

Hyperbolic Behavior Of Occupation Measures Between Neighboring Policies In CMDPs

Alexander Zadorojnyi and Guy Even

Abstract—We study the change in the steady state probabilities as the controller of a Markov Decision Process (MDP) shifts from one deterministic policy to another by gradually changing the selected action in a single state. We prove that the steady state probability for each state-action pair is a hyperbolic Möbius transformation. In particular, this implies that the change is monotone. The same holds also for the cost function in the discounted cost and expected average cost models. We extend this result also to constrained MDPs with an arbitrary number of constraints.

I. INTRODUCTION

Every stationary policy of a Markov Decision Process (MDP) induces an occupation measure over pairs of states and actions. Consider a controller that shifts from one deterministic policy to another by gradually changing its action in a single state. In this paper we investigate the question of how the occupation measure and the cost behave as result of such a change in the policy.

We prove that each component of the occupation measure is a hyperbolic Möbius transformation over the reals (i.e., a rectangular hyperbola) of the amount by which the controller shifts from one deterministic policy to the other. In particular, this implies that each component of the occupation measure changes monotonically. The proof also carries over to the cost function both in the discounted cost and the expected average cost models. The result extends to constrained Markov Decision Processes (CMDPs) with an arbitrary number of constraints.

Related Work: The question of the effect of the discount factor on the optimal cost of a CMDP was studied by Altman et al. in [2]. They proved that, when feasible, the cost is a piecewise rational function of the discount factor. Note that each rational function is the quotient of two high degree polynomials. This result extended a previous result of Smallwood for MDPs [9].

Previous Work: This paper was motivated by [10] where a new type of algorithm was presented for finding optimal policies in MDPs. Loosely speaking, the algorithm scans the polytope of all feasible occupation measures by adding an extra constraint. The value of the extra constraint is gradually changed, so that the intersection with the polytope cuts “slices” of the polytope. In each such slice, an optimal solution is computed. Finally, the cheapest solution among these optimal solutions is returned as the optimal occupation measure of the MDP. It turns out that the optimal solutions in the slices form a path between deterministic policies that disagree in a single state. In each move along the path, only one action is changed in a single state. Thus it is of interest

to investigate the change in the occupation measure as one moves between two neighboring deterministic policies.

We conclude with an application. Following [10], a uniqueness condition is obtained by applying a random perturbation. This uniqueness condition is required for the algorithm presented in [10], and we extend it to CMDPs. The proof presented here is simplified thanks to the monotonicity result. In addition, we perturb the extra cost vector instead of the optimized cost vector; this has the advantage of not affecting the optimum cost.

Organization: In Sec II we overview the definitions of MDPs, CMDPs, and their linear programming formulations. In Sec. III we define vertex and edge policies. In the case of an MDP, a vertex policy is simply a deterministic policy, and an edge policy is a policy with a single randomization. In Sec. IV we prove the result for MDPs. In Sec. V we prove the result for CMDPs. In Sec. VI we apply the result to obtain a uniqueness property. Conclusions are presented in Sec. VII.

II. BACKGROUND

In this section we briefly overview the topics of MDPs, CMDPs, and their linear programming formulations. See [1], [8] for more material on these topics.

A. Definition of MDP and CMDP.

An MDP is a 4-tuple $\langle X, U, P, c \rangle$, where $X = \{0, \dots, n-1\}$ is a finite set of states, $U = \{0, \dots, k-1\}$ is a finite set of actions, $P : X^2 \times U \rightarrow [0, 1]$ is a transition probability function, and $c : X \times U \rightarrow \mathbb{R}$ is a cost function. The probability of the transition from state x to state y when the action u is chosen is specified by the function P and denoted by $P(y|x, u)$. The cost associated with selecting the action u when in state x equals $c(x, u)$. We often refer to the cost function as a vector $c \in \mathbb{R}^{nk}$.

Time is discrete, and in each time unit t , let x_t denote the random variable that equals the state in time t . We assume an initial distribution for initial state at time $t = 0$. Similarly let u_t denote the random variable that equals the action selected in time t . The sequence of states $\{x_t\}_{t=1}^{\infty}$ defines an infinite random walk over the set of states X .

