UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE

INSTITUTO METRÓPOLE DIGITAL

IMD1103 - APRENDIZADO POR REFORÇO

CHARLES ANDRYE GALVAO MADEIRA

Relatório 1

Relatório descritivo de análise experimental com métodos de diferenças temporais tabulares

DOUGLAS FELIPE DE LIMA SILVA

NATAL-RN

2024

1. Contexto da atividade

Este relatório apresenta os resultados das análises experimentais de métodos de diferenças temporais (Q-learning e SARSA) nos ambientes simulados: Blackjack, Cliff Walking, Frozen Lake e Taxi que fazem parte do conjunto Toy Text do Gymnasium e proporcionam cenários para a aplicação de algoritmos de aprendizado por reforço tabular. O objetivo deste trabalho é avaliar e comparar o desempenho dos algoritmos Q-learning e SARSA em diferentes configurações de hiperparâmetros A análise visa identificar as melhores práticas e as configurações de hiperparâmetros mais eficazes para cada ambiente, destacando o impacto de cada variável nos resultados finais.

2. Objetivos

- **I.** Avaliar o desempenho dos métodos Q-learning e SARSA nos ambientes Blackjack, Cliff Walking, Frozen Lake e Taxi
- **II.** Investigar como variações na taxa de aprendizado, taxa de exploração, número de episódios de treinamento e uso de histórico afetam o desempenho dos algoritmos
- **III.** Determinar quais configurações de hiperparâmetros proporcionam os melhores resultados em cada ambiente
- **IV.** Utilizar gráficos e animações para demonstrar os resultados

3. Metodologia

1) Inicialização do Ambiente e dos Agentes

a) Os ambientes utilizados foram Blackjack-v1, CliffWalking-v0, FrozenLake-v1, e Taxi-v3 do Gymnasium

b) Classes de agentes foram implementadas utilizando os algoritmos

- **i)** **Q-learning**: Algoritmo que atualiza a política do agente diretamente pela maximização da função valor
- **ii)** **SARSA**: Algoritmo que atualiza a política do agente considerando a ação futura escolhida pelo próprio agente

2) Configuração dos Hiperparâmetros Funções de Treinamento

a) Foram analisados os hiperparametros learning_rates, n_episodes_list, start_epsilons, epsilon_decays, final_epsilons e use_history_options. Todas as combinações de hiperparametros foram utilizadas para treinamento e avaliação dos resultados

b) Foram definidas funções train_agent_qlearning e train_agent_sarsa adaptadas para o treinamento dos agentes com o método específico

3) Execução dos Experimentos

a) Iteração sobre todas as combinações de hiperparâmetros

b) Treinamento dos agentes em paralelo (multiprocessing) para acelerar o processo de experimentação

c) Obtenção dos resultados de desempenho (recompensas médias) para cada configuração de hiperparâmetros

4) **Análise dos Resultados**

a) **Estatísticas**: Foram calculadas a média e desvio padrão das recompensas para cada configuração

b) **Visualização dos Resultados**

i) **Heatmap**: Representação das médias das recompensas para cada combinação de hiperparâmetros

ii) **Gráficos de Linha**: Representação dos melhores e piores resultados para cada configuração

iii) **Melhor Resultado**: Gráficos individuais para o melhor resultado obtido no heatmap e gráfico de linha

iv) **Políticas e Valores de Estado**: Gráficos que representam valores de estado e política aprendida pelo agente do melhor agente treinado

v) **Animação:** GIF renderizado para representar a atuação do agente no ambiente

5) **Link dos Experimentos**

BlackJack e FrozenLake

Cliff Walking e Taxi

4. Resultados

1. Black Jack

Intervalos de Hiperparamtros

learning_rates = [0.01, 0.05, 0.1]

n_episodes_list = [5000,10000,20000]

start_epsilons = [1.0, 0.5]

epsilon_decays = [0.1, 0.01]

final_epsilons = [0.1, 0.01]

use_history_options = [True, False]

A análise dos resultados mostra que a taxa de aprendizado (LR) tem uma influência significativa no desempenho dos algoritmos. Taxas de aprendizado mais baixas (0.01) resultaram em recompensas médias ligeiramente melhores para ambos os algoritmos, SARSA e Q-learning. O aument da taxa de aprendizado para 0.1 resultou em uma piora na recompensa média, indicando uma piora no desempenho do agente. Quanto ao número de episódios (NE), aumentar para 100000 não mostrou melhorias significativas em comparação com 10000 ou 50000 episódios, isso pode sugerir que há um ponto de saturação onde mais episódios não contribuem para um melhor aprendizado A variação dos hiperparâmetros de exploração, incluindo o epsilon inicial (SE), o decaimento de epsilon (ED) e o epsilon final (FE), não demonstrou impactos muito claros no desempenho, indicnado queesses hiperparâmetros podem ser menos críticos em comparação com a taxa de aprendizado e o número de episódios Em relação ao uso de histórico, todos os resultados com melhor média de recompensa foram sem utilização de histórico, sugerindo que para o ambiente Blackjack o estado atual é suficiente para a tomada de decisão e a inclusão de histórico adicional confere vantagens ao agente.

