Scientific Experiments in Reinforcement Learning

Scott M. Jordan
University of Massachusetts
sjordan@cs.umass.edu

As a research community, one of our primary goals is to develop algorithms that will allow us to solve challenging real-world problems. However, our existing algorithms are difficult to apply, and we do not understand when they will likely work. As scientists, it is our job to produce knowledge that will let us better solve both existing and new problems. Furthermore, RL is evaluated empirically, so the knowledge we gain can only be as insightful as the experiments we conduct. The most common reinforcement learning (RL) experiment is performance evaluation, which compares the performance of several algorithms to establish that one method is superior to the others. Unfortunately, these experiments are often improperly set up, leading to results that are not reproducible and inconsequential [Henderson et al., 2018]. While many have proposed improvements for performance evaluations [Whiteson et al., 2011, Balduzzi et al., 2018, Jordan et al., 2020, Agarwal et al., 2021], performance evaluations suffer from an inescapable limitation: *they can only show which algorithm(s) work well, but not why*. To continue developing principled approaches to solve new problems, we must design experiments to improve our understanding of RL algorithms, not performance on benchmark problems.

This limitation of performance evaluation can also cause researchers to focus on developing new methods based on misunderstood concepts. For example, the Categorical DQN algorithm [Bellemare et al., 2017] was presented as a better method to approximate value functions, but the only experiments in the paper were performance evaluations. Based on claims of superior performance, researchers proposed other distributional approximations, all purporting to achieve superior performance [Dabney et al., 2018b,a, Barth-Maron et al., 2018, Yang et al., 2019]. However, later works showed that the distributional value representation only benefits neural networks and stems from providing a richer set of features throughout learning, not an improved representation of value [Lyle et al., 2019, Dabney et al., 2021]. These later experiments provide more value than any experiment showing improved performance because they give insights into what properties are essential for successful learning and can be generalized beyond a specific benchmark setting.

If we want our experiments to produce knowledge that can be generalized, we need to switch from asking competitive questions, e.g., does algorithm X outperform algorithm Y, to questions that are scientific [Hooker, 1995] in that they aim to further our understanding of how each algorithm works. This scientific style of experimentation should be the primary form of experimentation in papers. Performance evaluation should serve as a sanity check to confirm that the algorithms continue to perform well beyond carefully controlled experiments. Thankfully, there are many works performing experiments that further our understanding. For example, Tucker et al. [2018] conducted experiments to measure sources of variance and showed that action-dependent control variates produce little variance reduction over state-dependent ones, Linke et al. [2020] studied how the dynamics of different optimizers impact exploration when using intrinsic motivation, and Ghosh and Bellemare [2020] examined the representational properties for stable off-policy temporal difference learning. By measuring properties other than performance, these works produced valuable insights that go beyond a single algorithm and can be used to design new methods to meet specific challenges. When researchers conduct this type of experiment, designing algorithms for new problems becomes easier.

If we want to keep making progress and leverage the talents and creativity of our large community, we should stop competing to find the best algorithm and start working together to understand the necessary properties to solve sequential decision-making problems.

References

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G Bellemare. Deep reinforcement learning at the edge of the statistical precipice. *Advances in Neural Information Processing Systems*, 2021.

David Balduzzi, Karl Tuyls, Julien Pérolat, and Thore Graepel. Re-evaluating evaluation. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems, NeurIPS.*, pages 3272–3283, 2018.

Gabriel Barth-Maron, Matthew W. Hoffman, David Budden, Will Dabney, Dan Horgan, Dhruva TB, Alistair Muldal, Nicolas Heess, and Timothy P. Lillicrap. Distributed distributional deterministic policy gradients. In *6th International Conference on Learning Representations, ICLR*. OpenReview.net, 2018.

Marc G. Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning, ICML*, volume 70 of *Proceedings of Machine Learning Research*, pages 449–458. PMLR, 2017.

Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning, ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 1104–1113. PMLR, 2018a.

Will Dabney, Mark Rowland, Marc G. Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 2892–2901. AAAI Press, 2018b.

Will Dabney, André Barreto, Mark Rowland, Robert Dadashi, John Quan, Marc G. Bellemare, and David Silver. The value-improvement path: Towards better representations for reinforcement learning. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI*, pages 7160–7168. AAAI Press, 2021.

Dibya Ghosh and Marc G. Bellemare. Representations for stable off-policy reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning, ICML*, volume 119 of *Proceedings of Machine Learning Research*. PMLR, 2020.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18)*, pages 3207–3214, 2018.

John N. Hooker. Testing heuristics: We have it all wrong. *Journal of Heuristics*, 1(1):33–42, 1995.

Scott M. Jordan, Yash Chandak, Daniel Cohen, Mengxue Zhang, and Phillip S. Thomas. Evaluating the performance of reinforcement learning algorithms. In *Proceedings of the 37th International Conference on Machine Learning, ICML*, Proceedings of Machine Learning Research. PMLR, 2020.

Cam Linke, Nadia M. Ady, Martha White, Thomas Degris, and Adam White. Adapting behavior via intrinsic reward: A survey and empirical study. *J. Artif. Intell. Res.*, 69:1287–1332, 2020. doi: 10.1613/jair.1.12087. URL `https://doi.org/10.1613/jair.1.12087`.

Clare Lyle, Marc G. Bellemare, and Pablo Samuel Castro. A comparative analysis of expected and distributional reinforcement learning. In *The Thirty-Third AAAI Conference on Artificial Intelligence*, pages 4504–4511. AAAI Press, 2019.

George Tucker, Surya Bhupatiraju, Shixiang Gu, Richard E. Turner, Zoubin Ghahramani, and Sergey Levine. The mirage of action-dependent baselines in reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning, ICML*, pages 5022–5031, 2018.

Shimon Whiteson, Brian Tanner, Matthew E. Taylor, and Peter Stone. Protecting against evaluation overfitting in empirical reinforcement learning. In *2011 IEEE Symposium on Adaptive Dynamic Programming And Reinforcement Learning, ADPRL*, pages 120–127, 2011.

Derek Yang, Li Zhao, Zichuan Lin, Tao Qin, Jiang Bian, and Tie-Yan Liu. Fully parameterized quantile function for distributional reinforcement learning. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems, NeurIPS*, pages 6190–6199, 2019.