Solução de reinforcement learning para problema de transporte de objeto

1. Modelagem do MDP:

(a) Apresente a modelagem de estados considerada, bem como a quantidade de estados presentes no MDP. Inclua na contagem os estados não-válidos;
Estados: posição x do agente, posição y do agente, pegou o objeto, posição x do objeto, posição y do objeto.
Quantidade de estados presentes (incluindo os não válidos) = 7 x 6 x 2 x 7 x 6 = 3582

(b) Apresente a modelagem das ações que o agente pode executar;
Ações: Ir para cima, ir para baixo, ir para a direita, ir para a esquerda e ficar onde está.

(c) Apresente a modelagem da função de recompensa, com as situações em que o agente é recompensado bem como a magnitude da recompensa. Justifique as suas escolhas.
-0,5 quando o agente se desloca para qualquer célula válida;
0 para posição de captura do objeto (para incentivar a captura do objeto);
0 quando o agente se desloca para a base (sem o objeto entrar na base);
1 para quando o objeto entra na base;
Essa política visa favorecer o agente capturar o objeto e depois seguir até a base junto com o objeto.

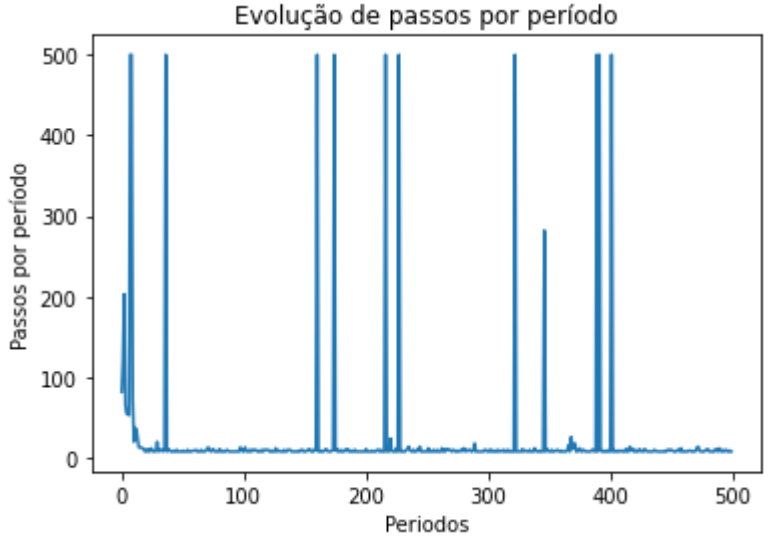
2. Configuração dos Experimentos

(a) Apresente os valores de taxa de aprendizagem (alfa) e fator de desconto (gamma) do algoritmo de aprendizagem Q-Learning;
Alfa = 0,3
Gama = 0.5

(b) Apresente as configurações do horizonte de aprendizagem, que é representado pela quantidade máxima de passos de tempo por episódios, quantidade máxima de episódios, e política de exploração ao longo do tempo;
Quantidade máxima de passos por episódio = 500
Quantidade máxima de episódios = 500
Probabilidade de escolher q-value aleatório (exploração) = 10%
Probabilidade de escolher melhor q-value (exploração) = 90%

3. Resultados Experimentais

(a) Apresente a curva de convergência, representada pela quantidade de passos (timesteps)necessários para resolver a tarefa ao longo do tempo (episódios).



(b) Apresente o tempo de processamento necessário para resolver o problema.