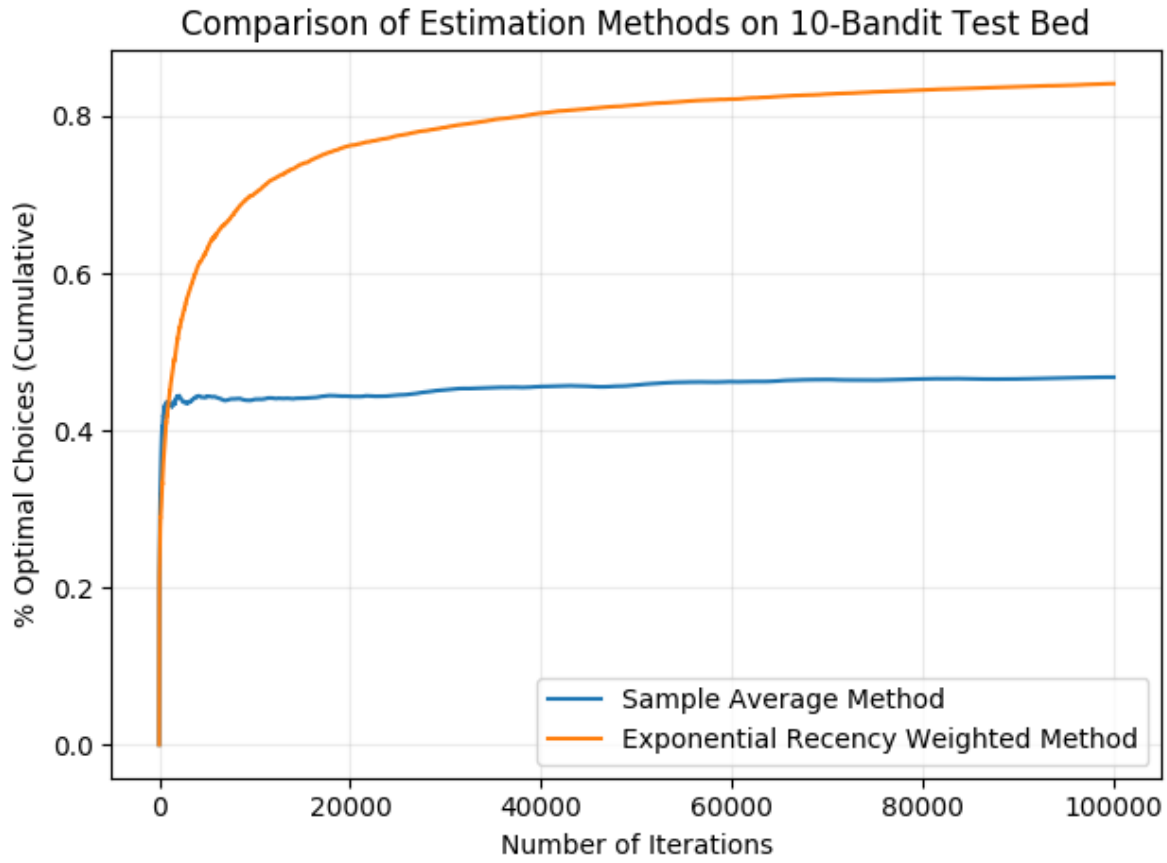
Exercise 2.5 (programming)

Q

Design and conduct an experiment to demonstrate the difficulties that sample-average methods have for non-stationary problems. Use a modified version of the 10-armed testbed in which all the $q_*(a)$ start out equal and then take independent random walks (say by adding a normally distributed increment with mean zero and standard deviation 0.01 to all the $q_*(a)$ on each step). Prepare plots like Figure 2.2 for an action-value method using sample averages, incrementally computed, and another action-value method using a constant step-size parameter, $\alpha = 0.1$. Use $\varepsilon = 0.1$ and longer runs, say of 10,000 steps.

A

This is a programming exercise. For the relevant code please see the repo.



Exercise 2.6: Mysterious Values

Q

The results shown in Figure 2.3 should be quite reliable because they are averages over 2000 individual, randomly chosen 10-armed bandit tasks. Why, then, are there oscillations and spikes in the early part of the curve for the optimistic method? In other words, what might make this method perform particularly better or worse, on average, on particular early steps?

A

At some point after step 10, the agent will find the optimal value. It will then choose this value greedily. The small step-size parameter (small relative to the initialisation value of 5) means that the estimate of the optimal value will converge slowly towards its true value.

It is likely that this true value is less than 5. This means that, due to the small step size, one of the sub-optimal actions will still have a value close to 5. Thus, at some point, the agent begins to act sub-optimally again.

Exercise 2.7: Unbiased Constant Step Trick

Q

In most of this chapter we have used sample averages to estimate action values because sample averages do not produce the initial bias that constant step sizes do (see the analysis in (2.6)). However, sample averages are not a completely satisfactory solution because they may perform poorly on non-stationary problems. Is it possible to avoid the bias of constant step sizes while retaining their advantages on non-stationary problems? One way is to use a step size of

$$\beta_t \doteq \alpha/\bar{o}_t, \tag{2}$$

where $\alpha > 0$ is a conventional constant step size and $\bar{o}_t$ is a trace of one that starts at 0:

$$\bar{o}_{t+1} = \bar{o}_t + \alpha(1 - \bar{o}_t) \tag{3}$$

for $t \geq 1$ and with $\bar{o}_1 \doteq \alpha$.

Carry out an analysis like that in (2.6) to show that β_t is an exponential recency-weighted average *without initial bias*.

A

Consider the answer to Exercise 2.4. There is no dependence of Q_k on Q_1 for $k > 1$ since $\beta_1 = 1$. Now it remains to show that the weights in the remaining sum decrease as we look further into the past. That is

$$w_i = \beta_i \prod_{k=i+1}^{n} (1 - \beta_k) \tag{4}$$

increases with i for fixed n. For this, observe that

$$\frac{w_{i+1}}{w_i} = \frac{\beta_{i+1}}{\beta_i(1 - \beta_{i+1})} = \frac{1}{1 - \alpha} > 1 \tag{5}$$

where we have assumed $\alpha < 1$. If $\alpha = 1$ then $\beta_t = 1 \; \forall \, t$.

Exercise 2.8: UCB Spikes

Q

In Figure 2.4 the UCB algorithm shows a distinct spike in performance on the 11th step. Why is this? Note that for your answer to be fully satisfactory it must explain both why the reward increases on the 11th step and why it decreases on the subsequent steps. Hint: if $c = 1$, then the spike is less prominent.

A

In the first 10 steps the agent cycles through all of the actions because when $N_t(a) = 0$ then a is considered maximal. On the 11th step the agent will most often then choose greedily. The agent will continue to choose greedily until $\ln(t)$ overtakes $N_t(a)$ for one of the other actions, in which case the agent begins to explore again hence reducing rewards.

Note that, in the long run, $N_t = O(t)$ and $\ln(t)/t \to 1$. So this agent is 'asymptotically greedy'.

Exercise 2.9

Q

Show that in the case of two actions, the soft-max distribution is the same as that given by the logistic, or sigmoid, function often used in statistics and artificial neural networks.

A

Let the two actions be denoted by 0 and 1. Now

$$\mathbb{P}(A_t = 1) = \frac{e^{H_t(1)}}{e^{H_t(1)} + e^{H_t(0)}} = \frac{1}{1 + e^{-x}}, \tag{6}$$

where $x = H_t(1) - H_t(0)$ is the relative preference of 1 over 0.

Exercise 2.10

Q

Suppose you face a 2-armed bandit task whose true action values change randomly from time step to time step. Specifically, suppose that, for any time step, the true values of actions 1 and 2 are respectively 0.1 and 0.2 with probability 0.5 (case A), and 0.9 and 0.8 with probability 0.5 (case B). If you are not able to tell which case you face at any step, what is the best expectation of success you can achieve and how should you behave to achieve it? Now suppose that on each step you are told whether you are facing case A or case B (although you still don't know the true action values). This is an associative search task. What is the best expectation of success you can achieve in this task, and how should you behave to achieve it?

A

I assume the rewards are stationary.

One should choose the action with the highest expected reward. In the first case, both action 1 and 2 have expected value of 0.5, so it doesn't matter which you pick.

In the second case one should run a normal bandit method separately on each colour. The expected reward from identifying the optimal actions in each case is 0.55.

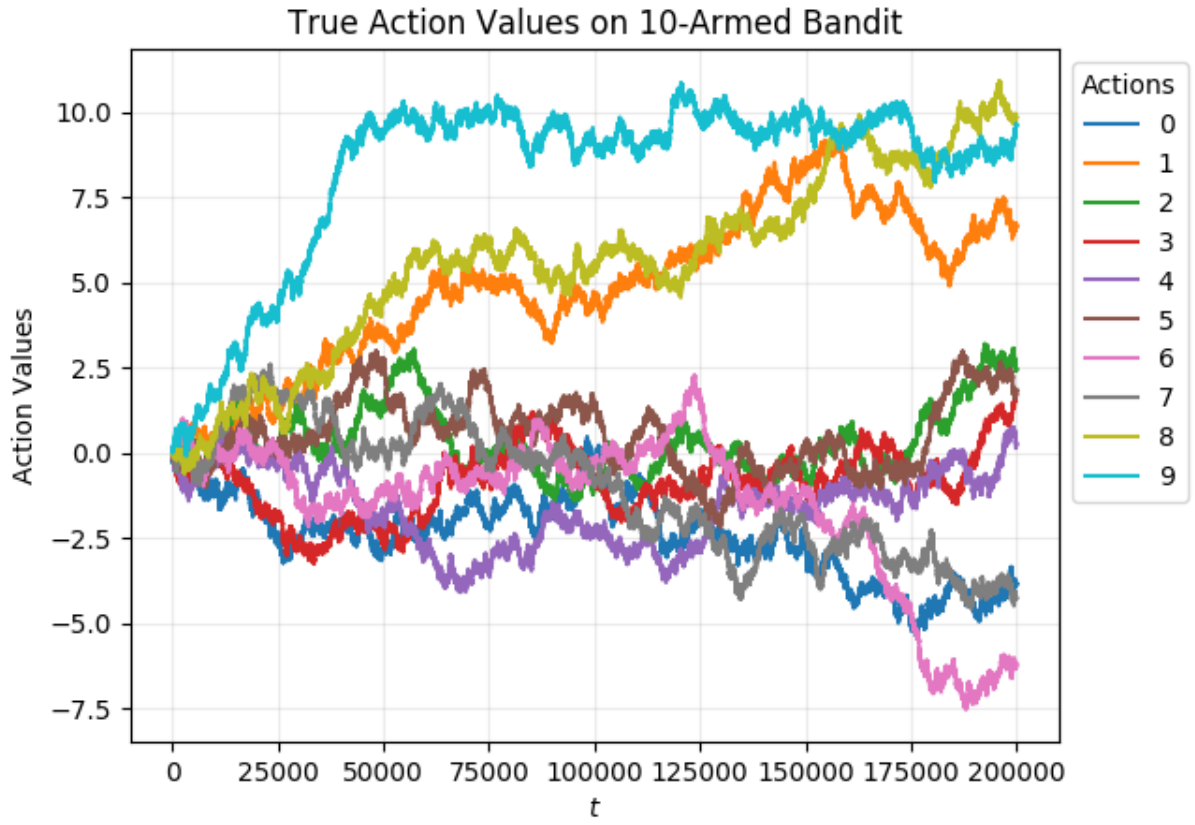
Exercise 2.11 (programming)

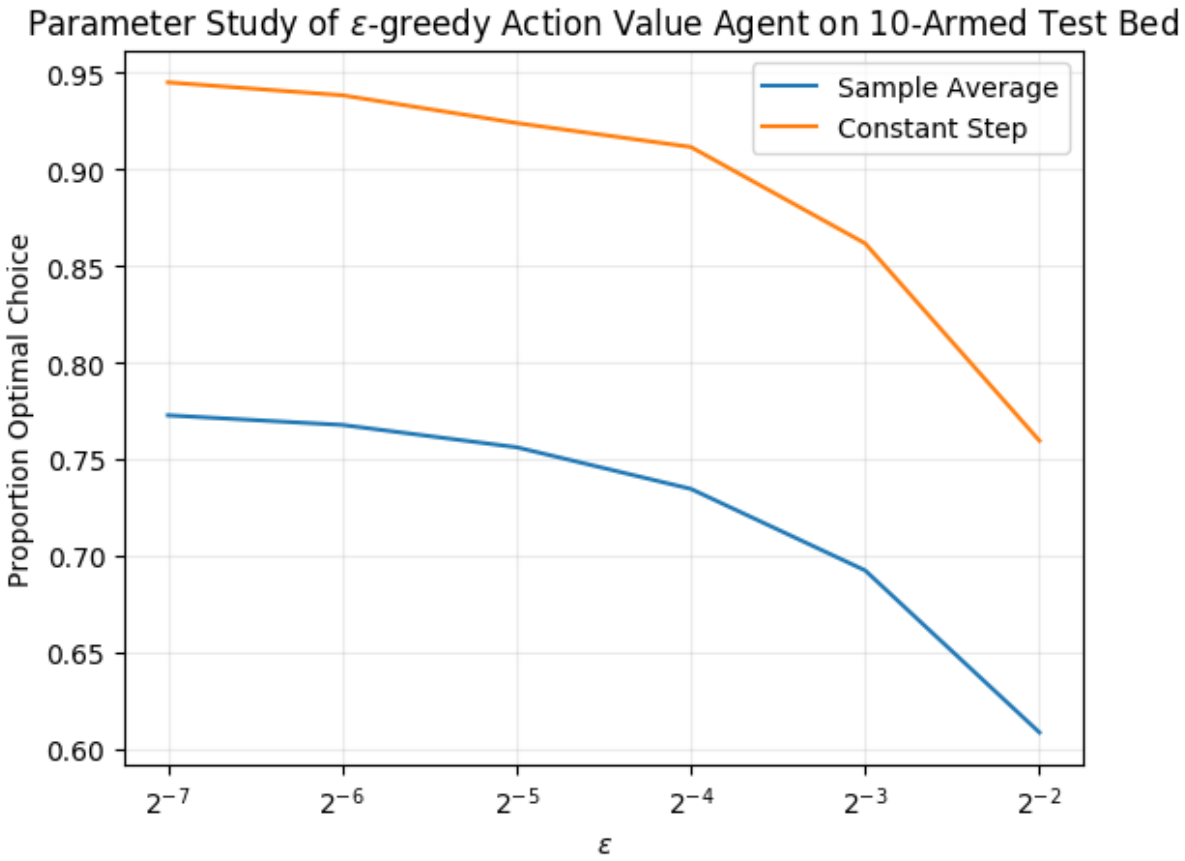
Q

Make a figure analogous to Figure 2.6 for the non-stationary case outlined in Exercise 2.5. Include the constant-step-size ε-greedy algorithm with $\alpha = 0.1$. Use runs of 200,000 steps and, as a performance measure for each algorithm and parameter setting, use the average reward over the last 100,000 steps.

