Neural Network and Deep Learning
Homework 5

Federico Agostini

1 Introduction

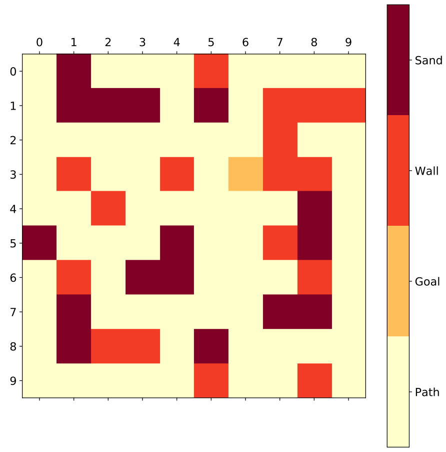
In this report reinforcment learning is explored training an agent on a 10 × 10 grid. The agent can perform 5 moves (move up, down, left, right or stay) in order to reach the SARSA and Q-learning policies are tested, along with different values for the learning rate α and the discount factor γ.

2 The environment

The enviroment is modified with respect to the simple grid proposed during the laboratory; in particular, as it can be seen in Fig. 1, it has different blocks:

- *Path*: standard blocks for the agent (no reward, possible to walk through)
- *Goal*: goal to reach (+1 reward for each timestep the agents remains on it)
- *Wall*: impassable obstacles (-1 reward, impossible to go through); also boundaries enter in this category
- *Sand*: crossable obstacles (-0.75 reward, possible to pass through)

