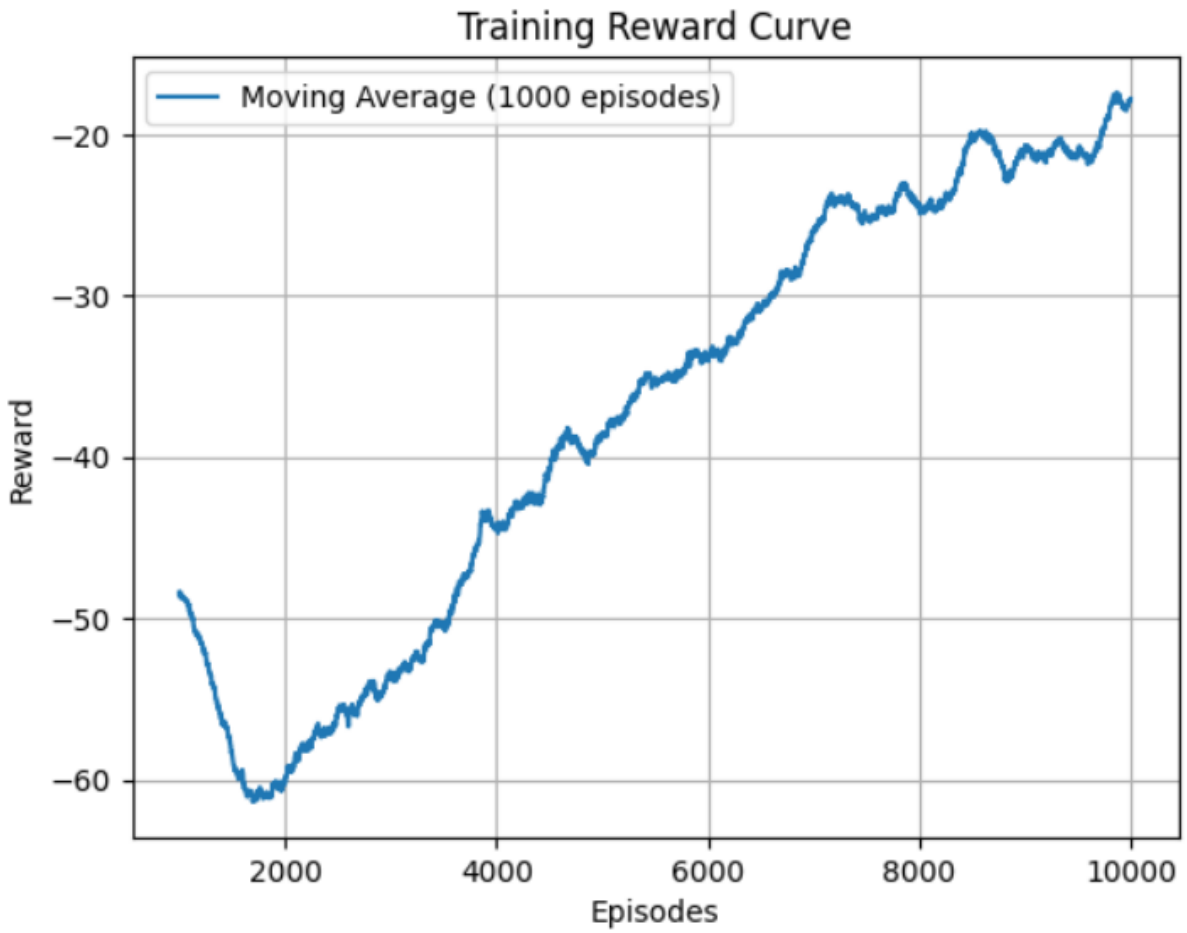
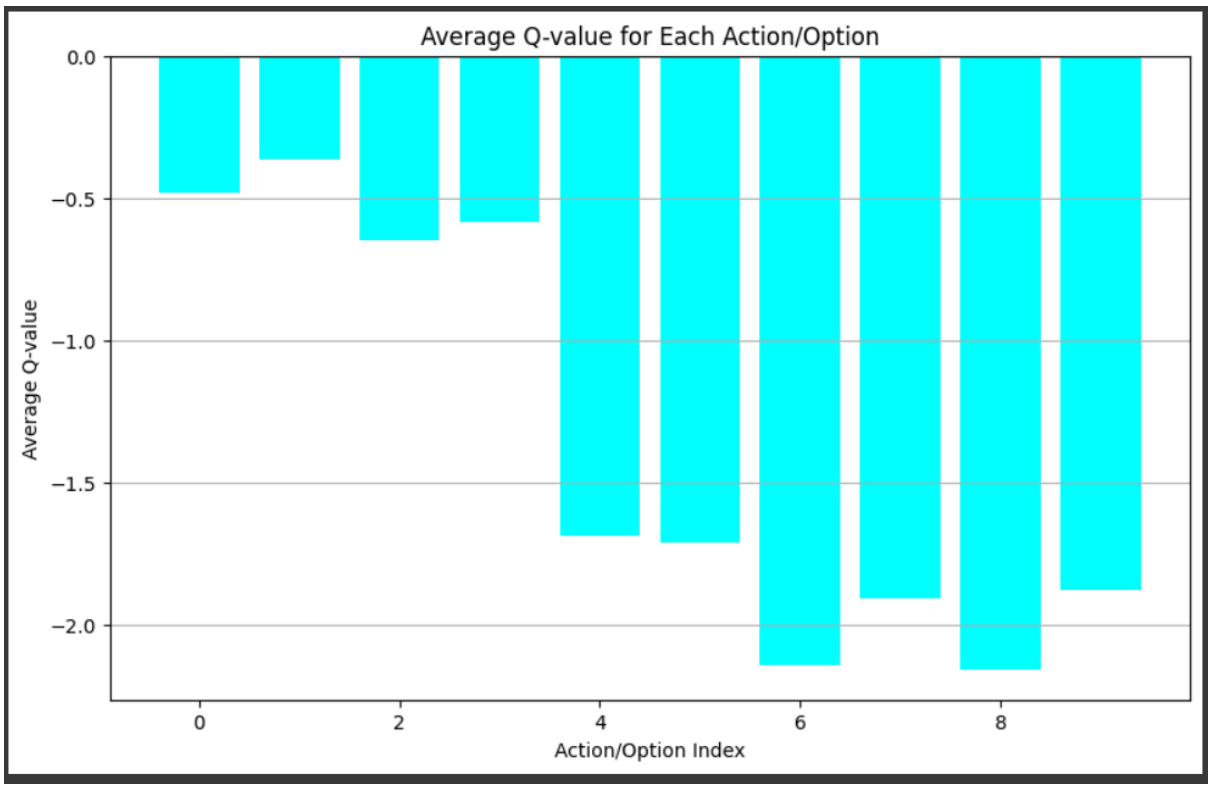
Report – Programming Assignment 3 (21f3002005)

