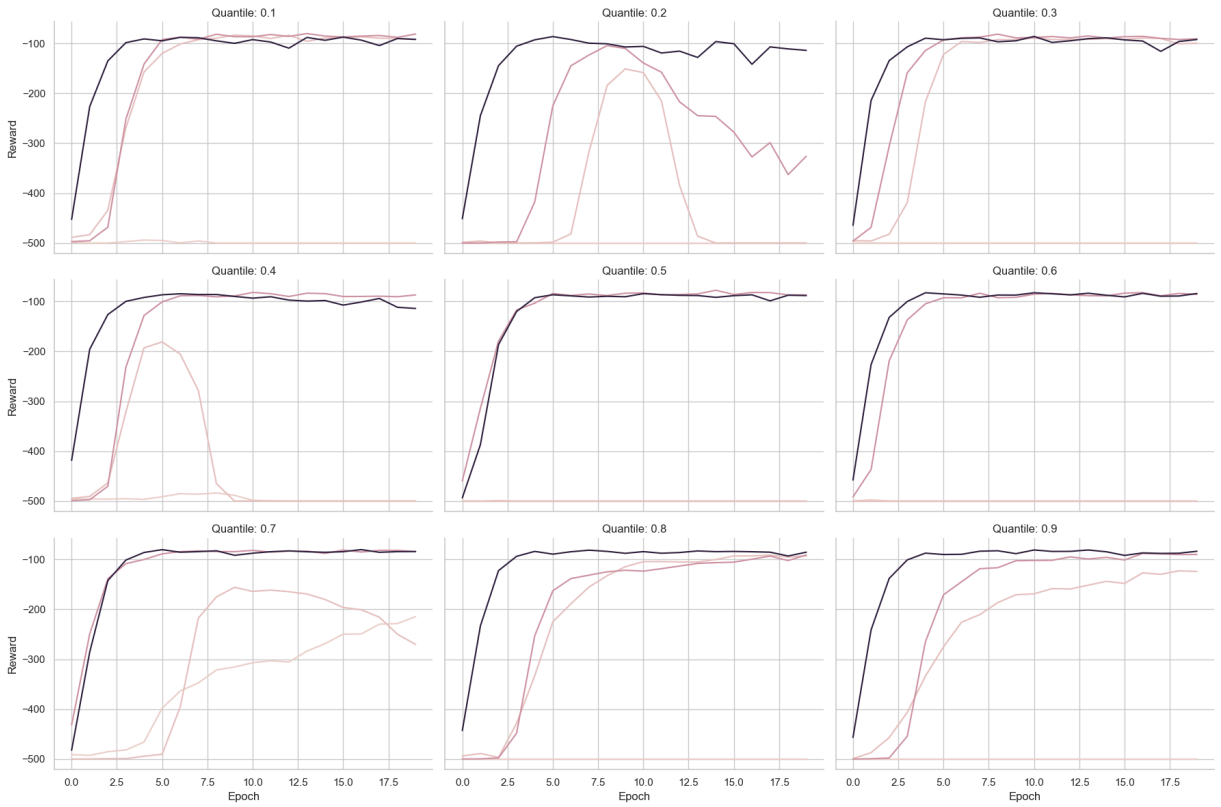
Задание 2

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

В данной работе рассматривается **Acrobot-v1**.

In [2]:



На графиках рассматриваются 2 гиперпараметра: квантиль и количество траекторий. Чем темнее график, тем больше траекторий, значения которые ипользовались: 30, 100, 300, 1000. Разные квантили представлены на разных графиках (от 0.1 до 0.9 включительно с шагом 0.1). Гиперпараметры нейросети и обучение не рассматривались, потому что иначе пространство гиперпараметров будет слишком большим (количество слоёв, количество нейронов, разные активации, разные оптимизаторы, разные learning rates итд). Для нейросети использовалась сеть с 3 слоями по 64 нейрона, LeakyReLU, Adam, lr 1e-4, epsilon = 1/N.

Агенты со средней наградой больше -100 считаются решившими задачу, это можно видеть по записанным видео (не включённым в отчёт). Для малого числа траекторий почти никакой квантиль не спасает, чем больше траекторий, тем

лучше. Видно, что для даже для низких квантилей, при большом числе траекторий задача решается, но решение нестабильно и засоряется плохими траекториями что ведёт к временному падению средней награды. Начиная с квантиля 0.5 решение для большого числа траекторий решение стабильно. С более высоким квантилем обучение происходит быстрее, так как меньше примеров нужно добавлять в нейросеть.

