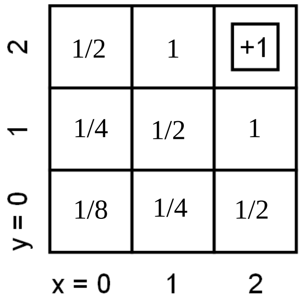
CS 6300 **HW06: Functional Approximation** **Due March 22, 2022**

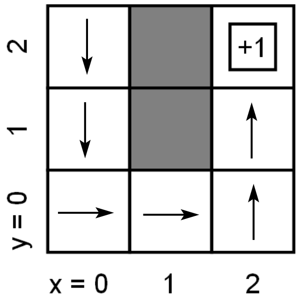
Please use the LaTeX template to produce your writeups. See the Homework Assignments page on the class website for details. Hand in through gradescope.

1 Functional Approximation

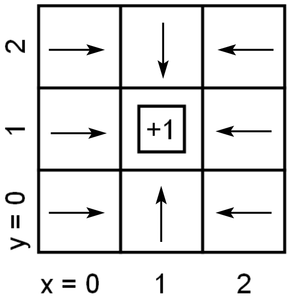
For the following gridworld problems, the agent can take the actions N, S, E, W, which move the agent one square in the respective directions. There is no noise, so these actions always take the agent in the direction attempted, unless that direction would lead off the grid or into a blocked (grey) square, in which case the action does nothing. The boxed +1 squares also permit the action X which causes the agent to exits the grid and enter the terminal state. The reward for all transitions are zero, except the exit transition, which has reward +1. Assume a discount of 0.5.



(A)



(B)



(C)

1. Fill in the optimal values for grid (A) (hint: this should require very little calculation).

See the diagram above. One way to derive the optimal values is as policy evaluation with an optimal policy (there are several different optimal policies due to symmetry). The value function is simplified for this problem.

$$\begin{aligned} V^*(s) &= \sum_{s'} T(s,\pi^*(s),s')\left[R(s,\pi^*(s),s') + \gamma V^*(s')\right] \\ &= R(s,\pi^*(s),s') + \frac{1}{2}V^*(s') \end{aligned}$$

Values propagate out from the terminal state.

$$\begin{aligned} V^*(1,2) &= R((1,2),\text{right},(2,2)) + \frac{1}{2}V^*(2,2) \\ &= 1 + \frac{1}{2}\cdot 0 = 1 \\ V^*(2,1) &= 1 \quad \text{similarly} \end{aligned}$$

The neighbors to these states have the value:

$$\begin{aligned}
V^*(0,2) &= R((0,2),\text{right},(1,2)) + \frac{1}{2}V^*(1,2) \\
&= 0 + \frac{1}{2}\cdot 1 = \frac{1}{2} \\
V^*(1,1) &= \frac{1}{2} \quad \text{similarly} \\
V^*(2,0) &= \frac{1}{2} \quad \text{similarly}
\end{aligned}$$

The neighbors to these states have the value:

$$\begin{aligned}
V^*(0,1) &= R((0,1),\text{up},(0,2)) + \frac{1}{2}V^*(0,2) \\
&= 0 + \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{4} \\
V^*(1,0) &= \frac{1}{4} \quad \text{similarly}
\end{aligned}$$

Finally,

$$\begin{aligned}
V^*(0,0) &= R((0,0),\text{up},(0,1)) + \frac{1}{2}V^*(0,1) \\
&= 0 + \frac{1}{2}\cdot\frac{1}{4} = \frac{1}{8}
\end{aligned}$$

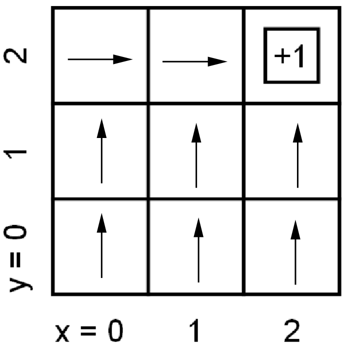
2. Specify the optimal policy for grid (B) and for grid (C) by placing an arrow in each empty square.

   See the figures above. Note that for (C) there is not a unique optimal policy. One of the optimal policies is shown.

   Imagine we have a set of real-valued features $f_i(s)$ for each non-terminal state $s = (x, y)$, and we wish to approximate the optimal utility values $V^*(s)$ by $V(s) = \sum_i w_i \cdot f_i(s)$ (linear feature-based approximation).

