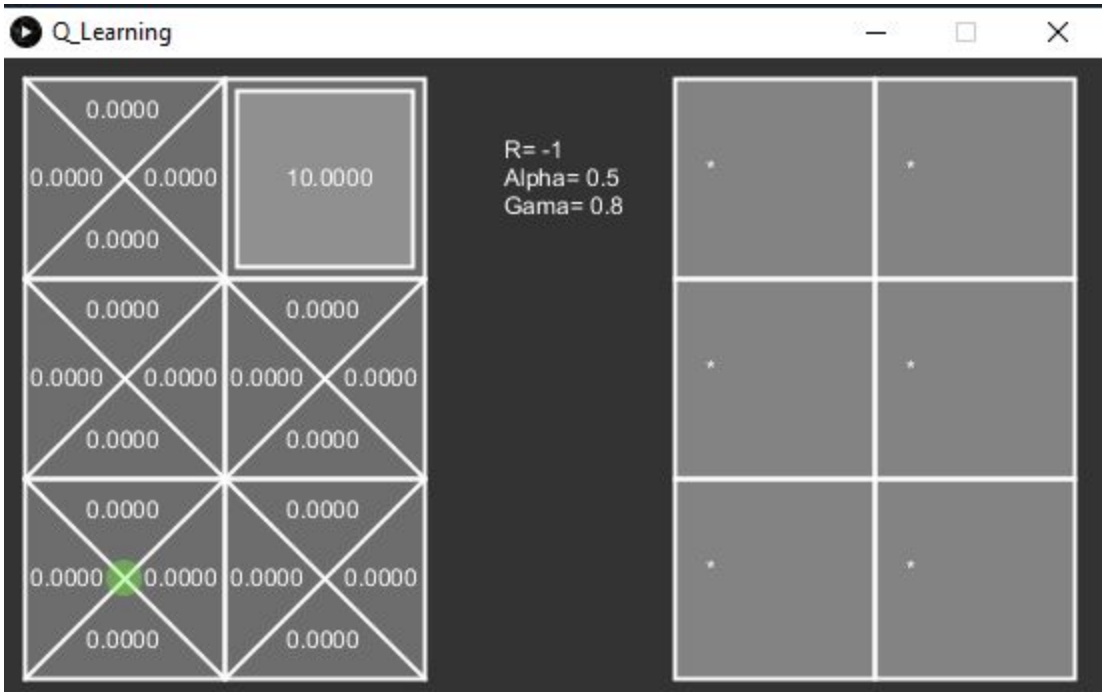
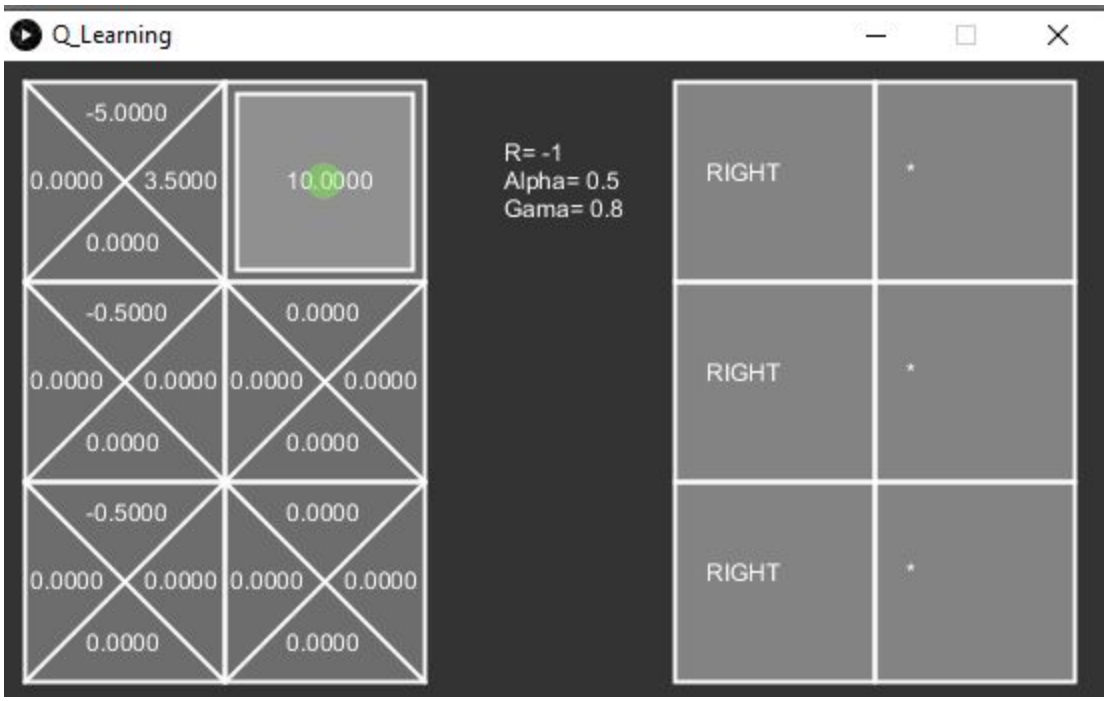
Considere o ambiente 3 x 2 onde a posição 6 é o estado terminal:

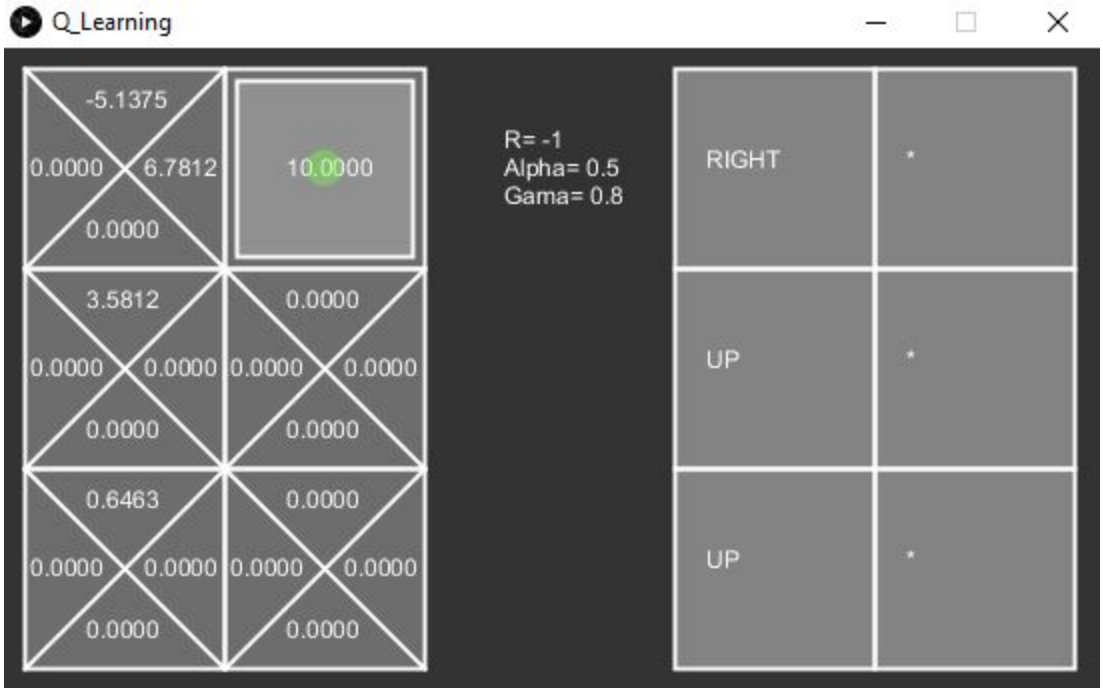
- Assuma que as ações **UP**, **DOWN**, **LEFT** e **RIGHT** são determinísticas;
- Recompensas:
    - **+10** no estado 6
    - **-10** se bater na parede
    - **-1** nos outros casos
- Aplicar o Q-learning sequencialmente usando as seguintes trajetórias:
    - Estado **inicial 1**, sequência **U**, **U**, **U**, **R**
    - Estado **inicial 5**: sequência **R**, **R**, **L**, **U**
- Inicialize a matriz Q com zeros e assuma **alpha = 0.5** e **gamma = 0.8**

