Robust Optimization 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

Recent advances in solutions to Hybrid MDPs with discrete and continuous state and action spaces have significantly extended the class of MDPs for which exact solutions can be derived, albeit at the expense of a restricted transition noise model. In this paper, we work around limitations of previous solutions by adopting a robust optimization approach in which Nature is allowed to adversarially determine transition noise within pre-specified confidence intervals. This allows one to derive an optimal policy with an arbitrary (user-specified) level of success probability and significantly extends the class of transition noise models for which Hybrid MDPs can be solved. This work also significantly extends results for the related "chance-constrained" approach in stochastic hybrid control to accommodate state-dependent noise. We demonstrate our approach working on a variety of hybrid MDPs taken from AI planning, operations research, and control theory, noting that this is the first time optimal robust solutions have been automatically derived for such problems.

1 Introduction

Many real-world sequential decision-making problems are naturally modeled with both discrete and continuous (hybrid) state and action spaces. When state transitions are stochastic, these problems can be modeled as Hybrid Markov Decision Processes (HMDPs), which have been studied extensively in AI planning [Boyan and Littman, 2001; Feng *et al.*, 2004; Li and Littman, 2005; Kveton *et al.*, 2006; Marecki *et al.*, 2007; Meuleau *et al.*, 2009; Zamani *et al.*, 2012] as well as control theory [Henzinger *et al.*, 1997; Hu *et al.*, 2000; De Schutter *et al.*, 2009] and operations research [Puterman, 1994]. However, all previous solutions to hybrid MDPs either take an approximation approach or restrict stochastic noise on continuous transitions to be state-independent or discretized (i.e., requiring continuous transitions to be a finite mixture over deterministic transitions).

Unfortunately, each of these assumptions can be quite limiting in practice when strong *a priori* guarantees on performance are required in the presence of general forms of state-dependent noise. For example, in a UAV NAVIGATION problem [], a human controller must be aware of all positions from which a UAV with a given amount of fuel reserves can return to its landing strip with high probability of success given known areas of (state-dependent) turbulence and weather events. In a SPACE TELESCOPE CONTROL problem [], one must carefully manage inertial moments and rotational velocities as the telescope maneuvers between different angular orientations and zoom positions, where noise margins increase when the telescope is in unstable positions (extended zooms). And lastly, in a RESERVOIR CONTROL problem, one must manage reservoir levels to ensure a sufficient water supply for a population while avoiding overflow conditions subject to uncertainty over daily rainfall amounts. In all of these problems, there is no room for error: a UAV crash, a space telescope spinning uncontrollably, or a flooded reservoir can all cause substantial physical, monetary, and/or environmental damage. What is needed are robust solutions to these problems that are cost-optimal while guaranteed not to exceed a prespecified margin of error.

To achieve cost-optimal robust solutions we build on ideas used in the chance-constrained control literature [Schwarm and Nikolaou, 1999; Li *et al.*, 2002; Ono and Williams, 2008; Blackmore *et al.*, 2011] that maintain confidence intervals on (multivariate) noise distributions and ensure that all reachable states are within these noise margins. However, all previous methods restrict either to linear systems with Gaussian uncertainty and state-independent noise or otherwise resort to approximation techniques. Furthermore, as these works are all inherently focused on control from a given initial state, they are unable to prove properties such as *robust controllability*, i.e., what states have a policy that can achieve a given cost with high certainty over some horizon?

In this work, we adopt a robust optimization receding horizon control approach in which Nature is allowed to adversarially determine transition noise w.r.t. constrained non-deterministic transitions in HMDPs. This permits us to find optimal robust solutions for a wide range of non-deterministic HMDPs and allows us to answer questions of *robust controllability* in very general state-dependent continuous noise settings. Altogether, this work significantly extends previous results in both the HMDP literature in AI and robust hybrid control literature and permits the solution of a new class of robust HMDP control problems.

2 Non-deterministic Hybrid MDPs

We first formally introduce the framework of Hybrid (discrete and continuous) Markov decision processes with non-deterministic continuous noise (ND-HMDPs) by extending the HMDP framework of [Zamani *et al.*, 2012]. The optimal solution for this model is then defined via robust dynamic programming.

2.1 Factored Representation

An HMDP is modelled using state variables $(\vec{b},\vec{x}) = (b_1,\ldots,b_a,x_1,\ldots,x_c)$ where each $b_i \in \{0,1\}$ $(1 \le i \le a)$ represents a discrete boolean variable and each $x_j \in \mathbb{R}$ $(1 \le j \le c)$ is continuous. To model continuous uncertainty in ND-HMDPs we additionally define intermediate noise variables $\vec{n} = n_1,\ldots,n_e$ where each $n_l \in \mathbb{R}$ $(1 \le l \le e)$. Both discrete and continuous actions are represented in the set $A = \{a_1(\vec{y}_1),\ldots,a_p(\vec{y}_p)\}$ where each action $a(\vec{y}) \in A$ references a (possibly empty) vector of continuous parameters $\vec{y} \in \mathbb{R}^{|\vec{y}|}$; we say an action is discrete if it has no continuous parameters ($|\vec{y}| = 0$), otherwise it is continuous.

