Домашняя работа #6 RL

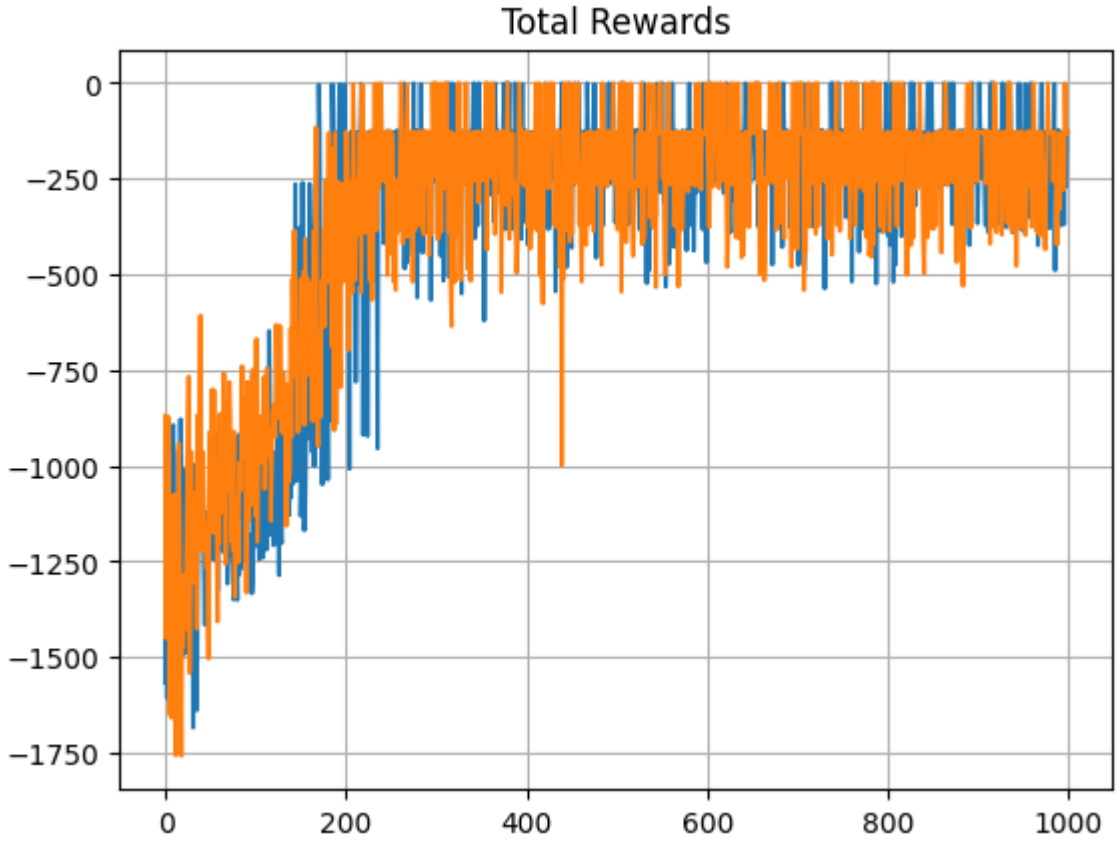
Введение

Первое задание

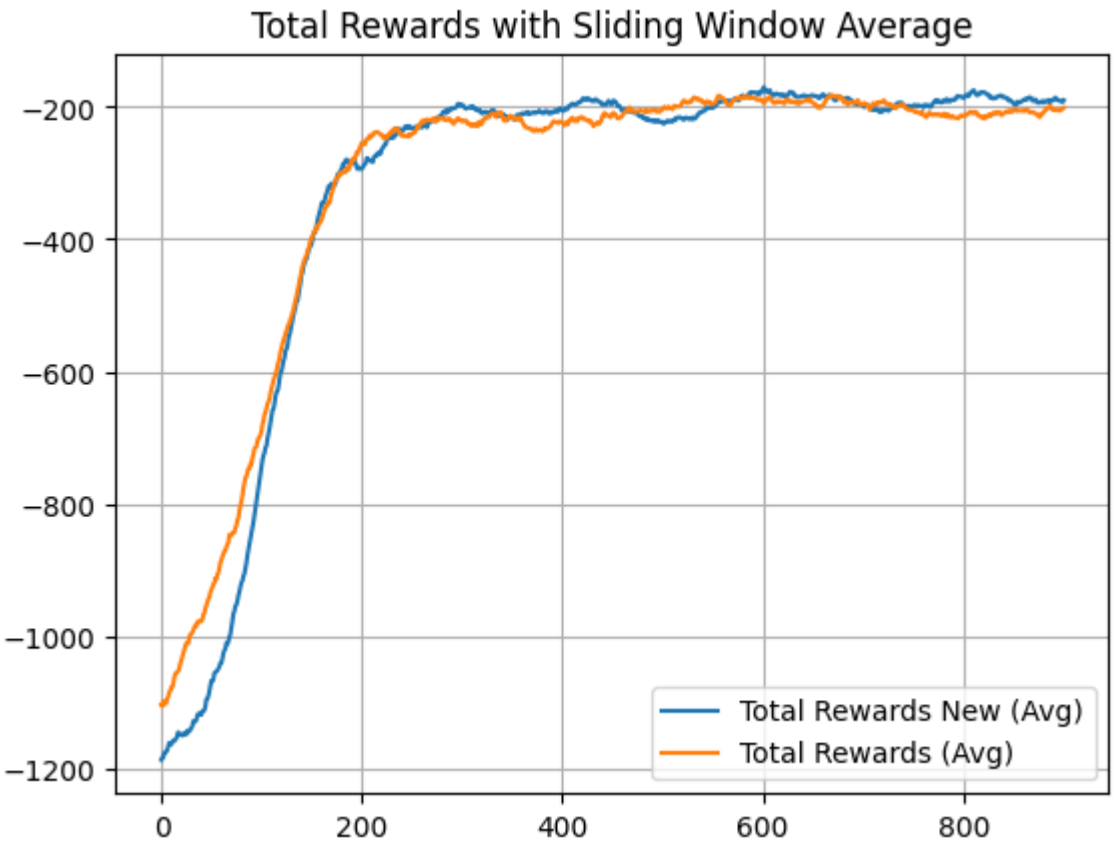
Обучение

Эксперимент 1

После переписывания b_advantages я запустил модель с теми же параметрами:



Оранжевый график здесь - это новый метод. В целом по ощущениям он более сглажен.