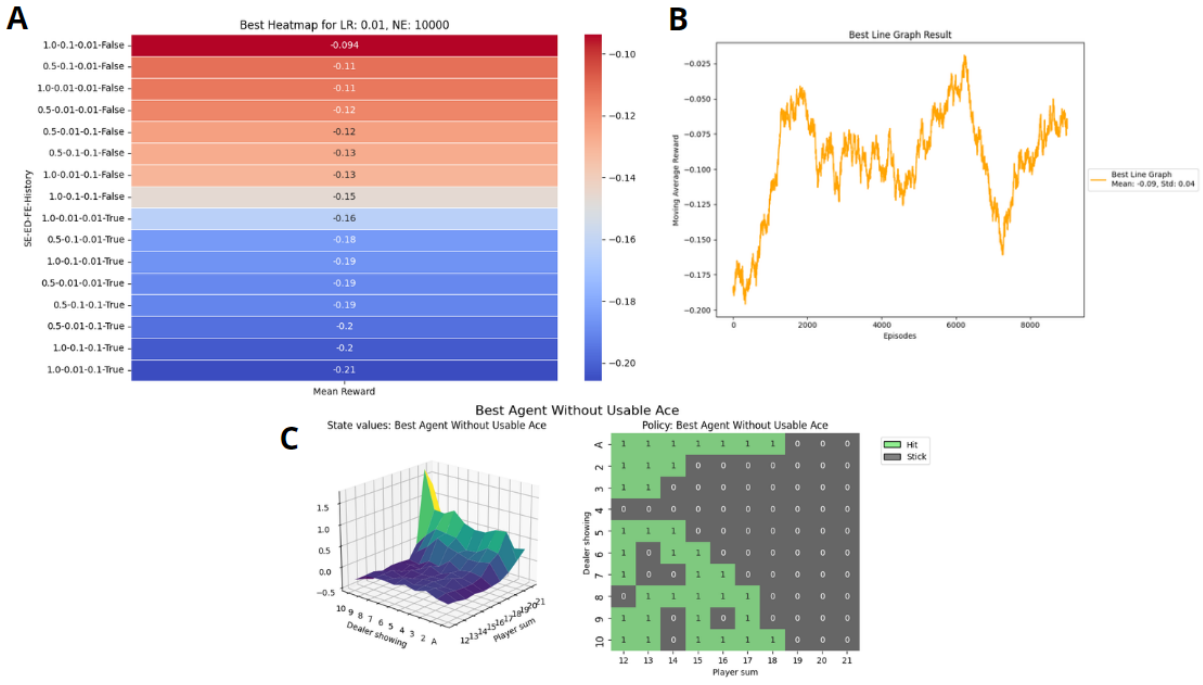

Figura A: Melhor Heatmap Individual para o Método Q-learning. O eixo Y representa as combinações de epsilon inicial (SE), decaimento de epsilon (ED), epsilon final (FE) e o uso de histórico (True ou False).

Figura B: Melhor Gráfico de Linha Individual para o Método Q-learning. A linha laranja mostra a recompensa média em movimento ao longo dos episódios.

Figura C: State Values e Policy do Melhor Agente para o Método Q-learning

O gráfico 3D à esquerda mostra os valores dos estados, enquanto o heatmap à direita mostra a política do agente. Cada célula do heatmap indica a ação preferida (Hit ou Stick) para uma combinação específica de soma do jogador e carta mostrada pelo dealer, destacando a eficácia da política aprendida pelo agente.

