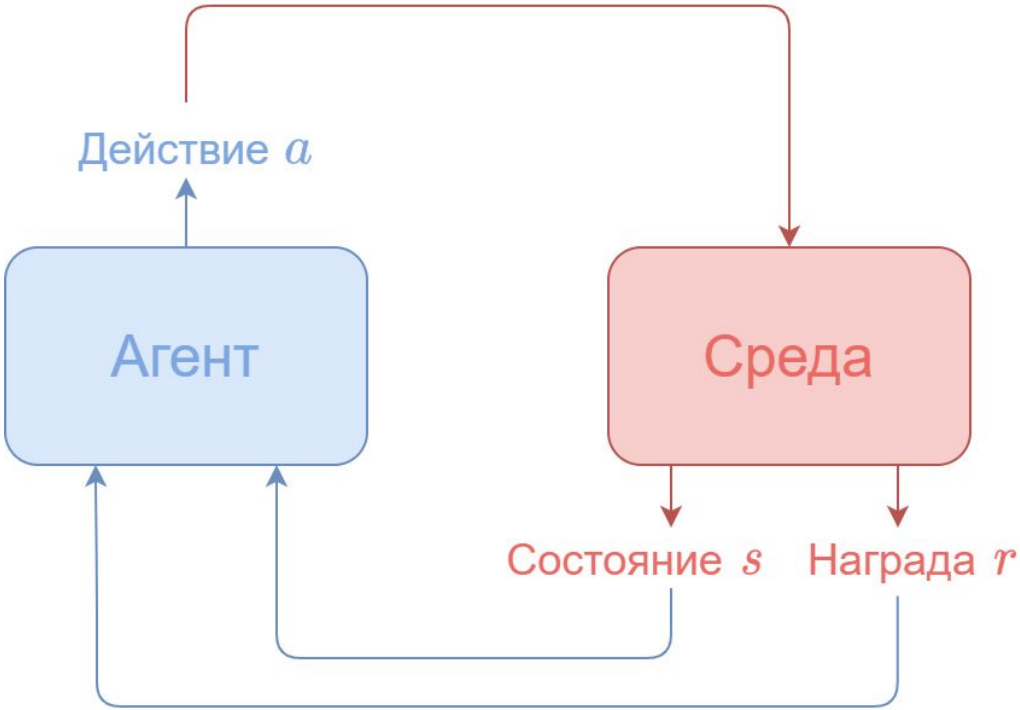
Кооперация в мультиагентном обучении с подкреплением

Егоров Владимир Сергеевич
Научный руководитель:
к.т.н. Браславский Павел Исаакович
Научный консультант:
Шпильман Алексей Александрович

НИУ ВШЭ Спб,
2020

Санкт-Петербургская школа физико-математических и компьютерных наук

Обучение с подкреплением



Мультиагентные среды

- Кооперативные (Starcraft 2 Team Mode [1])
- Соревновательные (Go [2])
- Смешанные (Sequential Social Dilemmas [3])

Изменение награды агента для продвижения кооперации в смешанных средах (Cooperative Reward Shaping, CRS).

Функции для оценки общего благосостояния (Social Welfare, SW):

- **Взвешенная сумма наград агентов [4, 5]**
  - Обучение набора политик разной степени кооперации [6]
  - Иерархическое обучение с подкреплением [7]
    - Максимизация суммы наград мета-агентом
- **Минимум наград агентов**

[1] Vinyals O. et al. – 2017., [2] Silver D. et al. – 2016., [3] Leibo J. Z. et al. – 2017., [4] Peysakhovich A., Lerer A. – 2018., [5] Hughes E. et al. – 2018., [6] Wang W. et al. – 2019., [7] Zheng S. et al. – 2020.

Проблемы

- Выбор функции Social Welfare
  - CRS ограничен функциями, коммутативными с мат. ожиданием (только сумма).
    - Цель агента — максимизация ожидания кумулятивной награды
  - Использование минимума приводит к непредсказуемому поведению.
- Артефакты, связанные с продолжительностью жизни агентов
  - Цель агентов — максимизация продолжительности жизни
  - Агент избегает получение негативной награды от остальных агентов, завершая свою жизнь раньше других.

