Reinforcement Learning for Information Retrieval

Alexander Kuhnle
alexander.kuhnle@blueprism.com
Blue Prism AI Labs
London, UK

Miguel Aroca-Ouellette
miguel.aroca@blueprism.com
Blue Prism AI Labs
London, UK

Anindya Basu
anindya.basu@blueprism.com
Blue Prism AI Labs
London, UK

Murat Sensoy
murat.sensoy@blueprism.com
Blue Prism AI Labs
London, UK

John Reid
john.reid@blueprism.com
Blue Prism AI Labs
London, UK

Dell Zhang*
dell.z@ieee.org
Blue Prism AI Labs
London, UK

ABSTRACT

There is strong interest in leveraging *reinforcement learning* (RL) for *information retrieval* (IR) applications including search, recommendation, and advertising. Just in 2020, the term "reinforcement learning" was mentioned in more than 60 different papers published by ACM SIGIR. It has also been reported that Internet companies like Google and Alibaba have started to gain competitive advantages from their RL-based search and recommendation engines. This full-day tutorial gives IR researchers and practitioners who have no or little experience with RL the opportunity to learn about the fundamentals of modern RL in a practical *hands-on* setting. Furthermore, some representative applications of RL in IR systems will be introduced and discussed. By attending this tutorial, the participants will acquire a good knowledge of modern RL concepts and standard algorithms such as REINFORCE and DQN. This knowledge will help them better understand some of the latest IR publications involving RL, as well as prepare them to tackle their own practical IR problems using RL techniques and tools. Please refer to the tutorial website (https://rl-starterpack.github.io/) for more information.

CCS CONCEPTS

• **Computing methodologies** → **Reinforcement learning**; • **Information systems** → **Information retrieval**.

KEYWORDS

Markov decision process, Deep Q-Networks, policy gradient, actor-critic methods, search engines, recommender systems, computational advertising

ACM Reference Format:
Alexander Kuhnle, Miguel Aroca-Ouellette, Anindya Basu, Murat Sensoy, John Reid, and Dell Zhang. 2021. Reinforcement Learning for Information Retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3404835.3462813

*Dell Zhang is the corresponding author of this tutorial. He is currently on leave from Birkbeck, University of London, and works full-time for Blue Prism AI Labs.

SIGIR '21, July 11–15, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8037-9/21/07…$15.00
https://doi.org/10.1145/3404835.3462813

1 MOTIVATION

Reinforcement learning (RL) [37] is an area of machine learning which is concerned with optimal decision making over time in a dynamic environment.

Recent years have witnessed rapid development and great success of methods combining reinforcement learning with *deep neural networks* [8, 19], e.g., in AlphaGo [34]. Unsurprisingly, many *information retrieval* (IR) researchers and practitioners have become interested in applying reinforcement learning techniques to solve challenging decision-making problems in IR systems. Just in 2020, the term "reinforcement learning" was mentioned in more than 60 different papers published by ACM SIGIR[1].

The growing popularity of reinforcement learning in the field of IR is attributed to not only the technology push but also the demand pull. Because of the wide usage of web and mobile apps, modern IR systems for search, recommendation, and advertising [5] have become more *personalized* and *interactive*. In these scenarios, traditional IR approaches which assume user preferences being static and maximize immediate user satisfaction no longer work well. RL is a promising approach to tackling the problems of personalization and interactivity by capturing users' evolving interests and optimizing their long-term engagement [54]. It has also been reported that Internet companies like Google and Alibaba have started to gain competitive advantages from their RL-based search and recommendation engines [3].

Compared to the other two basic machine learning paradigms — supervised and unsupervised learning — reinforcement learning is an area which people in the IR community are relatively less familiar with. Therefore, it seems beneficial and timely to provide a tutorial on reinforcement learning for IR which explains the fundamentals of modern reinforcement learning and illustrates how these techniques can be utilized to address IR problems like learning to rank.

