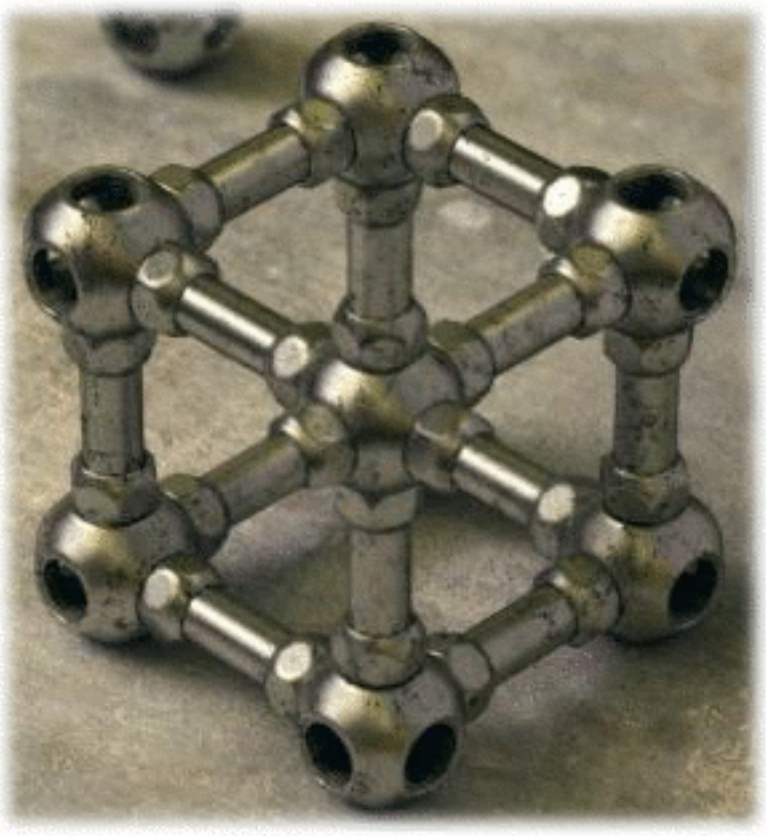

Factored Markov Decision Processes with Imprecise Probabilities: a multilinear solution [1]

Authors:

Karina Valdivia Delgado
Leliane Nunes de Barros
Fabio Gagliardi Cozman

[1] This work was supported by Fapesp Project LogProb, grant 2008/03995-5, São Paulo, Brazil.

Factored Markov Decision Processes with Imprecise Probabilities: a multilinear solution

Karina Valdivia Delgado,
Leliane Nunes de Barros and Fabio Gagliardi Cozman
Universidade de São Paulo,
São Paulo, Brazil.

Abstract

There are efficient solutions to planning problems modeled as a Markov Decision Process (MDP) envolving a reasonable number of states. However, known extensions of MDP are more suited to represent practical and more interesting applications, such as: (i) an MDP where states are represented by state variables, called a factored MDP; (ii) an MDP where probabilities are not completely known, called an MDPIP. In this work, we are interested in exploring efficient algorithms to solve *factored MDPIPs*.

Introduction

The main approach to solve a probabilistic planning problem is modeling it as a Markov decision process (MDP) (Bonet and Geffner 2005). Originally proposed by the community of decision theory, MDPs (Puterman 1994) provide an elegant mathematical framework for representing and solving sequential decision problems under uncertainty in completely observable environments.

An MDP models the interaction between an agent and its environment: at every stage, the agent decides to execute an action (with probabilistic effects) that will produce a future state and a reward. The agent's goal is to maximize the reward gained over a sequence of action choices.

Since acquiring the probability distribution of models from humans is difficult and often subjective, we should try to deal with imprecise probabilities in order to represent incomplete, ambiguous or conflicting expert believes about transitions of states. A Markov Decision Process with Imprecise Probabilities (MDPIP) (White III and Eldeib 1994) is a generalization of an MDP that allows the representation of imprecise probabilities. However, there are no efficient solutions for MDPIPs (Shirota et al. 2007).

Since, efficient solutions based on linear programming are known for factored MDPs, it seems that an MDPIP modeled in this way can be solved in a more efficient form. The goal of this Ph.D. project is to define the idea of a *factored MDPIP* and investigate different forms to solve this new problem, for which we have not found solutions in the literature.

