МИНОБРНАУКИ РОССИИ
САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)
Кафедра МОЭВМ

ОТЧЕТ

по лабораторной работе №3

по дисциплине «Обучение с подкреплением»

Тема: Реализация SAC для среды Flappy Bird

Студент гр. 0306 _____ Голубев А.Н.

Преподаватель _____ Глазунов С.А.

Санкт-Петербург
2025 г.

Цель работы.

Реализация SAC для среды Flappy Bird. Исследование зависимости результатов от изменения значения коэффициента контроля энтропии.

Задание.

1. Реализация SAC
2. Изменение значения alpha для контроля энтропии
3. Реализация автоматической настройки alpha

Выполнение работы.

1. Реализация SAC

Проверка работы алгоритма проводилась на среде Flappy Bird.

В пространство наблюдений данной среды входят:

- горизонтальное положение последней трубы
- вертикальное положение последней верхней трубы
- вертикальное положение последней нижней трубы
- горизонтальное положение следующей трубы
- вертикальное положение следующей верхней трубы
- вертикальное положение следующей нижней трубы
- горизонтальное положение второй следующей трубы
- вертикальное положение второй следующей верхней трубы
- вертикальное положение второй следующей нижней трубы
- вертикальное положение игрока
- вертикальная скорость игрока
- вращение игрока

В пространство действий входят:

- 0 — ничего не делать
- 1 — взмахнуть крыльями

Награды:

- 0.1 — птица жива
- 1.0 — птица пролетела через трубу
- -1.0 — птица умерла
- -0.5 — птица коснулась вершины экрана

Алгоритм SAC состоит из нескольких ключевых компонент:

- Две Q-сети (критики)

- Сеть политики (актор)
- Целевые Q-сети
- Replay буффер

Сеть политики (актор) определяет поведение агента путем выбора действий.

Q-сети нужны для оценки действий актора. Две сети используется для уменьшения склонности к завышению оценок.

Целевые сети используются для стабилизации обучения путем предоставления более консистентной ссылки для оценки значений. Целевые сети периодически обновляются, чтобы постепенно отслеживать весы основных сетей.

