Reinforcement Learning, EL2805

Computer Lab 1

November 27, 2020

Jules Callens, Marius Leblanc

1 The Maze and the Random Minotaur - Problem 1

a)

State space : For all (i_1, j_1) and (i_2, j_2) in the limit of the maze and with (i_1, j_1) not a wall, $((i_1, j_1), (i_2, j_2)) \in S$. (i_1, j_1) is the position of the player and (i_2, j_2) is the position of the minotaur.

Action space : $A = \{Right, Left, Up, Down, Stay\}$.

Time Horizon : T is finite, it is a finite time-horizon problem.

Probability : First with the action *Right* :
If the minotaur is not next to the limit of the maze:

$$\begin{aligned}
Pr(((i_1, j_1 + \delta), (i_2, j_2 + 1))|((i_1, j_1), (i_2, j_2)), Right) &= \frac{1}{4} \\
Pr(((i_1, j_1 + \delta), (i_2, j_2 - 1))|((i_1, j_1), (i_2, j_2)), Right) &= \frac{1}{4} \\
Pr(((i_1, j_1 + \delta), (i_2 + 1, j_2))|((i_1, j_1), (i_2, j_2)), Right) &= \frac{1}{4} \\
Pr(((i_1, j_1 + \delta), (i_2 - 1, j_2))|((i_1, j_1), (i_2, j_2)), Right) &= \frac{1}{4}
\end{aligned}$$

With $\delta = 0$ if $(i_1, j_1 + 1)$ is a wall or a limit of the maze and $\delta = 1$ in the other case.

If the minotaur is next to the right limit of the maze:

$$\begin{aligned}
Pr(((i_1, j_1 + \delta), (i_2, j_2 - 1))|((i_1, j_1), (i_2, j_2)), Right) &= \frac{1}{3} \\
Pr(((i_1, j_1 + \delta), (i_2 + 1, j_2))|((i_1, j_1), (i_2, j_2)), Right) &= \frac{1}{3} \\
Pr(((i_1, j_1 + \delta), (i_2 - 1, j_2))|((i_1, j_1), (i_2, j_2)), Right) &= \frac{1}{3}
\end{aligned}$$

With $\delta = 0$ if $(i_1, j_1 + 1)$ is a wall or a limit of the maze and $\delta = 1$ in the other case.

If the minotaur is next to the right and the bottom limits of the maze:

$$\begin{aligned}
Pr(((i_1, j_1 + \delta), (i_2, j_2 - 1))|((i_1, j_1), (i_2, j_2)), Right) &= \frac{1}{2} \\
Pr(((i_1, j_1 + \delta), (i_2 - 1, j_2))|((i_1, j_1), (i_2, j_2)), Right) &= \frac{1}{2}
\end{aligned}$$

With $\delta = 0$ if $(i_1, j_1 + 1)$ is a wall or a limit of the maze and $\delta = 1$ in the other case.

It is the same for the other limits of the maze for the minotaur. For the other actions you do the same thing but for the $Left$ it is $-\delta$, with the $Down$ it is $i_1 + \delta$, with the Up it is $i_1 - \delta$ and with the $Stay$ you do not change the player position.

If the minotaur can stay, you have to add 1 to the denominator of the fraction and add one probability for the minotaur to stay.

Reward : If the player is on the same position than the minotaur then the reward is -100. If you go in a wall or in a limit of the maze then your reward is -100. If you are on the exit of the maze without the minotaur then your reward is 100. Finally, if you do an action your reward is -1 to force the player to be fast.

b) On the figure 1 below you can see the optimal policy for $T = 20$.

On the figure 2 below you can see that if the minotaur cannot stay then if the goal is attainable ($T \geq 15$) the player will always reach the end. Else the player will more and more reach the end as T increases because he will take less and less risk. The difference between when the minotaur can stay and when he cannot is due to the fact that if we think of the maze as a chessboard, the player will start on a white case when the minotaur will start on a black case. When the minotaur move he will always change the color of the case under him, so if the player never stay he will also change the color of the case under him, and then the minotaur will not be able to catch the player. But if the minotaur can stay,

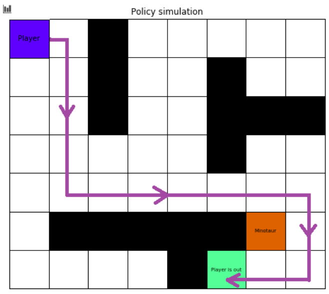


Figure 1: Optimal policy for $T = 20$ from the start

then he will randomly change or not the color under him, so the player will not be able to always predict the color under the minotaur and will not always win.