После сглаживания картина похожая. Оранжевый более стабильный.

Вывод

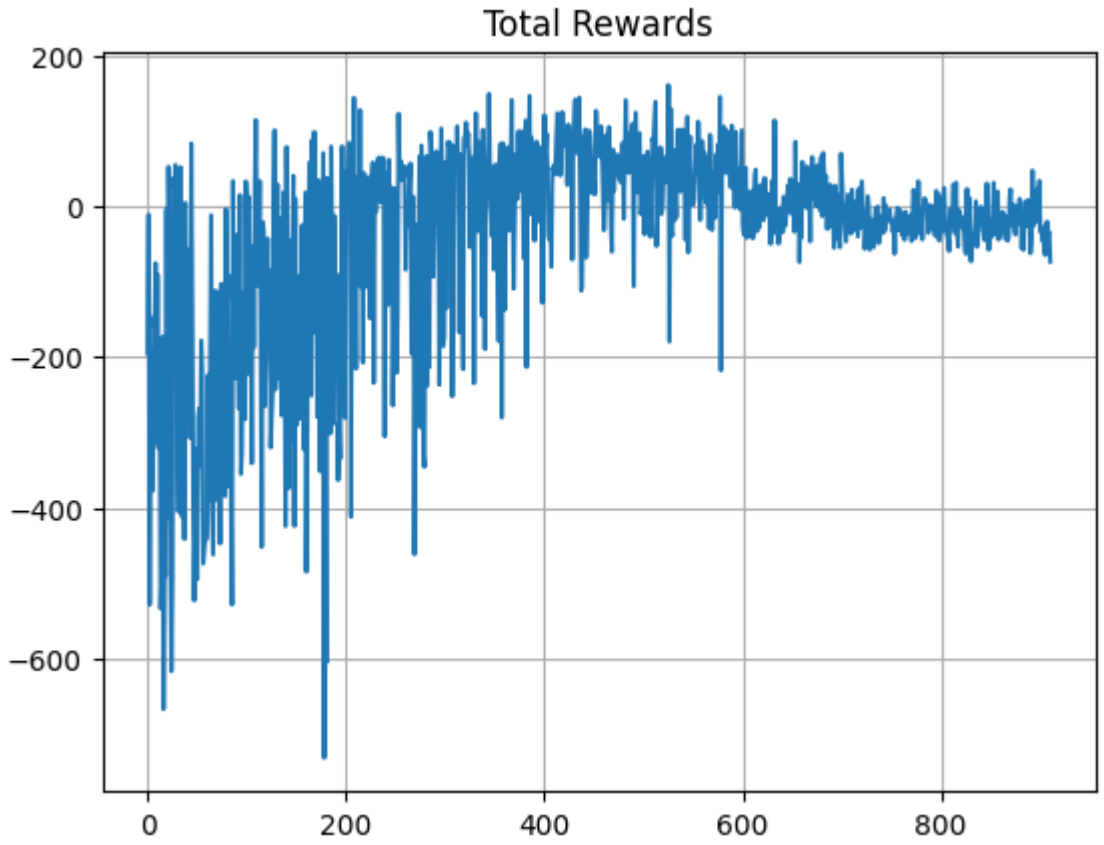
Новый метод позволяет достичь большей стабильности.