A (stationary) policy¹ is a function $\pi : X \times U \rightarrow [0, 1]$ such that $\sum_{u \in U} \pi(x, u) = 1$, for every $x \in X$. A policy controls the action selected in each state as follows: the probability of selecting action u in state x equals $\pi(x, u)$. If for a state x and an action u the policy π satisfies $\pi(x, u) = 1$, then we

¹By the general theory of MDPs and CMDPs [8], [1], under our conditions there exists an optimal stationary policy. Therefore we restrict our attention to such policies.

say that π is *deterministic* in state x . In this case we abuse notation and write $\pi(x) = u$. If there exists an action u such that $0 < \pi(x, u) < 1$, then we say that π is *randomized* in state x . A *deterministic* policy is a policy that is deterministic in all states.

The goal is to find a policy π that minimizes the cost $C(\pi)$ defined below. We consider two cost models: discounted cost and expected average cost, defined below.

Discounted cost model: In the discounted cost model, the parameter $\beta \in (0, 1)$ specifies the rate in which future costs are reduced. Let $P^\pi(x_t = x, u_t = u)$ denote the probability of the event $x_t = x$ and $u_t = u$ when the policy is π . The expected cost $E_t^\pi[c(x_t, u_t)]$ equals

$$E_t^\pi[c(x_t, u_t)] = \sum_{x \in X, u \in U} c(x, u) \cdot P^\pi(x_t = x, u_t = u).$$

The infinite horizon discounted expected cost $C(\pi)$ is defined by

$$C(\pi) \triangleq (1 - \beta) \cdot \sum_{t=0}^{\infty} \beta^t \cdot E_t^\pi[c(x_t, u_t)]. \quad (1)$$

Expected average cost model: In the expected average cost model, the cost $C(\pi)$ is defined by

$$C(\pi) \triangleq \lim_{T \rightarrow \infty} \left(\frac{\sum_{t=0}^{T-1} E_t^\pi[c(x_t, u_t)]}{T} \right). \quad (2)$$

It can be shown that this limit exists for every stationary policy [8].

Occupation measures: Every policy π induces a limit probability measure over the state-action pairs. We call this probability measure the *occupation measure* corresponding to π and denote it by ρ_π . The definition of ρ_π depends on the cost model. In the discounted cost model

$$\rho_\pi(x, u) \triangleq (1 - \beta) \cdot \sum_{t=0}^{\infty} \beta^t \cdot P^\pi(x_t = x, u_t = u).$$

In the expected average cost model

$$\rho_\pi(x, u) \triangleq \lim_{T \rightarrow \infty} \left(\frac{\sum_{t=0}^{T-1} P^\pi(x_t = x, u_t = u)}{T} \right).$$

In both cost models, $C(\pi) = \sum_{x, u} c(x, u) \cdot \rho_\pi(x, u)$.

Definition of CMDP: A Constrained MDP is an MDP with additional constraints. Each additional constraint is specified by a cost function $d_i : X \times U \rightarrow \mathbb{R}$, where $D_i(\pi)$ is defined in the same manner as $C(\pi)$. We denote the number of additional constraints by L . We first focus on the case that the additional constraints are equality constraints, i.e., $D_i(\pi) = b_c(i)$, for $i = 1, \dots, L$. We denote the constrained MDP with the constraints $\{D_i(\pi) = b_c(i)\}_{i=1}^L$ by CMDP.

A policy π is *feasible* if it satisfies all the constraints $D_i(\pi) = b_c(i)$, for $i = 1, \dots, L$. The optimization problem is to find a feasible policy π that minimizes $C(\pi)$.

Irreducibility assumption: An MDP is *irreducible* if every policy π induces an irreducible Markov chain. In the sequel we assume that the MDP is irreducible.

B. Linear Programming Formulation of CMDPs

Following [3], [5], [7] we consider linear programming formulations of MDPs and CMDPs. The linear program define the polytope of feasible occupation measures, namely, the set of occupation measures that are induced by feasible policies.

Consider the CMDP $= \langle X, U, P, c, d_1, \dots, d_L, b_c(1), \dots, b_c(L) \rangle$. We represent each cost function c and d_i as a column vector in \mathbb{R}^{nk} indexed by pairs in $X \times U$, namely, $c_{x,u} = c(x, u)$.