Постановка задачи

Цель проекта заключается в создании нового метода для продвижения кооперации в смешанных мультиагентных средах, который решает проблемы, связанные с продолжительностью жизни агентов и функцией Social Welfare

Задачи:

- Разработать новый метод продвижения кооперации
- Реализовать среду, в которой агент влияет на продолжительность своей жизни
- Оценить метод в средах и сравнить его с Cooperative Reward Shaping

Lawmaker

Оценка благосостояния *i*-го агента — оценка его политики $\pi_i : \mathsf{P}(\mathcal{A}_i \mid s)$:

$R_i(s) = \sum_{t=0}^{\infty} \gamma^t r_i$ — кумулятивная награда *i*-го агента

$V_{\pi_i}(s) = \mathbb{E}_{\pi_i}[R_i(s)]$ — функция ценности политики *i*-го агента

Оценка общего благосостояния агентов (Social Welfare, SW):

Cooperative Reward Shaping:

$$V^{SW} = \mathop{\mathbb{E}}_{\pi}\left[\sum_{t=0}^{\infty}\left(\gamma^t SW(\mathbf{r}_t)\right)\right]$$

Lawmaker:

$$V^{SW} = SW\left(\mathop{\mathbb{E}}_{\pi}\left[\sum_{t=0}^{\infty}\left(\gamma^t \mathbf{r}_t\right)\right]\right)$$

Задача мета-агента Lawmaker:

Максимизировать Social Welfare V^{SW}

Функция оценки действий агентов

Social Advantage - оценка влияния действий на Social Welfare:

$$A^{SW}(s, \mathbf{a}) = SW(\mathbf{V}'(s, \mathbf{a})) - SW(\mathbf{V}(s))$$

$V_{\pi_i}(s)$ — функция ценности *i*-го агента

$V'_{\pi_i}(s, \mathbf{a})$ — функция ценности *i*-го агента при условии, что следующие действия будут **a**

Следующая задача - оценить вклад в Social Advantage индивидуальных агентов:

$A^{SW}(s, \mathbf{a}) = \sum_i A_i^{SW}(s, a_i)$ - делаем предположение, что Social Advantage раскладывается по-агентно

Архитектура Lawmaker

Архитектура Actor-Critic:

$$\pi_i^{\oplus}(s) = \frac{\pi_i(s) \cdot \pi_i^{SW}(s)}{\sum_{a'} \pi_i(s) \cdot \pi_i^{SW}(s)}$$

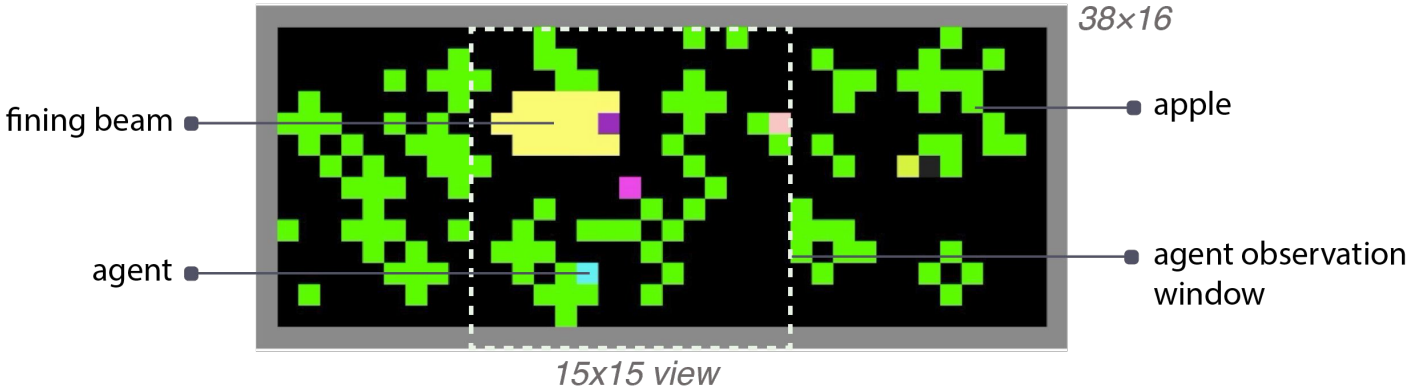
$\pi_i^{SW}(s)$ максимизирует $A_i^{SW}(s, a_i)$

π_i максимизирует личную награду i-го агента

Среды

Harvest [8]