1. Estado **inicial 1**, sequência **U**, **U**, **U**, **R:**

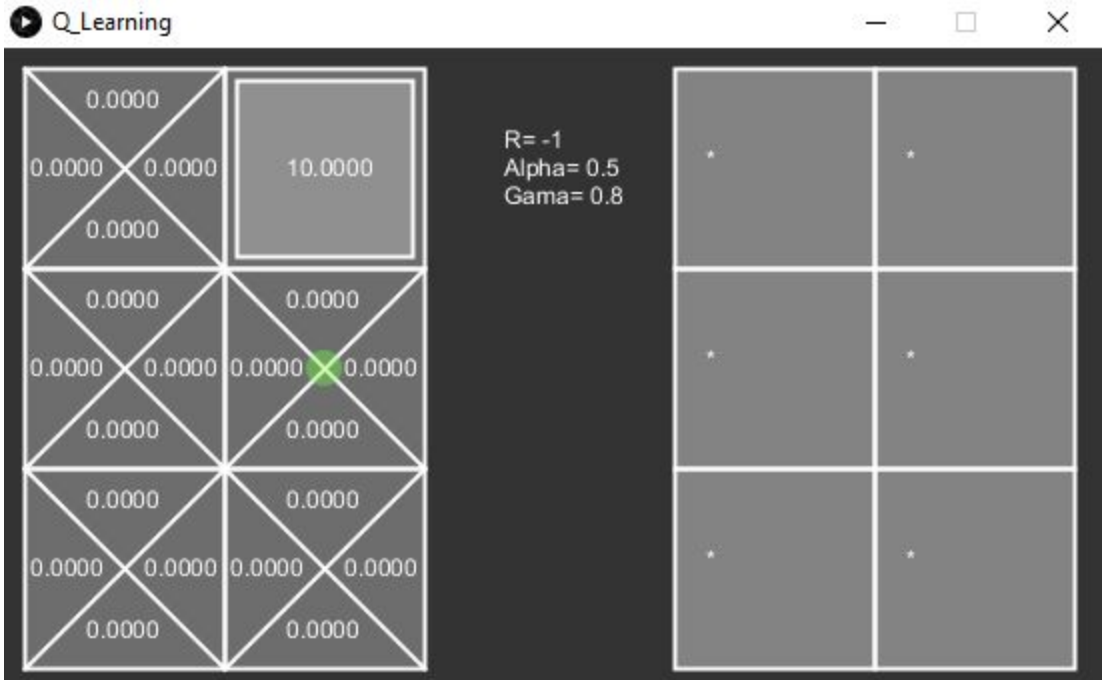


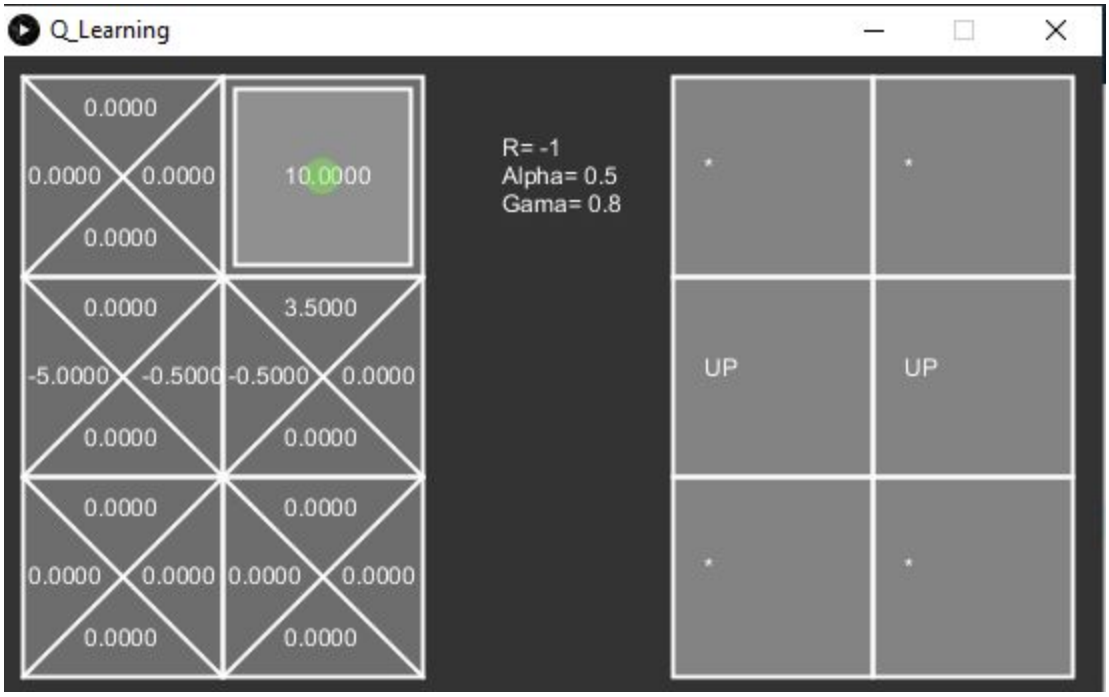


Ao executar a sequência a cima (**U**, **U**, **U**, **R**), em apenas uma única interação não é o suficiente para convergir nos resultados, porem, após 5 interações o sistema consegue gerar uma política e um caminho para o estado terminal.

