Министерство науки и высшего образования Российской Федерации

Федеральное государственное автономное образовательное

учреждение высшего образования

«Московский физико-технический институт

(национальный исследовательский университет)»

УТВЕРЖДЕНО

Проректор по учебной работе

А.А. Воронов

«\_\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_ 2024 г.

# ПРОГРАММА

по дисциплине: **Обучение с подкреплением**

по направлению подготовки:

**03.03.01 «Прикладные математика и физика»**

физтех-школа: **ФПМИ**

кафедра: **математических основ управления**

курс: 4, 5

семестры: 8, 10

лекции – 30 часов Диф. зачет/экзамен – 8, 10 семестры

практические (семинарские)   
занятия – нет

лабораторные занятия – 15

ВСЕГО АУДИТОРНЫХ ЧАСОВ – 45 Самостоятельная работа

– 60 часов

# Программу составили:

Юдин Н.Е., Иванов С.М.

Программа принята на заседании

кафедры математических основ управления

12 января 2024 года

Заведующий кафедрой А.В. Гасников

1. Введение в курс. Постановка задачи обучения с подкреплением. Табличные методы обучения с подкреплением.

Кросс-энтропийный метод (CEM).

Динамическое программирование. Value Iteration, Policy Iteration.

Библиотека OpenAI gym. Реализация табличного кросс-энтропийного метода.

Метод зеркального спуска в обучении с подкреплением.

2. Элементы теории принятия решений и случайных процессов. Q-обучение и его вариации.

Марковский процесс принятия решений. Оптимизационная формализация. Двойственность Фенхеля-Рокафеллара.

Deep Q-Network (DQN) и его модификации.

Distributional RL. Categorical DQN (c51), Quantile Regression DQN (QR-DQN).

Анализ сложности алгоритма Q-обучения.

3. Policy gradient-подход с натуральным градиентом и схемы «актёр-критик».

Внутренняя мотивация для исследования среды.

Подход Advantage Actor-Critic (A2C).

Оценивание по методу REINFORCE.

Trust-Region Policy Optimization (TRPO).

Generalized Advantage Estimation (GAE). Proximal Policy Optimization (PPO).

Методы вида Natural Policy Gradient с энтропийной регуляризацией, их глобальная сходимость.

4. Задачи непрерывного управления и имитационное обучение.

Непрерывное управление.

Имитационное обучение. Обратное обучение с подкреплением.

Monte Carlo Tree Search. AlphaZero, MuZero.

Linear Quadratic Regulator (LQR). Model-based RL.

**Литература**

*Основная*

1. Botev Z. I. et al. The cross-entropy method for optimization //Handbook of statistics. – Elsevier, 2013. – Т. 31. – С. 35-59.

2. Szita I., Lörincz A. Learning Tetris using the noisy cross-entropy method //Neural computation. – 2006. – Т. 18. – №. 12. – С. 2936-2941.

3. Sutton R. S., Barto A. G. Reinforcement learning: An introduction. – MIT press, 2018.

4. Mnih V. et al. Playing atari with deep reinforcement learning //arXiv preprint arXiv:1312.5602. – 2013.

5. Van Hasselt H., Guez A., Silver D. Deep reinforcement learning with double q-learning //Proceedings of the AAAI conference on artificial intelligence. – 2016. – Т. 30. – №. 1.

6. Schaul T. et al. Prioritized experience replay //arXiv preprint arXiv:1511.05952. – 2015.

7. Fortunato M. et al. Noisy networks for exploration //arXiv preprint arXiv:1706.10295. – 2017.

8. Wang Z. et al. Dueling network architectures for deep reinforcement learning //International conference on machine learning. – PMLR, 2016. – С. 1995-2003.

9. Bellemare M. G., Dabney W., Munos R. A distributional perspective on reinforcement learning //International Conference on Machine Learning. – PMLR, 2017. – С. 449-458.

10. Dabney W. et al. Distributional reinforcement learning with quantile regression //Proceedings of the AAAI Conference on Artificial Intelligence. – 2018. – Т. 32. – №. 1.

11. Burda Y. et al. Exploration by random network distillation //arXiv preprint arXiv:1810.12894. – 2018.

12. Pathak D. et al. Curiosity-driven exploration by self-supervised prediction //International conference on machine learning. – PMLR, 2017. – С. 2778-2787.

13. Mnih V. et al. Asynchronous methods for deep reinforcement learning //International conference on machine learning. – PMLR, 2016. – С. 1928-1937.

14. Schulman J. et al. Trust region policy optimization //International conference on machine learning. – PMLR, 2015. – С. 1889-1897.

15. Schulman J. et al. High-dimensional continuous control using generalized advantage estimation //arXiv preprint arXiv:1506.02438. – 2015.

16. Schulman J. et al. Proximal policy optimization algorithms //arXiv preprint arXiv:1707.06347. – 2017.

17. Lillicrap T. P. et al. Continuous control with deep reinforcement learning //arXiv preprint arXiv:1509.02971. – 2015.

18. Fujimoto S., Hoof H., Meger D. Addressing function approximation error in actor-critic methods //International conference on machine learning. – PMLR, 2018. – С. 1587-1596.