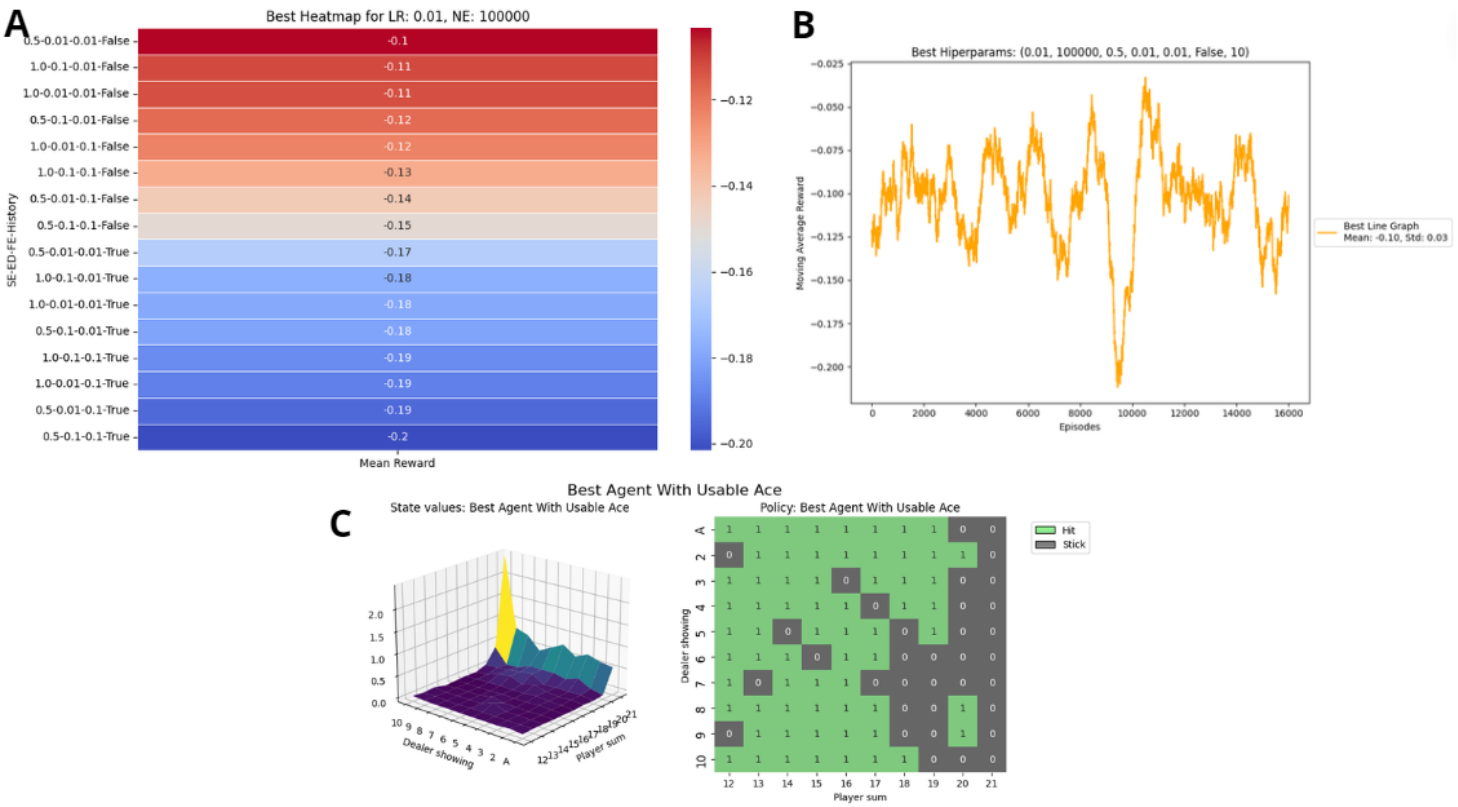

Figura A: Melhor Heatmap Individual para o Método SARSA. O eixo Y representa as combinações de epsilon inicial (SE), decaimento de epsilon (ED), epsilon final (FE) e o uso de histórico (True ou False).
Figura B: Melhor Gráfico de Linha Individual para o Método SARSA. A linha laranja mostra a recompensa média em movimento ao longo dos episódios.
Figura C: State Values e Policy do Melhor Agente para o Método SARS. O gráfico 3D à esquerda mostra os valores dos estados, enquanto o heatmap à direita mostra a política do agente. Cada célula do heatmap indica a ação preferida (Hit ou Stick) para uma combinação específica de soma do jogador e carta mostrada pelo dealer, destacando a eficácia da política aprendida pelo agente.

2. Frozen Lake

Intervalo de Hiperparâmetros

learning_rates = [0.01, 0.05, 0.1]
n_episodes_list = [5000,10000,20000]
start_epsilons = [1.0, 0.5]
epsilon_decays = [0.1, 0.01]
final_epsilons = [0.1, 0.01]
use_history_options = [True, False]

A análise dos resultados mostra que a taxa de aprendizado (LR) tem influência no desempenho dos algoritmos. Embora para a maioria das combinasções de hiperparametros a recompensa média tenha sido zero, taxas de aprendizado mais baixas (0.01) mostraram recompensas médias mais altas em ambos os algoritmos, principalmente em combinação com um número maior de episódios, indicando que o aumento no número de episódios (NE) resultou em recompensas médias maiores particularmente em Q-learning, sugerindo que com mais o agente explore melhor o ambiente e aprenda políticas mais eficazes. A variação dos hiperparâmetros de exploração SE, ED e FE, teve sua interpretação comprometida ua vez que a maioria das combinações apresentaram recompensa média 0. Assim como para o Black Jack, o uso de histórico não apresentou influência significativa no desempenho do agente, sugerindo que para o ambiente Frozen Lake, o estado atual também é suficiente para a tomada de decisão. Comparando os dois algoritmos, o Q-learning mostrou um desempenho superior ao SARSA para o melhor agente, com recompensas médias mais altas. O Q-learning mostrou-se relativamente equivalente ao SARSA em termos de desempenho geral com , especialmente com taxas de aprendizado mais baixas e um maior número de episódios, embora o Q-learning tenha apresentado uma recompensa média de 0.9 com LR = 0.01 e NE = 5000, valor bem próximo ao melhor resultado de 0.99 com LR = 0.01 e NE = 20000, alcançando uma recompensa média semelhante com 4 vezes menos NE, fenômeno que não é observado no SARSA.