Второе задание

Обучение

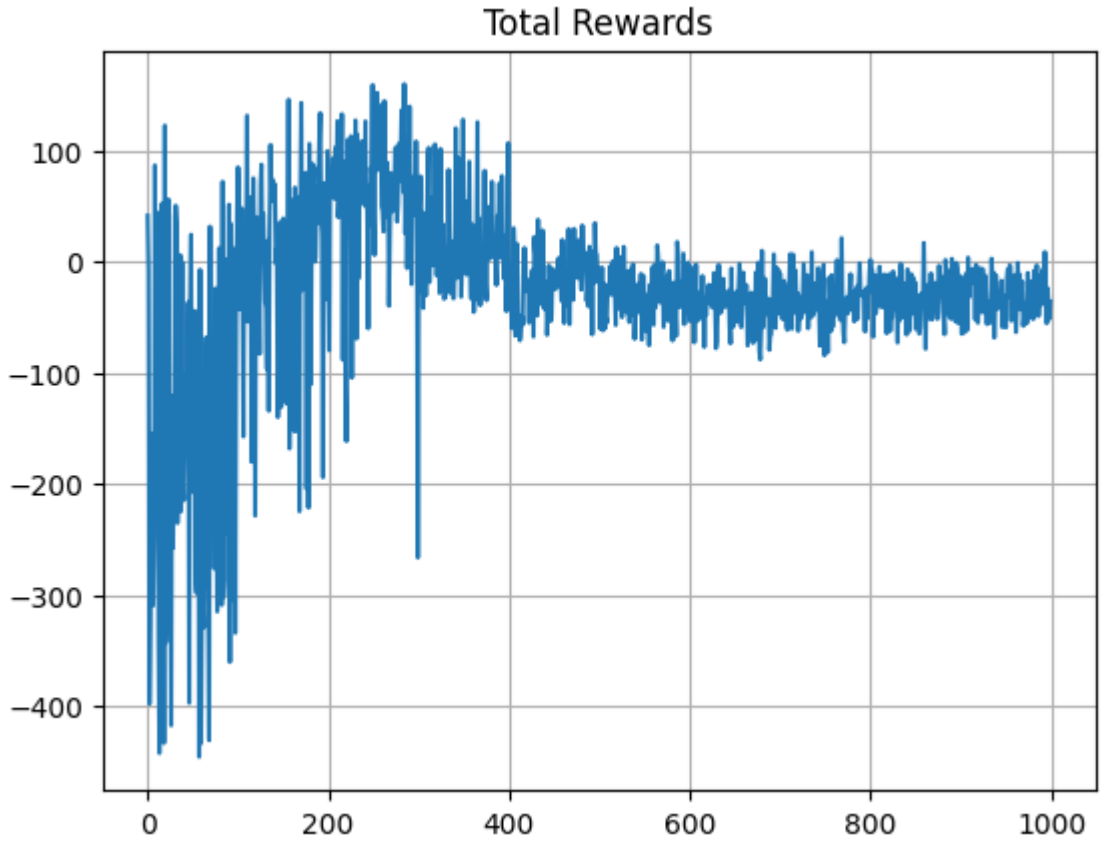
Эксперимент 1

После первого эксперимента с мультиразмерными действиями получились следующие результаты:



Модель достигает очень неплохих значений >100, но в какой-то момент попадает в локальный оптимум и застревает около нуля. Попробуем это исправить. Начал изучать как в PPO не сваливаться в локальных минимумах.

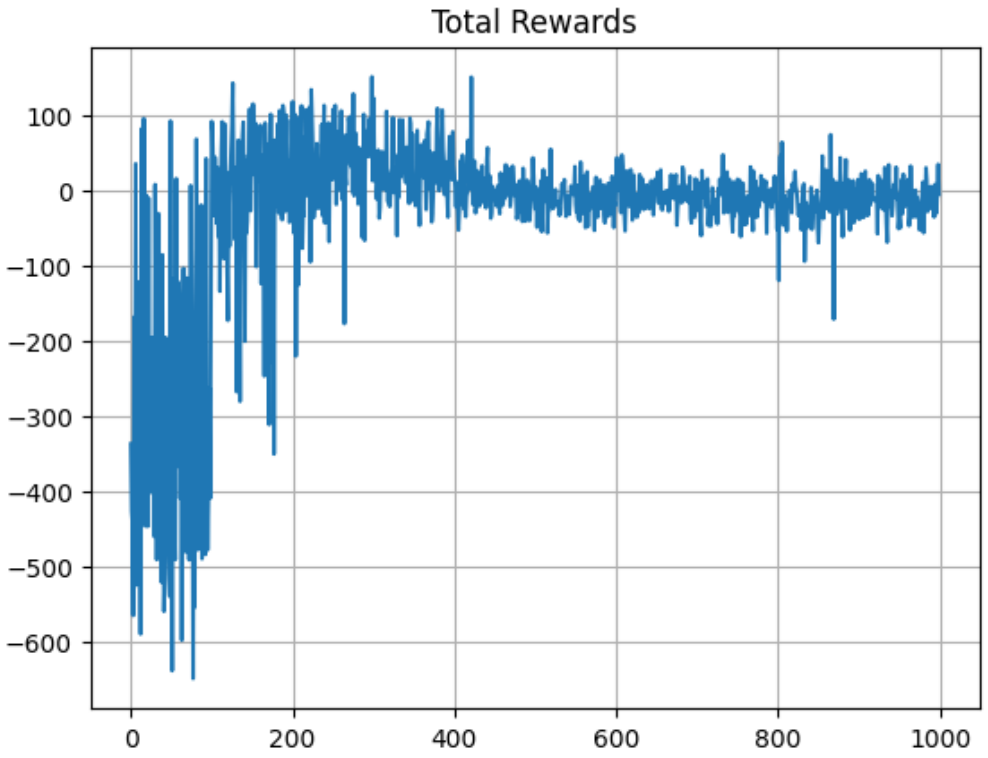
Эксперимент 2



Попытка увеличить эпсилон до 0.5 не привела к успеху.

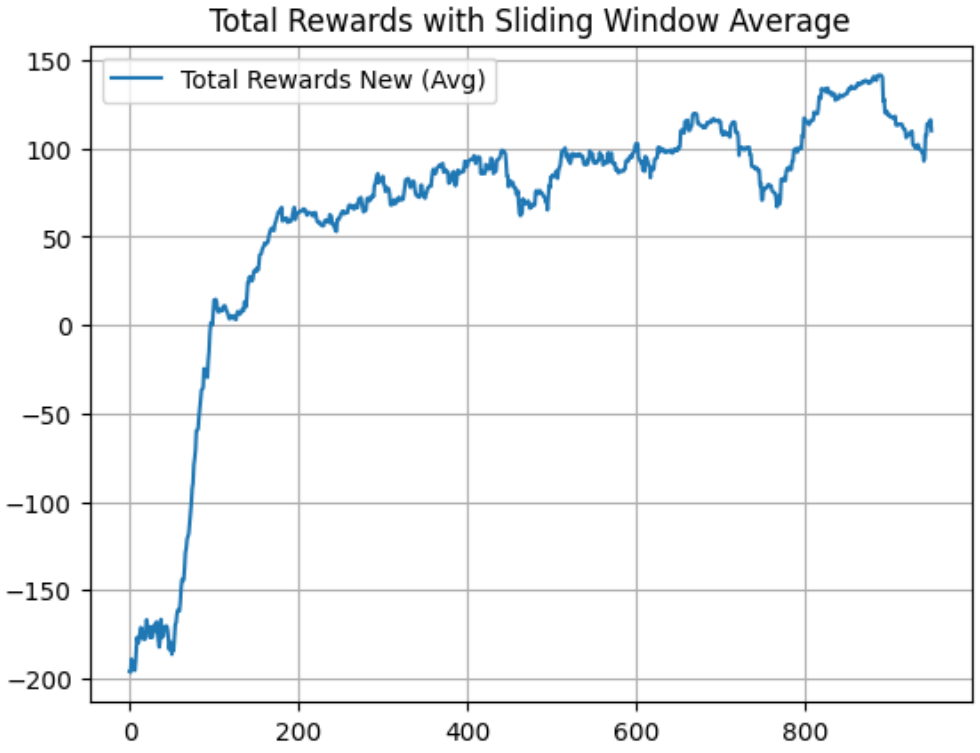
Эксперимент 3

Попробую увеличить гамму с 0.9 до 0.95. Возможно инопланетянину так будет легче понимать что +100 баллов это очень важно.