Given a current state $(\vec{b},\vec{x})$ and next state $(\vec{b}',\vec{x}')$ and an executed action $a(\vec{y})$ at the current state, a real-valued reward function $R(\vec{b},\vec{x},\vec{b}',\vec{x}',a,\vec{y})$ specifies the immediate reward obtained at the current state. The probability of the next state $(\vec{b}',\vec{x}')$ is defined by a joint state transition model $P(\vec{b}',\vec{x}'|\vec{b},\vec{x},a,\vec{y},\vec{n})$ which depends on the current state, action and noise. In a factored setting, we do not typically represent the transition distribution jointly but rather we factorize it into a dynamic Bayes net (DBN) [] as follows:

$$P(\vec{b}',\vec{x}'|\vec{b},\vec{x},a,\vec{y},\vec{n}) = \prod_{i=1}^{a} P(b'_i|\vec{b},\vec{x},\vec{b}',\vec{x}',a,\vec{y},\vec{n}) \prod_{j=1}^{c} P(x'_j|\vec{b},\vec{x},\vec{b}',\vec{x}',a,\vec{y},\vec{n}) \quad (1)$$

Here we allow synchronic arcs under the condition that the DBN forms a proper directed acyclic graph (DAG). For binary variables b_i $(1 \le i \le a)$, $P(b'_i|\vec{b},\vec{x},\vec{b}',\vec{x}',\vec{x},a,\vec{y},\vec{n})$ are defined as general conditional probability functions (CPFs), which are not necessarily tabular since they may condition on inequalities over continuous variables. For continuous variables x_j $(1 \le j \le c)$, the CPFs $P(x'_j|\vec{b},\vec{x},\vec{b}',\vec{x}',a,\vec{y},\vec{n})$ are represented with *piecewise linear equations* (PLEs) that may have piecewise conditions which are arbitrary logical combinations of $\vec{b}$, $\vec{b}'$ and linear inequalities over $\vec{x}$, $\vec{x}'$, and $\vec{n}$. Examples of PLEs will follow shortly.

In general, we assume that for each intermediate continuous noise variable n_l $(1 \le l \le e)$ a non-deterministic noise interval constraint function $N(n_l|\vec{b},\vec{x},a,\vec{y})$ has been defined that represents a range covering α of the probability mass for n_l and evaluates to $-\infty$ for legal values of n_l and $+\infty$ otherwise. The reason for the $\pm\infty$ evaluation is simple: in a robust solution to HMDPs with non-deterministic noise constraints, Nature will attempt to adversarially minimize the reward the agent can achieve and hence we let $N(n_l|\vec{b},\vec{x},a,\vec{y})$ take the value $+\infty$ for illegal values of n_l to ensure Nature will never choose illegal assignments of n_l when minimizing.

As an intuitive example, if $P(n_l|\vec{b},\vec{x},a,\vec{y}) = \mathcal{N}(n_l;\mu;\sigma^2)$ is a simple Normal distribution with mean μ and variance σ^2 and we let $\alpha = 0.95$ then we know that that the 95% of the probability mass lies within $\mu \pm 2\sigma$, hence

$$N(n_l|\vec{b},\vec{x},a,\vec{y}) = \begin{cases} \mu - 2\sigma \le n_l \le \mu + 2\sigma : & -\infty \\ \text{otherwise} : & +\infty \end{cases}.$$

To make the ND-HMDP framework concrete, we now introduce a running example used throughout the paper:

Example (RESERVOIR CONTROL). *The problem of when and how much to drain reservoirs so as to maximize electricity revenue over time while avoiding reservoir overflow and underflow is a major problem in the OR literature [Mahootchi, 2009; Yeh, 1985]. Here we focus on the problem of deciding between* drain *or* not drain *for each reservoir to gain maximum reward over time. While the amount of rain effects the reservoir water levels, it is uncertain due to weather changes.*

For simplicity, we work with a one-level water reservoir. The reward function is used to prevent overflow and underflow by assigning penalty to water levels outside the safe range. For the $drain$ action this is formally defined as:

$$R = \begin{cases} (200 \le l_1 \le 4500) \wedge (200 \le l'_1 \le 4500) & : 1 \\ \text{otherwise} & : -\infty \end{cases}$$

Electricity is generated during the $drain$ action, so a reward equal to the new water level l'_1 is assigned for the safe range of $no\text{-}drain$ action. We assume rainfall occurs on 4 subsequent days, followed by 4 dry days which is modelled by a counter of 3 boolean variables d_i $(i \in \{1,2,3\})$. The piecewise dynamics and noise function are defined using the following PLEs:

$$P(l'_1|l_1,n,d_i,d'_i,a=drain) = \delta\left(l'_1 - (n + l_1 - 2000)\right)$$
$$P(l'_1|l_1,n,d_i,d'_i,a=no-drain) = \delta\left(l'_1 - (n + l_1)\right)$$
$$n = \begin{cases} (d_1) \wedge (1200 \le n \le 2000) & : -\infty \\ \neg(d_1) \wedge (0 \le n \le 400) & : -\infty \\ \text{otherwise} & : +\infty \end{cases}$$

For rainy days, legal noise bounds are on the higher water levels while on dry days it occurs on lower water levels. Here use of the $\delta[\cdot]$ function ensures that the continuous CPF over x' integrates to 1.

A policy $\pi(\vec{b},\vec{x})$ specifies the action $a(\vec{y}) = \pi(\vec{b},\vec{x})$ to take at state $(\vec{b},\vec{x})$. In a robust solution to HMDPs with non-deterministic noise constraints, 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 $\gamma,\ 0 \le \gamma \le 1$, the expected sum of discounted rewards over horizon $h \in H$ $(H \ge 0)$ is maximized subject to Nature's adversarial attempt to choose value minimizing assignments of the noise variables. The value function V w.r.t. Π^* in this case is defined via a recursive expectation

$$V^{\Pi^{*,H}}(\vec{b},\vec{x}) = \min_{\vec{n}} \max\Big(N(n_1|\vec{b},\vec{x},\Pi^{*,H}),\ldots, \max\Big(N(n_e|\vec{b},\vec{x},\Pi^{*,H}), E_{\Pi^{*,H}}\Big[r^h + \gamma V^{\Pi^{*,H-1}}(\vec{b}',\vec{x}')\Big|\vec{b}_0,\vec{x}_0\Big]\Big)\cdots\Big)$$

where r^h is the reward obtained at horizon h following policy Π^* and using Nature's minimizing choice of $\vec{n}$ at each h.