[1] https://scholar.google.co.uk/scholar?as_ylo=2020&q=%22%20reinforcement+learning%22+source:SIGIR

2 OBJECTIVES

We hope that this tutorial will equip the participants with a good knowledge of reinforcement learning which helps them understand the latest IR publications involving reinforcement learning and enables them to tackle their own IR problems in practice using reinforcement learning.

2.1 Intended Audience

The target audience of this tutorial are IR researchers and practitioners with no or little experience with RL, who would like to study RL in a practical *hands-on* setting and learn about its recent applications in IR systems. To really benefit from the tutorial, the participants should be familiar with basic IR[2] concepts and be comfortable with Python[3] programming using Jupyter[4] notebooks.

2.2 Outline of the Topics

This tutorial consists of two main parts.

- In the first part, we will introduce the most important RL concepts and algorithms, including Exploitation vs Exploration, Markov Decision Processes (MDP) [37], Multi-Armed Bandit (MAB) [7, 18, 36], Q-Learning [40, 41], Deep Q-Network (DQN) [11, 24], Policy Gradient (such as REINFORCE [44]), and Actor-Critic [17] methods (such as DPG [35]).
- In the second part, we will illustrate a few representative applications of RL in IR, e.g., learning to rank [43, 45, 48, 51, 52], relevance feedback [25], query reformulation [29], IRGAN [38], recommender systems [3, 4, 22, 47, 55, 57], and computational advertising [56].

3 RELEVANCE TO THE COMMUNITY

There have been several tutorials on "machine learning for information retrieval" in CIKM-2008, SIGIR-2008 and SIGIR-2011 [33], and also a tutorial about "deep learning for information retrieval'' in SIGIR-2016 [21]. For the applications of (multi-armed) bandit — a simplified form of reinforcement learning — in interactive IR and recommender systems, two tutorials have been given in ICTIR-2017 [6] and RecSys-2020 [1] respectively. In the related research field of natural language processing (NLP), a tutorial entitled "deep reinforcement learning for natural language processing" has been provided in ACL-2018 [39]. However, to the best of our knowledge, so far there has not been any tutorial on major international IR conferences dedicated to "reinforcement learning for information retrieval", though a workshop on "deep reinforcement learning for information retrieval" was recently held at SIGIR-2020 [53] and will run again this year[5]. Furthermore, this tutorial distinguishes itself from the above related tutorials by its emphasis on the combination of theoretical concepts with practical *hands-on* exercises.

3.1 Previous Offerings

This tutorial is developed from an internal seminar series at Blue Prism AI Labs. It has been presented as a full-day tutorial at the BCS-IRSG Search Solutions (SS) conference[6] on November 24, 2020, and later at the 43rd European Conference on Information Retrieval (ECIR)[7] [12, 13] on March 28, 2021.

Compared to its previous versions, this tutorial for SIGIR-2021 will include deeper theoretical/mathematical materials, add an introduction to bandit algorithms, expand the section about actor-critic methods, cover several more recent IR papers using RL (e.g., [4, 15, 22, 47]), and discuss how the very latest advances in RL such as *offline reinforcement learning* [16, 20, 46, 49, 59] are going to affect the field of IR.

4 FORMAT AND SCHEDULE

This full-day tutorial **RL4IR** will be delivered *online* as part of the ACM SIGIR-2021 virtual conference that will take place on July 11–15, 2021.

The plan is that the tutorial will comprise *pre-recorded videos* (of about 6 hours in total) and two *live sessions* (each of up to 1.5 hours); those live sessions will be devoted to hands-on practices as well as Q&A using a video conferencing software (e.g., Zoom) and they will be spread apart to cater for audiences across the globe.

The *tentative* list of pre-recorded videos is as follows.

