A Strongly Polynomial Algorithm for Controlled Queues

Alexander Zadorojniy[*] Guy Even[†] Adam Shwartz[‡]

April 19, 2009

Abstract

We consider the problem of computing optimal policies of finite-state, finite-action Markov Decision Processes (MDPs). A reduction to a continuum of constrained MDPs (CMDPs) is presented such that the optimal policies for these CMDPs constitute a path in a graph defined over the deterministic policies. This path contains, in particular, an optimal policy of the original MDP. We present an algorithm based on this new approach that finds this path and thus an optimal policy. In the general case this path might be exponentially long in number of states and actions. We prove that the length of this path is polynomial if the MDP satisfies a coupling property. Thus we obtain a strongly polynomial algorithm for MDPs that satisfy the coupling property. We prove that discrete time versions of controlled $M/M/1$ queues induce MDPs that satisfy the coupling property. The only previously known polynomial algorithm for controlled $M/M/1$ queues in the expected average cost model is based on linear programming (and is not known to be strongly polynomial). Our algorithm works both for the discounted and expected average cost models, and the running time does not depend on the discount factor.

Keywords: Markov Decision Process (MDP), Constrained Markov Decision Process (CMDP), Controlled Queues, Linear Programming, $M/M/1$ Queue, Optimization.

1 Introduction

The problem of designing a strongly polynomial algorithm for finding an optimal policy in a Markov Decision Process (MDP) has been a long standing open problem [4]. The parameters of an MDP are: n - the number of states, k - the number of actions, and B - the length of the input in bits. In the discounted cost model there is an addition parameter $\beta < 1$ called the discount factor. Recently, Ye [26] presented a strongly polynomial combinatorial algorithm for the discounted cost model. This algorithm is based on a predictor-corrector interior-point algorithm.

The well known algorithms for solving MDPs are: value iteration, policy iteration, and linear programming [6, 9, 12, 18]. The running times of the value iteration and policy iteration algorithms in the discounted cost model are polynomial in n, k, B and $1/(1-\beta)$ [12, 26]. The dependence on $1/(1-\beta)$ implies that the algorithm is not strongly polynomial (e.g., when $\beta = 1 - 2^{-n}$). The only nontrivial upper bound on the number of iterations of the policy iteration algorithm (for two actions) that does not depend on the discount factor is $O(2^n/n)$ [14].

In the expected average cost model, the only polynomial algorithm is based on a reduction discovered nearly 50 years ago to linear programming [5, 7, 13]. Linear programming is not known to have strongly polynomial algorithms [20]. Hence the problem of developing a strongly polynomial algorithm for MDPs remains open in the expected average cost model.

[*] School of Electrical Engineering, Tel-Aviv Univ., Tel-Aviv 69978, Israel. sasha@eng.tau.ac.il

[†] School of Electrical Engineering, Tel-Aviv Univ., Tel-Aviv 69978, Israel. guy@eng.tau.ac.il

[‡] Department of Electrical Engineering, Technion, Haifa 32000, Israel. adam@ee.technion.ac.il

Contribution. We introduce a new approach for solving MDPs in the discounted cost model and expected average cost model. The approach is based on adding an artificial constraint with parameter α to obtain a constrained MDP denoted by CMDP(α). We consider the whole range of values for α so that it also includes the value that an optimal policy of the MDP attains. Our approach is based on a new structural lemma that proves that the set of optimal policies of CMDP(α) (for all values of α) constitutes a path in a graph over the deterministic policies. We present an algorithm that finds all the deterministic policies along the path. The optimal policy of the MDP is simply the min-cost policy along this path. We can not rule out the possibility that this path may be exponentially long, and hence the running time of this algorithm might be exponential.

We overcome the problem of a long path by introducing a coupling property. We prove that, if the coupling property holds and if a specific artificial constraint is chosen, then the length of the path is polynomial (i.e., $n \cdot k$). Hence the algorithm becomes strongly polynomial. We prove that the coupling property is satisfied in discrete versions of controlled birth-death processes such as single server controlled $M/M/1$ queues. Such controlled birth-death processes are among the most studied examples of MDPs [25, 1, 10, 23].

When the coupling property holds, the running time of the algorithm is $O(n^4 \cdot k^2)$. This running time holds both in the discounted cost model and the expected average cost model. This compares with the running time of Ye's algorithm which is $O(n^4 \cdot k^4 \cdot \log(nk/(1-\beta)))$. Thus, in addition to coping with the expected average cost model, we reduce the running time in the discounted cost model.

Organization. In Sec. 2 we briefly overview definitions related to MDPs and CMDPs. In Sec. 3 and 4 we present two properties: uniqueness and coupling. We prove that uniqueness can be obtained by randomly perturbing the cost vector. We prove that the coupling property holds in discrete time controlled $M/M/1$ queues. In Sec. 5 we study the structure of optimal policies of CMDP(α), for all values of α. Lemma 17 proves that these optimal policies are a path in a graph over the deterministic policies. In Sec. 6 we present a new algorithm for computing an optimal policy of an MDP. In Sec. 7 we present a strongly polynomial algorithm that works under the assumption that the coupling property holds. We conclude with a discussion of the assumptions that the MDP is irreducible and satisfies the uniqueness property.

2 Background

In this section we briefly overview the topics of MDPs, CMDPs, and their linear programming formulations. See [1, 18, 19, 22, 23] for more material on these topics.

2.1 Definition of MDP and CMDP.

An MDP is a 4-tuple $\langle X, U, P, c\rangle$, where $X = \{0, \ldots, n-1\}$ is a finite set of *states*, $U = \{0, \ldots, k-1\}$ is a finite set of *actions*, $P : X^2 \times U \to [0, 1]$ is a *transition probability function*, and $c : X \times U \to \mathbb{R}$ 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 $c \in \mathbb{R}^{nk}$.

An MDP is a generalization of a Markov chain, where in a Markov chain there is only one possible action in each state. For simplicity, we assume that the initial state is fixed and we denote it by x_0. In fact, Assumption 1 implies that the initial state does not affect the optimal policy. In the discounted cost model, one could could assume any initial probability distribution over the states.

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

A (stationary) policy[1] 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 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.

Definition 1 *A policy π is* strictly 1-randomized *if: (i) It is deterministic in all states but one state. (ii) Let x denote the state in which π is not deterministic. Then, the set $\{u : \pi(x,u) > 0\}$ contains only two actions.*

The goal is to find a policy 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 initial state equals x_0 and the (randomized) 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

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

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

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

It can be shown that this limit exists for every stationary policy [18].

Definition of CMDP. A constrained MDP is an MDP with an additional input consisting of a cost function $d : X \times U \rightarrow \mathbb{R}$ and a parameter α. The cost $D(\pi)$ of π is defined similarly to $C(\pi)$ in both models based on $E_t^{\pi}[d(x_t,u_t)] = \sum_{x \in X, u \in U} d(x,u) \cdot P^{\pi}(x_t = x, u_t = u)$. The additional input defines the constraint $D(\pi) = \alpha$ that a feasible policy must satisfy. The optimization problem in CMDP(α) is to find a policy π that satisfies the constraint $D(\pi) = \alpha$ and minimizes $C(\pi)$.

Occupation measures. Every policy π induces a 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 $\rho(x,u) \triangleq (1-\beta) \cdot \sum_{t=0}^{\infty} \beta^t \cdot P^{\pi}(x_t = x, u_t = u)$. In the expected average cost model $\rho(x,u) \triangleq \lim_{T \rightarrow \infty} \left(\frac{\sum_{t<T} P^{\pi}(x_t = x, u_t = u)}{T} \right)$.

Given an occupation measure $\rho(x,u)$ over $X \times U$, 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 if $\sum_{u'} \rho(x,u') = 0$, then one may define $\pi^{\rho}(x,u)$ arbitrarily as long as $\sum_{u} \pi^{\rho}(x,u) = 1$.)

We refer to an occupation measure ρ as deterministic (resp., strictly 1-randomized) if $\rho = \rho_\pi$ for a deterministic (resp., strictly 1-randomized) policy π.

[1] By the general theory of MDPs and CMDPs [18, 1], under our conditions there exists an optimal stationary policy. Therefore we restrict our attention to such policies.

Irreducibility Assumption.

Definition 2 (Irreducibility) *An* MDP *is* irreducible *if every deterministic policy π induces an irreducible Markov chain.*

Throughout the paper we assume the following.

Assumption 1 *We assume that the* MDP *is irreducible.*

2.2 Linear Programming Formulation of CMDPs

In this section we formulate MDP and CMDP(α) as linear programs. We denote the linear program corresponding to MDP (resp., CMDP(α)) by LP (resp. LP(α)). The linear program LP is of the form $\min\{c^T \cdot \rho \mid A\rho = b, \rho \geq 0\}$. The linear program LP$(\alpha)$ is of the form $\min\{c^T \cdot \rho \mid A \cdot \rho = b, d^T \cdot \rho = \alpha, \rho \geq 0\}$ (the transpose of a row vector v is denoted by v^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.

Given an MDP $\langle X, U, P, c\rangle$, where $X = \{0, 1, \ldots, n-1\}$ and $U = \{0, 1, \ldots, k-1\}$, we represent the cost function c as a column vector in $\mathbb{R}^{nk}$ indexed by pairs in $X \times U$, namely, $c_{x,u} = c(x, u)$. We begin with the LP formulation in the discounted cost model.

