BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2023
http://excel.fit.vutbr.cz

Search and Explore: Symbiotic Policy Synthesis in POMDPs

Bc. Filip Macák* Supervisor: Assoc. Prof. Milan Češka Accepted to CAV'23 (Core A*) [1]

Abstract
This work focuses on the synthesis of finite-state controllers (FSCs) for partially observable Markov decision processes (POMDPs). We consider POMDPs with indefinite-horizon specifications i.e. our focus is on long-term planning. The current state-of-the-art solutions for this problem show their weaknesses even on simple POMDPs. We propose an anytime algorithm that tightly integrates the inductive synthesis of FSCs with the belief exploration. We observe significant improvements in the quality of the produced FSCs, and the synthesis run-time. Finally, we achieve significantly lower memory consumption and make important observations on the size of the FSCs. With our results, we strengthen the position of formal methods for the POMDP synthesis problem which emerges in AI, robotics and even software verification.

*xmacak07@stud.fit.vutbr.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

Partially observable Markov decision processes (POMDPs) [2] provide an important model for sequential decision-making under uncertainty and limited state observability. They are widely used in many areas, including AI, robotics, and software verification. The synthesis problem for POMDPs asks to find a policy (i.e. strategy) ensuring the given specification. In this paper, we focus on indefinite-horizon specifications (i.e. time-unbounded and undiscounted) that are important for long-term planning. The problem of finding the optimal policy for these types of specifications is undecidable and thus approximate solutions are needed [3]. We focus on the synthesis of finite-state controllers (FSCs) providing compact, easy-to-use and interpretable polices [4].

Belief-based methods emerged as one of the state-of-the-art approaches. They build on the exploration of the belief space describing probability distributions over POMDP states. As the belief space might be huge or possibly infinite, an approximation of the unexplored belief space is required. Cut-offs [5], implemented in the tool called Storm [6], represent a recent approximation technique supporting indefinite-horizon specifications. Various point-based approximations, notably the SARSOP [7] algorithm, are very successful for short-term planning (i.e. discounted specifications), but perform poorly for long-term decision-making.

Recently, an approach based on **inductive synthesis of FSCs** has been introduced [8] and implemented in the tool Paynt [9]. The method searches for optimal FSCs in iteratively increasing families of candidate FSCs. This approach is able to find competitive FSCs, but struggles to handle large POMDPs or synthesis problems that require FSC to use a lot of memory.

Alternatively, various simulation-based and reinforcement learning methods [10] can be used to plan in large or unknown POMDPs. However, these approaches are data-intensive and do not provide safety and performance guarantees.

In this work, we propose a **symbiotic algorithm combining the strengths of inductive synthesis and belief exploration**. It builds on two novel ideas: i) FSCs obtained from the inductive synthesis can improve the approximation of the belief-space, ii) policies obtained from the belief-space exploration can improve the inductive search strategy. The proposed algorithm is able to outperform state-of-the-art methods on a wide range of benchmark models from both AI and formal method communities.

2. Symbiotic Synthesis Algorithm

Our ideas rely on the fact that a policy found via one approach can boost the other approach. The key observation is that such a policy is beneficial even when it is sub-optimal in terms of the given objective.

As demonstrated in Figure 1, the FSCs found by the inductive synthesis are used as cut-offs approximating

the unexplored belief space. This corresponds to executing the policy represented by the given FSC from frontier beliefs (i.e. unexplored beliefs reachable from the explored ones in one step). Such FSCs provide in many cases significantly better approximation than default cut-offs.

Inductive synthesis uses information from the fully observable (MDP) policies to steer the search. We propose using reference policies obtained from belief-based methods to improve the search strategy. In particular, we prioritise search towards a subfamily considering only actions proposed by the reference policy. Moreover, we use the information from the reference policy to infer observations where adding memory might be beneficial.

We experimentally showed that the proposed ideas improve both of these approaches individually. The natural step is to use the improved inductive synthesis FSCs for belief exploration and the improved belief exploration to further improve inductive synthesis i.e. to alternate between the approaches by closing the synthesis loop. The workflow and architecture of Saynt, the proposed symbiotic algorithm, is illustrated in Figure 2. Although the proposed integration seems straightforward, various technical obstacles had to be addressed, e.g., obtaining a compact controller from the finite approximation of the belief space, figuring out how best to use the reference policies for inductive search, and developing an interplay between the exploration and search phases that minimises the overhead.

