17 Wang HR, Zariphopoulou T, Zhou XY. Exploration versus exploitation in reinforcement learning: A stochastic control approach. arXiv preprint arXiv: 1812.01552, 2018.
18 Auer P. Using confidence bounds for exploitation-exploration trade-offs. The Journal of Machine Learning Research, 2003, 3(3): 397–422.
19 Bellman RE. Dynamic Programming. Princeton: Princeton University Press, 1957.
20 Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science, 2006, 313(5786): 504–507. [doi: 10.1126/science.1127647]
21 马骋乾, 谢伟, 孙伟杰. 强化学习研究综述. 指挥控制与仿真, 2018, 40(6): 68–72. [doi: 10.3969/j.issn.1673-3819.2018.06.015]
22 Watkins CJCH, Dayan P. *Q*-learning. Machine Learning, 1992, 8(3–4): 279–292.
23 Rummery GA, Niranjan M. On-line Q-learning Using Connectionist Systems. Cambridge: University of Cambridge, 1994.
24 Mnih V, Kavukcuoglu K, Silver D, *et al*. Playing atari with deep reinforcement learning. arXiv preprint arXiv: 1312.5602, 2013.
25 Mnih V, Kavukcuoglu K, Silver D, *et al*. Human-level control through deep reinforcement learning. Nature, 2015, 518(7540): 529–533. [doi: 10.1038/nature14236]
26 Van Hasselt H, Guez A, Silver D. Deep reinforcement learning with double Q-learning. Proceedings of the 30th AAAI Conference on Artificial Intelligence. Phoenix, AZ, USA. 2016. 2094–2100.
27 Schaul T, Quan J, Antonoglou I, *et al*. Prioritized experience replay. Proceedings of the 2016 International Conference on Learning Representations. San Juan, UT, USA. 2016. 1–21.
28 Wang ZY, Schaul T, Hessel M, *et al*. Dueling network architectures for deep reinforcement learning. Proceedings of the 33rd International Conference on International Conference on Machine Learning. New York, NY, USA. 2016. 1995–2003.
29 Fortunato M, Azar MG, Piot B, *et al*. Noisy networks for exploration. arXiv preprint arXiv: 1706.10295, 2017.
30 Bellemare MG, Dabney W, Munos R. A distributional perspective on reinforcement learning. Proceedings of the 34th International Conference on Machine Learning. Sydney, NSW, Australia. 2017. 449–458.
31 Hessel M, Modayil J, Van Hasselt H, *et al*. Rainbow: Combining improvements in deep reinforcement learning. Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, LA, USA. 2018. 3215–3222.
32 Williams RJ. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine Learning, 1992, 8(3–4): 229–256. [doi: 10.1007/BF00992696]
33 Schulman J, Levine S, Moritz P, *et al*. Trust region policy optimization. Proceedings of the 31st International Conference on Machine Learning. Lille, France. 2015. 1889–1897.
34 Schulman J, Wolski F, Dhariwal P, *et al*. Proximal policy optimization algorithms. arXiv preprint arXiv: 1707.06347, 2017.
35 Konda VR, Tsitsiklis JN. On actor-critic algorithms. SIAM Journal on Control and Optimization, 2003, 42(4): 1143–1166.
36 Mnih V, Badia AP, Mirza M, *et al*. Asynchronous methods for deep reinforcement learning. Proceedings of the 33rd International Conference on International Conference on Machine Learning. New York, NY, USA. 2016. 1928–1937.
37 Silver D, Lever G, Heess N, *et al*. Deterministic policy gradient algorithms. Proceedings of the 31st International Conference on International Conference on Machine Learning. Beijing, China. 2014. 387–395.
38 Lillicrap TP, Hunt JJ, Pritzel A, *et al*. Continuous control with deep reinforcement learning. 4th International Conference on Learning Representations. San Juan, UT, USA. 2016.
39 Fujimoto S, Van Hoof H, Meger D. Addressing function approximation error in actor-critic methods. Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden. 2018. 1587–1596.
40 Haarnoja T, Zhou A, Abbeel P, *et al*. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. Proceedings of the 35th International Conference on Machine Learning. Long Beach, CA, USA. 2018. 1861–1870.
41 Arulkumaran K, Deisenroth MP, Brundage M, *et al*. Deep reinforcement learning: A brief survey. IEEE Signal Processing Magazine, 2017, 34(6): 26–38. [doi: 10.1109/MSP.2017.2743240]
42 Gomez F, Schmidhuber J. Evolving modular fast-weight networks for control. Proceedings of the 15th International Conference on Artificial Neural Networks: Formal Models and their Applications. Berlin, Germany. 2005. 383–389.
43 Koutník J, Cuccu G, Schmidhuber J, *et al*. Evolving large-

