Proceedings Paper:
Dimitrova, R., Fu, J. and Topcu, U. (2016) Robust optimal policies for Markov decision processes with safety-threshold constraints. In: 2016 IEEE 55th Conference on Decision and Control (CDC). 2016 IEEE 55th Conference on Decision and Control (CDC), 12-14 Dec 2016, Las Vegas, NV, USA. IEEE , pp. 7081-7086. ISBN 9781509018383

https://doi.org/10.1109/cdc.2016.7799360

Robust optimal policies for Markov decision processes with safety-threshold constraints

Rayna Dimitrova[†], Jie Fu[*], Ufuk Topcu[‡]

***Abstract*— We study the synthesis of robust optimal control policies for Markov decision processes with transition uncertainty (UMDPs) and subject to two types of constraints: (i) constraints on the worst-case, maximal total cost and (ii) safety-threshold constraints that bound the worst-case probability of visiting a set of error states. For maximal total cost constraints, we propose a state-augmentation method and a two-step synthesis algorithm to generate deterministic, memoryless optimal policies given the reward to be maximized. For safety threshold constraints, we introduce a new cost function and provide an approximately optimal solution by a reduction to an uncertain Markov decision process under a maximal total cost constraint. The safety-threshold constraints require memory and randomization for optimality. We discuss the use and the limitations of the proposed solution.**

I. Introduction

Markov decision processes (MDPs) are important for modeling and control synthesis of stochastic systems. In practice, the transition kernels of Markov decision process (MDP)s are often estimated from data or have unknown but bounded parameters. For safety-critical systems, for instance a robot with bounded resources, such as battery, the system not only needs to perform reasonably well with respect to its task, but also must not exhaust the resource under worst-case uncertainty. Such resource constraints are modeled by defining a cost function and a bound on the total (discounted or non-discounted) cost in the worst-case realization of model uncertainty. The control objective is to synthesize a policy that maximizes the expected total reward in the worst case, while satisfying the cost constraint.

For MDPs with known transition kernel, similar synthesis problems have been extensively studied in planning for constrained MDP or cost-sensitive MDPs [1], [7], [13], [9], [12]. Different definitions of cost constraints lead to different solution approaches: For constraints on the expected total cost, formulations based on convex optimization were proposed in [1], for constraints on the maximal total cost under any possible execution, a dynamic programming approach was developed in [7], [13]. However, none of these methods is robust in the presence of modeling uncertainty. On the other hand, for MDPs with uncertain parameters, robust MDPs have been extensively studied [10], [8], [15]. Recently, robust control of MDPs has been extended to handle expressive temporal logic constraints [16], [3], [14]. A robust adaptive control design for uncertain MDP is developed [6] based on a robust(min-max) value iteration with estimates of transition probabilities. These approaches are based on robust policy or value iteration.

Our notion of cost constraints is most similar to that in [7], [13] in which the authors define *maximal total cost constraints* (or minimax constraints) so as to bound the total cost under the worst-case path of a policy. Inspired by their work, we use a state-augmentation method to transform planning in uncertain MDPs with maximal total cost constraints into a robust dynamic programming problem. After applying this transformation, robust policy and value iteration procedures become applicable. It is noted that the penalty method in [13] does not exclude the chance of constraint violation: One has to assign reward $-\infty$ to do so. We propose a two-step method: In the first step, the feasible set of memoryless policies is computed, by computing a strategy in a safety game on the graph of the uncertain MDP [4]. Within the set of feasible policies, we compute the optimal policy with respect to the given reward criterion.

In addition to maximal total cost constraints, we also consider *safety-threshold constraints*. These are a special case of expected total cost constraints, where we wish to enforce a bound on the probability of reaching a set of error states. We investigate the problem of computing a robust optimal policy under safety-threshold constraints. We propose a method to approximate the optimal solution by introducing a new cost function and applying the two-step algorithm. This approach is inherently conservative, since under safety-threshold constraints memory and randomization are needed to achieve optimality. We discuss this conservativeness and highlight its effects using several examples.

II. Preliminaries and problem formulation

A. Definitions

For a finite set $X = \{1, \ldots, |X|\}$, $\mathcal{D}(X)$ denotes the probability simplex in $\mathbb{R}^{|X|}$. Given a distribution $\mu \in \mathcal{D}(X)$, $\mathsf{Supp}(\mu) = \{x \in X \mid \mu(x) \neq 0\}$ is the support of μ.

We define uncertain Markov decision process (UMDP)s following the notation and formulation in [15]. A UMDP is a tuple $\mathcal{M} = (S, A, \mathcal{P}, \mu_0, r, c)$ where $S = \{1, \ldots, n\}$ and $A = \{1, \ldots, m\}$ are finite sets of states and actions respectively (n is the number of states and m is the number of actions in $\mathcal{M}$). μ_0 is the initial state distribution. $r : S \times A \times S \to \mathbb{R}$ and $c : S \times A \times S \to \mathbb{R}$ are reward and cost functions, respectively. The method presented herein naturally extends to handle constraints with respect to multiple cost functions, but for simplicity of the

[†]MPI-SWS, Kaiserslautern and Saarbrücken, Germany.
[*]Department of Electrical and Computer Engineering, Worcester Polytechnic Institute, Worcester, MA.
[‡]Department of Aerospace Engineering and Engineering Mechanics, University of Texas at Austin, TX.

presentation we consider UMDPs with a single cost function. The transition probability function is uncertain and captured by an *ambiguity set*

$$\mathcal{P} = \{P \in \mathcal{D}(S)^{n\times m} : \exists \xi \in \Xi \text{ such that } P(\cdot \mid s,a) = p(\xi; s, a), \forall (s,a) \in S \times A\},$$

where $P(\cdot \mid s,a)$ represents the probabilities of reaching states in S from s after action a being taken, $\Xi \subset \mathbb{R}^q$ is the set of uncertain parameters of size q, and $p(\xi; s, a)$ is an affine function from Ξ to $\mathcal{D}(S)$. We make the following assumption regarding the set of uncertain parameters.

