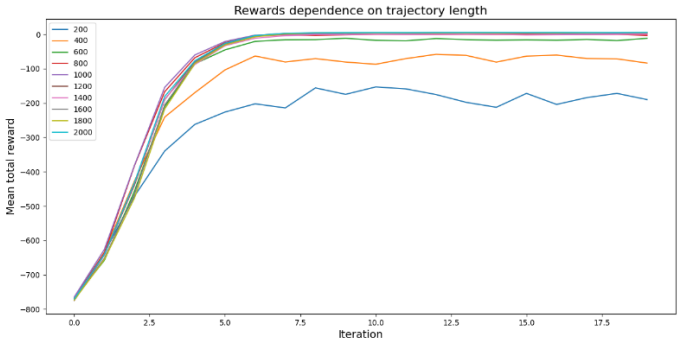
Задание 1.

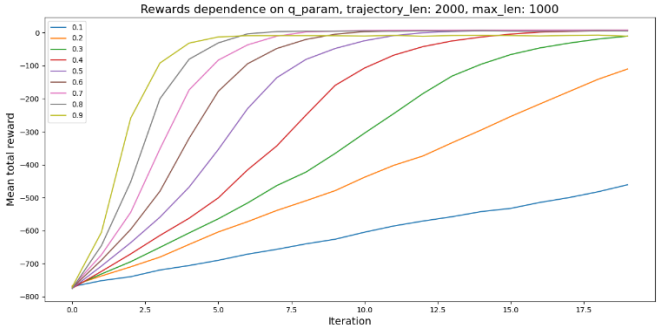
Эксперимент: зависимость награды от длинны траекторий.



Вывод.

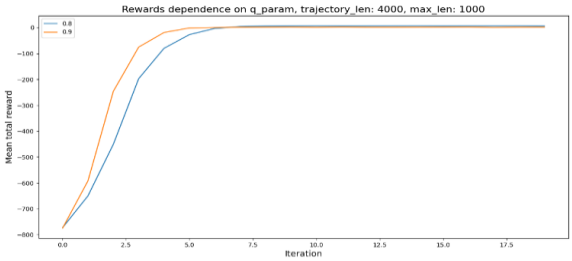
В данной задаче количество степеней свободы гораздо больше чем в лабиринте, следовательно, нам необходимо увеличить количество траекторий, поскольку многие из них не довозят пассажира до финальной точки. По графику видно, что уже после 800 мы приближаемся к положительной награде и примерно после 10 итерации рост прекращается.

Эксперимент: зависимость награды от границы квантиля.



Вывод.

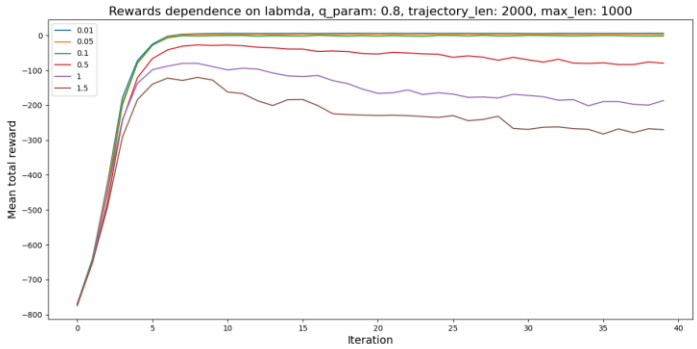
Квантили в диапазоне от 0.8 до 0.4 приходят к положительной награде на дистанции в 20 итераций. Мне кажется, что причина по которой 0.9 квантиль остается с отрицательной наградой, так же в количестве степеней свободы в данной задачи. Элитные траектории, которые остаются после отсечки 0.9 квантилем, не включают в себя все маршруты. Это можно исправить, расширив диапазон квантиля до 0.8 или увеличив количество траекторий trajectory_len, например, до 4000 как видно на след. графике.



Финальные гиперпараметры: q_param: 0.8, trajectory_len: 2000, max_len: 1000, iteration_n: 20.
Максимальная средняя награда в процессе обучения 5.75.

Задание 2.

