zbMATH Open
THE FIRST RESOURCE FOR MATHEMATICS

den Hengst, Floris; **François-Lavet, Vincent**; **Hoogendoorn, Mark**; **van Harmelen, Frank**
Planning for potential: efficient safe reinforcement learning. (English) Zbl 07570157
Mach. Learn. 111, No. 6, 2255-2274 (2022)

Summary: Deep reinforcement learning (DRL) has shown remarkable success in artificial domains and in some real-world applications. However, substantial challenges remain such as learning efficiently under safety constraints. Adherence to safety constraints is a hard requirement in many high-impact application domains such as healthcare and finance. These constraints are preferably represented symbolically to ensure clear semantics at a suitable level of abstraction. Existing approaches to safe DRL assume that being unsafe leads to low rewards. We show that this is a special case of symbolically constrained RL and analyze a generic setting in which total reward and being safe may or may not be correlated. We analyze the impact of symbolic constraints and identify a connection between expected future reward and distance towards a goal in an automaton representation of the constraints. We use this connection in an algorithm for learning complex behaviors safely and efficiently. This algorithm relies on symbolic reasoning over safety constraints to improve the efficiency of a subsymbolic learner with a symbolically obtained measure of progress. We measure sample efficiency on a grid world and a conversational product recommender with real-world constraints. The so-called Planning for Potential algorithm converges quickly and significantly outperforms all baselines. Specifically, we find that symbolic reasoning is necessary for safety during and after learning and can be effectively used to guide a neural learner towards promising areas of the solution space. We conclude that RL can be applied both safely and efficiently when combined with symbolic reasoning.

MSC:

68T05 Learning and adaptive systems in artificial intelligence

Keywords:

reinforcement learning; safety constraints; symbolic planning; reward shaping

Software:

AlphaZero; DeepSynth

Full Text: DOI

References:

[1] Alshiekh, M., Bloem, R., Ehlers, R., Könighofer, B., Niekum, S., \& Topcu, U. (2018). Safe reinforcement learning via shielding. In Proceedings of the AAAI conference on artificial intelligence (Vol. 32).

[2] Andreas, J., Klein, D., \& Levine, S. (2017). Modular multitask reinforcement learning with policy sketches. In Proceedings of the 36th international conference on machine learning vonference (pp. 166-175). PMLR.

[3] Baier, Christel; Katoen, Joost-Pieter, Principles of model checking (2008), MIT press · Zbl 1179.68076

[4] Bellemare, Marc; Srinivasan, Sriram; Ostrovski, Georg; Schaul, Tom; Saxton, David; Munos, Remi, Unifying count-based exploration and intrinsic motivation, Advances in neural information processing systems, 29, 1471-1479 (2016)

[5] Bloem, R., Könighofer, B., Könighofer, R., \& Wang, C. (2015). Shield synthesis. In International conference on tools and algorithms for the construction and analysis of systems (pp. 533-548). Springer. · Zbl 1420.68119

[6] Brafman, R. I., De Giacomo, G., \& Patrizi, F. (2018). LTLf/LDLf non-Markovian rewards. In Proceedings of the AAAI conference on artificial intelligence (Vol. 32).

[7] Burda, Y., Edwards, H., Pathak, D., Storkey, A., Darrell, T., \& Efros, A. A. (2019). Large-scale study of curiosity-driven learning. In International conference on learning representations.

[8] Camacho, A., Chen, O., Sanner, S., \& McIlraith, S. A. (2017). Non-markovian rewards expressed in LTL: Guiding search via reward shaping. In Tenth annual symposium on combinatorial search.

[9] Camacho, A., Icarte, R. T., Klassen, T. Q., Valenzano, R. A., \& McIlraith, S. A. (2019). Ltl and beyond: Formal languages for reward function specification in reinforcement learning. In Proceedings of the 28th joint conference on artificial intelligence (Vol. 19, pp. 6065-6073).

[10] De Giacomo, Giuseppe; Iocchi, Luca; Favorito, Marco; Patrizi, Fabio, Foundations for restraining bolts: Reinforcement learning

© 2023 FIZ Karlsruhe GmbH

with LTLf/LDLf restraining specifications, In Proceedings of the International Conference on Automated Planning and Scheduling, 29, 128-136 (2019)

[11] De Giacomo, Giuseppe; Favorito, Marco; Iocchi, Luca; Patrizi, Fabio, Imitation learning over heterogeneous agents with restraining bolts, In Proceedings of the International Conference on Automated Planning and Scheduling, 30, 517-521 (2020)

[12] den Hengst, F., Hoogendoorn, M., Van Harmelen, F., \& Bosman, J. (2019). Reinforcement learning for personalized dialogue management. In International conference on web intelligence (pp. 59-67). IEEE/WIC/ACM.

[13] den Hengst, Floris; Grua, Eoin Martino; el Hassouni, Ali; Hoogendoorn, Mark, Reinforcement learning for personalization: A systematic literature review, Data Science, 3, 1, 107-147 (2020)· doi:10.3233/DS-200028

[14] Dulac-Arnold, G., Mankowitz, D., \& Hester, T. (2019). Challenges of real-world reinforcement learning. In ICML workshop on real-life reinforcement learning. · Zbl 07465677

[15] Fu, J., \& Topcu, U. (2014). Probably approximately correct mdp learning and control with temporal logic constraints. In Proceedings of robotics: Science and systems (Vol. 10).