The effect of "max'ing" in each of the previously defined $N(n_l|\vec{b},\vec{x},a,\vec{y})$ $(1 \leq l \leq e)$ with the value function is one of the major insights and contributions of this paper. We noted before that Nature will never choose an illegal value of n_l where $N(n_l|\vec{b},\vec{x},a,\vec{y}) = +\infty$, instead it will choose a legal value of n_l for which $N(n_l|\vec{b},\vec{x},a,\vec{y}) = -\infty$ which when "max'ed" in with the value function effectively vanishes owing to the indentity $\max(v,-\infty) = v$ for all $v > -\infty$.

Finally, by leveraging the simple union bound, we can easily prove that that a policy will achieve $V^{\Pi^{*,H}}$ with at least $1-H(1-\alpha)$ probability since the probability of encountering a noise value outside the confidence interval is only $(1-\alpha)$ at any time step. Hence for a success probability of at least β, one should choose $\alpha = 1 - \frac{1-\beta}{H}$, e.g., $\beta = 0.95$ success probability requires an $\alpha = 0.99$ for $H = 5$.

2.2 Robust Dynamic Programming

We extend the value iteration dynamic programming algorithm [Bellman, 1957] and specifically the form used for HMDPs in [Zamani *et al.*, 2012] to a robust dynamic programming (RDP) algorithm for ND-HMDPs that may be considered a continuous action generalization of zero-sum alternating turn Markov games [Littman, 1994]. Initializing $V^0(\vec{b},\vec{x}) = 0$ the algorithm builds the h-stage-to-go value function $V^h(\vec{b},\vec{x})$.

The quality $Q_a^h(\vec{b},\vec{x},\vec{y},\vec{n})$ of taking action $a(\vec{y})$ in state $(\vec{b},\vec{x})$ with noise parameters $\vec{n}$ and acting so as to obtain $V^{h-1}(\vec{b}',\vec{x}')$ thereafter is defined as the following:

$$Q_a^h(\vec{b},\vec{x},\vec{y},\vec{n}) = \max\Big(N(n_1|\vec{b},\vec{x},\Pi^{*,H}),\ldots,\max\Big(N(n_e|\vec{b},\vec{x},\Pi^{*,H}),$$
$$\sum_{\vec{b}'}\int\prod_{i=1}^{a} P(b_i'|\vec{b},\vec{x},\vec{b}',\vec{x}',a,\vec{y},\vec{n})\prod_{j=1}^{c} P(x_j'|\vec{b},\vec{x},\vec{b}',\vec{x}',a,\vec{y},\vec{n})$$
$$\Big[R(\vec{b},\vec{x},\vec{b}',\vec{x}',a,\vec{y}) + \gamma V^{h-1}(\vec{b}',\vec{x}')d\vec{x}'\Big]\Big)\cdots\Big) \quad (2)$$

Here the noise constraints $N(\vec{n},\vec{b},\vec{x})$ are "max'ed" in with the value function to ensure Nature chooses a legal setting of n_l, effectively reducing each $\max$ to an identity operation.

Next, given $Q_a^h(\vec{b},\vec{x},\vec{y},\vec{n})$ as above for each $a \in A$, we can proceed to define the h-stage-to-go value function assuming that the agent attempts to maximize value subject to Nature's adversarial choice of value-minimizing noise:

$$V^h(\vec{b},\vec{x}) = \max_{a\in A}\max_{\vec{y}\in\mathbb{R}^{|\vec{y}|}}\min_{\vec{n}\mathbb{R}^{|\vec{e}|}}\Big\{Q_a^h(\vec{b},\vec{x},\vec{y},\vec{n})\Big\} \quad (3)$$

The optimal policy at horizon h can also be determined using the Q-function as below:

$$\pi^{*,h}(\vec{b},\vec{x}) = \arg\max_{a\in A}\arg\max_{\vec{y}\in\mathbb{R}^{|\vec{y}|}}\min_{\vec{n}\mathbb{R}^{|\vec{e}|}} Q_a^h(\vec{b},\vec{x},\vec{y},\vec{n}) \quad (4)$$

For finite-horizon HMDPs the optimal value function and policy are obtained up to horizon H. For infinite horizons where the optimal policy has finitely bounded value then value iteration terminates when two values are equal in subsequent horizons ($V^h = V^{h-1}$). In this case $V^\infty = V^h$ and $\pi^{*,\infty} = \pi^{*,h}$.

Up to this point we have only provided the abstract mathematical framework for ND-HMDPs and RDP. Fortuitously though, we can leverage the continuous $\max$ (and analoguously defined $\min$) operations and symbolic DP approach of [Zamani *et al.*, 2012] in order to compute RDP via (2) and (3) exactly in closed-form. We discuss this next.

3 Robust Symbolic Dynamic Programming

In order to compute the equations above, we propose a *robust symbolic dynamic programming* (RSDP) approach building on the work of [**?**; Sanner *et al.*, 2011]. This requires a value iteration algorithm described in Algorithm 1 (VI) and the regression subroutine described in Algorithm 2. In what follows we show how the techniques of SDP can be extended to compute RDP exactly in closed-form as discussed in the last section.

