

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

ОТЧЕТ

по лабораторной работе №2

по дисциплине «Обучение с подкреплением»

Тема: Реализация РРО для среды MountainCarContinious-v0

Студент гр. 0306

Голубев А.Н.

Преподаватель

Глазунов С.А.

Санкт-Петербург
2025 г.

Цель работы.

Реализация PPO для среды MountainCarContinuous-v0. Исследование зависимости результатов от параметров и добавления нормализации преимуществ.

Задание.

1. Реализация PPO
2. Добавление нормализации
3. Изменение длины траектории
4. Подбор оптимального значения clip ratio
5. Сравнение обучения при разных количествах эпох

Выполнение работы.

1. Реализация PPO

Основными созданными сущностями для алгоритма PPO являются нейронные сети Actor и Critic.

Actor рассчитывает потенциальное действие, а Critic оценивает его выгоду (advantage) на основании полученной награды. Таким образом, Critic корректирует дальнейшие действия Actor.

Входными параметрами для Actor, так как он выбирает действие, являются Observation Space среды MountainCarContinuous-v0, описание которых представлено на рисунке 1, а на выходе численный показатель движения влево или вправо в зависимости от знака числа в диапазоне от [-1 до 1]: движение влево (индекс 0) или движения вправо (индекс 1).

Observation Space

The observation is a `ndarray` with shape `(2,)` where the elements correspond to the following:

Num	Observation	Min	Max	Unit
0	position of the car along the x-axis	-Inf	Inf	position (m)
1	velocity of the car	-Inf	Inf	position (m)

Рис. 1 – Пространство наблюдений в среде MountainCarContinuous-v0

Данная реализация алгоритма регулируется набором параметров:

- Число итераций (iterations)
- Количество шагов для сбора одного батча (steps)
- Число эпох (epochs)
- Размер батча (mini_batch_size)
- Коэффициент дисконтирования (gamma)
- GAE lambda – параметр, определяющий баланс между немедленным и будущим вознаграждением
- Коэффициент обрезки PPO (clip_ratio)
- Коэффициент потери значения (value_coef)

- Коэффициент энтропийного бонуса (entropy_coef)
- Шаг обучения при градиентном спуске (lr)

2. Добавление нормализации

Нормализация преимуществ в алгоритме PPO позволяет стабилизировать обучение и улучшить эффективность использования данных.

В рамках данной реализации алгоритма нормализация происходит в конце метода `__returns_and_advantages()` класса `PPOAgent`.

Сравнение результатов обучения с нормализацией и без представлено на рисунках 2 — 3.

```
Iteration: 24 | Loss: -0.6032 | Mean reward: -206.84904323534352 | Time spent: 1.83 s
Iteration: 25 | Loss: -0.1990 | Mean reward: -201.55297871516186 | Time spent: 1.88 s
Iteration: 26 | Loss: 0.0818 | Mean reward: -225.0447029790878 | Time spent: 1.94 s
Iteration: 27 | Loss: -0.3750 | Mean reward: -206.53539544849446 | Time spent: 1.92 s
Iteration: 28 | Loss: -0.1297 | Mean reward: -226.7285993815664 | Time spent: 1.86 s
Iteration: 29 | Loss: 0.1472 | Mean reward: -221.2277192950199 | Time spent: 1.89 s
Iteration: 30 | Loss: -0.6374 | Mean reward: -205.01321696939695 | Time spent: 1.81 s
Iteration: 31 | Loss: -0.5038 | Mean reward: -195.07944347854541 | Time spent: 1.84 s
Iteration: 32 | Loss: 0.2070 | Mean reward: -211.57030372309828 | Time spent: 1.89 s
Iteration: 33 | Loss: -0.1390 | Mean reward: -220.69147996586221 | Time spent: 1.85 s
Iteration: 34 | Loss: -0.6281 | Mean reward: -208.53779375490478 | Time spent: 1.92 s
Iteration: 35 | Loss: -0.0945 | Mean reward: -228.43166441639605 | Time spent: 1.91 s
Iteration: 36 | Loss: -0.4931 | Mean reward: -224.9224120612757 | Time spent: 1.94 s
Iteration: 37 | Loss: -0.7388 | Mean reward: -203.49495051623396 | Time spent: 2.01 s
Iteration: 38 | Loss: 0.5407 | Mean reward: -232.2865845597906 | Time spent: 1.88 s
Iteration: 39 | Loss: 39.5945 | Mean reward: -74.72419184781566 | Time spent: 1.88 s
Iteration: 40 | Loss: 1.5708 | Mean reward: -211.00023618230506 | Time spent: 1.87 s
Iteration: 41 | Loss: -0.2821 | Mean reward: -199.0323265109626 | Time spent: 1.84 s
Iteration: 42 | Loss: -0.4141 | Mean reward: -219.8329318098902 | Time spent: 1.85 s
Iteration: 43 | Loss: -0.2933 | Mean reward: -210.49961504071172 | Time spent: 1.93 s
Iteration: 44 | Loss: -0.4771 | Mean reward: -211.46279798061136 | Time spent: 1.89 s
Iteration: 45 | Loss: 0.0881 | Mean reward: -224.96232941436978 | Time spent: 1.93 s
Iteration: 46 | Loss: -0.8895 | Mean reward: -192.50291800668856 | Time spent: 1.89 s
Iteration: 47 | Loss: 0.4272 | Mean reward: -218.93173085293836 | Time spent: 1.94 s
Iteration: 48 | Loss: -0.5789 | Mean reward: -198.1098188338215 | Time spent: 1.86 s
Iteration: 49 | Loss: -0.2294 | Mean reward: -207.591478951351 | Time spent: 1.85 s
Iteration: 50 | Loss: -0.1058 | Mean reward: -222.0927500072103 | Time spent: 1.82 s
Iteration: 51 | Loss: 0.4561 | Mean reward: -230.60544772509203 | Time spent: 1.91 s
Iteration: 52 | Loss: -0.7292 | Mean reward: -213.9224534790485 | Time spent: 1.87 s
Iteration: 53 | Loss: -0.2322 | Mean reward: -219.03042173535255 | Time spent: 1.86 s
Iteration: 54 | Loss: -0.1184 | Mean reward: -210.75531463603932 | Time spent: 1.91 s
Iteration: 55 | Loss: -0.4462 | Mean reward: -207.4731783512224 | Time spent: 1.86 s
Iteration: 56 | Loss: -0.4197 | Mean reward: -215.73589199042468 | Time spent: 1.9 s
Iteration: 57 | Loss: -0.2487 | Mean reward: -215.9163060639307 | Time spent: 1.91 s
Iteration: 58 | Loss: -0.2488 | Mean reward: -204.38983486480402 | Time spent: 1.84 s
Iteration: 59 | Loss: 0.0805 | Mean reward: -218.10703074738603 | Time spent: 1.91 s
Iteration: 60 | Loss: -0.2814 | Mean reward: -210.0339484928812 | Time spent: 1.95 s
Iteration: 61 | Loss: -0.6502 | Mean reward: -189.2170526588036 | Time spent: 1.86 s
Iteration: 62 | Loss: -0.0195 | Mean reward: -194.75794638469122 | Time spent: 1.89 s
Iteration: 63 | Loss: -0.2522 | Mean reward: -200.7123251483544 | Time spent: 1.82 s
Iteration: 64 | Loss: 0.1574 | Mean reward: -207.77831166407472 | Time spent: 1.91 s
Iteration: 65 | Loss: -0.4603 | Mean reward: -206.1444484134564 | Time spent: 1.86 s
Iteration: 66 | Loss: -0.4148 | Mean reward: -207.03866040729142 | Time spent: 1.87 s
Iteration: 67 | Loss: -0.3760 | Mean reward: -201.68305751882514 | Time spent: 1.91 s
Iteration: 68 | Loss: -0.3586 | Mean reward: -204.73313910211664 | Time spent: 1.9 s
Iteration: 69 | Loss: -0.4061 | Mean reward: -196.22273018148297 | Time spent: 1.91 s
Iteration: 70 | Loss: 0.2656 | Mean reward: -232.84354391408363 | Time spent: 1.9 s
Iteration: 71 | Loss: -0.4822 | Mean reward: -198.63516245031687 | Time spent: 1.9 s
Iteration: 72 | Loss: -0.1150 | Mean reward: -202.8277453388093 | Time spent: 1.88 s
Iteration: 73 | Loss: -0.1090 | Mean reward: -222.9651534328708 | Time spent: 1.86 s
Iteration: 74 | Loss: 17.7716 | Mean reward: -51.29217388203041 | Time spent: 1.91 s
Iteration: 75 | Loss: 1.8960 | Mean reward: -228.19657235204429 | Time spent: 1.83 s
Iteration: 76 | Loss: -0.4912 | Mean reward: -207.05258421488253 | Time spent: 1.85 s
Iteration: 77 | Loss: -0.1245 | Mean reward: -221.23448689255602 | Time spent: 1.88 s
```