- Пять агентов
- Сбор яблок

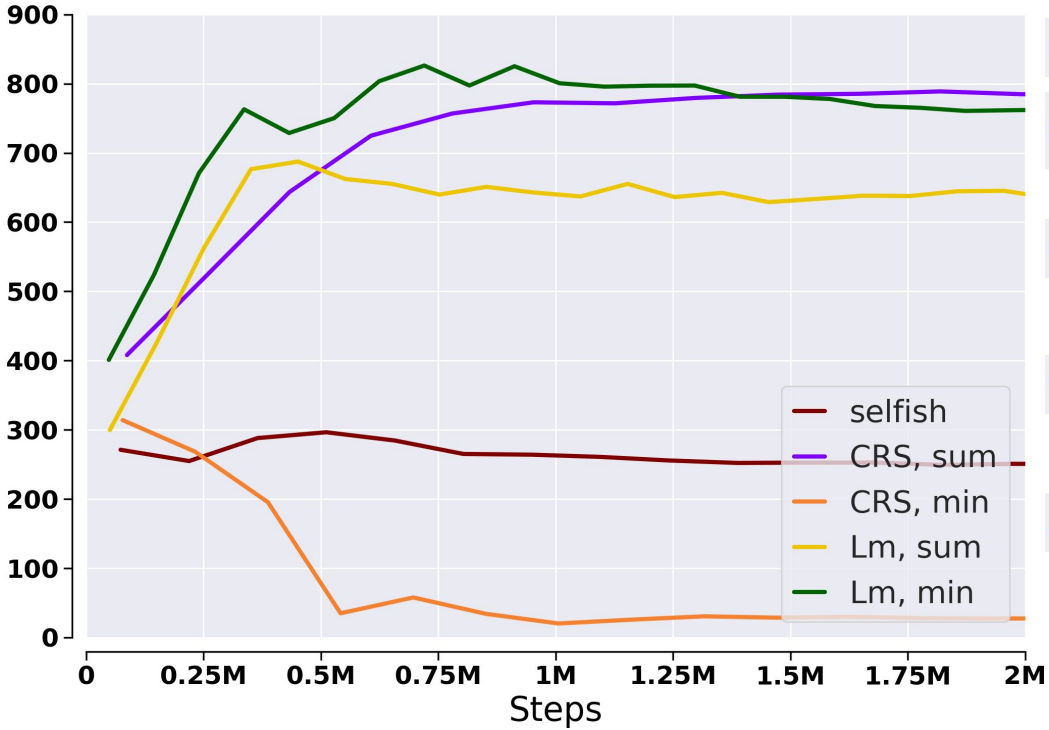


- Фиксированная продолжительность жизни
- Скорость вырастания яблок зависит от количества яблок вокруг

[8] Hughes E. et al. Inequity aversion resolves intertemporal social dilemmas //arXiv preprint arXiv:1803.08884. – 2018.

Результаты, Harvest

Общее количество яблок за жизнь агентов



selfish — Эгоистичный бэйзлайн

CRS, sum — CRS с суммой как функцией SW

CRS, min — CRS с минимумом как функцией SW

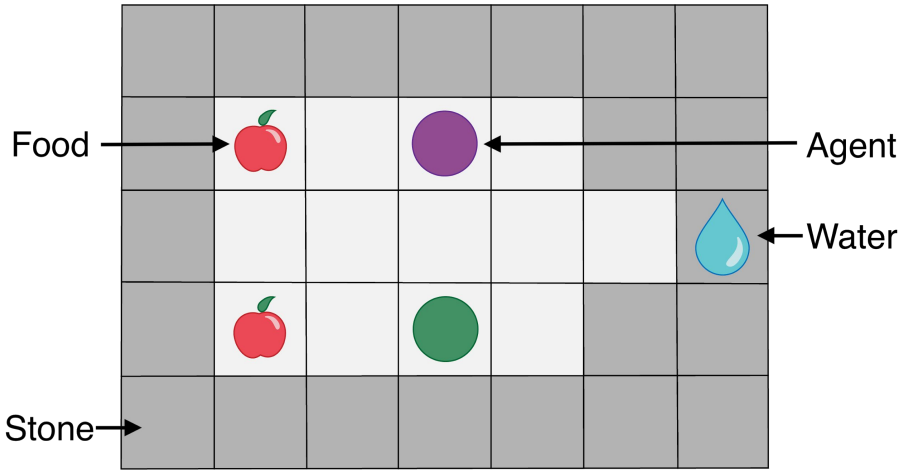
Lm, sum — Lawmaker с суммой как функцией SW

Lm, min — Lawmaker с минимумом как функцией SW

Среды

Eldorado (основано на [9])