In general we define *all* symbolic functions to be represented in *case* form [Boutilier *et al.*, 2001] for which a binary "cross-sum" operation can be defined as follows:

$$\begin{cases}\phi_1: & f_1\\ \phi_2: & f_2\end{cases} \oplus \begin{cases}\psi_1: & g_1\\ \psi_2: & g_2\end{cases} = \begin{cases}\phi_1\wedge\psi_1: & f_1+g_1\\ \phi_1\wedge\psi_2: & f_1+g_2\\ \phi_2\wedge\psi_1: & f_2+g_1\\ \phi_2\wedge\psi_2: & f_2+g_2\end{cases}$$

Here ϕ_i and ψ_j are logical formulae defined over the state $(\vec{b},\vec{x})$ and can include arbitrary logical $(\wedge,\vee,\neg)$ combinations of boolean variables and *linear* inequalities $(\geq,>,\leq,<)$ over continuous variables – we call this *linear case form* (LCF). The f_i and g_j are restricted to be *linear* functions. Similarly operations such as $\ominus$ and $\otimes$ may be defined with operations applied to LCF functions yielded LCF results.

ZAHRA TODO: define casemax here, then introduce remaining operations as used to compute results for Reservoir running example getting from V^0 to V^1... showing whatever you feel is most important by the end of page 4

NOTE: need to rewrite a lot this section from scratch... try not to be too verbose... we've already said why we do each operation so you just have to remark on the running example and what the result of each iteration says... don't explain the operations again unless something is really non-obvious... assume the reader can easily figure out the mechanical details of the derivation

While explaining the steps of the VI algorithm we show that all operations required to compute the optimal policy is supported by case representation and defined symbolically.

Initially the value function V^h is assigned to 0. For every horizon the h-stage-to-go value functions $V^h(\vec{b},\vec{x})$ is computed. To follow the steps, we use the second iteration of the Rover example here. For simplification, we omit the boolean variable b of taking a picture and only use one noisy continuous variable x. We now perform steps 1–4 for $h = 2$.

(1) For every action, the function Q_a^h is computed. Line 6 refers to Algorithm 2 which has the main steps below:

Algorithm 1: VI(CSA-MDP, H) $\longrightarrow (V^h, \pi^{*,h})$

1 **begin**
2 $\quad V^0 := 0, h := 0$
3 $\quad$ **while** $h < H$ **do**
4 $\quad\quad h := h + 1$
5 $\quad\quad$ **foreach** $a(\vec{y}) \in A$ **do**
6 $\quad\quad\quad Q_a^h(\vec{y}, \vec{n}) :=$ Regress $(V^{h-1}, a, \vec{y})$
7 $\quad\quad\quad Q_a^h(\vec{y}) := \min_{\vec{n}} Q_a^h(\vec{y}, \vec{n})$ *//Stochastic* min
8 $\quad\quad\quad Q_a^h := \max_{\vec{y}} Q_a^h(\vec{y})$ *// Continuous* max
9 $\quad\quad\quad V^h :=$ casemax$_a$ Q_a^h // casemax *all* Q_a
10 $\quad\quad\quad \pi^{*,h} := \arg\max_{(a,\vec{y})} Q_a^h(\vec{y})$
11 $\quad\quad$ **if** $V^h = V^{h-1}$ **then**
12 $\quad\quad\quad$ break *// Terminate if early convergence*
13
14 $\quad$ **return** $(V^h, \pi^{*,h})$
15 **end**

Algorithm 2: Regress($V, a, \vec{y}$) $\longrightarrow Q$

1 **begin**
2 $\quad Q =$ Prime (V) *// All* $b_i \to b_i'$ *and all* $x_i \to x_i'$
3 $\quad$ **ZAHRA TODO: Need to change a little now that we allow synchronic arcs... change to**
4 $\quad$ **for all** $v' \in Q$
5 $\quad$ ***child variables must come parents in order***
6 $\quad$ **if** $(v' = x_j')$ **then do continuous regression**
7 $\quad$ **else if** $(v' = b_i')$ **then do discrete regression**
8 $\quad$ **... just need to reformat what is below to fit this new control flow** *// Continuous regression marginal integration*
9 $\quad$ **for** *all* x_j' *in* Q **do**
10 $\quad\quad Q := \int Q \otimes P(x_j'|\vec{b}, \vec{b}', \vec{x}, a, \vec{y}, \vec{n})\, d_{x_j'}$
11 $\quad$ *// Discrete regression marginal summation*
12 $\quad$ **for** *all* b_i' *in* Q **do**
13 $\quad\quad Q := \left[Q \otimes P(b_i'|\vec{b}, \vec{x}, a, \vec{y}, \vec{n})\right]|_{b_i'=1}$
14 $\quad\quad\quad \oplus \left[Q \otimes P(b_i'|\vec{b}, \vec{x}, a, \vec{y}, \vec{n})\right]|_{b_i'=0}$
15 $\quad Q := R(\vec{b}, \vec{x}, a, \vec{y}) \oplus (\gamma \cdot Q)$
16 $\quad$ // max*-in noise variables*
17 $\quad$ **for** *all* n_k *in* Q **do**
18 $\quad\quad Q_a^h(\vec{y}, \vec{n}) :=$ casemax$_{n_k}$ $(Q, N(n_k, b_i', x_j'))$
19 $\quad$ **return** Q
20 **end**

(1a) Priming the current state variables (b_i, x_j) to build the next states (b_i', x_j') in the value function. This indicates that any occurrence of the current state in V^h is *symbolically substituted* with the next state variables that is $V'^h = V^h\sigma$ where $\sigma = \{b_i \setminus b_i', x_j \setminus x_j'\}$ for all values of i and j. For the second iteration this step is equal to the following:

$$Q = V'^1 = \begin{cases} -40 \le x' \le 40 : & 10 \\ x > 40 \vee x' < -40 : & -\infty \end{cases}$$

(1b) Performing Regression for continuous variable x in line 5 and boolean variable b in lines 8–9. *Boolean restriction* $f|_{b=v}$ assigns the value $v \in \{0,1\}$ to any occurrence of b in f.

