Reinforcement Learning

Stefano Perenzoni

Stephen Moran

**Grade E – D**

**1.1**

S set of states of the environment that the agent can reach. It’s a grid world. The number of total states is given by number of rows by number of columns.

A all possible actions the agent can carry in a state. 5 possible actions (left, right, up, down, stay). Some might be not available while the agent is in some particular states.

**2.2**

The agent is not able of completing the game

**4.1**

Trying different hyper parameters. Observe the number of episodes.

Number of episodes limited by episode\_max = 10000 AND diff < threshold (10^-6)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Learning rate | Discount | Number of episodes | Total steps | Total reward | Notes |
| 0.1 | 0.1 | 285 | 25878 |  | **Default configuration**  Update speed balanced  Reward prioritization balanced |
| 0.0001 | 0.1 | 5180 | 468670 |  | **Slow learning** |
| 0 | 0.1 | 1 | 56 |  | **Not learning**  Q and Q\_old are the same 🡪 diff doesn’t change and the algorithm concludes |
| 0.8 | 0.1 | 30 | 2708 |  | **High variance, fast update** Last episode diff goes from 3.66\*10^-4 to 1.91\*10^-7 triggering the threshold and the stopping criteria. |
| 0.1 | 1 | 1511 | 137996 | -43 | **Low variance & high long-term return**  Update speed balanced  Long term return is preferred to instant reward |
| 0.8 | 1 | 224 | 20371 | -226 | **High variance & high long-term return** |