CONSTRAINED MARKOV DECISION PROCESSES WITH TOTAL COST CRITERIA: LAGRANGIAN APPROACH AND DUAL LINEAR PROGRAM

Eitan ALTMAN
INRIA
2004 Route des Lucioles, B.P.93
06902 Sophia-Antipolis Cedex
France

Submitted: May 1995, revised: January and September 1997

Abstract

The aim of this paper is to investigate the Lagrangian approach and a related Linear Programming (LP) that appear in constrained Markov decision processes (CMDPs) with a countable state space and total expected cost criteria (of which the expected discounted cost is a special case). We consider transient MDPs and MDPs with uniform Lyapunov functions, and obtain for these an LP which is the dual of another one that has been shown to provide the optimal values and stationary policies [3, 4]. We show that there is no duality gap between these LPs under appropriate conditions. In obtaining the Linear Program for the general transient case, we establish, in particular, a calculation approach for the value function of the CMDP based on finite state approximation. Unlike previous approaches for state approximations for CMDPs (most of which were derived for the contracting framework), we do not need here any Slater type condition. We finally present another type of LP that allows the computation of optimal mixed stationary-deterministic policies.

1 Introduction

The problem of minimizing a single objective (with no constraints) in a Markov decision process (the total expected cost, or the expected average cost) can be handled by solving a system of Dynamic Programming equations, known as the Bellman optimality equations. These transform the problem of minimization over the class of all policies into a set of coupled minimization problems over the (much smaller) sets of actions. These Dynamic Programming equations may be the starting point for obtaining an LP formulation. Under suitable conditions, the value function is the *largest* "superharmonic functions": these are functions that satisfy some optimality inequalities (obtained directly from the optimality equations) for all states and actions.

This approach is in the basis of the derivation of the LPs by Kallenberg [53] and Hordijk and Kallenberg [50].

In the case of constrained MDPs, one can still derive directly this LP by using a Lagrangian approach, and applying some minmax theorem. Indeed, the Lagrangian approach allows us to transform a constrained control problem to an equivalent minmax non-constrained control problem. If a saddle point property holds, then the problem is transformed to a maxmin problem, which can be solved using an LP. This derivation of the dual LP for CMDPs was obtained by Altman and Spieksma [8] for the finite case. In this paper we follow this path to obtain an LP for CMDPs with a countable state space with the total expected cost criteria (of which the expected discounted cost is a special case).

A possible way to obtain a saddle point property for the Lagrangian, is to establish some convexity and compactness properties of the set of performance measures achievable by different classes of policies. This is closely related to the convexity and compactness properties of the occupation measures, analyzed by Derman [26], Kallenberg [53], Hordijk and Kallenberg [50] and Feinberg [34] for the finite state and action spaces; and by Krylov [56] (who studied controlled diffusion processes), Borkar [17, 18], Altman and Shwartz [2, 5], Spieksma [73] and Feinberg and Sonin [37, 38], for the infinite state space. We follow a somewhat different approach in this paper, by considering mixed policies (these correspond to an additional initial randomization between policies). By doing so, the costs become affine in the initial randomization, and the set of achievable costs is seen to be convex.

We consider two types of MDPs: the transient MDPs, for which the total expected time spent in each state is finite under any policy, and MDPs with a uniform Lyapunov function. The latter type of MDP has been shown to be equivalent to absorbing MDPs (defined in [48, 53] under some continuity conditions, and is more general than contracting MDPs. All three types of MDPs: contacting, absorbing and transient, are equivalent for the finite state space, as was shown in Kallenberg [53]; this is however not the case in the countable state space.

In the case of the transient MDPs, we assume that the immediate costs are nonnegative. For the different MDPs, we show that the Linear Program has the same value as the constrained MDP. However, it is only for the MDP with uniform Lyapunov function that the Linear Program is shown to be solvable, i.e. the supremum is attained by some solution.

In obtaining the Linear Program for the general transient case, we establish, in particular, a calculation approach for the value function of CMDP based on finite state approximation. Unlike previous approaches for state approximations for CMDPs (most of which were derived for the contracting framework, see [1, 2] and [4] Ch. 8), we do not need here any Slater type condition.

The Linear Program that we obtain in this paper is dual to the one obtained in [3, 4] for solving constrained MDPs with total expected cost criterion. We call the LP in [3, 4] the primal one, and the one we described above - the dual. The value of the primal LP was shown in [3, 4], under appropriate conditions, to be equal to the one of CMDP. We thus conclude that under the appropriate conditions, there is no duality gap between the primal and dual LP.

The primal LP enables us not only to obtain the value of the CMDP, but also to construct an optimal stationary policy.

The Lagrangian approach turns out to be not only a tool for obtaining an LP formulation, but has its own merits. It turns out to be very useful for sensitivity analysis and for obtaining asymptotical properties of constrained MDPs. Indeed, the Lagrangian approach that we present here allowed us to obtain in [4] theorems for approximations of the value and policies for CMDPs, which we applied for the study of the convergence in the discount factor (especially, in the neighborhood of 1), in the horizon as well as the study of state truncation techniques [4]. In particular, it allowed us to obtain an estimate of the approximation error.

Instead of focusing on stationary policies, one may consider another class of policies which is appealing for CMDPs: the mixed stationary-deterministic policies. These are policies obtained by an additional initial randomization over some set of policies (in our case - the stationary-deterministic ones). According to the outcome of the randomization, one stationary-deterministic policy is chosen, and the controller follows it forever. We shall show in this paper that this class of mixed stationary-deterministic policies is indeed dominating, i.e. it dominates the performance measures obtained by the class of all policies. This enables us to obtain an alternative Linear Programming formulation for solving CMDPs, different than the primal and dual LPs discussed above. The new LP is related to one which was introduced in [6], where instead of a probabilistic mixing between policies, a time-sharing mechanism was proposed (see also Ross [64]). A similar LP formulation was obtained by Feinberg [34] for the case of finite state and action spaces. In general, this approach requires a huge number of decision variables, even for the case of a finite state space. However, there are special applications in which one can eliminate a-priori many solutions, where this approach can be extremely efficient, and used even for problems with infinite state space (see [6]). The mixed stationary-deterministic policies have another nice property: the LP that we obtain by restricting to them was shown (in the case of finite state and action spaces) to have the same form for both the expected discounted and the expected average cost (see [34, 74]) for the general multi-chain case. This property enabled us to establish the convergence of the values and policies of expected discounted cost CMDPs to those of expected average cost, as the discount factor tends to one, see [74].

The structure of the paper is as follows. In the next section we present the model and notations. We obtain in Section 3 the dual Linear Program for the non-constrained MDP. Then we present in Section 4 the set of achievable vector performance measures. We introduce the Lagrange approach in Section 5, and use these results in Section 7 to establish the convergence of the values of MDPs with truncated state spaces to the value of our constrained MDP. In Section 6 we obtain the dual (infinite) LP for the (original) constrained MDP. A second LP approach is presented in Section 8. Finally, we present in the Appendix (Section 9) some extensions to and corrections of results in [3] on the continuity of occupation measures, and obtain similar results as in [3] for MDPs with uniform Lyapunov functions. We end this introduction by a short survey on the Lagrangian approach and on Mathematical Programming techniques in MDPs.

The Lagrange techniques in MDPs were first introduced by Beutler and Ross [15, 16] for the case of a single constraint. It allowed them to characterize the structure of optimal policies for the constrained problem, but it did not provide computational tools. This approach was extended by Sennott [69, 70] to the countable state space. The use of Lagrangian techniques for several constraints is quite recent (see e.g. Arapostathis et al. [11] Piunovskiy [60], and Altman and Spieksma [8]).

The use of LPs techniques for solving MDPs started already in the beginning of the sixties, with the pioneering work of D'Epenoux [24, 25], who considered the expected discounted cost case, and of De Ghellinck [21] and Manne [59] who studied the expected average cost (with the unichain condition). The analysis via LPs, of the expected average cost with the general multi-chain ergodic structure, has been presented by Denardo and Fox [23] and Denardo [22]. Hordijk and Kallenberg [49] presented a single LP for solving the multi-chain expected average problem. For an extensive description and survey of LP techniques for the non-constrained MDPs, see Kushner and Kleinman [57], Heilmann [43, 44], Arapostathis et al. [11], Puterman [61] and Kallenberg [53]. An important contribution to generalization of the LP techniques to infinite state and action spaces is due to Lasserre [58] who applied functional analytical tools, using the theory of infinite dimension LPs (due to Anderson and Nash [10]). Lasserre handles both the primal and dual LPs, establishes conditions for their solvability and for the absence of a duality gap, and presents conditions for the optimality of a stationary policy that is obtained using the solution to the primal LP. This work was extended in Hernández-Lerma and Lasserre [46] and Hernández-Lerma and Hernández-Hernández [45] to the case of non-countably infinite state and action spaces, and in Hordijk and Lasserre [51] - to the multi-chain expected average case. An alternative approach to derive the LP was obtained by Altman and Shwartz [5], Altman [2, 3] and Spieksma [73], using probabilistic techniques, and these were obtained directly for the constrained MDPs.

Ross and Varadarajan [65, 66] have considered problems where a constraint is imposed on the actual sample-path cost. They solved it using an LP approach. An interesting feature of this formulation is that ϵ-optimal stationary policies exist (in the finite state and action case) even under the general multi-chain ergodic structure. This is in contrast with the problem where all costs are defined through expectations. Moreover, the computation of the value and the ϵ-optimal policy is much simpler than for the problem with expected costs.

There are other alternative ways to make the costs more sensitive to deviations from the expectation. One way to achieve this goal is to have some *additional cost related to the variance*. Sobel [71] proposed to maximize the mean-variance ratio with constraints on the mean. Other approaches were proposed and analyzed in Filar and Lee [41], Kawai [55], Bayal-Gursoy and Ross [13] and Filar et al. [40]. A unified approach which extends the above ones was presented by Huang and Kallenberg [52] and solved using an algorithm based on parametric-Linear Programming. The case of infinite state space was analyzed by Altman and Shwartz [5]. Other recent papers in this topic are Sobel [72] and White [76].

Another way to penalize deviations of the costs from the expectation is to introduce some constraints on the *rate of convergence*. This approach was investigated by Altman and Zeitouni [9] using a mathematical Programming formulation. Bounds were obtained by using an LP.

The LP approach, in particular, and Mathematical Programming approaches, in general have been used also in the case of more than one controller (i.e. stochastic games), see e.g. the survey by Raghavan and Filar [62]. Recently, Altman and Shwartz [7] obtained characterization and existence results on stationary Nash equilibria in stochastic games with constraints, based on sensitivity analysis of LPs.

2 The model

$\mathbf{X}$ is a countable state space and A is a metric set of actions. We denote by $\mathsf{A}(x) \subset \mathsf{A}$ the compact set of actions available at state x; denote $\mathcal{K} = \{(x, a) : x \in \mathbf{X}, a \in \mathsf{A}(x)\}$, equipped with its Borel σ-field $\mathbb{K}$. $\mathcal{P}_{xay}$ is the probability of moving from state x to y if action a is chosen. $c : \mathcal{K} \to \mathbb{R}$ and $d : \mathcal{K} \to \mathbb{R}^K$ are immediate costs. (K, which will correspond to the number of constraints, is some given integer.) With some abuse of notation, we denote $c(x, \gamma) = \int c(x, a)\gamma(da)$ for any probability measure γ over $\mathsf{A}(x)$, with a similar definition for $d(x, \gamma)$.

We make throughout the assumptions that, for every state x,

$$c(x, \cdot) \text{ and } d^k(x, \cdot), k = 1, ..., K, \text{ are continuous on } \mathsf{A}(x), \tag{1}$$

$$\begin{array}{l}\text{the transition probabilities are continuous on } \mathsf{A}(x), \text{ i.e., if } a(n) \to a, \\ \quad \lim_{n\to\infty} \mathcal{P}_{xa(n)y} = \mathcal{P}_{xay}, \forall y \in \mathbf{X}.\end{array} \tag{2}$$

(Note that a slightly stronger continuity condition was assumed in [3]. However, the current version suffices for the results in [3] to hold, as follows from Theorem 9.1 in the Appendix). Define a history at time t to be a sequence of previous states and actions, as well as the current state: $h_t = (x_1, a_1, ..., x_{t-1}, a_{t-1}, x_t)$. Let $\mathbf{H}_t$ be the set of all possible histories of length t equipped with its Borel σ-algebra. A policy u is a sequence $u = (u_1, u_2, ...)$ where $u_t : \mathbf{H}_t \to M_1(\mathsf{A})$ is a measurable function that assigns to any history of length t, a probability measure over the set of actions, where $M_1(G)$ stands for the set of probability measures over a set G endowed with the topology of weak convergence of measures. If the history h_t is observed at time t, then the controller chooses an action within $\mathcal{A}$ with probability $u_t(\mathcal{A}|h_t)$, where $\mathcal{A}$ is any subset in $\mathsf{A}(x_t)$. Let U denote the set of all policies. We denote by U_M, U_S and U_D the Markov, stationary, and stationary-deterministic policies respectively. For $w \in U_S$, the state process becomes a stationary Markov chain whose transition probabilities are denoted by $P_{xy}(w)$.

