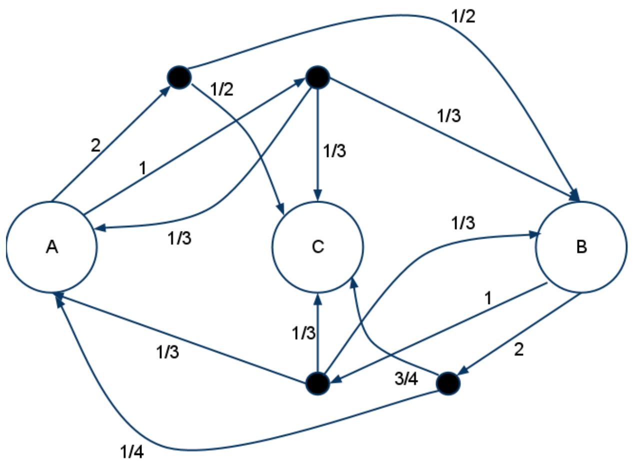
1 Policy Iteration (16 pts)

Consider the MDP below with states $X \in \{A, B, C\}$. State C is terminal. From states A and B one can take one of two actions $a \in \{1, 2\}$:

- Taking action 1 yields state A, B or C with equal probability 1/3.
- Taking action 2 from state A yields state B or C with equal probability 1/2.
- Taking action 2 from state B yields state A with probability $1/4$ and state C with probability $3/4$.

The reward function depends only on the starting state, and is $R(A) = R(B) = -1$. The discount is 1.



1. (8pts) Assume the initial policy is $\pi(A) = 1$, $\pi(B) = 2$. Perform two iterations of policy evaluation to determine the value functions $V_2^\pi(A)$ and $V_2^\pi(B)$.

   Answer: The policy evaluation equation is particularized to the reward function.

$$\begin{aligned} V_{i+1}^\pi(s) &= \sum_{s'} T(s,\pi(s),s')\left[R(s,\pi(s),s') + \gamma V_i^\pi(s')\right] \\ &= R(s) + \sum_{s'} T(s,\pi(s),s') V_i^\pi(s') \end{aligned}$$

   Applied to the different states, and noting that $V^\pi(C) = 0$,

$$\begin{aligned} V_{i+1}^\pi(A) &= R(A) + T(A,1,A)V_i^\pi(A) + T(A,1,B)V_i^\pi(B) + T(A,1,C)V^\pi(C) \\ &= -1 + \frac{1}{3}V_i^\pi(A) + \frac{1}{3}V_i^\pi(B) \\ V_{i+1}^\pi(B) &= R(B) + T(B,2,A)V_i^\pi(A) + T(B,2,C)V^\pi(C) \\ &= -1 + \frac{1}{4}V_i^\pi(A) \end{aligned}$$

The first iteration has initial values of zero.

$$\begin{aligned} V_1^\pi(A) &= -1 \\ V_1^\pi(B) &= -1 \end{aligned}$$

For the second iteration,

$$\begin{aligned} V_2^\pi(A) &= -1 + \frac{1}{3}V_1^\pi(A) + \frac{1}{3}V_1^\pi(B) \\ &= -1 - \frac{1}{3} - \frac{1}{3} = -5/3 \\ V_2^\pi(B) &= -1 + \frac{1}{4}V_1^\pi(A) \\ &= -1 - \frac{1}{4} = -5/4 \end{aligned}$$

2. (8pts) Next execute a policy improvement step to determine a better policy.

   Answer: The equation for policy extraction for $\gamma = 1$ is:

$$\pi_{i+1}(s) = \arg\max_a \sum_{s'} T(s,a,s')[R(s,a,s') + V_i^{\pi_i}(s')]$$

In state A, the two different actions 1 and 2 yield:

$$\begin{aligned} \pi_3(A) &= \arg\max_{a\in\{1,2\}} \begin{cases} R(A) + T(A,1,A)V_2^\pi(A) + T(A,1,B)V_2^\pi(B) + T(A,1,C)V^\pi(C) \\ R(A) + T(A,2,B)V_2^\pi(B) + T(A,2,C)V^\pi(C) \end{cases} \\ &= \arg\max_{a\in\{1,2\}} \begin{cases} -1 + (1/3)*(-5/3) + (1/3)*(-5/4) \\ -1 + (1/2)*(-5/4) \end{cases} \\ &= 2 \end{aligned}$$

In state B, the two different actions 1 and 2 yield:

$$\begin{aligned} \pi_3(B) &= \arg\max_{a\in\{1,2\}} \begin{cases} R(B) + T(B,1,A)V_2^\pi(A) + T(B,1,B)V_2^\pi(B) + T(B,1,C)V^\pi(C) \\ R(B) + T(B,2,A)V_2^\pi(A) + T(B,2,C)V^\pi(C) \end{cases} \\ &= \arg\max_{a\in\{1,2\}} \begin{cases} -1 + (1/3)*(-5/3) + (1/3)*(-5/4) \\ -1 + (1/4)*(-5/3) \end{cases} \\ &= 2 \end{aligned}$$

2 Q-Learning (16 pts)

For the previous MDP, you will perform Q-learning instead. The following episodes are observed, where each trial is of the form s, a, r, where s is the current state, a is the action from that state, and r is the reward. The destination state s' is on the next line.

A, 1, -1	B, 2, -1
A, 1, -1	B, 2, -1
B, 2, -1	A, 1, -1
C	C

1. (12 pts) Perform Q-learning using both episodes. The learning rate is 1/2, and the discount is still 1.