A

This is a programming exercise. For the relevant code please see the repo.




Parameter Study of ε-greedy Action Value Agent on 10-Armed Test Bed
Sample Average
Constant Step
Proportion Optimal Choice
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
2^{-7}
2^{-6}
2^{-5}
2^{-4}
2^{-3}
2^{-2}
ε

3 Finite Markov Decision Processes

Exercise 3.1

Q

Devise three example tasks of your own that fit into the MDP framework, identifying for each its states, actions, and rewards. Make the three examples as *different* from each other as possible. The framework is abstract and flexible and can be applied in many different ways. Stretch its limits in some way in at least one of your examples.

A

1. Simple example is a robot that hoovers a room. The state can be how much dust there is on the ground and where the robot is (including it's orientation). Actions can be to move and hoover. The reward can be the amount by which it reduces dust in the room on that action. This is Markov because all that is important from the previous state to the future is where the dust is left (maybe also how much).

2. Outlandish example is a football coach. Actions are playing strategies. Rewards are goals. State is current score, team fitness, etc..

3. financial trader. State is their current holdings on an asset. Reward is money from a trade. Actions are buy/sell. Maybe not markov because the environment may change predictably based on information from multiple steps ago.

Exercise 3.2

Q

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

A

The main thing about the MDP is that Markov property. There are tasks where this does not hold. For instance, in Poker, the previous states will determine what is in the deck and what is not. This does not obey Markov property.

Exercise 3.3

Q

Consider the problem of driving. You could define the actions in terms of the accelerator, steering wheel, and brake, that is, where your body meets the machine. Or you could define them farther out—say, where the rubber meets the road, considering your actions to be tire torques. Or you could define them farther in—say, where your brain meets your body, the actions being muscle twitches to control your limbs. Or you could go to a really high level and say that your actions are your choices of *where* to drive. What is the right level, the right place to draw the line between agent and environment? On what basis is one location of the line to be preferred over another? Is there any fundamental reason for preferring one location over another, or is it a free choice?

A

- The natural distinction depends on the task. If the task is to go from one location to another, the actions might be in terms of directing the car and altering its speed.
- This also depends on what we consider to be the decision making part here. If we consider the decisions to be made by the brain of a human driving, then their physical body will form part of the environment – if they break their leg then it will effect their goal.
- Actions have to be things that the agent can actually control. Take the example of an autonomous vehicle. One might consider that the agent has complete control over the car's break, accelerator and steering wheel. These operations would form the actions for the agent.

Exercise 3.4

Q

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

A

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

Table 1: Transition table

Exercise 3.5

Q

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

A

$$\sum_{s' \in \mathcal{S}^+} \sum_{r \in \mathcal{R}} p(s', r|s, a) = 1 \tag{7}$$

Exercise 3.6

Q

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 of this task?

A

Note that the pole will fall eventually with probability 1

$$G_t = -\gamma^{T-t}. \tag{8}$$

Whereas in the continuing case the value is

$$G_t = -\sum_{k \in \mathcal{K}} \gamma^{k-t}, \tag{9}$$

where $\mathcal{K}$ is the set of times after t at which the pole falls over.

Exercise 3.7

Q

Imagine that you are designing a robot to run a maze. You decide to give it a reward of $+1$ for escaping from the maze and a reward of zero at all other times. The task seems to break down naturally into episodes–the successive runs through the maze–so you decide to treat it as an episodic task, where the goal is to maximize expected total reward (3.7). 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?

A

If the agent keeps going randomly, it will reach the end of the maze with probability 1, so the value of G under most strategies is 1. What you actually want is for the thing to leave the maze as quickly as possible.

Note also that there are some instances in which the agent might just get stuck in a loop. You would have to impose another rule to put the agent into a terminal state here.

Exercise 3.8

Q

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, $\ldots$, G_5? Hint: Work backwards.

A

$G_0 = 2$, $G_1 = 3$, $G_2 = 2$, $G_3 = \frac{1}{2}$, $G_4 = \frac{1}{8}$, $G_5 = 0$.

Exercise 3.9

Q

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?

A

$$G_1 = 7\frac{\gamma}{1-\gamma} \tag{10}$$

$$G_0 = 2 + 7\frac{\gamma}{1-\gamma} \tag{11}$$

Exercise 3.10

Q

Prove (3.10).

A

Take $r \in \mathbb{C}$ with $|r| < 1$, then

$$\begin{aligned} S_N &\doteq \sum_{i=0}^{N} r^i \\ rS_N - S_N &= r^{N+1} - 1 \\ S_N &= \frac{1 - r^{N+1}}{1 - r} \\ S &\doteq \lim_{N \to \infty} S_N = \frac{1}{1 - r} \end{aligned}$$

Exercise 3.11

Q

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)?

A

$$\mathbb{E}_\pi[R_{t+1}|S_t = s] = \sum_a \pi(a|s) \sum_{s',r} rp(s', r|s, a) \tag{12}$$

Exercise 3.12

Q

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 neighbouring states, valued at $+2.3$, $+0.4$, -0.4, and $+0.7$. (These numbers are accurate only to one decimal place.)

A

$$\begin{aligned} v_\pi(\texttt{center}) &\doteq \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)[r + \gamma v_\pi(s')] \\ &= \frac{1}{4} \times 0.9 \times \sum_{s'} v_\pi(s') \\ &= \frac{1}{4} \times 0.9 \times 3.0 \\ &= 0.675 \end{aligned}$$

Exercise 3.13

Q

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.

A

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

Exercise 3.14

Q

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?

A

We choose actions by relative (additive) values. Add c to all states and

$$G_t \mapsto G_t + \frac{c}{1-\gamma} \tag{13}$$

so relative values unchanged.

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

Exercise 3.15

Q

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.

A

Let terminal time be T. In this case we have

$$G_t \mapsto G_t + c\frac{1-\gamma^T}{1-\gamma}, \tag{14}$$

so if the agent can procrastinate termination, then all else being equal it will increase v_π.

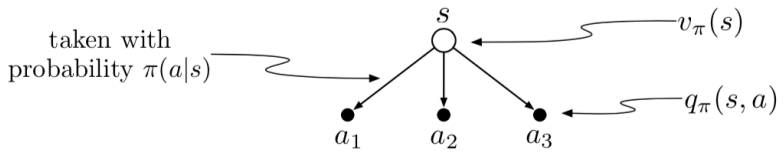
Suppose that we have an episodic task with one state S and two actions A_0, A_1. A_0 takes agent to terminal state with reward 1, while A_1 takes agent back to S with reward 0. In this case the agent should terminate to maximise reward.

If we add 1 to each reward then the return for doing A_1 forever is $\frac{1}{1-\gamma}$. Which can be bigger than 2 if we choose a discount factor smaller than $\frac{1}{2}$.

Exercise 3.16

Q

Exercise 3.16 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. □

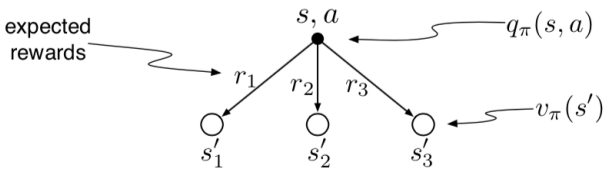
A

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

Exercise 3.17

Q

Exercise 3.17 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. □

A

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

Exercise 3.18

Q

Draw or describe the optimal state-value function for the golf example.

A

(Using the pictures.) Optimal state value gives values according to `driver` when off the green, then according to `putter` on the green.

Optimal policy is to use `driver` when off green and `putter` when on green.

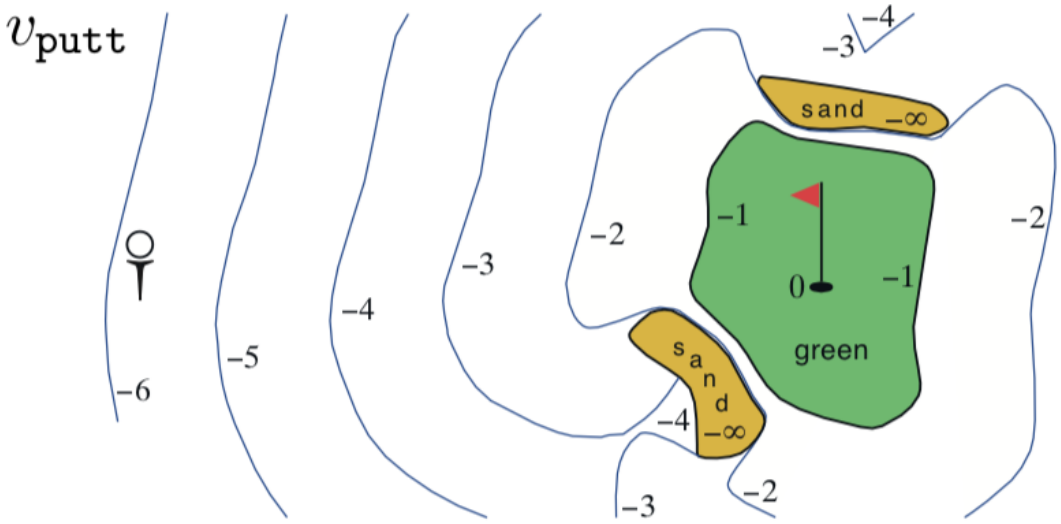
Exercise 3.19

Q

Draw or describe the contours of the optimal action-value function for putting, $q_*(s, \texttt{putter})$, for the golf example.

A

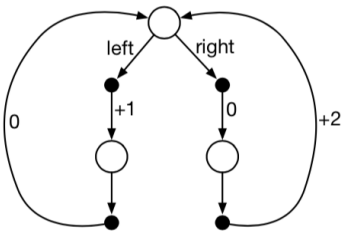
Should be the same as the value for the policy that always uses the `putter`.



Exercise 3.20

Q

Exercise 3.20 Consider the continuing MDP shown on 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 $\gamma = 0.5$? □



A

When $\gamma = 0$, $v_{\pi_{\texttt{left}}}$ is optimal. When $\gamma = 0.5$, they are both optimal. When $\gamma = 0.9$, $v_{\pi_{\texttt{right}}}$ is optimal.

Exercise 3.21

Q

Give the Bellman equation for q_* for the recycling robot.

A

This is just writing out the equation below, filling in the values given in the robot example.

$$q_*(s,a) = \sum_{s',r} p(s',r|s,a)[r + \gamma \max_{a'} q_*(s',a')] \tag{15}$$

Exercise 3.22

Q

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.

A

All actions tak the agent to A' with reward 10. We can see that $\gamma = 16.0/17.8 = 0.9$. This means that

$$v = 10 + 16 \times 0.9 = 24.4. \tag{16}$$

This is using the following framework

$$v_*(s) = \max_a \sum_{s',r} p(s',r|s,a)[r + \gamma v_*(s')]. \tag{17}$$

Exercise 3.23

Q

Give an equation for v_* in terms of q_*

A

$$v_*(s) = \sum_a \pi^*(a|s) q_*(s,a) \tag{18}$$

Exercise 3.24

Q

Give an equation for q_* in terms of v_* and the world's dynamics $p(s', r|s, a)$.

A

$$\begin{aligned} q_*(s,a) &= \mathbb{E}[R_{t+1} + \gamma v_*(S_{t+1}) | S_t = s, A_t = a] && (19) \\ &= \sum_{s',r} (r + \gamma v_*(s')) p(s', r|s, a) && (20) \end{aligned}$$

Exercise 3.25

Q

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

A

This is just any policy that acts greedily w.r.t. the optimal action-value function.

$$\pi_*(a|s) = \frac{\mathbb{1}\{a = \text{argmax}_{a'} \, q_*(a', s)\}}{\sum_a \mathbb{1}\{a = \text{argmax}_{a'} \, q_*(a', s)\}} \tag{21}$$

Exercise 3.26

Q

Give an equation for π_* in terms of v_* and the world's dynamics $p(s', r|s, a)$.

A

This is just the answer to 3.25 with the answer to 3.24 substituted in for q_*.

4 Dynamic Programming

Exercise 4.1

Q

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

A

$q_\pi(11, \texttt{down}) = -1$ since goes to terminal state. $q_\pi(7, \texttt{down}) = -15$.

Exercise 4.2

Q

In Example 4.1, suppose a new state 15 is added to the gridworld just below state 13, and its actions, `left`, `up`, `right`, and `down`, take the agent to states 12, 13, 14, and 15, respectively. Assume that the transitions from the original states are unchanged. What, then, is $v_\pi(15)$ for the equiprobable random policy? Now suppose the dynamics of state 13 are also changed, such that action down from state 13 takes the agent to the new state 15. What is $v_\pi(15)$ for the equiprobable random policy in this case?

A

$v_\pi(15) = -20$ if dynamics unchanged. If dynamics changed then apparently the state value is the same, but you would need to verify Bellman equations for all states for this.

Exercise 4.3

Q

What are the equations analogous to (4.3), (4.4), and (4.5) for the action-value function q_π and its successive approximations by a sequence of functions $q_0, q_1, q_2, \ldots$?

A

$$q_{k+1}(s,a) = \sum_{s',r} p(s', r|s, a) \left[r + \gamma \sum_{a'} \pi(a'|s) q_k(s', a') \right] \tag{22}$$

Exercise 4.4

Q

The policy iteration algorithm on the previous page has a subtle bug in that it may never terminate if the policy continually switches between two or more policies that are equally good. This is ok for pedagogy, but not for actual use. Modify the pseudocode so that convergence is guaranteed.

A

One problem is that the argmax_a has ties broken arbitrarily, this means that the same value function can give rise to different policies.

The way to solve this is to change the algorithm to take the whole set of maximal actions on each step and see if this set is stable and see if the policy is stable with respect to choosing actions from this set.

Exercise 4.5

Q

How would policy iteration be defined for action values? Give a complete algorithm for computing q_*, analogous to that on page 80 for computing q_*. Please pay special attention to this exercise, because the ideas involved will be used throughout the rest of the book.

A

We know that

$$v_\pi(s) = \sum_{a \in \mathcal{A}(s)} \pi(a|s) q_\pi(s, a) \tag{23}$$

so we know that

$$q_\pi(s, \pi'(s)) \geq \sum_{a \in \mathcal{A}(s)} \pi(a|s) q_\pi(s, a) \tag{24}$$

if π' is greedy with respect to π. So we know the algorithm still works for action values.

All there is now is to substitute the update for the action-value update and make the policy greedy with respect to the last iteration's action-values. Also need to make sure that the argmax_a is done consistently.

Exercise 4.6

Q

Suppose you are restricted to considering only policies that are ε-soft, meaning that the probability of selecting each action in each state, s, is at least $\varepsilon/|\mathcal{A}(s)|$. Describe qualitatively the changes that would be required in each of the steps 3, 2, and 1, in that order, of the policy iteration algorithm for v_π (page 80).

