Московский государственный технический университет им. Н.Э. Баумана

Факультет «Головной учебно-исследовательский и методический центр

профессиональной реабилитации лиц с ограниченными возможностями здоровья (инвалидов)»

Кафедра «Системы обработки информации и управления»



**Лабораторная работа 5**

**по дисциплине «Методы машинного обучения в АСОИУ»**

**"** **Обучение на основе временных различий"**

СТУДЕНТ:

студент группы ИУ5Ц-21М

Москалик А.А.

ПРЕПОДАВАТЕЛЬ:

Гапанюк Ю.Е.

Москва, 2024

# 

# Цель лабораторной работы: ознакомление с базовыми методами обучения с подкреплением на основе временных различий.

# Задание: на основе рассмотренного на лекции примера реализуйте следующие алгоритмы:

# - SARSA

# - Q-обучение

# - Двойное Q-обучение

# для любой среды обучения с подкреплением (кроме рассмотренной на лекции среды Toy Text / Frozen Lake) из библиотеки Gym (или аналогичной библиотеки).

# Ход работы

# SARSA

# Инициализация среды

# 

# Дискретизация состояния

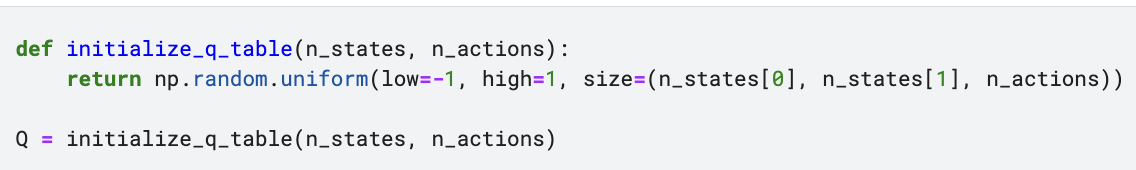
# Дискретизация состояния — это процесс преобразования непрерывного пространства состояний в дискретное пространство состояний. Это делается для того, чтобы уменьшить количество возможных состояний до управляемого количества, которое можно эффективно использовать в алгоритмах обучения с подкреплением, таких как SARSA. Дискретизация позволяет агенту работать с конечным набором состояний, что упрощает процесс обновления и хранения значений функции ценности (Q-значений).

# 

# Инициализация Q-таблицы

# Инициализация Q-таблицы — это процесс создания и начальной настройки таблицы Q-значений, которая используется в алгоритмах обучения с подкреплением. Q-таблица представляет собой двумерный массив (или трехмерный массив, если дискретизированное пространство состояний многомерное), где каждая ячейка хранит оценку полезности или ожидаемого вознаграждения для каждой комбинации состояния и действия.

# Начальная инициализация Q-таблицы может выполняться случайными значениями, нулями или другими константами, в зависимости от выбранной стратегии. Цель этой таблицы — постепенно обновляться на основе опыта агента, чтобы в итоге содержать значения, которые направляют агента к максимальному суммарному вознаграждению.

****

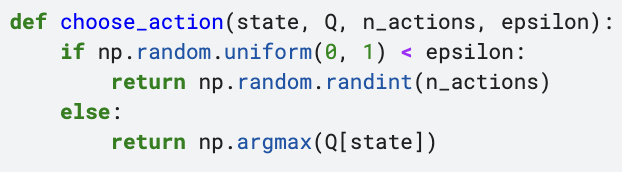
**Функция выбор действия**

Функция выбора действия в контексте обучения с подкреплением — это метод, который определяет, какое действие агент должен предпринять, находясь в определенном состоянии. Эта функция играет ключевую роль в процессе обучения, так как она напрямую влияет на исследование и использование среды агентом.

Основные аспекты функции выбора действия:

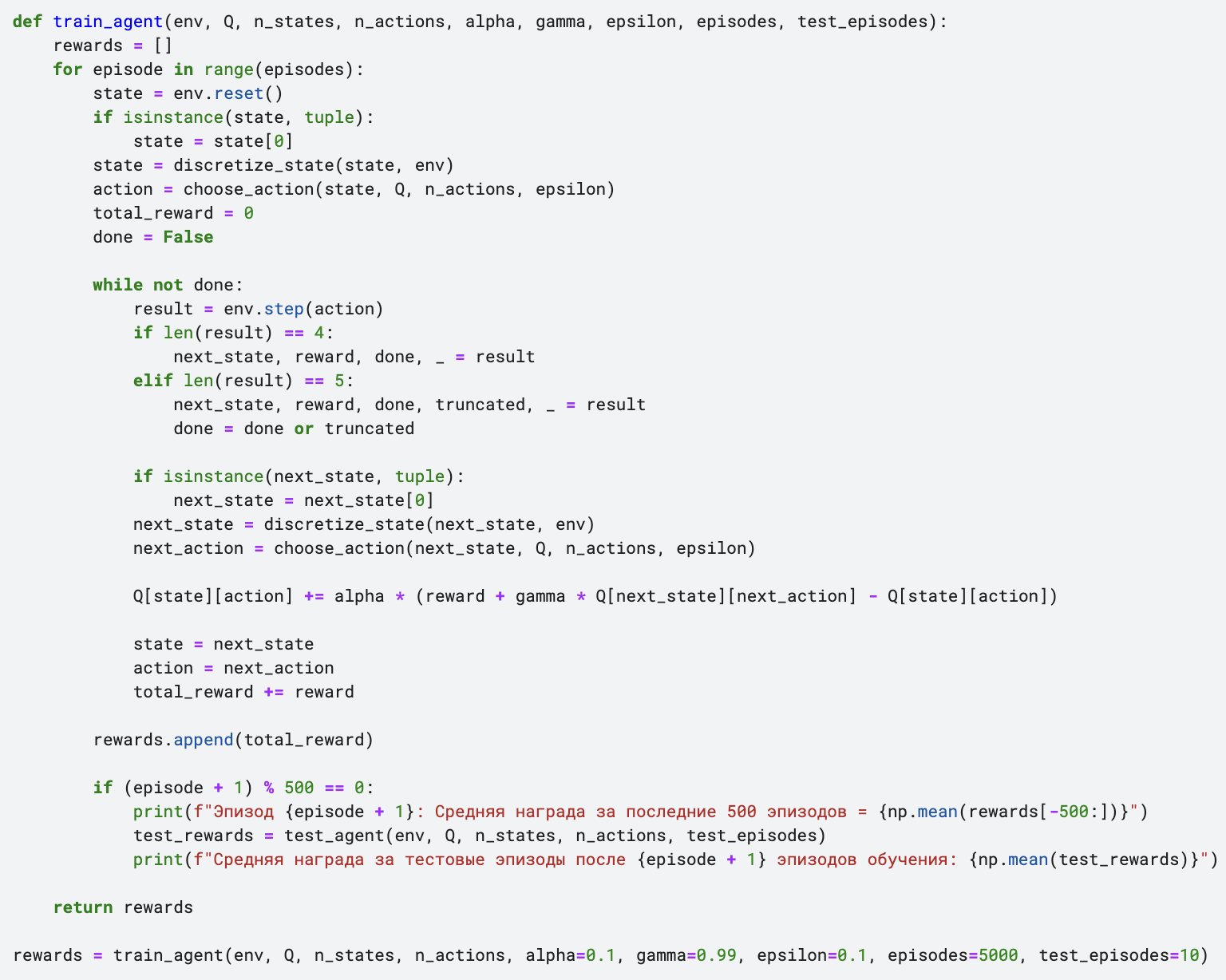
- Исследование (Exploration): Агент пробует разные действия, чтобы исследовать среду и узнать больше о возможных вознаграждениях и последствиях своих действий. Это важно на начальных этапах обучения, когда агент еще не накопил достаточно информации о среде.

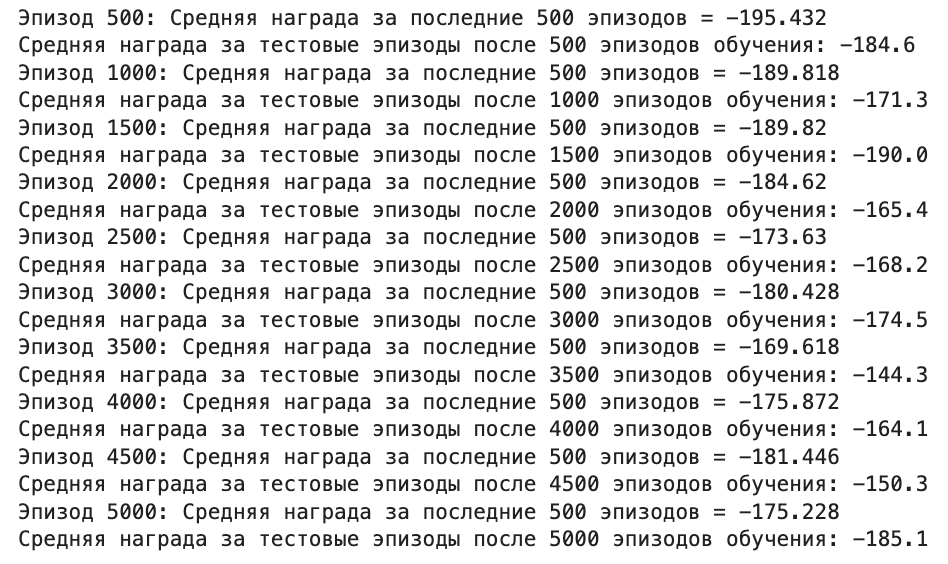
- Использование (Exploitation): Агент выбирает действия, которые, по его мнению, принесут наибольшее вознаграждение на основе уже накопленного опыта. Это важно на более поздних этапах обучения, когда агент стремится максимизировать свое вознаграждение.

