Explaining my files for better project assessment.

Files Name	Description
part_2_files	This folder contains Part 2 pickle, console output, and final renders.
|———deterministic_double_q_table.pkl	Deterministic Environment Double Q **Pickle File**
|———deterministic_q_table.pkl	Deterministic Environment Q **Pickle File**
|———hyperparams_1_stochastic_q_table.pkl	Stochastic Environment Q **Pickle File** for Hyperparams Set 1
|———hyperparams_2_stochastic_q_table.pkl	Stochastic Environment Q **Pickle File** for Hyperparams Set 2
|———hyperparams_3_stochastic_q_table.pkl	Stochastic Environment Q **Pickle File** for Hyperparams Set 3
|———hyperparams_4_stochastic_q_table.pkl	Stochastic Environment Q **Pickle File** for Hyperparams Set 4
|———hyperparams_5_stochastic_q_table.pkl	Stochastic Environment Q **Pickle File** for Hyperparams Set 5
|———hyperparams_6_stochastic_q_table.pkl	Stochastic Environment Q **Pickle File** for Hyperparams Set 6
|———hyperparams_base_stochastic_q_table.pkl	Stochastic Environment Q **Pickle File** for Hyperparams Initial set
|———stochastic_double_q_table.pkl	Stochastic Environment Double Q **Pickle File**
|——— evaluate_agent_deterministic_DoubleQ_console_output.txt	**Evaluation logs** for debugging purposes for Deterministic Env Double Q
|———evaluate_agent_deterministic_Q_console_output.txt	**Evaluation logs** for debugging purposes for Deterministic Env Q
|———evaluate_agent_stochastic_DoubleQ_console_output.txt	**Evaluation logs** for debugging purposes for Stochastic Env Double Q
|———evaluate_agent_stochastic_Q_console_output.txt	**Evaluation logs** for debugging purposes for Stochastic Env Q
|———final_evaluate_agent_deterministic_double_q.txt	Final **episode 1 render** for the Deterministic Environment Double Q
|———final_evaluate_agent_deterministic_Q.txt	Final **episode 1 render** for the Deterministic Environment Q
|———final_evaluate_hyperparams_1_stochastic_q.txt	Final **episode 1 render** for the Stochastic Environment Q Hyperparams Set 1
|———final_evaluate_hyperparams_2_stochastic_q.txt	Final **episode 1 render** for the Stochastic Environment Q Hyperparams Set 2
|———final_evaluate_hyperparams_3_stochastic_q.txt	Final **episode 1 render** for the Stochastic Environment Q Hyperparams Set 3
|———final_evaluate_hyperparams_4_stochastic_q.txt	Final **episode 1 render** for the Stochastic Environment Q Hyperparams Set 4
|———final_evaluate_hyperparams_5_stochastic_q.txt	Final **episode 1 render** for the Stochastic Environment Q Hyperparams Set 5
|———final_evaluate_hyperparams_6_stochastic_q.txt	Final **episode 1 render** for the Stochastic Environment Q Hyperparams Set 6
|———final_evaluate_hyperparams_base_stochastic_q.txt	Final **episode 1 render** for the Stochastic Environment Q Hyperparams Initial Set
|———final_evaluate_stochastic_double_q.txt	Final **episode 1 render** for the Stochastic Environment Double Q
|———train_agent_deterministic_DoubleQ_console_output.txt	**Training logs** for debugging purposes for Deterministic Env Double Q
|———train_agent_deterministic_Q_console_output.txt	**Training logs** for debugging purposes for Deterministic Env Q
|———train_agent_stochastic_DoubleQ_console_output.txt	**Training logs** for debugging purposes for Stochastic Env Double Q
|———train_agent_stochastic_Q_console_output.txt	**Training logs** for debugging purposes for Stochastic Env Q
part_3_files	This folder contains Part 3 pickle
|———part_3_q_table.pkl	Stock Market Q-Table Pickle File
NVDA.csv	Given In Assessmemt
ssaurav_assignment1_part1.ipynb	**Assessment Part 1 Code**
ssaurav_assignment1_part2.ipynb	**Assessment Part 2 Code**
ssaurav_assignment1_part3.ipynb	**Assessment Part 3 Code**

CSE 4/546: Reinforcement Learning
Spring 2025

Instructor: Alina Vereshchaka

Assignment 1 - Defining & Solving RL Environments

 # References

- https://gymnasium.farama.org/environments/toy_text/
- https://gymnasium.farama.org/tutorials/gymnasium_basics/environment_creation/#sphx-glr-tutorials-gymnasium-basics-environment-creation-py
- https://cs.stanford.edu/people/karpathy/reinforcejs/

Part 1: Defining RL Environments [30 points]

Describe the deterministic and stochastic environments, including their sets of actions, states, rewards, main objectives, etc.

For this assignment, I have chosen an Autonomous Drone Delivery environment, where a drone must pick up two packages and deliver them to their respective destinations. There are two different scenarios in this environment:

Deterministic Environment

A deterministic environment means that the outcomes of actions are fixed and predictable. The same action in the same state always leads to the same result.

Challenges in the Deterministic Environment:

- The environment is fixed, meaning all elements (drone, packages, delivery locations, and obstacles) remain the same in every run.
- There is a tornado, which acts as a no-fly zone, meaning the drone cannot pass through it.
- The tornado is surrounded by wind that pushes the drone in a specific direction. If the drone enters a wind-affected cell, it has a higher probability of moving in the wind's direction.
- And two birds act as obstacles, each imposing a negative reward.
- Since the environment does not change dynamically, the drone can learn an optimal path and follow it every time without unexpected disturbances.

Stochastic Environment

A stochastic environment means that outcomes are random to some degree. The same action in the same state may lead to different results each time.

Challenges in the Stochastic Environment:

- Random Placement: Every time the environment is initialised, the drone, packages, delivery locations, birds, tornado, and wind zones appear in random positions on the grid.
- Dynamic Elements:
  - The tornado moves around the grid, changing its location at every time step.
  - The wind direction also shifts, making navigation harder.
  - There are birds, which might collide with the drone.
- Increased Difficulty: The drone must learn and adapt to different scenarios in each run instead of relying on a fixed strategy.
- This makes the task more complex, as the drone cannot follow a fixed path. Instead, it must learn to make decisions on the fly based on changing conditions.

Environment Setup

The environment is represented as a 6x6 grid that models a city layout where the drone performs deliveries.

Grid Properties:

- Size: 6x6 cells.
- Obstacles: No-fly zones (static obstacles) that the drone cannot cross.
- Goal: The drone must pick up both packages and deliver them to their respective destinations.
- Actions Available:

The drone can take the following actions:

- Up
- Down
- Left
- Right
- Pick up a package
- Drop off a package

Rewards System:

The drone earns or loses points based on its actions and performance.

Event	Reward/Penalty
Successfully delivering a package	**100**
Picking up a package	**25**
Entering a no-fly zone (tornado)	**-100**
Getting hit by a bird	**-50**
Being pushed by the wind (any direction: up, down, left, right)	**-10**
Taking a step (movement cost)	**-2**
Dropping a package at the wrong location	**-50**
Attempting to pick up when not at a package location	**-25**
Moving out of the grid (invalid move)	**-25**
Repeating the same move unnecessarily	**-50**

Note: Some rewards are missing in the current checkpoint and will be added in Part 2 to finalise the reward system for optimisation. This includes penalties for failed drop-offs and failed pick-ups.

Terminal State

The drone successfully delivers both packages and earns maximum rewards.

Provide visualisations of your environments.

How did you define the stochastic environment?

In my stochastic environment, all objects are placed randomly at the start, and some elements continue to change as the drone moves. This creates an unpredictable environment, making it more challenging for the drone to complete its task.

Key Features of the Stochastic Environment:

Random Initialization:

At the beginning of each simulation, the positions of the drone, two packages, two delivery locations, tornado, wind zones, and birds are all placed randomly on the grid.

This ensures that the drone does not start in the same location every time and must adapt to different scenarios.

Dynamic Tornado and Wind Movement:

The tornado, which acts as a no-fly zone, moves randomly during the simulation.
The wind zones surrounding the tornado also change their position, increasing the difficulty of navigation.
If the drone enters a wind zone, it has a higher probability of being pushed in the wind's direction.

Real-Time Decision Making:

Because the tornado and wind move unpredictably, the drone cannot rely on a fixed path like in a deterministic environment.
Instead, it must constantly analyse the environment and make real-time decisions to avoid obstacles while still reaching the delivery destinations efficiently.

What is the difference between deterministic and stochastic environments?

Deterministic Environment:

- Everything in the environment is fixed and predictable.
- The drone, packages, delivery locations, tornado, and wind zones stay in the same positions every time the simulation starts.
- The tornado is a no-fly zone, and the wind always pushes the drone in the same direction.
- The drone can learn an optimal path and follow it without surprises.

Stochastic Environment:

- Everything in the environment is random and changes over time.
- The drone, packages, delivery locations, birds, tornado, and wind zones are placed randomly at the start.
- During the drone's movement, the tornado moves randomly with wind directions, making navigation harder.
- The drone must make real-time decisions instead of following a fixed path.

Main Difference:

In the deterministic environment, the outcome of every action is always the same, making it easier to plan a path.
In the stochastic environment, randomness makes each run different, forcing the drone to adapt to new challenges every time.

Safety in AI: Write a brief review (~5 sentences) explaining how you ensure the safety of your environments. E.g. how do you ensure that agent choose only actions that are allowed, that agent is navigating within defined state-space, etc.

To ensure the safety of my environment, I apply strict checks to prevent illegal actions and enforce correct behavior:

Boundary Checks: Before moving up, down, left, or right, the drone checks if the movement stays within the grid boundaries. If an action would move it outside, it is not allowed.

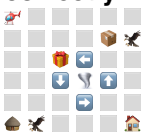
Valid Pick-Up Action: The drone can only pick up a package if it is in the same cell as the package. If it tries to pick up from an empty location, it receives a penalty.

Valid Drop-Off Action: The drone can only drop off a package if it is in the correct destination cell. If it drops a package at the wrong location, the package is moved to a new box, and the drone gets a negative reward.

No-Fly Zones & Obstacles: The drone is not allowed to enter tornado zones or other restricted areas. If it tries, it gets a large penalty.

By enforcing these rules, I ensure that the agent only takes legal actions, stays within the grid, and follows correct pick-up and drop-off procedures while learning to navigate efficiently. 🚁✅

Random Drone movement in deterministic env for 10+ steps to show that the environment logic is defined correctly.



```
Drone moved to (1, 0) Accumulated
Reward: 0
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
```

```
Attempted pickup failed
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
Attempted pickup failed
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
Drone moved to (2, 0) Accumulated
Reward: 0
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
```

```
Drone moved to (3, 0) Accumulated
Reward: 0
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
```

```
Drone moved to (2, 0) Accumulated
Reward: 0
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
```

```
Attempted dropoff failed (no package
carried)
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
Drone moved to (3, 0) Accumulated
Reward: 0
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
```

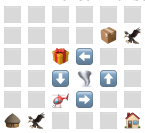
```
Drone moved to (3, 1) Accumulated
Reward: 0
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
```

```
Drone moved to (3, 2) Accumulated
Reward: -10
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
```

```
Attempted pickup failed
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
Drone moved to (3, 3) Accumulated
Reward: -110
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
```

```
Drone moved to (2, 3) Accumulated
Reward: -120
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
```

```
Attempted dropoff failed (no package
carried)
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
Attempted pickup failed
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
Drone moved to (2, 4) Accumulated
Reward: -120
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
```



```
Drone moved to (1, 4) Accumulated
Reward: -120
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
```

```
Attempted dropoff failed (no package
carried)
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
Drone moved to (1, 5) Accumulated
Reward: -170
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
```

```
Drone moved to (1, 4) Accumulated
Reward: -170
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
```

```
Attempted dropoff failed (no package
carried)
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
Attempted dropoff failed (no package
carried)
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
Drone moved to (2, 4) Accumulated
Reward: -170
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
```

```
Drone moved to (3, 4) Accumulated
Reward: -180
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
```

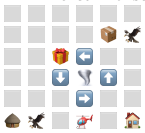
```
Drone moved to (3, 5) Accumulated
Reward: -180
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
```

```
Attempted dropoff failed (no package
carried)
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
Attempted dropoff failed (no package
carried)
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
Drone moved to (4, 5) Accumulated
Reward: -180
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
```

```
Drone moved to (3, 5) Accumulated
Reward: -180
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
```

```
Drone moved to (3, 4) Accumulated
Reward: -190
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
```

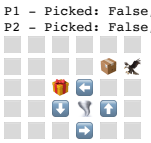
```
Attempted pickup failed
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
Drone moved to (3, 5) Accumulated
Reward: -190
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
```



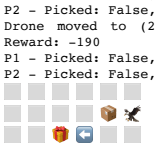
```
Attempted dropoff failed (no package
carried)
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
Drone moved to (2, 5) Accumulated
Reward: -190
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
```

```
Attempted pickup failed
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
Drone moved to (2, 4) Accumulated
Reward: -190
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
```

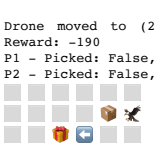
```
Attempted pickup failed
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
Drone moved to (2, 5) Accumulated
Reward: -190
```



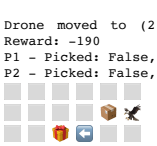
```
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
```



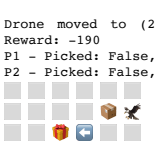
```
Attempted pickup failed
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
Drone moved to (2, 4) Accumulated
Reward: -190
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
```



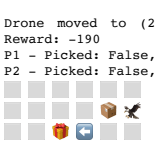
```
Drone moved to (2, 5) Accumulated
Reward: -190
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
```



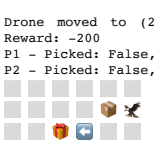
```
Drone moved to (2, 4) Accumulated
Reward: -190
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
```



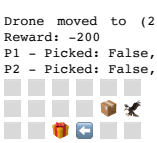
```
Drone moved to (2, 5) Accumulated
Reward: -190
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
```



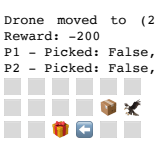
```
Drone moved to (2, 4) Accumulated
Reward: -190
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
```



```
Drone moved to (2, 3) Accumulated
Reward: -200
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
```

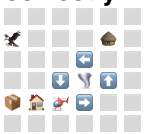


```
Drone moved to (2, 4) Accumulated
Reward: -200
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
```



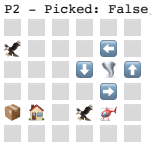
```
Drone moved to (2, 5) Accumulated
Reward: -200
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
```

Random Drone movement in stochastic env for 10+ steps to show that the environment logic is defined correctly.



Attempted dropoff failed (no package carried)
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
Drone moved to (3, 4) Accumulated Reward: -50
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False

Drone moved to (4, 4) Accumulated Reward: -50
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False



Attempted dropoff failed (no package carried)
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
Drone moved to (5, 4) Accumulated Reward: -50
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False

Attempted dropoff failed (no package carried)
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
Drone moved to (4, 4) Accumulated Reward: -150
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False

Drone moved to (3, 4) Accumulated Reward: -200
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False

Attempted dropoff failed (no package carried)
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
Drone moved to (4, 4) Accumulated Reward: -210
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False

Drone moved to (4, 3) Accumulated Reward: -210
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False

Drone moved to (5, 3) Accumulated Reward: -210

P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False

Attempted dropoff failed (no package carried)
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
Drone moved to (4, 3) Accumulated Reward: -210
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False

Attempted dropoff failed (no package carried)
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
Attempted dropoff failed (no package carried)
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
Picked up package 1
P1 - Picked: True, Delivered: False
P2 - Picked: False, Delivered: False
Drone moved to (4, 4) Accumulated Reward: -210
P1 - Picked: True, Delivered: False
P2 - Picked: False, Delivered: False

Dropped package 1 incorrectly
Attempted dropoff failed (no package carried)
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
Drone moved to (4, 3) Accumulated Reward: -210
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False

Drone moved to (5, 3) Accumulated Reward: -210
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False

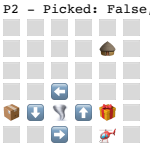
Drone moved to (5, 4) Accumulated Reward: -210
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False

Drone moved to (5, 3) Accumulated Reward: -210
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False

Drone moved to (5, 4) Accumulated Reward: -210
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False

Drone moved to (4, 4) Accumulated Reward: -210
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False

Drone moved to (4, 5) Accumulated Reward: -210
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False



Drone moved to (3, 5) Accumulated Reward: -220
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False

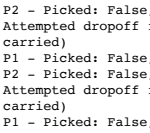
Drone moved to (2, 5) Accumulated Reward: -230
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False

Attempted dropoff failed (no package carried)
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
Attempted pickup failed
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
Attempted dropoff failed (no package carried)
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
Drone moved to (3, 5) Accumulated Reward: -240
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False

Drone moved to (4, 5) Accumulated Reward: -240
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False

Drone moved to (5, 5) Accumulated Reward: -240
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False

Attempted dropoff failed (no package carried)
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
Attempted pickup failed
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
Attempted dropoff failed (no package carried)
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
Drone moved to (4, 5) Accumulated Reward: -240
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False



Attempted dropoff failed (no package carried)
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
Attempted dropoff failed (no package carried)
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
Drone moved to (5, 4) Accumulated Reward: -240
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False

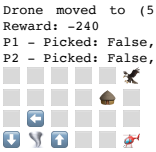
Attempted dropoff failed (no package carried)
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
Attempted dropoff failed (no package carried)
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
Attempted dropoff failed (no package carried)
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
Attempted dropoff failed (no package carried)
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
Drone moved to (4, 4) Accumulated Reward: -240
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False

Drone moved to (4, 3) Accumulated Reward: -240
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False

Drone moved to (4, 4) Accumulated Reward: -240
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False

Drone moved to (4, 3) Accumulated Reward: -240
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False

Attempted dropoff failed (no package carried)
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False
Drone moved to (5, 3) Accumulated Reward: -240
P1 - Picked: False, Delivered: False
P2 - Picked: False, Delivered: False



Part 2 [Total: 40 points] - Applying Tabular Methods

Deterministic Environment:

First, I applied **Q-learning** with the hyper parameters below, and the Q-table is saved in a **ssaurav_assignment1_final/deterministic_q_table.pkl** file, as shown in the screenshot.

```
# ---- My hyperparameters ----
hyperparams = {
    'alpha': 0.01,
    'gamma': 0.95,
    'epsilon': 1.0,
    'epsilon_decay': 0.995,
    'epsilon_min': 0.01,
    'episodes': 10000,
    'max_steps': 1000
}

print("Training a Q-Learning agent ...")

Q_deterministic_Q, rewards_deterministic_Q, eps_history_deterministic_Q = train_agent_deterministic_Q(env, hyperparams, render=True)
```

✓ 20.5s

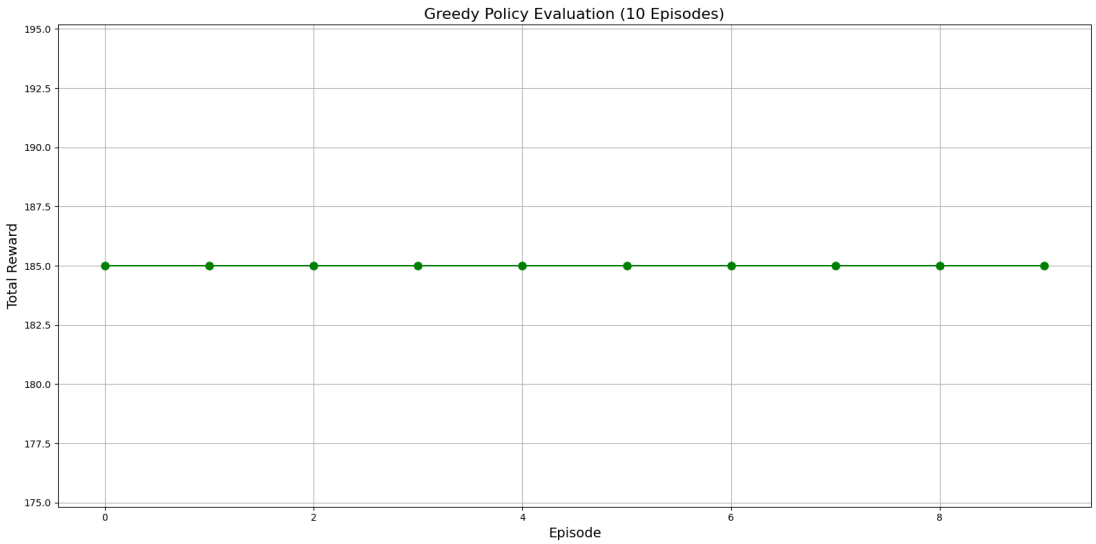
```
Training a Q-Learning agent ...
Episode 100/10000 | Eps: 0.6058 | Success in last 100: 0 %
Episode 200/10000 | Eps: 0.3670 | Success in last 100: 0 %
Episode 300/10000 | Eps: 0.2223 | Success in last 100: 0 %
Episode 400/10000 | Eps: 0.1347 | Success in last 100: 4 %
Episode 500/10000 | Eps: 0.0816 | Success in last 100: 13 %
Episode 600/10000 | Eps: 0.0494 | Success in last 100: 26 %
Episode 700/10000 | Eps: 0.0299 | Success in last 100: 21 %
Episode 800/10000 | Eps: 0.0181 | Success in last 100: 43 %
Episode 900/10000 | Eps: 0.0110 | Success in last 100: 69 %
Episode 1000/10000 | Eps: 0.0100 | Success in last 100: 77 %
Episode 1100/10000 | Eps: 0.0100 | Success in last 100: 86 %
Episode 1200/10000 | Eps: 0.0100 | Success in last 100: 95 %
Episode 1300/10000 | Eps: 0.0100 | Success in last 100: 96 %
Episode 1400/10000 | Eps: 0.0100 | Success in last 100: 98 %
Episode 1500/10000 | Eps: 0.0100 | Success in last 100: 93 %
Episode 1600/10000 | Eps: 0.0100 | Success in last 100: 98 %
Episode 1700/10000 | Eps: 0.0100 | Success in last 100: 91 %
Episode 1800/10000 | Eps: 0.0100 | Success in last 100: 86 %
Episode 1900/10000 | Eps: 0.0100 | Success in last 100: 100 %
Episode 2000/10000 | Eps: 0.0100 | Success in last 100: 87 %
Episode 2100/10000 | Eps: 0.0100 | Success in last 100: 97 %
Episode 2200/10000 | Eps: 0.0100 | Success in last 100: 93 %
Episode 2300/10000 | Eps: 0.0100 | Success in last 100: 88 %
Episode 2400/10000 | Eps: 0.0100 | Success in last 100: 95 %
...
Task complete count: 8408
Episode 10000/10000 | Eps: 0.0100 | Success in last 100: 87 %

Q-table saved to deterministic_q_table.pkl
```

```
# Evaluating the agent using a greedy policy
evaluation_rewards_deterministic_Q = evaluate_agent_determ
                                                    episod

# Plotting greedy policy evaluation results
plt.figure(figsize=(16,8))
plt.plot(evaluation_rewards_deterministic_Q, marker='o', l
plt.xlabel("Episode", fontsize=14)
plt.ylabel("Total Reward", fontsize=14)
plt.title("Greedy Policy Evaluation (10 Episodes)", fontsi
plt.grid(True)
plt.tight_layout()
plt.show()
```

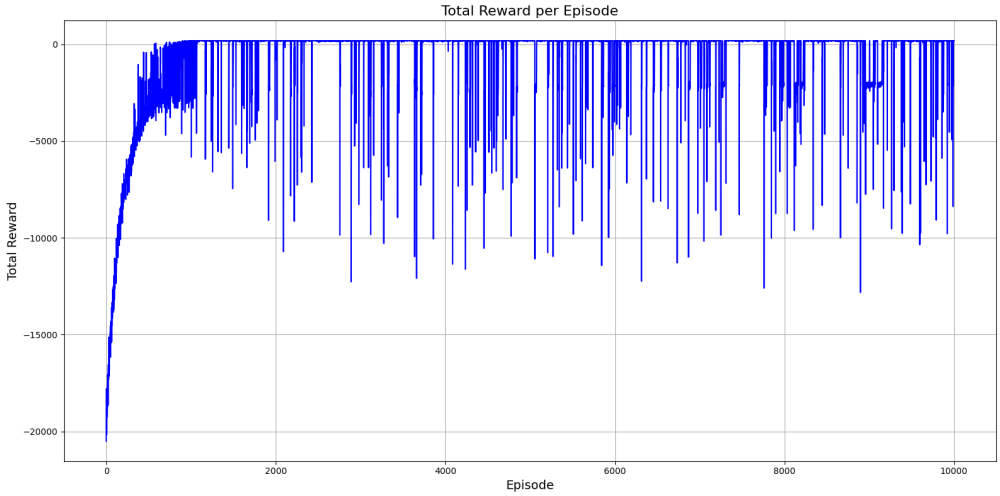
✓ 0.1s

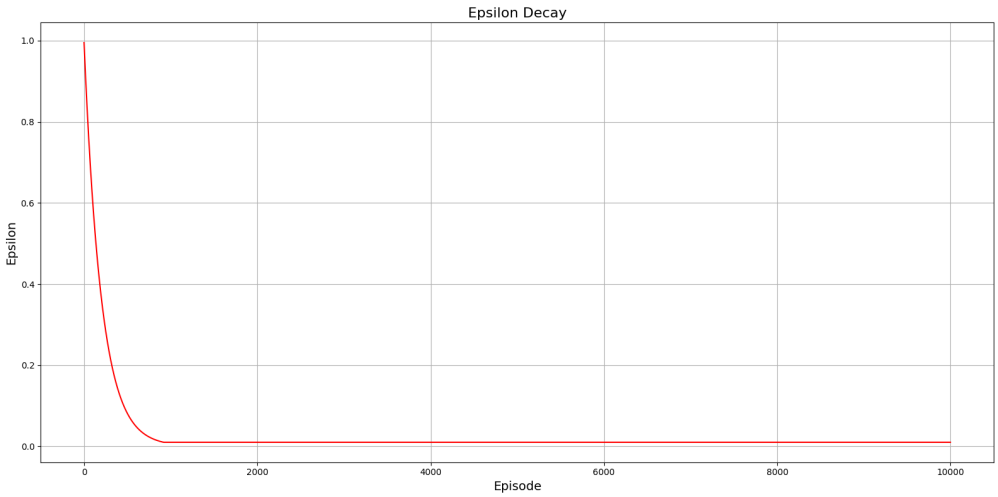
```
Evaluation Episode 1: Steps: 25 | Total Reward: 185
Evaluation Episode 2: Steps: 25 | Total Reward: 185
Evaluation Episode 3: Steps: 25 | Total Reward: 185
Evaluation Episode 4: Steps: 25 | Total Reward: 185
Evaluation Episode 5: Steps: 25 | Total Reward: 185
Evaluation Episode 6: Steps: 25 | Total Reward: 185
Evaluation Episode 7: Steps: 25 | Total Reward: 185
Evaluation Episode 8: Steps: 25 | Total Reward: 185
Evaluation Episode 9: Steps: 25 | Total Reward: 185
Evaluation Episode 10: Steps: 25 | Total Reward: 185
Task complete count: 10
```

I ran the evaluation for 10 episodes and achieved a maximum reward of 185 in 25 steps with a 100% successful task completion rate.



Graph showing total rewards per episode and the epsilon decay trend.

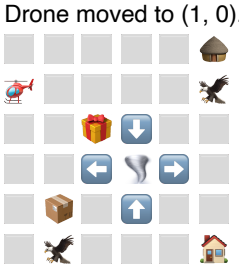




Render each step in greedy approach.

--- Evaluation Episode 1 starting ---

Drone moved to (1, 0). Step reward: -2



Evaluation Episode 1 - Step 1

Drone moved to (1, 1). Step reward: -2

Drone moved to (2, 1). Step reward: -2

Drone moved to (3, 1). Step reward: -2

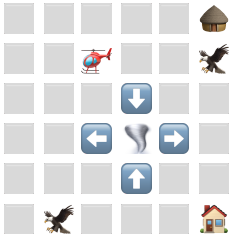
Drone moved to (4, 1). Step reward: -2

Picked up package 2 for 25 reward
Drone moved to (3, 1). Step reward: -2

Drone moved to (2, 1). Step reward: -2

Drone moved to (2, 2). Step reward: -2

Picked up package 1 for 25 reward
Drone moved to (1, 2). Step reward: -2



Evaluation Episode 1 - Step 11

Drone moved to (1, 3). Step reward: -2

Drone moved to (0, 3). Step reward: -2

Drone moved to (0, 4). Step reward: -2

Drone moved to (0, 5). Step reward: -2

Dropped package_1 incorrectly. Penalty -50
Delivered package_2 for +100 reward
Picked up package 1 for 25 reward
Drone moved to (0, 4). Step reward: -2

Drone moved to (1, 4). Step reward: -2

Drone moved to (2, 4). Step reward: -2

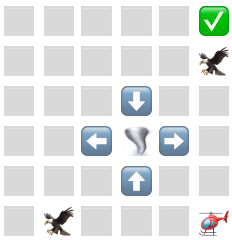
Drone moved to (2, 5). Step reward: -2

Evaluation Episode 1 - Step 21

Drone moved to (3, 5). Step reward: -2

Drone moved to (4, 5). Step reward: -2

Drone moved to (5, 5). Step reward: -2



Delivered package_1 for +100 reward
Task complete: All packages delivered 😎

Secondly, I applied **Double Q-learning** with the hyperparameters below, and the Double Q-table is saved in a **deterministic_double_q_table.pkl** file, as shown in the screenshot.

```python
# Choosing environment type: deterministic
deterministic = True
env = Environment(0, 0, stochastic=(not deterministic))

# Setting hyperparameters
hyperparams = {
    'alpha': 0.01,
    'gamma': 0.95,
    'epsilon': 1.0,
    'epsilon_decay': 0.995,
    'epsilon_min': 0.01,
    'episodes': 5000,
    'max_steps': 1000
}

print("Training Q-Learning agent ...")
(Q1_deterministic_DoubleQ, Q2_deterministic_DoubleQ), rewards_deterministic_DoubleQ, eps_history_deterministic_DoubleQ = train_agent_deterministic_DoubleQ(env, hyperparams, render=True)
```

✓ 12.0s Python

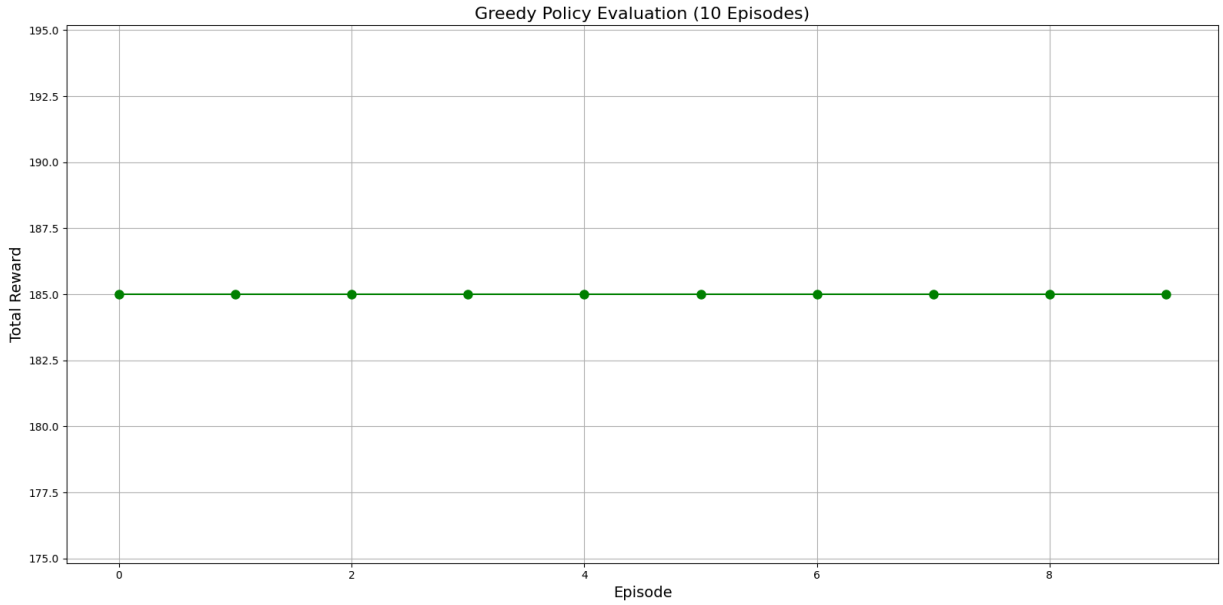
```
Training Q-Learning agent ...
Episode 100/5000 | Epsilon: 0.6058 | Success in last 100: 0 %
Episode 200/5000 | Epsilon: 0.3670 | Success in last 100: 0 %
Episode 300/5000 | Epsilon: 0.2223 | Success in last 100: 0 %
Episode 400/5000 | Epsilon: 0.1347 | Success in last 100: 7 %
Episode 500/5000 | Epsilon: 0.0816 | Success in last 100: 8 %
Episode 600/5000 | Epsilon: 0.0494 | Success in last 100: 22 %
Episode 700/5000 | Epsilon: 0.0299 | Success in last 100: 41 %
Episode 800/5000 | Epsilon: 0.0181 | Success in last 100: 54 %
Episode 900/5000 | Epsilon: 0.0110 | Success in last 100: 69 %
Episode 1000/5000 | Epsilon: 0.0100 | Success in last 100: 72 %
Episode 1100/5000 | Epsilon: 0.0100 | Success in last 100: 73 %
Episode 1200/5000 | Epsilon: 0.0100 | Success in last 100: 73 %
Episode 1300/5000 | Epsilon: 0.0100 | Success in last 100: 68 %
Episode 1400/5000 | Epsilon: 0.0100 | Success in last 100: 70 %
Episode 1500/5000 | Epsilon: 0.0100 | Success in last 100: 75 %
Episode 1600/5000 | Epsilon: 0.0100 | Success in last 100: 72 %
Episode 1700/5000 | Epsilon: 0.0100 | Success in last 100: 79 %
Episode 1800/5000 | Epsilon: 0.0100 | Success in last 100: 91 %
Episode 1900/5000 | Epsilon: 0.0100 | Success in last 100: 89 %
Episode 2000/5000 | Epsilon: 0.0100 | Success in last 100: 90 %
Episode 2100/5000 | Epsilon: 0.0100 | Success in last 100: 92 %
Episode 2200/5000 | Epsilon: 0.0100 | Success in last 100: 82 %
Episode 2300/5000 | Epsilon: 0.0100 | Success in last 100: 88 %
Episode 2400/5000 | Epsilon: 0.0100 | Success in last 100: 93 %
...
Task complete count: 3812

Episode 5000/5000 | Epsilon: 0.0100 | Success in last 100: 100 %
Double Q-tables saved to deterministic_double_q_table.pkl
```

