Московский государственный технический университет

им. Н.Э. Баумана

Факультет «Информатика и системы управления»

Кафедра «Системы обработки информации и управления»



**Лабораторная работа № 4**

**По курсу «методы машинного обучения в АСОИУ»**

**«Реализация алгоритма Policy Interation»**

**Выполнила:**

студентка ИУ5-23М

Светашева Ю.В.

**Проверил:**

Балашов А.М.

Подпись:

29.04.2024

Москва, 2024

**Описание задания**

На основе рассмотренного на лекции примера реализуйте алгоритм Policy Iteration для любой среды обучения с подкреплением (кроме рассмотренной на лекции среды Toy Text / Frozen Lake) из библиотеки [Gym](https://www.gymlibrary.dev/) (или аналогичной библиотеки).

**Задача, которую решает агент**

Агент должен найти оптимальную политику, которая максимизирует общую награду за время работы в среде.

В среде Taxi-v3 агент должен перемещать такси так, чтобы подбирать и доставлять пассажиров к месту назначения, максимизируя при этом награду. Алгоритм итерации политики помогает агенту найти оптимальную стратегию для выполнения этой задачи.

Задача агента в среде Taxi-v3

**Описание среды**

Среда Taxi-v3 представляет собой сетку 5x5, на которой агент (такси) может перемещаться в четырех направлениях: вверх, вниз, влево или вправо. В городе есть несколько возможных мест посадки и высадки пассажиров.

Цель: Агент должен управлять такси, чтобы подобрать пассажира в одном месте и отвезти его к месту назначения, получая награду за успешное выполнение задачи.

Действия: Агент может выбрать одно из пяти действий:

Вверх (0)

Вниз (1)

Влево (2)

Вправо (3)

Подобрать или высадить пассажира (4)

**Состояния**: Состояния определяют положение такси, местоположение пассажира и место назначения.

**Награда**: Агент получает положительную награду за успешную посадку и доставку пассажира к месту назначения и отрицательную награду (наказание) за неудачные действия, например, выход за границы карты.

Процесс работы кода

**Инициализация**: Агент начинает с случайной политики — стратегии, определяющей действия для каждого состояния.

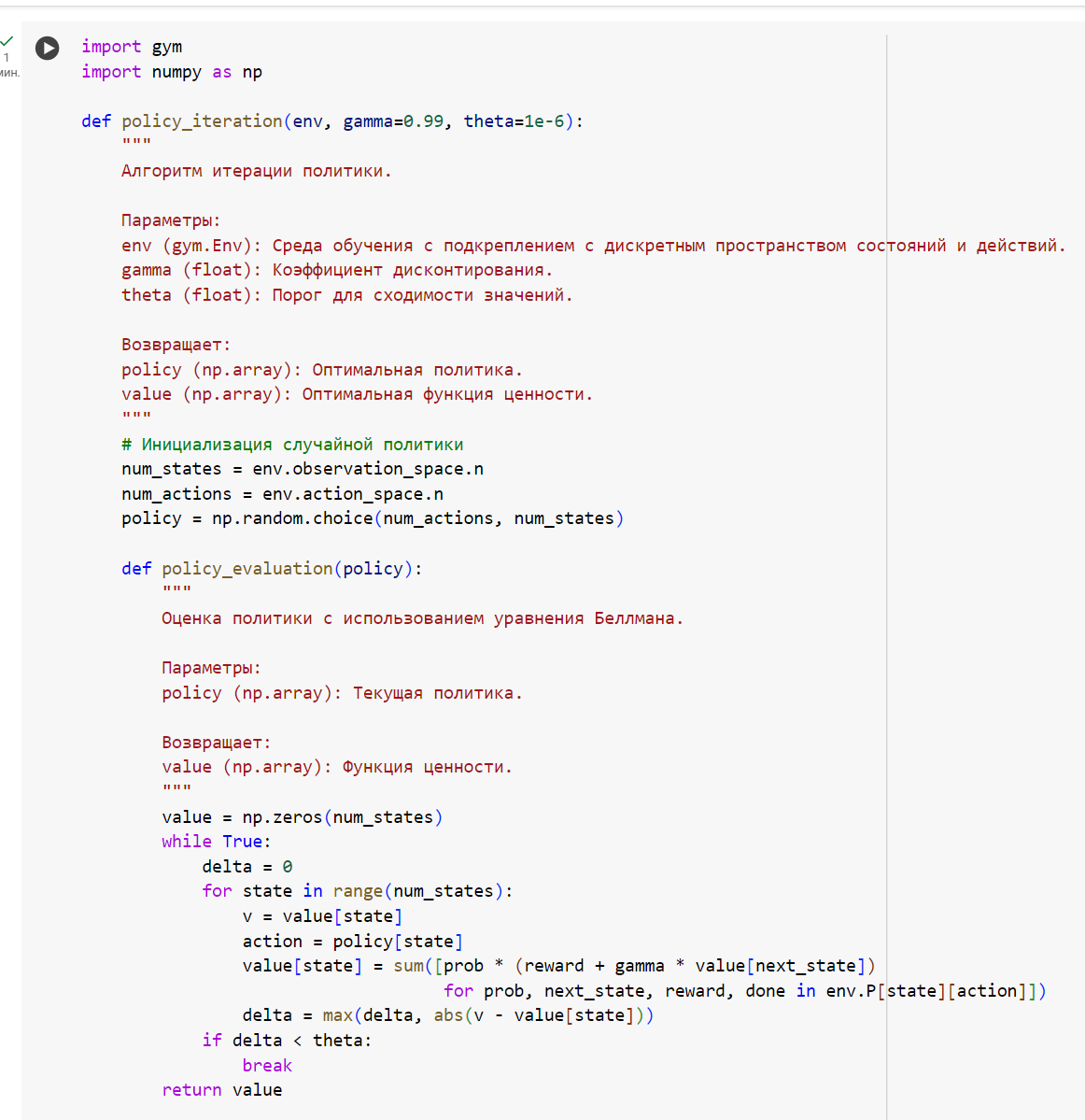
**Цикл итерации политики:** Агент оценивает текущую политику, вычисляя функцию ценности, и улучшает политику на основе этой функции ценности. Цикл продолжается до достижения сходимости.

