Задание 1

Анализ задачи

Задача “Taxi-v3” содержит 6 действий и 500 состояний. Из этого можно вывести, что необходимо увеличить количество траекторий при оценке политики, чтобы протестировать агента на всех возможных начальных состояниях.

Пусть q = 0.9, тогда если всего возможно 500 начальных состояний (для простоты опустим, что достижимых начальных состояния всего 404, так как на решение задачи это практически не влияет), чтобы в элитные траектории попала хотя бы одна траектория для каждого начального состояния, минимальное количество траекторий должно составлять 5000.

Baseline

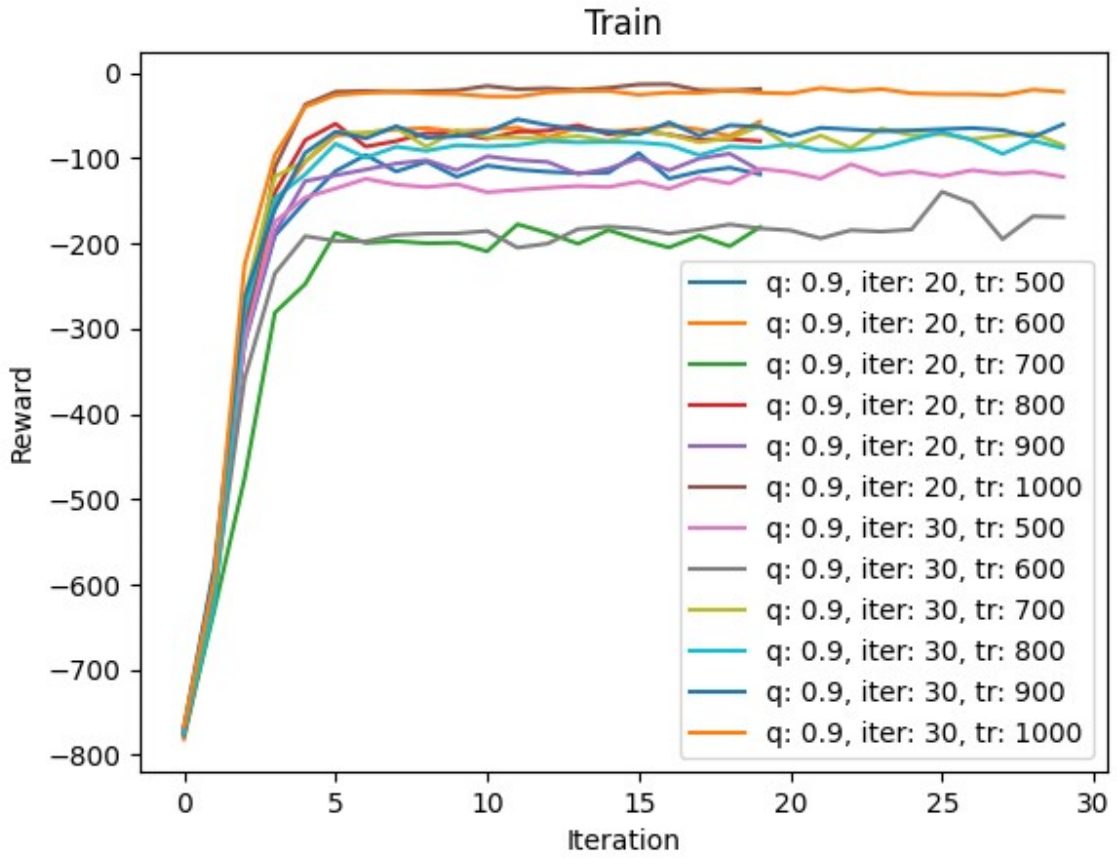
В качестве бейзлайна были использованы гиперпараметры с семинара:

q_param = 0.9

iterations_n =20

trajectory_n = 100

Для поиска оптимальных гиперпараметров я перебрал все возможные комбинации значений в определенном диапазоне, результаты экспериментов представлены на графике:



Для лучшей читаемости на график выведена только часть экспериментов.

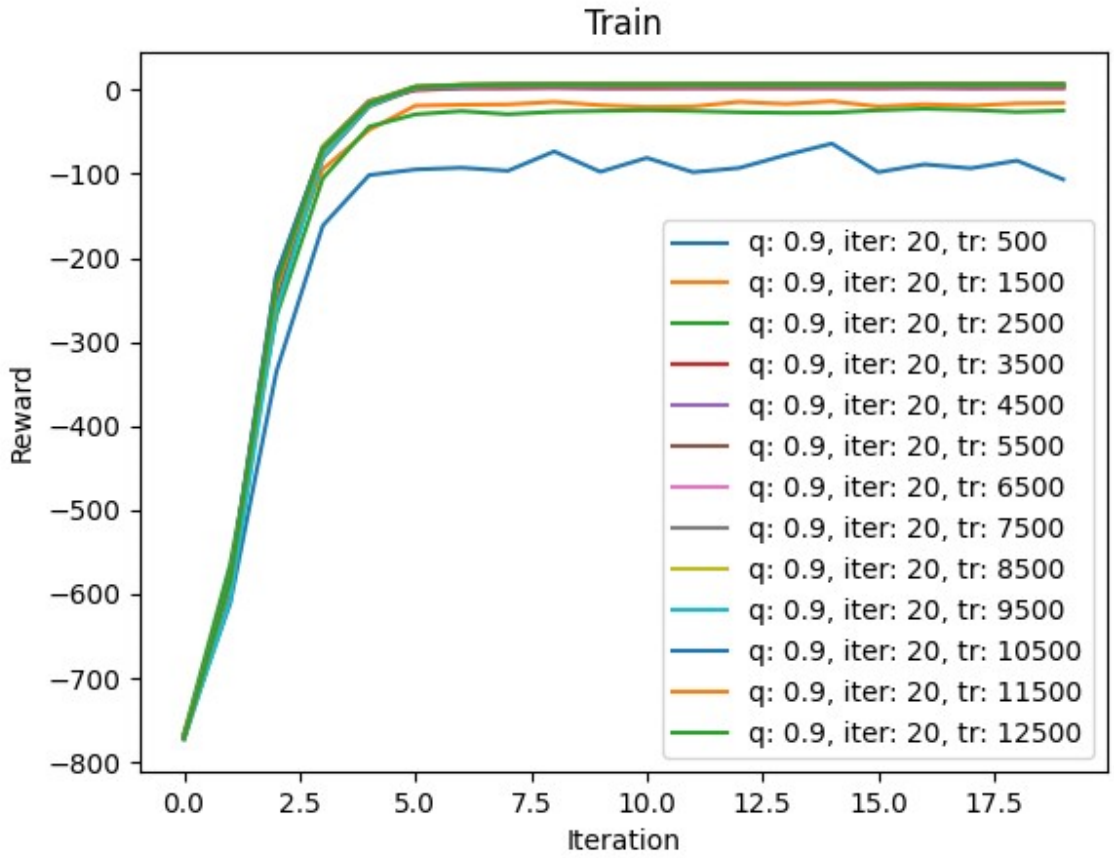
Из графика можно сделать следующие выводы:

- агент достигает верхнего порога качество приблизительно за 5-8 итераций
- после достижения потолка качество меняется не стабильно при малом количестве изученных траекторий, и стабилизируется при увеличении их количества

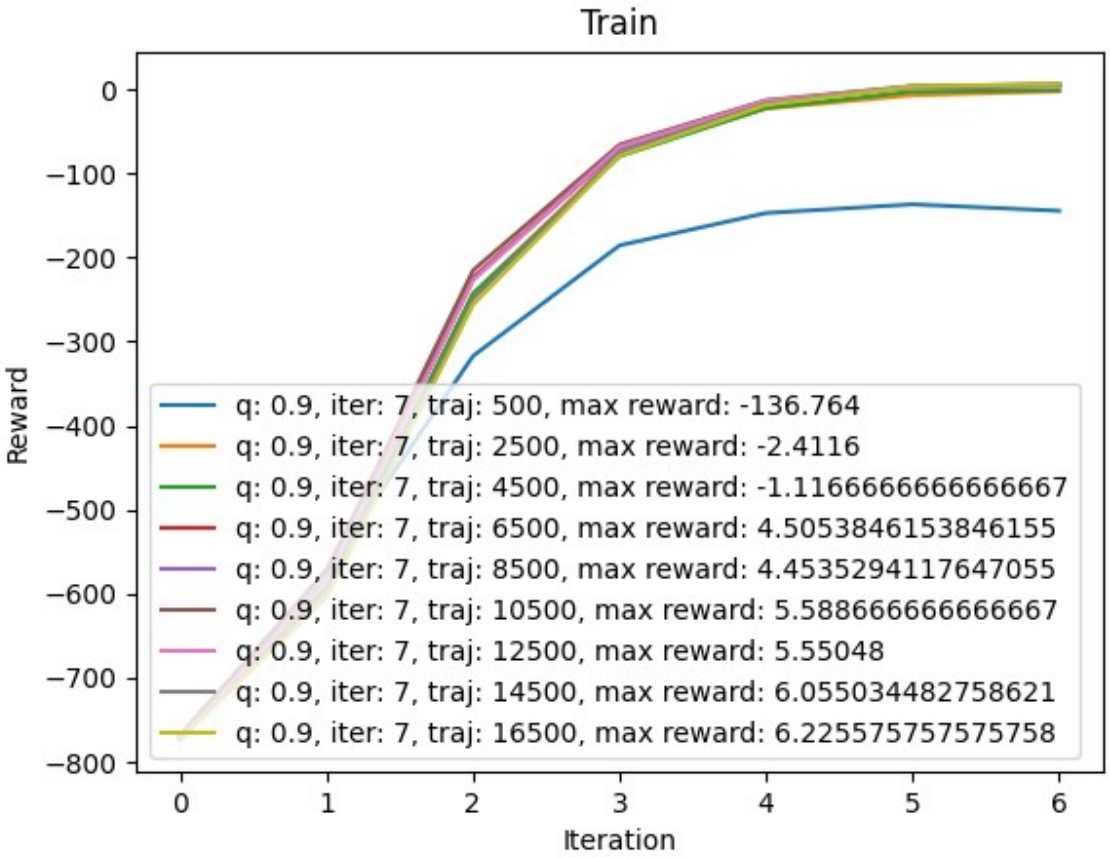
Из второго утверждения можно сделать вывод, что агент выбирает действие случайно. Можно предположить, что так происходит из-за того, что при обучении он не изучил все возможные варианты траекторий. Данное предположение может быть правдивым потому, что при q = 0.9 и trajectory_n = 500 в элитных траекториях окажутся траектории, использующие только приблизительно 10% начальных состояний.

Подбор гиперпараметров

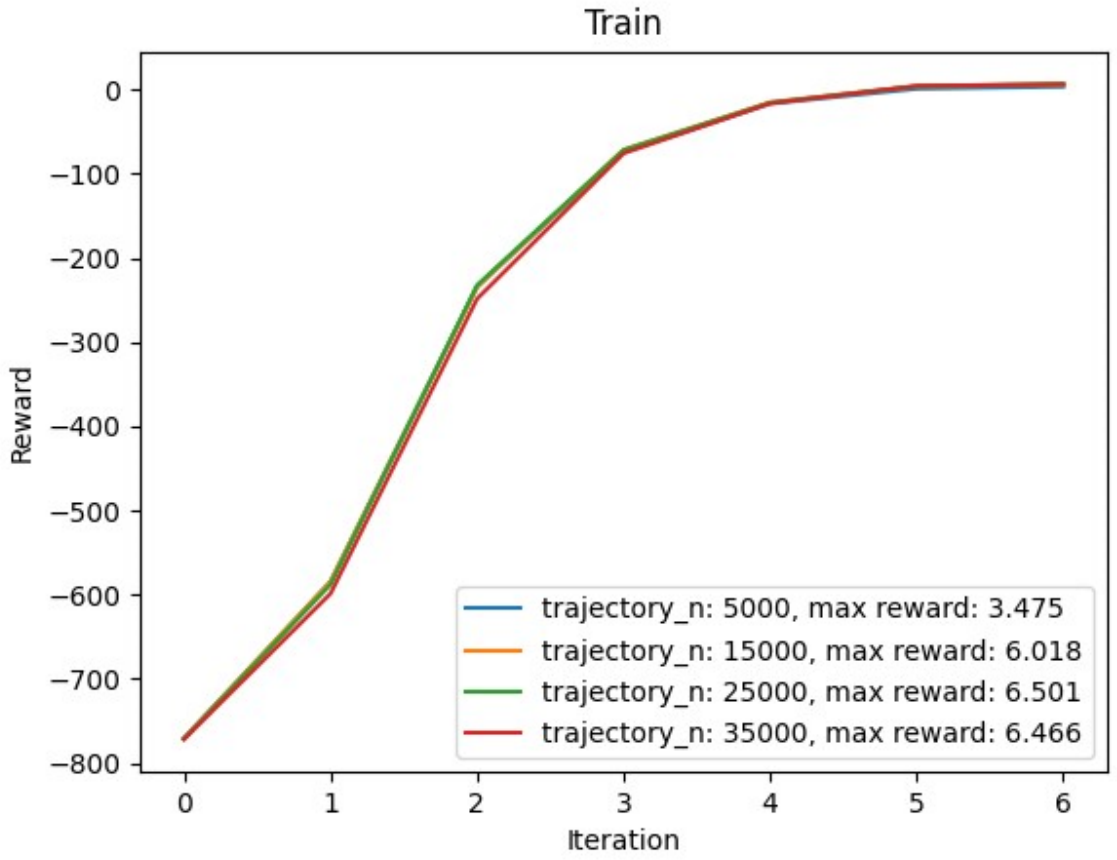
Попробуем увеличить количество анализируемых траекторий. Результаты эксперимента представлены на графике:



Видно, что при trajectory_n > 2500 обучение стабилизируется и качество улучшается. Примем iteration_n = 7 и обучим агента заново, отобразив максимальную награду:



Видно, что 7 итераций достаточно для стабилизации обучения и достижения определенного порога качества при заданных гиперпараметрах. Попробуем еще сильнее увеличить количество траекторий:



Видно, что графики практически совпадают, значит мы достигли определенного максимума стабильности и качества. Примем итоговые гиперпараметры следующими:

q_param = 0.9

iterations_n = 7

trajectory_n = 15000

Если вспомнить, что поле игры имеет размер 5x5, за каждый шаг до финиша начисляется награда -1, а за доставку пассажира +20, то оптимальная средняя награда будет около 8.5, так как минимальная награда будет равна 2 (когда такси находится в локации Y, пассажир в локации G, а точка назначения в локации R), а максимальная 15 (такси и пассажир находятся в Y, а пункт назначения в R). Улучшение качества до максимального может быть достигнуто с помощью сглаживания, что будет рассмотрено в следующем задании.