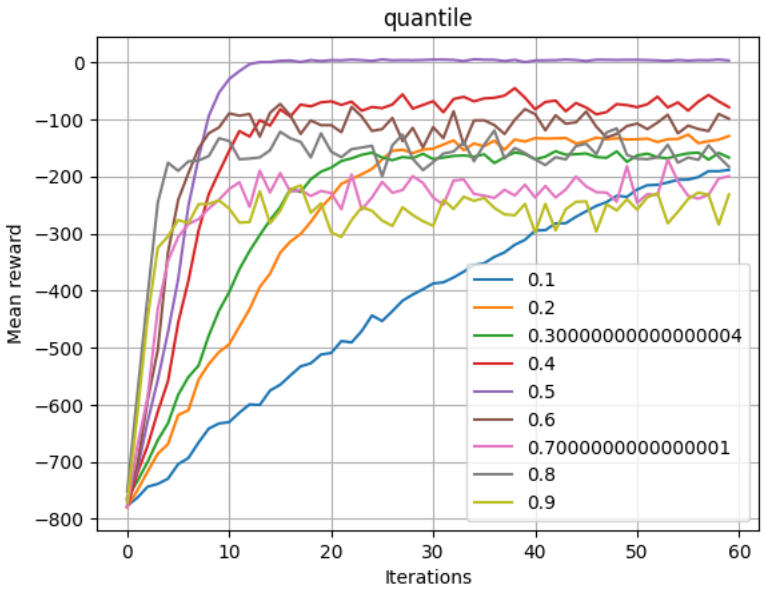
Отчёт о практике №1

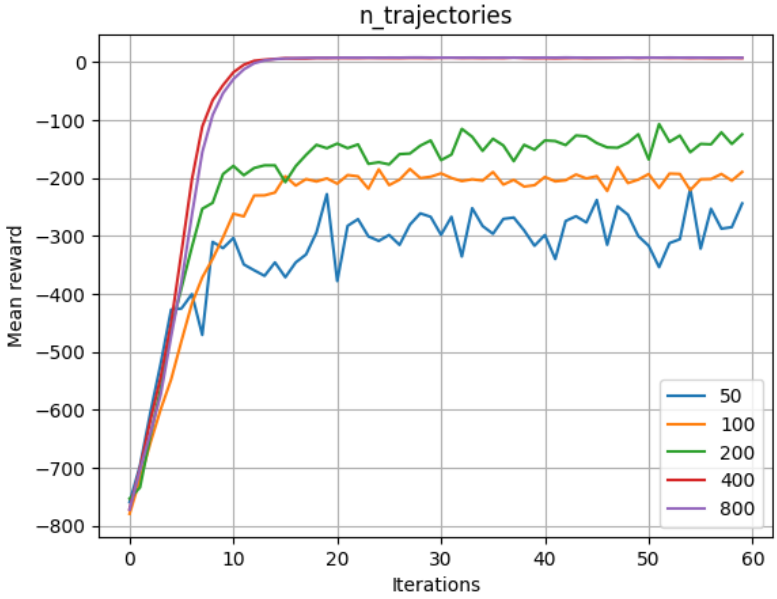
Дмитрий Торшин

Задача 1

Основным отличием от семинара является работа с новой средой: Такси. Она является стохастической, т. е. изначально такси появляется в случайной точке, пассажир также может быть случайно в одной из 4х точек, как и место, куда его можно отвезти. С другой стороны и само поле сильно отличается от лабиринта, ведь вместо одной траектории теперь есть свобода выбора, и нет единственного правильного решения.

Всё это приводит к тому, что параметры, которые подходили к лабиринту, теперь не являются актуальными. Во время проведения первых экспериментов, я обнаружил, что несмотря на то, что автомобиль решает поставленную перед ним задачу, но делает это неоптимальным способом, нарезая по ходу движения разнообразные зигзаги и круги. При этом график обучения показывал, что модель перестала учиться и вышла на горизонтальную асимптоту. Данное явление я решил объяснить недостатком *exploration*. Алгоритм кросс-энтропии сильно зависит от того, повезло ли нам на первых итерациях найти правильный путь, ведь остальные веса будут зануляться. Решить проблему легко — чем больше траекторий в итерации будет учитываться — тем лучше. Однако это приводит к увеличение времени выполнения и интереснее будет разобраться с другим параметром — оптимальным значением квантиля.





