RL Lab 1

Kildo Alkas Alias 971106-7430
Andrej Wilczek 880707-7477

November 2020

1 Introduction

2 Problem 1: The Maze and the Random Minotaur

2.1 a)

S = $\{(xP, yP, xM, yM)\}, 0 \leq xP, xM \leq 6, 0 \leq yP, yM \leq 7$
Unobtainable states such as walls are not included in the state space.

A = {up,down,left,right,stay}
Actions are for the player. Actions for the Minotaur are the same if it is permitted to stay, otherwise A lacks the action "stay".

$r(s, a) = -100$ if the action puts the player in an unobtainable state.
$r(s = (x_{goal}, y_{goal}), a = stay) = 0$
$r(s, a) = -1$ for all actions that gives a obtainable non-terminal state.

The time horizon is T.

$P(s' \mid s, a) = 1 * P_{Minotaur}$

The transition probabilities for the player are deterministic and are always equal to one if the action results in an obtainable state.
If the player chooses an action that transitions to a unobtainable state the probability that the player remains in its current state is equal to 1.
The transition probabilities for the Minotaur is a random walk and the probability is always equal to $1/\sum a$ for all actions, a, that result in an obtainable state.

2.2 b)

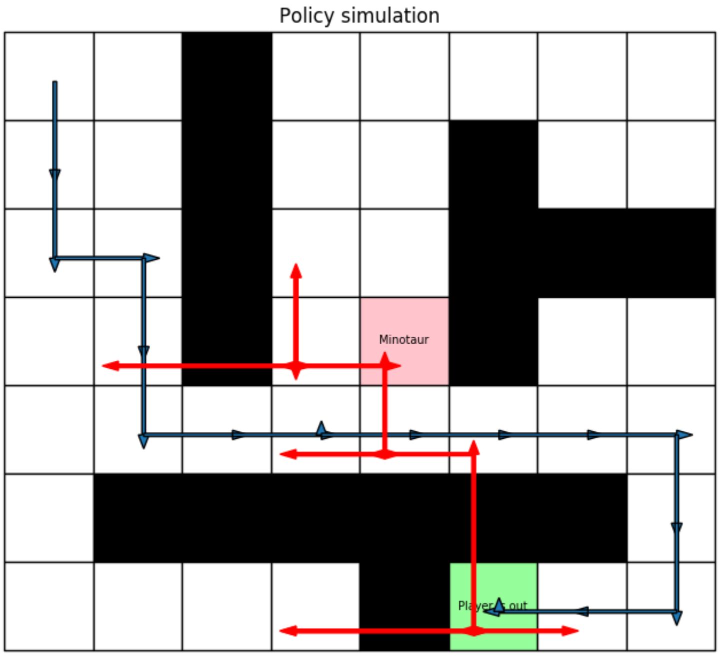

Figure 1: Illustration of the optimal policy for one simulation. Blue for player and red for Minotaur.

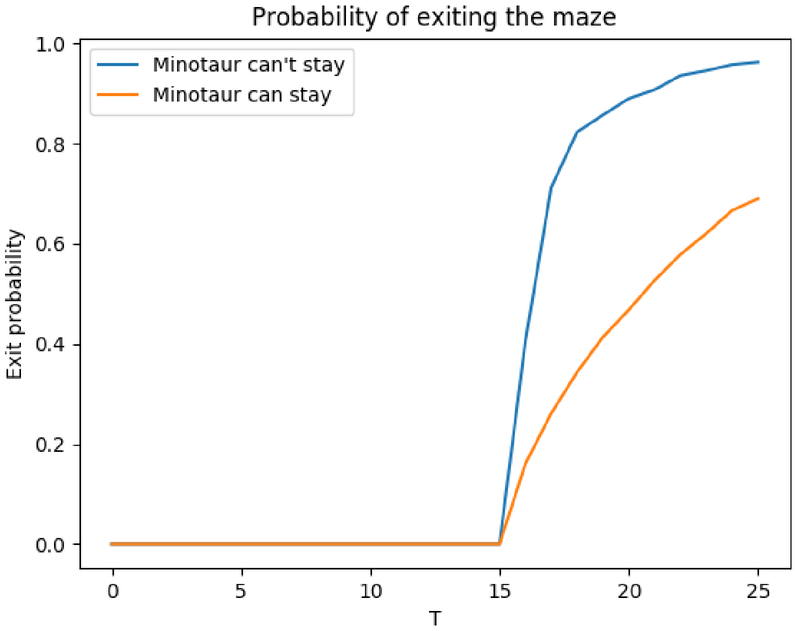

Figure 2: Probability of exit the maze under the optimal policy as a function of the time horizon.