Continuous integration of $\int Q(x_j') \otimes P(x_j'|\cdots)dx_j'$ where results in the following according to the rules of integration:

$$\int f(x_j') \otimes \delta[x_j' - h(\vec{z})]dx_j' = f(x_j')\{x_j'/h(\vec{z})\}$$

Here $P(x_j'|\cdots)$ is in the form of $\delta[x_j' - h(\vec{z})]$ ($h(\vec{z})$ which is a case statement and $\vec{z}$ does not contain x_j'). The latter operation in the result indicates that any occurrence of x_j' in $f(x_j')$ is substituted with the case statement $h(\vec{z})$. For our example this step results in the following intermediate Q-value:

$$\begin{cases} -40 \le x' \le 40 : & 10 \\ x > 40 \vee x' < -40 : & -\infty \end{cases} \otimes \delta\left(x' - (x + n + a)\right) =$$

$$\begin{cases} -40 \le (x + a + n) \le 40 : & 10 \\ (x + n + a) > 40 \vee (x + n + a) < -40 : & -\infty \end{cases}$$

(1c) Multiplying the regression by the discount factor and adding the reward function in line 10. *Unary operations* such as scalar multiplication $\gamma \cdot Q$ (and also negation $-Q$) on case statements Q is performed by applying the operation to each Q_i ($1 \le i \le k$) while adding the reward is a binary $\oplus$:

$$Q = V'^1 = \begin{cases} -40 \le (x + n + a) \le 40 : & 20 \\ (x + n + a) > 40 \vee (x + n + a) < -40 : & -\infty \\ x > 40 \vee x < -40 : & -\infty \end{cases}$$

(1d) Maximizing this result with the noise function in line 13. This step incorporates noise into the regressed Q-function consequently for each noise variable. Each noise variable assigns -∞ for legal values inside the boundary range +∞ for illegal values defined by the noise model $N(\vec{n}, \vec{b}, \vec{x})$. By maximizing in n_k all illegal values will remain +∞ since this is the maximum value compared to any other value and all legal values will be replaced by the regressed Q-value defined in step (1c) -∞ is less than any other Q-value so it is omitted in the maximization. The Rover example is redefined with this noise variable as below:

$$Q = \begin{cases} -5 \le n \le 5 : & +\infty \\ (-40 \le (x + n + a) \le 40) \wedge \neg(-5 \le n \le 5) : & 20 \\ \neg(-40 \le (x + n + a) \le 40) \wedge \neg(-5 \le n \le 5) : & -\infty \\ (x > 40 \vee x < -40) \wedge (-5 \le n \le 5) : & +\infty \\ (x > 40 \vee x < -40) \wedge \neg(-5 \le n \le 5) : & -\infty \end{cases}$$

(2) Naturally a noisy process aims to minimize the noise to reach robustness Thus the regressed stochastic $Q_a^h(\vec{y}, \vec{n})$ from Algorithm 2 is now minimized over the noise variables $\vec{n}$ in line 7. Intuitively this continuous minimization will never choose +∞ as there is always some value

smaller which insures that the transitioned model never chooses illegal values. All legal Q-values are considered in the minimization step to find the value corresponding to the minimum noise. Each partition i of this intermediate Q is considered for a continuous minimization separately with the final result a $\mathrm{casemin}$ on all the individual minimum results $\mathrm{casemin}_i \min_n \phi_i(\vec{b},\vec{x},\vec{n}) f_i(\vec{b},\vec{x},\vec{n})$. We demonstrate the steps of this algorithm for the second partition of Q defined as:

$$\min_n (-40 \leq (x+n+a) \leq 40) \wedge \neg(-5 \leq n \leq 5) : 20$$

For each partition the logical constraints are used to derive the (a) lower bound on n (LB = -5, $-40-a-x$); (b) upper bound on n (UB = 5, $40-a-x$) and (c) constraints independent of n (IND). In case of several bounds on n the maximum of all lower bounds and the minimum of all upper bounds is desired. A *symbolic case maximization* on two case statements of $(\phi_i : f_i)$ and (ψ_i, g_i) where ($i \in \{1,2\}$) is performed below.

$$\mathrm{casemax} = \begin{cases} \phi_1 \wedge \psi_1 \wedge f_1 > g_1 : f_1 \\ \phi_1 \wedge \psi_1 \wedge f_1 \leq g_1 : g_1 \\ \phi_1 \wedge \psi_2 \wedge f_1 > g_2 : f_1 \\ \phi_1 \wedge \psi_2 \wedge f_1 \leq g_2 : g_2 \\ \phi_2 \wedge \psi_1 \wedge f_2 > g_1 : f_2 \\ \phi_2 \wedge \psi_1 \wedge f_2 \leq g_1 : g_1 \\ \phi_2 \wedge \psi_2 \wedge f_2 > g_2 : f_2 \\ \phi_2 \wedge \psi_2 \wedge f_2 \leq g_2 : g_2 \end{cases}$$

Thus the bounds are defined as below:

$$LB = \begin{cases} a+x < -35 : & -40-x-a \\ a+x \geq -35 : & -5 \end{cases}$$
$$UB = \begin{cases} a+x < 35 : & 5 \\ a+x \geq -35 : & 40-a-x \end{cases}$$