Относительно стало лучше. Луноход крутится вокруг нуля. Попробую понять останавливается ли он выводя каждую +100 награду.

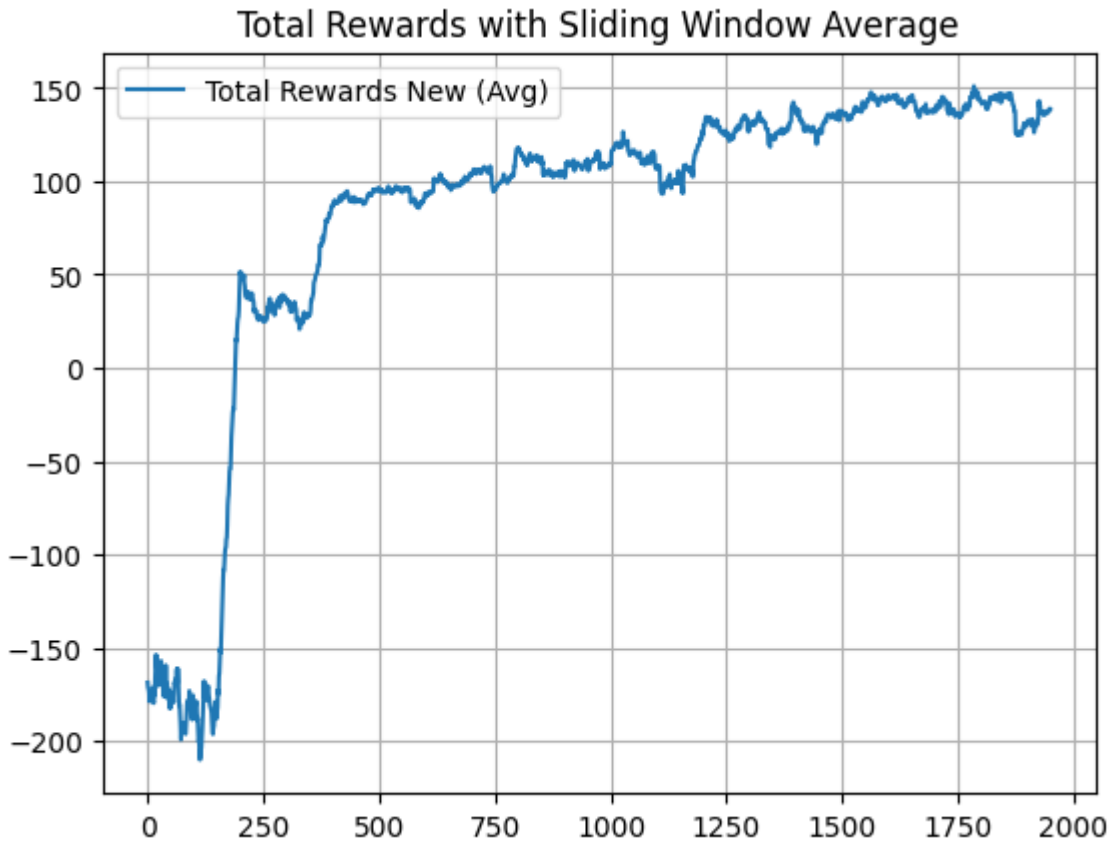
Эксперимент 4



def __init__(self, state_dim, action_dim, gamma=0.99, batch_size=128,epsilon=0.5, epoch_n=50, pi_lr=1e-3, v_lr=5e-3)
Поменял параметры на вышеприведенные модель стала лучше обучаться. Замечен прогресс.

Эксперимент 5

Увеличил количество траекторий в эпизоде до 200.



Считаю модель обученной)

Вывод

Хорошо, что отказался от идеи структурировать награду, и создавать буфер траекторий где инопланетянин сел. Получилось достичь решения гиперпараметрами.

