HW 1

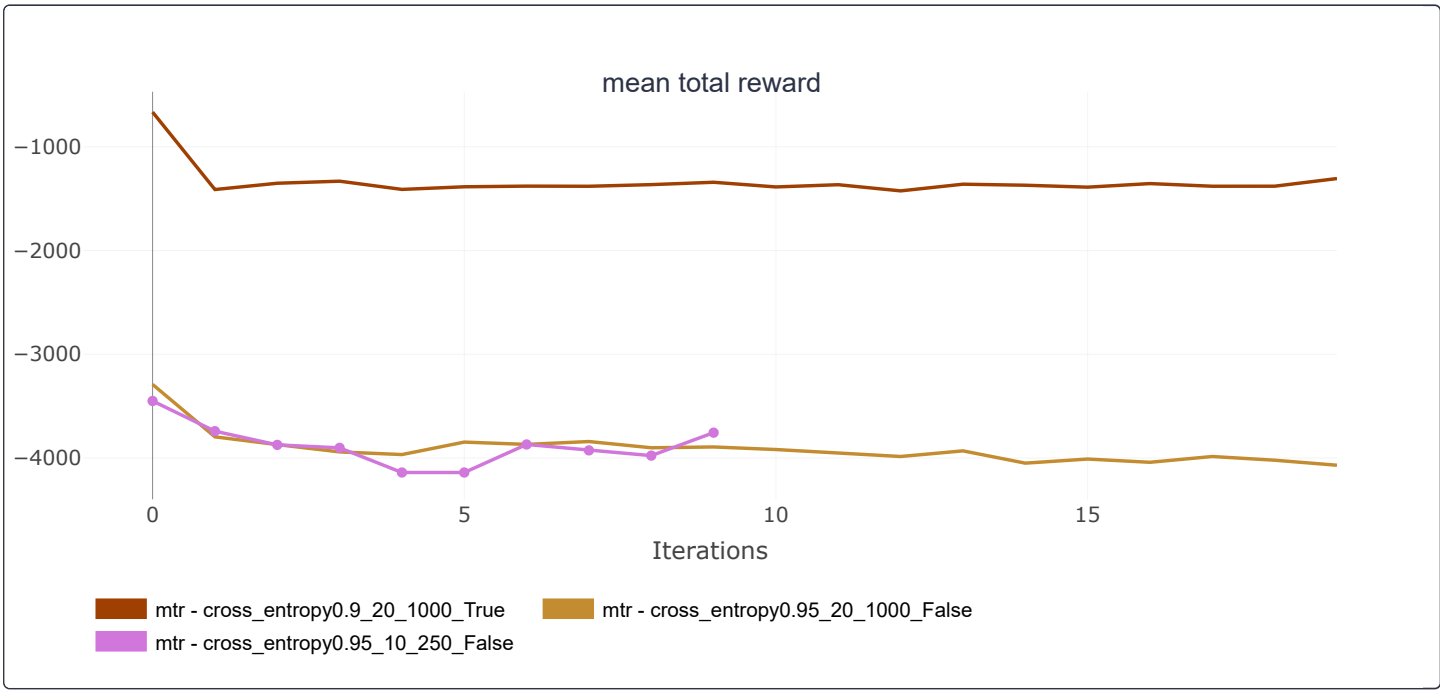
Код находится в гитхабе

Зайдите, пожалуйста, в дискорд. Было несколько вопросов, и я не до конца разобрался правильно ли я делаю или нет. (

кросс-энтропия

В процессе исследования исследовался алгоритм кросс-энтропии для задачи Такси. Было проведено множество экспериментов, исследовано множество гипер-параметров. На графике ниже показаны результаты. По оси абсцисс изображены итерации, по оси ординат -- среднее по итерации средних по траекториям награда. В названии графиков содержаться гиперпараметры прогона. По-порядку:

1. лямбда-параметр
2. количество итерации
3. количество траекторий
4. булево значение, которое отвечает за логирование в ClearML.



По результатам экспериментов, было замечено, что квантиль после нескольких итераций не понимается выше отметки -1000 (1000 -- это максимальное количество шагов, в методе .get_tajectory). Можно предположить, что в результате обучения агент "не хочет" выбирать действия, за которые накладывается бОльший штраф (подбор пассажира, высадка пассажира). Поэтому решает, что выгоднее просто перемещаться по полю, получая штраф -1.

