Problem 1

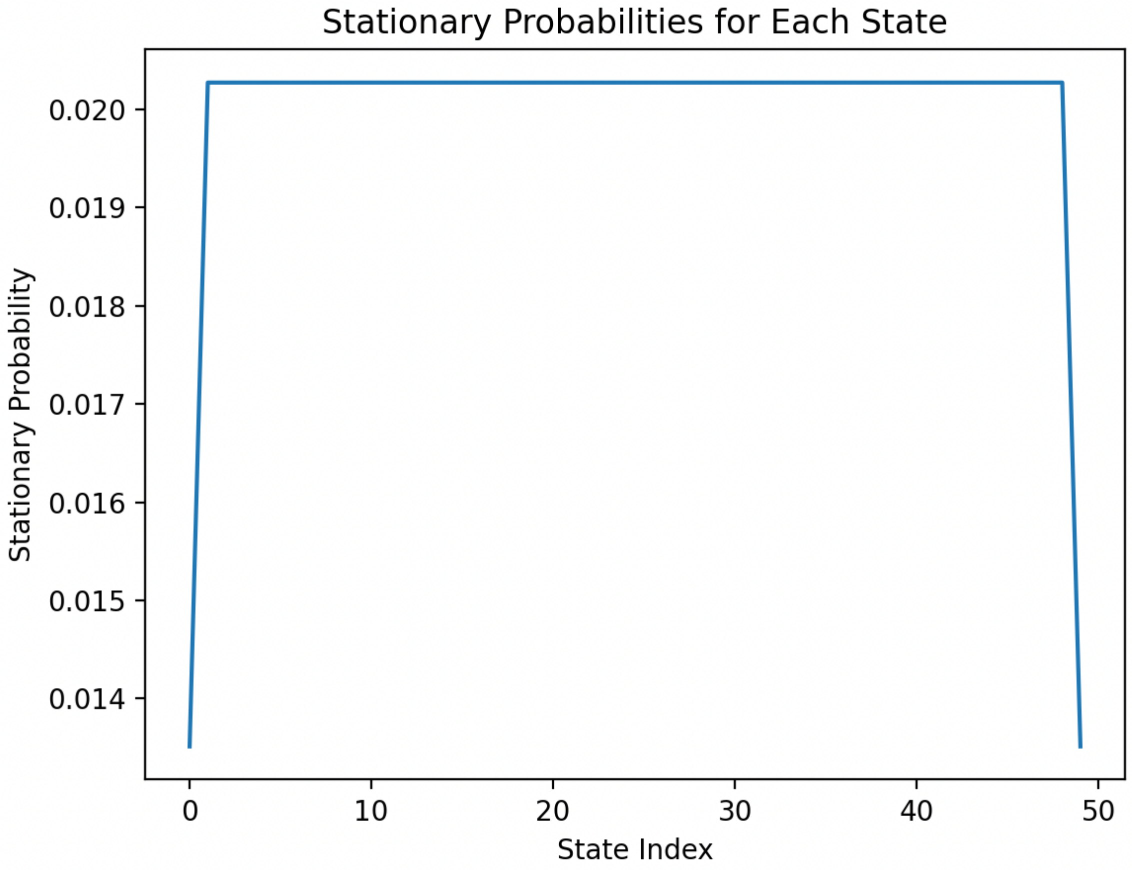
I completed the reading assignment of chapters 1, 2, 3, 4, and 5 of the course textbook: Sutton and Barto, Reinforcement Learning: An introduction, 2nd edition, The MIT Press.

Also, I watch the first 5 lectures of Prof. David Silver to deepen my understanding of the course.

