A* Algorithms for Dec-POMDPs*

Wietze Koops
wietze.koops@ru.nl

Radboud University, Nijmegen, The Netherlands

Introduction. Partially observable Markov decision processes (POMDPs) [6] formalize sequential decision making in a stochastic environment, where the decision maker (agent) cannot fully observe the state of the environment. Decentralized POMDPs (Dec-POMDPs) [9] generalize this to a multi-agent setting, where each agent locally decides which action to take, and each agent receives local observations of the state. As such, Dec-POMDPs provide a way to model settings where multiple agents cooperate to achieve a common goal, but where communication is costly or lossy. Applications include multi-UAV search [16], wireless sensor networks [15], bandwidth allocation [5], and maintenance problems [2].

The decision problem underlying solving Dec-POMDPs exactly or ϵ-optimally over a finite horizon is NEXP-hard [1,13]. Nevertheless, in many practical cases it is possible to exploit structure to do better than these complexity results suggest. Concretely, in this thesis we present three algorithms building on multi-agent A* (MAA*) [14,10], an algorithm that finds policies by exploring a search tree. The first algorithm, recursive small-step multi-agent A* (RS-MAA*), is an exact algorithm. The second algorithm, policy-finding multi-agent A* (PF-MAA*), is designed to find good policies for high horizons fast. The third algorithm, terminal reward multi-agent A* (TR-MAA*), aims to find upper bounds on the optimal value for high horizons.

Our experiments show excellent performance of all three algorithms on a wide range of standard benchmarks. We extend the horizon for which exact solving is possible on all hard benchmarks[1]. In addition, PF-MAA* finds superior policies for several benchmarks for high horizons, while TR-MAA* certifies the near-optimality of these policies by finding close upper bounds.

Small-step multi-agent A.* Classical multi-agent A* (MAA*) [14] uses a search tree whose outdegree is double exponential in the stage t. In particular, the running time of classical MAA* is always double exponential in the horizon h.

Our first insight is that this double exponential outdegree can be avoided by only fixing one action of one agent at once. This leads to a constant outdegree, at the cost of increasing the height of the search tree from linear to exponential in the horizon h. We call this novel search tree the small-step search tree, and the corresponding algorithm small-step multi-agent A*. Small-step MAA* is the basis for all three of our algorithms.

---

* Abstract of a MSc thesis supervised by Sebastian Junges and Nils Jansen. Parts of the thesis were presented earlier at IJCAI 2023 [7] and IJCAI 2024 [8].

[1] All benchmarks for which existing algorithms did not already scale to horizon 500.

RS-MAA.* Recursive small-step multi-agent A* (RS-MAA*) is our exact algorithm. Besides the small-step search tree, the main ingredient for RS-MAA* is the use of tight, recursive heuristics. In the context of A*, heuristics should overestimate the value of the best policy among a group of policies. Existing heuristics in the literature [11] overestimate the value by assuming that agents can use all joint observations or all but the last joint observation to decide how to act, rather than only their own local observations. Instead, our new heuristics correspond to revealing a small, fixed number of joint observations. We solve the corresponding problem by recursively solving Dec-POMDPs with a smaller horizon. To make this more scalable, we exploit the anytime nature of A* algorithms, which allows for returning an upper bound on the heuristic value at any point. In addition, we reuse lossless clustering [12], which allows grouping together local observation histories without loss in policy value.

PF-MAA.* Policy-finding multi-agent A* (PF-MAA*) is our policy finding algorithm. A challenge that PF-MAA* must address is that the size of policies, when presented as functions of the local observation history, grows exponentially with the horizon. To limit the size of the policies in our search space, we consider policies that only use the last k local observations to decide how to act. Our first contribution for PF-MAA* is the development of a lossless clustering of these windows of k observations. Secondly, to limit the time spent searching, we only expand the most promising partial policies, and prune other partial policies. Thirdly, we introduce novel heuristics, designed to guide the search towards the best policies. These heuristics use a precise estimate for the near future using a low-horizon Dec-POMDP, but a rough estimate for the further future.

TR-MAA.* Finally, terminal reward multi-agent A* (TR-MAA*) is our algorithm that finds upper bounds for high horizons. TR-MAA* also reuses lossless clustering [12]. However, the recursive heuristics used by RS-MAA* are often too expensive for high horizons (due to the curse of history). To make the heuristic more scalable, we regularly reveal the true state of the Dec-POMDP.

