Acadêmico: Leonardo Fiedler

Implemente uma solução via reinforcement learning para o problema de transporte de objeto e
apresente um relatório endereçando os seguintes aspectos da solução:

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;

1. Estados possíveis
   * Posição do agente
   * Posição do objeto
   * Está agarrado ao objeto

(b) Apresente a modelagem das ações que o agente pode executar

1. Lista de ações do agente

   * Mover para cima
   * Mover para a esquerda
   * Mover para a direita
   * Mover para baixo
   * Permanecer na mesma célula

(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.

R.: O agente tem 2 momentos em que é recompensado:

1. Quando chega até o objeto
2. Quando está agarrado com o objeto e chega a base

A ideia de criar este mecanismo é para acelerar o reconhecimento e a compreensão do agente sobre o cenário. Pois quando não está com o objeto, ao chegar nas alças esquerda ou direita, o agente é recompensado. A partir daí, o agente é recompensado apenas quando o objeto chega na base.

Quando o agente chega até o objeto a recompensa é de 40, já quando chega ao estado final, é 100. A ideia é criar uma tendência nas laterias, mas não muito alta para que o agente não repita o mesmo trajeto e possa explorar outras possibilidades e na base a recompensa é maior justamente para direcioná-lo para lá.

2. Configuração dos Experimentos

(a) Apresente os valores de taxa de aprendizagem (alfa) e fator de desconto (gamma) do algoritmo de aprendizagem Q-Learning;

1. Alfa: 0.9 - Aqui a ideia é focar em ações mais recentes e com o tempo, passe a otimizar o melhor caminho mais rapidamente
2. Gamma: 0.9 - A escolha se deu para que o algoritmo considere futuras recompensas, com o intuito de reduzir o número de episódios

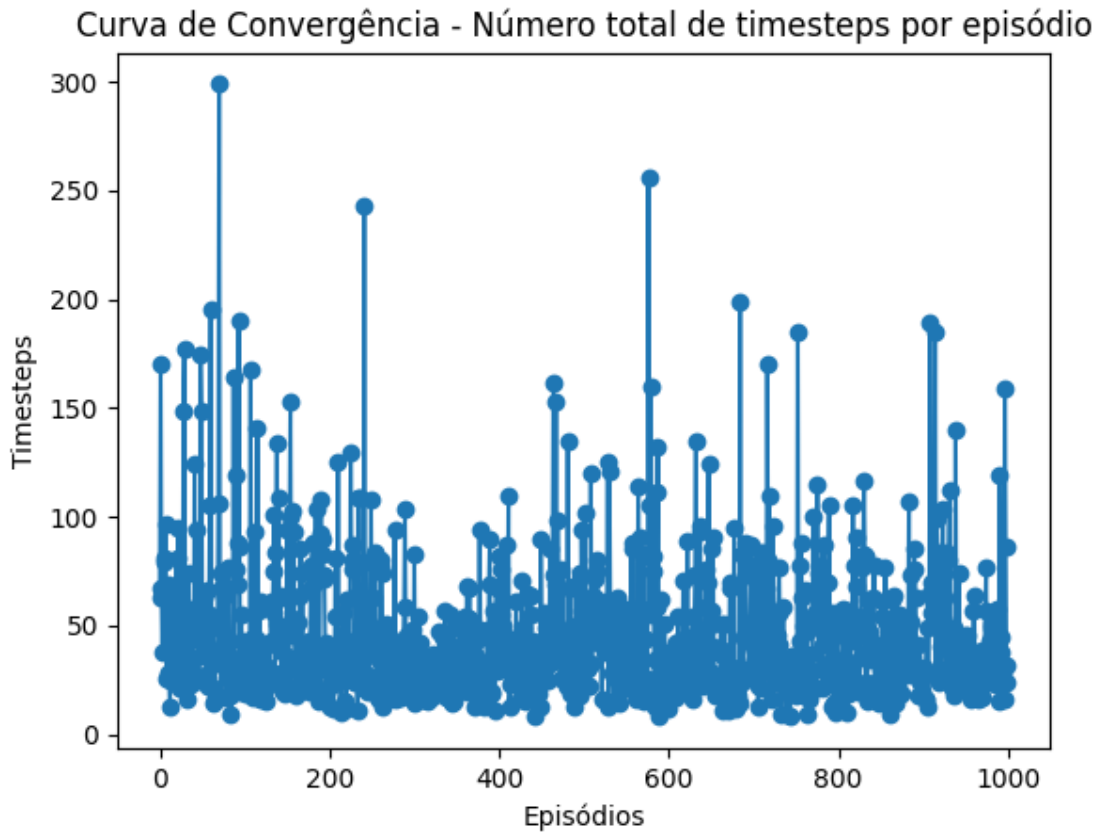
(b) Apresente as configurações do horizonte de aprendizagem, que é representado pela quantidade máxima de passos de tempo por episódios, quantidade máxima de episódios, e política de exploração ao longo do tempo;

1. Horizonte de aprendizado: 300 - Por conta da ação de parado na mesma célula, as primeiras iterações necessitam de maior atenção, portanto, mais passos gasta-se para chegar ao objetivo e explorar o cenário
2. Quantidade máximo de episódios: 1000
3. Política de exploração ao longo do tempo: 0.3 - A ideia é que o algoritmo, ao passar do tempo, tente pegar mais vezes o melhor caminho do que uma ação aleatória

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.



Para os tempos, pode-se considerar:

1. Visualização apenas do menor caminho ao final: 5 segundos
2. Sem qualquer tipo de visualização gráfica e sem cálculo do menor caminho: 4 segundos

OBS: Para a visualização gráfica, foi adicionado um tempo de 500ms a cada exibição, com intuito de ser plausível a visualização do passo a passo.