Home Assignment 1. Roman Myronenko

1 1 Efficient Routing MDP

(a)

$r_s \in \{-5, -0.5, 0, 2\}$

$r_s = -5$.

The best policy will be -10. It will be clearly seen if we move Right Up to the red square 7 and terminate the game with the reward of -10. Any other strategy will lead to a worse result.

$r_s = 0.5$.

The best policy is +2.5, because if we choose the aim of reaching green square 33 with 5 steps, the outcome will be +2.5. There is a worse result in any other strategy.

$r_s = 0$.

There is no negative reward for the regular move, but the optimal policy is still with the shortest path to 33.

$r_s = 2$.

In this case, the optimal policy depends on the discount factor γ. If the discount factor γ *is* small enough, it would be best to take the shortest path. But if γ is close to 1, then it would be better to prolong the journey and receive the award for it.

(b)

$r_s \in \{-0.5, 0\}$,

In this case, the optimal policy will give the shortest path to the green square.

Let $r_s = -0.5$, then: $v_\pi(2) = r_s(1 + \gamma + \gamma^2 + \gamma^3 + \gamma^4) + r_g\gamma^5 = -0.5(1 + 0.9 + 0.9^2 + 0.9^3 + 0.9^4) + 5 \cdot 0.9^5 = 0.9049$

$v_\pi(13) = -5$ $v_\pi(21) = -0.5(1 + 0.9) +$

$5 \cdot 0.9^2 = 3.1$ $v_\pi(32) = r_s + r_r\gamma = -0.5 -$

$5 \cdot 0.9 = -5$

(c)

$r_e = -5$.

If we move always right, then after 2 moves we meet barrier 14 and finish the game. In this case find the policy, which terminates the game as fast as possible.

$r_e = -0.5$.

The optimal policy - the one with the shortest path. It is: move down to the middle lane, then move right till the green square (2 - 3 - 9 - 15 - 21 - 27 - 33).

$r_e = 0$.

We still need the shortest path because of the discount factor. The same policy as in previous example.

$r_e = 2$.

In this case we will search for the longest path. If we move from 2 down to 4, then right. This path will have 2 more moves, and it's the longest path possible. In this case we depend on γ.

(d)

Based on the (a), the optimal path from state 2 have the value:

$$v_i(2) = r_s \left(\sum_{k=0}^{4} \gamma^k \right) + 5\gamma^5 = 5\gamma^5$$

Based on (c), there should be exactly 6 moves to get to the green square, so:

$$v_e(2) = r_e \left(\sum_{k=0}^{5} \gamma^k \right) + 5\gamma^6$$

The optimal path using efficient actions will be strictly more rewarding if:

$$v_e(2) > v_i(2)$$

$$r_e \left(\sum_{k=0}^{5} \gamma^k \right) + 5\gamma^6 > 5\gamma^5$$

$$r_e > \frac{5(\gamma^5 - \gamma^6)}{\sum_{k=0}^{5} \gamma^k}$$

For $\gamma = 0.9$ it will be approximately $r_e > 0.0630$.

(e)

There are some exceptions with the possibility of states in reaching the green square.

- Using only *efficient* actions - states {5,17}.
- Using only *inefficient* actions - only state 33.

(f)

By the definition of the value function for the policy π:

$$v_\pi(s) = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s \right]$$

Let's add constant *c* to each reward R_t:

$$\begin{aligned}
(v_\pi)_{\text{new}}(s) &= \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k \left(R_{t+k+1} + c \right) \middle| S_t = s \right] \\
&= \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} + c \sum_{k=0}^{\infty} \gamma^k \middle| S_t = s \right] \\
&= \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \middle| S_t = s \right] + c \sum_{k=0}^{\infty} \gamma^k
\end{aligned}$$

Taking into account that $\sum_{k=0}^{\infty} \gamma^k = \frac{\gamma}{1-\gamma}$, we can see, that each value function will increase on the same constant $\frac{c\gamma}{1-\gamma}$, regardless of the state.

Conclusion: Adding the constant to each reward is not going to change the optimal policy.

(g)

Based on the discussion in Quora, we can seva fuel by:

- avoiding stops or slow-downs
- moving at the speed, that is optimal for engine productivity

So, for the most sustainable route I would propose the rules: • Avoid cities. The biggest city - the largest is negative reward for the route. City has a lot of turns, stops, speed limit zones. All of these would require us to change the speed frequently, which will cause huge fuel consumption.

- Negative reward for the turns and changes of direction. Usually, we slow down before the turn. The larger the angle of the turn - the bigger penalty.
- Highways are the best for sustainable driving - give them high positive reward.
- Penalize traffic jam situations. In the modern maps we can receive the information about traffic jams in a real time.