Figure 1: *Environment used to train the agent.*



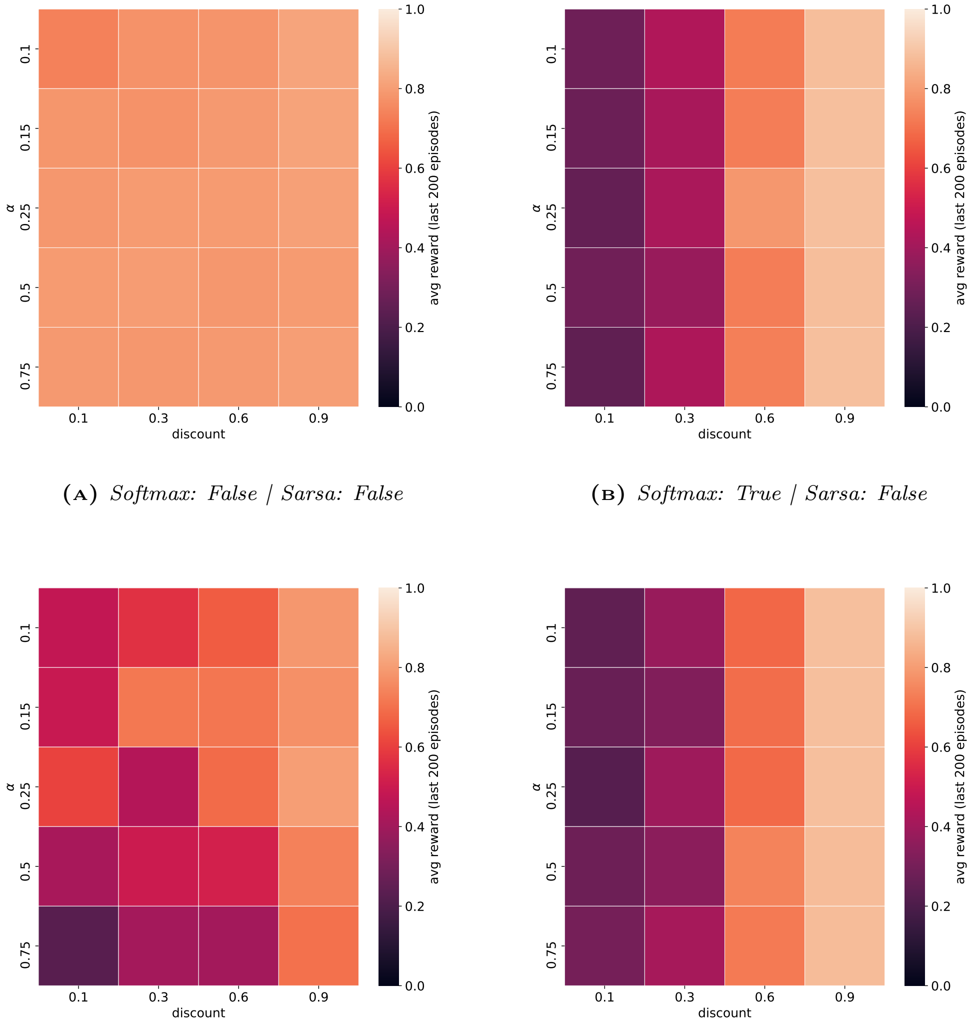
3 Training

Training is done with 2000 episodes each one of length 50; ϵ-greedy action selection is set to decrease from 0.8 to 0.001 evenly as the number of episodes increases. Discount γ and learning rate α are repectivly in `[0.1, 0.3, 0.6, 0.9]` and `[0.1, 0.15, 0.25, 0.5, 0.75]`. In addition, SARSA algorithm and Softmax policy are explored.

Fig. 2 displays the average reward in the last 200 episodes for the different combinations of the parameters. It can be noticed that if we sample using the Softmax, discount factor plays tha major role, while the learning rate does not influence the score; in the case where SARSA is used without Softmax, instead, both α and γ change the outcome; at last, if neither the softmax nor SARSA are active, an average reward near to 1 is always achieved.

Focusing on the parameters $\alpha = 0.25$ and $\gamma = 0.9$, Fig. 3 shows the average reward as function of the episode. When Softmax is not used, the average reward increases with a linear trend as function of the episode, while if Softmax is set to `True` it grows up faster and then oscillates between 0.8 and 1.

Figure 2: *Heatmaps of the average reward in the last 200 episodes for different combinations of the parameters α and γ.*



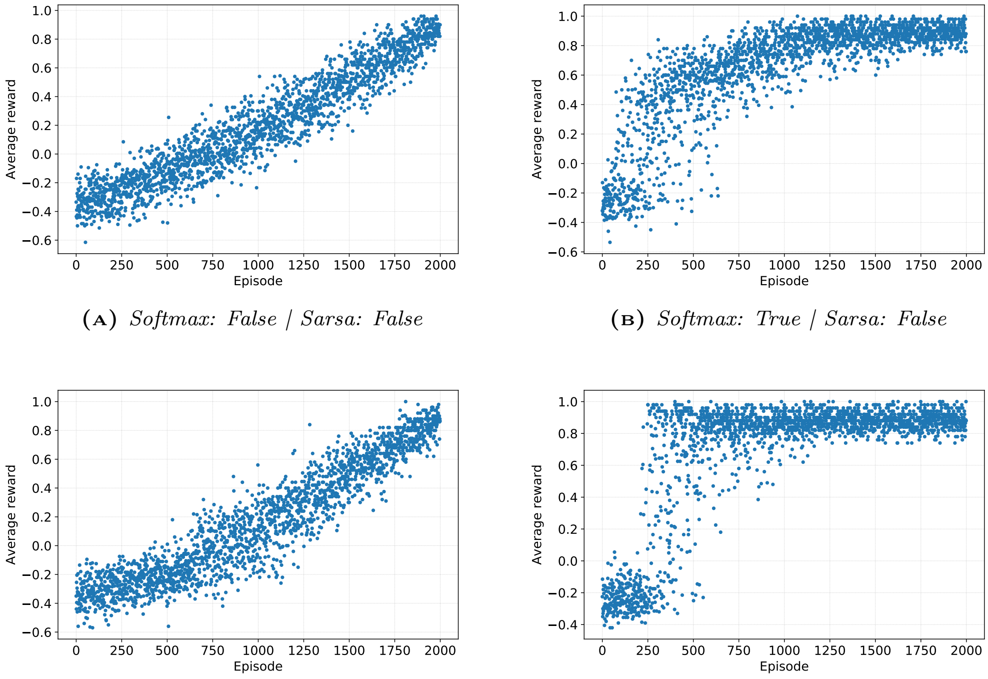
(A) *Softmax: False / Sarsa: False*

(B) *Softmax: True / Sarsa: False*

(C) *Softmax: False / Sarsa: True*

(D) *Softmax: True / Sarsa: True*

Figure 3: *Average reward as function of the episode. Learning rate is set to 0.25 and discount factor to 0.9.*



(A) *Softmax: False | Sarsa: False*

(B) *Softmax: True | Sarsa: False*

(C) *Softmax: False | Sarsa: True*

(D) *Softmax: True | Sarsa: True*

4 Testing

Trained agent is then tested in order to reach the goal starting from two different positions. Fig. 4 and 5 shows that the chosen path to reach the goal may be different depending on the parameters choisen during the training procedure.