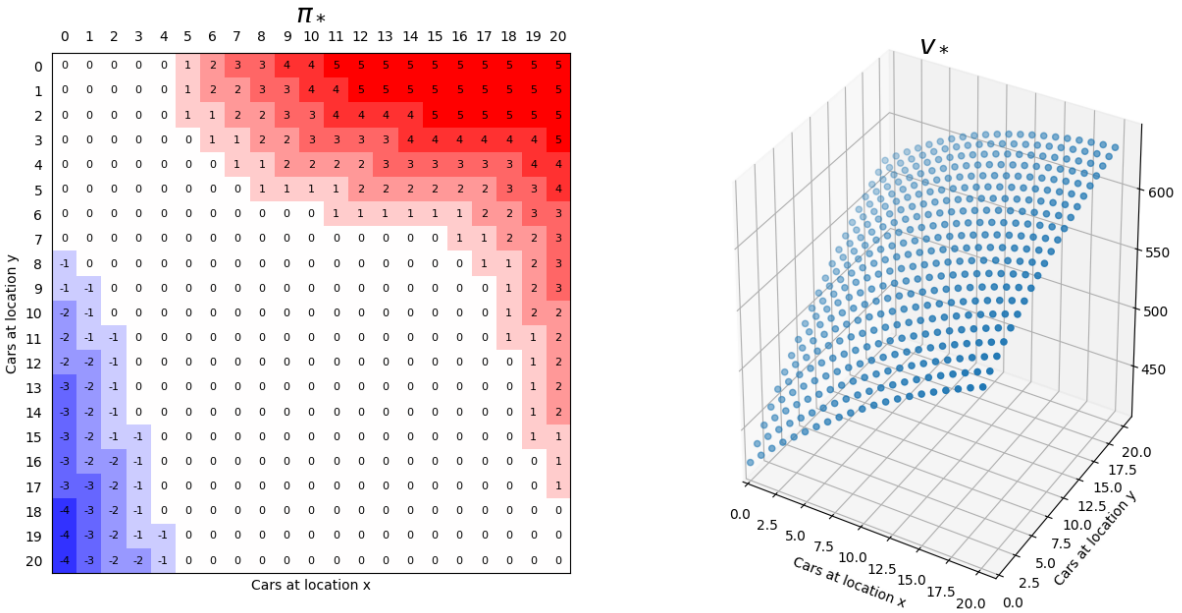
A

1. No change (but need policy to be able to be stochastic of course)
2. Need to re-write the Bellman update $v(s) \longleftarrow \sum_{a \in \mathcal{A}(s)} \pi(a|s) \sum_{s',r} p(s', r|s, a) \left[r + \gamma v(s')\right]$
3. Construct a greedy policy that puts weight on the greedy actions but is ε-soft. Be careful with the consistency of the argmax.

Exercise 4.7 (programming): Jack's Car Rental

Example 4.2: Jack's Car Rental Jack manages two locations for a nationwide car rental company. Each day, some number of customers arrive at each location to rent cars. If Jack has a car available, he rents it out and is credited \$10 by the national company. If he is out of cars at that location, then the business is lost. Cars become available for renting the day after they are returned. To help ensure that cars are available where they are needed, Jack can move them between the two locations overnight, at a cost of \$2 per car moved. We assume that the number of cars requested and returned at each location are Poisson random variables, meaning that the probability that the number is n is $\frac{\lambda^n}{n!}e^{-\lambda}$, where λ is the expected number. Suppose λ is 3 and 4 for rental requests at the first and second locations and 3 and 2 for returns. To simplify the problem slightly, we assume that there can be no more than 20 cars at each location (any additional cars are returned to the nationwide company, and thus disappear from the problem) and a maximum of five cars can be moved from one location to the other in one night. We take the discount rate to be $\gamma = 0.9$ and formulate this as a continuing finite MDP, where the time steps are days, the state is the number of cars at each location at the end of the day, and the actions are the net numbers of cars moved between the two locations overnight. Figure 4.2 shows the sequence of policies found by policy iteration starting from the policy that never moves any cars. ■

First we reproduce the original results.



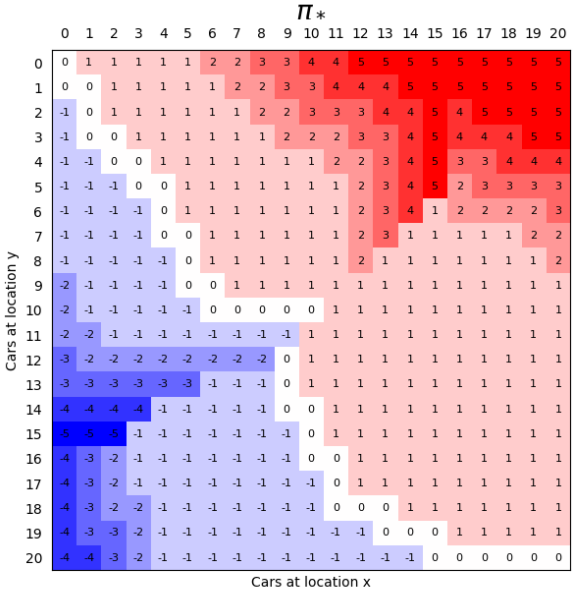
Q

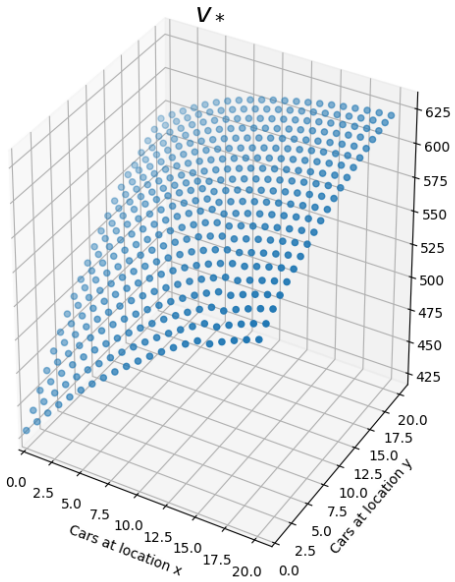
Write a program for policy iteration and re-solve Jack's car rental problem with the following changes. One of Jack's employees at the first location rides a bus home each night and lives near the second location. She is happy to shuttle one car to the second location for free. Each additional car still costs \$2, as do all cars moved in the other direction. In addition, Jack has limited parking space at each location. If more than 10 cars are kept overnight at a location (after any moving of cars), then an additional cost of \$4 must be incurred to use a second parking lot (independent of how many cars are kept there). These sorts of nonlinearities and arbitrary dynamics often occur in real problems and cannot easily be handled by optimisation methods other than dynamic programming. To check your program, first replicate the results given for the original problem. If your computer is too slow

for the full problem, cut all the numbers of cars in half.

A

This is a programming exercise. For the relevant code please see the repo.





Exercise 4.8

Q

Why does the optimal policy for the gambler's problem have such a curious form? In particular, for capital of 50 it bets it all on one flip, but for capital of 51 it does not. Why is this a good policy?

A

Since the coin is biased against us, we want to minimize the number of flips that we take. At 50 we can win with probability 0.4. At 51 if we bet small then we can get up to 52, but if we lose then we are still only back to 50 and we can again with with probability 0.4. (There is a whole paper on this problem called how to gamble if you must.)

Exercise 4.9 (programming): Gambler's Problem

Q

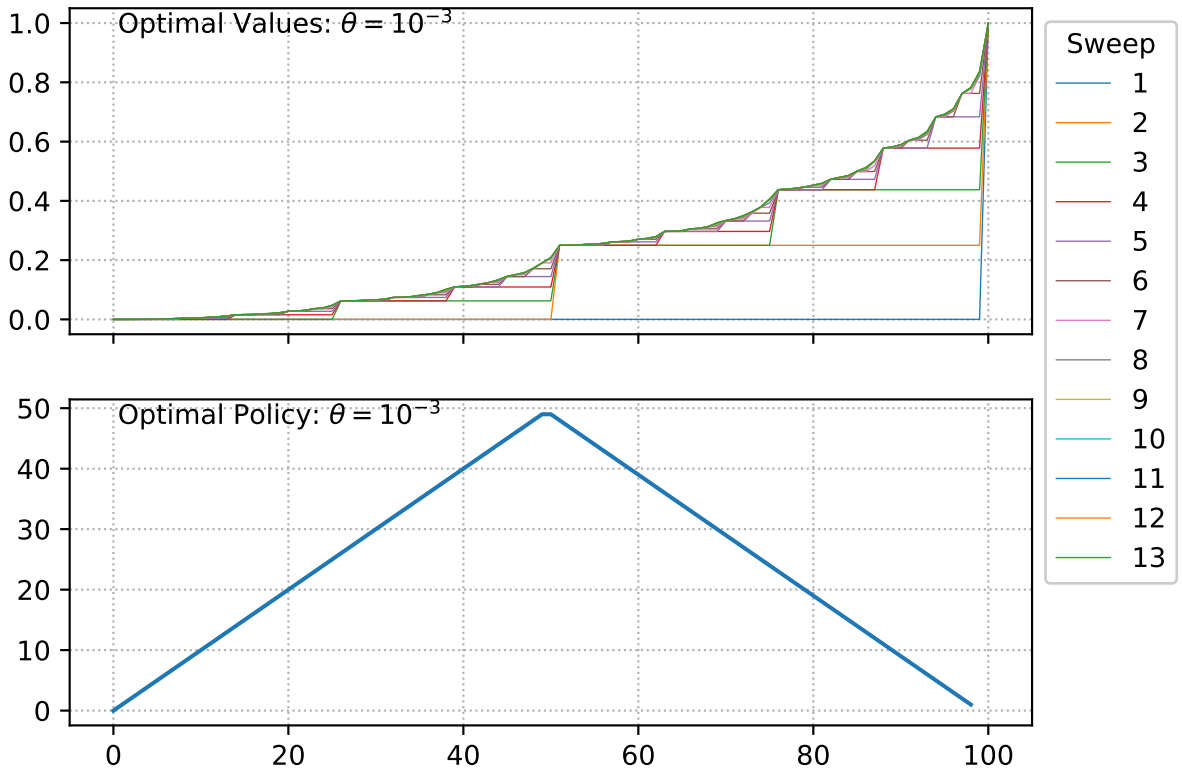
Implement value iteration for the gambler's problem and solve it for $p_h = 0.25$ and $p_h = 0.55$. In programming, you may find it convenient to introduce two dummy states corresponding to termination with capital of 0 and 100, giving them values of 0 and 1 respectively. Show your results graphically, as in Figure 4.3. Are your results stable as $\theta \to 0$?

A

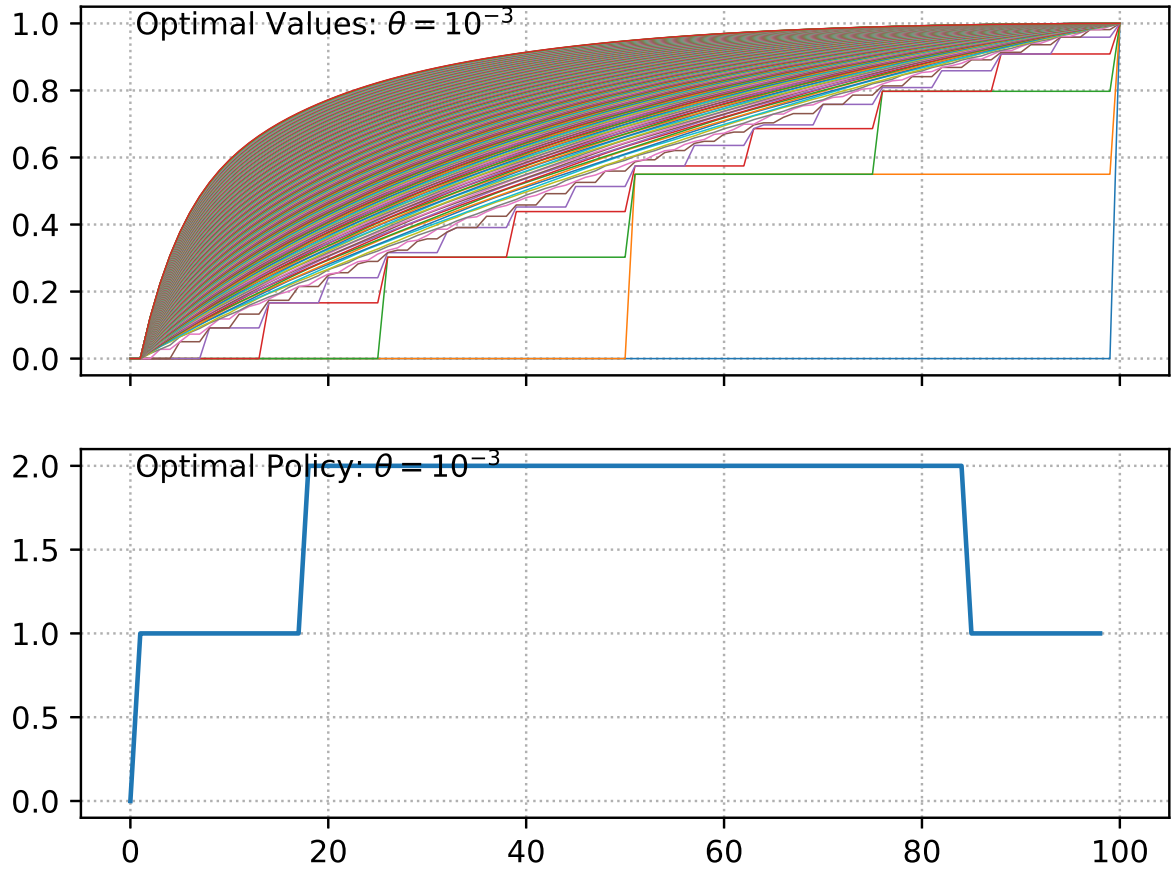
This is a programming exercise. For the relevant code please see the repo.

The process was stable as $\theta \to 0$ for $\mathbb{P}(\texttt{win}) < 0.5$.

$\mathbb{P}(\texttt{win}) = 0.25$



$$\mathbb{P}(\texttt{win}) = 0.55$$



Exercise 4.10

Q

What is the analog of the value iteration update (4.10) for action values, $q_{k+1}(s, a)$?

A

$$q_{k+1} = \max_{a'} \sum_{s',r} p(s', r|s, a) \left[r + \gamma q_k(s', a') \right] \tag{25}$$

5 Monte-Carlo Methods

Exercise 5.1

Q

Consider the diagrams on the right in Figure 5.1. Why does the estimated value function jump up for the last two rows in the rear? Why does it drop off for the whole last row on the left? Why are the frontmost values higher in the upper diagrams than in the lower?