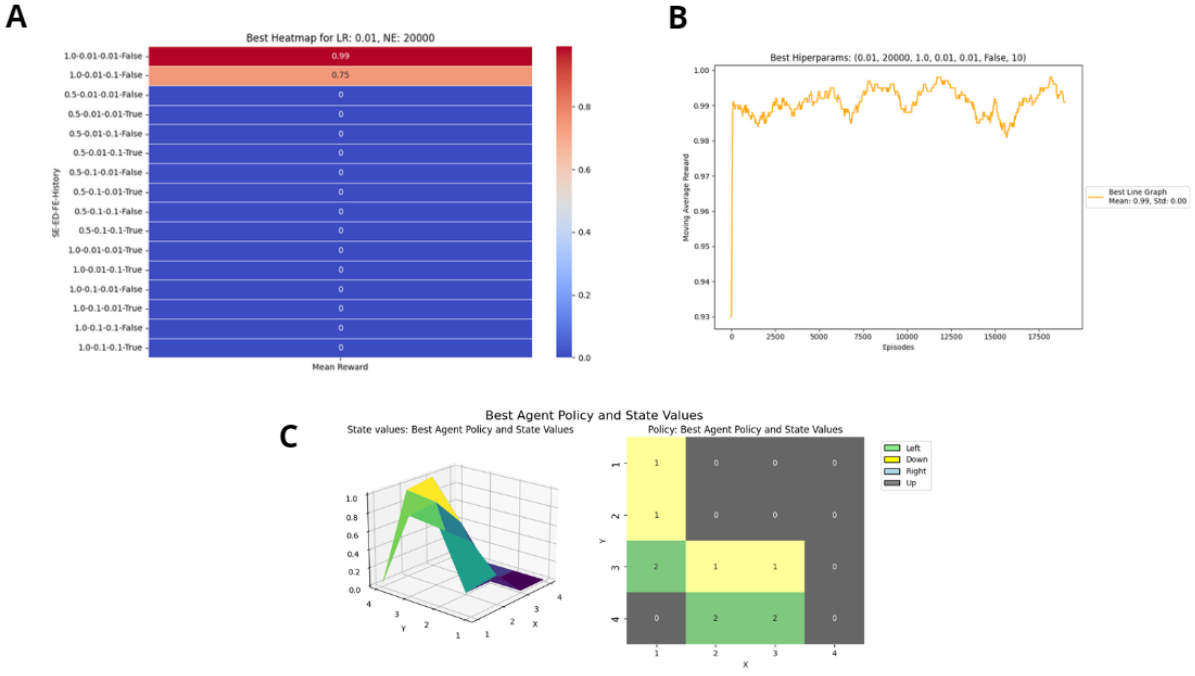

Figura A: Melhor Heatmap Individual para o Método Q-learning. O eixo Y representa as combinações de epsilon inicial (SE), decaimento de epsilon (ED), epsilon final (FE) e o uso de histórico (True ou False).
Figura B: Melhor Gráfico de Linha Individual para o Método Q-learning. A linha laranja mostra a recompensa média ao longo dos episódios.
Figura C: State Values e Policy do Melhor Agente para o Método Q-learning. O gráfico 3D à esquerda mostra os valores dos estados, enquanto o heatmap à direita mostra a política do agente. Cada célula do heatmap indica a ação preferida (Left, Down, Right, Up) para uma determinada posição no ambiente, destacando a eficácia da política aprendida pelo agente.

