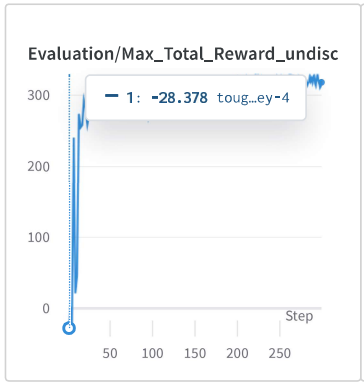
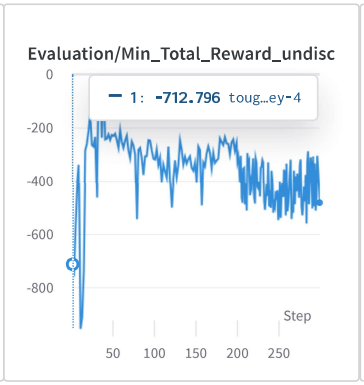
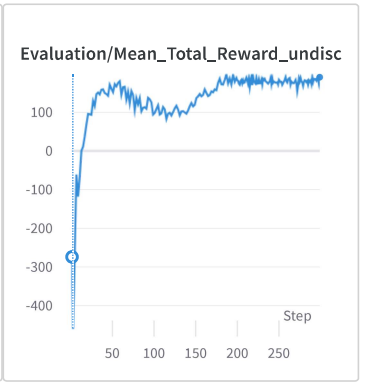
Отчет под Д32 DLRL'23

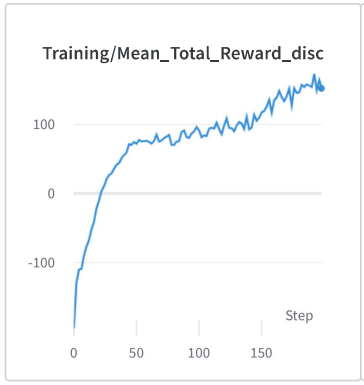
Pavel Nakaznenko

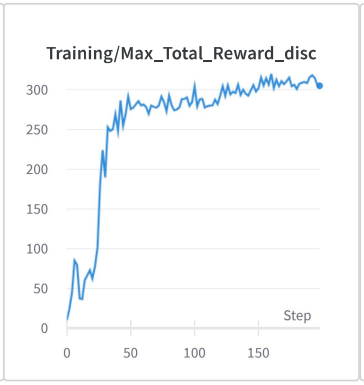
LunarLander-v2











--trajectory_n 500

--max_trajectory_len 5000

--episode_n 500

--gamma_q 0.9

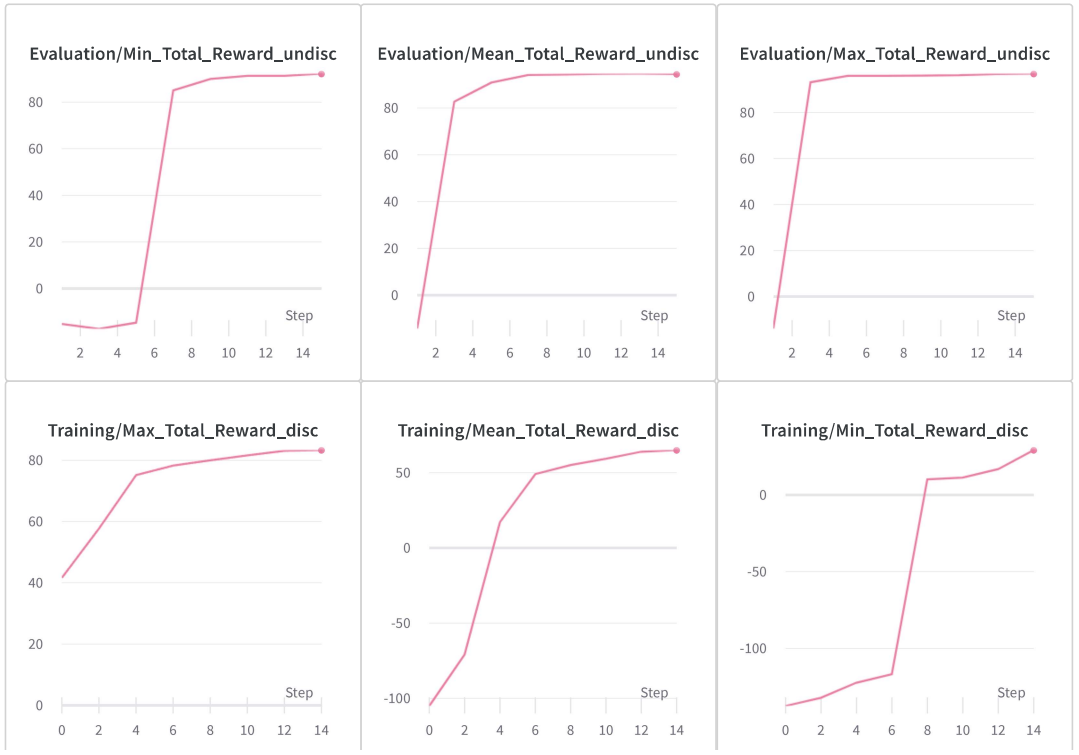
--lr 0.001

--exploration 0.0

--gamma_discount 1.0

Поскольку была выбрана высокая элитарность и отсутствие шума, было принято решение увеличить количество эпизодов. Дисконтирование reward (в том числе с gamma > 1) не показало себя эффективным, потому что reward функция и так довольно неплохо подсказывает направление для улучшения. В отличии от MountainCar

MountainCarContinuous-V0



--trajectory_n 500

--max_trajectory_len 1000

--episode_n 100

--gamma_q 0.99

--delta_q 0.2 - квантиль для накапливаемого между эпизодами пула элитных траекторий, отобранных по квантилю gamma_q

--lr 0.001

--exploration 0.5

--gamma_discount 1.0

В этой среде очень "непрощающая" reward функция. Ближайший локальный оптимум - стоять и ничего не делать (total_reward = 0). Чтобы избежать этого локального оптимума, мы добавляем много шума и делаем высокую элитарность при большом количестве траекторий. Дисконтировать действия здесь не имеет смысла, т.к. таковое дисконтирование заложено в смысл reward функции: штрафовать за все, кроме финального успеха. Данная среда хорошо показывает на сколько плохо подходит наш метод для подобного рода задач.