It will often be useful to extend the definition of a policy $u = (u_1, u_2, ...)$ so as to allow u_t to depend not only on h_t, but also on some initial randomizing mechanism. More precisely, for any class of policies $G \subset U$, equipped with some σ-algebra $\mathcal{G}$ that include singletons (sets that contain a single policy), we define $\overline{M}(G)$ to be the class of mixed policies generated by G; we call these *mixed-G policies.* A mixed-G policy is identified with a distribution q over G; the controller first uses q to choose some policy $u \in G$, and then proceeds with that policy from time 1 onwards. A policy as above that uses a distribution q is denoted by $\hat{q}$. Define $\mathcal{U} := \overline{M}(U_D)$.

A related way to obtain mixed policies is given in Dynkin and Yushkevich [31]; they consider mixing between the strategic measures generated by policies in U, rather than the mixing between the policies themselves.

In our definition of mixed policies we assume some measurable structure, i.e. that together with G there is given some σ-algebra $\mathcal{G}$ of sets in G, that include singletons (sets that contain a single policy), so that a probability on G is well defined. To that end, we let for each $x \in \mathbf{X}$, $\mathcal{B}(\mathsf{A}(x))$ denote the set of Borel subsets of $\mathsf{A}(x)$. $M_1(\mathsf{A}(x))$ is the space of probability measures over $\mathcal{B}(\mathsf{A}(x))$ endowed with the topology of weak convergence, and it is a linear Hausdorff compact metric space. (Since $\mathsf{A}(x)$ is compact metric, it is also separable, and hence the set of probability measures over $\mathcal{B}(\mathsf{A}(x))$ is tight. By Prohorov's Theorem, this implies the compactness of $M_1(\mathsf{A}(x))$). Both U_S and U_M can be represented as the set of functions that have some countable set I as range, and a countable product of compact sets (of measures) $\prod_{i\in I} M_1(\mathsf{A}_I)$ as image. Tychonov's Theorem implies that $\prod_{i\in I} M_1(\mathsf{A}_i)$ is also convex, compact in the topology of weak convergence, are metrizable, and they have as extreme points the sets U_D, and the set of pure Markov policies, respectively. We thus have a metric topology for U_D, U_S and U_M. We now define the Borel sets $\mathcal{B}_M$ of U_M, and the σ-algebra $\mathcal{G}_M$ generated by them. The set of mixed Markov strategies $\overline{M}(U_M)$ is identified with the compact set of probability measures (compact in the topology of weak convergence) on the space $(U_M, \mathcal{G}_M)$. We define similarly $\overline{M}(U_S)$ and $\overline{M}(U_D) = \mathcal{U}$. (It follows that U_D is indeed a measurable set, so that $\mathcal{U}$ is well defined. The measurability can be established using the argument in Borkar (1994) p. 178.)

As usual [11, 18, 31, 47], any given distribution β for the initial state (at time 1) and a policy u, determine a unique probability measure P_β^u over the space of trajectories of the states and actions. This defines the stochastic processes X_t and A_t of the states and actions. We denote by E_β^u the corresponding expectation operator. P_β^u is then a measurable function on the space of policies, see [33].

When β is concentrated on some state x we shall use the notation P_x^u (and E_x^u) instead of P_β^u (and E_β^u, respectively). β is then the Dirac measure $\beta(\mathcal{X}) = \delta_x(\mathcal{X}) = 1\{x \in \mathcal{X}\}$, for any $\mathcal{X} \subset \mathbf{X}$.

For a given set $\mathcal{M}$, define the hitting time of the set $\mathcal{M}$ as

$$T_{\mathcal{M}} \stackrel{\text{def}}{=} \min\{n > 1 : X_n \in \mathcal{M}\}.$$

We shall often omit $\mathcal{M}$ from the notation.

We denote $p_\beta^u(t;\mathcal{X}) := P_\beta^u(X_t \in \mathcal{X}, T > t)$ and, with some abuse of notation, $p_\beta^u(t;\overline{\mathcal{K}}) := P_\beta^u((X_t, A_t) \in \overline{\mathcal{K}}, T > t)$ for $\mathcal{X} \subset \mathbf{X}, \overline{\mathcal{K}} \subset \mathcal{K}$.

Define the occupation measure $f_{tc}(\beta, u)$ on $(\mathcal{K}, \mathbb{K})$ related to the total cost criterion by

$$f_{tc}(\beta, u; \overline{\mathcal{K}}) = \sum_{t=1}^{\infty} p_\beta^u(t; \overline{\mathcal{K}}), \qquad \overline{\mathcal{K}} \subset \mathcal{K}.$$

Define

$$\mathbf{L}_{\overline{U}}(\beta) = \bigcup_{u\in\overline{U}} \{f_{tc}(\beta,u)\} \text{ for any } \overline{U} \subset U \cup \overline{M}(U_M). \tag{3}$$

We set $\mathbf{L}(\beta) = \mathbf{L}_U(\beta)$. A class of policies $\overline{U}$ is called complete for the total cost criterion (for a given initial distribution β) if $\mathbf{L}_{\overline{U}}(\beta) = \mathbf{L}(\beta)$.

With some abuse of notation, we use f_{tc} also for the restriction over $\mathbf{X}$, i.e. $f_{tc}(\beta, u; \mathcal{X}) := f_{tc}(\beta, u; \cup_{x\in\mathcal{X}}(x, \mathsf{A}(x)))$ for any $\mathcal{X} \subset \mathbf{X}$.

For any policy u and initial distribution β, the total expected cost is defined as

$$C_{tc}(\beta,u) = \overline{\lim_{s\to\infty}} \sum_{t=1}^{s} E^u_\beta c(X_t, A_t) 1\{T > t\}. \tag{4}$$

The cost functions $D^k_{tc}(\cdot,\cdot)$, related to the immediate costs d^k are defined similarly. $C_{tc}(u) : \mathbf{X} \to \mathbb{R}$ will denote the function (or vector) whose x entry is $C_{tc}(x, u)$.

For a fixed vector $V = (V_1, ..., V_K)$ of real numbers, we define the constrained control problem **COP** as:

Find a policy that minimizes $C_{tc}(\beta, u)$ subject to $D_{tc}(\beta, u) \leq V$.

Here, and throughout, we use the notation $q_1 \leq q_2$ between two vectors $q_1, q_2 \in \mathbb{R}^K$ to mean componentwize ordering, i.e. $q_1(j) \leq q_2(j),\ j = 1, ..., K$. Similarly, for any two measures q_1, q_2 defined on the same measurable space $(\Omega, \mathcal{F})$, $q_1 \leq q_2$ means $q_1(B) \leq q_2(B)$ for any $B \in \mathcal{F}$. We denote by $\langle q_1, q_2\rangle$ the scalar product between two vectors.

The set of policies satisfying the constraints is called feasible. If this set is nonempty then the constrained problem is said to be feasible (we shall use a similar terminology for Linear Programs). Let $C_{tc}(\beta)$ be the value of the above problem (if the feasible set of policies is empty then we set $C_{tc}(\beta) = \infty$). If a feasible policy u^* achieves the minimum, i.e. $C_{tc}(\beta) = C_{tc}(\beta, u^*)$, then it is called optimal.

We present some definitions [3]:

Definition 2.1 *(Transient and absorbing policies)*
Fix an initial distribution β. Consider a partition of $\mathbf{X}$ to two disjoint sets $\mathbf{X}'$ and $\mathcal{M}$. A policy u is said to be $\mathbf{X}'$-transient, if $\sum_{t=1}^\infty P^u_\beta(X_t = x, T > t) < \infty$ for any $x \in \mathbf{X}$. It is called $\mathbf{X}'$-absorbing or, absorbing to $\mathcal{M}$, if $\sum_{t=1}^\infty P^u_\beta(T > t) < \infty$.

Definition 2.2 *(Transient and absorbing MDPs)*
Consider a partition of $\mathbf{X}$ to two disjoint sets $\mathbf{X}'$ and $\mathcal{M}$. The MDP is called an $\mathbf{X}'$-transient MDP (absorbing to $\mathcal{M}$, respectively) if all policies are $\mathbf{X}'$-transient (absorbing to $\mathcal{M}$, respectively). An $\mathbf{X}$-transient MDP is called a transient MDP.

When using $\mathbf{X}'$-transient MDPs, we shall assume that the immediate costs are nonnegative (we shall relax this condition for MDPs with uniform Lyapunov functions and contracting MDPs, defined below).

Sufficient conditions for determining that an MDP is $\mathbf{X}'$-transient or absorbing to $\mathcal{M}$ will be presented in Corollary 9.1.

We define next the uniform Lyapunov approach, due to [48] (see also [4, 11, 20]).

Definition 2.3 *(Uniform Lyapunov function)*
Consider a partition of $\mathbf{X}$ to the disjoint sets $\mathbf{X}'$ and $\mathcal{M}$. Define

$$\nu(x,a) = \max(|c(x,a)|, |d^1(x,a)|, ..., |d^K(x,a)|).$$

A function $\mu : \mathbf{X} \to [1, \infty)$ is said to be a uniform Lyapunov function if (i)-(iii) below hold:
M1(i) For all $(x,a) \in \mathcal{K}$,

$$\nu(x,a) + 1 + \sum_{y \in \mathbf{X}'} \mathcal{P}_{xay}\mu(y) \leq \mu(x). \tag{5}$$

M1(ii) For each $x \in \mathbf{X}$ the mapping $a \to \sum_{y \in \mathbf{X}'} \mathcal{P}_{xay}\mu(y)$ is continuous in $\mathsf{A}(x)$ (i.e. if $a_n \to a$ then $\sum_{y \in \mathbf{X}'} \mathcal{P}_{x,a_n,y}\mu(y)$ converges to $\sum_{y \in \mathbf{X}'} \mathcal{P}_{xay}\mu(y)$).
M1(iii) For each $x \in \mathbf{X}$ and $u \in U_D$,

$$\lim_{n \to \infty} \sum_{y \in \mathbf{X}} [{}_{\mathcal{M}}P(u)]^n_{xy}\mu(y) = 0, \tag{6}$$

where ${}_{\mathcal{M}}P(u)$ is the Taboo probability matrix obtained by replacing columns $j \in \mathcal{M}$ by columns with zero entries.

If such a function μ exists then the MDP is said to have a uniform Lyapunov function.

Several sufficient and necessary conditions for an MDP to have a uniform Lyapunov function are given in [20, 4]. It is further shown there that an MDP with a uniform Lyapunov function is absorbing, and that an absorbing MDP has a uniform Lyapunov function if the expected life time is continuous over the stationary policies.

For any functions $q : \mathbf{X} \to \mathbb{R}$, $Q : \mathbf{X} \times \mathbf{X} \to \mathbb{R}$ and $\mu : \mathbf{X} \to [1, \infty)$, define their μ-norms

$$\|q\|_\mu = \sup_{x \in \mathbf{X}} \frac{q(x)}{\mu(x)}, \qquad \|Q\|_\mu = \sup_{x \in \mathbf{X}} \frac{\sum_{y \in \mathbf{X}} Q_{xy}\mu(y)}{\mu(x)}. \tag{7}$$

We say that q and Q are μ-bounded if $\|q\|_\mu < \infty$ and $\|Q\|_\mu < \infty$, respectively. We define F^μ to be the set of functions from $\mathbf{X}$ to $\mathbb{R}$ having finite μ-norm over $\mathbf{X}$, and M^μ to be the set of nonnegative measures over $\mathbf{X}$ given by $\mathsf{M}^\mu := \{q : \sum_{x \in \mathbf{X}} q(x)\mu(x) < \infty\}$ (we shall use the notation $\langle q, \mu \rangle$ for $\sum_{x \in \mathbf{X}} q(x)\mu(x)$.)

We shall say that a function $f : \mathcal{K} \to \mathbb{R}$ is in $\overline{\mathsf{F}^\mu}$ if the function whose x entry is $\sup_{a \in \mathsf{A}(x)} |f(x,a)|$, is in F^μ. Similarly, a nonnegative measure q defined on $\mathcal{K}$ is said to be in $\overline{\mathsf{M}^\mu}$ if the measure $\overline{q}$ is in M^μ, where $\overline{q}(x) := q(x, \mathsf{A}(x))$.

Definition 2.4 *(Contracting MDPs)*
Let $\mathbf{X}'$ and $\mathcal{M}$ be two disjoint sets of states with $\mathbf{X} = \mathbf{X}' \cup \mathcal{M}$. An MDP is said to be contracting (on $\mathbf{X}'$) if there exist a scalar $\xi \in [0,1)$ and a vector $\mu : \mathbf{X} \to [1, \infty)$, such that for all $x \in \mathbf{X}, a \in \mathsf{A}(x)$,

$$\sum_{y \notin \mathcal{M}} \mathcal{P}_{xay} \mu(y) \leq \xi \mu(x). \tag{8}$$

When using contracting MDPs, we shall make the following assumptions on the initial distribution, the transition probabilities and the costs (instead of the assumptions below Definition 2.2):

- $\langle \beta, \mu \rangle < \infty$;
- The transition probabilities are μ-continuous, i.e. if $a(n) \to a$ then

$$\lim_{n \to \infty} \sum_{y \in \mathbf{X}'} \left| \mathcal{P}_{xa(n)y} - \mathcal{P}_{xay} \right| \mu(y) = 0; \tag{9}$$

- c and $d^k, k = 1, ..., K$ are in $\overline{\mathsf{F}^\mu}$.

