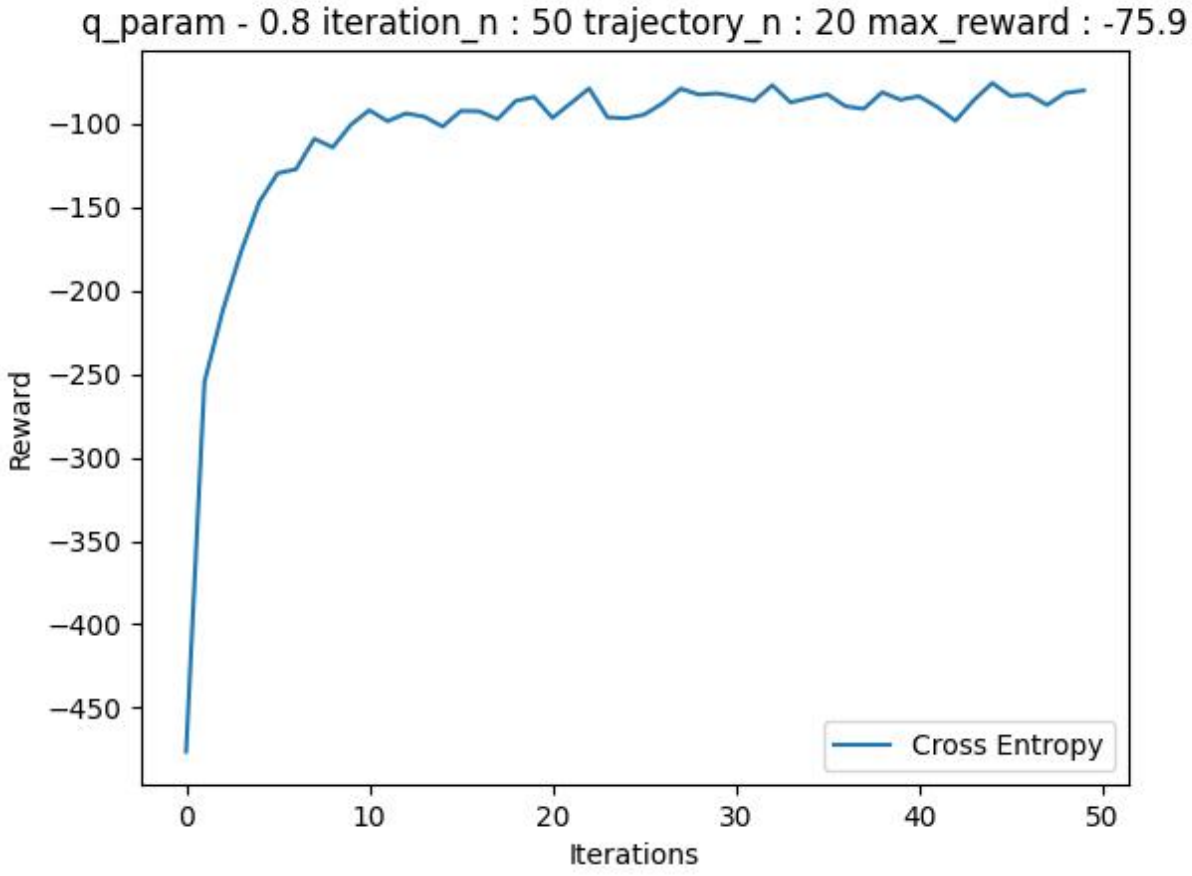
Отчет по домашнему заданию.

Задание 1: Пользуясь алгоритмом Кросс-Энтропии для конечного пространства действий обучить агента решать Acrobot-v1 или LunarLander-v2 на выбор. Исследовать гиперпараметры алгоритма и выбрать лучшие.

Я выбрал Acrobot-v1, тут основная задача, перекинуть вторую часть палки за черную линию.

Начнем с основных параметров, взятых с семинара.