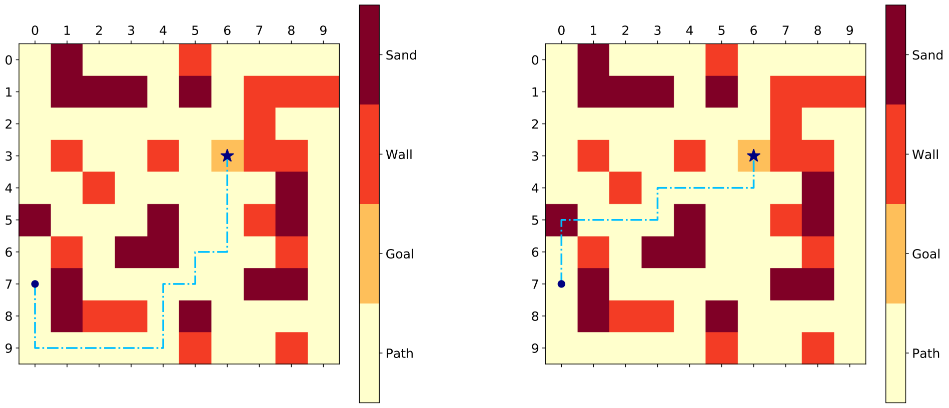
5 Different environments

Simulations are repeated chenging the environment, keeping only the sand (Fig. 6a), walls (Fig. 6b) or removing every obstacle (Fig. 6c).

The average reward of the last 200 episodes as function of the learning rate and discount follows a similar trend as before, if the obstacles exists and are all the same kind (Fig. 7 and 8), while without obstacles higer values are reached even for smaller values of the discount (Fig.9).

Having only walls leads the agent to learn the same path independently on the parameters, since the road is constrained by the environment. On the other hand, if the obstacles are crossable or they are removed, different pathway could be chosen to resolve the problem (Fig. 10).

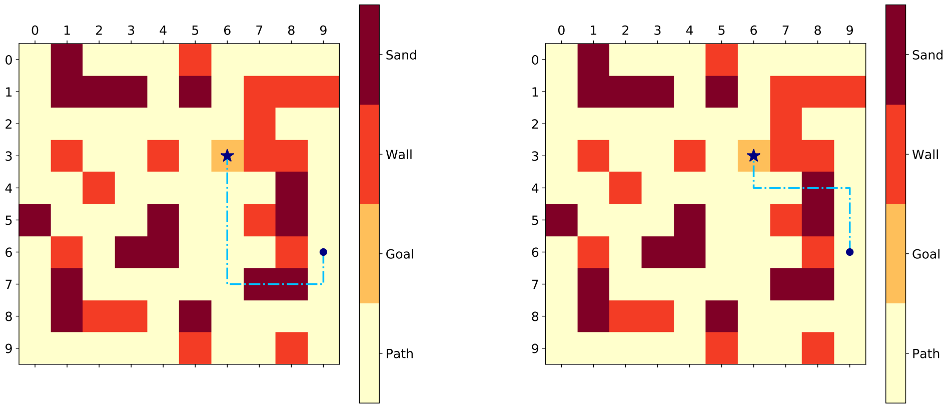
Figure 4: *Path chosen by the agent to reach the goal starting from the blue dot (star is the final position). Learning rate is set to 0.25 and discount factor to 0.9.*



(A) *If Sarsa is used without Softmax, the agent avoids the sand and prefers a longer journey to reach the goal.*

(B) *With different combiantions of Sarsa and Softmax, the agent chooses a shorter path even going through the sand.*