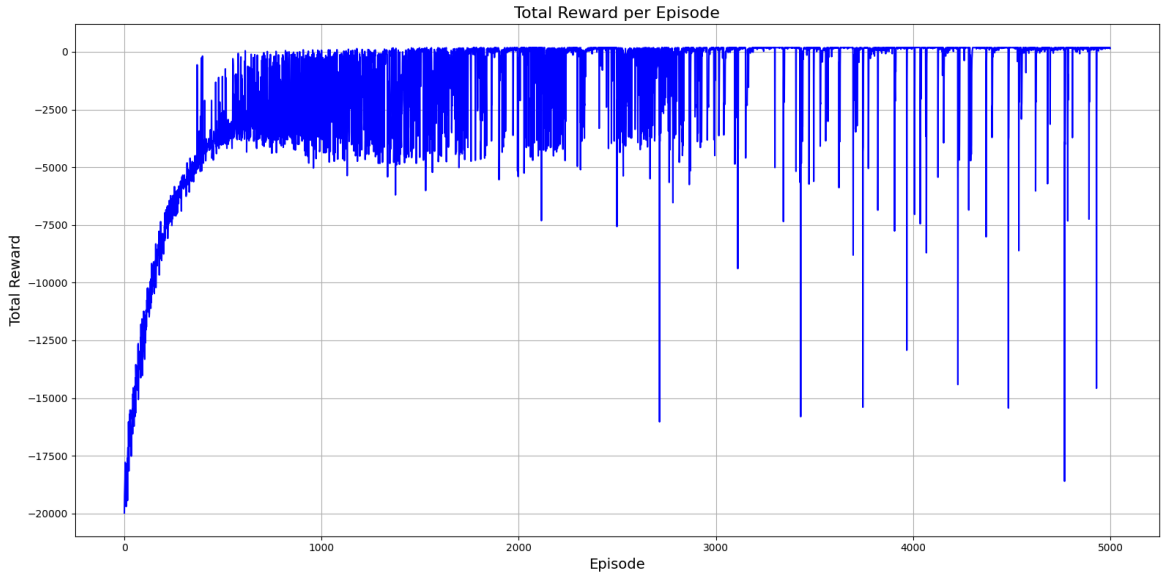
```python
# Evaluating the trained agent using a greedy policy

deterministic = True
env = Environment(0, 0, stochastic=(not deterministic))

q_table_filename = "deterministic_double_q_table.pkl"

evaluation_rewards_deterministic_DoubleQ = evaluate_agent_deterministic_DoubleQ(e
                                                                                q_table_filename=q_tabl
                                                                                episodes=10, max_steps=

# Plotting greedy policy evaluation results
plt.figure(figsize=(16,8))
plt.plot(evaluation_rewards_deterministic_DoubleQ, marker='o', linestyle='-', col
plt.xlabel("Episode", fontsize=14)
plt.ylabel("Total Reward", fontsize=14)
plt.title("Greedy Policy Evaluation (10 Episodes)", fontsize=16)
plt.grid(True)
plt.tight_layout()
plt.show()
```

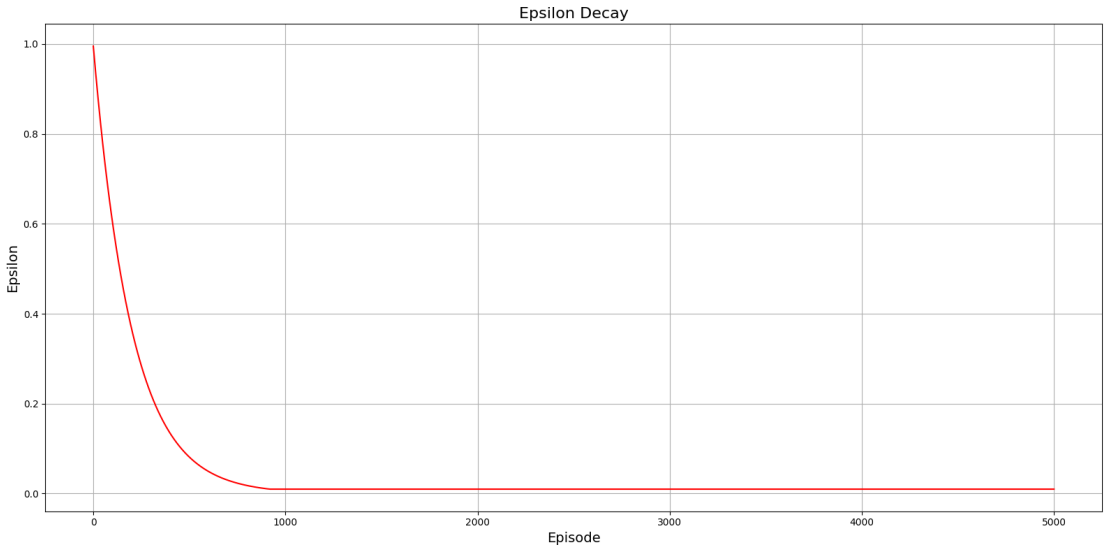
✓ 0.1s

```
Evaluation Episode 1: Steps: 25 | Total Reward: 185
Evaluation Episode 2: Steps: 25 | Total Reward: 185
Evaluation Episode 3: Steps: 25 | Total Reward: 185
Evaluation Episode 4: Steps: 25 | Total Reward: 185
Evaluation Episode 5: Steps: 25 | Total Reward: 185
Evaluation Episode 6: Steps: 25 | Total Reward: 185
Evaluation Episode 7: Steps: 25 | Total Reward: 185
Evaluation Episode 8: Steps: 25 | Total Reward: 185
Evaluation Episode 9: Steps: 25 | Total Reward: 185
Evaluation Episode 10: Steps: 25 | Total Reward: 185
Task complete count: 10
```

I ran the evaluation for 10 episodes and achieved a maximum reward of 185 in 25 steps with a 100% successful task completion rate.





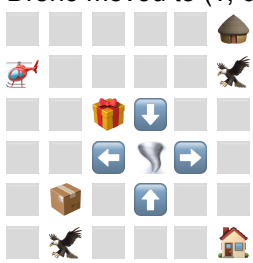


Render each step in greedy approach.

Evaluation Episode 1: Steps: 25 | Total Reward: 185
Task complete count: 1

--- Evaluation Episode 1 starting ---

Drone moved to (1, 0). Step reward: -2



Evaluation Episode 1 - Step 1
Drone moved to (2, 0). Step reward: -2

Drone moved to (2, 1). Step reward: -2

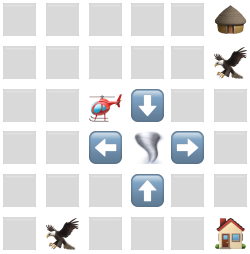
Drone moved to (3, 1). Step reward: -2

Drone moved to (4, 1). Step reward: -2

Picked up package 2 for 25 reward
Drone moved to (3, 1). Step reward: -2

Drone moved to (2, 1). Step reward: -2

Drone moved to (2, 2). Step reward: -2



Picked up package 1 for 25 reward
Drone moved to (1, 2). Step reward: -2

Evaluation Episode 1 - Step 11
Drone moved to (1, 3). Step reward: -2

Drone moved to (0, 3). Step reward: -2

Drone moved to (0, 4). Step reward: -2

Drone moved to (0, 5). Step reward: -2

Dropped package_1 incorrectly. Penalty -50
Delivered package_2 for +100 reward
Picked up package 1 for 25 reward
Drone moved to (0, 4). Step reward: -2

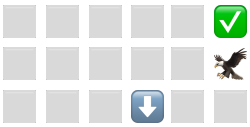
Drone moved to (1, 4). Step reward: -2

Drone moved to (2, 4). Step reward: -2

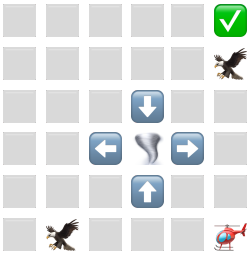
Drone moved to (2, 5). Step reward: -2

Evaluation Episode 1 - Step 21
Drone moved to (3, 5). Step reward: -2

Drone moved to (4, 5). Step reward: -2

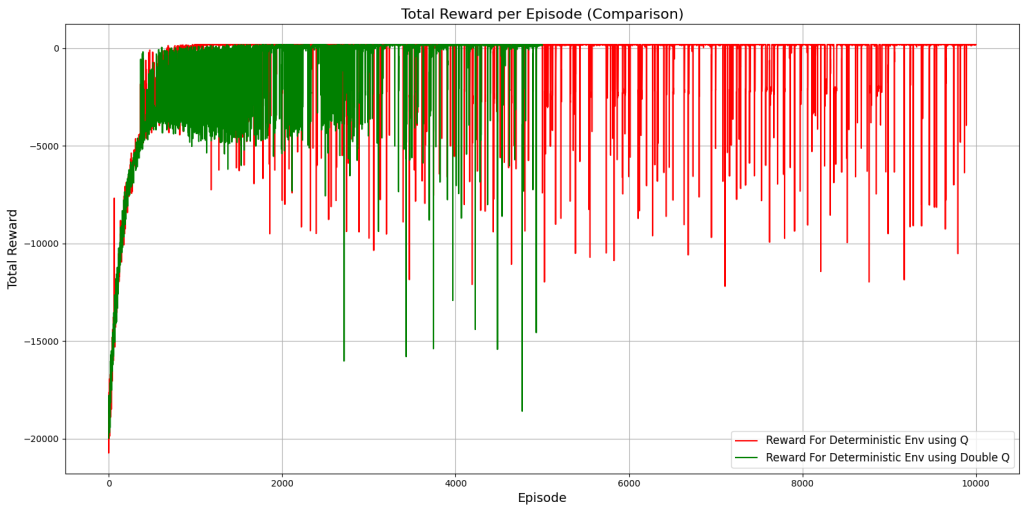


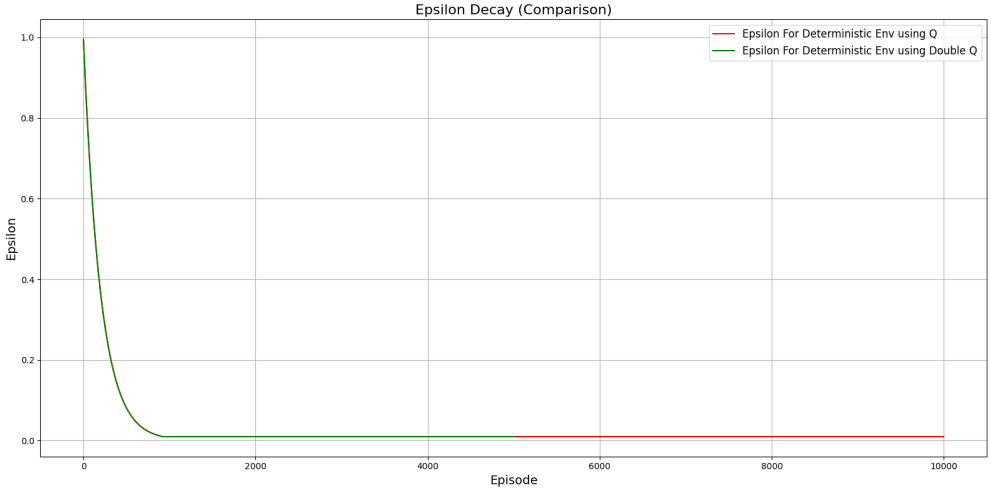
Drone moved to (5, 5). Step reward: -2

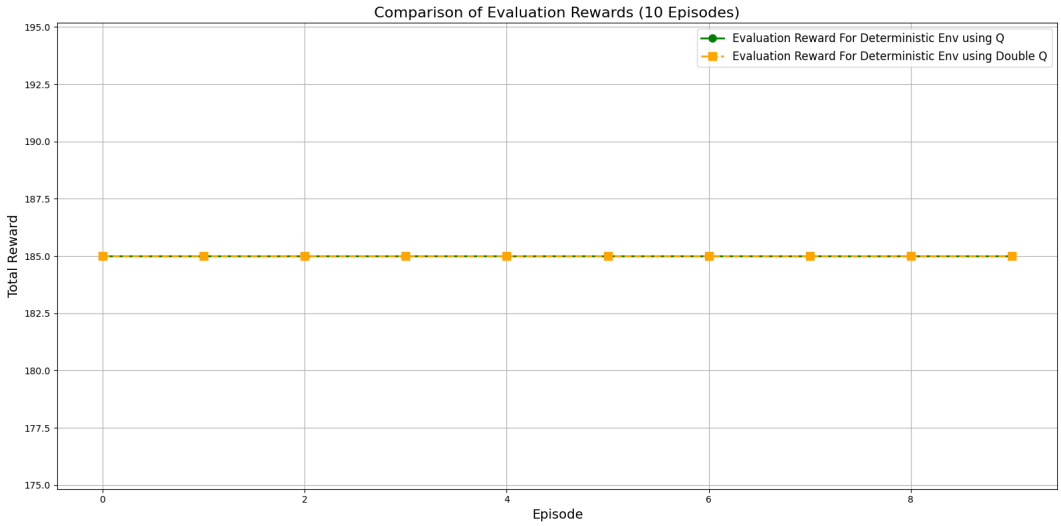


Delivered package_1 for +100 reward
Task complete: All packages delivered 😎

Comparing Q-learning and Double Q-learning in a deterministic environment.







Stochastic Environment:

First, I applied Q-learning using the **seven hyperparameters** listed below, and the Q-table is saved in the following files:

- ssaurav_assignment1_final/part_2_files/hyperparams_1_stochastic_q_table.pkl
- ssaurav_assignment1_final/part_2_files/hyperparams_2_stochastic_q_table.pkl
- ssaurav_assignment1_final/part_2_files/hyperparams_3_stochastic_q_table.pkl
- ssaurav_assignment1_final/part_2_files/hyperparams_4_stochastic_q_table.pkl
- ssaurav_assignment1_final/part_2_files/hyperparams_5_stochastic_q_table.pkl
- ssaurav_assignment1_final/part_2_files/hyperparams_6_stochastic_q_table.pkl
- ssaurav_assignment1_final/part_2_files/hyperparams_base_stochastic_q_table.pkl

Improvement Trend:

Hyperparameters were improved in the following order:

Base → Hyperparams 6 → Hyperparams 1.

Hyperparams_base

```
# Hyperparameter Base

hyperparams_base = {
    'alpha': 0.005,
    'gamma': 0.95,
    'epsilon': 1.0,
    'epsilon_decay': 0.9999,
    'epsilon_min': 0.01,
    'episodes': 10000,
    'max_steps': 1000
}


# Setting environment to stochastic mode
deterministic = False
env = Environment(0, 0, stochastic=(not deterministic))

print("Training Q-Learning agent (Stochastic Environment with Sensor-Augmented State) ...")
Q_stochastic_Q, rewards_stochastic_Q, eps_history_stochastic_Q = train_agent_stochastic_Q(env, hyperparams_base, hyperparams_name="hyperparams_base", render=False)
```

✓ 1m 19.6s

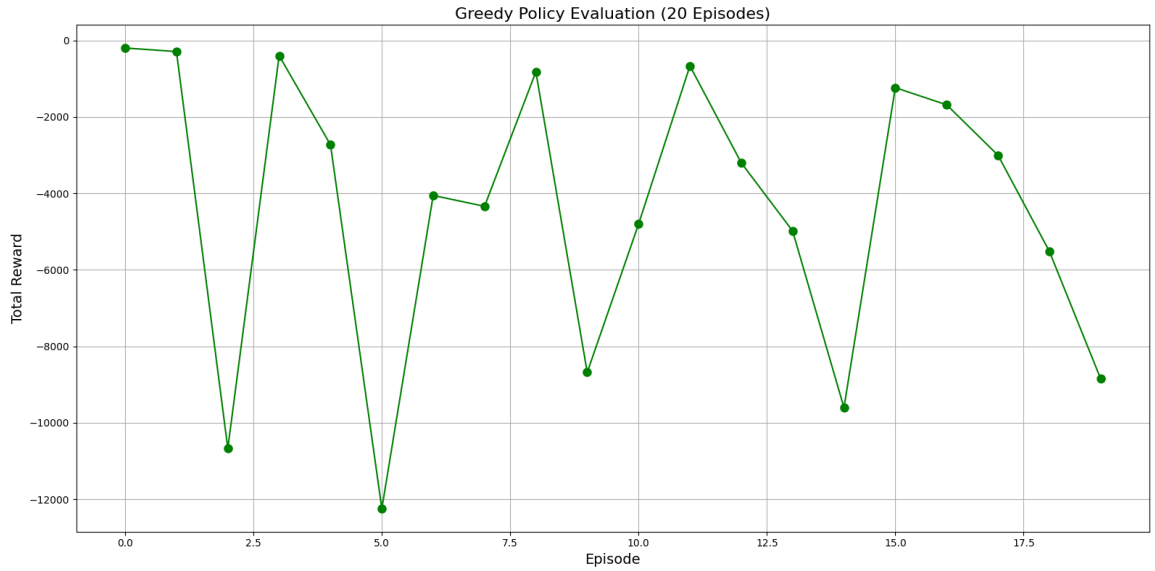
```
Training Q-Learning agent (Stochastic Environment with Sensor-Augmented State) ...
Episode 1000/10000 | Eps: 0.9048 | Success in last 1K: 0.8 %
Episode 2000/10000 | Eps: 0.8187 | Success in last 1K: 1.7 %
Episode 3000/10000 | Eps: 0.7408 | Success in last 1K: 4.2 %
Episode 4000/10000 | Eps: 0.6703 | Success in last 1K: 9.7 %
Episode 5000/10000 | Eps: 0.6065 | Success in last 1K: 14.5 %
Episode 6000/10000 | Eps: 0.5488 | Success in last 1K: 22.4 %
Episode 7000/10000 | Eps: 0.4966 | Success in last 1K: 32.5 %
Episode 8000/10000 | Eps: 0.4493 | Success in last 1K: 45.4 %
Episode 9000/10000 | Eps: 0.4066 | Success in last 1K: 56.0 %
Task complete count: 2519
Episode 10000/10000 | Eps: 0.3679 | Success in last 1K: 64.7 %

Q-table saved to hyperparams_base_stochastic_q_table.pkl
```

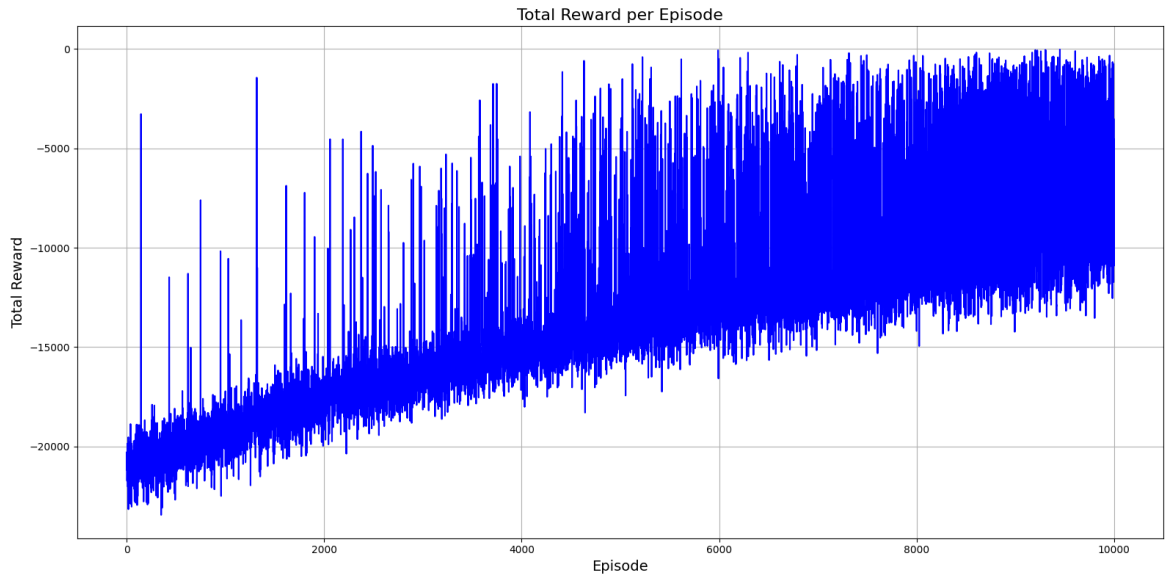
```python
# Evaluating the trained agent
deterministic = False
env = Environment(0, 0, stochastic=(not deterministic))  # Setting up the environment as deterministic

# Loading the trained Q-table
q_table_filename = "hyperparams_base_stochastic_q_table.pkl"

evaluation_rewards_stochastic_Q = evaluate_agent_stochastic_Q(env, q_table_filename=q_table_filename,
                                                              episodes=20, max_steps=1000, render=True)

# Plotting greedy policy evaluation
plt.figure(figsize=(16,8))
plt.plot(evaluation_rewards_stochastic_Q, marker='o', linestyle='-', color='green', markersize=8)
plt.xlabel("Episode", fontsize=14)
plt.ylabel("Total Reward", fontsize=14)
plt.title("Greedy Policy Evaluation (20 Episodes)", fontsize=16)
plt.grid(True)
plt.tight_layout()
plt.show()
```

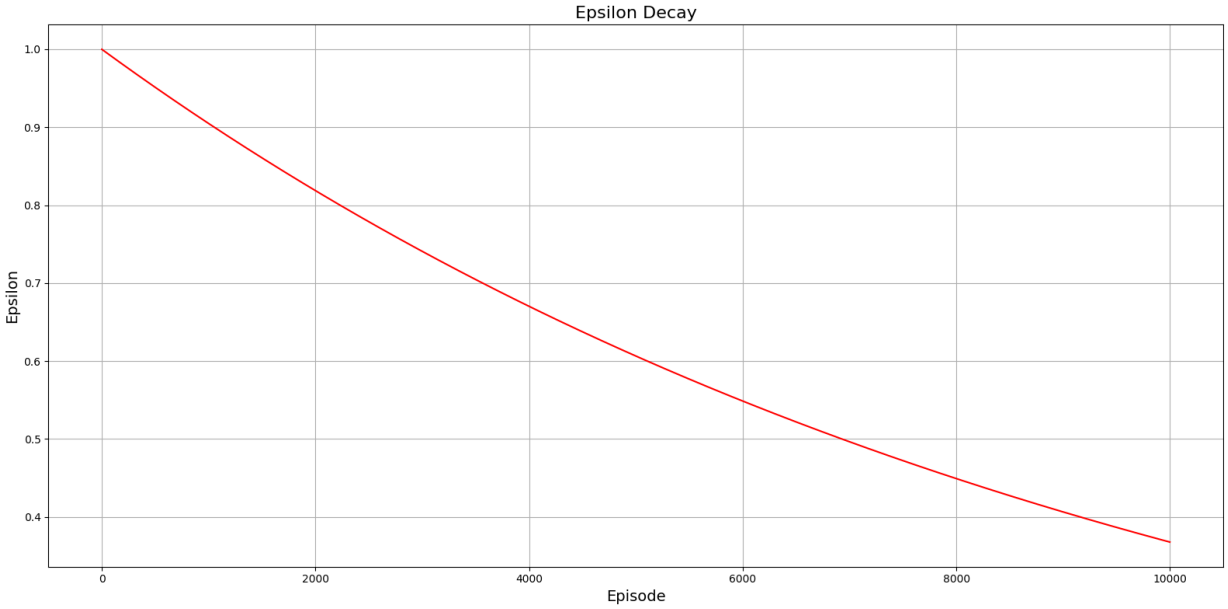
✓ 0.7s

```
Evaluation Episode 1: Steps: 94 | Total Reward: -198
Evaluation Episode 2: Steps: 74 | Total Reward: -292
Evaluation Episode 3: Steps: 1000 | Total Reward: -10667
Evaluation Episode 4: Steps: 218 | Total Reward: -399
Evaluation Episode 5: Steps: 635 | Total Reward: -2730
Evaluation Episode 6: Steps: 1000 | Total Reward: -12243
Evaluation Episode 7: Steps: 1000 | Total Reward: -4054
Evaluation Episode 8: Steps: 1000 | Total Reward: -4338
Evaluation Episode 9: Steps: 400 | Total Reward: -822
Evaluation Episode 10: Steps: 1000 | Total Reward: -8680
Evaluation Episode 11: Steps: 1000 | Total Reward: -4803
Evaluation Episode 12: Steps: 151 | Total Reward: -667
Evaluation Episode 13: Steps: 447 | Total Reward: -3201
Evaluation Episode 14: Steps: 765 | Total Reward: -4983
Evaluation Episode 15: Steps: 1000 | Total Reward: -9601
Evaluation Episode 16: Steps: 221 | Total Reward: -1236
Evaluation Episode 17: Steps: 223 | Total Reward: -1684
Evaluation Episode 18: Steps: 461 | Total Reward: -3003
Evaluation Episode 19: Steps: 1000 | Total Reward: -5514
Evaluation Episode 20: Steps: 1000 | Total Reward: -8851
Task complete count: 11
```

Out of 20, only 11 were successfully delivered.



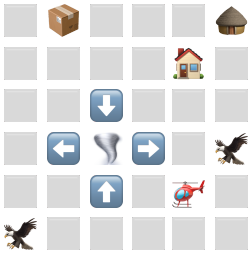




Evaluation Episode 1: Steps: 98 | Total Reward: -114
Task complete count: 1

--- Evaluation Episode 1 starting ---

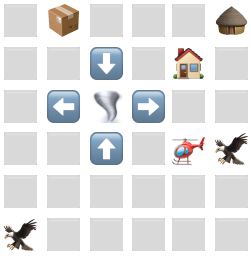
Drone moved to (4, 4). Step reward: -2



Evaluation Episode 1 - Step 1
Picked up package 1 for 25 reward
Drone moved to (3, 4). Step reward: -2

Drone moved to (2, 4). Step reward: -2

State (2, 4, 1, 0, 0, 'destination_1', 0, 0, 0, 0, -1, 0, -1) not found or has default Q-values. Choosing random valid action.
Drone moved to (3, 4). Step reward: -2



Drone moved to (2, 4). Step reward: -2

State (2, 4, 1, 0, 0, 'destination_1', 0, 0, 0, 0, -1, 0, -1) not found or has default Q-values. Choosing random valid action.
Attempted pickup failed. Penalty -25
State (2, 4, 1, 0, 0, 'destination_1', 0, 0, 0, 0, -1, 0, -1) not found or has default Q-values. Choosing random valid action.
Drone moved to (2, 5). Step reward: -2

Drone moved to (2, 4). Step reward: -2

State (2, 4, 1, 0, -1, 'destination_1', 0, 0, 0, 0, 0, 0, -1) not found or has default Q-values. Choosing random valid action.
Drone moved to (1, 4). Step reward: -10

Drone moved to (0, 4). Step reward: -2

Evaluation Episode 1 - Step 11
Drone moved to (0, 3). Step reward: -2

Drone moved to (0, 4). Step reward: -2

Drone moved to (0, 3). Step reward: -2

Drone moved to (1, 3). Step reward: -10

Drone moved to (2, 3). Step reward: -2

Drone moved to (3, 3). Step reward: -2

Action 1 repeated 4 times. Switching to a new action.
Dropped package_1 incorrectly. Penalty -50
Picked up package 1 for 25 reward
Drone moved to (4, 3). Step reward: -2

Drone moved to (3, 3). Step reward: -2

Evaluation Episode 1 - Step 21
Drone moved to (4, 3). Step reward: -2

Drone moved to (3, 3). Step reward: -10

Drone moved to (2, 3). Step reward: -10

Drone moved to (3, 3). Step reward: -10

Drone moved to (3, 2). Step reward: -2

Drone moved to (3, 1). Step reward: -2

Drone moved to (3, 2). Step reward: -10

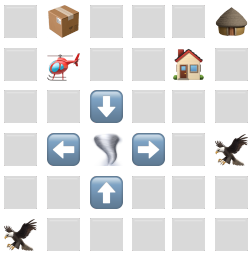
Drone moved to (3, 1). Step reward: -10

Drone moved to (2, 1). Step reward: -10

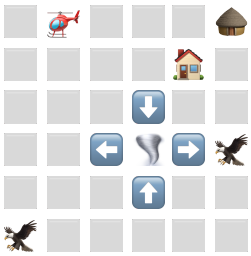
Drone moved to (2, 0). Step reward: -2

Evaluation Episode 1 - Step 31
Drone moved to (2, 1). Step reward: -2

Drone moved to (1, 1). Step reward: -2



Drone moved to (0, 1). Step reward: -2



Picked up package 2 for 25 reward
Drone moved to (1, 1). Step reward: -2

Drone moved to (2, 1). Step reward: -2

Drone moved to (2, 0). Step reward: -2

Drone moved to (2, 1). Step reward: -2

Drone moved to (2, 0). Step reward: -2

Drone moved to (2, 1). Step reward: -2

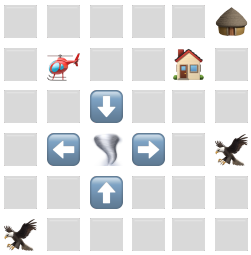
Evaluation Episode 1 - Step 41
Drone moved to (2, 0). Step reward: -2

Drone moved to (2, 1). Step reward: -2

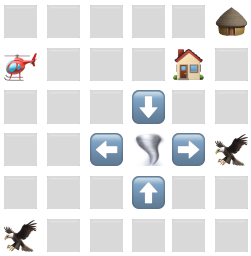
Drone moved to (2, 0). Step reward: -2

Drone moved to (2, 1). Step reward: -2

Drone moved to (1, 1). Step reward: -2



Drone moved to (1, 0). Step reward: -2



Drone moved to (0, 0). Step reward: -2

Drone moved to (0, 1). Step reward: -2

Drone moved to (1, 1). Step reward: -2

Drone moved to (2, 1). Step reward: -2

Evaluation Episode 1 - Step 51
Drone moved to (1, 1). Step reward: -2

Drone moved to (2, 1). Step reward: -2

Drone moved to (2, 0). Step reward: -2

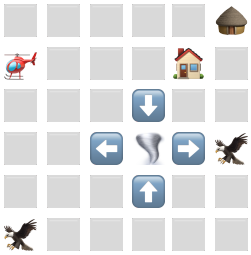
Drone moved to (2, 1). Step reward: -2

Drone moved to (2, 0). Step reward: -2

Drone moved to (2, 1). Step reward: -2

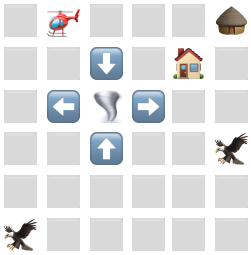
Drone moved to (1, 1). Step reward: -2

Drone moved to (1, 0). Step reward: -2

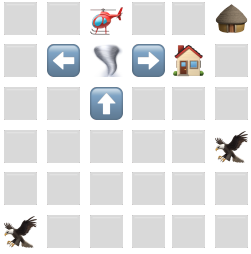


Drone moved to (0, 0). Step reward: -2

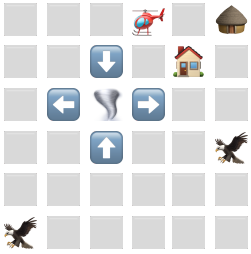
Drone moved to (0, 1). Step reward: -2



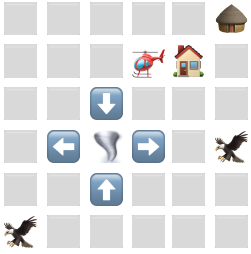
Evaluation Episode 1 - Step 61
Drone moved to (0, 2). Step reward: -10