Contracting MDPs are a subclass of absorbing MDPs, which are subclass of transient MDPs (see [3, 4]). Finally, we define dominating policies:

Definition 2.5 *(Dominating policies)*
A class of policies $\overline{U}$ is said to be a dominating class of policies for **COP** *for a given initial distribution β if for any policy $u \in U$ there exists a policy $\overline{u} \in \overline{U}$ such that*

$$C_{tc}(\beta, \overline{u}) \leq C_{tc}(\beta, u), \qquad \textit{and} \qquad D_{tc}(\beta, \overline{u}) \leq D_{tc}(\beta, u). \tag{10}$$

When (10) holds, we say that $\overline{u}$ dominates u.

One may restrict **COP** to U_M without loss of optimality, since this class is dominating (see [3] Theorem 2.1).

3 Non-constrained control: Dynamic and Linear Programming

We describe in this section the well known dynamic Programming formulation for solving unconstrained MDPs, and use it to obtain an equivalent LP formulation and a finite-state approximation scheme. This approach has been developed starting from Shapely [68], in the wider context of Markov games. For more detailed presentation, algorithmic procedures, and references, see e.g. [4, 11, 31, 61]. Introduce the dynamic Programming inequality:

$$\phi(x) \geq \min_{a \in \mathsf{A}(x)} \left[c(x,a) + \sum_{y \in \mathbf{X}'} \mathcal{P}_{xay} \phi(y) \right], \qquad x \in \mathbf{X}. \tag{11}$$

Theorem 3.1 *(Dynamic Programming: the transient case [4, 11, 61])*
Consider the transient framework (Definition 2.1 and nonnegative immediate costs) Then
(i) The optimal value $C_{tc}(x), x \in \mathbf{X}$, is the smallest (componentwize) nonnegative solution of (11).
(ii) For any state x, let $\mathcal{A}(x)$ be the set of actions that achieve the minimum of $[c(x,a) + \sum_{y \in \mathbf{X}'} \mathcal{P}_{xay} C_{tc}(y)]$. Any stationary policy g that chooses actions within $\mathcal{A}(x), x \in \mathbf{X}$ w.p.1 (i.e. for which $g_x(\mathcal{A}(x)) = 1$ for all $x \in \mathbf{X}$) is optimal.
(iii) The optimal value $C_{tc}(x), x \in \mathbf{X}$, achieves (11) with strict equality.

In the following example we show that, indeed, (11) may have several solutions larger than the value. This example will be used later to show that in the dual LP, one has to restrict the decision variables to some special set of linear functions.

Example 3.1 *(On the necessity of the restrictions on ϕ)*
Consider a discrete time queueing model. At each time period t, there may be an arrival of a customer with probability λ, or a departure from the queue with probability μ'. We assume that $\mu' > \lambda$. The arrivals and departures in different time periods are independent. The state space is the set of integers, and a state x has the meaning that there are x customers in the queue. There is no control here (thus, we may assume that $\mathsf{A}(x) = \{a\}$ contains a dummy control action a at all states). We wish to compute $\mathcal{T}$:= the expected time it takes the queue to empty (i.e. to reach state 0). In other words, we wish to compute the total expected cost with respect to the immediate cost $c(x) = c(x,a) = 1\{x \neq 0\}$ till the set $\mathcal{M} = \{0\}$ is reached. The solution satisfies

$$\begin{aligned} \mathcal{T}(0) &= 0 \\ \mathcal{T}(x) &= 1 + \mu' \mathcal{T}(x-1) + (1 - \mu' - \lambda)\mathcal{T}(x) + \lambda \mathcal{T}(x+1), \end{aligned} \tag{12}$$

and, in particular, it satisfies (11). (12) is a difference equation whose solution is given by

$$const \cdot \left[\left(\frac{\mu'}{\lambda} \right)^x - 1 \right] + \frac{x}{\mu' - \lambda}. \tag{13}$$

The expected time to reach 0 is $\mathcal{T}(x) = x/(\mu' - \lambda)$. But any constant other than 0 in (13) yields another solution of (11), so $\mathcal{T}(x)$ is not the unique solution, nor the smallest one. It is, however, the smallest nonnegative solution of (13) (and of (11)).

■

Next, we present the case of uniform Lyapunov function:

Theorem 3.2 *(Dynamic Programming: the case of uniform Lyapunov function [4, 48])*
Consider an MDP with a uniform Lyapunov function. Then
(i) The optimal value $C_{tc}(x), x \in \mathbf{X}$, is the unique solution of (11) in the class of μ-bounded functions; moreover, it achieves (11) with strict equality.
(ii) For any state x, let $\mathcal{A}(x)$ be the set of actions that achieve the minimum of $[c(x,a) + \sum_{y \in \mathbf{X}'} \mathcal{P}_{xay} C_{tc}(y)]$. Any stationary policy g that chooses actions within $\mathcal{A}(x), x \in \mathbf{X}$ w.p.1 (i.e. for which $g_x(\mathcal{A}(x)) = 1$ for all $x \in \mathbf{X}$) is optimal.

Remark 3.1 Consider Example 3.1. If we choose the μ norm to be $\mu(x) = \alpha^x$, for some $1 < \alpha < \mu'/\lambda$, then it can be seen that we obtain a contracting Markov chain with respect to this norm and to the set $\mathcal{M} = \{0\}$. The MDP therefore has a uniform Lyapunov function. We see that $\mathcal{T}$ is indeed the unique μ-bounded solution of (12) (and of (11)).

Definition 3.1 *(Superharmonic functions)*
Fix some $\mathbf{X}' \subset \mathbf{X}$. A function $\phi : \mathbf{X} \to \mathbb{R}$ is called superharmonic (for the total cost criterion) if it satisfies for all $x \in \mathbf{X}$ and $a \in \mathsf{A}(x)$:

$$\phi(x) \leq c(x,a) + \sum_{y \in \mathbf{X}'} \mathcal{P}_{xay} \phi(y). \tag{14}$$

Theorem 3.3 *(The value and superharmonic functions)*
(i) Consider the transient framework (Definition 2.1 and nonnegative immediate cost). Let ϕ be a superharmonic function. If for some optimal stationary policy g,

$$\overline{\lim_{t \to \infty}} E_x^g \phi(X_t) 1\{T > t\} \leq 0 \tag{15}$$

then the value satisfies $C_{tc} \geq \phi$ (componentwize!).
(ii) Consider an MDP with a uniform Lyapunov function. Then the value C_{tc} is the largest superharmonic function among the μ-bounded functions.

Proof: (i) From Theorem 3.1 it follows that that C_{tc} is a super-harmonic function. Choose a superharmonic function ϕ and let g be an optimal stationary policy satisfying (15). Then

$$\phi(x) \quad \leq \quad c(x,g) + \sum_{y \in \mathbf{X}'} \mathcal{P}_{xgy} \phi(y) = c(x,g) + E_x^g \phi(X_2) 1\{T > 2\}$$

$$
\begin{aligned}
&\leq \; c(x,g) + E_x^g\left[c(X_2,A_2)1\{T>2\} + E_{X_2}^g(\phi(X_3)1\{T>2\})\right] \\
&= \; c(x,g) + E_x^g c(X_2,A_2)1\{T>2\} + E_x^g(\phi(X_3)1\{T>3\}) \\
&\leq \; \ldots \; \leq \sum_{t=1}^n E_x^g c(X_t,A_t)1\{T>t\} + E_x^g \phi(X_{n+1})1\{T>n+1\}.
\end{aligned}
$$

(i) is established by taking $n \to \infty$. (A proof of a more general statement can be found in Lemma 3.6 in [36].)
(ii) follows as the proof of (i), by the dominated convergence theorem, since for any policy g, $\lim_{t\to\infty} E_x^g \mu(X_t)1\{T>t\} = 0$ (see e.g. [4, 20, 48]). ■

Motivated by Theorem 3.3, we introduce the following infinite Linear Program (which is dual to the one presented in [3] in the case of no constraints), with decision variables $\phi(y), y \in \mathbf{X}$.

$$
\begin{aligned}
\mathbf{DP}(\beta): \quad & \text{Find } \phi^* := \sup_\phi \langle \beta, \phi \rangle \text{ s.t.} \\
\phi(x) \;\leq\; & c(x,a) + \sum_{y\in\mathbf{X}'} \mathcal{P}_{xay}\phi(y), \qquad x \in \mathbf{X}, a \in \mathsf{A}(x).
\end{aligned}
$$

For the transient case (where the immediate cost is assumed to be nonnegative), we may further add nonnegativity constraints on ϕ without loss of optimality. Indeed, if ϕ is feasible for $\mathbf{DP}(\beta)$, then it is easily seen that ϕ' given by $\phi'(y) = \max(\phi(y), 0)$, $y \in \mathbf{X}$, is also feasible, and it clearly dominates ϕ.

We begin by considering the MDP with a uniform Lyapunov function. Theorem 3.3 implies the following:

Theorem 3.4 *(The dual Linear Program for MDPs with a uniform Lyapunov function)*
Consider $\mathbf{DP}(\beta)$ where the decision variables are restricted to the set $\phi \in \mathsf{F}^\mu$. Then for any initial distribution β (with $\langle \beta, \mu \rangle < \infty$), $\mathbf{DP}(\beta)$ is feasible; its value equals $C_{tc}(\beta)$ and $\phi(x) = C_{tc}(x), x \in \mathbf{X}$ is an optimal solution.

A similar statement could be obtained for the transient case, when restricting to functions for which the condition (15) from Theorem 3.3 (i) holds. However, the above condition may be difficult to verify. We therefore adopt an alternative approach, and identify a simple sub-class of functions satisfying that condition. Of course, if we restrict the LP to a subclass of functions, we risk to obtain only a lower bound on the optimal value. Our choice of functions ϕ will turn, however, to be rich enough, to obtain the same value as the one obtained by the larger class of policies satisfying (15). We begin by considering the absorbing case with *unbounded costs.*

Theorem 3.5 *(The dual Linear Program, absorbing case)*
Assume that the MDP is $\mathbf{X}'$-absorbing, and the immediate cost are nonnegative (not necessarily bounded). Consider $\mathbf{DP}(\beta)$ where the decision variables ϕ are all (nonnegative) bounded functions. Then for any initial distribution β, $\mathbf{DP}(\beta)$ is feasible and its value equals $C_{tc}(\beta)$.

Proof: Denote by $C^1(\beta)$ the value of $\mathbf{DP}(\beta)$ restricted to bounded ϕ. It follows that for any bounded function ϕ, eq. (15) holds for all policies: $\lim_{t\to\infty} E_x^g \phi(X_t)1\{T > t\} = 0$. Theorem 3.3 (i) implies that

$$C^1(x) \le C_{tc}(x) \tag{16}$$

for all x.

Let $\mathbf{X}_n$ be an increasing sequence of *finite* sets of states converging to $\mathbf{X}$. Consider **COP** with an immediate cost $c_n(x,a) = c(x,a)1\{x \in \mathbf{X}_n\}$; denote by $C_{tc}^n(\beta,u)$ the corresponding total expected cost, and by $C_{tc}^n(\beta)$ the corresponding optimal value. For any policy u and intial distribution β, $C_{tc}^n(\beta,u)$ is increasing in n, and hence so is $C_{tc}^n(\beta)$. Denote by $C_{tc}^*(\beta)$ the limit of $C_{tc}^n(\beta)$ as n tends to infinity. By Theorem 3.1, we have

$$C_{tc}^n(x) = \min_{a\in\mathsf{A}(x)} \left[c_n(x,a) + \sum_{y\in\mathbf{X}'} \mathcal{P}_{xay} C_{tc}^n(y) \right], \qquad x \in \mathbf{X}, \tag{17}$$

which implies that for all $x \in \mathbf{X}$ and $a \in \mathsf{A}(x)$ we have

$$C_{tc}^n(x) \le c_n(x,a) + \sum_{y\in\mathbf{X}'} \mathcal{P}_{xay} C_{tc}^n(y) \le c(x,a) + \sum_{y\in\mathbf{X}'} \mathcal{P}_{xay} C_{tc}^n(y). \tag{18}$$

Thus C_{tc}^n is a bounded super-harmonic function. Hence

$$C_{tc}^*(x) \le C^1(x) \le C_{tc}(x). \tag{19}$$

Let a_n be a minimizing action in (17) (it is, of course, a function of x), and let a^* be some limit point (as $n \to \infty$) obtained by diagonalization. By Fatou's Lemma, applied to (17), we get

$$C_{tc}^*(x) \ge c(x,a^*) + \sum_{y\in\mathbf{X}'} \mathcal{P}_{xa^*y} C_{tc}^*(y), \qquad x \in \mathbf{X}.$$

It then follows from Theorem 3.1 (i) that $C_{tc}(x) \le C_{tc}^*(x)$, and hence, by (19), we have $C_{tc}(x) = C^1(x)$. ■

Next, we introduce the transient case:

Theorem 3.6 *(The dual Linear Program, transient case)*
Assume that the MDP is $\mathbf{X}'$-transient, and the immediate cost are nonnegative. Consider $\mathbf{DP}(\beta)$ where the decision variables ϕ are all (nonnegative) functions that vanish outside of some finite set of states. In other words, for each ϕ in this class, there exists some finite $\mathbf{Y} \subset \mathbf{X}$ such that

$$\phi(x) = 0 \text{ for all } x \notin \mathbf{Y}. \tag{20}$$

Then for any initial distribution β, $\mathbf{DP}(\beta)$ is feasible and its value equals $C_{tc}(\beta)$.