A

- Policy is to hit unless $S \geq 20$. So you run a rik of going bust if you have 12-19, but you most likely win when you stick on 20 or 21
- Drops off because dealer has a usable ace
- Frontmost higher because you're less likely to go bust, but you still might get to 20 or 21 (π always hits here).

Exercise 5.2

Q

Suppose every-visit MC was used instead of first-visit MC on the blackjack task. Would you expect the results to be very different? Why or why not?

A

Results would be the same because this game is memoryless (cards are drawn with replacement).

Exercise 5.3

Q

What is the backup diagram for Monte Carlo estimation of q_π?

A

The same as the one shown in the book for state valus, only we have state-action pairs instead of states.

Exercise 5.4

Q

What is the equation analogous to (5.6) for *action* values $Q(s, a)$ instead of state values $V(s)$, again given returns generated using b?

A

We condition on taking action a in state s.

$$q_\pi(s, a) = \mathbb{E}_\pi[\rho_{t+1:T-1} G_t | S_t = s, A_t = s]$$

with returns generated from b. We estimate this quantity by

$$Q(s,a) = \frac{\sum_{t \in \mathcal{T}(s,a)} \rho_{t+1:T-1} G_t}{\sum_{t \in \mathcal{T}(s,a)} \rho_{t+1:T-1}}$$

where $\mathcal{T}(s,a)$ now contains timestamps of visits to state-action pairs.

Exercise 5.5

Q

In learning curves such as those shown in Figure 5.3 error generally decreases with training, as indeed happened for the ordinary importance-sampling method. But for the weighted importance-sampling method error first increased and then decreased. Why do you think this happened?

A

When there are fewer episodes the importance sampling ratios will be zero with higher probability since the behaviour policy will stick on values smaller than 20 (since it is random). Zero happens to be close to $v_\pi(s)$.

This effect lessens as we get more diversity in the episode trajectories.

Then after this the error reduces because the variance in the estimator reduces.

Exercise 5.6

Q

The results with Example 5.5 and shown in Figure 5.4 used a first-visit MC method. Suppose that instead an every-visit MC method was used on the same problem. Would the variance of the estimator still be infinite? Why or why not?

A

Yes, all terms in the sum are ≥ 0 and there woud just be more of them.

Exercise 5.7

Q

Modify the algorithm for first-visit MC policy evaluation (Section 5.1) to use the incremental implementation for sample averages described in Section 2.4.

A

Algo is the same apart from

- Initialise $V(s) = 0 \quad \forall s \in S$
- Don't need *Returns(s)* lists.
- Remove the last two lines and put in

$$V(S_t) \leftarrow V(S_t) + \frac{1}{T-t}[G_t - V(S_t)]$$

Exercise 5.8

Q

Derive the weighted-average update rule (5.8) from (5.7). Follow the pattern of the derivation of the unweighted rule (2.3).

A

Have $C_0 = 0$, $C_n = \sum_{k=1}^{n} W_k$ and

$$V_{n+1} = \frac{\sum_{k=1}^{n} W_k G_k}{C_n}.$$

Therefore,

$$\begin{aligned} C_n V_{n+1} &= \sum_{k+1}^{n-1} W_k G_k + W_k G_k && (26) \\ &= C_{n-1} V_n + W_n G_n && (27) \\ &= (C_n - W_n) V_n + W_n G_n. && (28) \end{aligned}$$

Finally

$$V_{n+1} = V_n + \frac{W_n}{C_n}[G_n - V_n].$$

Exercise 5.9

Q

In the boxed algorithm for off-policy MC control, you may have been expecting the W update to have involved the importance-sampling ratio $\pi(A_t|S_t)$, but instead it involves $1/b(A_t|S_t)$. Why is this nevertheless correct?

A

π is greedy, so

$$\pi(a|s) = \mathbb{1}\{a = \operatorname*{argmax}_{a'} Q(s, a')\}.$$

Exercise 5.10 (programming): Racetrack

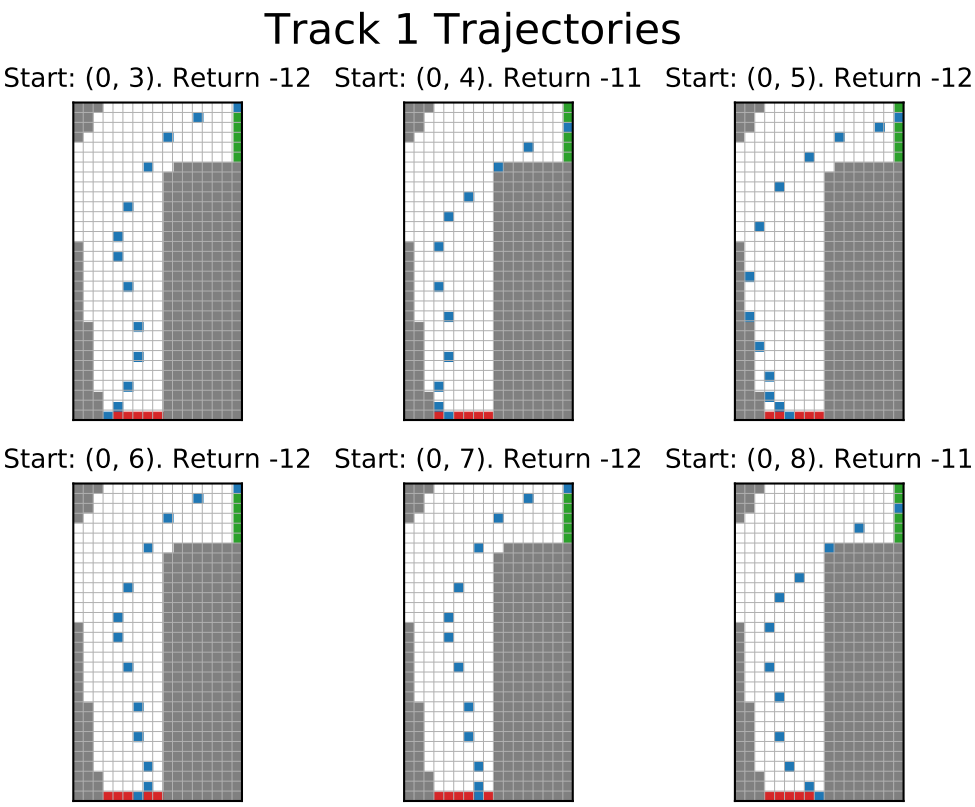
Q

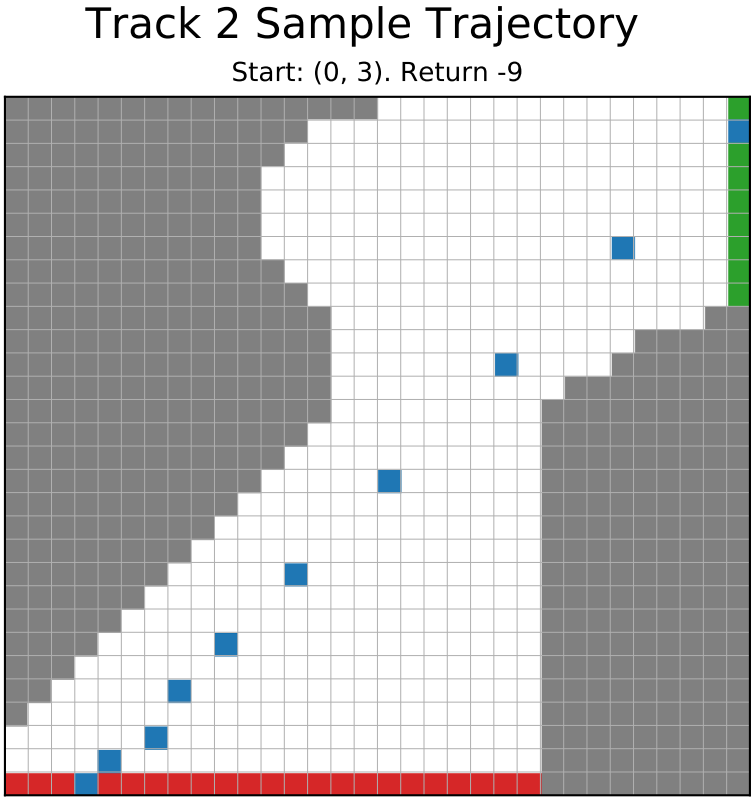
Consider driving a race car around a turn like those shown in Figure 5.5. You want to go as fast as possible, but not so fast as to run off the track. In our simplified racetrack, the car is at one of a discrete set of grid positions, the cells in the diagram. The velocity is also discrete, a number of grid cells moved horizontally and vertically per time step. The actions are increments to the velocity components. Each may be changed by +1, -1, or 0 in each step, for a total of nine (3×3) actions. Both velocity components are restricted to be nonnegative and less than 5, and they cannot both be zero except at the starting line. Each episode begins in one of the randomly selected start states with both velocity components zero and ends when the car crosses the finish line. The rewards are -1 for each step until the car crosses the finish line. If the car hits the track boundary, it is moved back to a random position on the starting line, both velocity components are reduced to zero, and the episode continues. Before updating the car's location at each time step, check to see if the projected path of the car intersects the track boundary. If it intersects the finish line, the episode ends; if it intersects anywhere else, the car is considered to have hit the track boundary and is sent

back to the starting line. To make the task more challenging, with probability 0.1 at each time step the velocity increments are both zero, independently of the intended increments. Apply a Monte Carlo control method to this task to compute the optimal policy from each starting state. Exhibit several trajectories following the optimal policy (but turn the noise off for these trajectories).

A

This is a programming exercise. For the relevant code please see the repo.





*Exercise 5.11

Q

Modify the algorithm for off-policy Monte Carlo control (page 110) to use the idea of the truncated weighted-average estimator (5.10). Note that you will first need to convert this equation to action values.

A

...

6 Temporal-Difference Learning

Exercise 6.1

Q

If V changes during the episode, then (6.6) only holds approximately; what would the difference be between the two sides? Let V_t denote the array of state values used at time t in the TD error (6.5) and in the TD update (6.2). Redo the derivation above to determine the additional amount that must be added to the sum of TD errors in order to equal the Monte Carlo error.

A

Write

$$\delta_t \doteq R_{t+1} + \gamma V_t(S_{t+1}) - V_t(S_t),$$

then

$$\begin{aligned}
G_t - V_t(S_t) &= R_{t+1} + \gamma G_{t+1} - V_t(S_t) && (29)\\
&= \delta_t + \gamma[G_{t+1} - V_{t+1}(S_{t+1})] + \gamma[V_{t+1}(S_{t+1}) - V_t(S_{t+1})] && (30)\\
&= \delta_t + \gamma[G_{t+1} - V_{t+1}(S_{t+1})] + \alpha\gamma[R_{t+2} + \gamma V_t(S_{t+2}) - V_t(S_{t+1})] && (31)\\
&\vdots \\
&= \sum_{k=t}^{T-1} \gamma^{k-t}\delta_k + \alpha \sum_{k=t}^{T-2} \gamma^{k-t+1}[R_{k+2} + \gamma V_k(S_{k+2}) - V_k(S_{k+1})]. && (32)
\end{aligned}$$

Exercise 6.2

Q

This is an exercise to help develop your intuition about why TD methods are often more efficient than Monte Carlo methods. Consider the driving home example and how it is addressed by TD and Monte Carlo methods. Can you imagine a scenario in which a TD update would be better on average than a Monte Carlo update? Give an example scenario—a description of past experience and a current state—in which you would expect the TD update to be better. Here's a hint: Suppose you have lots of experience driving home from work. Then you move to a new building and a new parking lot (but you still enter the highway at the same place). Now you are starting to learn predictions for the new building. Can you see why TD updates are likely to be much better, at least initially, in this case? Might the same sort of thing happen in the original task?

A

TD updates incorporate prior information. Suppose we had a good value estimate for a trajectory $\tau = S_1, S_2, \ldots, S_T$, then if we try to estimate the trajectory $\tau = S_0, S_1, S_2, \ldots, S_T$ using MC then we need to see multiple episodes of this to get a good estimate of $V(S_0)$, not leveraging the info we already have on τ. TD would use info on τ to back up the value of S_0 and hence converge much quicker. The key differences here are bootstrapping and online learning.

Exercise 6.3

Q

From the results shown in the left graph of the random walk example it appears that the first episode results in a change in only $V(A)$. What does this tell you about what happened on the first episode? Why was only the estimate for this one state changed? By exactly how much was it changed?

A

All states apart from the terminal states were initialised to the same value (the terminal states must be initialised to 0) and the reward for non-terminal transitions is 0, so the TD(0) updates do nothing on the first pass to states that cannot lead directly to termination.

In the first run, the agent terminated on the left.

$$\begin{aligned} V_1(A) &= V_0(A) + \alpha[0 + \gamma \times 0 + V_0(A)] && (33)\\ &= (1-\alpha)V_0(A) && (34)\\ &= 0.9 \times 0.5 && (35)\\ &= \frac{9}{20}. && (36) \end{aligned}$$

The value of the estimate for A reduced by $\alpha V_0(A) = 0.05$.

Exercise 6.4

Q

The specific results shown in the right graph of the random walk example are dependent on the value of the step-size parameter, α. Do you think the conclusions about which algorithm is better would be affected if a wider range of α values were used? Is there a different, fixed value of α at which either algorithm would have performed significantly better than shown? Why or why not?

A

- General arguments given earlier about the benefits of TD are independent of α
- Increases in α make the curve more
- Decreases in α make the curve more smooth but make it converge slower.
- We see enough of a range here to decide between the two methods

*Exercise 6.5

Q

In the right graph of the random walk example, the RMS error of the TD method seems to go down and then up again, particularly at high αs. What could have caused this? Do you think this always occurs, or might it be a function of how the approximate value function was initialized?

A

The state C happens to have een initialised to its true value. As training starts, updates occur on outer states (making them more accurate individually) which makes the error across all states reduce. This happens until the residual inaccuracies in the outer states propagate to C. The higher values of α make this effect more pronounced, because the value estimate for C changes more readily in these cases.

Exercise 6.6

Q

In Example 6.2 we stated that the true values for the random walk example are $\frac{1}{6}$, $\frac{2}{6}$, $\frac{3}{6}$, $\frac{4}{6}$, and $\frac{5}{6}$ for states A through E. Describe at least two different ways that these could have been computed. Which would you guess we actually used? Why?

A

Could have used DP, but probably did the following calculation.

First note that

$$V(s) = \mathbb{E}[\mathbb{1}\{\text{terminate on right from } S\}] = \mathbb{P}(\text{terminate on right from } S).$$

Also recognise that symmetry now implies $V(C) = 0.5$. Now

$$\begin{aligned} V(E) &= \frac{1}{2} \times 1 + \frac{1}{2} \times V(D) \\ &= \frac{1}{2} + \frac{1}{4}[V(C) + V(E)], \end{aligned}$$

so $V(E) = \frac{5}{6}$. We then get $V(D) = \frac{4}{6}$ and we can calculate the other states in the same way.