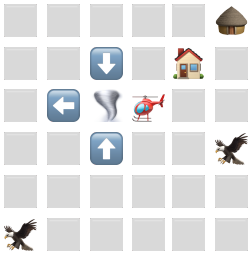
Drone moved to (0, 3). Step reward: -2



Drone moved to (1, 3). Step reward: -2



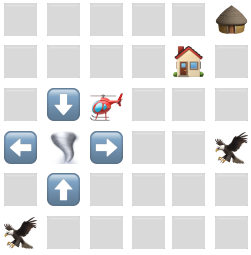
Drone moved to (2, 3). Step reward: -10



State (2, 3, 1, 1, -1, 0, 'destination_1', -1, -1, 0, -1, 0, 0) not found or has default Q-values. Choosing random valid action.
Drone moved to (1, 3). Step reward: -2

Drone moved to (2, 3). Step reward: -2

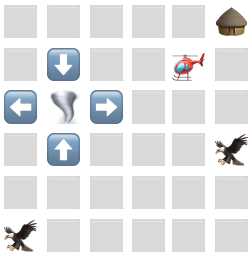
Drone moved to (2, 2). Step reward: -2



Drone moved to (1, 2). Step reward: -2

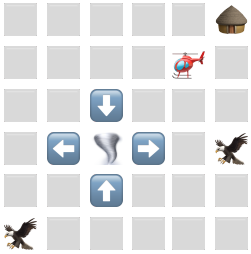
Drone moved to (1, 3). Step reward: -2

Drone moved to (1, 4). Step reward: -2



Evaluation Episode 1 - Step 71
State (1, 4, 1, 1, 0, 0, 'destination_2', 0, 'destination_1', 0, 0, 0, 0) not found or has default Q-values. Choosing random valid action.
Delivered package_1 for +100 reward
Dropped package_2 incorrectly. Penalty -50
Picked up package 2 for 25 reward
Drone moved to (1, 5). Step reward: -2

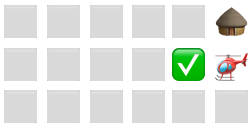
Drone moved to (1, 4). Step reward: -2

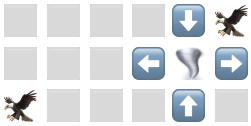


Drone moved to (1, 5). Step reward: -2

Drone moved to (1, 4). Step reward: -2

Drone moved to (1, 5). Step reward: -2





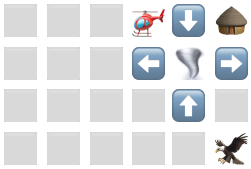
Drone moved to (1, 4). Step reward: -2

State (1, 4, 2, 1, 0, 0, 'destination_2', 0, 'destination_1', 0, 0, -1, 0) not found or has default Q-values. Choosing random valid action.
Drone moved to (1, 5). Step reward: -2

Drone moved to (1, 4). Step reward: -2

Evaluation Episode 1 - Step 81
State (1, 4, 2, 1, 0, 0, 'destination_2', 0, 'destination_1', 0, 0, -1, 0) not found or has default Q-values. Choosing random valid action.
Dropped package_2 incorrectly. Penalty -50
Picked up package 2 for 25 reward
State (1, 4, 2, 1, 0, 0, 'destination_2', 0, 'destination_1', 0, 0, -1, 0) not found or has default Q-values. Choosing random valid action.
Drone moved to (1, 3). Step reward: -2

Drone moved to (0, 3). Step reward: -2



Drone moved to (0, 2). Step reward: -2

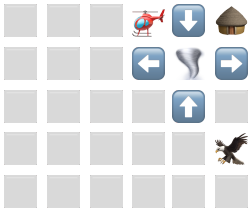
Drone moved to (0, 3). Step reward: -2

Drone moved to (0, 2). Step reward: -2

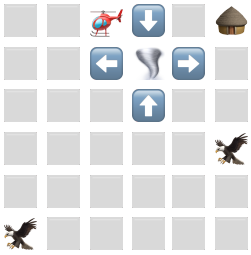
Drone moved to (0, 3). Step reward: -2

Drone moved to (0, 2). Step reward: -2

Drone moved to (0, 3). Step reward: -2



Evaluation Episode 1 - Step 91
Drone moved to (0, 2). Step reward: -2



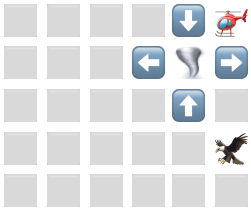
Drone moved to (0, 3). Step reward: -2

Drone moved to (0, 2). Step reward: -2

Drone moved to (0, 3). Step reward: -2

Drone moved to (0, 4). Step reward: -10

Drone moved to (0, 5). Step reward: -2



Delivered package_2 for +100 reward
Task complete: All packages delivered 😎

Hyperparams_6

Changing alpha from 0.005 to 0.001.

```python
# Hyperparameter set 6 changing alpha = 0.001

hyperparams_6 = {
    'alpha': 0.001,
    'gamma': 0.95,
    'epsilon': 1.0,
    'epsilon_decay': 0.9999,
    'epsilon_min': 0.01,
    'episodes': 10000,
    'max_steps': 1000
}


# Setting environment to stochastic mode
deterministic = False
env = Environment(0, 0, stochastic=(not deterministic))

print("Training Q-Learning agent (Stochastic Environment with Sensor-Augmented State) ...")
Q_stochastic_Q, rewards_stochastic_Q, eps_history_stochastic_Q = train_agent_stochastic_Q(env, hyperparams_6, hyperparams_name= 'hyperparams_6' ,render=False)
```

✓ 1m 22.7s

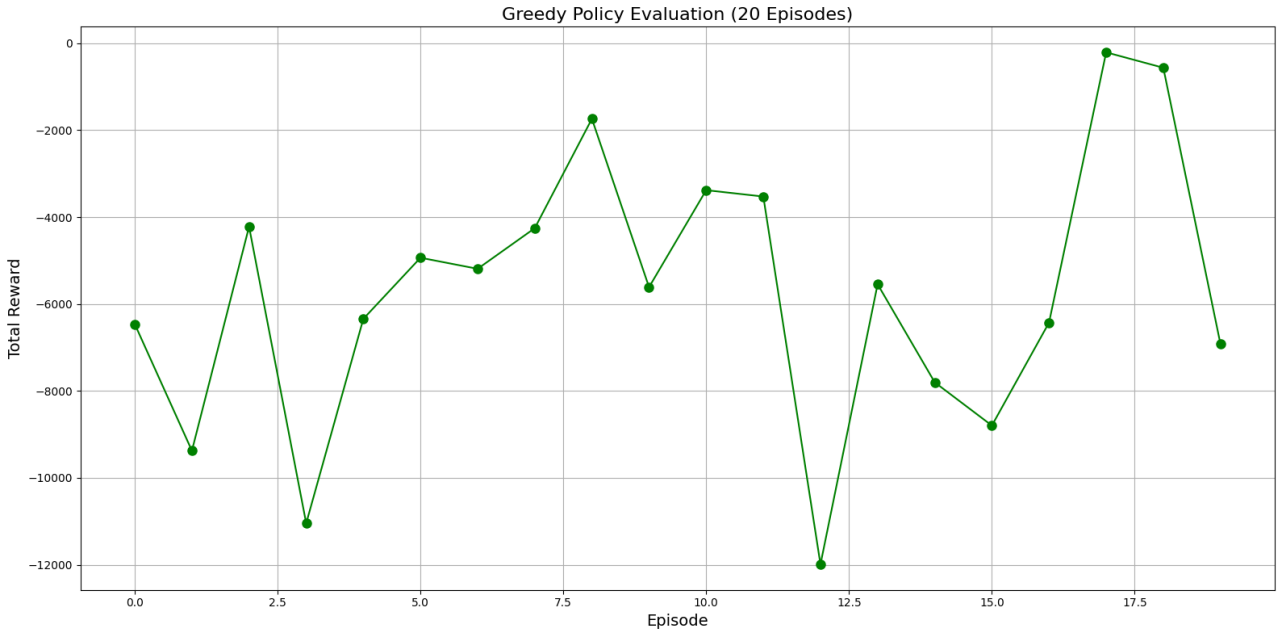
```
Training Q-Learning agent (Stochastic Environment with Sensor-Augmented State) ...
Episode 1000/10000 | Eps: 0.9048 | Success in last 1K: 1.4 %
Episode 2000/10000 | Eps: 0.8187 | Success in last 1K: 1.7 %
Episode 3000/10000 | Eps: 0.7408 | Success in last 1K: 3.6 %
Episode 4000/10000 | Eps: 0.6703 | Success in last 1K: 5.9 %
Episode 5000/10000 | Eps: 0.6065 | Success in last 1K: 9.1 %
Episode 6000/10000 | Eps: 0.5488 | Success in last 1K: 13.1 %
Episode 7000/10000 | Eps: 0.4966 | Success in last 1K: 23.7 %
Episode 8000/10000 | Eps: 0.4493 | Success in last 1K: 29.5 %
Episode 9000/10000 | Eps: 0.4066 | Success in last 1K: 42.9 %
Task complete count: 1782
Episode 10000/10000 | Eps: 0.3679 | Success in last 1K: 47.3 %

Q-table saved to hyperparams_6_stochastic_q_table.pkl
```

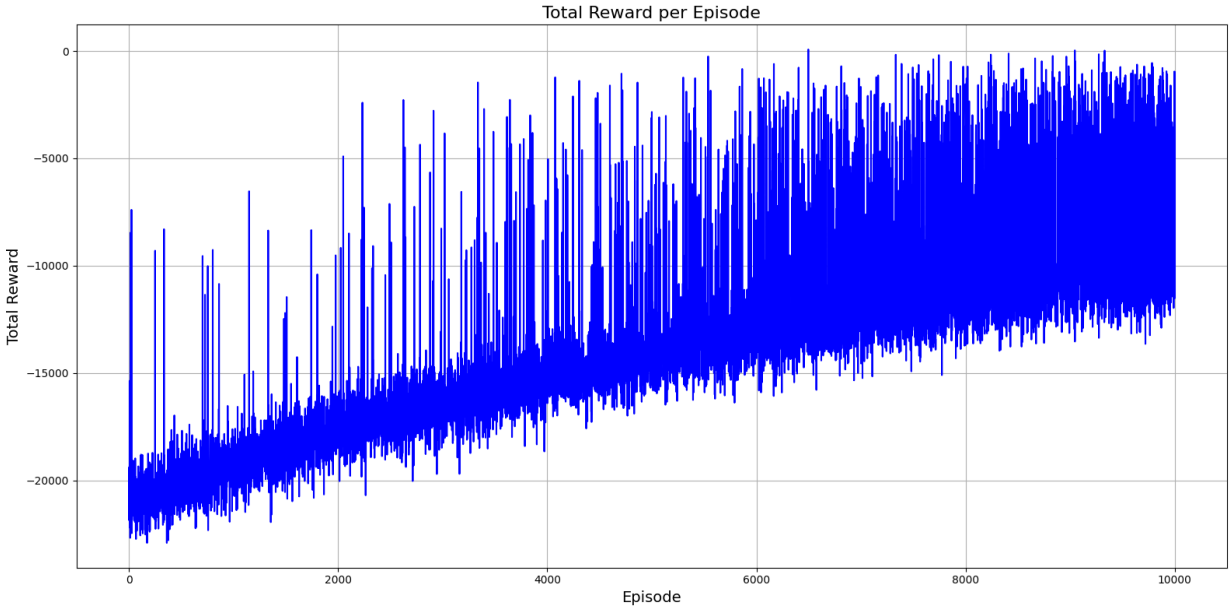
```python
# Evaluating the trained agent

deterministic = False
env = Environment(0, 0, stochastic=(not deterministic))  # Setting up the environment as stochastic

# Loading the trained Q-table
q_table_filename = "hyperparams_6_stochastic_q_table.pkl"

evaluation_rewards_stochastic_Q = evaluate_agent_stochastic_Q(env, q_table_filename= q_table_filename,
                                                              episodes=20, max_steps=1000, render=True)

# Plotting greedy policy evaluation
plt.figure(figsize=(16,8))
plt.plot(evaluation_rewards_stochastic_Q, marker='o', linestyle='-', color='green', markersize=8)
plt.xlabel("Episode", fontsize=14)
plt.ylabel("Total Reward", fontsize=14)
plt.title("Greedy Policy Evaluation (20 Episodes)", fontsize=16)
plt.grid(True)
plt.tight_layout()
plt.show()
```

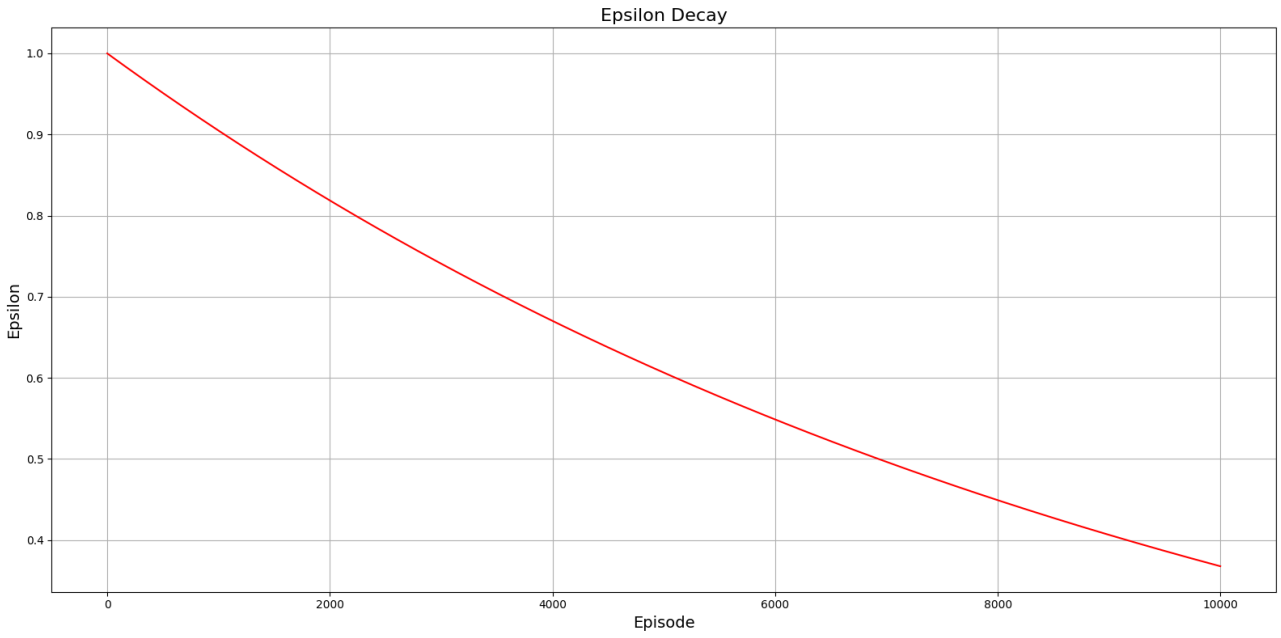
✓ 0.7s

```
Evaluation Episode 1: Steps: 1000 | Total Reward: -6285
Evaluation Episode 2: Steps: 1000 | Total Reward: -10884
Evaluation Episode 3: Steps: 1000 | Total Reward: -4495
Evaluation Episode 4: Steps: 809 | Total Reward: -6333
Evaluation Episode 5: Steps: 1000 | Total Reward: -8207
Evaluation Episode 6: Steps: 1000 | Total Reward: -8958
Evaluation Episode 7: Steps: 1000 | Total Reward: -7293
Evaluation Episode 8: Steps: 771 | Total Reward: -3958
Evaluation Episode 9: Steps: 1000 | Total Reward: -4393
Evaluation Episode 10: Steps: 652 | Total Reward: -3526
Evaluation Episode 11: Steps: 493 | Total Reward: -2934
Evaluation Episode 12: Steps: 1000 | Total Reward: -12397
Evaluation Episode 13: Steps: 418 | Total Reward: -3309
Evaluation Episode 14: Steps: 1000 | Total Reward: -6243
Evaluation Episode 15: Steps: 1000 | Total Reward: -6349
Evaluation Episode 16: Steps: 1000 | Total Reward: -5130
Evaluation Episode 17: Steps: 1000 | Total Reward: -9250
Evaluation Episode 18: Steps: 1000 | Total Reward: -4951
Evaluation Episode 19: Steps: 1000 | Total Reward: -4672
Evaluation Episode 20: Steps: 1000 | Total Reward: -4546
Task complete count: 5
```

Out of 20, only 5 were successfully delivered.







Evaluation Episode 1: Steps: 418 | Total Reward: -3687
Task complete count: 1

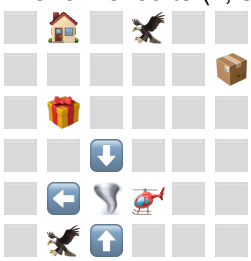
--- Evaluation Episode 1 starting ---

Drone moved to (5, 1). Step reward: -50

Evaluation Episode 1 - Step 1
Drone moved to (5, 2). Step reward: -2

Drone moved to (5, 3). Step reward: -2

Action 3 repeated 4 times. Switching to a new action.
Drone moved to (4, 3). Step reward: -10



Drone moved to (4, 4). Step reward: -2

Drone moved to (4, 3). Step reward: -10

Drone moved to (5, 3). Step reward: -2

Drone moved to (4, 3). Step reward: -10

Drone moved to (4, 4). Step reward: -2

Drone moved to (4, 3). Step reward: -2

Drone moved to (4, 2). Step reward: -2

Evaluation Episode 1 - Step 11
Drone moved to (4, 3). Step reward: -2

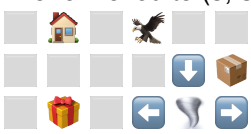
Drone moved to (3, 3). Step reward: -2

Drone moved to (3, 4). Step reward: -2

Drone moved to (3, 3). Step reward: -2

Drone moved to (4, 3). Step reward: -2

Drone moved to (3, 3). Step reward: -2



Drone moved to (3, 2). Step reward: -2

State (3, 2, 0, 0, 'package_1', 0, 0, 0, 'destination_2', -1, 0, 0, 0) not found or has default Q-values. Choosing random valid action.
Drone moved to (3, 3). Step reward: -2

Drone moved to (3, 2). Step reward: -2

Drone moved to (4, 2). Step reward: -2

Evaluation Episode 1 - Step 21
Drone moved to (4, 1). Step reward: -2

Drone moved to (4, 0). Step reward: -2

Drone moved to (4, 1). Step reward: -10

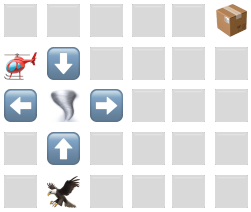
Drone moved to (4, 0). Step reward: -2

Drone moved to (3, 0). Step reward: -2

Drone moved to (2, 0). Step reward: -2

Drone moved to (1, 0). Step reward: -2

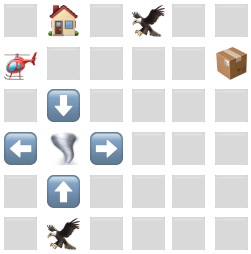
Action 0 repeated 4 times. Switching to a new action.
Drone moved to (2, 0). Step reward: -2



Drone moved to (1, 0). Step reward: -2

Drone moved to (2, 0). Step reward: -2

Evaluation Episode 1 - Step 31
Drone moved to (1, 0). Step reward: -2



Drone moved to (1, 1). Step reward: -2

Drone moved to (0, 1). Step reward: -2

Drone moved to (1, 1). Step reward: -2

Drone moved to (1, 0). Step reward: -2

Drone moved to (2, 0). Step reward: -2

Drone moved to (2, 1). Step reward: -2

Picked up package 1 for 25 reward
Drone moved to (2, 0). Step reward: -2

Drone moved to (3, 0). Step reward: -2

Evaluation Episode 1 - Step 41
Drone moved to (2, 0). Step reward: -2

Drone moved to (3, 0). Step reward: -2

Drone moved to (2, 0). Step reward: -2

Drone moved to (3, 0). Step reward: -2

Drone moved to (2, 0). Step reward: -2

Drone moved to (3, 0). Step reward: -2

Drone moved to (2, 0). Step reward: -2

Drone moved to (3, 0). Step reward: -2

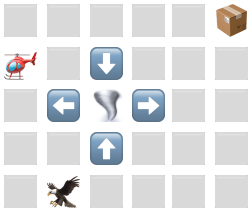
Drone moved to (2, 0). Step reward: -2

Drone moved to (2, 1). Step reward: -2

Evaluation Episode 1 - Step 51
Drone moved to (2, 0). Step reward: -2

Drone moved to (3, 0). Step reward: -2

Drone moved to (2, 0). Step reward: -2



Drone moved to (3, 0). Step reward: -2

Drone moved to (2, 0). Step reward: -2

Drone moved to (2, 1). Step reward: -2

Drone moved to (2, 2). Step reward: -10

Drone moved to (2, 1). Step reward: -2

Drone moved to (2, 2). Step reward: -10

Drone moved to (2, 1). Step reward: -10

Evaluation Episode 1 - Step 61
Drone moved to (3, 1). Step reward: -2

Drone moved to (3, 0). Step reward: -2

Drone moved to (2, 0). Step reward: -2

Drone moved to (2, 1). Step reward: -2

Drone moved to (2, 2). Step reward: -2

Drone moved to (3, 2). Step reward: -2

Drone moved to (3, 1). Step reward: -2

Drone moved to (3, 2). Step reward: -2

Drone moved to (3, 1). Step reward: -2

Drone moved to (4, 1). Step reward: -2

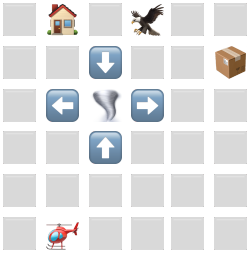
Evaluation Episode 1 - Step 71
Drone moved to (5, 1). Step reward: -50

Drone moved to (4, 1). Step reward: -2

Drone moved to (5, 1). Step reward: -50

Drone moved to (4, 1). Step reward: -2

Drone moved to (5, 1). Step reward: -50



Drone moved to (4, 1). Step reward: -2

Drone moved to (5, 1). Step reward: -50

Drone moved to (4, 1). Step reward: -2

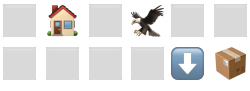
Drone moved to (5, 1). Step reward: -50

Drone moved to (4, 1). Step reward: -2

Evaluation Episode 1 - Step 81
Dropped package_1 incorrectly. Penalty -50
Drone moved to (5, 1). Step reward: -50

Drone moved to (5, 2). Step reward: -2

Drone moved to (4, 2). Step reward: -2



Drone moved to (4, 3). Step reward: -2

Drone moved to (3, 3). Step reward: -2

Drone moved to (3, 2). Step reward: -2

Drone moved to (3, 3). Step reward: -2

Drone moved to (3, 2). Step reward: -2

Drone moved to (4, 2). Step reward: -2

Evaluation Episode 1 - Step 91
Drone moved to (5, 2). Step reward: -2

Drone moved to (4, 2). Step reward: -2

Drone moved to (5, 2). Step reward: -2

Drone moved to (4, 2). Step reward: -10

State (4, 2, 0, 0, 0, 'destination_2', -1, 'package_1', -1, -1, -1, 0, -1) not found or has default Q-values. Choosing random valid action.
Drone moved to (3, 2). Step reward: -2

Drone moved to (4, 2). Step reward: -10

State (4, 2, 0, 0, 0, 'destination_2', -1, 'package_1', -1, -1, -1, 0, -1) not found or has default Q-values. Choosing random valid action.
Drone moved to (4, 3). Step reward: -10

Drone moved to (5, 3). Step reward: -2

Drone moved to (4, 3). Step reward: -2

Drone moved to (4, 2). Step reward: -2

Evaluation Episode 1 - Step 101
Drone moved to (5, 2). Step reward: -2

Drone moved to (4, 2). Step reward: -2

Drone moved to (5, 2). Step reward: -2

Drone moved to (5, 3). Step reward: -2

Drone moved to (4, 3). Step reward: -2

Drone moved to (3, 3). Step reward: -2

Drone moved to (3, 4). Step reward: -2

Drone moved to (3, 3). Step reward: -2

Drone moved to (3, 4). Step reward: -2

Drone moved to (3, 3). Step reward: -2

Evaluation Episode 1 - Step 111
Drone moved to (3, 4). Step reward: -2

Drone moved to (3, 3). Step reward: -2

Drone moved to (3, 4). Step reward: -2

Drone moved to (3, 3). Step reward: -2

Drone moved to (4, 3). Step reward: -2

Drone moved to (3, 3). Step reward: -2

Drone moved to (4, 3). Step reward: -2

Drone moved to (3, 3). Step reward: -2

Drone moved to (3, 2). Step reward: -2

Drone moved to (4, 2). Step reward: -2

Evaluation Episode 1 - Step 121
Drone moved to (5, 2). Step reward: -2

Drone moved to (4, 2). Step reward: -2

Drone moved to (4, 3). Step reward: -2

Drone moved to (3, 3). Step reward: -2

Drone moved to (4, 3). Step reward: -2

Drone moved to (3, 3). Step reward: -10

Drone moved to (4, 3). Step reward: -2

Drone moved to (3, 3). Step reward: -2

Drone moved to (4, 3). Step reward: -2

Drone moved to (3, 3). Step reward: -2

Evaluation Episode 1 - Step 131
Drone moved to (4, 3). Step reward: -2

Drone moved to (3, 3). Step reward: -2

Drone moved to (3, 4). Step reward: -2

Drone moved to (3, 3). Step reward: -2

Drone moved to (4, 3). Step reward: -2

Drone moved to (3, 3). Step reward: -2

Drone moved to (3, 4). Step reward: -2

Drone moved to (3, 3). Step reward: -2

Drone moved to (3, 4). Step reward: -2

Drone moved to (3, 3). Step reward: -10

Evaluation Episode 1 - Step 141
Drone moved to (4, 3). Step reward: -10

Drone moved to (5, 3). Step reward: -2

Drone moved to (5, 2). Step reward: -2

Drone moved to (4, 2). Step reward: -2

Drone moved to (5, 2). Step reward: -2

Drone moved to (4, 2). Step reward: -2

State (4, 2, 0, 0, 0, 'destination_2', 0, 'package_1', 0, -1, -1, 0, 0) not found or has default Q-values. Choosing random valid action.
Attempted pickup failed. Penalty -25
State (4, 2, 0, 0, 0, 'destination_2', 0, 'package_1', 0, -1, -1, 0, 0) not found or has default Q-values. Choosing random valid action.
Drone moved to (4, 3). Step reward: -2

Drone moved to (5, 3). Step reward: -2

Drone moved to (5, 4). Step reward: -2

Evaluation Episode 1 - Step 151
Drone moved to (5, 3). Step reward: -2

Drone moved to (5, 2). Step reward: -2

Drone moved to (4, 2). Step reward: -2

Drone moved to (4, 1). Step reward: -2

State (4, 1, 0, 0, 0, 0, 'destination_2', 0, 'package_1', -1, 0, -1, 0) not found or has default Q-values. Choosing random valid action.
Drone moved to (4, 0). Step reward: -2

Drone moved to (4, 1). Step reward: -2

Drone moved to (5, 1). Step reward: -50

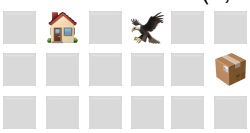
Drone moved to (5, 2). Step reward: -2

Drone moved to (4, 2). Step reward: -10

State (4, 2, 0, 0, 0, 'destination_2', -1, 'package_1', -1, -1, -1, 0, -1) not found or has default Q-values. Choosing random valid action.
Drone moved to (3, 2). Step reward: -10

Evaluation Episode 1 - Step 161
Drone moved to (3, 1). Step reward: -2

Drone moved to (4, 1). Step reward: -2

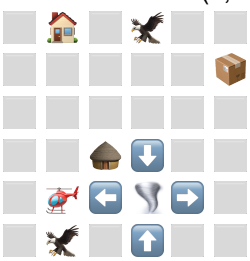


Drone moved to (5, 1). Step reward: -50

Drone moved to (5, 2). Step reward: -2

Drone moved to (4, 2). Step reward: -2

Drone moved to (4, 1). Step reward: -2



State (4, 1, 0, 0, 0, 0, 'destination_2', 0, 'package_1', -1, 0, -1, 0) not found or has default Q-values. Choosing random valid action.
Drone moved to (4, 2). Step reward: -100

Drone moved to (5, 2). Step reward: -2

Drone moved to (4, 2). Step reward: -10

State (4, 2, 0, 0, 0, 'destination_2', -1, 'package_1', -1, -1, -1, 0, -1) not found or has default Q-values. Choosing random valid action.
Drone moved to (3, 2). Step reward: -10

Evaluation Episode 1 - Step 171
Drone moved to (2, 2). Step reward: -2

Drone moved to (2, 3). Step reward: -2

Drone moved to (1, 3). Step reward: -2

Drone moved to (1, 4). Step reward: -2

Drone moved to (1, 5). Step reward: -2

Picked up package 2 for 25 reward
Drone moved to (2, 5). Step reward: -2

Drone moved to (2, 4). Step reward: -2

Drone moved to (2, 3). Step reward: -2

Drone moved to (2, 2). Step reward: -2

Evaluation Episode 1 - Step 181

Action 2 repeated 4 times. Switching to a new action.
Drone moved to (3, 2). Step reward: -2

State (3, 2, 0, 1, 0, 0, 0, 0, 'destination_2', -1, 'package_1', -1, -1) not found or has default Q-values. Choosing random valid action.
Drone moved to (3, 1). Step reward: -2

Drone moved to (4, 1). Step reward: -2

State (4, 1, 0, 1, 0, 0, 'destination_2', 0, 'package_1', 0, 0, -1, 0) not found or has default Q-values. Choosing random valid action.
Drone moved to (4, 2). Step reward: -2

State (4, 2, 0, 1, 0, 'destination_2', 0, 'package_1', 0, 0, -1, 0, 0) not found or has default Q-values. Choosing random valid action.
Drone moved to (4, 1). Step reward: -2

State (4, 1, 0, 1, 0, 0, 'destination_2', 0, 'package_1', 0, 0, -1, 0) not found or has default Q-values. Choosing random valid action.
Drone moved to (4, 0). Step reward: -2

Drone moved to (4, 1). Step reward: -2

State (4, 1, 0, 1, 0, 0, 'destination_2', 0, 'package_1', 0, 0, -1, 0) not found or has default Q-values. Choosing random valid action.
Drone moved to (4, 2). Step reward: -2

State (4, 2, 0, 1, 0, 'destination_2', 0, 'package_1', 0, 0, -1, 0, 0) not found or has default Q-values. Choosing random valid action.
Drone moved to (4, 3). Step reward: -2

Drone moved to (4, 2). Step reward: -2

Evaluation Episode 1 - Step 191
State (4, 2, 0, 1, 0, 'destination_2', -1, 'package_1', 0, 0, -1, 0, 0) not found or has default Q-values. Choosing random valid action.
Dropped package_2 incorrectly. Penalty -50
State (4, 2, 0, 0, 0, 'destination_2', -1, 'package_1', 'package_2', 0, -1, 0, 0) not found or has default Q-values. Choosing random valid action.
Attempted dropoff failed (no package carried). Penalty -50
State (4, 2, 0, 0, 0, 'destination_2', -1, 'package_1', 'package_2', 0, -1, 0, 0) not found or has default Q-values. Choosing random valid action.

