Assignment – 3

Defining and Solving Reinforcement Learning Task

Nikhil Saji

Vishnu Krishnakumar Menon

Part – 1:

Environment:

- Treasure Hunt Grid World

States:

- The environment is represented by a grid with 25 states, i.e. {S1 = (0,0), S2 = (0,1), S3= (0,2),……, S25 = (4,4)}

Actions:

- The agent can move in four directions i.e. Up, Down, Left, Right

Rewards:

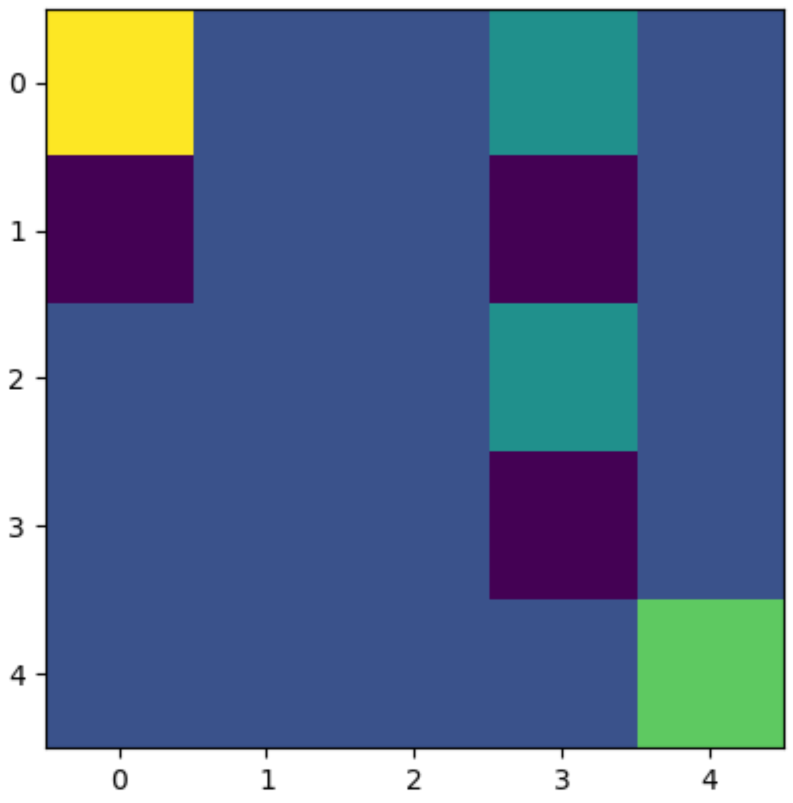
- The environment rewards the agent on three bases.
- If it lands on trap it is rewarded with –20 points
- When it lands on treasure it is rewarded with +20 points. Once reward is collected, it is removed from the environment.
- For all other movements it is rewarded with –0.5 points.
- When it reaches the final target, the environment rewards the agent with +100 points.

Main Objective:

- The agent aims to maximize rewards by collecting treasures and reaching the goal within the time limit.

Grid View (Visualization of the environment):

- Treasures: Represented by a positive value (Teal).
- Traps: Represented by a negative value (Dark purple).
- Goal: A distinct value indicating the target state. (Green)
- Agent: Its current position is visualized dynamically. (Yellow)



Safety in AI:

- The agent's actions are bounded within the grid dimensions, preventing it from moving outside the grid.
- State transitions are strictly constrained by the grid's layout and the defined action space.
- The environment resets if the agent reaches the goal or exceeds the step limit, ensuring predictable termination conditions.
- The agent is rewarded if it finds the treasure, hence avoiding unintended incentives.
- The environment penalizes the agent heavily for unsafe or undesired actions.

Part – 2 :

SARSA METHOD:

SARSA is an on-policy reinforcement learning algorithm that updates the action-value function, Q(s, a), based on the agent's chosen and observed state-action pairs. It helps the agent learn an optimal policy in grid-world problems by maximizing rewards and avoiding traps through Q-value updates.

- Key features:
  1. SARSA learns a policy by interacting with the environment.
  2. It Updates the Q-value using future estimates thus by making it computationally efficient.

Parameters

- Episodes: The agent was trained over 500 episodes to allow sufficient exploration and learning.
- Alpha: value was set to 0.1 to ensure a gradual update of the Q-values, which helps maintain stability.
- Gamma: Initially set to 0.9, prioritizing long-term rewards.
- Epsilon: set to 1.0 to encourage exploration at the start
- Epsilon decay: Set to 0.99, gradually reducing exploration to focus on exploitation as learning progresses.
- Minimum epsilon: Fixed at 0.1 to retain minimal exploration even in later stages.
- Maximum Timesteps: To prevent episodes from running indefinitely, the limit is set to 100.

ADVANTAGES OF SARSA

- It learns the policy that the agent is currently following, making it more stable in environments with unpredictable rewards.
- This approach is effective in situations where exploration must align with the updates to the policy.

DISADVANTAGES OF SARSA

- Convergence may be slower when using on-policy methods compared to off-policy methods like Q-learning.
- Insufficient exploration may result in less effective policies.

HYPERPARAMETER INFLUENCE

- **Discount Factor:** The value assigned to future rewards relative to immediate rewards is determined by this factor. A higher value, like 0.99, promotes the agent’s ability to learn strategies for the long term, whereas lower values, such as 0.7, emphasize short-term benefits. In this case, a value of γ=0.99 produced the best outcomes, as it motivated the agent to focus on achieving the goal rather than on immediate rewards.

- **Epsilon Decay:** The initial value 1.0 allows the agent to explore the environment randomly. A balance between exploration and exploitation is crucial to avoid local optima. The best epsilon decay rate =0.975, ensured sufficient exploration before stabilizing into exploitation.
- **Episodes and Maximum Timesteps:** The environment underwent training for 500 episodes, resulting in the convergence of rewards. To prevent unnecessary computations in situations where early goal achievement or traps occur, each episode was limited to 100 timesteps.

I chose the hyper parameters Epsilon Decay and Discount Factor(gamma) for hyper paramter tuning. Three different values were experimented for both hyper parameters Gamma= `[0.7, 0.89, 0.99]` and Epsilon Decay = `[0.93, 0.975, 0.999]`. After hyperparameter tuning the best hyperparameters obtained were:

- Gamma(discount factor) = 0.99
- Epsilon Decay = 0.975

RESULTS

Best Setup:

- Alpha = 0.1
- Gamma = 0.99
- Epsilon = 1.0
- Epsilon Decay = 0.975
- Minimum Epsilon = 0.1
- Maximum timestep per episode = 100

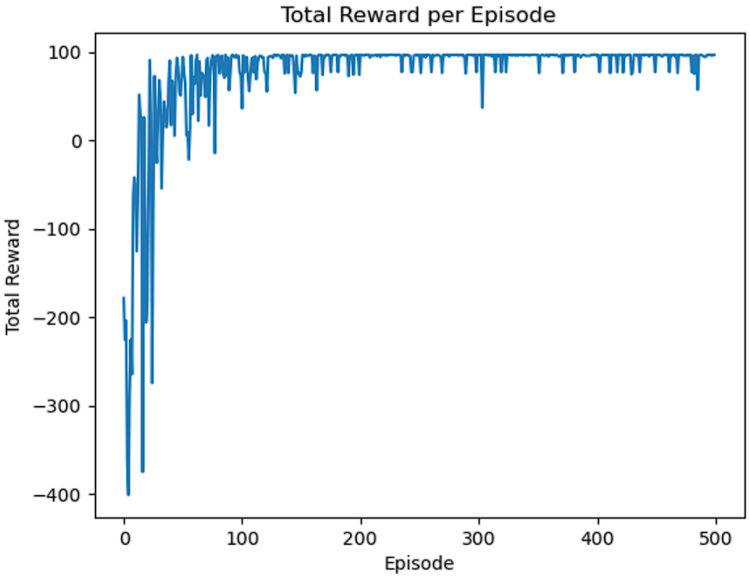
The agent developed an optimal policy after completing 500 episodes. The Q-values were updated effectively to reflect the dynamics of the environment.

- Average Reward (Greedy Policy): 96.5

PLOTS

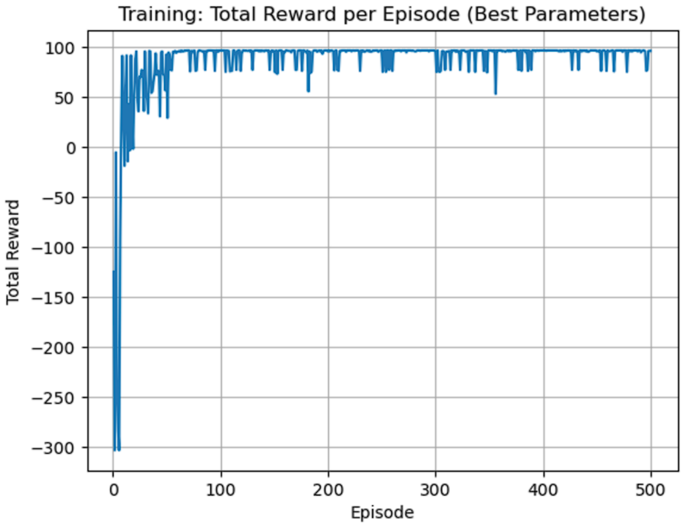
Total Reward per Episode

Trained Q-table



Rewards steadily increased over episodes, with occasional drops from exploration. After around 100 episodes, the agent achieved a high-reward policy.