Assumption II.1. *The set Ξ is polyhedral, i.e.,*

$$\Xi = \{\xi \in \mathbb{R}^q : a_l \xi + b_l \geq 0 \; \forall l = 1, \ldots, L\}$$

where L is the number of constraints on the uncertain parameters ξ and q is the dimension of ξ.

We overload the term UMDP to refer to tuples of the form $\mathcal{M} = (S, A, \mathcal{P}, \mu_0)$ and also of the form $\mathcal{M} = (S, A, \mathcal{P}, \mu_0, r)$, i.e., for the cases when we do not have reward and/or cost functions. In the case when the ambiguity set is a singleton $\mathcal{P} = \{P\}$, we obtain a conventional MDP, which we denote by $M = (S, A, P, \mu_0)$.

The labeled digraph $G = (S, E)$ of an MDP $M = (S, A, P, \mu_0)$ is defined such that $(s, a, s') \in E$ if and only if $P(s' \mid s, a) \neq 0$.

We assume the following about the ambiguity set $\mathcal{P}$.

Assumption II.2. *For any $P, P' \in \mathcal{P}$, the graph of the MDP $M = (S, A, P, \mu_0)$ is the same as that of the MDP $M' = (S, A, P', \mu_0)$. That is, all distributions defined by $\mathcal{P}$ have the same support, which defines precisely the graph structure.*

Under Assumption II.2, the graph induced by a UMDP is uniquely defined, and we denote it by $G = (S, E)$.

A *policy* for a UMDP $\mathcal{M} = (S, A, \mathcal{P}, \mu_0, r, c)$ is a sequence of functions $\pi = (\pi_t)_{t\in I}$ where $I = [0, T]$ for $T < \infty$ if we consider a finite horizon, and $T = \infty$ if we consider an infinite horizon, and $\pi_t : S \to \mathcal{D}(A)$ is a probability distribution over the action space A according to which the next action is chosen. A policy is *memoryless* if $\pi_t = \pi_{t'}$ for all $t, t' \in I$. We denote a memoryless policy simply by $\pi = (\pi)_{t\in I}$ by slightly overloading the notation. A policy is *deterministic* if $\pi_t : S \to A$. Thus, a memoryless and deterministic policy is a function $\pi : S \to A$.

Given a transition kernel P, a policy π induces a Markov chain $(s_t, a_t)_{t\in I}$. The *expected total reward* of $(s_t, a_t)_{t\in I}$ under a discounting factor $\gamma \in (0, 1]$ is

$$V_r^{P,\pi,\mu_0} = \mathbb{E}^{P,\pi}\left[\sum_{t\in I} \gamma^t r(s_t, a_t, s_{t+1}) \mid s_0 \sim \mu_0\right].$$

B. Problem formulation

For a UMDP the expected total reward is uncertain, as it depends on the uncertain transition kernel. Therefore, we consider the synthesis of policies that are robust to such an uncertainty, in the sense that we are looking for a policy that maximizes the worst-case expected reward

$$V_r^{\pi,\mu_0} = \inf_{P\in\mathcal{P}} V_r^{P,\pi,\mu_0}. \tag{1}$$

To bound the expected total reward for the infinite-horizon case when $\gamma = 1$, we assume that there exists a set of sink states [1]. A non-empty subset of sink states are *accepting* and a path under any policy must eventually reach an accepting sink state. The rewards of the self-loop transitions in all accepting sink states are 0.

Similarly, the total cost associated with a policy π also depends on the unknown transition kernel. Our goal is to synthesize policies for which the worst-case *maximal total non-discounted cost* is below a given threshold. The reason for considering non-discounted cost is that constraints typically represent limited energy and time resources. Resource consumption is not discounted, and typically the required policies should bound the total cost for *all possible executions* resulting from the policy under *all possible realizations of the uncertainty in the model*. Formally, the maximal total cost of a policy π is defined by the upper bound

$$V_c^{\pi,\mu_0} = \sup_{P\in\mathcal{P}} V_c^{P,\pi,\mu_0} \quad \text{, where}$$

$$V_c^{P,\pi,\mu_0} = \sup_{(S_t,A_t)_{t\in I}\in \mathsf{Supp}((s_t,a_t)_{t\in I})} \left[\sum_{t\in I} c(S_t, A_t, S_{t+1})\right] \tag{2}$$

and where $(S_t, A_t)_{t\in I}$ is a sample in the Markov chain $(s_t, a_t)_{t\in I}$ induced by π in the MDP with transition kernel P.

The main problem we study is stated as follows.