We denote the linear program corresponding to CMDP by LP. The linear program LP is of the form $\min\{c^t \cdot \rho \mid A\rho = b, A_c\rho = b_c, \rho \geq 0\}$. The occupation measure is the variable of the linear program LP and is represented by the column vector $\rho \in \mathbb{R}^{nk}$ indexed by pairs in $X \times U$, namely, the component $\rho_{x,u}$ denotes the value of the occupation measure $\rho(x, u)$. The matrix A_c has L rows, where the i th row of A_c is simply the vector d_i^t . (We denote the transpose of a row vector v by v^t .) The column vector b_c equals $(b_c(1), \dots, b_c(L))^t$. The matrix A and the vector b in the linear programs depend on the number of states, actions, transition probabilities and the cost model. We begin with the discounted cost model.

Discounted cost model: Define the matrix A as follows. For each action $u \in U$, let $P(u)$ denote the $n \times n$ square matrix whose entries are defined by $P(u)_{y,x} \triangleq P(y|x, u)$. The matrix A is an $n \times (nk)$ matrix obtained by concatenating the square matrices $I - \beta P(u)$, namely, $A = [I - \beta P(0) \dots I - \beta P(k-1)]$. The column vector $b \in \mathbb{R}^n$ is defined by $b = (1 - \beta, 0, \dots, 0)^t$, where the zeroth coordinate corresponds to the initial state.

Expected average cost model: In the expected average cost model, the matrix A is obtained as follows. First, consider the $(n+1) \times (nk)$ matrix obtained by adding a row consisting of ones to the concatenation of the matrices $I - P(u)$. The rank of this matrix is n by the Peron-Frobenius Theorem [6], hence we may omit one row without reducing the rank. Let A denote the $n \times (nk)$ matrix obtained after removing a dependent row. The vector b is a unit vector, where the coordinate of the one corresponds to the row consisting of ones in A . Note that the all-ones constraint implies that $\sum_{x,u} \rho(x, u) = 1$.

The followings propositions and theorem were proved for various cost models in [3], [4], [5], [7]. A more recent textbook proof appears in [1, Theorem 3.3].

Proposition 1: If π is a feasible policy of CMDP, then ρ_π is a feasible solution of LP.

Given a feasible solution ρ of LP, the policy π^ρ induced by ρ is defined by $\pi^\rho(x, u) \triangleq \rho(x, u) / \sum_{u'} \rho(x, u')$. (Note that the irreducibility assumption defined below rules out the possibility of $\sum_{u'} \rho(x, u') = 0$.)

Proposition 2: If ρ is a feasible solution of LP, then π^ρ is a feasible policy of CMDP, and $c^t \cdot \rho = C(\pi^\rho)$.

The following theorem shows that an optimal policy of CMDP can be found by solving LP.

Theorem 3: If ρ^* is an optimal solution of LP, then π^{ρ^*} is an optimal policy of CMDP.

The matrix A (in both cost models) has the following important property. For every policy π , the projection of A onto the columns in the set $\{(x, u) : \pi(x, u) > 0\}$ is of rank n . This follows from Gersgorin's Theorem in the discounted cost model and from the Peron-Frobenius Theorem in the expected average cost model [6].

III. VERTEX AND EDGE POLICIES

The support of a policy π is the set $\{(x, u) : \pi(x, u) > 0\}$. We denote the support by $\text{support}(\pi)$.

Definition 4: A policy π is strictly r -randomized if its support contains exactly $n + r$ state-action pairs.

Note that a strictly zero-randomized policy is simply a deterministic policy.

Terminology: The linear programming formulation of a CMDP defines the polytope of all feasible occupation measures. We refer to a policy that induces a basic feasible solution (bfs) as a *vertex* policy. Two basic feasible solutions are *adjacent* if their bases differ in a single column. Two vertex policies are *neighbors* if the corresponding occupation measures are adjacent. A policy that induces an occupation measure on an edge of the polytope is called an *edge* policy. In the case of an MDP the vertex policies are exactly the deterministic policies, and the edge policies are the strictly 1-randomized policies.