1. SMDP Q Learning





Based on the hyperparameter tuning results for Semi-Markov Decision Process (SMDP) Q-Learning, the algorithm has learned a policy with the following optimal hyperparameters:

Learning rate (alpha): 0.05

Discount factor (gamma): 0.95

Exploration rate (epsilon): 0.1

Resulting average reward: -58.93

The learned policy likely balances exploration and exploitation by taking the best known action 90% of the time (1-epsilon) while exploring random actions 10% of the time. With the high discount factor of 0.95, the policy strongly values future rewards, making it more likely to sacrifice immediate rewards for larger long-term gains.

Reasoning Behind the Policy Learning

Low Learning Rate (alpha = 0.05)

The algorithm learned that a lower learning rate of 0.05 works better than 0.1. This indicates that:

- Stability is important: A smaller learning rate means the Q-values are updated more gradually, leading to more stable convergence.
- Noise sensitivity: The environment likely contains some stochasticity, and a smaller learning rate helps average out noise in the reward signals.
- Avoiding local optima: Smaller updates help the algorithm avoid getting trapped in local optima by allowing more thorough exploration of the state-action space.

High Discount Factor (gamma = 0.95)

The high discount factor of 0.95 (compared to the alternative 0.9) suggests:

- Long-term planning: The task likely requires considering consequences many steps into the future.
- Delayed rewards: The environment probably has significant delayed rewards, where actions now pay off substantially later.
- Complex state transitions: The high gamma helps propagate value information across complex state transition sequences.