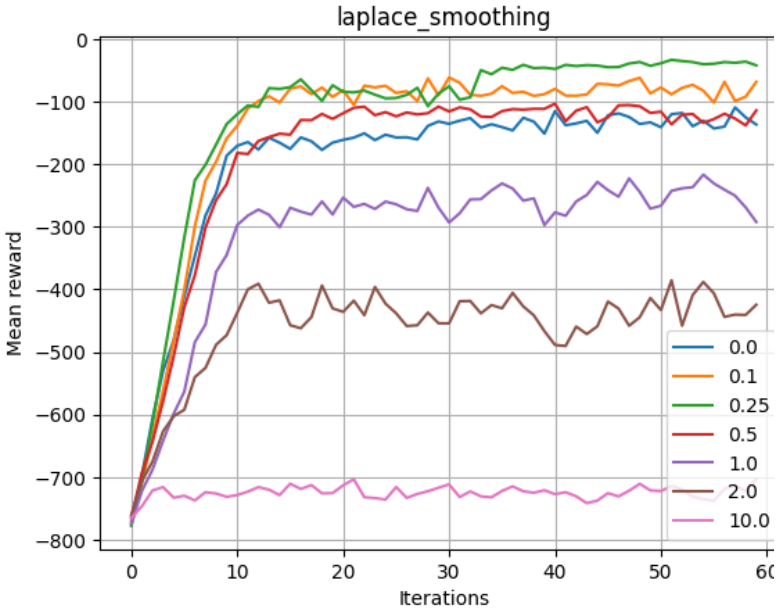
Результаты по квантилям(число траекторий фиксировано) подтвердили вывод о том, что выбирая мало количество траекторий(чем больше квантиль, тем их меньше), модель просто переобучается из-за случайного семплирования. А слишком малые значения наоборот ведут к тому, что постоянно присутствуют плохие траектории, от которых не получается избавиться.

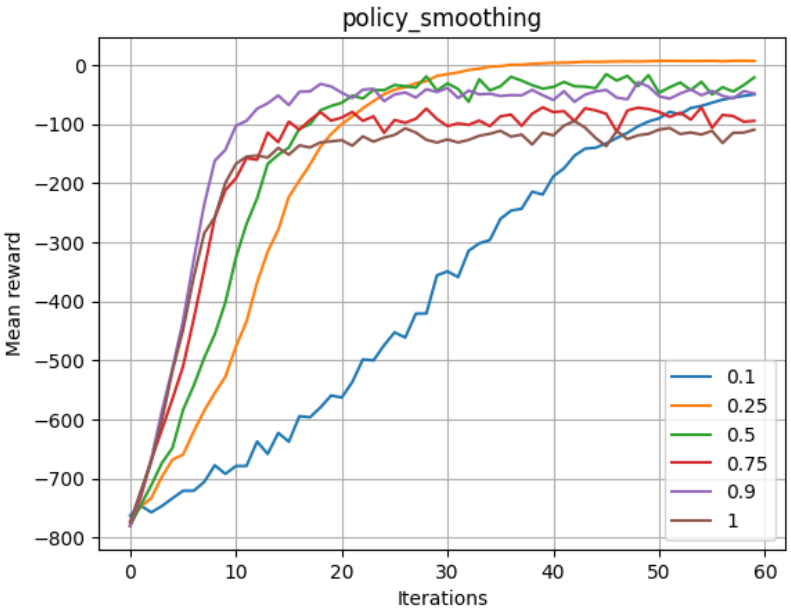
Была выбрана квантиль 0.5 и далее для неё я просто увеличил число траекторий(что самое важно — стартовых) и модель смогла сойтись к оптимальному решению. Однако для дальнейших экспериментов я решил остановиться на числе в 200 траекторий.

Задача 2

Здесь нужно было реализовать 2 метода сглаживания — Лапласа и политики. По графикам обучения выше видно, что начиная с какого-то момента результаты модели начинают колебаться около горизонтальных прямых и не могут «нащупать»

поступательное движение для того, чтобы сойтись к оптимальному решению. Методы сглаживания я проверял независимо друг от друга.