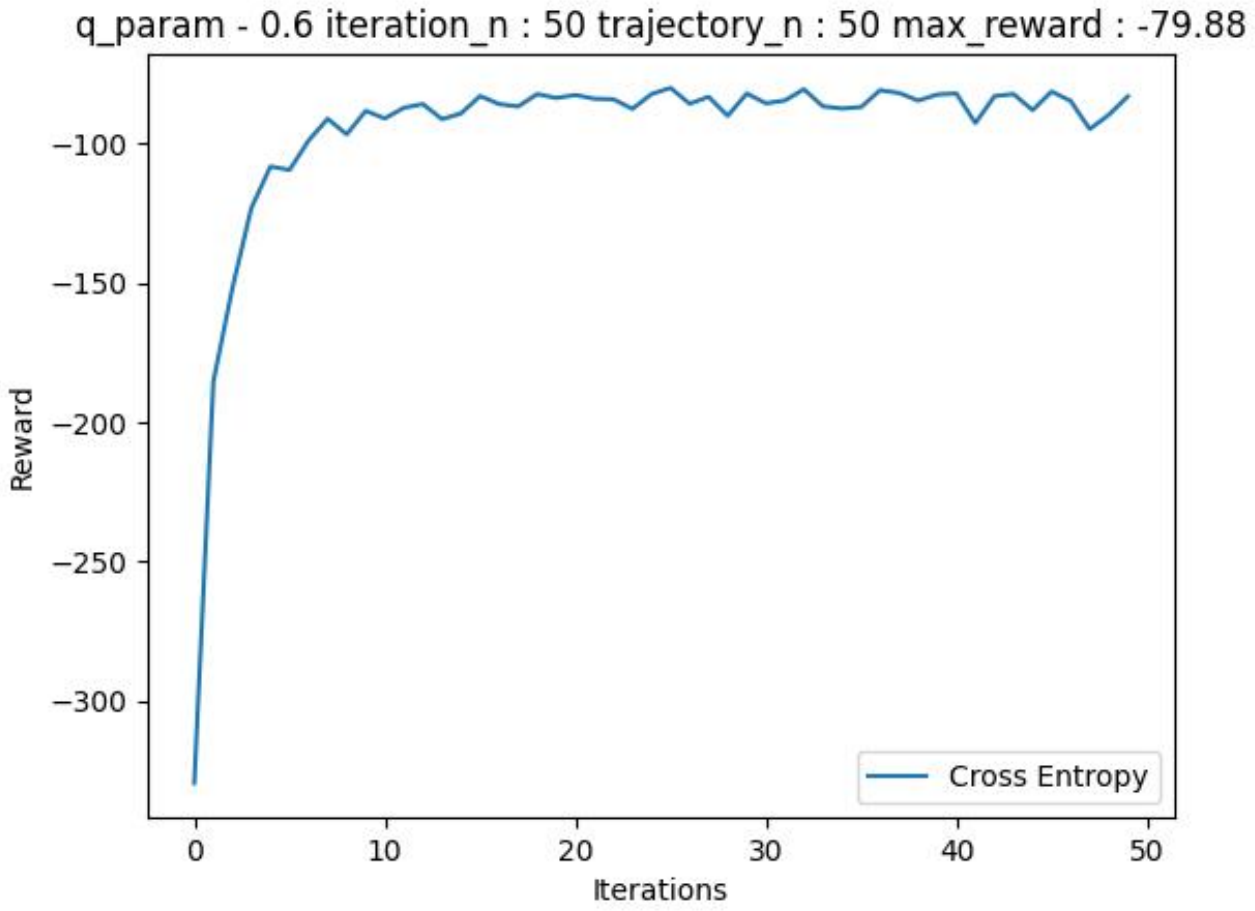
episode_n = 50
trajectory_n = 20
trajectory_len = 500
q_param = 0.8



