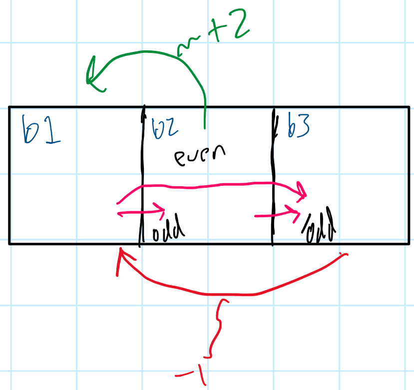
Hw2

Thursday, February 6, 2025 7:21 PM

1 a) The reward function represents the expected reward received from executing an action from a given state.
The state-action function Q is similar, but represents the expected reward from starting at a given state, and taking an action, and continuing with a greedy policy with respect to Q. So, Q represent expected total reward throughout a process, and the reward function only represents the reward for one step.

b) $$Q(s,a) = R(s,a) + \gamma \sum_{s'} T(s' \mid s,a) V(s')$$

2)



a) $S = \{b_1, b_2, b_3\}$

$$A = \{Roll, Reset\}$$

assuming fair die

$$T_{roll} = \begin{array}{c|ccc} & b_1 & b_2 & b_3 \\ \hline b_1 & 0 & 0.5 & 0.5 \\ b_2 & 0 & 0 & 1 \\ b_3 & 0 & 0 & 1 \end{array} \qquad T_{reset} = \begin{array}{c|ccc} & b_1 & b_2 & b_3 \\ \hline b_1 & 1 & 0 & 0 \\ b_2 & 1 & 0 & 0 \\ b_3 & 1 & 0 & 0 \end{array}$$

$$T_{roll,unfair} = \begin{bmatrix} 0 & p & 1-p \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

$$R(s,a,s') = \begin{cases} +2 & \text{if } a = \text{reset and } s = b_2 \\ -1 & \text{if } a = \text{reset and } s = b_3 \end{cases}$$

$$R = \begin{array}{cl} \begin{bmatrix} 0 \\ +2 \\ -1 \end{bmatrix} & \begin{array}{l} b_1 \\ \text{reset } b_2 \\ \text{reset } b_3 \end{array} \end{array}$$

b)

$$U_0 = \begin{array}{cl} \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} & \begin{array}{l} b_1 \\ b_2 \\ b_3 \end{array} \end{array}$$

☆ assuming $p = 0.5$ for this problem

Guessing $\pi = \begin{bmatrix} \text{roll} \\ \text{reset} \\ \text{reset} \end{bmatrix}$

$$U^\pi(s) = R(s, \pi(s)) + \gamma \sum_{s'} T(s, \pi(s), s') U^\pi(s)$$

$$U^\pi(b_1) = 0 + (p \cdot U^\pi(b_2) + (1-p) U^\pi(b_3))$$

$$U^\pi(b_2) = 2 + \gamma(U^\pi(b_1)) \quad \rightarrow$$

$U^\pi(b_2) = 2 + \gamma(U^\pi(b_1)) \rightarrow$

$U^\pi(b_3) = -1 + \gamma(U^\pi(b_1)) \rightarrow$

$\rightarrow b_1 = (\rho(2+\gamma b_1) + (1-\rho)(-1+\gamma b_1))\cdot\gamma$

$b_1 = (2\rho + \gamma\rho b_1 + \gamma b_1 - 1 + \rho - \gamma\rho b_1)\cdot\gamma$

$b_1 = 2\rho\gamma + \cancel{\gamma^2\rho b_1} + \gamma^2 b - \gamma + \rho\gamma - \cancel{\gamma^2\rho b_1}$

$b_1 - \gamma^2 b_1 = 2\rho\gamma - \gamma + \rho\gamma$

$b_1 = \dfrac{3\rho\gamma - \gamma}{1-\gamma^2} \xrightarrow[\gamma=0.95]{\rho=0.5} b_1 = 4.872$

$U^\pi(b_1) = 4.872$

plug in $\rightarrow U^\pi(b_2) = 6.63$

$U^\pi(b_3) = 3.63$

Check that these values satisfy $U^*(s) = \max_a \left(R(s,a) + \gamma\sum_{s'} T(s'|s,a)\,U^*(s')\right)$

$$b_1 \rightarrow \left.\begin{array}{l} \text{roll: } 0 + 0.95(0.5(6.63) + 0.5(3.63)) = 4.874 \\ \text{reset: } 0 + 0.95(4.872) = \text{less than } 4.874 \end{array}\right\} \max = 4.874$$

$$b_2 \rightarrow \left.\begin{array}{l} \text{roll: } 0 + 0.95(3.63) = \text{less than } 3.63 \\ \text{reset: } 2 + 0.95(4.872) = 6.63 \end{array}\right\} \max = 6.63$$

$b_3 \rightarrow$ roll: $0 + 0.95(3.63) =$ less than 3.63
reset: $-1 + 0.95(4.872) = 3.63$ $\Big\}$ max $= 3.63$

- All of these values are the same, except for s_1. This optimal value is very close though, likely resulting from a rounding error. So, the policy $\pi = \begin{Bmatrix} \text{roll} \\ \text{reset} \\ \text{reset} \end{Bmatrix}$ is optimal

c) I will use the same policy as (b) because in the case of all evens, reward will be minimized by repeatedly reaching b_3 and resetting. Similarly, with all odd rolls reward will be maximized by resetting only on b_2, with +2 reward. I am able to compute bounds by using the expression found earlier for $U^\pi(b_1)$ and varying the probability (p) term.

$$U^\pi(b_1) = \frac{3p\gamma - \gamma}{1-\gamma^2}$$

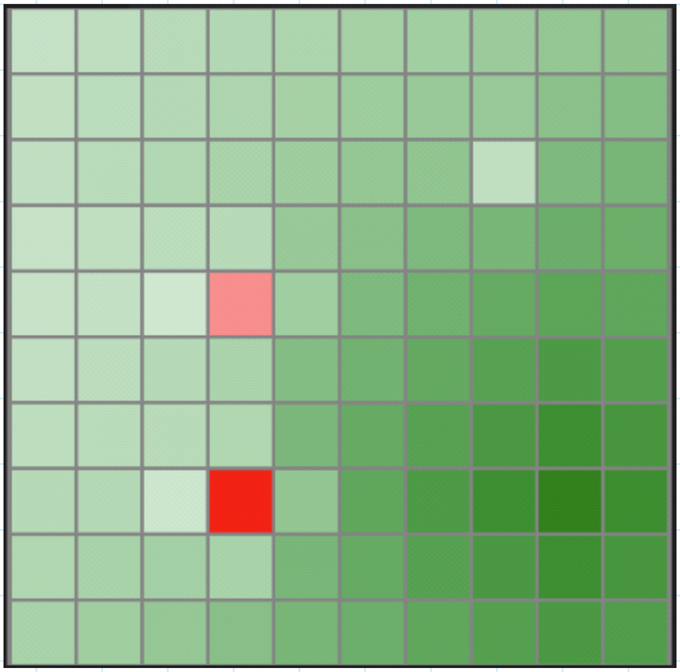
, where $p=0$ simulates all-even rolls, and $p=1$ simulates all-odd rolls.

$$U^\pi(b_1)\Big|_{\substack{p=0 \\ \gamma=0.95}} = -9.744$$

$$U^\pi(b_1)\Big|_{\substack{p=1 \\ \gamma=0.95}} = 19.487$$

So, discounted score $-9.744 \le U(b_1) \le 19.487$

3)



4) Passed autograder with n = 7