Discounted cost model. We 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, $A = [I - \beta P(0) \ldots I - \beta P(k-1)]$. The column vector $b \in \mathbb{R}^n$ is defined by $b = (1-\beta, 0 \ldots, 0)^T$, where the zeroth coordinate corresponds to the initial state.

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 $X \times U$. For a state-action pair (x, u), the component $\rho_{x,u}$ denotes the value of the occupation measure $\rho(x, u)$.

Expected average cost model. In the expected average cost model, the matrix A is an $(n+1) \times (nk)$ matrix obtained by adding a row $\vec{1}$ consisting of ones to the concatenation of the matrices $I - P(u)$. The vector b is a unit vector, where the coordinate of the one corresponds to the row $\vec{1}$ in A. Note that the constraint $\vec{1} \cdot \rho = 1$ implies that $\rho(x, u)$ is a probability distribution.

The following theorem was proved for various cost models in [5, 6, 7, 13]. A more recent textbook proof appears in [1, Theorem 3.3].

Theorem 1 (equivalence of $CMDP(\alpha)$ and $LP(\alpha)$) CMDP(α) *is feasible if and only if* LP(α) *is feasible. Moreover, if ρ^* is an optimal solution of* LP(α)*, then π^{ρ^*} is an optimal policy of* CMDP(α).

3 The Uniqueness Property

Consider an MDP and a fixed cost function $d(\cdot)$.

Definition 3 (Uniqueness) *An* MDP *satisfies the* uniqueness *property if the following holds for every $\alpha \in \mathbb{R}$: If π^* is deterministic and optimal for* CMDP(α) *and if $\pi \neq \pi^*$ is any stationary policy, then either $D(\pi) \neq \alpha$ or $C(\pi) > C(\pi^*)$.*

Uniqueness has the following geometric interpretation. Consider the polytope generated by all the deterministic occupation measure (i.e., the feasible solutions of LP). Intersect this polytope with a hyperplane $d^T \cdot \rho = \alpha$ to obtain the feasible solution of LP(α). If this intersection has an optimal solution that is a deterministic occupation measure, then this optimal solution is unique.

The following proposition follows from the fact that every basic feasible solution of $\text{LP}(\alpha)$ is either deterministic or strictly 1-randomized (Theorem 7).

Proposition 2 *An* MDP *satisfies the uniqueness property if for every $\alpha \in \mathbb{R}$, every deterministic policy π^*, and every deterministic or strictly 1-randomized policy $\pi \neq \pi^*$, if π^* is optimal for* $\text{CMDP}(\alpha)$*, then π is not optimal for* $\text{CMDP}(\alpha)$.

Uniqueness is, in a sense, a generic property, that is, it holds for most values of the parameters. We show this by adding a small random perturbation $\varepsilon \in \mathbb{R}^{nk}$ to the cost vector c to obtain the perturbed cost vector $c_\varepsilon = c + \varepsilon$. Given any positive μ_1 and μ_2, we choose the components of the vector ε randomly and independently so that the cost differs from that of the original model by at most μ_2, and the probability that uniqueness does not hold is at most μ_1.

Let $C_\varepsilon(\pi)$ denote the cost of a policy π with respect to the perturbed cost vector c_ε. Define each coordinate ε_i of ε by $\varepsilon_i \triangleq \frac{r_i}{2^{p_1}} \cdot 2^{-p_2}$, where p_1, p_2 are positive integers and r_i is uniformly distributed over the set $\{0, \ldots, 2^{p_1} - 1\}$. The following lemma proves that a random perturbation meets the requirements while increasing the length of each component of the cost vector c by $O(n \cdot \log k + \log \frac{1}{\mu_1 \cdot \mu_2})$ bits. This is done by choosing appropriate values for p_1, p_2.

Lemma 3 *If $p_1 \geq \log_2 \frac{k^{3n}}{\mu_1}$ and $p_2 \geq \log_2(nk/\mu_2)$, then (1) the uniqueness property holds with probability at least $1 - \mu_1$, and (2) for every policy π, $|C(\pi) - C_\varepsilon(\pi)| \leq \mu_2$.*

Proof: We prove part (1) as follows. Fix a realization of the vector ε, and suppose that c_ε does not obtain uniqueness for $\text{CMDP}(\alpha)$. This implies that there exists a deterministic policy π that is optimal with respect to the perturbed cost c_ε and is not unique. Let ρ_π denote the occupation measure corresponding to π. Since ρ_π is not the only optimal solution of $\text{LP}(\alpha)$ (with respect to the perturbed cost vector c_ε), there exists a basic feasible solution (bfs) ρ that is also optimal (with respect to the same perturbed cost vector c_ε). Since both ρ_π and ρ are optimal, it follows that

$$c_\varepsilon \cdot \rho_\pi = c_\varepsilon \cdot \rho. \tag{3}$$

We conclude that the event that perturbation by ε fails implies the existence of an α and a pair $\rho_\pi \neq \rho$ of occupation measures that satisfy: (1) $d \cdot \rho_\pi = d \cdot \rho = \alpha$, (2) ρ_π is induced by a deterministic policy π, (3) ρ is a bfs of $\text{LP}(\alpha)$, and (4) $c_\varepsilon \cdot \rho_\pi = c_\varepsilon \cdot \rho$. Since ε is random, the quantities $c_\varepsilon,\ \pi,\ \rho_\pi,\ \rho$, which depend on ε, are random as well. By the proof of Theorem 7, every bfs corresponds to a deterministic or strictly 1-randomized policy.

Let R_α denote the collection of all pairs (ρ_1, ρ_2) of bfs of $\text{LP}(\alpha)$ such that ρ_1 corresponds to a deterministic policy. By Theorem 7, ρ_2 corresponds either to a deterministic or to a strictly 1-randomized policy. Note that R_α does not depend on ε and is not a random set. Let $R = \bigcup_\alpha R_\alpha$.

We claim that $|R| < k^{3n}$. There are k^n deterministic policies, thus we need to consider at most k^n values of α. For each α, there are at most $k^n + \binom{k^n}{2} < k^{2n}$ basic feasible solutions of $\text{LP}(\alpha)$. Indeed, a basic feasible solution is either deterministic or strictly 1-randomized. We now bound the number of strictly 1-randomized basic feasible solutions of $\text{LP}(\alpha)$. Every strictly 1-randomized policy is a convex combination of two deterministic policies that disagree in a single state (there are less than $\binom{k^n}{2}$ such pairs). For each such pair of deterministic policies, at most one convex combination induces a bfs of $\text{LP}(\alpha)$. This follows from Proposition 14, since if every convex combination is optimal, then none is an extreme point of $\text{LP}(\alpha)$. Therefore, the number of strictly 1-randomized basic feasible solutions is bounded by $\binom{k^n}{2}$, and $|R| < k^{3n}$ as claimed.

Consider a pair $(\rho_1, \rho_2) \in R$. Without loss of generality, ρ_1 and ρ_2 disagree in the first coordinate. Let c_ε^1 denote the first coordinate of c_ε and let c_ε^{-1} denote the vector c_ε with the first coordinate removed, so that $c_\varepsilon = (c_\varepsilon^1, c_\varepsilon^{-1})$. We use identical notation for any vector. The equation $c_\varepsilon \rho_1 = c_\varepsilon \rho_2$ implies that

$$c_\varepsilon^1 \rho_1^1 + c_\varepsilon^{-1} \cdot \rho_1^{-1} = c_\varepsilon^1 \rho_2^1 + c_\varepsilon^{-1} \cdot \rho_2^{-1}.$$

Now

$$P(c_\varepsilon \cdot \rho_1 = c_\varepsilon \cdot \rho_2) = P(c_\varepsilon^1 \cdot (\rho_1^1 - \rho_2^1) = c_\varepsilon^{-1} \cdot (\rho_2^{-1} - \rho_1^{-1}))$$
$$\leq 2^{-p_1}$$

The last line follows from the fact that, given ρ_1, ρ_2 and ε^{-1}, the event $c_\varepsilon^1 \cdot (\rho_1^1 - \rho_2^1) = c_\varepsilon^{-1} \cdot (\rho_2^{-1} - \rho_1^{-1}))$ occurs for at most one value of ε^1.

We now bound the probability that perturbation fails, namely, Eq. 3 holds. Since $(\rho_\pi, \rho) \in R$,

$$\begin{aligned} P(c_\varepsilon \cdot \rho_\pi = c_\varepsilon \cdot \rho) &\leq P\big(c_\varepsilon \cdot \rho_1 = c_\varepsilon \cdot \rho_2 \text{ for some } (\rho_1, \rho_2) \in R\big) \\ &\leq \sum_{(\rho_1, \rho_2) \in R} P(c_\varepsilon \cdot \rho_1 = c_\varepsilon \cdot \rho_2) \\ &\leq k^{3n} 2^{-p_1}. \end{aligned}$$

We conclude that if $p_1 \geq \log_2 \frac{k^{3n}}{\mu_1}$ then the probability of non uniqueness is bounded by μ_1.

Part (2) requires that the perturbation does not change the cost of the optimal policy by more than μ_2. It suffices to show that, for every occupation measure ρ, $|(c_\varepsilon - c) \cdot \rho| \leq \mu_2$. Since ρ is an occupation measure, it follows that $|(c_\varepsilon - c) \cdot \rho| \leq \sum_i \varepsilon_i \leq n \cdot k \cdot 2^{-p_2}$. Hence, part (2) holds if $p_2 \geq \log_2(nk/\mu_2)$. ■

In the light of Lemma 3 we assume the following throughout the paper.