Problem 1. *Given a UMDP $\mathcal{M} = (S, A, \mathcal{P}, \mu_0, r, c)$, and a discounting factor γ, and an upper bound η on the maximal total cost, compute a policy π^* that satisfies the conditions*

$$V_r^{\pi^*,\mu_0} \geq V_r^{\pi,\mu_0} \text{ for all policies } \pi \text{ with } V_c^{\pi,\mu_0} \leq \eta \tag{3}$$

and

$$V_c^{\pi^*,\mu_0} \leq \eta. \tag{4}$$

Next we present a method for computing a solution to Problem 1, i.e., for computing a robust optimal policy in a UMDP with a constraint on the maximal total cost.

III. Main results

We assume that $r(s, a, s') \geq 0$ and $c(s, a, s') \geq 0$ for all $(s, a, s') \in S \times A \times S$, and that $\eta \geq 0$. In order to compute a policy enforcing the cost constraint in the UMDP, we augment the state space of the UMDP with a cost state variable h whose domain is $[0, \eta]$. For simplicity, we give the definition of cost-augmented MDP. The construction extends naturally to UMDPs using the definition of ambiguity sets for the transition function.

[1] A state s is a sink state if for all $a \in A$, $P(s \mid s, a) = 1$.

A. Cost-augmented uncertain MDP

We first provide the construction of cost-augmented MDP for the non-discounted case, i.e., $\gamma = 1$. We then show how the construction can be extended to discounted MDPs.

Definition 1 (Cost-augmented MDP: Non-discounted case)**.** *Given an MDP $M = (S, A, P, \mu_0, r, c)$, a discounting factor $\gamma = 1$, and an upper bound on the cost η, a cost-augmented Markov decision process (aug-MDP) $\widetilde{M}$ is*

$$\widetilde{M} = (\widetilde{S}, A, \widetilde{P}, \widetilde{\mu}_0, \gamma, \widetilde{r})$$

with components defined as follows.

- $\widetilde{S} = S \times H \cup \{\mathsf{sink}\}$ *is the augmented state space. An augmented state $(s, h) \in \widetilde{S}$ consists of a discrete state s and a cost state $h \in [0, \eta]$.* sink *is a new sink state.*
- $\widetilde{P}$ *is the transition probability function defined as follows: 1) For a given state (s, h) and an action a, in case that there exists s' such that $P(s' \mid s, a) \neq 0$, if $h' = h - c(s, a, s') \geq 0$, let $\widetilde{P}((s', h') \mid (s, h), a) = P(s' \mid s, a)$; otherwise the system transits to the new sink state* sink *by letting $\widetilde{P}(\mathsf{sink} \mid (s, h), a) = \sum_{(s' \in S : h - c(s,a,s') < 0)} P(s' \mid s, a)$. 2) For any s' such that $P(s' \mid s, a) = 0$, $\widetilde{P}((s', h') \mid (s, h), a) = 0$ for any pair $h, h' \in H$. 3) Lastly, $\widetilde{P}(\mathsf{sink} \mid \mathsf{sink}, a) = 1$ for any $a \in A$.*
- $\widetilde{\mu}_0 \in \mathcal{D}(\widetilde{S})$ *is defined such that $\widetilde{\mu}_0((s, \eta)) = \mu_0(s)$ for all $s \in S$ and $\widetilde{\mu}_0((s, h)) = 0$ for any $h \neq \eta$.*
- $\widetilde{r} : \widetilde{S} \times A \times \widetilde{S} \to \mathbb{R}$ *is the cost function defined such that $\widetilde{r}((s, h), a, (s', h')) = r(s, a, s')$ for any $h, h' \in H$ and $\widetilde{r}(\widetilde{s}, a, \mathsf{sink}) = 0$ for any $\widetilde{s} \in \widetilde{S}$.*

Given a UMDP $\mathcal{M} = (S, A, \mathcal{P}, \mu_0, r, c)$ we can define analogously the corresponding UMDP augmented with a cost state $\widetilde{\mathcal{M}} = (\widetilde{S}, A, \widetilde{\mathcal{P}}, \widetilde{\mu}_0, \widetilde{r})$, and term it cost-augmented uncertain Markov decision process (aug-UMDP).

In cases when $\gamma \neq 1$, one can transform the discounted MDP to a non-discounted MDP. It is proven in [5] that for any policy, the expected total rewards are the same in the resulting non-discounted MDP and the original discounted MDP. It is also straightforward to show that the maximal total costs are the same because the transformation does not affect the total cost/reward for a single path, though it affects the probability measure of paths under a given policy.

B. Constraint satisfaction

Proposition 1. *Given aug-UMDP $\widetilde{\mathcal{M}}$ and a policy π, let $\widetilde{M}^{\pi} = (\widetilde{s}_t, \widetilde{a}_t)_{0 \leq t < \infty}$ be the induced uncertain Markov chain. If $\sup_{P \in \widetilde{\mathcal{P}}} \mathbb{P}^{P, \widetilde{\mu}_0}(\exists t : \ 0 \leq t < \infty, \widetilde{s}_t = \mathsf{sink}) = 0$, then the maximal cost constraint* (4) *is satisfied.*

Proof. The proof directly follows from the construction of the aug-UMDP: To violate the maximal total cost constraint (4), a path must visit the state sink. Since the policy π avoids sink with probability 1 under any possible $P \in \mathcal{P}$, it holds that π enforces the constraint (4). □

Under Assumption II.2, we can compute a memoryless policy from the graph $\widetilde{G} = (\widetilde{S}, \widetilde{E})$ of the given aug-UMDP which ensures satisfaction of (4). To this end, we compute a function $f_{\mathsf{safe}} : \widetilde{S} \to 2^A$ that maps each state to a set of *"safe actions"* from that state. For any state (s, h) where f_{safe} is defined, a policy only taking an action from $f_{\mathsf{safe}}(s, h)$ is guaranteed to satisfy the constraint. The procedure for computing f_{safe} is given in Algorithm 1. Note that the input of Algorithm 1 is a general labeled graph $G = (S, E)$. For computing f_{safe}, we apply Alg. 1 to the graph $\widetilde{G} = (\widetilde{S}, \widetilde{E})$ of the given aug-UMDP and set of sink states $S_{\mathsf{sink}} = \{\mathsf{sink}\}$.

