Multi-Agent Thompson Sampling for Bandits with Sparse Neighbourhood Structures

Timothy Verstraeten[1], Eugenio Bargiacchi[1], Pieter J.K. Libin[1], Jan Helsen[1], Diederik M. Roijers[1,2], and Ann Nowé[1]

[1] Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium
[2] HU University of Applied Sciences Utrecht, Utrecht, 3584CS, the Netherlands

This is an extended abstract of the article that was accepted and published in Nature Scientific Reports: https://doi.org/10.1038/s41598-020-62939-3

Multi-agent coordination is prevalent in many real-world applications, such as traffic light control, warehouse commissioning and wind farm control. Often, such settings can be formulated as coordination problems in which agents have to cooperate in order to optimize a shared team reward. Handling multi-agent settings is challenging, as the size of the joint action space scales exponentially with the number of agents in the system. Therefore, an approach that directly considers all agents' actions jointly is computationally intractable. This has made such coordination problems the central focus in the planning literature. Fortunately, in real-world settings agents often only directly affect a limited set of neighbouring agents. This means that the global reward received by all agents can be decomposed into local components that only depend on small subsets of agents. Exploiting such loose couplings is key in order to keep multi-agent decision problems tractable.

In this work, we consider learning to coordinate in multi-agent systems. While most of the literature only considers approximate reinforcement learning methods for learning in multi-agent systems, it has recently been shown that it is possible to achieve theoretical bounds on the regret (i.e., how much reward is lost due to learning). In this work, we use the multi-agent multi-armed bandit problem definition, and improve upon the state of the art. Specifically, we propose the multi-agent Thompson sampling (MATS) algorithm [5, 6], which exploits loosely-coupled interactions in multi-agent systems.[3] The loose couplings are formalized as a *coordination graph*, which defines for subsets of agents whether their actions depend on each other. We assume the graph structure is known beforehand, which is the case in many real-world applications with sparse agent interactions (e.g., wind farm control). Additionally, our method leverages the exploration-exploitation mechanism of Thompson sampling (TS). TS has been shown to be highly competitive to other popular methods, e.g., the Upper Confidence Bound algorithm [3]. Recently, theoretical guarantees on its regret have been established, which renders the method increasingly popular in the literature. Additionally, due to its Bayesian nature, problem-specific priors

[3] The source code of MATS is available at github.com/Svalorzen/AI-Toolbox. [1]

can be specified, which has strong relevance in many practical fields, such as advertisement selection and influenza mitigation.

We provide a finite-time Bayesian regret analysis and prove that the upper regret bound of MATS is low-order polynomial in the number of actions of a single agent for sparse coordination graphs. This is a significant improvement over the exponential bound of classic TS, which is obtained when the coordination graph is ignored. Moreover, we show that MATS improves upon the state-of-the-art algorithms, Multi-Agent Upper Confidence Exploration (MAUCE) [2] and Sparse Cooperative Q-Learning (SCQL) [4], in various synthetic settings. Although MATS and MAUCE have similar theoretical guarantees, we found that MATS consistently outperforms both MAUCE and SCQL empirically. We argue that the high performance of MATS is due to the ability to seamlessly include domain knowledge about the reward distributions and treat the problem parameters as unknowns. To highlight the power of this property, we introduced a novel setting with skewed reward distributions. As MAUCE only supports symmetric exploration bounds, it is challenging to correctly assess the amount of exploration needed to solve this task. In contrast, MATS has the ability to exploit the shape of the reward distribution to achieve more targeted exploration. Finally, we demonstrate the practical benefits of MATS on a realistic wind farm control task. As wind passes through the farm, downstream turbines observe a significantly lower wind speed. This is known as the *wake effect*, which is due to the turbulence generated behind operational turbines. Wake redirection is a control mechanism where turbines' rotors are misaligned to deflect wake away from the wind farm. While a misaligned turbine produces less energy on its own, the group's total productivity is increased. Physically, the wake effect reduces over long distances, and thus, turbines tend to only influence their neighbours. We can use this domain knowledge to define groups of agents and organize them in a graph structure. We demonstrate that MATS achieves state-of-the-art performance on the wind farm control task.

References

1. Bargiacchi, E., Roijers, D.M., Nowé, A.: AI-Toolbox: A C++ library for reinforcement learning and planning (with python bindings). Journal of Machine Learning Research **21**(102), 1–12 (2020), http://jmlr.org/papers/v21/18-402.html
2. Bargiacchi, E., Verstraeten, T., Roijers, D., Nowé, A., van Hasselt, H.: Learning to coordinate with coordination graphs in repeated single-stage multi-agent decision problems. In: International Conference on Machine Learning (2018)
3. Chapelle, O., Li, L.: An empirical evaluation of Thompson sampling. In: Advances in Neural Information Processing Systems (NIPS). vol. 24, pp. 2249–2257 (2011)
4. Kok, J.R., Vlassis, N.: Sparse cooperative q-learning. In: Proc. of the 21st International Conference on Machine Learning. New York, NY, USA (2004)
5. Verstraeten, T., Bargiacchi, E., Libin, P.J.K., Helsen, J., Roijers, D.M., Nowé, A.: Multi-agent Thompson sampling for bandit applications with sparse neighbourhood structures. Sci. Rep. **10** (2020). https://doi.org/10.1038/s41598-020-62939-3
6. Verstraeten, T., Bargiacchi, E., Libin, P.J.K., Roijers, D.M., Nowé, A.: Thompson sampling for factored multi-agent bandits. In: Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems. pp. 2029–2031 (2020)