The path illustrated in Figure 1 is:

(0, 0, 6, 5), (1, 0, 6, 3), (2, 0, 6, 5), (2, 1, 6, 6), (3, 1, 6, 7), (4, 1, 5, 7), (4, 2, 6, 7), (4, 3, 6, 6), (4, 4, 6, 7), (4, 5, 6, 6), (4, 5, 6, 5), (4, 6, 6, 6), (4, 7, 4, 6), (5, 7, 3, 6), (6, 7, 3, 7), (6, 6, 1, 7), (6, 5, 3, 7), (6, 5, 4, 7), (6, 5, 4, 6), (6, 5, 4, 5), (6, 6, 4, 4)

2.3 c)

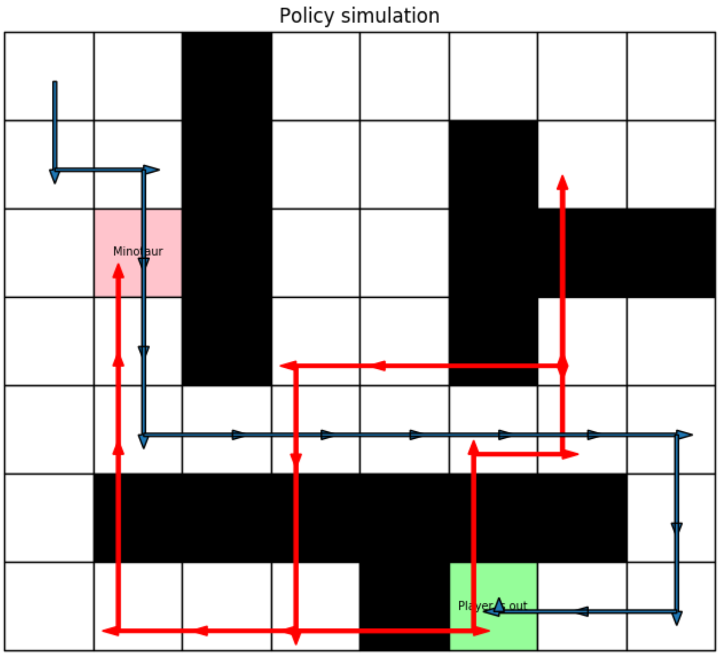

Figure 3: Animation of the path. blue for player and red for Minotaur

Path of the player and Minotaur:

[(0, 0, 6, 5), (1, 0, 4, 5), (1, 1, 4, 6), (2, 1, 3, 6), (3, 1, 1, 6), (4, 1, 3, 6), (4, 2, 3, 4), (4, 3, 3, 3), (4, 4, 4, 3), (4, 5, 6, 3), (4, 6, 6, 5), (4, 7, 6, 3), (5, 7, 6, 2), (6, 7, 6, 1), (6, 6, 4, 1), (6, 5, 3, 1), (6, 5, 2, 1)]

3 Bank robber

3.1 a)

S = $\{(xR, yR, xP, yP)\}, 0 \leq xR, xP \leq 2, 0 \leq yR, yP \leq 5,$
$\forall s$ where $(xR, yR) \neq (1, 2)$
The police station is considered as an unobtainable state for the robber.

A = {up,down,left,right,stay}
Actions are for the robber. Actions for the police are the same except they lack the action "stay".

$r(s, a) = -100$ if the action puts the player in an unobtainable state..
$r(s = (x_{bank}, y_{bank}), a = stay) = 10$
$r(s, a) = -50$ if action a leads to a state s' where $xR = xP$ and $yR = yP$.

The time horizon, T,is infinite.

$P(s' \mid s, a) = 1 * P_{Police}$

The transition probabilities for the robber are deterministic and are always equal to one if the action results in an obtainable state.
If the robber chooses an action that transitions to a unobtainable state the probability that the robber remains in its current state is equal to 1.
The transition probabilities for the police is either $\frac{1}{2}$ or $\frac{1}{3}$ depending on the current state. If the state is such that $xR = xP$ or $yR = yP$ the police has three possible actions as defined in the lab instructions and the probability of each action is $P_{Police}(a) = \frac{1}{3}$, if $xR \neq xP$ and $yR \neq yP$ the police has two possible actions and the probabilities are $P_{Police}(a) = \frac{1}{2}$ for all actions that leads to an obtainable state.

