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

1 Introduction

Probabilistic planning models – *Markov Decision Processes* (MDPs) and *Partially Observable Markov Decision Processes* (POMDPs) – have been extensively studied by Artificial Intelligence communities for planning under uncertainty. The conventional criterion for these probabilistic planning models is to find policies that minimize the expected cumulative cost, which we call the *Minimizing Expected Cost criterion* (MEC-criterion). While such a policy is good in the expected case, there is a chance that it might result in an exorbitantly high cumulative cost. Therefore, it is not suitable in high-stake planning problems, where exorbitantly high cumulative costs should be avoided. With the above motivation in mind, Yu *et al.* [1998] introduced the *Risk-Sensitive criterion* (RS-criterion), where the objective is to find a policy that maximizes the probability that the cumulative cost of possible execution trajectories is less than an user-defined initial cost threshold θ_0. Also, if we consider the initial cost threshold θ_0 as the cost budget that the system holds at the beginning, then the above objective probabilities, namely *reachable probabilities*, are the probabilities of the agent achieving its goal without exceeding its cost budget. By combining *goal-directed* MDPs and POMDPs with the RS-criterion, the corresponding risk-sensitive probabilistic planning models – *Risk-Sensitive MDPs* (RS-MDPs) [Yu *et al.*, 1998; Hou *et al.*, 2014] and *Risk-Sensitive POMDPs* (RS-POMDPs) [Hou *et al.*, 2016] – can be formalized. For RS-MDPs and RS-POMDPs, the objectives are to respectively find a policy π that maximizes the probability $Pr(c^{\mathcal{T}(s_0,\pi)} \leq \theta_0)$ and $\sum_s b_0(s) \cdot Pr(c^{\mathcal{T}(s,\pi)} \leq \theta_0)$, where $c^{\mathcal{T}(s,\pi)}$ is the cumulative cost of a trajectory formed by executing policy π from state s.

In this ongoing work, I formally define RS-MDPs and RS-POMDPs and introduce various algorithms for RS-MDPs and RS-POMDPs with different assumptions (e.g., with zero costs and with cost observations).

2 Current Progress

In our recent papers [Hou *et al.*, 2014; 2016], we show that the optimal policies for RS-MDPs and RS-POMDPs are not stationary with respect to the original states. The optimal action choice also depends on the cost threshold $\theta = \theta_0 - c^{\mathcal{T}}(t)$, where $c^{\mathcal{T}}(t)$ is the accumulated cost thus far up to the current time step t. The cost threshold θ is actually the wealth level or remainder cost budget of the system if the initial cost threshold θ_0 is considered as the cost budget at the beginning.

For optimal policies of RS-MDPs and RS-POMDPs, all essential information, which can be extracted from the execution history, is the cost threshold. Therefore, instead of considering only the state, the optimal decision of RS-MDPs and RS-POMDPs consider both the state and the cost threshold. We call a pair of state and cost threshold as an *augmented state* (s, θ), and the augmented state space is $\widehat{\mathbf{S}} : \mathbf{S} \times \mathbf{\Theta}$, where $\mathbf{\Theta}$ is the set of all possible cost thresholds. For RS-POMDPs, an augmented belief state is a probability distribution on augmented states $(s, \theta) \in \widehat{\mathbf{S}}$, and $\widehat{\mathbf{B}}$ is the set of all possible augmented belief states. Instead of an regular MDP or POMDP policy, an *RS-MDP policy* $\widehat{\mathbf{S}} \rightarrow \mathbf{A}$ or *RS-POMDP policy* $\widehat{\mathbf{B}} \rightarrow \mathbf{A}$ can always give the optimal solution. For RS-POMDPs, since the optimal decision depends on the cost threshold and the system needs to consider it, the issue of whether the actual costs can or cannot be observed is important. If the cost can be observed, the system knows exactly its cost threshold, and the augmented belief state space is reduced to $\mathbf{B} \times \mathbf{\Theta}$, where $\mathbf{B}$ is the set of all regular POMDP belief states. We distinguish the two situations whether costs can or cannot be observed for RS-POMDP [Hou *et al.*, 2016], and provide corresponding algorithms for both situations.

For RS-MDPs and RS-POMDPs, we can build corresponding *augmented MDPs* and *augmented POMDPs* based on augmented states, where the actions, transitions, and observations correspond to their counterpart in the original MDP and POMDP, respectively, and the reward function is 1 for transitions that transition into augmented states with goal states and non-negative cost thresholds, and 0 otherwise. Because reachable probabilities of any augmented states with negative costs are 0, it is only necessary to consider the possible cost threshold values inside the interval $[0, \theta_0]$, then the number of augmented states in the augmented MDPs or POMDPs is finite. The fundamental *Value Iteration* (VI) style algorithm can be applied on augmented MDPs and POMDPs. Solving an RS-MDP or RS-POMDP is actually equivalent to solving its corresponding augmented MDP or POMDP, respectively.