****

**Обучение агента**

Обучение агента в контексте машинного обучения, особенно в обучении с подкреплением, — это процесс, в ходе которого агент взаимодействует с окружающей средой и накапливает опыт, чтобы улучшить свои действия и стратегии с целью максимизации получаемого вознаграждения.

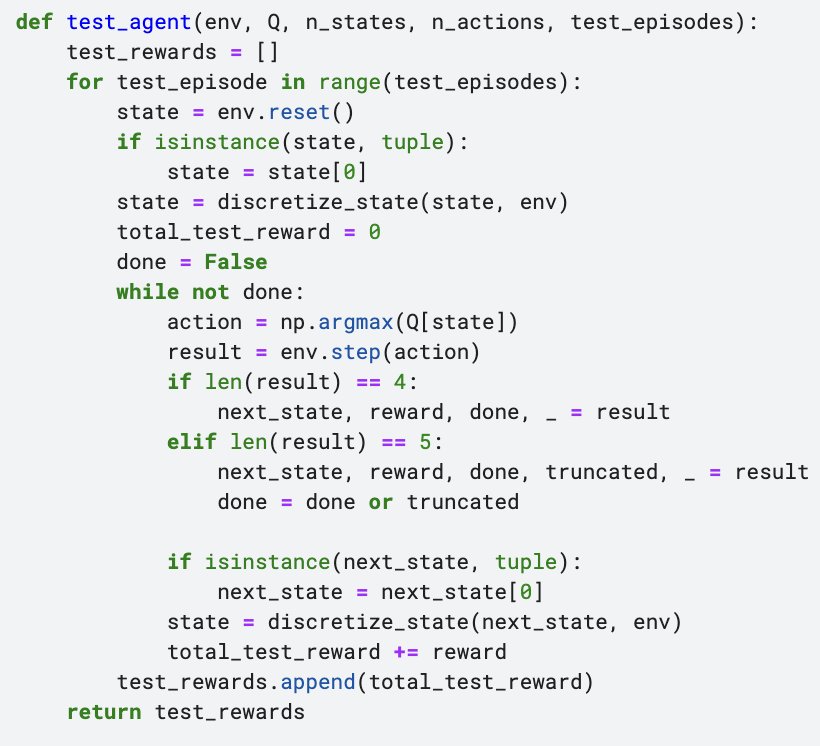
****

****

Эти результаты показывают, как средняя награда за эпизоды изменялась в процессе обучения агента с использованием алгоритма SARSA.

**Тестирование агента**

Тестирование агента — это процесс оценки производительности обученного агента в среде. В ходе тестирования агент выполняет действия, основываясь на своей стратегии (обычно на основе Q-таблицы или другой модели), но не обучается и не обновляет свои параметры. Цель тестирования — измерить, насколько эффективно агент выполняет поставленную задачу, используя знания, полученные во время обучения. Тестирование позволяет оценить, как хорошо агент усвоил оптимальную стратегию и насколько эффективно он может достигать целей в среде.

****

**Построение графика наград**

Построение графика наград — это визуализация процесса обучения агента с целью отслеживания его прогресса. График наград показывает, как изменяются получаемые агентом вознаграждения по мере прохождения эпизодов обучения.

