Thompson Sampling for Factored Multi-Agent Bandits

Extended Abstract

Timothy Verstraeten
Vrije Universiteit Brussel
Brussels, Belgium
tiverstr@vub.be

Eugenio Bargiacchi
Vrije Universiteit Brussel
Brussels, Belgium
ebargiac@vub.be

Pieter J.K. Libin
Vrije Universiteit Brussel
Brussels, Belgium
plibin@vub.be

Diederik M. Roijers
HU University of Applied Sciences
Utrecht, Netherlands
diederik.yamamoto-roijers@hu.nl

Ann Nowé
Vrije Universiteit Brussel
Brussels, Belgium
anowe@vub.be

CCS CONCEPTS

• **Computing methodologies** → **Multi-agent reinforcement learning**; **Sequential decision making**; • **Mathematics of computing** → *Probability and statistics*;

KEYWORDS

Multi-Agent Thompson Sampling, multi-agent multi-armed bandits

ACM Reference Format:
Timothy Verstraeten, Eugenio Bargiacchi, Pieter J.K. Libin, Diederik M. Roijers, and Ann Nowé. 2020. Thompson Sampling for Factored Multi-Agent Bandits. In *Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), Auckland, New Zealand, May 9–13, 2020,* IFAAMAS, 3 pages.

Multi-agent decision making is prevalent in many real-world applications, such as wind farm control [11], traffic light control [12] and warehouse commissioning [5]. In these settings, agents need to cooperate to maximize a shared team reward [3].

Coordination in multi-agent settings is challenging, due to the combinatorial increase in terms of the number of agents. Therefore, it is computationally intractable to consider all agents' actions jointly. Fortunately, in many real-world settings, an agent is only directly influenced by a small subset of neighboring agents. In this case, the team reward can be factorized over the groups of agents that influence each other. Such a sparse factorization must be exploited to keep multi-agent decision problems tractable.

In this work, we consider learning to coordinate in multi-agent multi-armed bandit problems (Section 1), and propose the Multi-Agent Thompson Sampling (MATS) algorithm (Section 2), which exploits sparse interactions in multi-agent systems. We assume that the groups of interacting agents are known beforehand, which is often the case in real-world applications.

Our method uses the exploration-exploitation mechanism of Thompson sampling (TS). TS has been shown to achieve high empirical performance [4]. Moreover, TS is a Bayesian method, which allows for the specification of prior knowledge through belief distributions. We argue that this is an important property to have in many practical applications, such as influenza mitigation [7, 8].

Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), May 9–13, 2020, Auckland, New Zealand.

We compare MATS against Sparse Cooperative Q-Learning (SCQL) and Multi-Agent Upper Confidence Exploration (MAUCE) on two synthetic settings, i.e., Bernoulli 0101-Chain and Gem Mining (Section 3). MATS improves upon the state of the art with respect to accuracy, learning speed and computational speed.

1 PROBLEM DESCRIPTION

In this work, we adopt the multi-agent multi-armed bandit (MAMAB) setting [2]. A MAMAB is similar to the multi-armed bandit formalism [10], but considers multiple agents factored into groups. When the agents have pulled a joint arm, each group receives a reward. The goal shared by all agents is to maximize the total sum of rewards. Formally,

Definition 1.1. A multi-agent multi-armed bandit (MAMAB) is a tuple $\langle \mathcal{D}, \mathcal{A}, f \rangle$ where

- $\mathcal{D}$ is the set of m enumerated agents. This set is factorized into ρ, possibly overlapping, subsets of agents $\mathcal{D}^e$.
- $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_m$ is the set of joint actions, or joint arms, which is the Cartesian product of the sets of actions $\mathcal{A}_i$ for each of the m agents in $\mathcal{D}$. We denote $\mathcal{A}^e$ as the set of local joint actions, or local arms, for the group $\mathcal{D}^e$.
- $f(\boldsymbol{a})$ is a stochastic function providing a global reward when a joint arm, $\boldsymbol{a} \in \mathcal{A}$, is pulled. The global reward function is decomposed into ρ noisy, observable and independent local reward functions, i.e., $f(\boldsymbol{a}) = \sum_{e=1}^{\rho} f^e(\boldsymbol{a}^e)$. A local function f^e only depends on the local arm $\boldsymbol{a}^e$ of the subset of agents in $\mathcal{D}^e$.

We denote the mean reward of a joint arm as $\mu(\boldsymbol{a}) = \sum_{e=1}^{\rho} \mu^e(\boldsymbol{a}^e)$. For simplicity, we refer to the i^{th} agent by its index i.

