Lab 1

EL2805 *November 25, 2022*

Kaj Munhoz Arfvidsson
19980213-4032

Erik Anderberg
19960806-7857

---

1 Problem 1

a) The problem can be formulated as an MDP as follows:

State space - The full state space is consisting of all possible combinations of the player's and the minotaur's state spaces, i.e. $\mathcal{S} = \mathcal{S}_p \times \mathcal{S}_m$. The minotaur's state space is quite straight forward, $\mathcal{S}_m = X \times Y$ whereas the player's state space is $\mathcal{S}_p = \{s_p \in X \times Y | s_p \neq \textit{occupied}\}$. Here $X = \{0, 1, ..., 7\}$ and $Y = \{0, 1, ..., 6\}$ are the coordinate axes of the maze. With 16 occupied squares this gives a total of $(8 \cdot 7 - 16) \cdot 7 \cdot 8 = 2240$ states, plus the terminal states *lost* for when the player is dead and *won* for when the player reached the exit without being intercepted by the Minotaur.

Action space - All possible combinations of actions the player can take and the Minotaur's movements: $(a_p, a_m) \in \mathcal{A} = \{up, down, left, right, stay\} \times \{up, down, left, right\}$.

Rewards - As the objective is to reach goal while avoiding obstacles and the Minotaur the rewards are $-\infty$ for going to a square with a wall or the minotaur, 0 for the goal square, and -1 for any other square.

Transition probabilities - The player's movement is deterministic (has the probability $P(s'|s, a_p) = 1$ if a_p is allowed, and 0 otherwise). If a_p is allowed, the probability for the Minotaur's movement is uniformly distributed over all allowed moves, i.e. $P = 0.25$ if all moves are allowed, $P = 0.5$ if only two moves are allowed and so on. The resulting transition probabilities $P(s'|s, a)$ are therefore uniformly distributed over all allowed states s'.

b) If the player and Minotaur take turns moving some things need to be modified to model it as an MDP:

State space - As before, but with an extra parameter $s_t \in \{\textit{player, minotaur}\}$ representing who's turn it is to move.

Action space - As before, except the Minotaur now has also has the action *stay*.

Rewards - The reward $r(s, a)$ will be:

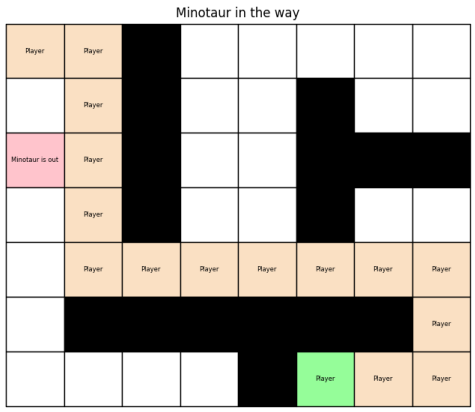
- $-\infty$ if the player and the Minotaur are on the same square ($s = \textit{lost}$) or if the resulting state s' if a is applied is illegal (outside the maze for example).
- 0 if $s = \textit{won}$.
- -1 otherwise.

Transition probabilities - The transition probabilities now also depend on who's turn it is as follows:

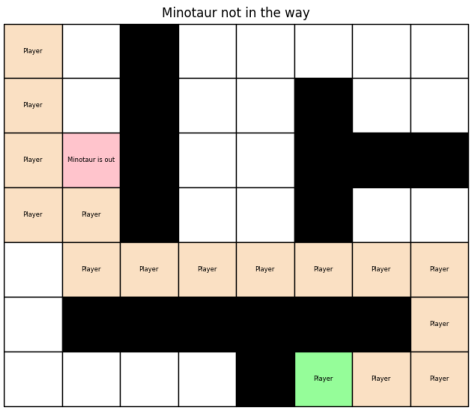
- $P(s'|s, a) = 1$ if $s_t = \textit{player}$, $s'_t = \textit{minotaur}$, $a_m = \textit{stay}$ and a_p is allowed.
- $P(s'|s, a) = 0$ if $(s_t = \textit{player} \wedge a_m \neq \textit{stay}) \vee (s_t = \textit{minotaur} \wedge (a_p \neq \textit{stay} \vee a_m = \textit{stay})) \vee (s_t = s'_t)$.

- $P(s'|s,a)$ is uniformly distributed over all states s' if $s_t = minotaur$, $s'_t = player$, $a_p = stay$, $a_m \neq stay$ and a_m is allowed.

c) In the figures below we can see that the player can adapt to where the minotaur is standing, choosing the optimal path to reach the exit.



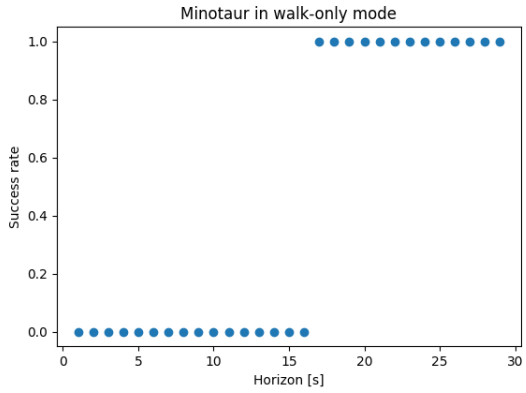
(a) Minotaur is fixed to a coordinate where the player usually visit.