- RL Basics (MDP etc.) [37]
- Bandits [7, 18, 36, 37]
- Tabular Q-Learning [37, 40, 41]
- Deep Q Network (DQN) [11, 24]
- IR applications using DQN [55–57]
- Policy Gradient (REINFORCE) [37, 44]
- IR applications using REINFORCE [3, 22, 25, 29, 38, 43, 45, 48, 51, 52]
- Actor-Critic [17, 35]
- IR applications using Actor-Critic [4, 47]
- Recent developments & outlook for research [2, 9, 10, 14, 16, 20, 23, 26–28, 30–32, 42, 46, 49, 50, 58, 59]

5 SUPPORT MATERIALS

The tutorial consists of a mix of presentations and short practical sessions with exercises or examples to experiment with. The following materials will be provided to tutorial attendees.

- Presentation slides (pdf).
- Examples/exercises as Jupyter notebooks (on Google Colab[8]).
- Easy-to-modify implementations of basic RL algorithms to get started.
- A curated webpage/repository on GitHub[9] with resources and references, including all the above materials required for this tutorial.

6 PRESENTER BIOGRAPHIES

Alexander Kuhnle is a Research Engineer at Blue Prism AI Labs, working on computer vision and learning from demonstrations. He recently finished his PhD at the University of Cambridge on visually grounded language understanding and machine learning

[2] https://nlp.stanford.edu/IR-book/information-retrieval-book.html
[3] https://www.python.org/
[4] https://jupyter.org/
[5] https://drl4ir.github.io/
[6] https://irsg.bcs.org/SearchSolutions/2020/sse2020.php
[7] https://www.ecir2021.eu/
[8] https://colab.research.google.com/
[9] https://rl-starterpack.github.io/

evaluation methodology. Over the last three years, he has also been the main developer and maintainer of Tensorforce[10], an open-source framework for applied deep reinforcement learning.

Miguel Aroca-Ouellette is a Research Engineer at Blue Prism AI Labs. He received his M.S. from the California Institute of Technology and worked at Google before moving to London in 2020. He has worked with machine learning models both in research and in bringing them to large-scale production capacity. His current work is on using kernel methods and graphical models for Intelligent Document Processing.

Anindya Basu is a Research Engineer at Blue Prism AI Labs. He worked on mobile applications for Samsung Electronics before moving to London for an MSc from University College London (UCL). Since then, he has worked on machine learning applications in recommender systems as well as robotic vision and actuation. He is currently working on deep learning models and bandit algorithms for visual understanding in Robotic Process Automation.

Murat Sensoy is a Senior Research Scientist at Blue Prism AI Labs. He was previously a Visiting Scholar at UCL, an Associate Professor at Ozyegin University, and a Postdoctoral Research Fellow at the University of Aberdeen. He received his PhD degree in Computer Engineering at Bogazici University in 2008. He developed semantic reasoning mechanisms for sensor networks, which are used by US Army Research Lab and IBM Research.

John Reid is a Staff Research Scientist at Blue Prism AI Labs and oversees the Intelligent Document Processing team there. He received his PhD in Bayesian statistics from the University of Cambridge and worked applying ML and Bayesian statistics to problems in computational biology at the University's Biostatistics Unit. He has taught on the Cambridge Computational Biology MPhil. He has also been a Turing Fellow at the Alan Turing Institute and a Research Fellow at UCL. He was the author of (probably) the first game-playing RL-agent to play autonomously online.

Dell Zhang[11] is a Reader in Computer Science at Birkbeck College, University of London (on leave) and a Staff Research Scientist at Blue Prism AI Labs. He is a Senior Member of ACM, a Senior Member of IEEE, and a Fellow of RSS. He got his PhD from the Southeast University (SEU) in Nanjing, China, and then worked as a Research Fellow at the Singapore-MIT Alliance (SMA) until he moved to the UK in 2005. His main research interests include Machine Learning, Information Retrieval, and Natural Language Processing. He has published 100+ papers, received multiple best paper awards, and won several prizes from international data science competitions. He has been giving lectures to both undergraduate and postgraduate students in Birkbeck and UCL.