*Exercise 6.7

Q

Design an off-policy version of the $TD(0)$ update that can be used with arbitrary target policy π and covering behavior policy b, using at each step t the importance sampling ratio $\rho_{t:t}$(5.3).

A

Let G_t be returns from an episode generated by b. Then

$$\begin{aligned} v_\pi(s) &= \mathbb{E}[\rho_{t:T-1} G_t | S_t = s] \\ &= \mathbb{E}[\rho_{t:T-1} R_{t+1} + \gamma \rho_{t:T-1} G_{t+1} | S_t = s] \\ &= \rho_{t:t} \mathbb{E}[R_{t+1} | S_t = s] + \gamma \rho_{t:t} \mathbb{E}[\rho_{t+1:T-1} G_{t+1} | S_t = s] \\ &= \rho_{t:t} \left(\mathbb{E}[R_{t+1} | S_t = s] + \mathbb{E}[\rho_{t+1:T-1} G_t | S_t = s] \right) \\ &= \rho_{t:t} \left(r(s, A_t) + v_\pi(S_{t+1}) \right). \end{aligned}$$

So the update for off-policy TD(0) (by sampling approximation) is

$$V(S_t) \leftarrow V(S_t) + \alpha \left[\rho_{t:t} R_{t+1} + \rho_{t:t} \gamma V(S_{t+1}) - V(S_t) \right]. \tag{37}$$

Exercise 6.8

Q

Show that an action-value version of (6.6) holds for the action-value form of the TD error $\delta_t = R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)$, again assuming that the values don't change from step to step.

A

Write $\delta_t \doteq R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)$. Then the Monte-Carlo error is

$$\begin{aligned} G_t - Q(S_t, A_t) &= R_{t+1} + \gamma G_{t+1} - Q(S_t, A_t) \\ &= \delta_t - \gamma[Q(S_{t+1}, A_{t+1}) + G_{t+1}] \\ &= \sum_{k=t}^{T-1} \gamma^{k-t} \delta_k \end{aligned}$$

Exercise 6.9 (programming): Windy Grid World with King's Moves

Q

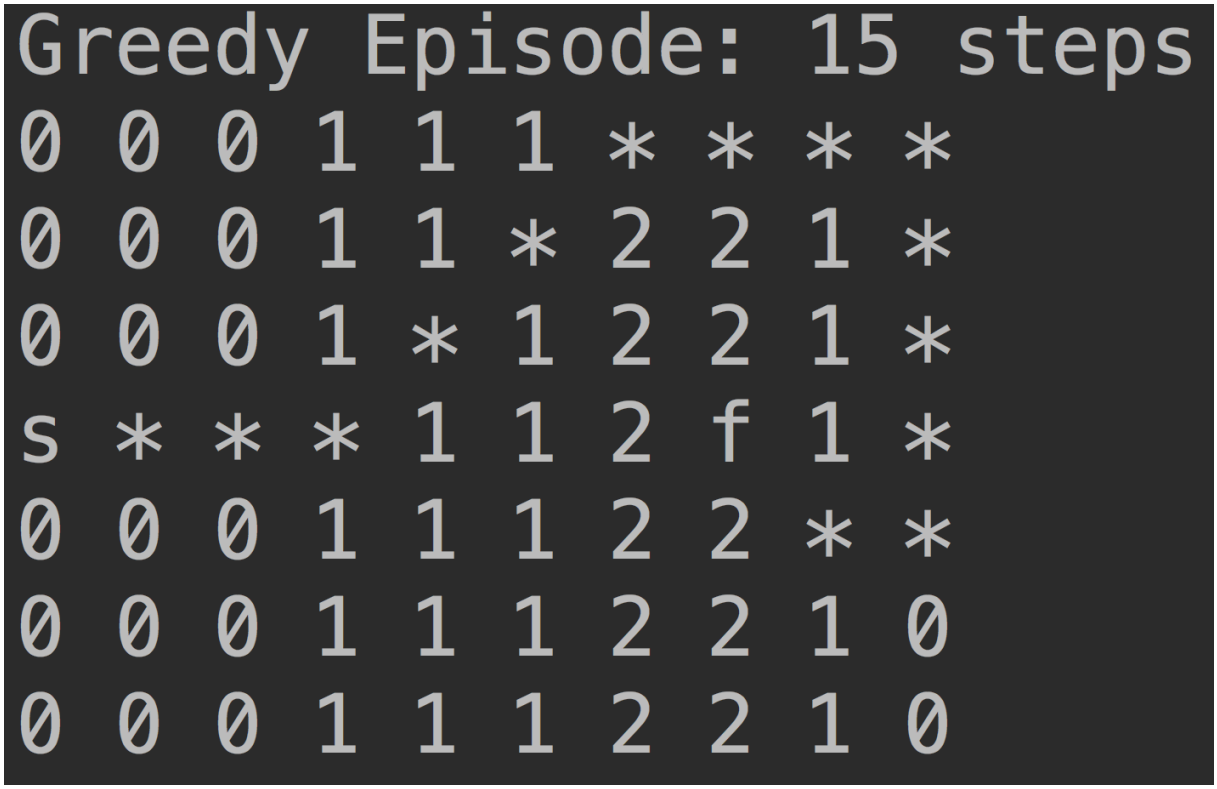
Re-solve the windy gridworld assuming eight possible actions, including the diagonal moves, rather than the usual four. How much better can you do with the extra actions? Can you do even better by including a ninth action that causes no movement at all other than that caused by the wind?

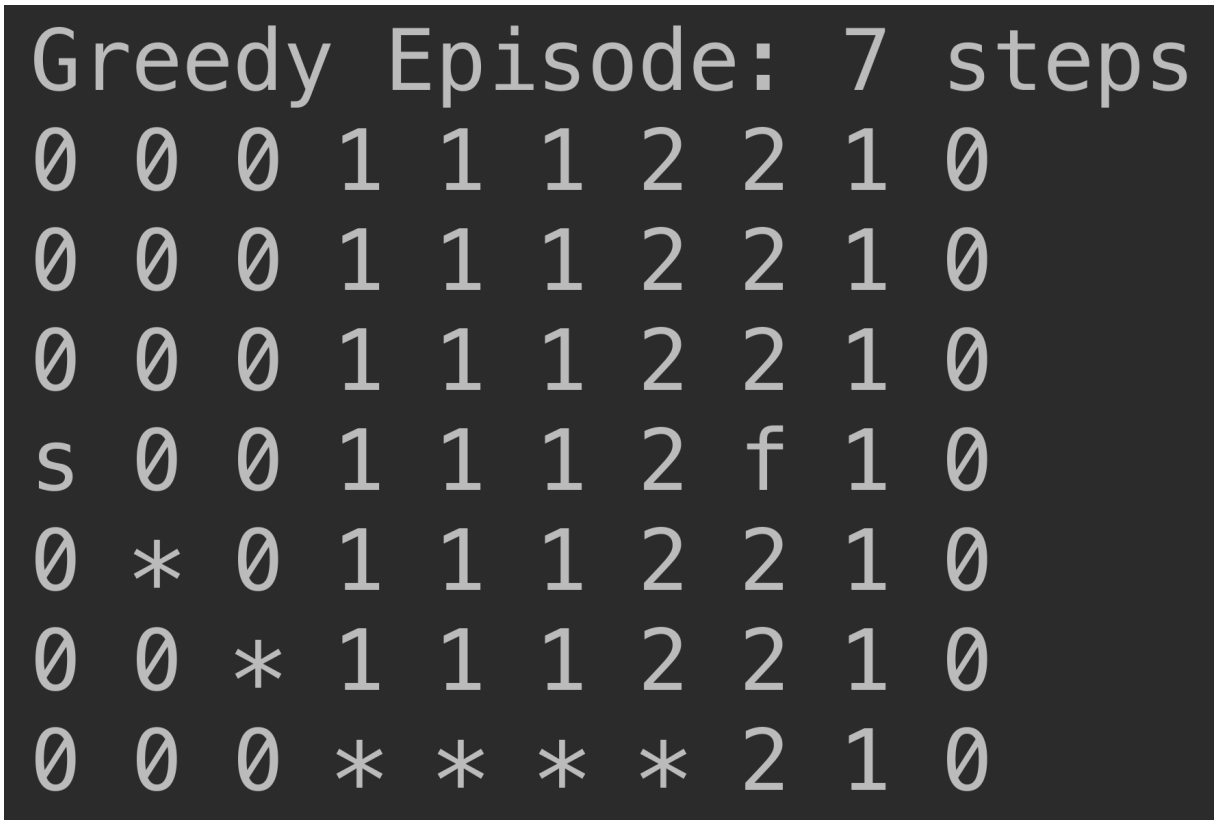
A

This is a programming exercise. For the relevant code please see the repo.

Optimal trajectory is now 7 steps, rather than 15. Including the do-nothing action is not helpful in this example because the wind blows vertically and the goal position is not vertically separated from the start position. It could be useful in other wind environments though.

Below are the optimal trajectories for the book example (no diagonal moves) and for the exercise (king's moves). The numbers represent the wind strength in that position.





Exercise 6.10 (programming): Stochastic Wind

Q

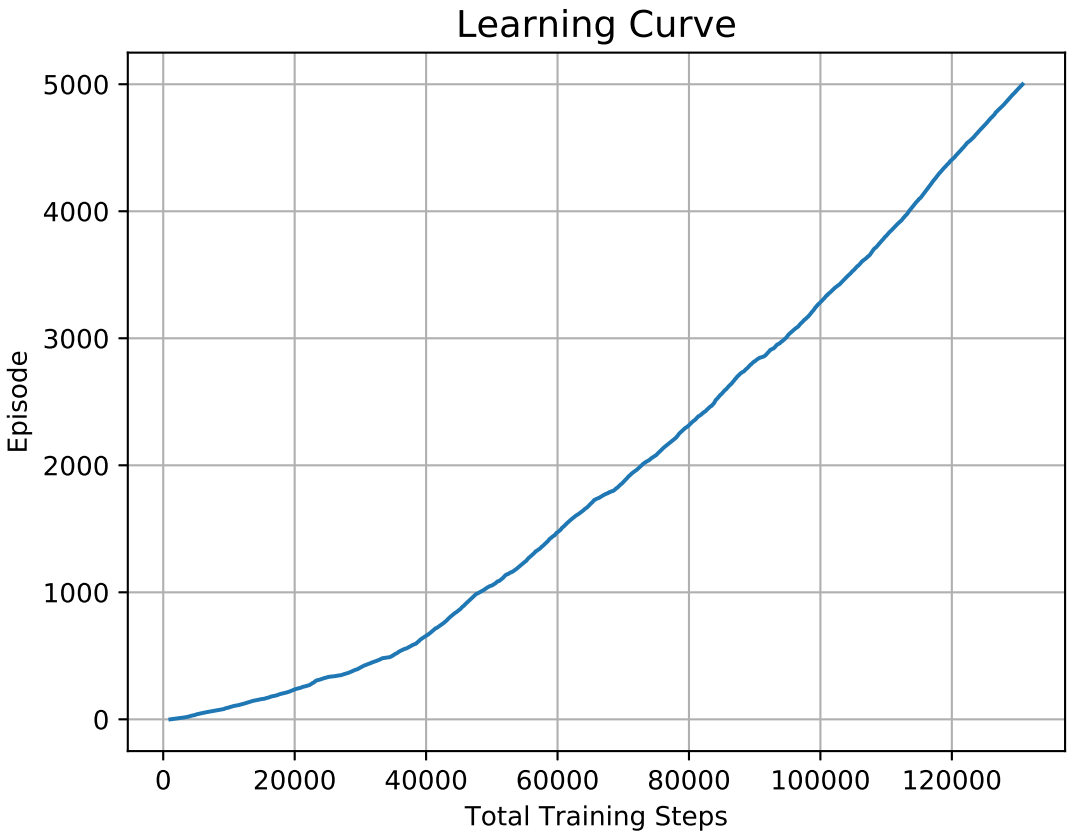
Re-solve the windy gridworld task with King's moves, assuming that the effect of the wind, if there is any, is stochastic, sometimes varying by 1 from the mean values given for each column. That is, a third of the time you move exactly according to these values, as in the previous exercise, but also a third of the time you move one cell above that, and another third of the time you move one cell below that. For example, if you are one cell to the right of the goal and you move left, then one-third of the time you move one cell above the goal, one-third of the time you move two cells above the goal, and one-third of the time you move to the goal.

A

This is a programming exercise. For the relevant code please see the repo.

Greedy trajectory and learning curve shown below. Note that although the gradient of the learning curve becomes constant (so the algorithm converges), the greedy episode shown suffers from stochasticity in the wind.

```
Greedy Episode: 26 steps
0 0 0 1 1 1 2 2 1 0
0 0 0 1 1 1 2 2 1 0
0 0 0 1 1 1 2 2 * *
s 0 0 1 1 1 2 f * *
0 * * 1 1 1 2 * * *
0 0 0 1 1 1 * * 1 0
0 0 0 * * * 2 2 * 0
```



Exercise 6.11

Q

Why is Q-learning considered an *off-policy* control method?

A

The returns are sampled as if the agent followed the greedy policy with respect to Q.

Exercise 6.12

Q

Suppose action selection is greedy. Is Q-learning then exactly the same algorithm as Sarsa? Will they make exactly the same action selections and weight updates?

A

Yes (?)

Exercise 6.13

Q

What are the update equations for Double Expected Sarsa with an ε-greedy target policy?

A

Expected SARSA has the update

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left[R_{t+1} + \gamma \mathbb{E}[Q(S_{t+1}) | S_{t+1}] - Q(S_t, A_t) \right]$$

The update for S_t, A_t is

$$R_{t+1} + \gamma \sum_a \pi(a|S_{t+1}) Q(S_{t+1}, a) - Q(S_t, A_t).$$

Double expected SARSA would be keeping two Q arrays and updating one of them each timestep, chosen with equal probability.

For an ε-greedy policy we would increment $Q_1(S_t, A_t)$ by

$$\alpha \left[R_{t+1} + \gamma \left(\frac{\varepsilon}{|\mathcal{A}(a)|} \sum_a Q_2(S_{t+1}, a) + (1 - \varepsilon) \max_a \{ Q_2(S_{t+1}, a) \} \right) - Q_1(S_t, A_t) \right] \tag{38}$$

and the same with 1 and 2 reversed.

Exercise 6.14

Q

Describe how the task of Jack's Car Rental (Example 4.2) could be reformulated in terms of after-states. Why, in terms of this specific task, would such a reformulation be likely to speed convergence?

A

One might have coded this up initially with the states as the number of cars in each garage each evening. The agent then takes some action (moves some cars) and we transition stochastically to some state.

An alternative would be to introduce the number of cars in the morning (after the agent has moved cars) as an afterstate. This is because the agent is able to deterministically change the environment from evening to next morning (before rentals or returns).