Problem 2

Part a

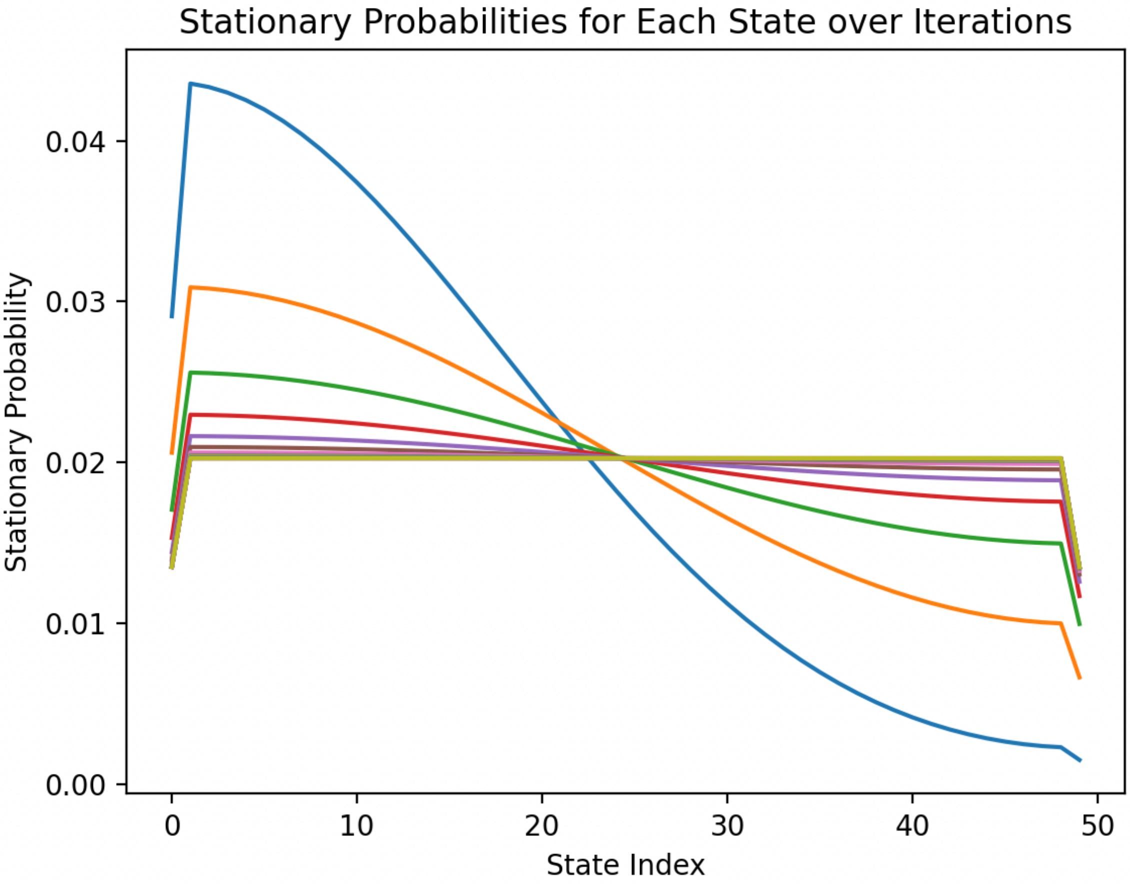
Programming solution in python notebook file. Below is the figure with the resulting stationary probabilities per state.



Part b

Programming solution in python notebook file. Below is the figure with the resulting stationary probabilities per state. For a better visualization the first n shown in plot is 500 and not 0.

From these results, we can observe that both method a and method b converge to the same stationary probabilities.



Xiyana Veroska Figuera Michal
20225398

Problem 3

Part a

Considering any two value functions $V_1(s)$ and $V_2(s)$

We need to prove that:

$$d\big(T(X_1), T(X_2)\big) \leq \ \gamma \, d(X_1, X_2)$$

More specifically we need to prove that:

$$\|T^*V_1(s) - T^*V_2(s)\|_\infty \leq \ \gamma \, \|V_1(s) - V_2(s)\|_\infty$$

We have

$$\|T^*V_1(s) - T^*V_2(s)\|_\infty =$$

$$\left\| max_a\{R(s,a) + \gamma \sum_{s'} p(s'|s,a)\, V_1(s')\} - max_{\tilde{a}}\{R(s,\tilde{a}) + \gamma \sum_{s'} p(s'|s,\tilde{a})\, V_2(s')\} \right\|_\infty$$

$$\leq \left\| max_a \left\{ R(s,a) + \gamma \sum_{s'} p(s'|s,a)\, V_1(s') - R(s,a) + \gamma \sum_{s'} p(s'|s,a)\, V_2(s') \right\} \right\|_\infty$$

$$\leq \ \left\| max_a\{\gamma \sum_{s'} p(s'|s,a)\, [V_1(s') - V_2(s')] \right\|_\infty$$

