Home Assignment 1

Pavlo Bilinskyi

April 2023

0.1 1 Efficient Routing MDP

0.1.1 (a)

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

Let's start with $r_s = -5$**.**

Clearly, the best policy will be to 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 outcome because the only positive reward can be received on green square 33, but the car must make at least 5 moves to get there, and the reward will be -20 in this case, which still is worse than our initial score -10.

Let's consider the case $r_s = 0.5$**.** Now when the moves don't have such a great penalty, it's better to get to the green square and receive the positive reward.

But the path must be the shortest possible because each move is still penalized.

In this case, the optimal policy will be the one which gives *the shortest path* to the green square 33 and avoids the barriers.

Why is that so? Because there must be at least 5 steps on grey squares before the car will reaches the green square. Therefore, in any case, we will get the inevitable penalty $5r_s$.

There exists a policy $\pi(s)$, which will yield the reward $5r_s + r_g = -2.5 + 5 = 2.5$, and this is the best outcome. Let's define it:

$$\pi(s) = \begin{cases} \text{RD}, & s = 3k - 1 \\ \text{RD}, & s = 3k \\ \text{RU}, & s = 3k + 1 \\ \text{RU}, & s = 3k + 2 \end{cases} \tag{1}$$

where $k = 1, 3, 5, \ldots, 11$, and RD, RU are the notations for the actions "move Right Down" and "move Right Up".

Starting from state 2, we will get the trajectory:

$$2, \text{RD}, -0.5, 9, \text{RD}, -0.5, 16, \text{RU}, -0.5, 21, \text{RD}, -0.5, 28, \text{RU}, -0.5, 33, \text{RD}, +5, STOP$$

which will yield the maximal reward 2.5.

Let's consider the case $r_s = 0$.

Now there is no negative reward for the regular move, but the optimal policy is still with the shortest path to 33. The reason is the *discount factor* γ, and the longest the path to 33, the less reward will be received.

So, the policy $\pi(s)$ from the previous example still remains the optimal policy as long as $\gamma < 1$.

Let's consider the case $r_s = 2$.

Now the optimal policy depends on the discount factor γ. If it's 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 (+2 for each additional grey square).

In our case $\gamma = 0.9$, and the longest path is the best.

Uniqueness of the optimal policy

If $r_s = -5$, the optimal policy is to move to 7, and it's unique (only one action).

If $r_s = -0.5$, the optimal policy is not unique. There are a few policies that have the shortest path to 33. For example, we could move $21 \to 26 \to 33$ or $21 \to 28 \to 33$, both ways leading from 21 to 33 with an equal outcome.

If $r_s = 0$ - the same as in the previous case.

If $r_s = 2$ there is only one policy with the longest possible path - unique.

Dependence on γ

The policy doesn't depend on γ case when $r_s \in \{-5, -0.5\}$.

In the case of $r_s = 0$ it only matters whether $\gamma < 1$ or not.

In the case of $r_s = 2$ it depends on the exact value of γ.

0.1.2 (b)

For the values $r_s \in \{-0.5, 0\}$, the optimal policy will give the shortest path to the green square (for example, $\pi(\cdot)$, defined in 1. In other cases - no.

It was already explained above. Let $r_s = -0.5$, then:

$$v_\pi(2) = r_s(1 + \gamma + \gamma^2 + \gamma^3 + \gamma^4) + r_g\gamma^5 = -0.5(1 + 0.9 + 0.9^2 + 0.9^3 + 0.9^4) + 5 \cdot 0.9^5 = 0.9049$$
$$v_\pi(13) = -5$$
$$v_\pi(21) = -0.5(1 + 0.9) + 5 \cdot 0.9^2 = 3.1$$
$$v_\pi(32) = r_s + r_r\gamma = -0.5 - 5 \cdot 0.9 = -5$$

0.1.3 (c)

Using the same logic as in (a).

$r_e = -5$. Like in (a), we should find the policy, which terminates the game as fast as possible. It is: move always right. After 2 moves we meet barrier 14 and finish the game.

$r_e = -0.5$. As in (a), 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.

$r_e = 2$. Now we will search for the longest path: 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.

