TD learning (SARSA and Q-learning)

1 Problem

For th maze 1 and the maze 2 shown in Fig.1 and Fig.2, the shortest path from the start (upper left) to the goal (lower right) is searched.

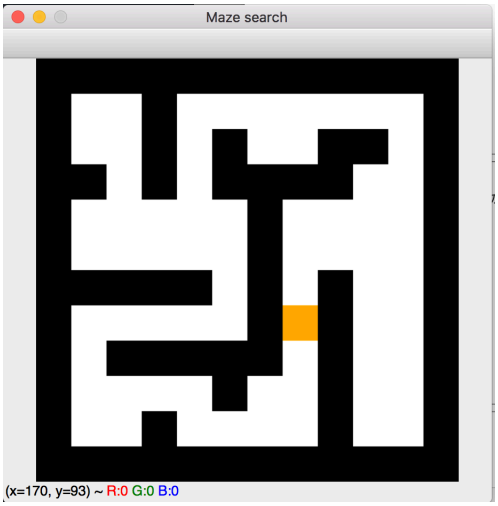

Fig. 1 Maze 1

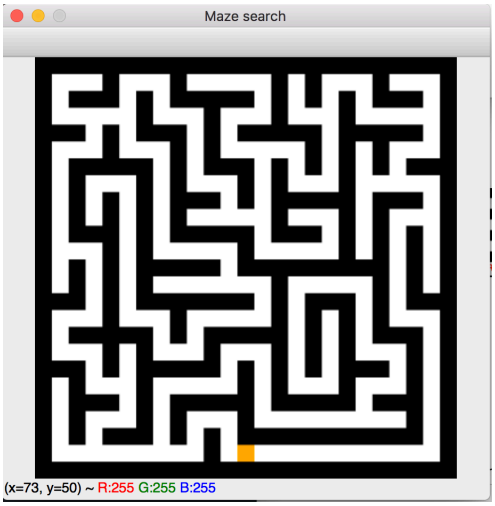

Fig. 2 Maze 2

2 Algorithm

Train the agent for automatic maze search by the SARSA and the Q-learning.

The agent is located at the start in the initial state, selects actions, obtains rewards r, updates the evaluation function V, and reaches the goal as the training episode. By repeating the learning episode, the agent's path converges (but not necessarily) to the shortest path. Stop the training when the performance is good enough.

The action selection uses the ϵ-greedy method. With the probability of $1-\epsilon$,

$$\arg\max_{x',y'}(r(x,y,x',y') + \gamma V(x',y')), \tag{1}$$

, where x, y is agent location and x', y' is masses to which agents can move next, and randomly select with the probability of ϵ.

The reward is $r = 1$ if (x', y') is the goal, and $r = -1$ otherwise.

In the SARSA, update the evaluation function as follow:

$$V(x, y) = V(x, y) + \alpha(r + \gamma V(x', y') - V(x, y)). \tag{2}$$

In the Q-learning, update the evaluation function as follow:

$$V(x, y) = V(x, y) + \alpha(r + \gamma \max V(x', y') - V(x, y)) \tag{3}$$

The parameters are set as $\gamma = 0.95, \alpha = 0.1, \epsilon = 0, 0.1, 0.5$. The evaluation function is initialized as $V(x, y) = 0$.