ACKNOWLEDGMENTS

We thank the tutorials chairs and anonymous reviewers for their constructive comments on the tutorial proposal. We are also grateful to the audience of the previous versions of this tutorial for their valuable feedback which has been helpful in improving the tutorial.

REFERENCES

[1] Andrea Barraza-Urbina and Dorota Glowacka. 2020. Introduction to Bandits in Recommender Systems. In *Fourteenth ACM Conference on Recommender Systems (RecSys '20)*. Association for Computing Machinery, New York, NY, USA, 748–750. https://doi.org/10.1145/3383313.3411547

[2] Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. 2018. Exploration by Random Network Distillation. *arXiv:1810.12894 [cs, stat]* (Oct. 2018). arXiv:1810.12894 [cs, stat]

[3] Minmin Chen, Alex Beutel, Paul Covington, Sagar Jain, Francois Belletti, and Ed H. Chi. 2019. Top-K Off-Policy Correction for a REINFORCE Recommender System. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (WSDM '19)*. Association for Computing Machinery, New York, NY, USA, 456–464. https://doi.org/10.1145/3289600.3290999

[4] Xu Chen, Yali Du, Long Xia, and Jun Wang. 2021. Reinforcement Recommendation with User Multi-Aspect Preference. In *Proceedings of The Web Conference 2021 (WWW '21)*. Association for Computing Machinery, Virtual Event.

[5] Hector Garcia-Molina, Georgia Koutrika, and Aditya Parameswaran. 2011. Information Seeking: Convergence of Search, Recommendations, and Advertising. *Commun. ACM* 54, 11 (Nov. 2011), 121–130. https://doi.org/10.1145/2018396.2018423

[6] Dorota Glowacka. 2017. Bandit Algorithms in Interactive Information Retrieval. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR '17)*. Association for Computing Machinery, New York, NY, USA, 327–328. https://doi.org/10.1145/3121050.3121108

[7] Dorota Glowacka. 2019. Bandit Algorithms in Information Retrieval. *Foundations and Trends in Information Retrieval* 13, 4 (May 2019), 299–424. https://doi.org/10.1561/1500000067

[8] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press.

[9] David Ha and Jürgen Schmidhuber. 2018. World Models. *arXiv:1803.10122 [cs, stat]* (March 2018). https://doi.org/10.5281/zenodo.1207631 arXiv:1803.10122 [cs, stat]

[10] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. 2020. Dream to Control: Learning Behaviors by Latent Imagination. *arXiv:1912.01603 [cs]* (March 2020). arXiv:1912.01603 [cs]

[11] Matteo Hessel, Joseph Modayil, Hado Van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver. 2018. Rainbow: Combining Improvements in Deep Reinforcement Learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[12] Djoerd Hiemstra, Marie-Francine Moens, Josiane Mothe, Raffaele Perego, Martin Potthast, and Fabrizio Sebastiani (Eds.). 2021. *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 – April 1, 2021, Proceedings, Part I*. Lecture Notes in Computer Science, Vol. 12656. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-030-72113-8

[13] Djoerd Hiemstra, Marie-Francine Moens, Josiane Mothe, Raffaele Perego, Martin Potthast, and Fabrizio Sebastiani (Eds.). 2021. *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 – April 1, 2021, Proceedings, Part II*. Lecture Notes in Computer Science, Vol. 12657. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-030-72240-1

[14] Jonathan Ho and Stefano Ermon. 2016. Generative Adversarial Imitation Learning. *Advances in Neural Information Processing Systems* 29 (2016).

[15] Eugene Ie, Vihan Jain, Jing Wang, Sanmit Narvekar, Ritesh Agarwal, Rui Wu, Heng-Tze Cheng, Morgane Lustman, Vince Gatto, Paul Covington, Jim McFadden, Tushar Chandra, and Craig Boutilier. 2019. Reinforcement Learning for Slate-Based Recommender Systems: A Tractable Decomposition and Practical Methodology. *arXiv:1905.12767 [cs, stat]* (May 2019). arXiv:1905.12767 [cs, stat]

[16] Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. 2020. MOReL: Model-Based Offline Reinforcement Learning. *Advances in Neural Information Processing Systems* 33 (2020), 21810–21823.