Assumption 2 *The* MDP *satisfies the uniqueness property.*

4 The Coupling Property

Definition 4 *Two deterministic policies are* neighbors *if they disagree in a single state.*

Definition 5 *Given a deterministic policy π and an action $j \neq \pi(i)$, the neighbor policy $\pi^{i,j}$ is defined by:*

$$\forall x \in X: \quad \pi^{i,j}(x) \triangleq \begin{cases} j & \textit{if } x = i \\ \pi(x) & \textit{otherwise.} \end{cases}$$

Thus two deterministic policies π and τ are *neighbors* if there exists a state i and an action j such that $\tau = \pi^{i,j}$.

Suppose that for every state i, there is a linear order over U. We denote the linear order over U corresponding to state i by $\leq_i$. In addition, we consider the natural linear order over the set of states $X = \{0, \ldots, n-1\}$.

The polynomial algorithm in Sec. 7 for finding an optimal policy depends on a property that we call the coupling property defined below.

Definition 6 (coupling property) *The* coupling property *holds with respect to the linear orders $\{\leq_i\}_{i \in X}$ if for every deterministic policy π, every state i, and every action j,*

$$\pi(i) \leq_i j \quad \Rightarrow \quad \forall x < i \, \forall u \in U : \rho_\pi(x, u) \leq \rho_{\pi^{i,j}}(x, u).$$

4.1 Examples of MDPs with The Coupling Property

In this section we present a "one dimensional" MDP, and prove that it satisfies the coupling property in the expected average cost model. We begin with a controlled nonabsorbing random walk. We then continue with a one dimensional MDP that corresponds to a discrete time controlled M/M/1 queue.

4.1.1 A controlled nonabsorbing random walk

A controlled nonabsorbing random walk is a simple example of an MDP that satisfies the coupling property. We formally describe it below.

The MDP has n states $\{0, \dots, n-1\}$. For $i < n-1$ there is a transition from state i to state $i+1$ with probability $P(i+1|i,j) \in (0,1)$. For $i > 0$ there is a transition from state i to state $i-1$ with probability $P(i-1|i,j) = 1 - P(i+1|i,j)$. For $i = 0$ there is a self-loop $P(0|0,j) = 1 - P(1|0,j)$, and similarly, for state $n-1$ there is a self-loop $P(n-1|n-1,j) = 1 - P(n-2|n-1,0)$.

We assume that all $P(i+1|i,j)$ transition probabilities are positive. Hence the MDP is irreducible.

The linear orders $\leq_i$ are defined as follows for each state $i \geq 1$.

$$j' \leq_i j'' \quad \Leftrightarrow \quad P(i-1|i,j') \leq P(i-1|i,j'').$$

Namely, the transition from state i to its left neighbor $i-1$ is not more likely under the action j' than under the action j''. The linear order $\leq_i$ is defined arbitrarily for $i = 0$.
The proof of the following lemma appears in Appendix B.

Lemma 4 *The coupling property holds for a controlled nonabsorbing random walk.*

4.1.2 A controlled discrete-time M/M/1 queue

We now consider a discrete-time version of a controlled $M/M/1$ queue obtained from a continuous-time controlled $M/M/1$ queue by a technique called uniformization (see Appendix A). A discrete controlled M/M/1 queue is similar to the controlled nonabsorbing random walk with the addition of self-loops in each state. Formally, the set of states is $\{0, \dots, n-1\}$. For $i < n-1$ there is a transition from state i to state $i+1$ with probability $P(i+1|i,j) \in (0,1)$. For $i > 0$ there is a transition from state i to state $i-1$ with probability $P(i-1|i,j) \in (0,1)$. In addition, for every state i, there is a self-loop with probability $P(i|i,j)$. Assumption 1 holds by the reduction from the continuous M/M/1 queue.

We assume that the actions do not affect the arrival rates, hence the probabilities $P(i+1|i,j)$ do not depend on the action j. Moreover, the reduction from an M/M/1 queue implies that, for all states i, the transitions from state i to state $i+1$ have the same probability. We therefore denote $P(i+1|i,j)$ simply by q. This means that the control only affects the service rates, and hence only the probabilities $P(i-1|i,j)$ and $P(i|i,j)$ depend on the action j.

For each state $i \geq 1$, the linear order $\leq_i$ in the discrete controlled M/M/1 queue is defined as follows:

$$j' \leq_i j'' \quad \Leftrightarrow \quad P(i-1|i,j') \leq P(i-1|i,j'').$$

We prove the following lemma for the expected average cost model. The same lemma can be proved if the control affects the arrival rates and does not affect the service rate.
The proof of the following lemma appears in Appendix B.

Lemma 5 *The coupling property holds for the controlled discrete time M/M/1 queue.*

5 Structure of Optimal Policies

5.1 Deterministic Policies

Notation. Given a policy π, let I_π denote the set of pairs (i,j) for which $\pi(i,j) > 0$. These pairs define columns of the matrix A. Let B_π denote the submatrix of A consisting of the projection of A to the columns in I_π. Let ρ_π denote the occupation measure corresponding to the policy π. Let $\tilde{\rho}_\pi$ denote the vector obtained by projecting ρ_π to coordinates in I_π.

The next proposition proves that, under Assumption 1, the mapping $\pi \mapsto \rho_\pi$ between deterministic policies and the corresponding occupation measure is one-to-one.

Proposition 6 *If π is a deterministic policy for* $\text{CMDP}(\alpha)$*, then: (i) $\tilde{\rho}_\pi$ is the unique solution of the equations $B_\pi \cdot \tilde{\rho} = b$, and (ii) the rank of B_π is n.*

Proof: Part (ii) follows from part (i). We now prove part (i). By the definition of I_π, if $(x,u) \notin I_\pi$, then $\rho_\pi(x,u) = 0$. Hence $A \cdot \rho_\pi = b$ if and only if $B_\pi \cdot \tilde{\rho}_\pi = b$. In the model of discounted cost, the matrix B_π is invertible by Gersgorin's Theorem [8], hence uniqueness follows.

In the model of expected average cost, if the MDP satisfies the Assumption 1, then by the Peron-Frobenius Theorem [8], the system $B_\pi \cdot \tilde{\rho}_\pi = b$ has a unique solution, and the proposition follows. ■

5.2 Properties of Optimal Policies

The following theorem, proved for the various cost models in [5, 6, 7, 13], states that, if $\text{CMDP}(\alpha)$ is feasible, there always exists an optimal policy that is either deterministic or strictly 1-randomized. The theorem is stated in terms of the occupation measure (i.e., the optimal solution of the $\text{LP}(\alpha)$). This theorem and its proof are an extension of the theorem that every MDP has an optimal policy that is deterministic.

Theorem 7 *If $LP(\alpha)$ is feasible, then there exists an optimal solution ρ^* of $LP(\alpha)$ that is deterministic or strictly 1-randomized.*

Proof: The rank of the constraints in $\text{LP}(\alpha)$ is at most $n+1$. This implies that in every basic feasible solution (bfs) there are at most $n+1$ nonzero variables. Fix an optimal bfs ρ^*. By Assumption 1, $\sum_u \rho^*(x,u) > 0$, for each state x. Hence, for each state x, except perhaps for one, $\rho(x,u)$ is positive for exactly one action, and the theorem follows. ■

5.3 Policies Along An Edge

Notation. Let π^0 and π^1 denote two deterministic policies that disagree in a single state. Let $\pi^q \triangleq q \cdot \pi^1 + (1-q) \cdot \pi^0$, for $0 \leq q \leq 1$. Note that π^q is a strictly 1-randomized policy if $0 < q < 1$. We say that a policy π *agrees with the zeros* of policy π^* if $\pi(x,u) = 0$ whenever $\pi^*(x,u) = 0$.

Let $A^{x,u}$ denote the column of A corresponding to $x \in X$ and $u \in U$. Complementary slackness implies the following optimality condition.

Proposition 8 *Let ρ and w denote feasible solutions of* $\text{LP}(\alpha)$ *and the dual linear program DLP, respectively. The following two conditions are equivalent: (1) ρ and w are optimal. (2) For every $x \in X$ and $u \in U$, either $\rho(x,u) = 0$ or the dual constraint is tight (i.e., $w^T \cdot A^{x,u} = c(x,u)$).*

Proposition 9 ([28]) *Let π^* denote an optimal policy for* $\text{CMDP}(\alpha^*)$*. Let π denote a policy that agrees with the zeros of π^*. Then, π is an optimal policy for* $\text{CMDP}(D(\pi))$*.*

Proof: Let $\rho^* = \rho_{\pi^*}$ and $\rho = \rho_\pi$. Note that $\rho^*(x,u) = 0$ implies that $\rho(x,u) = 0$. Let w^* denote a dual optimal solution of $\text{LP}(\alpha^*)$. By Proposition 8 it follows that, for every (x,u), either $\rho^*(x,u) = 0$ or the dual constraint is tight (i.e., $(w^*)^T \cdot A^{x,u} = c(x,u)$). Note that w^* is also a feasible solution of the DLP corresponding to $\text{CMDP}(D(\pi))$. It follows that ρ and w^* also satisfy the optimality condition, and hence, by Proposition 8, ρ is optimal, as required. ■

Proposition 10 *For every two policies π' and π'', such that $D(\pi') < D(\pi'')$, there exists a policy π such that $D(\pi') < D(\pi) < D(\pi'')$.*

Proof: Denote $\pi^q \triangleq q \cdot \pi'' + (1-q) \cdot \pi'$. Define $D(q) \triangleq D(\pi^q)$. Since $D(\pi)$ is continuous in π [28], it follows that $D(q)$ is continuous in q. It follows that the image of $D(q)$ over the interval $[0,1]$ contains the interval $[D(\pi'), D(\pi'')]$. ■

Proposition 9 and the proof of Proposition 10 imply the following.