scale neural networks for vision-based reinforcement learning. Proceedings of the 15th Annual Conference on Genetic and Evolutionary Computation. Amsterdam, the Netherland. 2013. 1061–1068.

44 Kakade S. A natural policy gradient. Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic. Cambridge, UK. 2001. 1531–1538.

45 Heess N, Dhruva TB, Sriram S, *et al*. Emergence of locomotion behaviours in rich environments. arXiv preprint arXiv: 1707.02286, 2017

46 Duan Y, Schulman J, Chen X, *et al*. RL^2: Fast reinforcement learning via slow reinforcement learning. arXiv preprint arXiv: 1611.02779, 2016.

47 Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks. Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia. 2017. 1126–1135.

48 Gupta A, Mendonca R, Liu YX, *et al*. Meta-reinforcement learning of structured exploration strategies. Proceedings of the 32nd International Conference on Neural Information Processing Systems. Red Hook, NY, USA. 2018. 5307–5316.

49 Mendonca R, Gupta A, Kralev R, *et al*. Guided meta-policy search. Advances in Neural Information Processing Systems 32. Vancouver, Canada. 2019.

50 OpenAI: Benchmarks for spinning up implementations. https://spinningup.openai.com/en/latest/spinningup/bench.html#hopper-pytorch-versions. 2018.

51 Foerster JN. Deep multi-agent reinforcement learning [Ph.D Thesis]. Oxford: University of Oxford, 2018.

52 Chen X, Deng XT. Settling the complexity of two-player nash equilibrium. 2006 47th Annual IEEE Symposium on Foundations of Computer Science. Berkeley, CA, USA. 2006. 261–272.

53 Busoniu L, Babuska R, De Schutter B. A comprehensive survey of multiagent reinforcement learning. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2008, 38(2): 156–172.

54 Foerster JN, Assael YM, De Freitas N, *et al*. Learning to communicate with deep multi-agent reinforcement learning. Proceedings of the 30th International Conference on Neural Information Processing Systems. Red Hook, NY, USA. 2016. 2145–2153.

55 杜威, 丁世飞. 多智能体强化学习综述. 计算机科学, 2019, 46(8): 1–8. [doi: 10.11896/j.issn.1002-137X.2019.08.001]

56 Littman ML. Markov games as a framework for multi-agent reinforcement learning. In: Cohen WW, Hirsh H, eds. Machine Learning Proceedings 1994. Amsterdam: Elsevier, 1994. 157–163. [doi: 10.1016/B978-1-55860-335-6.50027-1]

57 Hu JL, Wellman MP. Nash Q-learning for general-sum stochastic games. Journal of Machine Learning Research, 2004, 4(6): 1039–1069. [doi: 10.1162/1532443041827880]

58 Littman ML. Friend-or-foe Q-learning in general-sum games. Proceedings of the Eighteenth International Conference on Machine Learning. San Francisco, CA, USA. 2001. 322–328.

59 Tan M. Multi-agent reinforcement learning: Independent vs. Cooperative agents. Proceedings of the Tenth International Conference. Amherst, MA, USA. 1993. 330–337. [doi: 10.1016/B978-1-55860-307-3.50049-6]

60 Lowe R, Wu Y, Tamar A, *et al*. Multi-agent actor-critic for mixed cooperative-competitive environments. Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA. 2017. 6382–6393.

61 Foerster JN, Farquhar G, Afouras T, *et al*. Counterfactual multi-agent policy gradients. The 32th AAAI Conference on Artificial Intelligence. New Orleans, LA, USA. 2018. 2974–2982.