Background

This section contains a brief review of the linear programming formulation for MDPs and factored MDPs. We also show how an MDPIP can be formulated as a non-linear programming.

MDP

Formally, an MDP is defined by the tuple $\mathcal{M} = \langle T, S, A, R, P \rangle$ where: T is a countable set of stages, S is a finite set of states, A is a finite set of actions, R is the reward function and P defines the transition probabilities. Let $V^*(s)$ be the optimal value of the state $s \in S$, based on the value of the possible successor states $s' \in S$, for an agent that wants to maximize his expected reward. The *Bellman Optimality equation* is:

$$V^*(s) = max_{a \in A}\{R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a)V^*(s')\} \quad (1)$$

The formulation of an MDP problem as a linear programming is given by (Manne 1960)(α is the state relevance):

$$\min_{V^*} \quad : \quad \sum_{s} \alpha(s)V^*(s) \quad (2)$$

$$s.t. \quad : \quad V^*(s) \geq R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a)V^*(s'), \forall s \in S, a \in A.$$

MDPIP

An MDPIP models an agent that at every stage decides to execute an action (with probabilistic effects) that will produce a future state and a reward that also depends on the choices of nature (w.r.t. the probability imprecisions). The agent's goal is to maximize the reward gained over a sequence of choices of actions assuming, for example, that the nature chooses either to minimize agent's reward (*maxmin* criteria) or to maximize agent's reward (*maxmax* criteria). Formally, the definition of an MDPIP, described by the tuple $\mathcal{M} = \langle T, S, A, R, K \rangle$, follows the definition for an MDP plus the credal conditional sets[1] $K_a(s'|s)$, represented by linear inequations, to express all possible probability distributions (Cozman 2000). The *Bellman Optimality equation* for MDPIPs that adopt the *maxmin* criteria is:

$$V^*(s) = \max_{a \in A} \min_{P(s'|s,a) \in K_a(s'|s)} \{R(s) + \gamma \sum_{s' \in S} P(s'|s,a)V^*(s')\} \quad (3)$$

[1] A credal set $K(X)$ contains a set of probability distributions for variable X.

The Equation (3) can be reduced to a bilevel programming problem (Shirota et al. 2007):

$$\begin{aligned} \min_{V^*} \quad & : \quad \sum_s \alpha(s)V^*(s) \qquad (4) \\ s.t. \quad & : \quad V^*(s) \geq R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a)V^*(s'), \forall s \in S, a \in A. \\ & \quad P \in argmin \sum_{s' \in S} P(s'|s,a)V^*(s'). \\ & \quad s.t. : P(s'|s,a) \in K_a(s'|s) \end{aligned}$$

This bilevel problem can be transformed in an equivalent multilinear program (Shirota et al. 2007):

$$\begin{aligned} \min_{V^*,P} \quad & : \quad \sum_s \alpha(s)V^*(s) \qquad (5) \\ s.t. \quad & : \quad V^*(s) \geq R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a)V^*(s'), \forall s \in S, a \in A \\ & \quad P(s'|s,a) \in K_a(s'|s). \end{aligned}$$

Some algorithms for solving MDPIPs are based on dynamic programming (White III and Eldeib 1994) (Givan, Leach, and Dean 2000; Trevizan, Cozman, and de Barros 2007). Some of them only solve special cases of MDPIPs. We do not give details about these algorithms since the main focus of this project is to solve general MDPIPs using approximate methods based on optimization (linear or nonlinear programming), instead of using dynamic programming.

Factored MDPs

Much of the recent work in AI has focused on factored structured representations of MDPs and their efficient solutions. In a factored MDP, the states are described using state variables X_i for i=1..n; the transitions are represented in compact form by using Dynamic Bayesian Networks (DBN); the value function and the reward are also factored (Guestrin 2003). Although a factored MDP becomes exponential on the number of state variables, its representation explores the behavior of the state variables in the state transition model (Guestrin 2003).