Proof: Denote again by $C^1(\beta)$ the value of $\mathbf{DP}(\beta)$ restricted to ϕ that satisfy (20). Since the MDP is $\mathbf{X}'$-transient, it follows that for any function ϕ satisfying (20), eq. (15) holds for all policies: $\lim_{t\to\infty} E_x^g \phi(X_t) 1\{T > t\} = 0$. Theorem 3.3 (i) implies that

$$C^1(x) \le C_{tc}(x) \tag{21}$$

for all x.

Let there be a sequence of finite sets of states $\mathbf{X}_n$, increasing to $\mathbf{X}$. Consider a sequence $\mathbf{COP}_n$ of truncated problems where $\mathbf{COP}_n$ differs from the original $\mathbf{COP}$ by the fact that the process is restricted to $\mathbf{X}_n$. This is done by altering the transition probabilities and the costs:

$$\mathcal{P}^n_{xay} = \begin{cases} \mathcal{P}_{xay} & \text{if } x, y \in \mathbf{X}_n, \\ 1 & \text{if } x \notin \mathbf{X}_n, y = 0, \\ 0 & \text{otherwize,} \end{cases} \qquad c_n(x,a) = c(x,a) 1\{x \in \mathbf{X}_n\}. \tag{22}$$

Here 0 is some arbitrary (possibly new) state which is not in $\mathbf{X}'$. Denote by $C^n_{tc}(\beta, u)$ the corresponding total expected cost, and by $C^n_{tc}(\beta)$ the corresponding optimal value. For any policy u and intial distribution β, $C^n_{tc}(\beta, u)$ is increasing in n, and so is $C^n_{tc}(\beta)$. By Theorem 3.1, we have

$$C^n_{tc}(x) = \min_{a \in \mathsf{A}(x)} \left[c_n(x,a) + \sum_{y \in \mathbf{X}'} \mathcal{P}^n_{xay} C^n_{tc}(y) \right], \qquad x \in \mathbf{X}, \tag{23}$$

which implies that for all $x \in \mathbf{X}$ and $a \in \mathsf{A}(x)$ we have

$$C^n_{tc}(x) \le c_n(x,a) + \sum_{y \in \mathbf{X}'} \mathcal{P}^n_{xay} C^n_{tc}(y) \le c(x,a) + \sum_{y \in \mathbf{X}'} \mathcal{P}_{xay} C^n_{tc}(y). \tag{24}$$

Thus C^n_{tc} is a super-harmonic function that vanishes outside of $\mathbf{X}_n$. Denote by $C^*_{tc}(\beta)$ the (increasing) limit of $C^n_{tc}(\beta)$ as n tends to infinity. Hence,

$$C^*_{tc}(x) \le C^1(x) \le C_{tc}(x). \tag{25}$$

Let a_n be a minimizing action in (23) and let a^* be some limit point obtained by diagonalization. By Fatou's Lemma, applied to (23), we get

$$C^*_{tc}(x) \ge c(x,a^*) + \sum_{y \in \mathbf{X}'} \mathcal{P}_{xa^*y} C^*_{tc}(y), \qquad x \in \mathbf{X}.$$

It then follows from Theorem 3.1 (i) that $C_{tc}(x) \le C^*_{tc}(x)$, and hence, by (25), we have $C_{tc}(x) = C^1(x)$. ■

Remark 3.2 *(On the solvability of the dual Program)*
Unlike the case of a uniform Lyapunov function, the sup in $\mathbf{DP}(\beta)$ *need not be achieved in the setting of Theorem 3.5 or 3.6 (i.e. the dual LP need not be solvable); we cannot expect* $\phi(x) = C_{tc}(x), x \in \mathbf{X}$ *to be an optimal solution since there is no reason to expect* $C_{tc}(x)$ *to be bounded.*

Remark 3.3 *(State truncation)*
We have in fact presented in the proof of Theorem 3.6 a state truncation procedure, that enables to approximate the value of a non constrained MDP with countable state space by an MDP with a finite state space. The policy that chooses at state x the action $a^* = a^*(x)$ is optimal for the MDP, due to Theorem 3.1 (ii). Since this policy is the limit of policies optimal for the truncated MDPs, we conclude that also the policies converge. A different approach to state truncation is presented in [1, 2, 4] for the contracting framework.

Some restrictions in the dual LP, such as the one on bounded functions ϕ in the absorbing case (Theorem 3.5), or on functions ϕ converging to zero for the general transient case (Theorem 3.6) are indeed necessary. This can be seen from our Example 3.1. Any function among (13) are feasible for the dual LP. The supremum over all these unbounded functions is infinity, and not $\mathcal{T}$.

4 Set of achievable costs

Define for any $\overline{U} \subset U \cup \overline{M}(U_M)$ the set of achievable vector performance measures:

$$\mathbf{M}^{tc}_{\overline{U}}(\beta) = \cup_{u \in \overline{U}} \{(C_{tc}(\beta, u), D^k_{tc}(\beta, u), k = 1, ..., K)\}, \tag{26}$$

and set $\mathbf{M}^{tc}(\beta) := \mathbf{M}^{tc}_{U}(\beta) \cup \mathbf{M}^{tc}_{\overline{M}(U_M)}(\beta)$. Define

$$\mathbf{Q}_{tc}(\beta) = \left\{ \begin{array}{c} \rho \in M(\mathcal{K}) : \sum_{y \in \mathbf{X}} \int_{\mathsf{A}(y)} \rho(y, da)(\delta_x(y) - \mathcal{P}_{yax} 1\{x \in \mathbf{X}'\}) = \beta(x), x \in \mathbf{X} \\ \\ \rho(x, \mathsf{A}(x)) < \infty \text{ for all } x \in \mathbf{X}, \end{array} \right\}$$

where $M(\mathcal{K})$ is the set of nonnegative measures over $\mathcal{K}$.

$$\mathbf{Q}^{\mathrm{b}}_{tc}(\beta) \stackrel{\mathrm{def}}{=} \text{ the subset of finite measures among } \mathbf{Q}_{tc}(\beta),$$

$$\mathbf{Q}^{\nu}_{tc}(\beta) \stackrel{\mathrm{def}}{=} \{q \in \mathbf{Q}^{\mathrm{b}}_{tc}(\beta) : \langle q, \nu \rangle < \infty\}, \qquad \mathbf{Q}^{\mu}_{tc}(\beta) \stackrel{\mathrm{def}}{=} \mathbf{Q}_{tc}(\beta) \cap \overline{\mathsf{M}^{\mu}}. \tag{27}$$

(27) are defined for MDPs with a uniform Lyapunov function and for contracting MDPs (the parameters ν and μ are introduced in Definitions 2.3 and 2.4, respectively). For any sets B, B_1, B_2 of measures (equipped with a given topology) on some measurable space, define

- coB := the closed convex hull of B;
- $\min B$:= the set of minimal elements in B, i.e. $\rho \in \min B$ if there does not exist some $\rho' \in B$, $\rho' \leq \rho$ (i.e. $\rho'(\mathcal{A}') \leq \rho(\mathcal{A}')$ for any $\mathcal{A}'$), such that $\rho'(\mathcal{A}) < \rho(\mathcal{A})$ for some $\mathcal{A}$;
- $B_1 \prec B_2$ if $\forall \rho_2 \in B_2$ there exists $\rho_1 \in B_1$ such that $\rho_1 \leq \rho_2$.

Define also $\mathbf{V}_{tc}(\beta), \mathbf{V}_{tc}^{\nu}(\beta), \mathbf{V}_{tc}^{\mu}(\beta)$ and $\mathbf{V}_{tc}^{\mathrm{b}}(\beta)$ by

$$\bigcup_{\rho} \{(\langle \rho, c\rangle, \langle \rho, d^1\rangle, \langle \rho, d^2\rangle, ..., \langle \rho, d^K\rangle)\}, \tag{28}$$

where the union is taken over ρ in $\mathbf{Q}_{tc}(\beta), \mathbf{Q}_{tc}^{\nu}(\beta), \mathbf{Q}_{tc}^{\mu}(\beta)$ and $\mathbf{Q}_{tc}^{\mathrm{b}}(\beta)$, respectively.

The next characterization of achievable costs follows from the linear representation of the cost with respect to the occupation measures (see Lemma 9.3 (i) in the Appendix, and [3] Section 6), together with the characterization of the set of occupation measures achievable by different set of policies (see Theorem 9.2 (i) in the appendix and [3] Theorem 5.2).

Theorem 4.1 *(Characterization of the sets of achievable costs)*
(i) For transient MDPs with nonnegative immediate costs, $\mathbf{M}^{tc}(\beta)$ *is convex, and*

$$\min \mathbf{V}_{tc}(\beta) = \mathbf{M}_{U_S}^{tc}(\beta) \prec \mathbf{M}_{U_M}^{tc}(\beta) = \mathbf{M}^{tc}(\beta) \subset \mathbf{V}_{tc}(\beta).$$

(ii) For MDPs with a uniform Lyapunov function, $\mathbf{M}_{U_S}^{tc}(\beta)$ *is convex and compact, and satisfies*

$$\mathbf{M}_{\mathcal{U}}^{tc}(\beta) = \mathbf{M}^{tc}(\beta) = \mathbf{M}_{U_S}^{tc}(\beta) = co\mathbf{M}_{U_D}^{tc}(\beta) = \mathbf{V}_{tc}^{\nu}(\beta) = \mathbf{V}_{tc}^{\mathrm{b}}(\beta) = \min \mathbf{V}_{tc}(\beta).$$

For contracting MDPs, the above holds, in particular, with $\mathbf{V}_{tc}^{\mu}(\beta)$ *replacing* $\mathbf{V}_{tc}^{\nu}(\beta)$.

5 Constrained control: Lagrangian approach

We now go back to our constrained control problem. We use a standard Lagrangian approach for convex Programming to show that
(i) **COP** is equivalent to solving a non-constrained sup-inf problem;
(ii) the sup and inf can be interchanged under suitable conditions; the inf in the inf-sup problem is in fact achieved by some policy which is optimal for **COP**.
(iii) Under the Slater conditions, the sup is also obtained as max, that is, "optimal" policies and Lagrange multipliers exist for the sup-inf problem, and they satisfy the Kuhn-Tucker conditions.

The main result is presented in the following theorem. Its derivation is independent of the theory we developed in [3] for the achievable sets of occupation measure.

Theorem 5.1 *(The Lagrangian)*
Consider either
– the transient framework (Definition 2.1 and nonnegative immediate cost), or
– MDPs with a uniform Lyapunov function
(i) The value function satisfies

$$C_{tc}(\beta) = \inf_{u\in U} \sup_{\lambda\geq 0} J_{tc}^{\lambda}(\beta, u) = \inf_{u\in U_M} \sup_{\lambda\geq 0} J_{tc}^{\lambda}(\beta, u) = \inf_{u\in \overline{M}(U_M)} \sup_{\lambda\geq 0} J_{tc}^{\lambda}(\beta, u), \tag{29}$$

where

$$J_{tc}^{\lambda}(\beta,u) := C_{tc}(\beta,u) + \langle\lambda, D_{tc}(\beta,u) - V\rangle = \sum_{t=1}^{\infty} E_{\beta}^{u} j^{\lambda}(X_t,A_t) - \langle\lambda, V\rangle \tag{30}$$
$$j^{\lambda}(x,a) := c(x,a) + \langle\lambda, d(x,a)\rangle.$$

(ii) A policy u^* is optimal for **COP** *if and only if $C_{tc}(\beta) = \sup_{\lambda\geq 0} J_{tc}^{\lambda}(\beta,u^*)$.*
(iii) The value satisfies

$$C_{tc}(\beta) = \sup_{\lambda\geq 0} \min_{u\in\overline{M}(U_M)} J_{tc}^{\lambda}(\beta,u) = \sup_{\lambda\geq 0} \min_{u\in U_D} J_{tc}^{\lambda}(\beta,u). \tag{31}$$

Moreover, there exists some $u^* \in \mathcal{U}$ such that

$$C_{tc}(\beta) = \inf_{u\in\mathcal{U}} \sup_{\lambda\geq 0} J_{tc}^{\lambda}(\beta,u) = \sup_{\lambda\geq 0} J_{tc}^{\lambda}(\beta,u^*), \tag{32}$$

and u^* is optimal for **COP**.

In order to prove the theorem we need the following (e.g. Aubin [12], 1993, p. 126):

Lemma 5.1 *(Minimax Theorem)*
Let G_1, G_2 be convex subsets of linear topological spaces, and let G_1 be compact. Consider a function $\Psi : G_1 \times G_2 \to \mathbb{R}$ such that
– for each $g_2 \in G_2$, $g_1 \to \Psi(g_1,g_2)$ is convex and lower semi-continuous, and
– for each $g_1 \in G_1$, $g_2 \to \Psi(g_1,g_2)$ is concave.
Then there exists some $g_1^* \in G_1$ such that

$$\inf_{G_1}\sup_{G_2} \Psi(g_1,g_2) = \sup_{G_2} \Psi(g_1^*,g_2) = \sup_{G_2}\inf_{G_1} \Psi(g_1,g_2).$$

We are now ready to prove Theorem 5.1.
Proof of Theorem 5.1: (i) The first equality in (29) is standard: if $u \in U$ is feasible (i.e., it satisfies the constraints $D_{tc}(\beta,u) \leq V$) then

$$\sup_{\lambda\geq 0} J_{tc}^{\lambda}(\beta,u) = C_{tc}(\beta,u). \tag{33}$$

If $u \in U$ is not feasible then it is easily seen that

$$\sup_{\lambda\geq 0} J_{tc}^{\lambda}(\beta,u) = \infty. \tag{34}$$

We conclude that

$$\inf_{u\in U}\sup_{\lambda\geq 0} J_{tc}^{\lambda}(\beta,u) = \inf_{u\in U, D_{tc}(\beta,u)\leq V}\ \sup_{\lambda\geq 0} J_{tc}^{\lambda}(\beta,u) = \inf_{u\in U, D_{tc}(u)\leq V} C_{tc}(\beta,u) = C_{tc}(\beta).$$

Similarly, let $C'_{tc}(\beta) := \inf C_{tc}(\beta, u)$ over the set $\{u \in U_M : D_{tc}(\beta, u) \le V\}$. Then we have

$$C'_{tc}(\beta) = \inf_{u \in U_M} \sup_{\lambda \ge 0} J^{\lambda}_{tc}(\beta, u).$$

However, since the Markov policies are dominating (see e.g. Theorem 4.1 or [3] Theorem 2.1), it follows that

$$C'_{tc}(\beta) = C_{tc}(\beta). \tag{35}$$

This establishes the second equality. The third equality follows by the same arguments. This establishes (i). Moreover, (33) and (34) imply (ii).

