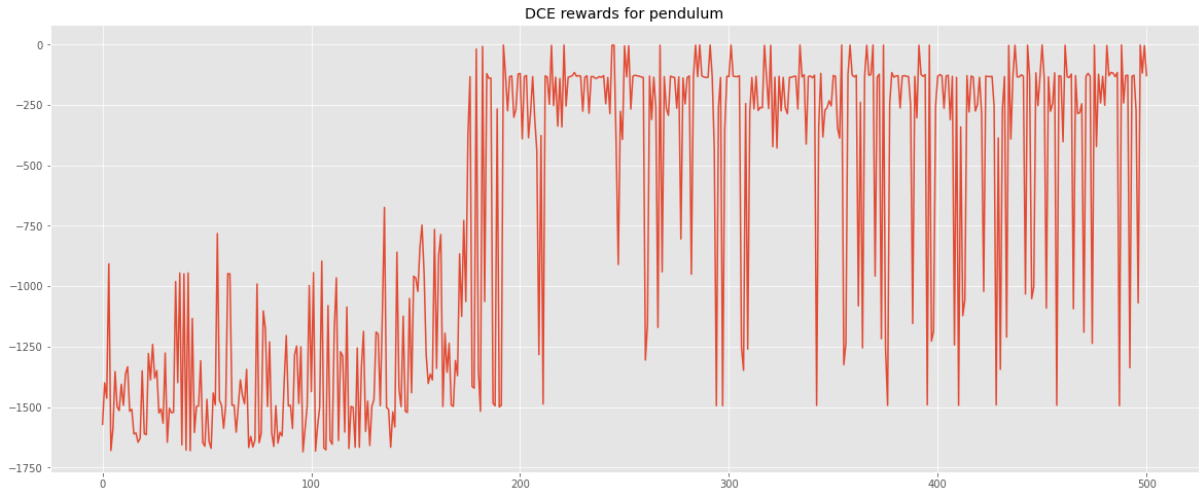
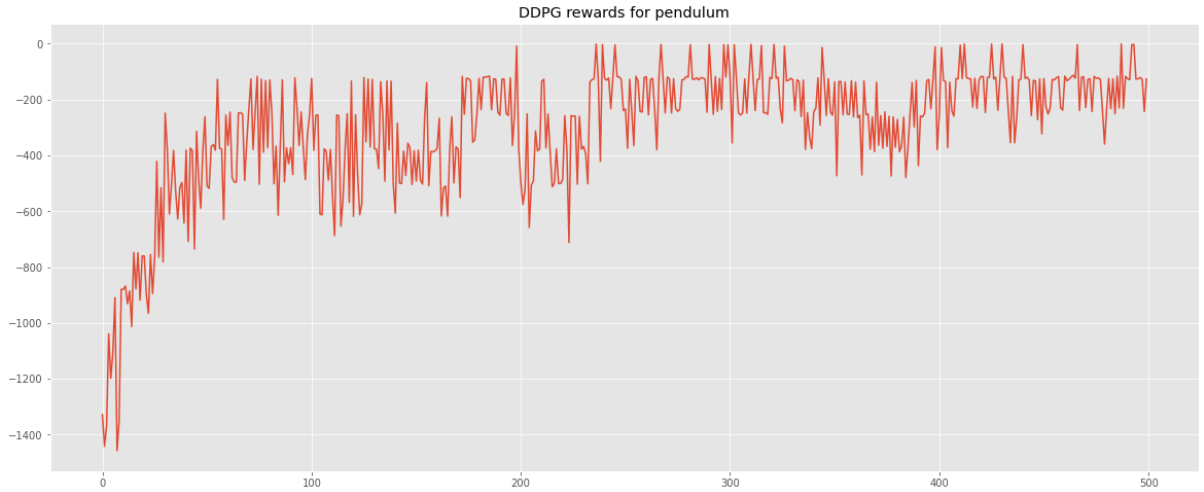
Pendulum

Основная сложность для меня заключалась в том, чтобы в принципе заставить deep-cross-entropy (DCE) работать вообще хоть как то. После моих наработок по второму дз, сетка решала задачу на максимум ~ -200.



