Search and Explore: Symbiotic Policy Synthesis in POMDPs

Bc. Filip Macák

Supervisor: Assoc. Prof. Milan Češka Consultant: Ing. Roman Andriushchenko

Accepted to CORE A* conference CAV'23 [1]

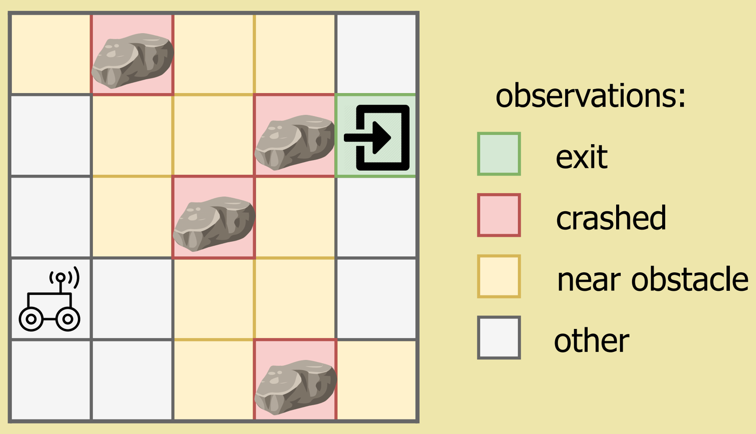
Excel@FIT 2023

BRNO UNIVERSITY OF TECHNOLOGY FACULTY OF INFORMATION TECHNOLOGY

Problem Formulation

Partially observable Markov decision processes (POMDPs)

- important model for sequential decision-making under uncertainty and limited observability
- widely used in many areas including AI, robotics, or software verification

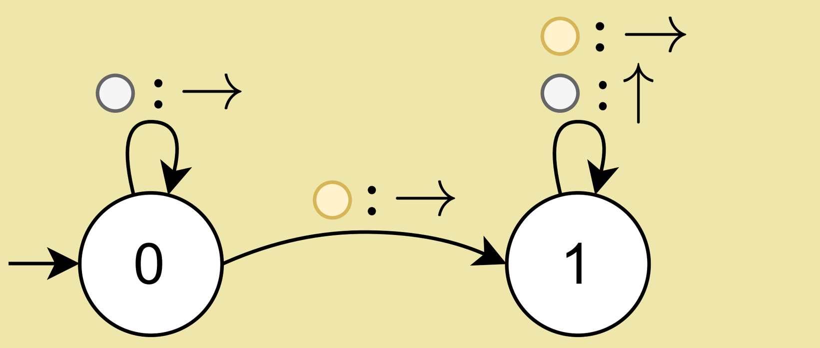


Robot's specification:

- minimise the number of steps to reach the exit
- keep the probability of crashing below 1%

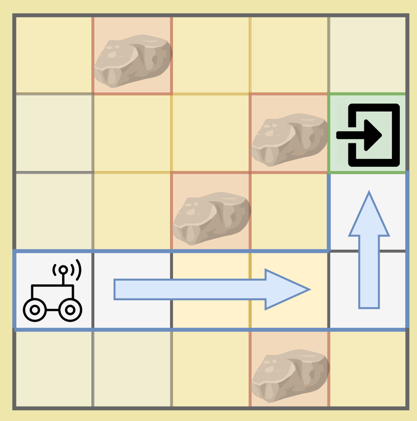
Offline synthesis problem: find the optimal policy (i.e. strategy) for the given specification.

- **indefinite-horizon** - no discounts, long-term goals, finding optimal policy is **undecidable**
- focus on **small, easy-to-execute and interpretable policies**
- we seek for optimal **finite-state controllers (FSCs)** within the given memory bounds



FSC picks the optimal action and updates memory based on current observation and memory node.