Рис. 2 — результаты без нормализации

Как можно заметить, в случае отсутствия нормализации средняя награда в среднем принимает значения ниже -200 и со временем не увеличивается (за исключением тех итераций, где машина достигает флага, из-за чего к награде прибавляется 100 очков).

Iteration: 129	Loss: -0.1844	Mean reward: -41.656723612421004	Time spent: 1.88 s
Iteration: 130	Loss: 0.6035	Mean reward: -43.74880970850795	Time spent: 1.93 s
Iteration: 131	Loss: 0.1430	Mean reward: -44.5213093193261	Time spent: 1.87 s
Iteration: 132	Loss: 0.2305	Mean reward: -37.26944666098181	Time spent: 1.92 s
Iteration: 133	Loss: -0.0773	Mean reward: -40.50172107839214	Time spent: 1.93 s
Iteration: 134	Loss: -0.0058	Mean reward: -40.18172271704222	Time spent: 1.87 s
Iteration: 135	Loss: -0.2016	Mean reward: -39.35499448829528	Time spent: 1.86 s
Iteration: 136	Loss: -0.0348	Mean reward: -39.898528820465955	Time spent: 1.9 s
Iteration: 137	Loss: 0.2536	Mean reward: -39.24405599450196	Time spent: 1.89 s
Iteration: 138	Loss: -0.1306	Mean reward: -37.722142927587655	Time spent: 1.84 s
Iteration: 139	Loss: 0.1706	Mean reward: -36.877363872448235	Time spent: 1.87 s
Iteration: 140	Loss: 0.4025	Mean reward: -38.798614837432055	Time spent: 1.87 s
Iteration: 141	Loss: 0.0472	Mean reward: -40.25571020505212	Time spent: 1.91 s
Iteration: 142	Loss: 0.1100	Mean reward: -42.48272393560197	Time spent: 1.88 s
Iteration: 143	Loss: 0.1571	Mean reward: -37.508349737893624	Time spent: 1.87 s
Iteration: 144	Loss: 0.0517	Mean reward: -36.55315055077516	Time spent: 1.88 s
Iteration: 145	Loss: 0.1499	Mean reward: -39.081067700221176	Time spent: 1.83 s
Iteration: 146	Loss: 0.1630	Mean reward: -36.93060275482153	Time spent: 1.89 s
Iteration: 147	Loss: 0.3190	Mean reward: -39.538370832714534	Time spent: 1.89 s
Iteration: 148	Loss: 0.2812	Mean reward: -39.520807635545644	Time spent: 1.85 s
Iteration: 149	Loss: 0.0213	Mean reward: -39.043481733432564	Time spent: 1.88 s
Iteration: 150	Loss: 0.0990	Mean reward: -38.471340699682166	Time spent: 1.9 s
Iteration: 151	Loss: 0.2925	Mean reward: -37.762217107505776	Time spent: 1.9 s
Iteration: 152	Loss: 0.1943	Mean reward: -37.49490127529115	Time spent: 1.95 s
Iteration: 153	Loss: -0.0950	Mean reward: -36.09536094569781	Time spent: 1.82 s
Iteration: 154	Loss: -0.0431	Mean reward: -35.535455461187595	Time spent: 1.9 s
Iteration: 155	Loss: -0.3742	Mean reward: -35.1356721208975	Time spent: 1.91 s
Iteration: 156	Loss: -0.2419	Mean reward: -38.20630394581303	Time spent: 1.84 s
Iteration: 157	Loss: 0.0231	Mean reward: -37.90563318002264	Time spent: 1.83 s
Iteration: 158	Loss: 0.0291	Mean reward: -32.64551632119577	Time spent: 1.9 s
Iteration: 159	Loss: 0.4384	Mean reward: -34.4084386081409	Time spent: 1.88 s
Iteration: 160	Loss: 0.0565	Mean reward: -34.35936493457103	Time spent: 1.86 s
Iteration: 161	Loss: 0.3921	Mean reward: -35.977873200091175	Time spent: 1.83 s
Iteration: 162	Loss: 0.0271	Mean reward: -34.60545349351257	Time spent: 1.86 s
Iteration: 163	Loss: 0.3246	Mean reward: -31.680321801934394	Time spent: 1.95 s
Iteration: 164	Loss: 0.2805	Mean reward: -32.109831579733275	Time spent: 1.89 s
Iteration: 165	Loss: -0.0093	Mean reward: -34.54477557828681	Time spent: 1.88 s
Iteration: 166	Loss: 0.6740	Mean reward: -34.30789467373141	Time spent: 1.89 s
Iteration: 167	Loss: 0.4813	Mean reward: -34.60191746950138	Time spent: 1.83 s
Iteration: 168	Loss: 0.2479	Mean reward: -34.840420339068196	Time spent: 1.89 s
Iteration: 169	Loss: 0.7217	Mean reward: -33.54076987019097	Time spent: 1.96 s
Iteration: 170	Loss: -0.1477	Mean reward: -35.23501669177458	Time spent: 1.84 s
Iteration: 171	Loss: 0.1211	Mean reward: -33.03988867932562	Time spent: 1.84 s
Iteration: 172	Loss: -0.0522	Mean reward: -33.46522200439256	Time spent: 1.85 s
Iteration: 173	Loss: -0.0624	Mean reward: -32.83659791299138	Time spent: 1.85 s
Iteration: 174	Loss: 0.4164	Mean reward: -32.652203366604155	Time spent: 1.92 s
Iteration: 175	Loss: 0.0664	Mean reward: -30.72264624877793	Time spent: 1.87 s
Iteration: 176	Loss: 0.1005	Mean reward: -33.851717391652	Time spent: 1.88 s
Iteration: 177	Loss: 0.3236	Mean reward: -33.72421855762402	Time spent: 1.9 s
Iteration: 178	Loss: 0.0877	Mean reward: -32.30954867419753	Time spent: 1.87 s
Iteration: 179	Loss: 0.1450	Mean reward: -33.558706599746685	Time spent: 1.85 s
Iteration: 180	Loss: 0.6046	Mean reward: -33.27457281487586	Time spent: 1.9 s
Iteration: 181	Loss: -0.3085	Mean reward: -29.087201450698153	Time spent: 1.86 s