Corollary 11 ([28]) *If π^{q^*} is an optimal policy for* CMDP(α^*) *and $q^* \in (0,1)$, then, for each $\alpha \in [\inf_{0 \leq q \leq 1} D(\pi^q), \sup_{0 \leq q \leq 1} D(\pi^q)]$, there exists $q_\alpha \in [0,1]$ such that π^{q_α} is an optimal policy for* CMDP(α).

Consider the strictly 1-randomized policy $\pi^{1/2} = (\pi^0 + \pi^1)/2$. Then $I_{\pi^{1/2}}$ is the set of pairs (i,j) for which $\pi^{1/2} > 0$.

Let $B(d)$ denote the $(n+1) \times (n+1)$ square matrix obtained by first augmenting the matrix A with the row d^T, and then projecting the augmented matrix on the columns in $I_{\pi^{1/2}}$.

Lemma 12 *The following three conditions are equivalent:*

(i) $D(\pi^0) = D(\pi^1)$.

(ii) $B(d)$ *is not of full rank.*

(iii) $D(\pi^q) = D(\pi^0)$, *for all* $q \in [0,1]$.

Proof: (i) => (ii). Fix $\alpha = D(\pi^0)$. The occupation measures ρ_{π^0} and ρ_{π^1} (induced by the deterministic policies π^0 and π^1, respectively) are distinct feasible solutions of LP(α). Hence, both $\tilde{\rho}_{\pi^0}$ and $\tilde{\rho}_{\pi^1}$ are distinct solutions of the system of equations $B(d) \cdot \tilde{\rho} = \binom{b}{\alpha}$. This implies that $B(d)$ is not of full rank.

(ii) => (iii). Both policies π^0 and π^1 induce occupation measures that are feasible solutions of LP. By convexity, for every $q \in [0,1]$, the occupation measure ρ_{π^q} is also a feasible solution of LP. Since B has rank n, if $B(d)$ is not of full rank, the last row (corresponding to the constraint $d^T \cdot \rho = \alpha$) depends on the other rows. Hence, every occupation measure ρ that is a feasible solution of LP and whose support is contained in $I_{\pi^{1/2}}$ has the same cost $d^T \cdot \rho$. This implies that $D(\pi^q) = D(\pi^0)$, for all $q \in [0,1]$, as required. Finally, the implication (iii) => (i) is trivial, and the lemma follows. ■

Proposition 13 *If $D(\pi^0) \neq D(\pi^1)$, then $C(\pi^q)$ is linear in $D(\pi^q)$ over the range $q \in [0,1]$.*

Proof: We consider two cases:

1. Suppose $B(d)$ is of full rank. In the model of discounted cost, $B(d)$ is an $(n+1) \times (n+1)$ square matrix, and thus invertible. Hence, $\tilde{\rho}_{\pi^q} = B(d)^{-1} \cdot (b, D(\pi^q))^T$. Therefore, $C(\pi^q) = \tilde{c} \cdot \tilde{\rho}_{\pi^q} = \tilde{c} \cdot B(d)^{-1} \cdot (b, D(\pi^q))^T$, and $C(\pi^q)$ is linear in $D(\pi^q)$, as required. In the model of expected average cost, one needs to remove first a dependent row from $B(d)$ to make it square and thus invertible.

2. If $B(d)$ is not of full rank, then by Lemma 12, $D(\pi^0) = D(\pi^1)$, a contradiction.

■

Proposition 14 *Fix a value of α. Consider the set of policies $E_{(0,1)} \triangleq \{\pi^q : 0 < q < 1\}$. Exactly one of the following cases holds:*

1. *Every policy in $E_{(0,1)}$ is an optimal policy of* CMDP(α).

2. *No policy in $E_{(0,1)}$ is an optimal policy of* CMDP(α).

3. *Exactly one policy in $E_{(0,1)}$ is an optimal policy of* CMDP(α).

Proof:

If $B(d)$ is of full rank then by Proposition 13, either exactly one policy in $E_{(0,1)}$ is an optimal policy of $\text{CMDP}(\alpha)$ or no policy in $E_{(0,1)}$ is an optimal policy of $\text{CMDP}(\alpha)$.

If $B(d)$ is not of full rank, then by Lemma 12, $D(\pi^q) = D(\pi^0)$, for $q \in [0, 1]$ and thus, either every policy in $E_{(0,1)}$ is a feasible policy of $\text{CMDP}(\alpha)$ or no policy in $E_{(0,1)}$ is a feasible policy of $\text{CMDP}(\alpha)$. By Proposition 9, if one policy in $E_{(0,1)}$ is an optimal policy of $\text{CMDP}(\alpha)$, then every policy in $E_{(0,1)}$ is optimal as well. ■

In the following lemmas, we abbreviate, and refer to a policy π as optimal if it is an optimal policy of $\text{CMDP}(D(\pi))$.

Lemma 15 *Let $q^* \in (0, 1)$. If π^{q^*} is an optimal strictly 1-randomized policy, then the function $D(q) \triangleq D(\pi^q)$ is strictly monotone in the interval $q \in [0, 1]$.*

Proof: The function $D(q)$ is continuous because the policy π^q is continuous in q, and $D(\pi)$ is continuous in π. If $D(q)$ is not strictly monotone, then there exist $q' < q''$ such that $D(q') = D(q'')$. By Proposition 9 each of the policies $\pi^{q'}$ and $\pi^{q''}$ is optimal for $\text{CMDP}(\alpha)$, where $\alpha = D(q')$. By the uniqueness assumption (Assumption 2), neither $\pi^{q'}$ or $\pi^{q''}$ is deterministic. Hence $0 < q' < q'' < 1$.

Let ρ' (resp. ρ'') denote the occupation measure that corresponds to the policy $\pi^{q'}$ (resp. $\pi^{q''}$). We first prove that $\rho' \neq \rho''$. Assume that π^0 and π^1 disagree in state s, and, without loss of generality, assume that $\pi^0(s) = 0$ and $\pi^1(s) = 1$. By Assumption 1, both occupation measures ρ' and ρ'' assign positive probability to state s. However, the ratios $\rho'(s, 0)/\rho'(s, 1) \neq \rho''(s, 0)/\rho''(s, 1)$.

On the other hand, since the support of ρ' and ρ'' are equal, it follows that the bases corresponding to ρ' and ρ'' are the same. Hence, ρ' and ρ'' are different solutions of the system $\tilde{B} \cdot \rho = \begin{pmatrix} b \\ \alpha \end{pmatrix}$, where $\tilde{B}$ is the basis matrix. We consider two cases. If $\tilde{B}$ is invertible, then we have immediately a contradiction. If $\tilde{B}$ is not invertible, then by Lemma 12, $D(\pi^0) = D(\pi^1) = \alpha$. Therefore, both π^0 and π^1 are feasible policies of $\text{CMDP}(\alpha)$. On the other hand, both π^0 and π^1 are optimal, hence $C(\pi^0) = C(\pi^1)$, a contradiction to the uniqueness assumption (Assumption 2). ■

Lemma 16 *Let $\pi' \neq \pi''$ denote two distinct optimal policies of $\text{CMDP}(\alpha')$ and $\text{CMDP}(\alpha'')$, respectively. If π' and π'' are either deterministic or strictly 1-randomized, then $\alpha' \neq \alpha''$.*

Proof: Assume for the sake of contradiction that $D(\pi') = D(\pi'')$. Recall that by definition $\alpha' = D(\pi')$ and $\alpha'' = D(\pi'')$. Since both π' and π'' are optimal, it follows that $C(\pi') = C(\pi'')$. If either π' or π'' is deterministic, then the lemma follows from the uniqueness assumption. If both policies are strictly 1-randomized, then let π' (resp. π'') be a convex combination of two deterministic policies π^0 and π^1 (resp. τ^0 and τ^1). By Lemma 15, D increases along the edge between π^0 and π^1 (resp. τ^0 and τ^1). Without loss of generality, $D(\tau^0) \leq D(\pi^0) \leq D(\pi')$. It follows that Assumption 2 is violated for $\alpha = D(\pi^0)$. ■

5.4 Graph Representation

Definition 7 (policy graph) *The* policy graph *is a graph $G = (V, E)$, where V is the set of deterministic policies, and E is the set of pairs of neighboring deterministic policies (i.e. policies that disagree in exactly one state).*

In the case of two actions $k = 2$, the policy graph is isomorphic to the n dimensional hypercube. In the general case, the policy graph is isomorphic to the Cartesian product of n copies of the complete graph over k vertices.

We consider the edge (π^0, π^1) between neighboring deterministic policies as a representation of all convex combinations $\pi^q = (1-q) \cdot \pi^0 + q \cdot \pi^1$ of π^0 and π^1. In such a case we say that π^q *belongs* to the edge (π^0, π^1).

Let Γ denote the set of deterministic or strictly 1-randomized feasible policies for CMDP(α), for all values of α. Let $\Gamma^* \subseteq \Gamma$ denote the subset of optimal policies in Γ. In light of Proposition 9, Γ^* consists of vertices (i.e. deterministic policies) and edges (i.e. deterministic and strictly 1-randomized policies).

Lemma 17 *The set Γ^* is a path in the policy graph G.*

Proof: Let G^* denote the subgraph of G that consists of the vertices and edges in Γ^*. The proof consists of the following stages: (1) Prove that G^* is connected. (2) Prove that the degree of every vertex in G^* is at most two.