Proposition 5: If π is a vertex (edge, resp.) policy then $\text{support}(\pi) \leq n + L$ ($\text{support}(\pi) \leq n + L + 1$, resp.). If π^0 and π^1 are neighboring vertex policies, then $|\text{support}(\pi^0) \cup \text{support}(\pi^1)| \leq n + 1 + L$.

Proof: The occupation measure ρ_π is a basic feasible solution, hence its support is (contained in) a basis. The rank of the constraints in LP is $n + L$. Hence $\text{support}(\pi) \leq n + L$, as required. The difference between bases of adjacent basic feasible solutions is one column, hence $|\text{support}(\pi^0) \cup \text{support}(\pi^1)| \leq n + 1 + L$. ■

The following proposition is based on the fact that if there exists an optimal solution to a linear program, then there exists a bfs solution that is optimal.

Proposition 6: For every feasible policy π there exists a vertex policy π^* such that $C(\pi^*) \leq C(\pi)$.

IV. MDP: OCCUPATION MEASURES ALONG AN EDGE

Notation: Let π^0 and π^1 denote two neighboring deterministic policies. Let $q \in [0, 1]$. Let $\pi^q \triangleq q \cdot \pi^1 + (1 - q) \cdot \pi^0$. Note that, for $0 < q < 1$, the policy π^q is a strictly 1-randomized policy. Let ρ^q denote the occupation measure induced by the policy π^q .

The following lemma proves that every component of ρ^q is a hyperbolic Möbius transformation as a function of $q \in [0, 1]$.

Lemma 7: For every $x \in X$ and $u \in U$, there exist constants $a_{x,u}, g_{x,u}, e, f \in \mathbb{R}$ such that

$$\rho^q(x, u) = \frac{a_{x,u} \cdot q + g_{x,u}}{e \cdot q + f}.$$

Moreover, the constants e and f do not depend on x or u .

Proof: Project the $n \times nk$ matrix A to the $(n + 1)$ columns of state-action pairs in $\text{support}(\pi^0) \cup \text{support}(\pi^1)$.

Let \hat{A} denote the $n \times (n + 1)$ matrix obtained from this projection. Similarly, let $\hat{\rho} \in \mathbb{R}^{n+1}$ denote the projection of the occupation measure ρ to $\text{support}(\pi^0) \cup \text{support}(\pi^1)$. Let x^* denote the state in which π^0 and π^1 disagree. Let $u_0 = \pi^0(x^*)$ and $u_1 = \pi^1(x^*)$. We first claim that the occupation measure ρ_{π^q} is a solution of the following system S of constraints:

$$\hat{A} \cdot \hat{\rho} = b, \quad (3)$$

$$\rho \geq 0, \quad (4)$$

$$\forall (x, u) \notin \text{support}(\pi^0) \cup \text{support}(\pi^1) : \rho(x, u) = 0, \quad (5)$$

$$q \cdot \rho(x^*, u_0) - (1 - q) \cdot \rho(x^*, u_1) = 0. \quad (6)$$

Obviously, ρ^q satisfies the constraints in S except Eq. 6. As for Eq. 6, let $\rho^q(x^*) \triangleq \sum_u \rho^q(x^*, u)$. Note that

$$\begin{aligned} \rho^q(x^*, u_0) &= \rho^q(x^*) \cdot \pi^q(x^*, u_0) \\ &= \rho^q(x^*) \cdot (1 - q). \end{aligned}$$

Similarly,

$$\begin{aligned} \rho^q(x^*, u_1) &= \rho^q(x^*) \cdot \pi^q(x^*, u_1) \\ &= \rho^q(x^*) \cdot q. \end{aligned}$$

By the irreducibility assumption $\rho^q(x^*) > 0$, thus it follows that ρ^q satisfies Eq. 6, as claimed.

We consider the following two cases.