0.1.4 (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:

$$
\begin{aligned}
v_e(2) &> v_i(2) \\
r_e \left(\sum_{k=0}^{5} \gamma^k \right) + 5\gamma^6 &> 5\gamma^5 \\
r_e &> \frac{5(\gamma^5 - \gamma^6)}{\sum_{k=0}^{5} \gamma^k}
\end{aligned}
$$

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

0.1.5 (e)

Almost all states have possibility to reach green square with using either efficient or inefficient actions exclusively. But there are exceptions.

Using only *efficient* actions - states $\{5, 17\}$. Because there are barriers from the right side and the bottom, and any move will terminate the game.

Using only *inefficient* actions - only state 33. This is the only state, where any move will lead to the barrier.

0.1.6 (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: No. Adding the constant to each reward is not going to change the optimal policy.

0.1.7 (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.

0.2 2 Value Iteration Theorem

0.2.1 (a)

The proof is analogical to the proof for the backup operator B:

$$\begin{aligned}
\|B_\pi V - B_\pi V'\| &= \left\| \mathbf{E}_\pi \left[R(s,a) + \gamma \sum_{s' \in S} p\left(s'|s,a\right) V(s') \right] - \mathbf{E}_\pi \left[R(s,a) + \gamma \sum_{s' \in S} p\left(s'|s,a\right) V'(s') \right] \right\| \\
&= \left\| \mathbf{E}_\pi \left[\gamma \sum_{s' \in S} p\left(s'|s,a\right) \left[V(s') - V'(s') \right] \right] \right\| \\
&= \gamma \left\| \sum_{a \in A} \pi(a|s) \sum_{s' \in S} p\left(s'|s,a\right) \left[V(s') - V'(s') \right] \right\| \\
&\leq \gamma \left\| \sum_{a \in A} \pi(a|s) \cdot \|V - V'\| \sum_{s' \in S} p\left(s'|s,a\right) \right\| \\
&= \gamma \|V - V'\|
\end{aligned}$$

0.2.2 (b)

Idea 1 how to prove uniqueness: By the Banach fixed point theorem, contraction operator has a unique fixed point.

Idea 2 how to prove uniqueness: From the contrary, suppose that there are two fixed points: V and V', such that $B_\pi V = V$ and $B_\pi V' = V'$. Using the contraction property:

$$\|B_\pi V - B_\pi V'\| \leq \gamma \|V - V'\|$$

Using fixed point property, we have:

$$\|V - V'\| \leq \gamma \|V - V'\|$$

Given $0 \leq \gamma < 1$, the only possible case is $\|V - V'\| = 0$, which contradicts the assumption that V and v' - different fixed points. Q.E.D.

The fixed point of B_π is the **value function** v_π, correnspondent to the policy π. It follows from the Bellman equation for the value function.

0.2.3 (c)

No, V_π and V are not necessarily equal.

The equality $V_\pi = V$ implies that V is the optimal value function of the MDP (based on the Bellman optimality condition).

If V hasn't yet converged to V^*, they are not equal: $V \neq V_\pi$.

0.2.4 (d)

Both operators are equal: $BV = B_\pi V$.

Greedy policy is deterministic, in that case:

$$B_\pi V = R(s, \pi(s)) + \gamma \sum_{s' \in S} p(s'|s, \pi(s)) V(s')$$

However, action $a = \pi(s)$ was determined by maximizing the quantity $R(s,a) + \gamma \sum_{s' \in S} p(s'|s,a) V(s')$ over $a \in A$. So, we can rewrite:

$$B_\pi V = \max_{a \in A} \left[R(s,a) + \gamma \sum_{s' \in S} p(s'|s,a) V(s') \right] = BV$$

and this shows, that B_π and B are identical operators in the case of greedy policy π.

0.2.5 (e)

Skip.

0.2.6 (f)

The estimation $\|B^k V - B^k V'\| \leq \gamma^k \|V - V'\|$ can be proven by continuously applying the contraction property:

$$\begin{aligned}
\|B^k V - B^k V'\| &= \|B(B^{k-1}V) - B(B^{k-1}V')\| \leq \\
&\leq \gamma \|B^{k-1}V - B^{k-1}V'\| = \gamma \|B(B^{k-2}V) - B(B^{k-2}V')\| \leq \\
&\leq \gamma^2 \|B^{k-2}V - B^{k-2}V'\| \leq \\
&\leq \ldots \leq \\
&\leq \gamma^k \|V - V'\|
\end{aligned}$$

0.2.7 (g)

Skip.

0.2.8 (h)

Skip.