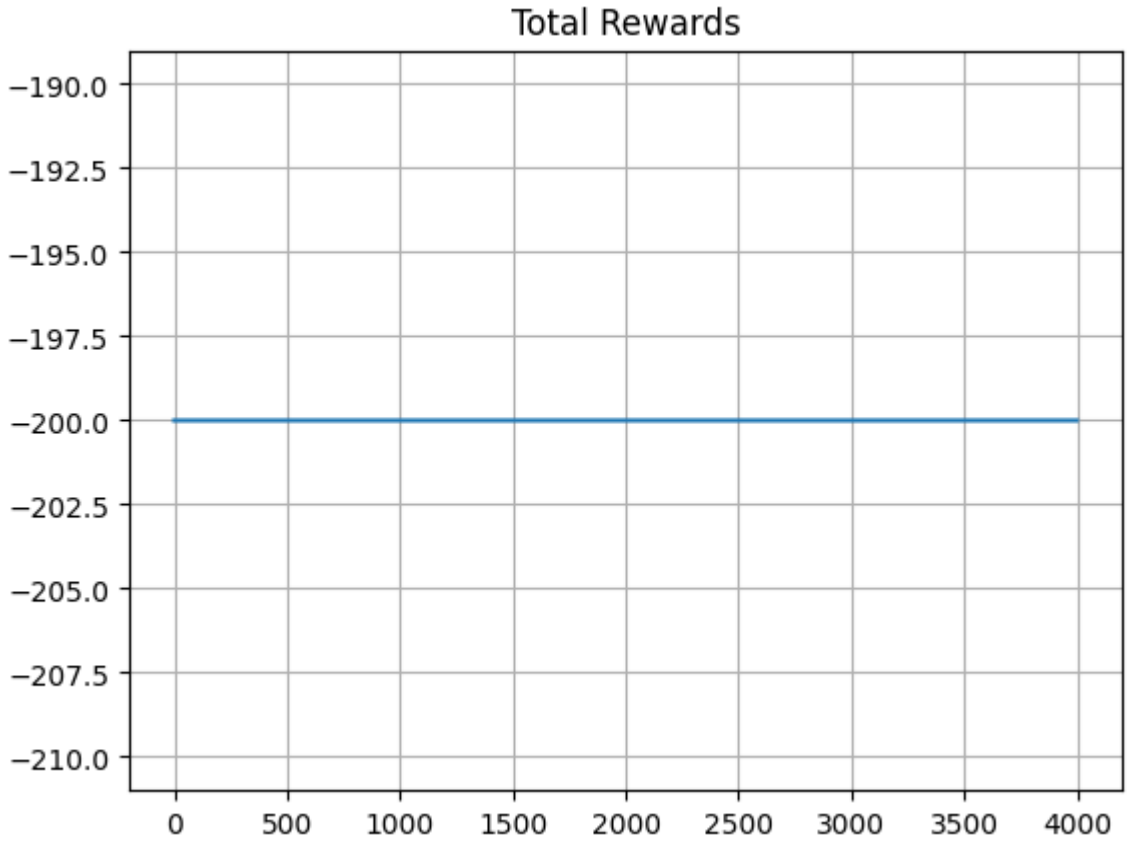
Третье задание

Обучение

Эксперимент 1

В третьем необходимо было обучить акробота. Для начала я переписал код для него, и зная акробота так как работал с ним в предыдущих дз сразу понимал приблизительно

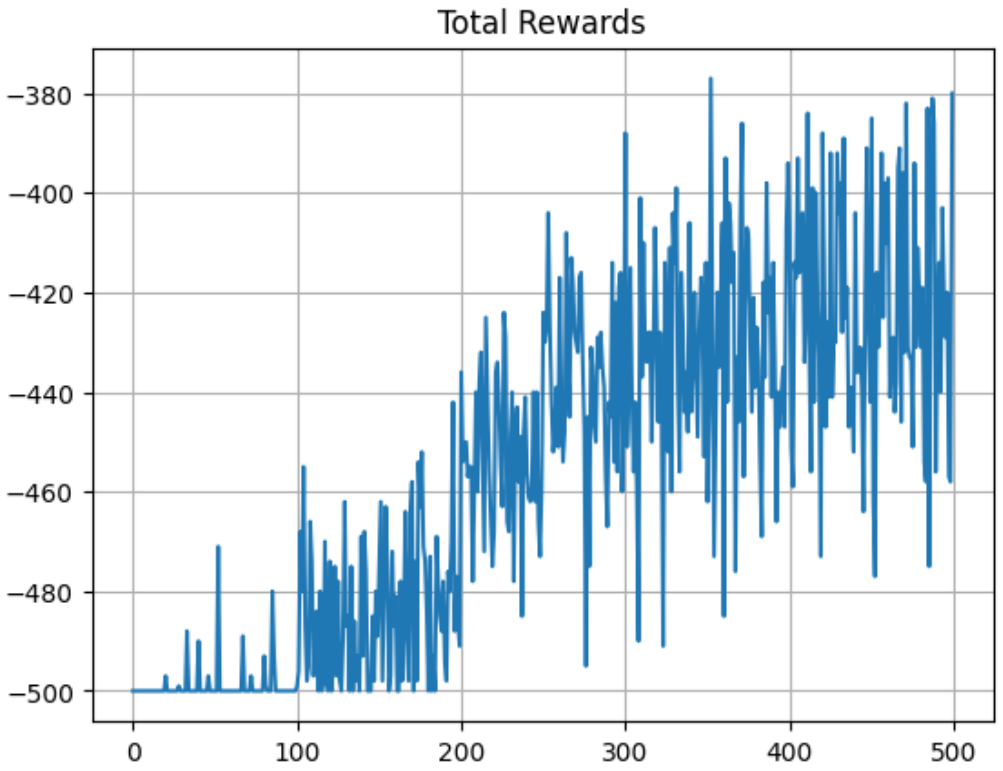
какие параметры выставить. gamma -0.999 epsilon=0.5, epoch_n=30, pi_lr=1e-3, v_lr=5e-3. После первого прогона модель вообще не обучилась, и я понял что у меня всего 10 траекторий и 10 эпизодов. Сделал 200 траекторий, и 20 эпизодов.