[16] Gaon, Maor; Brafman, Ronen, Reinforcement learning with non-markovian rewards, In Proceedings of the AAAI Conference on Artificial Intelligence, 34, 3980-3987 (2020)· doi:10.1609/aaai.v34i04.5814

[17] Grzes, M., \& Kudenko, D. (2008). Plan-based reward shaping for reinforcement learning. In International IEEE conference intelligent systems (Vol. 2, pp. 10-22). IEEE.

[18] Gu, S., Holly, E., Lillicrap, T., \& Levine, S. (2017). Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In 2017 IEEE international conference on robotics and automation (ICRA) (pp. 3389-3396). IEEE.

[19] Hasanbeig, M., Abate, A., \& Kroening, D. (2020). Cautious reinforcement learning with logical constraints. In Proceedings of the 19th international conference on autonomous agents and multiagent systems (pp. 483-491). · Zbl 1455.68190

[20] Hasanbeig, M., Jeppu, N. Y., Abate, A., Melham, T., \& Kroening, D. (2021).Deepsynth: Automata synthesis for automatic task segmentation in deep reinforcement learning. In The 35th AAAI conference on artificial intelligence, AAAI (Vol. 2, p. 36).

[21] Hessel, M., Modayil, J., Van Hasselt, H., Schaul, T., Ostrovski, G., Dabney, W., Horgan, D., Piot, B., Azar, M., \& Silver, D. (2018). Rainbow: Combining improvements in deep reinforcement learning. In Proceedings of the AAAI conference on artificial intelligence (Vol. 32).

[22] Icarte, R. T., Klassen, T., Valenzano, R., \& McIlraith, S. (2018). Using reward machines for high-level task specification and decomposition in reinforcement learning. In Proceedings of the 37th international conference on machine learning conference (pp. 2107-2116).

[23] Illanes, L., Yan, X., Icarte, R. T., \& McIlraith, S. A. (2020). Symbolic plans as high-level instructions for reinforcement learning. In Proceedings of the international conference on automated planning and scheduling (Vol. 30, pp. 540-550).

[24] Junges, S., Jansen, N., Dehnert, C., Topcu, U., \& Katoen, J.-P.. (2016). Safety-constrained reinforcement learning for mdps. In International conference on tools and algorithms for the construction and analysis of systems (pp. 130-146). Springer.

[25] Könighofer, B., Lorber, F., Jansen, N., \& Bloem, R. (2020). Shield synthesis for reinforcement learning. In International symposium on leveraging applications of formal methods (pp. 290-306). Springer.

[26] Mazala, R. (2002). Infinite games (pp. 23-38). Springer. ISBN 978-3-540-36387-3. · Zbl 1021.91502

[27] Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., \& Riedmiller, M. (2013). Playing atari with deep reinforcement learning. In NIPS deep learning workshop.

[28] Mnih, Volodymyr; Kavukcuoglu, Koray; Silver, David; Rusu, Andrei A.; Veness, Joel; Bellemare, Marc G.; Graves, Alex; Riedmiller, Martin; Fidjeland, Andreas K.; Ostrovski, Georg, Human-level control through deep reinforcement learning, Nature, 518, 7540, 529-533 (2015)· doi:10.1038/nature14236

[29] Ng, A. Y., Harada, D., \& Russell, S. (1999). Policy invariance under reward transformations: Theory and application to reward shaping. In Proceedings of the 16th international conference on machine learning (pp. 278-287).

[30] Pnueli, A. (1977). The temporal logic of programs. In 18th Annual symposium on foundations of computer science (pp. 46-57). IEEE.

[31] Pnueli, A., \& Rosner, R. (1989). On the synthesis of a reactive module. In ACM SIGPLAN-SIGACT (pp. 179-190). · Zbl 0686.68015

[32] Silver, David; Hubert, Thomas; Schrittwieser, Julian; Antonoglou, Ioannis; Lai, Matthew; Guez, Arthur; Lanctot, Marc; Sifre, Laurent; Kumaran, Dharshan; Graepel, Thore, A general reinforcement learning algorithm that masters chess, shogi, and go through self-play, Science, 362, 6419, 1140-1144 (2018) · Zbl 1433.68320 · doi:10.1126/science.aar6404

[33] Sutton, Richard S.; Barto, Andrew G., Reinforcement learning: An introduction (2018), MIT press · Zbl 1407.68009

[34] Tomic, S., Pecora, F., \& Saffiotti, A. (2020). Learning normative behaviors through abstraction. In Proceedings of the 24th European conference on artificial intelligence.

[35] Watkins, Christopher JCH; Dayan, Peter, Q-learning, Machine Learning, 8, 3-4, 279-292 (1992) · Zbl 0773.68062

[36] Wen, M., Ehlers, R., \& Topcu, U. (2015). Correct-by-synthesis reinforcement learning with temporal logic constraints. In 2015 IEEE/RSJ international conference on intelligent robots and systems (IROS) (pp. 4983-4990). RSJ/IEEE.

[37] Wiering, Marco; Van Otterlo, Martijn, Reinforcement learning, Adaptation, learning, and optimization, 12, 3 (2012)· doi:10.1007/978-3-642-27645-3$_1$

[38] Zhang, H., Gao, Z., Zhou, Y., Zhang, H., Wu, K., \& Lin, F. (2019). Faster and safer training by embedding high-level knowledge into deep reinforcement learning. arXiv preprint. arXiv:1910.09986

© 2023 FIZ Karlsruhe GmbH

This reference list is based on information provided by the publisher or from digital mathematics libraries. Its items are heuristically matched to zbMATH identifiers and may contain data conversion errors. It attempts to reflect the references listed in the original paper as accurately as possible without claiming the completeness or perfect precision of the matching.

© 2023 FIZ Karlsruhe GmbH