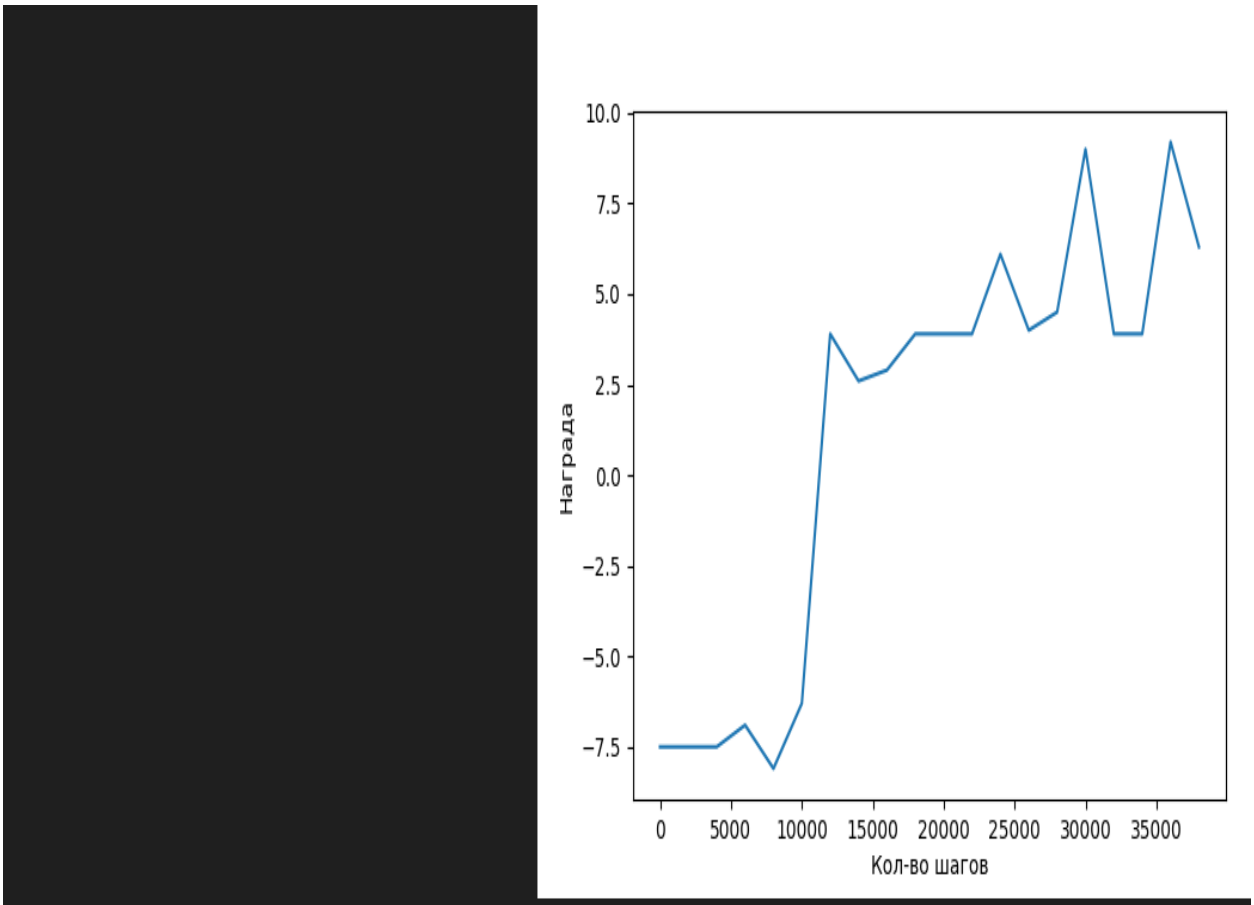
В обучении с помощью SAC важную роль играет параметр alpha. Параметр alpha – это коэффициент регуляризации энтропии, также известный как параметр температуры. Его задача — контролировать баланс между ожидаемой выгодой и энтропией, мерой случайности в политике. Чем больше его значение, тем более актор склонен к исследованию в процессе обучения. Меньшее значение наоборот побуждает актора чаще использовать уже изученные данные.

Есть два вида алгоритма SAC. Первый использует фиксированное значение alpha, второй — меняет его в процессе обучения.

В данной работе было рассмотрено два подхода.

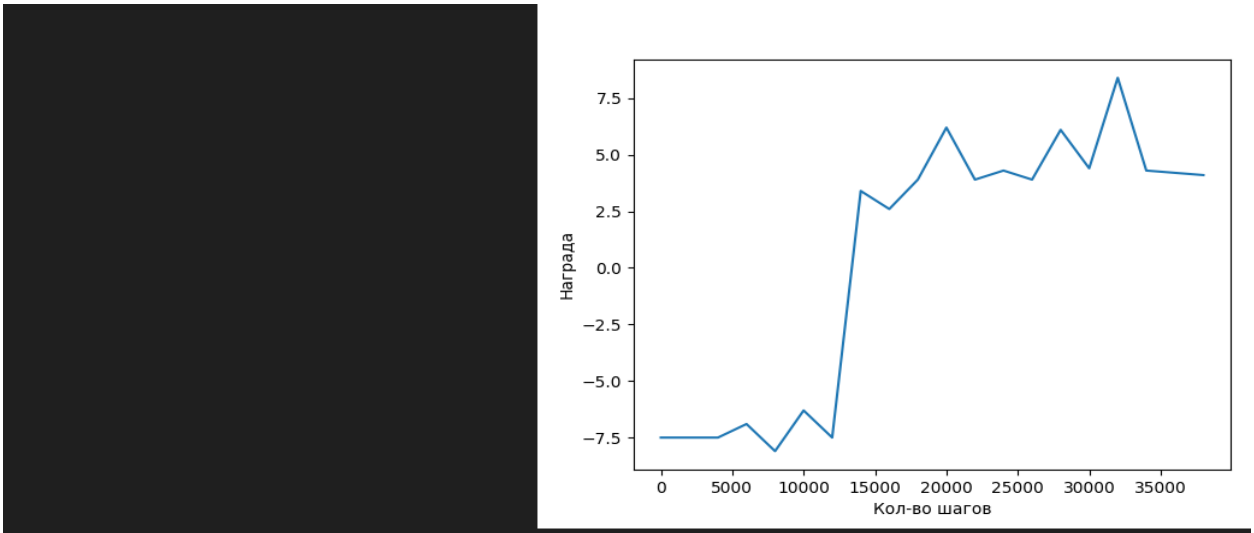
2. Изменение значения alpha для контроля энтропии