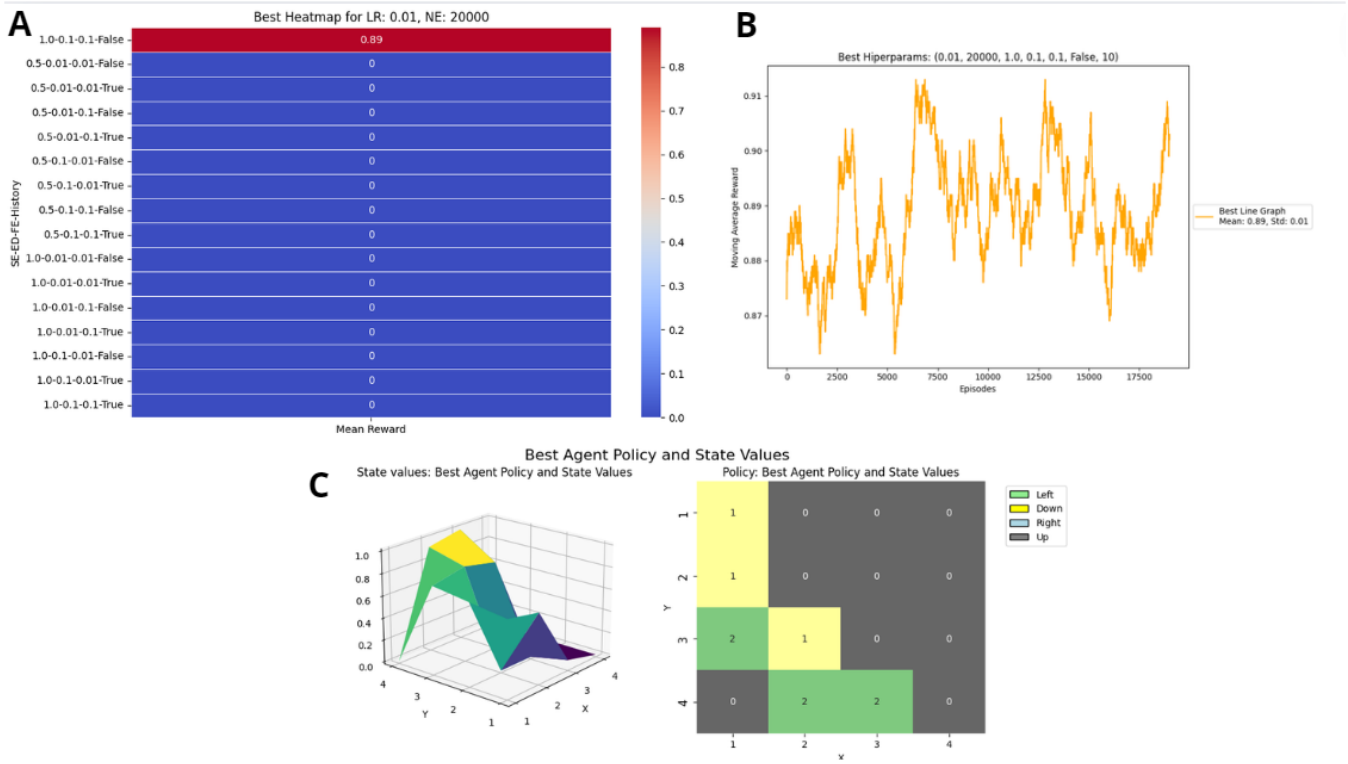

Figura A: Melhor Heatmap Individual para o Método SARSA. O eixo Y representa as combinações de epsilon inicial (SE), decaimento de epsilon (ED), epsilon final (FE) e o uso de histórico (True ou False).
Figura B: Melhor Gráfico de Linha Individual para o Método SARSA. A linha laranja mostra a recompensa média ao longo dos episódios.
Figura C: State Values e Policy do Melhor Agente para o Método SARSA. O gráfico 3D à esquerda mostra os valores dos estados, enquanto o heatmap à direita mostra a política do agente. Cada célula do heatmap indica a ação preferida (Left, Down, Right, Up) para uma determinada posição no ambiente, destacando a eficácia da política aprendida pelo agente.

3. Cliff Walking

Intervalo de Hiperparametros

learning_rates = [0.01, 0.05, 0.1] # Reduced for faster testing
n_episodes_list = [1000, 5000, 10000] # Reduced for faster testing
start_epsilons = [1.0, 0.5]
epsilon_decays = [0.1, 0.01]
final_epsilons = [0.1, 0.01]

Os resultados obtidos indicam que taxas de aprendizado mais baixas (0.01) resultaram em recompensas médias piores para ambos os algoritmos enquanto a taxa de aprendizado em 0.1 promoveu uma performance ligeiramente melhor, mas ainda negativa. O aumento do número de episódios para 5000 mostrou leve melhoria em relação a 1000 episódios, entretanto dobrar o número de episódios para 1000 em NE = 0.1 práticamente não mudou a média das recompensas, uma melhoria mais significativa foi observada com o aumento de NE para LR 0.05 e 0.01. Mais episódios proporcionaram uma maior oportunidade de aprendizagem, mas a convergência ainda foi limitada. Avaliando a variação dos hiperparâmetros de exploração SE, ED e FE a melhor combinação encontrada foi respectivamente 1.0 – 0.1 - 0.01, embora combinações distintas também tenham apresentado bons resultados. Nesse ambiente especifico não foi possível avaliar a influencia do histórico devido a limitações no colab, por algum motivo ao executar a célula sempre ficava travada. Comparando os 2 métodos, SARSA mostrou uma vantagem sobre Q-learning em termos de recompensas médias, mostrando resultados variando entre -20 e -44 enquanto Q-learning apresentou resultados entre -23 e -2.6^{+04}. Avaliando os valores das recompensas médias, aparentemente ambos os algoritmos tiveram dificuldades com as penalidades do ambiente, resultando em recompensas negativas.