In this case we would speed convergence by reducing the number of action-values to be calculated. For instance, we can now evaluate $(10, 0)$ moving one car and $(9, 1)$ moving no cars as the same afterstate $(9, 1)$.

7 n-step TD Prediction

Exercise 7.1

Q

In Chapter 6 we noted that the Monte Carlo error can be written as the sum of TD errors (6.6) if the value estimates don't change from step to step. Show that the n-step error used in (7.2) can also be written as a sum TD errors (again if the value estimates don't change) generalizing the earlier result.

A

Write

$$G_{t:t+n} = \sum_{i=1}^{n} \gamma^{i-1} R_{t-i} + \gamma^n V(S_{t+n}),$$

then have the TD error

$$\delta_t \doteq R_{t+1} + \gamma V(S_{t+1}) - V(S_t).$$

So the n-step error can be written as

$$\begin{aligned} G_{t:t+n} - V(S_t) &= R_{t+1} + \gamma \sum_{i=1}^{n-1} \gamma^{i-1} R_{t+1+i} + \gamma^n V(S_{t+n}) - V(S_t) \\ &= \delta_t + \gamma \left(G_{t+1:t+n} - V(S_{t+1}) \right) \\ &\vdots \\ &= \sum_{k=t}^{t+n-1} \gamma^{k-t} \delta_k \end{aligned}$$

Exercise 7.2 (programming)

Q

With an n-step method, the value estimates *do* change from step to step, so an algorithm that used the sum of TD errors (see previous exercise) in place of the error in (7.2) would actually be a slightly different algorithm. Would it be a better algorithm or a worse one? Devise and program a small experiment to answer this question empirically.

A

...

Exercise 7.3

Q

Why do you think a larger random walk task (19 states instead of 5) was used in the examples of this chapter? Would a smaller walk have shifted the advantage to a different value of n? How about the change in left-side outcome from 0 to -1 made in the larger walk? Do you think that made any difference in the best value of n?

A

- Smaller walk would have shifted advantage to smaller n because when $n \geq \frac{\#\text{states}-1}{2}$ (#states is odd) the algorithm updates all states visited by the terminal reward. This means that the algorithm only makes value changes of size α, since the values are no longer bootstrapped or backed up.
- The addition of the -1 reward on the left favours smaller values of n, because in longer episodes the larger values of n will have to update many states by the terminal reward (now -1 rather than 0) thus increasing variance

Exercise 7.4

Q

Prove that the n-step return of Sarsa (7.4) can be written exactly in terms of a novel TD error, as

$$G_{t:t+n} = Q_{t-1}(S_t, A_t) + \sum_{k=t}^{\min(t+n,T)-1} \gamma^{k-1}[R_{k+1} + \gamma Q_k(S_{k+1}, A_{k+1}) - Q_{k-1}(S_k, A_k)].$$

A

Denote

$$G_{t:t+n} \doteq \sum_{i=1}^{n} \gamma^{i-1} R_{t+i} + \gamma^n Q_{t+n-1}(S_{t+n}, A_{t+n})$$

for $n \geq 1$ and $0 \leq t < T - n$ and with $G_{t:t+n} = G_t$ if $t + n > T$.

Set $\tau = \min(t + n, T) - 1$ and observe that

$$\begin{aligned}
&\sum_{k=t}^{\tau} \gamma^{k-t}[R_{t+1} + \gamma Q_k(S_{k+1}, A_{k+1}) - Q_{k-1}(S_k, A_k)] \\
&= \sum_{k=t}^{\tau} \gamma^{k-t} R_{t+1} + \gamma \sum_{k=t}^{\tau} \gamma^{k-t} Q_k(S_{k+1}, A_{k+1}) - \sum_{k=t}^{\tau} \gamma^{k-t} Q_{k-1}(S_k, A_k) \\
&= G_{t:t+n} - \mathbb{1}\{t + n < T\}\gamma^n Q_{t+n-1}(S_{t+n}, A_{t+n}) + \gamma^\tau Q_\tau(S_\tau, A_\tau) - Q_{t-1(S_t,A_t)} \\
&= G_{t:t+n} - Q_{t-1}(S_t, A_t).
\end{aligned}$$

Exercise 7.5

Q

Write the pseudocode for the off-policy state-value prediction algorithm described above.

A

The update that we use is the same as the n-step TD update, only multiplied by the importance sampling ratio and with the control variate added.

$$G_{t:h} \doteq \rho_t(R_{t+1} + \gamma G_{t+1:h}) + (1 - \rho_t)V_{h-1}(S_t)$$

The algorithm is therefore the same, but with these steps replacing the old returns calculations, using the latest available value function.

If you take the recursion relation literally then we should get a control variate for each of the intermediary states $(t + 1, \ldots, t + n)$, but I don't think that this is what is intended as this would just increase variance further.

Exercise 7.6

Q

Prove that the control variate in the above equations does not change the expected value of the return.

A

In (7.13) we have

$$\begin{aligned}\mathbb{E}[(1-\rho_t)V_{h-1}(S_t)] &= \mathbb{E}_b[(1-\rho_t)V_{h-1}(S_t)] \\ &= \mathbb{E}_b[(1-\rho_t)]\mathbb{E}_b[V_{h-1}(S_t)] \\ &= 0.\end{aligned}$$

In the second case (7.14) we have

$$\begin{aligned}&\mathbb{E}_b[\bar{V}_{h-1}(S_{t+1}) - \rho_{t+1}Q_{h-1}(S_{t+1},A_{t+1})|S_{t+1}] \\ &= \sum_a \pi(a|S_{t+1})Q_{h-1}(S_{t+1},a) - \sum_a b(a|S_{t+1})\frac{\pi(a|S_{t+1})}{b(a|S_{t+1})}Q_{h-1}(S_{t+1},a) \\ &= 0\end{aligned}$$

*Exercise 7.7

Q

Write the pseudocode for the off-policy action-value prediction algorithm described immediately above. Pay particular attention to the termination conditions for the recursion upon hitting the horizon or the end of episode.

A

...

Exercise 7.8

Q

Show that the general (off-policy) version of the n-step return (7.13) can still be written exactly and compactly as the sum of state-based TD errors (6.5) if the approximate state value function does not change.

A

Update target is

$$G_{t:h} = \rho_t(R_{t+1} + \gamma G_{t+1:h}) + (1-\rho_t)V_{h-1}(S_t).$$

Assume state-value function does not change and introduce the TD error

$$\delta_t \doteq R_{t+1} + \gamma V(S_{t+1}) - V(S_t).$$

Then

$$\begin{aligned} G_{t:h} - V(S_t) &= \rho_t \left(R_{t+1} + \gamma G_{t+1:h} - V(S_t) \right) \\ &= \rho_t \left(R_{t+1} + \gamma [G_{t+1:h} - V(S_{t+1})] + \gamma V(S_{t+1}) - V(S_t) \right) \\ &= \rho_t \delta_t + \rho_t \gamma [G_{t+1:h} - V(S_{t+1})] \\ &\vdots \\ &= \sum_{i=t}^{\min(h,T)-1} \rho_{t:i} \gamma^{i-t} \delta_i \end{aligned}$$

Exercise 7.9

Q

Repeat the above exercise for the action version of the off-policy n-step return (7.14) and the Expected Sarsa TD error (the quantity in brackets in Equation 6.9).

A

Action-value update is

$$G_{t:h} = R_{t+1} + \gamma \rho_{t+1} \left(G_{t+1:h} - Q_{h-1}(S_{t+1}, A_{t+1}) \right) - \gamma \bar{V}_{h-1}(S_h)$$

where

$$\bar{V}_h \doteq \sum_a \pi(a|S_h) Q(S_h, a)$$

and we assume that the action value function does not change between iterations. Define

$$\delta_t \doteq R_{t+1} + \gamma \bar{V}(S_{t+1}) - Q(S_t, A_t).$$

Then

$$\begin{aligned} G_{t:h} - Q(S_t, A_t) &= \delta_t + \gamma \rho_{t+1} (G_{t+1:h} - Q(S_{t+1}, A_{t+1})) \\ &\vdots \\ &= \sum_{i=t}^{\min(h,T)-1} \gamma^{i-t} \rho_{t+1:i} \delta_i \end{aligned}$$

where we enforce the convention that $\rho_{a:b} = 1$ if $a > b$.

Exercise 7.10 (programming)

Q

Devise a small off-policy prediction problem and use it to show that the off-policy learning algorithm using (7.13) and (7.2) is more data efficient than the simpler algorithm using (7.1) and (7.9).

A

...

Exercise 7.11

Q

Show that if the approximate action values are unchanging, then the tree-backup return (7.16) can be written as a sum of expectation-based TD errors:

$$G_{t:t+n} = Q(S_t, A_t) + \sum_{k=t}^{\min(t+n,T)-1} \delta_k \prod_{i=t+1}^{k} \gamma\pi(A_i|S_i),$$

where $\delta_t \doteq R_{t+1} + \gamma\bar{V}_t(S_{t+1}) - Q(S_t, A_t)$ and $\bar{V}_t$ is given by (7.8).

A

Assume action-values unchanging. The recursion formula for tree backup is

$$G_{t:t+n} = R_{t+1} + \gamma \sum_{a \neq A_{t+1}} \pi(a|S_{t+1})Q(S_{t+1}, a) + \gamma\pi(A_{t+1}, S_{t+1})G_{t+1:t+n}.$$

Define

$$\delta_t \doteq R_{t+1} + \gamma\bar{V}(S_{t+1}) - Q(S_t, A_t)$$

where

$$\bar{V}_h \doteq \sum_a \pi(a|S_h)Q(S_h, a).$$

Then

$$\begin{aligned} G_{t:t+n} - Q(S_t, A_t) &= R_{t+1} + \gamma\bar{V}(S_{t+1}) - \gamma\pi(A_{t+1}, S_{t+1})Q(S_{t+1}, A_{t+1}) \\ &\quad - Q(S_t, A_t) + \gamma\pi(A_{t+1}, S_{t+1})G_{t+1:t+n} \\ &= \delta_t - \gamma\pi(A_{t+1}, S_{t+1})[G_{t+1:t+n} - Q(S_{t+1}, A_{t+1})] \\ &\vdots \\ &= \sum_{i=1}^{\min(t+n,T)-1} \delta_i \prod_{j=t+1}^{i} \gamma\pi(A_j, S_j) \end{aligned}$$

where we define the product operator to have the behaviour $\prod_a^b[\cdot] = 1$ for $a > b$.

8 Planning and Learning with Tabular Methods

Exercise 8.1

Q

The non-planning method looks particularly poor in Figure 8.3 because it is a one-step method; a method using multi-step bootstrapping would do better. Do you think one of the multi-step bootstrapping methods from Chapter 7 could do as well as the Dyna method? Explain why or why not.

A

Dyna updates using all past experience so quickly synthesises this into an optimal trajectory. n-step bootstrapping might be slower because it only states visited in the last n steps.

Exercise 8.2

Q

Why did the Dyna agent with exploration bonus, Dyna-Q+, perform better in the first phase as well as in the second phase of the blocking and shortcut experiments?

A

Increased exploration means Dyna-Q+ finds the optimal policy quicker than Dyna-Q. Dyna-Q may find a trajectory that works but is suboptimal and then have to wait a long time for it to take enough exploratory actions to find an optimal policy.

Exercise 8.3

Q

Careful inspection of Figure 8.5 reveals that the difference between Dyna-Q+ and Dyna-Q narrowed slightly over the first part of the experiment. What is the reason for this?

A

Dyna-Q+ will take suboptimal actions in order to explore (when τ gets large). Dyna-Q will not do this so has better asymptotic performance.

Exercise 8.4 (programming)

Q

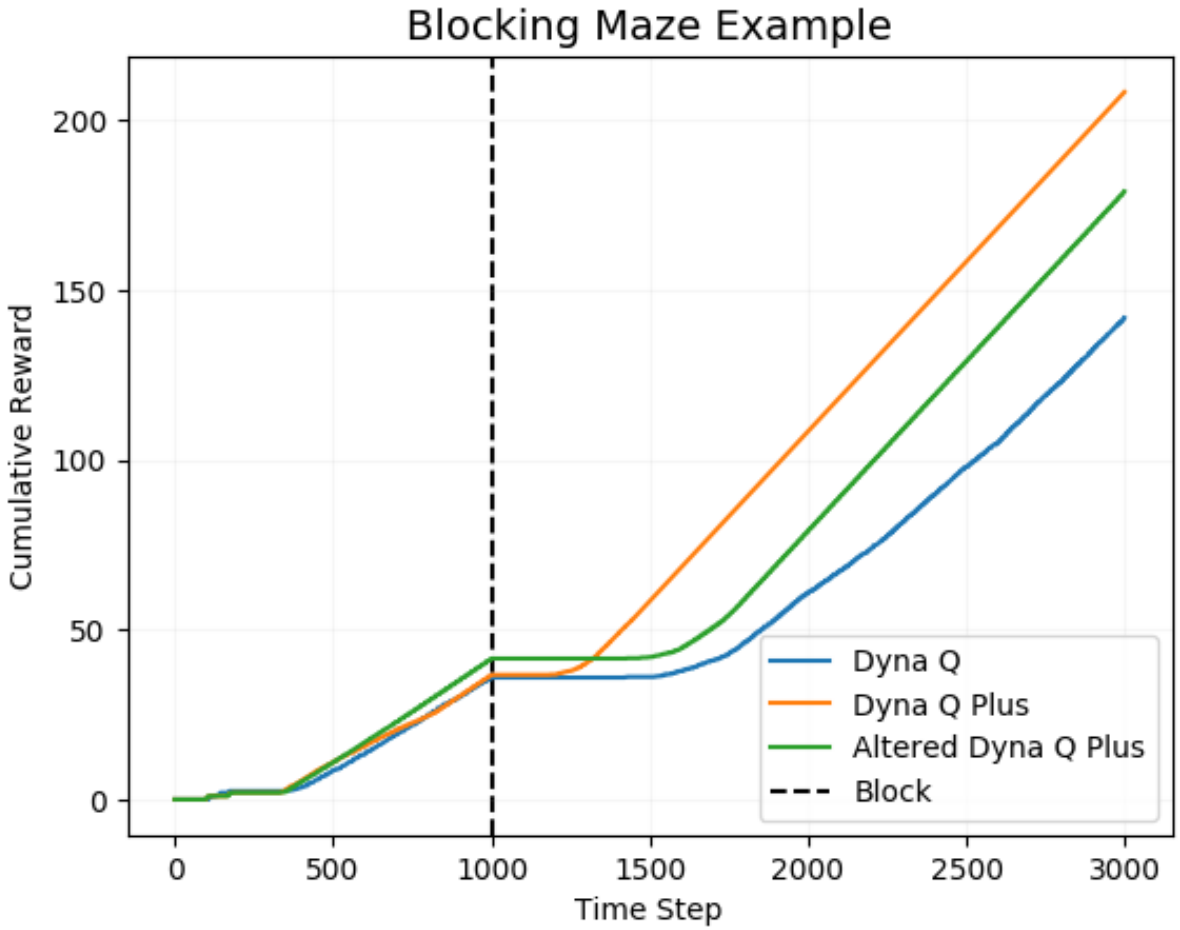
The exploration bonus described above actually changes the estimated values of states and actions. Is this necessary? Suppose the bonus $\kappa\sqrt{\tau}$ was used not in updates, but solely in action selection. That is, suppose the action selected was always that for which $Q(S_t, a) + \kappa\sqrt{\tau(S_t, a)}$ was maximal. Carry out a gridworld experiment that tests and illustrates the strengths and weaknesses of this alternate approach.