Input: A labeled graph $G = (S, E)$ with $E \subseteq S \times A \times S$.
A set of sink states (nodes) $S_{\mathsf{sink}} \subseteq S$.
Output: A function $f_{\mathsf{safe}} : S \to 2^A$.
Initialize $W_0 := S \setminus S_{\mathsf{sink}}$, $f_{\mathsf{safe}}(s) := \emptyset$ for all $s \in S$;
while *True* **do**
  $W_{i+1} := W_i$;
  for $s \in W_i$ **do**
    $f_{\mathsf{safe}}(s) := \{a \in A \mid \forall (s, a, s') \in E : s' \in W_{i+1}\}$;
    if $f_{\mathsf{safe}}(s) = \emptyset$ **then** $W_{i+1} := W_{i+1} \setminus \{s\}$;
  end
  if $W_{i+1} = W_i$ **then** break;
end
return f_{safe};

Algorithm 1: Almost sure constraint satisfaction.

The Lemma below follows from the construction of f_{safe}.

Lemma 1. *Under Assumption II.2, let $f_{\mathsf{safe}} : \widetilde{S} \to 2^A$ be the function obtained by Algorithm 1. The set of memoryless policies feasible with respect to the cost constraint* (4) *is $\Pi = \{\pi : S \times A \to [0, 1] \mid \pi(s, a) > 0 \Leftrightarrow a \in f_{\mathsf{safe}}(s)\}$.*

We can use the function f_{safe} to compute the feasible solutions to Problem 1 in the set of memoryless polices.

C. A two-step solution for UMDP

Now it is clear that given the function f_{safe} in the cost-augmented UMDP we can prune in each state all the actions not allowed by this function and compute a robust optimal policy in the UMDP after this modification. Formally, the revised UMDP $\widetilde{\mathcal{M}}_{\mathsf{R}} = (\widetilde{S}, A, \widetilde{\mathcal{P}}_{\mathsf{R}}, \widetilde{\mu}_0, \widetilde{r})$ is such that $P_{\mathsf{R}} \in \widetilde{\mathcal{P}}_{\mathsf{R}}$ if and only if there exists a $P \in \widetilde{\mathcal{P}}$ such that for any $\widetilde{s}, a$, if $a \in f_{\mathsf{safe}}(\widetilde{s})$ then $P_{\mathsf{R}}(\cdot \mid \widetilde{s}, a) = P(\cdot \mid \widetilde{s}, a)$, otherwise $P_{\mathsf{R}}(\cdot \mid \widetilde{s}, a)$ is the zero vector. Then, the robust optimal policy in $\widetilde{\mathcal{M}}_{\mathsf{R}}$ is a solution to Problem 1 for the given UMDP.

The robust optimal policy for an uncertain MDP can be computed using various methods [10], [16], [15].

IV. Safety-threshold constraints

In this section, we study the synthesis problem subject to safety-threshold constraints. The goal is to bound the probability of visiting a set of error/unsafe states while maximizing the expected total reward. We show that the solution approach proposed for Problem 1 serves as an approximate solution to this problem.

A. Cost function for safety-threshold constraints

Let $\mathcal{M} = (S, A, \mathcal{P}, \mu_0, r)$ be a UMDP and let $S_{err} \subseteq S$ be a given a set of error states. We assume that each $s \in S_{err}$ is a sink state and not initial, i.e., $\mu_0(s) = 0$.

For an MDP with transition kernel P and a policy π we define the *safety value* V_s^{P,π,μ_0} of π to be the probability of reaching S_{err} under the policy π in M. Formally,

$$V_s^{P,\pi,\mu_0} = \sum_{t \in I} \mathbb{P}^{P,\pi} \left[s_t \in S_{err} \mid s_0 \sim \mu_0, \forall t' < t, s_{t'} \notin S_{err} \right]. \tag{5}$$

Then, the safety-value of a policy π in a UMDP is defined by the upper bound (i.e., worst case)

$$V_s^{\pi,\mu_0} = \sup_{P \in \mathcal{P}} V_s^{P,\pi,\mu_0}. \tag{6}$$

Problem 2 (Robust optimality under a safety threshold)**.** *Given a UMDP $\mathcal{M} = (S, A, \mathcal{P}, \mu_0, r)$ and a real-valued constant $\eta \in [0, 1]$, compute a policy π^* such that*

$$V_r^{\pi^*,\mu_0} \geq V_r^{\pi,\mu_0} \text{ for all policies } \pi \text{ with } V_s^{\pi,\mu_0} \leq \eta \tag{7}$$

and

$$V_s^{\pi^*,\mu_0} \leq \eta. \tag{8}$$

We now show how we can solve Problem 2 conservatively by reducing it to Problem 1 using the following cost function.

