Scaling Cooperative Online Planning under Partial Observability for Many Agents⋆

Maris F.L. Galesloot[0009−0002−5112−8584]

Radboud University, Nijmegen, The Netherlands
maris.galesloot@ru.nl

Introduction. Partially observable Markov decision processes (POMDPs, [6]) are a framework for sequential decision making, able to model many problems [8, 12, 21]. However, this model is intractable in general [11]. Online planning is a practical approach to tackle POMDPs under limited computational and time budgets, focusing the available resources on the reachable and most promising parts of the solution space [9]. These methods compute approximations of the values of actions and distributions over the states of the system, where the latter are known as *beliefs* [6]. In particular, POMCP is a simple yet efficient online planning algorithm that achieves good performance in large POMDPs [17]. Problems with multiple agents, such as teams of mobile robots or autonomous surveillance systems, can be modelled by multi-agent POMDPs (MPOMDPs, [13]) by assuming noiseless free communication [14]. A particular challenge that makes solving MPOMDPs even harder than POMDPs is the combinatorial number of actions and observations that grow exponentially with the number of agents [15]. This increased complexity makes a naive full-width application of online planning algorithms ineffective as the reachable solution space increases drastically.
To mitigate this issue, we can exploit the locality of interactions between the agents, often captured by so-called *coordination graphs* (CG, [5]). In particular, by estimating the action value for subsets of agents instead of all agents based on such graphs [1]. The main concepts are to factorise the value estimates over the action space of subsets of agents in the *factored statistics* (FS-POMCP) variant and, additionally, to factorise the observation space in the *factored trees* (FT-POMCP) variant. However, this does not directly address the issue of scaling the belief-state estimation when many agents are involved. Additionally, it complicates the selection of actions as all local combinations must be considered. Therefore, to develop scalable methods, we must exploit the given structure as much as possible. In this work, we investigate *how to **scale online MPOMDP planning** when **many agents** are involved.* Furthermore, we study *static graphs as heuristics to problems where agents move and coordinate dynamically.*

Contributions. In this thesis, we *i)* introduce new algorithm variants equipped for achieving high returns in large MPOMDPs; *ii)* address the scaling issues caused by (a) large observation spaces and (b) dense cyclic CGs; and *iii)* evaluate various algorithm combinations empirically on a set of diverse benchmarks.

⋆ Abstract of an MSc thesis, supervised by Dr. Nils Jansen, Dr. Sebastian Junges, and Dr. Thiago D. Simão.

Algorithm variants. We introduce algorithms based on the so-called *particle filter tree* (Sparse-PFT, [10]) that searches over approximations of the belief. Instead of full-width expansion of the joint action space, we generalise over local action spaces by maintaining sets of local statistics (FS-PFT) or building separate trees (FT-PFT), for each combination of agents given by the CG [1].

Environment	Multi-Agent RockSample			
Nr. of Agents	3	4	5	6
FS-PFT	-5.8 ± 1.3	-5.2 ± 1.2	**4.2** ± 0.9	2.6 ± 0.6
FS-W-POMCP	-2.9 ± 0.8	0.5 ± 0.5	0.1 ± 0.8	**6.9** ± 1.1
FS-POMCPOW	-2.9 ± 0.9	**4.9** ± 0.8	**3.3** ± 1.1	0.4 ± 1.4
FT-W-POMCP	1.7 ± 0.6	**3.6** ± 0.7	-0.2 ± 0.4	-1.5 ± 0.6
W-POMCP	**8.4** ± 1.3	-1.5 ± 1.1	0.0 ± 0.0	OOR

Fig. 1. Best-performing planning algorithms on MARS. Factored algorithms use *variable elimination*, which performed best.

Belief estimation. We extend (factored) POMCP variants to incorporate *weighted particle filtering* in (factored) W-POMCP [20]. In these variants, the belief nodes are enriched with the addition of observation probabilities [19]. In FS-POMCPOW, we also gradually increase the number of allowed expansions of action and belief nodes by *double progressive widening* [4, 19]. For FT variants, we propose an ensemble of belief approximations that consider local observation probabilities. We sample from the ensemble proportional to the likelihood of each filter. The likelihood is a statistic that reflects if said filter likely contains particles that generated the true observation [7]. We employ this method in FT-W-POMCP and FT-PFT. FT-POMCP uses an unweighted variant.