Drone moved to (5, 2). Step reward: -2

Drone moved to (5, 3). Step reward: -2

Drone moved to (5, 2). Step reward: -2

Drone moved to (5, 3). Step reward: -2

Drone moved to (5, 2). Step reward: -2

Drone moved to (5, 3). Step reward: -2

Drone moved to (5, 2). Step reward: -2

Drone moved to (5, 3). Step reward: -2

Evaluation Episode 1 - Step 201
Drone moved to (5, 2). Step reward: -2

Drone moved to (5, 3). Step reward: -2

Drone moved to (5, 2). Step reward: -2

Attempted dropoff failed (no package carried). Penalty -50
Attempted dropoff failed (no package carried). Penalty -50
Attempted dropoff failed (no package carried). Penalty -50
Action 5 repeated 4 times. Switching to a new action.
Attempted dropoff failed (no package carried). Penalty -50
Attempted dropoff failed (no package carried). Penalty -50
Attempted dropoff failed (no package carried). Penalty -50
Action 5 repeated 4 times. Switching to a new action.
Move attempted out of grid boundaries. Penalty -25
Evaluation Episode 1 - Step 211
Attempted dropoff failed (no package carried). Penalty -50

Attempted dropoff failed (no package carried). Penalty -50
Action 5 repeated 4 times. Switching to a new action.
Attempted dropoff failed (no package carried). Penalty -50
Attempted dropoff failed (no package carried). Penalty -50
Attempted dropoff failed (no package carried). Penalty -50
Action 5 repeated 4 times. Switching to a new action.
Move attempted out of grid boundaries. Penalty -25
Attempted dropoff failed (no package carried). Penalty -50
Attempted dropoff failed (no package carried). Penalty -50
Action 5 repeated 4 times. Switching to a new action.
Attempted dropoff failed (no package carried). Penalty -50
Attempted dropoff failed (no package carried). Penalty -50
Evaluation Episode 1 - Step 221
Attempted dropoff failed (no package carried). Penalty -50
Action 5 repeated 4 times. Switching to a new action.
Drone moved to (4, 2). Step reward: -2

Drone moved to (3, 2). Step reward: -2

State (3, 2, 0, 0, -1, 0, 0, 0, 'destination_2', 0, 'package_1', 'package_2', 0) not found or has default Q-values.
Choosing random valid action.
Attempted dropoff failed (no package carried). Penalty -50
State (3, 2, 0, 0, -1, 0, 0, 0, 'destination_2', 0, 'package_1', 'package_2', 0) not found or has default Q-values.
Choosing random valid action.
Drone moved to (2, 2). Step reward: -10

Drone moved to (2, 1). Step reward: -10

Drone moved to (1, 1). Step reward: -2

Drone moved to (0, 1). Step reward: -2

Drone moved to (1, 1). Step reward: -10

Drone moved to (1, 2). Step reward: -10

Evaluation Episode 1 - Step 231
Drone moved to (1, 3). Step reward: -10

Drone moved to (1, 4). Step reward: -2

Drone moved to (0, 4). Step reward: -2

Drone moved to (1, 4). Step reward: -2

Drone moved to (1, 3). Step reward: -2

Drone moved to (1, 4). Step reward: -2

Drone moved to (0, 4). Step reward: -2

Drone moved to (1, 4). Step reward: -2

Drone moved to (1, 3). Step reward: -2

Drone moved to (1, 4). Step reward: -2

Evaluation Episode 1 - Step 241
Drone moved to (1, 3). Step reward: -10

Drone moved to (1, 4). Step reward: -10

Drone moved to (1, 5). Step reward: -2

Drone moved to (1, 4). Step reward: -2

Drone moved to (0, 4). Step reward: -2

Drone moved to (0, 5). Step reward: -2

Drone moved to (1, 5). Step reward: -2

Drone moved to (1, 4). Step reward: -2

Drone moved to (1, 3). Step reward: -10

Drone moved to (1, 4). Step reward: -10

Evaluation Episode 1 - Step 251
Drone moved to (1, 5). Step reward: -2

Drone moved to (1, 4). Step reward: -10

Drone moved to (1, 5). Step reward: -2

Drone moved to (2, 5). Step reward: -2

Drone moved to (2, 4). Step reward: -10

Drone moved to (2, 3). Step reward: -10

Drone moved to (3, 3). Step reward: -2

Drone moved to (2, 3). Step reward: -100

Attempted pickup failed. Penalty -25
Attempted pickup failed. Penalty -25
Evaluation Episode 1 - Step 261
Attempted pickup failed. Penalty -25
Action 4 repeated 4 times. Switching to a new action.
Drone moved to (1, 3). Step reward: -2

Drone moved to (1, 2). Step reward: -2

Attempted pickup failed. Penalty -25
Attempted pickup failed. Penalty -25
Attempted pickup failed. Penalty -25
Action 4 repeated 4 times. Switching to a new action.
Drone moved to (1, 3). Step reward: -10

Drone moved to (2, 3). Step reward: -10

Drone moved to (3, 3). Step reward: -2

Drone moved to (4, 3). Step reward: -2

Evaluation Episode 1 - Step 271
Action 1 repeated 4 times. Switching to a new action.
Drone moved to (4, 2). Step reward: -2

State (4, 2, 0, 0, 0, 'destination_2', 0, 'package_1', 'package_2', 0, -1, 0, 0) not found or has default Q-values. Choosing random valid action.
Picked up package 2 for 25 reward
State (4, 2, 0, 1, 0, 'destination_2', 0, 'package_1', 0, 0, -1, 0, 0) not found or has default Q-values. Choosing random valid action.
Drone moved to (4, 3). Step reward: -2

Drone moved to (3, 3). Step reward: -2

Drone moved to (3, 4). Step reward: -2

Drone moved to (3, 5). Step reward: -2

Drone moved to (3, 4). Step reward: -2

Drone moved to (3, 5). Step reward: -2

Drone moved to (3, 4). Step reward: -2

Drone moved to (3, 5). Step reward: -2

Evaluation Episode 1 - Step 281
Drone moved to (3, 4). Step reward: -2

Drone moved to (4, 4). Step reward: -2

Drone moved to (4, 3). Step reward: -2

Drone moved to (4, 2). Step reward: -2

State (4, 2, 0, 1, 0, 'destination_2', 0, 'package_1', 0, 0, -1, 0, 0) not found or has default Q-values. Choosing random valid action.
Drone moved to (3, 2). Step reward: -2

State (3, 2, 0, 1, 0, 0, -1, 0, 'destination_2', 0, 'package_1', 0, 0) not found or has default Q-values. Choosing random valid action.
Drone moved to (2, 2). Step reward: -2

Attempted pickup failed. Penalty -25
Attempted pickup failed. Penalty -25
Attempted pickup failed. Penalty -25
Action 4 repeated 4 times. Switching to a new action.
Drone moved to (2, 1). Step reward: -2

Evaluation Episode 1 - Step 291
Drone moved to (3, 1). Step reward: -2

Drone moved to (4, 1). Step reward: -2

State (4, 1, 0, 1, 0, -1, 'destination_2', 0, 'package_1', 0, 0, -1, 0) not found or has default Q-values. Choosing random valid action.
Drone moved to (4, 2). Step reward: -2

State (4, 2, 0, 1, 0, 'destination_2', 0, 'package_1', 0, 0, -1, 0, 0) not found or has default Q-values. Choosing random valid action.
Drone moved to (3, 2). Step reward: -2

State (3, 2, 0, 1, 0, -1, 0, 0, 'destination_2', 0, 'package_1', 0, 0) not found or has default Q-values. Choosing random valid action.
Attempted pickup failed. Penalty -25
State (3, 2, 0, 1, 0, -1, 0, 0, 'destination_2', 0, 'package_1', 0, 0) not found or has default Q-values. Choosing random valid action.
Drone moved to (3, 1). Step reward: -2

Drone moved to (4, 1). Step reward: -2

State (4, 1, 0, 1, 0, 0, 'destination_2', 0, 'package_1', 0, 0, -1, 0) not found or has default Q-values. Choosing random valid action.
Drone moved to (4, 0). Step reward: -2

Drone moved to (4, 1). Step reward: -2

State (4, 1, 0, 1, 0, 0, -1, 0, 'package_1', 0, 0, -1, 0) not found or has default Q-values. Choosing random valid action.
Dropped package_2 incorrectly. Penalty -50
Evaluation Episode 1 - Step 301
Picked up package 1 for 25 reward
Picked up package 2 for 25 reward
Drone moved to (4, 2). Step reward: -10

Drone moved to (4, 1). Step reward: -10

Attempted pickup failed. Penalty -25
Attempted pickup failed. Penalty -25
Attempted pickup failed. Penalty -25
Action 4 repeated 4 times. Switching to a new action.
Drone moved to (3, 1). Step reward: -2

Drone moved to (2, 1). Step reward: -2

Drone moved to (2, 2). Step reward: -2

Evaluation Episode 1 - Step 311
Drone moved to (2, 3). Step reward: -2

Drone moved to (2, 2). Step reward: -2

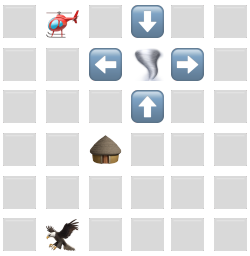
Drone moved to (2, 1). Step reward: -2

Drone moved to (1, 1). Step reward: -2

Drone moved to (0, 1). Step reward: -2

Delivered package_1 for +100 reward
Dropped package_2 incorrectly. Penalty -50
Picked up package 2 for 25 reward
Drone moved to (0, 2). Step reward: -2

Drone moved to (0, 1). Step reward: -2



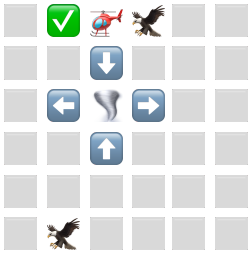
Drone moved to (0, 0). Step reward: -2

Evaluation Episode 1 - Step 321
Drone moved to (0, 1). Step reward: -2

Drone moved to (0, 2). Step reward: -2

Attempted pickup failed. Penalty -25
Attempted pickup failed. Penalty -25
Attempted pickup failed. Penalty -25
Action 4 repeated 4 times. Switching to a new action.
Move attempted out of grid boundaries. Penalty -25
Attempted pickup failed. Penalty -25
Attempted pickup failed. Penalty -25
Action 4 repeated 4 times. Switching to a new action.
Drone moved to (1, 2). Step reward: -2

Drone moved to (0, 2). Step reward: -2



Evaluation Episode 1 - Step 331

State (0, 2, 2, 1, -1, -1, -1, 'destination_1', 0, -1, 0, -1, 0) not found or has default Q-values. Choosing random valid action.
Dropped package_2 incorrectly. Penalty -50
State (0, 2, 2, 0, -1, -1, -1, 'destination_1', 'package_2', -1, 0, -1, 0) not found or has default Q-values. Choosing random valid action.
Drone moved to (1, 2). Step reward: -2

State (1, 2, 2, 0, 'destination_1', 'package_2', -1, -1, 0, 0, -1, -1, 0) not found or has default Q-values. Choosing random valid action.
Attempted dropoff failed (no package carried). Penalty -50
State (1, 2, 2, 0, 'destination_1', 'package_2', -1, -1, 0, 0, -1, -1, 0) not found or has default Q-values. Choosing random valid action.
Attempted pickup failed. Penalty -25
State (1, 2, 2, 0, 'destination_1', 'package_2', -1, -1, 0, 0, -1, -1, 0) not found or has default Q-values. Choosing random valid action.
Drone moved to (1, 1). Step reward: -100

Drone moved to (0, 1). Step reward: -10

Drone moved to (0, 2). Step reward: -2

Picked up package 2 for 25 reward
Move attempted out of grid boundaries. Penalty -25
Move attempted out of grid boundaries. Penalty -25
Evaluation Episode 1 - Step 341
Move attempted out of grid boundaries. Penalty -25
Action 0 repeated 4 times. Switching to a new action.
Attempted pickup failed. Penalty -25
Move attempted out of grid boundaries. Penalty -25

Move attempted out of grid boundaries. Penalty -25
Action 0 repeated 4 times. Switching to a new action.
Drone moved to (0, 1). Step reward: -2

Drone moved to (0, 2). Step reward: -2

Move attempted out of grid boundaries. Penalty -25
Move attempted out of grid boundaries. Penalty -25
Move attempted out of grid boundaries. Penalty -25
Action 0 repeated 4 times. Switching to a new action.
Move attempted out of grid boundaries. Penalty -25
Evaluation Episode 1 - Step 351
Move attempted out of grid boundaries. Penalty -25
Move attempted out of grid boundaries. Penalty -25
Action 0 repeated 4 times. Switching to a new action.
Drone moved to (1, 2). Step reward: -10

Drone moved to (1, 3). Step reward: -2

Drone moved to (1, 4). Step reward: -2

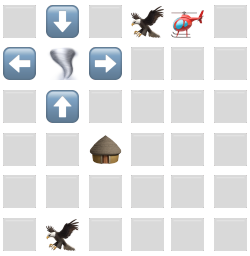
Drone moved to (1, 3). Step reward: -10

Drone moved to (1, 4). Step reward: -10

Drone moved to (1, 5). Step reward: -2

Drone moved to (0, 5). Step reward: -2

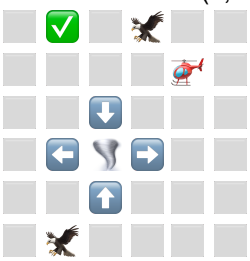
Drone moved to (0, 4). Step reward: -2



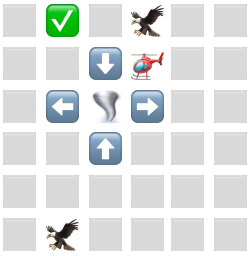
Evaluation Episode 1 - Step 361
Drone moved to (1, 4). Step reward: -2

Drone moved to (1, 3). Step reward: -2

Drone moved to (1, 4). Step reward: -2



Drone moved to (1, 3). Step reward: -2



Drone moved to (1, 4). Step reward: -2

Drone moved to (1, 3). Step reward: -2

Drone moved to (1, 4). Step reward: -2

Drone moved to (1, 3). Step reward: -2

Drone moved to (1, 4). Step reward: -2

Drone moved to (1, 3). Step reward: -2

Evaluation Episode 1 - Step 371
Drone moved to (1, 4). Step reward: -2

Drone moved to (1, 3). Step reward: -2

Drone moved to (1, 4). Step reward: -2

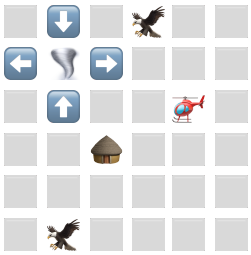
Drone moved to (1, 3). Step reward: -2

Drone moved to (1, 4). Step reward: -2

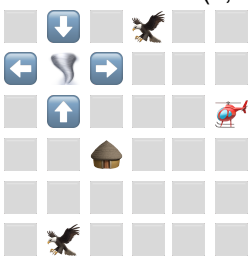
Drone moved to (1, 5). Step reward: -2

Drone moved to (2, 5). Step reward: -2

Drone moved to (2, 4). Step reward: -2



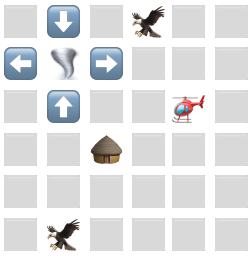
Drone moved to (2, 5). Step reward: -2



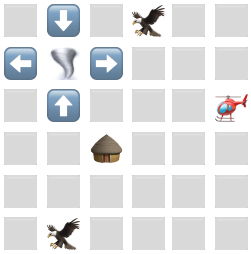
Drone moved to (2, 4). Step reward: -2

Evaluation Episode 1 - Step 381
Drone moved to (2, 5). Step reward: -2

Drone moved to (2, 4). Step reward: -2



Drone moved to (2, 5). Step reward: -2



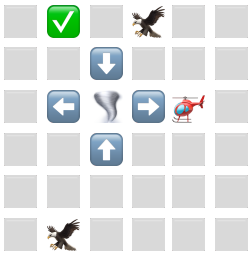
Drone moved to (2, 4). Step reward: -2

Drone moved to (2, 5). Step reward: -2

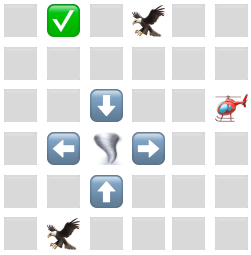
Drone moved to (2, 4). Step reward: -2

Drone moved to (2, 5). Step reward: -2

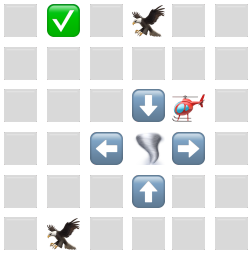
Drone moved to (2, 4). Step reward: -2



Drone moved to (2, 5). Step reward: -2



Drone moved to (2, 4). Step reward: -2



Evaluation Episode 1 - Step 391
Drone moved to (1, 4). Step reward: -2

Drone moved to (1, 3). Step reward: -2

Drone moved to (2, 3). Step reward: -2

Drone moved to (2, 2). Step reward: -2

Drone moved to (2, 1). Step reward: -2

Drone moved to (3, 1). Step reward: -2

Drone moved to (3, 0). Step reward: -2

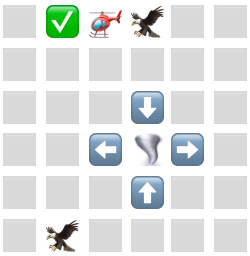
Drone moved to (3, 1). Step reward: -2

Drone moved to (2, 1). Step reward: -2

Drone moved to (1, 1). Step reward: -2

Evaluation Episode 1 - Step 401
Drone moved to (1, 2). Step reward: -2

Drone moved to (0, 2). Step reward: -2



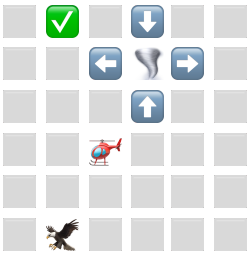
Attempted pickup failed. Penalty -25
Attempted pickup failed. Penalty -25
Attempted pickup failed. Penalty -25
Action 4 repeated 4 times. Switching to a new action.
Attempted pickup failed. Penalty -25
Attempted pickup failed. Penalty -25
Attempted pickup failed. Penalty -25
Action 4 repeated 4 times. Switching to a new action.
Drone moved to (0, 3). Step reward: -50

Drone moved to (1, 3). Step reward: -2

Evaluation Episode 1 - Step 411
Attempted pickup failed. Penalty -25
Attempted pickup failed. Penalty -25
Attempted pickup failed. Penalty -25
Action 4 repeated 4 times. Switching to a new action.
Drone moved to (2, 3). Step reward: -10

Drone moved to (3, 3). Step reward: -2

Drone moved to (3, 2). Step reward: -2



Delivered package_2 for +100 reward
Task complete: All packages delivered 😎

Hyperparams_5 -> Changing alpha from 0.001 to 0.01.

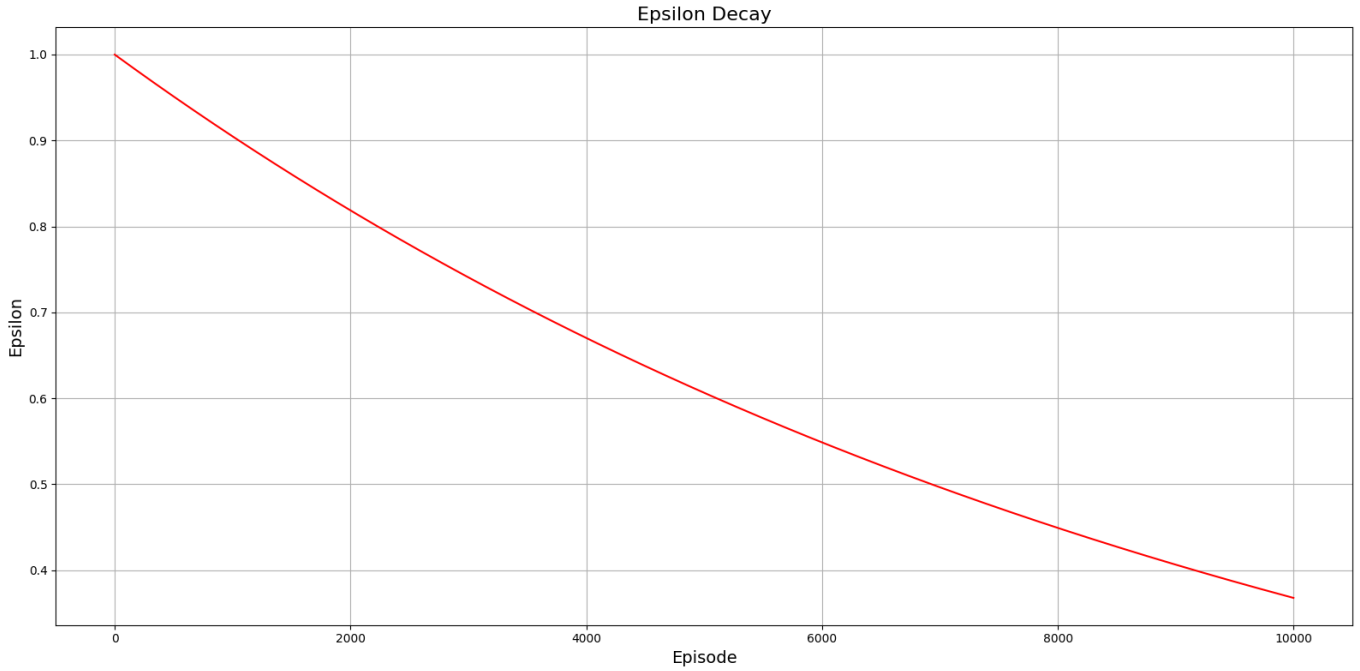
```python
# Hyperparameter set 5 changing alpha = 0.01

hyperparams_5 = {
    'alpha': 0.01,
    'gamma': 0.95,
    'epsilon': 1.0,
    'epsilon_decay': 0.9999,
    'epsilon_min': 0.01,
    'episodes': 10000,
    'max_steps': 1000
}


# Setting environment to stochastic mode
deterministic = False
env = Environment(0, 0, stochastic=(not deterministic))

print("Training Q-Learning agent (Stochastic Environment with Sensor-Augmented State) ...")
Q_stochastic_Q, rewards_stochastic_Q, eps_history_stochastic_Q = train_agent_stochastic_Q(env, hyperparams_5,hyperparams_name='hyperparams_5', render=False)
```

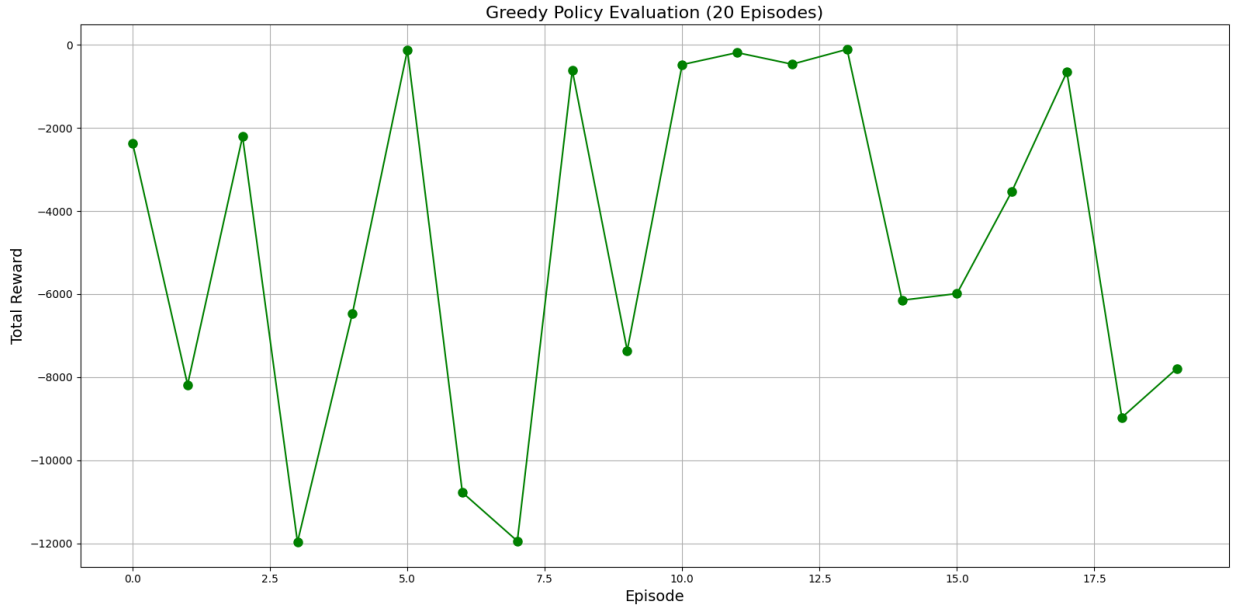
✓ 1m 17.6s

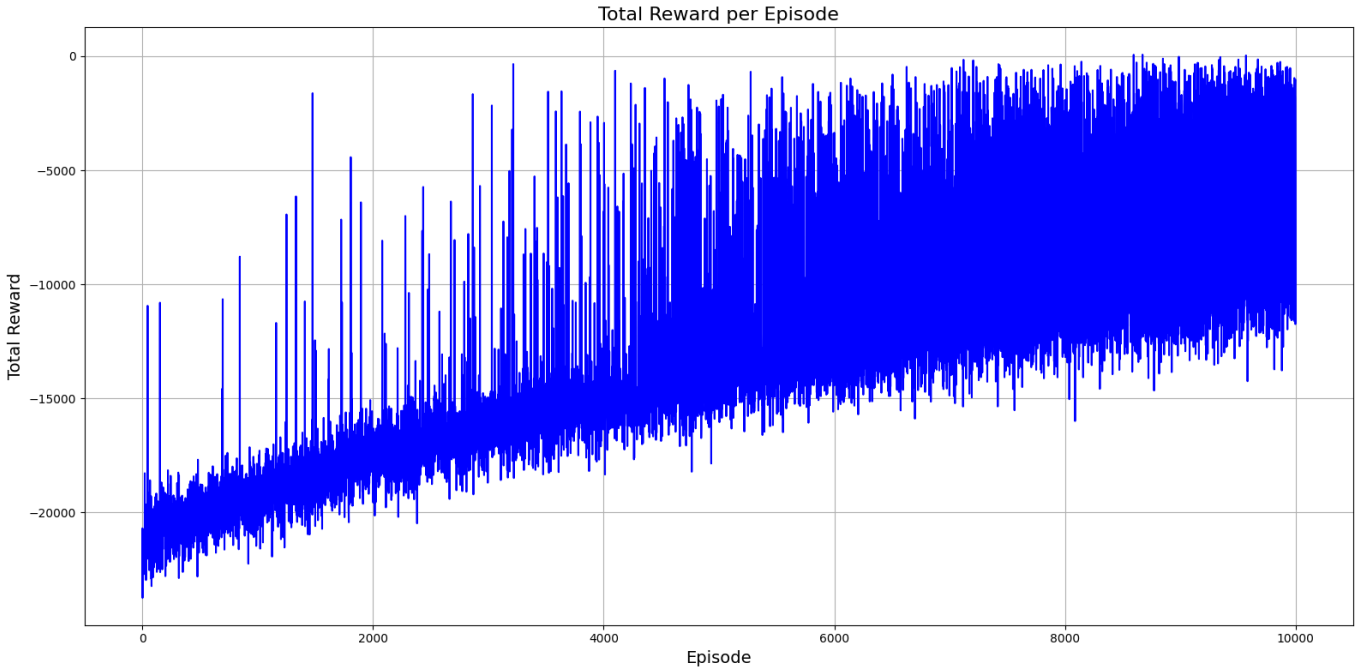
```
Training Q-Learning agent (Stochastic Environment with Sensor-Augmented State) ...
Episode 1000/10000 | Eps: 0.9048 | Success in last 1K: 0.7 %
Episode 2000/10000 | Eps: 0.8187 | Success in last 1K: 2.1 %
Episode 3000/10000 | Eps: 0.7408 | Success in last 1K: 3.7 %
Episode 4000/10000 | Eps: 0.6703 | Success in last 1K: 9.6 %
Episode 5000/10000 | Eps: 0.6065 | Success in last 1K: 17.5 %
Episode 6000/10000 | Eps: 0.5488 | Success in last 1K: 28.3 %
Episode 7000/10000 | Eps: 0.4966 | Success in last 1K: 38.7 %
Episode 8000/10000 | Eps: 0.4493 | Success in last 1K: 47.4 %
Episode 9000/10000 | Eps: 0.4066 | Success in last 1K: 60.0 %
Task complete count: 2765
Episode 10000/10000 | Eps: 0.3679 | Success in last 1K: 68.5 %

Q-table saved to hyperparams_5_stochastic_q_table.pkl
```

Out of 20, only 12 were successfully delivered.

```python
# Evaluating the trained agent
deterministic = False
env = Environment(0, 0, stochastic=(not deterministic))  # Setting up the environment as stochastic

# Loading the trained Q-table
q_table_filename = "hyperparams_5_stochastic_q_table.pkl"
evaluation_rewards_stochastic_Q = evaluate_agent_stochastic_Q(env, q_table_filename=q_table_filename,
                                                              episodes=20, max_steps=1000, render=True)

# Plotting greedy policy evaluation
plt.figure(figsize=(16,8))
plt.plot(evaluation_rewards_stochastic_Q, marker='o', linestyle='-', color='green', markersize=8)
plt.xlabel("Episode", fontsize=14)
plt.ylabel("Total Reward", fontsize=14)
plt.title("Greedy Policy Evaluation (20 Episodes)", fontsize=16)
```







Evaluation Episode 1: Steps: 255 | Total Reward: -1403
Task complete count: 1

--- Evaluation Episode 1 starting ---

State (3, 1, 0, 0, 0, 0, 'package_2', 0, 0, 0, 0, -1, 'package_1') not found or has default Q-values. Choosing random valid action.
Drone moved to (2, 1). Step reward: -2

Evaluation Episode 1 - Step 1
Drone moved to (2, 2). Step reward: -10

Drone moved to (3, 2). Step reward: -2

State (3, 2, 0, 0, 0, 'package_2', -1, 0, 0, 0, -1, 'package_1', 0) not found or has default Q-values. Choosing random valid action.
Drone moved to (4, 2). Step reward: -2

Picked up package 1 for 25 reward
Drone moved to (3, 2). Step reward: -2

State (3, 2, 1, 0, 0, 'package_2', -1, 0, 0, 0, -1, 0, 0) not found or has default Q-values. Choosing random valid action.
Attempted pickup failed. Penalty -25
State (3, 2, 1, 0, 0, 'package_2', -1, 0, 0, 0, -1, 0, 0) not found or has default Q-values. Choosing random valid action.
Drone moved to (4, 2). Step reward: -2

Drone moved to (3, 2). Step reward: -2

State (3, 2, 1, 0, 0, 'package_2', -1, 0, 0, 0, -1, 0, 0) not found or has default Q-values. Choosing random valid action.
Drone moved to (3, 3). Step reward: -2

Drone moved to (3, 2). Step reward: -2

Evaluation Episode 1 - Step 11
State (3, 2, 1, 0, 0, 'package_2', -1, 0, 0, 0, -1, 0, 0) not found or has default Q-values. Choosing random valid action.
Drone moved to (3, 1). Step reward: -2

State (3, 1, 1, 0, 0, 0, 'package_2', 0, 0, 0, 0, -1, 0) not found or has default Q-values. Choosing random valid action.
Drone moved to (3, 0). Step reward: -2

Drone moved to (3, 1). Step reward: -2

State (3, 1, 1, 0, 0, 0, 'package_2', 0, 0, 0, 0, -1, 0) not found or has default Q-values. Choosing random valid action.
Drone moved to (3, 2). Step reward: -2

State (3, 2, 1, 0, 0, 'package_2', 0, 0, 0, 0, -1, 0, 0) not found or has default Q-values. Choosing random valid action.
Drone moved to (3, 3). Step reward: -2

Drone moved to (3, 2). Step reward: -2

State (3, 2, 1, 0, 0, 'package_2', -1, 0, 0, 0, -1, 0, 0) not found or has default Q-values. Choosing random valid action.
Drone moved to (2, 2). Step reward: -10

Drone moved to (2, 3). Step reward: -2

State (2, 3, 1, 0, -1, 0, 0, 'package_2', 0, -1, 0, 0, 0) not found or has default Q-values. Choosing random valid action.
Drone moved to (2, 2). Step reward: -10

Drone moved to (3, 2). Step reward: -2

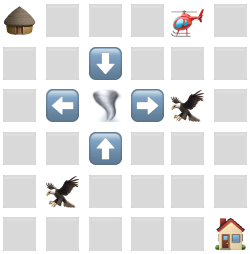
Evaluation Episode 1 - Step 21
Drone moved to (2, 2). Step reward: -10

Drone moved to (2, 3). Step reward: -2

State (2, 3, 1, 0, -1, 0, 0, 'package_2', 0, -1, 0, 0, 0) not found or has default Q-values. Choosing random valid action.
Attempted pickup failed. Penalty -25
State (2, 3, 1, 0, -1, 0, 0, 'package_2', 0, -1, 0, 0, 0) not found or has default Q-values. Choosing random valid action.
Drone moved to (1, 3). Step reward: -2

Attempted pickup failed. Penalty -25
Attempted pickup failed. Penalty -25
Attempted pickup failed. Penalty -25
Action 4 repeated 4 times. Switching to a new action.
Drone moved to (1, 4). Step reward: -2

Drone moved to (0, 4). Step reward: -2



Drone moved to (1, 4). Step reward: -2

Evaluation Episode 1 - Step 31
Drone moved to (1, 5). Step reward: -2

Drone moved to (1, 4). Step reward: -2

Drone moved to (1, 5). Step reward: -2

Drone moved to (2, 5). Step reward: -2

Drone moved to (1, 5). Step reward: -2

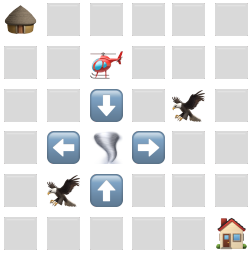
Drone moved to (1, 4). Step reward: -2

Drone moved to (1, 3). Step reward: -2

Drone moved to (1, 4). Step reward: -2

Drone moved to (1, 3). Step reward: -2

Drone moved to (1, 2). Step reward: -2

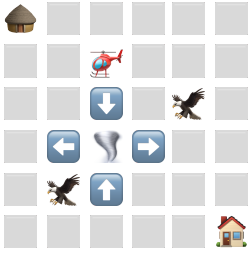


Evaluation Episode 1 - Step 41
Drone moved to (0, 2). Step reward: -2

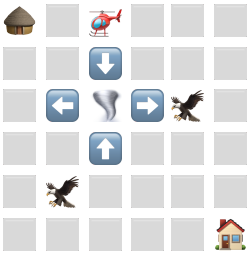
Drone moved to (1, 2). Step reward: -2

Drone moved to (1, 3). Step reward: -2

Drone moved to (1, 2). Step reward: -2



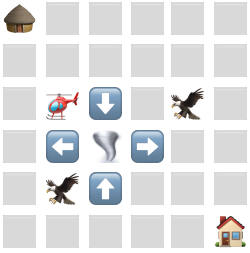
Drone moved to (0, 2). Step reward: -2



Drone moved to (0, 1). Step reward: -2

Drone moved to (1, 1). Step reward: -2

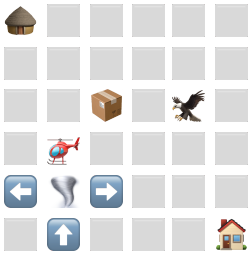
Drone moved to (2, 1). Step reward: -2



Drone moved to (2, 0). Step reward: -2

Drone moved to (2, 1). Step reward: -2

Evaluation Episode 1 - Step 51
Drone moved to (3, 1). Step reward: -10

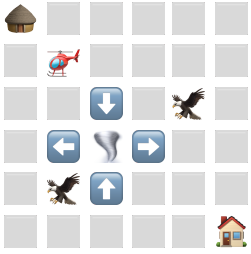


Drone moved to (2, 1). Step reward: -10

Drone moved to (1, 1). Step reward: -2

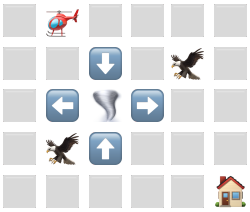
Drone moved to (0, 1). Step reward: -2

Drone moved to (1, 1). Step reward: -2



Drone moved to (0, 1). Step reward: -2

Drone moved to (1, 1). Step reward: -2



Drone moved to (0, 1). Step reward: -2

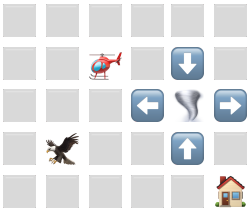
Drone moved to (1, 1). Step reward: -2

Drone moved to (0, 1). Step reward: -2

Evaluation Episode 1 - Step 61
Drone moved to (1, 1). Step reward: -2

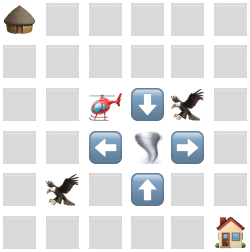
Drone moved to (2, 1). Step reward: -2

Drone moved to (2, 2). Step reward: -2



Picked up package 2 for 25 reward
Drone moved to (2, 3). Step reward: -2

Drone moved to (2, 2). Step reward: -2



Drone moved to (1, 2). Step reward: -2

Drone moved to (2, 2). Step reward: -2

Drone moved to (1, 2). Step reward: -2

Drone moved to (2, 2). Step reward: -2

Evaluation Episode 1 - Step 71
Drone moved to (1, 2). Step reward: -2

Drone moved to (2, 2). Step reward: -2

Drone moved to (2, 3). Step reward: -2

Drone moved to (1, 3). Step reward: -2

Drone moved to (1, 2). Step reward: -2

Drone moved to (2, 2). Step reward: -2

Drone moved to (2, 3). Step reward: -2

Drone moved to (2, 2). Step reward: -2

Drone moved to (2, 3). Step reward: -10

Drone moved to (2, 2). Step reward: -2

Evaluation Episode 1 - Step 81
Drone moved to (2, 3). Step reward: -2

Drone moved to (2, 2). Step reward: -2

Drone moved to (1, 2). Step reward: -2

Drone moved to (2, 2). Step reward: -2

Drone moved to (1, 2). Step reward: -2

Drone moved to (2, 2). Step reward: -2

Drone moved to (1, 2). Step reward: -2

Drone moved to (1, 3). Step reward: -2

Drone moved to (1, 4). Step reward: -2

Drone moved to (1, 3). Step reward: -2

Evaluation Episode 1 - Step 91
Drone moved to (1, 2). Step reward: -2

Drone moved to (2, 2). Step reward: -10

Drone moved to (2, 1). Step reward: -10

Drone moved to (3, 1). Step reward: -2

Drone moved to (3, 0). Step reward: -2

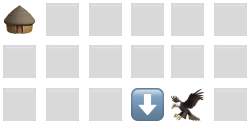
Drone moved to (3, 1). Step reward: -2

Drone moved to (3, 2). Step reward: -2

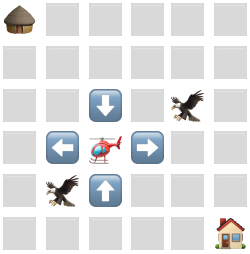
Drone moved to (4, 2). Step reward: -2

Drone moved to (5, 2). Step reward: -2

Drone moved to (4, 2). Step reward: -2

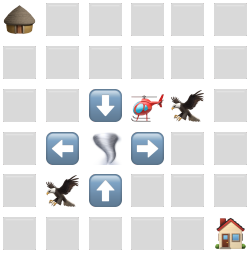


Evaluation Episode 1 - Step 101
Drone moved to (3, 2). Step reward: -100



Drone moved to (3, 3). Step reward: -2

Drone moved to (2, 3). Step reward: -2



Attempted pickup failed. Penalty -25
Attempted pickup failed. Penalty -25
Attempted pickup failed. Penalty -25
Action 4 repeated 4 times. Switching to a new action.
Attempted pickup failed. Penalty -25
Attempted pickup failed. Penalty -25
Attempted pickup failed. Penalty -25
Action 4 repeated 4 times. Switching to a new action.
Attempted pickup failed. Penalty -25
Evaluation Episode 1 - Step 111
Attempted pickup failed. Penalty -25
Attempted pickup failed. Penalty -25
Action 4 repeated 4 times. Switching to a new action.
Dropped package_1 incorrectly. Penalty -50
Dropped package_2 incorrectly. Penalty -50
Picked up package 1 for 25 reward
Picked up package 2 for 25 reward
Action 4 repeated 4 times. Switching to a new action.
Attempted pickup failed. Penalty -25
Attempted pickup failed. Penalty -25
Attempted pickup failed. Penalty -25
Action 4 repeated 4 times. Switching to a new action.
Attempted pickup failed. Penalty -25
Attempted pickup failed. Penalty -25
Evaluation Episode 1 - Step 121

Attempted pickup failed. Penalty -25
Action 4 repeated 4 times. Switching to a new action.
Drone moved to (2, 2). Step reward: -2

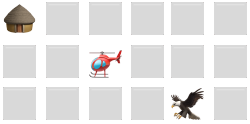
Drone moved to (2, 3). Step reward: -2

Attempted pickup failed. Penalty -25
Attempted pickup failed. Penalty -25
Attempted pickup failed. Penalty -25
Action 4 repeated 4 times. Switching to a new action.
Drone moved to (1, 3). Step reward: -2

Drone moved to (1, 2). Step reward: -2

Drone moved to (2, 2). Step reward: -2

Drone moved to (1, 2). Step reward: -2



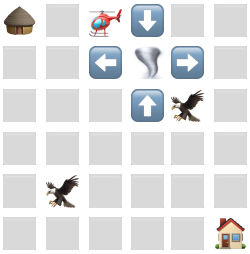
Evaluation Episode 1 - Step 131
Drone moved to (2, 2). Step reward: -2

Drone moved to (1, 2). Step reward: -2

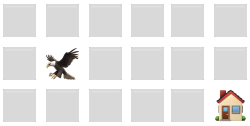
Drone moved to (0, 2). Step reward: -2

Drone moved to (0, 3). Step reward: -2

Drone moved to (0, 2). Step reward: -2



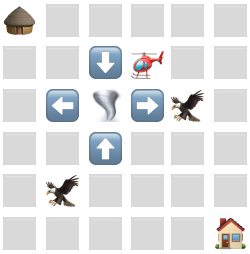
Drone moved to (0, 3). Step reward: -2



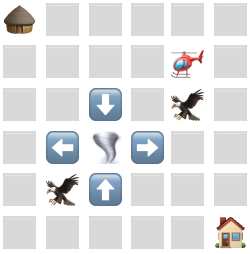
Drone moved to (1, 3). Step reward: -100

Attempted pickup failed. Penalty -25
Attempted pickup failed. Penalty -25
Attempted pickup failed. Penalty -25
Evaluation Episode 1 - Step 141
Action 4 repeated 4 times. Switching to a new action.
Drone moved to (0, 3). Step reward: -2

Drone moved to (1, 3). Step reward: -2



Drone moved to (1, 4). Step reward: -2



Drone moved to (1, 3). Step reward: -2

Drone moved to (0, 3). Step reward: -2

Drone moved to (0, 2). Step reward: -2

Drone moved to (0, 3). Step reward: -10

Drone moved to (0, 4). Step reward: -2

Drone moved to (0, 3). Step reward: -2

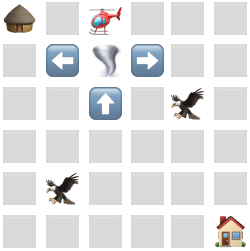
Drone moved to (0, 4). Step reward: -2

Evaluation Episode 1 - Step 151
Drone moved to (0, 3). Step reward: -2

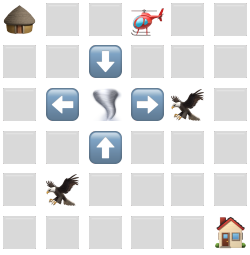
Drone moved to (0, 4). Step reward: -2

Drone moved to (0, 3). Step reward: -2

Drone moved to (0, 2). Step reward: -10

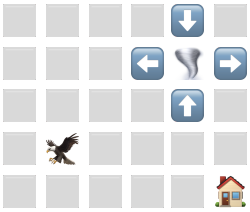


Drone moved to (0, 3). Step reward: -2



Drone moved to (0, 2). Step reward: -2

Drone moved to (0, 3). Step reward: -2



Drone moved to (0, 4). Step reward: -10

Drone moved to (0, 3). Step reward: -2

Drone moved to (0, 2). Step reward: -2

Evaluation Episode 1 - Step 161
Drone moved to (0, 3). Step reward: -2

Drone moved to (0, 2). Step reward: -2

Drone moved to (0, 3). Step reward: -2

Drone moved to (1, 3). Step reward: -2

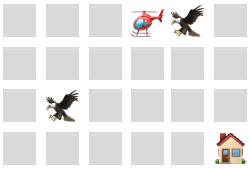
Drone moved to (1, 4). Step reward: -2

Drone moved to (1, 3). Step reward: -2

Drone moved to (1, 4). Step reward: -2

Drone moved to (1, 3). Step reward: -10

Drone moved to (2, 3). Step reward: -10



Drone moved to (2, 4). Step reward: -50

Evaluation Episode 1 - Step 171
Drone moved to (3, 4). Step reward: -2

Drone moved to (3, 5). Step reward: -2

Drone moved to (3, 4). Step reward: -2

Drone moved to (3, 5). Step reward: -2

Drone moved to (3, 4). Step reward: -2

Drone moved to (3, 3). Step reward: -2

Drone moved to (3, 2). Step reward: -2

Drone moved to (3, 3). Step reward: -2

Drone moved to (2, 3). Step reward: -2

Attempted pickup failed. Penalty -25
Evaluation Episode 1 - Step 181
Attempted pickup failed. Penalty -25
Attempted pickup failed. Penalty -25
Action 4 repeated 4 times. Switching to a new action.
Drone moved to (2, 2). Step reward: -2

Drone moved to (3, 2). Step reward: -2

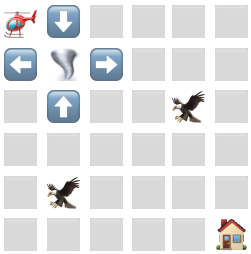
Drone moved to (3, 1). Step reward: -2

Drone moved to (2, 1). Step reward: -10

Drone moved to (2, 0). Step reward: -10

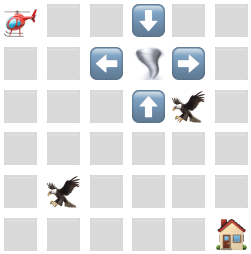
Drone moved to (1, 0). Step reward: -10

State (1, 0, 1, 1, -1, 'destination_2', -1, -1, -1, -1, -1, 0, -1) not found or has default Q-values. Choosing random valid action.
Drone moved to (0, 0). Step reward: -2



Dropped package_1 incorrectly. Penalty -50
Delivered package_2 for +100 reward
Evaluation Episode 1 - Step 191
Picked up package 1 for 25 reward
Drone moved to (0, 1). Step reward: -2

Drone moved to (0, 0). Step reward: -2



Drone moved to (1, 0). Step reward: -2

Drone moved to (1, 1). Step reward: -2

Drone moved to (0, 1). Step reward: -2

Drone moved to (0, 0). Step reward: -2

Drone moved to (0, 1). Step reward: -2

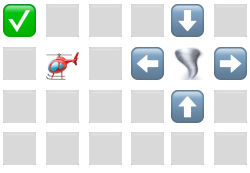
Drone moved to (0, 0). Step reward: -2

Drone moved to (1, 0). Step reward: -2

Evaluation Episode 1 - Step 201
Drone moved to (1, 1). Step reward: -2

Drone moved to (0, 1). Step reward: -2

Drone moved to (1, 1). Step reward: -2



Drone moved to (0, 1). Step reward: -2

Drone moved to (1, 1). Step reward: -2

Drone moved to (0, 1). Step reward: -2

Drone moved to (1, 1). Step reward: -2

Drone moved to (1, 2). Step reward: -10

Drone moved to (1, 3). Step reward: -2

Drone moved to (1, 4). Step reward: -2

Evaluation Episode 1 - Step 211
Drone moved to (1, 3). Step reward: -2

Drone moved to (1, 4). Step reward: -2

Drone moved to (0, 4). Step reward: -2

Drone moved to (0, 5). Step reward: -2

Drone moved to (0, 4). Step reward: -2

Drone moved to (1, 4). Step reward: -2

Drone moved to (1, 3). Step reward: -2

Drone moved to (2, 3). Step reward: -2

Drone moved to (1, 3). Step reward: -2

Drone moved to (0, 3). Step reward: -2

Evaluation Episode 1 - Step 221
Drone moved to (1, 3). Step reward: -2

Drone moved to (2, 3). Step reward: -2

Drone moved to (1, 3). Step reward: -2

Drone moved to (0, 3). Step reward: -2

Drone moved to (0, 2). Step reward: -2

Drone moved to (0, 3). Step reward: -2

Drone moved to (0, 2). Step reward: -2

Drone moved to (1, 2). Step reward: -2

Drone moved to (2, 2). Step reward: -2

Drone moved to (3, 2). Step reward: -10

Evaluation Episode 1 - Step 231
Action 1 repeated 4 times. Switching to a new action.
Dropped package_1 incorrectly. Penalty -50
Drone moved to (3, 1). Step reward: -2

Drone moved to (3, 0). Step reward: -2

Drone moved to (3, 1). Step reward: -2

Drone moved to (2, 1). Step reward: -2

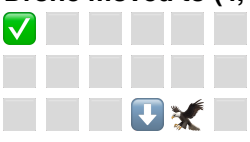
Drone moved to (2, 2). Step reward: -2

Drone moved to (3, 2). Step reward: -2

Picked up package 1 for 25 reward
Drone moved to (4, 2). Step reward: -2

Drone moved to (5, 2). Step reward: -2

Evaluation Episode 1 - Step 241
Drone moved to (4, 2). Step reward: -2



Drone moved to (3, 2). Step reward: -2

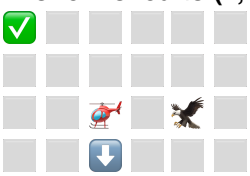
Drone moved to (3, 1). Step reward: -2

Drone moved to (3, 0). Step reward: -2

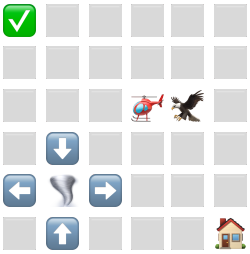
Drone moved to (2, 0). Step reward: -2

Drone moved to (2, 1). Step reward: -2

Drone moved to (2, 2). Step reward: -2



Drone moved to (2, 3). Step reward: -2



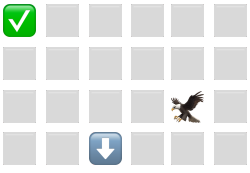
Drone moved to (3, 3). Step reward: -2

Drone moved to (3, 4). Step reward: -2

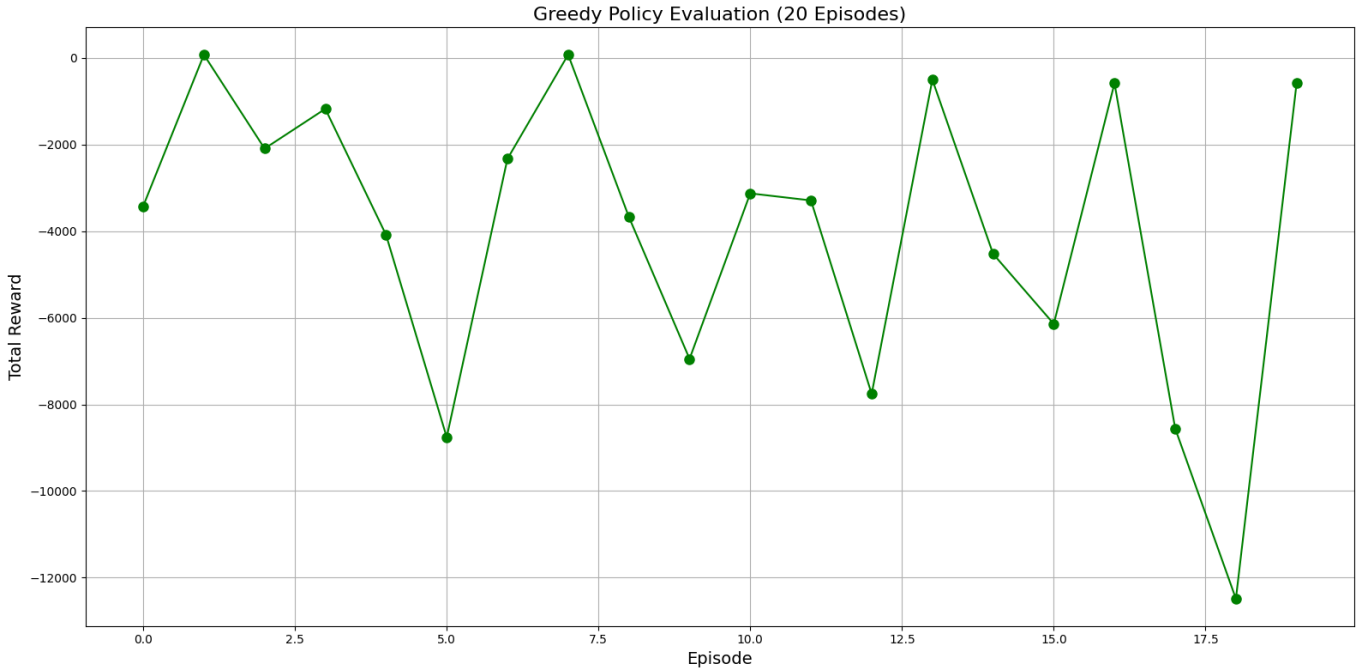
Evaluation Episode 1 - Step 251
Drone moved to (4, 4). Step reward: -2

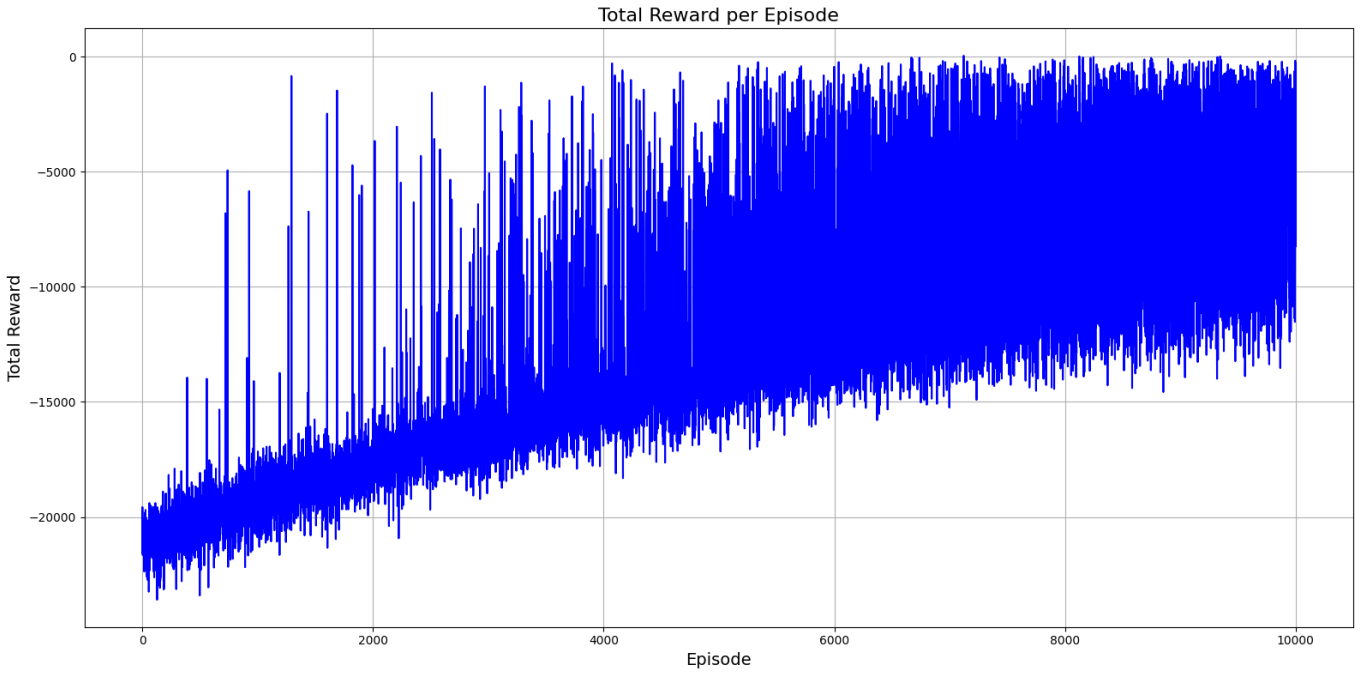
Drone moved to (5, 4). Step reward: -2

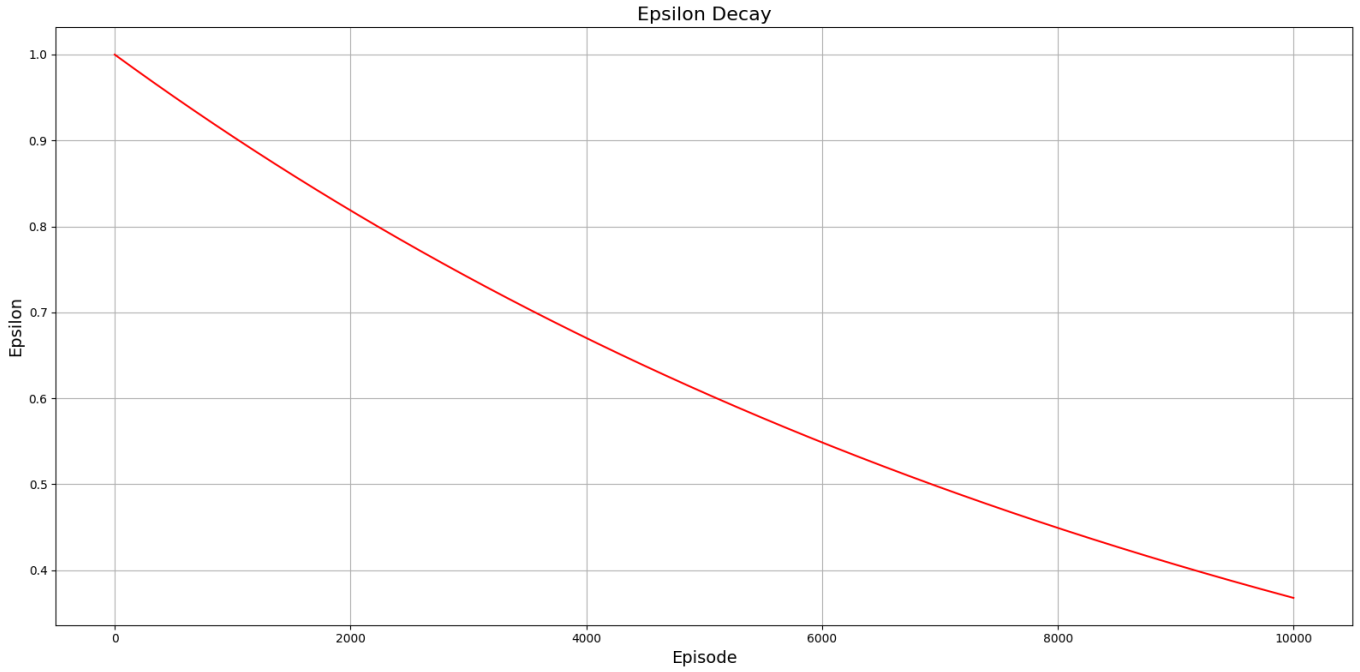
Drone moved to (5, 5). Step reward: -2



Delivered package_1 for +100 reward
Task complete: All packages delivered 😎


Greedy Policy Evaluation (20 Episodes)
Total Reward
Episode
0
−2000
−4000
−6000
−8000
−10000
−12000
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5


Total Reward per Episode
Total Reward
Episode
0
−5000
−10000
−15000
−20000
0
2000
4000
6000
8000
10000


Epsilon Decay
Epsilon
Episode
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0
2000
4000
6000
8000
10000

hyperparams_4 -> Changing alpha from 0.01 to 0.1.

```python
# Hyperparameter set 4 changing alpha = 0.1

hyperparams_4 = {
    'alpha': 0.1,
    'gamma': 0.95,
    'epsilon': 1.0,
    'epsilon_decay': 0.9999,
    'epsilon_min': 0.01,
    'episodes': 10000,
    'max_steps': 1000
}

# Setting environment to stochastic mode
deterministic = False
env = Environment(0, 0, stochastic=(not deterministic))

print("Training Q-Learning agent (Stochastic Environment with Sensor-Augmented State) ...")
Q_stochastic_Q, rewards_stochastic_Q, eps_history_stochastic_Q = train_agent_stochastic_Q(env, hyperparams_4, hyperparams_name='hyperparams_4',render=False)
```

✓ 1m 14.6s

```
Training Q-Learning agent (Stochastic Environment with Sensor-Augmented State) ...
Episode 1000/10000 | Eps: 0.9048 | Success in last 1K: 1.0 %
Episode 2000/10000 | Eps: 0.8187 | Success in last 1K: 1.5 %
Episode 3000/10000 | Eps: 0.7408 | Success in last 1K: 6.3 %
Episode 4000/10000 | Eps: 0.6703 | Success in last 1K: 13.4 %
Episode 5000/10000 | Eps: 0.6065 | Success in last 1K: 22.2 %
Episode 6000/10000 | Eps: 0.5488 | Success in last 1K: 39.9 %
Episode 7000/10000 | Eps: 0.4966 | Success in last 1K: 52.0 %
Episode 8000/10000 | Eps: 0.4493 | Success in last 1K: 66.1 %
Episode 9000/10000 | Eps: 0.4066 | Success in last 1K: 68.0 %
Task complete count: 3456
Episode 10000/10000 | Eps: 0.3679 | Success in last 1K: 75.2 %

Q-table saved to hyperparams_4_stochastic_q_table.pkl
```

Out of 20, only 15 were successfully delivered.

```python
# Evaluating the trained agent
deterministic = False
env = Environment(0, 0, stochastic=(not deterministic))  # Setting up the environment as stochastic

# Loading the trained Q-table
q_table_filename = "hyperparams_4_stochastic_q_table.pkl"

evaluation_rewards_stochastic_Q = evaluate_agent_stochastic_Q(env, q_table_filename=q_table_filename,
                                                              episodes=20, max_steps=1000, render=True)

# Plotting greedy policy evaluation
plt.figure(figsize=(16,8))
plt.plot(evaluation_rewards_stochastic_Q, marker='o', linestyle='-', color='green', markersize=8)
plt.xlabel("Episode", fontsize=14)
plt.ylabel("Total Reward", fontsize=14)
plt.title("Greedy Policy Evaluation (20 Episodes)", fontsize=16)
plt.grid(True)
plt.tight_layout()
plt.show()
```

✓ 0.6s

```
Evaluation Episode 1: Steps: 436 | Total Reward: -3424
Evaluation Episode 2: Steps: 53 | Total Reward: 72
Evaluation Episode 3: Steps: 1000 | Total Reward: -2095
Evaluation Episode 4: Steps: 226 | Total Reward: -1178
Evaluation Episode 5: Steps: 399 | Total Reward: -4086
Evaluation Episode 6: Steps: 1000 | Total Reward: -8759
Evaluation Episode 7: Steps: 213 | Total Reward: -2333
Evaluation Episode 8: Steps: 35 | Total Reward: 73
Evaluation Episode 9: Steps: 319 | Total Reward: -3671
Evaluation Episode 10: Steps: 750 | Total Reward: -6950
Evaluation Episode 11: Steps: 431 | Total Reward: -3129
Evaluation Episode 12: Steps: 1000 | Total Reward: -3292
Evaluation Episode 13: Steps: 1000 | Total Reward: -7758
Evaluation Episode 14: Steps: 127 | Total Reward: -505
Evaluation Episode 15: Steps: 721 | Total Reward: -4525
Evaluation Episode 16: Steps: 742 | Total Reward: -6143
Evaluation Episode 17: Steps: 176 | Total Reward: -587
Evaluation Episode 18: Steps: 986 | Total Reward: -8557
Evaluation Episode 19: Steps: 1000 | Total Reward: -12488
Evaluation Episode 20: Steps: 190 | Total Reward: -587
Task complete count: 15
```

Evaluation Episode 1: Steps: 70 | Total Reward: -449
Task complete count: 1

--- Evaluation Episode 1 starting ---

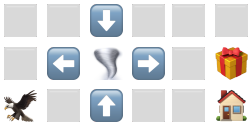
Drone moved to (2, 4). Step reward: -2

Evaluation Episode 1 - Step 1
Picked up package 2 for 25 reward
Drone moved to (2, 3). Step reward: -2

Drone moved to (2, 4). Step reward: -2

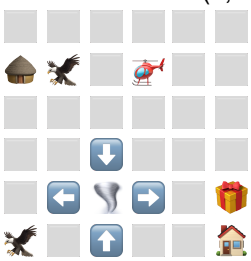
Drone moved to (1, 4). Step reward: -2

Drone moved to (2, 4). Step reward: -2

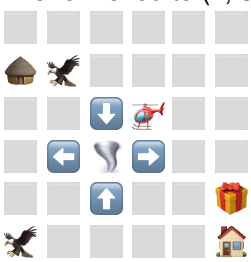


Drone moved to (2, 3). Step reward: -2

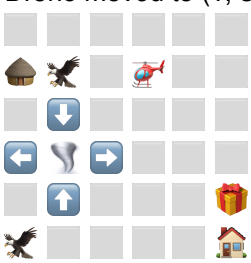
Drone moved to (1, 3). Step reward: -2



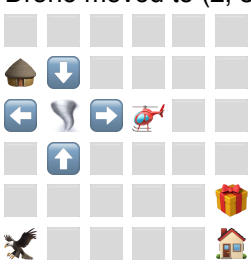
Drone moved to (2, 3). Step reward: -2



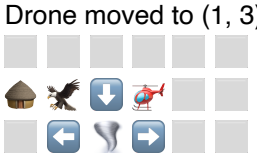
Drone moved to (1, 3). Step reward: -2

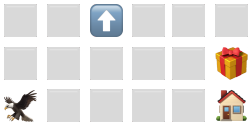


Drone moved to (2, 3). Step reward: -2



Evaluation Episode 1 - Step 11
Drone moved to (1, 3). Step reward: -2



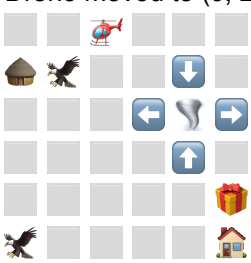


Drone moved to (0, 3). Step reward: -2

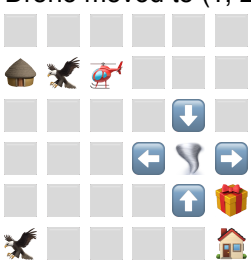
Drone moved to (0, 2). Step reward: -2

Drone moved to (0, 3). Step reward: -2

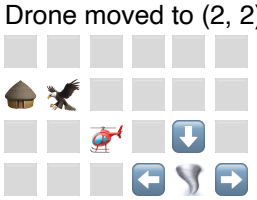
Drone moved to (0, 2). Step reward: -2



Drone moved to (1, 2). Step reward: -2



Drone moved to (2, 2). Step reward: -2



Drone moved to (1, 2). Step reward: -2

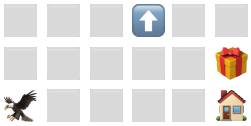
Drone moved to (2, 2). Step reward: -2

Drone moved to (2, 1). Step reward: -2

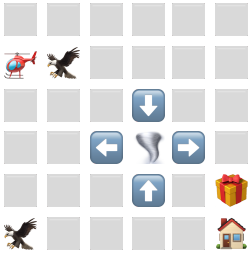
Evaluation Episode 1 - Step 21
Drone moved to (1, 1). Step reward: -50