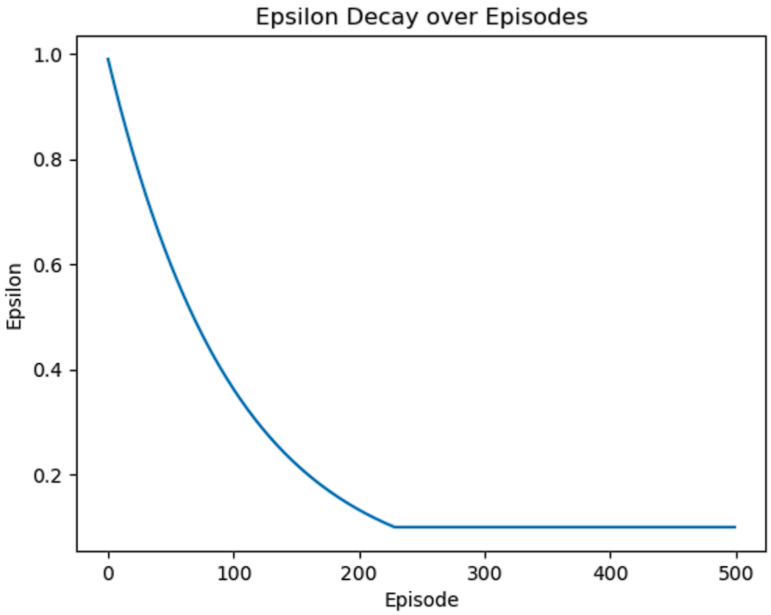
Total Reward per Episode (Best Parameters)



Training with optimal parameters demonstrated consistent learning and a high reward plateau, indicating that the agent effectively learned the environment.

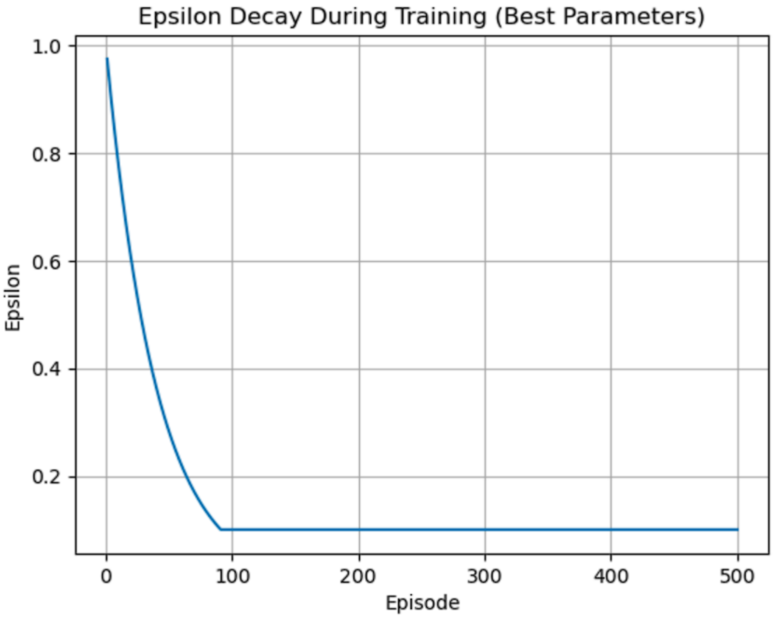
Epsilon Decay over Episodes

Trained Q-table



The epsilon value decreased logarithmically, which shifted the agent from exploration to exploitation.

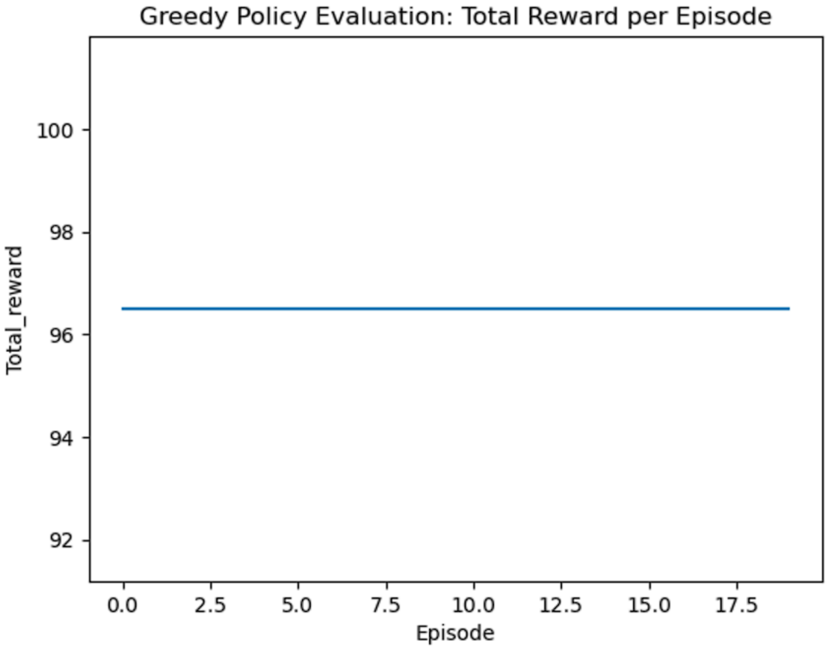
Epsilon Decay During Training (Best Parameters)



The epsilon decay plot reflects a smooth transition to exploitation.

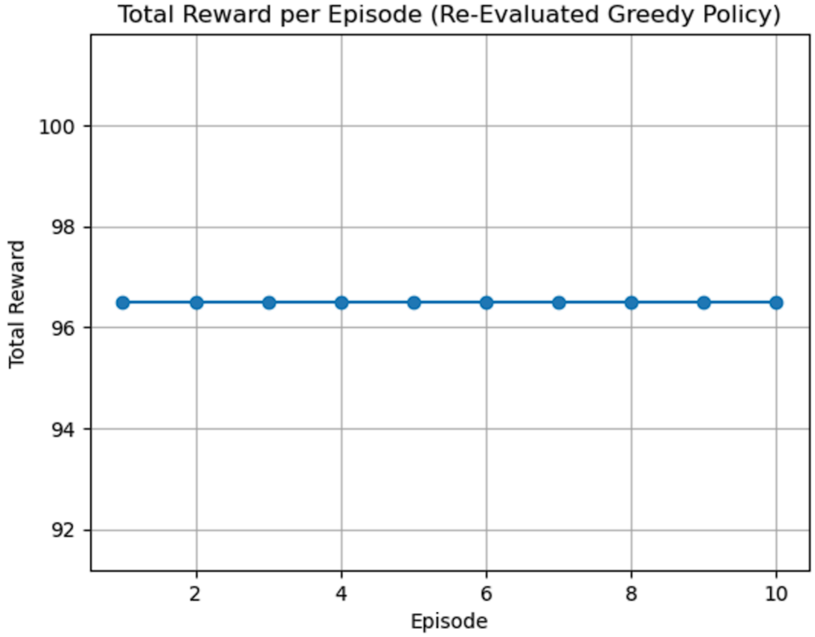
Greedy Policy Evaluation

Trained Q-table



The evaluation of the greedy policy showed consistently high rewards, indicating that the agent adhered to the optimal policy it learned.

Re-Evaluated Greedy Policy



Re-evaluating the policy confirmed consistent performance over multiple runs, validating the robustness of the learned policy.

Part – 3 :

Double Q-learning Method

Double Q-learning is a reinforcement learning method designed to reduce the overestimation bias present in Q-learning. Rather than having a single Q-table, it operates with two distinct Q-tables, QA and QB, which are updated in an alternating fashion. When selecting actions, the combined estimate of QA and QB is utilized.

Update Rule:

$QA(s,a) \leftarrow QA(s,a)+\alpha [r+\gamma QB(s',\text{argmax}\ QA(s',a)) - QA(s,a)]$

$QB(s,a) \leftarrow QB(s,a)+\alpha [r+\gamma QA(s',\text{argmax}\ QB(s',a)) - QB(s,a)]$

Advantages:

- Reduces overestimation bias.
- Enhances learning stability by preserving two separate Q-values.

Disadvantages:

- In comparison to standard Q-learning, memory requirements are increased twofold.
- Computationally heavier due to maintaining two Q-tables.

Derivation of the N-Step Double Q-Learning Update Rule:

To derive the update rule for N-step Double Q-Learning, we build upon the conventional one-step Double Q-Learning method by incorporating N-step returns. In Double Q-Learning, the Q-function is divided into two distinct estimates, QA and QB, in order to mitigate the bias of overestimation.

Update Rule for Double Q-Learning (One-Step):

$QA\ (st\ ,at\ ) \leftarrow QA\ (st\ ,at\ )+\alpha\ [\ rt+1\ +\gamma QB\ (st+1\ ,\text{argmax}\ QA\ (st+1\ ,a)) - QA\ (st\ ,at\ )\ ]$

$QB\ (st\ ,at\ ) \leftarrow QB\ (st\ ,at\ )+\alpha[\ rt+1\ +\gamma QA\ (st+1\ ,\text{argmax}\ QB\ (st+1\ ,a)) - QB\ (st\ ,at\ )\ ]$

Generalizing to N-step Returns:

For N-step returns, rather than directly using the Q-value of the subsequent state for bootstrapping, we take into account the total rewards accumulated over N steps, as well as performing bootstrapping after those N steps.

The N-step return $G_{t:t+n}$

$G_{t:t+n} = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots + \gamma^{n-1} R_{t+n} + \gamma^n QB\ (s_{t+n}\ ,\text{argmax}\ QA\ (s_{t+n}\ ,a))$

If QA is updated, we bootstrap using QB , and vice versa. The update rule becomes:

$QA\ (st\ ,at\ ) \leftarrow QA\ (st\ ,at\ )+\alpha[G_{t:t+n} - QA\ (st\ ,at\ )]$

$QB\ (st\ ,at\ ) \leftarrow QB\ (st\ ,at\ )+\alpha[G_{t:t+n} - QB\ (st\ ,at\ )]$

Recursive Definition of $G_{t:t+n}$

$G_{t:t+n} = R_{t+1} + \gamma(G_{t+1:t+n})$

This allows efficient computation by maintaining a buffer of rewards and states during each episode.