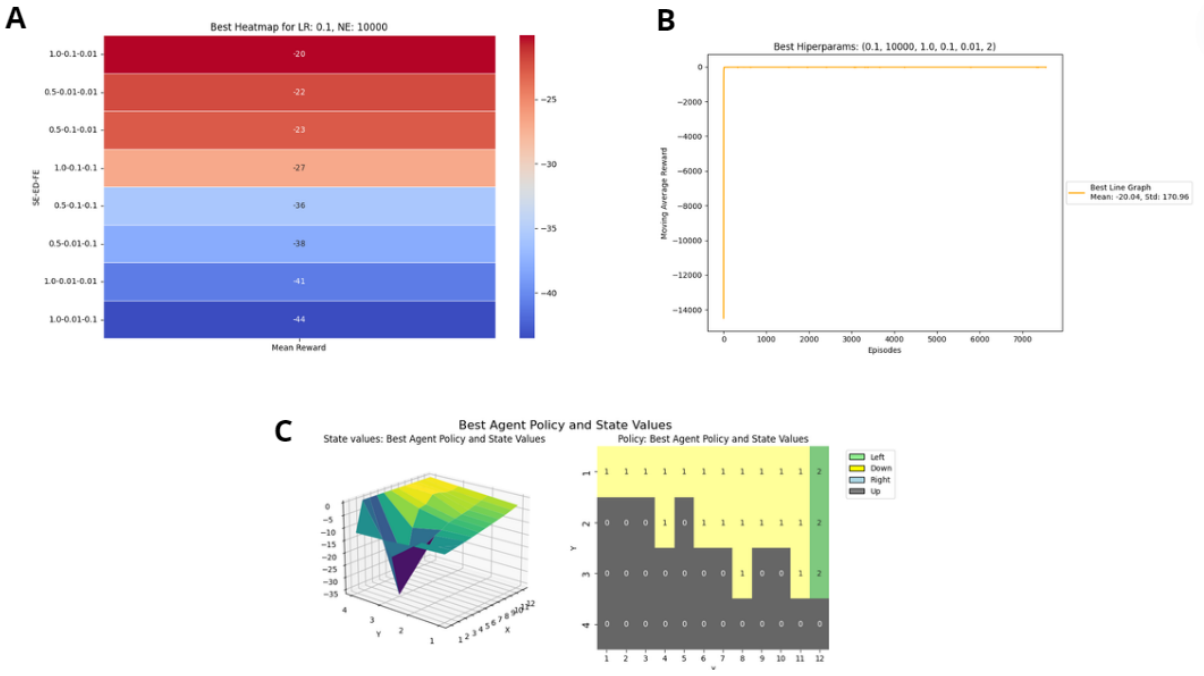

Figura A: Melhor Heatmap Individual para o Método Q-learning. O eixo Y representa as combinações de epsilon inicial (SE), decaimento de epsilon (ED), epsilon final (FE).
Figura B: Melhor Gráfico de Linha Individual para o Método Q-learning. A linha laranja mostra a recompensa média ao longo dos episódios.
Figura C: State Values e Policy do Melhor Agente para o Método Q-learning. O gráfico 3D à esquerda mostra os valores dos estados, enquanto o heatmap à direita mostra a política do agente. Cada célula do heatmap indica a ação preferida (Left, Down, Right, Up) para uma determinada posição no ambiente, destacando a eficácia da política aprendida pelo agente.

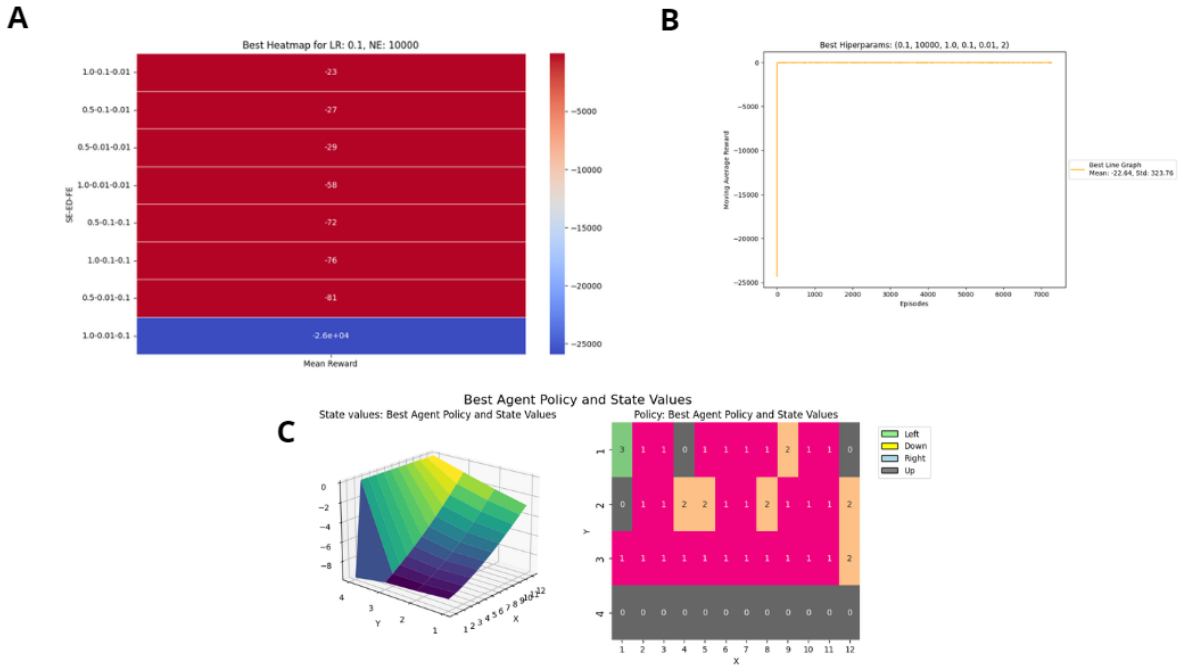

Figura A: Melhor Heatmap Individual para o Método SARSA. O eixo Y representa as combinações de epsilon inicial (SE), decaimento de epsilon (ED), epsilon final (FE) e o uso de histórico (True ou False).
Figura B: Melhor Gráfico de Linha Individual para o Método SARSA. A linha laranja mostra a recompensa média ao longo dos episódios.
Figura C: State Values e Policy do Melhor Agente para o Método SARSA. O gráfico 3D à esquerda mostra os valores dos estados, enquanto o heatmap à direita mostra a política do agente. Cada célula do heatmap indica a ação preferida (Left, Down, Right, Up) para uma determinada posição no ambiente, destacando a eficácia da política aprendida pelo agente.