Итог: Max_mean_reward = -75.9

Дальше увеличил кол-во траекторий и уменьших q_param

episode_n = 50
trajectory_n = 50
trajectory_len = 500
q_param = 0.6

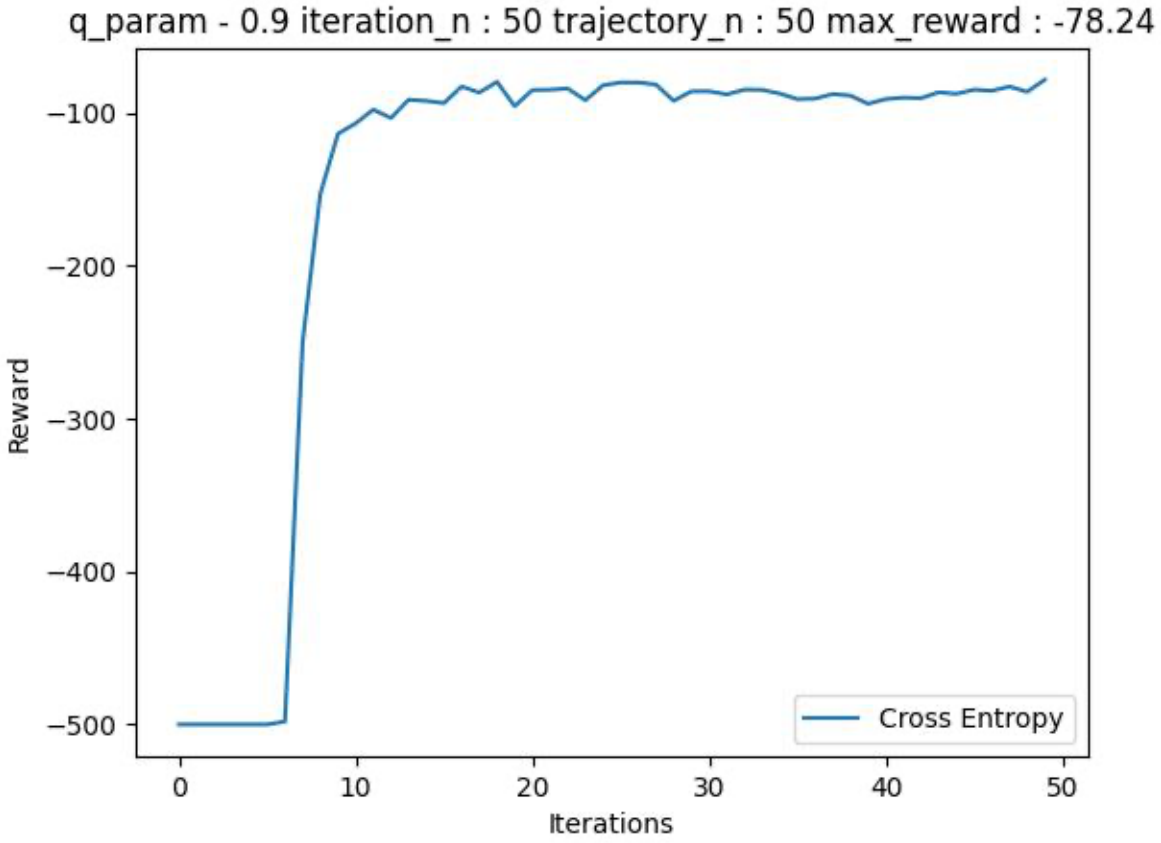


Итог: Max_mean_reward = -79.88

Далее увеличил q_param

episode_n = 50
trajectory_n = 50
trajectory_len = 500

q_param = 0.9



Итог: Max_mean_reward = -78.24

Итог по заданию: Не удалось получить более лучшие результаты по mean_reward, при усложнении модели mean_reward уменьшается. В любом случае, нейронная сеть научилась выполнять поставленную задачу.