Here we can use:

$$\|max_a(a,b)]\| \leq max_a(\|a\|, \|b\|)$$

And we get:

$$\leq \ \gamma max_a\{\sum_{s'} p(s'|s,a) \ \| \ [V_1(s') - V_2(s')]\|_\infty\}$$

Since the probability the does not change $\| \ [V_1(s') - V_2(s')]\|_\infty$, we can take it out,

$$\leq \ \gamma\| \ [V_1(s') - V_2(s')]\|_\infty \ max_a\{\sum_{s'} p(s'|s,a) \ \}$$

And we can finally prove the contraction because, $\sum_{s'} p(s'|s,a) = 1, \ \ \forall \ s, a$

$$\|T^*V_1(s) - T^*V_2(s)\|_\infty \leq \ \gamma \ \|V_1(s) - V_2(s)\|_\infty$$

Problem 4

Part a

3.6) From textbook unified definition of return (eq. 3.11),

$$G_t = \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$$

Where we can include either but not both $T = \infty$ or $\gamma = 1$

If we treat pole-balancing as an episodic task with discount and only reward of -1 upon failure.

$$G_t = 0 + (\gamma)(0) + (\gamma^2)(0) + \cdots + (\gamma^{T-t-1})(-1) = -\gamma^{T-t-1}$$

Which is the same as for the continuous case if $0 \leq \gamma < 1$, because K is the number of time steps before failure, thus in the episodic case we could make K = T-t

$$G_t = -\gamma^{K-1}$$

Part b

3.8) We can obtain the returns as follows

$G_5 = 0$ (because $T = 5$)

$G_4 = R_5 = 2$

$G_3 = R_4 + \gamma\, G_4 = 3 + \frac{1}{2}\, 2 = 4$

$G_2 = R_3 + \gamma\, G_3 = 6 + \frac{1}{2}\, 4 = 8$

$G_1 = R_2 + \gamma\, G_2 = 2 + \frac{1}{2}\, 8 = 6$

$G_0 = R_1 + \gamma\, G_1 = -1 + \frac{1}{2}\, 6 = 2$

Part c

3.9) To solve for G_1 we can harness that:

$$\sum_{n=0}^{\infty} ar^n = \frac{a}{1-r}$$

Then,

$$G_1 = \sum_{n=0}^{\infty} 7(0.9)^n = \frac{7}{1-0.9} = 70$$

And thus,

$$G_0 = R_1 + \gamma\, G_1 = 2 + (0.9)(70) = 65$$

Problem 5

Part a

4.1) We can use the bellman expectation equation as defined in David Silver’s slide 32 of MDP to obtain $q_\pi(11, down)$ and $q_\pi(7, down)$

$$q_\pi(s,a) = R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a\, v_\pi(s')$$

Here, because it is episodic $\gamma = 1$ and because an action in a state always leads to the same next state, $P_{ss'}^a = 1$

Then,

$$q_\pi(11, down) = -1 + v_\pi(T) = -1 + 0 = -1$$

$$q_\pi(7, down) = -1 + v_\pi(11) = -1 - 14 = -15$$

Part b

4.2) We can solve $v_\pi(15)$ using the bellman expectation equation as defined in David Silver's slide 33 of MDP

$$v_\pi(s) = \sum_{a \in A} \pi(a|s)\,[R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a\, v_\pi(s')]$$

Again $\gamma = 1$ because it is episodic and $\pi(a|s) = 0.25$ because there are 4 equiprobable actions. Also $P_{ss'}^a = 1$ because the action in a state always lead to the same next action.

$$v_\pi(15) = (0.25)[-1 + v_\pi(12)] + (0.25)[-1 + v_\pi(13)] + (0.25)[-1 + v_\pi(14)] + (0.25)[-1 + v_\pi(15)]$$

