Towards Scalable MDP Algorithms

Andrey Kolobov
Advisors: **Mausam** and **Daniel S. Weld**
{akolobov, mausam, weld}@cs.washington.edu
Dept of Computer Science and Engineering
University of Washington, Seattle WA-98195

1 Introduction

Since the emergence of Artificial Intelligence as a field, planning under uncertainty has been viewed as one of its crucial subareas. Accordingly, the current lack of scalability of probabilistic planning techniques is a major reason why the grand vision of AI has not been fulfilled yet. Besides being an obstacle to advances in AI, scalability issues also hamper the applicability of probabilistic planning to real-world problems.

A powerful framework for describing probabilistic planning problems is Markov Decision Processes (MDPs). Informally, an MDP specifies the objectives the agent is trying to achieve, the actions the agent can perform, and the states in which it can end up while working towards the objective. Solving an MDP means finding a policy, i.e. an assignment of actions to states, that allows the agent to achieve the objective. Optimal solution methods, those that look for the "best" policy according to some criterion, typically try to analyze all possible states or a large fraction of them. Since state space sizes of realistic scenarios can be astronomical, these algorithms quickly run out of memory. Fortunately, the mathematical structure of some classes of MDPs has allowed for inventing more efficient algorithms. This is the case for Stochastic Shortest Path (SSP) problems, whose mathematical properties gave rise to a family of algorithms called Find-and-Revise. When used in combination with a heuristic, the members of this family can find a near-optimal, or even optimal, policy while avoiding the analysis of many states. Nonetheless, the sheer number of states in real-world problems forces algorithms based on state-level analysis to exhaust their capabilities much too early, calling for a fundamentally different approach. Moreover, several expressive and potentially useful MDP classes are currently not known to have a mathematical structure for creating efficient approximation techniques.

This dissertation advances the state of the art in probabilistic planning in three complementary ways. For SSP MDPs, it derives a novel class of approximate algorithms based on *generalizing* state analysis information across many states. This information is stored in the form of *basis functions* that are generated automatically and the number of which is much smaller than the number of states. As a result, the proposed algorithms have a very compact memory footprint and arrive at high-quality solutions faster than their state-based counterparts. In a parallel effort, the dissertation also builds up mathematical apparatus for classes of MDPs that previously had no applicable approximation algorithms. The developed theory enables the extension of the powerful Find-and-Revise paradigm to these MDP classes as well, providing the first memory-efficient algorithms for solving them. Last but not least, the dissertation will apply the proposed theoretical techniques to a large-scale real-world problem, urban traffic routing being one of the candidates.

2 Approximate Algorithms for SSP MDPs

The optimal algorithms' lack of scalability has led researchers to consider at least three paradigms for approximately solving SSP MDPs. *Heuristic search* algorithms, when given the initial state of the MDP and a heuristic, use the latter to avoid analyzing many states that a-priori are not part of any good policy. However, the number of states that end up being analyzed is often still too high to fit in memory. *Determinization* techniques solve a relaxation of the original MDP with all or most of the probabilistic information removed. They are fast due to the use of classical planners for solving the determinization but have trouble with subtle probabilistic structure. *Function approximation* approaches store the state space analysis information in the form of a small number of *basis functions*. Unfortunately, man of these algorithms require carefully hand-crafted basis functions for good results, i.e. are not fully automatic.

This dissertation proposes a novel *generalization framework* that integrates all three paradigms, negating their drawbacks but retaining the advantages of each. Like function approximation, it uses a small number of special basis functions that associate some information, e.g., a numeric value, with every state. Aggregating values assigned by basis functions to a given state allows for reconstructing the state value function and hence obtaining a policy. Basis functions in this framework are conjunctions of variable values yielded by regressing the goal conjunction through a trajectory (a sequence of action outcomes) that reaches the goal from some state. Each such basis function has a non-zero value in all states where the former's conjunction of values is present. Since every basis function is, in essence, a precondition of some trajectory, it provides implicit reachability information for all states in which its value is nonzero at once. Moreover, the trajectories for deriving basis functions can be discovered as in determinization approaches, by using fast classical planners. This makes the basis functions problem specific, easy to generate automatically, and able to capture information about large

parts of the state space.

The dissertation describes three techniques that use the above generalization framework.