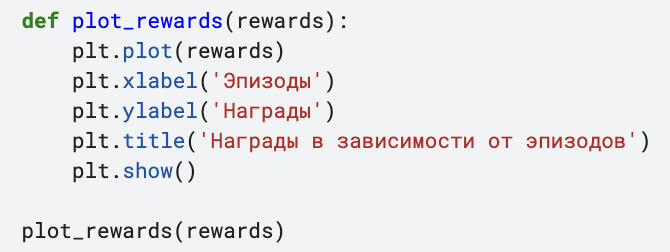




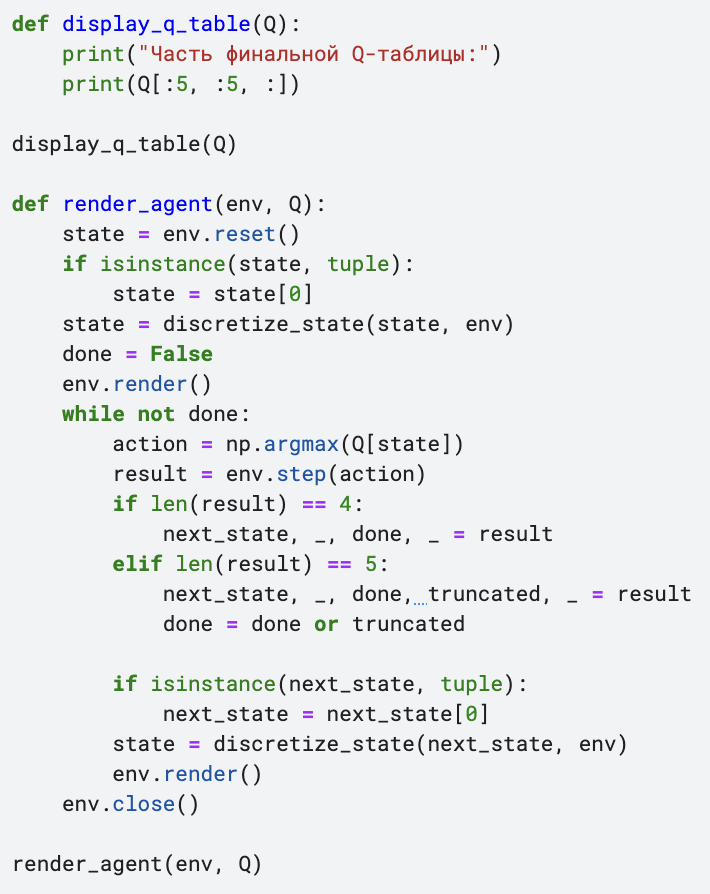
График наград может показать, как агент постепенно улучшает свою стратегию, если награды возрастают со временем.

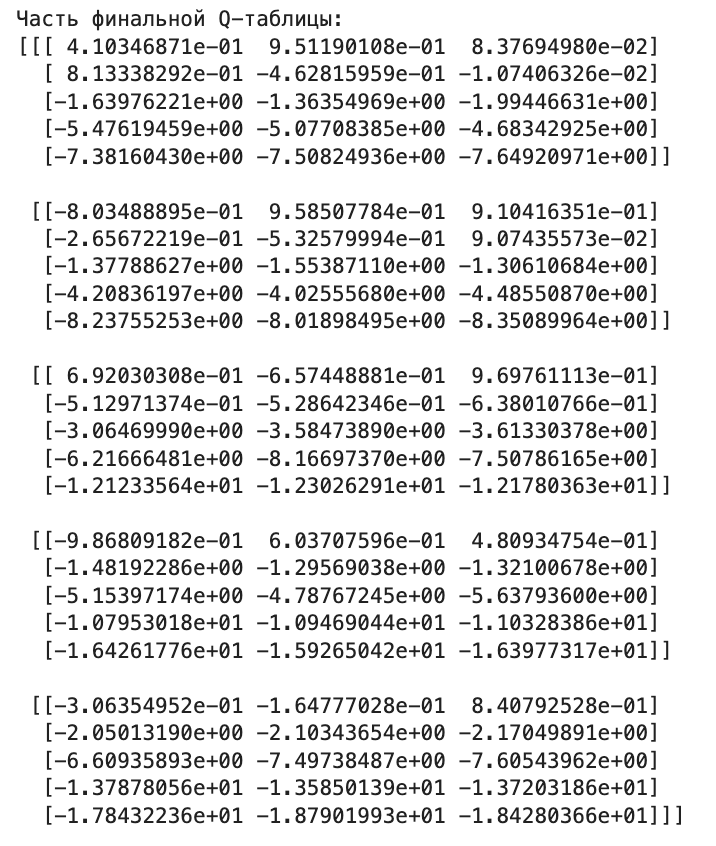
Если награды остаются низкими или сильно колеблются, это может указывать на необходимость настройки параметров обучения или выбора другого алгоритма.

**Вывод Q-таблицы и тестирование модели**

Вывод Q-таблицы — это процесс отображения значений Q-таблицы после завершения обучения. Q-таблица хранит оценки полезности для каждой комбинации состояния и действия.

Тестирование модели — это процесс оценки эффективности обученного агента в среде без дальнейшего обучения. Агент выполняет действия, основываясь на текущих значениях Q-таблицы, и собираются данные о его производительности.