Drone moved to (1, 0). Step reward: -2

Delivered package_2 for +100 reward
Drone moved to (2, 0). Step reward: -2



Drone moved to (1, 0). Step reward: -2



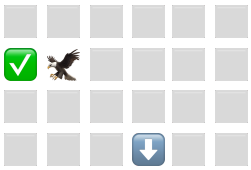
Drone moved to (2, 0). Step reward: -2

Drone moved to (2, 1). Step reward: -10

Drone moved to (3, 1). Step reward: -10

Drone moved to (4, 1). Step reward: -10

Drone moved to (4, 0). Step reward: -2



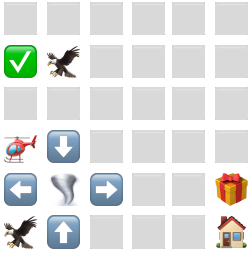
Evaluation Episode 1 - Step 31
Drone moved to (4, 1). Step reward: -2

Drone moved to (4, 0). Step reward: -2

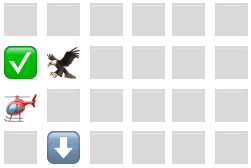
Drone moved to (3, 0). Step reward: -2

Drone moved to (2, 0). Step reward: -2

Drone moved to (3, 0). Step reward: -2



Drone moved to (2, 0). Step reward: -2



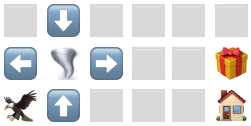
Drone moved to (2, 1). Step reward: -10

Drone moved to (1, 1). Step reward: -50

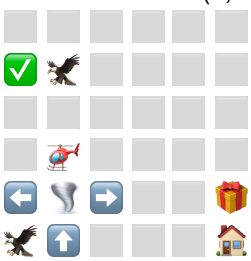
Drone moved to (2, 1). Step reward: -2

State (2, 1, 0, 2, 'destination_2', -1, 0, 0, 0, 0, 0, -1, 0) not found or has default Q-values. Choosing random valid action.
Attempted pickup failed. Penalty -25
Evaluation Episode 1 - Step 41
State (2, 1, 0, 2, 'destination_2', -1, 0, 0, 0, 0, 0, -1, 0) not found or has default Q-values. Choosing random valid action.
Attempted pickup failed. Penalty -25
State (2, 1, 0, 2, 'destination_2', -1, 0, 0, 0, 0, 0, -1, 0) not found or has default Q-values. Choosing random valid action.
Attempted pickup failed. Penalty -25
State (2, 1, 0, 2, 'destination_2', -1, 0, 0, 0, 0, 0, -1, 0) not found or has default Q-values. Choosing random valid action.
Drone moved to (1, 1). Step reward: -50

Drone moved to (2, 1). Step reward: -2



State (2, 1, 0, 2, 'destination_2', -1, 0, 0, 0, 0, 0, -1, 0) not found or has default Q-values. Choosing random valid action.
Attempted dropoff failed (no package carried). Penalty -50
State (2, 1, 0, 2, 'destination_2', -1, 0, 0, 0, 0, 0, -1, 0) not found or has default Q-values. Choosing random valid action.
Drone moved to (3, 1). Step reward: -10



Drone moved to (3, 2). Step reward: -10

Drone moved to (3, 1). Step reward: -10

Drone moved to (4, 1). Step reward: -10

Drone moved to (5, 1). Step reward: -2

Evaluation Episode 1 - Step 51

Drone moved to (5, 2). Step reward: -2

Drone moved to (5, 3). Step reward: -2

Drone moved to (4, 3). Step reward: -2

Drone moved to (3, 3). Step reward: -2

Drone moved to (2, 3). Step reward: -2

Drone moved to (3, 3). Step reward: -2

Drone moved to (3, 4). Step reward: -2

Drone moved to (3, 3). Step reward: -100

Drone moved to (4, 3). Step reward: -2

Drone moved to (4, 4). Step reward: -2

Evaluation Episode 1 - Step 61
Attempted dropoff failed (no package carried). Penalty -50
Attempted dropoff failed (no package carried). Penalty -50
Attempted dropoff failed (no package carried). Penalty -50
Action 5 repeated 4 times. Switching to a new action.
Drone moved to (3, 4). Step reward: -2

Drone moved to (3, 5). Step reward: -2

Drone moved to (4, 5). Step reward: -2

Picked up package 1 for 25 reward
Drone moved to (5, 5). Step reward: -2

Delivered package_1 for +100 reward
Task complete: All packages delivered 😎

hyperparams_3 -> Changing episodes from 10000 to 15000

```python
# Hyperparameter set 3 chaning episodes = 15000
hyperparams_3 = {
    'alpha': 0.1,
    'gamma': 0.95,
    'epsilon': 1.0,
    'epsilon_decay': 0.9999,
    'epsilon_min': 0.01,
    'episodes': 15000,
    'max_steps': 1000
}

# Setting environment to stochastic mode
deterministic = False
env = Environment(0, 0, stochastic=(not deterministic))

print("Training Q-Learning agent (Stochastic Environment with Sensor-Augmented State) ...")
Q_stochastic_Q, rewards_stochastic_Q, eps_history_stochastic_Q = train_agent_stochastic_Q(env, hyperparams_3, hyperparams_name='hyperparams_3', render=False)
```

✓ 1m 36.3s

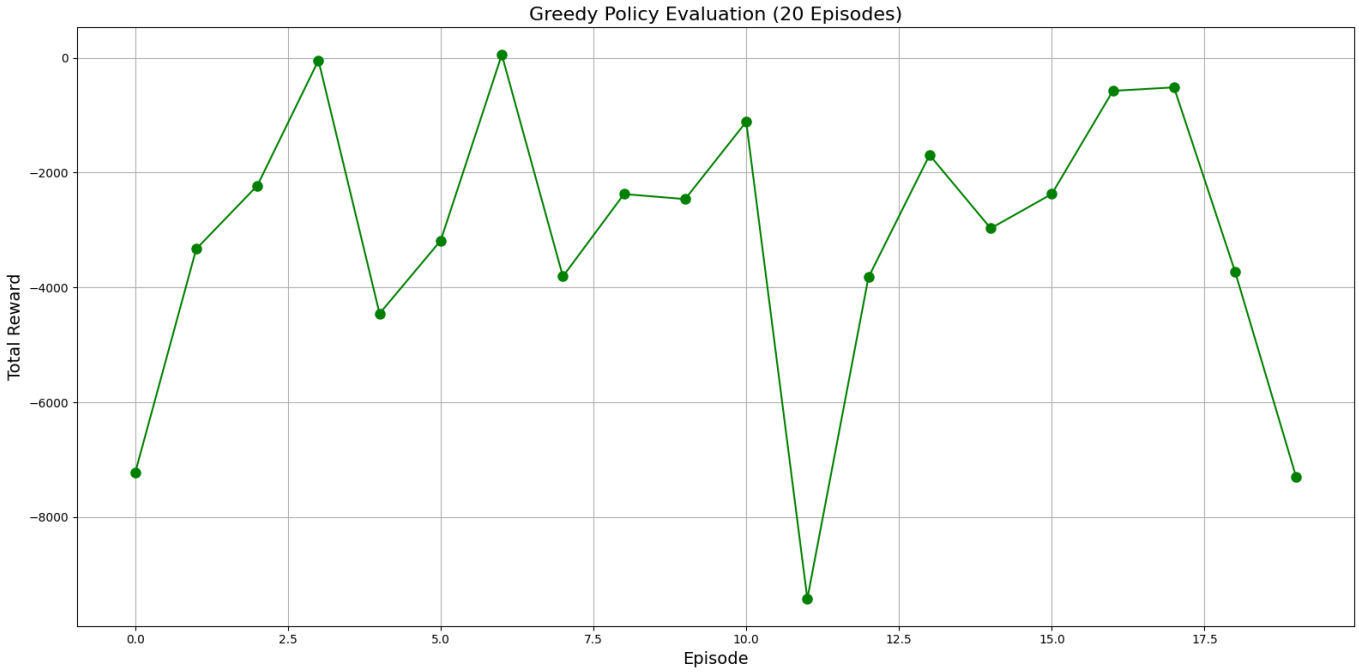
```
Training Q-Learning agent (Stochastic Environment with Sensor-Augmented State) ...
Episode 1000/15000 | Eps: 0.9048 | Success in last 1K: 0.7 %
Episode 2000/15000 | Eps: 0.8187 | Success in last 1K: 2.4 %
Episode 3000/15000 | Eps: 0.7408 | Success in last 1K: 6.4 %
Episode 4000/15000 | Eps: 0.6703 | Success in last 1K: 10.9 %
Episode 5000/15000 | Eps: 0.6065 | Success in last 1K: 25.5 %
Episode 6000/15000 | Eps: 0.5488 | Success in last 1K: 35.7 %
Episode 7000/15000 | Eps: 0.4966 | Success in last 1K: 53.0 %
Episode 8000/15000 | Eps: 0.4493 | Success in last 1K: 62.8 %
Episode 9000/15000 | Eps: 0.4066 | Success in last 1K: 70.8 %
Episode 10000/15000 | Eps: 0.3679 | Success in last 1K: 76.1 %
Episode 11000/15000 | Eps: 0.3329 | Success in last 1K: 78.1 %
Episode 12000/15000 | Eps: 0.3012 | Success in last 1K: 83.1 %
Episode 13000/15000 | Eps: 0.2725 | Success in last 1K: 85.3 %
Episode 14000/15000 | Eps: 0.2466 | Success in last 1K: 86.2 %
Task complete count: 7641
Episode 15000/15000 | Eps: 0.2231 | Success in last 1K: 87.1 %

Q-table saved to hyperparams_3_stochastic_q_table.pkl
```

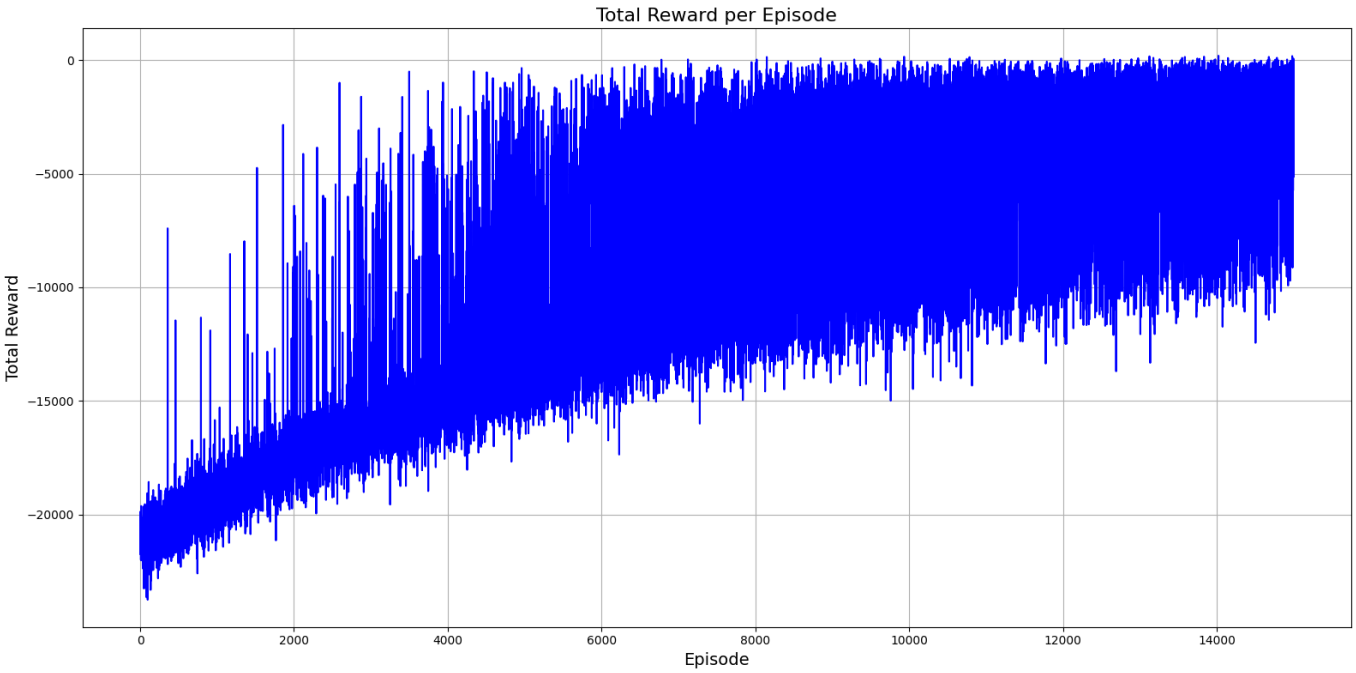
Out of 20, only 15 were successfully delivered.

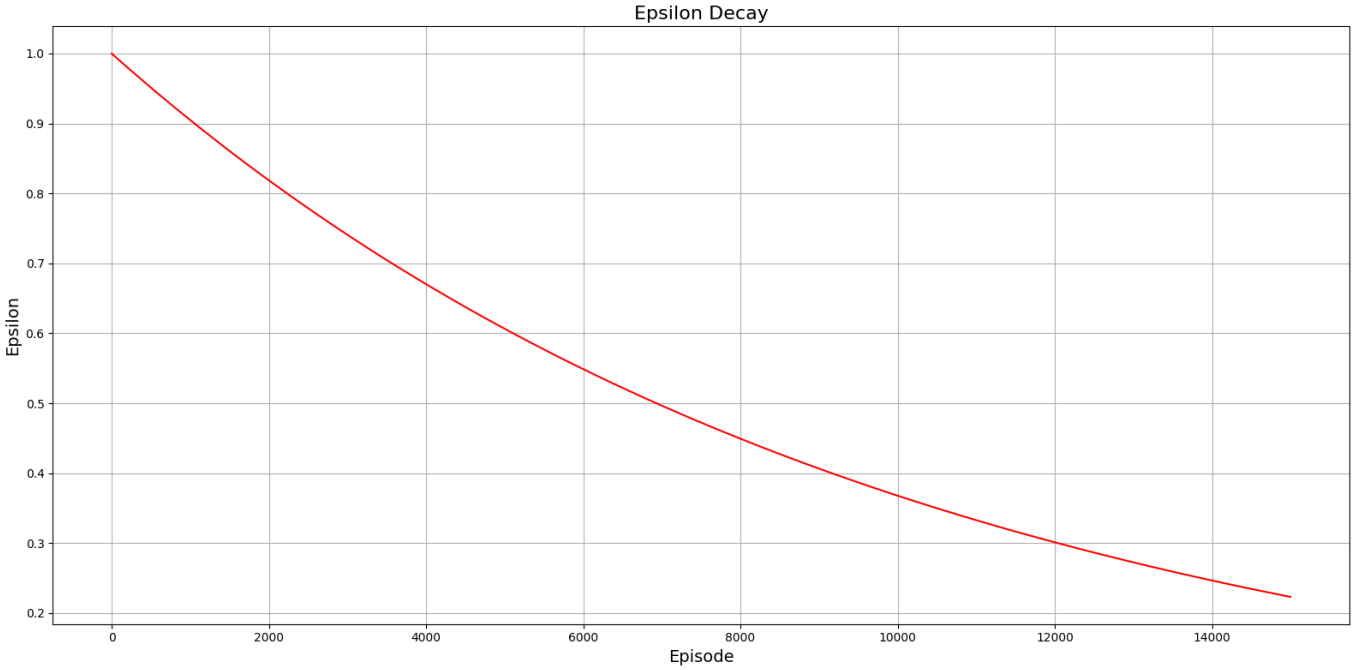
```python
# Evaluating the trained agent

deterministic = False
env = Environment(0, 0, stochastic=(not deterministic))  # Setting up the environment as stochastic

# Loading the trained Q-table
q_table_filename = "hyperparams_3_stochastic_q_table.pkl"

evaluation_rewards_stochastic_Q = evaluate_agent_stochastic_Q(env, q_table_filename=q_table_filename,
                                                              episodes=20, max_steps=1000, render=True)

# Plotting greedy policy evaluation
plt.figure(figsize=(16,8))
plt.plot(evaluation_rewards_stochastic_Q, marker='o', linestyle='-', color='green', markersize=8)
plt.xlabel("Episode", fontsize=14)
plt.ylabel("Total Reward", fontsize=14)
plt.title("Greedy Policy Evaluation (20 Episodes)", fontsize=16)
plt.grid(True)
plt.tight_layout()
plt.show()
```

✓ 0.7s

```
Evaluation Episode 1: Steps: 1000 | Total Reward: -7230
Evaluation Episode 2: Steps: 1000 | Total Reward: -3331
Evaluation Episode 3: Steps: 447 | Total Reward: -2227
Evaluation Episode 4: Steps: 133 | Total Reward: -47
Evaluation Episode 5: Steps: 789 | Total Reward: -4456
Evaluation Episode 6: Steps: 340 | Total Reward: -3185
Evaluation Episode 7: Steps: 51 | Total Reward: 52
Evaluation Episode 8: Steps: 731 | Total Reward: -3810
Evaluation Episode 9: Steps: 409 | Total Reward: -2375
Evaluation Episode 10: Steps: 370 | Total Reward: -2463
Evaluation Episode 11: Steps: 194 | Total Reward: -1116
Evaluation Episode 12: Steps: 1000 | Total Reward: -9422
Evaluation Episode 13: Steps: 406 | Total Reward: -3825
Evaluation Episode 14: Steps: 346 | Total Reward: -1695
Evaluation Episode 15: Steps: 1000 | Total Reward: -2972
Evaluation Episode 16: Steps: 367 | Total Reward: -2374
Evaluation Episode 17: Steps: 114 | Total Reward: -578
Evaluation Episode 18: Steps: 129 | Total Reward: -517
Evaluation Episode 19: Steps: 656 | Total Reward: -3728
Evaluation Episode 20: Steps: 1000 | Total Reward: -7306
Task complete count: 15
```


Greedy Policy Evaluation (20 Episodes)
Total Reward
Episode
0
−2000
−4000
−6000
−8000
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5


Total Reward per Episode
Total Reward
Episode
0
−5000
−10000
−15000
−20000
0
2000
4000
6000
8000
10000
12000
14000


Epsilon Decay
Epsilon
Episode
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0
2000
4000
6000
8000
10000
12000
14000

Evaluation Episode 1: Steps: 75 | Total Reward: -142
Task complete count: 1

--- Evaluation Episode 1 starting ---

Drone moved to (1, 1). Step reward: -2

Evaluation Episode 1 - Step 1
Drone moved to (1, 2). Step reward: -2

Drone moved to (1, 1). Step reward: -2

Drone moved to (1, 2). Step reward: -2

Drone moved to (1, 1). Step reward: -2

Drone moved to (1, 2). Step reward: -2

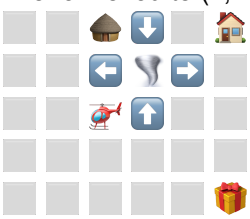
Drone moved to (1, 1). Step reward: -2

Drone moved to (1, 2). Step reward: -10

Drone moved to (1, 1). Step reward: -2

Drone moved to (2, 1). Step reward: -2

Drone moved to (2, 2). Step reward: -2



Evaluation Episode 1 - Step 11
Drone moved to (3, 2). Step reward: -2

Drone moved to (3, 1). Step reward: -2

Drone moved to (3, 2). Step reward: -2

Drone moved to (3, 1). Step reward: -2

Drone moved to (3, 2). Step reward: -2

Drone moved to (3, 1). Step reward: -2

Drone moved to (3, 2). Step reward: -2

Drone moved to (3, 1). Step reward: -2

Drone moved to (3, 2). Step reward: -2

Drone moved to (3, 1). Step reward: -2

Evaluation Episode 1 - Step 21
Drone moved to (3, 2). Step reward: -2

Drone moved to (3, 1). Step reward: -2

Drone moved to (3, 2). Step reward: -2

Drone moved to (4, 2). Step reward: -2

Drone moved to (4, 3). Step reward: -2

Drone moved to (4, 4). Step reward: -2

Drone moved to (4, 5). Step reward: -2

Picked up package 1 for 25 reward
Drone moved to (4, 4). Step reward: -2

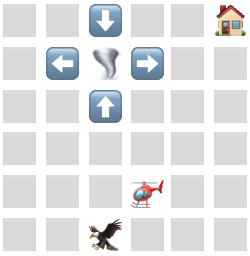
Drone moved to (4, 3). Step reward: -2

Evaluation Episode 1 - Step 31
Drone moved to (5, 3). Step reward: -2

Picked up package 2 for 25 reward
Drone moved to (5, 4). Step reward: -2

Drone moved to (4, 4). Step reward: -2

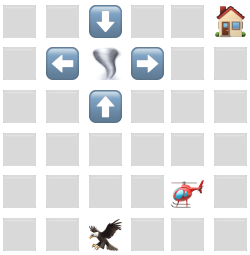
Drone moved to (4, 3). Step reward: -2



Drone moved to (5, 3). Step reward: -2

Drone moved to (5, 4). Step reward: -2

Drone moved to (4, 4). Step reward: -2



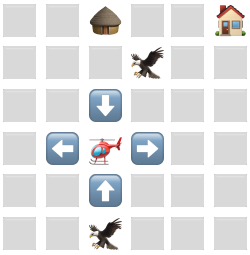
Drone moved to (4, 3). Step reward: -2

Drone moved to (4, 2). Step reward: -2

Evaluation Episode 1 - Step 41
Drone moved to (4, 1). Step reward: -2

Drone moved to (4, 2). Step reward: -2

Drone moved to (3, 2). Step reward: -100



Drone moved to (3, 1). Step reward: -2

Drone moved to (3, 0). Step reward: -2

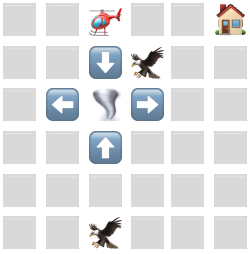
Drone moved to (2, 0). Step reward: -2

Drone moved to (2, 1). Step reward: -2

Drone moved to (1, 1). Step reward: -2

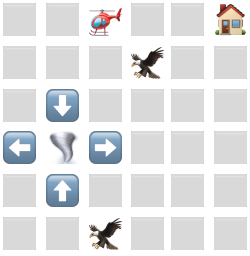
State (1, 1, 1, 1, 0, 0, 'destination_2', 0, 0, 0, 0, -1, 0) not found or has default Q-values. Choosing random valid action.
Drone moved to (0, 1). Step reward: -2

Drone moved to (0, 2). Step reward: -2



Evaluation Episode 1 - Step 51
Drone moved to (0, 1). Step reward: -2

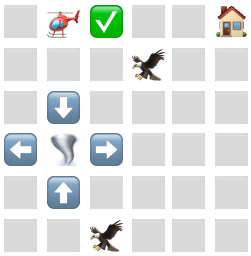
Drone moved to (0, 2). Step reward: -2



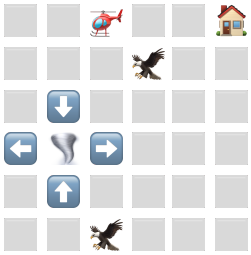
Dropped package_1 incorrectly. Penalty -50
Delivered package_2 for +100 reward
Picked up package 1 for 25 reward
Drone moved to (1, 2). Step reward: -2

State (1, 2, 1, 2, 0, 'destination_2', 0, 0, 0, -1, -1, 0, 0) not found or has default Q-values. Choosing random valid action.
Drone moved to (1, 1). Step reward: -10

Drone moved to (0, 1). Step reward: -2

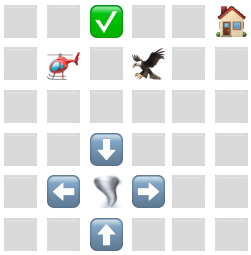


Drone moved to (0, 2). Step reward: -2



Drone moved to (1, 2). Step reward: -2

Drone moved to (1, 1). Step reward: -2

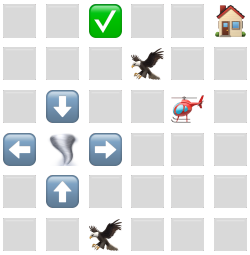


Evaluation Episode 1 - Step 61
Drone moved to (1, 2). Step reward: -2

Drone moved to (2, 2). Step reward: -2

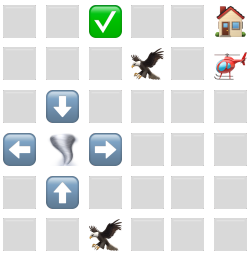
Attempted pickup failed. Penalty -25
Attempted pickup failed. Penalty -25
Attempted pickup failed. Penalty -25
Action 4 repeated 4 times. Switching to a new action.
Dropped package_1 incorrectly. Penalty -50
Picked up package 1 for 25 reward
Attempted pickup failed. Penalty -25
Action 4 repeated 4 times. Switching to a new action.
Drone moved to (2, 3). Step reward: -2

Drone moved to (2, 4). Step reward: -2

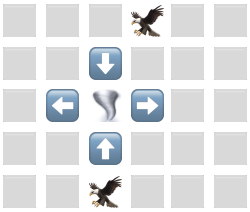


Evaluation Episode 1 - Step 71
Drone moved to (1, 4). Step reward: -2

Drone moved to (1, 5). Step reward: -2



Drone moved to (0, 5). Step reward: -2



Delivered package_1 for +100 reward
Task complete: All packages delivered 😎

hyperparams_2 -> Changing episodes from 15000 to 25000.

```python
# Hyperparameter set 2 chaning episodes = 25000
hyperparams_2 = {
    'alpha': 0.1,
    'gamma': 0.95,
    'epsilon': 1.0,
    'epsilon_decay': 0.9999,
    'epsilon_min': 0.01,
    'episodes': 25000,
    'max_steps': 1000
}

# Setting environment to stochastic mode
deterministic = False
env = Environment(0, 0, stochastic=(not deterministic))

print("Training Q-Learning agent (Stochastic Environment with Sensor-Augmented State) ...")
Q_stochastic_Q, rewards_stochastic_Q, eps_history_stochastic_Q = train_agent_stochastic_Q(env, hyperparams_2, hyperparams_name='hyperparams_2', render=False)
```

✓ 2m 9.0s

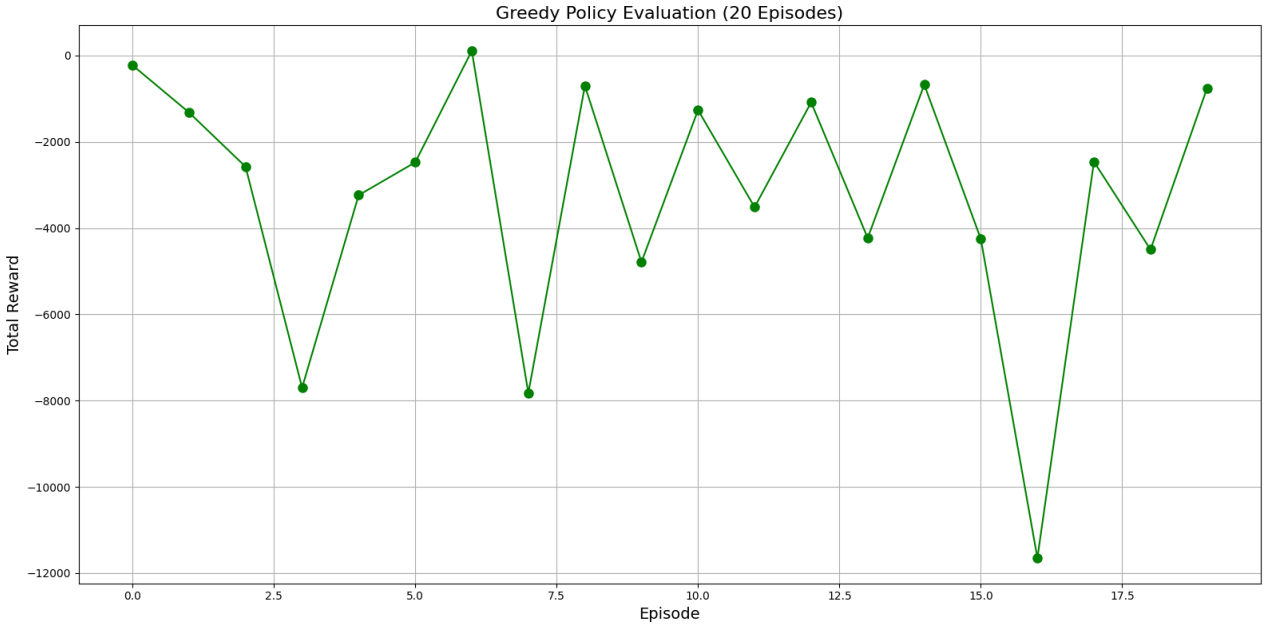
```
Training Q-Learning agent (Stochastic Environment with Sensor-Augmented State) ...
Episode 1000/25000 | Eps: 0.9048 | Success in last 1K: 1.9 %
Episode 2000/25000 | Eps: 0.8187 | Success in last 1K: 2.2 %
Episode 3000/25000 | Eps: 0.7408 | Success in last 1K: 5.9 %
Episode 4000/25000 | Eps: 0.6703 | Success in last 1K: 12.6 %
Episode 5000/25000 | Eps: 0.6065 | Success in last 1K: 24.2 %
Episode 6000/25000 | Eps: 0.5488 | Success in last 1K: 41.3 %
Episode 7000/25000 | Eps: 0.4966 | Success in last 1K: 51.3 %
Episode 8000/25000 | Eps: 0.4493 | Success in last 1K: 59.9 %
Episode 9000/25000 | Eps: 0.4066 | Success in last 1K: 70.8 %
Episode 10000/25000 | Eps: 0.3679 | Success in last 1K: 75.7 %
Episode 11000/25000 | Eps: 0.3329 | Success in last 1K: 79.4 %
Episode 12000/25000 | Eps: 0.3012 | Success in last 1K: 83.7 %
Episode 13000/25000 | Eps: 0.2725 | Success in last 1K: 88.0 %
Episode 14000/25000 | Eps: 0.2466 | Success in last 1K: 87.1 %
Episode 15000/25000 | Eps: 0.2231 | Success in last 1K: 86.4 %
Episode 16000/25000 | Eps: 0.2019 | Success in last 1K: 90.5 %
Episode 17000/25000 | Eps: 0.1827 | Success in last 1K: 91.2 %
Episode 18000/25000 | Eps: 0.1653 | Success in last 1K: 91.7 %
Episode 19000/25000 | Eps: 0.1496 | Success in last 1K: 91.1 %
Episode 20000/25000 | Eps: 0.1353 | Success in last 1K: 93.4 %
Episode 21000/25000 | Eps: 0.1224 | Success in last 1K: 91.6 %
Episode 22000/25000 | Eps: 0.1108 | Success in last 1K: 93.5 %
Episode 23000/25000 | Eps: 0.1002 | Success in last 1K: 94.3 %
Episode 24000/25000 | Eps: 0.0907 | Success in last 1K: 94.2 %
Task complete count: 16955
Episode 25000/25000 | Eps: 0.0821 | Success in last 1K: 93.6 %

Q-table saved to hyperparams_2_stochastic_q_table.pkl
```

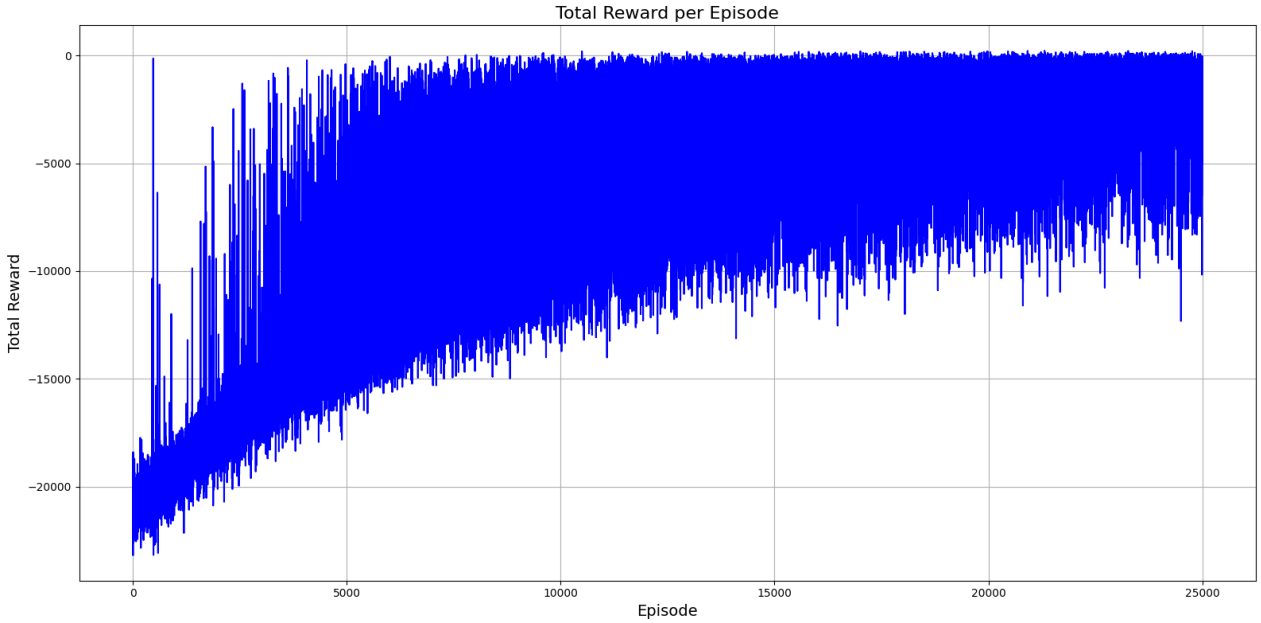
Out of 20, only 16 were successfully delivered.