- 1) For every $q \in (0, 1)$, constraint 6 does not depend on the constraints in Eq. 3. In this case we augment the matrix \hat{A} by the row corresponding to constraint 6 to obtain an invertible matrix \bar{A} . Similarly, augment the vector b by one zero to obtain \bar{b} . Thus, $\hat{\rho}_{\pi^q} = (\bar{A})^{-1} \cdot \bar{b}$. The inverse $(\bar{A})^{-1}$ equals $\frac{\text{adj}(\bar{A})}{\det(\bar{A})}$, where $\text{adj}(\bar{A})_{i,j}$ is the adjugate matrix and $\det(\bar{A})$ is the determinant. For simplicity², assume the state of the chain at time $t = 0$ is x_0 . This implies that the only nonzero component in b is the first component b_0 , therefore $\hat{\rho}_{\pi^q}(x, u) = b_0 \cdot \frac{\text{adj}(\bar{A})_{(x,u),1}}{\det(\bar{A})}$. We need to prove that both the adjugate $\text{adj}(\bar{A})_{(x,u),1}$ and the determinant $\det(\bar{A})$ are linear functions of q . Indeed, this follows by computing both values using the last row in which there are only two nonzero entries: q in the component (x^*, u_0) and $-(1 - q)$ in the component (x^*, u_1) . Note that the denominator $\det(\bar{A})$ does not depend on the pair (x, u) , as required.
- 2) There exists a $q' \in (0, 1)$ such that the constraint 6 depends on the constraints in Eq. 3. The support of π^0 is contained in the support of $\text{support}(\pi^0) \cup \text{support}(\pi^1)$, thus ρ_{π^0} is a solution of the system S without the constraint 6. By the dependency it follows that ρ_{π^0} also satisfies the constraint 6 for $q = q'$. However, by the irreducibility assumption, $\rho_{\pi^0}(x^*, u_0) > 0$. Since $\rho_{\pi^0}(x^*, u_1) = 0$, it follows that ρ_{π^0} cannot satisfy the constraint 6 for $q = q'$, a contradiction. ■

²The case of initial distribution not concentrated in a single state is proved similarly.

Corollary 8: $D(\pi^q)$ and $C(\pi^q)$ are hyperbolic Möbius transformations for $q \in [0, 1]$. Thus, $D(\pi^q)$ and $C(\pi^q)$ are monotone in q .

Proof: By Lemma 7 every component of ρ^q is a hyperbolic Möbius transformation of q with the same linear function of q in the denominator. It follows that every linear functional of ρ^q is also a hyperbolic Möbius transformation of q . Finally, hyperbolic Möbius transformations are monotone. ■

V. CMDP: POLICIES ALONG AN EDGE

In this section we extend the results of Section IV to MDPs under constraints, namely, CMDPs.

Instead of neighboring deterministic policies, let π^0 and π^1 denote two neighboring vertex policies. The policy π^q is the edge policy defined by $\pi^q \triangleq q \cdot \pi^1 + (1 - q) \cdot \pi^0$.

The following lemma is a generalization of Lemma 7 for CMDPs.

Lemma 9: For every $x \in X$ and $u \in U$, there exist constants $a_{x,u}, g_{x,u}, e, f \in \mathbb{R}$ such that

$$\rho^q(x, u) = \frac{a_{x,u} \cdot q + g_{x,u}}{e \cdot q + f}.$$

Moreover, the constants e and f do not depend on x or u .

The proof of Lemma 9 is similar to the proof of Lemma 7.

As in the case of MDPs, we summarize with the following corollary.

Corollary 10: $D(\pi^q)$ and $C(\pi^q)$ are hyperbolic Möbius transformations for $q \in [0, 1]$. Moreover, they are monotone in q .

VI. THE UNIQUENESS PROPERTY

Consider a CMDP with L equality constraints and an extra cost function $D(\pi)$. Let $\text{CMDP}(\alpha)$ denote the constrained MDP obtained by adding the constraint $D(\pi) = \alpha$ to CMDP.

The “natural” definition of uniqueness is that every optimal policy of $\text{CMDP}(\alpha)$ is unique. However, solving CMDPs or MDPs using the algorithm from [10] requires a different definition for uniqueness. On one hand, we need to strengthen the definition so that uniqueness holds for all values of α . On the other hand, it suffices to consider only optimal policies of $\text{CMDP}(\alpha)$ that are *vertex* policies of CMDP.

