# Метод ансамблирования алгоритмов обучения с подкреплением на основе иерархичности

Д.А. Козлов
*Самарский национальный исследовательский университет им. академика С.П. Королева*
Самара, Россия
djoade100@gmail.com

***Аннотация*—В статье предлагается алгоритм ансамблирования нескольких алгоритмов обучения с подкреплением. Предложенный подход действует в среднем эффективнее чем каждый из алгоритмов в ансамбле по отдельности. В статье рассматривается ансамбль из алгоритмов REDQ и SAC. Выходом из ансамбля является выход алгоритма, выбранного с помощью DQN. Возможно ансамблирование других алгоритмов и в другом количестве. Обучение с подкреплением является перспективной областью в машинном обучении. Важной нерешенной задачей обучения с подкреплением является обобщение сложных задач, и решение их при помощи мета-алгоритмов. Предлагаемый метод возможно использовать в сложных задачах, состоящих из многих подзадач, эффективные решения для которых могут предложить различные алгоритмы из ансамбля.**

***Ключевые слова— обучение с подкреплением, soft actor-critic, randomized ensembled double q-learning, deep q-learning, ансамбль, машинное обучение, мета-алгоритм***

## 1. Введение

Обучение с подкреплением – перспективная и динамично развивающаяся область машинного обучения, которая позволяет решать различные задачи в области робототехники [1], биологии [2], рекомендательных систем [3], высокочастотного трейдинга [4].

В ходе работы была выдвинута гипотеза о том, что, если управлять выходом ансамбля из нескольких алгоритмов обучения с подкреплением при помощи алгоритма обучения с подкреплением с дискретным выходом, выбирающим алгоритм для конкретного шага в зависимости от состояния, в которое попал агент, то такая концепция позволит использовать в конкретном состоянии тот алгоритм, который в данном состоянии даст наибольший суммарный доход по всей траектории (последовательность состояний и действий).

Предлагается использовать иерархичность алгоритмов обучения с подкреплением, в ходе которой становится возможным агрегирование выхода ансамбля алгоритмов, в том числе и множества сложно связанных ансамблей.

## 2. Связанные методы

Задача обучения с подкреплением ставится как задача, в которой агент взаимодействует со средой. Взаимодействуя со средой агент получает от неё четвёрку (s, a, r, s’), где s – состояние, некоторый вектор, характеризующий состояние, a – действие, которое совершил агент, тоже некоторый вектор, r – reward, награда, которую агент получил в результате своих действий, скалярное значение, s’ – следующее состояние, в которое мы попали из состояния s путём совершения действия a. Задача агента максимизировать суммарную награду, получаемую от среды.

Также необходимо отметить, что процессы, рассматриваемые в обучении с подкреплением, обладают свойством марковости, это означает, что в состоянии s, следующее состояние зависит только от текущего состояния s и предпринимаемого действия a.

### *A. Алгоритм REDQ*

В состав REDQ (Randomized Ensembled Double Q-Learning) [5] входят три составляющих, которые позволяют ему достигать высоких показателей: коэффициент отношения количества обновлений алгоритма к количеству полученных данных (UTD, update to data) >> 1; ансамбль Q-функций; целевая минимизация случайного подмножества Q-функций из ансамбля.

### *Б. Алгоритм SAC*

В структуре алгоритма SAC (Soft Actor-Critic) [6] сеть Actor стремится максимизировать ожидаемое вознаграждение, а также максимизировать энтропию. Это вынуждает алгоритм достигать наибольшей награды действуя как можно более хаотично. Данная особенность позволяет эффективно решать дилемму исследования-использования.

### *B. Алгоритм DQN*

Алгоритм DQN (Deep Q-Network) [7] содержит дуэльную сеть, представляющую две отдельные оценки: одну для функции ценности состояния и одну для функции преимущества действия, зависящей от состояния.

## 3. Предлагаемый метод

В рамках данной работы рассматривается алгоритм, включающий в себя алгоритм DQN как верхний по иерархии алгоритм или контролирующий алгоритм. Действия, предлагаемые им, будем называть мета-действиями.

Блок схема разработанной модели представлена на рисунке 1. Алгоритм включает в себя следующие шаги:

1. Получение от среды состояния, действия, награды, следующего состояния, мета-действия. Здесь, под получением от среды действия и мета-действия подразумевается некоторая абстракция, позволяющая получить прошлые действия.
2. Передача информации о состоянии и награде всем алгоритмам, информации о действии алгоритмам ансамбля, информации о мета-действии контролирующему алгоритму.
3. Обновление параметров каждого алгоритма.

4. Получение действия от каждого алгоритма и мета-действия. На основе дискретного мета-действия производится выбор одного из действий алгоритмов ансамбля в качестве выходного.

