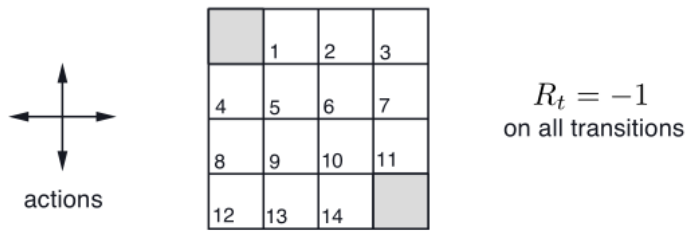
CSE – 426 Homework 10

Griffin Kent

1) In the example MDP in Figure 4.1 of the RL book, show how the policy evaluation algorithm calculates $v_\pi(1) = -1.7$ of the state 1, when $k = 2$.

Example 4.1 Consider the 4×4 gridworld shown below.



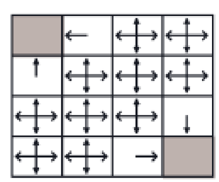
(Solution): In the policy evaluation algorithm, we will start by initializing all the states to 0 when $k = 0$. Then, using the Bellman Equation for updating the value of each state

$$v_{k+1}(s) \leftarrow \sum_a \pi(a|s) \sum_{s',r} P(s',r|s,a)[r + \gamma v_k(s')]$$

We can calculate the value of state $s = 1$ when $k = 1$. First, since the action space for this problem is defined as $A \triangleq \{up, down, left, right\}$, we can see that the probability of choosing any one of them is $\pi(a|s) = \frac{1}{4}$. Since any of these actions deterministically cause the corresponding state transitions, we can drop the $P(s',r|s,a)$ term. We will also take $\gamma = 1$ since all rewards are equally important. We are given the approximation of the state-value function for the random policy (with all actions equally likely) for $k = 1$ to be the following:

$k = 1$

0.0	-1.0	-1.0	-1.0
-1.0	-1.0	-1.0	-1.0
-1.0	-1.0	-1.0	-1.0
-1.0	-1.0	-1.0	0.0



Then, for $k = 2$, we have

$$v_\pi(1) = \frac{1}{4}\sum_a \sum_{s',r} P(s',r|s=1,a)[r + \gamma v_k(s')]$$

$$= \frac{1}{4}\Big[\Big(r + \gamma v_k\big(1(up)\big)\Big) + \Big(r + \gamma v_k\big(2(right)\big)\Big) + \Big(r + \gamma v_k\big(5(down)\big)\Big) + \Big(r + \gamma v_k\big(0(left)\big)\Big)\Big]$$

$$= \frac{1}{4}\big[\big(-1 + \gamma(-1)\big) + \big(-1 + \gamma(-1)\big) + \big(-1 + \gamma(-1)\big) + (-1 + \gamma \cdot 0)\big]$$

$$= \frac{1}{4}[(-2) + (-2) + (-2) - 1] \approx -1.7.$$

■

2) Use the Bellman equation for the action-value function $q_\pi(s,a)$ for a policy π to design a synchronous update formula for policy evaluation. Then use the big-O notation to find the time complexity of this update in one iteration.

[*Hints: Your formula must contain* $q_{k+1}(s,a)$ *and* $q_k(s,a)$.]

(Solution): The Bellman equation for the action-value function is defined as

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

Similar to the update formula for the state-value function, we can form the action-value function update formula as the following:

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

To determine the time complexity of this update for a single iteration of the policy evaluation algorithm, we need to know how many terms must be computed. From the derived equation above, we can see that the worst-case time complexity of updating a single action-value pair in big-O notation can be stated as

$$q_{k+1}(s,a) = O(s' \cdot r \cdot a).$$

Since a single iteration of the policy evaluation algorithm performs this update for all states $s \in S$ and $a \in A$, we can see that the time complexity for one iteration of the policy evaluation algorithm is

$$O((s'a)^s r).$$

■

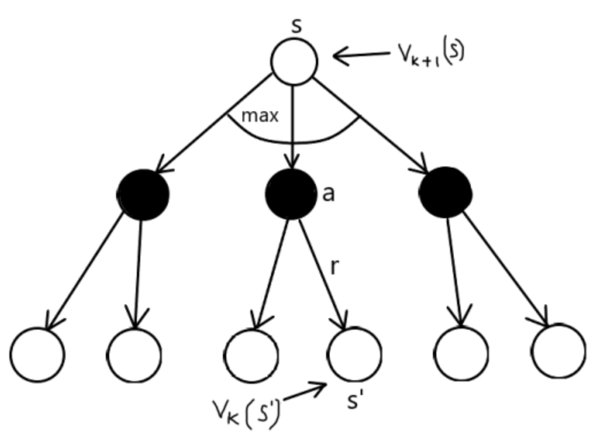
3) Draw the backup diagram for the synchronous value iteration equation in the lecture note, namely,

$$v_{k+1}(s) = \max_a \sum_{s',r} P(s',r|s,a)[r + \gamma v_k(s')].$$

Make sure you clearly mark the elements of the diagram with the max operator, the symbols s', r, $v_k(s')$, and $v_{k+1}(s)$.

[*Hints: Refer to Section 3.6 of the RL book for a close example.*]

(Solution):

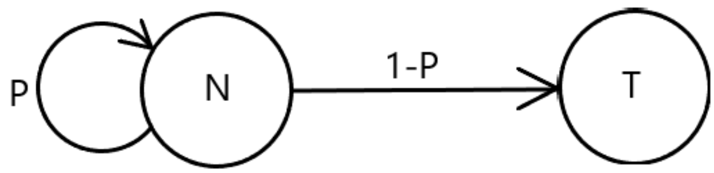


■

4) Exercise 5.5 of the RL book:

Exercise 5.5) Consider an MDP with a single nonterminal state and a single action that transitions back to the nonterminal state with probability p and transitions to the terminal state with probability $1-p$. Let the reward be $+1$ on all transitions and let $\gamma = 1$. Suppose you observe one episode that lasts 10 steps with a return of 10. What are the first-visit and every-visit estimators of the value of the nonterminal state? (Refer to page 92 of RL).

(Solution): To start, let's denote the terminal state as node T and the nonterminal state as node N. Then we can draw the MDP below:



The episode we are given can be written in the form $\{S_0, A_0, R_1, S_1, \ldots, S_{T-1}, A_{T-1}, R_T\}$ as the following:

$$E = \{N, P, 1, N, P, 1, N, P, 1, N, P, 1, N, P, 1, N, P, 1, N, P, 1, N, P, 1, N, P, 1, N, (1-P), 1\}$$

This is because, since there is only one nonterminal state, that state must have revisited itself 9 times before it moved to the terminal state.

Now, calculating the first-visit estimator $v_\pi(N)$ we will only count the first time that state N is visited: the returns of state N will be $Returns(N) = \{10\}$. Now taking the average of this single term will yield the first-visit estimator $v_\pi(N) = 10$.

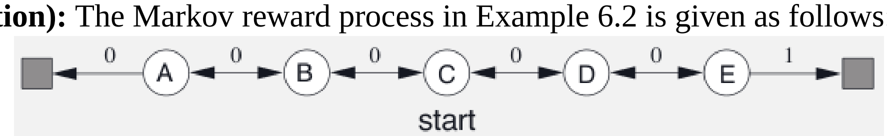
Now, calculating the every-visit estimator $v_\pi(N)$ will be similar to the process that we used for the first-visit estimator, however, we will count the returns for all of the visits to state N: the returns will then be $Returns(N) = \{1,2,3,4,5,6,7,8,9,10\}$. Taking the average of the returns will yield the every-visit estimator $v_\pi(N)$ =5.5.