Denote the connected components of G^* by $U_1, U_2, \ldots, U_s$. By continuity, the image of the function $D()$ over each connected component is an interval. Denote the image of U_i by I_i. By Lemma 16, the intervals $I_1, \ldots, I_s$ are pairwise disjoint. By Proposition 10 the union of the intervals $I_1 \cup \cdots \cup I_s$ is an interval. To avoid a contradiction, we conclude that G^* contains only a single connected component. Hence G^* is connected, as required.

If the degree of a vertex π is at least 3, consider three edges in Γ^* that are incident to π. By Lemma 15, $D(\pi)$ is strictly monotone as one travels along each of these edges incident to v. Moreover, for at least two edges, the slope of $D(\pi)$ as one approaches v has the same sign, namely, monotone increasing (or decreasing). Two such edges in Γ^* contain two optimal policies $\pi' \neq \pi'' \in \Gamma^*$ such that $D(\pi') = D(\pi'')$. This contradicts Lemma 16, and the lemma follows. ■

The next corollary follows from Lemma 16 and Lemma 17.

Corollary 18 *$D(\pi)$ is strictly monotone along the path Γ^*.*

6 A General Algorithm

In this section we present a general algorithm for computing optimal policies of irreducible MDPs that satisfy the uniqueness property. Although we can not prove that the running time of this algorithm is polynomial in general, in the next section we prove strong polynomiality of a variant when the coupling property holds.

6.1 Geometric Interpretation of The Algorithm

The algorithm is based on Lemma 17 that states that the set Γ^* of optimal deterministic and strictly 1-randomized policies form a path in the policy graph. Consider the polytope P generated by the deterministic occupation measures. We introduce a cost vector d. Let P_α denote the intersection of P with the hyperplane $d^T \cdot \rho = \alpha$. Let $\alpha_{\min}$ (resp., $\alpha_{\max}$) denote the minimum (resp., maximum) value of α for which P_α is not empty. For each $\alpha \in [\alpha_{\min}, \alpha_{\max}]$, the polytope P_α contains a single occupation measure $\rho(\alpha)$ that corresponds to a policy $\pi(\alpha) \in \Gamma^*$.

The algorithm assigns a zero-one cost vector d so that $\alpha_{\min} = 0$ and $\alpha_{\max} = 1$. Moreover, it is trivial to find the optimal deterministic policy π such that $D(\pi) = 0$. Given a prefix of Γ^* ending in a deterministic policy π, the algorithm finds the next deterministic policy τ along Γ^* as follows. First, note that τ must be a neighbor of π. Namely, there exists a state i and an action j such that $\tau = \pi^{i,j}$. This limits the number of candidates for τ to nk. Second, by Coro. 18, $D(\tau) > D(\pi)$. Thus if we depict the neighboring policies on a (C, D)-plane (see Fig. 1), then τ is simply the policy with the smallest slope.

The algorithm ends when all neighbors τ of π satisfy $D(\tau) \leq D(\pi)$. Thus, the algorithm has reached the last policy along Γ^*.

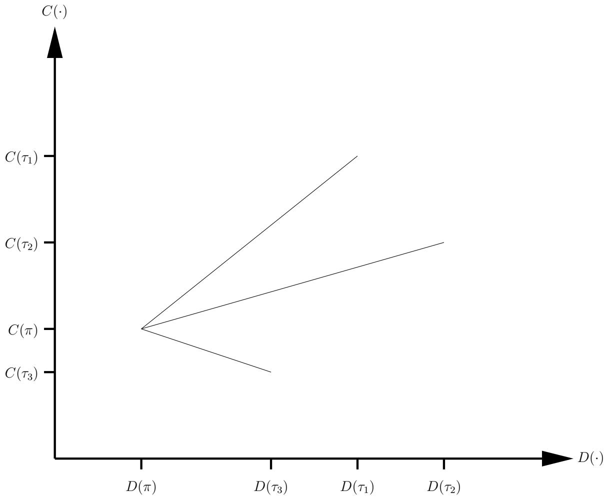

Figure 1: Suppose a prefix of Γ^* ends in a deterministic policy π. The algorithm has to compute the next policy along Γ^* among the neighboring policies τ_1, τ_2, τ_3. The costs $C(\cdot)$ and $D(\cdot)$ of each policy are depicted in the graph. The algorithm chooses the policy τ_3 since the segment between $(D(\pi), C(\pi))$ and $(D(\tau_3), C(\tau_3))$ has the smallest slope.

6.2 Notation

Given a deterministic policy π, we define the gradient $\nabla_{i,j}$ as follows:

$$\nabla_{i,j} \triangleq \frac{C(\pi^{i,j}) - C(\pi)}{D(\pi^{i,j}) - D(\pi)} \tag{4}$$

The parameters in the definition of $\nabla_{i,j}$ can be easily computed as follows. Recall that B_π denotes the projection of the columns of the matrix A on the pairs in I_π (i.e., the basis matrix corresponding to the basic feasible solution ρ_π). For a vector ρ_π, the projection to the coordinates in I_π is denoted by $\tilde{\rho}_\pi$. Since π is a deterministic policy, by Proposition 6 the corresponding occupation measure ρ_π when projected to I_π is the unique solution for $B_\pi \cdot \tilde{\rho}_\pi = b$. Hence $C(\pi) = \tilde{c}_\pi \cdot \tilde{\rho}_\pi$, $D(\pi) = \tilde{d}_\pi \cdot \tilde{\rho}_\pi$, and the analogous computations hold for $C(\pi^{i,j})$ and $D(\pi^{i,j})$.

6.3 Algorithm Description

The algorithm adds a new artificial cost function $D(\pi)$ specified by a cost vector $d \in \{0,1\}^{nk}$. The MDP with the constraint $D(\pi) = \alpha$ is denoted by CMDP(α). In the linear programming formulation, LP(α) is the linear program obtained by adding the constraint $d^T \cdot \rho = \alpha$ to LP. The algorithm computes the set Γ^* of optimal (deterministic or strictly 1-randomized) policies for CMDP(α), for every value of α. This set is found by computing Γ^*. Finally, an optimal policy for the MDP is chosen as a deterministic policy in Γ^* with minimum cost $C(\cdot)$.

A listing of the algorithm appears as Algorithm 1. In line 1, the algorithm assigns zero-one costs $d(i,j)$. For each state, one (arbitrary) action is assigned zero cost, and the other actions are assigned unit cost. In line 2, the initial policy is set. This policy simply chooses the zero cost action for each state. This initial policy π achieves the minimum value for $D(\pi)$. The path p begins with the initial policy as its starting point.

The algorithm builds the path p by adding a new edge in each iteration of the while-loop. The last policy (vertex) added to p is denoted by π. In each iteration of the while-loop the path p is augmented by a new edge $(\pi, \pi^{i,j})$. In line 4, this new edge $(\pi, \pi^{i,j})$ is chosen such that $(i,j) = \operatorname{argmin}\{\nabla_{i,j} \mid \forall (i,j) \text{ such that } D(\pi^{i,j}) > D(\pi)\}$. In line 5, the new edge is added to the path p. In line 6, the new endpoint of p is updated. In lines 7-8, the minimum cost policy along p is updated, if necessary. In line 11, a minimum cost policy is returned.

Algorithm 1 A heuristic for finding an optimal policy for the MDP $\min\{c \cdot \rho \mid A \cdot \rho = b\}$. We assume that the MDP is irreducible and satisfies the uniqueness property.

```
 1: Define
        $$d(i,j) \triangleq \begin{cases} 0 & \text{if } j = 0 \\ 1 & \text{otherwise.} \end{cases}$$
 2: Initialize:
        $\pi \leftarrow (0, \dots, 0)$        {$\pi$ chooses the "zero" action in each state}
        $\mathit{opt} \leftarrow \pi$        {best policy so far}
        $p \leftarrow \{\pi\}$        {path $p$ starts with $\pi$}
 3: while $exists\ (i,j)\ such\ that\ D(\pi^{i,j}) > D(\pi)$ do
 4:     $(i,j) \leftarrow \operatorname{argmin}\{\nabla_{i,j} \mid \forall (i,j)$ such that $D(\pi^{i,j}) > D(\pi)\}$
 5:     add the edge $(\pi, \pi^{i,j})$ to $p$
 6:     $\pi \leftarrow \pi^{i,j}$        {$\pi^{i,j}$ becomes the current endpoint of $p$}
 7:     if $C(\pi^{i,j}) < C(\mathit{opt})$ then
 8:         $\mathit{opt} \leftarrow \pi^{i,j}$        {$\mathit{opt}$ is the best policy so far}
 9:     end if
10: end while
11: return $\mathit{opt}$
```

6.4 Correctness

We now prove that Algorithm 1 finds an optimal policy. To prove this we prove that the algorithm computes Γ^*, the path of optimal solutions of $LP(\alpha)$ (for all values of α) in the policy graph.

Theorem 19 *The path p computed by the algorithm 1 equals Γ^*.*

Proof: We prove by induction on the number of iterations of the while-loop that p is a prefix of Γ^* in each iteration. Since the costs $d(x,u)$ are in $\{0,1\}$, it follows that for every policy τ, $D(\tau) \geq 0$. Hence, $LP(\alpha)$ is feasible only if $\alpha \geq 0$. Clearly the initial policy $\pi_0 = (0, \dots, 0)$ satisfies $D(\pi_0) = 0$. We claim that the initial policy is the only policy with $D(\pi) = 0$. Consider an optimal policy $\pi \neq \pi_0$. Consider a state x and action u for which $\pi(x,u) > 0$ while $\pi_0(x,u) = 0$. By the Assumption 1, $\rho_\pi(x,u) > 0$. Since $d(x,u) = 1$, it follows that $D(\pi) > 0$. We conclude that the initial policy is optimal for $\alpha = 0$. Moreover, the initial policy is the endpoint of the path Γ^* with smallest cost $D(\cdot)$, and the induction basis holds.