И оказалось что дело не в этом)

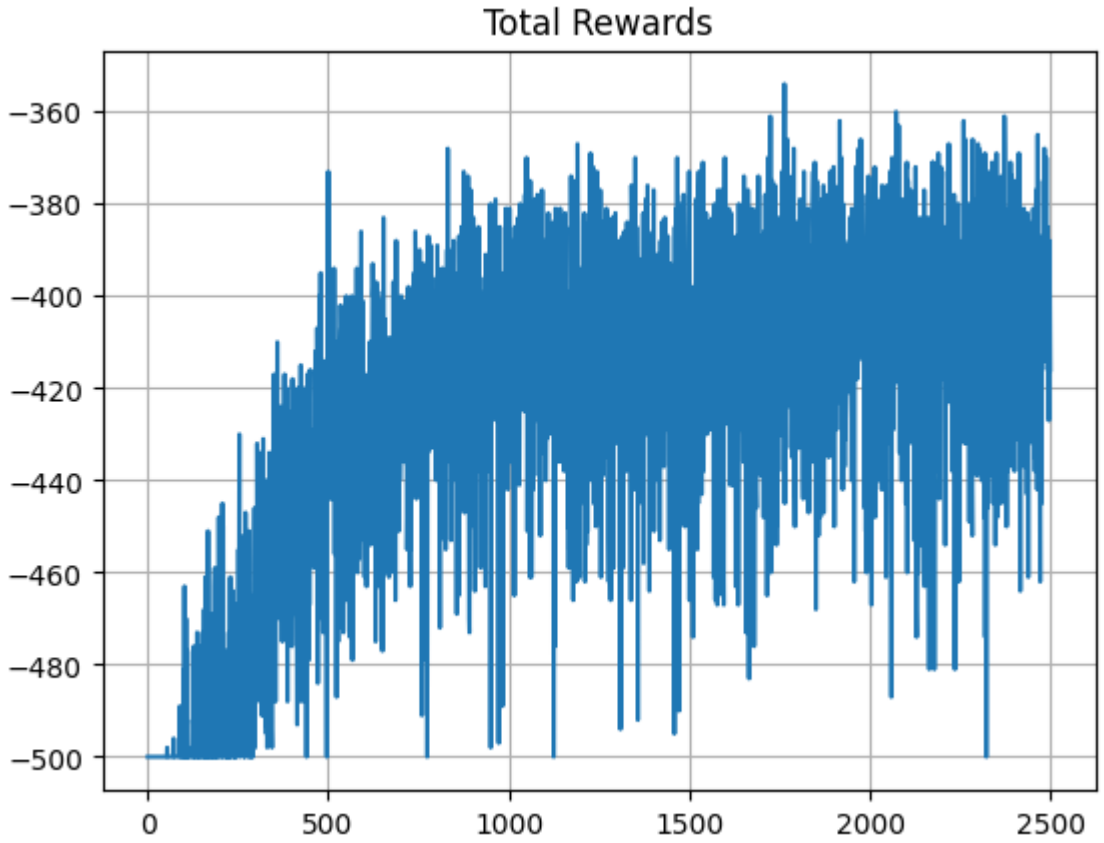
Эксперимент 2

Я задумался о параметрах в контексте ппо, и мне показалось что агент переобучается на неуспешных стратегиях, и поэтому решил ограничить батч сайз, и количество эпох. Вот что получилось:



Это что-то уже было похоже на обучение.
gamma=0.99, batch_size=32, epsilon=0.5, epoch_n=10, pi_lr=3e-4, v_lr=5e-3

Эксперимент 3

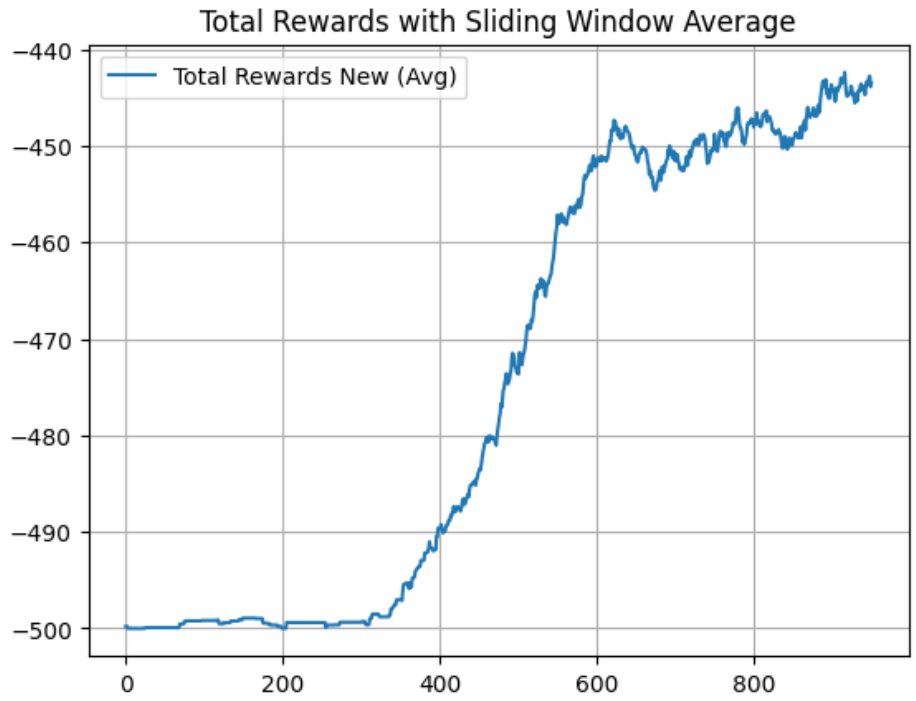


увидел количество траекторий. Все еще очень плохо.

Эксперимент 4

Тут я попал в тупик, и поиграл с большим количеством гиперпараметров в разных форматах. gamma=0.999, batch_size=128, epsilon=0.2, epoch_n=30, pi_lr=1e-4,