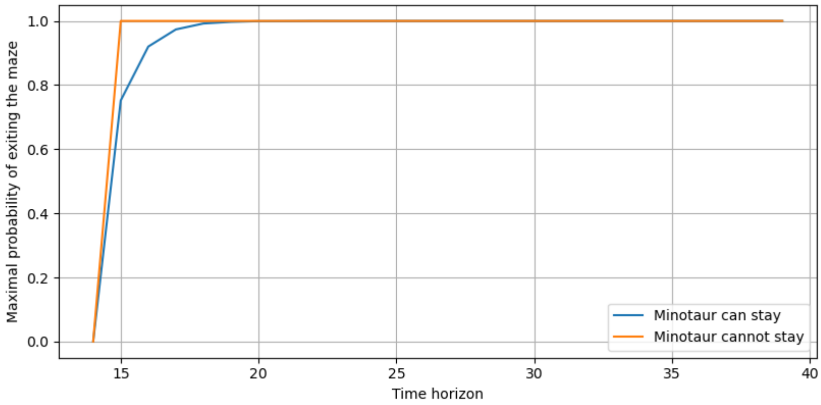


Figure 2: The maximal probability of exiting the maze as a function of T

c) State space : The new state space is $S' = S \cup \{Dead\}$, to take in count the fact that the player can die.

Action space : The action space is exactly the same as before because there is not new actions.

Time Horizon : The time horizon is exactly the same as before.

Probability : For the probability of transition, we have to add at each action and state the probability of dying from oldness. Which is a probability of $p = \frac{1}{30}$ at each time because we know that $\mathbb{E} = \frac{1}{p} = 30$. So it will become:

$$Pr(Dead|((i_1, j_1), (i_2, j_2)), Right) = p$$

And for all the other probabilities you multiply them by $1-p$, to have the sum of the probabilities equal to 1.

Reward : The reward is the same as before with one more reward which is when you die from oldness then the reward will be -100 to force him to hurry up.

The probability of getting out alive using this policy by simulating 10000 games with $T = 100$ is 0.59.

2 Robbing Banks - Problem 2

a)

State space : For all (i_1, j_1) and (i_2, j_2) in the limit of the city, $((i_1, j_1), (i_2, j_2) \in S$. (i_1, j_1) is the position of the player and (i_2, j_2) is the position of the police.

Action space : $A = \{Right, Left, Up, Down, Stay\}$.

Time Horizon : T is infinite, the rewards are discounted at rate λ.

Reward : Each time you are at a bank, and the police is not there, you collect a reward of 10 SEK. If the police catches you, you loose 50 SEK.

Probability : The police always chases the thief, but moves randomly in his direction. So, if the police on the same line (on the right of the thief for example), then:

$$Pr(((i_1+\delta_1, j_1+\delta_2), (i_2, j_2-1))|((i_1, j_1), (i_2, j_2)), a) = \frac{1}{3}$$
$$Pr(((i_1+\delta_1, j_1+\delta_2), (i_2-1, j_2))|((i_1, j_1), (i_2, j_2)), a) = \frac{1}{3}$$
$$Pr(((i_1+\delta_1, j_1+\delta_2), (i_2, j_2+1))|((i_1, j_1), (i_2, j_2)), a) = \frac{1}{3}$$

With a $\in$ A, any action of the thief. δ depends of the action chosen:

Action	δ_1	δ_2
Right	0	+1
Left	0	-1
Up	-1	0
Down	+1	0

If there is a limit in the direction chosen the thief stay: $(\delta_1, \delta_2) = (0, 0)$

If the police is on the same column (above the thief for example):

$$Pr(((i_1+\delta_1, j_1+\delta_2),(i_2-1,j_2))|((i_1,j_1),(i_2,j_2)),a) = \frac{1}{3}$$
$$Pr(((i_1+\delta_1, j_1+\delta_2),(i_2,j_2+1))|((i_1,j_1),(i_2,j_2)),a) = \frac{1}{3}$$
$$Pr(((i_1+\delta_1, j_1+\delta_2),(i_2,j_2-1))|((i_1,j_1),(i_2,j_2)),a) = \frac{1}{3}$$

With a $\in$ A, any action of the thief. δ depends of the action chosen as above.

