УДК 004

АЛГОРИТМ РЕГУЛЯРИЗАЦИИ ПРОКСИМАЛЬНОЙ ОПТИМИЗАЦИИ ПОЛИТИКИ

Асадулаев А.А. Университет ИТМО,
Научный руководитель – Доцент ФИТиП, Фильченков А.А..
Университет ИТМО

Установить эффективные границы в алгоритме Проксимальной Оптимизации Политики часто проблематично. Чтобы понять, как нам нужно ограничивать обновление политики на каждом этапе, нам нужны косвенные характеристики политики в дополнение к полученным вознаграждениям, поскольку вознаграждения не всегда доступны. В нашей статье мы изучаем способность обусловленности агента быть такой характеристикой.

Введение. В некоторых случаях обновления происходят небольшими шагами, что может привести к проблемам с исследованием среды агентом. Агенты могут недостаточно изучать свою среду или разрабатывать стратегии. Если агент попадает в ловушку в некоторых состояниях в начале обучения, опыт, полученный в других местах среды позже, может не привести к значительным изменениям политики, поскольку диапазон клипов стал слишком маленьким.

Основная часть. В сложных условиях, когда переход из одной области среды в другую не является «тривиальным», агент может застрять в локальном минимуме. В результате агент имеет продуктивную стратегию только в небольшой области среды, что приводит к непредсказуемым результатам политики для других состояний. Для решения данной проблемы мы пытаемся ответить на следующий вопрос: Является ли обусловливание приемлемой оценкой политики и может ли информация о формировании политики помочь сформировать более всеобъемлющую и надежную политику?

Мы проводим серию экспериментальных исследований, касающихся взаимосвязи между эффективностью политики и его обусловленностью. Поскольку оценка обусловленности с использованием Singular Value Decomposition (SVD) отнимает много времени, мы адаптировали методику, разработанную для оценки обусловленности генеративных моделей для агентов обучения с подкреплением. Наши эксперименты демонстрируют соответствие между обусловленностью агента и соотношением достигнутых вознаграждений.

Основываясь на этих наблюдениях, мы предлагаем алгоритм, который использует обусловленность агента для формирования надежной политики. В нашем алгоритме вносится регуляризация для алгоритма Проксимальной Оптимизации Политики. Предложенная регуляризация напрямую оценивает обусловленность агента и, исходя из этого, меняет политическую тенденцию. Мы демонстрируем все результаты задач непрерывного управления в таких средах PyBullet как Humanoid-v0, Hopper-v0, Ant-v0 и Reacher-v0.

Выводы. В результате наших экспериментов мы пришли к заключению что обусловленность для большинства сред связана с успешностью политики. На основе этого мы представляем улучшенную версию алгоритма Проксимальной Оптимизации Политики, который динамически адаптирует величину обновление политики на каждом этапе, в зависимости от обусловленности агента.

Асадулаев А.А. (автор) Подпись

Фильченков А.А. (научный руководитель) Подпись