Theoretically, the number of possible cost thresholds $|\mathbf{\Theta}|$ is finite, but it can be very large in practice because the cost threshold θ can be any real numbers. Liu and Koenig [2006] introduced MDPs with *utility functions*, that map cumulative rewards to utility values, and sought to find policies that maximize the expected utility. They introduced *Functional Value Iteration* (FVI), which finds optimal policies for MDPs with *Piecewise Linear* (PWL) utility functions. Marecki and Varakantham [2010] extended FVI to solve *finite-horizon* POMDPs with PWL utility functions. By considering the cost threshold as a continuous variable on the domain $[0, \theta_0]$, the RS-criterion is equivalent to an assumption that the system has a *step* utility function. Thus, solving RS-MDPs and RS-POMDPs is equivalent to solving MDPs and POMDPs with a step utility function, and the solutions are reachable probability functions $\mathbf{S} \rightarrow (\mathbf{\Theta} \rightarrow [0, 1])$ that map each state to a *Piecewise Constant* (PWC) function on the domain $[0, \theta_0]$. For both situations that the cost can or cannot be observed, we customized FVI to solve *goal-directed* RS-POMDPs and introduced *linear programming* techniques to prune dominated vectors. From the viewpoint of utility functions, all algorithms we introduced in this ongoing work can be used to solve MDPs and POMDPs with a specific type of utility function – utility functions with *constant tails*, which assume that the agent gets a constant utility if the cost threshold is smaller than a threshold. For example, MDPs with utility functions and *worst-case guarantees* [Ermon *et al.*, 2012] is a specific case of utility functions with constant tails.

For RS-MDPs and RS-POMDPs, if the model assumes all costs are positive, then the successor augmented states and belief states always have smaller cost thresholds. Thus, there exist no cycles between augmented states or belief states, and *Depth-First Search* (DFS) style algorithms can be introduced to traverse respectively the reachable augmented state or belief state space from the initial augmented state or belief state. DFS algorithms update the reachable probabilities for augmented states or belief states in the *reverse topology order*. For RS-MDPs, we introduced a *Dynamic Programming* (DP) style algorithm that traverses the augmented state space backwards form augmented states with original goal states and cost threshold 0. DP updates reachable probabilities for all augmented states from smaller cost thresholds to larger cost thresholds.

If RS-MDPs allow zero costs, the transition edges with zero costs can form cycles between augmented states with exact the same cost threshold. By adopting the idea from *Topological Value Iteration* (TVI) [Dai *et al.*, 2011], we introduced the TVI-DFS and TVI-DP algorithms, which are generalized versions of DFS and DP. Both TVI-DFS and TVI-DP identify the *Strongly Connected Components* (SSCs) formed by zero cost cycles and perform updates for all augmented states in one SCC simultaneously. For RS-POMDPs with zero costs, if the cost can be observed, a DP style algorithm can solve it by solve one regular POMDP for each possible cost threshold $\theta \in \mathbf{\Theta}$, from smaller cost thresholds to larger cost thresholds.

Generally, the DFS and DP style algorithms (include TVI-DFS and TVI-DP) are faster than VI and FVI. The reason is that VI and FVI need to perform updates for the entire augmented state or belief state space in every iteration. On the other hand, DFS and DP style algorithms only perform updates for each augmented state or belief state for a minimum number of iterations. In addition, DFS style algorithms only explore the reachable augmented state or augmented belief state space, which can be much smaller than the entire one.

3 Future Plan

As a next step, I plan to design an approximation algorithm – *Local Search* (LS) – that produces policies for RS-MDPs by adjusting the regular optimal policy based on the MEC-criterion. LS will first get the optimal MDP policy with the minimum expected cumulative cost, and then form an RS-MDP policy by assuming that every augmented state (s, θ) for a state s and all possible cost thresholds θ have the same action choice. Given this formed RS-MDP policy, namely *abstracted policy*, we can compute the set of reachable augmented states. Then, we iterate through this set of reachable augmented states, and for each augmented state in this set, we evaluate all actions and choose the action that maximizes the reachable probability under the assumption that every other augmented state uses the abstracted policy. Once a different action is chosen for an augmented state, we record it for this augmented state and now have a new abstracted policy. Once LS goes through all the reachable augmented states, we recompute the set of reachable augmented states again with the new updated abstracted policy, and the abstracted policy will keep improving through this process. We keep repeating this process until there are no new reachable augmented states can be found and the selected action for all reachable augmented states remain unchanged across subsequent iterations. This LS algorithm may not guarantee optimality but it is *anytime* and *memory-bounded*.

References

[Dai *et al.*, 2011] Peng Dai, Mausam, Daniel Weld, and Judy Goldsmith. Topological value iteration algorithms. *Journal of Artificial Intelligence*, 42(1):181–209, 2011.

[Ermon *et al.*, 2012] Stefano Ermon, Carla Gomes, Bart Selman, and Alexander Vladimirsky. Probabilistic planning with non-linear utility functions and worst-case guarantees. In *Proc. of AAMAS*, pages 965–972, 2012.

[Hou *et al.*, 2014] Ping Hou, William Yeoh, and Pradeep Varakantham. Revisiting risk-sensitive MDPs: New algorithms and results. In *Proc. of ICAPS*, pages 136–144, 2014.

[Hou *et al.*, 2016] Ping Hou, William Yeoh, and Pradeep Varakantham. Solving risk-sensitive POMDPs with and without cost observations. In *Proc. of AAAI*, 2016.

[Liu and Koenig, 2006] Yaxin Liu and Sven Koenig. Functional value iteration for decision-theoretic planning with general utility functions. In *Proc. of AAAI*, pages 1186–1193, 2006.

[Marecki and Varakantham, 2010] Janusz Marecki and Pradeep Varakantham. Risk-sensitive planning in partially observable environments. In *Proc. of AAMAS*, pages 1357–1368, 2010.

[Yu *et al.*, 1998] Stella Yu, Yuanlie Lin, and Pingfan Yan. Optimization models for the first arrival target distribution function in discrete time. *Journal of Mathematical Analysis and Applications*, 225:193–223, 1998.