Course 2, Module 5
Planning, Learning & Acting

CMPUT 397
Fall 2019

Any questions about course admin?

- Link for questions:
  - **http://www.tricider.com/brainstorming/3LEf4A3lOPB**

Mini-test

Q1) Given a choice between two actions, the agent should always pick the one with larger _____.

a) reward

b) return

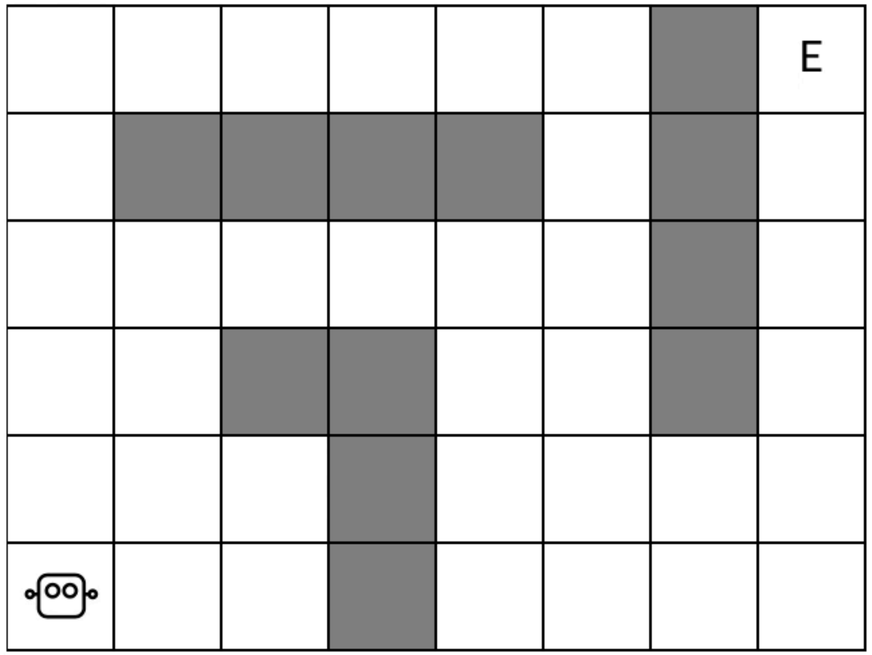
c) value

Q2) The goal of reinforcement learning can be seen as producing a _____, which maps from _____ to _____ .

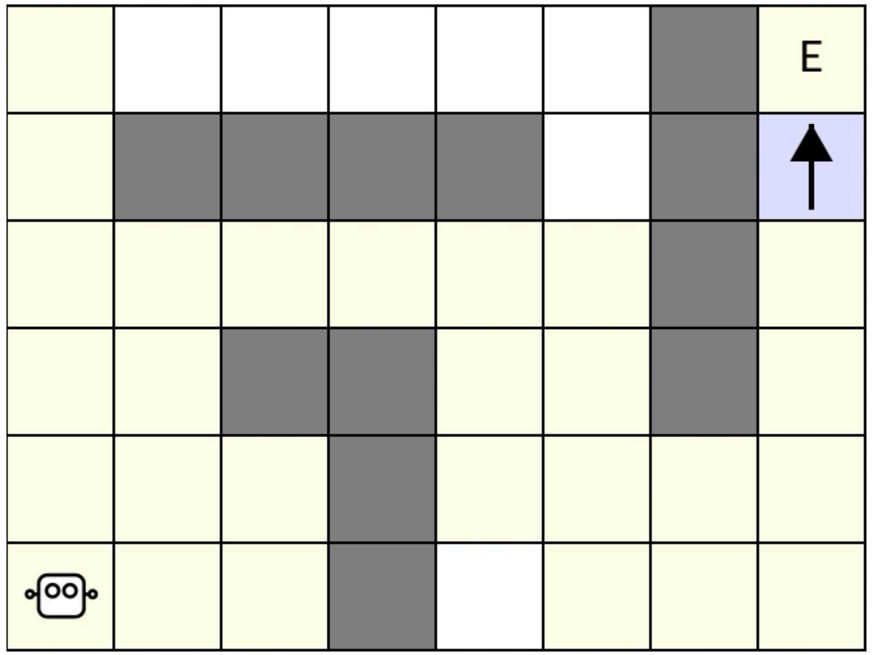
Q3) T F If a policy π is greedy with respect to its own value function, Vπ, then it is an optimal policy.