Experiments. Table 1 shows some empirical results. The first two columns show the highest horizon that we can solve exactly, compared to best result in the literature from GMAA*-ICE [10]. The next two columns show the value of the best policy we found, compared to the best literature values found by FB-HSVI [3] or GA-FSC [4]. The final two columns compare our upper bound to the MDP value, which is the only bound available in the literature scaling to high horizons.

Table 1. Empirical results (time limit: 30 minutes, memory limit: 16GB).

Setting	Exact (max. h)		Policy value ($h = 50$)		Upper bound ($h = 50$)	
Problem	Ours	lit.	Ours	lit.	Ours	MDP
DecTiger	12	6	81.0	80.7	101.3	1000
Grid	7	6	37.5	40.5	47.3	48.8
BoxPushing	5	4	1210	1201	1212	1306
Mars	9	4	125.8	128.9	132.4	145.0
Grid3x3	7	5	44.35	44.32	44.38	44.62

References

1. Bernstein, D.S., Givan, R., Immerman, N., Zilberstein, S.: The complexity of decentralized control of Markov decision processes. Mathematics of operations research **27**(4), 819–840 (2002)
2. Bhustali, P., Andriotis, C.P.: Assessing the optimality of decentralized inspection and maintenance policies for stochastically degrading engineering systems. In: BNAIC/BeNeLearn 2023: Joint International Scientific Conferences on AI and Machine Learning (2023)
3. Dibangoye, J.S., Amato, C., Buffet, O., Charpillet, F.: Optimally solving Dec-POMDPs as continuous-state MDPs: Theory and algorithms. Tech. rep., Tech. Rep. RR-8517, Inria (2014)
4. Eker, B., Akin, H.L.: Solving decentralized POMDP problems using genetic algorithms. AAMAS **27**(1), 161–196 (2013)
5. Hemmati, M., Yassine, A., Shirmohammadi, S.: A Dec-POMDP model for congestion avoidance and fair allocation of network bandwidth in rate-adaptive video streaming. In: SSCI. pp. 1182–1189. IEEE (2015)
6. Kaelbling, L.P., Littman, M.L., Cassandra, A.R.: Planning and Acting in Partially Observable Stochastic Domains. Artif. Intell. **101**(1-2), 99–134 (1998)
7. Koops, W., Jansen, N., Junges, S., Simão, T.D.: Recursive Small-Step Multi-Agent A* for Dec-POMDPs. In: IJCAI. pp. 5402–5410 (2023)
8. Koops, W., Junges, S., Jansen, N.: Approximate Dec-POMDP solving using Multi-Agent A*. In: IJCAI. pp. 6743–6751 (2024)
9. Oliehoek, F.A., Amato, C.: A Concise Introduction to Decentralized POMDPs. Briefs in Intelligent Systems, Springer (2016)
10. Oliehoek, F.A., Spaan, M.T.J., Amato, C., Whiteson, S.: Incremental clustering and expansion for faster optimal planning in Dec-POMDPs. JAIR **46**, 449–509 (2013)
11. Oliehoek, F.A., Spaan, M.T.J., Vlassis, N.: Optimal and approximate Q-value functions for decentralized POMDPs. JAIR **32**, 289–353 (2008)
12. Oliehoek, F.A., Whiteson, S., Spaan, M.T.J.: Lossless clustering of histories in decentralized POMDPs. In: AAMAS. pp. 577–584 (2009)
13. Rabinovich, Z., Goldman, C.V., Rosenschein, J.S.: The complexity of multiagent systems: the price of silence. In: AAMAS. pp. 1102–1103 (2003)
14. Szer, D., Charpillet, F., Zilberstein, S.: MAA*: A heuristic search algorithm for solving decentralized POMDPs. In: UAI (2005)
15. Zhang, J., Xu, L., Zhou, S., Ye, X.: A novel sleep scheduling scheme in green wireless sensor networks. J. Supercomput. **71**(3), 1067–1094 (2015)
16. Zhu, X., Vanegas, F., Gonzalez, F., Sanderson, C.: A multi-UAV system for exploration and target finding in cluttered and GPS-denied environments. In: International Conference on Unmanned Aircraft Systems. pp. 721–729. IEEE (2021)