CONSTRAINED MARKOV DECISION PROCESSES WITH TOTAL COST CRITERIA: OCCUPATION MEASURES AND PRIMAL LP

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

Submitted: August 1994, Revised: May 1995

Abstract

This paper is the third in a series on constrained Markov decision processes (CMDPs) with a countable state space and unbounded cost. In the previous papers we studied the expected average and the discounted cost. We analyze in this paper the total cost criterion. We study the properties of the set of occupation measures achieved by different classes of policies; we then focus on stationary policies and on mixed deterministic policies and present conditions under which optimal policies exist within these classes. We conclude by introducing an equivalent infinite Linear Program.

Keywords: Constrained Markov Decision Processes, countable state space, infinite horizon, unbounded cost, total cost criterion.

1 Introduction

Constrained Markov decision processes (CMDPs) arise in situations when the controller has more than one objective. A typical situation is when we want to minimize one type of cost while keeping other costs lower than some given bounds. Such problems frequently arise in computer networks and data communications, see Lazar [36], Spieksma and Hordijk [30], Nain and Ross [37], Ross and Chen [40], Altman and Shwartz [5] and Feinberg and Reiman [22]. The theory for solving constrained MDPs was developed by Derman and Klein [18], Derman [17], Derman and Veinott [19], Kallenberg [32], Hordijk and Kallenberg [29], Beutler and Ross [10, 11], Ross and Varadarajan [41], Ross [39], Altman and Shwartz [6, 7], Altman [2, 1], Spieksma [45], Sennott [43, 44], Borkar [15], Feinberg [21], Feinberg and Shwartz [25], Haviv [26], Piunovskiy [38].

This paper is the third in a series on CMDPs with a countable state space and unbounded cost. In [6] we developed the theory for expected average criteria, in [2] we studied the discount cost criteria, and in this paper we analyze the total cost criteria. We consider three types of Markov decision processes (MDPs): the transient MDPs, for which the total expected time spent in each state is finite under any policy, the absorbing MDPs, for which the total expected "life time" of the system is finite under any policy, and contracting MDPs. All three types of MDPs are equivalent for the finite state space, as was shown in [32]; this is however not the case in the countable state space.

We follow a methodology that is similar to the one used in our previous papers. As a first step, we analyze the set of occupation measures achievable by different classes of policies. We establish convexity and compactness properties of these sets. This type of analysis of occupation measure goes back to Derman [17], who also made use of it for studying constrained MDPs (for finite state and action spaces). It was further developed by Kallenberg and Hordijk [29, 32] and Feinberg [21] (who considered the semi-Markov case). The properties of occupation measures corresponding to the infinite state space were investigated by Krylov [34] (who studied controlled diffusion processes), Borkar [13, 14], Altman and Shwartz [2, 6], Spieksma [45] and Feinberg and Sonin [23, 24]. In all these papers, conditions are given for the stationary policies to achieve the same occupation measures as any other policy. For the expected average cost criterion there are cases and counter examples where stationary policies do not achieve all possible occupation measures. This may occur either due to a multi-chain ergodic structure (see Hordijk and Kallenberg [29, 32] for the case of finite state and actions), or, in the infinite case, due to non-tightness (see Borkar [14] Ch. 5, Altman and Shwartz [6], and Speiksma [45]).

In the context of constrained MDPs there is an interest not only in the class of stationary policies, but also in the class of mixed stationary-deterministic policies (where one adds an initial randomization to select a policy within the stationary-deterministic ones, and then uses only that policy). Indeed, it follows from Feinberg [21], who considered the finite state and action spaces, that a linear program can provide an optimal policy within that class.

For contracting and absorbing MDPs, we show that both stationary policies as well as mixed deterministic policies achieve the same occupation measures as those achieved by all policies. For the contracting case, this should not be surprising, since, as it was observed already by Van der Wal ([46] p. 101) that any contracting model can be transformed into an equivalent discounted model, for which the result on stationary policies was obtained by Borkar [13]. Surprisingly, this result turns not to hold for the more general transient MDPs. Indeed, counter-examples have been presented recently by Feinberg and Sonin [24]. However, we show in this paper that the set of stationary policies has the following property. For any occupation measure achievable by some policy u, there is a stationary policy that achieves an occupation measure that is smaller than or equal to that one achieved by u.