Handling Terminal States - If the episode ends before reaching N steps, the remaining rewards are cut off, and the return is calculated solely based on the rewards that were observed.

HYPERPARAMETER INFLUENCE

HYPERPARAMETER INFLUENCE

- **Discount Factor:** The value assigned to future rewards relative to immediate rewards is determined by this factor. A higher value, like 0.99, promotes the agent’s ability to learn strategies for the long term, whereas lower values, such as 0.7, emphasize short-term benefits. In this case, a value of γ=0.89 produced the best outcomes, as it motivated the agent to focus on achieving the goal rather than on immediate rewards.
- **Epsilon Decay:** The initial value 1.0 allows the agent to explore the environment randomly. A balance between exploration and exploitation is crucial to avoid local optima. The best epsilon decay rate =0.93, ensured sufficient exploration before stabilizing into exploitation.
- **Episodes and Maximum Timesteps:** The environment underwent training for 500 episodes, resulting in the convergence of rewards. To prevent unnecessary computations in situations where early goal achievement or traps occur, each episode was limited to 200 timesteps.

I chose the hyper parameters Epsilon Decay and Discount Factor(gamma) for hyper paramter tuning. Three different values were experimented for both hyper parameters Gamma= [0.7, 0.89, 0.99] and Epsilon Decay = [0.93, 0.975, 0.999]. After hyperparameter tuning the best hyperparameters obtained were:

- Gamma(discount factor) = 0.89
- Epsilon Decay = 0.93

RESULTS

Best Setup:

- Alpha = 0.1
- Gamma = 0.89
- Epsilon = 1.0
- Epsilon Decay = 0.93
- Minimum Epsilon = 0.1
- Maximum timestep per episode = 100

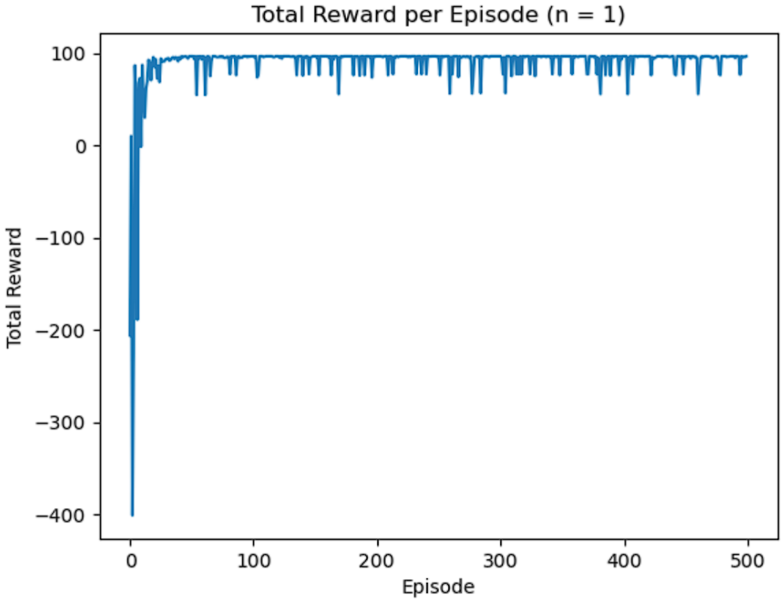
PLOTS & TABLES

N = 1

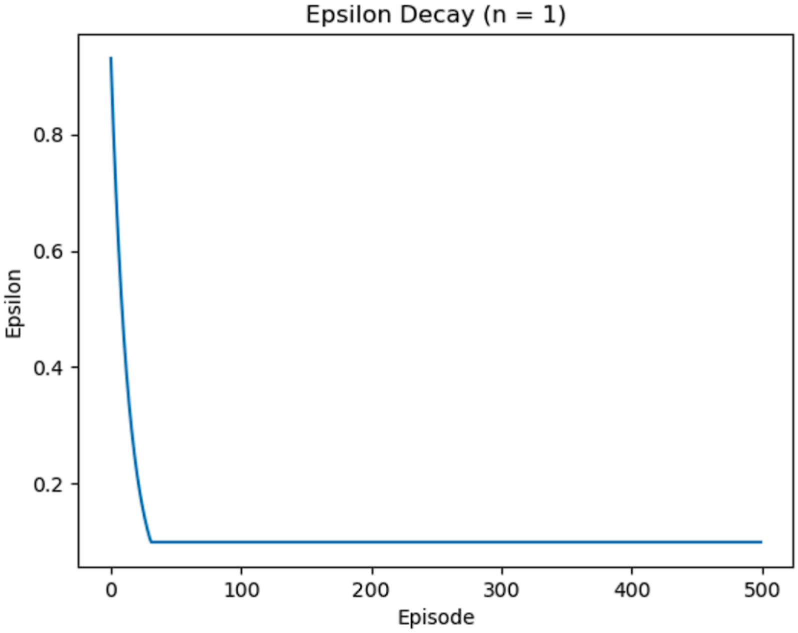
```
Trained Q_A: [[ 5.84849487e+00  2.30763690e+01 -1.65068497e+01  6.48163872e+00]
 [ 1.14181160e+01  2.98605664e+01 -2.68315894e-02  2.26209039e+00]
 [ 1.26793168e+01 -5.29436206e-01  3.58018587e+01  2.68998532e+00]
 [-2.80326921e-01  1.91035490e+00 -8.23284897e+00 -4.43528632e-01]
 [-1.66107114e-01 -3.88704891e-01  1.30399088e+01 -1.35857762e-01]
 [-9.58751908e-01 -2.55125293e-01  5.21833243e-01 -7.82994532e+00]
 [-4.74257494e-01 -4.83122765e-01  1.13870701e+01 -6.92605093e+00]
 [ 1.04119931e+01 -3.91968477e+00  4.58373310e+01 -2.81678030e-01]
 [-2.81254795e-01 -2.52582926e-01  3.90770935e+01 -2.28280469e-01]
 [-6.21938859e-02 -1.44152332e-01  5.88286077e+01 -2.00422750e+00]
 [-6.31097801e+00  1.80568064e+01 -3.09713913e-01 -3.99788643e-01]
 [-5.42306105e-01  4.48229869e+01 -2.97118540e-01 -3.40024208e-01]
 [ 1.37282671e+01  5.45231847e+01 -9.11965580e-02  1.51820385e+01]
 [-5.76717954e+00  6.47669301e+01  1.33999368e+00  1.47162790e+01]
 [ 1.92132982e+01  1.80600535e+01  8.16074797e+01  2.93612795e+01]
 [-5.53546547e-01 -3.22714633e-01 -2.36453098e-01 -2.70938355e-01]
 [-2.47184634e-01 -3.28741000e-01 -2.27086549e-01 -3.61887120e-01]
 [-2.64770340e-01 -3.80200250e+00  2.98924396e+00 -2.26199706e-01]
 [ 5.46022185e+00  6.22261348e+01 -4.21657596e-01  0.00000000e+00]
 [ 1.32682748e+01  2.37293801e+01  1.00000000e+02  1.43073536e+01]
 [-2.25650899e-01 -1.87267140e-01 -2.38962124e-01 -2.16816631e-01]
 [-1.29580214e-01  2.80545843e-01 -2.20577297e-01 -1.54452628e-01]
 [-9.72250000e-02  1.95761157e+01 -1.05756778e-01 -5.63425662e-02]
 [-5.29741532e+00  5.69532790e+01 -5.42275000e-02  3.11778075e-02]
 [ 0.00000000e+00  0.00000000e+00  0.00000000e+00  0.00000000e+00]]

Trained Q_B: [[ 5.45953400e+00  2.60500102e+01 -1.31222673e+01  7.55565439e+00]
 [ 8.47958887e+00  3.02279579e+01  2.68318434e-01  3.13264010e+00]
 [ 5.59188224e+00 -4.99612268e-01  3.93942972e+01  5.72923671e+00]
 [-5.50014061e-01  2.47998245e+00 -3.81660696e+00 -4.07911867e-01]
 [-3.55213876e-01 -2.08124507e-01  1.37624683e+01 -6.15648694e-01]
 [-6.66832109e-01 -6.62816142e-01  4.58740843e+00 -1.05188125e+01]
 [-4.94264891e-01 -4.50466785e-01  1.08031092e+01 -1.02847639e+01]
 [ 1.30126752e+01 -4.72867723e+00  4.51827977e+01  5.27338192e-01]
 [-1.35640917e-01  6.53172980e+00  2.56897367e+01 -1.41633169e-01]
 [-1.86144741e-01 -1.03228745e-01  4.35222846e+01 -6.91295379e+00]
 [-5.25574879e+00  1.17072604e+01 -3.21258484e-01 -2.46656609e-01]
 [-2.44059704e-01  3.98375322e+01 -3.75797657e-01 -3.59822997e-01]
 [ 2.07481835e+01  5.90097743e+01 -7.93267210e-02  1.58323681e+01]
 [-3.79622056e+00  6.82473435e+01  4.11292664e+00  1.37647181e+01]
 [ 1.07235585e+01  3.25890277e+01  8.40133963e+01  3.37265641e+01]
 [-1.98008953e-01 -2.43187215e-01 -3.20537642e-01 -2.20138079e-01]
 [-2.55678237e-01 -2.99454064e-01 -2.03471115e-01 -1.24423563e-01]
 [-9.92275000e-02 -3.82685522e+00  3.03306393e+00 -1.88796931e-01]
 [-4.00344767e-01  7.04510621e+01 -3.50769029e-01 -6.07461770e-02]
 [ 2.37438261e+01  3.46268587e+01  1.00000000e+02  7.77128864e+00]
 [-1.97311402e-01 -1.89510253e-01 -2.47776950e-01 -2.24581619e-01]
 [-9.95397003e-02  1.97885545e+00  0.00000000e+00 -1.05430289e-01]
 [-1.05633150e-01  1.39104226e+01 -1.01925887e-01 -1.11239601e-01]
 [-3.72200930e+00  6.51321560e+01  4.84596599e+00  1.32750978e+00]
 [ 0.00000000e+00  0.00000000e+00  0.00000000e+00  0.00000000e+00]]
```