Для данного эксперимента было выбрано несколько значений alpha: 0.05, 0.2, 0.8. Результаты представлены на рисунках 1 — 3.



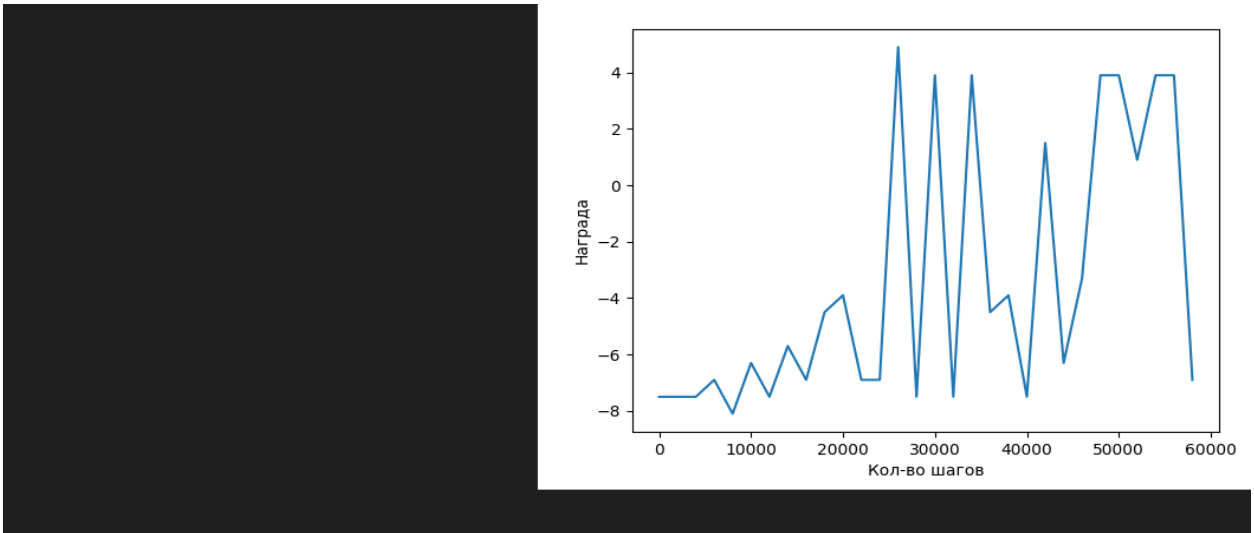

```
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  PORTS  GITLENS
reward: 2.600000000000002 | alpha: 0.05
reward: 2.900000000000002 | alpha: 0.05
reward: 3.8999999999999986 | alpha: 0.05
saved model
reward: 3.8999999999999986 | alpha: 0.05
reward: 3.8999999999999986 | alpha: 0.05
reward: 6.099999999999994 | alpha: 0.05
reward: 3.9999999999999982 | alpha: 0.05
reward: 4.499999999999964 | alpha: 0.05
saved model
reward: 8.999999999999984 | alpha: 0.05
reward: 3.8999999999999986 | alpha: 0.05
reward: 3.8999999999999986 | alpha: 0.05
reward: 9.199999999999983 | alpha: 0.05
reward: 6.299999999999994 | alpha: 0.05
saved model
```