Saynt works iteratively: in each iteration, it performs the inductive search and the belief exploration, both with a given timeout. It starts in the inductive mode, where it searches for the best memoryless FSCs $F_{\mathcal{I}}$. Afterwards, it runs the belief exploration and uses $F_{\mathcal{I}}$ as a cut-off in frontier beliefs. The result is a finite belief approximation that can be analysed with off-the-shelf MDP model checkers, yielding FSC $F_{\mathcal{B}}$ that selects for each explored belief the best action and executes $F_{\mathcal{I}}$ in frontier beliefs. FSC $F_{\mathcal{B}}$ is then used to guide the search during the subsequent inductive phase, where non-memoryless FSCs are considered. First, $F_{\mathcal{B}}$ is used to deduce the proper amount of memory to consider in candidate FSCs. Second, we prioritise the inductive search in a subfamily of FSCs that consider actions suggested by $F_{\mathcal{B}}$.

Saynt is an anytime synthesis algorithm that produces in each iteration two corresponding FSCs $F_{\mathcal{I}}$ and $F_{\mathcal{B}}$. This brings another advantage as it gives the user a choice between a very compact $F_{\mathcal{I}}$ and a slightly better, albeit much larger, $F_{\mathcal{B}}$.

3. Experimental Evaluation

We compare Saynt with the two state-of-the-art synthesis algorithms for POMDPs wrt. indefinite-horizon specifications, namely, with Storm [5] and Paynt [8]. We consider a wide range of benchmark models from AI and formal method communities.

First, we focus on comparing the quality of the resulting FSCs and the time needed to compute them. Figure 3 shows how the quality of the controllers improves over time for selected models. We observe that Saynt steadily outperforms both baselines in both quality and speed, as demonstrated in the plots. In some cases, we are able to achieve improvements in quality of up to 40%. We note that the distance to the (unknown) optimal values remains unclear. From the plots, we see that the benefits of Saynt are independent of which of the methods performs better. We also observe that the achieved improvement grows for larger and more complex POMDPs. When it comes to the run-times, the individual algorithms are sometimes faster until the first iteration of Saynt finishes. After that, Saynt typically provides better FSCs in a shorter time.

We also investigate the memory usage of the considered algorithms (see the last figure). The memory footprint of the inductive synthesis is insignificant and the memory usage of Saynt is heavily dominated by the belief exploration. Saynt reduces the memory footprint of Storm by a factor of 4. In fact, Storm usually used all of the available memory (64GB) before the 15-minute timeout was reached. Hence, the proposed integration enables an efficient belief space exploration for larger POMDPs.

4. Conclusions

We proposed Saynt, a symbiotic integration of the two main approaches for controller synthesis in POMDPs. Saynt substantially improves the value of the resulting controllers and provides an anytime, push-button synthesis algorithm allowing users to select the controller based on the trade-off between value, size, and the synthesis time.

Author's Contribution

Filip Macák significantly contributed to the formulation of the research ideas, designing methodological approaches as well as to their implementation and experimental evaluation. He also contributed to the writing of [1]. He would like to thank to the co-authors of [1] for their help with this research.

References

[1] Roman Andriushchenko, Milan Češka, Sebastian Junges, Joost-Pieter Katoen, and Filip Macák. Search and explore: Symbiotic policy synthesis in pomdps. Accepted to CAV'23.

[2] Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artif. Intell.*, 101(1-2):99–134, 1998.

[3] Omid Madani, Steve Hanks, and Anne Condon. On the undecidability of probabilistic planning and related stochastic optimization problems. *Artificial Intelligence*, 147(1):5–34, 2003.

[4] Blai Bonet, Hector Palacios, and Hector Geffner. Automatic derivation of finite-state machines for behavior control. In *AAAI*, 2010.

[5] Alexander Bork, Joost-Pieter Katoen, and Tim Quatmann. Under-approximating expected total rewards in POMDPs. In *TACAS (2)*, volume 13244 of *LNCS*, pages 22–40. Springer, 2022.

[6] Christian Dehnert, Sebastian Junges, Joost-Pieter Katoen, and Matthias Volk. A Storm is coming: A modern probabilistic model checker. In *CAV*, volume 10427 of *LNCS*, pages 592–600. Springer, 2017.

[7] Hanna Kurniawati, David Hsu, and Wee Sun Lee. SARSOP: Efficient point-based POMDP planning by approximating optimally reachable belief spaces. In *Robotics: Science and Systems*. MIT Press, 2008.

[8] Roman Andriushchenko, Milan Češka, Sebastian Junges, and Joost-Pieter Katoen. Inductive synthesis of finite-state controllers for POMDPs. In *UAI*, volume 180, pages 85–95. PMRL, 2022.

[9] Roman Andriushchenko, Milan Češka, Sebastian Junges, Joost-Pieter Katoen, and Šimon Stupinský. PAYNT: a tool for inductive synthesis of probabilistic programs. In *CAV*, volume 12759 of *LNCS*, pages 856–869. Springer, 2021.

[10] Julian Schrittwieser et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, dec 2020.