Tabular Dyna-Q

Number of actions taken: 0

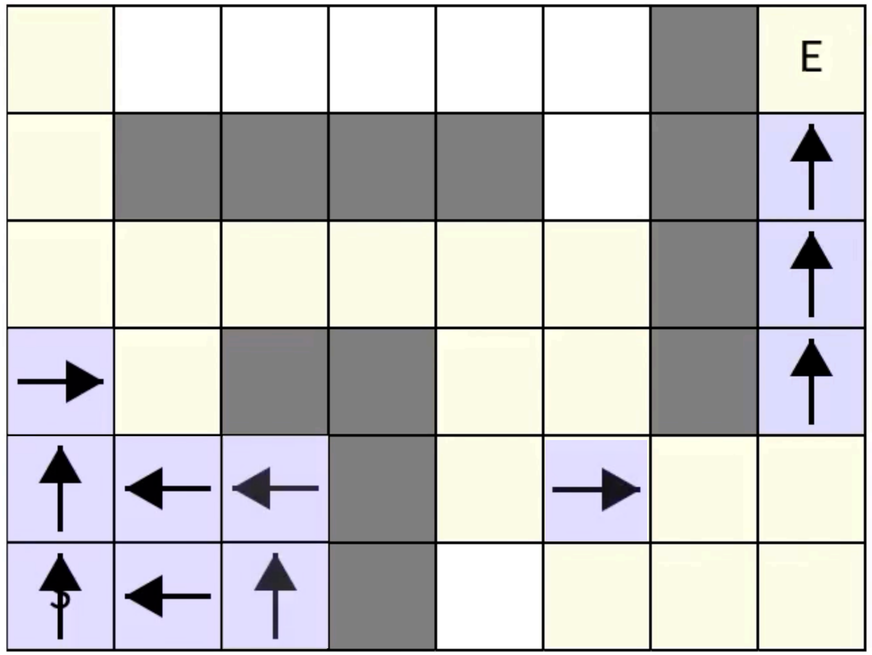


Number of actions taken: 184



Number of steps planned: 100

Number of actions taken: 185



Worksheet Questions

1. An agent observes the following two episodes from an MDP,

$$S_0 = 0, A_0 = 1, R_1 = 1, S_1 = 1, A_1 = 1, R_2 = 1$$

$$S_0 = 0, A_0 = 0, R_1 = 0, S_1 = 0, A_1 = 1, R_2 = 1, S_2 = 1, A_2 = 1, R_3 = 1$$

   and updates its deterministic model accordingly. What would the model output for the following queries:

(a) Model($S = 0, A = 0$):

(b) Model($S = 0, A = 1$):

(c) Model($S = 1, A = 0$):

(d) Model($S = 1, A = 1$):

2. An agent is in a 4-state MDP, $\mathcal{S} = \{1, 2, 3, 4\}$, where each state has two actions $\mathcal{A} = \{1, 2\}$. Assume the agent saw the following trajectory,

$$
\begin{aligned}
&S_0 = 1, A_0 = 2, R_1 = -1, \\
&S_1 = 1, A_1 = 1, R_2 = 1, \\
&S_2 = 2, A_2 = 2, R_3 = -1, \\
&S_3 = 2, A_3 = 1, R_4 = 1, \\
&S_4 = 3, A_4 = 1, R_5 = 100, \\
&S_5 = 4
\end{aligned}
$$

and uses Tabular Dyna-Q with 5 planning steps for each interaction with the environment.

(a) Once the agent sees S_5, how many Q-learning updates has it done with **real experience**? How many updates has it done with **simulated experience**?

$$
\begin{aligned}
&S_0 = 1, A_0 = 2, R_1 = -1, \\
&S_1 = 1, A_1 = 1, R_2 = 1, \\
&S_2 = 2, A_2 = 2, R_3 = -1, \\
&S_3 = 2, A_3 = 1, R_4 = 1, \\
&S_4 = 3, A_4 = 1, R_5 = 100, \\
&S_5 = 4
\end{aligned}
$$

(b) Which of the following are possible (or not possible) simulated transitions $\{S, A, R, S'\}$ given the above observed trajectory with a deterministic model and random search control?

i. $\{S = 1, A = 1, R = 1, S' = 2\}$

ii. $\{S = 2, A = 1, R = -1, S' = 3\}$

iii. $\{S = 2, A = 2, R = -1, S' = 2\}$

iv. $\{S = 1, A = 2, R = -1, S' = 1\}$

v. $\{S = 3, A = 1, R = 100, S' = 5\}$