   Answer: The equation for Q-learning is

$$\begin{aligned} Q(s_t, a_t) &= (1-\alpha)Q(s_t, a_t) + \alpha(R(s_t, a_t, s_{t+1}) + \gamma \max_{a'} Q(s_{t+1}, a')) \\ &= \frac{1}{2}Q(s_t, a_t) + \frac{1}{2}(-1 + \max_{a'} Q(s_{t+1}, a')) \end{aligned}$$

   All Q-states are initialized to zero. Running through the trials,

$$\begin{aligned} Q(A,1) &= \frac{1}{2} * 0 + \frac{1}{2}(-1+0) = -\frac{1}{2} \\ Q(A,1) &= \frac{1}{2} * -\frac{1}{2} + \frac{1}{2}(-1+0) = -\frac{3}{4} \\ Q(B,2) &= \frac{1}{2} * 0 + \frac{1}{2}(-1+0) = -\frac{1}{2} \\ Q(B,2) &= \frac{1}{2} * -\frac{1}{2} + \frac{1}{2}(-1+0) = -\frac{3}{4} \\ Q(B,2) &= \frac{1}{2} * -\frac{3}{4} + \frac{1}{2}(-1+0) = -\frac{7}{8} \\ Q(A,1) &= \frac{1}{2} * -\frac{3}{4} + \frac{1}{2}(-1+0) = -\frac{7}{8} \end{aligned}$$

2. (4 pts) Perform policy extraction afterwards.

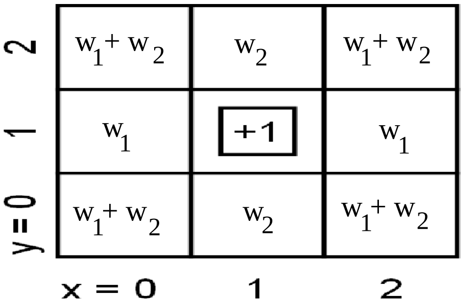
   Answer: The equation for policy extraction is:

$$\pi(s) = \arg\max_{a} Q(s, a)$$

   The Q-states not listed above are zero. Hence $\pi(A) = 2$ and $\pi(B) = 1$.

3 Functional Approximation (16 pts)

For the following gridworld problem, the agent can take the actions N, S, E, W, which move the agent one square in the respective directions. There is no noise, so these actions always take the agent in the direction attempted, unless into a wall in which case the agent stays put. The boxed +1 square is the terminal state. The reward for all transitions is zero, except the transition into the terminal state, which has reward +1. Assume a discount of 0.5.



	x = 0	1	2
2	$w_1 + w_2$	w_2	$w_1 + w_2$
1	w_1	$\boxed{+1}$	w_1
y = 0	$w_1 + w_2$	w_2	$w_1 + w_2$

Clearly there are a number of equivalent optimal policies. Suppose we represent the values by features $f_1(x, y) = (x - 1)^2$ and $f_2(x, y) = (y - 1)^2$. Solve for the weights w_1 and w_2 that yield an optimal policy.

The values according to the weights are placed in the cells above. Since actions are deterministic, the policy extraction equation is:

$$\pi(s) = \arg \max_{a \in \{N,S,E,w\}} R(s, a, s') + 0.5 * V(s')$$

For the squares directly adjacent to (1,1), the policy should move to (1,1). E.g.,

$$\begin{aligned} \pi(1,0) &= \arg\max \begin{cases} 1 & a = N \\ 0 + 0.5 * (w_1 + w_2) & a = E, W \\ 0 + 0.5 * w_2 & a = S \end{cases} \\ &= N \end{aligned}$$

This means $1 > 0.5 * (w_1 + w_2)$ and $1 > 0.5 * w_2$. The same holds for $\pi(1, 2)$. For the other squares adjacent to (1,1), e.g.,

$$\begin{aligned} \pi(0,1) &= \arg\max \begin{cases} 1 & a = E \\ 0 + 0.5 * (w_1 + w_2) & a = N, S \\ 0 + 0.5 * w_1 & a = W \end{cases} \\ &= E \end{aligned}$$

This means $1 > 0.5 * (w_1 + w_2)$ and $1 > 0.5 * w_1$. The same holds for $\pi(2, 1)$. For the remaining squares, they should just not move into the wall. E.g.,

$$\begin{aligned}\pi(0,0) &= \arg\max \begin{cases} 0 + 0.5 * w_1 & a = N \\ 0 + 0.5 * w_2 & a = E \\ 0 + 0.5 * (w_1 + w_2) & a = S, W \end{cases} \\ &= N \text{or} E\end{aligned}$$

This means $0.5 * w_1 > 0.5 * (w_1 + w_2)$ so that $0 > w_2$, or it means $0.5 * w_2 > 0.5 * (w_1 + w_2)$ so that $0 > w_1$. Either is optimal. The other diagonal states are similar.

$$\begin{aligned}\pi(0,2) &= \arg\max \begin{cases} 0 + 0.5 * w_1 & a = S \\ 0 + 0.5 * w_2 & a = E \\ 0 + 0.5 * (w_1 + w_2) & a = N, W \end{cases} \\ &= S \text{or} E\end{aligned}$$

This means $0.5 * w_1 > 0.5 * (w_1 + w_2)$ so that $0 > w_2$, or it means $0.5 * w_2 > 0.5 * (w_1 + w_2)$ so that $0 > w_1$. This is the same for all four diagonal squares.

Combining all results, this means $1 > 0.5 * (w_1 + w_2)$, $1 > 0.5w_1$, $1 > 0.5w_2$, and one of $w_1 < 0$ or $w_2 < 0$.