Рис. 3 — результаты при наличии нормализации

В случае наличия нормализации среднее значение награды значительно возросло. Также можно заметить, что в данном случае значения средней награды постепенно растут (на первой итерации средняя награда равнялась -115).

Все следующие эксперименты проводились с наличием нормализации.

3. Изменение длины траектории

Для эксперимента были выбраны следующие длины траекторий: 1024, 2048, 4096.

Результаты обучения представлены на рисунках 4 — 6.

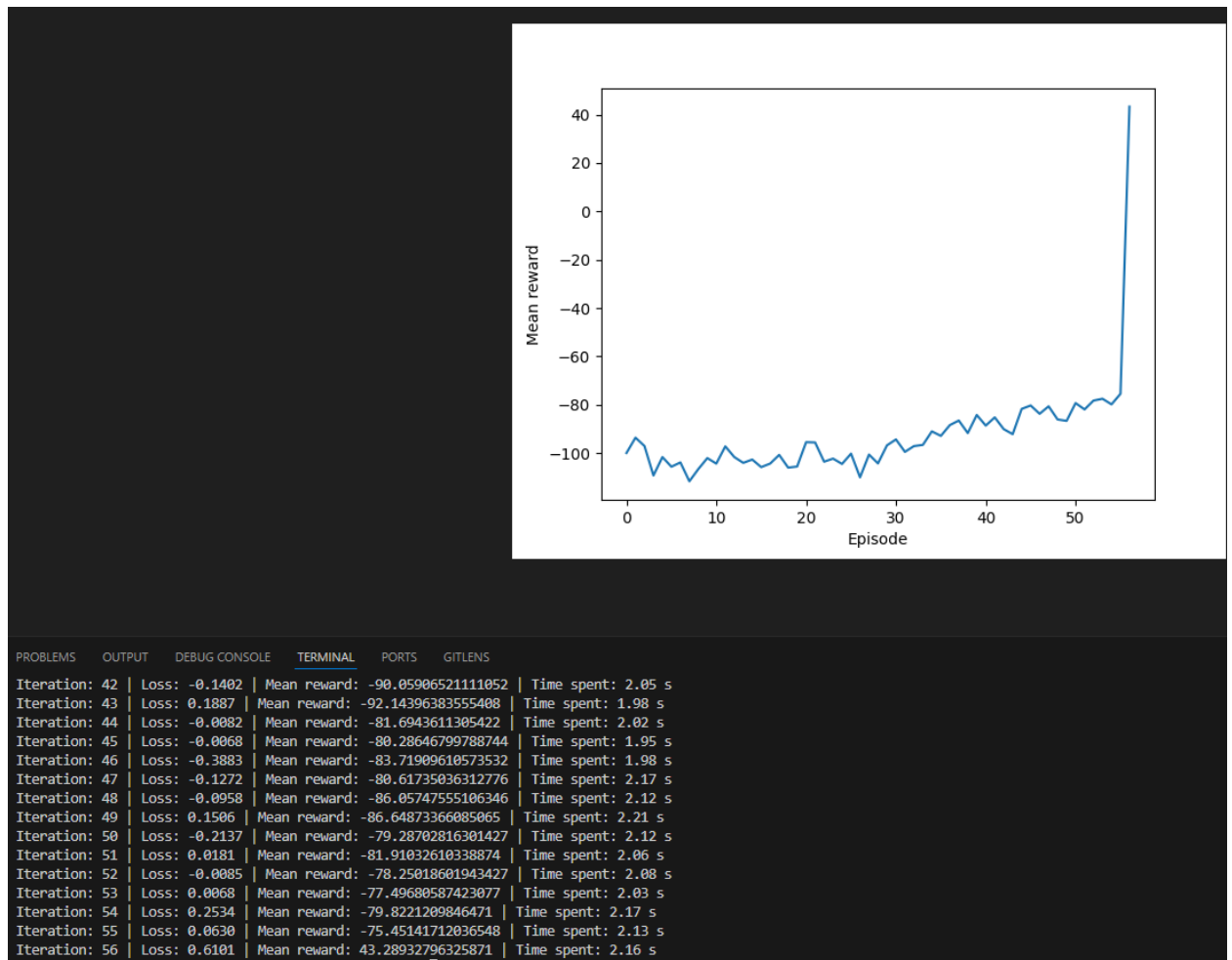


Рис. 4 — результат обучения при steps = 1024

