Robust Solutions for Hybrid MDPs with State-dependent Noise

Zahra Zamani
ANU & NICTA
Canberra, Australia
zahra.zamani@anu.edu.au

Scott Sanner
NICTA & ANU
Canberra, Australia
ssanner@nicta.com.au

Karina Valdivia Delgado
University of Sao Paulo
Sao Paulo, Brazil
kvd@ime.usp.br

Leliane Nunes de Barros
University of Sao Paulo
Sao Paulo, Brazil
leliane@ime.usp.br

Abstract

Continuous spaces with stochastic dynamics can represent a rich class of real world problems. While decision theoretic planning provides optimal solutions to restricted continuous domains, fully stochastic continuous states with continuous action spaces have not been solved without approximating techniques. Here we propose a symbolic dynamic programming (SDP) approach to provide an optimal closed-form solution for Stochastic Discrete and Continuous MDPs with piecewise linear dynamics. We show how noisy states are symbolically modelled and intuitively minimized in the value iteration algorithm with unknown state parameters. The proposed algorithm uses the extended algebraic decision diagrams (XADDs) as an efficient data structure for SDPs to demonstrate empirical results on problems such as the Inventory Control. The results show the existence of the first fully automated exact solution to the stochastic noisy continuous problem definitions.

1 Introduction

[Sanner *et al.*, 2011]

2 Robust Continuous State and Action MDPs

We first formally introduce the framework of Robust Continuous State and Action Markov decision processes (RCSA-MDPs) extended from CSA-MDPs [**?**]. The optimal solution is then defined by a Robust Dynamic Programming (RDP) approach.

2.1 Factored Representation

A RCSA-MDP is modelled using state variables $(\vec{b}, \vec{x}) = (b_1, \ldots, b_a, x_1, \ldots, x_b)$ where each $b_i \in \{0, 1\}$ $(1 \leq i \leq a)$ represents discrete boolean variables and each $x_j \in \mathbb{R}$ $(1 \leq j \leq b)$ is continuous. Both discrete and continuous actions are represented by the set $A = \{a_1(\vec{y}_1), \ldots, a_p(\vec{y}_d)\}$, where $\vec{y}_k \in \mathbb{R}^{|\vec{y}_k|}$ $(1 \leq k \leq d)$ denote continuous parameters for action a_k.

Given a state $(\vec{b}, \vec{x})$ and an executed action $a(\vec{y})$ at this state, a joint state transition model $P(\vec{b}', \vec{x}' | \vec{b}, \vec{x}, a, \vec{y})$ specifies the probability of the next state $(\vec{b}', \vec{x}')$ and a reward function $R(\vec{b}, \vec{x}, a, \vec{y})$ specifies the immediate reward at this state. To model uncertainty in RCSA-MDPs we assume an error ϵ bounded by some convex region on the state. A noise model $N(\vec{x})$ is defined on the continuous variables.

A policy $\pi(\vec{b}, \vec{x})$ at this state specifies the action $a(\vec{y}) = \pi(\vec{b}, \vec{x})$ to take at this state. An optimal sequence of finite horizon policies $\Pi^* = (\pi^{*,1}, \ldots, \pi^{*,H})$ is desired such that given the initial state $(\vec{b}_0, \vec{x}_0)$ at $h = 0$ and a discount factor $\gamma,\ 0 \leq \gamma \leq 1$, the expected sum of discounted rewards over horizon $h \in H; H \geq 0$ is maximized:

$$V^{\Pi^*}(\vec{b}, \vec{x}) = E_{\Pi^*}\left[\sum_{h=0}^{H} \gamma^h \cdot r^h \Big| \vec{b}_0, \vec{x}_0\right]. \qquad (1)$$

where r^h is the reward obtained at horizon h following the optimal policy.

Similar to the dynamic Bayes net (DBN) structure of CSA-MDPs [**?**] we assume *synchronic arcs* (variables that condition on each other in the same time slice) from $\vec{b}$ to $\vec{x}$ but not within the binary $\vec{b}$ or continuous variables $\vec{x}$.Thus the factorized joint transition model is defined as

$$P(\vec{b}', \vec{x}' | \vec{b}, \vec{x}, a, \vec{y}) = \prod_{i=1}^{n} P(b_i' | \vec{b}, \vec{x}, a, \vec{y}) \prod_{j=1}^{m} P(x_j' | \vec{b}, \vec{b}', \vec{x}, a, \vec{y}) N(\vec{x}).$$

For binary variables b_i $(1 \leq i \leq a)$ the conditional probability $P(b_i' | \vec{b}, \vec{x}, a, \vec{y})$ is defined as conditional probability functions (CPFs). For continuous variables x_j $(1 \leq j \leq b)$, the CPFs $P(x_j' | \vec{b}, \vec{b}', \vec{x}, a, \vec{y})$ are represented with *piecewise linear equations* (PLEs) that condition on the action, current state, and previous state variables with piecewise conditions that may be arbitrary logical combinations of $\vec{b}$, $\vec{b}'$ and linear inequalities over $\vec{x}$. The noise function $N(\vec{x})$ is in form of a bounded error depending only on the current state.

As a simple example, consider the following CPF forms for discrete and continuous variables:

We allow the reward function $R(\vec{b}, \vec{x}, a, \vec{y})$ to be a general piecewise linear function (boolean or linear conditions and

linear values) or a piecewise quadratic function of univariate state and a linear function of univariate action parameters. These constraints ensure piecewise linear boundaries that can be checked for consistency using a linear constraint feasibility checker, which we will later see is crucial for efficiency.

2.2 Robust Dynamic Programming

3 Symbolic Dynamic Programming (SDP)

4 Empirical Results

5 Related Work

6 Concluding Remarks

References

[Sanner *et al.*, 2011] Scott Sanner, Karina Valdivia Delgado, and Leliane Nunes de Barros. Symbolic dynamic programming for discrete and continuous state mdps. In *Proceedings of the 27th Conference on Uncertainty in AI (UAI-2011)*, Barcelona, 2011.