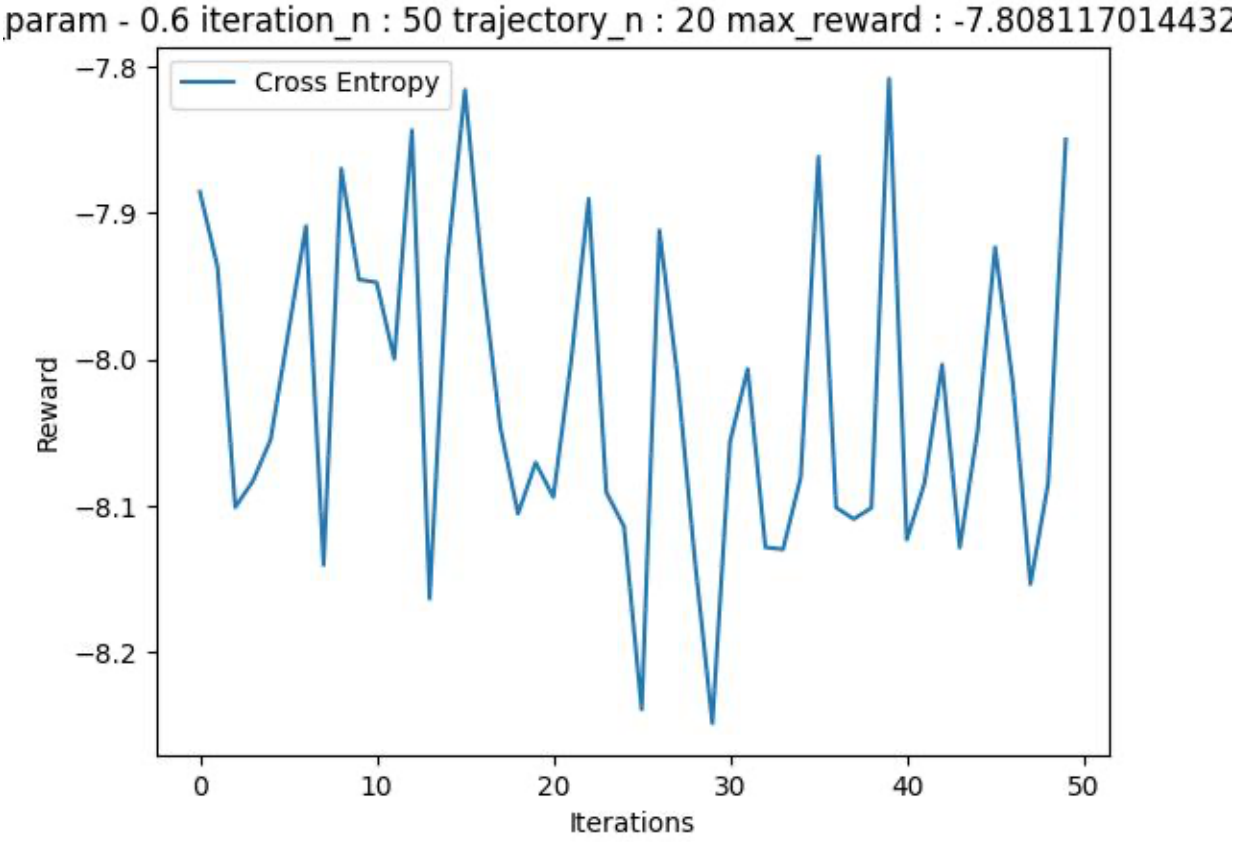
Задание 2: Реализовать алгоритм Кросс-Энтропии для непрерывного пространства действий. Обучить агента решать Pendulum-v1 или MountainCarContinuous-v0 на выбор. Исследовать гиперпараметры алгоритма и выбрать лучшие.

Я выбрал MountainCarContinuous-v0, тут задача завести машину на гору. Тут меняется задача, если в задании 1, мы использовали классификацию, то в задании 2, мы должны использовать

регрессию. Также чтобы вписаться в интервал [-1,1], использую функцию тангенса, вместо софтмакса. Добавил изменение политик с учетом epsilon.