The minima points of upper and lower bounds are evaluated for the leaf value which equals to substituting the bounds instead of the noise variable n in the leaf function. The minimum of these two evaluations are then stored, note that in our example the leaf is a constant 20 value which is not effected by this step.
Natural constraints on bounds $LB \leq UB$ and the IND constraints are also considered for the minimization on a single partition to obtain:

$$Q = \begin{cases} (-40 \leq x \leq 40) \wedge (-45 \leq (x+a) \leq 45) : & 20 \\ (-40 \leq x \leq 40) \wedge \neg(-45 \leq (x+a) \leq 45) : & +\infty \\ (x > 40 \vee x < -40) : & +\infty \end{cases}$$

The final result of a continuous minimization is a $\mathrm{casemin}$ over all partitions which results in the following Q-value:

$$Q = \begin{cases} (-40 \leq x \leq 40) \wedge (-35 \leq (x+a) \leq 35) : & 20 \\ (-40 \leq x \leq 40) \wedge \neg(-35 \leq (x+a) \leq 35) : & -\infty \\ (x > 40 \vee x < -40) : & -\infty \end{cases}$$

3 The resulting Q-value with minimal noise is maximized over the continuous action parameter in line 8; a symbolic continuous maximization operation, the major contribution of [Zamani *et al.*, 2012]. The resulting Q_a^h can be determined as the following:

$$Q = \begin{cases} (-40 \leq x \leq 40) : & 20 \\ (x > 40 \vee x < -40) : & -\infty \end{cases}$$

4 A discrete $\mathrm{casemax}$ on the set of discrete actions for all Q-functions defines the final V and the optimal policy is defined as the $\arg\max$ over the set of discrete and continuous actions on Q. In our example the final value $V^h = Q^h$ because there is only one single discrete action.

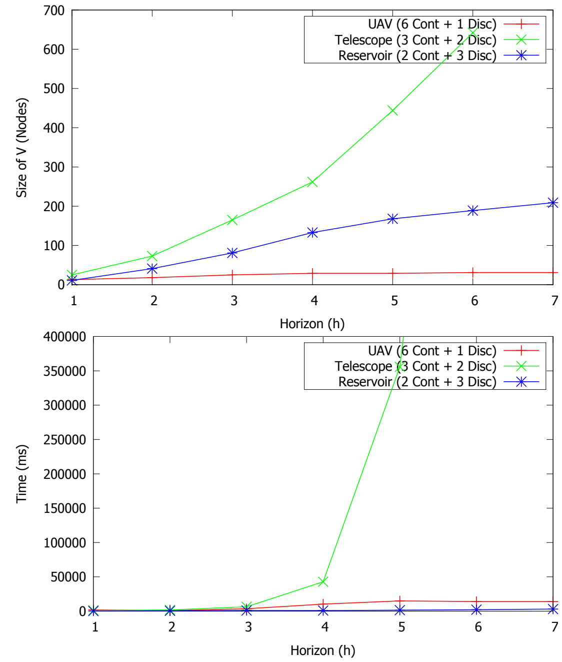

Figure 2: Space and elapsed time (between current and previous horizon) vs. horizon.

To implement the case statements efficiently with continuous variables, extended Algebraic Decision diagrams (XADDs) are used from [Sanner *et al.*, 2011] which is extended from ADDs [Bahar *et al.*, 1993]. Unreachable paths can be pruned in XADDs using LP solvers and all operations including the continuous minimization can be defined using XADDs by treating each path from root to leaf node as a single case partition with conjunctive constraints, $\min_n$ is performed at each leaf subject to these constraints and all path $\min_n$'s are then accumulated via the $\mathrm{casemin}$ operation to obtain the final result.

4 Empirical Results

We evaluated RH-MDP using XADDs on the RESERVOIR CONTROL problem used as a running example, UAV NAVIGATION problem and SPACE TELESCOPE CONTROL problem — described below.[1]

[1]While space limitations prevent a self-contained descrip-

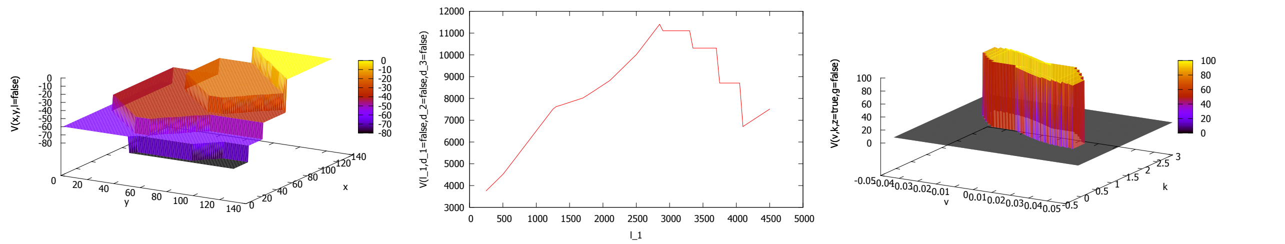

Figure 1: *(left)* $V^4(x, y, l = false)$UAV NAVIGATION problem; *(middle)* $V^7(l_1, d_1 = false, d_2 = false, d_3 = false)$ RESERVOIR CONTROL problem; *(right)* $V^4(k, v, z = true, g = false)$ SPACE TELESCOPE CONTROL problem.