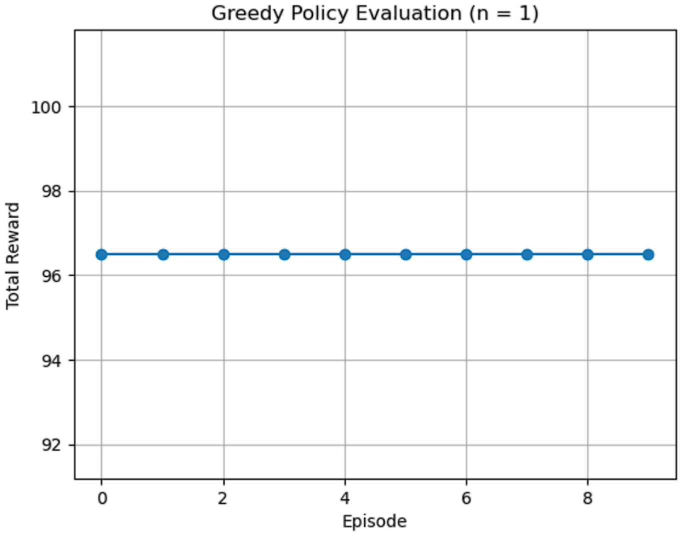
Total Reward per Episode



Epsilon Decay



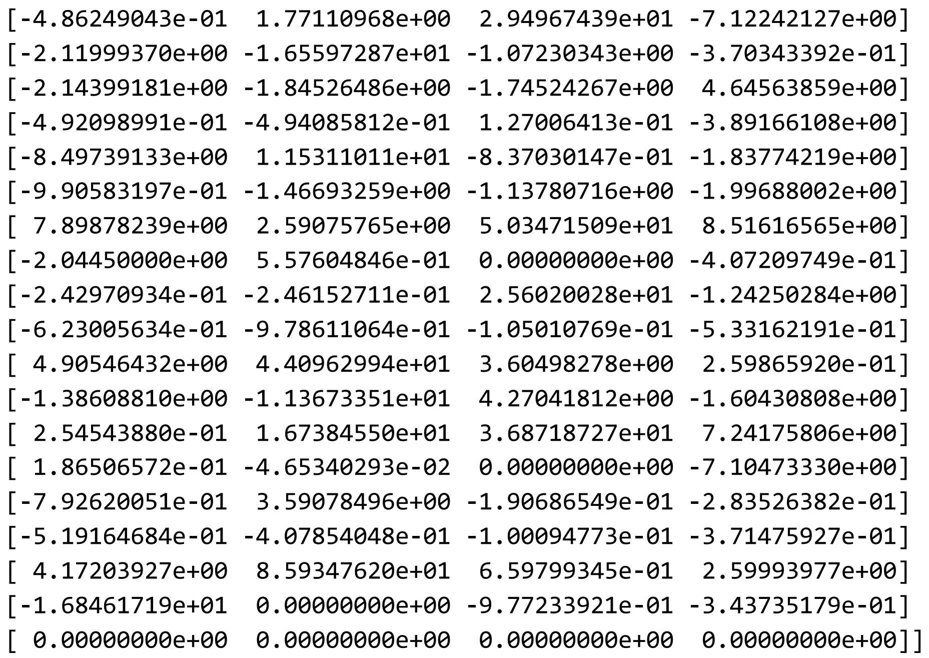
GREEDY POLICY EVALUATION



N = 2

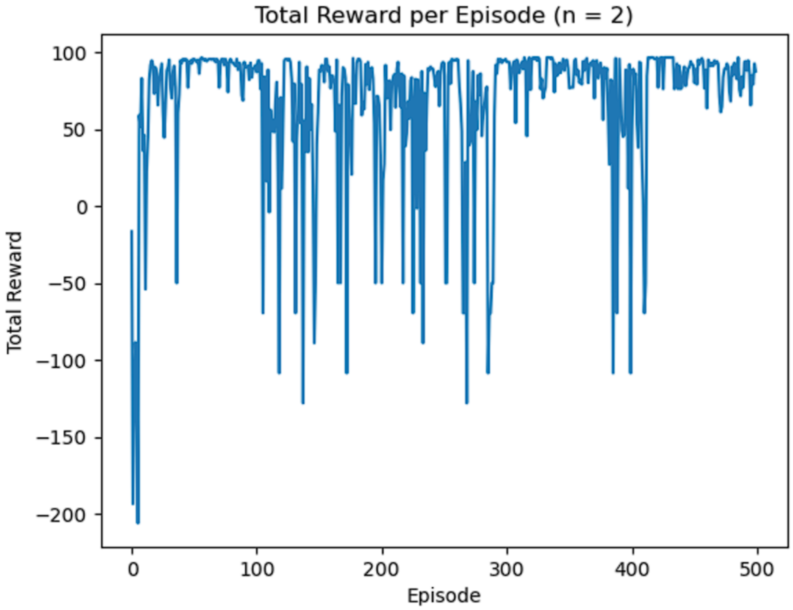
```
Trained Q_A: [[-4.69764639e+00 -8.12645795e-01 -1.25410606e+01 -1.49381841e+00]
 [ 2.88067816e+00 -6.86719804e-01 -2.35659344e+00 -1.64095627e+00]
 [-2.21117731e+00 -3.66288479e-01  2.88488861e+01  3.10423659e+00]
 [-1.80594329e+00 -9.35050028e-01 -8.69250497e+00 -1.21357407e+00]
 [-1.49976382e+00 -1.25181635e+00 -3.01318408e+00 -3.41551007e+00]
 [-1.69644181e+00 -1.68399751e+00 -1.91147344e+00 -6.31606851e+00]
 [ 5.98111812e-01  1.41265729e+01  2.18731321e+01 -9.50409985e+00]
 [-1.80372835e+00 -7.30158873e+00 -2.57165634e+00 -3.91771179e+00]
 [-2.06621098e+00 -1.68150397e+00 -1.48450801e+00  1.02441196e+00]
 [-1.87445487e-01 -1.76268600e+00  1.56669309e-01 -7.30390096e+00]
 [-3.95072026e+00  4.98717148e+00 -4.70849794e-01 -1.79550000e-01]
 [-3.00815832e+00 -2.34714012e+00 -1.49014261e+00 -1.71729850e+00]
 [-3.23873415e+00  3.35567177e+00  4.69513525e+01  5.76809659e+00]
 [-5.64006643e+00 -2.98353821e-01 -2.04450000e+00 -4.53046278e-01]
 [-1.80719677e+00  1.83692196e-01  2.37840701e+01 -1.33422360e+00]
 [-2.32837993e+00 -5.46571494e-01 -8.25914933e-01 -5.95404338e-01]
 [ 4.68022989e+00  5.00463188e+01  8.45344540e+00  7.54133816e+00]
 [-1.95854515e+00 -1.21941161e+01 -9.68668396e-01 -1.75748072e+00]
 [-4.46042323e-02  4.96092736e-01  2.73450396e+01  5.59154448e+00]
 [-2.98645458e-01  2.96408673e-02  0.00000000e+00 -4.64234307e+00]
 [ 5.72987672e-01  1.89088675e+00 -2.00618181e-01 -4.59181932e-01]
 [-4.91237089e-01 -1.21233278e-02 -2.72561522e-01 -5.30997359e-01]
 [ 8.22975267e+00  8.69305483e+01 -1.34863958e-01  1.83889866e+00]
 [-1.82003534e+01  0.00000000e+00 -5.21885003e-01  5.26342211e-02]
 [ 0.00000000e+00  0.00000000e+00  0.00000000e+00  0.00000000e+00]]

Trained Q_B: [[-1.15645064e+00  1.58207615e+00 -1.37319022e+01 -3.42934495e+00]
 [-1.26758914e+00 -2.75099121e+00 -2.06391663e+00 -2.61232367e+00]
 [-1.92348645e+00  8.22503622e-01  2.45248340e+01 -2.70125554e-01]
 [-6.39882019e-01 -1.40164621e+00 -7.92821544e+00 -8.89394147e-01]
 [-1.13164296e+00 -1.33597806e+00 -1.58121236e+00 -5.34552575e-01]
 [-3.40775277e+00 -1.13869071e+00 -9.40221675e-01 -2.07232214e+00]
```