Рис. 1 — результаты обучения (alpha = 0.05)




```
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  PORTS  GITLENS
reward: 3.8999999999999986 | alpha: 0.2
saved model
reward: 6.199999999999994 | alpha: 0.2
reward: 3.8999999999999986 | alpha: 0.2
reward: 4.299999999999997 | alpha: 0.2
reward: 3.8999999999999986 | alpha: 0.2
reward: 6.099999999999994 | alpha: 0.2
saved model
reward: 4.399999999999997 | alpha: 0.2
reward: 8.399999999999986 | alpha: 0.2
reward: 4.299999999999997 | alpha: 0.2
reward: 4.1999999999999975 | alpha: 0.2
reward: 4.099999999999998 | alpha: 0.2
saved model
```

Рис. 2 — результаты обучения (alpha = 0.2)




```
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  PORTS  GITLENS
reward: -3.899999999999998 | alpha: 0.8
saved model
reward: -7.499999999999998 | alpha: 0.8
reward: 1.5 | alpha: 0.8
reward: -6.299999999999999 | alpha: 0.8
reward: -3.299999999999998 | alpha: 0.8
reward: 3.8999999999999986 | alpha: 0.8
saved model
reward: 3.8999999999999986 | alpha: 0.8
reward: 0.9000000000000004 | alpha: 0.8
reward: 3.8999999999999986 | alpha: 0.8
reward: 3.8999999999999986 | alpha: 0.8
reward: -6.899999999999999 | alpha: 0.8
saved model
```

Рис. 3 — результаты обучения (alpha = 0.8)

По результатам обучения можно заметить, что при наименьшем значении alpha положительные значения награды начали появляться раньше. Более того, при росте alpha возросла частота отрицательных частей. Можно предположить, что это связано с тем, что из-за высокой энтропии при большем значении alpha актор предпринимает более «смелые» решения в поисках оптимальной политики, и это приводит к более низким результатам.

3. Реализация автоматической настройки alpha

Первоначальная цель увеличения стандартного вознаграждения за счет энтропии политики состоит в том, чтобы стимулировать исследования еще недостаточно хорошо изученного состояния (отсюда высокая энтропия). И наоборот, для состояний, где уже выработана политика, близкая к оптимальной, было бы предпочтительнее уменьшить энтропийный бонус политики, чтобы она не сбилась с пути поиска, поскольку ее поощряют к высокой энтропии.

Оптимальное значение alpha вычисляется по формуле на рисунке 4.

$$\alpha_t^* = \operatorname{argmin}_{\alpha_t} \mathbb{E}_{a_t \sim \pi_t^*} \left[-\alpha_t \log \pi_t^*(a_t | s_t; \alpha_t) - \alpha_t \mathcal{H} \right],$$

Рис. 4 — формула оптимального значения alpha

где H представляет целевую энтропию, или, другими словами, желаемую нижнюю границу ожидаемой энтропии политики в распределении траектории $(s_t , a_t) \sim \rho_\pi$, вызванную последним. В качестве эвристического параметра для определения целевой энтропии используется размерность пространства действий задачи.

Результаты обучения представлены на рисунках 5 — 6.