3. If our features are $f_1(x, y) = x$ and $f_2(x, y) = y$, give values of w_1 and w_2 for which a one-step look-ahead policy extracted from V will be optimal in grid (A).

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



From the feature function $V(s) = w_1 x + w_2 y$, the values of the state are shown explicitly in the left figure above by substituting particular grid integers x and y. We want that policy extraction yields one of the optimal policies, such as the right figure above.

$$\begin{aligned}\pi(s) &= \arg\max_a \sum_{s'} T(s,a,s')\left[R(s,a,s') + \gamma V(s')\right] \\ &= \arg\max_a \left[R(s,a,s') + \frac{1}{2}V(s')\right]\end{aligned}$$

Considering state (0,0), any action into one of the walls results in the same state with a value 0. To move out of the state, then $w_1 > 0$ or $w_2 > 0$. State s' will be determined by whichever of w_1 and w_2 is larger.

$$\begin{aligned}\pi(0,0) &= \arg\max_a \left[0 + \frac{1}{2}V(s')\right] \\ &= \arg\max_a \begin{cases} \frac{1}{2}V(0,1) = \frac{1}{2}w_2 & a = \text{up} \\ \frac{1}{2}V(1,0) = \frac{1}{2}w_1 & a = \text{right} \end{cases}\end{aligned}$$

For example, if $w_2 > w_1$, then $a =$ up and $s' = (0,1)$. From state (0,1), to continue with the policy in the right diagram, then $2w_2 > w_1 + w_2$, or $w_2 > w_1$ which is the same condition as before. Reaching state (0,2), it is required that $w_1 + 2w_2 > 2w_2$, or $w_1 > 0$. Flipping the magnitudes such that $w_1 > w_2 > 0$ results in a different policy going along the x-axis edge. The final answer is that $w_1 > 0$ and $w_2 > 0$.

4. Can we represent the actual optimal values V^* for grid (A) using these two features? Why or why not?

   No, because square (0,0) will always have a value of zero.

5. For each of the feature sets listed below, state which (if any) of the grid MDPs above can be 'solved', in the sense that we can express some (possibly non-optimal) values which produce optimal one-step look-ahead policies.

(a) $f_1(x, y) = x$ and $f_2(x, y) = y$.

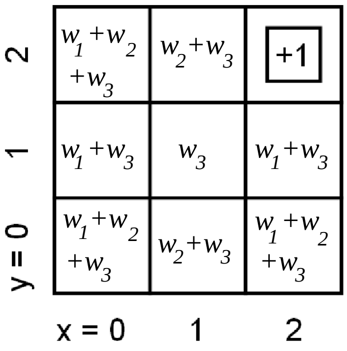
For (A), it works as demonstrated above. For (B) and (C), the feature values are the same as in the left figure above.

For (B), to move from (0,2) to (0,1) we must have $2w_2 < w_2$, or $w_2 < 0$. To move from (2,0) to (2,1), we must have $w_2 + 2w_1 > 2w_1$, or $w_2 > 0$, a contradiction. (C) is similar.

(b) For each (i, j), a feature $f_{i,j}(x, y) = 1$ if $(x, y) = (i, j)$, 0 otherwise.

For all three, set $w_{i,j} = V^*(i, j)$, the optimal value function.

(c) $f_1(x, y) = (x - 1)^2$, $f_2(x, y) = (y - 1)^2$, and $f_3(x, y) = 1$.



	x = 0	1	2
2	w_1+w_2 $+w_3$	w_2+w_3	+1
1	w_1+w_3	w_3	w_1+w_3
y = 0	w_1+w_2 $+w_3$	w_2+w_3	w_1+w_2 $+w_3$

For (A), the feature values are shown above. They will be the same for (B) except for its blacked-out squares. States (0,0), (0,2), and (2,0) have the same value. So do states (0,1) and (2,1), as well as states (1,0) and (1,2). (A) and (B) fail because states along the optimal paths cannot be distinguished.

For (C), state (2,2) has the value $w_1 + w_2 + w_3$. w_3 is a bias added in all the squares, and hence its value is arbitrary. For simplification, suppose $w_3 = 0$. Then $w_1 < 0$ and $w_2 < 0$ from squares (0,1)/(2,1) and (1,0)/(1,3), and $w_1 + w_2 < 0$ for squares (0,0)/(0,2)/(2,0). The final answer is $w_2 < 0$ and $w_1 < 0$.