UAV NAVIGATION: The state consist of UAVs continuous position x and y. In a given time step, the UAV may move a continuous distance $ax \in [-40, 40]$ and $ay \in [-40, 40]$. The turbulence introduces a noise nx and ny respectively in the movement, given by:

$$nx = \begin{cases} (y \geq 50 + x) \wedge (nx \leq -20) \wedge (nx \geq 20) & : legal \\ (y < 50 + x) \wedge (nx \leq -5) \wedge (nx \geq 5) & : legal \\ else & : illegal \end{cases}$$

$$ny = \begin{cases} (y \geq 50 + x) \wedge (ny \leq -20) \wedge (ny \geq 20) & : legal \\ (y < 50 + x) \wedge (ny \leq -5) \wedge (ny \geq 5) & : legal \\ else & : illegal \end{cases}$$

The UAV goal is to achieve the region $x + y > 200$. It receives a reward penalty $(-\infty)$ for being in positions from which a UAV with a given amount of fuel reserves cannot return to its landing strip and, if the UAV is not in the goal position $(\neg l)$, the action cost is 20:

$$R = \begin{cases} (l) \wedge (x \leq 130) \wedge (y \leq 130) \wedge (x \geq 0) \wedge (y \geq 0) & : 0 \\ (\neg l) \wedge (x \leq 130) \wedge (y \leq 130) \wedge (x \geq 0) \wedge (y \geq 0) & : -20 \\ else & : -\infty \end{cases}$$

SPACE TELESCOPE CONTROL: We have extended the problem of slewing a space telescope in order to look a new objective given in [Löhr *et al.*, 2012]. This problem has six actions $a_0, \cdots, a_5$ that change the angle k and the angular rate v. The transition function for a_5 action, when $v < 1\ \frac{deg}{seg}$ and the $z = false$ is:

$$\begin{aligned} k' &= (k + 40.55 * v) \\ v' &= (2/3v + n) \\ z' &= (true), \end{aligned}$$

Note that we assume a noise in the transition function of the angular rate for a_5, since this action is the only one that changes the zoom of the telescope during the slew. The noise is given by:

$$n = \begin{cases} \neg(z) \wedge (n \leq 0.04 * v) \wedge (n \geq -0.04 * v) & : legal \\ else & : illegal \end{cases}$$

The reward is

$$R = \begin{cases} (z) \wedge (v \leq 0.02) \wedge (k \leq 1.683) \wedge (v \geq -0.02) \wedge (k \geq 1.283) & : 100 \\ else & : -cost \end{cases}$$

where cost is 0 for action a_0, 1 for actions a_i $i \in \{1, 2, 3, 4\}$ and 10 for action a_5.

5 Related Work

This work extends results in HMDP in AI [Boyan and Littman, 2001; Feng *et al.*, 2004; Li and Littman, 2005; Kveton *et al.*, 2006; Marecki *et al.*, 2007; Meuleau *et al.*, 2009; Zamani *et al.*, 2012] and hybrid system control literature [Henzinger *et al.*, 1997; Hu *et al.*, 2000; De Schutter *et al.*, 2009] to handled state-dependent noise.

In the hybrid control literature, a challenging topic is to solve the controllability problem that is NP hard [Blondel and Tsitsiklis, 1999]. A hybrid system is called hybrid controllable if, for any pair of valid states, there exists at least one permitted control sequence (correct control-laws) between them [Tittus and Egardt, 1998; Yang and Blanke, 2007]. Another challenging topic for stochastic hybrid systems, a class of hybrid systems that allows uncertainty, is tried to maximize the probability that the execution will remain in safe states as long as possible [Hu *et al.*, 2000]. This work is related with both topics, however we want to answer a slightly different question, called the robust controllability problem: what states have a policy to achieve a goal (that can be modeled as a reward or cost function) with high certainty over some horizon? To the authors knowledge, in the control area there are few results to answer a similar question except in the chance-constrained predictive stochastic sub-area, that finds the optimal sequence of control inputs subject to the constraint that the probability of failure must be below a user-specified threshold [Blackmore *et al.*, 2011]. However all the previous work in this sub-are is focused on linear systems subject to Gaussian uncertainty and state-independence noise [Schwarm and Nikolaou, 1999; Li *et al.*, 2002; Ono and Williams, 2008; Blackmore *et al.*, 2011] or resort to approximation techniques [Blackmore *et al.*, 2010]. Different from them, our approach is not approximated and can solve problems with state-dependent noise that is crucial in settings like autonomous vehicles, satellite maneuvers and environmental control systems.

tion of all domains, we note that all Java source code and a human/machine readable file format for all domains needed to reproduce the results in this paper can be found online at `http://code.google.com/p/xadd-inference`.

6 Concluding Remarks

This work has combined symbolic techniques and data structures from the HMDP literature in AI with techniques from chance-constrained control theory to provide optimal robust solutions to a range of problems with general continuous transitions and state-dependent noise for which no general exact closed-form solutions previously existed. Using these techniques we were able to find optimal policies and answer questions of robust controllability for a variety of highly risk-sensitive applications from AI planning, control theory, and operations research such as UAV NAVIGATION, SPACE TELESCOPE CONTROL, and RESERVOIR CONTROL. Among many potential avenues for future work, combining this receding horizon control approach with focused search techniques as in HAO* [Meuleau *et al.*, 2009] should preserve our strong robust optimality guarantees while substantially increasing the scalability of our approach in exchange for restricting solution optimality to a known set of initial states.

References

[Bahar *et al.*, 1993] R. Iris Bahar, Erica Frohm, Charles Gaona, Gary Hachtel, Enrico Macii, Abelardo Pardo, and Fabio Somenzi. Algebraic Decision Diagrams and their applications. In *IEEE /ACM International Conference on CAD*, 1993.