[17] Vijay Konda and John Tsitsiklis. 1999. Actor-Critic Algorithms. *Advances in Neural Information Processing Systems* 12 (1999).

[18] Tor Lattimore and Csaba Szepesvári. 2020. *Bandit Algorithms*. Cambridge University Press.

[19] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep Learning. *Nature* 521, 7553 (2015), 436–444.

[20] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. 2020. Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems. *arXiv:2005.01643 [cs, stat]* (Nov. 2020). arXiv:2005.01643 [cs, stat]

[21] Hang Li and Zhengdong Lu. 2016. Deep Learning for Information Retrieval. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '16)*. Association for Computing Machinery, New York, NY, USA, 1203–1206. https://doi.org/10.1145/2911451.2914800

[22] Jiaqi Ma, Zhe Zhao, Xinyang Yi, Ji Yang, Minmin Chen, Jiaxi Tang, Lichan Hong, and Ed H. Chi. 2020. Off-Policy Learning in Two-Stage Recommender Systems. In *Proceedings of The Web Conference 2020 (WWW '20)*. Association for Computing Machinery, New York, NY, USA, 463–473. https://doi.org/10.1145/3366423.3380130

[10] https://github.com/tensorforce/tensorforce

[11] https://www.dcs.bbk.ac.uk/~dell/

[23] Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. 2016. Asynchronous Methods for Deep Reinforcement Learning. *arXiv:1602.01783 [cs]* (Feb. 2016). arXiv:1602.01783 [cs]

[24] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, and Georg Ostrovski. 2015. Human-Level Control through Deep Reinforcement Learning. *Nature* 518, 7540 (2015), 529.

[25] Ali Montazeralghaem, Hamed Zamani, and James Allan. 2020. A Reinforcement Learning Framework for Relevance Feedback. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 59–68. https://doi.org/10.1145/3397271.3401099

[26] Arun Nair, Praveen Srinivasan, Sam Blackwell, Cagdas Alcicek, Rory Fearon, Alessandro De Maria, Vedavyas Panneershelvam, Mustafa Suleyman, Charles Beattie, Stig Petersen, Shane Legg, Volodymyr Mnih, Koray Kavukcuoglu, and David Silver. 2015. Massively Parallel Methods for Deep Reinforcement Learning. *arXiv:1507.04296 [cs]* (July 2015). arXiv:1507.04296 [cs]

[27] Soroush Nasiriany, Vitchyr Pong, Steven Lin, and Sergey Levine. 2019. Planning with Goal-Conditioned Policies. *Advances in Neural Information Processing Systems* 32 (2019).

[28] Andrew Y. Ng and Stuart J. Russell. 2000. Algorithms for Inverse Reinforcement Learning.. In *ICML*, Vol. 1. 2.

[29] Rodrigo Nogueira and Kyunghyun Cho. 2017. Task-Oriented Query Reformulation with Reinforcement Learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 574–583. https://doi.org/10.18653/v1/D17-1061

[30] Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. 2017. Curiosity-Driven Exploration by Self-Supervised Prediction. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70 (ICML'17)*. JMLR.org, Sydney, NSW, Australia, 2778–2787.