4. Taxi

Intervalo de Hiperparâmetros

learning_rates = [0.01, 0.05, 0.1]

n_episodes_list = [1000, 5000,10000]

start_epsilons = [1.0, 0.5]

epsilon_decays = [0.1, 0.01]

final_epsilons = [0.1, 0.01]

A analise dos resultados indica que taxas de aprendizado mais baixas (0.01) resultaram em recompensas médias piores para ambos os algoritmos, o aumento na taxa de aprendizado para 0.1 resultou em uma performance substancialmente melhor para todos os números de episódios comparando com LR 0.05 e 0.01, mas ainda negativa, exceto para LR 0.1 e NE 10000. Aumentar o número de episódios para 5000 mostrou melhorias significativas em relação a 1000 episódios, indicando que mais episódios proporcionaram uma maior oportunidade de aprendizagem do agente, observa-se no gráfico de linha a evolução constante do agente e sua convergencia. A variação dos hiperparâmetros de exploração não apresentou tendências aparentes, indicando que os parametros mais criticos foram LR e NE. Nesse ambiente especifico também não foi possível avaliar a influencia do histórico devido a limitações no colab, por algum motivo ao executar a célula sempre ficava travada. Ambos os algoritmos tiveram dificuldades em superar as penalidades do ambiente, tendo em vista a complexidade maior do ambiente, com obstáculos presentes e fatores não deterministicos como a variação da possição do passageiro e ponto de desembarque, entretanto o Q-learning parece ser mais resistente a variações de hiperparâmetros, enquanto SARSA mostrou variações mais amplas no desempenho.

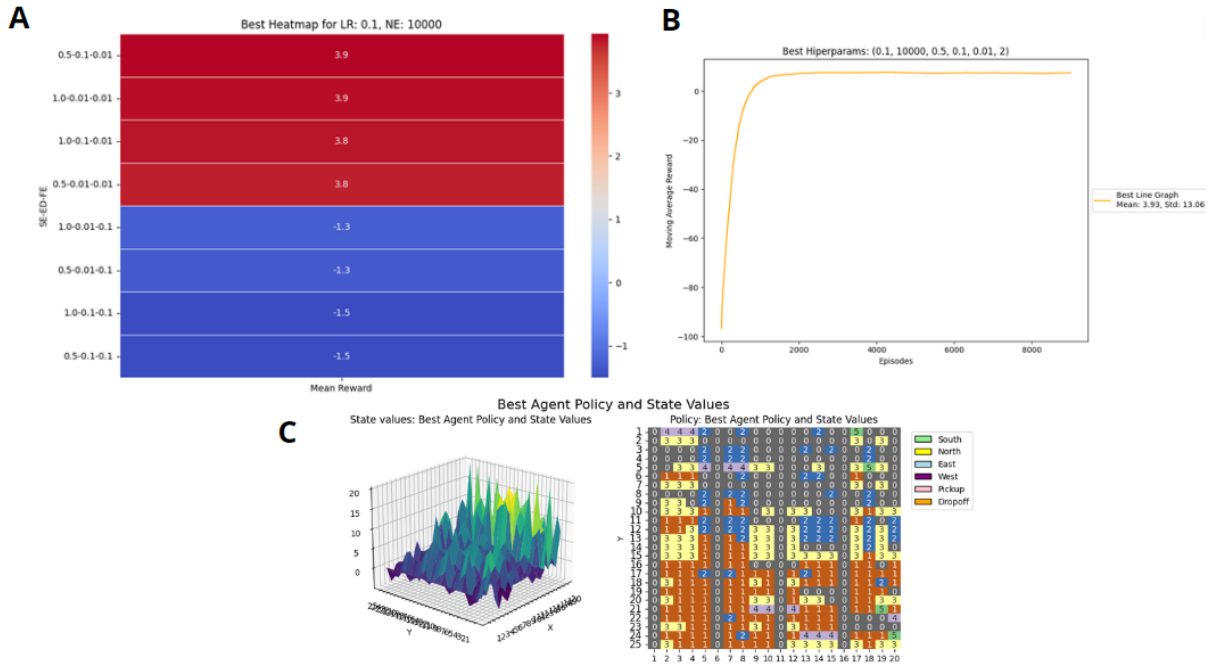

Figura A: Melhor Heatmap Individual para o Método Q-learning. O eixo Y representa as combinações de epsilon inicial (SE), decaimento de epsilon (ED), epsilon final (FE).

Figura B: Melhor Gráfico de Linha Individual para o Método Q-learning. A linha laranja mostra a recompensa média ao longo dos episódios.

Figura C: State Values e Policy do Melhor Agente para o Método Q-learning. O gráfico 3D à esquerda mostra os valores dos estados, enquanto o heatmap à direita mostra a política do agente. Cada célula do heatmap indica a ação preferida (South, North, West, East, Pick Up e Dropoff) para uma determinada posição no ambiente, destacando a eficácia da política aprendida pelo agente.