■

5) (Graduate Only) Exercise 6.6 of the RL book. The exercise requires two ways to solve for the Bellman equation for the state-value function, and here we are asking for just one: use the vectorized Bellman equation in the lecture notes to solve a linear system for $v_\pi(s)$, $s \in \{A, B, C, D, E\}$.

Exercise 6.6) 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 needed? Why?

(Solution): The Markov reward process in Example 6.2 is given as follows:



Where transitions between states have equal probabilities (0.5 since there are only two possible transitions from any given state). Since this task is undiscounted, we have $\gamma = 1$. The vectorized Bellman equation is defined as

$$\boldsymbol{v}_\pi = \boldsymbol{r}_\pi + \gamma P_\pi \boldsymbol{v}_\pi$$

Where

$$\boldsymbol{r}_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} P(s', r|s, a)r,$$

$$P_\pi(s, s') = \sum_a \pi(a|s) \sum_r P(s', r|s, a).$$

We can then solve for the state-values $v_\pi(s)$ by solving the above linear system.
First, we can see that the matrix of state-to-state transition probabilities P_π will be defined as the following:

$$P_\pi = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \end{array} \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{bmatrix} 0 & 0.5 & 0 & 0 & 0 \\ 0.5 & 0 & 0.5 & 0 & 0 \\ 0 & 0.5 & 0 & 0.5 & 0 \\ 0 & 0 & 0.5 & 0 & 0.5 \\ 0 & 0 & 0 & 0.5 & 0 \end{bmatrix} \end{array}.$$

Similarly, we can see that the vector of expected immediate rewards $\boldsymbol{r}_\pi$ is:

$$\boldsymbol{r}_\pi = \begin{bmatrix} r_\pi(A) \\ r_\pi(B) \\ r_\pi(C) \\ r_\pi(D) \\ r_\pi(E) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0.5 \end{bmatrix}.$$

The Bellman linear system then becomes

$$\boldsymbol{v}_\pi(s) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0.5 \end{bmatrix} + \begin{bmatrix} 0 & 0.5 & 0 & 0 & 0 \\ 0.5 & 0 & 0.5 & 0 & 0 \\ 0 & 0.5 & 0 & 0.5 & 0 \\ 0 & 0 & 0.5 & 0 & 0.5 \\ 0 & 0 & 0 & 0.5 & 0 \end{bmatrix} \cdot \begin{bmatrix} v_\pi(A) \\ v_\pi(B) \\ v_\pi(C) \\ v_\pi(D) \\ v_\pi(E) \end{bmatrix}$$

Which yields

$$\begin{bmatrix} v_\pi(A) \\ v_\pi(B) \\ v_\pi(C) \\ v_\pi(D) \\ v_\pi(E) \end{bmatrix} = \begin{cases} 0.5v_\pi(B) \\ 0.5v_\pi(A) + 0.5v_\pi(C) \\ 0.5v_\pi(B) + 0.5v_\pi(D). \\ 0.5v_\pi(C) + 0.5v_\pi(E) \\ 0.5 + 0.5v_\pi(D) \end{cases}$$

Starting by substituting $v_\pi(A) = \frac{1}{2}v_\pi(B)$ into equation (2), we have

$$v_\pi(B) = 0.25v_\pi(B) + 0.5v_\pi(C)$$

$$\rightarrow v_\pi(B) = \frac{2}{3}v_\pi(C).$$

Continuing this chain of substitutions, we obtain the following:

$$v_\pi(C) = \frac{3}{4}v_\pi(D),$$

$$v_\pi(D) = \frac{4}{5}v_\pi(E),$$

$$v_\pi(E) = \frac{5}{6}.$$

Since we've finally obtained a scalar value for $v_\pi(E)$, we can now back-substitute to obtain the results:

$$v_\pi(E) = \frac{5}{6},$$
$$v_\pi(D) = \frac{4}{6},$$
$$v_\pi(C) = \frac{3}{6},$$
$$v_\pi(B) = \frac{2}{6},$$
$$v_\pi(A) = \frac{1}{6}.$$

Thus, we have obtained the true values for the states.

■