The induction step is proved as follows. Let π denote the last policy added to p. Let $\pi^{i,j}$ denote the next policy added to p, namely, $(i,j) \leftarrow \operatorname{argmin}\{\nabla_{i,j} \mid \forall (i,j)$ such that $D(\pi^{i,j}) > D(\pi)\}$. Let $\pi^{\hat{\imath},\hat{\jmath}}$ denote the next policy along Γ^* after π. We wish to prove that $(i,j) = (\hat{\imath},\hat{\jmath})$.

Assume for the sake of contradiction that $(i,j) \neq (\hat{\imath},\hat{\jmath})$. By Coro. 18, $D(\pi^{\hat{\imath},\hat{\jmath}}) > D(\pi)$. Let $D' = \min\{D(\pi^{\hat{\imath},\hat{\jmath}}), D(\pi^{i,j})\}$. Since the cost $D(\tau)$ is a continuous function of the policy τ, the cost D' is obtained in two policies: π_1 along the edge between π and $\pi^{i,j}$ and π_2 along the edge between π and $\pi^{\hat{\imath},\hat{\jmath}}$. For example, $\pi_1 = \pi^{i,j}$ and π_2 is a convex combination of π and $\pi^{\hat{\imath},\hat{\jmath}}$. The policy π_2 is also an optimal policy. However, by Proposition 13 and the definition of (i,j), it follows that $C(\pi_1) \leq C(\pi_2)$.

This contradicts the optimality of π_2 (if $C(\pi_1) < C(\pi_2)$) or the uniqueness of the solution (if $C(\pi_1) = C(\pi_2)$). Hence, $(i,j) = (\hat{\imath}, \hat{\jmath})$, which completes the induction step.

We now prove that when the algorithm terminates, p cannot be augmented anymore, and hence, p equals Γ^*, as required. Indeed, if p is a proper prefix of Γ^*, then the cost $D(\cdot)$ increases from π to the next deterministic policy in Γ^*. In this case, the algorithm would not have terminated yet because the set $\{(i,j) : D(\pi^{i,j}) > D(\pi)\}$ is not empty. ■

Corollary 20 *Algorithm 1 computes an optimal policy of the* MDP.

Proof: The MDP has an optimal policy π^* that is deterministic. This policy is also in Γ^*. By Theorem 19, π^* appears in the sequence of policies scanned by the algorithm. ■

Let $|\Gamma^*|$ denote the number of deterministic policies in Γ^*.

Proposition 21 *The complexity of Algorithm 1 is $O(|\Gamma^*| \cdot n^3 \cdot k)$.*

Proof: In each deterministic policy (vertex along Γ^*), the algorithm checks at most $n \cdot k$ options for the next policy. The running time of each check is dominated by matrix inversion. Matrix inversion is applied to a basis matrix that is obtained from a adjacent basis, namely, a change in a single column. By the Sherman-Morrison Formula [16], the inverse matrix can be computed in time $O(n^2)$. Thus, the complexity of the algorithm is $O(|\Gamma^*| \cdot n^3 \cdot k)$, as required. ■

7 A Strongly Polynomial Algorithm: when coupling property holds

7.1 Notation

For every deterministic policy τ, let $\rho_{\min}(\tau) \triangleq \min\{\rho_\tau(i,\tau(i)) : i \in X\}$. Similarly, $\rho_{\max}(\tau) \triangleq \max\{\rho_\tau(i,\tau(i)) : i \in X\}$. Let $\rho_{\min} \triangleq \min_\tau \rho_{\min}(\tau)$ and $\rho_{\max} \triangleq \max_\tau \rho_{\max}(\tau)$, where the minimum and maximum are taken only over deterministic policies. Assumption 1 implies that $\rho_{\min} > 0$.

The algorithm uses a parameter R that satisfies

$$R \geq \frac{1+\rho_{\max}}{\rho_{\min}}. \tag{5}$$

There is no need to precisely compute the right hand side of Eq. 5; instead we use an upper bound based on Assumption 1 as follows. Obviously $\rho_{\max} < 1$. In the expected average cost model, we lower bound $\rho_{\min}$ by $\tilde{\rho}_{\min} = p_{\min}^n$, where $p_{\min}$ is the minimum nonzero transition probability in the MDP. This lower bound holds simply by considering all paths of length n with nonzero transition probabilities to a given state. In the discounted cost model, we lower bound $\rho_{\min}$ by $\tilde{\rho}_{\min} = (1-\beta) \cdot \beta^{n-1} \cdot p_{\min}^{n-1}$. The algorithm uses the following value for R:

$$R \triangleq \max\left\{\frac{2}{\tilde{\rho}_{\min}}, nk\right\}. \tag{6}$$

7.2 Algorithm Description

A listing of the algorithm appears as Algorithm 2. The algorithm works under the additional assumption that the coupling property holds. The algorithm is a variation of Algorithm 1. The only difference is in the definition of the new artificial cost constraint $d^T \cdot \rho = \alpha$. This definition now relies on the linear orders $\leq_i$ and the parameter R. The costs $d(i,j)$ are exponential functions of i and j.

In line 1, the algorithm sorts the actions in each state, namely, it computes the linear orders $\leq_i$ over U for each state $i \in X$. In line 2, costs $d(i,j)$ are assigned to each pair $(i,j) \in X \times U$. In line 3, the initial policy is set. This policy simply chooses the first action (according to the order $\leq_i$) for each state i. This initial policy π achieves the minimum value for $D(\pi)$. The remaining lines are identical to corresponding lines in Algorithm 1.

Algorithm 2 A strongly polynomial algorithm for finding an optimal policy for the MDP $\min\{c \cdot \rho \mid A \cdot \rho = b\}$. We assume that the MDP is irreducible and satisfies both the uniqueness and coupling properties.

```
Sort the actions for each state $i \in X$ according to the order $\leq_i$. Let $j_0^i \leq_i j_1^i \leq_i \cdots \leq_i j_{k-1}^i$ denote
  the actions sorted according to the order $\leq_i$.
Define $d(i, j_\ell^i) \triangleq R^{k\cdot(n-i)+\ell}$.
Initialize:

        $\pi \leftarrow (j_0^0, \ldots, j_0^{n-1})$        {$\pi$ chooses the "first" action in each state}
        $opt \leftarrow \pi$                                {best policy so far}

while $exists\ (i,j)\ such\ that\ D(\pi^{i,j}) > D(\pi)$ do
    $(i,j) \leftarrow \mathrm{argmin}\{\nabla_{i,j} \mid \forall (i,j)$ such that $D(\pi^{i,j}) > D(\pi)\}$
    $\pi \leftarrow \pi^{i,j}$
    if $C(\pi^{i,j}) < C(opt)$ then
        $opt \leftarrow \pi^{i,j}$                          {$opt$ is the best policy so far}
    end if
end while
return $opt$
```

7.3 Correctness

The following lemma is used both for proving the correctness and running time of Algorithm 2. Consider two neighboring deterministic policies π and τ. By Lemma 15, it follows that the cost $D(\cdot)$ is strictly monotone along the edge in the policy graph from π to τ. The following lemma determines whether $D(\cdot)$ increases or decreases along the edge (π, τ). The lemma relies on both Assumption 1 and the coupling assumption.

For two actions j_1 and j_2 we say that $j_1 <_i j_2$ if $j_1 \leq_i j_2$ and $j_1 \neq j_2$.

Lemma 22 *Let $\pi \neq \tau$ denote two neighboring deterministic policies, i.e., $\tau = \pi^{i,j}$. Then, $\pi(i) <_i \tau(i)$ if and only if $D(\pi) < D(\tau)$.*

Proof: We first prove that $\pi(i) <_i \tau(i)$ implies that $D(\pi) < D(\tau)$. The proof is based on a coupling argument and the exponential growth of $d(x,u)$ as x is closer to the initial state 0. By definition,

$$D(\tau) - D(\pi) = \sum_{x \in X, u \in U} d(x,u) \cdot (\rho_\tau(x,u) - \rho_\pi(x,u)). \tag{7}$$

We partition this difference into three parts:

$$\delta_1 \triangleq \sum_{x<i} \sum_{u \in U} d(x,u) \cdot (\rho_\tau(x,u) - \rho_\pi(x,u)).$$
$$\delta_2 \triangleq \sum_{x>i} \sum_{u \in U} d(x,u) \cdot (\rho_\tau(x,u) - \rho_\pi(x,u)).$$
$$\delta_3 \triangleq \sum_{u \in U} d(i,u) \cdot (\rho_\tau(i,u) - \rho_\pi(i,u)).$$

Since $\pi(i) <_i \tau(i)$, the coupling property states that $\rho_\pi(x,u) \leq \rho_\tau(x,u)$, for every $x < i$ and every $u \in U$. Therefore, $\delta_1 \geq 0$.