3.2 b)

As expected the value function evaluated in the initial state (see Figure 6) is increasing for increasing values of lambda. This is because we discount the reward less and therefore receive a larger reward for future state-actions pairs. As a result of this we can see that the behaviour for low values of lambda (see Figure 5)is more focused on short-term rewards as opposed to the behaviour for larger values of lambda (see Figure 4) which emphasises long term rewards more.

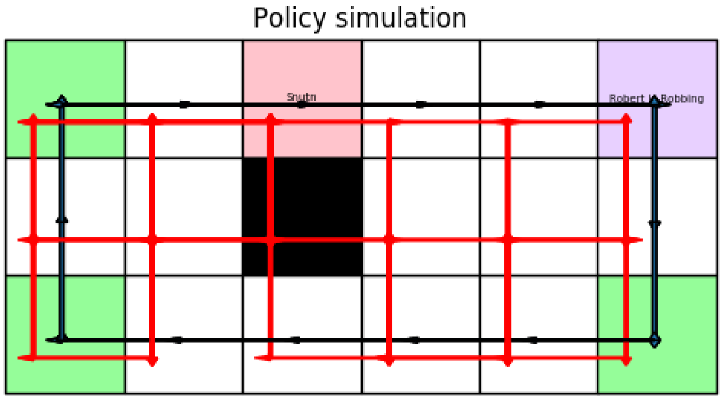

Figure 4: Illustration of optimal policy for 100 time steps with $\lambda = 0.88$.

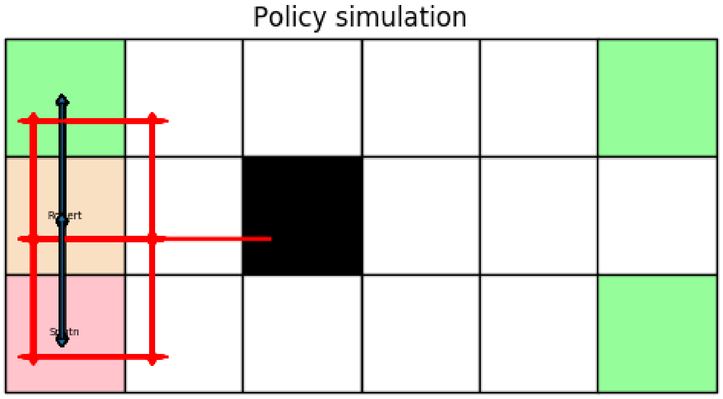

Figure 5: Illustration of optimal policy for 100 time steps with $\lambda = 0.7$.

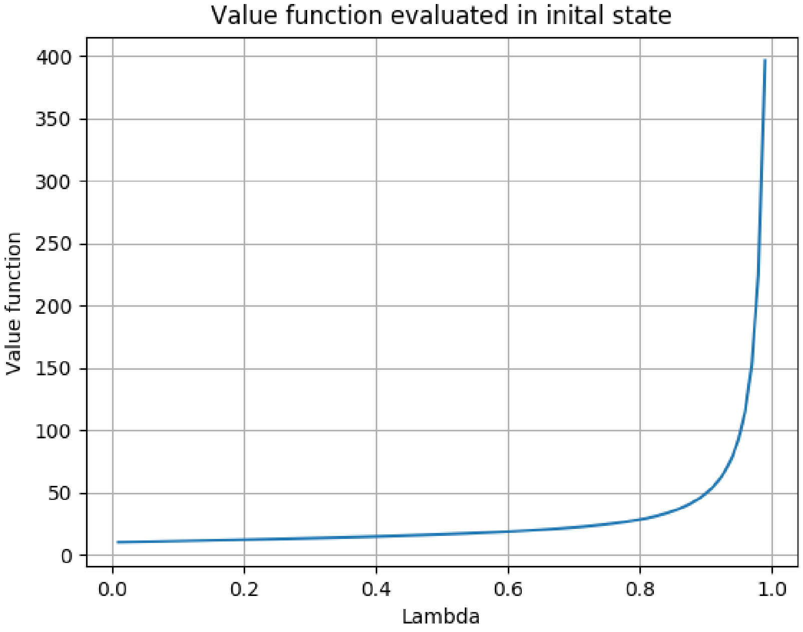

Figure 6: The value function evaluated in the initial state for different values of lambda.