Eliminating cycles. Both action selection methods, *variable elimination* and *maxplus*, endure a high complexity in cyclic graphs with dense coordination structures and many agents. We consider only *current maximal* possible contribution of the local value predictions. We extract a *maximum spanning tree* (MST) based on the maximal possible contribution of each edge to the maximisation. The error introduced is bounded by the sum of weights of the removed edges [16].

Results. We show empirically that the MST extension is essential in settings with dense cyclic coordination (fig. 2). Furthermore, our algorithm variants are the best-performing on multi-agent RockSample (MARS) [18, 2]. On this benchmark (fig. 1), we can also see the positive effect of a static graph as a heuristic for a problem with dynamic coordination, resulting in adequate performance. Furthermore, the results affirm that value decomposition improves planning performance even when the problem is not neatly factored [3].

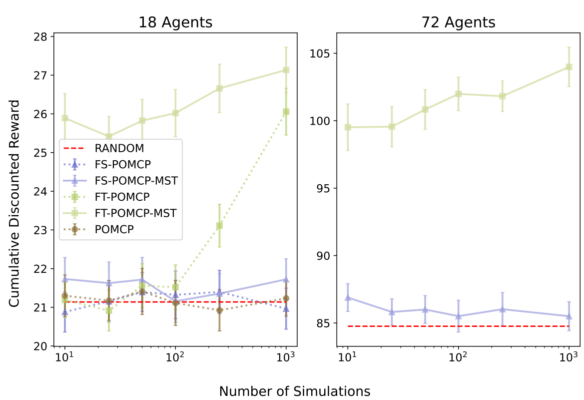

Fig. 2. Returns across the number of agents and simulations on SysAdmin using unweighted belief estimation and *variable elimination* with (solid) and without (dotted) our MST extension. Variants without the MST ran out of memory in the 72-agent setting.

Conclusion. Our extensions to existing online planning algorithms tackle many-agent MPOMDPs efficiently, achieving high performance. Future work consists of learning factored value estimates offline and finding a suitable graph structure algorithmically without prior knowledge of the topology.