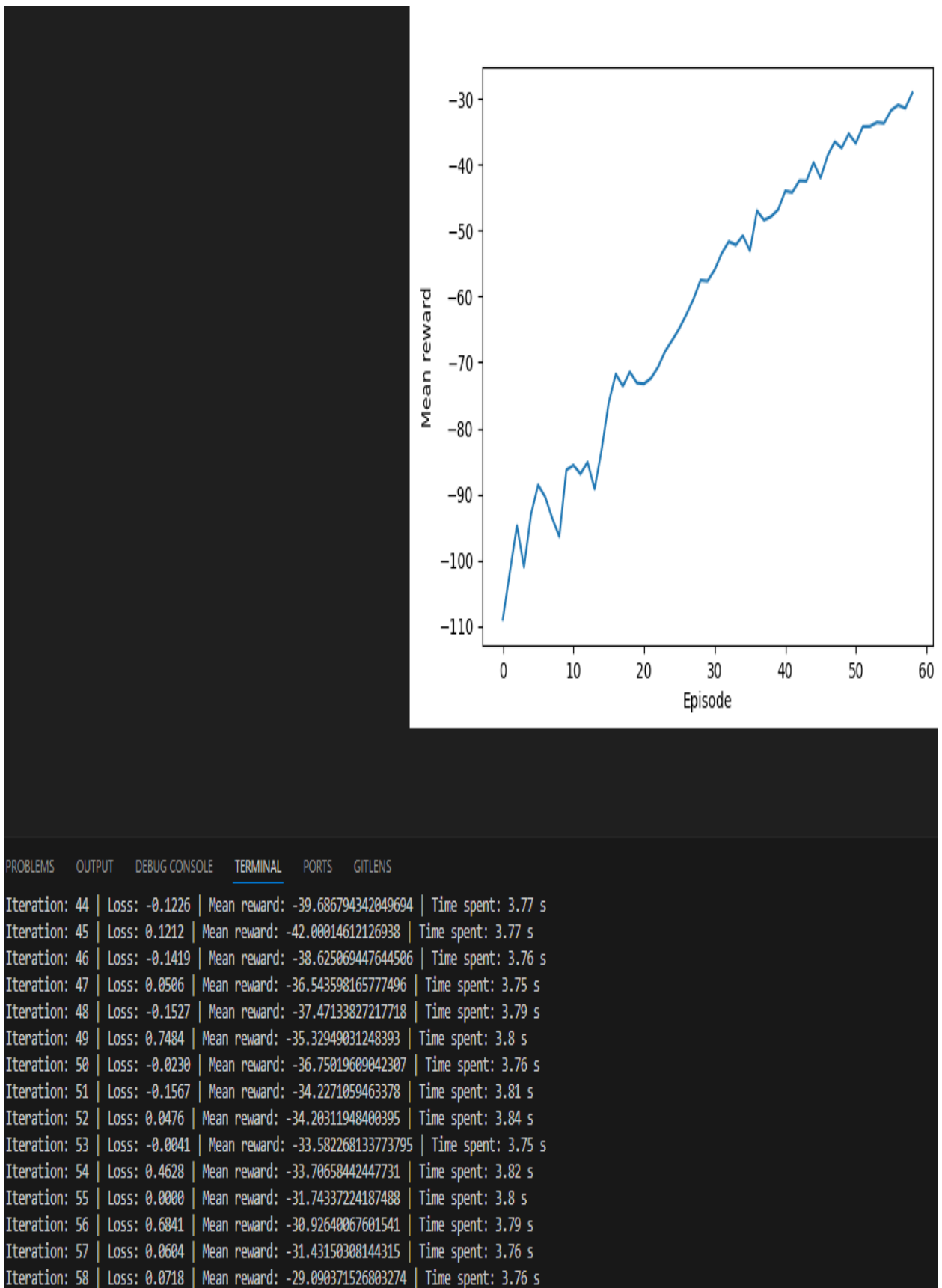


Рис. 5 — результат обучения при steps = 2048

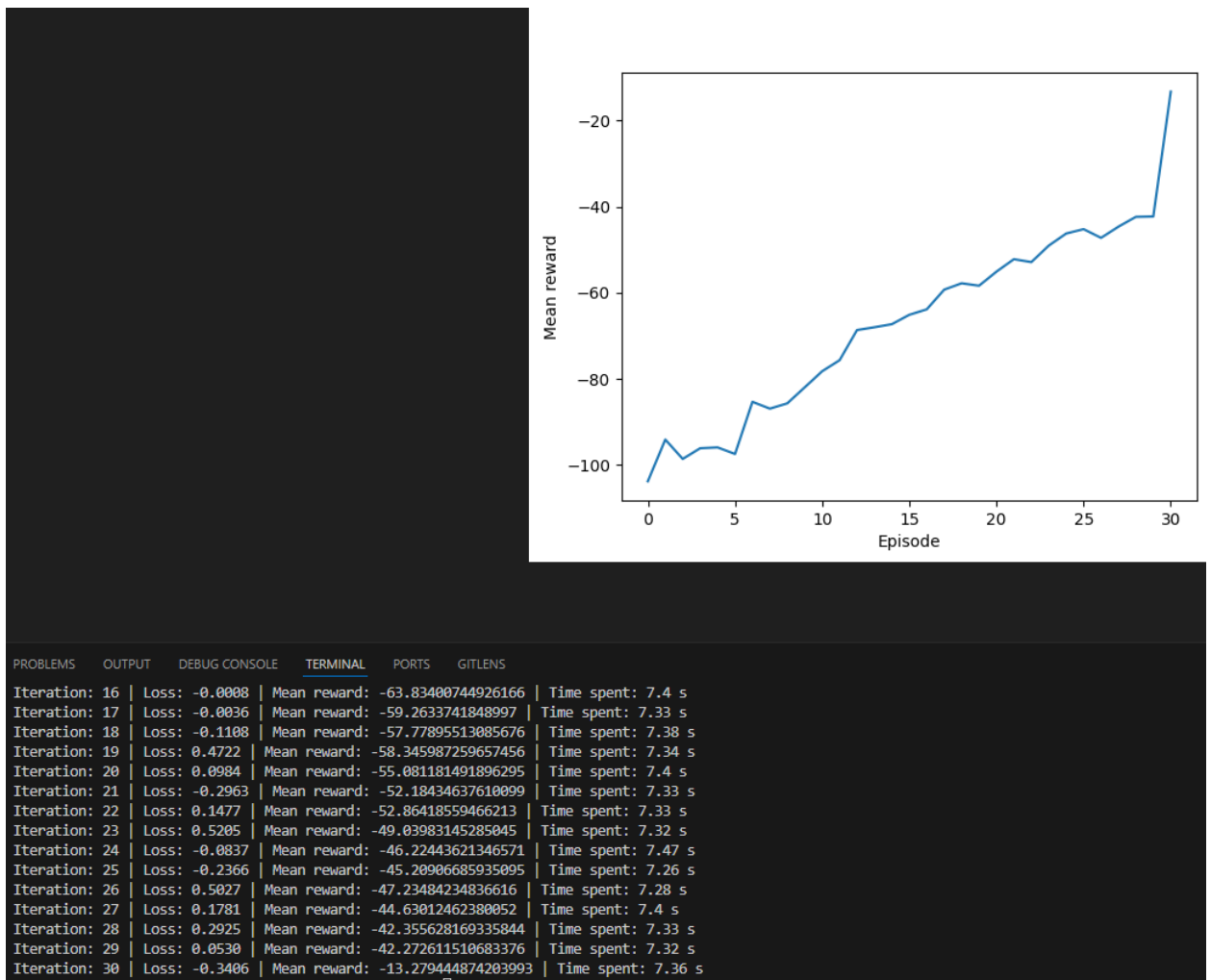


Рис. 6 — результат обучения при steps = 4096

По результатам обучения можно заметить следующее:

- При увеличении steps время, затрачиваемое на одну итерацию, тоже растет
- По мере увеличения steps уменьшается резкость и частота скачков в значениях средней награды

4. Подбор оптимального значения clip ratio

Для данного эксперимента были выбраны следующие значения clip ratio: 0.08, 0.2, 0.5. Результаты продемонстрированы на рисунках 7 — 9.