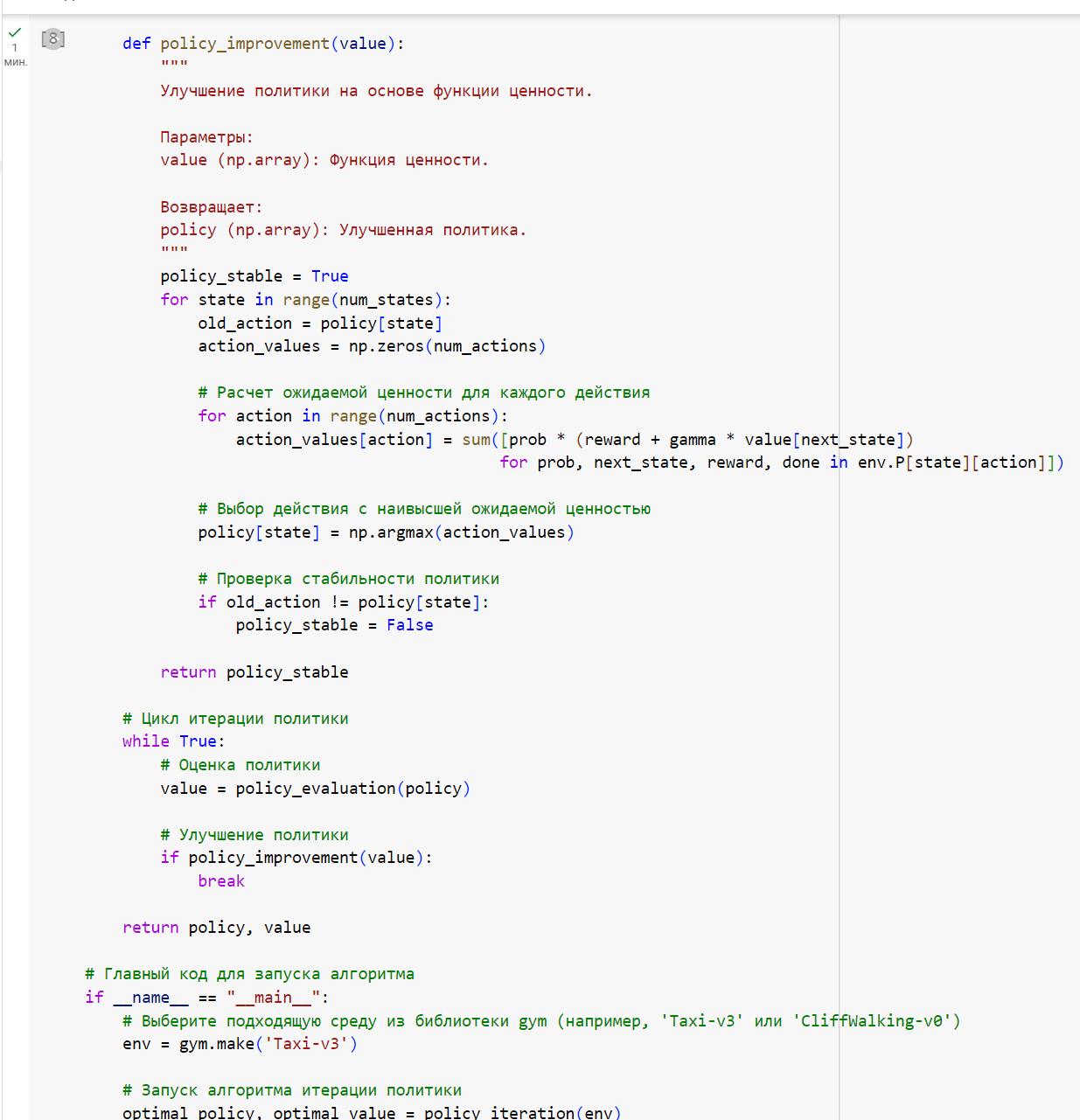
Тестирование: После достижения оптимальной политики агент тестирует свою стратегию, перемещая такси по городу и максимизируя награду.

