Efficient Exploration Using Expert Knowledge
CS748 - Project Proposal

Ishank Juneja - 16D070012

1 Introduction

Reinforcement Learning (RL) is a very promising framework for designing intelligent-agents that perform desired tasks while interacting with potentially complex environments. This is partly due to the fact that its objective of maximizing returns by taking desirable actions closely mimics the process of animal learning. Recently, RL has seen numerous success stories in the field of game playing. Examples include Atari games, Go and video games like DOTA. However, the adoption of RL outside of simulated environments is not yet widespread.
A part of the problem with adopting RL can be attributed to the requirement that an agent start exploring its environment from scratch every time its designer wishes for it to perform a new task. Starting tabula-rasa, and learning new behavior through pure exploration requires large sequences of sample points. This requirement imposed on artificial agents is far from our everyday experience. In the case of humans, our prior learning guides our learning of new related tasks. While learning a task, our exploration is biased by our skill set acquired up to that point. For instance consider a person learning to ride a bicycle. Initially it takes them a long time to internalize knowledge about maintaining balance and controlling the handle-bar. However, when the same person later learns to ride a motorcycle, it takes them very little exploration in learning to ride it. The aim of this project is to take a step towards achieving this goal for artificial agents.

2 Related Work

There has been extensive work towards making agents learn desired behaviours within a reasonable sample horizon. Some early work includes the paradigm of reward-shaping (Ng et al. (1999)). Under this scheme, expert knowledge about the problem is incorporated into a shaping reward function which is added on top of the sparse rewards from the agent's environment. The additional shaping rewards encourage the agent to learn desirable behaviours faster. The problem with this technique is the difficulty of incorporating prior knowledge into the form of an effective shaping function. Another approach that works by the designer controlling the reward function is reward-search (Singh et al. (2009)). This is a meta-learning method which performs a brute-force search on the space of possible rewards to learn a reward scheme under which an agent can learn the desired behaviour the quickest. Although great in principle, this approach is infeasible outside of small tabular domains due to a search space exponential in the number of state action pairs (s, a). Further there is no direct way of incorporating expert knowledge into the reward-search framework.
Some other strategies for efficient exploration learn distributions of latent variables during the deployed agents lifetime. Examples of these methods include Hausman et al. (2018) and Gupta et al. (2018). The learnt generative models are then used to bias future exploration undertaken by the agent. Empirically these techniques have been shown to be quite effective, however their successes are domain specific and they don't provide any universal guarantees for improved performance.
Although the motivation for this work is aligned with that of the aforementioned works, the view taken for the problem is more closely aligned with the idea of transfer learning. As per Torrey & Shavlik (2009) "Transfer learning is the improvement of learning in a new task through the transfer of knowledge from a related task that has already been learned." In particular, through this work, we wish to reduce required exploration to achieve good performance by *injecting expert knowledge*.

Probably Approximately Correct (PAC) - RL

In this work, the structure of the expert knowledge to be incorporated into exploration is taken to be a good policy π^e such that $||V^*(s) - V^{\pi^e}(s)||_\infty < \Delta_{\pi^e}$. Where V^* and V^{π^e} are the value functions for the optimal policy and π^e respectively and Δ_{π^e} is a constant. Keeping this in mind a natural choice for a strategy to build upon would be the PAC-RL framework.
An agent's learning algorithm qualifies as PAC-RL if for any Markov Decision Process (MDP) $\{S, A, T, R, \gamma\}$, on halting the algorithm outputs a policy π such that

$$\mathbb{P}(||V^* - V^\pi||_\infty < \epsilon) > 1 - \delta$$

To qualify as PAC-RL, the algorithm must also do this efficiently - In time steps polynomial in $|S|, |A|, 1/\epsilon, 1/\delta, 1/(1-\gamma)$ and R_{max} (the largest reward).
Over the years, a few efficient PAC-RL strategies have been proposed. Explicit Explore or Exploit - $\mathbf{E}^3$ (Kearns & Singh (1998)) is one of the early strategies that popularised PAC-RL. $\mathbf{E}^3$ computes explore/exploit policies valid for a fixed number of steps during exploration. The algorithm is somewhat complicated with numerous rules and sub-routines. In contrast *R-MAX* (Brafman & Tennenholtz (2002)) is a simpler strategy that computes a parameter c based on the requirements ϵ and δ. Then for each state-action pair (s, a) explored fewer than c times, it sets $Q(s, a) = R_{max}/(1-\gamma)$. Action elimination (Even-Dar et al. (2003)) borrows ideas from the optimal arm selection problem in multi armed bandits. Particularly it builds upon the successive arm elimination algorithm. Most recently, Model Based Interval Estimation-MBIE (Strehl & Littman (2005)) was proposed as a viable candidate for the PAC-RL problem. It works by maintaining confidence intervals on the probability distributions associated with each state-action pair. Further, it provides an exploration bonus to insufficiently explored states. Both of these being ideas that mirror the UCB (Upper Confidence Bound) arm selection algorithm for multi armed bandits.