(iii) We shall apply Lemma 5.1 where G_1 stands for the convex and compact set $\overline{M}(U_M)$ (for a discussion on the compactness, see [3, 4]), and G_2 stands for the convex set $\{\lambda \ge 0\}$. $J^{\lambda}(\beta, u) : G_1 \times G_2 \to \mathbb{R}$ is affine in both its arguments, and thus in particular, convex in its first argument and concave in the second. $J^{\lambda}(\beta, u)$ is lower semi-continuous in u (see Lemma 6.3 in [3] and Lemma 9.3 in the Appendix). Hence by Lemma 5.1, we have

$$\sup_{\lambda \ge 0} \min_{u \in \overline{M}(U_M)} J^{\lambda}_{tc}(\beta, u) = \inf_{u \in \overline{M}(U_M)} \sup_{\lambda \ge 0} J^{\lambda}_{tc}(\beta, u) = C_{tc}(\beta) \tag{36}$$

where the last equality follows from (29); this establishes the first equality. For fixed λ, $J^{\lambda}_{tc}(\beta, u)$ is minimized by a policy in U_D (by Theorem 3.1 and 3.2), i.e.

$$\min_{u \in \overline{U}} J^{\lambda}_{tc}(\beta, u) = \min_{u \in U_D} J^{\lambda}_{tc}(\beta, u). \tag{37}$$

for any class of policies $\overline{U}$ that contains U_D. The proof of (31) is established by combining this with eq. (36).

By combining (36) with (37), we get

$$C_{tc}(\beta) = \sup_{\lambda \ge 0} \min_{u \in \mathcal{U}} J^{\lambda}_{tc}(\beta, u). \tag{38}$$

(32) is obtained by applying again Lemma 5.1 with G_1 as the convex and compact set $\mathcal{U}$, and G_2 as the convex set $\{\lambda \ge 0\}$. Here again, $J^{\lambda}_{tc}(\beta, u) : G_1 \times G_2 \to \mathbb{R}$ is affine in both its arguments, and thus in particular, convex in its first argument and concave in the second; $J^{\lambda}_{tc}(\beta, u)$ is lower semi-continuous in u (see Lemma 6.3 in [3] and Lemma 9.3 in the Appendix). This implies, in particular, the existence of $u^* \in \mathcal{U}$ that minimizes the Lagrangian $J^{\lambda}_{tc}(\beta, u)$. By part (ii) of the theorem it is also optimal for **COP**. ∎

Corollary 5.1 *(Dominance of $\mathcal{U}$)*
Consider the transient framework, (Definition 2.1) and nonnegative immediate costs. Then $\mathcal{U}$ is a dominating class of policies.

Proof: Choose an arbitrary policy v. Consider a new **COP** with the same state and action spaces, the same transition probabilities, and with $K+1$ constraints. The immediate costs $\{\tilde{c}, \tilde{d}^1, ..., \tilde{d}^{K+1}\}$ are given in terms of the immediate costs corresponding to the original **COP**:

$$\tilde{c} = 0, \qquad \tilde{d}^k = d^k, k = 1, ..., K, \qquad \tilde{d}^{K+1} = c.$$

We set

$$\tilde{V}_k := \tilde{D}_{tc}(\beta, v), \ \ k = 1, ..., K+1.$$

The new **COP** is feasible, since the policy v is feasible (by definition of $\tilde{V}_k$). By the same arguments as in the first part of the proof of Theorem 5.1, it follows from (32) that there exists a feasible policy $\overline{u}$ among $\mathcal{U}$ for the new **COP**. This implies that $\overline{u}$ dominates v (for the original **COP**). ■

For the contracting framework, the statement corresponding to Corollary 5.1 has been established in [3] Theorem 7.1. For MDPs with a uniform Lyapunov function, it is established in Lemma 9.3 in the Appendix.

Next, we consider the existence of maximizing Lagrangians.

Theorem 5.2 *(The Lagrangian: Slater condition)*
Consider either the transient framework or MDPs with a uniform Lyapunov function. If there exists some policy u for which $D_{tc}(\beta, u) < V$ then there exist nonnegative Lagrange multipliers $\lambda^* = \{\lambda_1^*, ..., \lambda_K^*\}$ such that

$$C_{tc}(\beta) = \min_{u \in \mathcal{U}} J_{tc}^{\lambda^*}(\beta, u) = \min_{u \in U_D} J_{tc}^{\lambda^*}(\beta, u). \tag{39}$$

Moreover, any optimal policy u^* satisfies the Kuhn-Tucker conditions:

$$\lambda_k^*(D_{tc}^k(\beta, u^*) - V_k) = 0, \qquad k = 1, ..., K.$$

Proof: $J_{tc}^{\lambda}(\beta, u)$ is a convex function over the convex set $\mathcal{U}$, and $D_{tc}^k(u)$ are lower semi-continuous in $\mathcal{U}$. By a standard minimax theory (see e.g. Rockafellar [63] p. 45, and Theorems 17 and 18 in p. 41) it follows that there exist nonnegative Lagrange multipliers $\lambda^* = \{\lambda_1^*, ..., \lambda_K^*\}$ such that

$$\min_{u \in \mathcal{U}} J_{tc}^{\lambda^*}(\beta, u) = \sup_{\lambda \geq 0} \min_{u \in \mathcal{U}} J_{tc}^{\lambda}(\beta, u) = \min_{u \in \mathcal{U}} \sup_{\lambda \geq 0} J_{tc}^{\lambda}(\beta, u),$$

which equals $C_{tc}(\beta)$, according to Theorem 5.1 (iii). The second equality in (39) follows from the fact that for fixed λ, $J^{\lambda}(\beta, u)$ is minimized by a policy in U_D (by Theorem 3.1 (ii) and 3.2 (ii)). The Kuhn-Tucker conditions follow from standard arguments (similar to the proof of part (i) of Theorem 5.1, see Rockafellar [63], Theorem 15). ■

Since stationary policies are dominating (Theorem 7.1 in [3] and Lemma 9.3 in the Appendix) under the suitable conditions, Theorem 5.1 implies the following Corollary.

Corollary 5.2 *(Saddle point)*
Consider either the transient framework (Definition 2.1 and nonnegative immediate cost), or an MDP with a uniform Lyapunov function. Then for any class of policies $\overline{U}$ *that contains either* U_S *or* $\mathcal{U}$*, we have*

$$C_{tc}(\beta) = \sup_{\lambda \geq 0} \min_{u \in \overline{U}} J^{\lambda}_{tc}(\beta, u) = \min_{u \in \overline{U}} \sup_{\lambda \geq 0} J^{\lambda}_{tc}(\beta, u) = \sup_{\lambda \geq 0} J^{\lambda}_{tc}(\beta, u^*) \tag{40}$$

for some $u^* \in \overline{U}$*.*

Proof: The proof for $\mathcal{U}$ is straightforward. We establish the claims for U_S. The first equality follows exactly as (35). That $C_{tc}(\beta) = \sup_{\lambda \geq 0} J^{\lambda}_{tc}(\beta, u^*)$ follows from Theorem 5.1 (ii) and the dominance of U_S. Combining this with the first equality in (29) implies the last equality in (40). ■

6 The dual LP

Next, we introduce the Linear Program which is the dual to the one we considered in [3]. Consider the DP with decision variables $\phi(y), y \in \mathbf{X}$ and $\lambda \in \mathbb{R}^K_+$ ($\mathbb{R}^K_+$ are vectors in $\mathbb{R}^K$ whose entries are nonnegative).

$$\mathbf{DP_1}(\beta): \qquad \text{Find } \Theta^*(\beta) := \sup_{\phi,\lambda} \langle \beta, \phi \rangle - \langle \lambda, V \rangle \text{ s.t.}$$
$$\phi(x) \leq c(x,a) + \langle \lambda, d(x,a) \rangle + \sum_{y \in \mathbf{X}'} \mathcal{P}_{xay} \phi(y), \quad x \in \mathbf{X}, a \in \mathsf{A}(x).$$

We begin by considering the MDP with uniform Lyapunov function. Combining Theorem 3.4 with Theorem 5.1, we show that the value of $\mathbf{DP_1}(\beta)$ equals the value of **COP**. This, together with Theorem 9.2 in the Appendix, implies that there is no duality gap between the Linear Program $\mathbf{LP}^{\mu}_{\mathbf{1}}(\beta)$ (presented already in [3] for the contracting framework, and extended to MDPs with uniform Lyapunov function in Theorem 9.2 in the Appendix), and its dual Program $\mathbf{DP_1}(\beta)$.

Theorem 6.1 *(The dual LP for MDPs with uniform Lyapunov function)*
Consider an MDP with a uniform Lyapunov function. Consider $\mathbf{DP_1}(\beta)$ *restricted to* $\psi \in \mathsf{F}^{\mu}$*.* $\mathbf{DP_1}(\beta)$ *is feasible if and only if* **COP** *is feasible. The value of* $\mathbf{DP_1}(\beta)$ *equals* $C_{tc}(\beta)$ *and* $\phi(x) = C_{tc}(x), x \in \mathbf{X}$ *is an optimal solution.*

A similar result is obtained for the absorbing and transient cases with nonnegative immediate costs. By combining Theorems 3.5, 3.6 and 5.1 we have:

Theorem 6.2 *(The dual LP, the absorbing and transient case: unbounded costs)*
Consider either
(i) an $\mathbf{X}'$*-absorbing MDP with nonnegative immediate costs, and with* $\mathbf{DP_1}(\beta)$ *restricted to bounded* ϕ*; or*
(ii) an $\mathbf{X}'$*-transient MDP with nonnegative immediate costs, with* $\mathbf{DP_1}(\beta)$ *restricted to* ϕ *satisfying (20).*
Then $\mathbf{DP_1}(\beta)$ *is feasible (*$\phi = 0$ *is a feasible solution). Moreover, the value of* $\mathbf{DP_1}(\beta)$ *equals* $C_{tc}(\beta)$*.*

7 State truncation

We consider in this section transient MDPs with nonnegative immediate costs. We already showed in Remark 3.3 that the value of a non-constrained MDP can be computed as the limit of the (increasing sequence of) values of the MDPs with truncated spaces (as described in (22)). In other words, we showed that

$$\lim_{n\to\infty} C_{tc}^n(\beta) = \sup_{n\in\mathbb{N}} C_{tc}^n(\beta) = C_{tc}(\beta)$$

where $C_{tc}^n(\beta)$ is the value of the MDP truncated to the finite set $\mathbf{X}_n$. Moreover, we showed that the optimal policies converge.

It should thus not be surprising that a similar result holds for the constrained MDP. Indeed, for any $\lambda \geq 0$ we have by Remark 3.3:

$$\lim_{n\to\infty} J_{tc}^{\lambda,n}(\beta) = \sup_{n\in\mathbb{N}} J_{tc}^{\lambda,n}(\beta) = J_{tc}^{\lambda}(\beta) \tag{41}$$

where

$$J_{tc}^{\lambda,n}(\beta) = \inf_{u\in U} J_{tc}^{\lambda,n}(\beta,u), \qquad J_{tc}^{\lambda}(\beta) = \inf_{u\in U} J_{tc}^{\lambda}(\beta,u)$$

and where $J_{tc}^{\lambda,n}(\beta,u)$ is the Lagrangian defined in (30) corresponding to the nth truncated MDP. According to Corollary 5.2, we have

$$C_{tc}^n(\beta) = \min_{u\in U}\sup_{\lambda\geq 0} J_{tc}^{\lambda,n}(\beta,u), \qquad C_{tc}(\beta) = \sup_{\lambda\geq 0}\min_{u\in U} J_{tc}^{\lambda}(\beta,u).$$

Combining this with (41), we have

$$C_{tc}(\beta) = \sup_{\lambda\geq 0}\sup_{n\in\mathbb{N}}\inf_{u\in U} J_{tc}^{\lambda,n}(\beta,u) = \sup_{n\in\mathbb{N}}\sup_{\lambda\geq 0}\inf_{u\in U} J_{tc}^{\lambda,n}(\beta,u) = \sup_{n\in\mathbb{N}} C_{tc}^n(\beta).$$

This establishes the convergence of the values for the state-truncated **COP** to the value of **COP**. Unlike previous approaches for state approximations for **COP** (most of which were derived for the contracting framework, see [4] and [1, 2]), we do not need here any Slater type condition.