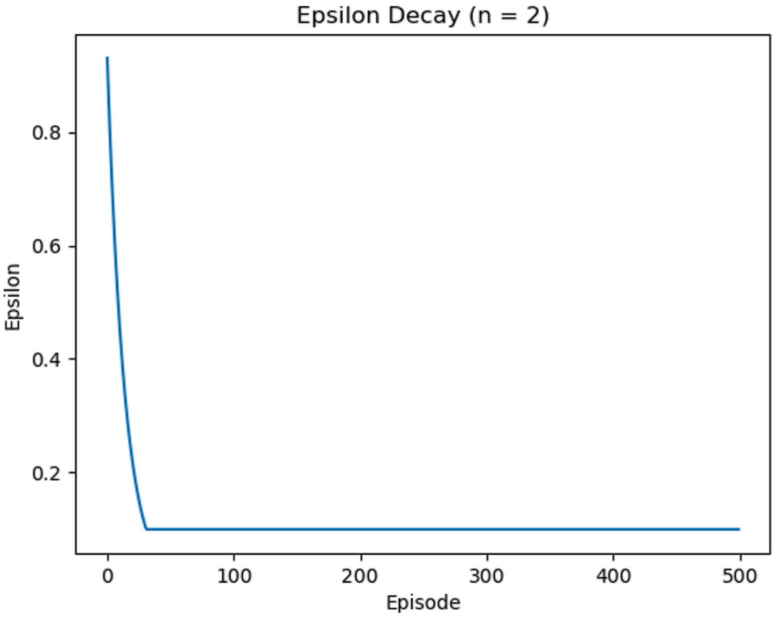


```
 [-4.86249043e-01  1.77110968e+00  2.94967439e+01 -7.12242127e+00]
 [-2.11999370e+00 -1.65597287e+01 -1.07230343e+00 -3.70343392e-01]
 [-2.14399181e+00 -1.84526486e+00 -1.74524267e+00  4.64563859e+00]
 [-4.92098991e-01 -4.94085812e-01  1.27006413e-01 -3.89166108e+00]
 [-8.49739133e+00  1.15311011e+01 -8.37030147e-01 -1.83774219e+00]
 [-9.90583197e-01 -1.46693259e+00 -1.13780716e+00 -1.99688002e+00]
 [ 7.89878239e+00  2.59075765e+00  5.03471509e+01  8.51616565e+00]
 [-2.04450000e+00  5.57604846e-01  0.00000000e+00 -4.07209749e-01]
 [-2.42970934e-01 -2.46152711e-01  2.56020028e+01 -1.24250284e+00]
 [-6.23005634e-01 -9.78611064e-01 -1.05010769e-01 -5.33162191e-01]
 [ 4.90546432e+00  4.40962994e+01  3.60498278e+00  2.59865920e-01]
 [-1.38608810e+00 -1.13673351e+01  4.27041812e+00 -1.60430808e+00]
 [ 2.54543880e-01  1.67384550e+01  3.68718727e+01  7.24175806e+00]
 [ 1.86506572e-01 -4.65340293e-02  0.00000000e+00 -7.10473330e+00]
 [-7.92620051e-01  3.59078496e+00 -1.90686549e-01 -2.83526382e-01]
 [-5.19164684e-01 -4.07854048e-01 -1.00094773e-01 -3.71475927e-01]
 [ 4.17203927e+00  8.59347620e+01  6.59799345e-01  2.59993977e+00]
 [-1.68461719e+01  0.00000000e+00 -9.77233921e-01 -3.43735179e-01]
 [ 0.00000000e+00  0.00000000e+00  0.00000000e+00  0.00000000e+00]]
```

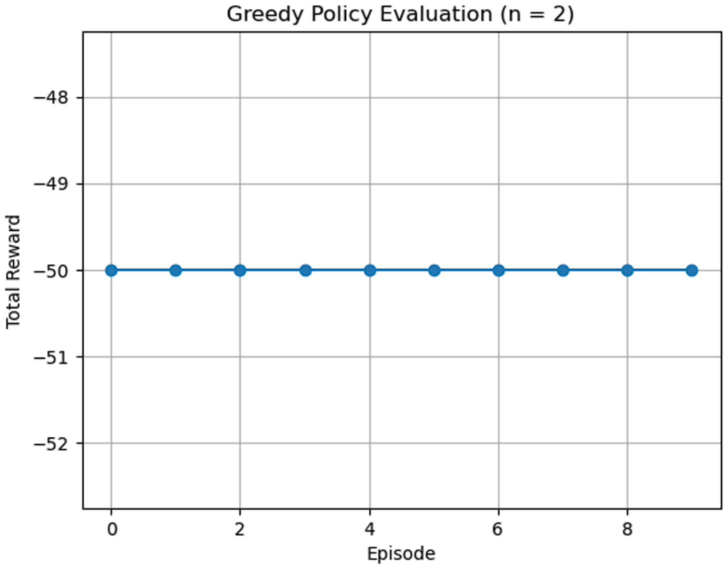
TOTAL REWARDS PER EPISODE



EPSILON DECAY



GREEDY EVALUVATION POLICY

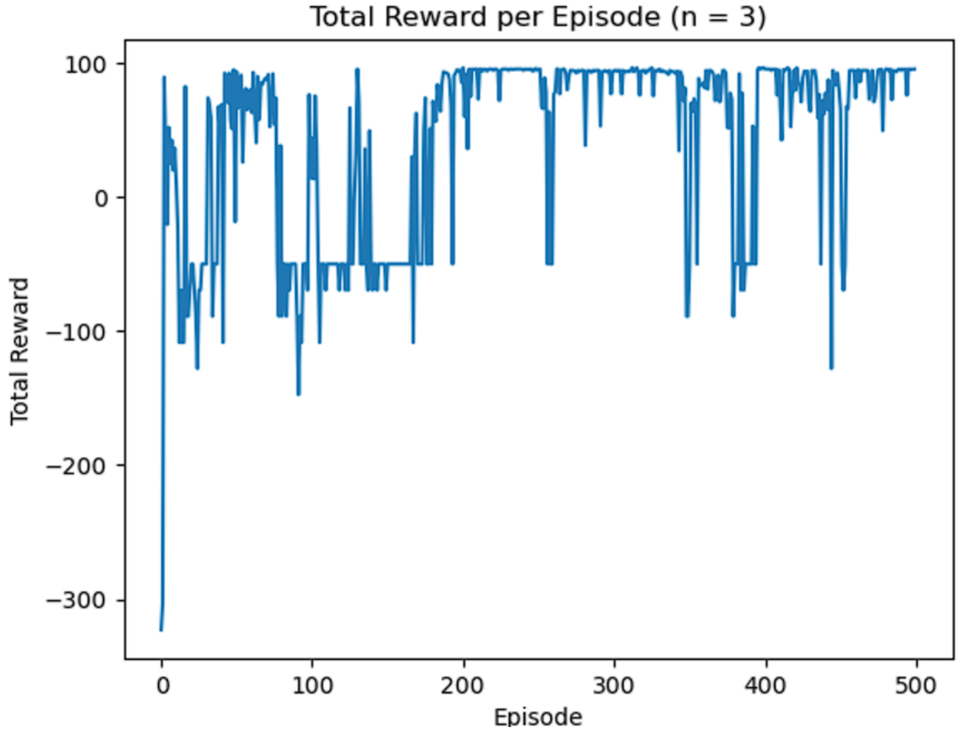


N = 3

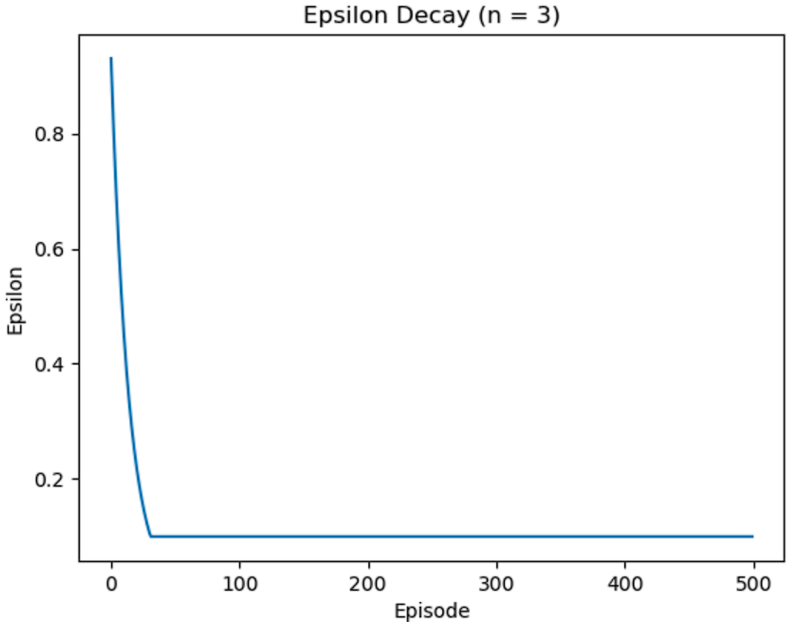
```
Trained Q_A: [[ -3.4108915   -2.47243042 -17.45557362  -3.76737823]
 [ -4.85215436  -2.77262948   4.16050884  -4.12065288]
 [ -3.62756872  -2.95986659  -3.41847032  -3.14842138]
 [ -4.86310122  -6.57705016 -14.12663254  -4.83101599]
 [ -4.05431618  -2.95741307  -3.30660751  -4.04082241]
 [ -3.78632254  -2.61980393  -0.83884685  -7.408383  ]
 [ -4.23325601  -3.87227195   0.14570054 -15.79681625]
 [ -4.13554303 -14.95613004   1.87796893  -1.30268588]
 [ -3.3653273   -0.92477152  -0.134105    -1.55697023]
 [ -2.78710406  -3.01369917  15.42855927  -7.49904288]
 [-10.19353002  -3.62756741  13.2572468   -4.88254516]
 [ -1.53791137  -2.77021884   0.06147973  -4.10713944]
 [ -3.91342571  -3.40793027   0.27075794  -0.91422354]
 [ -7.49674645  19.35847174 -13.37282535  -3.11996296]
 [ -7.36864276  -3.32778683  -3.49930432  -7.34042056]
 [ -2.15555535  -2.96368605   3.10898244  -2.18630407]
 [ -1.64453037  -1.22802647   0.8896183   24.04898552]
 [ -0.85560374 -16.99465691  15.34039943  -1.91203012]
 [ -4.42051248  -1.0908711   -0.37494224   4.33693014]
 [  0.09456834  -0.27967979   0.           5.06544187]
 [ -1.62791431   4.13385835  -2.49519622  -2.06707563]
 [ -2.20630259  59.72305542  -2.32169391  -2.25146928]
 [ -0.98011589  13.89961537   3.99445009  -1.9237933 ]
 [-12.38612532   0.          -2.47690763  -0.90651395]
 [  0.           0.           0.           0.        ]]

Trained Q_B: [[ -6.59311613  -3.05686519 -19.2929042   -4.78345801]
 [ -3.13697137  -3.12876233   6.3160204   -5.3664707 ]
 [ -3.53028176  -5.30492507  -4.54146929  -3.63487751]
 [ -6.59496347  -4.12982365 -13.76003746  -4.22033019]
 [ -4.12643357  -3.98051901  -4.77140055  -6.11666948]
 [ -2.35126849  -4.3895101   -2.00480316 -12.40183762]
 [ -3.45545324  -3.89474239   0.60548143 -18.40277259]
 [ -3.73703046 -14.38623017   0.5635436   -1.58852162]
 [ -0.57013687  -4.29303099  -5.15289799  -2.36028682]
 [ -4.54872063  -3.70284442   7.80420655 -15.08319941]
 [-10.60844581  -2.75679699  11.39143691  -5.31123627]
 [ -3.59024521  -0.6938652    0.69866188  -5.25197762]
 [ -2.77346123  -5.3920371   -2.83781747   1.02300208]
 [ -5.8783956    4.74321028   4.30947177  -2.85540523]
 [ -1.79263177  -5.82569082  -4.42209787  -7.00374589]
 [ -3.85560152  -2.34051337   3.68267012  -2.90303476]
 [ -0.70842028  -1.07807265  -1.66050828  26.08908683]
 [ -2.51268535 -10.49220188  -1.31691192  -1.69120928]
 [ -3.63318034  -2.28234394  -0.84297881   0.71126629]
 [ -1.00073584  -1.53165245   0.          -6.36724947]
 [  1.33567338  -0.68664669  -1.88097365  -1.86016672]
 [ -2.19072943  63.75210327  -2.07986915  -1.77410647]
 [ -0.58086473   4.91383525  -1.16240063  -1.16906557]
 [-16.55834686   0.           1.4466767   -0.67538338]
 [  0.           0.           0.           0.        ]]
```

