УДК 004.83

ОПТИМИЗАЦИЯ АГЕНТОВ ОБУЧЕНИЯ С ПОДКРЕПЛЕНИЕМ МЕТОДАМИ ОПТИМАЛЬНОГО ТРАНСПОРТА

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

Введение. Область обучения с подкреплением активно развивается в различных сферах. Обучение с подкреплением – это обучение методом проб и ошибок, в котором агент взаимодействует со средой, чтобы обучить стратегию поведения, которая максимизирует полученную награду. Большинство алгоритмов обучения основаны на дисконтированной формулировке проблемы. Награда, получаемая после k шагов взвешивается с учётом текущего состояния и номера k. Таким образом, дисконтированный фактор уменьшает вклад наград в долгосрочной перспективе[1]. Однако во многих практических задачах, дисконтированный Марковский процесс противоречит ественным мертикам. Также в дисконтированной постановке используется гиперпараметр, который настраивается эмпирическим путем и может сильно повлиять на результат работы алгоритма. В отличие от дисконтированной постановки, средняя постановка фокусируется на продолжительном среднем результате, максимизируя среднюю награду за шаг с равными весами.

Основная часть. Из Блэквелловской теории оптимальности марковских процессов[2] известно, что если дисконтированное значение очень близко к единице, то можно достичь оптимальности в средней постановке, однако на практике алгоритмы часто из-за этого ломаются, когда дисконтированное значение близко к единице.
Средняя постановка Марковских процессов получила меньше внимания по сравнению с дисконтированной формулировкой. Поэтому был разработан алгоритм в постановке средней награды с использованием нейронных сетей. Алгоритм основан на двойственной задаче линейного программирования для задачи обучения с подкреплением. В работе представлены:

1) Задача оптимизации средней награды с множителями Лагранжа.
2) Процедура оптимизации для двойственной задачи.
3) Новый model-based алгоритм средней награды, основанный на нейронных сетях.

Выводы. Был разработан алгоритм средней награды. А также было проведено тестирование метода на множестве контрольных сред.

Список использованных источников:

1. Andrychowicz M., Raichuk A., Stańczyk P., Orsini M., Girgin S., Marinier R., Hussenot L., Geist M., Pietquin O., Michalski M., Gelly S., Bachem O. What Matters In On-Policy Reinforcement Learning? A Large-Scale Empirical Study // arXiv:2006.05990. - 2020
2. Blackwell D. Discrete dynamic programming // The Annals of Mathematical Statistics. - 1962. - С. 719-726.