[31] Alexander Pritzel, Benigno Uria, Sriram Srinivasan, Adrià Puigdomènech, Oriol Vinyals, Demis Hassabis, Daan Wierstra, and Charles Blundell. 2017. Neural Episodic Control. *arXiv:1703.01988 [cs, stat]* (March 2017). arXiv:1703.01988 [cs, stat]

[32] Adam Santoro, Ryan Faulkner, David Raposo, Jack Rae, Mike Chrzanowski, Theophane Weber, Daan Wierstra, Oriol Vinyals, Razvan Pascanu, and Timothy Lillicrap. 2018. Relational Recurrent Neural Networks. *arXiv:1806.01822 [cs, stat]* (June 2018). arXiv:1806.01822 [cs, stat]

[33] Luo Si and Rong Jin. 2011. Machine Learning for Information Retrieval. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '11)*. Association for Computing Machinery, New York, NY, USA, 1293–1294. https://doi.org/10.1145/2009916.2010167

[34] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, and Marc Lanctot. 2016. Mastering the Game of Go with Deep Neural Networks and Tree Search. *Nature* 529, 7587 (2016), 484–489.

[35] David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. 2014. Deterministic Policy Gradient Algorithms. In *International Conference on Machine Learning*. PMLR, 387–395.

[36] Aleksandrs Slivkins. 2019. Introduction to Multi-Armed Bandits. *Foundations and Trends in Machine Learning* 12, 1-2 (Nov. 2019), 1–286. https://doi.org/10.1561/2200000068

[37] Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction* (second ed.). MIT Press.

[38] Jun Wang, Lantao Yu, Weinan Zhang, Yu Gong, Yinghui Xu, Benyou Wang, Peng Zhang, and Dell Zhang. 2017. IRGAN: A Minimax Game for Unifying Generative and Discriminative Information Retrieval Models. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17)*. Association for Computing Machinery, New York, NY, USA, 515–524. https://doi.org/10.1145/3077136.3080786

[39] William Yang Wang, Jiwei Li, and Xiaodong He. 2018. Deep Reinforcement Learning for NLP. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*. Association for Computational Linguistics, Melbourne, Australia, 19–21. https://doi.org/10.18653/v1/P18-5007

[40] Christopher JCH Watkins and Peter Dayan. 1992. Q-Learning. *Machine learning* 8, 3-4 (1992), 279–292.

[41] Christopher John Cornish Hellaby Watkins. 1989. Learning from Delayed Rewards. (1989).

[42] Greg Wayne, Chia-Chun Hung, David Amos, Mehdi Mirza, Arun Ahuja, Agnieszka Grabska-Barwinska, Jack Rae, Piotr Mirowski, Joel Z. Leibo, Adam Santoro, Mevlana Gemici, Malcolm Reynolds, Tim Harley, Josh Abramson, Shakir Mohamed, Danilo Rezende, David Saxton, Adam Cain, Chloe Hillier, David Silver, Koray Kavukcuoglu, Matt Botvinick, Demis Hassabis, and Timothy Lillicrap. 2018. Unsupervised Predictive Memory in a Goal-Directed Agent. *arXiv:1803.10760 [cs, stat]* (March 2018). arXiv:1803.10760 [cs, stat]

[43] Zeng Wei, Jun Xu, Yanyan Lan, Jiafeng Guo, and Xueqi Cheng. 2017. Reinforcement Learning to Rank with Markov Decision Process. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17)*. Association for Computing Machinery, New York, NY, USA, 945–948. https://doi.org/10.1145/3077136.3080685

[44] Ronald J. Williams. 1992. Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning. *Machine Learning* 8, 3-4 (1992), 229–256.

[45] Long Xia, Jun Xu, Yanyan Lan, Jiafeng Guo, Wei Zeng, and Xueqi Cheng. 2017. Adapting Markov Decision Process for Search Result Diversification. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17)*. Association for Computing Machinery, New York, NY, USA, 535–544. https://doi.org/10.1145/3077136.3080775

[46] Teng Xiao and Donglin Wang. 2021. A General Offline Reinforcement Learning Framework for Interactive Recommendation. In *The Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI*.