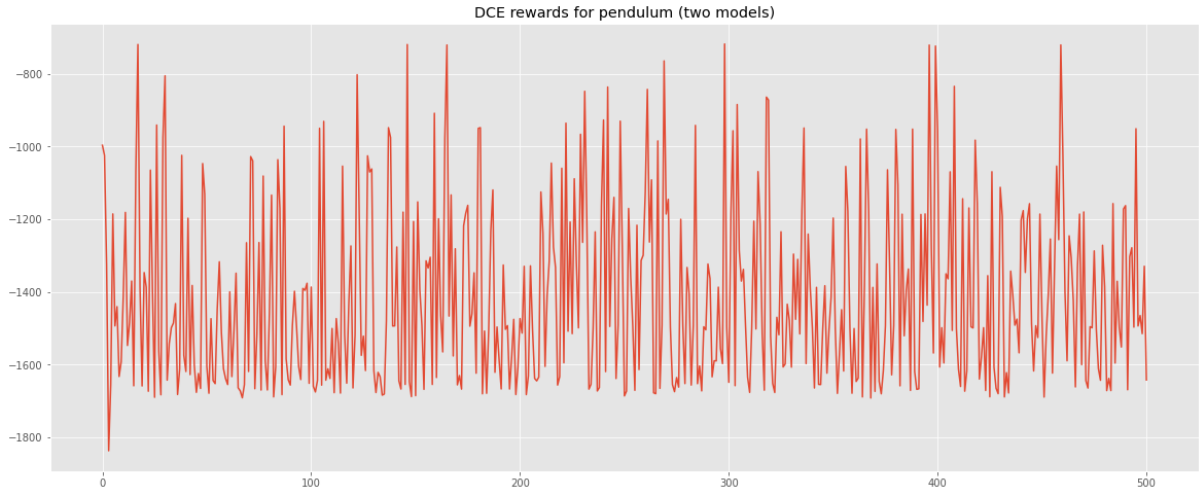
Награды для ddpg:



Собственно, DDPG мне кажется лучшим методом для решения этой задачи в силу нескольких причин:

1) как видно на графиках, отрицательные выбросы гораздо меньше
2) само обучение занимает меньшее количество времени (ибо в DCE также генерится какое-то количество траекторий за эпоху)

Возможно, стоило также добавить в dce вторую модель, как в double dqn и как делали для ddpg. Я попробовал и стало хуже:



Поэтому я удалил, и оставил все как и было. По итогу: ddpg как на занятии дает гораздо лучшие результаты. К тому же, для dce мне я сделал бОльшую нейронную сеть

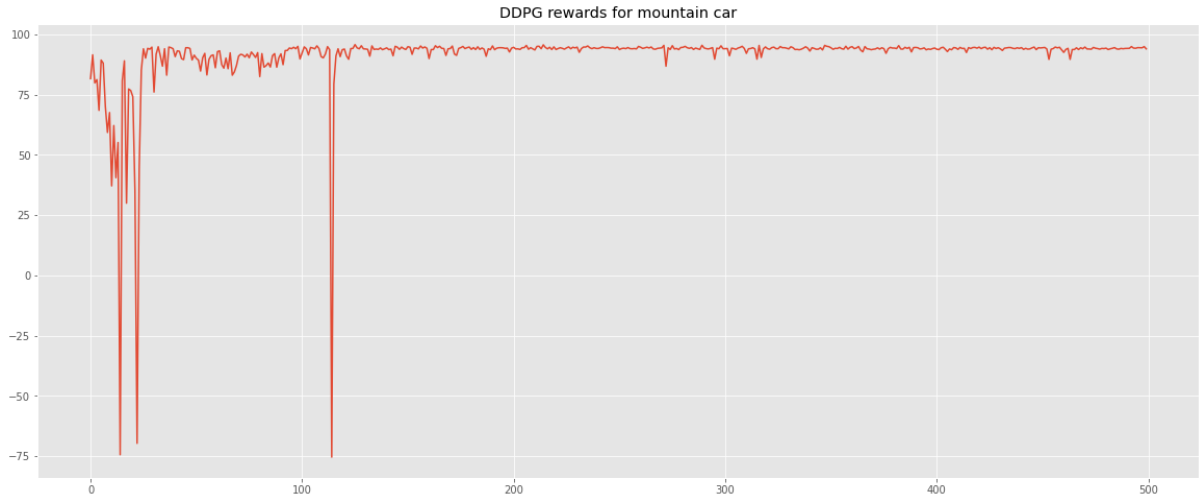
```
def __init__(self, state_dim, action_dim):
    super(Net_dce, self).__init__()
    self.fc1 = nn.Linear(state_dim, 512)
    self.fc2 = nn.Linear(512, 256)
    self.fc3 = nn.Linear(256, 128)
    self.fc4 = nn.Linear(128, action_dim)
    self.relu = nn.ReLU(True)
    self.tanh = nn.Tanh()

def forward(self, x):

    x = self.relu(self.fc1(x))
    x = self.relu(self.fc2(x))
    x = self.relu(self.fc3(x))
    x = self.tanh(self.fc4(x))

    return x
```

Но работало все равно так себе по сравнению с ddpg. Вероятнее всего стоило дальше продолжать играться с шумом, который мы добавляем к весам dce напрямую (гиперпараметр сигма). Я пробовал постоянный шум, пробовал постепенно уменьшающийся шум, пробовал на первые 150 эпох сигма=2, потом на 150 сигма=1, на оставшиеся 200 эпох - постоянно уменьшающийся до 0.02. Также пробовал менять количество траекторий за эпоху, и процент тех, которые отбираются как элитные, но помогло не очень.

Также возможно для DCE стоит прописать early_stopping то есть прерывание обучения, если например среднее значение всех наград за все обучение становится больше -150

MountainCarContinuous

Собственно никаких дополнений к ddpg делать опять же не пришлось, и работало более менее хорошо



Результат для dce был похуже так же в силу того, что как минимум требуется гораздо больше времени для того, чтобы обучить агента