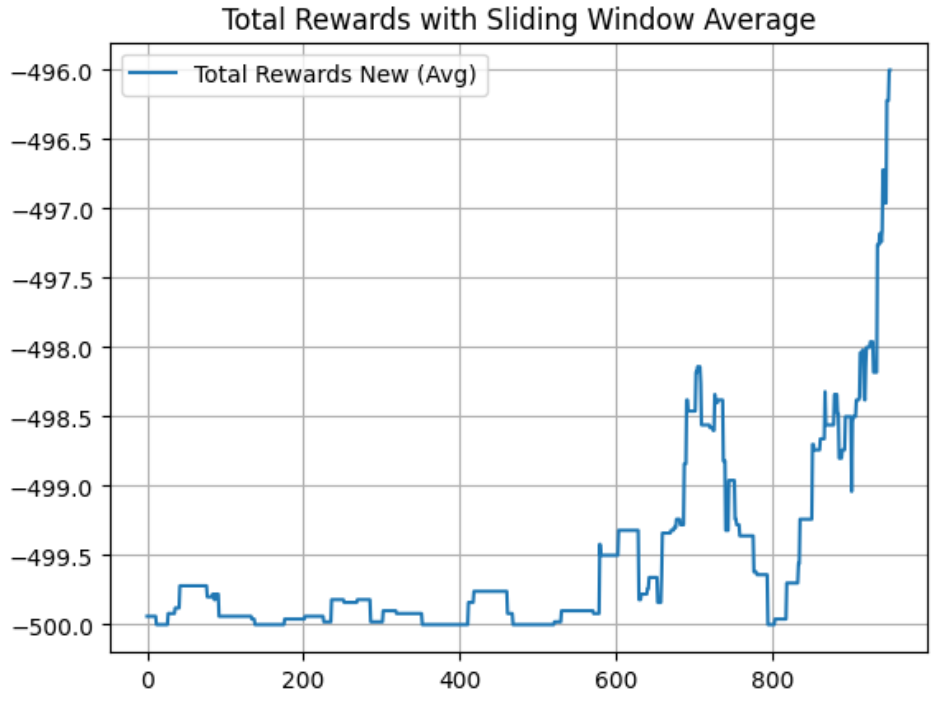
v_lr=5e-4. С этим набором параметров я получил вот такой результат:



Было видно какое-то обучение. Я решил увеличить rl до pi_lr=5e-4, v_lr=1e-3 и посмотреть что выйдет. Вышло то же самое. Посоветовали уменьшить количество нейронов. Пробую.

Эксперимент 5

Уменьшил количество нейронов до 16 в каждом слое. Тестирую со следующими параметрами:
gamma=0.999, batch_size=256, epsilon=0.2, epoch_n=30, pi_lr=1e-4, v_lr=5e-4.



Результат нулевой.

Эксперимент 6

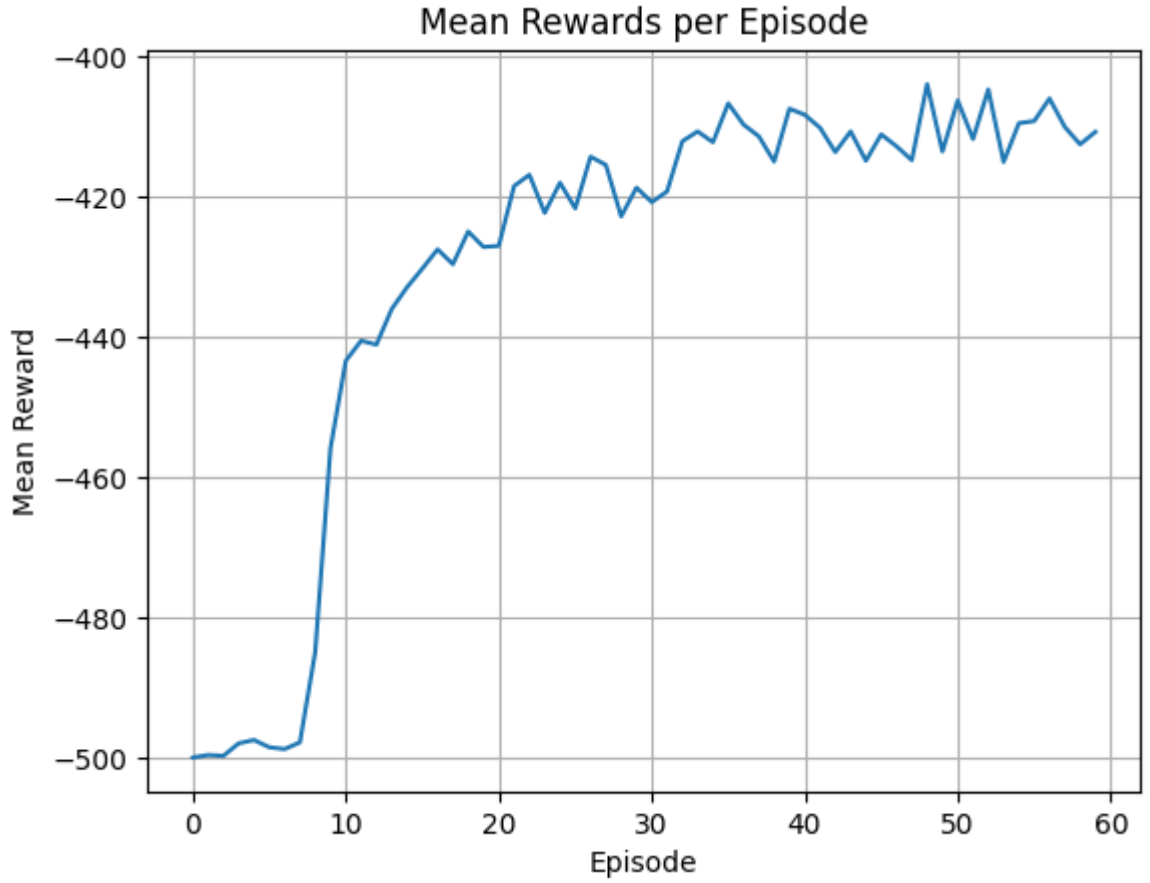
Переписал модель на ELU. Модель не обучилась. Решил бросить эксперимент с функциями активации.

Эксперимент 7

На этом этапе я перестал записывать коэффициенты. После бессоной ночи и мучения с разнообразными параметрами, я поменял количество нейронов на 256, и добавил энтропию. После запуска мне удалось достичь результатов в районе -170 ограничив длину траектории 200 на изначальных параметрах.