[47] Xin Xin, Alexandros Karatzoglou, Ioannis Arapakis, and Joemon M. Jose. 2020. Self-Supervised Reinforcement Learning for Recommender Systems. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 931–940. https://doi.org/10.1145/3397271.3401147

[48] Jun Xu, Zeng Wei, Long Xia, Yanyan Lan, Dawei Yin, Xueqi Cheng, and Ji-Rong Wen. 2020. Reinforcement Learning to Rank with Pairwise Policy Gradient. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 509–518. https://doi.org/10.1145/3397271.3401148

[49] Mengjiao Yang and Ofir Nachum. 2021. Representation Matters: Offline Pretraining for Sequential Decision Making. *arXiv:2102.05815 [cs]* (Feb. 2021). arXiv:2102.05815 [cs]

[50] Yaodong Yang and Jun Wang. 2021. An Overview of Multi-Agent Reinforcement Learning from Game Theoretical Perspective. *arXiv:2011.00583 [cs]* (March 2021). arXiv:2011.00583 [cs]

[51] Wei Zeng, Jun Xu, Yanyan Lan, Jiafeng Guo, and Xueqi Cheng. 2018. Multi Page Search with Reinforcement Learning to Rank. In *Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR '18)*. Association for Computing Machinery, New York, NY, USA, 175–178. https://doi.org/10.1145/3234944.3234977

[52] Wei Zeng, Weiijie Yu, Jun Xu, Lan, Yanyan, and Cheng, Xueqi. 2020. Imitation Learning to Rank. *Journal of Chinese Information Processing* 34, 1 (2020), 97–105.

[53] Weinan Zhang, Xiangyu Zhao, Li Zhao, Dawei Yin, Grace Hui Yang, and Alex Beutel. 2020. Deep Reinforcement Learning for Information Retrieval: Fundamentals and Advances. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 2468–2471. https://doi.org/10.1145/3397271.3401467

[54] Xiangyu Zhao, Changsheng Gu, Haoshenglun Zhang, Xiaobing Liu, Xiwang Yang, and Jiliang Tang. 2019. Deep Reinforcement Learning for Online Advertising in Recommender Systems. *arXiv:1909.03602 [cs]* (Sept. 2019). arXiv:1909.03602 [cs]

[55] Xiangyu Zhao, Liang Zhang, Zhuoye Ding, Long Xia, Jiliang Tang, and Dawei Yin. 2018. Recommendations with Negative Feedback via Pairwise Deep Reinforcement Learning. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18)*. Association for Computing Machinery, New York, NY, USA, 1040–1048. https://doi.org/10.1145/3219819.3219886

[56] Xiangyu Zhao, Xudong Zheng, Xiwang Yang, Xiaobing Liu, and Jiliang Tang. 2020. Jointly Learning to Recommend and Advertise. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '20)*. Association for Computing Machinery, New York, NY, USA, 3319–3327. https://doi.org/10.1145/3394486.3403384

[57] Guanjie Zheng, Fuzheng Zhang, Zihan Zheng, Yang Xiang, Nicholas Jing Yuan, Xing Xie, and Zhenhui Li. 2018. DRN: A Deep Reinforcement Learning Framework for News Recommendation. In *Proceedings of the 2018 World Wide Web Conference (WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 167–176. https://doi.org/10.1145/3178876.3185994

[58] Ming Zhou, Yong Chen, Ying Wen, Yaodong Yang, Yufeng Su, Weinan Zhang, Dell Zhang, and Jun Wang. 2019. Factorized Q-Learning for Large-Scale Multi-Agent Systems. In *Proceedings of the First International Conference on Distributed Artificial Intelligence (DAI '19)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3356464.3357707

[59] Lixin Zou, Long Xia, Pan Du, Zhuo Zhang, Ting Bai, Weidong Liu, Jian-Yun Nie, and Dawei Yin. 2020. Pseudo Dyna-Q: A Reinforcement Learning Framework for Interactive Recommendation. In *Proceedings of the 13th International Conference on Web Search and Data Mining (WSDM '20)*. Association for Computing Machinery, New York, NY, USA, 816–824. https://doi.org/10.1145/3336191.3371801