In a second step, we relate the cost to the occupation measure, and show that the stationary policies and the mixed deterministic policies are dominant, i.e. their performance is as good as the performance of the set of all policies. This is obtained either in case of nonnegative costs, or for the contracting framework. We further show that if the CMDP is feasible then an optimal stationary policy as well as an optimal mixed stationary-deterministic policy existy.

Finally, we show that the CMDP is equivalent to an infinite Linear Program (LP); i.e. it has the same value, and there is a one to one correspondence between feasible (or optimal) solutions of the LP and of the CMDP.

2 The model

Define the set $\{\mathbf{X}, \mathsf{A}, \mathcal{P}, c, d\}$ where

- $\mathbf{X}$ is a countable state space. Generic notation for states will be x, y, z.
- A is a metric set of actions. We denote by $\mathsf{A}(x)$ the compact set of actions available at state x, and set $\mathcal{K} = \{(x,a) : x \in \mathbf{X}, a \in \mathsf{A}(x)\}$. A generic notation for an action will be a.
- $\mathcal{P}$ are the transition probabilities; thus $\mathcal{P}_{xay}$ is the probability to move from state x to y if action a is chosen.
- $c : \mathcal{K} \to \mathbb{R}$ is an immediate cost. This cost will be related to a cost functional which we shall minimize.
- $d : \mathcal{K} \to \mathbb{R}^K$ is a K-dimensional vector of immediate costs, related to some constraints. 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 following assumption:

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

$$\begin{array}{l}\text{The transition probabilities are continuous, i.e. if } a(n)\to a \\ \text{in } \mathsf{A}(x) \text{ then} \\ \quad \lim_{n\to\infty}\sum_{y\in\mathbf{X}}|\mathcal{P}_{xa(n)y}-\mathcal{P}_{xay}|=0.\end{array} \tag{2}$$

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. 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. ($M_1(G)$ 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 was 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 $\mathsf{A}(x_t)$. The class of all policies defined as above is denoted by U. We introduce the following classes of policies:

- U_M:= Markov policies, for which for any t, u_t is only a function of x_t (and not of the whole history). We may identify a Markov policy with a map $u:\{(x,t)\to M_1(\mathsf{A}(X)), x\in\mathbf{X}, t=1,2,...\}$.
- U_S:= stationary policies, which are subset of U_M; w is stationary if w_t does not depend on t. We shall identify (with some abuse of notation) a stationary policy with a map $w:\{x\to M_1(\mathsf{A}(x)), x\in\mathbf{X}\}$. Under any stationary policy w, the state process becomes a Markov chain with transition probabilities $P_{xy}(w)=\int\mathcal{P}_{xay}w_x(da)$.
- U_D:= stationary deterministic policies, which are subset of U_S; a policy g is stationary deterministic if the action it chooses at state x is a function of x. g is thus identified with a map $g:\mathbf{X}\to\mathsf{A}$.

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. In particular, for any class of policies $G\subset U$, 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 $\mathcal{U}:=\overline{M}(U_D)$.

In the above definition we implicitly 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. We shall sometime include $\mathcal{G}$ in the notation, i.e. denote by $\overline{M}(G,\mathcal{G})$ the class of mixed-G strategies with respect to $\mathcal{G}$, and identify them by all probability measures on $(G,\mathcal{G})$.

For each x, let $\mathcal{B}(\mathsf{A}(x))$ 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 $M_1(\mathsf{A}(x))$).