Definition 11 (Vertex uniqueness): $\text{CMDP}(\alpha)$ satisfies *vertex uniqueness* with respect to CMDP if, for every optimal policy π^* of $\text{CMDP}(\alpha)$ that is a vertex policy of CMDP, and for every feasible policy π of $\text{CMDP}(\alpha)$,

$$\pi \neq \pi^* \Rightarrow C(\pi) > C(\pi^*).$$

Definition 12 (Uniqueness): A CMDP and an extra cost function $D(\pi)$ satisfy the *uniqueness property* if, for every $\alpha \in \mathbb{R}$, $\text{CMDP}(\alpha)$ satisfies *vertex uniqueness* with respect to CMDP.

We say that α is *dangerous* if the hyperplane $d^t \cdot \rho = \alpha$ intersects a vertex of the polytope of the feasible occupation measures of CMDP. Uniqueness requires that $\text{CMDP}(\alpha)$ satisfy vertex uniqueness with respect to CMDP if α is dangerous.

Proposition 13: The number of dangerous values of α is bounded by the number of vertex policies of CMDP which is bounded by $k^n \cdot \sum_{r=0}^L \binom{kn-n}{r} \leq L \cdot n^L \cdot k^{n+L}$.

Proof: The number of dangerous values of α is bounded by the number of vertices of the polytope of the feasible occupation measures of CMDP, i.e., basic feasible solutions of LP. Each basic feasible solution is determined by the basis that contains $n + r$ columns for $r \leq L$. In fact, the irreducibility assumption requires that there is at least one column per state in a basis. Hence, the number of basic feasible solutions is bounded by $k^n \cdot \sum_{r=0}^L \binom{kn-n}{r}$, as required. ■

The random Perturbation: Uniqueness is obtained by adding a small random perturbation $\varepsilon \in \mathbb{R}^{nk}$ to the extra cost vector d . We require that the perturbed instance satisfies uniqueness with probability at least $1 - \mu$, for $\mu > 0$.

Let the coordinates ε_i of ε be independent random variables uniformly chosen from the set of integers $\{0 \dots 2^p - 1\}$ (or any set of 2^p distinct reals). The following lemma proves that a random perturbation meets the requirements while increasing the length of each component of the extra cost vector d by $O(L \log n + (n + L) \cdot \log k + \log \frac{1}{\mu}) = \tilde{O}(n + k) + \log \frac{1}{\mu}$ bits. This is done by choosing an appropriate value for p .

Let $d_\varepsilon = d + \varepsilon$. Let $\text{CMDP}_\varepsilon(\alpha)$ (resp. $\text{LP}_\varepsilon(\alpha)$) denote the constrained MDP (resp. linear program) obtained by adding the constraint $d_\varepsilon^t \cdot \rho = \alpha$ to CMDP (resp. LP). Let $D_\varepsilon(\pi)$ denote the cost of a policy π induced by the vector d_ε .

Lemma 14: If $p \geq \log_2 \frac{(L \cdot n^L \cdot k^{n+L})^3}{\mu}$, then CMDP and the cost function $D_\varepsilon(\pi)$ satisfy the uniqueness property with probability at least $1 - \mu$.

Proof: A realization of the vector ε is *bad* if there exists an α such that $\text{CMDP}_\varepsilon(\alpha)$ does not satisfy vertex uniqueness with respect to CMDP. In particular, there exist two feasible policies $\pi^* \neq \pi$ of $\text{CMDP}_\varepsilon(\alpha)$ that satisfy: (i) both π and π^* are optimal policies of $\text{CMDP}_\varepsilon(\alpha)$, and (ii) π^* is a vertex policy of CMDP. Let ρ_{π^*} (resp. ρ_π) denote the occupation measure corresponding to π^* (resp. π). Since π^* is a vertex policy of CMDP, ρ_{π^*} is a bfs of LP.

Without loss of generality, we may assume that π is a vertex policy of $\text{CMDP}_\varepsilon(\alpha)$, and thus, $\rho(\pi)$ is a bfs of $\text{LP}(\alpha)$.

Let R denote the set of pairs (ρ_1, ρ_2) of occupation measures such that (i) ρ_1 and ρ_2 are feasible solutions of LP. (ii) ρ_1 is a bfs of LP, and (iii) there exists an edge or vertex policy π of LP such that $\rho_2 = \rho_\pi$.

