Summary

title Regret analysis of constrained irreducible MDPs with reset action
author Takashi Watanabe

In this thesis, we study regret analysis of a reinforcement learning on constrained irreducible Markov Decision Processes (MDPs) with reset action. In reinforcement learning(RL), a learner interacts with an environment with trading off the exploration to collect information about the environment and the exploitation of gathered information. To evaluate the performance of RL algorithms on MDPs, the regret framework is one of typical frameworks that have been studied (Regret is the notion of the difference between the gain of an optimal policy and accumulative reward of an algorithm). For weakly-communicating MDPs, there have been many studies which analyze the regret of specific algorithm (a weakly-communicating MDP is an MDP which induces Markov-chain which has a single closed irreducible class and a set of states which is transient under all stationary policies for some stationary policies)[1, 2, 3, 4]. Among these studies, common approaches can divide rougly into the two directions: optimism in the face of uncertainty(OFU) and posterior sampling. In the OFU approach, the learner has optimistic estimates of the value function and executes the optimistic policy. On the other hand, in the posterior sampling approach the learner has a Bayesian distribution over MDPs and executes the optimal policy of a sampled MDP. In this paper, we focus on the OFU approach. There are a lot of extensions of MDPs, and one of them is Constrained MDPs(CMDPs). CMDP has multiple reward functions, one of which corresponds to that of MDPs, and the others correspond to constraints. A learner tries to learn a policy which satisfies all constraints of the CMDP and gets more accumulative reward of objective reward function. For irreducible CMDPs (the word irreducible means that for any deterministic policy, Markov-chain induced by the policy and CMDP is irreducible), we can use linear programming problem to compute optimal stationary policies. This means that the solution of the linear programming problem corresponds to stationary policy one to one. In this thesis, we extend the regret framework to constrained case and study CMDPs with reset action through this extended regret framework. We introduce an algorithm, Constrained-UCRL, which is motivated by UCRL2

algorithm[1]. Constrained-UCRL uses confidence intervals like UCRL2, and solves a linear programming problem to compute policy instead of extended value iteration in UCRL2. We show that Constrained-UCRL achieves regret bounds $\tilde{O}(SA^{\frac{1}{2}}T^{\frac{3}{4}})$ up to logarithmic factors with high probability for both the gain and the constraint violations.

Section 2 presents the preliminary of this thesis. We describe the definition of MDP, policy, and other related notion. Also we describe the corresponding definition in the case of CMDP.

Section 3 presents the description of Constrained-UCRL algorithm. As we mentioned above, Constrained-UCRL is motivated by UCRL2 algorithm, the algorithm which is shown to have sublinear regret bounds with high probability for weakly-communicating MDPs. The main difference between Constrained-UCRL and UCRL2 is the way to compute optimistic policy. UCRL2 uses extended value iteration to compute it. However, in CMDP settings, we face the difficulties in the execution of (extended) value iteration due to the existence of multiple rewards which need to consider simultaneously. Then instead of (extended) value iteration, Constrained-UCRL uses the linear programming.

At first of Section 3, we mention the five assumptions in this thesis: Irreducibility of MDPs/CMDPs, deterministic rewards, the existence of reset action for CMDPs, satisfiability of CMDPs, and the given information to the learner. Then we describe the pseudo-code of Constrained-UCRL, and state the main theorem which show that Constrained-UCRL algorithm has sublinear regret bounds with high probability for both the gain and the constraint violations. In MDP settings, sublinear regret (bound) is the condition which indicates that an algorithm behaves like the optimal policy after sufficient timesteps. This theorem implies that there is an algorithm that learn some kinds of CMDPs practically like the case of weakly-communicating MDPs.

Section 4 is the preparation of Section 5. In this section, we show the relationship between the linear programming problem in the pseudo-code of Constrained-UCRL algorithm and the execution policy. To compute the execution policy, we take three steps: modifying confidence interval, solving a linear programming problem corresponds to modified confidence interval, modifying the reset probability (the probabilty of selecting the reset action) of computed policy. The purpose of modifying confidence interval is to limit the interval which includes irreducible CMDPs only. Due to this modification, we can connect the modified confidence interval with the linear programming problem for an irreducible CMDP. Though we do not get the linear programming problem corresponds to the modified confidence interval, we can get it through introducing new variables. After solving the linear programming problem, if there is no solution of the linear programming problem, then Constrained-UCRL algorithm terminates and returns false, which means that Constrained-UCRL fail to learn true CMDP. Otherwise, through computing over the solution of the problem we get an optimistic policy of modified confidence interval. Indeed, we take the other step, modifying computed policy, to get small bias span through control the least probability of selecting reset action in modified policy. Bias span is the span of bias vector, where bias vector denotes the state-dependent accumulative rewards. Since in irreducible MDPs and CMDPs the gain of a stationary policy becomes to state-independent, the effect of bias becomes smaller and smaller with thinking of expected average behavior. But, in Constrained-UCRL algorithm, we change policy at the start of each episode so that we need to consider the effect of bias.

Section 5 presents the proof of the main theorem. This analysis is mainly followed in a way of the regret analysis of UCRL2. We evaluate the stochastic upper bound of T-step accumulative regret which corresponds to i-th constraint for specific i, t. This evaluation is separated into three parts: evaluation between the gain of true optimal policy on true CMDP and that of estimated policy on estimated CMDP, evaluation between the gain of estimated policy on estimated CMDP and that on true CMDP, evaluation between the gain of estimated policy on true CMDP and the actual accumulative reward during execution of Constrained-UCRL. In the first evaluation, we use the result of optimism of estimated policy in modified confidence interval. In the second evaluation, we use the small bias span of estimated policy and CMDP, and the upper bound of L_1-deviation of transition probability function against true CMDP. In the third evaluation, we use Azuma-Hoeffding Inequality[5, 6], which stochastically evaluates the sum of random variables which compose a martingale-difference sequence. On the other hand, we evaluate the upper bound of the probability that Constrained-UCRL algorithm terminates until timestep T, which means Constrained-UCRL algorithm fail to learn true CMDP. Putting the result of these analyses together, we show that the theorem holds.

Section 6 concludes this thesis.

The contribution of this thesis is summarized as follows. At first, we give a way to extend the regret framework for the non-constrained case to the constrained case. This extended regret framework matches the original framework in the case that there is no constraint. Then, we show the analysis of Constrained-UCRL algorithm with this extended regret framework. The main theorem shows that Constrained-UCRL algorithm has sublinear regret bounds with high probability for both the gain and the constraint violations. This implies that there is an algorithm that learn some kinds of CMDPs practically like the case of weakly-communicating MDPs.

References

[1] Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600, 2010.

[2] Aristide Tossou, Debabrota Basu, and Christos Dimitrakakis. Near-optimal optimistic reinforcement learning using empirical bernstein inequalities. *arXiv preprint arXiv:1905.12425*, 2019.

[3] Peter L Bartlett and Ambuj Tewari. Regal: A regularization based algorithm for reinforcement learning in weakly communicating mdps. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 35–42. AUAI Press, 2009.

[4] Ronan Fruit, Matteo Pirotta, Alessandro Lazaric, and Ronald Ortner. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In *International Conference on Machine Learning*, pages 1573–1581, 2018.

[5] Kazuoki Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series*, 19(3):357–367, 1967.

[6] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.