Probabilistic Planning with Risk-Sensitive Criterion

Ping Hou
Department of Computer Science
New Mexico State University
Las Cruces, NM 88003, USA
phou@cs.nmsu.edu

Introduction

Probabilistic planning models and, in particular, *Markov Decision Processes* (MDPs), *Partially Observable Markov Decision Processes* (POMDPs) and *Decentralized Partially Observable Markov Decision Processes* (Dec-POMDPs) have been extensively used by AI and Decision Theoretic communities for planning under uncertainty. Typically, the solvers for probabilistic planning models find policies that minimize the expected cumulative cost (or, equivalently, maximize the expected cumulative reward). While such a policy is good in the expected case, there is a small chance that it might result in an exorbitantly high cost. Therefore, it is not suitable in high-stake planning problems, where exorbitantly high costs should be avoided.

With this motivation in mind, Yu, Lin, and Yan (1998) introduced the *Risk-Sensitive criterion* (RS-criterion) for MDPs, where the objective is to find a policy π that maximizes the probability $Pr(c^{\mathcal{T}(\pi)}(s_0) \leq \theta_0)$, where $c^{\mathcal{T}(\pi)}(s_0)$ is the cumulative cost of the policy and θ_0 is the cost threshold. They combine MDPs with the RS-criterion to formalize *Risk-Sensitive MDPs* (RS-MDPs) and introduced a *Value Iteration* (VI) like algorithm to solve a typical type of RS-MDPs. Liu and Koenig (2006) generalized RS-MDPs by mapping the MDP rewards to risk-sensitive utility functions and sought to find policies that maximize the expected utility—an RS-MDP is a specific case, where the utility function is a step function. They introduced *Functional Value Iteration* (FVI), which finds optimal policies for general utility functions by approximating it as piecewise linear (PWL) functions.

Unfortunately, algorithms like VI and FVI cannot scale to large problems as they need to perform Bellman updates for all states and all break points of their utility function in each iteration. As such, more efficient algorithms can be developed to take advantage of structure in RS-MDPs.

In my work, I introduced various algorithms for RS-MDPs with different assumptions (e.g., MDPs with dead ends and MDPs with zero or negative cost cycles). In addition to RS-MDPs, POMDPs and Dec-POMDP can also be combined with RS-criterion to formalize *Risk-Sensitive POMDPs* (RS-POMDPs) and *Risk-Sensitive Dec-POMDPs* (RS-Dec-POMDPs). Algorithms can also be designed for RS-POMDPs and RS-Dec-POMDPs with different assumptions.

Current Progress

In our recent paper (Hou, Yeoh, and Varakantham 2014), we formally defined *Risk-Sensitive MDPs* (RS-MDPs) and show that the optimal policy for RS-MDPs is not stationary in the original state space. So an *MDP policy* $\pi : \mathbf{S} \rightarrow \mathbf{A}$, namely a mapping from states to actions, is not always optimal. Instead, the execution history can be compctly represented by a cost threshold value $\theta = \theta_0 - c^{\mathcal{T}(\pi)}(s_0, t)$, namely the amount of unused cost, where θ_0 is the initial cost threshold and $c^{\mathcal{T}(\pi)}(s_0, t)$ is the accumulated cost thus far up to the current time step t. Therefore, instead of an MDP policy, an *RS-MDP policy* $\pi : \mathbf{S} \times \mathbf{\Theta} \rightarrow \mathbf{A}$, which is a mapping of augmented states $(s, \theta \mid s \in \mathbf{S}, \theta \in \mathbf{\Theta})$ to actions $a \in \mathbf{A}$, can give an optimal solution to an RS-MDP.

Based on the augmented state (s, θ), we can build a *augmented MDP*, where the actions and transitions correspond to their counterpart in the original MDP and the reward function is 1 for transitions that transition into a goal state and 0 otherwise, which is an important property of MAXPROB MDPs (Kolobov et al. 2011). Since the cost threshold value θ could be real number, the number of augmented states (s, θ) could be infinite and the augmented MDP is actually an *MDP with Continuous State Spaces* (Marecki, Koenig, and Tambe 2007). From a decision-theoretic view, the solution for an RS-MDP can be represented with step utility functions, which is the mapping from cost threshold value to the reachable probability, for each state. Each augmented state actually corresponds to a point in the utility function, and the number of break points in the utility function is countable. So, the reachable probability of all break points together completely describes the entire utility function.