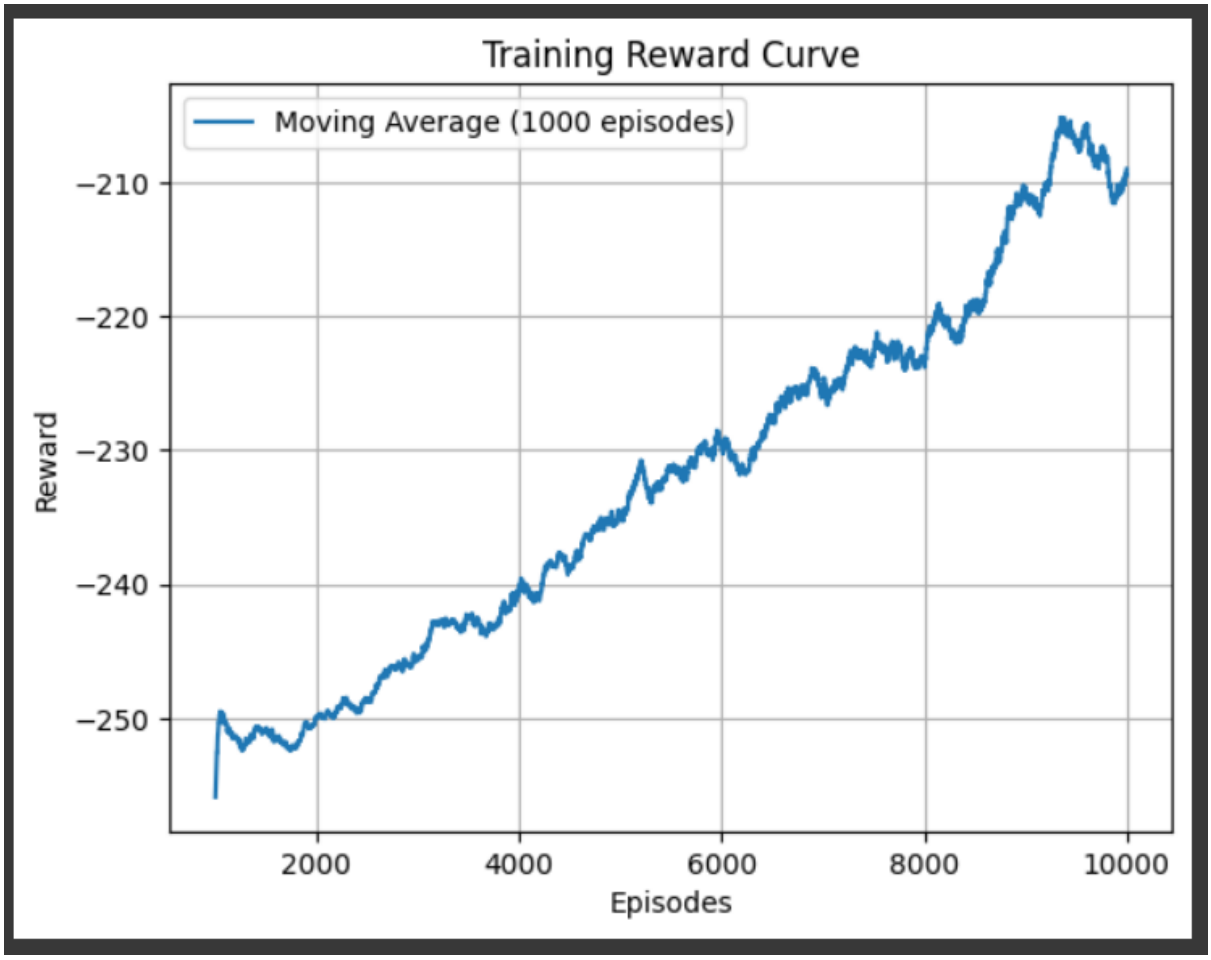
Moderate Exploration Rate (epsilon = 0.1)