62 Sunehag P, Lever G, Gruslys A, *et al*. Value-decomposition networks for cooperative multi-agent learning based on team reward. The 17th International Conference on Autonomous Agents and Multiagent Systems. Stockholm, Sweden. 2018. 2085–2087.

63 Rashid T, Samvelyan M, De Witt CS, *et al*. QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning. Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden. 2018. 4295–4304.

64 Yang YD, Luo R, Li MN, *et al*. Mean field multi-agent reinforcement learning. arXiv preprint arXiv: 1802.05438, 2018.

65 Finn C. Learning to learn with gradients [Ph.D Thesis]. Berkeley: University of California, 2018.

66 Wang JX, Kurth-Nelson Z, Tirumala D, *et al*. Learning to reinforcement learn. arXiv preprint arXiv: 1611.05763, 2016.

67 Vuorio R, Sun SH, Hu HX, *et al*. Multimodal model-agnostic meta-learning via task-aware modulation. Advances in Neural Information Processing Systems (NIPS

2019). Vancouver, BC, Canada. 2019. 1–12.
68 Nagabandi A, Clavera I, Liu SM, *et al*. Learning to adapt in dynamic, real-world environments through meta-reinforcement learning. arXiv preprint arXiv: 1803.11347v6, 2019.
69 Duan Y, Andrychowicz M, Stadie BC, *et al*. One-shot imitation learning. Proceedings of the 31st Conference on Neural Information Processing Systems. Long Beach, CA, USA. 2017. 1087–1098.
70 Finn C, Yu TH, Zhang TH, *et al*. One-shot visual imitation learning via meta-learning. arXiv preprint arXiv: 1709.04905, 2017.
71 Yu TH, Finn C, Dasari S, *et al*. One-shot imitation from observing humans via domain-adaptive meta-learning. 6th International Conference on Learning Representations. Vancouver, BC, Canada. 2018. [doi: 10.15607/RSS.2018.XIV.002]
72 Xu K, Ratner E, Dragan A, *et al*. Learning a prior over intent via meta-inverse reinforcement learning. arXiv preprint arXiv: 1805.12573, 2018.
73 Xie AN, Singh A, Levine S, *et al*. Few-shot goal inference for visuomotor learning and planning. Proceedings of the 2nd Conference on Robot Learning. Zurich, Switzerland. 2018. 40–52.
74 Rakelly K, Zhou A, Quillen D, *et al*. Efficient off-policy meta-reinforcement learning via probabilistic context variables. arXiv preprint arXiv: 1903.08254v1, 2019.
75 Silver D, Schrittwieser J, Simonyan K, *et al*. Mastering the game of go without human knowledge. Nature, 2017, 550(7676): 354–359. [doi: 10.1038/nature24270]
76 Silver D, Hubert T, Schrittwieser J, *et al*. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. Science, 2018, 362(6419): 1140–1144. [doi: 10.1126/science.aar6404]
77 Browne CB, Powley E, Whitehouse D, *et al*. A survey of Monte Carlo tree search methods. IEEE Transactions on Computational Intelligence and AI in Games, 2012, 4(1): 1–43. [doi: 10.1109/tciaig.2012.2186810]
78 Neller TW, Hnath S. Approximating optimal dudo play with fixed-strategy iteration counterfactual regret minimization. Proceedings of the 13th International Conference on Advances in Computer Games. Berlin, Germany. 2012. 170–183. [doi: 10.1007/978-3-642-31866-5_15]
79 OpenAI Five 2016–2019. https://openai.com/projects/five/.
80 Mülling K, Kober J, Kroemer O, *et al*. Learning to select and generalize striking movements in robot table tennis. The International Journal of Robotics Research, 2013, 32(3): 263–279. [doi: 10.1177/0278364912472380]
81 谢天瑞. 基于强化学习的 D2D 智能组网 [硕士学位论文]. 北京: 北京邮电大学, 2018.
82 Mirowski P, Pascanu R, Viola F, *et al*. Learning to navigate in complex environments. arXiv preprint arXiv: 1611.03673, 2017.
83 Banino A, Barry C, Uria B, *et al*. Vector-based navigation using grid-like representations in artificial agents. Nature, 2018, 557(7705): 429–433. [doi: 10.1038/s41586-018-0102-6]
84 Liu CL, Tomizuka M. Algorithmic safety measures for intelligent industrial co-robots. 2016 IEEE International Conference on Robotics and Automation. Stockholm, Sweden. 2016. 3095–3102.
85 Liu CL, Tomizuka M. Designing the robot behavior for safe human-robot interactions. In: Wang Y, Zhang YM, eds. Trends in Control and Decision-Making for Human-Robot Collaboration Systems. Cham: Springer, 2017. 241 –270. [doi: 10.1007/978-3-319-40533-9_11]
86 Bazzan ALC, Klügl F. Introduction to intelligent systems in traffic and transportation. Synthesis Lectures on Artificial Intelligence and Machine Learning, 2014, 7(3): 1–137.
87 杨文臣, 张轮, Zhu F. 多智能体强化学习在城市交通网络信号控制方法中的应用综述. 计算机应用研究, 2018, 35(6): 1613–1618. [doi: 10.3969/j.issn.1001-3695.2018.06.003]
88 Chu TS, Wang J, Codecà L, *et al*. Multi-agent deep reinforcement learning for large-scale traffic signal control. IEEE Transactions on Intelligent Transportation Systems, 2020, 21(3): 1086–1095.
89 Belletti F, Haziza D, Gomes G, *et al*. Expert level control of ramp metering based on multi-task deep reinforcement learning. IEEE Transactions on Intelligent Transportation Systems, 2018, 19(4): 1198–1207. [doi: 10.1109/TITS.2017.2725912]
90 Tesla vehicle deliveries and autopilot mileage statistics. https://lexfridman.com/tesla-autopilot-miles-and-vehicles/.
91 Fridman L, Terwilliger J, Jenik B. DeepTraffic: Crowdsourced hyperparameter tuning of deep reinforcement learning systems for multi-agent dense traffic navigation. Proceedings of the 32nd Conference on Neural Information Processing Systems. Montréal, QC, Canada. 2018. [doi: 10.5281/zenodo.2530457]
92 https://selfdrivingcars.mit.edu/deeptraffic-about/.
93 O'Kelly M, Sinha A, Namkoong H, *et al*. Scalable end-to-end autonomous vehicle testing via rare-event simulation.