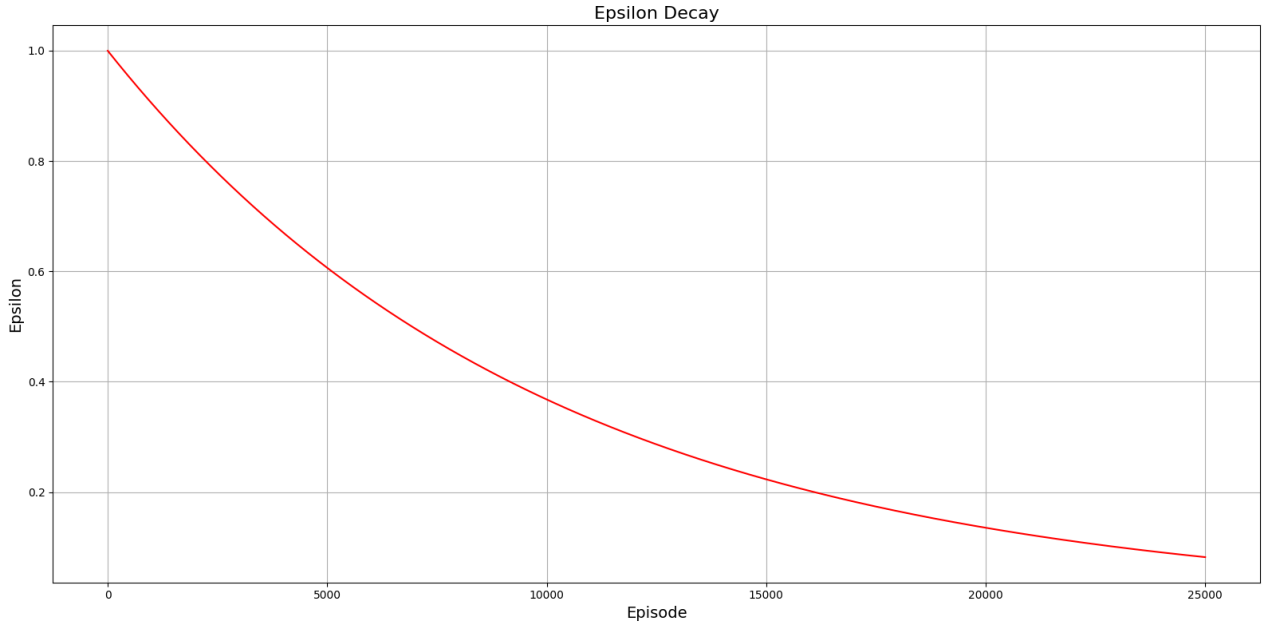
```python
# Evaluating the trained agent

deterministic = False
env = Environment(0, 0, stochastic=(not deterministic))  # Setting up the environment as stochastic

# Loading the trained Q-table
q_table_filename = "hyperparams_2_stochastic_q_table.pkl"

evaluation_rewards_stochastic_Q = evaluate_agent_stochastic_Q(env, q_table_filename=q_table_filename,
                                                              episodes=20, max_steps=1000, render=True)

# Plotting greedy policy evaluation
plt.figure(figsize=(16,8))
plt.plot(evaluation_rewards_stochastic_Q, marker='o', linestyle='-', color='green', markersize=8)
plt.xlabel("Episode", fontsize=14)
plt.ylabel("Total Reward", fontsize=14)
plt.title("Greedy Policy Evaluation (20 Episodes)", fontsize=16)
plt.grid(True)
plt.tight_layout()
plt.show()
```

✓ 1.0s

```
Evaluation Episode 1: Steps: 64 | Total Reward: -225
Evaluation Episode 2: Steps: 216 | Total Reward: -1320
Evaluation Episode 3: Steps: 209 | Total Reward: -2578
Evaluation Episode 4: Steps: 1000 | Total Reward: -7701
Evaluation Episode 5: Steps: 1000 | Total Reward: -3238
Evaluation Episode 6: Steps: 548 | Total Reward: -2477
Evaluation Episode 7: Steps: 67 | Total Reward: 108
Evaluation Episode 8: Steps: 821 | Total Reward: -7824
Evaluation Episode 9: Steps: 176 | Total Reward: -702
Evaluation Episode 10: Steps: 352 | Total Reward: -4790
Evaluation Episode 11: Steps: 335 | Total Reward: -1264
Evaluation Episode 12: Steps: 353 | Total Reward: -3516
Evaluation Episode 13: Steps: 314 | Total Reward: -1086
Evaluation Episode 14: Steps: 614 | Total Reward: -4232
Evaluation Episode 15: Steps: 178 | Total Reward: -676
Evaluation Episode 16: Steps: 1000 | Total Reward: -4254
Evaluation Episode 17: Steps: 1000 | Total Reward: -11649
Evaluation Episode 18: Steps: 370 | Total Reward: -2462
Evaluation Episode 19: Steps: 428 | Total Reward: -4499
Evaluation Episode 20: Steps: 171 | Total Reward: -755
Task complete count: 16
```


Greedy Policy Evaluation (20 Episodes)
Total Reward
Episode
0
−2000
−4000
−6000
−8000
−10000
−12000
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5


Total Reward per Episode
Total Reward
Episode
0
−5000
−10000
−15000
−20000
0
5000
10000
15000
20000
25000


Epsilon Decay
Epsilon
Episode
1.0
0.8
0.6
0.4
0.2
0
5000
10000
15000
20000
25000

Evaluation Episode 1: Steps: 64 | Total Reward: -116
Task complete count: 1

--- Evaluation Episode 1 starting ---

Drone moved to (5, 2). Step reward: -2

Evaluation Episode 1 - Step 1
Drone moved to (5, 3). Step reward: -2

Picked up package 1 for 25 reward
Drone moved to (5, 4). Step reward: -2

Drone moved to (4, 4). Step reward: -2

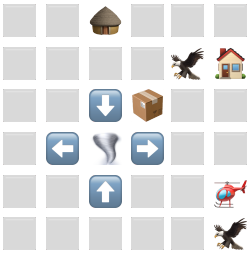
Drone moved to (4, 3). Step reward: -2

Drone moved to (4, 4). Step reward: -2

Drone moved to (4, 3). Step reward: -2

Drone moved to (4, 4). Step reward: -2

Drone moved to (4, 5). Step reward: -2



Drone moved to (4, 4). Step reward: -2

Evaluation Episode 1 - Step 11
Drone moved to (4, 3). Step reward: -2

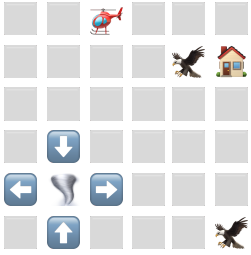
Drone moved to (3, 3). Step reward: -2

Drone moved to (2, 3). Step reward: -2

Picked up package 2 for 25 reward
Drone moved to (1, 3). Step reward: -2

Drone moved to (1, 2). Step reward: -2

Drone moved to (0, 2). Step reward: -2



Dropped package_1 incorrectly. Penalty -50
Delivered package_2 for +100 reward

Picked up package 1 for 25 reward
Drone moved to (0, 1). Step reward: -2

Evaluation Episode 1 - Step 21
Drone moved to (1, 1). Step reward: -2

Drone moved to (0, 1). Step reward: -2

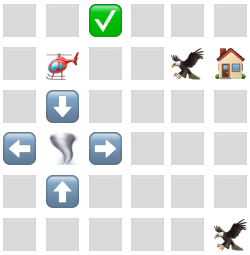
Drone moved to (1, 1). Step reward: -2

Drone moved to (0, 1). Step reward: -2

Drone moved to (1, 1). Step reward: -2

Drone moved to (0, 1). Step reward: -2

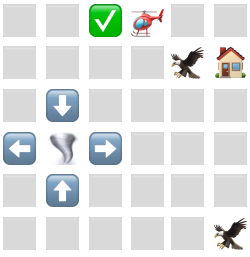
Drone moved to (1, 1). Step reward: -2



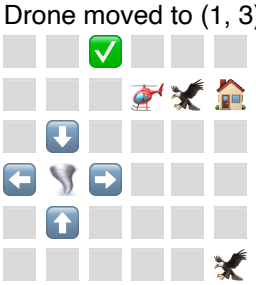
Drone moved to (1, 2). Step reward: -2

Drone moved to (0, 2). Step reward: -2

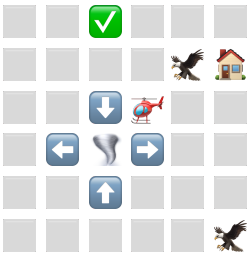
Drone moved to (0, 3). Step reward: -2



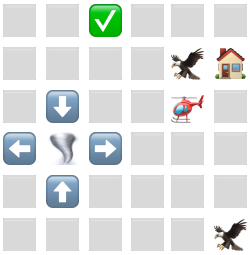
Evaluation Episode 1 - Step 31
Drone moved to (1, 3). Step reward: -2



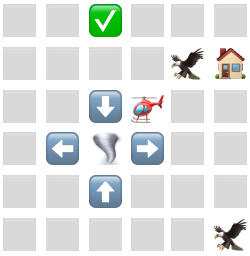
Drone moved to (2, 3). Step reward: -2



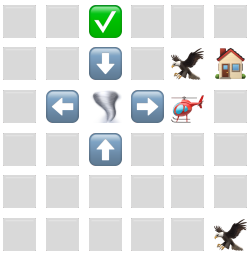
Drone moved to (2, 4). Step reward: -2



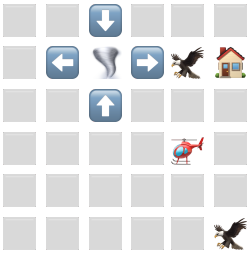
Drone moved to (2, 3). Step reward: -2



Drone moved to (2, 4). Step reward: -2

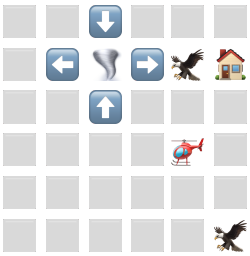


Drone moved to (3, 4). Step reward: -2



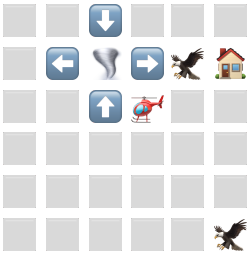
Drone moved to (2, 4). Step reward: -2

Drone moved to (3, 4). Step reward: -2

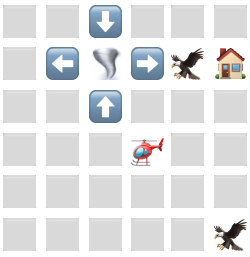


Drone moved to (2, 4). Step reward: -2

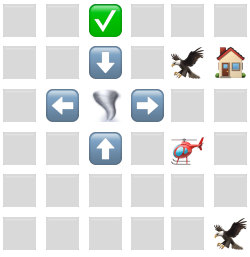
Drone moved to (2, 3). Step reward: -2



Evaluation Episode 1 - Step 41
Drone moved to (3, 3). Step reward: -2

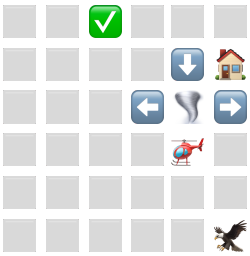


Drone moved to (3, 4). Step reward: -2



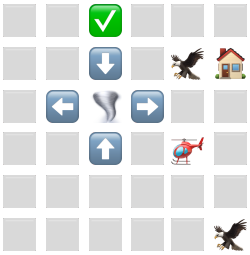
Drone moved to (4, 4). Step reward: -2

Drone moved to (3, 4). Step reward: -10

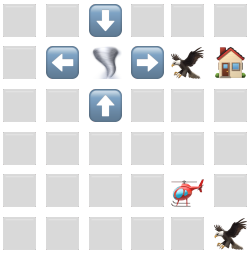


Drone moved to (3, 3). Step reward: -10

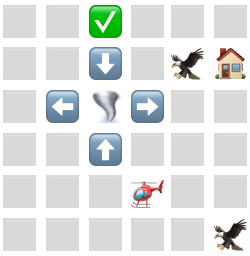
Drone moved to (3, 4). Step reward: -2



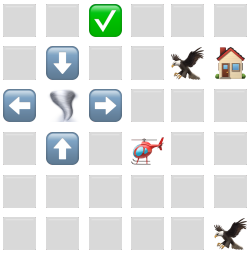
Drone moved to (4, 4). Step reward: -2



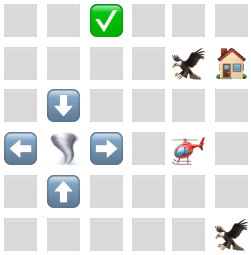
Drone moved to (4, 3). Step reward: -2



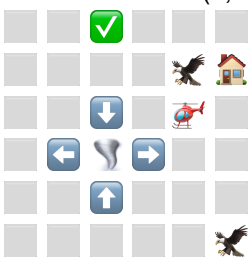
Drone moved to (3, 3). Step reward: -2



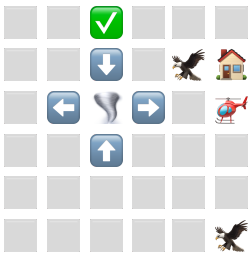
Drone moved to (3, 4). Step reward: -2



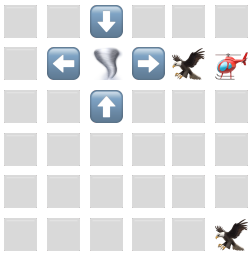
Evaluation Episode 1 - Step 51
Drone moved to (2, 4). Step reward: -2



Attempted pickup failed. Penalty -25
Attempted pickup failed. Penalty -25
Attempted pickup failed. Penalty -25
Action 4 repeated 4 times. Switching to a new action.
Attempted pickup failed. Penalty -25
Attempted pickup failed. Penalty -25
Attempted pickup failed. Penalty -25
Action 4 repeated 4 times. Switching to a new action.
Attempted pickup failed. Penalty -25
Attempted pickup failed. Penalty -25
Attempted pickup failed. Penalty -25
Evaluation Episode 1 - Step 61
Action 4 repeated 4 times. Switching to a new action.
Drone moved to (2, 5). Step reward: -2



Drone moved to (1, 5). Step reward: -2



Delivered package_1 for +100 reward
Task complete: All packages delivered 😎

 # Best hyperparameter set Rank 1
hyperparams_1 -> Changing episodes from 25000 to 50000.

```python
# Best hyperparameter set Rank 1
hyperparams_1 = {
    'alpha': 0.1,
    'gamma': 0.95,
    'epsilon': 1.0,
    'epsilon_decay': 0.9999,
    'epsilon_min': 0.01,
    'episodes': 50000,
    'max_steps': 1000
}

# Setting environment to stochastic mode
deterministic = False
env = Environment(0, 0, stochastic=(not deterministic))

print("Training Q-Learning agent (Stochastic Environment with Sensor-Augmented State) ...")
Q_stochastic_Q, rewards_stochastic_Q, eps_history_stochastic_Q = train_agent_stochastic_Q(env, hyperparams_1, hyperparams_name='hyperparams_1', render=False)
```

✓ 2m 59.1s

```
Training Q-Learning agent (Stochastic Environment with Sensor-Augmented State) ...
Episode 1000/50000 | Eps: 0.9048 | Success in last 1K: 0.9 %
Episode 2000/50000 | Eps: 0.8187 | Success in last 1K: 3.2 %
Episode 3000/50000 | Eps: 0.7408 | Success in last 1K: 6.6 %
Episode 4000/50000 | Eps: 0.6703 | Success in last 1K: 12.4 %
Episode 5000/50000 | Eps: 0.6065 | Success in last 1K: 27.7 %
Episode 6000/50000 | Eps: 0.5488 | Success in last 1K: 41.0 %
Episode 7000/50000 | Eps: 0.4966 | Success in last 1K: 53.2 %
Episode 8000/50000 | Eps: 0.4493 | Success in last 1K: 61.8 %
Episode 9000/50000 | Eps: 0.4066 | Success in last 1K: 70.0 %
Episode 10000/50000 | Eps: 0.3679 | Success in last 1K: 74.4 %
Episode 11000/50000 | Eps: 0.3329 | Success in last 1K: 81.3 %
Episode 12000/50000 | Eps: 0.3012 | Success in last 1K: 82.7 %
Episode 13000/50000 | Eps: 0.2725 | Success in last 1K: 86.8 %
Episode 14000/50000 | Eps: 0.2466 | Success in last 1K: 87.0 %
Episode 15000/50000 | Eps: 0.2231 | Success in last 1K: 88.2 %
Episode 16000/50000 | Eps: 0.2019 | Success in last 1K: 88.1 %
Episode 17000/50000 | Eps: 0.1827 | Success in last 1K: 91.3 %
Episode 18000/50000 | Eps: 0.1653 | Success in last 1K: 91.1 %
Episode 19000/50000 | Eps: 0.1496 | Success in last 1K: 91.9 %
Episode 20000/50000 | Eps: 0.1353 | Success in last 1K: 92.2 %
Episode 21000/50000 | Eps: 0.1224 | Success in last 1K: 92.8 %
Episode 22000/50000 | Eps: 0.1108 | Success in last 1K: 93.4 %
Episode 23000/50000 | Eps: 0.1002 | Success in last 1K: 94.5 %
Episode 24000/50000 | Eps: 0.0907 | Success in last 1K: 95.5 %
...
Task complete count: 41060
Episode 50000/50000 | Eps: 0.0100 | Success in last 1K: 96.6 %

Q-table saved to hyperparams_1_stochastic_q_table.pkl
```

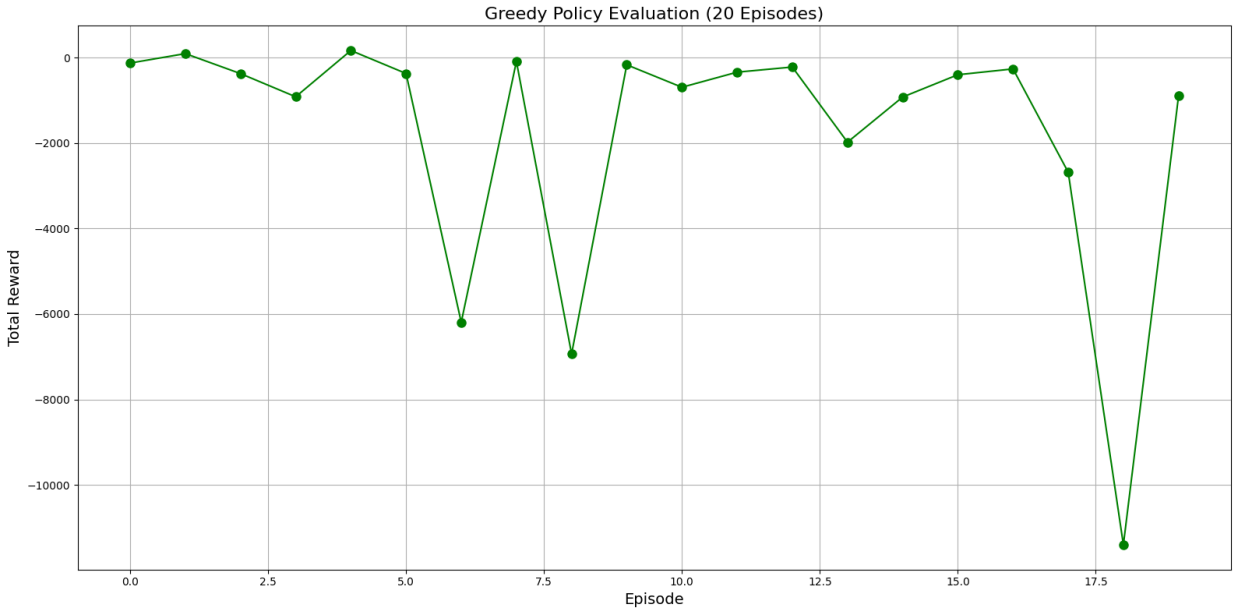
Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

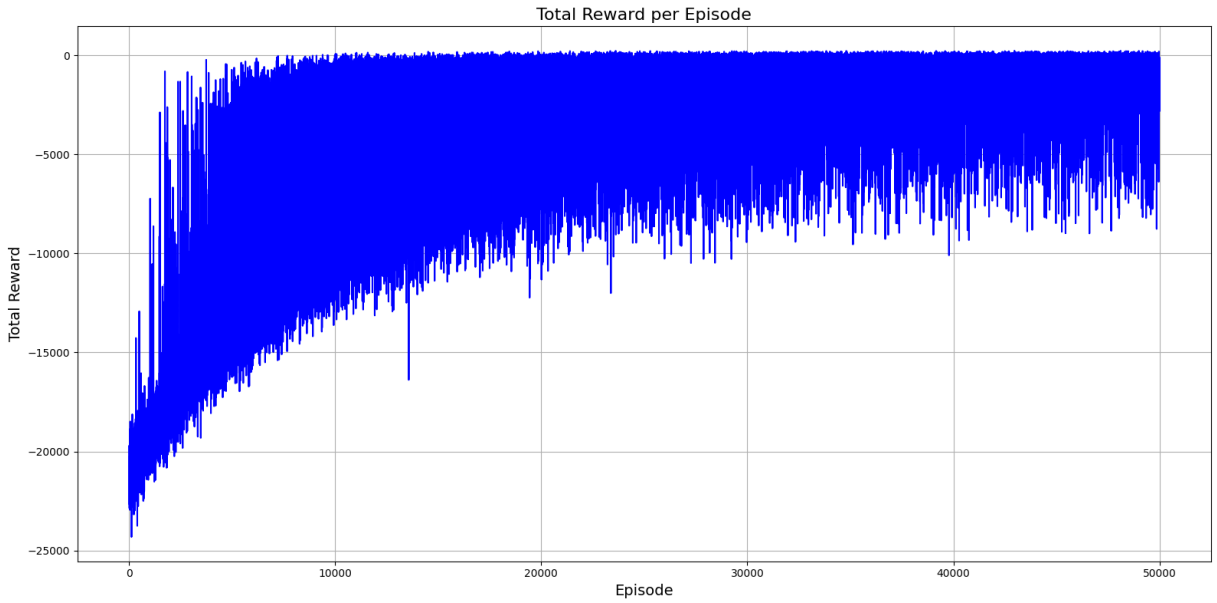
Out of 20, 18 were successfully delivered.

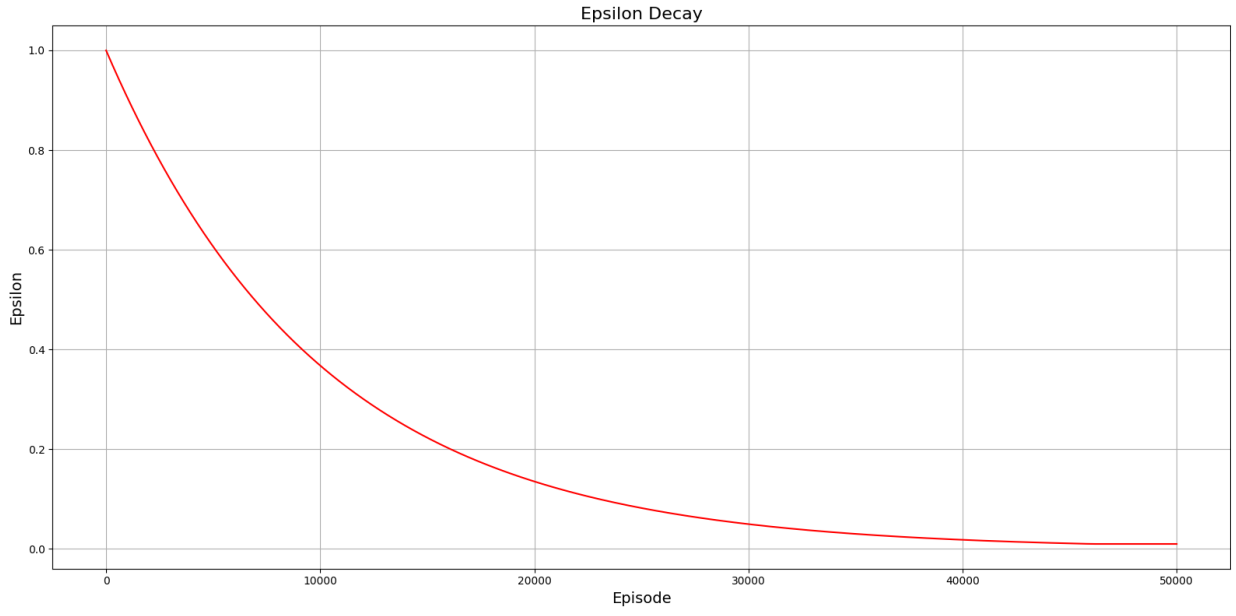
```python
# Evaluating the trained agent
deterministic = False
env = Environment(0, 0, stochastic=(not deterministic))  # Setting up the environment as stochastic

# Loading the trained Q-table
q_table_filename = "hyperparams_1_stochastic_q_table.pkl"


evaluation_rewards_stochastic_Q = evaluate_agent_stochastic_Q(env, q_table_filename=q_table_filename,
                                                              episodes=20, max_steps=1000, render=True)

# Plotting greedy policy evaluation
plt.figure(figsize=(16,8))
plt.plot(evaluation_rewards_stochastic_Q, marker='o', linestyle='-', color='green', markersize=8)
plt.xlabel("Episode", fontsize=14)
plt.ylabel("Total Reward", fontsize=14)
plt.title("Greedy Policy Evaluation (20 Episodes)", fontsize=16)
plt.grid(True)
plt.tight_layout()
plt.show()
```

✓ 1.1s

```
Evaluation Episode 1: Steps: 134 | Total Reward: -129
Evaluation Episode 2: Steps: 55 | Total Reward: 92
Evaluation Episode 3: Steps: 157 | Total Reward: -376
Evaluation Episode 4: Steps: 110 | Total Reward: -918
Evaluation Episode 5: Steps: 35 | Total Reward: 165
Evaluation Episode 6: Steps: 77 | Total Reward: -373
Evaluation Episode 7: Steps: 951 | Total Reward: -6194
Evaluation Episode 8: Steps: 43 | Total Reward: -91
Evaluation Episode 9: Steps: 1000 | Total Reward: -6934
Evaluation Episode 10: Steps: 121 | Total Reward: -170
Evaluation Episode 11: Steps: 304 | Total Reward: -695
Evaluation Episode 12: Steps: 195 | Total Reward: -342
Evaluation Episode 13: Steps: 124 | Total Reward: -222
Evaluation Episode 14: Steps: 463 | Total Reward: -1980
Evaluation Episode 15: Steps: 132 | Total Reward: -924
Evaluation Episode 16: Steps: 135 | Total Reward: -402
Evaluation Episode 17: Steps: 186 | Total Reward: -266
Evaluation Episode 18: Steps: 434 | Total Reward: -2683
Evaluation Episode 19: Steps: 1000 | Total Reward: -11400
Evaluation Episode 20: Steps: 115 | Total Reward: -899
Task complete count: 18
```


Greedy Policy Evaluation (20 Episodes)
Total Reward
Episode
0
−2000
−4000
−6000
−8000
−10000
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5


Total Reward per Episode
Total Reward
Episode
0
−5000
−10000
−15000
−20000
−25000
0
10000
20000
30000
40000
50000


Epsilon Decay
Epsilon
Episode
1.0
0.8
0.6
0.4
0.2
0.0
0
10000
20000
30000
40000
50000

Evaluation Episode 1: Steps: 29 | Total Reward: 176
Task complete count: 1

--- Evaluation Episode 1 starting ---

Drone moved to (3, 2). Step reward: -2

Evaluation Episode 1 - Step 1
Drone moved to (2, 2). Step reward: -2

Picked up package 1 for 25 reward
Drone moved to (2, 1). Step reward: -2

Drone moved to (1, 1). Step reward: -2

Delivered package_1 for +100 reward
Drone moved to (0, 1). Step reward: -2

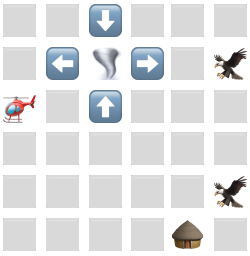
Picked up package 2 for 25 reward
Drone moved to (0, 2). Step reward: -2

Drone moved to (1, 2). Step reward: -2

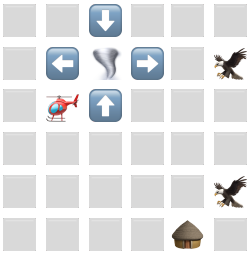
Drone moved to (2, 2). Step reward: -10

Evaluation Episode 1 - Step 11
Drone moved to (2, 1). Step reward: -10

Drone moved to (2, 0). Step reward: -2



Drone moved to (2, 1). Step reward: -2



Drone moved to (3, 1). Step reward: -2

Drone moved to (4, 1). Step reward: -2

Drone moved to (5, 1). Step reward: -2

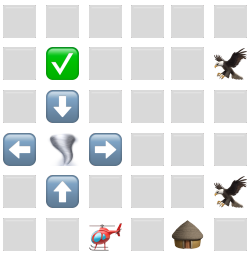
Drone moved to (5, 2). Step reward: -2

Drone moved to (5, 1). Step reward: -2

Drone moved to (5, 2). Step reward: -2

Drone moved to (5, 1). Step reward: -2

Evaluation Episode 1 - Step 21
Drone moved to (5, 2). Step reward: -2



Drone moved to (4, 2). Step reward: -2

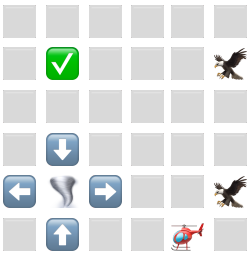
Drone moved to (4, 1). Step reward: -10

Drone moved to (4, 2). Step reward: -2

Drone moved to (4, 3). Step reward: -2

Drone moved to (4, 4). Step reward: -2

Drone moved to (5, 4). Step reward: -2



Delivered package_2 for +100 reward
Task complete: All packages delivered 😎

Applying Double Q-learning in a stochastic environment.

```python
# Best hyperparams 1
hyperparams = {
                'alpha': 0.1,
                'gamma': 0.95,
                'epsilon': 1.0,
                'epsilon_decay': 0.9999,
                'epsilon_min': 0.01,
                'episodes': 50000,
                'max_steps': 1000
            }


deterministic = False  # Set to False for stochastic mode.
env = Environment(0, 0, stochastic=(not deterministic))


print("Training Q-Learning agent (Stochastic Environment with Sensor-Augmented State) ...")
(Q1_stochastic_DoubleQ, Q2_stochastic_DoubleQ), rewards_stochastic_DoubleQ, eps_history_stochastic_DoubleQ = train_agent_stochastic_DoubleQ(env, hyperparams, render=False)
```