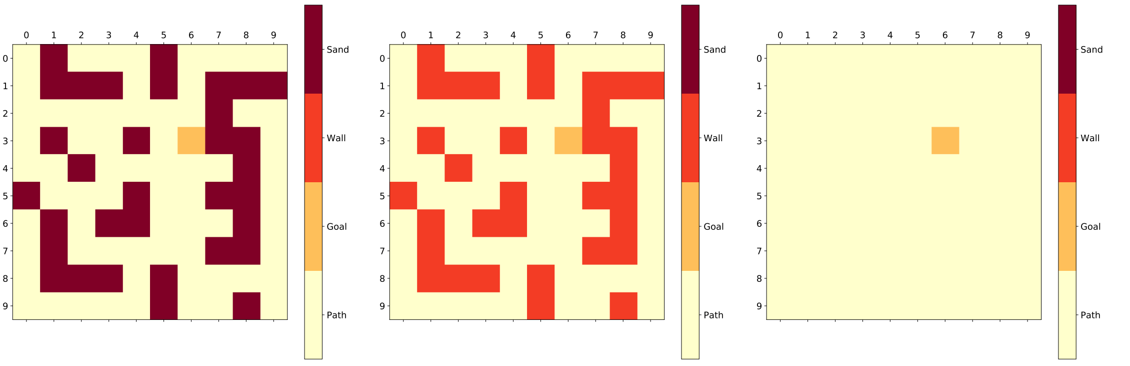
Figure 5: *Path chosen by the agent to reach the goal starting from the blue dot (star is the final position). Learning rate is set to 0.25 and discount factor to 0.9.*



(A) *If Sarsa is used with Softmax, the agent prefers a longer path and crosses two sand blocks.*

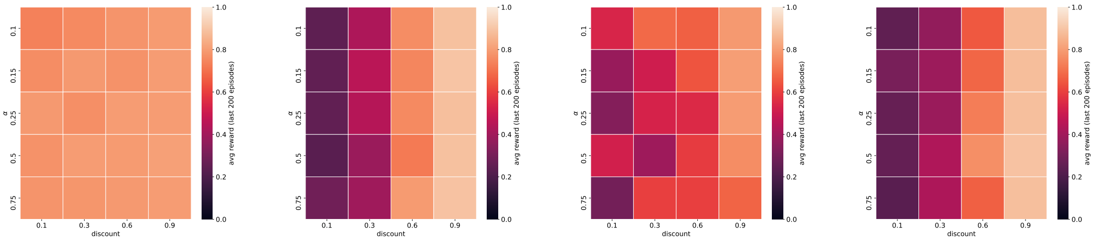
(B) *With different combiantions of Sarsa and Softmax, the agent chooses a shorter path and goes through only one sand.*

Figure 6: *Different environments used in simulations.*



(A) (B) (C)

Figure 7: *Heatmaps of the average reward in the last 200 episodes for different combinations of the parameters α and γ. The environment contains only sand obstacles.*



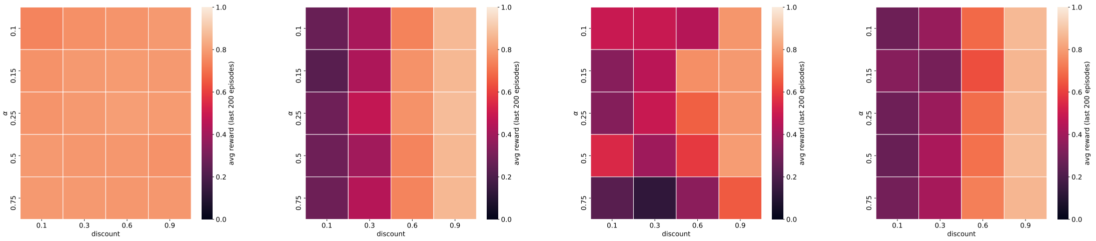
(A) *Softmax: False Sarsa: False*

(B) *Softmax: True Sarsa: False*

(C) *Softmax: False Sarsa: True*

(D) *Softmax: True Sarsa: True*

Figure 8: *Heatmaps of the average reward in the last 200 episodes for different combinations of the parameters α and γ. The environment contains only wall obstacles.*



