УДК 004.83

РЕШЕНИЕ ЗАДАЧИ ОБУЧЕНИЯ С ПОЛКРЕПЛЕНИЕМ В СРЕДНЕЙ ПОСТАНОВКЕ МЕТОДАМИ ЛИНЕЙНОГО ПРОГРАММИРОВАНИЯ

Артеменко А.О. (Университет ИТМО)
Научный руководитель – аспирант Асадулаев А.А.
(Университет ИТМО)

Введение. Большинство алгоритмов обучения с подкреплением оптимизируют **дисконтированную** награду, то есть пытаются получить награду как можно быстрее, что способствует ускорению сходимости и снижению дисперсии оценок. Однако такой подход неравнозначно оценивает состояния, что может негативно отражаться на сходимости и достижимости в некоторых задачах[1]. Несмотря на то, что дисконтированная постановка является уместной в определенных сферах, таких как финансовые задачи, во многих инженерных задачах будущие вознаграждения рассматриваются равнозначно, и предпочтение отдается постановке долгосрочного **среднего вознаграждения**[2]. Предлагается посмотреть на эту задачу с новой стороны, и не сводить готовые решения от дисконтированной постановки к средней, а разработать новый подход основываясь на изначальной постановке задачи среднего подкрепления с использованием методов **линейного программирования**.

Основная часть. Данный алгоритм получается из преобразования постановки задачи и перехода к двойственной формулировке методами линейного программирования. В ходе обучения строится модель среды, которая ускоряет обучение политики, а также даёт возможность обучаться на исторических данных без взаимодействия со средой. Поиск оптимальной политики может также быть расписан, как двойственная задача линейного программирования, в которой выбор действия описывается как стационарное распределение состояний-действий.

Выводы. В области обучения с подкреплением выбор алгоритмов для решения задачи среднего обучения крайне ограничен, и у всех у них есть свои недостатки. Был разработан алгоритм средней награды. А также было проведено тестирование метода на множестве контрольных сред.

Список использованных источников:

[1] Amit, R., Meir, R., and Ciosek, K. Discount factor as a regularizer in reinforcement learning. In International conference on machine learning, pp. 269–278. PMLR, 2020.

[2] Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. Optimality and approximation with policy gradient methods in markov decision processes. In Conference on Learning Theory, pp. 64–66. PMLR, 2020.