- RETRASE [Kolobov *et al.*, 2009], a planner that explores the state space in a series of trials using the current greedy policy (similarly to RTDP), updating basis function values with an analogue of Bellman backup to associate probabilistic information with them. Experimental results demonstrate that RETRASE outperforms the winners of past International Probabilistic Planning Competitions on many problems.
- GOTH [Kolobov *et al.*, 2010a], a heuristic that sets basis functions' values to the costs of trajectories that these basis functions were derived from, turning the latter into informative state value estimates. Empirical data shows GOTH to save planners more memory than another very popular and effective heuristic, FF.
- SIXTHSENSE [Kolobov *et al.*, 2010b], a machine learning algorithm for identifying dead ends in MDPs. Dead ends are states from which the goal cannot be reached, preventing the derivation of basis functions for them as above. Present in many MDPs, such states can be plentiful, wasting significant computational resources. To capture information about them, SIXTHSENSE uses a special type of basis functions called *nogoods*. Contrary to the guarantees given by ordinary basis functions, the presence of a nogood in a state means that the goal is impossible to reach from that state. SIXTHSENSE can serve as a submodule in nearly any planner and construct nogoods with a fast statistical learning technique. The experiments indicate that SIXTHSENSE makes planners appreciably faster and more frugal with memory.

3 Beyond Stochastic Shortest-Paths Problems

Although SSP MDPs model many important scenarios, this class's restrictions leave out plenty of interesting problems as well. Consider, for instance, a setting in which the agent is looking for a policy that maximizes the probability of reaching the goal from some initial state, independently of such a policy's cost. An MDP in which the agent incurs zero cost for any action and receives a reward of 1 for attaining the goal naturally describes this situation. Unfortunately, in general it will not be an SSP because its zero-cost actions may form loops. While seemingly innocuous, the presence of such structures in the state space of an MDP invalidates a fundamental property of SSP problems, that Bellman backup, the essential operator in algorithms for solving SSP MDPs, is always guaranteed to converge to the optimal solution. Zero-cost loops give rise to multiple suboptimal solutions that Bellman backup may converge to. Since SSP approximation algorithms (e.g., those falling under the Find-and-Revise framework) indirectly rely on the optimality of Bellman backup's convergence as well, they cannot be used successfully on problems like above.

As this dissertation demonstrates, a promising way to extend approximation algorithms beyond SSP MDPs is to invent an operator that converges to the unique optimal solution on a broader class of problems. The first paper exploring this possibility [Kolobov *et al.*, 2011] introduces a new problem type, Generalized Stochastic Shortest-Paths (GSSP), which properly contains such important classes as SSP, positive-bounded, and negative MDPs. The paper develops a mathematical apparatus for analyzing GSSPs and proposes an operator provably having the desired optimal convergence property. It also empirically demonstrates the successful performance on GSSP problems of a heuristic search algorithm with this operator in lieu of Bellman backup.

Planned future work in this direction will extend the analysis to handle even more general MDP classes.

4 Applications

Benchmark performance of a new technique is, unfortunately, not a reliable predictor of its performance in the real world. As a truly challenging scalability test, this dissertation will employ theoretical concepts developed in it in handling a large realistic application. The choice of a particular application largely depends on data availability. However, one exciting possibility is tackling the problem of traffic routing in a large urban area. A recent paper on this topic by IBM researchers reported that a routing system based on relatively unsophisticated deterministic planning algorithms could reduce the average travel time in the city of Stockholm by as much as 62%. Clearly, more careful traffic modeling with probabilistic MDPs and scalable algorithms to solve them would yield even more impressive savings. Besides the average travel time, probabilistic techniques could help optimize for other important criteria such as travel time variance, and thus allow for a finer control of a city's traffic situation.

5 Conclusions

Increasing the scalability of MDP algorithms is a central problem in AI. This dissertation approaches it in two ways. For SSP MDPs, it develops a class of approximation algorithms that combines existing successful solution paradigms in a novel way. These algorithms represent state space information compactly with the help of basis functions, and consume little computational resources as a result, while producing high-quality policies. For more general MDP types, the theoretical guarantees underlying SSP algorithms typically do not hold. This dissertation makes it possible to extend efficient approximation algorithms to these classes as well by developing an appropriate mathematical framework. Future work will concentrate on developing efficient algorithms for yet more general MDP classes and on applying the techniques proposed here to large-scale real-world applications such as urban traffic routing.

References

[Kolobov *et al.*, 2009] A. Kolobov, Mausam, and D. Weld. ReTrASE: Integrating paradigms for approximate probabilistic planning. In *IJCAI'09*, 2009.

[Kolobov *et al.*, 2010a] A. Kolobov, Mausam, and D. Weld. Classical planning in MDP heuristics: with a little help from generalization. In *ICAPS'10*, 2010.

[Kolobov *et al.*, 2010b] A. Kolobov, Mausam, and D. Weld. SixthSense: Fast and reliable recognition of dead ends in MDPs. In *AAAI'10*, 2010.

[Kolobov *et al.*, 2011] A. Kolobov, Mausam, D. Weld, and H. Geffner. Heuristic search for generalized stochastic shortest path MDPs. In *ICAPS'11*, 2011.