RL Final exam

Alessandro Lodi (274425)

Part a

We assume that all actions are possible in any state so when the agent tries to perform an illegal action (e.g. pass a wall) it remains in the same state.

---

a. $S = \{(x, y, tool) \mid x \in [1, 4], y \in [1, 4], tool \in [0, 1] \; with \; x, y, tool \in \mathbb{N}\}$ where x and y represent the agent's position on the grid and $tool$ indicates whether the agent has taken the cooking tool. So the number of states $|S| = 32$. The state (4,4,1) is the absorbing state.

---

b. $A = \{left, right, up, bottom\}$ the agent as required can only move in 4 directions so $|A| = 4$

---

c. With S and A defined as above the dimensionality of the transition function is $32x32x4$

---

d. (Marked with "a. ACTION" in the spreadsheet)

---

e. The reward function is set to -1 for all states and actions except for:

- $R((4, 1, 1)|(4, 2, 1), right) = 10$
- $R((1, 3, 1)|(1, 2, 0), up) = 10$

and we can set $\gamma = 1$.

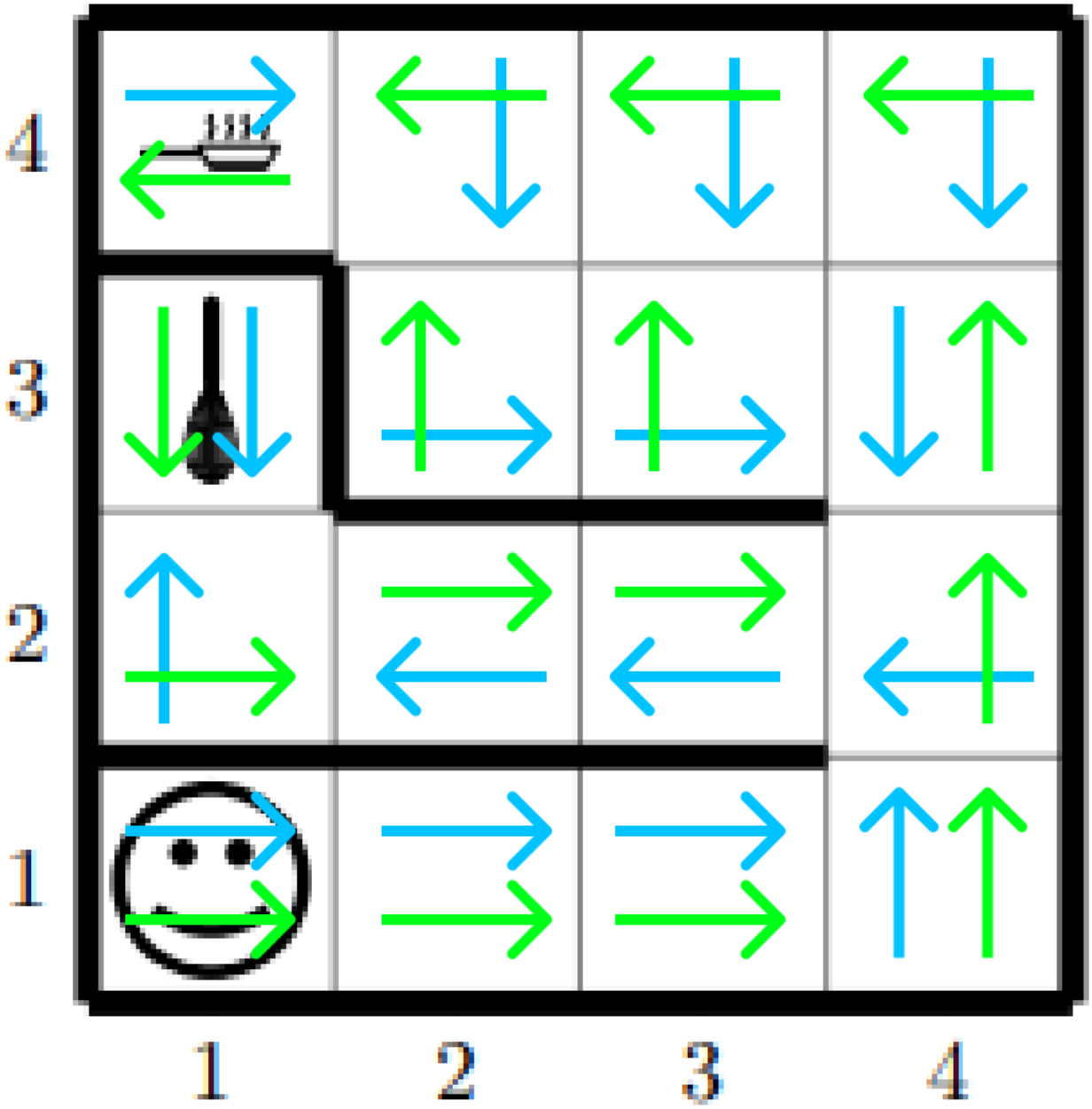
---

f. No because for each $\gamma > 0$ the reward of the actions that picks the tool and cooks will propagate in the value function

---

g. $|A|^{|S|} = 4^{32}$

h. In the image below the blue lines are associated with the states with $tool = 0$ while the green lines with states with $tool = 1$



---

i. Deterministic

---

i. Generally speaking stochastic policies can be better when we have a partially observable state or when the environment is stochastic. So in this case we have no advantages.

Part b

a. (Marked with "b. ACTION" in the spreadsheet)

---

b. In that case the probability of the optimal action is 50% so the agent will have no advantage in choosing other actions. Futhermore the remaining actions are opposed to each other and this does not make possible having an optimal policy that uses the wrong directions

---

c. Yes, the value of the optimal policy will change because the policy evaluation depends on the transition matrix