If the police is not on the same column nor on the same line (on the top right of the thief for example):

$$Pr(((i_1+\delta_1, j_1+\delta_2),(i_2-1,j_2))|((i_1,j_1),(i_2,j_2)),a) = \frac{1}{2}$$
$$Pr(((i_1+\delta_1, j_1+\delta_2),(i_2,j_2-1))|((i_1,j_1),(i_2,j_2)),a) = \frac{1}{2}$$

With a $\in$ A, any action of the thief. δ depends of the action chosen as above.

We can define all the probabilities left without loss of generality as in the three example above.

b) The average maximal reward earned is given by the value function. As we can see on the figure below the bigger the lambda is the bigger is the reward: it is normal because lambda is the discount factor : without an important discount the reward is less important.

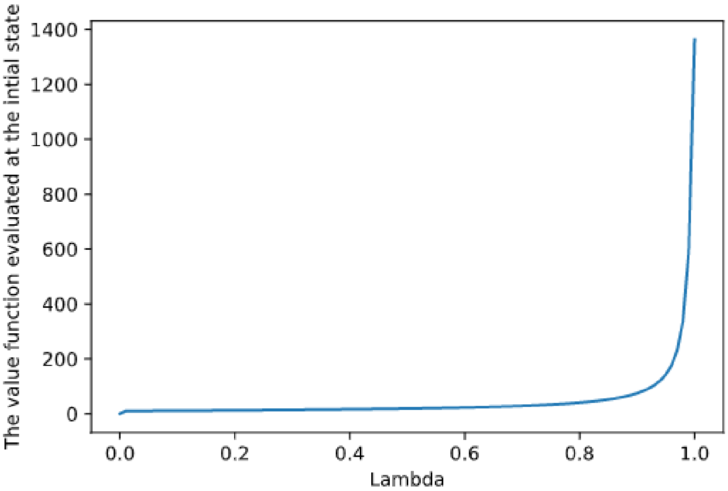
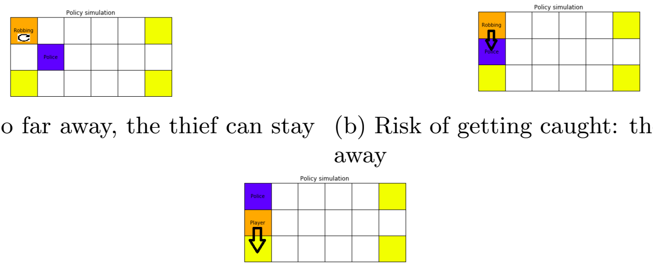


Figure 3: The value function evaluated at the initial state

Optimal policy: The best choice for the thief depend on two factor: the position of the police and the discount factor. To maximise his gain, the thief wants to never get caught but the direction of his escape depend of λ: if $\lambda > 0.05$, he has the ”time” to go to an other bank, else he stay close to the first bank.

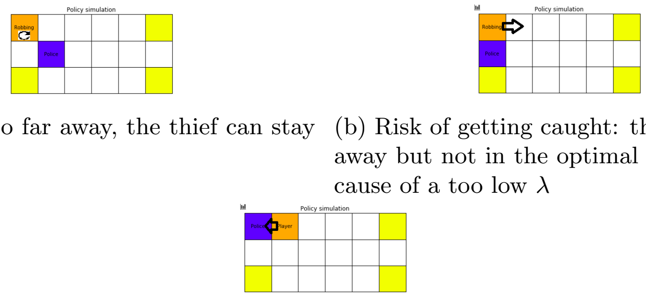


(a) Police too far away, the thief can stay in the bank

(b) Risk of getting caught: the thief move away

(c) The thief go to the closest bank which is the farthest of the police

Figure 4: Plots of the best policy for $\lambda > 0.05$



(a) Police too far away, the thief can stay in the bank

(b) Risk of getting caught: the thief move away but not in the optimal direction because of a too low λ