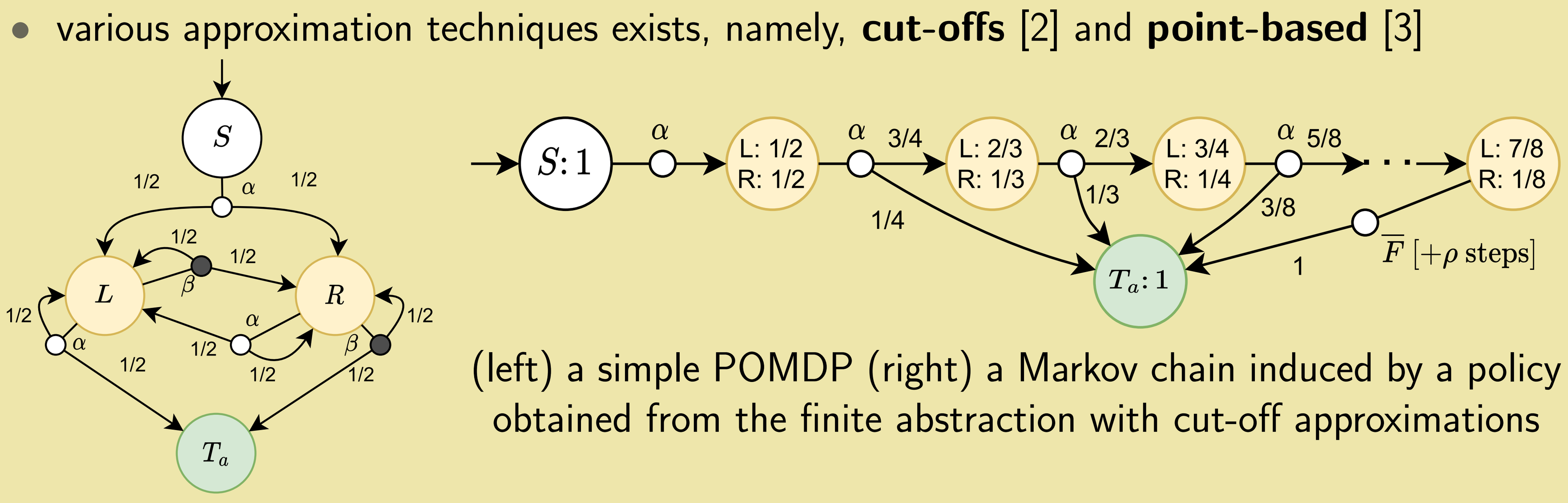
Execute FSC ⟹



State-of-the-Art Methods and Their Limitations

Belief-based methods

- beliefs: probability distribution over states of a POMDP
- construct and analyse the reachable belief space, which might be huge/infinite
- various approximation techniques exists, namely, **cut-offs** [2] and **point-based** [3]



(left) a simple POMDP (right) a Markov chain induced by a policy obtained from the finite abstraction with cut-off approximations

- **limitations:**
  - existing cut-offs (implemented in the tool STORM [2]) are not sufficient – even some small POMDPs may require to explore a large belief space, leading to a poor performance
  - point-based methods, notably SARSOP [3], perform poorly for long-term planning, i.e. when a high discount factor is needed

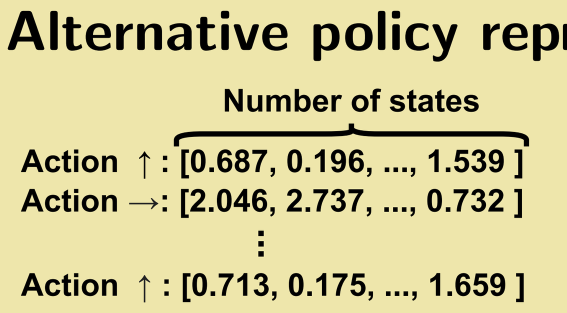
Inductive synthesis of FSCs

- inductive exploration of the family of candidate FSCs using fully-observable abstraction and counter-examples
- iterative expansion of the family by adding memory to suitable observations
- implemented in the tool PAYNT [4]
- **limitations:**
  - for large POMDPs, the family size is huge and its exploration is expensive
  - the family size grows exponentially with the memory added to FSC – if a lot of memory is needed, exploration becomes computationally intractable

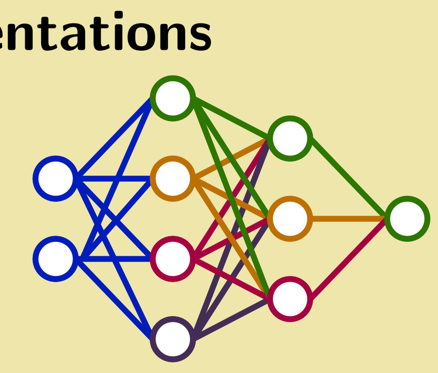
Simulation-based and reinforcement learning methods [5]

- aim at problems where the underlying POMDP is not known or is prohibitively large – require interaction with the environment (simulations) and are typically data-intensive
- typically used in online planning: in the given time choose the best action for the current state – the cost and efficiency of sampling limits the performance

Alternative policy representations



Set of alpha-vectors



Neural network

Difficult to interpret, execute, and verify – problematic in safety-critical applications

Key Contribution

Symbiotic integration of the belief-space exploration and the inductive synthesis that scales for larger POMDPs while providing safe, easy-to-use and interpretable controllers. This work strengthens the position of formal methods for the POMDP synthesis problem.

