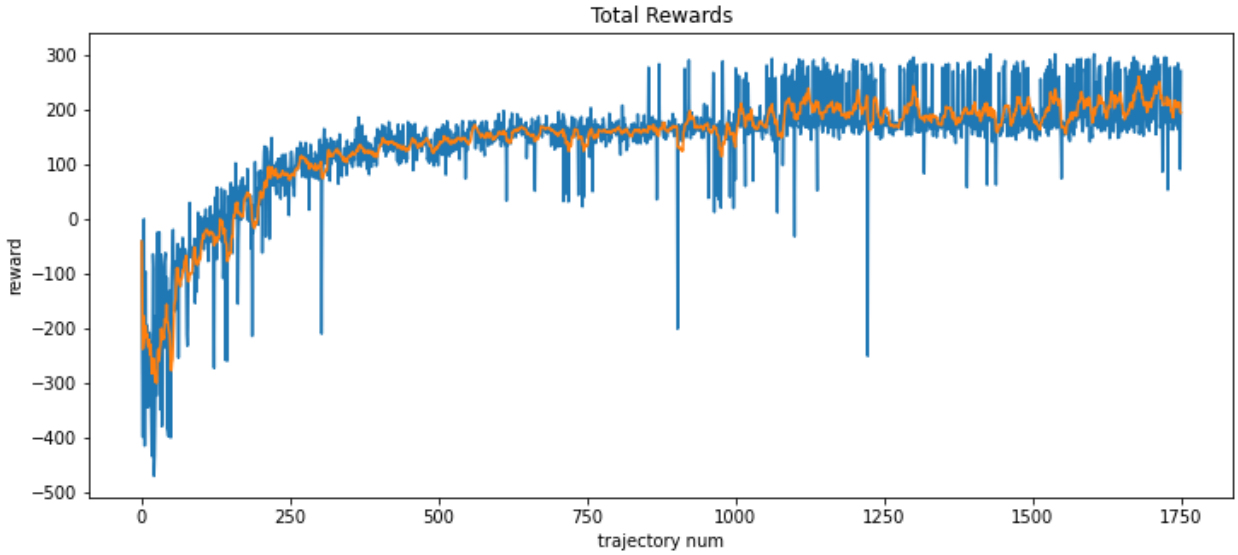
Задание 1. Сравнить кривые обучения алгоритма с “новым” способом и “старым” способом (из практики) на задаче Pendulum.

Эксперимент: Сравнение adantage метода без returns со старым способом.

Вывод.
Примерно с 400 итерации награда алгоритмов становится очень похожа, но новый способ обучается быстрее чем старый. Для наглядности я немного сгладил награды.

Задание 2. Модифицировать PPO для работы в средах с многомерным пространством.



Вывод.
В начале обучения алгоритм сходится в награде чуть выше 100, примерно после 1000 траектории, средняя награда начинает достигать 200 и выше.

Задание 3.