In order to define a probability space for mixed policies, we first assume that the sets $\mathsf{A}(x)$ are finite for all x. Then, for any time t, the set of histories $\mathbf{H}_t$ is countable, so that the set $\mathbf{H} = \cup_t \mathbf{H}_t$ is countable. Let $x : \mathbf{H} \to \mathbf{X}$ be the projection that assigns $x(h_t) = y$ if $h_t = (x_1, a_1, ..., x_{t-1}, a_{t-1}, x_t)$, and $x_t = y$. U can be identified with all functions which have the countable set $\mathbf{H}$ as range, and a countable product of compact sets $\prod_{h\in\mathbf{H}} M_1(\mathsf{A}(x(h)))$ as image. Therefore, Tychonov's Theorem implies that $\prod_{h\in\mathbf{H}} M_1(\mathsf{A}(x(h)))$ is also convex, compact set in the topology of weak convergence and it is metrizable by virtue of Theorem 4.14 in Royden [42]. Moreover, it is easily seen that the extreme points of U are the pure policies, i.e. those which do not use any randomization at any time (in response to any history). We thus obtained a metric topology for U. Moreover, we can now define the Borel sets $\mathcal{B}_U$ of U, and the σ-algebra $\mathcal{G}_U$ generated by them. They include in particular all singletons. The set of mixed strategies $\overline{M}(U)$ is now identified with the set of probability measures on the space $(U, \mathcal{G}_U)$.

The above topology and σ-field $\mathcal{G}_U$ do not extend to the case when $\mathsf{A}(x)$ are not finite, since the sets $\mathbf{H}_t$ are then not countable. However, we may still use a similar construction for U_M, U_S and U_D. 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. The same considerations as above show that U_S and U_M are 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 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 $\overline{M}(U_D) = \mathcal{U}$.

Finally, for the class of policies U, one can consider the discrete σ-algebra $\mathcal{G}_U^D$ (which is generated by singletons), and define $\overline{M}(U, \mathcal{G}_U^D)$ with respect to that σ-algebra.

Any given distribution β for the initial state (at time 1) and a policy u define 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. The construction of the probability space for $u \in U$ is standard, see e.g. Hinderer [27]. For mixed policies, the construction is done similarly. Moreover, for any mixed policy, the probability space for the state and action processes can be chosen to be the same as the one obtained by some equivalent policy in U. This was established for the more general setting of MDPs with several controllers (stochastic games) by Kuhn [35], Aumann [8] and Bernhard [9].

When β is concentrated on some state x (i.e. $\beta = \delta_x$), we shall use the notation P_x^u instead of P_β^u. Denote $p_\beta^u(t;x) = P_\beta^u(X_t = x)$ and $p_\beta^u(t;x,\mathcal{A}) = P_\beta^u(X_t = x, A_t \in \mathcal{A})$, $\mathcal{A} \subset \mathsf{A}(x)$. We have for all $\beta \in M_1(\mathbf{X})$ and policies u,

$$p_\beta^u(t;x) = p_\beta^u(t;x,\mathsf{A}(x)),$$

and for $t > 1$,

$$p_\beta^u(t;x) = \sum_y \int_{\mathsf{A}(y)} p_\beta^u(t-1;y,da)\mathcal{P}_{yax} = \int_{\mathcal{K}} p_\beta^u(t-1;d\kappa)\mathcal{P}_{\kappa x}. \tag{3}$$

Next, we define the cost criteria that will appear in the constrained control problem. For any policy u and initial distribution β, the total cost is defined as

$$C_{tc}(\beta,u) = \sum_{t=1}^{\infty} E_\beta^u c(X_t, A_t). \tag{4}$$

The costs $D^k(\beta,u)$ related to the immediate costs d^k are defined similarly. 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. $q_1(j) \leq q_2(j)$, $j = 1, ..., K$. The set of policies satisfying the constraints are called feasible. Let $C(\beta)$ be the value of the above problem. (If the feasible set of policies is empty then we set $C(\beta) = \infty$). If a feasible policy u^* achieves the minimum, i.e. $C(\beta) = C_{tc}(\beta, u^*)$ then it is called optimal.

Definition 2.1 *(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*

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

We shall need the following theorem which shows that the Markov policies are dominating. It is based on a theorem by Derman and Strauch [20], extended by Hordijk [28]. It further implies the convexity of the set of some occupation measures, which will be defined later.

Theorem 2.1 *(Sufficiency of Markov policies)*
(i) Choose any initial distribution β, and any $\gamma \in M_1(U_M)$. Let $\hat{\gamma}$ be the corresponding policy in $\overline{M}(U_M)$. Then there exists some $v \in U_M$ such that for all t,

$$p_\beta^{\hat{\gamma}}(t; \cdot, \cdot) = p_\beta^v(t; \cdot, \cdot) \tag{5}$$

(ii) Choose any initial distribution β, and a distribution γ over U with a discrete support, i.e. $\gamma \in M_1(U, \mathcal{G}_U^D)$. Let $\hat{\gamma}$ be the corresponding policy in $\overline{M}(U, \mathcal{G}_U^D)$. Then there exists some $v \in U_M$ such that for all t, (5) holds.