✓ 3m 19.8s

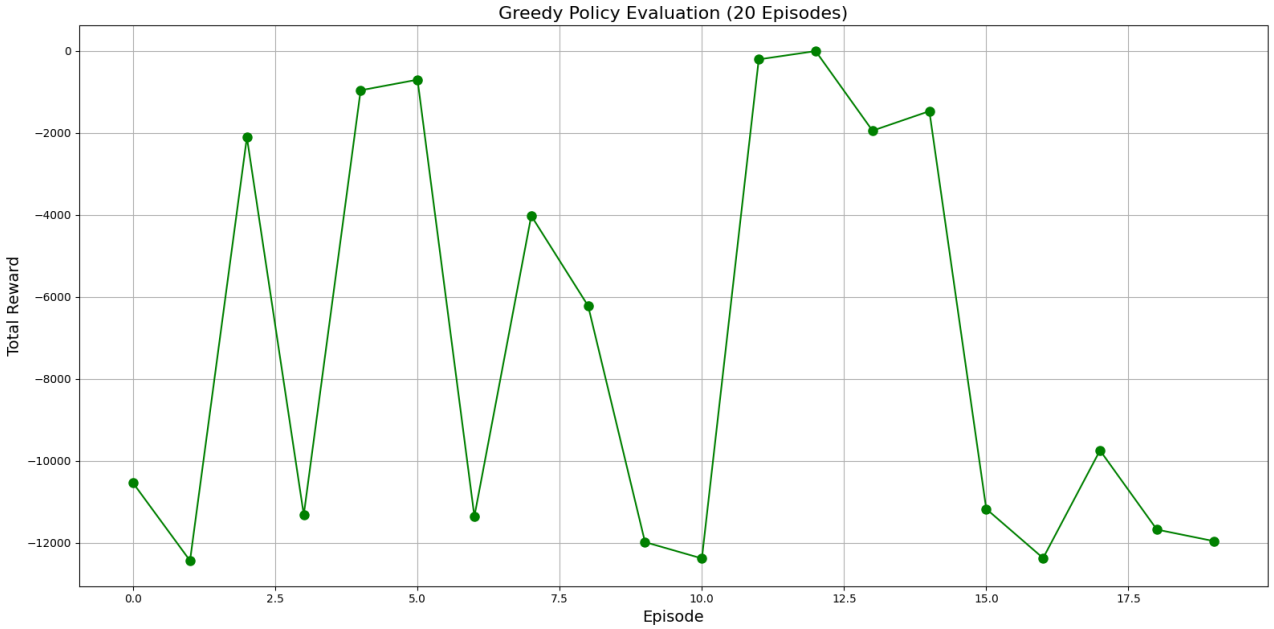
```
Training Q-Learning agent (Stochastic Environment with Sensor-Augmented State) ...
Episode 1000/50000 | Epsilon: 0.9048 | Success in last 1K: 0.8 %
Episode 2000/50000 | Epsilon: 0.8187 | Success in last 1K: 1.7 %
Episode 3000/50000 | Epsilon: 0.7408 | Success in last 1K: 4.0 %
Episode 4000/50000 | Epsilon: 0.6703 | Success in last 1K: 9.9 %
Episode 5000/50000 | Epsilon: 0.6065 | Success in last 1K: 19.4 %
Episode 6000/50000 | Epsilon: 0.5488 | Success in last 1K: 35.5 %
Episode 7000/50000 | Epsilon: 0.4966 | Success in last 1K: 48.0 %
Episode 8000/50000 | Epsilon: 0.4493 | Success in last 1K: 59.3 %
Episode 9000/50000 | Epsilon: 0.4066 | Success in last 1K: 69.8 %
Episode 10000/50000 | Epsilon: 0.3679 | Success in last 1K: 73.5 %
Episode 11000/50000 | Epsilon: 0.3329 | Success in last 1K: 79.4 %
Episode 12000/50000 | Epsilon: 0.3012 | Success in last 1K: 83.3 %
Episode 13000/50000 | Epsilon: 0.2725 | Success in last 1K: 84.1 %
Episode 14000/50000 | Epsilon: 0.2466 | Success in last 1K: 87.4 %
Episode 15000/50000 | Epsilon: 0.2231 | Success in last 1K: 89.7 %
Episode 16000/50000 | Epsilon: 0.2019 | Success in last 1K: 90.1 %
Episode 17000/50000 | Epsilon: 0.1827 | Success in last 1K: 89.2 %
Episode 18000/50000 | Epsilon: 0.1653 | Success in last 1K: 90.4 %
Episode 19000/50000 | Epsilon: 0.1496 | Success in last 1K: 92.3 %
Episode 20000/50000 | Epsilon: 0.1353 | Success in last 1K: 91.9 %
Episode 21000/50000 | Epsilon: 0.1224 | Success in last 1K: 93.2 %
Episode 22000/50000 | Epsilon: 0.1108 | Success in last 1K: 94.6 %
Episode 23000/50000 | Epsilon: 0.1002 | Success in last 1K: 93.3 %
Episode 24000/50000 | Epsilon: 0.0907 | Success in last 1K: 94.5 %
...
Task complete count: 40700
Episode 50000/50000 | Epsilon: 0.0100 | Success in last 1K: 96.1 %

Double Q-tables saved to stochastic_double_q_table.pkl
Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...
```

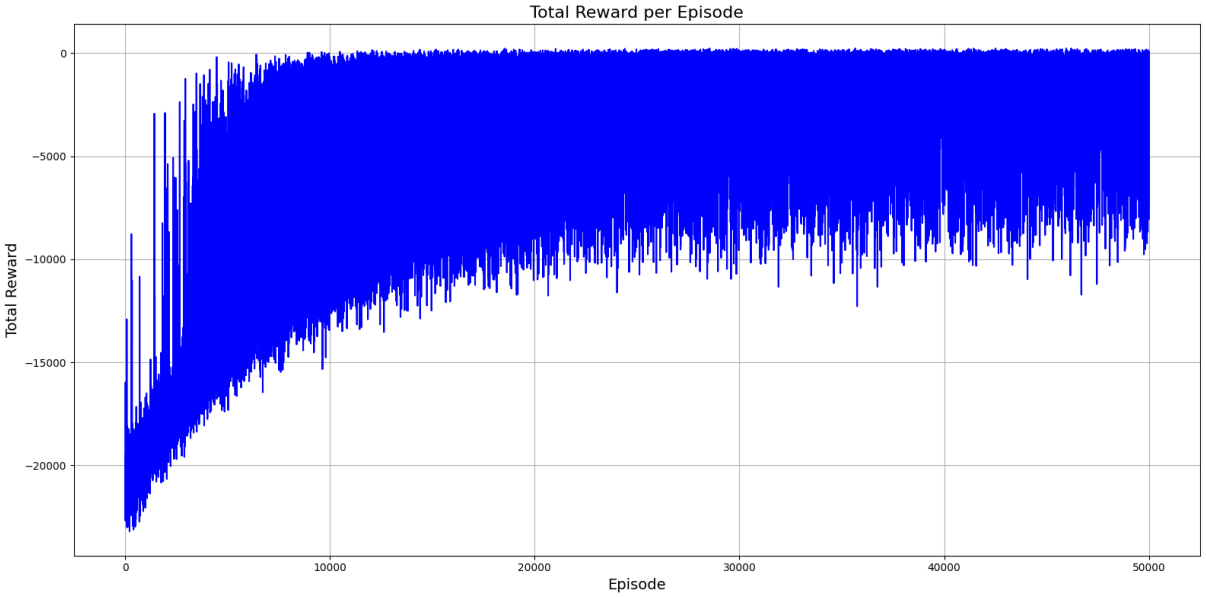
Out of 20, only 9 were successfully delivered.

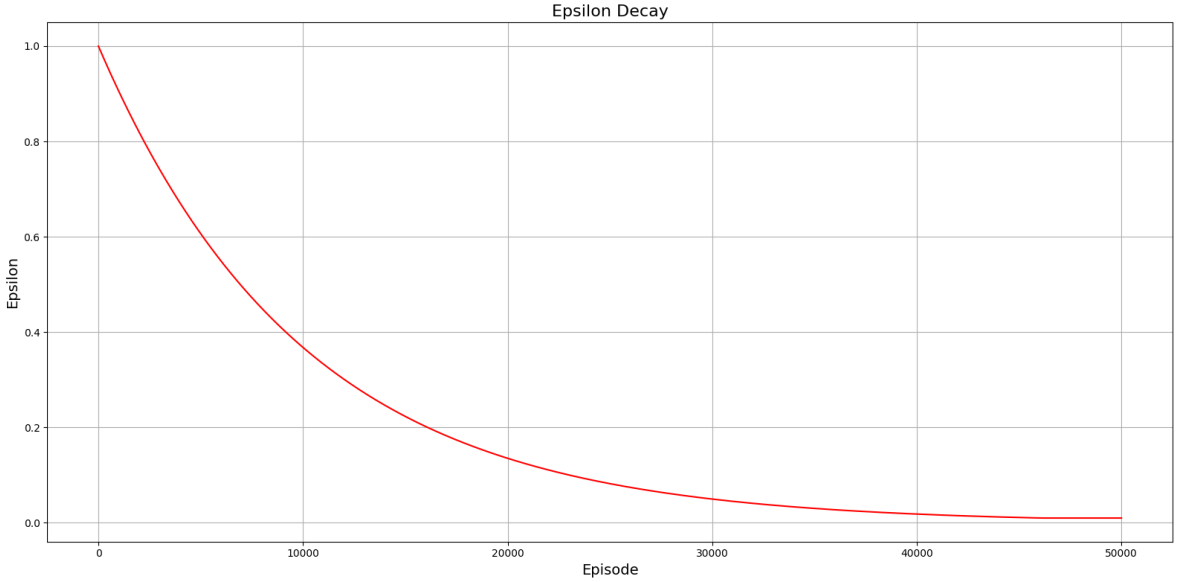
```python
# Evaluating the trained agent using a greedy policy

deterministic = False  # Set to False for stochastic mode.
env = Environment(0, 0, stochastic=(not deterministic))

q_table_filename = "stochastic_double_q_table.pkl"

evaluation_rewards_stochastic_DoubleQ = evaluate_agent_stochastic_DoubleQ(env,
                                                q_table_filename=q_table_filename,
                                                episodes=20, max_steps=1000, render=True)

# Plotting greedy policy evaluation results
plt.figure(figsize=(16,8))
plt.plot(evaluation_rewards_stochastic_DoubleQ, marker='o', linestyle='-', color='green', markersize=8)
plt.xlabel("Episode", fontsize=14)
plt.ylabel("Total Reward", fontsize=14)
plt.title("Greedy Policy Evaluation (20 Episodes)", fontsize=16)
plt.grid(True)
plt.tight_layout()
plt.show()
```

✓ 1.7s

```
Evaluation Episode 1: Steps: 1000 | Total Reward: -10544
Evaluation Episode 2: Steps: 1000 | Total Reward: -12437
Evaluation Episode 3: Steps: 147 | Total Reward: -2109
Evaluation Episode 4: Steps: 1000 | Total Reward: -11318
Evaluation Episode 5: Steps: 227 | Total Reward: -962
Evaluation Episode 6: Steps: 149 | Total Reward: -703
Evaluation Episode 7: Steps: 1000 | Total Reward: -11357
Evaluation Episode 8: Steps: 258 | Total Reward: -4023
Evaluation Episode 9: Steps: 516 | Total Reward: -6222
Evaluation Episode 10: Steps: 1000 | Total Reward: -11984
Evaluation Episode 11: Steps: 1000 | Total Reward: -12382
Evaluation Episode 12: Steps: 68 | Total Reward: -208
Evaluation Episode 13: Steps: 61 | Total Reward: -5
Evaluation Episode 14: Steps: 201 | Total Reward: -1947
Evaluation Episode 15: Steps: 167 | Total Reward: -1471
Evaluation Episode 16: Steps: 1000 | Total Reward: -11174
Evaluation Episode 17: Steps: 1000 | Total Reward: -12374
Evaluation Episode 18: Steps: 1000 | Total Reward: -9747
Evaluation Episode 19: Steps: 1000 | Total Reward: -11678
Evaluation Episode 20: Steps: 1000 | Total Reward: -11959
Task complete count: 9
```


Greedy Policy Evaluation (20 Episodes)
Total Reward
Episode
0
−2000
−4000
−6000
−8000
−10000
−12000
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5


Total Reward per Episode
Total Reward
Episode
0
−5000
−10000
−15000
−20000
0
10000
20000
30000
40000
50000


Epsilon Decay
Epsilon
Episode
1.0
0.8
0.6
0.4
0.2
0.0
0
10000
20000
30000
40000
50000

Evaluation Episode 1: Steps: 47 | Total Reward: -47
Task complete count: 1

--- Evaluation Episode 1 starting ---

Drone moved to (4, 1). Step reward: -2

Evaluation Episode 1 - Step 1
Drone moved to (3, 1). Step reward: -2

Drone moved to (2, 1). Step reward: -2

Drone moved to (1, 1). Step reward: -2

Action 0 repeated 4 times. Switching to a new action.
Attempted dropoff failed (no package carried). Penalty -50
Drone moved to (0, 1). Step reward: -2

Drone moved to (0, 0). Step reward: -2

Picked up package 2 for 25 reward
Drone moved to (0, 1). Step reward: -2

Drone moved to (1, 1). Step reward: -2

Delivered package_2 for +100 reward
Evaluation Episode 1 - Step 11
Drone moved to (1, 0). Step reward: -2

Drone moved to (1, 1). Step reward: -2

Drone moved to (1, 0). Step reward: -2

Drone moved to (1, 1). Step reward: -2

Drone moved to (1, 0). Step reward: -2

Drone moved to (1, 1). Step reward: -2

Drone moved to (1, 0). Step reward: -2

Drone moved to (1, 1). Step reward: -2

Drone moved to (0, 1). Step reward: -2

Drone moved to (0, 0). Step reward: -2

Evaluation Episode 1 - Step 21
Drone moved to (0, 1). Step reward: -2

Drone moved to (0, 0). Step reward: -2

Drone moved to (0, 1). Step reward: -2

Drone moved to (0, 0). Step reward: -2

Drone moved to (0, 1). Step reward: -2

Drone moved to (0, 0). Step reward: -2

Drone moved to (0, 1). Step reward: -2

Drone moved to (0, 0). Step reward: -2

Drone moved to (0, 1). Step reward: -2

Drone moved to (0, 2). Step reward: -2

Evaluation Episode 1 - Step 31
Drone moved to (1, 2). Step reward: -2

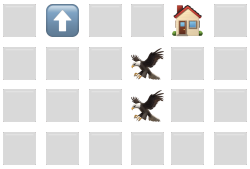
Drone moved to (1, 3). Step reward: -2

Drone moved to (2, 3). Step reward: -2

Drone moved to (2, 4). Step reward: -2

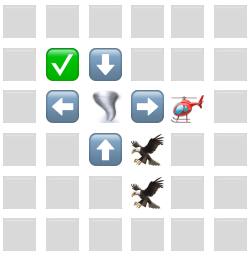
Attempted pickup failed. Penalty -25
Attempted pickup failed. Penalty -25
Attempted pickup failed. Penalty -25
Action 4 repeated 4 times. Switching to a new action.
Attempted dropoff failed (no package carried). Penalty -50
Attempted pickup failed. Penalty -25
Attempted pickup failed. Penalty -25
Evaluation Episode 1 - Step 41
Action 4 repeated 4 times. Switching to a new action.
Drone moved to (1, 4). Step reward: -2

Drone moved to (1, 5). Step reward: -2



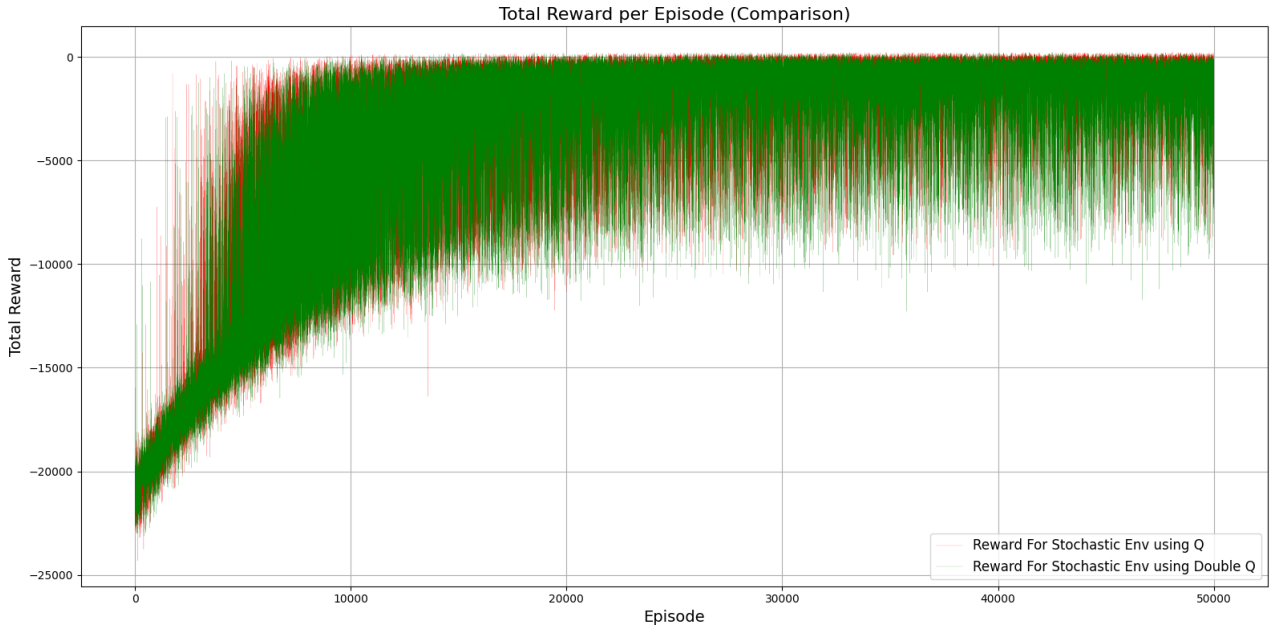
Picked up package 1 for 25 reward
Drone moved to (1, 4). Step reward: -2

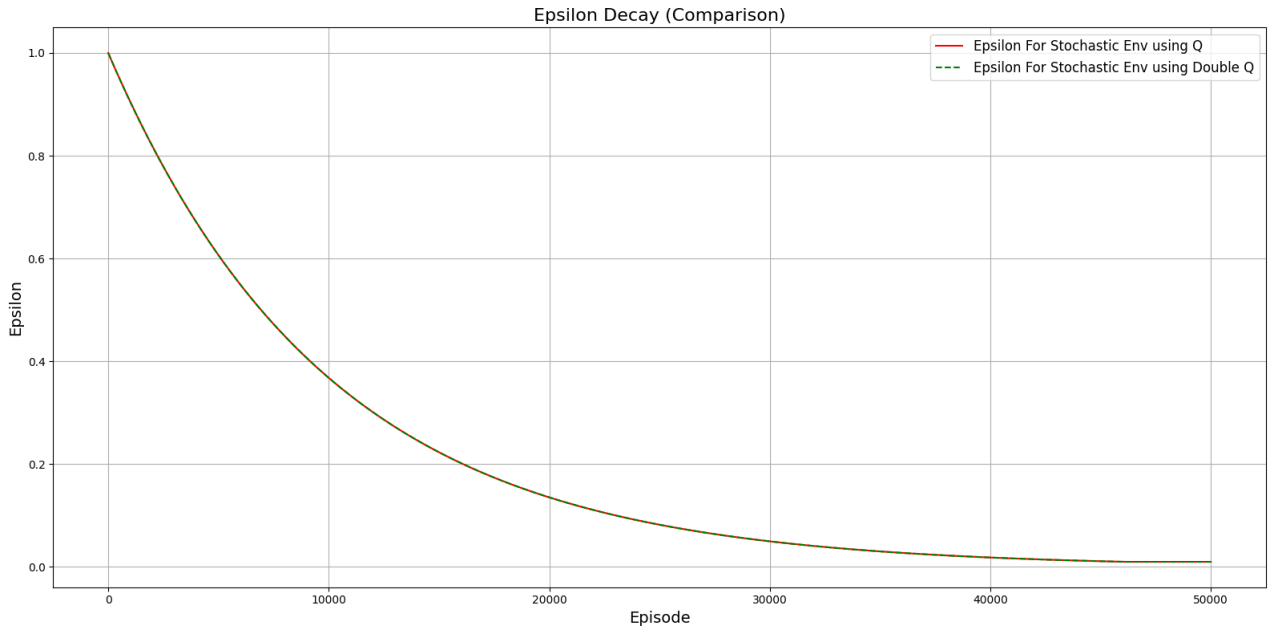
Drone moved to (2, 4). Step reward: -2

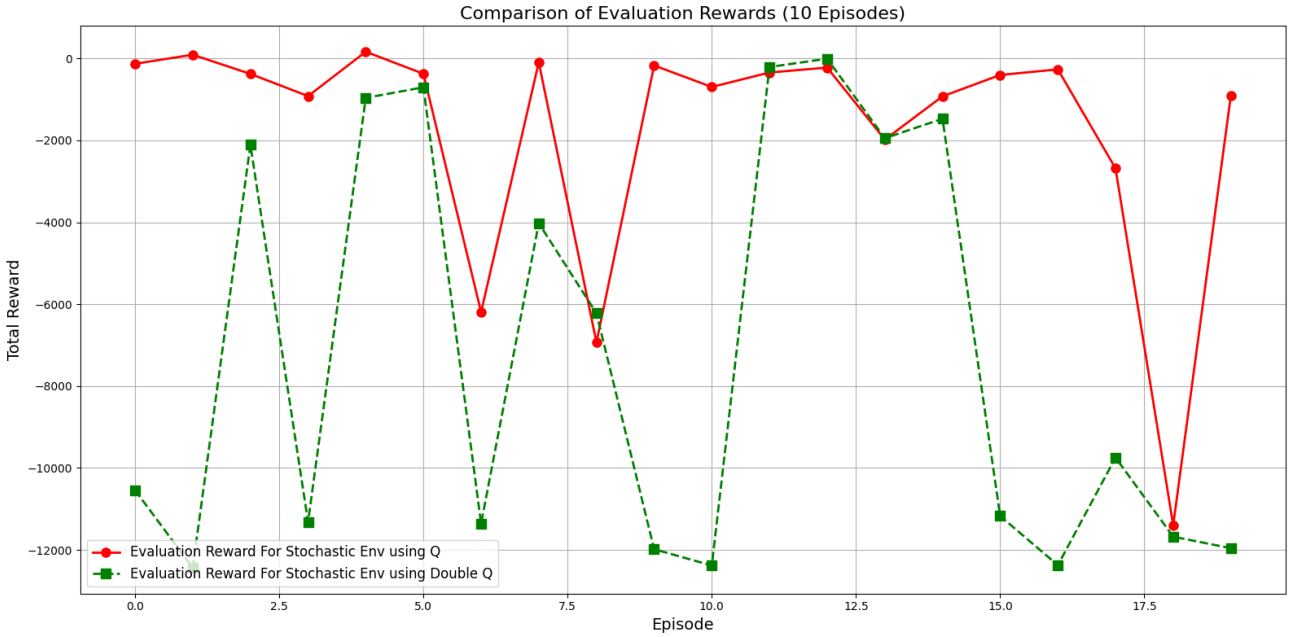


Delivered package_1 for +100 reward
Task complete: All packages delivered 😎

Comparison of Q and Double Q in Stochastic Environment







Tabular Methods Explanation:

1. Q-Learning

Q-Learning is a model-free reinforcement learning algorithm that updates the Q-values based on the maximum possible reward from the next state. It follows the Bellman equation and updates the Q-values using the following formula:
Q-Learning Update Rule **(Reference: Used from class notes)**

Key Features of Q-Learning

Q-learning (off-policy TD control) for estimating $\pi \approx \pi_*$

Algorithm parameters: step size $\alpha \in (0, 1]$, small $\varepsilon > 0$
Initialize $Q(s, a)$, for all $s \in \mathcal{S}^+, a \in \mathcal{A}(s)$, arbitrarily except that $Q(terminal, \cdot) = 0$

Loop for each episode:
  Initialize S
  Loop for each step of episode:
    Choose A from S using policy derived from Q (e.g., ε-greedy)
    Take action A, observe R, S'
    $Q(S, A) \leftarrow Q(S, A) + \alpha\big[R + \gamma \max_a Q(S', a) - Q(S, A)\big]$
    $S \leftarrow S'$
  until S is terminal

Target
Prediction
Immediate Reward
loss

- Off-policy: Learns from the max possible future reward, not necessarily the one taken.
- Converges to the optimal policy when trained long enough.
- Simple to implement and works well in deterministic environments.

2. Double Q-Learning

Double Q-learning is an improvement over Q-learning that reduces overestimation of Q-values, which can happen in standard Q-learning. Instead of using a single Q-table, Double Q-learning uses two Q-tables (Q1 and Q2) and randomly updates one at a time.
Double Q-Learning Update Rule **(Reference: Used from class notes)**

Double Q-learning, for estimating $Q_1 \approx Q_2 \approx q_*$

Algorithm parameters: step size $\alpha \in (0, 1]$, small $\varepsilon > 0$
Initialize $Q_1(s, a)$ and $Q_2(s, a)$, for all $s \in \mathcal{S}^+, a \in \mathcal{A}(s)$, such that $Q(terminal, \cdot) = 0$

Loop for each episode:
  Initialize S
  Loop for each step of episode:
    Choose A from S using the policy ε-greedy in $Q_1 + Q_2$
    Take action A, observe R, S'
    With 0.5 probabilility:
      $Q_1(S, A) \leftarrow Q_1(S, A) + \alpha\Big(R + \gamma Q_2\big(S', \operatorname{argmax}_a Q_1(S', a)\big) - Q_1(S, A)\Big)$
    else:
      $Q_2(S, A) \leftarrow Q_2(S, A) + \alpha\Big(R + \gamma Q_1\big(S', \operatorname{argmax}_a Q_2(S', a)\big) - Q_2(S, A)\Big)$
    $S \leftarrow S'$
  until S is terminal

Key Features of Double Q-Learning

- Reduces Q-value overestimation (common in standard Q-learning).
- More stable learning in stochastic environments.
- Uses two Q-tables, updating one at a time for better accuracy.

Tabular Methods Used: Q-Learning and Double Q-Learning

To solve the drone delivery problem, I used Q-learning and Double Q-learning, both of which are tabular reinforcement learning methods. These methods use a Q-table to store and update values for each state-action pair, helping the agent learn the best possible actions to take in different situations.

Comparison of Q-Learning and Double Q-Learning in a Deterministic Environment

In a deterministic environment, both Q-learning and Double Q-learning performed equally well. Even though I trained Double Q-learning with only half the steps (5000) compared to Q-learning (10000), it still achieved:

- Same final reward (185)
- Same number of steps (25) to reach the goal
- Same sequence of actions

Key Observations

Q-learning worked well because the environment is fully predictable, meaning the agent always transitions to the expected next state.

Double Q-learning showed no additional improvement because its main advantage—reducing overestimation errors—only matters in stochastic environments where transitions and rewards have randomness.

Since both algorithms found the same optimal policy, Q-learning was already sufficient for solving this task.

Comparison of Q-Learning and Double Q-Learning in a Stochastic Environment

In the stochastic environment, Q-learning outperformed Double Q-learning using the following hyperparameters:

```
hyperparams = {
    'alpha': 0.1,        # Learning rate
    'gamma': 0.95,       # Discount factor
    'epsilon': 1.0,      # Initial exploration rate
    'epsilon_decay': 0.9999,  # Decay rate for epsilon
    'epsilon_min': 0.01, # Minimum epsilon value
    'episodes': 50000,   # Number of training episodes
    'max_steps': 1000    # Maximum steps per episode
}
```

Performance Results (20 Evaluation Episodes)

Algorithm	Successful Deliveries (out of 20)
Q-Learning	**18**
Double Q-Learning	**9**

Even though both algorithms showed a 90%+ success rate in the last 1000 training episodes, Double Q-learning failed to maintain this performance during greedy testing.

Possible Reasons for Double Q-Learning's Poor Performance

Insufficient Training Episodes
- Since Double Q-learning updates two separate Q-tables, it might need more training episodes than Q-learning to converge properly.

High Variance in Updates
- The random selection of Q1 or Q2 for updates may have led to inconsistencies in learning in a stochastic setting.

Key Observations
- Q-learning learned faster and performed better in stochastic conditions.
- Double Q-learning failed to generalize well during testing, possibly due to insufficient training.

Final Thoughts

Algorithm	Best for	Limitations
Q-Learning	**Deterministic or mildly stochastic environments**	**Overestimates Q-values in highly stochastic settings**
Double Q-Learning	**Highly stochastic environments** where random transitions and rewards affect learning	**Requires more training to stabilize**

In my experiment:

- Q-learning was more effective for both deterministic and stochastic environments.
- Double Q-learning may require more episodes or additional tuning to perform well in stochastic settings.
- For future improvements, I could try increasing the training episodes for Double Q-learning and adjusting its learning rate to see if it can perform better in stochastic environments.

My Approach to Designing a Good Reward System

To train the drone efficiently, I designed a reward system that encourages correct actions and penalizes mistakes. Below is the complete reward system I implemented:

Event	Reward/Penalty
Successfully delivering a package	**100**
Picking up a package	**25**
Entering a no-fly zone (tornado)	**-100**
Getting hit by a bird	**-50**
Being pushed by the wind (any direction: up, down, left, right)	**-10**
Taking a step (movement cost)	**-2**
Dropping a package at the wrong location	**-50**
Attempting to pick up when not at a package location	**-25**
Moving out of the grid (invalid move)	**-25**
Repeating the same move unnecessarily	**-50**

Encouraging Task Completion

- The agent gets +100 when it successfully delivers a package.
- It also earns +25 for picking up a package, so it learns to grab packages efficiently.

Discouraging Mistakes

- Hitting obstacles like a tornado (-100) or birds (-50) teaches the drone to avoid dangerous areas.
- Wind pushes (-10) make the agent careful around no-fly zones.

- If it tries to drop a package in the wrong place, it loses -50 to encourage correct delivery.
- Repetition penalty (-50) prevents the agent from getting stuck in loops, forcing it to explore better routes.

Avoiding Random and Inefficient Movements

- Each step costs -2, so the agent learns to take the shortest possible route.
- Trying to pick up a package at the wrong location results in -25, preventing random pickup attempts.
- Going out of bounds results in -25, so the agent stays within the grid.
- If the agent keeps repeating the same move without progress, it gets a -50 penalty, helping it avoid getting stuck.

What I Learned from Testing Different Rewards

- At first, when I only rewarded package delivery, the drone struggled to learn efficient paths. It took unnecessary steps, often crashed into obstacles, and sometimes got stuck in loops.
- After adding step penalties (-2) and mistake penalties (-50 for repetition, -25 for wrong actions, etc.), the agent started moving smarter and more efficiently. It avoided obstacles, took better routes, and completed tasks faster without getting stuck.
- With this balanced reward system, the agent now learns proper movement, avoids risks, and successfully delivers packages while minimizing repeated actions.

Part 3 [Total: 30 points] - Solve Stock Trading Environment

For this task, I applied Q-learning with a state space of 4 observations using the following hyperparameters:

```
{
    'alpha': 0.005,
    'gamma': 0.95,
    'epsilon': 1.0,
    'epsilon_decay': 0.9995,
    'epsilon_min': 0.01,
    'episodes': 10,000
}
```

I experimented with various hyperparameter settings and observed a maximum reward of 79,672 during training. I consistently saved the best Q-table in a pickle file.

Additionally, I tested different values for the number of days considered, including 10 days, 5 days, and 2 days. The highest reward during testing was achieved when considering 2 days, so I finalized my training and testing parameters accordingly.

```
--- Training Episode 1 starting ---
Episode 1/10000 | Steps: 408 | Total Reward: -1138.6196516967636 | Epsilon: 1.000
New best Q-table saved with reward: -1138.620

--- Training Episode 2 starting ---
Episode 2/10000 | Steps: 408 | Total Reward: -954.6080502884979 | Epsilon: 0.999
New best Q-table saved with reward: -954.608

--- Training Episode 3 starting ---
Episode 3/10000 | Steps: 408 | Total Reward: -1136.2194049484046 | Epsilon: 0.999
Episode reward -1136.219 did not beat best reward -954.608

--- Training Episode 4 starting ---
Episode 4/10000 | Steps: 408 | Total Reward: -1091.3766035410185 | Epsilon: 0.998
Episode reward -1091.377 did not beat best reward -954.608

--- Training Episode 5 starting ---
Episode 5/10000 | Steps: 408 | Total Reward: -1186.3165176901214 | Epsilon: 0.998
Episode reward -1186.317 did not beat best reward -954.608

--- Training Episode 6 starting ---
Episode 6/10000 | Steps: 408 | Total Reward: -1269.3419368942045 | Epsilon: 0.997
Episode reward -1269.342 did not beat best reward -954.608
...

--- Training Episode 10000 starting ---
Episode 10000/10000 | Steps: 408 | Total Reward: 41900.917826033634 | Epsilon: 0.010
Episode reward 41900.918 did not beat best reward 79672.953
```

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

```
Initial Q-table (empty entries as I am using a dictionary for easier implementation):

Trained Q-table (for visited states):

0 [1302.99924443  383.73171147  372.08989362]
1 [1176.98695943 1172.48736257 1740.30077686]
2 [965.73580584 341.18676392 359.96181981]
3 [ 998.73571359  728.23741122 1471.30100796]
```