RL Exercise 1

Hadar Sharvit

April 19, 2022

1 Efficient Routing MDP

$\gamma = 0.9$, $r_g = +5$, $r_r = -5$, $start\ square = 2$, $end\ square = 33$, $Horizon = \infty$

1.1 Question A

- $r_s = -5$: the path of optimal policy is $2 \to 9 \to 16 \to 16 \to 21 \to 26$ or $28 \to 33$, as every grey square loses us point, so we wish to minimize the number of steps.
- $r_s = 0$: path would be the same, as the same logic applies.
- $r_s = 0$: as long as the path is finished in 33, any path is optimal, but the only options are the ones in the previous bullet point.
- $r_s = 2$: now every step gives us reward, so we wish to maximize the number of steps taken. in our specific case we still follow the same path, because otherwise we enter a red block

generally speaking the optimal policy does depend on γ. for example if $\gamma = 0$ we only care about the first reward, and in that case a maximal initial reward would be the one that governs the optimal policy.

1.2 Question B

I think that all of them yields the same path, i.e the shortest. taking $r_s = 0$, we now calculate the optimal value function using Bellman backup:

$$V_0^\pi(s) = (0, 0, ..., 0), \text{ and } \forall s \in S : V_k^\pi(s) = R(s, \pi(s)) + \gamma \sum_{s' \in S} P(s'|s, \pi(s)) V_{k-1}^\pi(s')$$

This is for π that is optimal. we will represent V as a matrix, corresponding to the 2D Grid World. Starting with V_1

$$\forall s \in S : V_1(s) = R(s, \pi(s)) + \gamma \sum_{s' \in S} 0 \to V_1 = \begin{bmatrix} -5 & -5 & -5 & -5 & -5 & -5 \\ 0 & 0 & -5 & 0 & 0 & 0 \\ 0 & 0 & -5 & 0 & 0 & +5 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -5 & 0 & -5 & -5 & 0 \\ -5 & -5 & -5 & -5 & -5 & -5 \end{bmatrix}$$

now for V_2

$$\forall s \in S : V_2(s) = R(s, \pi(s)) + \gamma \sum_{s' \in S} P(s'|s, \pi(s)) V_1^\pi(s')$$

$$V_2 = \begin{bmatrix} -5 & -5 & -5 & -5 & -5 & -5 \\ 0 & 0 & -5 & 0 & 0 & 0 \\ 0 & 0 & -5 & 0 & 0 & +5 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -5 & 0 & -5 & -5 & 0 \\ -5 & -5 & -5 & -5 & -5 & -5 \end{bmatrix}$$

2 Value Iteration Theorem

2.1 Question A

we will prove that $(B_\pi V)(s) = \mathbf{E}_{a\sim\pi}[R(s,a) + \gamma \sum_{s'\in S} p(s'|s,a)V(s')]$ is a contraction mapping:
denoting R_a and P_a as a reward vector and probability matrix for the action a, we can write

$$(B_\pi V)(s) = \mathbf{E}_{a\sim\pi}[R_a + \gamma P_a V] \tag{1}$$

and therefore,

$$\begin{aligned} ||B_\pi V_1 - B_\pi V_2||_\infty &= ||\mathbf{E}_{a\sim\pi}[R_a + \gamma P_a V_1] - \mathbf{E}_{a\sim\pi}[R_a + \gamma P_a V_2]||_\infty \\ &= ||\mathbf{E}_{a\sim\pi}[R_a + \gamma P_a V_1 - R_a - \gamma P_a V_2]||_\infty \\ &= \gamma||\mathbf{E}_{a\sim\pi}[P_a V_1 - P_a V_2]||_\infty \\ &= \gamma||\mathbf{E}_{a\sim\pi}[P_a(V_1 - V_2)]||_\infty \end{aligned} \tag{2}$$