Proof: (i) We write $p_\beta^{\hat{\gamma}}$ in an integral form:

$$p_\beta^{\hat{\gamma}}(t; \cdot, \cdot) = \int_{U_M} \gamma(du) P_\beta^u(X_t = \cdot, A_t \in \cdot).$$

Define v to be the Markov policy given by

$$v_t(\mathcal{A}|x) := \frac{\int \gamma(du) P_\beta^u(X_t = x, A_t \in \mathcal{A})}{\int \gamma(du) P_\beta^u(X_t = x)} \tag{6}$$

for all integers t, states x and $\mathcal{A} \subset \mathsf{A}(x)$, for which the denominator is nonzero. When it is zero, define $v_t(\cdot|x)$ to be an arbitrary probability measure over $\mathsf{A}(x)$. The proof follows by induction. (5) clearly holds for $t = 1$, since for any policy $u \in U_M$,

$$P_\beta^u(X_1 = x, A_1 \in \mathcal{A}) = \beta(x) u_1(\mathcal{A}|x),$$

and $\int \gamma(du) P(X_1 = x) = \beta(x)$; this implies

$$P_\beta^v(X_1 = x, A_1 \in \mathcal{A}) = \beta(x) v_1(\mathcal{A}|x) = \int \gamma(du) P_\beta^u(X_1 = x, A_1 \in \mathcal{A}).$$

Assume that (5) holds for some t, i.e.

$$\int \gamma(du) P_\beta^u(X_t = x, A_t \in \mathcal{A}) = P_\beta^v(X_t = x, A_t \in \mathcal{A}) = [\beta P(v_1) P(v_2) ... P(v_{t-1})]_x v_t(\mathcal{A}|x). \quad (7)$$

We show first that

$$\int \gamma(du) P_\beta^u(X_{t+1} = x) = [\beta P(v_1) P(v_2) ... P(v_t)]_x. \quad (8)$$

Since

$$P_\beta^u(X_{t+1} = x | X_t = y, A_t = a) = \mathcal{P}_{yax}, \qquad P_\beta^u - a.s.$$

for all $u \in U_M$, we obtain by conditioning on X_t, A_t and by (7), that the left-hand side of (8) equals

$$\sum_{y \in \mathbf{X}} [\beta P(v_1) P(v_2) ... P(v_{t-1})]_y \int_{\mathsf{A}(y)} \mathcal{P}_{yax} v_t(da|x).$$

This implies (8). Combining now (8) with (6), we get

$$\begin{aligned} \int \gamma(du) P_\beta^u(X_{t+1} = x, A_{t+1} \in \mathcal{A}) &= v_t(\mathcal{A}|x) \int \gamma(du) P_\beta^u(X_{t+1} = x) \\ = \quad & [\beta P(v_1) P(v_2) ... P(v_t)]_x v_t(\mathcal{A}|x) = P_\beta^v(X_{t+1} = x, A_{t+1} \in \mathcal{A}). \end{aligned}$$

This concludes the proof of (i). (ii) is obtained by the same arguments, see Derman and Strauch [20], Hordijk [28]. ■

Remark 2.1 *(The converse)*
An interesting question is whether some converse exists, i.e. whether we can describe any Markov policy (which uses randomizations) as a mixture of some policies within some class of simpler policies. The answer is positive, and that class can be taken as the class of pure Markov policies (which do not use randomizations). This was established by Kadelka [31].

3 Transient and Absorbing MDPs

Definition 3.1 *Fix an initial distribution β. A policy u is said to be $\mathbf{X}'$-transient where $\mathbf{X}' \subset \mathbf{X}$, if*

$$\sum_{t=1}^{\infty} p_\beta^u(t;x) < \infty \text{ for any } x \in \mathbf{X}'.$$

It is called $\mathbf{X}'$-absorbing if

$$\sum_{t=1}^{\infty} p_\beta^u(t;\mathbf{X}') < \infty.$$

Definition 3.2 *An MDP for which all policies are $\mathbf{X}'$-transient ($\mathbf{X}'$-absorbing) is called a $\mathbf{X}'$-transient ($\mathbf{X}'$-absorbing, respectively) MDP. A $\mathbf{X}$-transient MDP is called a transient MDP.*

Here are some properties of transient stationary policies.