References

1. Amato, C., Oliehoek, F.A.: Scalable planning and learning for multi-agent POMDPs. In: Bonet, B., Koenig, S. (eds.) Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA. pp. 1995–2002. AAAI Press (2015), http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9889
2. Cai, P., Luo, Y., Hsu, D., Lee, W.S.: Hyp-despot: A hybrid parallel algorithm for online planning under uncertainty. Int. J. Robotics Res. **40**(2-3) (2021). https://doi.org/10.1177/0278364920937074, https://doi.org/10.1177/0278364920937074
3. Castellini, J., Oliehoek, F.A., Savani, R., Whiteson, S.: Analysing factorizations of action-value networks for cooperative multi-agent reinforcement learning. Auton. Agents Multi Agent Syst. **35**(2), 25 (2021). https://doi.org/10.1007/s10458-021-09506-w, https://doi.org/10.1007/s10458-021-09506-w
4. Couetoux, A., Doghmen, H.: Adding double progressive widening to upper confidence trees to cope with uncertainty in planning problems. In: The 9th European Workshop on Reinforcement Learning (EWRL-9) (2011)
5. Guestrin, C., Venkataraman, S., Koller, D.: Context-specific multiagent coordination and planning with factored MDPs. In: Dechter, R., Kearns, M.J., Sutton, R.S. (eds.) Proceedings of the Eighteenth National Conference on Artificial Intelligence and Fourteenth Conference on Innovative Applications of Artificial Intelligence, July 28 - August 1, 2002, Edmonton, Alberta, Canada. pp. 253–259. AAAI Press / The MIT Press (2002), http://www.aaai.org/Library/AAAI/2002/aaai02-039.php
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). https://doi.org/10.1016/S0004-3702(98)00023-X, https://doi.org/10.1016/S0004-3702(98)00023-X
7. Katt, S., Oliehoek, F.A., Amato, C.: Bayesian reinforcement learning in factored POMDPs. In: Elkind, E., Veloso, M., Agmon, N., Taylor, M.E. (eds.) Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19, Montreal, QC, Canada, May 13-17, 2019. pp. 7–15. International Foundation for Autonomous Agents and Multiagent Systems (2019), http://dl.acm.org/citation.cfm?id=3331668
8. Kochenderfer, M., Amato, C., Chowdhary, G., How, J., Reynolds, H.: Decision Making Under Uncertainty: Theory and Application. MIT Lincoln Laboratory Series, MIT Press (2015), https://books.google.nl/books?id=hUBWCgAAQBAJ
9. Kocsis, L., Szepesvári, C.: Bandit based Monte-Carlo planning. In: ECML. Lecture Notes in Computer Science, vol. 4212, pp. 282–293. Springer (2006)
10. Lim, M.H., Becker, T.J., Kochenderfer, M.J., Tomlin, C.J., Sunberg, Z.N.: Optimality guarantees for particle belief approximation of POMDPs (2023)
11. Madani, O., Hanks, S., Condon, A.: On the undecidability of probabilistic planning and related stochastic optimization problems. Artif. Intell. **147**(1-2), 5–34 (2003)
12. Memarzadeh, M., Boettiger, C.: Adaptive management of ecological systems under partial observability. Biological Conservation **224**, 9–15 (2018)
13. Messias, J.V., Spaan, M.T.J., Lima, P.U.: Efficient offline communication policies for factored multiagent POMDPs. In: NIPS. pp. 1917–1925 (2011)
14. Oliehoek, F.A., Amato, C.: A Concise Introduction to Decentralized POMDPs. Springer Briefs in Intelligent Systems, Springer (2016). https://doi.org/10.1007/978-3-319-28929-8, https://doi.org/10.1007/978-3-319-28929-8

15. Pynadath, D.V., Tambe, M.: The communicative multiagent team decision problem: Analyzing teamwork theories and models. J. Artif. Intell. Res. **16**, 389–423 (2002). https://doi.org/10.1613/jair.1024, https://doi.org/10.1613/jair.1024
16. Rogers, A., Farinelli, A., Stranders, R., Jennings, N.R.: Bounded approximate decentralised coordination via the max-sum algorithm. Artif. Intell. **175**(2), 730–759 (2011). https://doi.org/10.1016/j.artint.2010.11.001, https://doi.org/10.1016/j.artint.2010.11.001
17. Silver, D., Veness, J.: Monte-carlo planning in large POMDPs. In: NIPS. pp. 2164–2172. Curran Associates, Inc. (2010)
18. Smith, T., Simmons, R.G.: Heuristic search value iteration for POMDPs. In: Chickering, D.M., Halpern, J.Y. (eds.) UAI '04, Proceedings of the 20th Conference in Uncertainty in Artificial Intelligence, Banff, Canada, July 7-11, 2004. pp. 520–527. AUAI Press (2004), https://dslpitt.org/uai/displayArticleDetails.jsp?mmnu=1&smnu=2&article_id=1150&proceeding_id=20
19. Sunberg, Z.N., Kochenderfer, M.J.: Online algorithms for POMDPs with continuous state, action, and observation spaces. In: de Weerdt, M., Koenig, S., Röger, G., Spaan, M.T.J. (eds.) Proceedings of the Twenty-Eighth International Conference on Automated Planning and Scheduling, ICAPS 2018, Delft, The Netherlands, June 24-29, 2018. pp. 259–263. AAAI Press (2018), https://aaai.org/ocs/index.php/ICAPS/ICAPS18/paper/view/17734
20. Thrun, S., Burgard, W., Fox, D.: Probabilistic robotics. Intelligent robotics and autonomous agents, MIT Press (2005)
21. Ulbrich, S., Maurer, M.: Probabilistic online POMDP decision making for lane changes in fully automated driving. In: ITSC. pp. 2063–2067. IEEE (2013)