Let $c_{safety} : S \times A \times S \to \mathbb{R}$ be the cost function such that, for $s, s' \in S$ and $a \in A$, we have, if $s \notin S_{err}$,

$$c_{safety}(s, a, s') = \sup_{P \in \mathcal{P}} \sum_{s'' \in S_{err}} P(s, a, s'')$$

and, if s is a sink state, $c_{safety}(s, a, s') = 0$. Intuitively, $c_{safety}(s, a, s')$ is the worst-case probability of entering a state in S_{err} when taking action a in state s. Note that the value $c_{safety}(s, a, s')$ does not depend on the successor state s', but is determined by the current state s and the action a.

The following proposition formalizes the relationship between the maximal total cost V_c^{π,μ_0} (used in Problem 1) based on the cost function c_{safety} defined above, and the safety value V_s^{π,μ_0} (used in Problem 2). More precisely, we show that $V_s^{\pi,\mu_0} \leq V_c^{\pi,\mu_0}$ for every policy π. Intuitively, the value V_s^{P,π,μ_0} is the probability measure of the set of paths in the resulting Markov chain that reach the set S_{err} of error states. This is the sum of the measures of all cones of the finite paths reaching S_{err}. The value V_c^{π,μ_0}, on the other hand, is the maximal sum of transition costs over all paths. Each such sum corresponds to the sum of the probabilities of the taken actions to enter S_{err} in one step. For each element of the sum, the probability of the path prefix is not accounted for in V_c^{π,μ_0}, which leads to an over-approximation of the safety value. We give a bound on the effect of this over-approximation through a bound on V_c^{π,μ_0}, provided that it is finite. This bound is the sum of the cost of all transitions entering S_{err}, and is easily obtained from the worst-case probability of actions entering S_{err}.

Proposition 2. *Let $\mathcal{M} = (S, A, \mathcal{P}, \mu_0, r)$ be a UMDP, and $S_{err} \subseteq S$ be a set of error states. Let $\mathcal{M}_{safety} = (S, A, \mathcal{P}, \mu_0, r, c_{safety})$ be the UMDP obtained from $\mathcal{M}$ by adding the cost function c_{safety} defined by S_{err}.*

Let π be a memoryless policy in $\mathcal{M}$ (and $\mathcal{M}_{safety}$). Then, $V_s^{\pi,\mu_0} \leq V_c^{\pi,\mu_0}$. Furthermore, if $V_c^{\pi,\mu_0} < \infty$, then $V_c^{\pi,\mu_0} \leq \sum_{(s,a,s'), s' \in S_{err}} \sup_{P \in \mathcal{P}} P(s, a, s')$, where the above sum is over edges (s, a, s') in the Markov chain M^π induced by the policy π.

Proof. Let $M = (S, A, P, \mu_0, r)$ be an MDP with $P \in \mathcal{P}$ and π be a memoryless policy inducing a Markov chain M^π.

We define functions $v_s, v_c : S \to \mathbb{R} \cup \{\infty\}$ as follows: For a sink state s, let $v_s(s) = v_c(s) = 0$; For a non-sink state s and $a = \pi(s)$, let $v_s(s) = \sum_{s' \in S_{err}} P(s, a, s') + \sum_{s' \in S \setminus S_{err}} P(s, a, s') \cdot v_s(s')$, and $v_c(s) = \max_{s' \in S, P(s,a,s') > 0} \left(c_{safety}(s, a, s') + v_c(s') \right)$.

Thus, $v_s(s)$ is the probability of visiting a state in S_{err} and $v_c(s)$ is the maximal total cost of a path ending in a sink state. The definitions imply that $V_s^{\pi,\mu_0} = \sum_{s \in S} \mu_0(s) \cdot v_s(s)$ and $V_c^{\pi,\mu_0} = \max_{s \in S, \mu_0(s) > 0} v_c(s)$. To show that $V_s^{\pi,\mu_0} \leq V_c^{\pi,\mu_0}$ it suffices to prove $v_s(s) \leq v_c(s)$ for all $s \in S$.

Given the definition of cost function c_{safety}, we have

$$\begin{aligned} v_c(s) &= \max_{s' \in S, P(s,a,s') > 0} \left(\sup_{P' \in \mathcal{P}} \sum_{s'' \in S_{err}} P'(s, a, s'') + v_c(s') \right), \\ &\geq \sum_{s' \in S_{err}} P(s, a, s') + \max_{s' \in S \setminus S_{err}, P(s,a,s') > 0} v_c(s'). \end{aligned}$$

On the other hand,

$$\begin{aligned} v_s(s) &= \sum_{s' \in S_{err}} P(s, a, s') + \sum_{s' \in S \setminus S_{err}} P(s, a, s') \cdot v_s(s') \\ &\leq \sum_{s' \in S_{err}} P(s, a, s') + \max_{s' \in S \setminus S_{err}, P(s,a,s') > 0} v_s(s'). \end{aligned}$$

Since $v_c(s) = v_s(s)$ for all sink states s, using backward induction, we have $v_s(s) \leq v_c(s)$ for all $s \in S$.