$$v_\pi(15) = (0.25)[-4 + v_\pi(12) + v_\pi(13) + v_\pi(14) + v_\pi(15)]$$

$$v_\pi(15) = (0.25)[-4 - 22 - 20 - 14 + v_\pi(15)]$$

$$v_\pi(15) = (0.25)(-60) + (0.25)v_\pi(15)$$

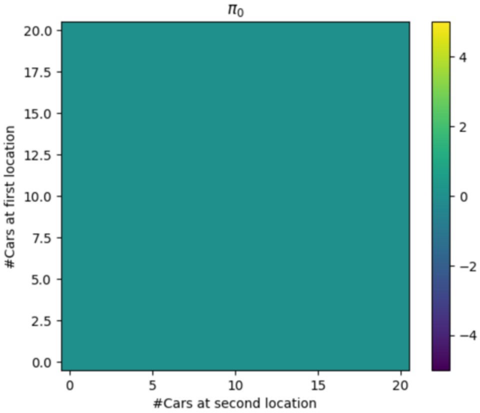
$$v_\pi(15) - (0.25)v_\pi(15) = -15$$

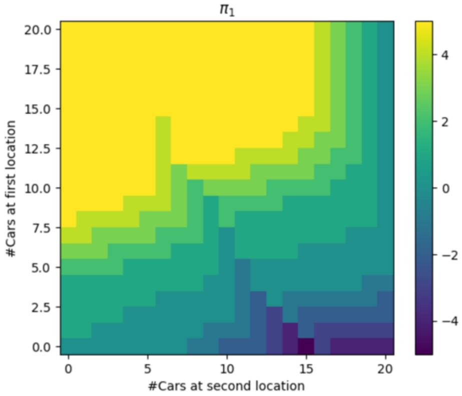
$$v_\pi(15) = \frac{-15}{0.75}$$

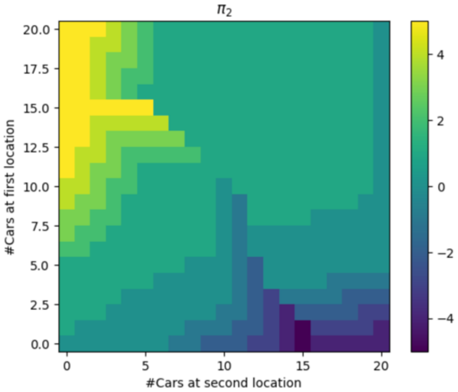
$$v_\pi(15) = -20$$

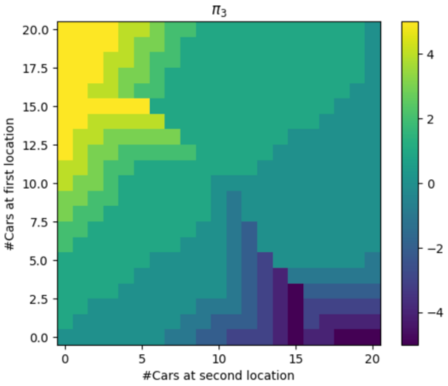
Now if the next state of state 13 was changed for the action down so that it takes agent to the state 15, the value of state 13 and 15 would remain to be -20 each. This is because since we are changing the next state from a non-terminal states to another non-terminal state, the reward of such non-terminal next step is the same (-1).

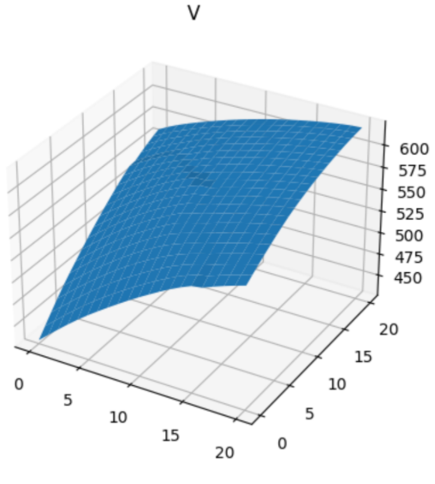
Problem 6

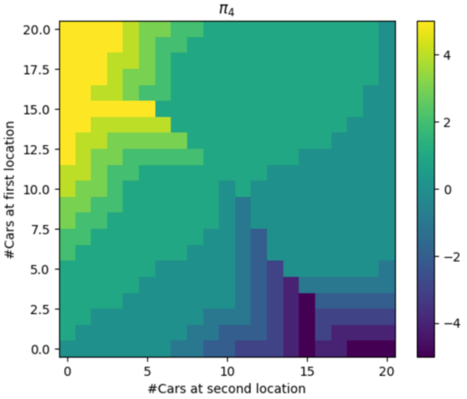

π_0


π_1


π_2


π_3


V


π_4

Xiyana Veroska Figuera Michal
20225398