Lemma 3.1 *(Stationary policies in transient and absorbing MDPs)*
Fix some initial distribution β on $\mathbf{X}'$ and a stationary $\mathbf{X}'$-transient policy w. Then
(i) $f^*(x) := \sum_{t=1}^{\infty} p_\beta^w(t;x)$ is the minimal solution to

$$f = \beta + fP(w), \qquad f \geq 0. \tag{9}$$

(where f and β are row vectors on $\mathbf{X}'$, and $P(w)$ is the restriction to $\mathbf{X}'$ of the transition probability matrix of the Markov chain corresponding to the stationary policy w).
(ii) Assume that the MDP is $\mathbf{X}'$-absorbing. Fix some policy u and define $g^*(x) := \sum_{t=1}^{\infty} p_\beta^u(t;x)$. If g^* satisfies (9), then $g^*(x) = f^*(x)$ for all $x \in \mathbf{X}'$.

Proof: (i) It follows easily that f^* is indeed a solution of (9). Iterating (9) we get for all integers n:

$$\begin{aligned} f &= \beta + (\beta + fP(w))P(w) = \beta + \beta P(w) + f[P(w)]^2 \\ &= \sum_{i=1}^{n-1} \beta[P(w)]^i + f[P(w)]^n = \sum_{t=1}^{n-1} p_\beta^w(t) + f[P(w)]^n \end{aligned} \tag{10}$$

(i) follows since the above holds for all n and since $f \geq 0$.
(ii) follows since $g^*[P(w)]^n$ converges to zero. Indeed, define $\mathbf{1} : \mathbf{X}' \to \mathbb{R}$ to be the function

whose entries are all 1. Since w is absorbing, $\langle g^*, \mathbf{1}\rangle < \infty$. (Here, and throughout, we use the notation $\langle q_1, q_2\rangle$ between two vectors to denote the scalar product.) For any integer n and $y \in \mathbf{X}$, the yth column of $[P(w)]^n$ is bounded by $\mathbf{1}$, so by the the generalized dominance convergence theorem ([42], Proposition 11.18),

$$\lim_{n\to\infty} g^*[P(w)]^n = g^*\left(\lim_{n\to\infty}[P(w)]^n\right) = 0. \tag{11}$$

To get the last equality, it suffices to show the following: let y be some state for which $g^*(y) > 0$. Then the yth row of the matrix $P^\infty = \overline{\lim}_{n\to\infty}[P(w)]^n$ is zero. Assume that for some z, $P^\infty_{yz} \neq 0$. There exists some time t for which $p^u_\beta(t; y) > 0$. Consider a policy v that behaves like policy u till time t and then behaves like the stationary policy w. Then

$$\sum_{s=1}^{\infty} p^v_\beta(s; z) \geq p^u_\beta(t; y) \sum_{n=0}^{\infty} [P^n(w)]_{yz} = \infty$$

This contradicts the fact that the MDP is absorbing. This shows that $P^\infty_{yz} = 0$ for all z, from which (11) follows. The proof now follows taking the limit as $n \to \infty$ in (10). ■

We study in this paper the total cost criterion for $\mathbf{X}'$-transient MDPs. Let $\mathcal{M}$ be the complement of $\mathbf{X}'$ in $\mathbf{X}$. We assume throughout that $c(x, a) = d^k(x, a) = 0, k = 1, ..., K$ for any $x \in \mathcal{M}$, and that the initial distribution β has zero mass on $\mathcal{M}$ ($\mathcal{M}$ may be empty). The total cost criterion has the meaning of the total expected cost till the set $\mathcal{M}$ is reached. The theory developed below may be applied for studying the constrained optimal control problems with finite horizon cost and with discounted cost (with finite or infinite horizon, see [4]).