By the above discussion and by applying union bound, it follows that

$$\begin{aligned} Pr(\varepsilon \text{ is bad}) &\leq Pr(\exists (\rho_1, \rho_2) \in R \text{ such that } d_\varepsilon^t \rho_1 = d_\varepsilon^t \rho_2) \\ &\leq \sum_{\text{bfs } \rho_1} Pr(\exists \rho_2 \text{ such that } (\rho_1, \rho_2) \in R \\ &\quad \text{and } d_\varepsilon^t \rho_1 = d_\varepsilon^t \rho_2). \end{aligned}$$

Consider ρ_2 in the last expression; it corresponds either to an edge or vertex policy π of CMDP. In Proposition 13 we bounded the number of vertex policies. We now wish to bound the number of edge policies that ρ_2 may correspond

to. If π is an edge policy, then it is the convex combination of two neighboring vertex policies π_0 and π_1 . By Corollary 10, given a (dangerous) α and two neighboring vertex policies π_0 and π_1 , there is at most one edge (or vertex) policy π “between” π_0 and π_1 that satisfies $d_\varepsilon^t \cdot \rho_\pi = \alpha$. Hence, for each bfs ρ_1 , the number of occupation measures ρ_2 such that $(\rho_1, \rho_2) \in R$ is bounded by $(L \cdot n^L \cdot k^{n+L})^2$.

Note that if ρ_1 and ρ_2 are given, then $Pr(d_\varepsilon^t \rho_1 = d_\varepsilon^t \rho_2) \leq 2^{-p}$. We conclude that

$$Pr(\varepsilon \text{ is bad}) \leq (L \cdot n^L \cdot k^{n+L})^3 \cdot 2^{-p}.$$

We conclude that if $p \geq \log_2 \frac{(L \cdot n^L \cdot k^{n+L})^3}{\mu}$, then the probability that uniqueness is not satisfied is bounded by μ . ■

VII. CONCLUSIONS

We characterized the nonlinear change of the occupation measure between neighboring vertex policies by hyperbolic Möbius transformations. An important consequence is the monotone change of each component. We applied this property to a random perturbation that achieves a uniqueness property with high probability.

The results presented here can be extended to the case when some of the constraints of the CMDP are inequality constraints, namely $A_c \rho \leq b_c$. Let ρ_s represents the vector ρ with additional slackness variables. In this case we rewrite the LP representing a given CMDP as $\min\{c^t \cdot \rho_s \mid A \rho_s = b, A_c \rho_s = b_c, \rho_s \geq 0\}$. Note, that only nonzero elements are added to matrix A_c .

REFERENCES

- [1] E. Altman, *Constrained Markov Decision Processes*, Chapman&Hall/CRC,1999.
- [2] E. Altman, A. Hordijk and L. C. M. Kallenberg , On the value in constrained control of Markov chains , *ZOR - Mathematical Methods of Operations Research*, Vol. 44, Issue 3, pp. 387-400, 1996.
- [3] G. De Ghellinck, Les problemes de decisions sequentielles. *Cahiers Centre d'Etudes Recherche Operationnelle*, pp. 161-179, 1960.
- [4] F. D'Epenoux, A probabilistic production and inventory problem. *Management Sci.*, pp. 98-108, 1963. (Translation of an article published in *Revue Francaise Recherche Operationnelle* (1960).)
- [5] C. Derman, On sequential decisions and Markov chains, *Management Sci.*, v9, pp. 16-24, 1962.
- [6] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, 2005.
- [7] A. S. Manne, Linear programming and sequential decisions. *Management Sci.*, pp. 259-267, 1960.
- [8] M. L. Puterman, *Markov Decision Processes*, John Wiley & Sons, 1994.
- [9] R. D. Smallwood, Optimum policy regions for Markov processes with discounting, *Operations Research* 14, pp. 658-669, 1966.
- [10] A. Zadorojniy, G. Even and A. Shwartz, A Strongly Polynomial Algorithm for Controlled Queues, Accepted to *Math. of Operations Research*, 2009.