Now suppose that $V_c^{P,\pi,\mu_0} < \infty$. Since the transition costs are non-negative, each infinite path from an initial state in M_{safety}^π contains only finitely many transitions with non-zero cost. Thus, for each such path τ in M_{safety}^π with total cost c_τ we have that $c_\tau \leq \sum_{(s,a,s'), s' \in S_{err}} \sup_{P \in \mathcal{P}} P(s, a, s')$, since if there is a repeated transition in the path τ, then a path containing a cycle with infinite cost can be constructed (a contradiction). We can then conclude that $V_c^{P,\pi,\mu_0} \leq \sum_{(s,a,s'), s' \in S_{err}} \sup_{P \in \mathcal{P}} P(s, a, s')$. Since the kernel P was arbitrarily chosen from $\mathcal{P}$, the claim follows. □

The proposition above implies that every feasible solution to Problem 1 is a feasible solution to Problem 2. The following example demonstrates that the converse does not hold. Hence, an optimal solution of Problem 1 can be sub-optimal for Problem 2. When the UMDP does not contain paths with unbounded costs, we can bound the difference between maximal total cost and expected total cost of a memoryless and deterministic policy by $\sum_{(s,a,s'), s' \in S_{err}} \sup_{P \in \mathcal{P}} P(s, a, s')$. Although this bound is rather coarse, in the cases when the probabilities on transitions entering error states (and their sum) are small, the difference $V_c^{\pi,\mu_0} - V_s^{\pi,\mu_0}$ is also small.

We now illustrate the difference between the safety value and the maximal total cost and the approximation the latter one induces with respect to Problem 2 using several examples. For simplicity of the presentation, in the examples we use MDPs, but since those are a special case of UMDPs, the same can be done for UMDPs.

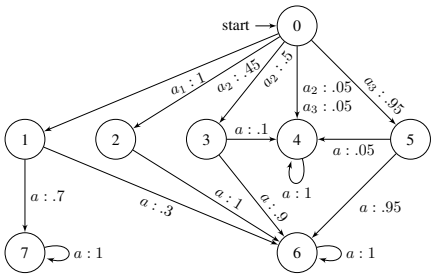

Fig. 1. The MDP $M = (S, A, P, \mu_0, r)$ used in Example 1 and Example 2.

Example 1. *Consider the MDP $M = (S, A, P, \mu_0, r)$ with transition probabilities shown in Figure 1. The initial distribution is such that $\mu_0(0) = 1$ and $\mu_0(s) = 0$ for all $s \neq 0$. The reward function is such that $r(s, a, s') = 1$ if $s \neq 6$ and $s' = 6$, and $r(s, a, s') = 0$ otherwise.*

Consider the set of error states $S_{err} = \{4\}$ and $\eta = 0.1$. By the definition of c_{safety} we have (omitting the subscript): $c(0, a_1, 1) = c(1, a, 6) = c(1, a, 7) = c(2, a, 6) = c(6, a, 6) = c(7, a, 7) = c(4, a, 4) = 0$, $c(0, a_2, 2) = c(0, a_2, 3) = c(0, a_2, 4) = c(0, a_3, 4) = c(0, a_3, 5) = c(5, a, 4) = c(5, a, 6) = 0.05$, $c(3, a, 4) = c(3, a, 6) = 0.1$.

Now, consider three memoryless deterministic policies $\pi_1, \pi_2, \pi_3 : S \to A$ with $\pi_i(0) = a_i$, $\pi_i(s) = a$ for $i \in \{1, 2, 3\}$ and $s \neq 0$. Their values are as follows:

	$V_r^{\cdot,\mu_0}$	$V_s^{\cdot,\mu_0}$	$V_c^{\cdot,\mu_0}$
π_1	*0.3*	*0*	*0*
π_2	*0.9*	*0.1*	*0.15*
π_3	*0.9025*	*0.0975*	*0.1*

This example demonstrates that solving Problem 2 approximately by formulating it as Problem 1 introduces conservatism. More specifically, for policy π_2 we have $V_s^{\pi_2,\mu_0} = 0.1 \leq \eta$ and $V_c^{\pi_2,\mu_0} = 0.15 > \eta$. Thus, policy π_2 does not satisfy the constraint of Problem 1, while it satisfies the one in Problem 2. Policy π_3, on the other hand, meets both constraints, and since it has value better than that of policy π_1, it is the optimal solution to Problem 1. (Note that in this example policy π_3 is also optimal for Problem 2. In Section V we will see an example where the optimal solution to Problem 1 is sub-optimal for Problem 2.)

For a known MDP $M = (S, A, \mathcal{P}, \mu_0, r)$, that is when the set $\mathcal{P}$ is a singleton, the approximation introduced by considering Problem 1 with the cost function c_{safety} precisely coincides with the approximation used in [11]. In [11] the probability of exiting a set of feasible (in our case, safe) states is over-approximated by the sum of probabilities of doing so in each step of the execution. For a known MDP, the cost $c(s, a, s')$ is exactly the probability of entering a state in S_{err}, and, in this case, we impose an upper bound on the sum of the probabilities for the individual steps by the total cost of the optimal policy for Problem 1.

As we saw in Section III-A, the computation of a robust optimal policy under a maximal total cost constraint can be precisely reduced to computing a robust optimal policy in a cost-augmented UMDP. Since for a UMDP with rectangular uncertainty the optimal expected reward can be achieved by a deterministic policy [15], the same holds after adding a maximal total cost constraint. This, however, is not the case for a safety threshold constraint (i.e., Problem 2), as it can be seen in the example we give below.

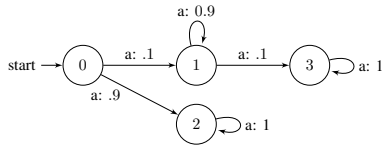

Fig. 2. The MDP $M = (S, A, P, \mu_0, r)$ used in Example 3.