We now bound δ_2 as follows. Note that, for every $x > i$ and $u \in U$, it follows that $d(x,u) \leq R^{k\cdot(n-i)-1}$. Hence,

$$\begin{aligned}\delta_2 &\geq \sum_{x>i}\sum_{u\in U} d(x,u)\cdot(0-\rho_\pi(x,u))\\ &\geq -\rho_{\max}\cdot(n-i)\cdot k\cdot R^{k\cdot(n-i)-1}\\ &> -R^{k\cdot(n-i)},\end{aligned}$$

where the last line follows from $\rho_{\max} < 1$ and $R \geq kn > k\cdot(n-i)$.

We now bound δ_3 as follows. Denote the index of $\pi(i)$ and $\tau(i)$ in the order $\leq_i$ as $\ell(\pi)$ and $\ell(\tau)$, respectively. The assumption $\pi(i) <_i \tau(i)$ implies that $0 \leq \ell(\pi) < \ell(\tau) < k$. Since π and τ are deterministic it follows that

$$\begin{aligned}\delta_3 &= d(i,j)\cdot\rho_\tau(i,j) - d(i,\pi(i))\cdot\rho_\pi(i,\pi(i))\\ &\geq R^{k\cdot(n-i)}\cdot(\rho_{\min}\cdot R^{\ell(\tau)} - \rho_{\max}\cdot R^{\ell(\pi)})\\ &\geq R^{k\cdot(n-i)},\end{aligned}$$

where the last line follows from $R \geq (1+\rho_{\max})/\rho_{\min}$ and $\ell(\tau) \geq \ell(\pi)+1$. Note that the second line requires that $\rho_\tau(i,j) > 0$. Indeed, Assumption 1 implies that $\rho_\tau(i,j) > 0$.

It follows that $\delta_1+\delta_2+\delta_3 > 0 - R^{k\cdot(n-i)} + R^{k\cdot(n-i)} = 0$, as required.

The converse direction is proved as follows. By contraposition, $D(\pi) < D(\tau)$ implies that $\pi(i) \leq_i \tau(i)$. We rule out equality (namely, $\pi(i) = \tau(i)$) since $\pi \neq \tau$ and $\tau = \pi^{i,j}$. ■

Corollary 23 *The initial policy $\pi_0 = (j_0^1,\ldots,j_0^n)$ in Algorithm 2 is an optimal policy of* CMDP(α_0) *for $\alpha_0 \triangleq D(\pi_0)$. Moreover, $LP(\alpha)$ is not feasible for $\alpha < \alpha_0$.*

Proof: Consider the policy τ of minimum cost $D(\cdot)$ in Γ^*. By Lemma 15, τ is a deterministic policy. Suppose, for the sake of contradiction, that $\tau \neq \pi_0$. Let i denote a state for which $\pi_0(i) \neq \tau(i)$. By the definition of π_0 it follows that $\pi_0(i) <_i \tau(i)$. Let $\pi = \tau^{i,\pi_0(i)}$. Note that π and τ satisfy the premises of Lemma 22. Hence $D(\pi) < D(\tau)$, contradicting the minimality of $D(\tau)$. It follows that π_0 is the unique policy in Γ^* whose cost is $D(\pi_0)$, as required. ■

Theorem 24 *Algorithm 2 returns an optimal deterministic policy of* MDP *if the coupling property holds.*

Proof: The proof follows the proof of Theorem 19 and Coro. 20 . The only modification, based on Coro. 23, is the justification that the initial policy is an endpoint of Γ^* . ■

7.4 Running Time

Proposition 25 *If the coupling property holds, then $|\Gamma^*| \leq n\cdot k$.*

Proof: By Coro. 18, the cost $D(\cdot)$ increases along Γ^*. Let π immediately precede τ along Γ^*. Suppose $\tau = \pi^{i,j}$. By Lemma 22, $\pi(i) <_i \tau(i)$. This implies that the length of the path is bounded by $n\cdot k$. ■

Corollary 26 *The complexity of Algorithm 2 is $O(k^2\cdot n^4)$.*

Proof: Follows directly from Propositions 21 and 25. ■

8 Discussion

We presented an algorithm for computing an optimal policy of an irreducible MDP. A variation of this algorithm runs in strongly polynomial time if the MDP satisfies the coupling property, e.g., a controlled discrete time M/M/1 queue. The algorithm is based on two assumptions: irreducibility and uniqueness.

The uniqueness property is shown in Lemma 3 to hold with high probability if the cost function is randomly perturbed. Therefore, if the MDP does not satisfy the uniqueness property, then the need for a random perturbation implies that the algorithm is a randomized ε-approximation algorithm.

The irreducibility assumption of the MDP (i.e., Assumption 1) is used several times in the proofs to show that the occupation measure is positive for every state. In the discounted cost model (i.e., $\beta < 1$), irreducibility can be replaced with the more relaxed assumption that, for every deterministic policy, every state is reachable from the initial state with positive probability.

Acknowledgments

We thank Boaz Patt-Shamir for many helpful conversations.
Adam Shwartz holds The Julius M. and Bernice Naiman Chair in Engineering. His research was supported in part by the fund for promotion of research, and by the fund for promotion of sponsored research at the Technion.

References

[1] E. Altman, *Constrained Markov decision processes*, Chapman & Hall/CRC, 1999.

[2] ______, *Applications of Markov decision processes in communication networks*, Handbook of Markov Decision Processes: Methods and Applications (E. Feinberg and A. Shwartz, eds.), 2002, pp. 489–536.

[3] F.J. Beutler and K.W. Ross, *Uniformization for semi-Markov decision processes under stationary policies*, Journal of Applied Probability (1987), 644–656.

[4] V.D. Blondel and J.N. Tsitsiklis, *A survey of computational complexity results in systems and control*, Automatica **36** (2000), 1249–1274.

[5] G. de Ghellinck, *Les problemes de decisions sequentielles*, Cahiers Centre Etudes Rech. Operationnelle **2** (1960), 161–179.

[6] F. d'Epenoux, *A probabilistic production and inventory problem*, Management Science (1963), 98–108.

[7] C. Derman, *On sequential decisions and Markov chains*, Management Science (1962), 16–24.

[8] R.A. Horn and C.R. Johnson, *Matrix analysis*, Cambridge university press, 2005.

[9] L. Kallenberg, *Finite state and action MDPs*, Handbook of Markov Decision Processes: Methods and Applications (E. Feinberg and A. Shwartz, eds.), 2002, pp. 21–87.

[10] M.Y. Kitaev and V.V. Rykov, *Controlled queueing systems*, CRC press, 1995.

[11] L. Kleinrock, *Queueing systems, volume I: theory*, John Wiley & Sons New York, 1975.

[12] M.L. Littman, T.L. Dean, and L.P. Kaelbling, *On the complexity of solving Markov decision problems*, Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, 1995, pp. 394–402.

[13] A.S. Manne, *Linear programming and sequential decisions*, Management Science (1960), 259–267.

[14] Y. Mansour and S. Singh, *On the complexity of policy iteration*, Uncertainty in Artificial Intelligence, vol. 99, 1999.

[15] N. Megiddo, *Method for solving stochastic control problems of linear systems in high dimension*, October 3 2006, US Patent 7,117,130.

[16] C.D. Meyer, *Matrix analysis and applied linear algebra*, Society for Industrial Mathematics, 2000.

[17] C.H. Papadimitriou and J.N. Tsitsiklis, *The complexity of Markov decision processes*, Mathematics of operations research (1987), 441–450.

[18] M.L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*, John Wiley & Sons, Inc. New York, NY, USA, 1994.

[19] S.M. Ross, *Introduction to probability models*, Academic press, 2000.

[20] A. Schrijver, *Theory of linear and integer programming*, Wiley, 2000.

[21] R.F. Serfozo, *An equivalence between continuous and discrete time Markov decision processes*, Operations Research (1979), 616–620.

[22] H.C. Tijms, *Stochastic modelling and analysis: a computational approach*, John Wiley & Sons, Inc. New York, NY, USA, 1986.

[23] ______, *Stochastic models: an algorithmic approach*, John Wiley & Sons, Inc. New York, NY, USA, 1994.

[24] P. Tseng, *Solving H-horizon, stationary Markov decision problems in time proportional to log (H)*, Operations Research Letters **9** (1990), no. 5, 287–297.

[25] M. Yadin and P. Naor, *On queueing systems with variable service capacities*, Nav. Res. Logist. Q **14** (1967), 43–53.

[26] Y. Ye, *A New Complexity Result on Solving the Markov Decision Problem*, Mathematics of Operations Research **30** (2005), no. 3, 733–749.

[27] A. Zadorojniy, *Constrained Markov Decision Processes with Application to Wireless Communications*, Master's thesis, M. Sc, Electrical Engineering, Technion, 2004.

[28] A. Zadorojniy and A. Shwartz, *Robustness of policies in constrained Markov decision processes*, IEEE Transactions on Automatic Control **51** (2006), no. 4, 635–638.

A Uniformization of A Birth-Death Process

Uniformization [3, 21, 22] is a technique that transforms a continuous time Markov process into a discrete one while preserving the occupation measure. In the case of a controlled Markov process, uniformization also preserves the optimality of policies and their costs. We apply uniformization to transform a controlled continuous time birth-death process [11] (i.e., one-dimensional queue) to a discrete MDP.

The queue dynamics are determined by service completions, which are independent exponentially-distributed random variables with rates $\mu(i,u)$, that depend on the state i and the action u, and an independent Poisson arrival process with rate λ (that does not depend on the state or the action). Equivalently, the inter-arrival times are exponentially distributed with rate λ.