A

This is a programming exercise. For the relevant code please see the repo.

The change means that exploration only takes into account the next action, not whole trajectories. In the Dyna-Q+ algorithm can explore whole new paths through the planning stage.

This is backed up by the results. The altered Dyna-Q+ learns the new path slower because it can't plan new trajectories.



Exercise 8.5

Q

How might the tabular Dyna-Q algorithm shown on page 164 be modified to handle stochastic environments? How might this modification perform poorly on changing environments such as considered in this section? How could the algorithm be modified to handle stochastic environments *and* changing environments?

A

- You could take frequency of occurrences of transitions to estimate the transition probability for the model. (These would be the MLE estimates.)
- Make expected updates when planning.

- This would would present an issue if the environment changed since the changes would just be reflected in changing transition probabilities (which could take a long time to reflect the change in the environment.)
- A solution to this could be to use an exploration bonus to encourage the agent to continue to select various states and keep the model up to date.
- A better solution would be to add some notion of confidence to the model estimates of the transition probabilities. Could model the probabilities like

  $$p(s, a, s') = \hat{p}(s, a, s')(1 - \sigma(\tau)) + \sigma(\tau)e,$$

  where $\hat{p}$ is the MLE estimate of the probabilities, e is the equiprobable estimate and $\sigma(\tau)$ is a sigmoid of the time since the state-action pair (s, a) was list visited.

Exercise 8.6

Q

The analysis above assumed that all of the *b* possible next states were equally likely to occur. Suppose instead that the distribution was highly skewed, that some of the *b* states were much more likely to occur than most. Would this strengthen or weaken the case for sample updates over expected updates? Support your answer.

A

If the transition probabilities are skewed then the expected updates perform the same while sample updates get accurate on the most probably outcomes very quickly. This strengthens the case for sample updates.

Exercise 8.7

Q

Some of the graphs in Figure 8.8 seem to be scalloped in their early portions, particularly the upper graph for $b = 1$ and the uniform distribution. Why do you think this is? What aspects of the data shown support your hypothesis?

A

In the case of the uniform distribution of updates and $b = 1$, the start state is visited roughly once every $|\mathcal{S}|$ updates. When this happens, the action values of the neighbourhood of the start state are updated and their undergo a greater change than when the states that are not in the neighbourhood are the start state are updated. Thus, when the policy is evaluated, the value of the start state changes a lot if the start state has been visited recently, and not so much otherwise (since the change comes from values backed up from states far away from the start state).

In the on-policy case, the start state is visited much more often (on average more than once every 10 updates, since $\mathbb{P}(\texttt{terminate}) = 0.1$) so it does not exhibit this behaviour. When b is larger there are more connections between states, so the neighbourhood of the start state is larger, so this feature is also reduced.

Exercise 8.8 (programming)

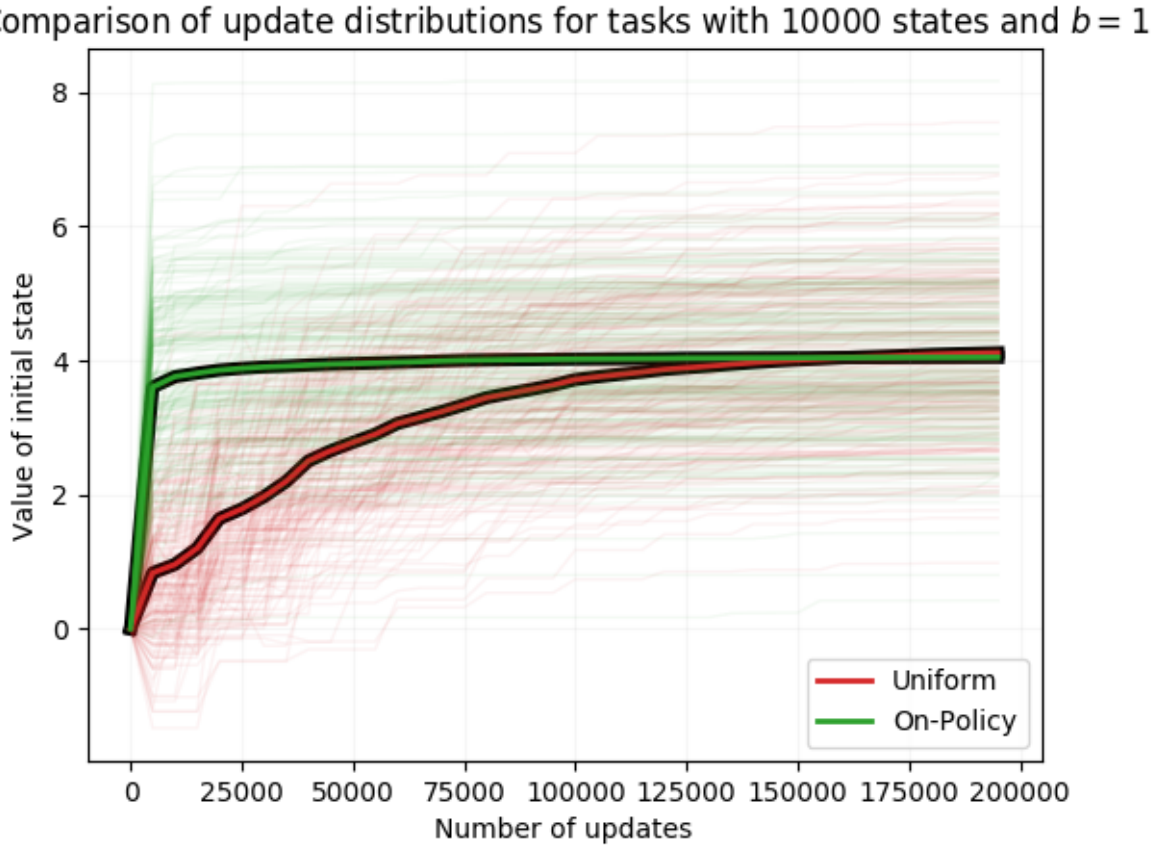
Q

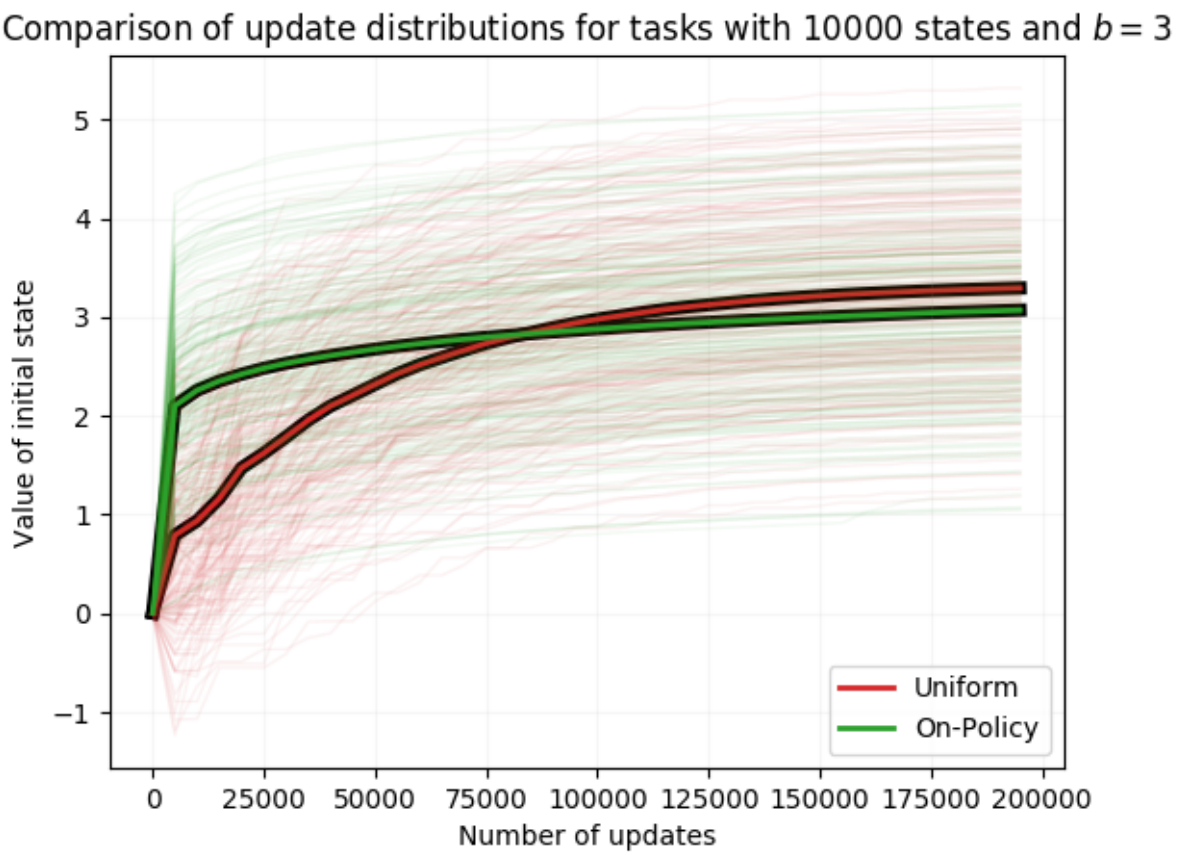
Replicate the experiment whose results are shown in the lower part of Figure 8.8, then try the same experiment but with $b = 3$. Discuss the meaning of your results.

A

This is a programming exercise. For the relevant code please see the repo.

Charts show averages of 200 runs. We see that in the $b = 3$ case the uniformly distributed updates overtakes the on-policy updates much quicker. This is due to the greater complexity of the state-space (the number of states on which the value of the starting state depends is exponential in b), of which the on-policy updates neglects large portions.




Comparison of update distributions for tasks with 10000 states and $b = 3$
Value of initial state
5
4
3
2
1
0
−1
0
25000
50000
75000
100000
125000
150000
175000
200000
Number of updates
Uniform
On-Policy

9 On-policy Prediction with Approximation

Exercise 9.1

Q

Show that tabular methods such as presented in Part I of this book are a special case of linear function approximation. What would the feature vectors be?

A

Write $\hat{V}(s, \mathbf{w}) = w$ and we get that $\nabla_{\mathbf{w}} \hat{V}(s, \mathbf{w}) = 1$ so we return to tabular TD learning. In this case the features are $x(s) = 1 \; \forall s \in \mathcal{S}$.

Exercise 9.2

Q

Why does (9.17) define $(n+1)^k$ distinct features for dimension k?

A

Each of the k terms can independently have one of $n+1$ exponents, hence the total number of features is $(n+1)^k$.

Exercise 9.3

Q

What n and $c_{i,j}$ produce the feature vectors $\mathbf{x}(s) = (1, s_1, s_2, s_1 s_2, s_1^2, s_2^2, s_1 s_2^2, s_1^2 s_2, s_1^2 s_2^2)$?

A

$n = 2$ and $c_i, j = C_{ij}$ where

$$C = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ 1 & 1 \\ 2 & 0 \\ 0 & 2 \\ 1 & 2 \\ 2 & 1 \\ 2 & 2 \end{pmatrix}$$

Exercise 9.4

Q

Suppose we believe that one of two state dimensions is more likely to have an effect on the value function than is the other, that generalization should be primarily across this dimension rather than along it. What kind of tilings could be used to take advantage of this prior knowledge?

A

Tiles that are thin along the dimension of interest and long across it. Rectangles, for instance.

Exercise 9.5

Q

Suppose you are using tile coding to transform a seven-dimensional continuous state space into binary feature vectors to estimate a state value function $\hat{v}(s, \mathbf{w}) \approx v_\pi(s)$. You believe that the dimensions do not interact strongly, so you decide to use eight tilings of each dimension separately (stripe tilings), for $7 \times 8 = 56$ tilings. In addition, in case there are some pairwise interactions between the dimensions, you also take all $\binom{7}{2} = 21$ pairs of dimensions and tile each pair conjunctively with rectangular tiles. You make two tilings for each pair of dimensions, making a grand total of $21 \times 2 + 56 = 98$ tilings. Given these feature vectors, you suspect that you still have to average out some noise, so you decide that you want learning to be gradual, taking about 10 presentations with the same feature vector before learning nears its asymptote. What step-size parameter α should you use? Why?

A

Each tiling is a partition, so each tiling has exactly one tile activated per state. This means that in our case the number of features is 98. We consider each of these equally likely because we are uninformed. We therefore take

$$\alpha = \frac{1}{10 \times 98} = \frac{1}{980}.$$

So that on average we see each feature 10 times before asymptote. [Note that this assumes a constant target.]

10 On-policy Control with Approximation

Exercise 10.1

Q

We have not explicitly considered or given pseudocode for any Monte Carlo methods or in this chapter. What would they be like? Why is it reasonable not to give pseudocode for them? How would they perform on the Mountain Car task?

A

- Monte Carlo is n-step Sarsa with $n \to \infty$
- This is same pseudocode as given, but with full episodes and G_t rather than $G_{t:t+n}$.
- Could have been very poor on the mountain car as may never have finished the first episode and does not learn within an episode (online)

Exercise 10.2

Q

Give pseudocode for semi-gradient one-step *Expected Sarsa* for control.

A

Expected sarsa is the same but the target is

$$\sum_{k=t}^{t+n-1} \gamma^{k-t} R_{k+1} + \sum_a \pi(a|S_{t+n}) q_{t+n-1}(S_{t+n}, a)$$

Exercise 10.3

Q

Why do the results shown in Figure 10.4 have higher standard errors at large n than at small n?

A

The longer the step length then the greater the variance in initial runs, this is because the agent needs to wait for n steps to start learning. Some initial episodes of high n cases could have been very poor.

Exercise 10.4

Q

Give pseudocode for a differential version of semi-gradient Q-learning.

A

Same as others but with the target

$$R_{t+1} - \bar{R}_{t+1} - \max_a \hat{q}(S_{t+1}, a, \mathbf{w}_t)$$

Exercise 10.5

Q

What equations are needed (beyond 10.10) to specify the differential version of TD(0)?