8 A second LP approach for optimal mixed policies

In this section we present an alternative LP formulation for **COP**. The decision variables will correspond to the probability measures over the space of all stationary-deterministic policies; in particular, this will mean for the case that the state and action spaces are finite, that the number of decision variables will be equal to the number of stationary-deterministic policies; this is in contrast with the previous LP approach for which the number of decision variables is typically much smaller: $\sum_{x \in \mathbf{X}} |\mathsf{A}(x)|$.

It follows from Corollary 5.1 (for the transient framework with nonnegative immediate costs) and from Altman [3] (for the contracting case) that $C_{tc}(\beta)$ is the value of **COP** restricted to $\mathcal{U}$:

$$\min_{u \in \mathcal{U}} C_{tc}(\beta, u) \text{ s.t. } D_{tc}(\beta, u) \leq V.$$

For the case of uniform Lyapunov function, the restriction to $\mathcal{U}$ will be justified in Lemma 9.3. **COP** restricted to $\mathcal{U}$ can be rewritten as a Linear Program:

$$\begin{aligned} \mathbf{LP_2}(\beta): \qquad & \min_{\gamma \in M_1(U_D)} \int C_{tc}(\beta, u) \gamma(du) \\ \text{s.t.} \qquad & \int D^k_{tc}(\beta, u) \gamma(du) \leq V^k, \quad k = 1, ..., K. \end{aligned} \tag{42}$$

This yields the following:

Theorem 8.1 *(Relation between* **COP** *and* $\mathbf{LP_2}(\beta)$*)*
Consider either the transient framework (with nonnegative immediate costs), or an MDP with a uniform Lyapunov function. Then
(i) **COP** *is feasible if and only if* $\mathbf{LP_2}(\beta)$ *is feasible (i.e. the set satisfying (42) is nonempty). If* $\mathbf{LP_2}(\beta)$ *is feasible then there exists an optimal policy in* $\mathcal{U}$ *for* **COP**.
(ii) The values of **COP** *and of* $\mathbf{LP_2}(\beta)$ *are equal.*
(iii) If γ *is a solution of* $\mathbf{LP_2}(\beta)$ *then the policy* $\hat{\gamma} \in \mathcal{U}$ *is optimal for* **COP**.

9 Appendix: extensions of results in [3]

We extend in this Section results that appeared in [3] on the continuity of occupation measures, and on the LP approach for absorbing MDPs. (We use the same notations as those in [3].) We also correct related minor mistakes that appeared in [3].

Lemma 9.1 *(Boundedness of the life time)*
(i) For $\mathbf{X}'$*-transient MDPs,* $\sup_{u \in U} f_{tc}(\beta, u; y) < \infty$ *for all* $y \in \mathbf{X}$.
(ii) For an $\mathbf{X}'$*-absorbing MDP,* $\sup_{u \in U} f_{tc}(\beta, u; \mathbf{X}) < \infty$.
(iii) For a contracting MDP, $\sup_{u \in U} \sum_{y \in \mathbf{X}'} f_{tc}(\beta, u; y) \mu(y) < \infty$*, and* $f_{tc}(\beta, u; \cdot)$ *are integrable w.r.t.* μ *uniformly over all policies.*

Proof: (i) and (ii) follow from a result on positive dynamic Programming, see e.g. p. 108 in [31]. It states the following: given a nonnegative costs c, for any $\epsilon > 0$, there exists some policy w s.t.

$$\sup_{u \in U} C_{tc}(x, u) < C_{tc}(x, w) + \epsilon, \quad \forall x \in \mathbf{X}.$$

Hence

$$\sup_{u \in U} C_{tc}(\beta, u) < C_{tc}(\beta, w) + \epsilon. \tag{43}$$

By choosing $c(x, a) = 1\{x = y\}$, we obtain (i). By considering $c(x, a) = 1\{x \in \mathbf{X}'\}$, we obtain (ii). The first part of (iii) was established in Lemma 4.1 in [3] and can be obtained alternatively by (43) with $c(x, a) = \mu(x)$ (see [73] Lemma 5.3(ii) and its proof). For the uniform integrability, we shall restrict, without loss of generality, to Markov policies. The uniform integrability can be obtained from the continuity of $C_{tc}(\beta, u)$ over U_M (Lemma 6.2 in [3]). Assume that $f_{tc}(\beta, u; \cdot)$ is not uniformly integrable w.r.t. μ over U_M. Then there exists some $\epsilon' > 0$, a sequence of compact sets $\mathbf{X}_n \subset \mathbf{X}$ that increases to $\mathbf{X}$ and a sequence of Markov policies $u(n)$ such that for all n,

$$\sum_{y \in \mathbf{X}} f_{tc}(\beta, u(n); y) c_n(y) > \epsilon', \qquad \text{where } c_n(y) \stackrel{\text{def}}{=} \mu(y) 1\{y \in \mathbf{X}' \backslash \mathbf{X}_n\}.$$

Choose some subsequence of $u(n)$ that converges to some Markov policy u. Then due to the continuity of the total expected cost and the linear representation of the cost ([3] Theorem 6.1), we have for any integer m,

$$\langle f_{tc}(\beta, u), c_m \rangle = \lim_{n \to \infty} \langle f_{tc}(\beta, u(n)), c_m \rangle \geq \lim_{n \to \infty} \langle f_{tc}(\beta, u(n)), c_n \rangle \geq \epsilon'.$$

But on the other hand, since c_m decreases to zero as $m \to \infty$, we have by the monotone convergence theorem

$$\lim_{m \to \infty} \langle f_{tc}(\beta, u), c_m \rangle = 0.$$

Hence, the uniform integrability is established by contradiction. ∎

The statements of the above theorem also hold for mixed policies (this follows from Theorem 2.1 in [3]). The sufficiency of Markov policies (Theorem 2.1 in [3]) together with (43) imply the following:

Corollary 9.1 *(Sufficient conditions for transient and absorbing MDPs)*
A sufficient condition for an MDP to be $\mathbf{X}'$-transient (resp., $\mathbf{X}'$-absorbing) is that every stationary policy is $\mathbf{X}'$-transient (resp., $\mathbf{X}'$-absorbing).

Next, we present a property of the contracting framework, related to the μ-continuity of the transition probabilities.

Lemma 9.2 *(μ-continuity and uniform integrability)*
The μ-continuity of the transition probabilities, defined in (9), together with the assumption that

$$\sum_{y \notin \mathcal{M}} \mathcal{P}_{xay} \mu(y) < \infty \tag{44}$$

for all $a \in \mathsf{A}(x)$, imply that
(i) For any fixed x, $\{\mathcal{P}_{xay}\}_a$ are integrable with respect to μ uniformly in a, i.e. for any sets $\mathbf{X}_n$ increasing to $\mathbf{X}$,

$$\lim_{n\to\infty} \sup_a \sum_{y \notin \mathbf{X}_n} \mathcal{P}_{xay} \mu(y) = 0.$$

(ii) Let $u^n \to u$ (weakly) where u and u^n are probability measures over $\mathsf{A}(x)$. Then

$$\lim_{n\to\infty} \sum_{y \in \mathbf{X}} |P_{xy}(u^n) - P_{xy}(u)| \, \mu(y) = 0. \tag{45}$$

Proof: Assume that (i) does not hold. Let $a(n)$ be a sequence of actions converging to some a such that

$$\lim_{n\to\infty} \sum_{y \notin \mathbf{X}_n} \mathcal{P}_{xa(n)y} \mu(y) = \delta > 0. \tag{46}$$

However, the μ-continuity of the transition probabilities implies that

$$\lim_{n\to\infty} \sum_{y \notin \mathbf{X}_n} |\mathcal{P}_{xa(n)y} - \mathcal{P}_{xay}| \, \mu(y) = 0. \tag{47}$$

Since assumption (44) implies that $\lim_{n\to\infty} \sum_{y \notin \mathbf{X}_n} \mathcal{P}_{xay}\mu(y) = 0$, we deduce from (47) that

$$\lim_{n\to\infty} \sum_{y \notin \mathbf{X}_n} \mathcal{P}_{xa(n)y} \mu(y) = 0,$$

which contradicts (46), thus establishing (i).
(ii) The weak convergence of u^n implies that $P_{xy}(u^n)$ converges to $P_{xy}(u)$ for all x and y. (i) implies that $P_{x\cdot}(u^n)$ are integrable with respect to μ uniformly in n. By combining these arguments one obtains

$$\lim_{n\to\infty} \sum_{y \in \mathbf{X}'} P_{xy}(u^n)\mu(y) = \sum_{y \in \mathbf{X}'} P_{xy}(u)\mu(y)$$

(this follows from [14] Theorem 5.4, p. 32). This implies (45) (by Scheffé's Lemma [77] p. 55). ■

Next we present definitions of lower semi-continuity.

Definition 9.1 *(Lower semi-continuity)*
Consider a class of policies $\overline{U}$ and some Borel space Y. Consider a measured-value function $f : \overline{U} \to M(Y)$, where $f(u; \mathcal{Y})$ stands for the measure it assigns to a Borel set $\mathcal{Y} \subset Y$ for $u \in \overline{U}$. It is said to be weakly l.s.c. on $\overline{U}$ if for any $u \in \overline{U}$ and any sequence $u^n \in \overline{U}$ that weakly converges to u, we have

$$\underline{\lim}_{n\to\infty} \langle f(u^n), c\rangle \geq \langle \underline{\lim}_{n\to\infty} f(u^n), c\rangle,$$

for any nonnegative function $c : Y \to \mathbb{R}$ which is lower semi-continuous.

The following theorem generalizes Lemma 5.1 (p. 59) in [3] and establishes the continuity of the occupation measures for MDPs with a uniform Lyapunov function.

Theorem 9.1 *(Properties of f_{tc})*
(i) For transient MDPs, the measures $f_{tc}(\beta, \cdot)$ defined on $\mathcal{K}$ are weakly l.s.c. over U_M. The state occupation measures $f_{tc}(\beta, u; \cdot)$ are weakly continuous over U_M. The above statements hold also for the policies $\overline{M}(U_M)$.
(ii) Consider an MDP with a uniform Lyapunov function. For any continuous function $c' : \mathcal{K} \to \mathbb{R}$ satisfying $|c'| \leq \nu + 1$, $\langle f_{tc}(\beta, \cdot), c'\rangle$ are continuous over U_M (thus the measures $f_{tc}(\beta, \cdot)$, defined over $\mathcal{K}$, are (weakly) continuous over U_M). The total cost is linear in the occupation measure:

$$C_{tc}(\beta, u) = \langle f_{tc}(\beta, u), c\rangle := \int_{\mathcal{K}} c(\kappa) f_{tc}(\beta, u; d\kappa) \tag{48}$$

for any $u \in U$. All the above statements hold also for the policies $\overline{M}(U_M)$.
(iii) Consider a contracting MDP. Then the state occupation measures $f_{tc}(\beta, \cdot)$ (defined on $\mathbf{X}$) are μ-continuous over U_M.

Proof: (i) Assume that $u^n \to u$, where $u^n, u \in U_M$ (i.e. for any $x \in \mathbf{X}$ and t, $u_t^n(x)$ converges weakly to $u_t(x)$). Then $P_{xy}(u_1^n) \to P_{xy}(u_1)$ for all $x, y \in \mathbf{X}$. By the bounded convergence theorem, this implies that $\sum_x \beta(x) P_{xy}(u_1^n) \to \sum_x \beta(x) P_{xy}(u_1)$ for all $y \in \mathbf{X}$. Moreover, the m step probabilities also converge, and for all integers m,

$$\lim_{n\to\infty} p_\beta^{u^n}(m; y, \mathcal{A}) = p_\beta^u(m; y, \mathcal{A}) \tag{49}$$

for all $y \in \mathbf{X}$, and for sets $\mathcal{A}$ for which $u^n(\mathcal{A}|y)$ converges to $u(\mathcal{A}|y)$. (This is established in eq. (20) in [3]; however it is mistakenly written in [3] that it holds for all sets $\mathcal{A}$.)

Denote

$$Z^n(m, y) := \int_{\mathsf{A}(y)} p_\beta^{u^n}(m; y, da) c'(y, a),$$

$$Z(m, y) := \int_{\mathsf{A}(y)} p_\beta^{u}(m; y, da) c'(y, a),$$

where c' is an arbitrary nonnegative l.s.c. function. Since $p_\beta^{u^n}(m)$ converges weakly to $p_\beta^u(m)$ (from (49)) for all m and y, it follows that

$$\underline{\lim}_{n\to\infty} Z^n(m,y) \geq Z(m,y),$$

see Doob [30] p. 133 (in the related theorem in Doob [30], the cost c' is assumed to be bounded. However, it can easily be seen that only boundedness from below is used in the proof of that theorem). Applying Fatou's Lemma ([67] Proposition 11.17) with respect to the (infinite) measure over $\mathbb{N} \times \mathbf{X}$ generated by $\mu_n(m,y) = \mu(m,y) = 1$ for all m and y, we obtain

$$\underline{\lim}_{n\to\infty} \langle f_{tc}(\beta, u^n), c' \rangle = \underline{\lim}_{n\to\infty} \sum_{m,y} Z^n(m,y) \geq \sum_{m,y} Z(m,y) = \langle f_{tc}(\beta,u), c' \rangle,$$

which concludes the proof of the first statement in (i). (The equalities above follow from the fact that the integrand is nonnegative, see Tonelli's Theorem in Royden, 1988, Thm 12.20). This implies all the statements in (i) concerning the l.s.c. in U_M.

