# Deep Reinforcement Learning Spring 2020 Assignment 1: Reading Summary

**Minghao Zhang**

## 1 Introduction

*Exploration and Exploitation Trade-off* is a classical and important problem in reinforcement learning area[1]. On the one hand, we need to learn how to use our current knowledge to improve policy better; on the other hand, we have to explore the unseen environment to make sure we can learn the optimal policy. There are lots of ways to improve the exploration ability and maintain good performance on exploitation meanwhile. Information based method[2], curiosity-driven method[3] and count-based method[4] are proposed in recent years to handle this difficulty with different intuitions. This paper shows that PSRL obeys performance guarantees intimately. Specifically, PSRL samples the context variable from the posterior distribution conditioned on the history as the conditional feature at the start of a training iteration and then computes the policy based on the sample, which, as they claim, is regarded as the uncertainty of the environment to guide exploration. They give the finite-time bounds on regret and show much analysis separated from the algorithm. Besides, thanks to the simple sampling-based algorithm, PSRL is computationally efficient relative to many optimistic methods. However, it also has some limitations and can be extended from many perspectives.

## 2 Algorithm Understanding

Briefly speaking, PRSL samples from the history to get a context which represents for the current experience, and then uses this knowledge to build a policy with the better ability of exploration. PSRL always selects policies according to the probability they are optimal since it can learn the uncertainty about each policy in a statistical efficient way through the posterior distribution. PSRL begins with a prior distribution over MDP. This can be understood as the initial uncertainty about this new unknown environment. At the start of each episode, PSRL samples an MDP from the posterior distribution conditioned on the current history. This step means we use the learned knowledge to compute the probability that each action in a given state will be the optimal one. Meanwhile, consideration of the history lets the agent know "what has already been learned" or "what the agent has known about certainly". Then it computes the new policy and executes it. In a different view, this algorithm can be regarded as a stochastic optimism which is simple and sufficient for learning. The reason for this is that random sampling from historical experience can approximately get a estimated MDP distribution and give us a chance to analyse without knowing the true MDP space in the expectation case. So it also provides a genernal view on performance and regret.

## 3 Insight of the Analysis

The main theorem of this paper is that the expected regret of this algorithm can be bounded by a finite time polynomial. This result holds for any prior distribution on MDPs, so it has a great generalization. Though it's more common to measure the performance with the worst case, this bound can show a more explicit performance of the agent. Furthermore, its corollary also gives a bound for the frequentist regret for any MDPs with a non-zero probability distribution. The term $\tau$ in the bound allows the optimization on the episode length to reduce the regret, contrast to the former work's non-optimizable constant term.

100th Conference on Neural Information Processing Systems (NeurIPS 2086), Beijing, China.

They also demonstrate that the sampled MDP leads a total regret that is an unbiased estimation of the regret led by the ground-truth MDP. This conclusion is led by a centric observation that the sampled MDP and the true MDP are identically distributed. This ensures an incremental improvement of the policy.
Besides, they provide more analysis of different views. They rewrite the regret into two terms: one step Bellman error under the sampled MDP and the randomness in the transitions of the true MDP. This decomposition has some similarities with the variational inference, which allows us to estimate the unreachable true MDP's properties. They also prove that the sampled Bellman operator concentrates around the true Bellman operator by introducing a confidence set. The proposed analysis bridges the value iteration process and this regret based on the posterior sampling computation, which makes this work completed.

## 4 Limitations and Extensions

Although this work inspires the community a lot in terms of how to use the gained experience to improve policy, it remains many limitations and has a huge space to be extended.
Firstly, this work didn't have any direct optimization of the policy or value function. Only depending on the historical changes to improve the policy may not have a near-optimal guarantee, i.e., it can play well in the expectation view but never achieve a near-optimal policy. To solve this problem, naively we can add the modern deep reinforcement learning algorithm to help improve policy with reward. Combining with replay buffer algorithms may also lead the great improvement and context property can extend this PSRL-based method to meta learning area[5][6].
Secondly, although it has an expectation guarantee, this algorithm just randomly samples the historical MDPs, without caring about the quality of the transitions. For instance, if we always sample the poor transition with no reward in a sparse reward environment, we may never have the chance to learn the right policy. The worst results may be extremely terrible. To overcome this shortcoming, we can consider a prioritized replay buffer[7] to sample those more efficient MDPs firstly, say, the one with high reward or big TD error (depend on the choice of the RL algorithm). This may also lower the regret bound simply because the sampling process supplies more advantages to the exploration strategy.
Last but not least, this algorithm has a very tricky problem: it needs to consider the implementation of the buffer. Ina large state space setting, we can not store all the history MDPs. If we just forget the former transitions, then we are still facing the exploration ability limitation since we still can't accurately learn the experience. Meanwhile, the needed memory will also be so large to apply in large environments. This may be the general weakness of sample-based RL and there is a trade-off between memory and performance which needs more human engineering.

## References


[1] Kun Shao, Zhentao Tang, Yuanheng Zhu, Nannan Li, and Dongbin Zhao. A Survey of Deep Reinforcement Learning in Video Games. *arXiv e-prints*, page arXiv:1912.10944, December 2019.

[2] Anirudh Goyal, Riashat Islam, Daniel Strouse, Zafarali Ahmed, Matthew Botvinick, Hugo Larochelle, Yoshua Bengio, and Sergey Levine. InfoBot: Transfer and Exploration via the Information Bottleneck. *arXiv e-prints*, page arXiv:1901.10902, January 2019.

[3] Adrien Laversanne-Finot, Alexandre Péré, and Pierre-Yves Oudeyer. Curiosity Driven Exploration of Learned Disentangled Goal Spaces. *arXiv e-prints*, page arXiv:1807.01521, July 2018.

[4] Georg Ostrovski, Marc G. Bellemare, Aaron van den Oord, and Remi Munos. Count-Based Exploration with Neural Density Models. *arXiv e-prints*, page arXiv:1703.01310, March 2017.

[5] Kate Rakelly, Aurick Zhou, Deirdre Quillen, Chelsea Finn, and Sergey Levine. Efficient Off-Policy Meta-Reinforcement Learning via Probabilistic Context Variables. *arXiv e-prints*, page arXiv:1903.08254, March 2019.

[6] Abhishek Gupta, Russell Mendonca, YuXuan Liu, Pieter Abbeel, and Sergey Levine. Meta-Reinforcement Learning of Structured Exploration Strategies. *arXiv e-prints*, page arXiv:1802.07245, February 2018.

[7] Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized Experience Replay. *arXiv e-prints*, page arXiv:1511.05952, November 2015.