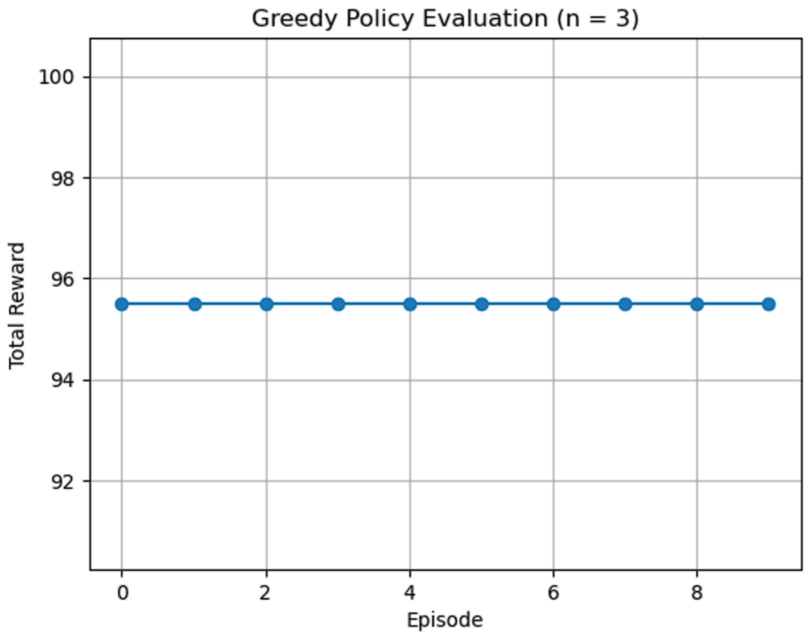
TOTAL REWARDS PER EPISODE



EPSION DECAY



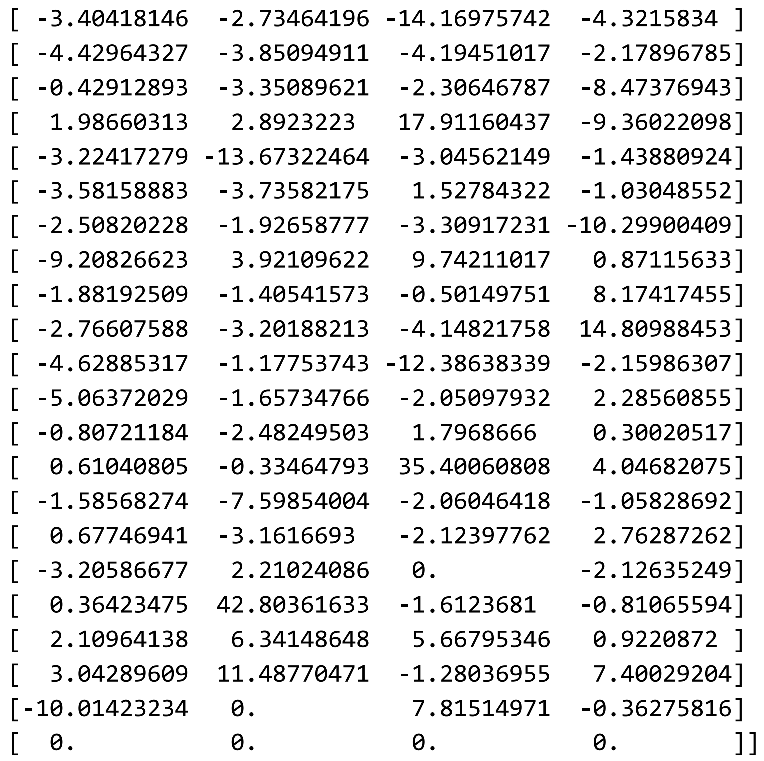
GREEDY EVALUATION POLICY



N = 4

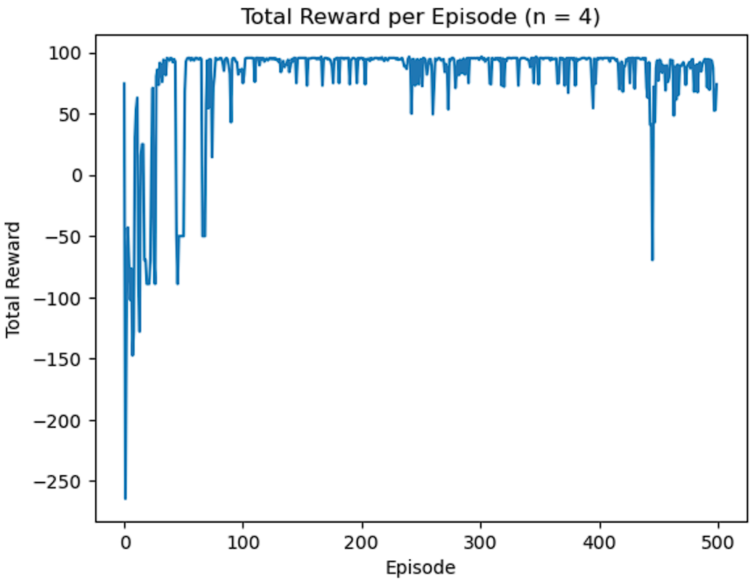
```
Trained Q_A: [[-5.39449505e+00  5.47157825e+00 -1.06383816e+01 -3.12529937e+00]
 [-1.00348645e+00 -8.61786311e-01 -3.43112601e+00 -2.88490051e+00]
 [-2.98605845e+00 -3.47408635e+00  3.91643870e+00 -5.93887921e+00]
 [-6.29786450e+00 -6.96821158e+00 -5.79964742e+00 -2.31047307e+00]
 [-3.93383374e+00 -3.90589504e+00 -4.86522570e+00 -6.63278522e+00]
 [-1.26218576e-02 -2.38994281e+00 -2.31742146e+00 -6.90039425e+00]
 [-8.28023045e-02 -3.95139331e+00  1.72886148e+01 -1.30096563e+01]
 [-2.47515891e+00 -1.63463278e+01 -2.72333480e+00 -2.47890672e+00]
 [-2.42925182e+00 -3.87038885e+00  2.04056332e-01 -5.22286750e+00]
 [-2.95976545e+00 -3.62128701e+00 -1.93436584e+00 -1.12176727e+01]
 [-9.92649773e+00  3.11425039e+00  1.22928696e+01 -1.89476731e-01]
 [-1.77831868e+00 -6.76522524e-01  2.32717681e+00  5.04870846e+00]
 [-9.66384237e-01 -3.26787879e+00 -5.86799005e-01  1.02380966e+01]
 [-8.92117383e+00 -4.53775115e+00  0.00000000e+00  3.65001571e-01]
 [-1.61031179e+00 -2.42786053e+00 -4.45857881e+00  5.42175385e+00]
 [ 5.02123355e-01 -2.35026424e+00  5.79298854e-01 -1.47462930e+00]
 [ 1.93093063e+00  7.98307212e-01  4.08857663e+01  4.39070923e+00]
 [-3.84135170e+00 -4.10907804e+00 -4.07571910e+00  3.51557717e-01]
 [-2.90011544e+00 -1.24080101e+00 -3.24041033e-02  1.87140567e-01]
 [-3.53720941e+00  7.34576019e+00  0.00000000e+00  0.00000000e+00]
 [ 2.72017245e+00  4.07869068e+01 -6.38018913e-01 -1.97063248e+00]
 [-5.35175819e-01  5.25448186e+00  4.10379704e+00  4.25446866e+00]
 [ 9.27795821e-01  1.33248025e+01  3.33353574e+00  1.69845071e+01]
 [-6.66536690e+00  0.00000000e+00 -1.17920693e+00 -8.27525620e-01]
 [ 0.00000000e+00  0.00000000e+00  0.00000000e+00  0.00000000e+00]]

Trained Q_B: [[ -2.45612734   2.31771894 -15.36446389  -1.21730577]
 [ -2.88228422   0.22091182  -1.17660488  -3.98732466]
 [ -4.99570893  -3.18228807   2.29091552   0.7160504 ]
```