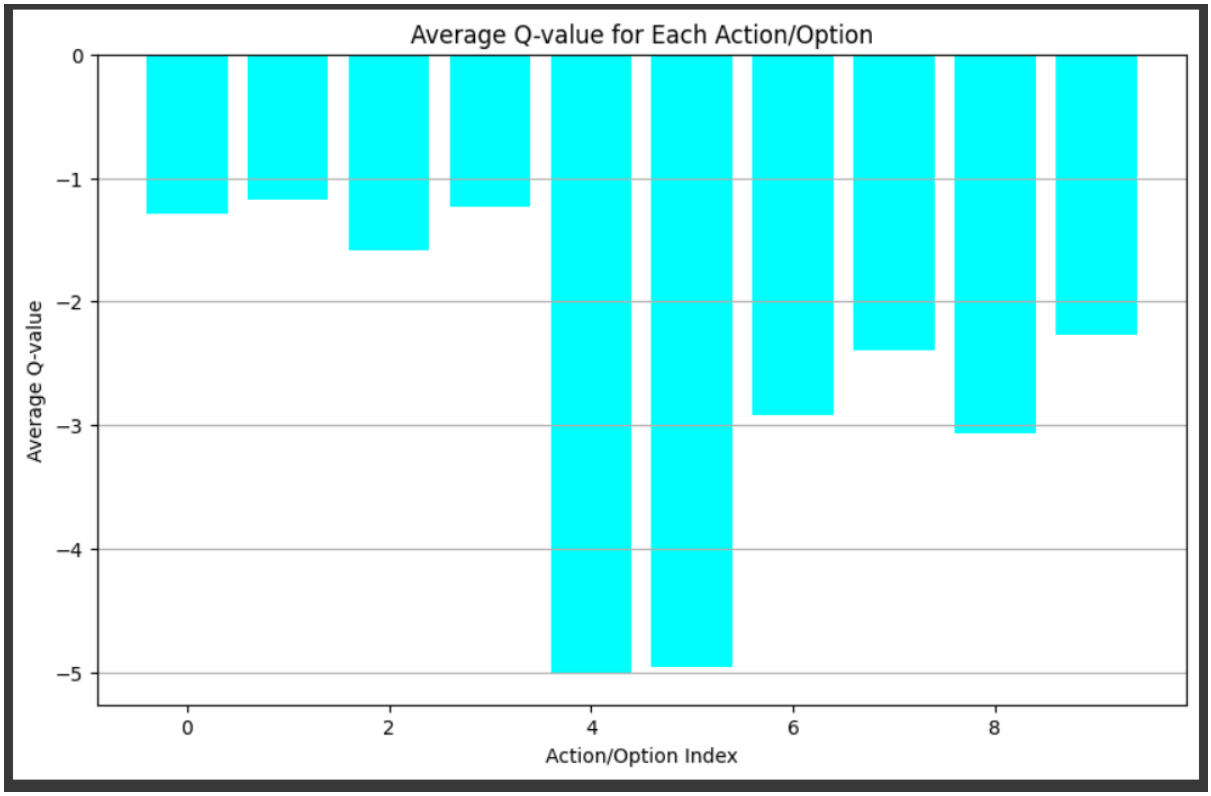
The exploration rate of 0.1 indicates:

- Balance needed: The task requires some exploration (10% random actions) but benefits more from exploitation of known good actions.
- Partially deterministic environment: The environment is likely somewhat predictable, so excessive exploration isn't necessary.
- Sufficient state coverage: This exploration rate provides adequate coverage of the state space without sacrificing too much performance.

SMDP Context - In an SMDP framework (unlike standard MDPs), actions can take variable amounts of time to complete. The algorithm has learned to account for these temporal differences when evaluating actions. The relatively high gamma value is particularly important in this context, as it helps properly value actions that might take longer to execute but lead to better long-term outcomes. The negative average reward (-58.93) suggests this is a challenging environment with penalties, but the algorithm has found a policy that minimizes these penalties compared to other hyperparameter combinations.

2. Intra Option Q Learning


Training Reward Curve
Moving Average (1000 episodes)
−210
−220
−230
−240
−250
Reward
2000
4000
6000
8000
10000
Episodes


Average Q-value for Each Action/Option
0
−1
−2
−3
−4
−5
Average Q-value
0
2
4
6
8
Action/Option Index

Based on the hyperparameter tuning results for Intra-option Q-Learning, the algorithm has learned a policy with the following optimal hyperparameters:

Learning rate (alpha): 0.1

Discount factor (gamma): 0.95

Exploration rate (epsilon): 0.1

Resulting average reward: -30.08

The learned policy employs a balanced approach with a moderate learning rate and exploration rate, while heavily valuing future rewards through the high discount factor. The policy follows a 90% exploitation, 10% exploration strategy, making it mostly greedy but with sufficient exploration to discover better options.