Результат:

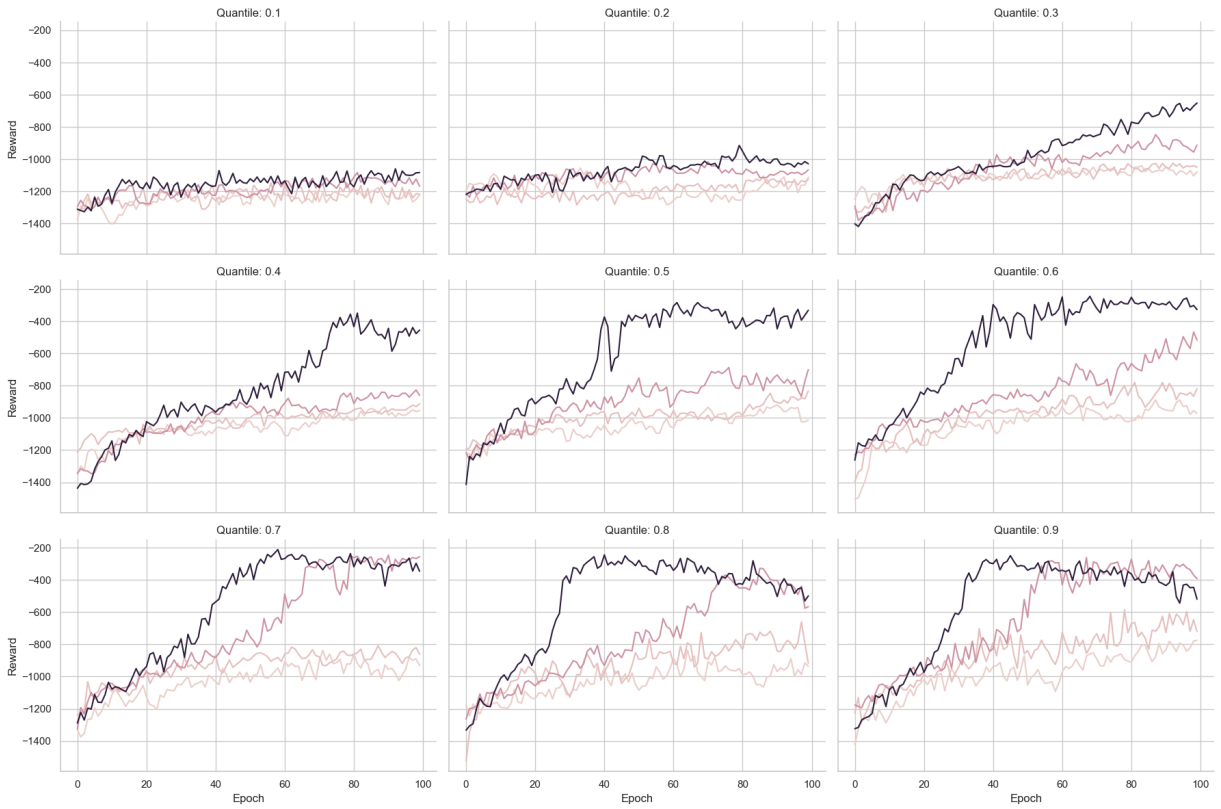
- Количество траекторий: 1000 (достаточно, но чем больше, тем лучше)
- Квинтиль: 0.9 (любое значение больше 0.5 подойдёт)

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.

В данной работе рассматривается **Pendulum-v1**.

In [7]:



На графиках рассматриваются 2 гиперпараметра: квантиль и количество траекторий. Чем темнее график, тем больше траекторий, значения которые

ипользовались: 30, 100, 300, 1000. Разные квантили представлены на разных графиках (от 0.1 до 0.9 включительно с шагом 0.1). Гиперпараметры нейросети примерно такие же, только epsilon уменьшается равномерно от 1 до 0.2 (как было показано на семинаре).

Среда считается решённой при награде выше -300. Для низких квантилей особого прогресса в решении задачи не было, нейросеть менее устойчива к мусорным данным и при низком квантиле слишком много шума мешают агенту выучить верную стратегию. Интересная ситуация с высокими квантилями и большим числом траекторий, там агент достигается оптимальной стратегии, но потом среднее значение падает. Анализ поведения агента показывает в силу особенности среды в ней всегда будут ситуации, когда маятник внизу и его надо раскачать. Видимо, эти данные мешают сети удерживать локальный минимум и сеть сбивается. Поэтому в данной задаче кажется важным количество примеров, которые попадают в обучение сети, либо более низкий квантиль, либо меньшее число траекторий.

Результат:

- Количество траекторий и квантиль: 1000 траекторий и квантиль 0.6, либо 300 траекторий и квантиль 0.7.

In [ ]: