2024 European Control Conference (ECC)
June 25-28, 2024. Stockholm, Sweden

Constrained Markov decision processes with uncertain transition probabilities

V Varagapriya[1], Vikas Vikram Singh[2], Abdel Lisser[3]

Abstract— **We consider a constrained Markov decision process with uncertain transition probabilities, where the uncertainty is driven by a single parameter which belongs to an interval. We model it using a robust optimization framework and show that it is equivalent to a bilinear programming problem. We propose a linear programming-based algorithm to compute its global optimal solution. The numerical experiments are performed on a well-known class of Markov decision problems called Garnets using our algorithm as well as Gurobi bilinear solver. We observe that for the case of dense transition probabilities, our algorithm outperforms Gurobi bilinear solver.**

I. INTRODUCTION

A Markov decision process (MDP) is a framework to model the decision-making of a dynamic system that has a predefined set of states and a set of available actions. At each time, the system is at some state, and an action is chosen by a decision maker; accordingly, a cost is incurred, and the system moves to the next state according to a controlled Markov chain. This process repeats over an infinite horizon. The decision maker aims to minimize the overall expected cost incurred. For an MDP problem with a finite number of states and actions, with known stationary costs and transition probabilities, a stationary deterministic optimal policy can be computed using various algorithms, such as value iteration, policy iteration, and linear programming (LP) methods ([13]). In many scenarios, multiple costs are incurred whenever a state is visited, and an action is chosen. Such cases are modelled using a constrained Markov decision process (CMDP) framework where the decision maker aims to minimize one type of overall cost subject to constraints on all other overall costs. For a CMDP problem with a finite number of states and actions, with known stationary costs and transition probabilities, a stationary randomized optimal policy can be computed by solving an LP problem [1].

However, the values of the model parameters, namely, the costs and the transition probabilities, are typically obtained via observation and historical data. Hence, the assumption in MDP/CMDP problems that their values are exactly known is erroneous and may result in policies that is not optimal in practice [9]. Thus, it is better to model them as uncertain parameters. Although a number of works have studied MDPs in this context, including, [15], [12], [5], [19], the studies on CMDPs with uncertain parameters are limited. In [7], a CMDP problem with known transition probabilities and uncertain cost vectors belonging to an interval uncertainty set is considered, and the resulting robust CMDP problem is shown to be equivalent to an LP problem. This result is extended in [17] to other convex uncertainty sets, and in each case, the resulting robust CMDP problem is shown to be equivalent to a convex optimization problem. On the other hand, in [18], the costs are defined using random vectors, and the CMDP problem is modelled as a joint chance constraint programming problem. This problem is then approximated using two second-order cone programming problems which give upper and lower bounds on the optimal value of the original problem. To the best of our knowledge, a reformulation of the CMDP problem under uncertain transition probabilities has not been considered in the literature.

In this paper, we consider a CMDP problem under a discounted cost criterion, with known costs and uncertain transition probabilities. Such a problem can be applied to a machine replacement problem where the transition probabilities need not be exactly known and are realized as the system progresses. We assume the uncertainty stems from a single parameter that belongs to an interval. We formulate it as a robust CMDP problem and show that it is equivalent to a bilinear programming (BP) problem when the policies are restricted to the stationary class. We construct an LP-based algorithm to compute the optimal policy of the robust CMDP problem by finding the global optimal solution of the BP problem. The numerical experiments are performed on randomly generated instances on a well-known class of MDP problems called Garnets ([2], [3]) using LP-based algorithm and existing Gurobi bilinear solver. In many instances, LP-based algorithm performs better than the Gurobi bilinear solver.

[1] Department of Mathematics, Indian Institute of Technology Delhi, Hauz Khas, 110016, New Delhi, India varagapriyav@gmail.com

[2] Department of Mathematics, Indian Institute of Technology Delhi, Hauz Khas, 110016, New Delhi, India vikassingh@iitd.ac.in

[3] Laboratory of Signals and Systems, University Paris Saclay, CNRS, CentraleSupelec, Bat Breguet, 3 Rue Joliot Curie, Gif-sur-Yvette, 91190, France abdel.lisser@l2s.centralesupelec.fr

The structure of the paper is as follows. Section II contains a CMDP model. In Section III, we present a BP formulation of a robust CMDP model and an LP-based algorithm to compute its optimal policy. Section IV contains numerical experiments on random CMDPs, and we conclude the paper in Section V.

II. Constrained Markov decision process

We consider a discrete-time infinite-horizon CMDP model under a discounted cost criterion. Let S and $A(s)$ denote a finite set of states and actions available at each state $s \in S$, respectively. Let the set of all state-action pairs be denoted by $\mathcal{K}$, i.e., $\mathcal{K} = \{(s,a) \mid s \in S, a \in A(s)\}$; $|\mathcal{K}|$ is the cardinality of $\mathcal{K}$. At time $t = 0$, suppose the system is at s_0 with probability $\gamma(s_0) > 0$, at which an action a_0 is taken by the decision maker. As a consequence, costs $c(s_0, a_0)$, $d^k(s_0, a_0)$, $k \in \mathbb{K} = \{1, 2, \dots, K\}$, are incurred. At $t = 1$, the system moves to a state s_1 with probability $p(s_1|s_0, a_0)$ and the same procedure repeats infinitely. We assume $\gamma(s) > 0$ for all $s \in S$, and the costs $c(s,a)$, $d^k(s,a)$, $k \in \mathbb{K}$ and the transition probabilities $p(s'|s,a)$ are stationary, i.e., they are time-invariant. The choice of an action at time t defined by a decision rule may depend on the history $h_t = (s_0, a_0, s_1, a_1, \dots, s_{t-1}, a_{t-1}, s_t)$ at time t. Whenever a history h_t is observed, an action a_t is taken at s_t according to a decision rule $f_t^h = f_t(h_t) \in \wp(A(s_t))$, where $\wp(A(s_t))$ denotes the set of probability distributions over the action set $A(s_t)$. Such a decision rule is called a history-dependent decision rule and a sequence $f^h = (f_t^h)_{t=0}^{\infty}$ is called a history-dependent policy. A policy is called stationary if there exists a decision rule f such that $f_t^h = f$ for all t. We denote a stationary policy, with some abuse in notations, by f and according to f, whenever the Markov chain visits a state s, an action $a \in A(s)$ is taken with probability $f(s,a)$. Let F_S be the set of all stationary policies. An optimal policy of a CMDP problem exists in the class of stationary policies [1]. In this paper, we restrict our attention to these stationary policies for simplicity. The cost vector $\big(c(s,a)\big)_{(s,a)\in\mathcal{K}}$ corresponds to the objective function and the cost vectors $\big(d^k(s,a)\big)_{(s,a)\in\mathcal{K}}$, $k \in \mathbb{K}$, constitute the costs on which constraints are imposed. The policy f and the initial distribution γ define a probability measure $\mathbb{P}_\gamma^f$ over the state and action trajectories (for details, see Section 2.1.6 of [13]), and $\mathbb{E}_\gamma^f$ denotes the expectation operator corresponding to $\mathbb{P}_\gamma^f$. We denote a random state-action pair at time t by $(\mathbb{X}_t, \mathbb{A}_t)$. For a given $f \in F_S$, γ, and a discount factor $\alpha \in (0,1)$, the expected discounted costs corresponding to the objective function and constraints are given by

$$C_\alpha(\gamma, f) = (1-\alpha)\sum_{t=0}^{\infty} \alpha^t \mathbb{E}_\gamma^f\big(c(\mathbb{X}_t, \mathbb{A}_t)\big), \quad (1)$$

$$D_\alpha^k(\gamma, f) = (1-\alpha)\sum_{t=0}^{\infty} \alpha^t \mathbb{E}_\gamma^f\big(d^k(\mathbb{X}_t, \mathbb{A}_t)\big),\ \forall\ k \in \mathbb{K}, \quad (2)$$

where $1-\alpha$ denotes the normalization constant that ensures the above costs do not become large when $\alpha \approx 1$. Let $\xi = \big(\xi_k\big)_{k\in\mathbb{K}}$ be the upper bounds on the expected discounted costs given by (2). Then, an optimal policy for a CMDP problem can be obtained by solving the following optimization problem

$$\begin{aligned} \min_{f \in F_S}\ & C_\alpha(\gamma, f) \\ \text{s.t.}\ & D_\alpha^k(\gamma, f) \le \xi_k,\ \forall\ k \in \mathbb{K}. \end{aligned} \quad (3)$$

It follows from the proof of Theorem 3.1 of [1] that the costs in (1), (2) can be written in matrix form as

$$C_\alpha(\gamma, f) = (1-\alpha)\gamma^{\mathrm{T}}\big(I - \alpha P_f\big)^{-1} c_f, \quad (4)$$

$$D_\alpha^k(\gamma, f) = (1-\alpha)\gamma^{\mathrm{T}}\big(I - \alpha P_f\big)^{-1} d_f^k,\ \forall\ k \in \mathbb{K}, \quad (5)$$

where $c_f = \big(c_f(s)\big)_{s\in S}$, $c_f(s) = \sum_{a\in A(s)} f(s,a)c(s,a)$, $d_f^k = \big(d_f^k(s)\big)_{s\in S}$, $d_f^k(s) = \sum_{a\in A(s)} f(s,a)d^k(s,a)$, $k \in \mathbb{K}$; T denotes the transposition. Furthermore, $P_f = \big(P_f(s'|s)\big)_{s,s'} \in \mathbb{R}^{|S|\times|S|}$ is the transition probability matrix induced by f, such that $P_f(s'|s) = \sum_{a\in A(s)} f(s,a)p(s'|s,a)$, and $I \in \mathbb{R}^{|S|\times|S|}$ is the identity matrix. Since the costs defined by (4), (5) are nonlinear functions of f, it is difficult to solve (3) in terms of f. It follows from [1] that the expected discounted costs can be written as linear functions of occupation measure $g_\alpha(\gamma, f) = \big\{g_\alpha(\gamma, f; s, a)\big\}_{(s,a)\in\mathcal{K}}$ defined by,

$$g_\alpha(\gamma, f; s, a) = \big((1-\alpha)\gamma^T\big(I - \alpha P_f\big)^{-1}\big)_s f(s,a), \quad (6)$$

where $\big((1-\alpha)\gamma^T\big(I-\alpha P_f\big)^{-1}\big)_s$ denotes the s^{th}-component of the vector $(1-\alpha)\gamma^T\big(I-\alpha P_f\big)^{-1}$. It follows from Theorem 3.2 of [1] that the set of occupation measures defined for stationary policies is equal to the set $\mathcal{Q}_\alpha(\gamma)$ defined by

$$\mathcal{Q}_\alpha(\gamma) = \Big\{\rho \in \mathbb{R}^{|\mathcal{K}|} \big| \sum_{(s,a)\in\mathcal{K}} \rho(s,a)\big(\delta(s',s) - \alpha p(s'|s,a)\big) = (1-\alpha)\gamma(s'),\ \forall\ s' \in S,\ \rho(s,a) \ge 0, \forall\ (s,a) \in \mathcal{K}\Big\}, \quad (7)$$

where $\delta(s', s)$ is the Kronecker delta, i.e., $\delta(s', s) = 1$ if $s' = s$, and 0, otherwise. As a result, $\rho(s,a) = g_\alpha(\gamma, f; s, a)$, for all $(s,a) \in \mathcal{K}$, $s' \in S$ and (3) can be equivalently written as the following LP problem

$$\min_{\rho \in \mathbb{R}^{|\mathcal{K}|}} \sum_{(s,a)\in\mathcal{K}} \rho(s,a)c(s,a)$$

$$\text{s.t.} \sum_{(s,a)\in\mathcal{K}} \rho(s,a)d^k(s,a) \le \xi_k, \ \forall\, k \in \mathbb{K}, \ \rho \in \mathcal{Q}_\alpha(\gamma).$$

If ρ^* is the optimal solution of the above LP problem, the optimal stationary policy f^* of (3) is given by $f^*(s,a) = \dfrac{\rho^*(s,a)}{\sum_{a\in A(s)} \rho^*(s,a)}$, for all $(s,a) \in \mathcal{K}$. Since we assume $\gamma(s) > 0$, for all $s \in S$, it follows from (7) that every vector $\rho \in \mathcal{Q}_\alpha(\gamma)$ satisfies $\sum_{a\in A(s)} \rho(s,a) > 0$, for all $s \in S$. Thus, the policy f^* is well-defined.

III. Robust CMDP model

We consider a CMDP problem with known costs and uncertain transition probabilities defined by a single uncertain parameter, u which belongs to an interval $[\underline{u}, \bar{u}]$. The expected discounted costs depend on u, and we denote them by $C^u_\alpha(\gamma, f)$, $D^{k,u}_\alpha(\gamma, f)$, for all $k \in \mathbb{K}$. The corresponding robust CMDP problem restricted to the class of stationary policies is defined as

$$\begin{aligned} &\min_{f\in F_S} \ \max_{u\in[\underline{u},\bar{u}]} \ C^u_\alpha(\gamma, f) \\ &\quad \text{s.t.} \max_{u\in[\underline{u},\bar{u}]} \ D^{k,u}_\alpha(\gamma, f) \le \xi_k, \ \forall\, k \in \mathbb{K}. \end{aligned} \tag{8}$$

Similar to (4) and (5), we can write

$$\begin{aligned} C^u_\alpha(\gamma, f) &= (1-\alpha)\gamma^{\mathrm{T}}\big(I - \alpha P^u_f\big)^{-1} c_f, \\ D^{k,u}_\alpha(\gamma, f) &= (1-\alpha)\gamma^{\mathrm{T}}\big(I - \alpha P^u_f\big)^{-1} d^k_f, \ \forall\, k \in \mathbb{K}, \end{aligned}$$

where P^u_f is the transition probability matrix, under u, induced by f, whose $(s,s')^{th}$ entry is defined by, $P^u_f(s'|s) = \sum_{a\in A(s)} f(s,a)p^u(s'|s,a)$. For each $(s,a) \in \mathcal{K}$, $s' \in S$, $p^u(s'|s,a)$ denotes the uncertain transition probability of moving to a state s' from a state s when an action $a \in A(s)$ is taken. We propose two specific uncertainty structures for the transition probabilities.

(S1) There exists an uncertain state, $\hat{s}$, such that the transition probabilities to all the states from all state-action pairs $(\hat{s}, a)$, $a \in A(\hat{s})$, are uncertain while other transition probabilities are deterministic. For all $(s,a) \in \mathcal{K}, s' \in S$, they are given by

$$p^u(s'|s,a) = \begin{cases} p(s'|s,a) + um(s'|s,a); & s = \hat{s}, \\ p(s'|s,a); & \text{otherwise}, \end{cases}$$

where $m(s'|\hat{s},a)$ are known scalars such that $\sum_{s'\in S} m(s'|\hat{s},a) = 0$, for all $a \in A(\hat{s})$.

(S2) Transition probabilities corresponding to all state-action pairs are uncertain such that for all $(s,a) \in \mathcal{K}, s' \in S$, $p^u(s'|s,a) = p(s'|s,a) + u\beta(s)m(s')$, where $\beta(s)$ and $m(s')$ are known scalars such that $\sum_{s'\in S} m(s') = 0$.

Under (S1) and (S2), for each $(s,a) \in \mathcal{K}$, $s' \in S$, $p(s'|s,a)$ is the observed transition probability while $m(s'|s,a)$, $m(s')$, and $\beta(s)$ are known scalars. We assume that the uncertain parameter, u and the known scalars are such that $p^u(s'|s,a) \ge 0$, $(s,a) \in \mathcal{K}$, $s' \in S$, i.e., they define transition probabilities. Thus, we can represent P^u_f in matrix form by $P^u_f = P_f + uM_f$, and under both (S1) and (S2), the rank of the matrix M_f is 1. Therefore, the matrix $\big(I - \alpha P^u_f\big)$ can be viewed as a sum of a full rank matrix, $\big(I - \alpha P^{\underline{u}}_f\big)$ and a rank 1 matrix, $\alpha(\underline{u} - u)M_f$. The inverse of such a matrix is given by the following Lemma 1.

Lemma 1. *([11]) For given non-singular matrices G and $G{+}H$, where H is a matrix of rank 1, $\big(G{+}H\big)^{-1} = G^{-1} - \dfrac{G^{-1}HG^{-1}}{1 + \mathrm{Tr}(HG^{-1})}$, where $\mathrm{Tr}(HG^{-1})$ denotes the trace of the matrix, HG^{-1} such that $1{+}\mathrm{Tr}(HG^{-1}) \ne 0$.*

Let $Q^u_f = \big(I - \alpha P^u_f\big)^{-1}$. The matrices $G = I - \alpha P^{\underline{u}}_f$ and $H = \alpha(\underline{u}{-}u)M_f$ satisfy the conditions of Lemma 1 and $Q^u_f = \big(G + H\big)^{-1}$. Thus,

$$Q^u_f = Q^{\underline{u}}_f + \frac{\alpha(u - \underline{u})}{1 - \alpha(u - \underline{u})\mathrm{Tr}\big(M_f Q^{\underline{u}}_f\big)} Q^{\underline{u}}_f M_f Q^{\underline{u}}_f.$$

Lemma 2. *For a given $f \in F_S$, the function $g : [\underline{u}, \bar{u}] \to \mathbb{R}$ defined as $g(u) = \dfrac{u - \underline{u}}{1 - \alpha(u - \underline{u})\mathrm{Tr}\big(M_f Q^{\underline{u}}_f\big)}$, is strictly increasing in u.*

Proof. This follows from the first-order derivative of g, $g'(u) = \dfrac{1}{\big(1 - \alpha(u - \underline{u})\mathrm{Tr}\big(M_f Q^{\underline{u}}_f\big)\big)^2} > 0.$ □

It follows from Lemma 2 that depending on the sign of the terms $\gamma^T Q^{\underline{u}}_f M_f Q^{\underline{u}}_f c_f$ and $\gamma^T Q^{\underline{u}}_f M_f Q^{\underline{u}}_f d^k_f$, $k \in \mathbb{K}$, the optimal solution of each inner optimization problem in (8) occurs at either endpoint of the interval $[\underline{u}, \bar{u}]$. We define $z = \max_{u\in\{\underline{u},\bar{u}\}} C^u_\alpha(\gamma, f)$ and equivalently write (8) as

$$\begin{aligned} &\min_{z, f\in F_S} \ z \\ &\text{s.t.} \ (1-\alpha)\gamma^T Q^u_f c_f \le z, \ \forall\, u \in \{\underline{u}, \bar{u}\}, \\ &(1-\alpha)\gamma^T Q^u_f d^k_f \le \xi_k, \ \forall\, k \in \mathbb{K}, \ u \in \{\underline{u}, \bar{u}\}. \end{aligned} \tag{9}$$

A. Bilinear programming formulation

Using the occupation measure corresponding to stationary policies, we show that (9) is equivalent to a BP problem.

Theorem 3. *The optimization problem (9) is equivalent to the following BP problem*

$$\min_{z,\, \rho_{\underline{u}}, \rho_{\bar{u}} \in \mathbb{R}^{|\mathcal{K}|}} \ z$$

$$\text{s.t.} \sum_{(s,a)\in\mathcal{K}} \rho_{\underline{u}}(s,a)c(s,a) \leq z, \ \rho_{\underline{u}} \in \mathcal{Q}^{\underline{u}}_{\alpha}(\gamma), \tag{10a}$$

$$\sum_{(s,a)\in\mathcal{K}} \rho_{\underline{u}}(s,a)d^k(s,a) \leq \xi_k, \ \forall\, k \in \mathbb{K}, \tag{10b}$$

$$\sum_{(s,a)\in\mathcal{K}} \rho_{\bar{u}}(s,a)c(s,a) \leq z, \ \rho_{\bar{u}} \in \mathcal{Q}^{\bar{u}}_{\alpha}(\gamma), \tag{10c}$$

$$\sum_{(s,a)\in\mathcal{K}} \rho_{\bar{u}}(s,a)d^k(s,a) \leq \xi_k, \ \forall\, k \in \mathbb{K}, \tag{10d}$$

$$\rho_{\underline{u}}(s,a)\Bigg(\sum_{a\in A(s)} \rho_{\bar{u}}(s,a)\Bigg) = \rho_{\bar{u}}(s,a)\Bigg(\sum_{a\in A(s)} \rho_{\underline{u}}(s,a)\Bigg), \quad \forall\, (s,a) \in \mathcal{K}, \tag{10e}$$

where $\mathcal{Q}^{u}_{\alpha}(\gamma)$, is defined as in (7), with $p(s'|s,a)$ replaced by $p^u(s'|s,a)$, for $u \in \{\underline{u}, \bar{u}\}$.

Proof. Let (z, f) be a feasible vector of (9). For each $(s,a) \in \mathcal{K}$, $u \in \{\underline{u}, \bar{u}\}$, define $\rho_u(s,a) = g_\alpha(\gamma, f, u; s, a)$, where $g_\alpha(\gamma, f, u; s, a)$ is defined by (6) using the transition probability matrix P^u_f. Since $\gamma(s) > 0$, for all $s \in S$, it follows from (6) that $\sum_{a\in A(s)} g_\alpha(\gamma, f, u; s, a) > 0$, for all $s \in S$. It follows from Theorem 3.2 of [1] that $\rho_{\underline{u}}$ and $\rho_{\bar{u}}$ satisfy (10a-b) and (10c-d), respectively, and for each $(s,a) \in \mathcal{K}$, $u \in \{\underline{u}, \bar{u}\}$, f satisfies $f(s,a) = \dfrac{\rho_u(s,a)}{\sum_{a\in A(s)} \rho_u(s,a)}$. Hence, ρ_u, $u \in \{\underline{u}, \bar{u}\}$, satisfies (10e). Therefore, $(z, \rho_{\underline{u}}, \rho_{\bar{u}})$ is a feasible vector of (10). Conversely, let $(z, \rho_{\underline{u}}, \rho_{\bar{u}})$ be a feasible vector of (10). Since $\gamma(s) > 0$, for all $s \in S$, equality constraints in (10a) and (10c) imply that $\sum_{a\in A(s)} \rho_u(s,a) > 0$, for all $s \in S$, $u \in \{\underline{u}, \bar{u}\}$. From (10e), define for each $(s,a) \in \mathcal{K}$, $u \in \{\underline{u}, \bar{u}\}$, $f(s,a) = \dfrac{\rho_u(s,a)}{\sum_{a\in A(s)} \rho_u(s,a)}$. Furthermore, since $\rho_{\underline{u}}$ and $\rho_{\bar{u}}$ satisfy (10a-b) and (10c-d), respectively, it follows from the proof of Theorem 3.2 of [1] that, $\{\rho_u(s,a)\}_{(s,a)\in\mathcal{K}} = \{g_\alpha(\gamma, f, u; s, a)\}_{(s,a)\in\mathcal{K}}$, $u \in \{\underline{u}, \bar{u}\}$. Therefore, (z, f) is a feasible vector of (9). □

Theorem 3 shows that the reformulation of (9) using occupation measures is a BP problem with bilinear terms limited to (10e). It is well known that a bilinear equality constraint can be approximated by a McCormick envelope ([10], [4]). In Section III-B, we propose an algorithm to solve (10), based on solving a sequence of LP problems. At every iteration of the algorithm, the McCormick envelope gets finer and produces a tighter lower bound. The algorithm eventually converges to a global optimal solution of (10).

B. LP-based algorithm

For notational simplicity, let $\rho_u(s) = \sum_{a\in A(s)} \rho_u(s,a)$, for all $s \in S$, $u \in \{\underline{u}, \bar{u}\}$. The constraints (10e) can be equivalently written as

$$\begin{aligned} &w_{\underline{u}}(s,a) = w_{\bar{u}}(s,a), \ w_{\underline{u}}(s,a) = \rho_{\underline{u}}(s,a)\rho_{\bar{u}}(s), \\ &w_{\bar{u}}(s,a) = \rho_{\bar{u}}(s,a)\rho_{\underline{u}}(s), \ \forall\, (s,a) \in \mathcal{K}. \end{aligned} \tag{11}$$

In order to approximate the bilinear constraints of (11), we construct lower and upper bounds for each term used in the product. From the proof of Theorem 3, for a given ρ_u, $u \in \{\underline{u}, \bar{u}\}$, there exists an $f \in F_S$ such that $\rho_u(s,a) = g_\alpha(\gamma, f, u; s, a)$, for all $(s,a) \in \mathcal{K}$. Hence, from (6) we have $\rho_u(s) = \big((1-\alpha)\gamma^T Q^u_f\big)_s$, for all $s \in S$. Define an $|S| \times |S|$ matrix, $P^u_{\min} = \big(p^u_{\min}(s'|s)\big)_{s,s'}$, where $p^u_{\min}(s'|s) = \min_{a\in A(s)} p^u(s'|s,a)$, for all $s, s' \in S$. Thus, $P^u_f \succeq P^u_{\min}$, for all $f \in F_S$, where $\succeq$ denotes componentwise inequality. Moreover, since $\big(I - \alpha P^u_f\big)$ and $\big(I - \alpha P^u_{\min}\big)$ are M-matrices, it follows from Theorem 1.8 of [6] that, $Q^u_f \succeq \big(I - \alpha P^u_{\min}\big)^{-1}$. Thus, a lower bound to $\rho_u(s)$, is given by

$$\rho^L_u(s) = \big((1-\alpha)\gamma^T \big(I - \alpha P^u_{\min}\big)^{-1}\big)_s. \tag{12}$$

Since $\sum_{s\in S} \rho_u(s) = 1$, an upper bound to $\rho_u(s)$ is given by

$$\rho^U_u(s) = 1 - \sum_{s'\neq s} \rho^L_u(s'), \ \forall\, s \in S. \tag{13}$$

Therefore, for each $s \in S$, $\rho_u(s) \in [\rho^L_u(s), \rho^U_u(s)]$, and thus, for each $(s,a) \in \mathcal{K}$, $\rho_u(s,a) \in [0, \rho^U_u(s)]$. We approximate (11) by the following McCormick envelope

$$\begin{aligned} &w_{\underline{u}}(s,a) = w_{\bar{u}}(s,a), \ w_{\underline{u}}(s,a) \leq \rho^U_{\bar{u}}(s)\rho_{\underline{u}}(s,a), \\ &w_{\underline{u}}(s,a) \leq \rho^L_{\bar{u}}(s)\rho_{\underline{u}}(s,a) + \rho^U_{\underline{u}}(s)\rho_{\bar{u}}(s) - \rho^U_{\underline{u}}(s)\rho^L_{\bar{u}}(s), \\ &w_{\underline{u}}(s,a) \geq \rho^L_{\bar{u}}(s)\rho_{\underline{u}}(s,a), \ w_{\bar{u}}(s,a) \leq \rho^U_{\underline{u}}(s)\rho_{\bar{u}}(s,a), \\ &w_{\underline{u}}(s,a) \geq \rho^U_{\bar{u}}(s)\rho_{\underline{u}}(s,a) + \rho^U_{\underline{u}}(s)\rho_{\bar{u}}(s) - \rho^U_{\underline{u}}(s)\rho^U_{\bar{u}}(s), \\ &w_{\bar{u}}(s,a) \leq \rho^L_{\underline{u}}(s)\rho_{\bar{u}}(s,a) + \rho^U_{\bar{u}}(s)\rho_{\underline{u}}(s) - \rho^U_{\bar{u}}(s)\rho^L_{\underline{u}}(s), \\ &w_{\bar{u}}(s,a) \geq \rho^U_{\underline{u}}(s)\rho_{\bar{u}}(s,a) + \rho^U_{\bar{u}}(s)\rho_{\underline{u}}(s) - \rho^U_{\bar{u}}(s)\rho^U_{\underline{u}}(s), \\ &w_{\bar{u}}(s,a) \geq \rho^L_{\underline{u}}(s)\rho_{\bar{u}}(s,a), \ \forall\, (s,a) \in \mathcal{K}. \end{aligned} \tag{14}$$

Therefore, an LP approximation of (10) is given by

$$\min_{z,\, \rho_{\underline{u}}, \rho_{\bar{u}}, w_{\underline{u}}, w_{\bar{u}}} \ z \quad \text{s.t.} \ (10a-d), \ (14). \tag{15}$$

The optimal value of (15) is a lower bound to the optimal value of (10). If an optimal solution of (15) satisfies (11), it is also an optimal solution of (10). This motivates us to consider the following stopping criterion,

$$\max_{(s,a)\in\mathcal{K}} |\rho_{\underline{u}}(s,a)\rho_{\bar{u}}(s) - \rho_{\bar{u}}(s,a)\rho_{\underline{u}}(s)| \leq \epsilon, \tag{16}$$

Algorithm 1 : LP-based algorithm

```
Input : All known parameters of (10), $\epsilon$
Output : Optimal solution of (10)
initialization : Fix $\rho^L_{\underline{u}}$, $\rho^U_{\underline{u}}$, $\rho^L_{\bar{u}}$, $\rho^U_{\bar{u}}$ from (12) and (13), $j = 0$, $B_j = \{[\rho^L_u, \rho^U_u]\}_{u \in \{\underline{u}, \bar{u}\}}$, $\bar{B}_j = \{\}$
solve $(15)_{B_j}$ and get $(z^*, \rho^*_{\underline{u}}, \rho^*_{\bar{u}})$
    ▷ $(15)_{B_j}$: LP problem (15) where (14) is constructed using bounds of $B_j$
while (16) is false for $(\rho^*_{\underline{u}}, \rho^*_{\bar{u}})$ do
    find $s'$ that satisfies (17)
    solve $(15)_{SR_1(B_j)}$
    solve $(15)_{SR_2(B_j)}$
    solve $(15)_{SR_3(B_j)}$
    solve $(15)_{SR_4(B_j)}$
        ▷ $(15)_{SR_i(B_j)}$: LP problem (15) where (14) is constructed using bounds of $SR_i(B_j)$; $SR_i(B_j)$ is obtained by splitting $B_j$ as in (SR$_i$), $i = 1, 2, 3, 4$
    $\mathbb{R} = \{\{SR_i(B_j)\}_{i=1}^4, \bar{B}_j\}$
    $B_{j+1} \in \arg\min_{R \in \mathbb{R}} z^*\big((15)_R\big)$    ▷ $z^*\big((15)_R\big)$: optimal value, $z^*$ of $(15)_R$
    $\bar{B}_{j+1} = \mathbb{R} \backslash B_{j+1}$
    $j \leftarrow j + 1$
    $\rho^*_{\underline{u}} \leftarrow \rho^*_{\underline{u}}\big((15)_{B_j}\big)$ ▷ $\rho^*_{\underline{u}}\big((15)_{B_j}\big)$: optimal solution, $\rho^*_{\underline{u}}$ of $(15)_{B_j}$
    $\rho^*_{\bar{u}} \leftarrow \rho^*_{\bar{u}}\big((15)_{B_j}\big)$ ▷ $\rho^*_{\bar{u}}\big((15)_{B_j}\big)$: optimal solution, $\rho^*_{\bar{u}}$ of $(15)_{B_j}$
end while
```

for some pre-fixed small $\epsilon > 0$, for the algorithm based on solving a sequence of LPs given by (15). For a sufficiently small ϵ, if an optimal solution $(z^*, \rho^*_{\underline{u}}, \rho^*_{\bar{u}}, w^*_{\underline{u}}, w^*_{\bar{u}})$ of (15) satisfies (16), we declare it to be an optimal solution of (10). However, if (16) is not satisfied, we split the feasible region of (10) at the decision variables ([14], [16]). For splitting, we select the state index, s', that contradicts the equality by the largest absolute value, i.e.,

$$s' \in \arg\max_{(s,a) \in \mathcal{K}} |\rho^*_{\underline{u}}(s,a)\rho^*_{\bar{u}}(s) - \rho^*_{\bar{u}}(s,a)\rho^*_{\underline{u}}(s)| \qquad (17)$$

and split $\rho_u(s')$ at the midpoint of the interval $[\rho^L_u(s'), \rho^U_u(s')]$, $u \in \{\underline{u}, \bar{u}\}$. This is one of the standard points for splitting ([4]). This gives four sub-regions, (SR$_i$), $i = 1, 2, 3, 4$. We intersect (10) with each sub-region, update the lower and upper bounds corresponding to state s', which satisfies (17) and keep the bounds other other states the same. We approximate each sub-problem by generating finer McCormick envelopes ([4]) using the updated lower and upper bounds. The optimal value of the LP approximation of i^{th}, $i = 1, 2, 3, 4$, sub-problem gives a lower bound to the optimal value of (10) in the sub-region (SR$_i$). If the optimal solution of the LP approximation of a sub-problem with the lowest optimal value satisfies (16), we stop and declare the optimal solution of the sub-problem as an optimal solution of (10). Otherwise, we split the feasible sub-region, corresponding to the lowest optimal value of (10) into four sub-regions and repeat the procedure. Selecting the sub-region with the lowest optimal value among all the sub-regions, for further splitting guarantees that we do not lose the global optimal solution of (10). We summarize the procedure in Algorithm 1.

IV. Numerical Experiments

We perform numerical experiments, using Algorithm 1, on a class of CMDPs, motivated from Garnets ([3], [18]). The experiments are performed in MATLAB using Yalmip toolbox ([8]) with Gurobi solver on an Intel(R) 64-bit Core(TM) i5-8250U CPU @ 1.60GHz with 8.0 GB RAM machine. We fix $\alpha = 0.7$ and γ to be a randomly generated probability distribution.

We fix the parameter tuple $(|S|, |A|, |B_F|)$, where $|A|$ is the number of actions (we assume that same actions are available at each state), and $|B_F|$ is the number of states reachable from every state-action pair. From each $(s,a) \in \mathcal{K}$, we choose $|B_F|$ states and denote them by $B_F^{(s,a)} = \{s^{i_{B_F}}\}_{i=1}^{|B_F|}$. For each (s,a), we randomly generate $|B_F|-1$ values and denote them by $\big(q^i\big)_{i=1}^{|B_F|-1}$. Following Section 4.2 of [18], the transition probabilities obtained via observation history are defined as

$$p(s'|s,a) = \begin{cases} q^i - q^{i-1}; s' = s^{i_{B_F}},\ i = 1, 2, \ldots, |B_F|, \\ 0; \text{otherwise}, \end{cases}$$

where $q^0 = 0$ and $q^{|B_F|} = 1$.

We assume that the observed transition probabilities are uncertain and follow the uncertainty structure (S1). We randomly pick a state $\hat{s} \in S$ and for each action $a \in A(\hat{s})$, we choose $0.25|B_F|$ states from $B_F^{(\hat{s},a)}$ such that the transition probability to these states from $\hat{s}$ is more than the transition probability for rest of the states of $B_F^{(\hat{s},a)}$. Let these states be denoted by $B_{0.25F}^{(\hat{s},a)}$. We generate the weights for (S1) as: $m(s'|\hat{s},a) \in \Big[-\min_{s'} p(s'|\hat{s},a), \min_{s'} p(s'|\hat{s},a)\Big]$ if $s' \in B_{0.25F}^{(\hat{s},a)}$, and $m(s'|\hat{s},a) = 0$, otherwise, such that $\sum_{s' \in B_{0.25F}^{(\hat{s},a)}} m(s'|\hat{s},a) = 0$. We assume $u \in [-1, 1]$. We fix $|S| = 500$, $|A|, |\mathbb{K}| = 10$. We randomly generate the components of $c \in \mathbb{R}^{5000}$ from $(50, 500)$, $d^k \in \mathbb{R}^{5000}$ from $(50, 200)$, and $\xi \in \mathbb{R}^{10}$ from $(100, 150)$. We consider two values of $|B_F|$: (i) 100 (20% reachable states) and (ii) 400 (80% reachable states). For the above data, we solve (10) using the existing bilinear solver and Algorithm 1 with $\epsilon = 10^{-7}$. For Algorithm 1, we terminate the execution of the while loop after a wall-clock time of 14400 seconds. We summarize the results in Figure 1. We conclude that the bilinear solver and Algorithm 1 give the same optimal value in each instance. In the instances marked with '$*$', the Algorithm 1 exceeded the fixed time and was terminated for $|B_F| = 100$. In the figure, we summarize the CPU time taken till the termination of execution for these cases. In terms of the CPU time taken, we observe that the bilinear solver performs better in 15 instances with $|B_F| = 100$, while Algorithm 1 performs better in 19 instances with $|B_F| = 400$. In addition, the bilinear solver takes less