Factored Transition Model In factored MDP, for each action a we define a DBN with two layers directed acyclic graph: one representing the actual state and other representing the next state. The nodes are denoted by X_i and X_i' for state variables in the actual state and next state, respectively. Edges are allowed *from* nodes in the first layer *into* the second layer, and also between nodes in the second layer. We denote by $Parents(X_i')$ the *parents* of X_i' in the graph. The graph is assumed endowed with the following Markov condition: a variable X_i' is conditionally independent of its nondescendants given its parents. This implies the following factorization of transition probabilities:

$$P(s'|s,a) = \prod_{i=1}^{n} P(x_i'|Parents(X_i'), a) \qquad (6)$$

Factored Value Function $V^*(s)$ can be approximated using a linear combination of basis functions $h_1, ..., h_k$, i.e.:

$$\widehat{V}(s) = \sum_{j=1}^{k} w_j h_j(s) \qquad (7)$$

A necessary condition to make efficient calculations (Koller and Parr 1999) is the scope of each basis function be restricted to some subset of state variables $C_i \subset S = \{X_1, ..., X_n\}$.

Factored Reward Function Like for the factored valued function, the scope of each local-reward function R_i should be restricted to some small subset of state variables $D_i \subset S = \{X_1, ..., X_n\}$.

$$R(s,a) = \sum_{i=1}^{k_R} R_i(D_i(a), a) \in \mathbb{R} \qquad (8)$$

Algorithms for solving factored MDPs Approximate linear programming (ALP) has emerged recently as one of the most promising methods for solving complex factored MDPs (Guestrin 2003):

$$\begin{aligned} \min_{w} \quad & : \quad \sum_s \alpha(s) \sum_{i=1}^{k} w_i h_i(s) \qquad (9) \\ s.t. \quad & : \quad \sum_{i=1}^{k} w_i h_i(s) \geq R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) \sum_{i=1}^{k} w_i h_i(s'), \\ & \quad \forall s \in S, a \in A. \end{aligned}$$

Refinements for the ALP approach have been developed over the past few years, e.g. the algorithm ALP-Reformulation (Guestrin 2003) that creates a new smaller set of equivalent constraints for the Problem (9), avoiding to enumerate the complete set of states. There are other efficient algorithms that use general techniques to solve linear problems with large number of constraints (Patruscu 2004; de Farias and Roy 2004; Dolgov and Durfee 2006).

Proposal

Our proposal is to define a model with all advantages of factored MDPs that also represents the imprecision over the transition probabilities, that is, a *factored MDPIP*. Thus, the goals of this project are: (i) to give a formal definition of a factored MDPIP (not found in the literature) and (ii) to propose a solution for a factored MDPIP, for which the algorithms for factored MDP can not be applied.

Factored MDPIP: definitions

In a *factored MDPIP*, the states are defined using state variables X_i for $i = 1..n$ and the transitions are represented by Dynamic Credal Networks (DNC). The value and reward functions are factored as for a factored MDP.

Factored Transition Model: Dynamic Credal Network A credal network (Cozman 2000) generalizes the concept of a Bayesian network, allowing each variable, for each configuration of its parents, be associated with a set of probability densities (credal sets), instead of a single density (Cozman 2000). A DCN has also two layers, one representing the actual state and other the next state. We assume the Markov condition to operate over all combinations of distributions, each one satisfying the factorization in Equation (6).

Algorithms for solving factored MDPIPs Using the factored value function and replacing it in Problem (4):

$$\min_{w} \quad : \quad \sum_s \alpha(s) \sum_{i=1}^{k} w_i h_i(s) \qquad (10)$$

$$
\begin{aligned}
s.t. \quad : \quad & \sum_{i=1}^{k} w_i h_i(s) \geq R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) \sum_{i=1}^{k} w_i h_i(s'), \\
& \forall s \in S, a \in A. \\
& P \in argmin \sum_{s' \in S} P(s'|s,a) \sum_{i=1}^{k} w_i h_i(s'). \\
& s.t. : P(s'|s,a) \in K_a(s'|s)
\end{aligned}
$$

where:

$$P(s'|s,a) = \prod_i P(x'_i | Parents(X'_i), a)$$

Note that: there are $|S| * |A|$ constraints in the first level and m_2 constraints in the second level of the bilevel problem; the constraints in the first level are non-linear; there are the same variables in the first level and the second level (the variables that correspond to probabilities). So, the Problem (10) is not a simple bilevel problem and most known algorithms for solving bilevel problems can not be used. Based on that, we claim that it is better to work with an equivalent multilinear problem:

$$
\begin{aligned}
\min_{w,P} \quad : \quad & \sum_{s} \alpha(s) \sum_{i} w_i h_i(\mathbf{s}) \qquad (11) \\
s.t. \quad : \quad & \sum_{i} w_i h_i(s) \geq R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) \sum_{i} w_i h_i(s'), \\
& \forall s \in S, a \in A. \\
& P(s'|s,a) \in K_a(s'|s)
\end{aligned}
$$

where:

$$P(s'|s,a) = \prod_i P(x'_i | Parents(X'_i), a)$$

There are $|S| * |A| + m_2$ constraints in the multilinear programming problem for factored MDPIP, the same for bilevel problem. However, there are more available techniques for solving multilinear problems than bilevel problems. Note that in a factored MDPIP, modeled as a multilinear programming, the objective function coefficients can be calculated in the same way as factored MDP, since this function has not changed. Although for an MDPIP the probabilities are variables, it is possible to calculate the constraints in an efficient way, as in factored MDP, since the constraints are of the same type. However, we are still working with the complete set of constraints ($|S| * |A| + m_2$) and the direct use of general non-linear solvers for factored MDPIPs, can only solve problems with small state space. Therefore, the challenge is: *how to reduce the number of constraints*.

The main idea of this proposal is to apply the ALP-Reformulation technique (Guestrin 2003) to solve factored MDPIPs. This is possible since the constraints of the multilinear program are the same type of the constraints in the linear program for factored MDPs, with the addition of the probability variables. Applying this technique, the set of constraints are replaced with an equivalent set of constraints avoiding to enumerate the complete set of states. Thus our claim is that, to solve factored MDPIPs, the most promising approach is to apply the ALP-Reformulation technique in the multilinear problem (11) and then use a solver that can work with many constraints.

To carry out our experiments we also want to solve problems described in PPDDL, that uses a representation very close to a factored MDPIPs (using the *probabilistic* and *oneof* operator (Trevizan, Cozman, and de Barros 2008)). The question is: how to obtain the basis functions for these domains? This is a challenge that we will have to face. One extra challenge we have to deal with is how to represent and take advantage of the initial and goal states in a factored MDPIP (Problem (11)).

References

Bonet, B., and Geffner, H. 2005. mGPT: A probabilistic planner based on heuristic search. In *JAIR*, volume 24, 933–944.

Cozman, F. G. 2000. Credal networks. *Artificial Intelligence* 120:199–233.

de Farias, D. P., and Roy, B. V. 2004. On constraint sampling in the linear programming approach to approximate dynamic programming. *Math. Oper. Res.* 29(3):462–478.

Dolgov, D. A., and Durfee, E. H. 2006. Symmetric approximate linear programming for factored MDPs with application to constrained problems. *Ann. Math. Artif. Intell.* 47(3-4):273–293.

Givan, R.; Leach, S.; and Dean, T. 2000. Bounded-parameter markov decision processes. *Artificial Intelligence* 122:71–109(39).

Guestrin, C. E. 2003. *Planning under uncertainty in complex structured environments*. Ph.D. Dissertation, Stanford University. Adviser-Daphne Koller.

Koller, D., and Parr, R. 1999. Computing factored value functions for policies in structured MDPs. In *IJCAI*, 1332–1339.

Manne, A. S. 1960. Linear programming and sequential decision models. In *Management Science*, volume 6(3), 259–267.

Patruscu, R.-E. 2004. *Linear Approximations for Factored Markov Decision Processes*. Ph.D. Dissertation, University of Waterloo.

Puterman, M. L. 1994. *Markov Decision Processes—Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc.

Shirota, R.; Cozman, F. G.; Trevizan, F. W.; de Campos, C. P.; and de Barros, L. N. 2007. Multilinear and integer programming for markov decision processes with imprecise probabilities. In *5th ISIPTA*, 395–404.

Trevizan, F. W.; Cozman, F. G.; and de Barros, L. N. 2007. Planning under Risk and Knightian Uncertainty. In *IJCAI*, 2023–2028.

Trevizan, F. W.; Cozman, F. G.; and de Barros, L. N. 2008. Mixed probabilistic and nondeterministic factored planning through MDPST. In *ICAPS*, Workshop: A Reality Check for Planning and Scheduling Under Uncertainty.

White III, C. C., and Eldeib, H. K. 1994. Markov decision processes with imprecise transition probabilities. *Oper. Res.* 42(4):739–749.