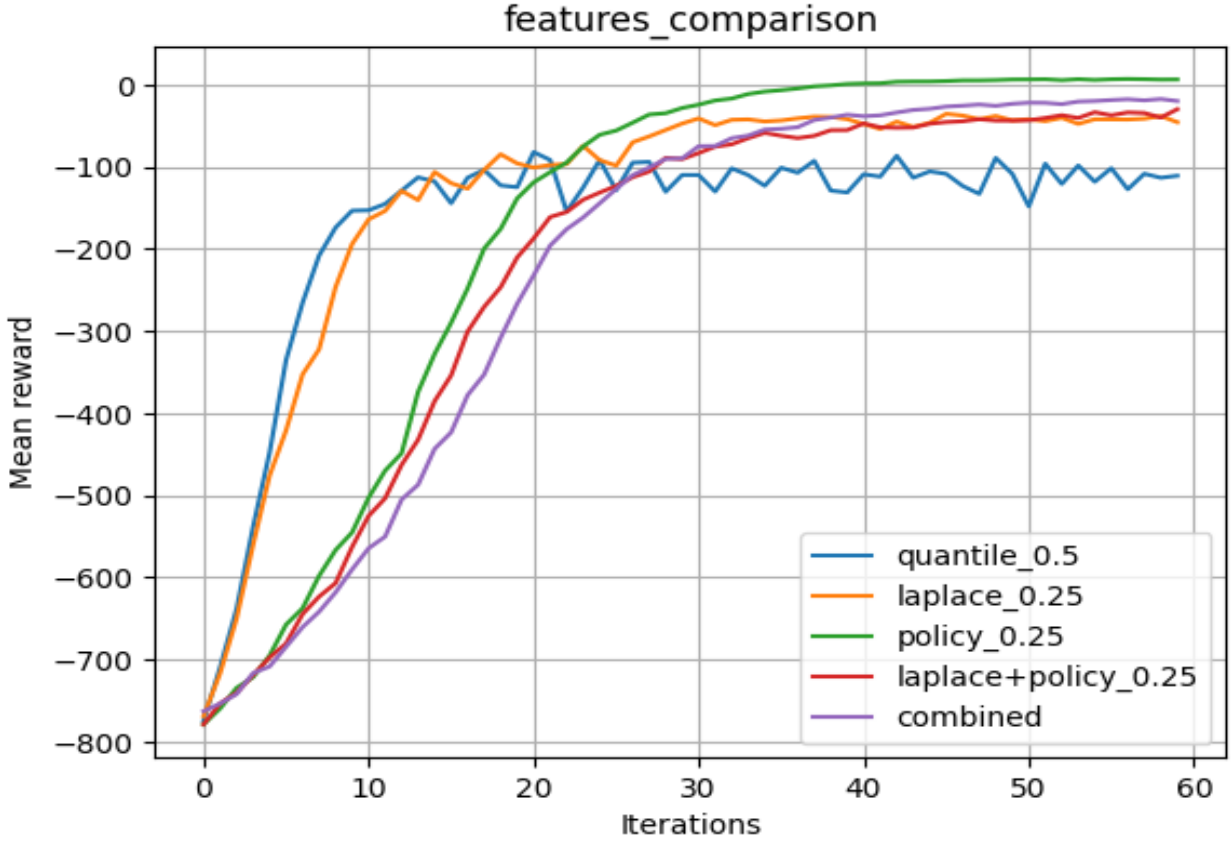




Кстати, сглаживание по Лапласу в случае значения гаммы 0 — просто вырождается (почему-то в лекции стояло условие, что гамма строго положительная). Основное сравнение как раз интересует с синей кривой. Высокие значения гаммы приводят к тому, что модель уже не в состоянии учиться(как было и при низких значениях квантили). Однако при низких значениях гаммы, получается добиться более высоких результатов + уменьшить дисперсию средней награды между итерациями.

Сглаживание политики — бомба! При значении 1 оно вырождается, поэтому сравниваем с коричневой кривой. Понижая, значения гаммы модели становятся всё лучше, а что самое важное, графики становятся гладкими. Чем-то напоминает момент как улучшение SGD. Однако при слишком сильно сглаживании модель перестаёт понимать, как ей учиться.

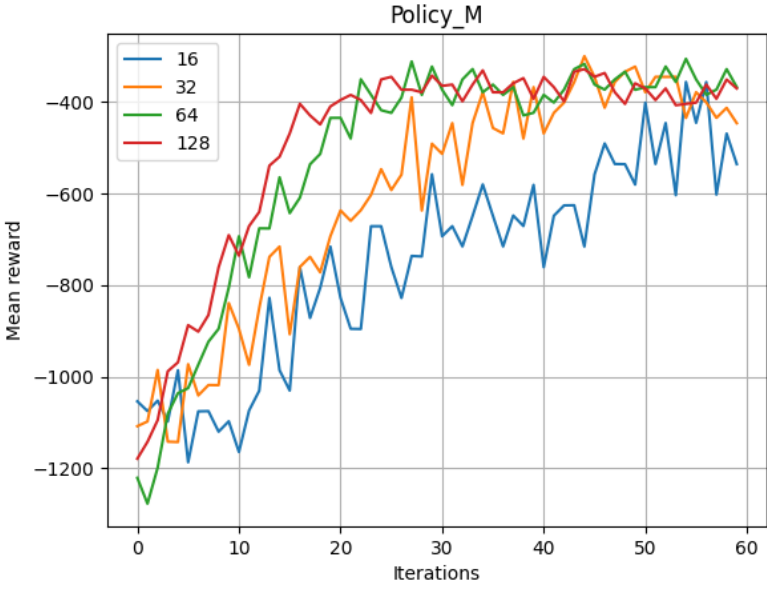
Потом я попробовал совместить идеи и посмотреть, что из этого получится:

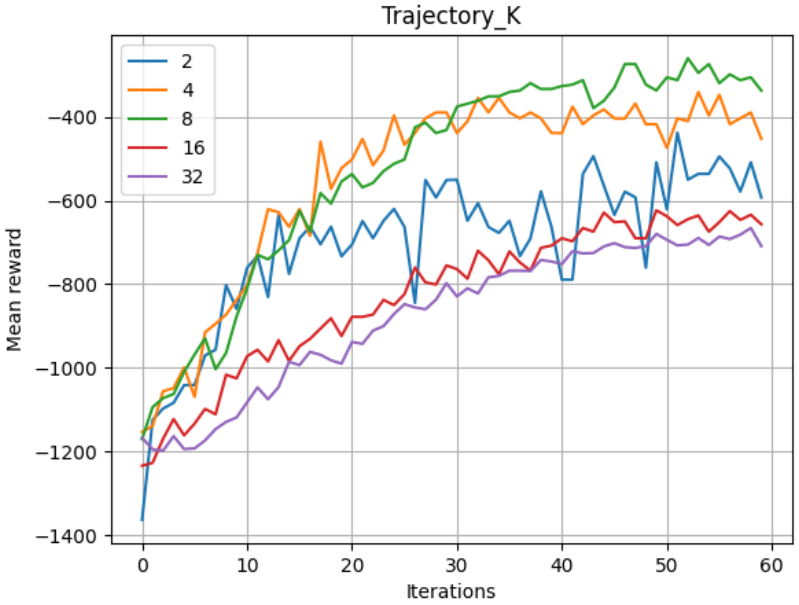


Тут видно, что именно сглаживание политики приносит наибольший эффект(как и выбор квантили). А разные методы сглаживания, не всегда хорошо сочетаются друг с другом. Рецепт такой. Сглаживание Лапласа — не нужно. Берём побольше траекторий и подбираем оптимальную пару: квантиль и гамму для сглаживания по политике.

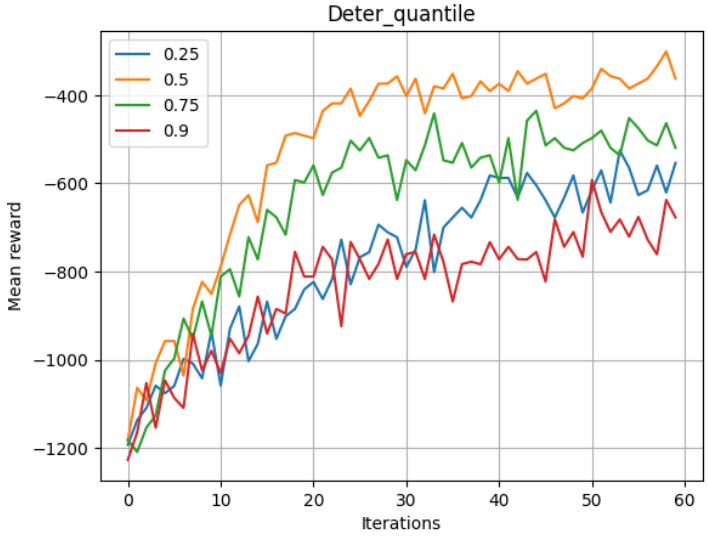
Задача 3

Результаты последнего метода расстраивают, но их легко объяснить. Наша модель сильно зависит от начального семплирования. Здесь, чтобы хоть какого-то остаться в рамках вычислительной сложности, мы вынуждены уменьшить число траекторий и число детерминированных политик. Жертвовать в первую очередь будем вторым, т. к. нам важно уметь корректно вычислять квантили и с одной стороны сразу откидывать явный мусор, с другой стороны иметь достаточный выбор для исследований. И здесь кроется жаба. Даже если мы умеем отбирать политики, то кто сказал, что все траектории этой политики хорошие? Наоборот, легко предположить, что там есть как и хорошие траектории так и плохие. После отсеивания по квантилям мы лишь выбрали тот набор детерминированных политик, в котором число хороших траекторий, больше чем плохих. И всё. После этого очевидно, что в «градиенте» обучения будет много шумов и плохих траекторий. Посмотрим на результаты привычным способом(фиксируем все параметры, кроме одного).





Чем больше политик — тем лучше мы умеем отличать оптимальные от неоптимальных. С другой стороны, взяв какую-то политику, чем больше траекторий из неё мы просемплируем — тем лучше оценим, хорошая она или плохая. Если брать только одну траекторию — теряется фишка метода, да и нам могло просто повести с начальной точкой. Увеличивая число траекторий приходим к тому, что попадается всё больше плохих траекторий внутрь хорошей политики и модель перестаёт учиться.



Видно, что результаты сильно далеко от того, что показывает стандартный метод кросс-энтропии. Я решил дать методу последний шанс, попробовав подобрать другое значение квантили(ведь её смысл изменился).

Однако результатов это не принесло.