сглаживание

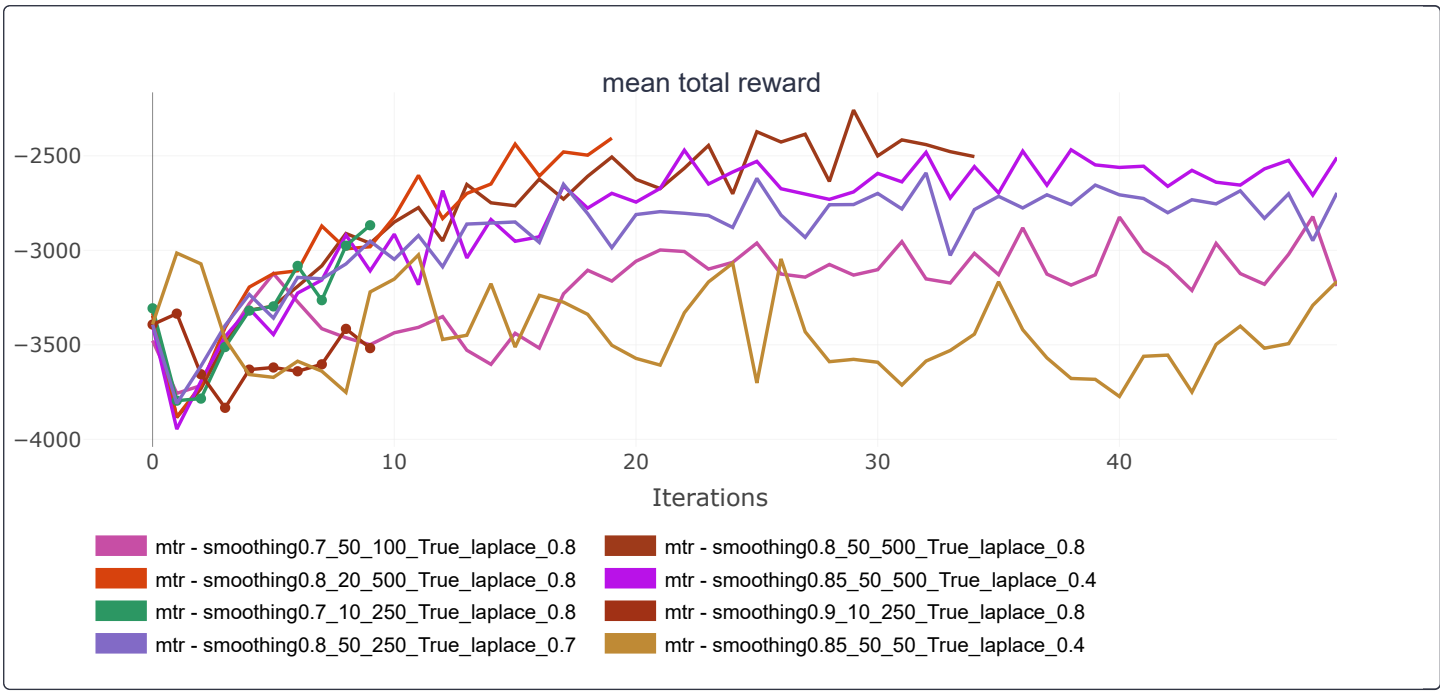
Сглаживание должно исправить проблему, поставленную в предыдущем абзаце, так как вероятность выбора действия никогда не будет равна 0.

сглаживание по Лапассу

Было замечено, что при уменьшении q-параметра, значения награды возрастают. Возможно, стоит использовать какое-то адаптивное значение.

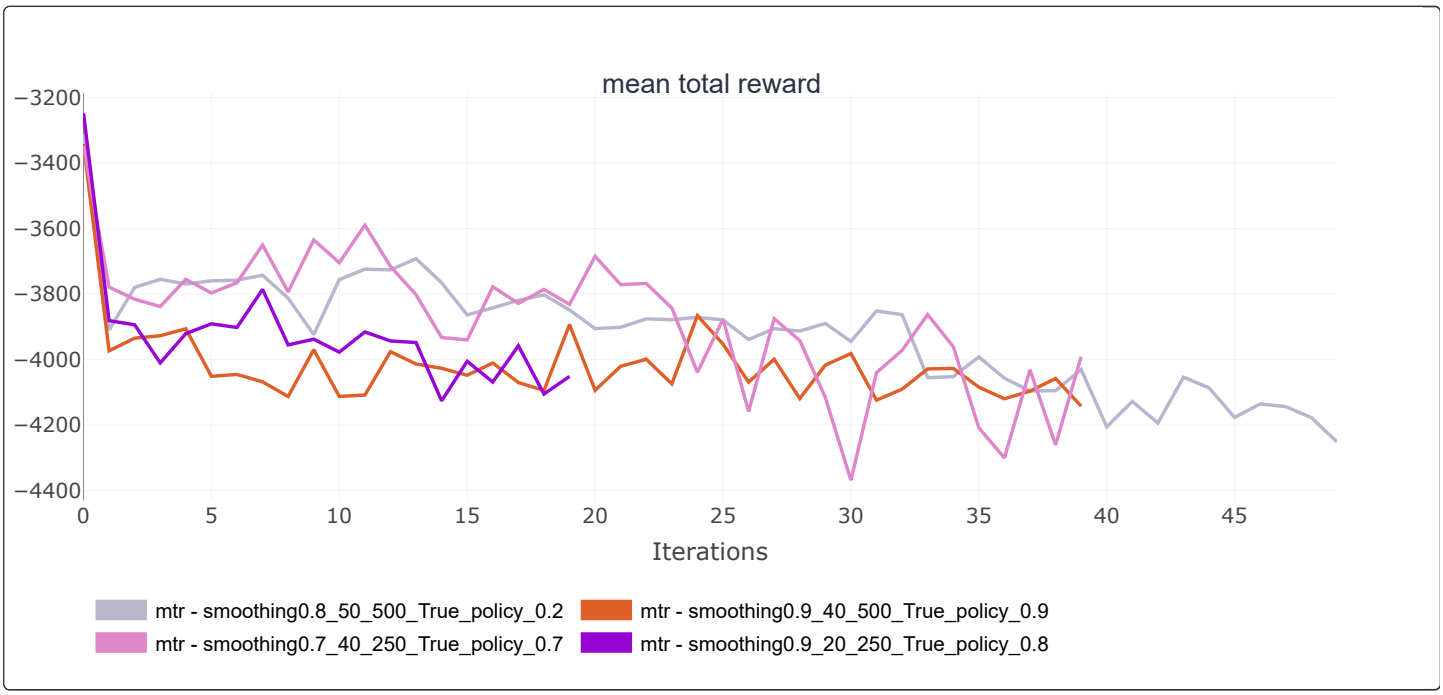
С уменьшением количества трайекторий возрастает дисперсия.