The objective is to minimize the expected cumulative regret [1], which is defined as:

$$\mathbb{E}\left[R(T,\pi)\right] \triangleq \mathbb{E}\left[\sum_{t=1}^{T} \mu(\boldsymbol{a}_*) - \mu(\boldsymbol{a}_t) \,\middle|\, \pi\right] \tag{1}$$

where $\boldsymbol{a}_*$ is the optimal joint arm and $\boldsymbol{a}_t$ is the joint arm pulled at time t.

Cumulative regret can be minimized by ignoring the factored structure of the MAMAB, e.g., by using vanilla TS. This leads to a combinatorial problem with respect to the number of agents. Therefore, it is highly beneficial to consider the sparse structure to tractably solve coordination in multi-agent systems.

2 METHODS

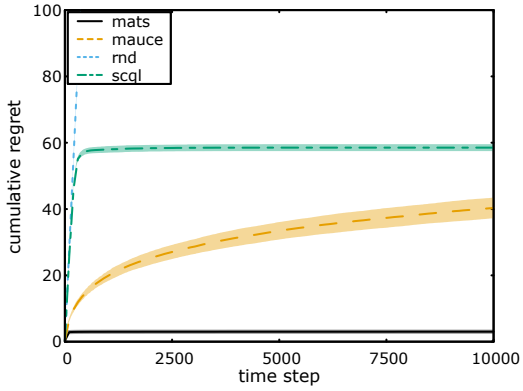
We propose the Multi-Agent Thompson Sampling (MATS) algorithm for decision making in factored multi-agent multi-armed bandit problems. Consider a MAMAB with groups $\mathcal{D}^e$. First, we define a set of priors over the local mean rewards $\mu^e(\boldsymbol{a}^e)$ of each action $\boldsymbol{a}$. Next, at each time step t, MATS draws a sample $\theta_t^e(\boldsymbol{a}^e)$ from the posterior for each group and local arm given the history of all past observations, $\mathcal{H}_{t-1}$. In the case of TS, we choose the arm with the highest sample. However, in our case, as the expected reward is decomposed into several local means, we have to pick the joint action $\boldsymbol{a}$ that maximizes the sum over samples $\theta_t^e(\boldsymbol{a}^e)$. Note that a single agent may have conflicting local optimal arms over the groups it is part of. To this end, we use variable elimination (VE), which computes the maximizing joint action without enumerating over all joint actions [6]. Finally, the joint arm that maximizes the total expected reward is pulled and a reward $f_t^e(\boldsymbol{a}_t^e)$ will be obtained for each group. MATS is formally described in Algorithm 1.

Data: Prior $Q^e_{\boldsymbol{a}^e}$ per group $\mathcal{D}^e$ and local action $\boldsymbol{a}^e$
$\mathcal{H}_0 \leftarrow \{\}$
for $t \in [1..T]$ **do**
  $\forall e \in [1..\rho], \boldsymbol{a}^e \in \mathcal{A}^e$:
    $\theta_t^e(\boldsymbol{a}^e) \sim Q^e_{\boldsymbol{a}^e}(\cdot \mid \mathcal{H}_{t-1})$
  $\boldsymbol{a}_t \leftarrow \arg\max_{\boldsymbol{a}} \sum_{e=1}^{\rho} \theta_t^e(\boldsymbol{a}^e)$ using VE
  $\langle f_t^e(\boldsymbol{a}_t^e)\rangle_{e=1}^{\rho} \leftarrow$ Pull joint arm $\boldsymbol{a}_t$
  $\mathcal{H}_t \leftarrow \mathcal{H}_{t-1} \cup \left\{\langle \boldsymbol{a}_t^e, f_t^e(\boldsymbol{a}_t^e)\rangle_{e=1}^{\rho}\right\}$
end

