Artificial Intelligence 230 (2016) 192–223

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

Real-time dynamic programming for Markov decision processes with imprecise probabilities

Karina V. Delgado*, Leliane N. de Barros, Daniel B. Dias, Scott Sanner

ARTICLE INFO

Article history:
Received 8 February 2014
Received in revised form 19 May 2015
Accepted 17 September 2015
Available online 25 September 2015

Keywords:
Probabilistic planning
Markov decision process
Robust planning

ABSTRACT

Markov Decision Processes have become the standard model for probabilistic planning. However, when applied to many practical problems, the estimates of transition probabilities are inaccurate. This may be due to conflicting elicitations from experts or insufficient state transition information. The Markov Decision Process with Imprecise Transition Probabilities (MDP-IPs) was introduced to obtain a robust policy where there is uncertainty in the transition. Although it has been proposed a symbolic dynamic programming algorithm for MDP-IPs (called SPUDD-IP) that can solve problems up to 22 state variables, in practice, solving MDP-IP problems is time-consuming. In this paper we propose efficient algorithms for a more general class of MDP-IPs, called Stochastic Shortest Path MDP-IPs (SSP MDP-IPs) that use initial state information to solve complex problems by focusing on reachable states. The (L)RTDP-IP algorithm, a (Labeled) Real Time Dynamic Programming algorithm for SSP MDP-IPs, is proposed together with three different methods for sampling the next state. It is shown here that the convergence of (L)RTDP-IP can be obtained by using any of these three methods, although the Bellman backups for this class of problems prescribe a minimax optimization. As far as we are aware, this is the first asynchronous algorithm for SSP MDP-IPs given in terms of a general set of probability constraints that requires non-linear optimization over imprecise probabilities in the Bellman backup. Our results show up to three orders of magnitude speedup for (L)RTDP-IP when compared with the SPUDD-IP algorithm.

1. Introduction

A Markov Decision Process (MDP) encodes the interaction between an agent and its environment: at every stage, the agent decides to execute an action (with probabilistic effects) which leads to a next state and yields a reward. The agent's objective is to maximize the expected reward though a sequence of actions. MDPs [40] have been used as the standard model for probabilistic planning problems where the uncertainty is represented by a state transition probability matrix for each possible action.

There are three important types of MDPs: finite horizon MDPs, where the agent has to act for $H \neq \infty$ steps; infinite horizon MDPs where the agent has to act for $H = \infty$ steps; and indefinite horizon MDPs where the agent has to act for a finite and unknown number of steps (also known as *Stochastic Shortest Path MDPs – SSP MDPs*). Some of the efficient algorithms to MDPs have sought to exploit the factored structure in their representation [11,31,46,9,30], as well as the initial state information of SSP MDPs with a focus on the solution quality of states via asynchronous dynamic programming [3,8].

* Corresponding author.
E-mail addresses: kvd@usp.br (K.V. Delgado), leliane@ime.usp.br (L.N. de Barros), dbdias@ime.usp.br (D.B. Dias), ssanner@nicta.com.au (S. Sanner).

http://dx.doi.org/10.1016/j.artint.2015.09.005
0004-3702/© 2015 Elsevier B.V. All rights reserved.

However, when addressing many real-world problems, it is simply impossible to obtain a precise representation of the transition probabilities in an MDP. There may be many reasons for this, including (a) imprecise or conflicting elicitations from experts [29], (b) insufficient data to form a basis for estimating precise transition models [23], (c) non-stationary transition probabilities due to insufficient state information, or (d) the occurrence of unpredictable events [51]. In all these situations, the agent can be seen making decisions against the adversarial choices of Nature.

Example 1.1. In an automated navigation system, the probabilities of reaching different locations after a movement, may change in the course of time depending on environmental factors (such as weather and road conditions), which can make the navigation more difficult and subject to failure. In general, it is hard to accurately model all these changes since they can include many external dependencies. In view of this, it would be better to have a policy that is optimized for a range of feasible probabilities so that it can be made robust against transition uncertainty.