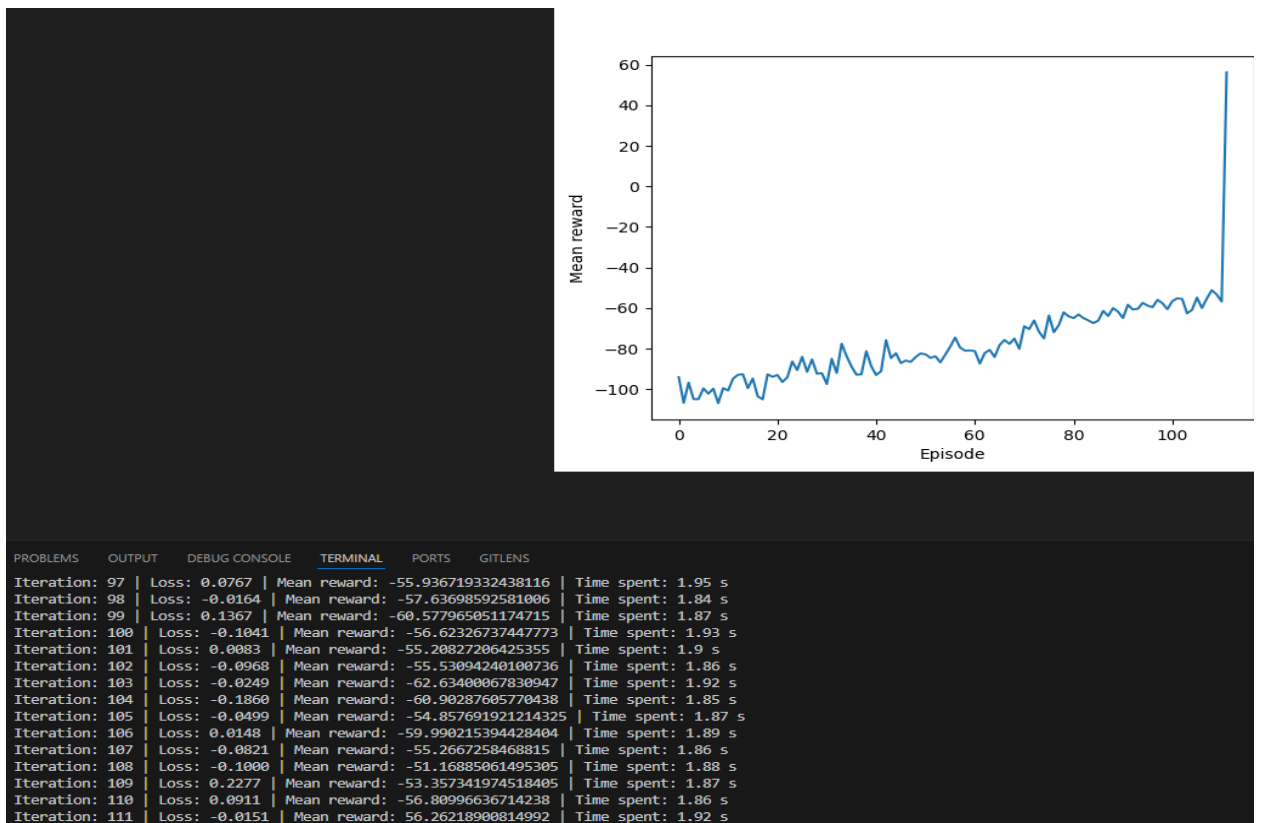


Рис. 7 — результат обучения при $\text{clip_ratio} = 0.08$

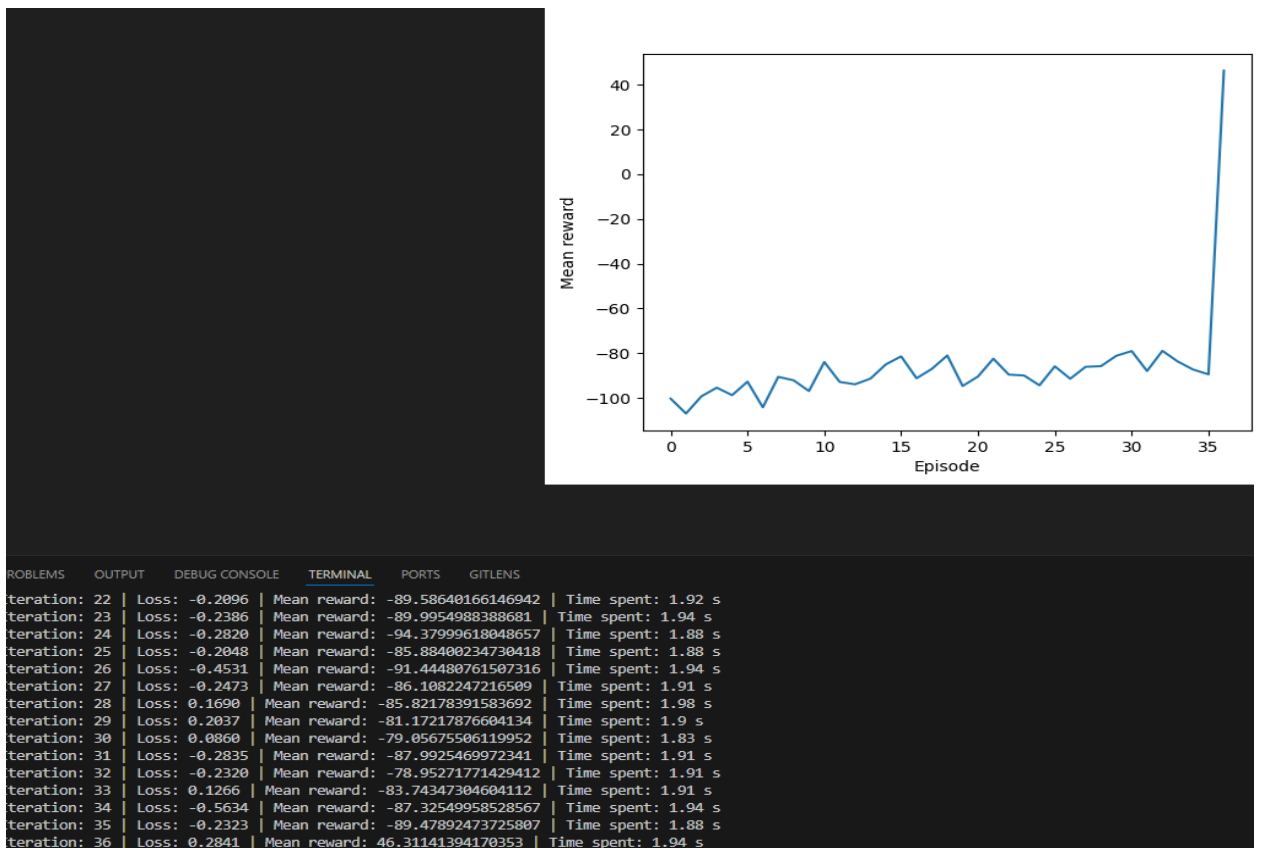


