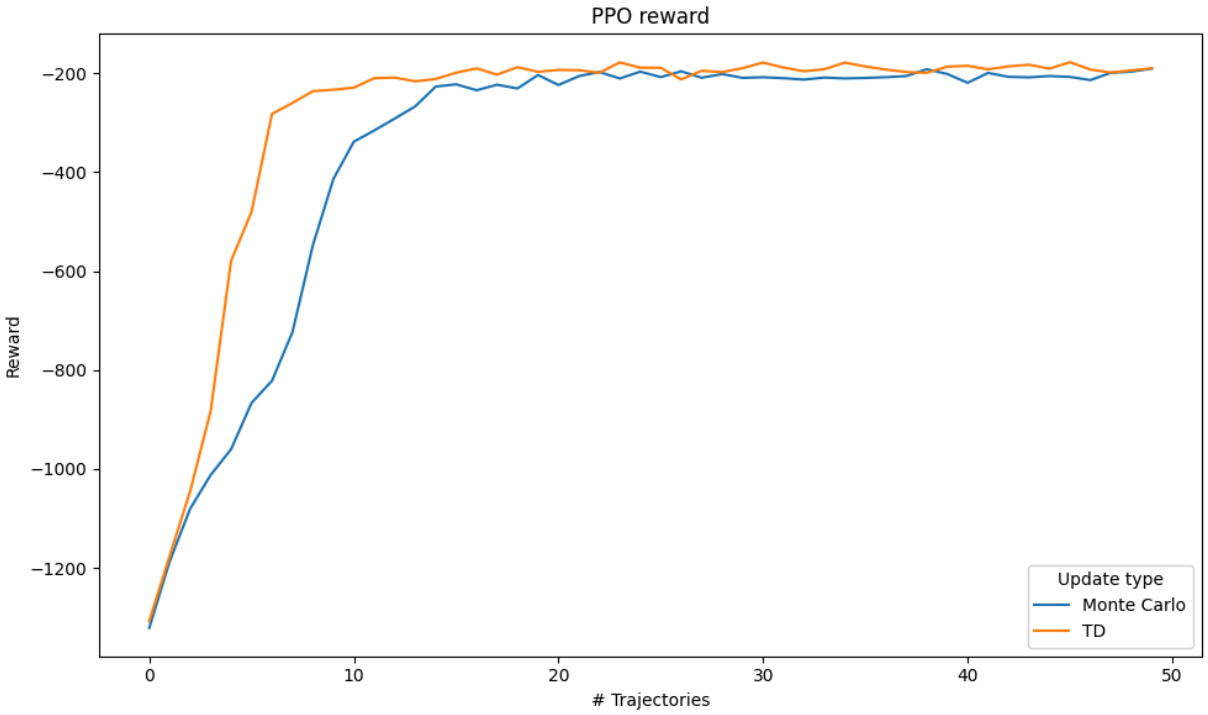
Неделя 6

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.

In [10]:

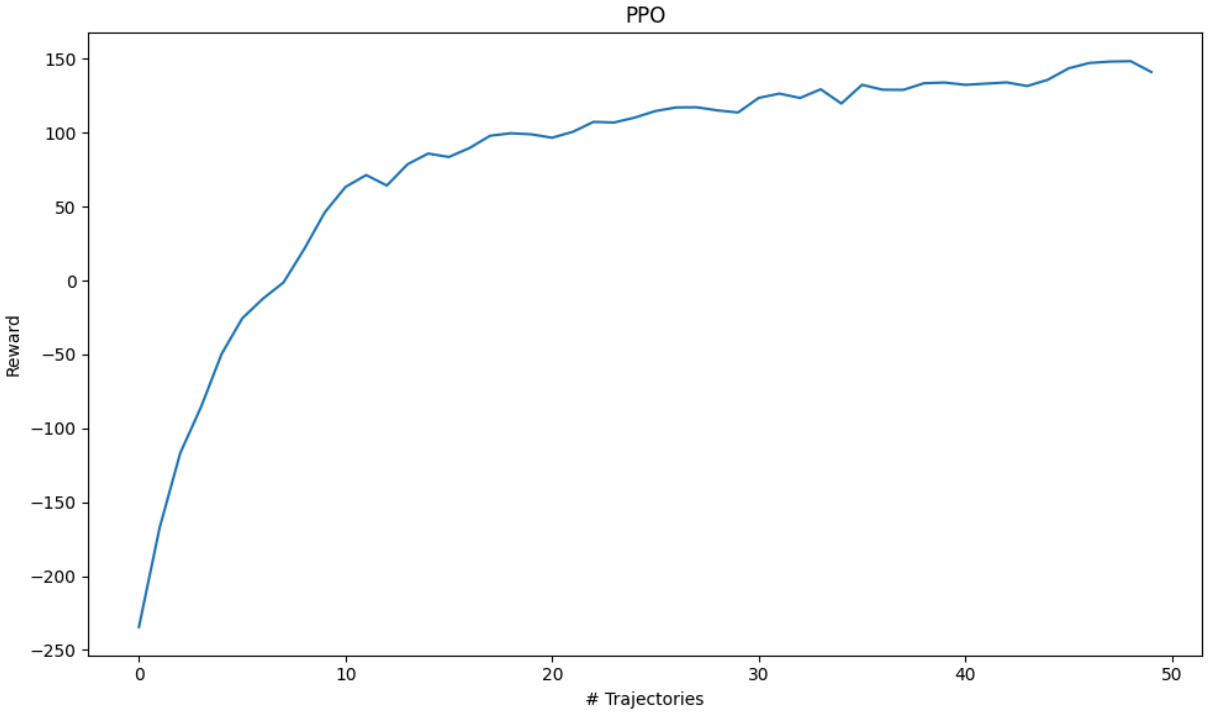


На данном графике изображён алгоритм PPO с вычислением Advantage по методу Монте-Карло (из лекций), а так же с помощью Q функции (на графике TD - temporal difference). Видно, что для задачи Pendulum обучение происходит немного быстрее новым способом. Полагаю, что среда достаточно простая и Q оценка довольно точная, что позволяет ускорить обучение.

Вывод:

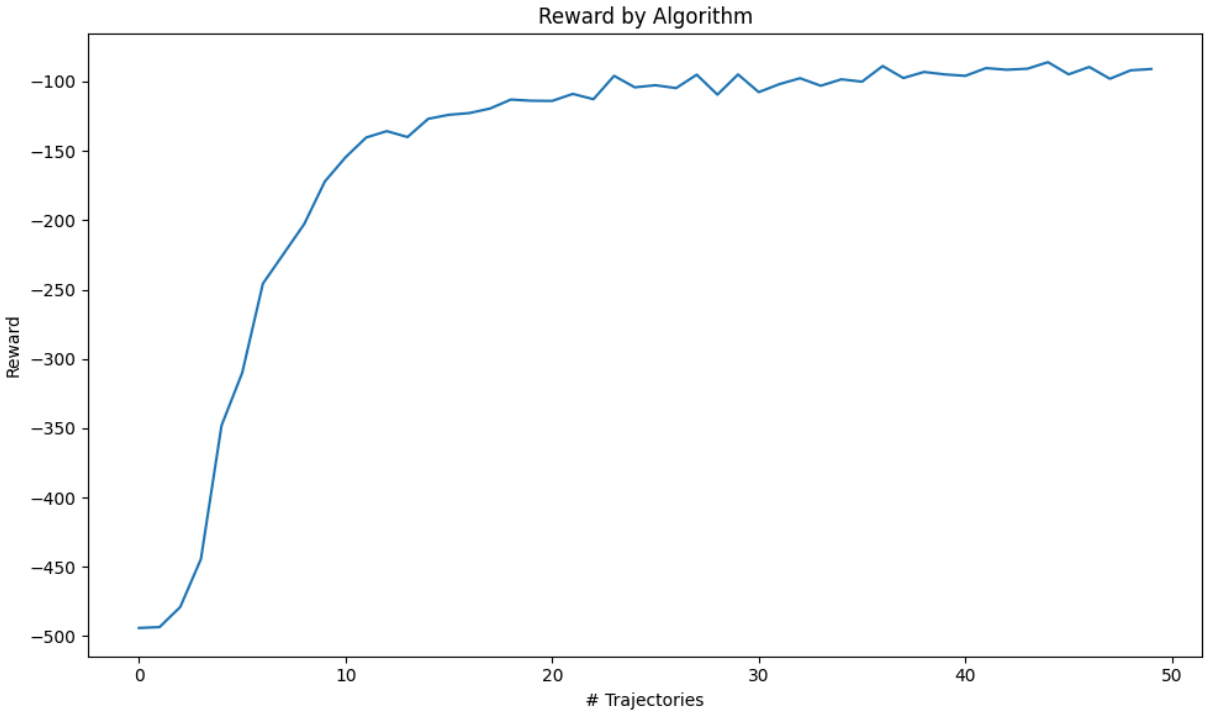
- Вычисление Advantage новым способом ускоряет обучение.

In [13]:



На графике представлена кривая обучения LinarLander continuous с помощью алгоритма PPO. Поскольку пространство действий двумерное, нейросеть возвращала 2 средних и 2 дисперсии для каждой размерности. В остальном алгоритм не изменялся.

In [15]:



Обучение Acrobot с помощью алгоритма PPO. Распределение в агенте было заменено на категориальное с 3 значениями, в остальном алгоритм не менялся.