[Bellman, 1957] Richard E. Bellman. *Dynamic Programming*. Princeton University Press, Princeton, NJ, 1957.

[Blackmore *et al.*, 2010] L. Blackmore, M. Ono, A. Bektassov, and B.C. Williams. A probabilistic particle-control approximation of chance-constrained stochastic predictive control. *Robotics, IEEE Transactions on*, 26(3):502 –517, 2010.

[Blackmore *et al.*, 2011] Lars Blackmore, Masahiro Ono, and Brian C. Williams. Chance-constrained optimal path planning with obstacles. *IEEE Transactions on Robotics*, 27(6):1080–1094, 2011.

[Blondel and Tsitsiklis, 1999] Vincent D. Blondel and John N. Tsitsiklis. Complexity of stability and controllability of elementary hybrid systems. *Automatica*, 35(3):479–489, 1999.

[Boutilier *et al.*, 2001] Craig Boutilier, Ray Reiter, and Bob Price. Symbolic dynamic programming for first-order MDPs. In *IJCAI-01*, pages 690–697, Seattle, 2001.

[Boyan and Littman, 2001] Justin Boyan and Michael Littman. Exact solutions to time-dependent MDPs. In *Advances in Neural Information Processing Systems NIPS-00*, pages 1026–1032, 2001.

[De Schutter *et al.*, 2009] B. De Schutter, W.P.M.H. Heemels, J. Lunze, and C. Prieur. Survey of modeling, analysis, and control of hybrid systems. In J. Lunze and F. Lamnabhi-Lagarrigue, editors, *Handbook of Hybrid Systems Control – Theory, Tools, Applications*, chapter 2, pages 31–55. Cambridge University Press, Cambridge, UK, 2009.

[Feng *et al.*, 2004] Zhengzhu Feng, Richard Dearden, Nicolas Meuleau, and Richard Washington. Dynamic programming for structured continuous markov decision problems. In *Uncertainty in Artificial Intelligence (UAI-04)*, pages 154–161, 2004.

[Henzinger *et al.*, 1997] Thomas A. Henzinger, Pei H. Ho, and Howard W. Toi. HYTECH: A Model Checker for Hybrid Systems. *International Journal on Software Tools for Technology Transfer*, 1(1-2):110–122, 1997.

[Hu *et al.*, 2000] J. Hu, John Lygeros, and S. Sastry. Towards a theory of stochastic hybrid systems. *Lecture Notes in Computer Science LNCS*, 1790:160–173, 2000.

[Kveton *et al.*, 2006] Branislav Kveton, Milos Hauskrecht, and Carlos Guestrin. Solving factored mdps with hybrid state and action variables. *Journal Artificial Intelligence Research (JAIR)*, 27:153–201, 2006.

[Li and Littman, 2005] Lihong Li and Michael L. Littman. Lazy approximation for solving continuous finite-horizon mdps. In *National Conference on Artificial Intelligence AAAI-05*, pages 1175–1180, 2005.

[Li *et al.*, 2002] Pu Li, Moritz Wendt, and GüNter Wozny. Brief a probabilistically constrained model predictive controller. *Automatica*, 38(7):1171–1176, 2002.

[Littman, 1994] Michael L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *ICML*, pages 157–163, 1994.

[Löhr *et al.*, 2012] Johannes Löhr, Patrick Eyerich, Thomas Keller, and Bernhard Nebel. A planning based framework for controlling hybrid systems. In *Proceedings of the Twenty-Second International Conference on Automated Planning and Scheduling*, 2012.

[Mahootchi, 2009] Masoud Mahootchi. *Storage System Management Using Reinforcement Learning Techniques and Nonlinear Models*. PhD thesis, University of Waterloo,Canada, 2009.

[Marecki *et al.*, 2007] Janusz Marecki, Sven Koenig, and Milind Tambe. A fast analytical algorithm for solving markov decision processes with real-valued resources. In *International Conference on Uncertainty in Artificial Intelligence IJCAI*, pages 2536–2541, 2007.

[Meuleau *et al.*, 2009] Nicolas Meuleau, Emmanuel Benazera, Ronen I. Brafman, Eric A. Hansen, and Mausam. A heuristic search approach to planning with continuous resources in stochastic domains. *Journal Artificial Intelligence Research (JAIR)*, 34:27–59, 2009.

[Ono and Williams, 2008] Masahiro Ono and Brian C. Williams. An efficient motion planning algorithm for stochastic dynamic systems with constraints on probability of failure. In *AAAI*, pages 1376–1382, 2008.

[Puterman, 1994] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, New York, 1994.

[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.

[Schwarm and Nikolaou, 1999] Alexander T. Schwarm and Michael Nikolaou. Chance-constrained model predictive control. *AIChE Journal*, 45(8):1743–1752, 1999.

[Tittus and Egardt, 1998] M. Tittus and B. Egardt. Control design for integrator hybrid systems. *Automatic Control, IEEE Transactions on*, 43(4):491 –500, apr 1998.

[Yang and Blanke, 2007] Zhenyu Yang and Mogens Blanke. A unified approach to controllability analysis for hybrid control systems. *Nonlinear Analysis: Hybrid Systems*, 1(2):212 – 222, 2007.

[Yeh, 1985] William G Yeh. Reservoir management and operations models: A state-of-the-art review. *Water Resources research*, 21,12:17971818, 1985.

[Zamani *et al.*, 2012] Z. Zamani, S. Sanner, and C. Fang. Symbolic dynamic programming for continuous state and action mdps. In *In Proceedings of the 26th AAAI Conference on Artificial Intelligence (***AAAI-12***)*, Toronto, Canada, 2012.