PPO. Отчет

Содержание

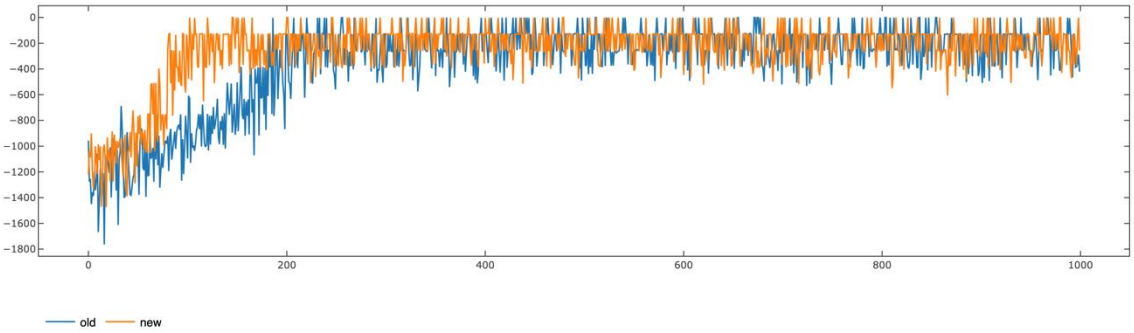
1. Как было сказано на занятиях. Advantage функцию в PPO можно считать и учить по-разному. В задании предлагается написать и исследовать другой способ делать это. А именно использовать представление A(s,a) = r + \gamma V(s') - V(s), где s' - следующее состояние. То есть returns в данном случае использовать не нужно. Необходимо сравнить кривые обучения алгоритма с этим “новым” способом и “старым” способом (из практики) на задаче Pendulum.

Игра: **Pendulum-v1**
Результаты можно посмотреть в ClearML

Параметры для PPO старого и нового алгоритма Advatage функции были выбраны следующие:

- **Episode n:** 50
- **Trajectory n:** 20
- **Gamma:** 0.9
- **Batch size:** 128
- **Epsilon:** 0.2
- **Epoch n:** 30
- **Pi learning rate:** 1e-4
- **V learning rate:** 5e-4

Reward на траекториях при обучении:



Вывод:
Обе реализации advantage функции показывают хорошие результаты. Однако новая реализация сходится и выходит на плато быстрее.

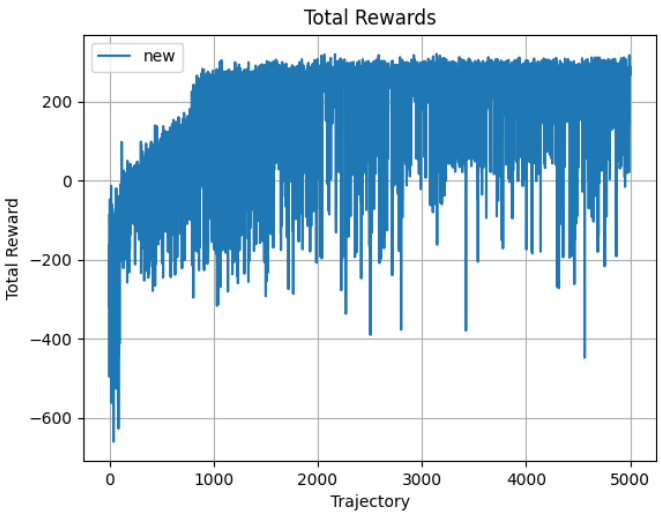
2. На практике мы написали PPO для случая одномерного пространства действий. Использование же его для многомерного пространства действий требует небольших технических изменений в коде (при этом содержательно ничего не меняется). Задание заключается в том, чтобы внести эти изменения (т.е. модифицировать PPO для работы в средах с многомерным пространством действий) и решить с его помощью LunarLander (результат должен быть больше 100). Для того, чтобы сделать LunarLander с непрерывным пространством действий нужно положить continuous=True (см. пояснения в Lunar Lander - Gym Documentation (gymlibrary.dev))

Игра: **LunarLander-v2.**
К сожалению, в виртуальном окружении к этой игре возникли проблемы с трекингом экспериментов в clear-ml (возможно из-за дополнительно установленных пакетов), поэтому логировал и трекал на локальную машину.

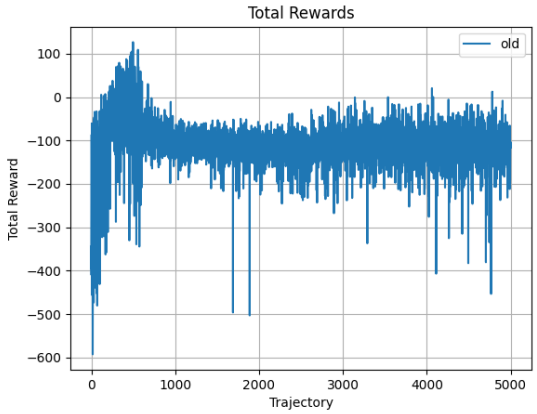
Исследованы следующие гиперпараметры (приведены те, которые дали лучший результат):