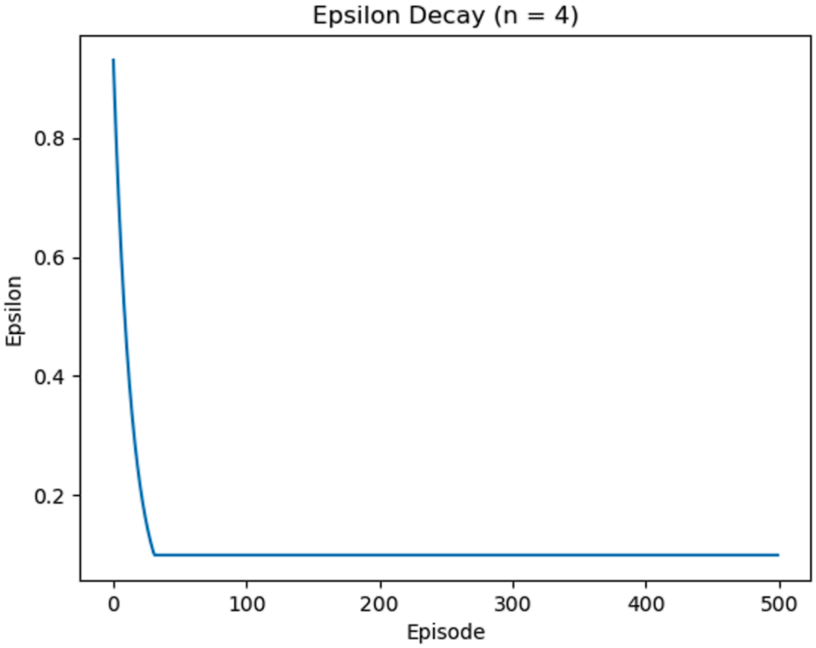


```
[ -3.40418146  -2.73464196 -14.16975742  -4.3215834 ]
[ -4.42964327  -3.85094911  -4.19451017  -2.17896785]
[ -0.42912893  -3.35089621  -2.30646787  -8.47376943]
[  1.98660313   2.8923223   17.91160437  -9.36022098]
[ -3.22417279 -13.67322464  -3.04562149  -1.43880924]
[ -3.58158883  -3.73582175   1.52784322  -1.03048552]
[ -2.50820228  -1.92658777  -3.30917231 -10.29900409]
[ -9.20826623   3.92109622   9.74211017   0.87115633]
[ -1.88192509  -1.40541573  -0.50149751   8.17417455]
[ -2.76607588  -3.20188213  -4.14821758  14.80988453]
[ -4.62885317  -1.17753743 -12.38638339  -2.15986307]
[ -5.06372029  -1.65734766  -2.05097932   2.28560855]
[ -0.80721184  -2.48249503   1.7968666    0.30020517]
[  0.61040805  -0.33464793  35.40060808   4.04682075]
[ -1.58568274  -7.59854004  -2.06046418  -1.05828692]
[  0.67746941  -3.1616693   -2.12397762   2.76287262]
[ -3.20586677   2.21024086   0.          -2.12635249]
[  0.36423475  42.80361633  -1.6123681   -0.81065594]
[  2.10964138   6.34148648   5.66795346   0.9220872 ]
[  3.04289609  11.48770471  -1.28036955   7.40029204]
[-10.01423234   0.           7.81514971  -0.36275816]
[  0.           0.           0.           0.        ]]
```