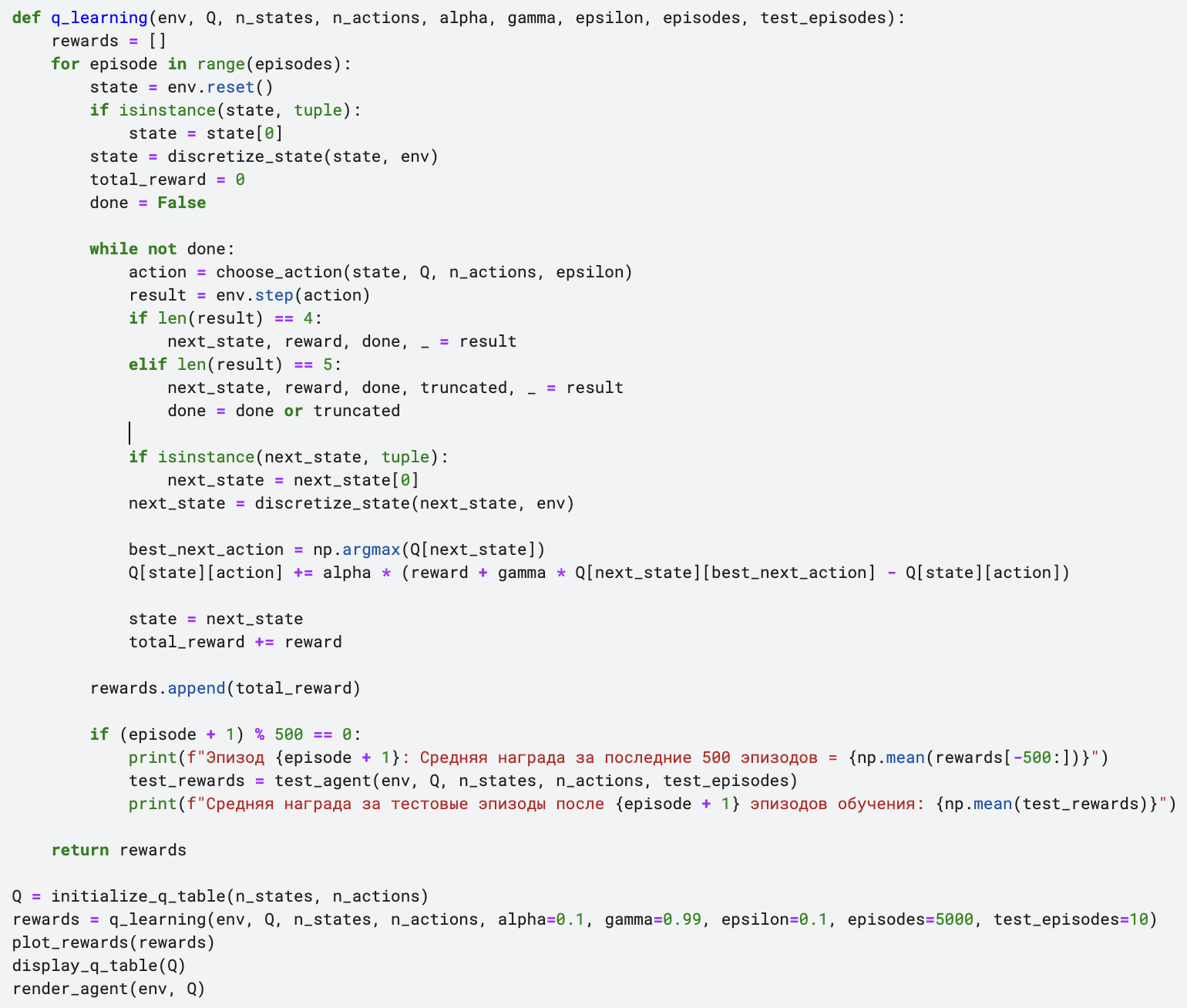


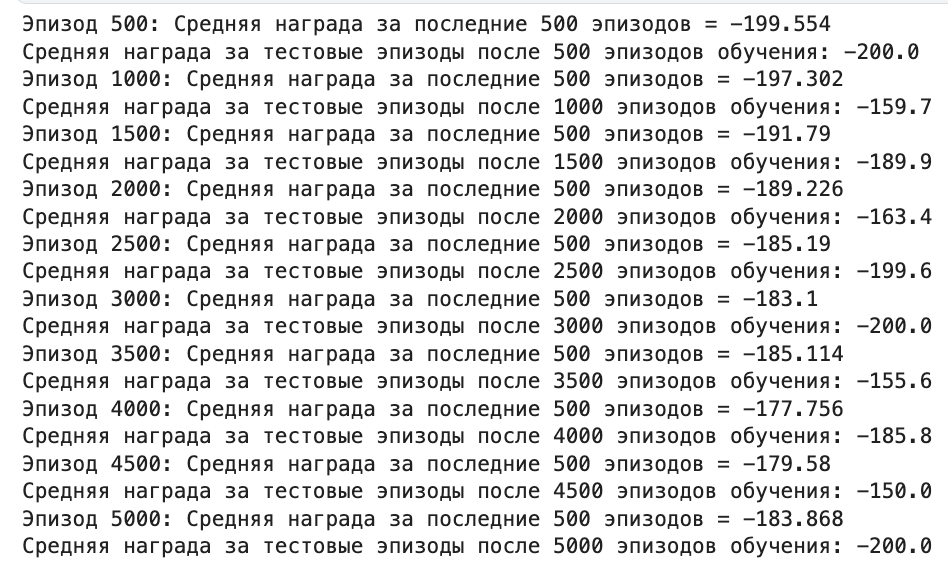


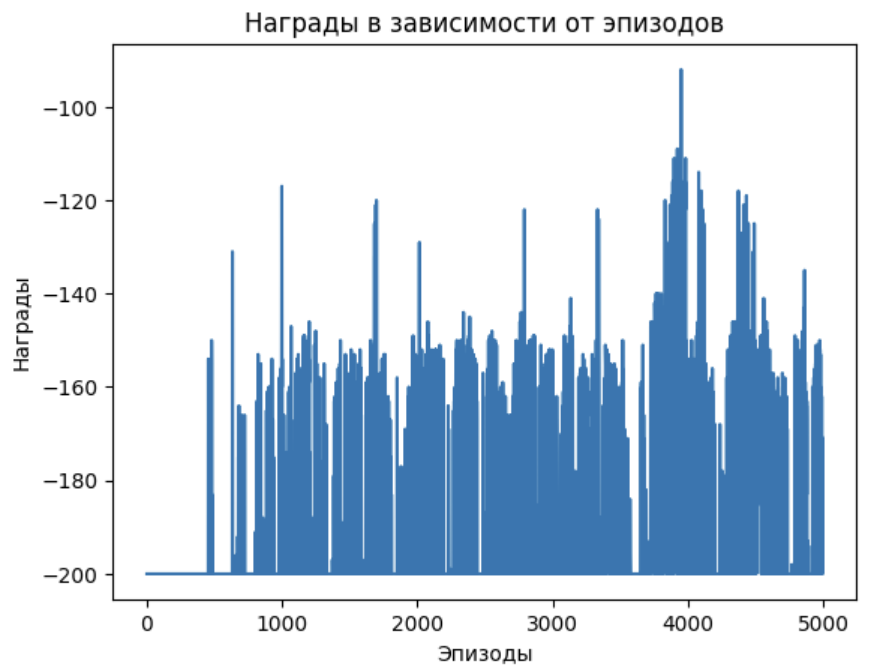
Алгоритм SARSA был успешно реализован и продемонстрировал свою эффективность в среде MountainCar-v0. Мы увидели, что агент способен обучаться и улучшать своё поведение с течением времени, что подтверждается ростом средней награды. Реализация SARSA и последующая визуализация результатов позволили нам глубже понять механизм работы алгоритмов обучения с подкреплением на основе временных различий.