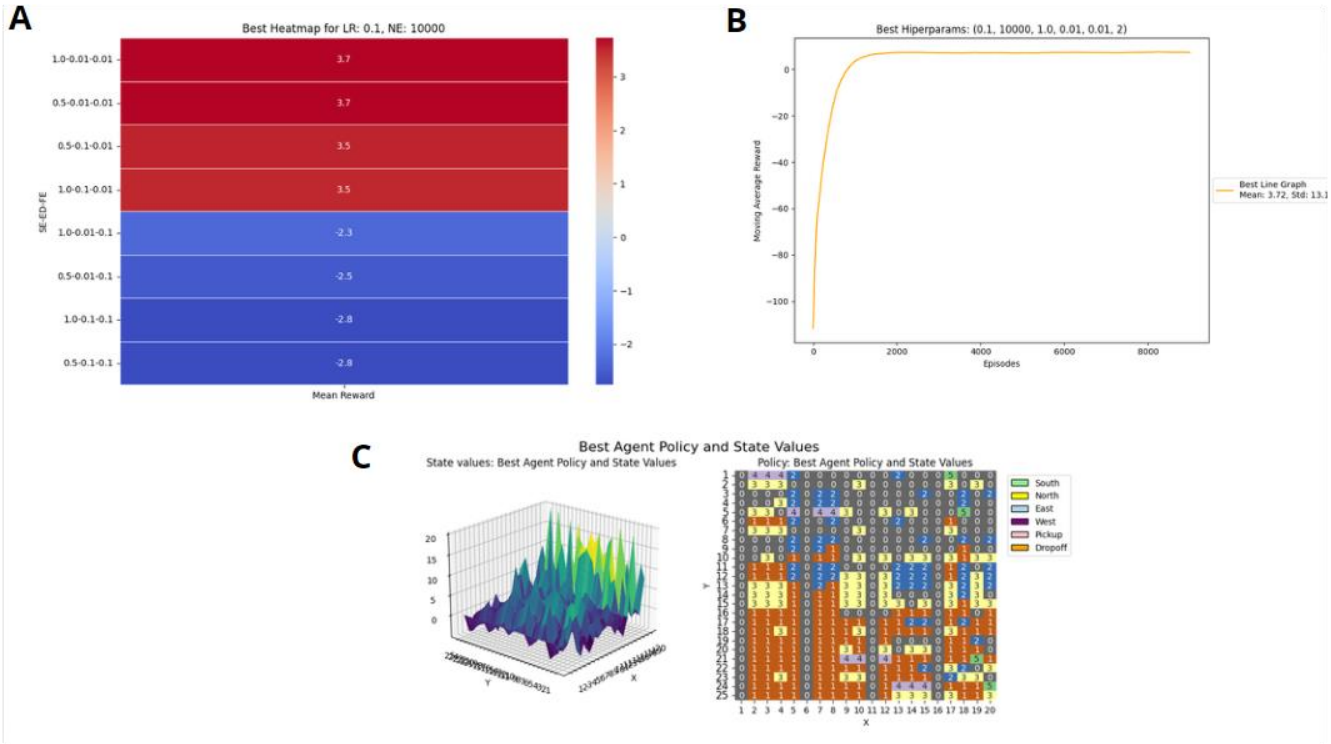

Figura A: Melhor Heatmap Individual para o Método SARSA. O eixo Y representa as combinações de epsilon inicial (SE), decaimento de epsilon (ED), epsilon final (FE).
Figura B: Melhor Gráfico de Linha Individual para o Método SARSA. A linha laranja mostra a recompensa média ao longo dos episódios.
Figura C: State Values e Policy do Melhor Agente para o Método SARSA. O gráfico 3D à esquerda mostra os valores dos estados, enquanto o heatmap à direita mostra a política do agente. Cada célula do heatmap indica a ação preferida (South, North, West, East, Pick Up e Dropoff) para uma determinada posição no ambiente, destacando a eficácia da política aprendida pelo agente.

5. Conclusões

A partir da análise em conjunto dos experimentos realizados com os métodos de diferenças temporais (Q-learning e SARSA) nos ambientes Blackjack, Cliff Walking, Frozen Lake e Taxi do Gymnasium é possível concluir que a taxa de aprendizado e o número de episódios foram os hiperpâremetros que mais tiveram influência no desempenho dos agentes. Os hiperparâmetros de exploração, como épsilon inicial (SE), decaimento de épsilon (ED) e épsilon final (FE), não demonstraram tendências ou padões notáveis, sendo observados com valores distintos mesmo com recompensa média parecidas. O uso de histórico, em geral, não afetou significativamente o desempenho dos agentes, indicando que o estado atual era suficiente para a tomada de decisão eficaz na maioria dos ambientes.

No desempenho geral, o Q-learning mostrou uma leve vantagem sobre o SARSA em termos de recompensas médias e variação dos valores das recompensas, se mostrando possivelmente mais resistente a variações de hiperparâmetros, enquanto SARSA exibiu maiores variações no desempenho, ambos os algoritmos enfrentaram dificuldades em ambientes mais complexos como Black Jack e Taxi, onde as penalidades e os fatores não determinísticos influenciaram negativamente as recompensas, mesmo assim, para ambos os algoritmos, foi possível treinar agentes que conseguiram concluir o objetivo do ambiente .

Os resultados dos experimentos demonstram a importância da seleção de hiperparâmetros para otimizar o desempenho dos algoritmos de aprendizado por reforço.