Example 1.2. In the field of genetics, a genetic regulatory network and a set of actions (therapeutic interventions) can be modeled as an MDP [20,38,13]. This procedure should prevent the network from moving into undesirable states associated with diseases. However, modeling the exact MDP transition probabilities may not be possible when there are few samples or even when exogenous events occur. In this case, a model with imprecise transitions can be used to compute a robust policy for therapeutic interventions.

The *Markov Decision Process with Imprecise transition Probabilities* (MDP-IP) was introduced [44,50] to accommodate optimal models of sequential decision-making in the presence of constraints on the transition probabilities. An MDP-IP is a sequential decision process endowed with a state space, actions and rewards like any MDP, but where transition probabilities may be imprecisely specified via the *parameterized state transition matrices*. For instance, the probability of moving from state s_1 to state s_2, after executing action a_1 can be given by a parameter p_1 that is subject to a constraint such as $0 \leq p_1 \leq 0.75$. That is, instead of a probability measure we have a set of probability measures, for a fixed state-action pair subject to a set of constraints φ over k parameters denoted by $\vec{p} \in [0,1]^k$.

While the MDP-IP establishes a solid framework for the real-world application of decision-theoretic planning, its solution is extremely time-consuming in practice [23] and involves a complex optimization problem rather than an MDP problem: the agent's goal is to maximize its expected reward during a sequence of actions, by taking into account the worst case scenario of imprecise transition probabilities.

The state-of-the-art algorithms for MDP-IPs are based on factored representations and *synchronous dynamic programming* [44,50,23]. The efficiency of the factored MDP-IP algorithm, SPUDD-IP [23], is due to the use of PADDs [23], algebraic decision diagrams with parameterized expressions on its leaves, which can efficiently compute an exact solution that yields speedups of up to two orders of magnitude with the current exact Value Iteration techniques for MDP-IPs. In this study, our concern is to make further improvements in the performance of approaches for solving MDP-IPs. In particular, we seek to explore asynchronous algorithms for a more general class of MDP-IP, called *Stochastic Shortest Path MDP-IP* (SSP MDP-IP).

An asynchronous dynamic programming algorithm for SSP MDPs [4] of particular interest has been the trial-based real-time dynamic programming (RTDP) [3] as is corroborated by a wide range of recent work [8,37,45,43]. Starting from the initial state, this approach updates sampled states during trials (runs), which are the result of simulating a greedy policy. RTDP algorithms have a number of distinct advantages for practical SSP MDP, which are as follows:

(a) *Anytime performance:* RTDP algorithms can be interrupted at any time, and generally yield a better solution the longer they are allowed to run; and
(b) *Optimality without exhaustive exploration:* By focusing on trial-based searches of reachable states from the initial state and using informed heuristics, RTDP algorithms can obtain an optimal policy while visiting (sampling) only a fraction of the state space.

LRTDP [8] is an extension of RTDP that adds a label to states s when all the states s' that are reachable with the greedy policy from s have their utility estimates changed by less than ϵ.

An efficient asynchronous dynamic programming algorithm for SSP MDP-IPs with enumerated states has been previously proposed although it is restricted to interval-based imprecision [14]. However, in general the problem is given in a factored form, i.e., in terms of state variables and in this case even if an interval-based imprecision is assumed for the variables, the previous algorithm is no longer applicable since there are multilinear parameterized joint transition probabilities and a nonlinear optimization over imprecise probabilities in the Bellman backup. This paper outlines the first asynchronous algorithm for SSP MDP-IPs in terms of a general set of constraints on probabilities that can also be applied to factored SSP MDP-IPs. The challenges of extending the RTDP and LRTDP algorithms to solve SSP MDP-IP problems are: (i) How to ensure the convergence of an asynchronous dynamic programming algorithm for SSP MDP-IPs? (ii) How to sample the next state in a trial given the imprecise probabilities? These can be addressed by making the following innovative contributions:

- We propose the first asynchronous algorithm for SSP MDP-IPs given in terms of a general set of probability constraints that requires nonlinear optimization over imprecise probabilities in the Bellman backup, called (L)RTDP-IP.