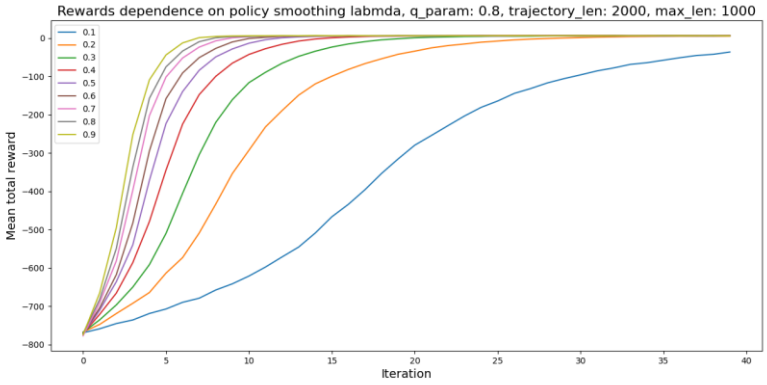
Эксперимент: зависимость награды от коэффициента сглаживания лямбда.



Вывод.

Сглаживание Лапласа приводит к положительной средней награде только при очень низкой лямбде. Причина заключается в том, что при росте лямбды увеличивается вероятность подобрать/сбросить пассажира в неправильном месте и получить сильно негативную награду (-10).

Эксперимент: зависимость награды от сглаживания по политике, с коэффициентом лямбда (0, 1]

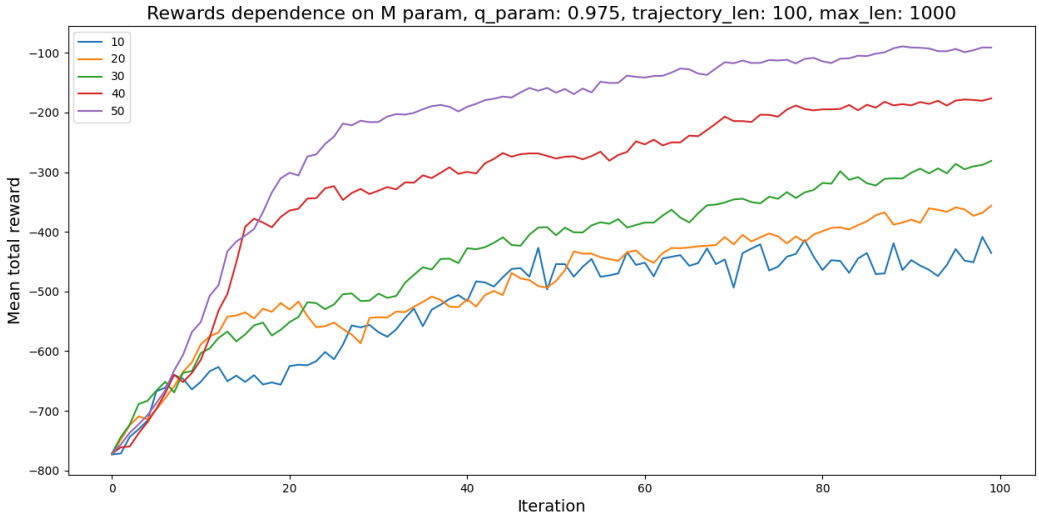


Вывод.

Данный метод увеличивает максимальную среднюю награду в процессе обучения до 6.73 по сравнению с 5.75 без сглаживания, за счет того, что мы передаем в новую политику часть информации от старой. При лямбде = 0.9 достигается максимальная награда и скорость сходимости, что говорит о том что в новую политику нужно передавать только малую часть старой.

Задание 3.

Эксперимент: зависимость награды от параметра M



Вывод.

В данной задаче поскольку мы семплим целые батчи я уменьшил количество траекторий в каждом батче до 100 и повысил значение квантиля до 0.975 с целью уменьшения общего количества траекторий подающегося в обучение. Обучение требует гораздо больше итераций из-за того, что мы не выбираем элитные траектории, а подаем целые батчи семплируя по средней награде батча и в выборку попадают не самые лучшие траектории. При значении параметра M = 50, удается получить сходимость в районе -90.