3 Problem Specification

As mentioned earlier, through this work, we aim to incorporate expert knowledge to perform more informed exploration. It is assumed that this knowledge is available in the form of a Δ_{π^e}-optimal policy $\pi^e(s)$. Precisely, for exploration on a given MDP $M = \{S, A, T, R, \gamma\}$ we wish to design a PAC-RL algorithm $\mathcal{A}$ with guarantees ϵ, δ such that

$$T^{\mathcal{A}}(\{S, A, T, R, \gamma, \epsilon, \delta, \pi^e\}) < T^{PAC-RL}(\{S, A, T, R, \gamma, \epsilon, \delta\})$$

For certain good policies π^e. Here $T^{\mathcal{A}}$ and T^{PAC-RL} are the time horizons required to provide the PAC ϵ, δ guarantees for the algorithms $\mathcal{A}$ and the base PAC-RL algorithm (on top of which $\mathcal{A}$ is built) respectively.

4 Proposed Approach and Plan

To achieve the stated goal, the proposed approach is to first try and develop an enhanced exploration strategy for the PAC - multi armed bandit problem (PAC-MAB). This is motivated by the course of work taken by Even-Dar et al. (2002) which was followed by the action-elimination algorithm for MDPs Even-Dar et al. (2003). In particular Even-Dar et al. (2002) provides a black-box strategy for converting a PAC-MAB algorithm to a PAC-RL algorithm. In case this does not work out an alternate strategy is to try and inject the optimal policy π^* into a PAC-RL exploration strategy. The decision on which of the available algorithms to choose will be taken at a later stage.
The procedure for developing expert knowledge injected strategies would be to try and *intelligently guess* ideas that may work and test their effectiveness via simulations.

Time	Target Tasks
2 weeks (Midterm)	Thorough Literature review Either a strategy for PAC-MABs or Informed decision on PAC framework to build upon
8 weeks	Initial results using optimal policy π^* as expert knowledge or results from MAB to RL
12 weeks (Final)	Extension of results to Δ-optimal (good) policies

References

Brafman, R. I. and Tennenholtz, M. R-max - a general polynomial time algorithm for near-optimal reinforcement learning. *J. Mach. Learn. Res.*, 3(3):213–231, March 2002. ISSN 1532-4435. doi: 10.1162/153244303765208377. URL https://doi.org/10.1162/153244303765208377.

Even-Dar, E., Mannor, S., and Mansour, Y. PAC bounds for multi-armed bandit and markov decision processes. In Kivinen, J. and Sloan, R. H. (eds.), *Computational Learning Theory, 15th Annual Conference on Computational Learning Theory, COLT 2002, Sydney, Australia, July 8-10, 2002, Proceedings*, volume 2375 of *Lecture Notes in Computer Science*, pp. 255–270. Springer, 2002. doi: 10.1007/3-540-45435-7_18. URL https://doi.org/10.1007/3-540-45435-7_18.

Even-Dar, E., Mannor, S., and Mansour, Y. Action elimination and stopping conditions for reinforcement learning. In Fawcett, T. and Mishra, N. (eds.), *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003, Washington, DC, USA*, pp. 162–169. AAAI Press, 2003. URL http://www.aaai.org/Library/ICML/2003/icml03-024.php.

Gupta, A., Mendonca, R., Liu, Y., Abbeel, P., and Levine, S. Meta-reinforcement learning of structured exploration strategies. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 5302–5311. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/7776-meta-reinforcement-learning-of-structured-exploration-strategies.pdf.

Hausman, K., Springenberg, J. T., Riedmiller, M., Heess, N., and Wang, Z. Learning an embedding space for transferable robot skills. 2018.

Kearns, M. J. and Singh, S. P. Near-optimal reinforcement learning in polynominal time. In Shavlik, J. W. (ed.), *Proceedings of the Fifteenth International Conference on Machine Learning (ICML 1998), Madison, Wisconsin, USA, July 24-27, 1998*, pp. 260–268. Morgan Kaufmann, 1998.

Ng, A. Y., Harada, D., and Russell, S. J. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99. Morgan Kaufmann Publishers Inc., 1999.

Singh, S., Lewis, R. L., and Barto, A. G. Where do rewards come from? In *Proceedings of the Annual Conference of the Cognitive Science Society*, 2009.

Strehl, A. L. and Littman, M. L. A theoretical analysis of model-based interval estimation. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, pp. 856–863, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1595931805. doi: 10.1145/1102351.1102459. URL https://doi.org/10.1145/1102351.1102459.

Torrey, L. and Shavlik, J. Transfer learning, 2009.