Example 2. *Consider the MDP from Example 1, again with set of error states $S_{err} = \{4\}$. Now, however, let $\eta = 0.05$. Clearly, the policy π_1 defined in Example 1 is the only deterministic solution to Problem 2 since $V_s^{\pi_1,\mu_0} = 0 < 0.05 = \eta$.*

Let π_4 be the randomized *policy such that $\pi_4(0, a_1) = 0.5$ and $\pi_4(0, a_3) = 0.5$. We have that $V_s^{\pi_4,\mu_0} = 0.5 \cdot 0 + 0.5 \cdot 0.0975 = 0.04875$ and thus, π_4 meets the threshold $\eta = 0.05$.*

Moreover, we have that $V_r^{\pi_4,\mu_0} = 0.5 \cdot 0.9025 + 0.5 \cdot 0.3 = 0.60125 > 0.3 = V_r^{\pi_1,\mu_0}$, meaning that π_4 has better expected reward than π_1. Thus, if we allow randomized policies, policy π_1 is not an optimal solution to Problem 2.

Note that $V_c^{\pi_4,\mu_0} = 0.1 > \eta$, i.e., π_4 is not feasible for Problem 1. Randomized policies are not more powerful than deterministic ones when considering Problem 1.

In Proposition 2 we established a bound on the difference between the expected total cost and the maximal total cost under the assumption that the latter one is finite. Below we give an example of an MDP where this is not a case.

Example 3. *Consider the MDP $M = (S, A, P, \mu_0, r)$ with transition probabilities given in Figure 2 and initial distribution where $\mu_0(0) = 1$ and $\mu_0(s) = 0$ for all $s \neq 0$. Let $S_{err} = \{3\}$ be a set of error states and let $r(0, a, 2) = 1$ and $r(s, a, s') = 0$ if $s \neq 0$ or $s' \neq 2$. For the single policy π in M we have $V_s^{\pi,\mu_0} = 0.1$ for the probability of reaching S_{err}. Since $c_{safety}(1, a, 1) = c_{safety}(1, a, 3) = 0.1$, every path of the form $s_0 s_1^k s_3$ has cost $(k + 1) \cdot 0.1$. Thus, $V_c^{\pi,\mu_0} = \infty$.*

In order to capture the safety value using the maximal total cost constraint more precisely, one way is to pre-compute all states from which an error state is reached almost surely for all possible policies and then make all such states error states as well. Formally, this set of states is $S_{unsafe} = \{s \mid \forall \pi \in \Pi, \exists P \in \mathcal{P}, \mathbb{P}^{P,\pi}(s_t \in S_{err} \mid s_0 = s) = 1\}$. Note that S_{unsafe} can be computed from the graph of a UMDP, using the algorithm described in [2, Teorem10.112].

In Example 3, we have $S_{unsafe} = \{1, 3\}$ and by revising $S_{err} = S_{unsafe}$, the maximal total cost becomes 0.1, which is the cost of transitioning to the new sink error state 1.

V. Examples

In this section we give two examples of UMDPs and illustrate the solutions to Problem 1.

Example 4. *Consider the UMDP whose nominal transition probabilities are shown in Figure 3. The only uncertain transition is action b in state 2 with uncertainty ± 0.05. We have initial distribution $\mu_0(1) = 1$ and $\mu_0(s) = 0$ for $s \neq 1$.*

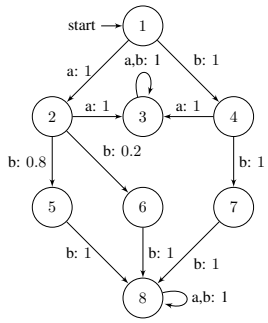

Fig. 3. The UMDP $\mathcal{M} = (S, A, \mathcal{P}, \mu_0, r, c)$ used in Example 4.

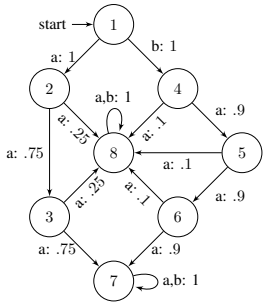

Fig. 4. The UMDP $\mathcal{M} = (S, A, \mathcal{P}, \mu_0, r)$ used in Example 5.

The reward function is such that $r(5, b, 8) = r(6, b, 8) = r(7, b, 8) = 16$ *and* $r(s, \sigma, s') = 0$ *for all other edges. That is, to collect the reward, the goal is to visit state* 8*. The discounting factor for the reward is* $\gamma = 0.5$*.*

Essentially, there are two possible policies in this UMDP that may reach state 8 *starting from the initial state. Intuitively, these two policies model two different paths to state* 8*, depending on the action chosen at the initial state. Let* π_a *be the policy such that* $\pi_a(1) = a$ *and* $\pi_a(s) = b$ *for all* $s \neq 1$*, and let* π_b *be the policy where* $\pi_b(s) = b$ *for all* s*.*

Since both policies π_a *and* π_b *reach state* 8 *after* 3 *transitions with the exact same probability, they have the same expected total reward. More precisely,*$V_r^{\pi_a} = V_r^{\pi_b} = 4$*.*