Integrating Inductive Synthesis and Belief Exploration

Builds on the two novel ideas

- use the FSCs obtained from inductive synthesis to improve the cut-offs in the belief-space
- use policies obtained from the explored belief space to accelerate the inductive search

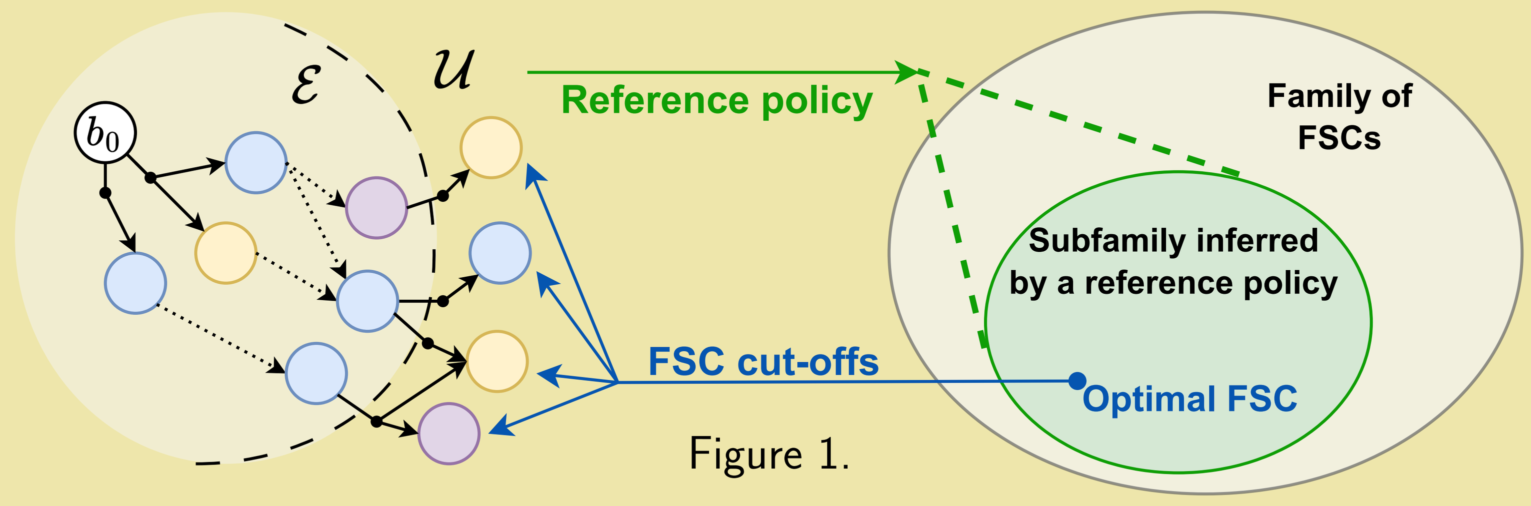


Figure 1.

A novel symbiotic synthesis algorithm Saynt

- closing the integration loop between STORM and PAYNT
- STORM provides reference policies for PAYNT, PAYNT provides cut-off FSCs for STORM
- iterative anytime synthesis algorithm – in each iteration two FSCs $F_{\mathcal{B}}$ and $F_{\mathcal{I}}$ are obtained