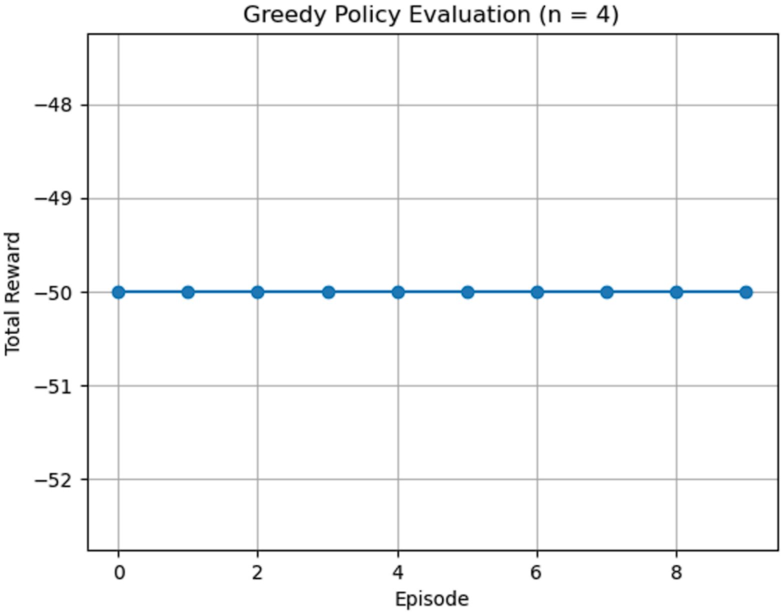
TOTAL REWARDS PER EPISODES



EPSILON DECAY



GREEDY EVALUATION POLICY

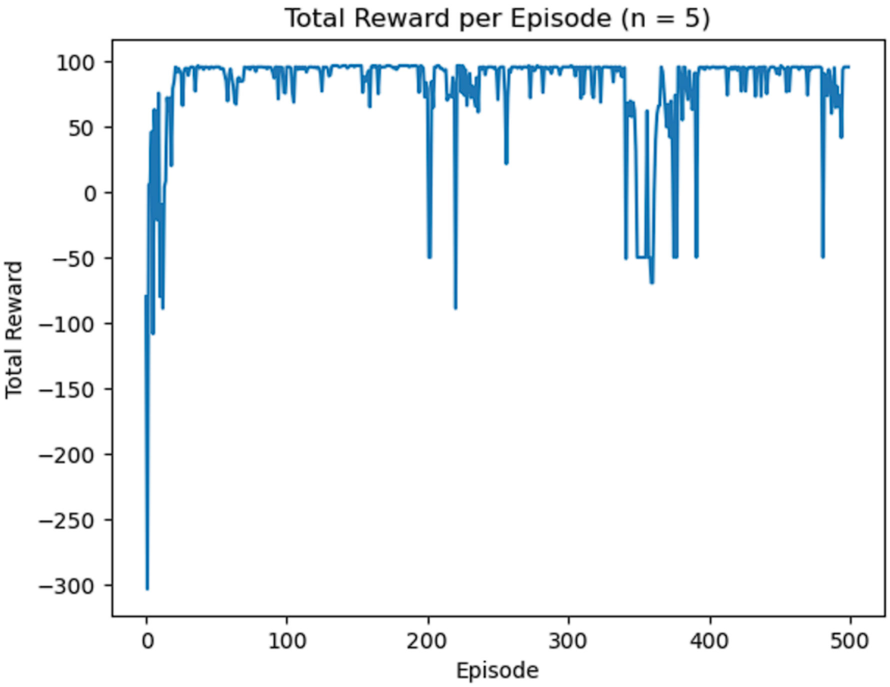


N = 5

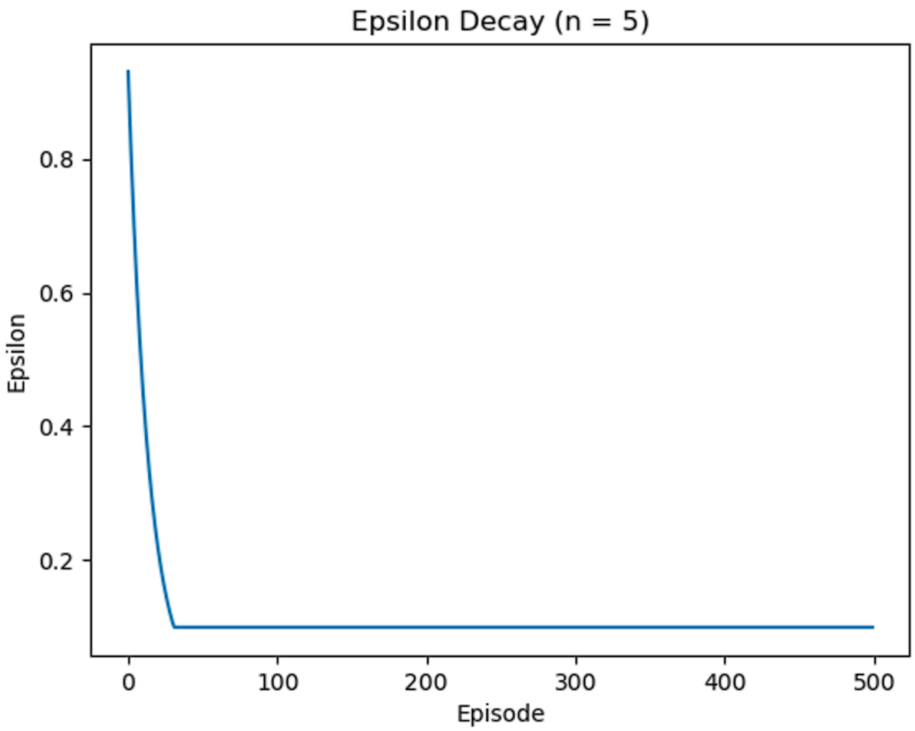
```
Trained Q_A: [[-4.67762594e+00  1.27073114e+01 -1.95542291e+01 -7.01437963e+00]
 [-3.20504134e+00 -9.14888931e-01 -4.48954835e+00 -3.96533537e+00]
 [-2.17452947e+00 -1.93675818e+00  1.97951294e+00 -2.33345768e+00]
 [-4.77851580e+00 -4.02697913e+00 -5.60332993e+00 -2.65289489e+00]
 [-2.79240772e+00 -3.25644374e+00 -3.80148219e+00 -2.79264070e+00]
 [-2.18298498e+00 -4.82628736e+00 -1.61517920e+00 -7.32907684e+00]
 [-2.52892466e+00 -4.57106433e-01 -3.05727632e+00 -1.41277388e+01]
 [-3.99918471e+00 -1.04660380e+01 -2.42482520e+00 -2.68769376e+00]
 [-8.14084689e-01 -3.10083712e+00 -3.05084492e+00 -2.06299700e+00]
 [-1.72193833e+00 -1.53102590e+00  1.08462005e+00 -2.15072457e+00]
 [-1.10115735e+01 -1.83440184e+00 -2.04146966e+00 -7.64392051e+00]
 [-2.49076387e+00 -3.48532694e+00  4.97736565e+01 -2.50153064e+00]
 [-2.80452889e+00 -2.47231230e+00 -2.99919152e+00 -1.42369089e+00]
 [-5.46202060e+00 -4.34074917e+00  2.61027565e+00 -1.85684852e+00]
 [-2.33636761e+00  6.75754011e-01 -1.57541412e+00  2.86746904e+00]
 [-4.41533612e+00  9.20127576e+00 -2.86159805e+00 -5.97819881e+00]
 [-3.87023496e+00 -2.41071791e+00  5.04488691e+00 -2.64123685e+00]
 [ 3.82606643e-02 -6.13870099e+00  9.68033188e-02 -3.30395149e+00]
 [-3.99428477e-01  7.67118732e-02 -1.86539668e+00  7.52919497e-01]
 [ 7.67118732e-02 -4.78452987e+00  0.00000000e+00 -3.52541412e+00]
 [-2.90091258e+00 -2.67675997e+00 -3.09344306e+00 -1.00842712e+00]
 [ 3.11666082e+00  1.19363342e+01 -2.86434246e+00 -2.32595654e+00]
 [-3.43893986e+00  1.58054683e+00 -1.74053511e+00 -2.65292215e+00]
 [-1.01276393e+01  0.00000000e+00 -1.92782291e+00  4.42700296e-01]
 [ 0.00000000e+00  0.00000000e+00  0.00000000e+00  0.00000000e+00]]
```

```
Trained Q_B: [[-7.86777740e+00  8.26940236e+00 -1.37043459e+01 -6.76692126e+00]
 [-2.10394442e+00 -2.17496036e+00 -4.11707179e+00 -3.84002281e+00]
 [-2.88039115e+00 -2.95784668e+00 -2.35660814e+00 -2.49517418e+00]
 [-6.63268270e+00 -3.29971680e+00 -1.05574815e+01 -3.82185224e+00]
 [-2.58155119e+00 -3.12879287e+00 -2.65534434e+00 -2.48229304e+00]
 [-4.50887689e+00 -2.26773060e+00 -2.46834038e+00 -7.43171853e+00]
 [-1.27337892e+00 -1.80393564e+00 -3.33291758e-01 -1.09831466e+01]
 [-2.55942420e+00 -1.10952262e+01 -2.50642664e+00 -4.48349781e+00]
 [ 1.92017985e+00  0.00000000e+00  0.00000000e+00 -6.85213866e-01]
 [-9.84841630e-01 -1.64512585e+00  3.98769620e+00 -4.18227145e+00]
 [-6.36056786e+00 -3.70218916e+00 -4.86037926e+00 -3.75372477e+00]
 [-3.34818227e+00 -1.75840386e+00  5.03063057e+01 -2.59368824e+00]
 [-1.66206720e+00 -4.48630310e+00 -2.77909559e+00 -6.73681159e-01]
 [-8.36246616e+00  0.00000000e+00 -4.15314514e+00 -2.40653401e+00]
 [-4.46344905e-01 -2.36174205e+00 -2.15118656e-01 -5.39995253e+00]
 [-5.65921500e+00  9.58511841e+00 -2.78470580e+00 -2.72126939e+00]
 [-2.64339749e+00 -3.04323687e+00  1.21415970e+01 -2.68778734e+00]
 [-3.41700778e+00  2.58462967e+00  6.22881426e+00  4.15342673e-02]
 [-2.00724571e-01 -2.99255375e-01 -2.00724571e-01 -3.32945338e+00]
 [-5.38471380e-01  0.00000000e+00  0.00000000e+00  0.00000000e+00]
 [-2.62162899e+00 -2.78993865e+00 -2.64759378e+00 -2.66203945e+00]
 [ 4.00619249e+00  1.95726651e+01 -1.91017508e+00 -2.58014920e+00]
 [-2.07722319e+00 -5.98124455e+00 -3.21445772e+00 -1.80592560e+00]
 [-3.95183828e-01  0.00000000e+00 -4.64230834e-01 -7.80965686e-01]
 [ 0.00000000e+00  0.00000000e+00  0.00000000e+00  0.00000000e+00]]
```

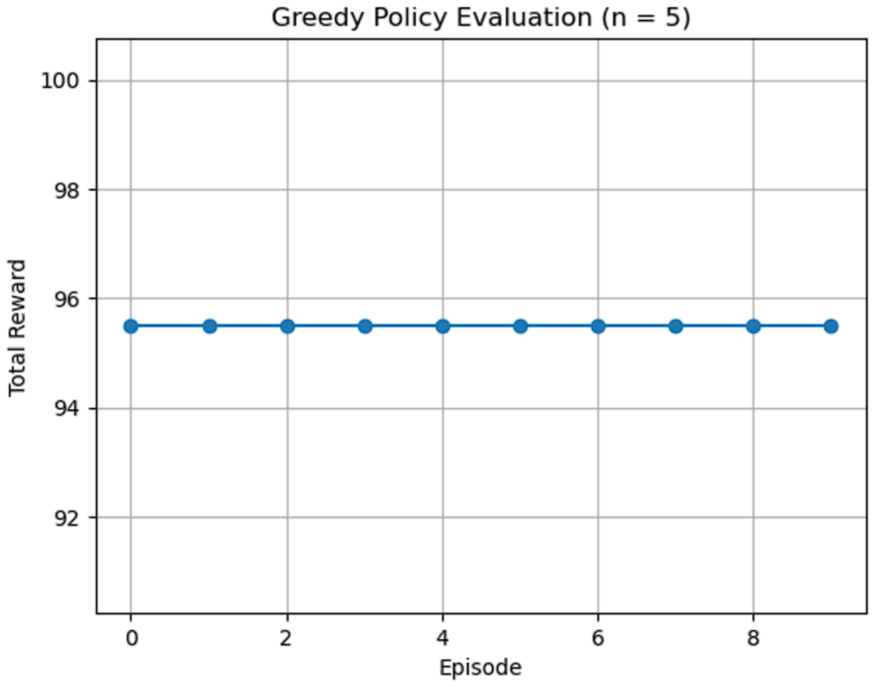
TOTAL REWARDS PER EPISODE



EPSILON DECAY



GREEDY EVALUATION POLICY



Following the analysis, n=3 was determined to be the **ideal value**. It provided the most effective balance between performance and computational expense. The agent successfully learned strategies that optimized rewards while keeping computational requirements within reasonable limits. This selection guaranteed consistent learning dynamics and an acceptable runtime for iterative training.

Comparison of SARSA and N-Step Double Q-Learning:

Hyperparameter Tuning:

- In SARSA Key parameters such as **learning rate (alpha)**, **discount factor (gamma)**, and **epsilon decay** significantly impacted the performance. Tuning hyperparameters for SARSA indicated that moderate values of γ (for instance, 0.9) along with slower rates of epsilon decay achieved the optimal balance between exploration and exploitation.
- In N-step Double Q learning The tuning process indicated that higher values of (γ)gamma (e.g., 0.89) and a slightly faster decay rate (e.g., 0.93) enabled the agent to learn long-term strategies while maintaining efficient exploration.

Performance:

- **SARSA** generally converged more slowly compared to N-Step Double Q-Learning. This is likely due to its on-policy nature, where the updates are limited by the current policy. In comparison, **N-Step Double Q-Learning** reached convergence more quickly, particularly for n=3, because it employs multi-step returns that transmit reward signals over extended periods.

Exploration vs Exploitation:

- The effectiveness of SARSA depended significantly on achieving a balance between exploration and exploitation because it is based on the existing policy.
- N-Step Double Q-Learning leveraged the separation of Q-tables (QA and QB) to better balance exploration and exploitation, particularly in sparse reward environments.

Reward per Episode:

- SARSA showed gradual improvement in total reward per episode but with more pronounced dips during exploration phases.
- N-Step Double Q-Learning showed a more efficient and quicker convergence when n=3, suggesting more effective policy learning.

Epsilon Decay:

- Both algorithms demonstrated considerable sensitivity to the rate of epsilon decay. SARSA needed a slower decay to explore more efficiently, whereas N-Step Double Q-Learning managed to balance exploration and exploitation even with quicker decay rates.

Team Contribution

Team Member	Assignment Part	Contribution (%)
Nikhil Saji	Part 1, 2, 3 & Bonus	50%
VISHNU KRISHNAKUMAR MENON	Part 1, 2, 3 & Bonus	50%