19. Haarnoja T. et al. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor //International conference on machine learning. – PMLR, 2018. – С. 1861-1870.

20. Finn C., Levine S., Abbeel P. Guided cost learning: Deep inverse optimal control via policy optimization //International conference on machine learning. – PMLR, 2016. – С. 49-58.

21. Ho J., Ermon S. Generative adversarial imitation learning //Advances in neural information processing systems. – 2016. – Т. 29.

22. Silver D. et al. Mastering the game of go without human knowledge //nature. – 2017. – Т. 550. – №. 7676. – С. 354-359.

23. Schrittwieser J. et al. Mastering atari, go, chess and shogi by planning with a learned model //Nature. – 2020. – Т. 588. – №. 7839. – С. 604-609.

24. Paternain S. et al. Constrained reinforcement learning has zero duality gap //Advances in Neural Information Processing Systems. – 2019. – Т. 32.

25. Ying L., Zhu Y. A Note on Optimization Formulations of Markov Decision Processes //arXiv preprint arXiv:2012.09417. – 2020.

26. Tiapkin D., Gasnikov A. Parallel stochastic mirror descent for MDPs //arXiv preprint arXiv:2103.00299. – 2021.

27. Jin C. et al. Is Q-learning provably efficient? //Advances in neural information processing systems. – 2018. – Т. 31.

28. Cen S. et al. Fast global convergence of natural policy gradient methods with entropy regularization //Operations Research. – 2021.

*Дополнительная литература*

1. Salimans T. et al. Evolution strategies as a scalable alternative to reinforcement learning //arXiv preprint arXiv:1703.03864. – 2017.

2. Gaier A., Ha D. Weight agnostic neural networks //Advances in neural information processing systems. – 2019. – Т. 32.

3. Kapturowski S. et al. Recurrent experience replay in distributed reinforcement learning //International conference on learning representations. – 2018.

4. Badia A. P. et al. Agent57: Outperforming the atari human benchmark //International Conference on Machine Learning. – PMLR, 2020. – С. 507-517.

5. Dabney W. et al. Implicit quantile networks for distributional reinforcement learning //International conference on machine learning. – PMLR, 2018. – С. 1096-1105.

6. Hessel M. et al. Rainbow: Combining improvements in deep reinforcement learning //Thirty-second AAAI conference on artificial intelligence. – 2018.

7. Houthooft R. et al. Vime: Variational information maximizing exploration //Advances in neural information processing systems. – 2016. – Т. 29.

8. Badia A. P. et al. Never give up: Learning directed exploration strategies //arXiv preprint arXiv:2002.06038. – 2020.

9. Engstrom L. et al. Implementation matters in deep policy gradients: A case study on PPO and TRPO //arXiv preprint arXiv:2005.12729. – 2020.

10. Kuznetsov A. et al. Controlling overestimation bias with truncated mixture of continuous distributional quantile critics //International Conference on Machine Learning. – PMLR, 2020. – С. 5556-5566.

11. Levine S. Reinforcement learning and control as probabilistic inference: Tutorial and review //arXiv preprint arXiv:1805.00909. – 2018.

12. Ha D., Schmidhuber J. Recurrent world models facilitate policy evolution //Advances in neural information processing systems. – 2018. – Т. 31.

13. Hafner D. et al. Mastering atari with discrete world models //arXiv preprint arXiv:2010.02193. – 2020.

14. Nachum O., Dai B. Reinforcement learning via fenchel-rockafellar duality //arXiv preprint arXiv:2001.01866. – 2020.

15. Li T. et al. Faster Algorithm and Sharper Analysis for Constrained Markov Decision Process //arXiv preprint arXiv:2110.10351. – 2021.

16. Li H. et al. Accelerating Primal-dual Methods for Regularized Markov Decision Processes //arXiv preprint arXiv:2202.10506. – 2022.

17. Jin Y., Sidford A. Efficiently solving MDPs with stochastic mirror descent //International Conference on Machine Learning. – PMLR, 2020. – С. 4890-4900.

18. Jin Y., Sidford A. Towards tight bounds on the sample complexity of average-reward mdps //International Conference on Machine Learning. – PMLR, 2021. – С. 5055-5064.

*Рекомендуемые интернет-ресурсы*

http://www.machinelearning.ru – профессиональный информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных.

http://shad.yandex.ru – сайт школы анализа данных Яндекса.

https://openai.com/blog/

https://lilianweng.github.io/posts/

https://distill.pub/

https://hackernoon.com/

Подписано в печать 19.01.2024. Формат 60  84 1/16. Усл. печ. л. 0,4

Уч.-изд. л. 0,4. Тираж 70 экз. Заказ № 84.

Федеральное государственное автономное образовательное

учреждение высшего образования   
«Московский физико-технический институт

(национальный исследовательский университет)»

141700, Московская обл., г. Долгопрудный, Институтский пер., 9

Тел. (495) 408-58-22, e-mail: [rio@mipt.ru](mailto:rio@mipt.ru)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Отдел оперативной полиграфии «Физтех-полиграф»

141700, Московская обл., г. Долгопрудный, Институтский пер., 9

Тел. (495) 408-84-30, e-mail: [polygraph@mipt.ru](mailto:polygraph@mipt.ru)