Параметры

episode_n = 50
trajectory_n = 20
trajectory_len = 500
q_param = 0.6
lr = 1e-2

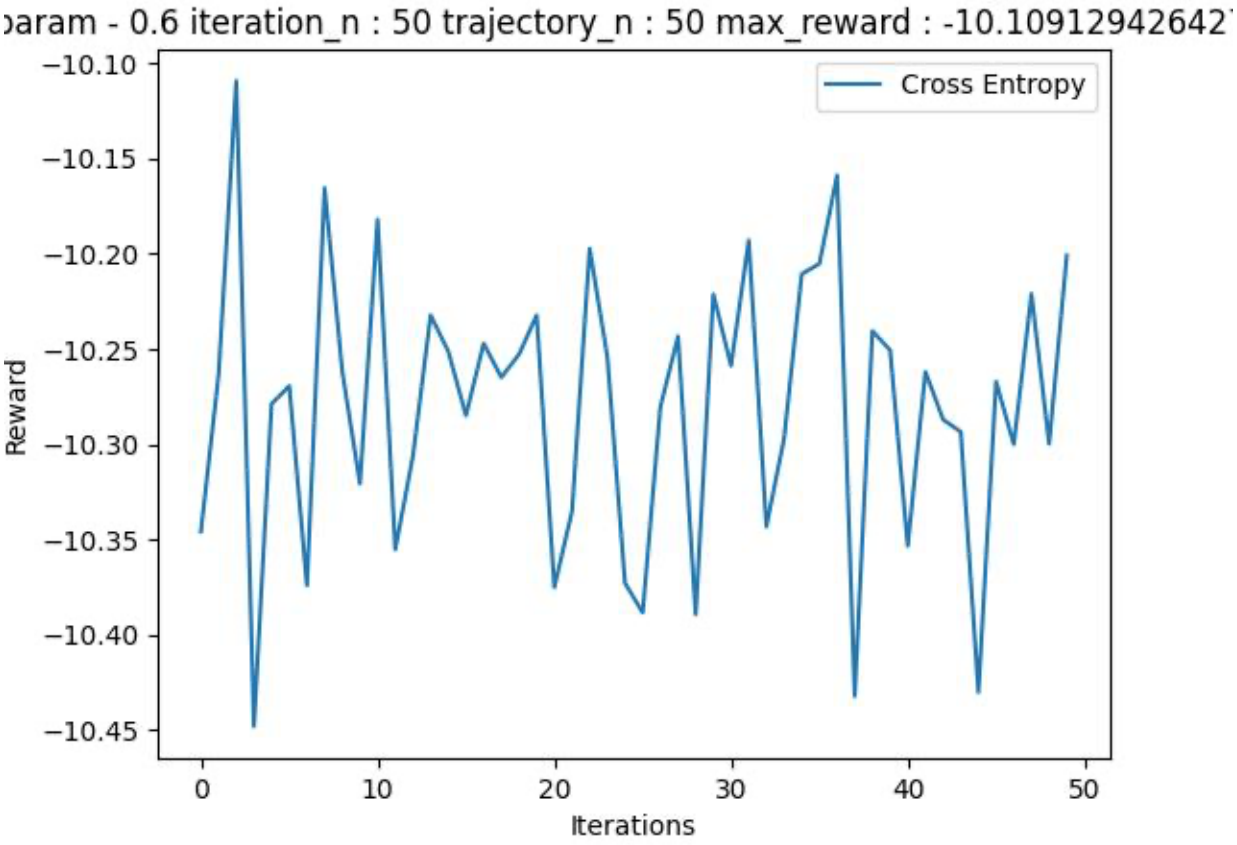


Итог: видно рандомное блуждание, max_reward = -7.8

Изменение параметров: lr, trajectory_n, trajectory_len не привели к улучшению результатов.

Params:

lr = 1e-1
episode_n = 50
trajectory_n = 50
trajectory_len = 1000
q_param = 0.6



Итог: max_reward = -10.109. Научить машинку взбираться на гору не получилось, я провел более 10 изменений параметров, ситуация не меняется. Мне кажется, что нейронная сеть, не пытается решить проблему увеличения поощрения, она пытается минимизировать получаемые штрафы.