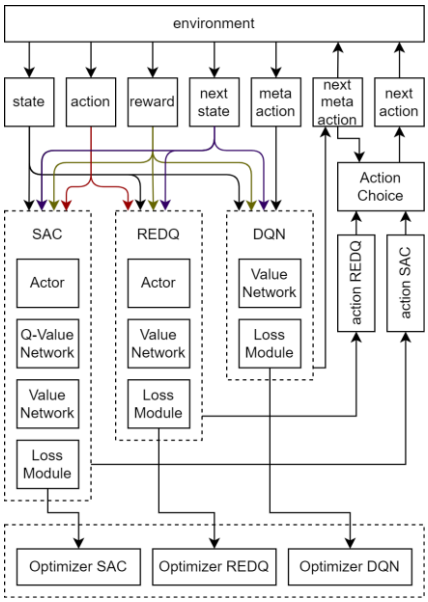


Рис. 1. Рассматриваемый алгоритм

## 4. Экспериментальные исследования

В экспериментах были протестированы реализации алгоритмов REDQ, SAC и реализация предложенного алгоритма в среде dm-control cheetah. Здесь агенту необходимо научиться передвигаться как можно быстрее, путём перемещения конечностей в двумерном пространстве. Каждый тест включал в себя $10^6$ шагов (взаимодействий со средой).

### *А. Результаты экспериментов*

Результативность полученного алгоритма выражается в уровне награды, которую способен достигнуть агент после N шагов обучения. На графике на рисунке 2 изображен график зависимости награды агента от шага обучения. Видно, что предложенный подход даёт более стабильный и гладкий уровень награды на каждом шаге. Значения награды усреднены при помощи экспоненциально взвешенного сглаживания с коэффициентом alpha равным 0,1

### *Б. Эффективность алгоритма*

Рассмотрим выборочную и временную эффективность предложенного алгоритма. Из графика на рисунке 2 видно, что алгоритм обладает большей выборочной эффективностью, чем SAC или REDQ. Выборочная эффективность оценивается по количеству шагов обучения необходимому для достижения некоторой награды. Таким образом, разработанный алгоритм достигает того же уровня награды что и SAC или REDQ за меньшее количество взаимодействий со средой. Временная эффективность, то есть чистое время, необходимое для обучения алгоритма, падает пропорционально количеству алгоритмов в ансамбле, так, если SAC или REDQ требовалось для $10^6$ шагов около 5–7 часов с использованием i7-9700K и RTX 3060, то разработанному алгоритму требуется около 10-11 часов на обучение в течение тех же $10^6$ шагов.

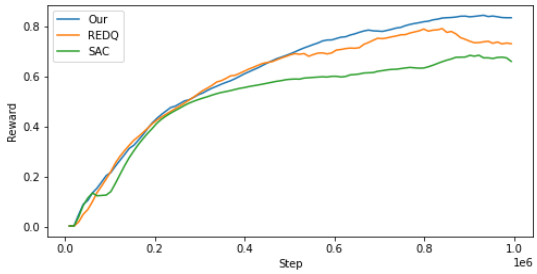


Рис. 2. Результат работы алгоритмов. Сравнение награды достигаемой алгоритмом к определенному шагу

## Заключение

В результате исследования был реализован алгоритм на основе предлагаемого метода. Предложенная реализация содержит ансамбль из алгоритмов REDQ и SAC. Выходом ансамбля управляет контролирующий алгоритм DQN. Реализация алгоритма выполнена при помощи фреймворка Pytorch RL и обладает высокой производительностью. Результаты исследований показывают, что реализованный алгоритм показывает эффективность выше, или в худшем случае, на уровне лучшего алгоритма в ансамбле.

В дальнейшем можно исследовать реализации предложенного алгоритма, отличающиеся от рассмотренного в этой работе: большей сложностью иерархии, бо́льшим количеством алгоритмов в ансамбле.

## Благодарности

Работа выполнена при поддержке Российского научного фонда (проект № 21-11-00321, https://rscf.ru/en/project/21-11-00321/).

## Литература

[1] Haarnoja, T. Soft Actor-Critic Algorithms and Applications / T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel, S. Levine — 2019. — DOI: 10.48550/arXiv.1812.05905.

[2] Mahmud, M. Applications of Deep Learning and Reinforcement Learning to Biological Data / M. Mahmud, M.S. Kaiser, A. Hussain, S. Vassanelli // IEEE Transactions on Neural Networks and Learning Systems. — 2018. — Vol. 29(6). — P. 2063–2079.

[3] Zheng, G. DRN: A Deep Reinforcement Learning Framework for News Recommendation / G. Zheng, F. Zhang, Z. Zheng, Y. Xiang, N.J. Yuan, X. Xie, Z. Li // Proceedings of the 2018 World Wide Web Conference: WWW '18. – 2018. – P. 167–176.

[4] Zhang, Z. Deep Reinforcement Learning for Trading / Z. Zhang, S. Zohren, S. Roberts // The Journal of Financial Data Science. – 2020. – Vol. 2(2). – P. 25–40.

[5] Chen, X. Randomized Ensembled Double Q-Learning: Learning Fast Without a Model / X. Chen, C. Wang, Z. Zhou, K. Ross // International Conference on Learning Representations. – 2021. DOI: 10.48550/arXiv.2101.05982.

[6] Haarnoja, T. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor / T. Haarnoja, A. Zhou, P. Abbeel, S. Levine // Proceedings of Machine Learning Research. – 2018. – Vol. 80. – P. 1856–1865. DOI: 10.48550/arXiv.1801.01290.

[7] Wang, Z. Dueling Network Architectures for Deep Reinforcement Learning / Z. Wang, T. Schaul, M. Hessel, H. van Hasselt, M. Lanctot, N. de Freitas // Proceedings of the 33rd International Conference on International Conference on Machine Learning. – 2016. – Vol. 48. – P. 1995–2003. DOI: 10.48550/arXiv.1511.06581.