Результаты показаны на графике:



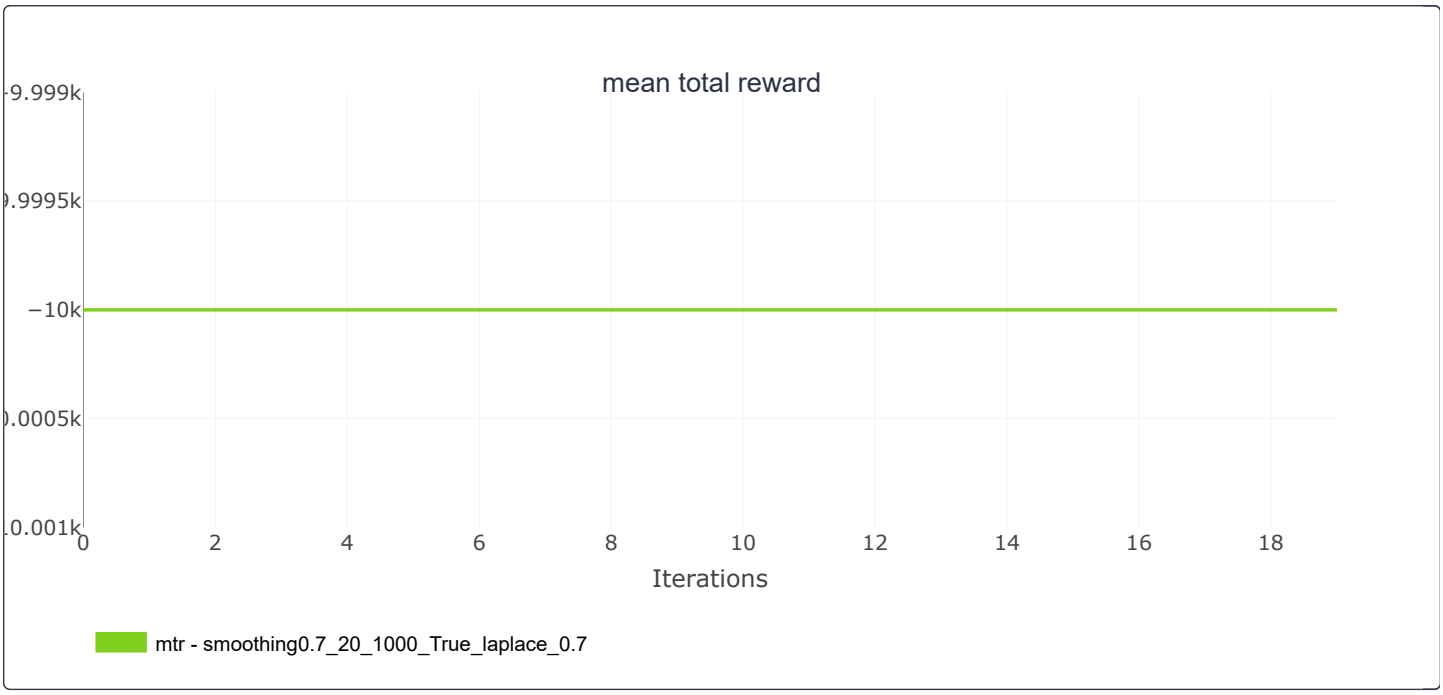
policy сглаживание

Для policy хороших результатов добится не удалось



deterministic polices

Детерменированные политики просто не получились


mean total reward
-9.999k
-9.9995k
−10k
-10.0005k
-10.001k
0
2
4
6
8
10
12
14
16
18
Iterations
mtr - smoothing0.7_20_1000_True_laplace_0.7