4 Contracting MDPs

Following Dekker and Hordijk [16] and Spieksma [45], we introduce the following μ-norm. For any functions $q : \mathbf{X} \to \mathbb{R}$, $Q : \mathbf{X} \times \mathbf{X} \to \mathbb{R}$, and $\mu : \mathbf{X} \to [1, \infty)$, we define

$$\|q\|_\mu = \sup_{x\in\mathbf{X}} \frac{q(x)}{\mu(x)}, \qquad \|Q\|_\mu = \sup_{x\in\mathbf{X}} \frac{\sum_{y\in\mathbf{X}} Q_{xy}\mu(y)}{\mu(x)}. \tag{12}$$

It is easily verified that μ is indeed a norm. In particular, it satisfies $\|Qq\|_\mu \leq \|Q\|_\mu \|q\|_\mu$, and for $Q^1, Q^2 : \mathbf{X}\times\mathbf{X} \to \mathbb{R}$ we have $\|Q^1Q^2\|_\mu \leq \|Q^1\|_\mu \|Q^2\|_\mu$. 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 norms, and M^μ to be the set of measures $\mathsf{M}^\mu := \{q \in M(\mathbf{X}) : E^q\mu < \infty\}$ (here, $E^q\mu := \sum_x q(x)\mu(x)$).

With some abuse of notation, we shall say that a function $f : \mathcal{K} \to \mathbb{R}$ is in F^μ if the function defined on $\mathbf{X}$ whose x entry is $\sup_{a\in\mathsf{A}(x)} f(x,a)$, is in F^μ. Similarly, a measure q on $\mathcal{K}$ is said to be in M^μ if the measure $\overline{q}$ is in M^μ, where $\overline{q}(x) := q(x, \mathsf{A}(x))$.

Definition 4.1 *(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 some scalar $\xi \in [0,1)$, a vector $\mu : \mathbf{X} \to [1,\infty)$, such that for all $x \in \mathbf{X}, a \in \mathsf{A}(x)$,

$$\sum_{y\notin\mathcal{M}} \mathcal{P}_{xay}\mu(y) \le \xi\mu(x). \tag{13}$$

When using contracting MDPs, we shall make the following assumptions on the initial distribution, the transition probabilities and the costs:

- *$\langle \beta, \mu \rangle < \infty$;*
- *The transition probabilities are μ-continuous, i.e. if $a(n) \to a$ in $\mathsf{A}(x)$ then*

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

- *$c(x,a)$ and $d^k(x,a), k = 1, ..., K$ are μ-bounded, i.e. $\exists \overline{b} < \infty$ s.t.*

$$\sup_{u\in U_D} \|c(\cdot,u)\|_\mu < \overline{b} \text{ and } \sup_{u\in U_D,k} \left\|d^k(\cdot,u)\right\|_\mu < \overline{b}$$

An alternative way to write (13) is by introducing the taboo probabilities. We define for any $Q : \mathbf{X}\times\mathbf{X} \to \mathbb{R}$

$$_{\mathcal{M}}Q_{xy} = \begin{cases} Q_{xy} & \text{if } y \notin \mathcal{M}, \\ 0 & \text{otherwize .} \end{cases} \tag{14}$$

We further define

$$_{\mathcal{M}}p^u_x(t;x) := P^u_\beta(X_t = x, X_s \notin \mathcal{M}, s = 1, ..., t).$$

(13) can be rewritten as

$$\sup_{w\in U_D} \|_{\mathcal{M}}P(w)\|_\mu \le \xi.$$

We now show that the contracting framework implies that the MDP is $\mathbf{X}'$-absorbing where $\mathbf{X}' = \mathbf{X} \setminus \mathcal{M}$, for suitable initial distributions.

Lemma 4.1 *(Rate of convergence)*
Consider the contracting MDP. The μ-norm of ${}_{\mathcal{M}}p^u_{(\cdot)}(t)$ converges to 0 in a geometric rate, uniformly over all $u \in U$, i.e. for any $x \in \mathbf{X}'$

$$ {}_{\mathcal{M}}p^u_x(t; \mathbf{X}') \leq \sum_{y \in \mathbf{X}'} {}_{\mathcal{M}}p^u_x(t; y)\mu(y) \leq \mu(x)\xi^{t-1}. $$