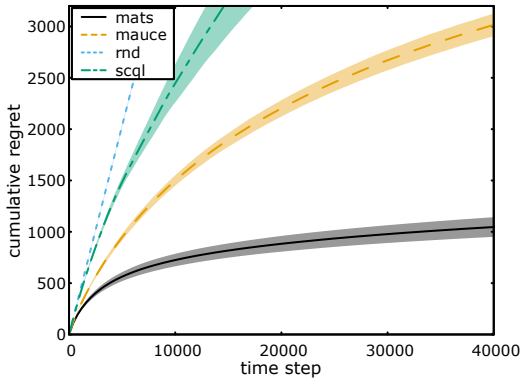
Algorithm 1: MATS

3 RESULTS

We demonstrate the performance of MATS on two benchmark settings, i.e., Bernoulli 0101-Chain and Gem Mining [2]. Bernoulli 0101-Chain is a coordination problem that considers a chain of agents with actions in $\{0, 1\}$, where the optimal joint action is an alternating sequence of zeroes and ones. Gem Mining considers a set of mines and nearby villages, where the village workers are sent to excavate the mines. As villages can be connected to the same mine, coordination between the villages is necessary to optimally allocate workers to mines. In both settings, the rewards are Bernoulli-distributed, per group of agents. We compare MATS against the state-of-the-art methods, MAUCE and SCQL, as well as against a random policy baseline (rnd). For MAUCE and SCQL, we set the exploration parameters to the values previously determined for our experimental settings [2]. For MATS, we use non-informative Jeffreys priors [9] to the Bernoulli likelihoods used in the experimental settings, which lead to Beta posteriors. The results are shown in Figure 1. We observe that in both settings, MATS consistently outperforms MAUCE as well as SCQL. We can see that MATS solves the Bernoulli 0101-Chain problem in only a few time steps, while MAUCE still pulls many sub-optimal actions after 10000 time steps (see Figure 1(a)). In the more challenging Gem Mining problem, the cumulative regret of MAUCE is three times as high as the cumulative regret of MATS around 40000 time steps (see Figure 1(b)).



(a) Bernoulli 0101-Chain



(b) Gem Mining

Figure 1: Cumulative normalized regret averaged over 100 runs for (a) Bernoulli 0101-Chain and (b) Gem Mining. Both the mean (line) and standard deviation (shaded area) are plotted.

We argue that MATS performs well due to the additional information about the reward and prior distributions. In contrast to MAUCE, which has fixed symmetric exploration bounds, MATS can adapt exploration of the arms based on the shape of the posteriors—in this case, a Beta distribution, which can be skewed. Additionally, exploration parameters need to be determined for MAUCE, which are challenging to choose based on prior knowledge about the data. In contrast, MATS uses direct general descriptions of the data, which are often available.

4 ACKNOWLEDGMENTS

The authors would like to acknowledge FWO (Fonds Wetenschappelijk Onderzoek) for their support through the SB grants of Timothy Verstraeten (#1S47617N), Eugenio Bargiacchi (#1SA2820N) and Pieter JK Libin (#1S31916N). Diederik M Roijers was a Postdoctoral Fellow with the FWO (grant #12J0617N). This research was supported by funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme.

REFERENCES

[1] Shipra Agrawal and Navin Goyal. 2013. Further optimal regret bounds for Thompson sampling. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Vol. 31. 99–107.

[2] Eugenio Bargiacchi, Timothy Verstraeten, Diederik M. Roijers, Ann Nowé, and Hado Hasselt. 2018. Learning to Coordinate with Coordination Graphs in Repeated Single-Stage Multi-Agent Decision Problems. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*. 482–490.

[3] Craig Boutilier. 1996. Planning, learning and coordination in multiagent decision processes. In *TARK 1996: Proceedings of the 6th conference on Theoretical aspects of rationality and knowledge*. 195–210.

[4] Olivier Chapelle and Lihong Li. 2011. An empirical evaluation of Thompson sampling. In *Advances in Neural Information Processing Systems (NIPS)*, Vol. 24. 2249–2257.

[5] Daniel Claes, Frans Oliehoek, Hendrik Baier, and Karl Tuyls. 2017. Decentralised Online Planning for Multi-Robot Warehouse Commissioning. In *Proceedings of the 16th Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*. International Foundation for Autonomous Agents and Multiagent Systems, 492–500.

[6] Carlos Guestrin, Daphne Koller, and Ronald Parr. 2001. Multiagent Planning with Factored MDPs. In *Advances in Neural Information Processing Systems (NIPS)*, Vol. 14. 1523–1530.

[7] Pieter J.K. Libin, Timothy Verstraeten, Diederik M. Roijers, Jelena Grujic, Kristof Theys, Philippe Lemey, and Ann Nowé. 2018. Bayesian Best-Arm Identification for Selecting Influenza Mitigation Strategies. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)*. Springer, 456–471.

[8] Pieter J.K. Libin, Timothy Verstraeten, Diederik M. Roijers, Wenjia Wang, Kristof Theys, and Ann NowÃĬ. 2019. Thompson Sampling for m-top Exploration. In *Proceedings of the IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*. 1414–1420.

[9] David Lunn, Chris Jackson, Nicky Best, David Spiegelhalter, and Andrew Thomas. 2012. *The BUGS book: A practical introduction to Bayesian analysis*. Chapman and Hall/CRC.

[10] William R. Thompson. 1933. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 25, 3/4 (1933), 285–294.

[11] Timothy Verstraeten, Ann Nowé, Jonathan Keller, Yi Guo, Shuangwen Sheng, and Jan Helsen. 2019. Fleetwide data-enabled reliability improvement of wind turbines. *Renewable and Sustainable Energy Reviews* 109 (2019), 428–437.

[12] Marco Wiering. 2000. Multi-agent reinforcement learning for traffic light control. In *Proceedings of the 17th International Conference on Machine Learning (ICML)*. 1151–1158.