The lower semi-continuity on $\overline{M}(U_M)$ follows from the lower semi-continuity on U_M, see Doob (1994) p. 133.

(ii) $E_x^u \hat{\mu}(X_n) 1\{T > n\}$ is continuous over U_M for all $x \in \mathbf{X}$ and $n \in \mathbb{N}$, where $\hat{\mu}$ is any function that is bounded by the Lyapunov function μ. To show that, we first note that $E_\beta^u \hat{\mu}(X_2) 1\{T > 2\}$ is continuous in $u \in U_M$; this follows from condition M1(ii) in the definition 2.3. We proceed by induction to show that

$$E_x^u \hat{\mu}(X_n) 1\{T > n\} = [({}_{\mathcal{M}}P(u_1)\, {}_{\mathcal{M}}P(u_2) \ldots {}_{\mathcal{M}}P(u_{n-1}))\, \hat{\mu}]_x \tag{50}$$

is continuous in u. Suppose (50) holds for some n. We have

$$E_x^u \hat{\mu}(X_{n+1}) 1\{T > n+1\} = \sum_{y \in \mathbf{X}'} {}_{\mathcal{M}}P_{xy}(u_1) Z_y(u)$$

where

$$Z_y(u) \stackrel{\text{def}}{=} [({}_{\mathcal{M}}P(u_2)\, {}_{\mathcal{M}}P(u_3) \ldots {}_{\mathcal{M}}P(u_n))\, \hat{\mu}]_x \,.$$

By iterating property M1(i) in Definition 2.3, it follows that $Z_y(u) \leq \mu(y)$ for any u and y. Moreover, the inductive assumption implies that Z_y is continuous in $u \in U_M$. Since $[{}_{\mathcal{M}}P(u_1)\mu]_x$ is continuous in u_1, it follows from the generalized dominated convergence that

$$[{}_{\mathcal{M}}P(u_1) Z(u)]_x = E_x^u \hat{\mu}(X_{n+1}) 1\{T > n+1\}$$

is continuous in $u \in U_M$, which establishes the induction.

It follows from property M1(i) that for all $t \geq 1$,

$$E_\beta^u c'(X_t, A_t) 1\{T > t\} + E_\beta^u \mu(X_{t+1}) 1\{T > t+1\} \leq E_\beta^u \mu(X_t) 1\{T > t\}. \tag{51}$$

By summing over $t \in n, ..., n+m$ we get

$$E^u_\beta \sum_{t=n}^{n+m} \left[c'(X_t, A_t) 1\{T > t\} \right] \le E^u_\beta \mu(X_n) 1\{T > n\}. \tag{52}$$

Taking the limit as $m \to \infty$, we get

$$E^u_\beta \sum_{t=n}^{\infty} \left[c'(X_t, A_t) 1\{T > t\} \right] \le \sup_{u \in U} E^u_\beta \mu(X_n) 1\{T > n\},$$

which tends to zero [4, 48].
Thus, $E^u_\beta \sum_{t=1}^{n} \left[(c'(X_t, A_t) 1\{T > t\} \right]$ converges to $C'_{tc}(\beta, u)$ uniformly in $u \in U_m$.

For any n, we shall show that $E^u_\beta \left[c'(X_n, A_n) 1\{T > n\} \right]$, and thus $E^u_\beta \sum_{t=1}^{n} \left[c'(X_t, A_t) 1\{T > t\} \right]$, are continuous in u. The proof is then established by combining this continuity with the uniform convergence (that we established in the previous paragraph).

Let $c'(x, u_t) \stackrel{\text{def}}{=} \int c'(x, a) u_t(da|x)$. We have

$$E^u_\beta \left[c'(X_t, A_t) 1\{T > t\} \right] = \sum_{x \in \mathbf{X}} P^u_\beta(X_t = x, T > t) c'(x, u_t)$$

We note that $c'(x, u_t)$ is continuous in u_t for every x, since it $c'(\cdot, \cdot)$ is continuous. Moreover, $P^u_\beta(X_t = x, T > t)$ are continuous over $u \in U_M$; this follows from (50). Since $c'(x, u_t) \le \nu(x, u_t) + 1 \le \mu(x)$ for any u and x, and since $\sum_{x \in \mathbf{X}} P^u_\beta(X_t = x, T > t)\mu(x)$ is continuous, it follows from the dominated convergence Theorem that $E^u_\beta \left[c'(X_t, A_t) 1\{T > t\} \right]$ is continuous in u. This establishes the statements on continuity for U_M.

Let γ^n be a sequence in $M_1(U_M)$ converging weakly to some γ. Let $\hat{\gamma}^n$ and $\hat{\gamma}$ be the corresponding policies in $\overline{M}(U_M)$. Then the continuity on $\overline{M}(U_M)$ follows since $f_{tc}(\beta, \cdot)$ are bounded and continuous functions on U_M, so that the weak convergence of γ^n implies (Theorem 2.1 in [14]):

$$\lim_{n \to \infty} f_{tc}(\beta, \hat{\gamma}^n) = \lim_{n \to \infty} \langle \gamma^n, f_{tc}(\beta, \cdot) \rangle = \langle \lim_{n \to \infty} \gamma^n, f_{tc}(\beta, \cdot) \rangle = f_{tc}(\beta, \hat{\gamma}).$$

Next, we establish (48). For the finite horizon cost, we have by Fubini's Theorem ([67] Theorem 12.19) $C^t_{tc}(\beta, u) = \langle \sum_{s=1}^{t} p^u_\beta(s), c \rangle$. Since $\sum_{s=1}^{t} p^u_\beta(s)$ converges (setwize) to $f_{tc}(\beta, u)$ when $t \to \infty$, we have by the monotone convergence Theorem

$$\lim_{t \to \infty} \langle \sum_{s=1}^{t} p^u_\beta(s), |c| \rangle = \langle f_{tc}(\beta, u), |c| \rangle.$$

Since

$$\langle f_{tc}(\beta, u), |c| \rangle \le \langle f_{tc}(\beta, u), \nu \rangle \le \langle \beta, \mu \rangle < \infty,$$

(the inequality before the last follows from (52) with $n = 1$) we have by the bounded convergence Theorem

$$C_{tc}(\beta, u) = \lim_{t\to\infty} C_{tc}^t(\beta, u) = \lim_{t\to\infty} \langle \sum_{s=1}^{t} p_\beta^u(s), c\rangle = \langle f_{tc}(\beta, u), c\rangle.$$

(iii) is proven in Lemma 5.1 (ii) in [3]. We mention some typos: in the 2nd, 3rd and 5th lines of the proof of Lemma 5.1 (ii) in [3], the argument of P_{xy} should be u_1. The argument y is missing twice in l. 15 of p. 61. We also mention that l. 2 of the proof of Lemma 5.1 (ii) in [3] follows from Lemma 9.2. ■

As an example where the occupation measures are l.s.c. in the policies, and not continuous, consider the following example.

Example 9.1 Consider $\mathbf{X}'$ to be the natural numbers, and assume that $\mathsf{A}(x) = \{0, 1\}$ for all x. For any state $x > 1$, action 0 results in a transition to state $x + 1$, and action 1 results in a transition to state 1. From state 1 we leave in one step the set $\mathbf{X}'$. Assume that we start at some state $y > 1$. For any n, a policy u^n that chooses action 1 for the first time at some instant larger than n achieves $f_{tc}(\beta, u^n; 1) = 1$. But the policy u that is obtained as the weak limit of u^n, i.e. the policy that always chooses action 0, achieves $f_{tc}(\beta, u; 1) = 0$.

Note that condition

$$\lim_{n\to\infty} \sup_{u\in U_M} \sum_{t=n}^{\infty} p_\beta^u(t; \mathbf{X}') = 0, \tag{53}$$

is not satisfied (this condition holds if the MDP has a uniform Lyapunov function, see [20, 4]). Indeed, $\sum_{t=n}^{\infty} p_\beta^{u^n}(t; 1) = 1$. Thus $\lim_{n\to\infty} \sup_{u\in U_M} \sum_{t=n}^{\infty} p_\beta^u(t; \mathbf{X}') \geq 1$. ■

In the above example, the MDP is transient. However, continuity may fail also for absorbing MDPs. In fact, in the Counter Example of Fisher & Ross [42], the total expected time till state 0 is reached is not continuous over the stationary policies. (This follows by combining the calculations in [42] with eq. (11.17) in the analysis in [73] of this example.)

Theorem 9.1 (ii) enables to extend other results from [3] for the case of uniform Lyapunov function. In particular, we have

Lemma 9.3 *(Uniform Lyapunov function: properties of the cost)*
Consider an MDP with a uniform Lyapunov function. Then any complete class of policies (see definition below (3)) is a dominating class of policies. If **COP** *is feasible, then there exist an optimal policy in U_S and in $\mathcal{U}$.*

Proof: The first part follows from part (48) and the definition of completeness. The proof of the second part is the same as the proof of Theorem 7.1 (iii) in [3]. ■

Our next goal is to obtain the primal LP solution for **COP** for MDPs with a uniform Lyapunov function. To that end, we need the following Lemma, whose proof follows immediately from the proof of Lemma 3.1 (i) in [3].

Lemma 9.4 *(Stationary policies in transient and absorbing MDPs)*
Fix some initial distribution β and a stationary $\mathbf{X}'$-transient policy w. Then $f^*(x) := \sum_{t=1}^{\infty} p_\beta^w(t;x)$ is the unique solution to

$$f = \beta + f\,{}_{\mathcal{M}}P(w), \qquad f \geq 0, \tag{54}$$

among f that satisfy $\lim_{n\to\infty} f[{}_{\mathcal{M}}P]^n(w) = 0$ (where f and β are row vectors on $\mathbf{X}$, and ${}_{\mathcal{M}}P(w)$ is the transition probability matrix of the Markov chain corresponding to the stationary policy w, in which we replace columns corresponding to states in $\mathcal{M}$ by zero's).

Theorem 9.2 *(Occupation measures and LP solution for* **COP** *in the case of a uniform Lyapunov function)*
Consider an MDP with a uniform Lyapunov function. Then
(i) $\mathbf{L}(\beta) = \mathbf{Q}_{tc}^{\mathrm{b}}(\beta) = \mathbf{Q}_{tc}^{\nu}(\beta)$, where $\mathbf{Q}_{tc}^{\nu}(\beta)$ is given in (27), and $\mathbf{L}(\beta)$ is defined below (3).
(ii) Consider the following $\mathbf{LP}_{\mathbf{1}}^{\mathrm{b}}(\beta)$: Find the infimum $\mathcal{C}^*$ of $\mathcal{C}(\rho) := \langle \rho, c\rangle$ subject to:

$$\mathcal{D}^k(\rho) := \langle \rho, d^k\rangle \leq V_k, k = 1, ..., K, \qquad \rho \in \mathbf{Q}_{tc}^{\mathrm{b}}(\beta). \tag{55}$$

Then (ii.1) $\mathcal{C}^* = C_{tc}(\beta)$.
(ii.2) For any $u \in U$, $\rho(u) := f_{tc}(\beta, u) \in \mathbf{Q}_{tc}^{\mathrm{b}}(\beta)$, $C_{tc}(\beta, u) = \mathcal{C}(\rho(u))$ and $D_{tc}(\beta, u) = \mathcal{D}(\rho(u))$; conversely, for any $\rho \in \mathbf{Q}_{tc}^{\mathrm{b}}(\beta)$, consider any stationary policy $w(\rho)$ for which $w_y(\mathcal{A}) = \rho(y, \mathcal{A})[\rho(y, \mathsf{A}(y))]^{-1}$ whenever the denominator is nonzero. Then $C_{tc}(\beta, w(\rho)) = \mathcal{C}(\rho)$ and $D_{tc}(\beta, w(\rho)) = \mathcal{D}(\rho)$.
(ii.3) $\mathbf{LP}_{\mathbf{1}}^{\mathrm{b}}(\beta)$ is feasible if and only if **COP** *is. Assume that* **COP** *is feasible. Then there exists an optimal solution ρ^* for $\mathbf{LP}_{\mathbf{1}}^{\mathrm{b}}(\beta)$, and the stationary policy $w(\rho^*)$ is optimal for* **COP**.

Proof: (i) That $\mathbf{L}(\beta) \subset \mathbf{Q}_{tc}(\beta)$ was proved in Theorem 5.2 (i) in [3]. It is established in [4, 20, 48] that for all $f \in \mathbf{L}(\beta)$, $\langle f, \nu\rangle < \infty$ and f is a finite measure. Hence $\mathbf{L}(\beta) \subset \mathbf{Q}_{tc}^{\mathrm{b}}(\beta) \subset \mathbf{Q}_{tc}^{\nu}(\beta)$.

It remains to prove that $\mathbf{Q}_{tc}^{\mathrm{b}}(\beta) \subset \mathbf{L}(\beta)$. For any $f \in \mathbf{Q}_{tc}^{\mathrm{b}}(\beta)$, there exists some stationary policy w such that (54) holds, see the proof of Theorem 5.2 (i) in [3] eq. (23). Now, ${}_{\mathcal{M}}P^n(w)$ converges to 0 pointwize (this follows from (6)). By the bounded convergence Theorem ([67] Proposition 11.18), this implies that $\lim_{n\to\infty} f\,{}_{\mathcal{M}}P^n(w) = 0$. Lemma 9.4 implies, finally, that $\mathbf{Q}_{tc}^{\mathrm{b}}(\beta) \subset \mathbf{L}(\beta)$.