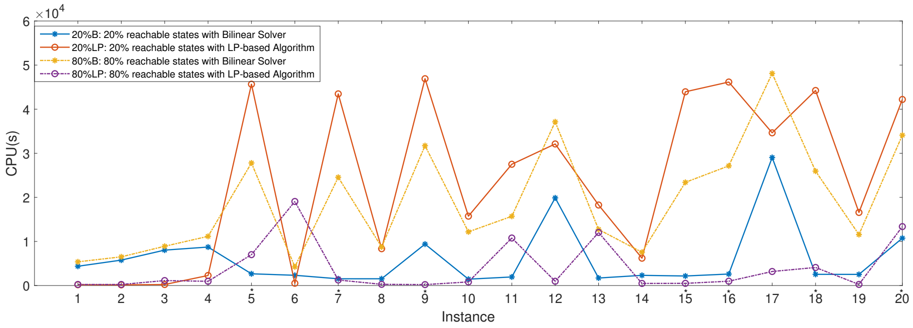

Fig. 1. Random CMDPs.

time to solve the problems with $|B_F| = 100$ than with $|B_F| = 400$ in each instance, thereby indicating that it depends on the problem size. In contrast, Algorithm 1 takes less time to solve problems with $|B_F| = 400$ than with $|B_F| = 100$ in 16 instances. This may be because a larger value of $|B_F|$, forces the components of m in a smaller interval. Hence, starting from a tight bound of the decision variables in the bilinear equality constraints and splitting solves the problem in relatively less time.

V. Conclusion

We consider a CMDP problem with uncertain transition probabilities defined by a single uncertain parameter. Under a stationary class of policies, the robust CMDP problem is equivalent to a BP problem. We propose an algorithm to compute its global optimal solution and compare the performance of our algorithm with the Gurobi bilinear solver on random CMDPs. In most cases, our algorithm performs better than Gurobi bilinear solver as the density of the transition probability matrix increases. For a dense matrix, the interval of uncertainty becomes small. Thus, starting from tight lower bounds, our algorithm converges faster compared to Gurobi bilinear solver, which is designed for general BP problems. The CMDP formulation in this paper considers the case where uncertainty in transition probabilities is driven by a single uncertain parameter. However, in many realistic situations, there could be multiple uncertain parameters. We plan to explore this direction in our future research.

Acknowledgement

The research of the first author was supported by CSIR, India. The research of the second author was supported by DST/CEFIPRA Project No. IFC/4117/DST-CNRS-5th call/2017-18/2.

References

[1] E. Altman. *Constrained Markov Decision Processes*. Chapman and Hall/CRC, London, 1999.

[2] T. W. Archibald, K. I. M. McKinnon, and L. C. Thomas. On the generation of Markov decision processes. *Journal of the Operational Research Society*, 46(3):354–361, 1995.

[3] Layla El Asri, Bilal Piot, Matthieu Geist, Romain Laroche, and Olivier Pietquin. Score-based inverse reinforcement learning. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, AAMAS '16, page 457–465, Richland, SC, 2016. International Foundation for Autonomous Agents and Multiagent Systems.

[4] Matteo Fischetti and Michele Monaci. A branch-and-cut algorithm for mixed-integer bilinear programming. *European Journal of Operational Research*, 282(2):506–514, 2020.

[5] Garud N Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280, 2005.

[6] Charles R Johnson and Ronald L Smith. Inverse M-matrices, II. *Linear algebra and its applications*, 435(5):953–983, 2011.

[7] Erim Kardes. A robust constrained Markov decision process model for admission control in a single server queue. In *Proceedings of the 2014 International Conference on Industrial Engineering and Operations Management*, pages 1606–1615, 2014.

[8] J. Löfberg. Yalmip : A toolbox for modeling and optimization in MATLAB. In *In Proceedings of the CACSD Conference*, pages 284–289, Taipei, Taiwan, 2004.

[9] Shie Mannor, Duncan Simester, Peng Sun, and John N Tsitsiklis. Bias and variance approximation in value function estimates. *Management Science*, 53(2):308–322, 2007.

[10] Garth P McCormick. Computability of global solutions to factorable nonconvex programs: Part i—convex underestimating problems. *Mathematical programming*, 10(1):147–175, 1976.

[11] Kenneth S Miller. On the inverse of the sum of matrices. *Mathematics magazine*, 54(2):67–72, 1981.

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

[13] Martin L. Puterman. *Markov Decision Process*. John Wiley & Sons, USA, 1st edition, 1994.

[14] Ignacio Quesada and Ignacio E Grossmann. A global optimization algorithm for linear fractional and bilinear programs. *Journal of Global Optimization*, 6:39–76, 1995.

[15] Jay K Satia and Roy E Lave Jr. Markovian decision processes with uncertain transition probabilities. *Operations Research*, 21(3):728–740, 1973.

[16] Hanif D Sherali and Amine Alameddine. A new reformulation-linearization technique for bilinear programming problems. *Journal of Global optimization*, 2:379–410, 1992.

[17] V Varagapriya, Vikas Vikram Singh, and Abdel Lisser. Constrained Markov decision processes with uncertain costs. *Operations Research Letters*, 50(2):218–223, 2022.

[18] V Varagapriya, Vikas Vikram Singh, and Abdel Lisser. Joint chance-constrained Markov decision processes. *Annals of Operations Research*, 322:1013–1035, 2023.

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