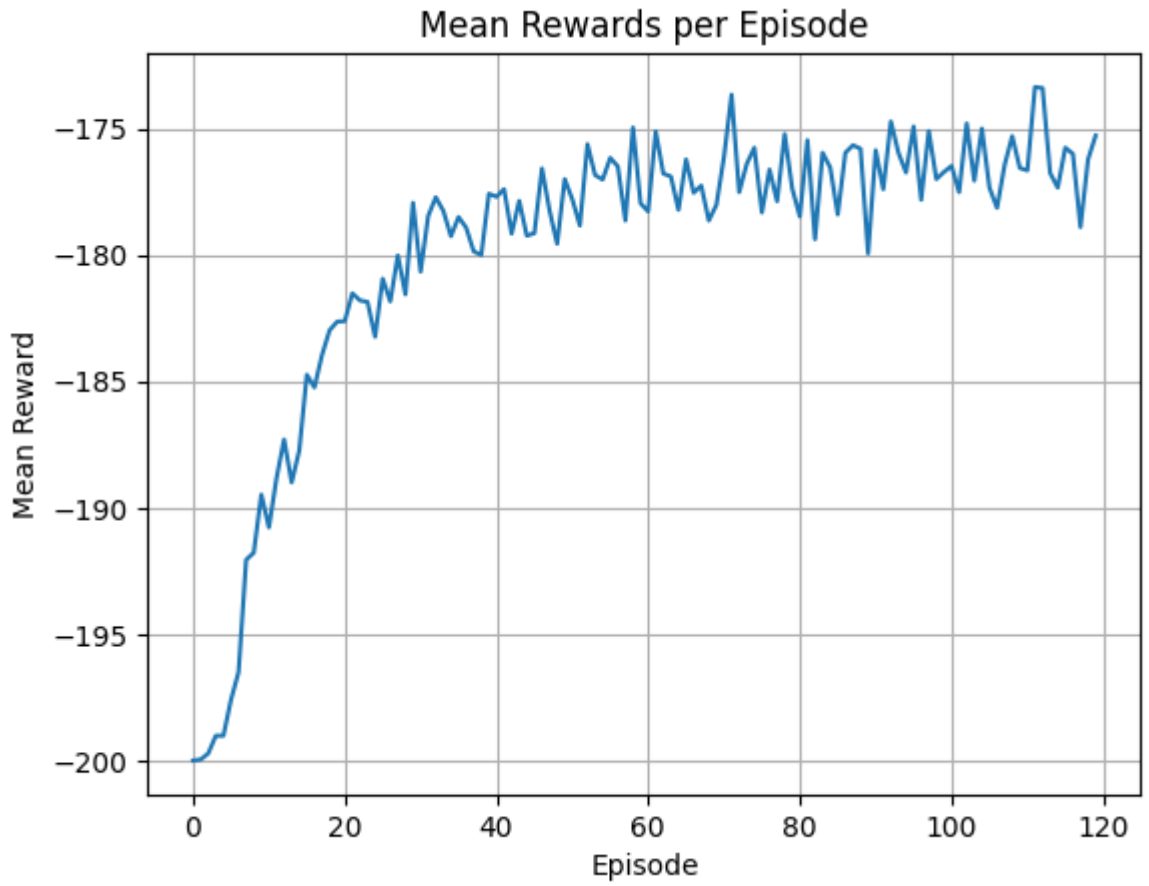
Эксперимент 8

Увеличил количество эпох. Поставил epsilon на 0.4, и e poch_n на 50. Поставил длину траекторий 500. Алгоритм как будто застрял на -400.



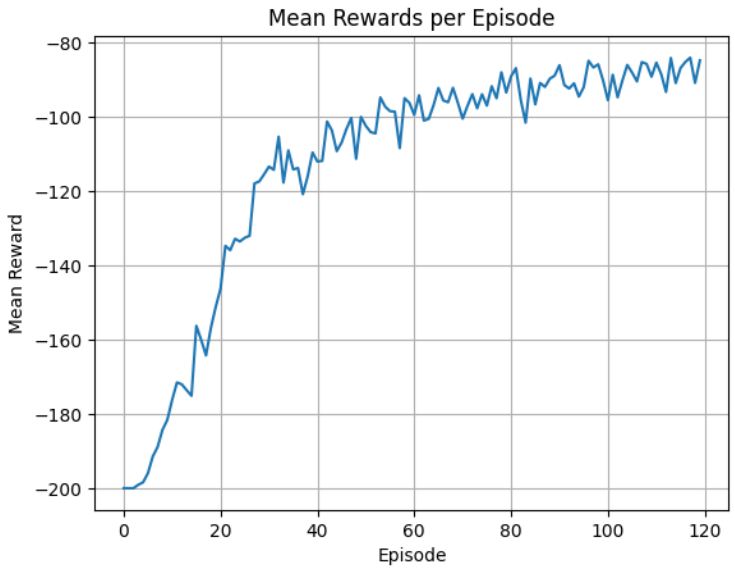
Эксперимент 9

Попробовал тоже самое, но с длиной эпохи 200:



Эксперимент 10

Обнаружил, что не сделал проверку на done)))))). После обучение пошло в разы быстрее и акробот решился)



Acrobot-v1 достаточно быстро обучился со следующими параметрами:

gamma=0.99, batch_size=64, epsilon=0.2, epoch_n=20, pi_lr=1e-4, v_lr=5e-4

Вывод

Я потратил порядка 20 часов играясь с параметрами, и кодом PPO для тренировки acrobot, но оказалась что вся проблема в том что я просто забыл проверку на done). Много нервов было потрачено, но зато и научился большому количеству новых вещей, и получше разобрал алгоритм. В целом, понял какие параметры на что влияют. Например, то что можно увеличивать exploration через batch_size или epsilon. А еще то как обучается PPO.