In (Hou, Yeoh, and Varakantham 2014), we show that the number of the break points of utility function is finite as long as the cost function does not form *negative cycles* in the original state space. By extracting those augmented states corresponding to break points, they together form the states of an augmented MDP with a finite state space. Starting from augmented states with an original goal state, a *Dynamic Programming* (DP) style algorithm, TVI-DP, is introduced to traverse the augmented state space backwards without generating the augmented MDP explicitly. The exploration of augmented state space can be stopped when a user-defined initial cost threshold θ_0 is reached or the reachable probability become convergence. Besides TVI-DP, if an RS-MDP user is only interested in the reachable probability and policy for a specific initial cost threshold θ_0, it is unnecessary to get the full solution. A *Depth-First Search* (DFS) style algorithm, TVI-DFS, is provided to traverse the augmented state space from the initial augmented state (s_0, θ_0), where s_0 is the initial state in the original MDP. TVI-DFS traverses only augmented states that are reachable from (s_0, θ_0). Those augmented states often correspond to points on segments of the utility function and, thus, might not correspond to break points in the utility function. TVI-DFS will stop the exploration when the cost threshold is less than 0. TVI-DFS also implicitly traverse another subset augmented MDP with finite states. Both TVI-DP and TVI-DFS use techniques from the Topological Value Iteration (TVI) (Dai et al. 2011) algorithm to identify and handle zero cost cycles in the state space.

Besides research on RS-MDPs, I have also studied another related problem—Uncertain MDPs, which are the MDPs with uncertainty and instability in their parameters. In (Hou, Yeoh, and Son 2014), I introduced a general algorithm framework that uses a reactive approach and allows off-the-shelf MDP algorithms to solve Uncertain MDPs by planning based on currently available information and replan if and when the problem changes.

Research Plan

The overall scope of my thesis is to develop efficient and scalable algorithms to optimize the RS-criterion in probabilistic planning problems. I now describe my short- and long-term plans.

Short-Term Plans

PVI-DP: If the original MDP includes negative cycles with discounted transition probability in the state space, then the number of break points of utility functions would be infinite. TVI-DFS and TVI-DP would suffer from this situation because the augmented state space that they would need to traverse would be infinitely large. Therefore, smarter search strategies are needed to handle the infinite augmented state space. For example, one possible algorithm is PVI-DP, which adopts ideas from *Prioritized Value Iteration* (PVI) (Wingate and Seppi 2005). Similar to TVI-DP, PVI-DP traverses the augmented state space backwards from the goal states. Every time PVI-DP expands a fringe augmented state, it performs value updates for its predecessors, and pushes them into a priority queue based on the update error. Next, PVI-DP pops the augmented state with the biggest update error, performs value updates on it, pushes it back, and expands it if it is fringe node. After repeating the previous process, the exploration of augmented state space would stop at the fringe nodes whose update error is sufficiently small, thereby reaching convergence. In addition, if the original state space can be separated into more than one *Strongly Connected Component* (SCC), then the traversal of augmented state space can follow the reverse topological order of SCCs in the original state space. Not only can it perform PVI-DP on the SCCs one by one, but it can also perform TVI-DP instead of PVI-DP on the SCCs without negative cycles, which should accelerate the algorithm.

PO-FVI: For RS-POMDPs, we plan to propose a POMDP style and reduced version of FVI called PO-FVI. The key idea of PO-FVI is to represent the solution of RS-POMDPs as a set of vector of utility functions, rather than representing the value function as a set of vector of numbers (commonly referenced to as α-vectors) in the original POMDP. RS-POMDPs have many interesting properties that are different than those in RS-MDPs, and we have started to develop better algorithms on top of PO-FVI.

Long-Term Plans

Besides the above tasks, I also plan to consider other planning models related to the RS-criterion. For example, how to solve RS-Dec-POMDPs; how to solve RS-MDPs when the underlying MDP model is an *MDP with reward discount factor* rather than a *Goal-Directed MDP*; and how to quickly solve an RS-criterion problem with *satisfaction* goals rather than *optimization* goals.

References

Dai, P.; Mausam; Weld, D.; and Goldsmith, J. 2011. Topological value iteration algorithms. *Journal of Artificial Intelligence* 42(1):181–209.

Hou, P.; Yeoh, W.; and Son, T. C. 2014. Solving uncertain MDPs by reusing state information and plans. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, 2285–2292.

Hou, P.; Yeoh, W.; and Varakantham, P. 2014. Revisiting risk-sensitive MDPs: New algorithms and results. In *Proceedings of the International Conference on Automated Planning and Scheduling (ICAPS)*, 136–144.

Kolobov, A.; Mausam; Weld, D. S.; and Geffner, H. 2011. Heuristic search for generalized stochastic shortest path MDPs. In *Proceedings of the International Conference on Automated Planning and Scheduling (ICAPS)*, 130–137.

Liu, Y., and Koenig, S. 2006. Functional value iteration for decision-theoretic planning with general utility functions. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, 1186–1193.

Marecki, J.; Koenig, S.; and Tambe, M. 2007. A fast analytical algorithm for solving Markov decision processes with real-valued resources. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2536–2541.

Wingate, D., and Seppi, K. D. 2005. Prioritization methods for accelerating MDP solvers. *Journal of Machine Learning Research* 6:851–881.

Yu, S.; Lin, Y.; and Yan, P. 1998. Optimization models for the first arrival target distribution function in discrete time. *journal of Mathematical Analysis and Applications* 225:193–223.