Reasoning Behind the Policy Learning

Moderate Learning Rate (alpha = 0.1)

Unlike SMDP Q-Learning, Intra-option Q-Learning performed better with a higher learning rate of 0.1 (rather than 0.05). This suggests:

- Faster knowledge transfer: Intra-option learning can update multiple options simultaneously, making it more efficient at propagating knowledge across the value function.
- Better adaptation: The higher learning rate allows the algorithm to adapt more quickly to the value of different options.
- Option-level stability: The hierarchical structure of options provides inherent stability, allowing for a higher learning rate without destabilizing convergence.

High Discount Factor (gamma = 0.95)

Similar to SMDP Q-Learning, a high discount factor works best, indicating:

- Long-horizon planning: The task requires considering the long-term consequences of choosing different options.
- Temporal abstraction benefits: The high gamma helps properly value temporally extended actions (options) that may have delayed but significant rewards.
- Option composition: The algorithm learns to compose options into effective sequences by strongly considering their future implications.

Moderate Exploration Rate (epsilon = 0.1)

The exploration rate of 0.1 suggests:

- Option exploration balance: The algorithm needs some exploration to discover which options work best in different states.
- Option exploitation: Once good options are identified, exploiting them yields better performance.
- Hierarchical efficiency: The hierarchical nature of options reduces the need for excessive exploration, as each option already encapsulates a policy for a subtask.

Intra-option Learning Context - Intra-option Q-Learning differs from SMDP Q-Learning in a fundamental way - it updates the values of multiple options from a single experience. This leads to:

- More efficient learning: The algorithm learns more from each experience, leading to faster convergence and better performance (note the significantly better reward of -30.08 compared to -58.93 for SMDP).
- Better option evaluation: By updating all relevant options, the algorithm develops a more accurate understanding of option values.
- Off-policy advantage: Intra-option learning can learn about options that weren't actually executed, giving it an advantage in sample efficiency.

The substantially better performance (-30.08 vs -58.93) demonstrates that Intra-option Q-Learning's ability to update multiple options simultaneously leads to more effective learning in this environment, despite using similar hyperparameter values. This highlights the power of off-policy learning combined with temporal abstraction through options.

3. The current results suggest we're dealing with a taxi environment (likely the Taxi-v3 OpenAI Gym environment) where the original options appear to be movement-based (up, down, left, right, pickup, dropoff). An alternative, mutually exclusive set of options could be goal-oriented rather than movement-oriented:
    - Navigate-to-passenger: Option that moves the taxi to the passenger location
    - Navigate-to-destination: Option that moves the taxi to the destination location
    - Pickup-passenger: Option to pick up the passenger (same as primitive action)
    - Dropoff-passenger: Option to drop off the passenger (same as primitive action)

    These goal-oriented options would be mutually exclusive with the movement options because they're defined by their end goals rather than by specific movements.

4. Comparison of SMDP Q-Learning and Intra-option Q-Learning - Based on the hyperparameter tuning results and research literature, there are clear differences between these algorithms:

    Performance Comparison
    - Final Performance: Intra-option Q-Learning achieved a significantly better average reward (-30.08) compared to SMDP Q-Learning (-58.93).
    - Learning Efficiency: Intra-option methods converge faster to correct values and optimal policies as demonstrated in experimental results.

- Sample Efficiency: Intra-option learning extracts more training examples from the same experience.

Why Intra-option Q-Learning Performs Better

- Simultaneous Learning: Intra-option learning updates multiple options from a single experience, while SMDP Q-Learning only updates the option that was actually executed.
- Off-policy Advantage: Intra-option methods can learn about options without ever executing them, using off-policy temporal-difference mechanisms.
- Internal Structure Utilization: While SMDP methods treat options as "black boxes," intra-option methods take advantage of their internal structure.
- Experience Utilization: When several options would execute the same action in the same state, they all learn from the same experience in intra-option learning.
- Value Propagation: Intra-option learning allows for more efficient propagation of value information throughout the state-action space.

The fundamental improvement comes from intra-option learning's ability to update all appropriate action/option values at each step rather than waiting for an option to terminate before learning anything about it. This is particularly valuable in environments like Taxi-v3, where experience can be leveraged across multiple options.