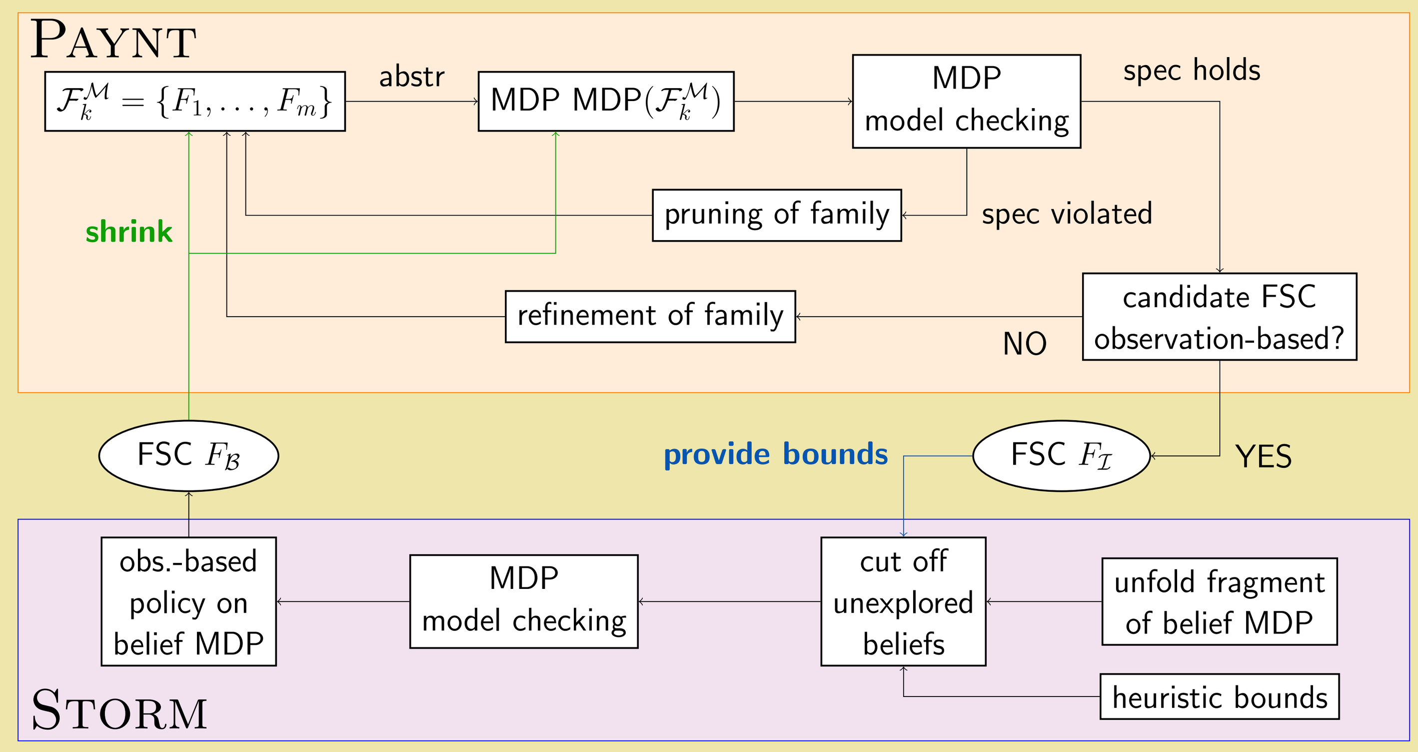


Figure 2.

Experiments

- performance of SAYNT is compared to STORM [2] and PAYNT [4], state-of-the-art tools for offline synthesis of FSCs for POMDPs with indefinite-horizon specifications
- **wide range of benchmark models** from AI and formal methods communities

The graphs shows how the quality of the controllers improves over time for selected models:

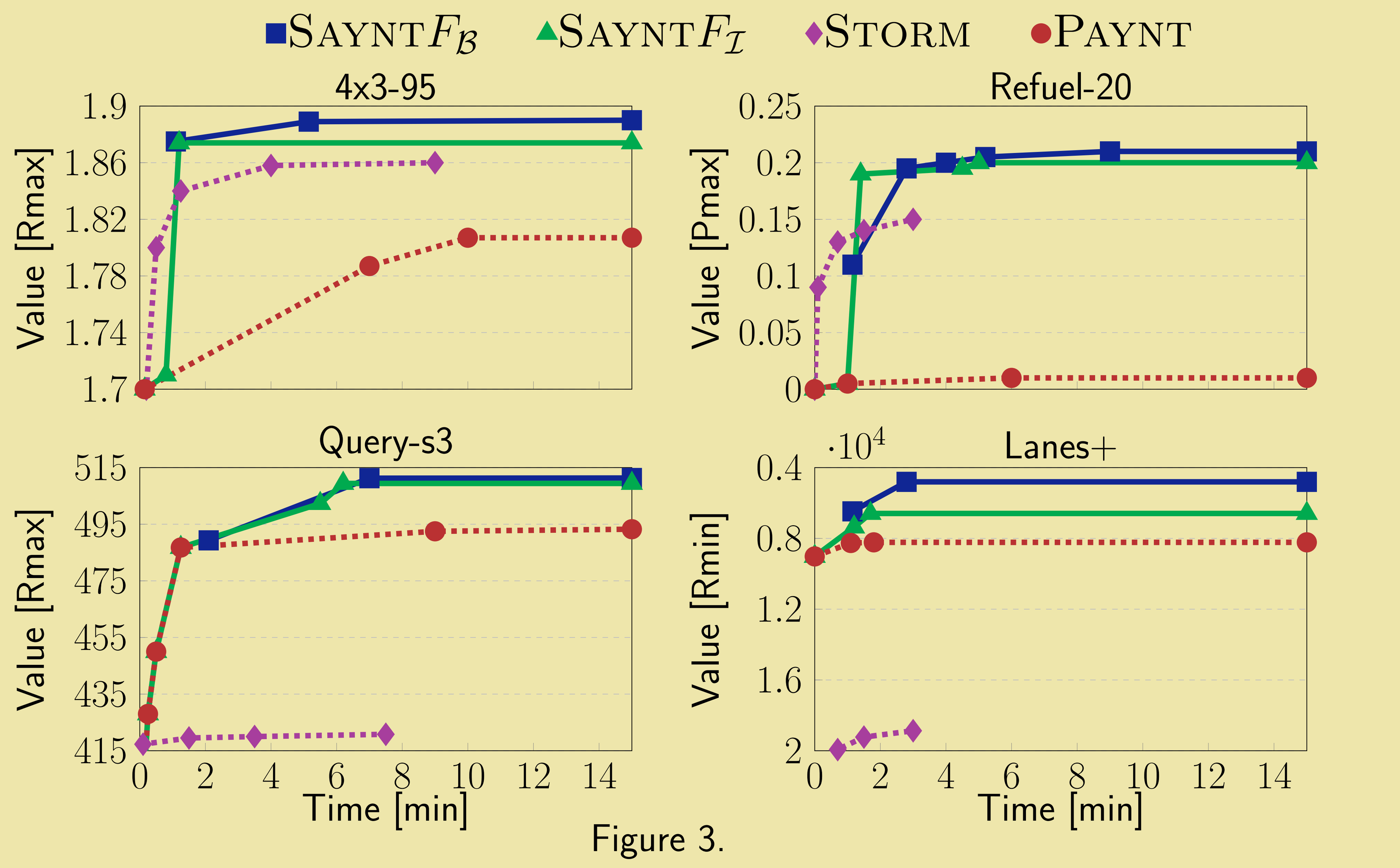
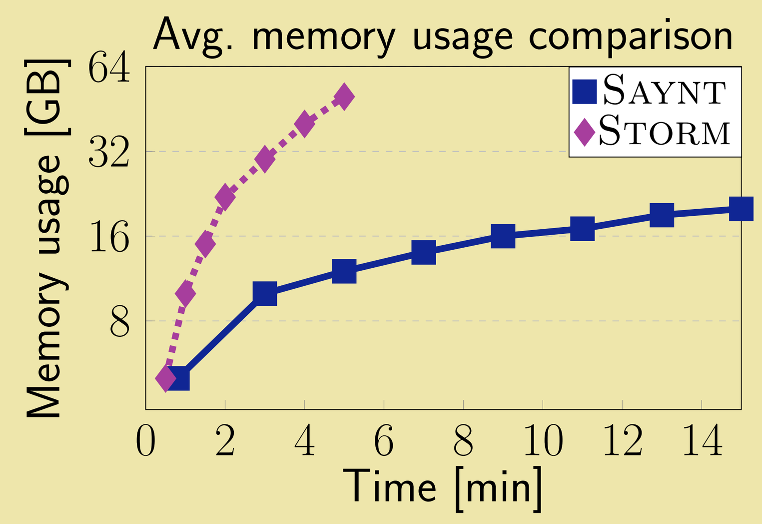


Figure 3.

Saynt steadily outperforms both baselines – the quality of improvements grows with the complexity of POMDPs and reaches up to 40%.

Saynt reduces the memory usage of Storm by a factor of 4 and thus allows an efficient belief-space exploration of larger POMDPs.

Saynt gives users a unique choice of which controller to use: smaller $F_{\mathcal{I}}$ or slightly better but much larger $F_{\mathcal{B}}$.



Acknowledgement and References

We would like to thank Alexander Bork, Sebastian Junges and Joost-Pieter Katoen for their help with this research.

[1] R. Andriushchenko, A. Bork, M. Češka, S. Junges, J.P. Katoen, and F. Macák. Search and Explore: Symbiotic Policy Synthesis in POMDPs. Accepted to CAV'23.

[2] A. Bork et al. Under-approximating expected total rewards in POMDPs. In TACAS'22.

[3] H. Kurniawati et al. SARSOP: Efficient point-based POMDP planning by approximating optimally reachable belief spaces. In Robotics: Science and Systems 2008.

[4] R. Andriushchenko et al. Inductive synthesis of finite-state controllers for POMDPs. In UAI'22.

[5] J. Schrittwieser et al. Mastering atari, go, chess and shogi by planning with a learned model. In Nature 2020