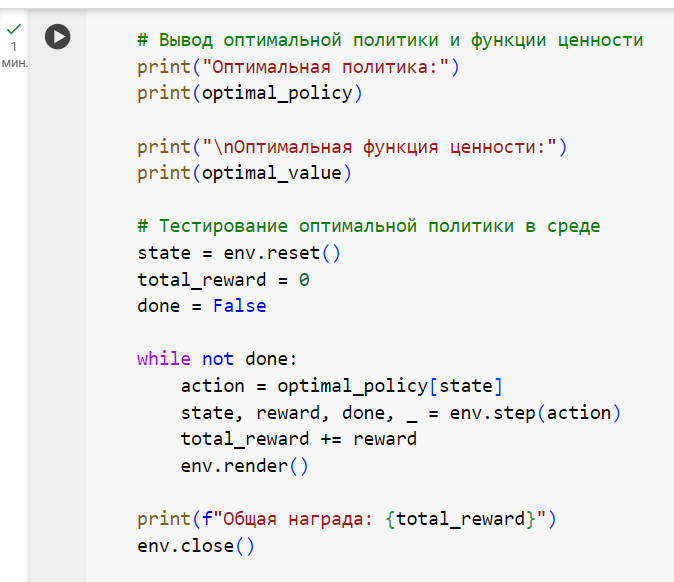
**Итоги**

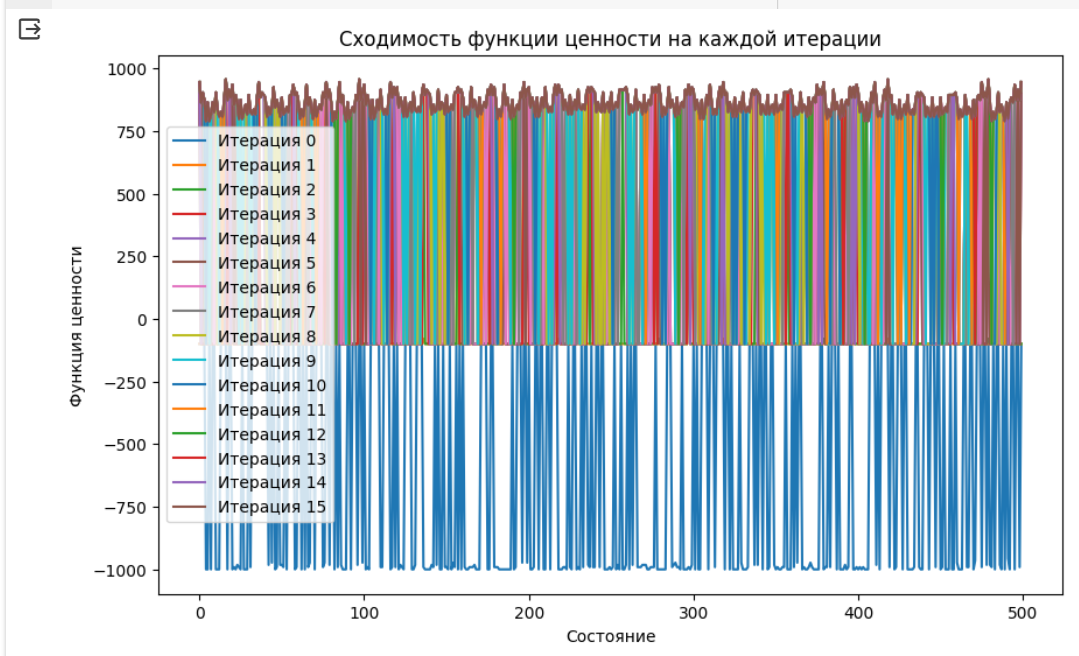
В результате код реализует алгоритм итерации политики для того, чтобы агент научился эффективно перемещать такси в среде Taxi-v3, доставляя пассажиров к месту назначения и максимизируя награду.