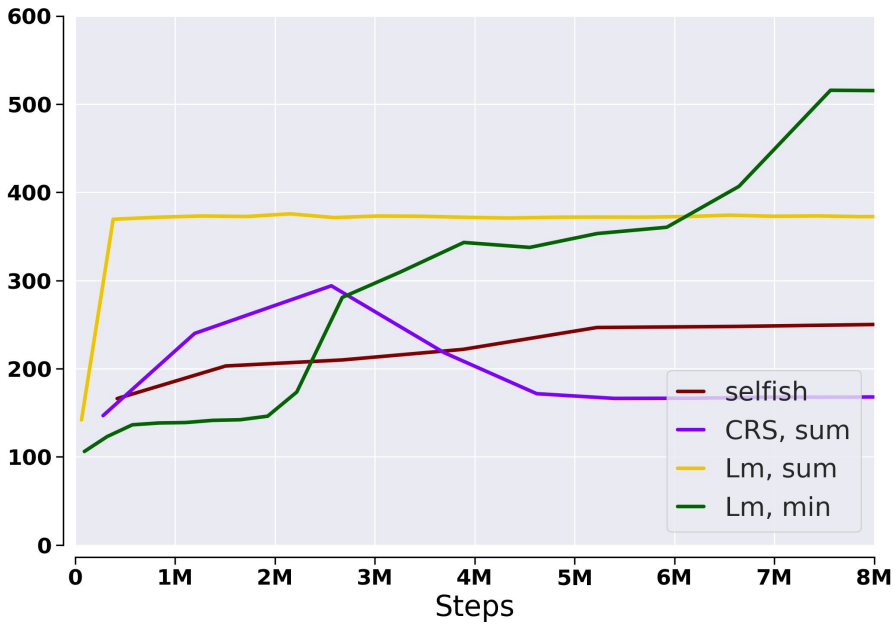
- Два агента
- Разреженная награда
- Сбор ресурсов
- Атака оппонента
- Положительная награда за шаг
- Отрицательная награда в конце жизни



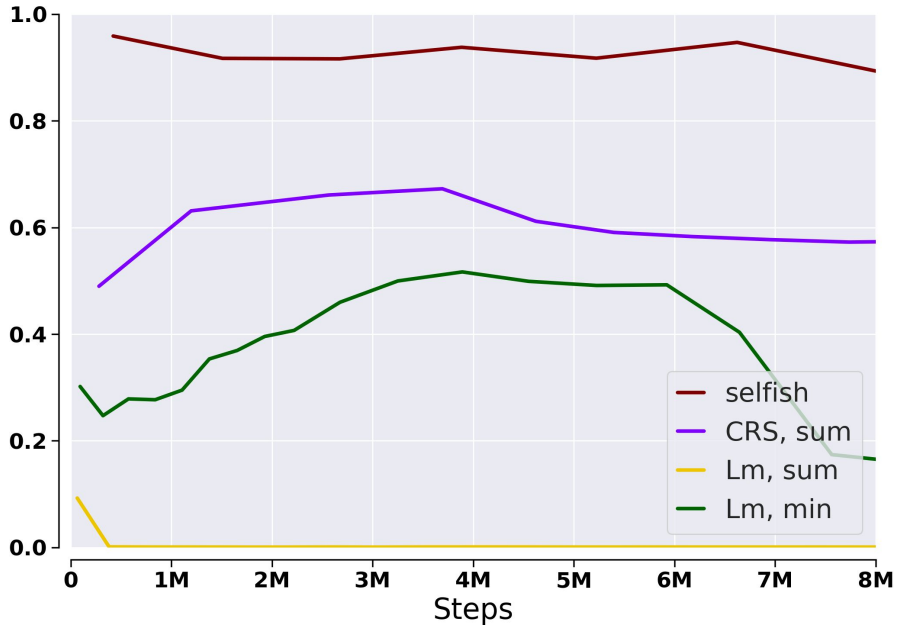
[9] Suarez J. et al. Neural mmo: A massively multiagent game environment for training and evaluating intelligent agents //arXiv preprint arXiv:1903.00784. – 2019.

Результаты, Eldorado

Суммарная продолжительность жизни



Процент атаки



Итоги

- Реализован метод продвижения кооперации Lawmaker, основанный на изменении политики агента и решающий проблемы, связанные с продолжительностью жизни и функцией Social Welfare
- Реализована среда Eldorado, в которой агент влияет на продолжительность своей жизни
- Эмпирически показана эффективность Lawmaker в сравнении с CRS в средах Harvest и Eldorado
- По результатам исследования подготовлена публикация

Код работы: https://github.com/vladimirrim/bachelor_thesis