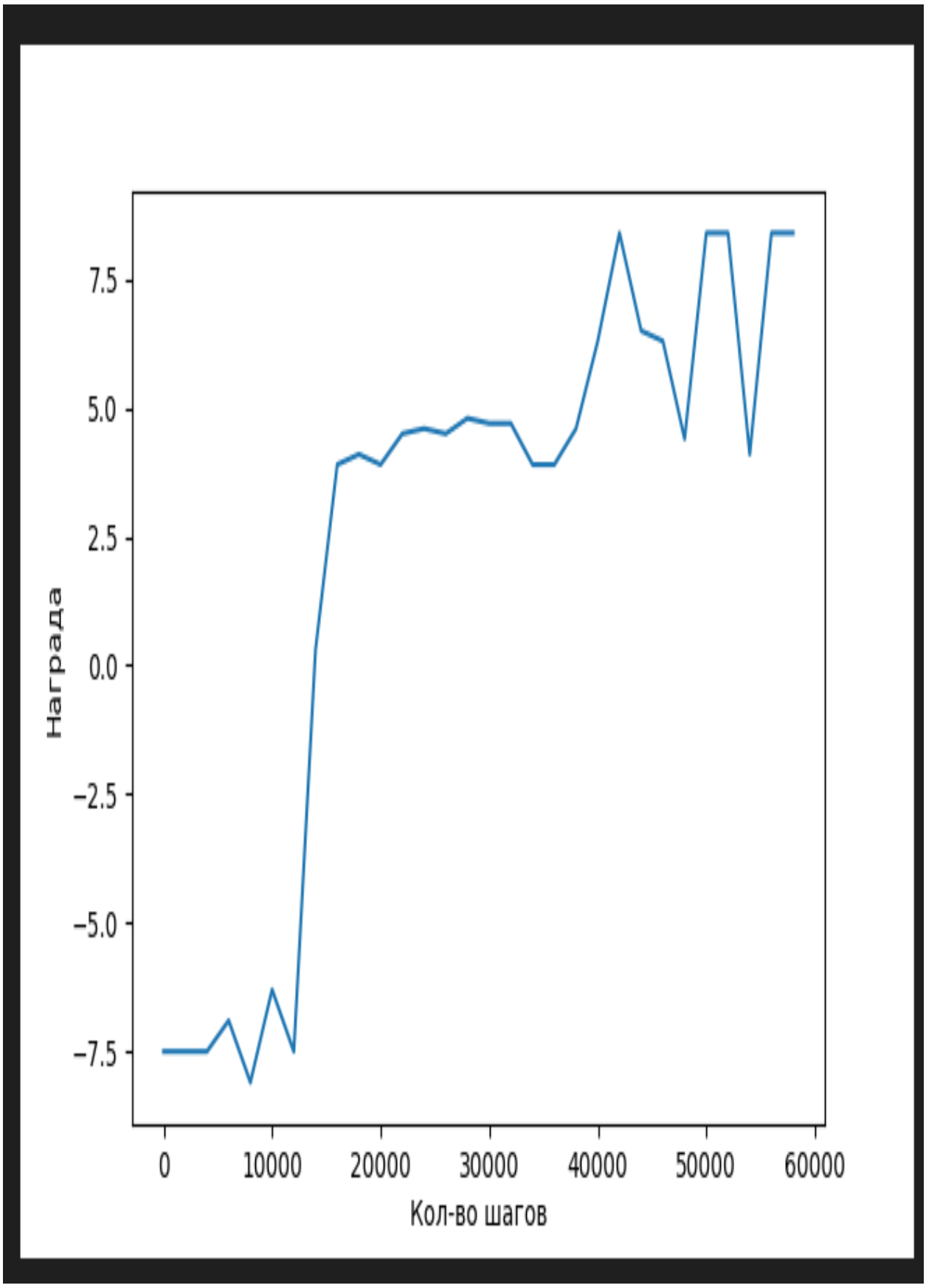


Рис. 5 — результат обучения (график)