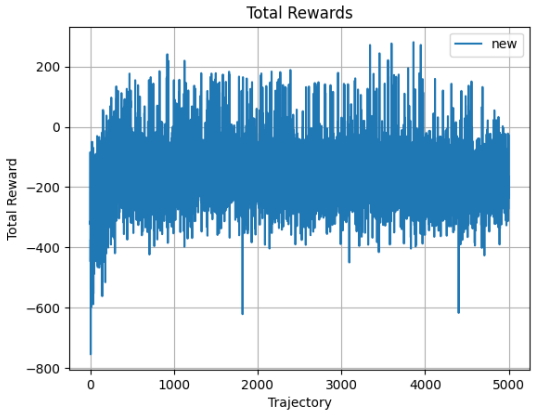
- **Episode n:** 50
- **Trajectory n:** 100
- **Gamma:** 0.99
- **Batch size:** 128
- **Epsilon:** 0.2
- **Epoch n:** 100
- **Pi learning rate:** 1e-4
- **V learning rate:** 5e-4
- **Advantage function:** new

Reward на траекториях при обучении:



Примеры неудачных запусков (Reward на траекториях при обучении):





Вывод:

При обучении на лучших гиперпараметрах ревард смог выйти на плато и обеспечить результат в районе +- 200.

Пример reward на последнем эпизоде:

- Episode: 49 - trajectory: 0: 298.04850823465097
- Episode: 49 - trajectory: 1: 30.5109868721201
- Episode: 49 - trajectory: 2: 38.85822971259918
- Episode: 49 - trajectory: 3: 257.59145315437786
- Episode: 49 - trajectory: 4: 192.79877122587445
- Episode: 49 - trajectory: 5: 159.4917831540259
- Episode: 49 - trajectory: 6: 299.0082929118528
- Episode: 49 - trajectory: 7: 279.87101733735767
- Episode: 49 - trajectory: 8: 276.44301653973673
- Episode: 49 - trajectory: 9: 287.2913271052229
- Episode: 49 - trajectory: 10: 302.05596116649946
- Episode: 49 - trajectory: 11: 282.8418261981436
- Episode: 49 - trajectory: 12: 36.05369477313738
- Episode: 49 - trajectory: 13: 286.88957682093235
- Episode: 49 - trajectory: 14: 293.7818199543438
- Episode: 49 - trajectory: 15: 310.13682032211597
- Episode: 49 - trajectory: 16: 268.93508521852254
- Episode: 49 - trajectory: 17: 227.69471846223254
- Episode: 49 - trajectory: 18: 283.6055068352467
- Episode: 49 - trajectory: 19: 252.12889763260884
- Episode: 49 - trajectory: 20: 279.39632960475024
- Episode: 49 - trajectory: 21: 274.15750575505194
- Episode: 49 - trajectory: 22: 289.1676266540035
- Episode: 49 - trajectory: 23: 293.2112610572964
- Episode: 49 - trajectory: 24: 240.96338803493782
- Episode: 49 - trajectory: 25: 307.281690211956
- Episode: 49 - trajectory: 26: 253.28019312023173
- Episode: 49 - trajectory: 27: 276.93046970953117
- Episode: 49 - trajectory: 28: 303.56084922383786

- Episode: 49 - trajectory: 29: 265.56128626973816
- Episode: 49 - trajectory: 30: 249.50713278294788
- Episode: 49 - trajectory: 31: 296.2453557865823
- Episode: 49 - trajectory: 32: 293.37214319509053
- Episode: 49 - trajectory: 33: 307.71598984093265
- Episode: 49 - trajectory: 34: 275.13458600479424
- Episode: 49 - trajectory: 35: 252.22223309835212
- Episode: 49 - trajectory: 36: 260.9878538933613
- Episode: 49 - trajectory: 37: 275.1839105546106
- Episode: 49 - trajectory: 38: 276.01678234406245
- Episode: 49 - trajectory: 39: 65.2801364101972
- Episode: 49 - trajectory: 40: 311.75109478767547
- Episode: 49 - trajectory: 41: 264.5498146808037
- Episode: 49 - trajectory: 42: 240.5736674393561
- Episode: 49 - trajectory: 43: 294.1433241893394
- Episode: 49 - trajectory: 44: 280.31559072084843
- Episode: 49 - trajectory: 45: 308.25365783440134
- Episode: 49 - trajectory: 46: 272.9618360937654
- Episode: 49 - trajectory: 47: 8.779743335374604
- Episode: 49 - trajectory: 48: 301.3219252412596
- Episode: 49 - trajectory: 49: 293.58952695726185
- Episode: 49 - trajectory: 50: 270.4428721038853
- Episode: 49 - trajectory: 51: -15.507204520529982
- Episode: 49 - trajectory: 52: 279.4353566166576
- Episode: 49 - trajectory: 53: 286.5994841426291
- Episode: 49 - trajectory: 54: 53.67460575845365
- Episode: 49 - trajectory: 55: 272.941676716466
- Episode: 49 - trajectory: 56: 307.56874183017305
- Episode: 49 - trajectory: 57: 64.87617010109781
- Episode: 49 - trajectory: 58: 259.30021741003515
- Episode: 49 - trajectory: 59: 267.0257700280539
- Episode: 49 - trajectory: 60: 253.15376127977547
- Episode: 49 - trajectory: 61: 251.47679638225233
- Episode: 49 - trajectory: 62: 102.3531864859477
- Episode: 49 - trajectory: 63: 253.15235883927122
- Episode: 49 - trajectory: 64: 287.74056230445456
- Episode: 49 - trajectory: 65: 19.551628211603003
- Episode: 49 - trajectory: 66: 270.3775812091259
- Episode: 49 - trajectory: 67: 251.55914708862323
- Episode: 49 - trajectory: 68: 271.95605478193914
- Episode: 49 - trajectory: 69: 274.7321337341608
- Episode: 49 - trajectory: 70: 278.6962584413565
- Episode: 49 - trajectory: 71: 284.85290173712
- Episode: 49 - trajectory: 72: 252.3119075468812
- Episode: 49 - trajectory: 73: 272.58744798968223
- Episode: 49 - trajectory: 74: 270.571094506558
- Episode: 49 - trajectory: 75: 254.2849482866414