The proof of (ii) is similar to the proof of Theorem 8.1 in [3]. ■

Some more comments on [3]: Note that Theorem 9.1, together with Theorem 6.1 in [3] imply, in particular, Lemma 6.1 in [3] and the first part of Lemma 6.3 in [3] (whose proof was incorrect). Part (ii) of Theorem 9.1 was implicitly assumed to hold in the proof of Theorem 5.2 (ii) of [3]. Finally, in p. 58 l. 5 of [3], the word "closed" is missing in the definition coB_1 in [3].

Acknowledgement: The author wishes to thank Prof. Eugene Feinberg for many fruitful suggestions.

References

[1] E. Altman (1993), "Asymptotic Properties of Constrained Markov Decision Processes", *ZOR - Methods and Models in Operations Research*, **37**, Issue 2, pp. 151-170.

[2] E. Altman (1994), "Denumerable constrained Markov Decision Processes and finite approximations", *Math. of Operations Research*, **19**, pp. 169-191.

[3] E. Altman (1996), "Constrained Markov decision processes with total cost criteria: occupation measures and primal LP", *ZOR - Methods and Models in Operations Research*, **43**, issue 1, pp. 45-72.

[4] E. Altman (1995), *Constrained Markov Decision Processes*, book in preparation.

[5] E. Altman and A. Shwartz (1991a), "Markov decision problems and state-action frequencies", *SIAM J. Control and Optimization.* **29**, pp. 786-809.

[6] E. Altman and A. Shwartz (1993), "Time-sharing policies for controlled Markov chains", *Operations Research*, **41**, pp. 1116-1124.

[7] E. Altman and A. Shwartz (1995), "Constrained Markov Games: Nash Equilibria", submitted to *Annals of dynamic games.*

[8] E. Altman and F. Spieksma (1995), "The Linear Program approach in Markov Decision Problems revisited", *ZOR - Methods and Models in Operations Research*, **42**, Issue 2, pp. 169-188.

[9] E. Altman and O. Zeitouni (1994), "Rate of convergence of empirical measures and costs in controlled Markov chains and transient optimality", *Math. of Operations Research*, **19**, pp. 955-974.

[10] J. Anderson and P. Nash (1987), *Linear Programming in Infinite-Dimentional Spaces*, Wiley, England.

[11] A. Arapostathis, V. S. Borkar, E. Fernández-Gaucherand, M. K. Ghosh and S. I. Marcus (1993), "Discrete-time controlled Markov processes with average cost criterion: a survey", *SIAM J. Control and Optimization*, **31**, pp. 282-344.

[12] J. P. Aubin (1993), *Optima and Equilibria, An Introduction to Nonlinear Analysis*, Springer-Verlag.

[13] M. Bayal-Gursoy and K. W. Ross (1992), "Variability sensitive Markov decision processes", *Math. of Operations Research*, **17**, pp. 558-571.

[14] P. Billingsley (1968), *Convergence of Probability Measures*, J. Wiley, New-York.

[15] F. J. Beutler and K. W. Ross (1985), "Optimal policies for controlled Markov chains with a constraint", *J. Mathematical Analysis and Applications* **112**, 236-252.

[16] F. J. Beutler and K. W. Ross (1986), "Time-Average Optimal Constrained Semi-Markov Decision Processes", *Advances of Applied Probability*, **18**, pp. 341-359.

[17] V. S. Borkar (1988), "A convex analytic approach to Markov decision processes", *Prob. Th. Rel. Fields*, **78**, pp. 583-602.

[18] V. S. Borkar (1990), *Topics in Controlled Markov Chains*, Longman Scientific & Technical.

[19] V. S. Borkar (1994), "Ergodic control of Markov Chains with constraints – the general case", *SIAM J. Control and Optimization*, **32**, pp. 176-186.

[20] R. Cavazos-Cadena and O. Hernaández-Lerma (1992), "Equivalence of Lyapunov stability criteria in a class of Markov decision processes", *Appl. Math. Optim.* **26**, pp. 113-137.

[21] G. T. De Ghellinck (1960), "Les problèmes de décisions séquentielles", *Cahiers du Centre de Recherche Opérationelle*, **2**, pp. 161-179.

[22] E. V. Denardo (1970a), "On linear programming in a Markov decision problem", *Management Science*, **16**, pp. pp. 281-288.

[23] E. V. Denardo and B. L. Fox (1968), "Multichain Markov renewal programs", *SIAM J. of Applied Math.*, **16**, pp. 468-487.

[24] F. D'Epenoux (1960), "Sur un problème de production et de stockage dans l'aléatoire", *Revue Française de Recherche Opérationelle*, **14**, pp. 3-16.

[25] F. D'Epenoux (1963), "A probabilistic production and inventory problem", *Management Science* **10**, 98-108.

[26] C. Derman (1970), *Finite State Markovian Decision Processes,* Academic Press.

[27] C. Derman and M. Klein (1965), "Some remarks on finite horizon Markovian decision models", *Operations research*, **13**, pp. 272-278.

[28] C. Derman and R. E. Strauch (1966), "On memoryless rules for controlling sequential control processes", *Ann. Math. Stat*$>$ **37**, pp. 276-278.

[29] C. Derman and A. F. Veinott, Jr. (1972), "Constrained Markov decision chains", *Management Science*, **19**, pp. 389-390.

[30] J. L. Doob (1994), *Measure Theory*, Springer-Verlag.

[31] E. Dynkin and A. Yushkevich (1979), *Controlled Markov Processes*, Springer-Verlag, Berlin.

[32] A. Federgruen (1979), "Geometric convergence of value-iteration in multichain Markov decision problems", *Adv. Appl. Prob.* **11**, pp. 188-217.

[33] E. A. Feinberg (1986), "Sufficient Classes of Strategoes in Discrete Dynamic Programming I: Decomposition of Randomized Strategies and Embedded Models", *SIAM Theory Probab. Appl.*, **31**, pp. 658-668.

[34] E. A. Feinberg (1995), "Constrained semi-Markov decision processes with average rewards", *ZOR - Methods and Models in Operations Research*, **39**, pp. 257-288.

[35] E. A. Feinberg and M. I. Reiman (1994), "Optimality of randomized trunk reservation", *Probability in the Engineering and Informational Sciences*, **8**, pp. 463-489.

[36] E. A. Feinberg and I. Sonin (1983), "Stationary and Markov policies in countable state dynamic programming", *Lecture notes in Mathematics*, **1021**, 111-129.

[37] E. A. Feinberg and I. Sonin (1993), "The existence of an equivalent stationary strategy in the case of discount factor equal one", Unpublished Draft.

[38] E. A. Feinberg and I. Sonin (1995), "Notes on equivalent stationary policies in Markov decision processes with total rewards", *ZOR - Methods and Models in Operations Research*, **44**, pp. 205-221.

[39] E. A. Feinberg E. and A. Shwartz (1995), "Constrained discounted dynamic programming", *Math. of Operations Research*, **20**, pp. 302-320.

[40] J. A. Filar and L. C. M Kallenberg and H. M. Lee (1989), "Variance-penalized Markov decision processes", *Math. of Operations Research*, **14**, pp. 147-161.

[41] J. A. Filar and H. M. Lee (1985), "Gain/Variability tradeoffs in undiscounted Markov decision Processes", *Proceedings of 24th Conference on Decision and Control IEEE*, pp. 1106-1112.

[42] L. Fisher and S. M. Ross (1968), "An example in denumerable decision processes", *Ann. Math. Stat.* **39**, pp. 674-675.

[43] W. R. Heilmann (1977), "Generalized linear programming in Markovian decision problems", *Bonner Math. Schriften*, **98**, pp. 33-39.

[44] W. R. Heilmann (1978), "Solving stochastic dynamic programming problems by linear programming – an annotated bibliography", *Z. Oper. Res.*, **22**, pp. 43-53.

[45] O. Hernández-Lerma and D. Hernández-Hernández (1994), "Discounted cost Markov decision processes on Borel spaces: the linear programming formulation", *J. of Math. Anal. and Appl.*, **183**, pp. 335-351.

[46] O. Hernández-Lerma and J. B. Lasserre (1994), "Linear Programming and average optimality on Borel spaces-unbounded costs", *SIAM J. Control and Optimization*, **32** pp. 480-500.

[47] K. Hinderer (1970), *Foundation of Non-Stationary Dynamic Programming with Discrete Time Parameter*, Vol. 33, Lecture Notes in Operations Research and Mathematical Systems, Springer-Verlag, Berlin.

[48] A. Hordijk (1977), *Dynamic Programming and Markov Potential Theory*, Second Edition, Mathematical Centre Tracts 51, Mathematisch Centrum, Amsterdam.

[49] A. Hordijk and L. C. M. Kallenberg (1979), "Linear programming and Markov decision chains", *Management Science*, **25**, pp. 352-362.

[50] A. Hordijk and L. C. M. Kallenberg (1984), "Constrained undiscounted stochastic dynamic programming", *Mathematics of Operations Research,* **9** pp. 276-289.

[51] A. Hordijk and J. B. Lasserre (1994), "Linear programming formulation of MDPs in countable state space: the multichain case", *ZOR - Methods and Models in Operations Research*, **40**, pp. 91-108.

[52] Y. Huang and L. C. M. Kallenberg (1994), "On finding optimal policies for Markov decision chains: A unifying framework for mean-variance tradeoffs", *Math. of Operations Research*, **19**, pp. 434-448.

[53] L. C. M. Kallenberg (1983), *Linear Programming and Finite Markovian Control Problems*, Mathematical Centre Tracts 148, Amsterdam.

[54] L. C. M. Kallenberg (1994), "Survey of linear programming for standard and nonstandard Markovian control problems, Part I: Theory", *ZOR - Methods and Models in Operations Research*, **40**, pp. 1-42.

[55] H. Kawai (1987), "A variance minimization problem for a Markov decision process", *European Journal of operations Research* **31**, pp. 140-145.

[56] N. Krylov (1985), "Once more about the connection between elliptic operators and Ito's stochastic equations", *Statistics and control of stochastic processes* Steklov Seminar 1984 (Krylov N. et al. eds), Optimization Software, New York, 69-101.

[57] H. Kushner and J. Kleinman (1971), "Mathematical programming and the control of Markov chains", *Internat. J. Control*, **13**, pp. 801-820.

[58] J. B. Lasserre (1994), "Average optimal stationary policies and Linear programming in countable state Markov decision processes", *J. Math. Anal. Appl.* **183**, pp. 233-249.

[59] A. S. Manne (1960), "Linear programming and sequential decisions", *Management Science*, **6**, pp. 259-267.

[60] A. B. Piunovskiy (1994), "Control of jump processes with constraints", *Avtomatika i telemekhanika*, **4** pp. 75-89.

[61] M. Puterman (1994), *Markov decision processes*, John Wiley & Sons, New York.

[62] T. E. S. Raghavan and J. A. Filar (1991), "Algorithms for Stochastic Games - A survey", *ZOR - Methods and Models in Operations Research*, **35**, pp. 437-472.

[63] R. T. Rockafellar (1989), *Conjugate Duality and Optimization*, Society for Industrial and Applied Mathematics, 2nd printing, Philadelphia.

[64] K. W. Ross (1989), "Randomized and past-dependent policies for Markov decision processes with multiple constraints", *Operations Research* **37**, pp. 474-477.

[65] K. Ross and R. Varadarajan (1989), 'Markov Decision Processes with Sample path constraints: the communicating case", *Operations Research*, **37**, pp. 780-790.

[66] K. Ross and R. Varadarajan (1991), "Multichain Markov Decision Processes with a Sample Path Constraint: A Decomposition Approach", *Math. of Operations Research*, **16**, pp. 195-207.

[67] H. L. Royden (1988), *Real Analysis*, 3rd Edition, Macmillan publishing Company, New York.

[68] L. S. Shapley (1953), "Stochastic Games", *proccedings Nat. Acad. of Science USA*, **39**, pp. 1095-1100.

[69] L. I. Sennott (1991), "Constrained discounted Markov decision chains", *Probability in the Engineering and Informational Sciences*, **5**, pp. 463-475.

[70] L. I. Sennott (1993), "Constrained average cost Markov decision chains", *Probability in the Engineering and Informational Sciences*, **7**, pp. 69-83.

[71] M. J. Sobel (1985), "Maximal mean/standard deviation ratio in undiscounted MDP", *OR Letters*, **4**, pp. 157-159.

[72] M. J. Sobel (1994), "Mean-variance tradeoffs in an undiscounted MDP", *Operations Research*, **42**, pp. 175-188.

[73] F. M. Spieksma (1990), *Geometrically Ergodic Markov Chains and the Optimal Control of Queues*, Ph.D. thesis, University of Leiden.

[74] M. Tidball and A. Altman (1996), "Continuity of optimal values and solutions of convex optimization, and constrained control of Markov chains", Submitted to *SIAM J. Control and Optimization*.

[75] J. Van Der Wal (1981), *Stochastic Dynamic Programming*, Mathematical Centre Tract 139, Mathematisch Centrum, Amsterdam.

[76] D. J. White (1994), "A mathematical programming approach to a problem in variance penalised Markov decision processes", *OR Spectrum* **15**, pp. 225-230.

[77] D. Williams (1992) *Probability and Martingales*, Cambridge University Press.