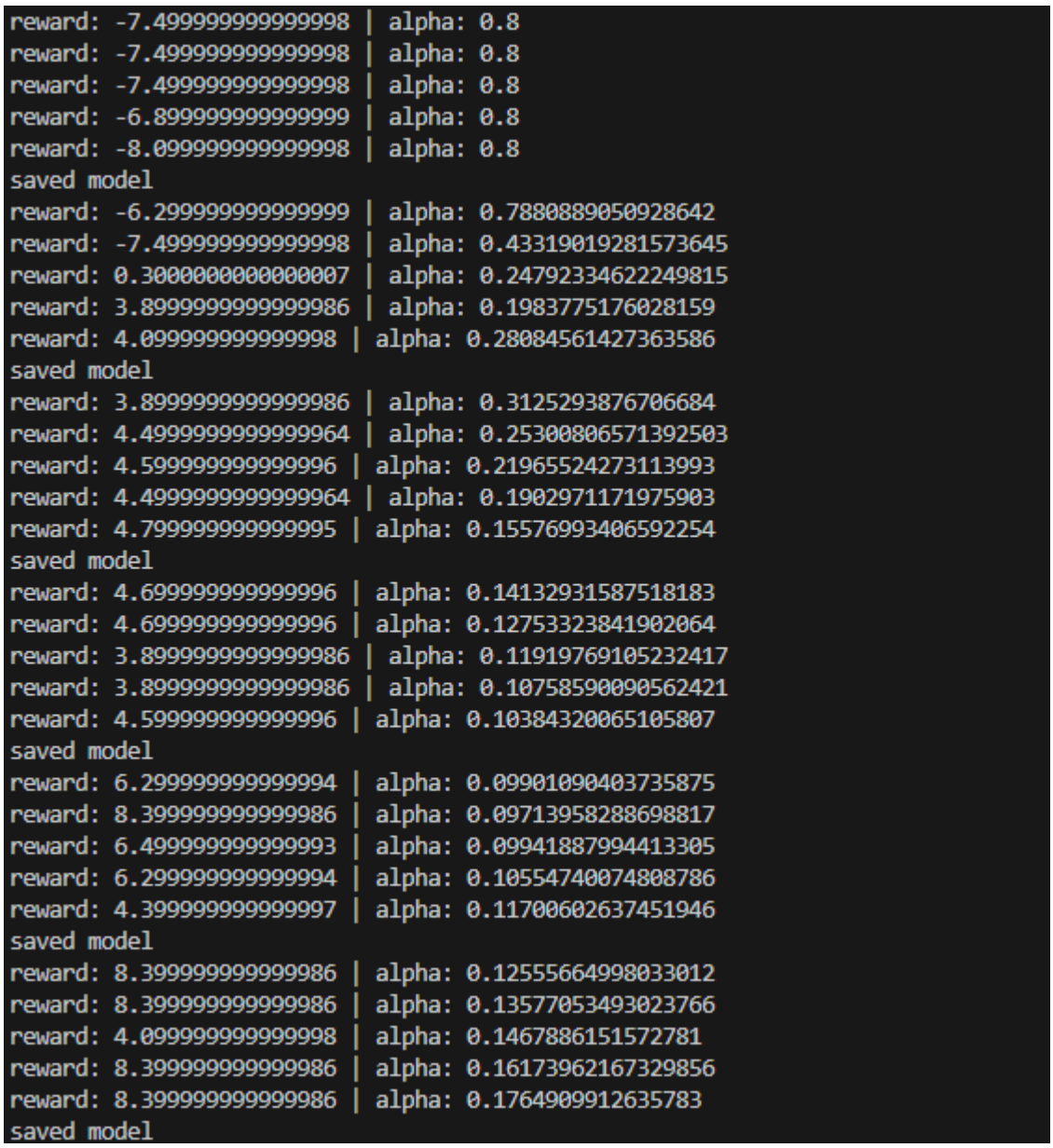


Рис. 6 — динамика изменения alpha

Можно заметить, что уже на ранних этапах обучения начал происходить спад значений alpha. Можно предположить, что актор достаточно быстро выработал политику, близкую к оптимальной, поэтому далее высокое значение энтропии не было нужно.

Выводы.

В ходе данной работы была разработана реализация алгоритма SAC. Было рассмотрено влияние параметра alpha на результаты обучения. Были реализованы две версии алгоритма: с постоянным и меняющимся значением alpha.