- Episode: 49 - trajectory: 76: 259.95335687423983
- Episode: 49 - trajectory: 77: 283.8852780560064
- Episode: 49 - trajectory: 78: 281.5614691003454
- Episode: 49 - trajectory: 79: 274.5926669300102
- Episode: 49 - trajectory: 80: 270.21947590118134
- Episode: 49 - trajectory: 81: 275.25261111572803
- Episode: 49 - trajectory: 82: 259.3658937661768
- Episode: 49 - trajectory: 83: 239.86385596886157
- Episode: 49 - trajectory: 84: 255.1961034517897
- Episode: 49 - trajectory: 85: 272.4170853723314
- Episode: 49 - trajectory: 86: 257.71395461367763
- Episode: 49 - trajectory: 87: 20.93712026729368
- Episode: 49 - trajectory: 88: 178.9069086514275
- Episode: 49 - trajectory: 89: 282.9493150513832
- Episode: 49 - trajectory: 90: 280.91494269749705
- Episode: 49 - trajectory: 91: 281.41282640850324
- Episode: 49 - trajectory: 92: 317.2595976486177
- Episode: 49 - trajectory: 93: 142.51225037377165
- Episode: 49 - trajectory: 94: 295.73671090642836
- Episode: 49 - trajectory: 95: 265.6062587585964
- Episode: 49 - trajectory: 96: 281.5836413497775
- Episode: 49 - trajectory: 97: 271.71457782588317
- Episode: 49 - trajectory: 98: 289.8902076140227
- Episode: 49 - trajectory: 99: 268.31767981806126

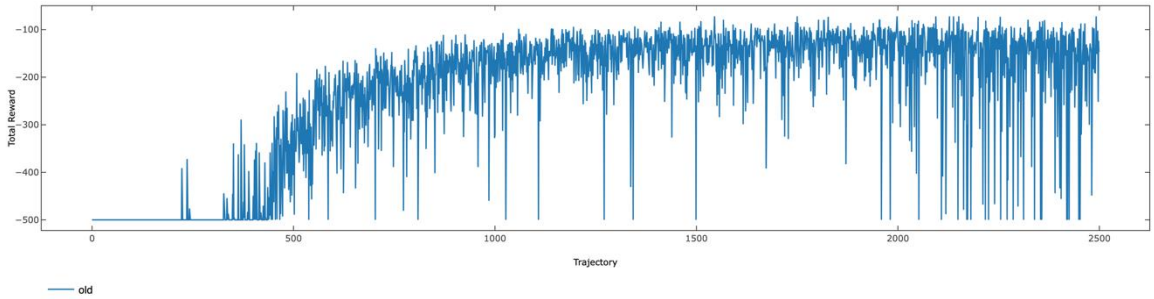
3. Написать PPO для работы в средах с конечным пространствам действий и решить Acrobot. Для решения можно использовать Categorial из torch.distributions (см. pytorch документацию).

Игра: **Acrobot-v1**
Результаты можно посмотреть в ClearML

Параметры для PPO старого и нового алгоритма Advatage функции были выбраны следующие:

- **Episode n:** 50
- **Trajectory n:** 50
- **Gamma:** 0.99
- **Batch size:** 128
- **Epsilon:** 0.5
- **Epoch n:** 100
- **Pi learning rate:** 1e-4
- **V learning rate:** 5e-2
- **Advantage function:** old

Reward на траекториях при обучении:



Вывод:
К данному алгоритму долго подбирались гиперпараметры, чтобы он начал обучаться, но все-таки был достигнуть успех. Reward вышел на плато и стал крутиться около -100, что является хорошим результатом для этого алгоритма.