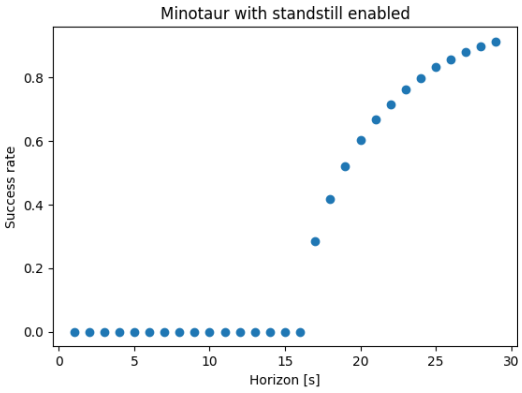
(b) Minotaur is fixed to a coordinate where the player would not usually do not visit.

d) If the Minotaur has to move, the player can make sure that the MD (Manhattan distance) is uneven, which makes it impossible for them to end up on the same square. Running the simulation with the initial state such that the MD is even, we see that the player stays still for one time step, ensuring that the MD is uneven for the rest of the simulation. As a result, the player can always escape given a sufficient time horizon to reach the exit.

If the Minotaur can stay the player loses the ability to guarantee an uneven MD. The probability of the player escaping therefore depends on the time horizon, as it decides how long the player can wait for the Minotaur to move. With a long enough time horizon the probability of the exit still being blocked goes towards zero.



(a) Minotaur must walk at every time step.



(b) Minotaur can either walk or stand still.

e) To programmatically simplify the new problem scenario a new state *poisoned* is introduced, $\mathcal{S} := \mathcal{S} \cup \{\textit{poisoned}\}$. Any state, except *lost*, can transition to *poisoned* and only the *lost* state can be reached from *poisoned*. This is reflected in the updated transition probabilities which are now changed to

- $P(s'|s,a) := 1$ if $s = \textit{poisoned} \wedge s' = \textit{lost}$
- $P(s'|s,a) := \beta$ if $s' = \textit{poisoned} \wedge s \in \mathcal{S} \setminus \{\textit{poisoned, lost}\}$
- $P(s'|s,a) := (1-\beta)P(s'|s,a)$ otherwise.

where $\beta = 1/30$ is the probability of being poisoned. Since *poisoned* will always result in *lost* in the following time step there is no need to update the rewards.

f) To test the new problem we simulate 10 000 games, first with the minotaur in walk-only mode and then also when the minotaur can stand still.

- Success rate with the minotaur in walk-only mode: $5956/10000 = 0.60$
- Success rate with the minotaur when standing still is enabled: $5313/10000 = 0.53$

g) On-policy means that the learning method uses the current policy, whereas off-policy means that another policy (possibly a random one) is used.

To converge Q-learning and SARSA both require all state-action pairs to be visited infinitely often, and the learning rate must approach zero at a suitable rate. The latter is guaranteed if α fulfills

$$\sum_{n=1}^{\infty} \alpha_n(a) = \infty \text{ and } \sum_{n=1}^{\infty} \alpha_n^2(a) < \infty,$$

i.e. α needs to be big enough to escape local maxima but small enough for the value functions to converge. SARSA also requires the learning policy to become greedy in the limit, for example by setting $\epsilon = 1/t$.

h) The state space can this time be expanded by adding a variable $K \in \{0,1\}$, indicating whether the player has picked up the keys or not, resulting in $s = (x_p, y_p, x_m, y_m, k)$. The action space is the same as before, but the reward is changed so that it's only zero at the exit if $k = 1$, i.e. the player has passed by point C and picked up the keys. It can be thought of as a maze with two floors, where the player starts on the top floor, the exit is on the bottom floor, and the only stairs are at C (and the Minotaur exists on both floors on the same square simultaneously). The transition probabilities change a bit more, and are now (where $\theta = 1/50$):

- $P(s'|s,a) = (1-\theta)(0.35 + 0.65 \cdot \frac{1}{n})$ if a is allowed and $MD(s') < MD(s)$, where $MD(s) = |x_p - x_m| + |y_p - y_m|$ is the Manhattan distance between the player and the Minotaur, and n is the number of available squares around the Minotaur.
- $P(s'|s,a) = (1-\theta)(1 - (0.35 + 0.65 \cdot \frac{1}{n}))$ if a is allowed and $MD(s') \geq MD(s)$.
- $P(s'|s,a) = \theta$ if $s' = \textit{lost}$
- $P(s'|s,a) = 1$ if $s = s' \in \{\textit{won,lost}\}$ for all actions.
- $P(s'|s,a) = 0$ otherwise.