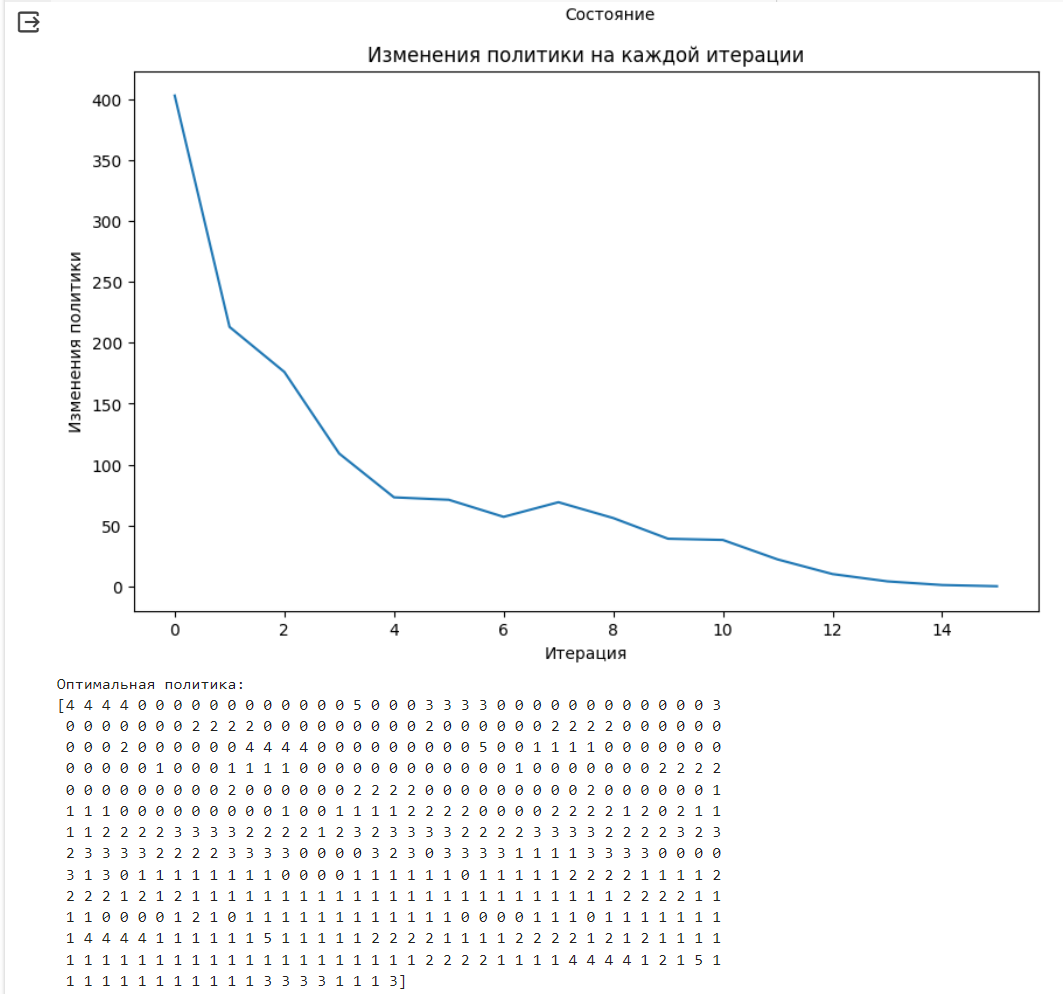
**Текст программы и экранные формы с примерами выполнения программы**

****

****

****

****

****

**Вывод**

На основе полученных результатов можем сделать вывод о том, что алгоритм итерации политики в данном коде реализует метод для обучения агента в среде с подкреплением. Это алгоритм, который использует сочетание оценок текущей политики и ее улучшений для нахождения оптимальной политики, которая максимизирует суммарную награду за время работы агента в среде.