A

Just need the semi-gradient update

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \alpha\delta_t \nabla_{\mathbf{w}_t} \hat{v}(S_t, \mathbf{w}_t)$$

where

$$\delta_t = R_{t+1} - \bar{R}_{t+1} + \hat{v}(S_{t+1}, \mathbf{w}_t) - \hat{v}(S_t, \mathbf{w}_t)$$

Exercise 10.6

Q

Consider a Markov reward process consisting of a ring of three states A, B, and C, with state transitions going deterministically around the ring. A reward of 1 is received upon arrival in A and otherwise the reward is 0. What are the differential values of the three states?

A

The average reward is $\bar{R} = \frac{1}{3}$. To calculate the differential return we have

$$V(A) = \sum_t (a_t - \bar{R})$$

where $a_i = \mathbb{1}\{i + 1 \equiv 0 \pmod 3\}$. This doesn't converge in the normal way, so to attempt to calculate it let's consider

$$V(A;\gamma) = \sum_t \gamma^t \left(a_t - \frac{1}{3}\right)$$

then, formally, we have

$$\lim_{\gamma \to 1} V(A;\gamma) = V(A).$$

Now

$$\begin{aligned} V(A;\gamma) &= -\frac{1}{3} - \frac{1}{3}\gamma + \frac{2}{3}\gamma^2 + \sum_{t=3}^{\infty} \gamma^t \left(a_t - \frac{1}{3}\right) \\ &= \frac{1}{3}(2\gamma^2 - \gamma - 1) + \gamma^3 \sum_{t=0}^{\infty} \gamma^t \left(a_t - \frac{1}{3}\right) \end{aligned}$$

so

$$\begin{aligned} V(A;\gamma) &= \frac{1}{3}\frac{2\gamma^2 - \gamma - 1}{1 - \gamma^3} \\ &= -\frac{1}{3}\frac{2\gamma + 1}{\gamma^2 + \gamma + 1} \end{aligned}$$

which leads to $V(A) = -\frac{1}{3}$.

Then we have

$$V(A) = -\frac{1}{3} + V(B) \quad \implies \quad V(B) = 0$$

and

$$V(B) = -\frac{1}{3} + V(C) \quad \implies \quad V(C) = \frac{1}{3}.$$

Exercise 10.7

Q

Suppose there is an MDP that under any policy produces the deterministic sequence of rewards 1, 0, 1, 0, 1, 0, . . . going on forever. Technically, this is not allowed because it violates ergodicity; there is no stationary limiting distribution μ_π and the limit (10.7) does not exist. Nevertheless, the average reward (10.6) is well defined; What is it? Now consider two states in this MDP. From A, the reward sequence is exactly as described above, starting with a 1, whereas, from B, the reward sequence starts with a 0 and then continues with 1, 0, 1, 0, ... The differential return (10.9) is not well defined for this case as the limit does not exist. To repair this, one could alternately define the value of a state as

$$v_\pi(s) \doteq \lim_{\gamma \to 1} \lim_{h \to \infty} \sum_{t=0}^{h} \gamma^t \left(\mathbb{E}_\pi[R_{t+1} | S_0 = s] - r(\pi) \right).$$

Under this definition, what are the values of states A and B?

A

Define

$$f(h) = \frac{1}{2h} \sum_{t=0}^{2h} \mathbb{1}\{t \equiv 0 \pmod 2\} = \frac{h+1}{2h}$$

then

$$\bar{R} = \lim_{h \to \infty} f(h/2) = \lim_{h \to \infty} f(h) = \frac{1}{2}.$$

Now to compute the differential state values we write

$$V(S; \gamma) = \lim_{h \to \infty} \sum_{t=0}^{h} \gamma^t \left(\mathbb{E}[R_{t+1} | S_0 = s] - \bar{R} \right)$$

then

$$\begin{aligned} V(A; \gamma) &= 1 - \bar{R} + \gamma V(A; \gamma) \\ V(A; \gamma) &= -\bar{R} + \gamma V(B; \gamma) \end{aligned}$$

so

$$V(A; \gamma) = \frac{1}{2}(1 - \gamma) - \gamma^2 V(A; \gamma)$$

and

$$\begin{aligned} V(A; \gamma) &= \frac{1}{2} \frac{1 - \gamma}{1 - \gamma^2} \\ &= \frac{1}{2(1 + \gamma)}. \end{aligned}$$

Finally, $V(A) = \lim_{\gamma \to 1} V(A; \gamma) = \frac{1}{4}$ and $V(B) = -\frac{1}{4}$.

Exercise 10.8

Q

The pseudocode in the box on page 251 updates $\bar{R}_{t+1}$ using δ_t as an error rather than simply $R_{t+1} - \bar{R}_{t+1}$. Both errors work, but using δ_t is better. To see why, consider the ring MRP of three states from Exercise 10.6. The estimate of the average reward should tend towards its true value of

$\frac{1}{3}$. Suppose it was already there and was held stuck there. What would the sequence of $R_{t+1} - \bar{R}_{t+1}$ errors be? What would the sequence of δ_t errors be (using (10.10))? Which error sequence would produce a more stable estimate of the average reward if the estimate were allowed to change in response to the errors? Why?

A

$\bar{R} = \frac{1}{3}$ fixed.

The sequence of errors from $R_t - \bar{R}_t$ starting in A would be

$$-\frac{1}{3}, -\frac{1}{3}, \frac{2}{3}, -\frac{1}{3}, -\frac{1}{3}, \frac{2}{3}, \dots$$

while the sequence of TD errors starting in A (taking differential values from Exercise 10.6) would be

$$0, 0, 0, 0, 0, 0, \dots$$

which is clearly of much lower variance and would therefore give more stable updates. Once $\bar{R}$ gets to the correct value it never leaves.

Exercise 10.9

Q

In the differential semi-gradient n-step Sarsa algorithm, the step-size parameter on the average reward, β, needs to be quite small so that $\bar{R}$ becomes a good long-term estimate of the average reward. Unfortunately, $\bar{R}$ will then be biased by its initial value for many steps, which may make learning inefficient. Alternatively, one could use a sample average of the observed rewards for $\bar{R}$. That would initially adapt rapidly but in the long run would also adapt slowly. As the policy slowly changed, $\bar{R}$ would also change; the potential for such long-term non-stationarity makes sample-average methods ill-suited. In fact, the step-size parameter on the average reward is a perfect place to use the unbiased constant-step-size trick from Exercise 2.7. Describe the specific changes needed to the boxed algorithm for differential semi-gradient n-step Sarsa to use this trick.

A

We define a parameter β and seed a sequence u_n with $u_0 = 0$. The under the if statement where $\tau \geq 0$ we place the following:

$$u \leftarrow u + \beta(1 - u)$$
$$\bar{R} \leftarrow \bar{R} + \frac{\beta}{\mu}(R - \bar{R})$$

11 *Off-policy Methods with Approximation

Exercise 11.1

Q

Convert the equation of n-step off-policy TD (7.9) to semi-gradient form. Give accompanying definitions of the return for both the episodic and continuing cases.

A

Tabular case is

$$V_{t+n}(S_t) = V_{t+n-1} + \alpha\rho_{t:t+n-1}[G_{t:t+n} - V_{t+n-1}(S_t)].$$

The semi-gradient weight update is

$$\mathbf{w}_{t+n} = \mathbf{w}_{t+n-1} + \alpha\rho_{t:t+n-1}[G_{t:t+n} - \hat{v}(S_t, \mathbf{w}_{t+n-1})]\nabla_{\mathbf{w}}\hat{v}(S_t, \mathbf{w}_{t+n-1}),$$

noting the occurrence of the $n$step TD Error

$$\delta_t^n = G_{t:t+n} - \hat{v}(S_t, \mathbf{w}_{t+n-1}).$$

We define the returns in the two cases

episodic $G_{t:t+n} = \sum_{i=t}^{t+n-1} \gamma_{i-t}R_{i+1} + \gamma^n\hat{v}(S_{t+n}, \mathbf{w}_{t+n-1})$

continuing $G_{t:t+n} = \sum_{i=t}^{t+n-1}(R_{i+1} - \bar{R}_i) + \hat{v}(S_{t+n}, \mathbf{w}_{t+n-1})$

where in each case $G_{t:h} = G_t$ if $h \geq T$.

*Exercise 11.2

Q

Convert the equations of n-step Q(σ) (7.11 and 7.17) to semi-gradient form. Give definitions that cover both the episodic and continuing cases.

A

The update is

$$\mathbf{w}_{t+n} = \mathbf{w}_{t+n-1} + \alpha[G_{t:t+n} - \hat{q}(S_t, A_t, \mathbf{w}_{t+n-1})]\nabla_{\mathbf{w}}\hat{q}(S_t, A_t, \mathbf{w}_{t+n-1}$$

with the following definitions of returns targets

Episodic

$$G_{t:h} = R_{t+1} + \gamma\left[\sigma_{t+1}\rho_{t+1} + (1-\sigma_{t+1})\pi(A_{t+1}|S_{t+1})\right]\left[G_{t:h} - \hat{q}(S_t, A_t, \mathbf{w}_{h-1})\right] + \gamma\bar{V}_{h-1}(S_{t+1})$$

Continuing

$$G_{t:h} = R_{t+1} - \bar{R}_t + \left[\sigma_{t+1}\rho_{t+1} + (1-\sigma_{t+1})\pi(A_{t+1}|S_{t+1})\right]\left[G_{t:h} - \hat{q}(S_t, A_t, \mathbf{w}_{h-1})\right] + \bar{V}_{h-1}(S_{t+1})$$

where

$$\bar{V}_i(s) = \sum_a \pi(a|s)\hat{q}(s, \mathbf{w}_i)$$

and $G_{h:h} = \hat{q}_(S_h, A_h, \mathbf{w}_{h-1})$ if $h < T$ while if $h = T$ we have $G_{T-1:T} = R_T$ in the episodic case and $G_{T-1:T} = R_T - \bar{R}_{T-1}$ in the continuing case.

Note that in each case the value functions are defined with respect to the relevant episodic discounted or continuing average excess return.

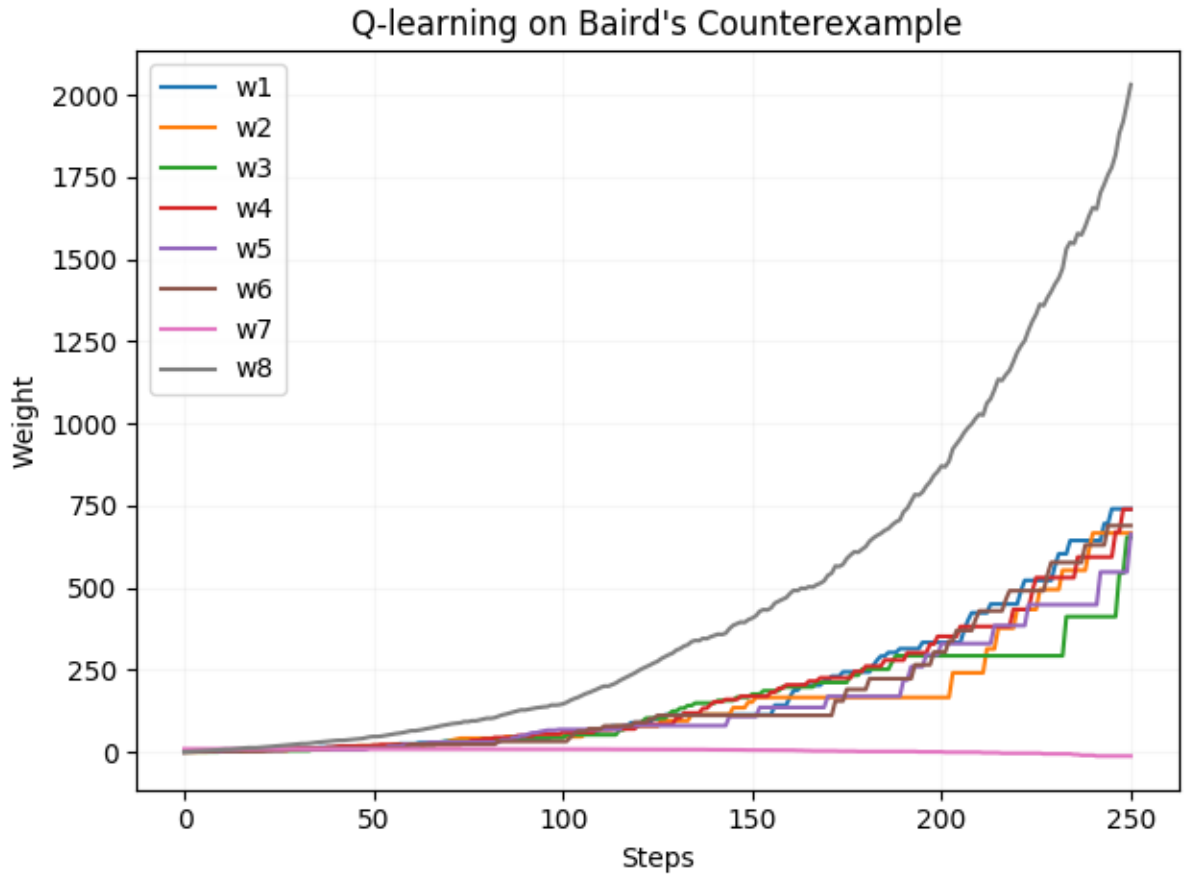
Exercise 11.3 (programming)

Q

Apply one-step semi-gradient Q-learning to Baird's counterexample and show empirically that its weights diverge.

A

This is a programming exercise. For the relevant code please see the repo.



Exercise 11.4

Q

Prove (11.24). Hint: Write the $\bar{\text{RE}}$ as an expectation over possible states s of the expectation of the squared error given that $S_t = s$. Then add and subtract the true value of state s from the error (before squaring), grouping the subtracted true value with the return and the added true value with

the estimated value. Then, if you expand the square, the most complex term will end up being zero, leaving you with (11.24).

A

Define

$$\overline{\mathrm{VE}}(\mathbf{w}) = \mathbb{E}_{s\sim\mu}[v_\pi(s) - \hat{v}(s,\mathbf{w})]$$

Now have the return error

$$\begin{aligned}
\overline{\mathrm{RE}} &\doteq \mathbb{E}\left[(G_t - \hat{v}(S_t,\mathbf{w}))^2\right] && (39)\\
&= \overline{\mathrm{VE}}(\mathbf{w}) + \mathbb{E}\left[(G_t - v_\pi(S_t))^2\right] + 2\mathbb{E}\left[(G_t - v_\pi(S_t))[v_\pi(S_t) - \hat{v}(S_t,\mathbf{w})]\right]. && (40)
\end{aligned}$$

The final term is

$$\begin{aligned}
\mathbb{E}\left[(G_t - v_\pi(S_t))[v_\pi(S_t) - \hat{v}(S_t,\mathbf{w})]\right] &= \mathbb{E}_{s\sim\mu}\left\{\mathbb{E}\left[(G_t - v_\pi(s))[v_\pi(s) - \hat{v}(s,\mathbf{w})]\right]|s\right\} && (41)\\
&= 0 && (42)
\end{aligned}$$