We now use the following fact. Given two independent exponential random variables X, Y with rates a and b resp. and a number $c > a + b$, we can construct the two variables as follows. Choose a Poisson process with rate c. When the process jumps, throw an (independent) 3-sided die. With probability a/c claim that X occurred. With probability b/c claim that Y occurred, and with probability $(c-a-b)/c$ neither occurred—in which case we repeat. This gives the same probabilistic behavior as the original variables.

Now denote by $v_i(u)$ the rate by which the queue leaves state i if the selected action is u. To apply the previous argument, set $v_i(u) = \lambda + \mu(i,u)$ as the total rate in state i. Let $v = \max_{i,u}(v_i(u))$. The transition rates of the uniformized process are now independent of the state and of the actions chosen in each state, and are equal to v. Once a transition occurred, its type is determined according to

$$P(j|i,u) = \begin{cases} \frac{\lambda}{v} & \text{for } j = i+1 \\ \frac{\mu(i,u)}{v} & \text{for } j = i-1 \\ 1 - \frac{\lambda+\mu(i,u)}{v} & \text{for } j = i \end{cases} \tag{8}$$

with the appropriate modifications at $i = 0$ and $i = n-1$. We have obtained a process that shares the probabilistic description of the original, except that self-loops were introduced. In the new process, times between events (including self-loops) are independent, identically distributed (exponential with rate v).

To obtain a discrete time process we observe the process at jump times (including times of self loops). Since inter-jump times are i.i.d., we can simply count the number of jumps, where the transition probabilities are determined by Equation 8, and obtain a discrete-time process. The properties of the inter-jump times imply that, under any deterministic policy, the occupation measure of the original process agrees with that of the discrete time process, and in particular all cost functionals agree as well.

B Proof That Coupling Property Holds

Proof of Lemma 4. We prove the lemma for the expected average cost model. We use the following abbreviated notation. Fix a deterministic policy π. For each state x, let $\rho(x)$ denote the occupation measure for state x under the policy π, namely, $\rho(x) \triangleq \sum_{u\in U} \rho_\pi(x,u)$. Let $\rho'(x) \triangleq \sum_{u\in U} \rho_{\pi^{i,j}}(x,u)$. Let $p(x) \triangleq P(x-1|x,\pi(x))$ and $q(x) \triangleq P(x+1|x,\pi(x))$. Similarly, let $p'(x) \triangleq P(x-1|x,\pi^{i,j}(x))$ and $q'(x) \triangleq P(x+1|x,\pi^{i,j}(x))$. For state 0, let $p(0) \triangleq P(0|0,\pi(0))$, and for state $n-1$ let $q(n-1) \triangleq P(n-1|n-1,\pi(n-1))$.

We claim that the following holds, for every state $x \geq 1$:

$$\rho(x) = \frac{q(x-1)\cdot q(x-2)\cdots q(0)}{p(x)\cdot p(x-1)\cdots p(1)} \cdot \rho(0). \tag{9}$$

The proof of 9 is similar to the analytic solution of the equilibrium probabilities of continuous time birth-death queuing systems [11]. We prove Eq. 9 by induction on x. The basis for $x = 1$ is equivalent to $\rho(1)\cdot p(1) = q(0)\cdot\rho(0)$. Indeed, the first constraint of LP implies that

$$\rho(0)\cdot p(0) + \rho(1)\cdot p(1) = \rho(0)\cdot p(0) + \rho(0)\cdot q(0).$$

This is in fact the balance equation [11, 23] that compares the probability of transitions entering state 0 with the probabilities of the transitions emanating from state 0.

Assume that Eq. 9 holds for $x \leq k$. The induction step for $x = k + 1$ uses the LP constraint for state k (i.e., the balance equation for state k). Namely,

$$\rho(k-1) \cdot q(k-1) + \rho(k+1) \cdot p(k+1) = \rho(k)(p(k) + q(k)).$$

Rearranging, we obtain,

$$\rho(k+1) = \frac{1}{p(k+1)} \cdot (\rho(k)(p(k) + q(k)) - \rho(k-1) \cdot q(k-1)).$$

Dividing $\rho(k)/\rho(k-1)$, and substituting according to Eq. 9 gives $\rho(k) \cdot p(k) = \rho(k-1) \cdot q(k-1)$. Therefore,

$$\begin{aligned}\rho(k+1) &= \frac{1}{p(k+1)} \cdot (\rho(k) \cdot q(k)) \\ &= \frac{q(k) \cdot q(k-1) \cdots q(0)}{p(k+1) \cdot p(k) \cdots p(1)} \cdot \rho(0),\end{aligned}$$

which completes the proof of Eq. 9.

Our goal is to prove that if $\pi(i) \leq_i j$, then $\rho_\pi(x) \leq \rho_{\pi^{i,j}}(x)$, for every $x < i$. Let $\gamma(x) \triangleq \frac{q(x-1) \cdot q(x-2) \cdots q(0)}{p(x) \cdot p(x-1) \cdots p(1)}$. Similarly, let $\gamma'(x)$ denote the above ratio with respect to the policy $\pi^{i,j}$.

We claim that for every state x, $\gamma(x) \geq \gamma'(x)$. Indeed, for $x < i$, $\gamma(x) = \gamma'(x)$ since the ratio differs only when $x \geq i$. For $x = i$ it follows that $\gamma(x)/\gamma'(x) = p'(i)/p(i) \geq 1$ since $\pi(i) \leq_i j$. For $x > i$ it follows that $\gamma(x)/\gamma'(x) = \frac{p'(i)}{p(i)} \cdot \frac{q(i)}{q'(i)} \geq 1$.

Recall first that since ρ is an occupation measure, it follows that $\sum_{x \in X} \rho(x) = 1$. Hence, by Eq. 9,

$$\begin{aligned}1 &= \sum_{x \in X} \rho(x) = \rho(0) \cdot \sum_{x \in X} \gamma(x) \\ 1 &= \sum_{x \in X} \rho(x) = \rho'(0) \cdot \sum_{x \in X} \gamma'(x).\end{aligned}$$

Since $\sum_{x \in X} \gamma(x) \geq \sum_{x \in X} \gamma'(x)$, it follows that $\rho(0) \leq \rho'(0)$. For every state $x < i$ we have $\gamma(x) = \gamma'(x)$, hence by Eq. 9 it follows that $\rho(x) \leq \rho'(x)$, as required. ■

Proof of Lemma 5. We use the same notation as in the proof of Lemma 4. We claim that the following holds, for every state $x \geq 1$:

$$\rho(x) = \frac{q^x}{p(x) \cdot p(x-1) \cdots p(1)} \cdot \rho(0). \tag{10}$$

The proof of Eq. 10 is by induction on x. The basis for $x = 1$ is equivalent to $\rho(1) \cdot p(1) = q \cdot \rho(0)$. Indeed, it holds because of the balance equation [11, 23]:

$$\rho(0) \cdot p(0) + \rho(1) \cdot p(1) = \rho(0) \cdot p(0) + \rho(0) \cdot q$$

that compares the probability of transitions entering state 0 with the probabilities of the transitions emanating from state 0. Note that this balance equation does not hold in the discounted cost model.

Assume that Eq. 10 holds for $x \leq k$, the induction step for $x = k + 1$ uses the balance equation for state k. Namely,

$$\rho(k-1) \cdot q + \rho(k) \cdot P(k|k, \pi(k)) + \rho(k+1) \cdot p(k+1) = \rho(k)(p(k) + P(k|k, \pi(k)) + q).$$

Rearranging, we obtain,

$$\rho(k+1) = \frac{1}{p(k+1)} \cdot (\rho(k)(p(k) + q) - \rho(k-1) \cdot q).$$

By dividing Eq. 10 for $\rho(k)$ and $\rho(k-1)$ it follows that $\rho(k) \cdot p(k) = \rho(k-1) \cdot q$. Therefore,

$$\begin{aligned}\rho(k+1) &= \frac{1}{p(k+1)} \cdot (\rho(k) \cdot q) \\ &= \frac{q^{(k+1)}}{p(k+1) \cdot p(k) \cdots p(1)} \cdot \rho(0),\end{aligned}$$

which completes the proof of Eq. 10.

Our goal is to prove that if $\pi(i) \leq_i j$, then $\rho_\pi(x) \leq \rho_{\pi^{i,j}}(x)$, for every $x < i$. Let $\gamma(x) \triangleq \frac{q^x}{p(x) \cdot p(x-1) \cdots p(1)}$. Similarly, let $\gamma'(x)$ denote the above ratio with respect to the policy $\pi^{i,j}$.

We claim that for every state x, $\gamma(x) \geq \gamma'(x)$. Indeed, for $x < i$, $\gamma(x) = \gamma'(x)$ since the ratio differs only when $x \geq i$. For $x \geq i$ it follows that $\gamma(x)/\gamma'(x) = p'(i)/p(i) \geq 1$ since $\pi(i) \leq_i j$.

Recall first that since ρ is an occupation measure, it follows that $\sum_{x \in X} \rho(x) = 1$. Hence, by Eq. 10,

$$\begin{aligned}1 &= \sum_{x \in X} \rho(x) = \rho(0) \cdot \sum_{x \in X} \gamma(x) \\ 1 &= \sum_{x \in X} \rho(x) = \rho'(0) \cdot \sum_{x \in X} \gamma'(x).\end{aligned}$$

Since $\sum_{x \in X} \gamma(x) \geq \sum_{x \in X} \gamma'(x)$, it follows that $\rho(0) \leq \rho'(0)$. For every state $x < i$ we have $\gamma(x) = \gamma'(x)$, hence by Eq. 10 it follows that $\rho(x) \leq \rho'(x)$, as required. ■