let A be the action a for which the expectation is maximal

$$\begin{aligned} \gamma||\mathbf{E}_{a\sim\pi}[P_a(V_1 - V_2)]||_\infty &\le \gamma||P_A(V_1 - V_2)||_\infty \\ &\le \gamma||P_A||_\infty||V_1 - V_2||_\infty \\ &= \gamma||V_1 - V_2||_\infty \end{aligned} \tag{3}$$

which means $B_\pi V$ is a contraction mapping

2.2 Question B

As we've seen in class, V_π is a fixed point of B_π, i.e $(B_\pi V_\pi)(s) = V_\pi(s)$. Furthermore, it is a unique fixed point from the contraction mapping theorem (as B_π is a contraction mapping)

2.3 Question C

Strictly discussing definitions, we know that $V^\pi = R^\pi(s) + \gamma \sum_{s'\in S} P^\pi(s'|s)V^\pi(s')$ where R^π and P^π are the expected reward and probability when $a \sim \pi$. this means that $V^\pi = V$ only in terms of expectation.

2.4 Questions D→H combined

$$\begin{aligned} ||V^\pi - V^*|| &= ||V^\pi - V_{n+1} + V_{n+1} - V^*|| \\ &\le ||V^\pi - V_{n+1}|| + ||V_{n+1} - V^*|| \end{aligned} \tag{4}$$

focusing on the first term

$$\begin{aligned} ||V^\pi - V_{n+1}|| &= ||B_\pi V^\pi - V_{n+1}|| \\ &= ||B_\pi V^\pi - BV_{n+1} + BV_{n+1} - V_{n+1}|| \\ &\le ||B_\pi V^\pi - BV_{n+1}|| + ||BV_{n+1} - V_{n+1}|| \end{aligned} \tag{5}$$

Where the first transition is because V^π is a fixed point of B_π. Now, since π is maximal over the actions using V_{n+1}, we know that $B_\pi V_{n+1} = BV_{n+1}$. we can also use the fact that $V_{n+1} = BV_n$:

$$\begin{aligned} ||V^\pi - V_{n+1}|| &\le ||B_\pi V^\pi - B_\pi V_{n+1}|| + ||BV_{n+1} - BV_n|| \\ &\le \gamma||V^\pi - V_{n+1}|| + \gamma||V_{n+1} - V_n|| \end{aligned} \tag{6}$$

where the last inequality is because B_π and B are contraction mappings. this gives

$$||V^\pi - V_{n+1}|| \le \frac{\gamma}{1-\gamma}||V_{n+1} - V_n|| \tag{7}$$

we similarly repeat the process for the second term

$$\begin{aligned} ||V^* - V_{n+1}|| &\le ||V^* - V_{n+2}|| + ||V_{n+2} - V_{n+1}|| \\ &= ||BV^* - BV_{n+1}|| + ||BV_{n+1} - BV_n|| \\ &\le \gamma||V^* - V_{n+1}|| + \gamma||V_{n+1} - V_n|| \end{aligned} \tag{8}$$

which yields

$$||V^* - V_{n+1}|| \leq \frac{\gamma}{1-\gamma}||V_{n+1} - V_n|| \tag{9}$$

we now use the fact that $||V_{n+1} - V_n|| \leq \epsilon(1-\gamma)/2\gamma$ in eq. (4)

$$\begin{aligned} ||V^\pi - V^*|| &\leq \frac{2\gamma}{1-\gamma}||V_{n+1} - V_n|| \\ &\leq \frac{2\gamma}{1-\gamma} \cdot \frac{1-\gamma}{2\gamma}\epsilon \\ &= \epsilon \end{aligned} \tag{10}$$

i.e, π is an ϵ-optimal policy, as needed

3 Frozen Lake MDP

3.1 Question A and B

code - see under vi_and_pi.py

3.2 Question C

As expected, when the dynamics of the world are stochastic, the number of iterations required increases, and the policy changes much more frequently until it finally converges