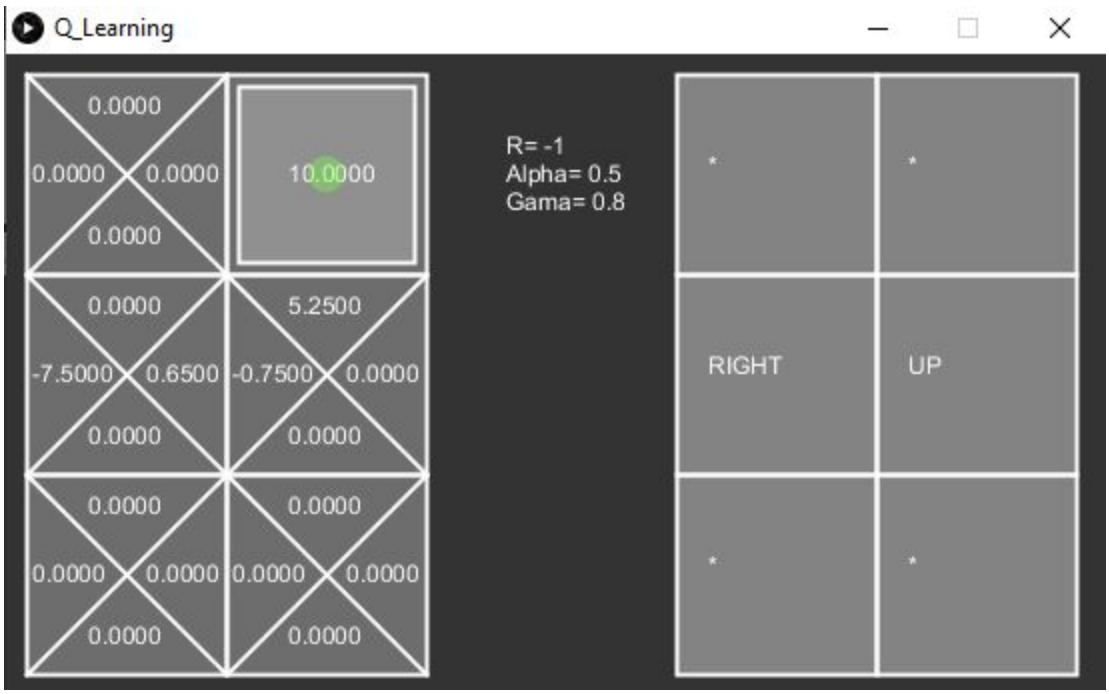


2. Estado **inicial 5**: sequência **L**, **L**, **R**, **U:**



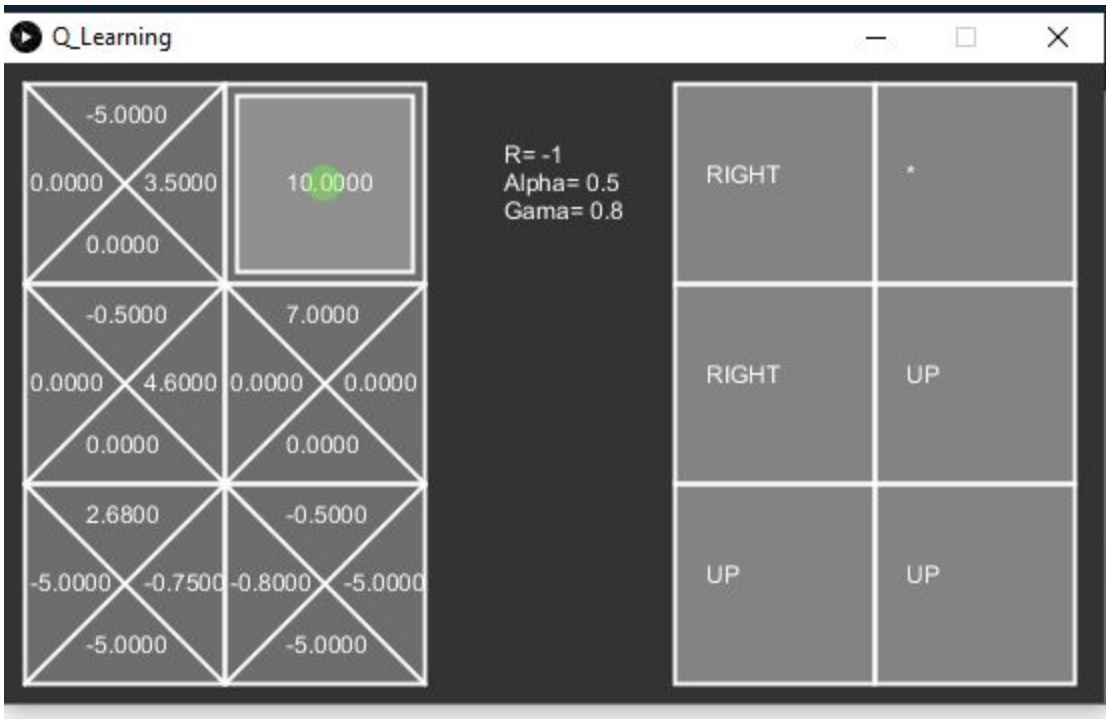


Assim como na sequência acima(**L**, **L**, **R**, **U**), ao executar a sequência a cima, em apenas uma única interação não é o suficiente para convergir nos resultados, porem com 2 interações já é o suficiente para se desenvolver uma política que leve ao estado terminal.



Em uma execução experimental onde o agente explora seguindo pelo caminho que tenha sempre um valor de utilidade maior para o estado atual, e em casos de empate e

resolvido verificando os estado de na ordem oraria(UP, RIGHT, DOWN, LEFT) o agente leva varia interações partindo de um estado inicial (estado 0), para alcançar o estado terminal 6.



O código foi desenvolvido em Python utilizando o Processing (Disponível https://processing.org/) e está em anexo ao exercício.