# МЕТА-ОБУЧЕНИЕ С ПОДКРЕПЛЕНИЕМ РОБОТИЗИРОВАННОЙ СИСТЕМЫ

**А. В. Сидоренко, К. А. Акула**

*Белорусский государственный университет, Минск, Беларусь*
*E-mail: sidorenkoa@yandex.ru, Ksenea1@gmail.com*

Проведен вычислительный эксперимент с использованием разработанной компьютерной программы по мета-обучению с подкреплением агентов в модели роботизированной системы. Показано, что по определенному значению функции вознаграждения в модели роботизированной системы можно определить прекращение процесса обучения.

*Ключевые слова: мета-обучение с подкреплением; компьтерная программа; модель; роботизированная система.*

***Введение.*** Мета-обучение с подкреплением (Meta Reinforcement Learning, meta-RL) представляет собой систему методов решения задач посредством комбинирования мета-обучения и обучения с подкреплением (RL). После обучения агента решению набора задач, он может решить новую задачу, разработав новый алгоритм обучения с подкреплением с его динамикой внутренней активности.

Хорошо разработанная модель мета-обучения должна распространяться на новые задачи или новые среды, которые ранее не встречались во время обучения. Процесс адаптации, представляющий собой мини-учебную сессию, происходит при тестировании с ограниченным воздействием новых конфигураций. Даже без какой-либо явной подстройки, например, без обратного распространения градиента для обучаемых переменных, модель мета-обучения автономно корректирует внутренние скрытые состояния для обучения.

***Мета-обучение с подкреплением.*** Целью обучения с подкреплением является создание параметризованной политики πθ, которая максимизирует дисконтированное будущее вознаграждение за состояния, выбранные из начального состояния распределения

$$L_T = E\left[\sum_{t=0}^{T} \gamma^t r_t | s_0 \sim p, a_t \sim \pi(s_t), s_{t+1} \sim P(s_{t,a_t})\right] \qquad (1)$$

где $r_t$ – значение вознаграждения, $\gamma$ – коэффициент дисконтирования, $P(s_t, a_t)$ – матрица переходов, $s$ – состояние, $a$ – действие, $\pi(s_t)$ – политика, E- Евклидово пространство.

При мета-обучении с подкреплением имеется распределение p(T): T → [0, 1] по задачам T = {$T_1$,..., $T_N$} и для любой среды, выбранной из этого дистрибутива, необходимо как можно скорее произвести действие. Возможность учета только различий между средами в распределении p(T) позволяет алгоритмам мета-обучения потенциально ис-

пользовать класс сред T, чтобы обучиться новой задаче. При этом Ti приблизительно равно p всего за 1 - 10 взаимодействий. Обучение мета-подкреплению можно представить при использовании следующей целевой функции

$$L_T = min \sum_T E_{\pi\Delta(\theta)}[L_T] \quad (2)$$

где Δ(θ) представляет метод обновления, который происходит в несколько приемов и собирает ограниченное количество опытов каждого Ti, для обновления θ.

Обычно из одного класса задач берутся задачи по обучению и тестированию. Допустим, имеется набор задач, каждая из которых сформулирована как Марковский процесс принятия решений (MDP), причем Mi принадлежит M. Марковский процесс принятия решений определяется четырьмя кортежами, Mi = (S, A, Pi, Ri). В таблице ниже приводятся соответствующие обозначения.

Таблица

**Марковский процесс принятия решений**

| Символ | Значение |
|---|---|
| S | Множество состояний |
| A | Множество действий |
| Pi:S×A×S→R+ | Функция вероятности перехода |
| Ri:S×A→R | Функция вознаграждения |

***Реализация мета-обучения с подкреплением.*** Общая конфигурация мета-обучения с подкреплением похожа на алгоритм обучения с подкреплением, за исключением того, что последнее вознаграждение $r_{t-1}$ и последнее действие t -1 также включены в наблюдение за политикой в дополнение к текущему состоянию $s_t$:

- В RL: $\pi_\theta(s_t) \rightarrow$ распределение по A
- В мета-RL: $\pi_\theta(a_{t-1}, r_{t-1}, s_t) \rightarrow$ распределение по A

Это приводит к тому, все изменения состояния действия хранятся и учитываются в модели, а политика может учитывать динамику между состояниями, вознаграждением и действиями в текущем MDP и соответствующим образом скорректировать свою стратегию. Поскольку политика является периодической, нет необходимости явно указывать последнее состояние в качестве входных данных. Процедура обучения работает следующим образом:

1. Ввести новый MDP, Mi приблизительно равно M.
2. Сбросить скрытое состояние модели.
3. Обновить вес модели, собрав несколько траекторий.
4. Повторить с первого этапа.

Разработанная нами компьютерная программа позволила провести вычислительный эксперимент по мета-обучению с подкреплением. На рисунке 1 приведены результаты вычислительного эксперимента, проведенного нами при мета-обучении с подкреплением системы агентов в модели роботизированной системы.

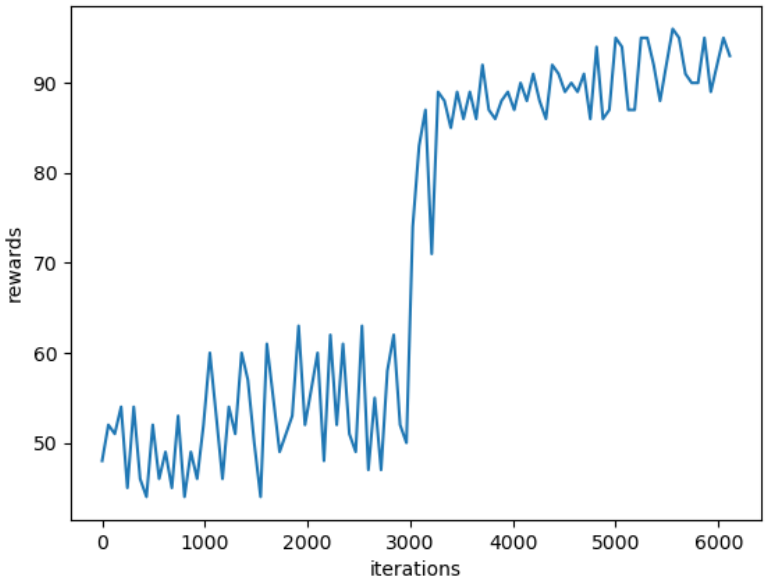


Рис. 1. График зависимости вознаграждения
от количества итераций в модели

Из графика видно, что при изменении количества итераций обучения системы от 0 до 3000 величина вознаграждения линейно возрастает с увеличением количества итераций. При количестве итераций, превышающих 3000, функция вознаграждения испытывает скачок, резко увеличивается, что позволяет сделать вывод о прекращении обучения.

## БИБЛИОГРАФИЧЕСКИЕ ССЫЛКИ

1. Vuong Q. et al. Meta reinforcement learning from observational data [Electronic resource]. – Mode of access: https://arxiv.org/abs/1909.11373. – Date of access: 12.03.2020.
2. Wang J. X. et al. Learning to reinforcement learn [Electronic resource]. – Mode of access: https://arxiv.org/abs/1611.05763. – Date of access: 13.03.2020.
3. Duan Y. et al. $RL^2$: Fast Reinforcement Learning via Slow Reinforcement Learning [Electronic resource]. Mode of access: https://arxiv.org/abs/1611.02779. – Date of access: 13.03.2020.