Proceedings of the 32nd International Conference on Neural Information Processing Systems. Red Hook, NY, USA. 2018. 9849–9860.

94 Tang XC, Qin ZW, Zhang F, *et al*. A deep value-network based approach for multi-driver order dispatching. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Anchorage, AK, USA. 2019. 1780–1790.

95 Xu Z, Li ZX, Guan QW, *et al*. Large-scale order dispatch in on-demand ride-hailing platforms: A learning and planning approach. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London, UK. 2018. 905–913.

96 Zhao XY, Xia L, Tang JL, *et al*. Deep reinforcement learning for search, recommendation, and online advertising: A survey. ACM SIGWEB Newsletter, 2019: 4. [doi: 10.1145/3320496.3320500]

97 Yin DW, Hu YN, Tang JL, *et al*. Ranking relevance in yahoo search. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, CA, USA. 2016. 323–332. [doi: 10.1145/2939672.2939677]

98 Wei W, Xu J, Lan YY, *et al*. Reinforcement learning to rank with markov decision process. Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. Shinjuku, Japan. 2017. 945–948. [doi: 10.1145/3077136.3080685]

99 Zhang S, Yao LN, Sun AX, *et al*. Deep learning based recommender system: A survey and new perspectives. ACM Computing Surveys, 2019, 52(1): 5. [doi: 10.1145/3285029]

100 Zhao XY, Zhang L, Ding ZY, *et al*. Recommendations with negative feedback via pairwise deep reinforcement learning. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London, UK. 2018. 1040–1048. [doi: 10.1145/3219819.3219886]

101 Wu D, Chen C, Yang X, *et al*. A multi-agent reinforcement learning method for impression allocation in online display advertising. arXiv preprint arXiv: 1809.03152, 2019.