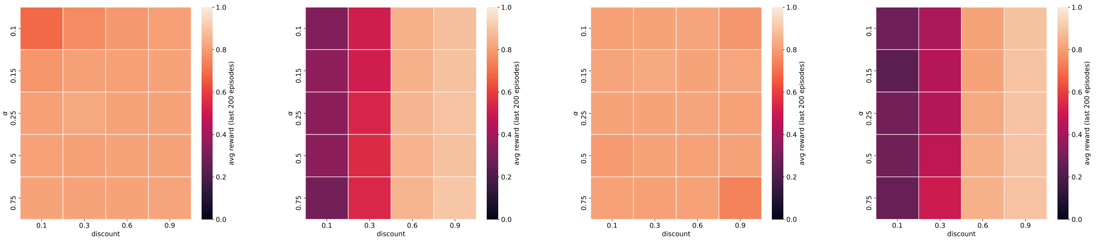
(A) *Softmax: False Sarsa: False*

(B) *Softmax: True Sarsa: False*

(C) *Softmax: False Sarsa: True*

(D) *Softmax: True Sarsa: True*

Figure 9: *Heatmaps of the average reward in the last 200 episodes for different combinations of the parameters α and γ. The environment contains no obstacles.*



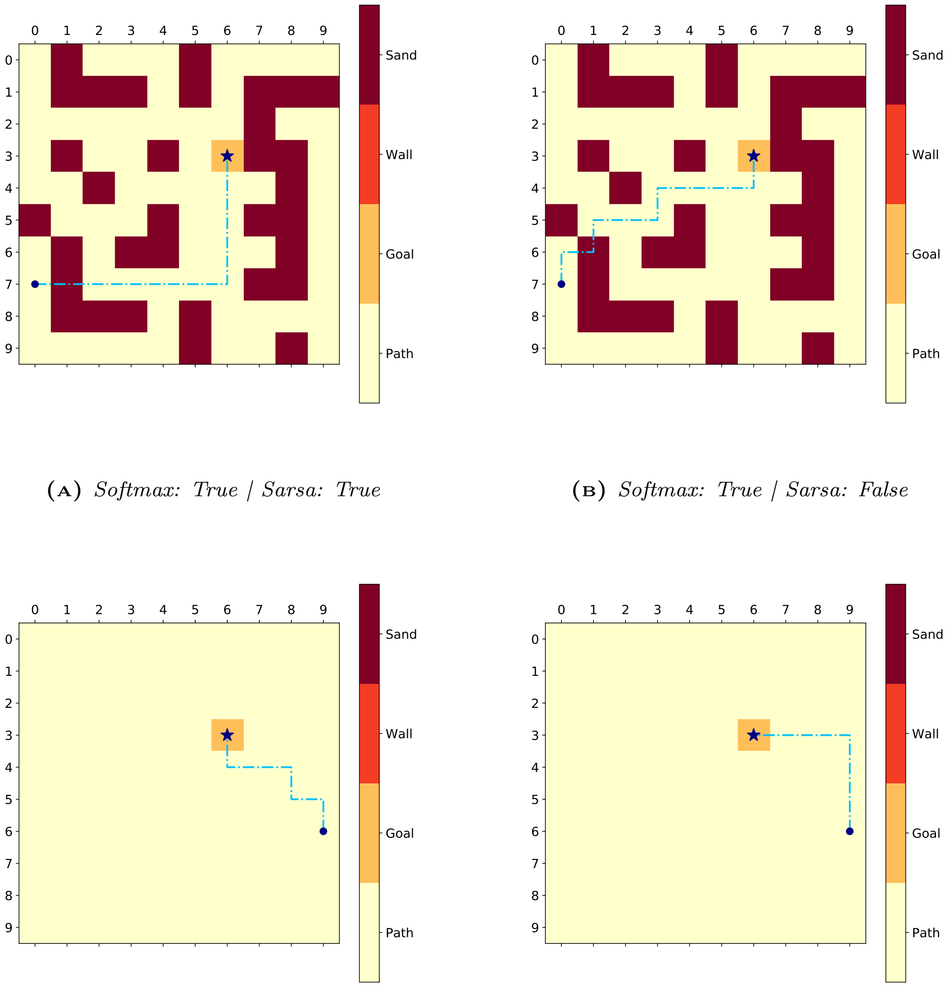
(A) *Softmax: False Sarsa: False*

(B) *Softmax: True Sarsa: False*

(C) *Softmax: False Sarsa: True*

(D) *Softmax: True Sarsa: True*

Figure 10: *Path chosen by the agent to reach the goal starting from the blue dot (star is the final position). Learning rate is set to 0.25 and discount factor to 0.9.*



(A) *Softmax: True / Sarsa: True*

(B) *Softmax: True / Sarsa: False*

(C) *Softmax: False / Sarsa: True*

(D) *Softmax: False / Sarsa: False*