The cost function is such that $c(2, b, 5) = 1$*,* $c(2, b, 6) = 5$*,* $c(8, b, 8) = 0$*,* $c(3, a, 3) = 0$ *and* $c(s, \sigma, s') = 3$ *for all other edges. The maximal total cost of policy* π_a *is* $V_c^{\pi_a} = 11$*, and the maximal total cost for* π_b *is* $V_c^{\pi_b} = 9$*.*

Let $\eta = 10$*. In this case, since* $V_c^{\pi_a} > \eta$ *and* $V_c^{\pi_b} < \eta$*, from these two policies only* π_b *satisfies the maximal cost constraint, and thus, policy* π_b *is the solution to Problem 1.*

Example 5. *Consider the UMDP with nominal transition probabilities shown in Figure 4 and uncertainty* ± 0.05*. We have initial distribution* $\mu_0(1) = 1$ *and* $\mu_0(s) = 0$ *for* $s \neq 1$*.*

The reward function is such that $r(3, a, 7) = r(6, a, 7) = 16$ *and* $r(s, \sigma, s') = 0$ *for all other edges. The reward is collected by visiting state* 7*. The discounting factor is* $\gamma = 0.5$*.*

There are two possible policies that may reach state 7*, starting from state* 1*. Let* π_a *be the policy that chooses action* a *in all states, and* π_b *the one that selects* b *in state* 1 *and* a *in all other states. Due to the discounting, the expected reward of* π_a *is higher than that of* π_b*, since, intuitively, policy* π_a *allows reaching the goal faster than policy* π_b*.*

Consider a set of error states $S_{err} = \{8\}$*, which defines the cost function* c_{safety}*, and a safety threshold* $\eta = 0.55$*. The cost of each transition is the maximal probability to enter state* 8*. Thus, policy* π_a *is more risky than* π_b*, in the sense that at each state it has probability* 0.3 *to enter* S_{err}*, while for policy* π_b *the risk at each state is* 0.15*.*

The maximal total cost of policy π_a *is* $V_c^{\pi_a} = 0.6$*, and of* π_b *is* $V_c^{\pi_b} = 0.45$*. Since* $V_c^{\pi_a} > \eta$ *and* $V_c^{\pi_b} < \eta$*, the robust optimal solution to Problem 1 is* π_b*. The worst case probability of* π_a *reaching* S_{err} *is* $V_s^{\pi_a} = 0.51 \leq \eta$*, meaning that* π_b *is sub-optimal for Problem 2, since* $V_r^{\pi_a} > V_r^{\pi_b}$*.*

VI. Conclusion

In this work we studied the synthesis of robust optimal control policies for uncertain MDPs subject to cost constraints. We focus on constraints bounding the worst-case maximal total cost and safety-threshold constraints defined by the probability of visiting error states. We proposed a method for robust policy synthesis that yields robust optimal solutions under maximal total cost constraints. Our technique readily applies in the case of safety-threshold constraints, but the resulting solution may be sub-optimal.

References

[1] Eitan Altman. *Constrained Markov decision processes*, volume 7. CRC Press, 1999.

[2] Christel Baier and Joost-Pieter Katoen. *Principles of model checking*. MIT Press, 2008.

[3] Michael Benedikt, Rastislav Lenhardt, and James Worrell. LTL model checking of interval Markov chains. *LNCS*, 7795:32–46, 2013.

[4] Julien Bernet, David Janin, and Igor Walukiewicz. Permissive strategies: from parity games to safety games. *RAIRO-Theoretical Informatics and Applications-Informatique Théorique et Applications*, 36(3):261–275, 2002.

[5] Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control, Vol. II*. Athena Scientific, 3rd edition, 2007.

[6] Luca F Bertuccelli, Brett Bethke, and Jonathan P How. Robust adaptive markov decision processes in multi-vehicle applications. In *Proc. ACC'09*, pages 1304–1309. IEEE, 2009.

[7] Richard C Chen and Gilmer L Blankenship. Dynamic programming equations for discounted constrained stochastic control. *IEEE Transactions on Automatic Control*, 49(5):699–709, 2004.

[8] Erick Delage and Shie Mannor. Percentile Optimization for Markov Decision Processes with Parameter Uncertainty. *Operations Research*, 58(1):203–213, 2009.

[9] D Kim, Kee-eung Kim, and Pascal Poupart. Cost-Sensitive Exploration in Bayesian Reinforcement Learning. *Advances in neural information processing systems*, pages 1–9, 2012.

[10] Arnab Nilim and Laurent El Ghaoui. Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

[11] Masahiro Ono, Marco Pavone, Yoshiaki Kuwata, and J. Balaram. Chance-constrained dynamic programming with application to risk-aware robotic space exploration. *Autonomous Robots*, 39(4):555–571, 2015.

[12] Martin Pecka and Tomas Svoboda. Safe Exploration Techniques for Reinforcement Learning An Overview. *LNCS*, pages 357–375, 2014.

[13] Alexey B. Piunovskiy. Dynamic programming in constrained Markov decision processes. *Control and Cybernetics*, 35(3):645–660, 2006.

[14] Alberto Puggelli, Wenchao Li, Alberto L. Sangiovanni-Vincentelli, and Sanjit A. Seshia. Polynomial-time verification of PCTL properties of MDPs with convex uncertainties. *LNCS*, 8044:527–542, 2013.

[15] Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.

[16] Eric M Wolff, Ufuk Topcu, and Richard M Murray. Robust control of uncertain markov decision processes with temporal logic specifications. In *Proc. CDC'12*, pages 3372–3379, 2012.