(c) The thief go to the closest bank

Figure 5: Plots of the best policy for $\lambda < 0.05$

3 Bank Robbing (Reloaded) - Problem 3

State space : For all (i_1, j_1) and (i_2, j_2) in the limit of the city, $((i_1, j_1), (i_2, j_2) \in S$. (i_1, j_1) is the position of the player and (i_2, j_2) is the position of the police.

Action space : $A = \{Right, Left, Up, Down, Stay\}$.

Time Horizon : T is infinite, the rewards are discounted at rate $\lambda = 0.8$.

Reward : For each round spent in the bank, you receive a reward of 1 SEK. The police walks randomly you lose 10 SEK.

Probability : are the same as in The Maze and the Random Minotaur

a)

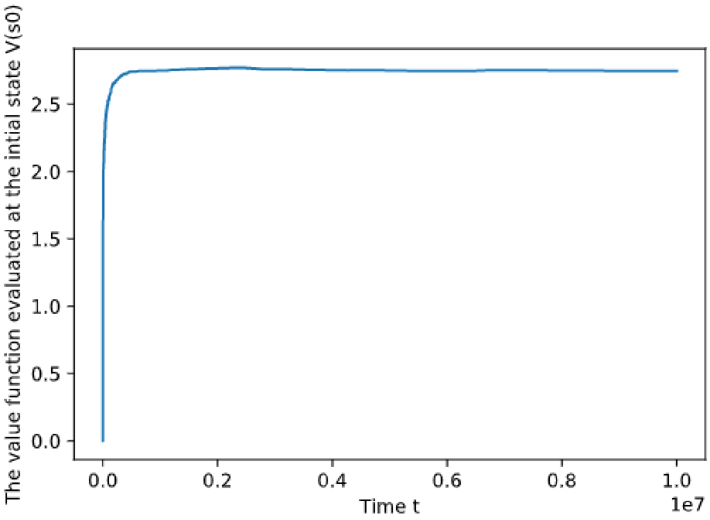


Figure 6: The value function with Q-Learning

On the figure 6 we can see that the value function converges toward 2.7 after 10 000 000 iterations. It shows also that we need many iterations (and so a lot of explorations) to know the best policy.

b)

SARSA algorithm seems to have a slower converge for many epsilon and even with a small epsilon (corresponding to a small amount of explorations) have better results. However it does not reach to Q-learning score.

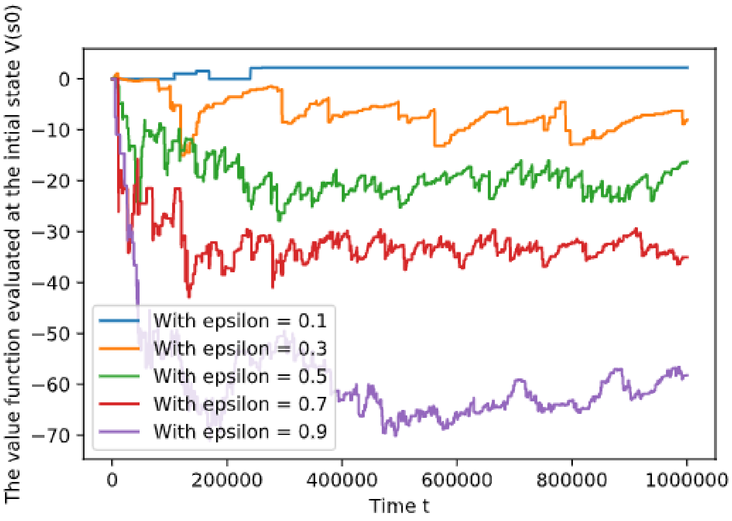


Figure 7: The value function with SARSA