The Role of Environment Access in Agnostic Reinforcement Learning

Akshay Krishnamurthy AKSHAYKR@MICROSOFT.COM
Microsoft Research

Gene Li GENE@TTIC.EDU
Toyota Technological Institute at Chicago

Ayush Sekhari SEKHARI@MIT.EDU
MIT

Editors: Nika Haghtalab and Ankur Moitra

Abstract

We study Reinforcement Learning (RL) in environments with large state spaces, where function approximation is required for sample-efficient learning. Departing from a long history of prior work, we consider the weakest possible form of function approximation, called agnostic policy learning, where the learner seeks to find the best policy in a given class Π, with no guarantee that Π contains an optimal policy for the underlying task. Although it is known that sample-efficient agnostic policy learning is not possible in the standard online RL setting without further assumptions, we investigate the extent to which this can be overcome with stronger forms of access to the environment. Specifically:

1. We show that even with a strong function approximation assumption called *policy completeness*, and *generative access*—perhaps the strongest possible access to the MDP—policy learning methods cannot achieve sample complexity guarantees that scale with the intrinsic complexity of exploration, as measured via the *coverability coefficient* [XFB+22] of the MDP. This resolves an open problem posed by [JLR+23] and shows, in a strong, information-theoretic sense, that policy learning methods cannot explore.
2. We study the μ-reset setting, where the learner can roll out from an exploratory reset distribution μ, and investigate whether error amplification can be controlled without policy completeness (which is required for classical results of PSDP [BKSN03] and CPI [KL02]). We show that agnostic policy learning is information-theoretically impossible. We also show algorithm-specific lower bounds for PSDP and CPI under the weaker condition of *policy class realizability*.
3. In light of these lower bounds, we introduce a new model of access called *hybrid resets*, which subsumes both local simulators (which is weaker than generative access) and μ-resets. We show that under hybrid resets, and when the reset distribution satisfies *pushforward concentrability* [XJ21], sample-efficient policy learning is possible in Block MDPs [JKA+17, DKJ+19] via a new algorithm. Since all of our lower bound constructions are Block MDPs, this indicates the significant power of hybrid reset access in agnostic policy learning. On a technical level, we introduce a new algorithmic tool called a *policy emulator* that allows us to efficiently evaluate various policies within a large class Π. Informally speaking, a policy emulator is the "minimal object" useful for solving policy learning. Instead of learning the Block MDP in a traditional model-based sense (which would require samples scaling with the observation space size), our algorithm leverages hybrid resets to construct a policy emulator in a statistically efficient manner.

Taken together, our results reveal intriguing interplays between function approximation and environment access in RL.[1]

1. Extended abstract. Full version appears as [arXiv:2504.05405, v1].
Authors are listed in alphabetical order of their last names.

Acknowledgments

We thank Dylan Foster, Sasha Rakhlin, Zeyu Jia, Cong Ma, Nathan Srebro, and Wen Sun for helpful conversations. AS acknowledges support from ARO through award W911NF-21-1-0328, as well as Simons Foundation and the NSF through award DMS-2031883.

References

[BKSN03] James Bagnell, Sham M Kakade, Jeff Schneider, and Andrew Ng. Policy search by dynamic programming. *Advances in Neural Information Processing Systems*, 2003.

[DKJ+19] Simon Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudik, and John Langford. Provably efficient RL with rich observations via latent state decoding. In *International Conference on Machine Learning*, 2019.

[JKA+17] Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, 2017.

[JLR+23] Zeyu Jia, Gene Li, Alexander Rakhlin, Ayush Sekhari, and Nathan Srebro. When is agnostic reinforcement learning statistically tractable? *arXiv:2310.06113*, 2023.

[KL02] Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *International Conference on Machine Learning*, 2002.

[XFB+22] Tengyang Xie, Dylan J Foster, Yu Bai, Nan Jiang, and Sham M Kakade. The role of coverage in online reinforcement learning. *arXiv:2210.04157*, 2022.

[XJ21] Tengyang Xie and Nan Jiang. Batch value-function approximation with only realizability. In *International Conference on Machine Learning*, 2021.