Рис. 8 — результат обучения при $\text{clip_ratio} = 0.2$

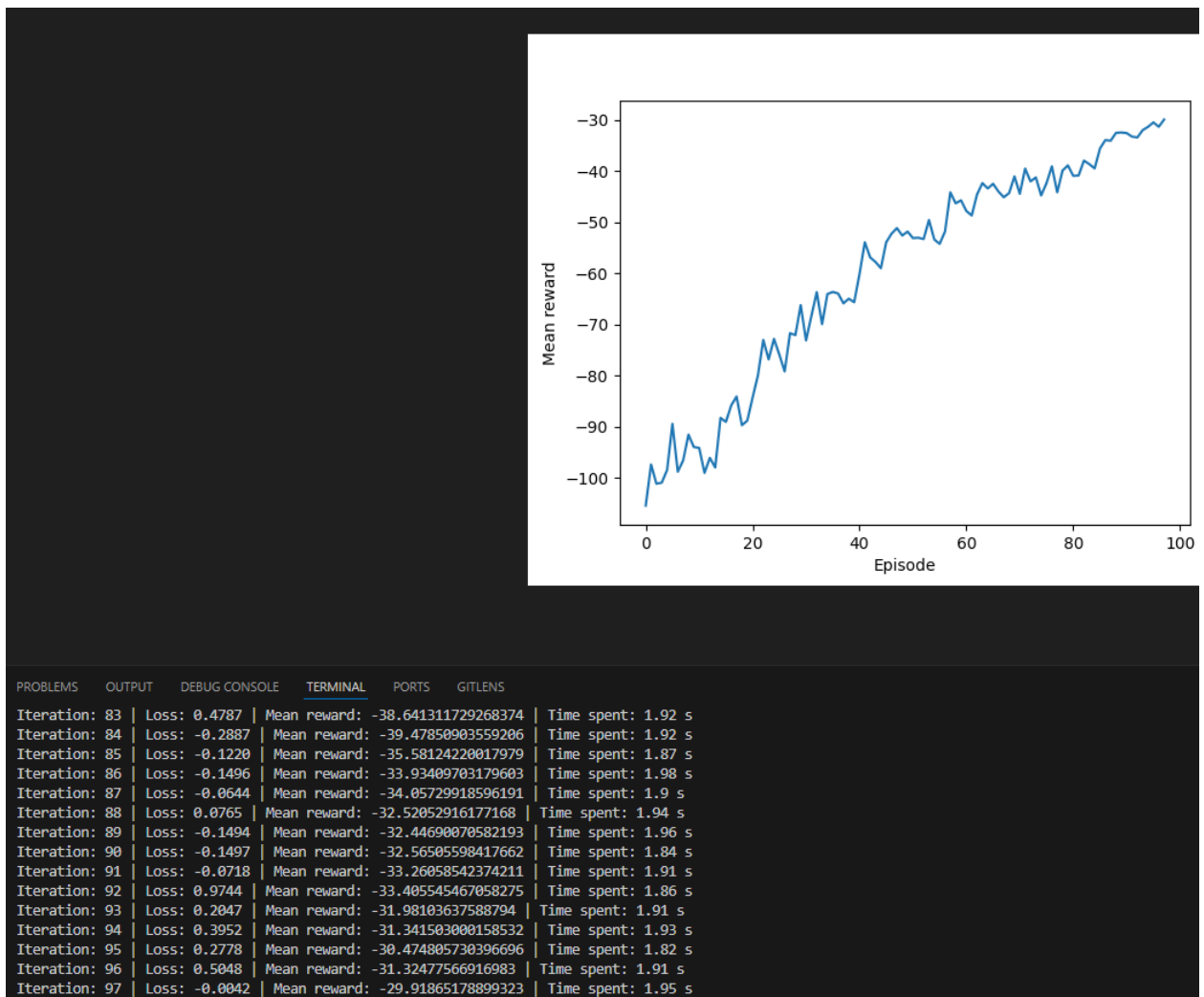


Рис. 9 — результат обучения при $\text{clip_ratio} = 0.5$

Можно сделать следующие выводы:

- Быстрее всего пород для окончания обучения был пройден при $\text{clip_ratio} = 0.2$
- Больше всего итераций потребовалось при $\text{clip_ratio} = 0.08$
- При $\text{clip_ratio} = 0.5$ обучение проходило без сильных скачков в среднем значении награды
- При $\text{clip_ratio} = 0.2$ машина в режиме тестирования чаще достигает цели (флага на холме)
- При $\text{clip_ratio} = 0.5$ машина в режиме тестирования относительно чаще предыдущих случаев находилась в районе дна котлована

Из всего вышеперечисленного можно сделать вывод, что оптимальным является значение `clip_ratio = 0.2`.

5. Сравнение обучения при разных количествах эпох

Были взяты следующие значения количества эпох: 10, 20, 30. Результаты представлены на рисунках 10 — 12.

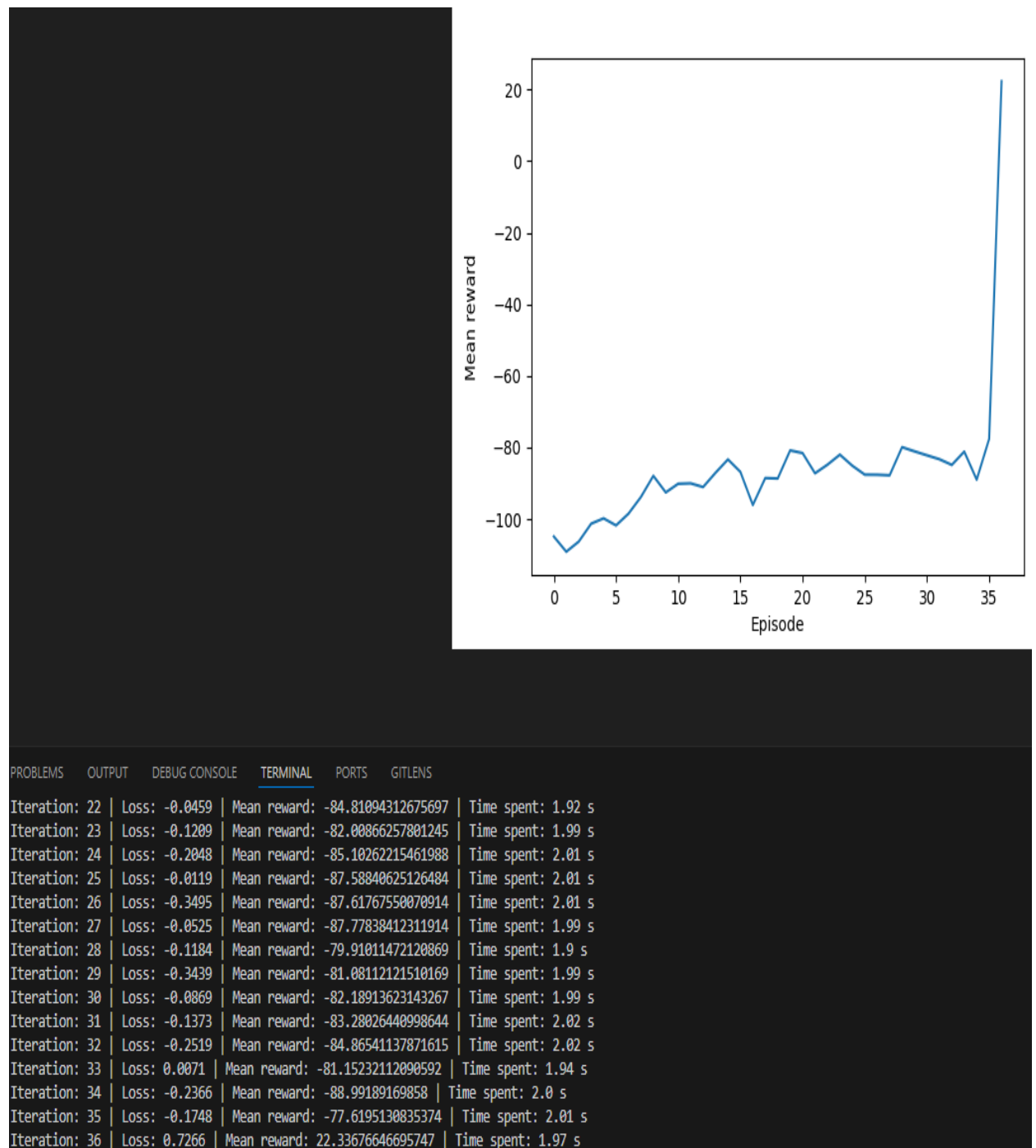


Рис. 10 — результат обучения при `epochs = 10`

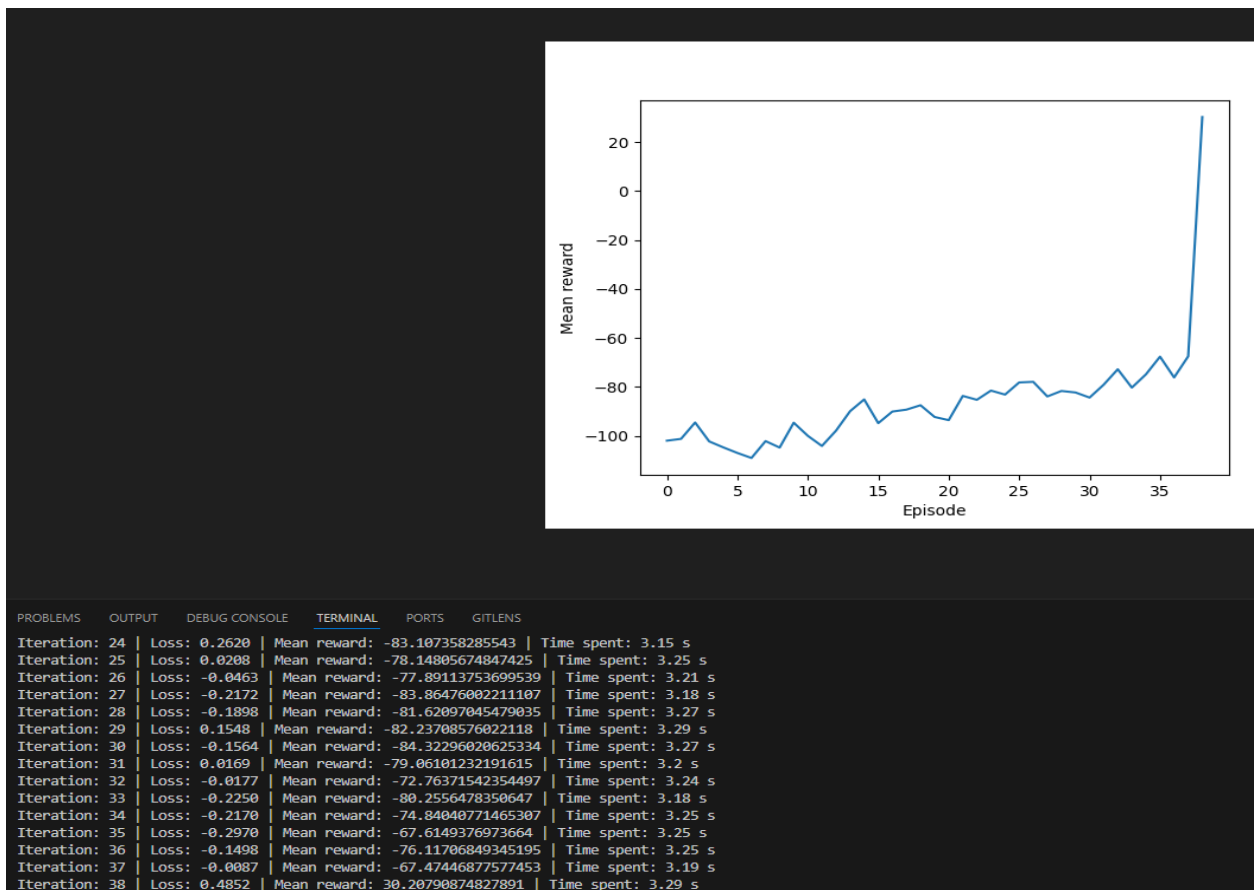


Рис. 11 — результат обучения при epochs = 20

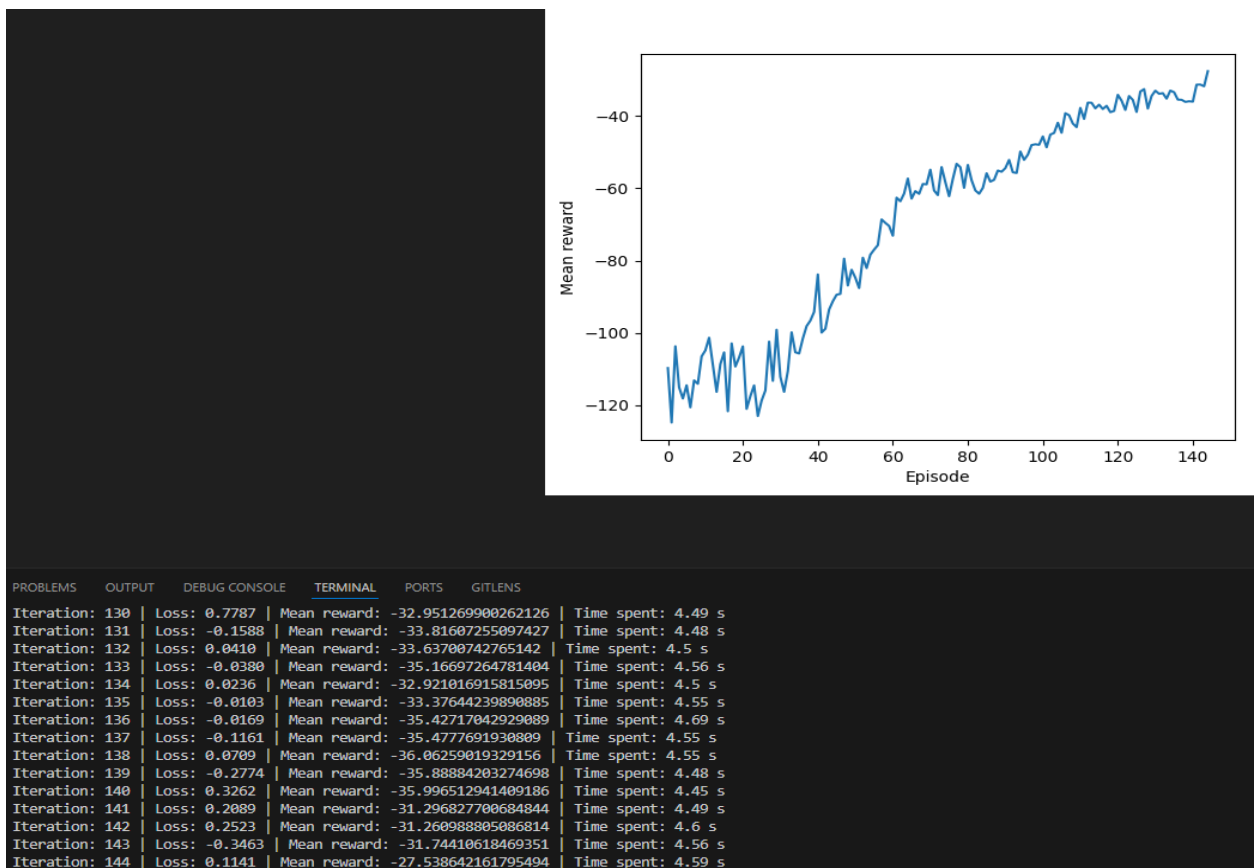


Рис. 12 — результат обучения при epochs = 30

По результатам можно заметить следующее:

- По мере увеличения числа эпох растет среднее время итерации
- При `epochs = 30` не было таких же резких скачков значений, как при `epochs = 10` и `epochs = 20`

Выводы.

Была выполнена реализация PPO для среды MountainCarContinuous-v0. Было проведено исследование влияния изменения параметров алгоритма на результаты, а также эффективность применения нормализации к значениям advantages.