Towards More Practical Reinforcement Learning

Travis Mandel[1]
Advisors: Emma Brunskill[2] and Zoran Popović[1]
[1]Center for Game Science, Computer Science & Engineering, University of Washington
[2]School of Computer Science, Carnegie Mellon University
tmandel@cs.washington.edu

Abstract

Reinforcement Learning is beginning to be applied outside traditional domains such as robotics, and into human-centric domains such as healthcare and education. In these domains, two problems are critical to address: We must be able to evaluate algorithms with a collection of prior data if one is available, and we must devise algorithms that carefully trade off exploration and exploitation in such a way that they are guaranteed to converge to optimal behavior quickly, while retaining very good performance with limited data. In this thesis, I examine these two problems, with an eye towards applications to educational games.

1 Introduction

Supervised machine learning has become pervasive, with broad impact both to computer science in general and to daily life. However, the end goal is usually not to learn to predict, but rather to learn to act intelligently. This falls under the umbrella of reinforcement learning, which asks how an agent should learn to act from its experience to optimize some reward signal. Despite this very general formulation, the real-world successes of reinforcement learning have been relatively few, and largely limited to robotics domains.

We believe this is not due to a lack of compelling applications: Domains such as education, healthcare, dialogue systems, and many more fit naturally into a reinforcement learning framework. However, there are two challenges that arise when developing reinforcement learning techniques for these domains. The first problem is the lack of principled evaluation methods that allow us to compare many different approaches on real-world problems, and have confidence in an algorithm before deploying it. The second is developing algorithms which tradeoff exploration and exploitation in a way that ensures good asymptotic performance, while also performing well empirically with limited data. In this thesis, we seek to address these two problems in a way that improves over state-of-the-art.

Our investigations span the applied and the theoretical: although we care about developing good general-purpose approaches, our immediate application is educational games. In these settings, we must choose how to intervene to increase player learning and engagement. For example, one game we have experimented on is a puzzle game called Refraction, which teaches kids how to multiply fractions by splitting laser beams. In this case, we need to choose right sequence of puzzles to give to students such that they complete the most concepts successfully. Making this decision is difficult because of the large amount of information we can collect about each student, most of which is not very useful for the task at hand. Refraction and our other educational games, such as Treefrog Treasure, have each been played by hundreds of thousands of players, providing ample opportunity for learning how to improve these games using data.

2 Evaluation

Motivation In all of machine learning, being able to evaluate different approaches is critical. We need an evaluation method to choose between different representations, optimize parameters within a representation, decide when to add new features, and convince ourselves that an approach will work well before deploying it in the real world.

In supervised learning, the problem of evaluation is trivial. Simple statistics such as accuracy can be computed and, when computed over a test dataset, do an excellent job of determining which of several disparate approaches is better at prediction without having to actually deploy the algorithm.

However, in reinforcement learning, how to measure the performance of a proposed algorithm is very difficult. One approach is to run it online in the true environment, but in most domains that is prohibitively expensive. A common approach in domains like robotics is to build a simulator based on our understanding of the rules that govern the world (i.e. physics), and evaluate approaches in the simulator to see which is most promising. However, one cannot in general assume such a simulator is of high quality. This becomes particularly apparent when interacting with humans, where we do not know simple rules that govern how a human will react to an educational or medical intervention. Instead, we seek to evaluate in an unbiased way using previously collected datasets. Work in contextual bandits has looked at this problem [Li *et al.*, 2011], but investigations in reinforcement learning have been limited. Popular methods in this area such as ECR [Tetreault and Litman, 2006], tended to be biased and inconsistent, and could be misleading in certain cases, as we showed in [Mandel *et al.*, 2014].

Past work In [Mandel *et al.*, 2014] we developed an unbiased data-driven methodology, incorporating ideas from prior work such as [Precup, 2000], for comparing different representations given a dataset. We demonstrated how this methodology allowed us to develop a good representation for a challenging high-dimensional reinforcement learning task. We deployed this policy in a trial with thousands of students and observed a 32% improvement over random and expert policies. We also observed that our estimates corresponded closely with reality.

In [Mandel *et al.*, 2015] we developed a new method for evaluation for the restricted case of multi-armed bandits. Compared to prior approaches, it is unbiased without needing to know the sampling distribution, and it is much more data-efficient, especially for deterministic algorithms.

3 Guarantees & Performance

In order to confidently deploy reinforcement learning algorithms into the real world, we need to have confidence that they will perform well far into the future. Specifically, we want guarantees that they will not just eventually identity the optimal policy, but converge (ideally quickly) to acting optimally. One robust way to formulate this is regret, and finding algorithms with optimal regret bounds is a problem that has been especially well-studied in the multi-armed bandit setting. However, a straightforward translation of popular bandit algorithms such as UCB to more realistic settings can cause poor empirical performance [Osband *et al.*, 2013; Chapelle and Li, 2011].

Past work In [Mandel *et al.*, 2015] we examined several assumptions made by typical bandit algorithms that are typically violated in real-world deployments. For example feedback may be delayed instead of returning immediately, or one may have access to an arbitrary prior dataset which they wish to incorporate without harming performance. We showed how these problems can cause issues in practice, but proposed an approach, inspired by work such as [Joulani *et al.*, 2013], to solving these problems that ensures good regret guarantees and good empirical performance. Another interesting problem tackled in this work is that of heuristics: One may wish to incorporate an arbitrary heuristic which can often help performance early on, but does not possess good theoretical guarantees. We showed how one can retain strong regret guarantees while using these heuristics to get an early benefit.

4 Future work

Evaluation We are currently working on ways to to do **nonstationary** policy evaluation in reinforcement learning. In the nonstationary case, the goal is not to evaluate a fixed policy, but instead to evaluate how an algorithm learns over time as it gathers data. Although this has been studied in contextual bandits, there hasn't been work addressing this problem in reinforcement learning. However, in reinforcement learning, balancing exploration and exploitation is still important, and it's important to be able to compare approaches on real problems. Preliminary results look promising in this case.

Guarantees & Performance We are currently examining Bayesian approaches for reinforcement learning which have been shown to have good empirical performance and good guarantees [Osband *et al.*, 2013]. Past work in this area did not do any generalization between states, limiting its empirical performance, especially in large domains. Typical approaches for Bayesian state aggregation are computationally expensive and typically come without meaningful theoretical guarantees. We are investigating new methods that are computationally efficient, are guaranteed to converge to acting optimally, and have good performance empirically.

We are also examining how best to cope with more realistic models of delay that arise in practice.

5 Conclusion

In this thesis, I look at a variety of problems that prevent reinforcement learning algorithms from being readily applicable to real-world problems such as educational games. Principal among these problems is developing good unbiased evaluation techniques, and carefully managing the exploration-exploitation tradeoff in such a way that we achieve good asymptotic guarantees and good empirical performance in realistic settings. We have made several contributions so far, but there remain further directions that merit investigation.

References

[Chapelle and Li, 2011] Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In *NIPS*, pages 2249–2257, 2011.

[Joulani *et al.*, 2013] Pooria Joulani, Andras Gyorgy, and Csaba Szepesvari. Online learning under delayed feedback. In *Proceedings of The 30th International Conference on Machine Learning*, pages 1453–1461, 2013.

[Li *et al.*, 2011] Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *WSDM*, pages 297–306. ACM, 2011.

[Mandel *et al.*, 2014] Travis Mandel, Yun-En Liu, Sergey Levine, Emma Brunskill, and Zoran Popovic. Offline policy evaluation across representations with applications to educational games. In *AAMAS*, pages 1077–1084. IFAAMAS, 2014.

[Mandel *et al.*, 2015] Travis Mandel, Yun-En Liu, Emma Brunskill, and Zoran Popovic. The queue method: Handling delay, heuristics, prior data, and evaluation in bandits. *AAAI*, 2015.

[Osband *et al.*, 2013] Ian Osband, Dan Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, pages 3003–3011, 2013.

[Precup, 2000] Doina Precup. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, page 80, 2000.

[Tetreault and Litman, 2006] Joel R Tetreault and Diane J Litman. Comparing the utility of state features in spoken dialogue using reinforcement learning. In *NAACL HLT*, 2006.