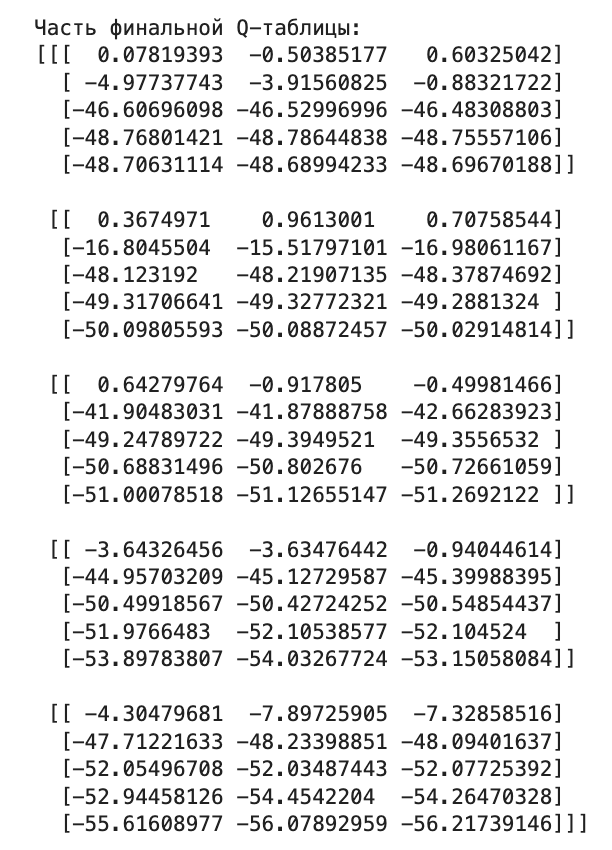
1. **Q-обучение**

Q-обучение — это метод обучения с подкреплением, при котором агент учится оптимальной стратегии, максимизируя ожидаемое суммарное вознаграждение. Агент обновляет Q-значения, используя максимальное ожидаемое вознаграждение от следующего состояния, независимо от выбранного действия.

****

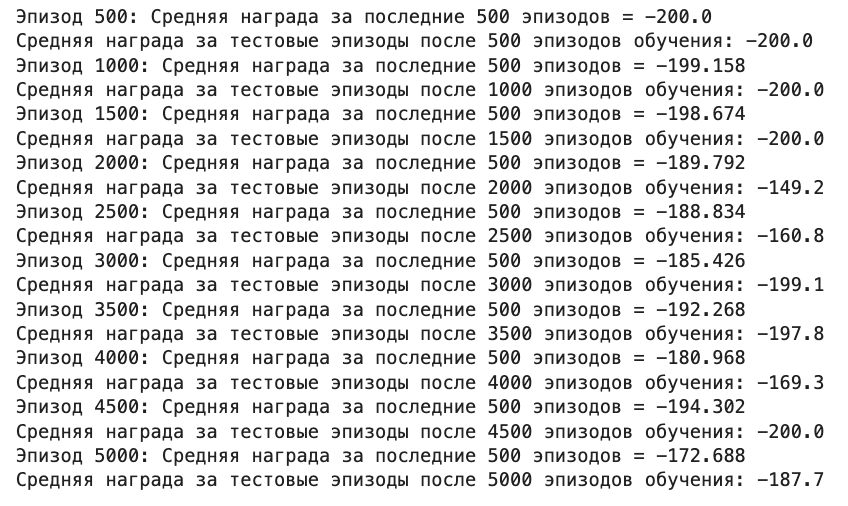
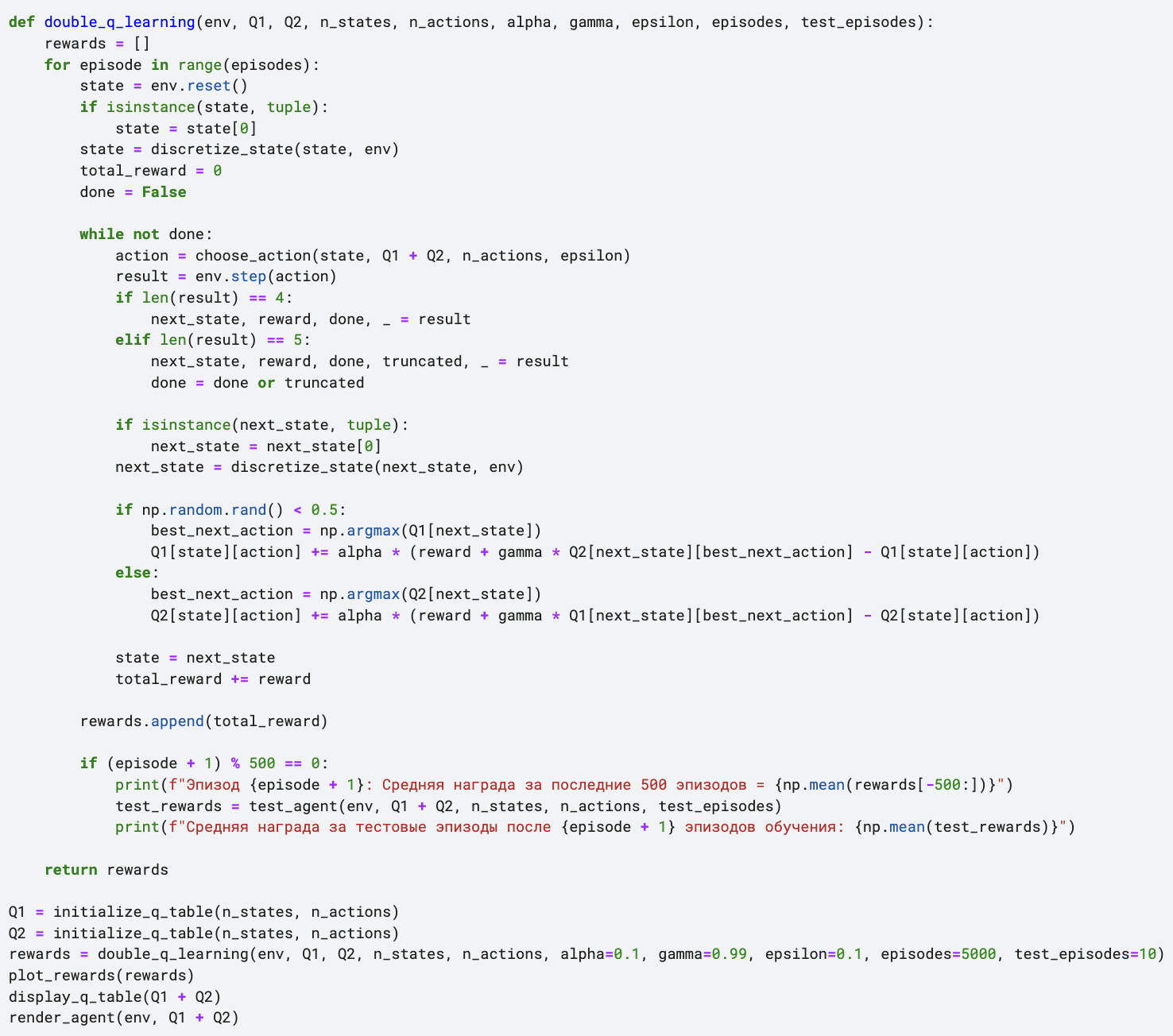
****

****

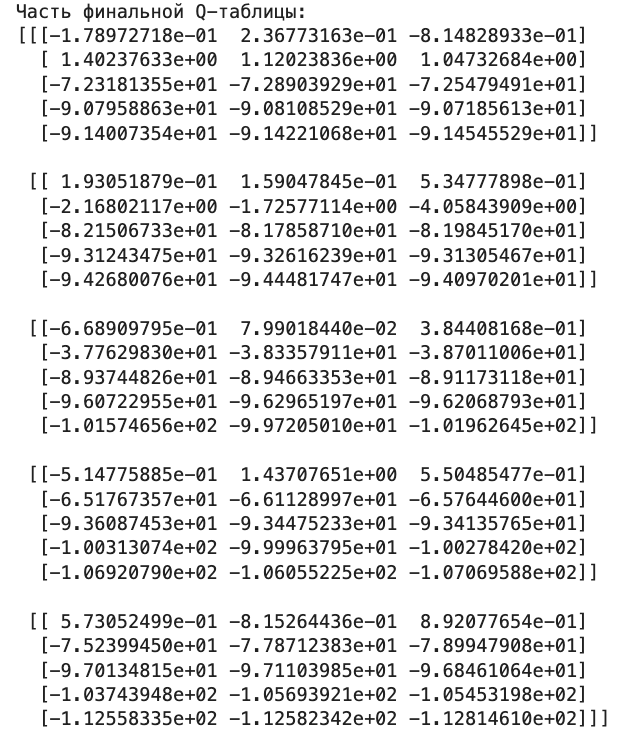
****

1. **Двойное Q-обучение**

Двойное Q-обучение — это метод обучения с подкреплением, который использует две отдельные Q-таблицы (Q1 и Q2) для уменьшения положительной предвзятости при оценке ожидаемого вознаграждения. В каждом шаге обучения одна из таблиц используется для выбора действия, а другая — для оценки вознаграждения, чередуя их на каждом шаге.

****

****

****

**Вывод:**

1. SARSA:

- Средние награды: Начиная с -199.668 и доходя до -182.998, SARSA показывает улучшение, но оно не является значительным.

- Тестовые награды: Большие колебания наград указывают на нестабильность стратегии. Например, тестовые награды после 2000 эпизодов улучшились до -172.9, но затем снова упали.

2. Q-обучение:

- Средние награды: Показатели улучшаются с -199.554 до -183.868, что свидетельствует о некотором прогрессе.

- Тестовые награды: Результаты также демонстрируют большие колебания, особенно в начале и конце обучения. Награды сильно варьируются от -200.0 до -150.0.

3. Двойное Q-обучение:

- Средние награды: Стабильное улучшение от -200.0 до -172.688. Это указывает на то, что агент последовательно улучшает свою стратегию.

- Тестовые награды: Результаты более стабильны по сравнению с другими алгоритмами. Например, тестовые награды после 2000 эпизодов улучшились до -149.2, и, несмотря на некоторые колебания, остаются стабильными.

Двойное Q-обучение обеспечивает наилучшие результаты благодаря своей устойчивости к переоценке вознаграждений и стабильности обучения.