Moreover,

$$ {}_{\mathcal{M}}p^u_\beta(t; \mathbf{X}') \leq \sum_{y \in \mathbf{X}'} {}_{\mathcal{M}}p^u_\beta(t; y)\mu(y) \leq \langle \beta, \mu \rangle \xi^{t-1}; $$

and

$$ \sum_{t=1}^{\infty} {}_{\mathcal{M}}p^u_\beta(t; \mathbf{X}') \leq \sum_{t=1}^{\infty} \sum_{y \in \mathbf{X}'} {}_{\mathcal{M}}p^u_\beta(t; y)\mu(y) \leq \frac{\langle \beta, \mu \rangle}{1 - \xi}, $$

which implies that the MDP is $\mathbf{X}'$-absorbing.

Proof: Choose any $u \in U$ and let $v = v(u)$ be the corresponding Markov policy given in Remark 2.1. Viewing ${}_{\mathcal{M}}p^u_{(\cdot)}(t; \cdot)$ as a matrix, we have

$$ \| {}_{\mathcal{M}}p^u_{\cdot}(t; \cdot) \|_\mu \leq \| {}_{\mathcal{M}}P(v_1) \|_\mu \, \| {}_{\mathcal{M}}P(v_2) \|_\mu \cdots \| {}_{\mathcal{M}}P(v_{t-1}) \|_\mu \leq \xi^{t-1}, $$

which implies

$$ {}_{\mathcal{M}}p^u_x(t; \mathbf{X}') \leq \sum_{y \in \mathbf{X}'} p^u_x(t; y)\mu(y) \leq \mu(x)\xi^{t-1}. $$

This concludes the proof. ■

Hence, contracting MDPs are a subclass of absorbing MDPs, which are subclass of transient MDPs. The converse needs not hold; if $\mathbf{X} = \mathbb{N}$ and $P_{n,n+1}(w) = 1$ for some $w \in U_S$, then w is transient but nonabsorbing. In the case of finite state and action spaces, Kallenberg showed in [32] that all the above MDPs are equivalent.

Lemma 3.1 can be strengthen:

Lemma 4.2 *(Uniqueness of a bounded solution)*

Consider a contracting MDP. Fix a stationary transient policy w on $\mathbf{X}'$. Fix some initial distribution β such that $\langle \beta, \mu \rangle < \infty$. Then $f(x) = \sum_{t=1}^{\infty} {}_{\mathcal{M}}p_{\beta}^{w}(t;x)$ is the unique μ-bounded solution of

$$f = \beta + fP(w). \tag{15}$$

Proof: Let f' be some μ-bounded solution of (15). Iterating (15), we get:

$$\begin{aligned} f'(y) &= \beta(y) + {}_{\mathcal{M}}p_{\beta}^{w}(2;y) + \sum_{x \in \mathbf{X}'} f'(x)[P^2(w)]_{xy} \\ &= \beta(x) + {}_{\mathcal{M}}p_{\beta}^{w}(2;x) + {}_{\mathcal{M}}p_{\beta}^{w}(3;x) + \sum_{x \in \mathbf{X}'} f'(x)[P^3(w)]_{xy} \\ &= \ldots = \beta(x) + {}_{\mathcal{M}}p_{\beta}^{w}(2;x) + \ldots + {}_{\mathcal{M}}p_{\beta}^{w}(t;x) + \sum_{x \in \mathbf{X}'} f'(x)[P^t(w)]_{xy} \end{aligned} \tag{16}$$

We have (as in Lemma 4.1)

$$\left| \sum_{x \in \mathbf{X}'} f'(x)[P^t(w)]_{xy} \right| \leq \mu(y) \left\| f' \right\|_{\mu} \left\| {}_{\mathcal{M}}P^t(w) \right\|_{\mu} \leq \mu(y) \left\| f' \right\|_{\mu} \xi^t \to 0$$

The proof is established by taking the limit as $t \to \infty$ in (16). ■

5 Occupation measure

For any given initial distribution β and policy u, define the occupation measure $f_{tc}(\beta, u; x, \cdot)$ related to the total cost criterion by

$$f_{tc}(\beta, u; x, \mathcal{A}) = \sum_{t=1}^{\infty} {}_{\mathcal{M}}p_{\beta}^{u}(t; x, \mathcal{A}), \qquad \mathcal{A} \subset \mathsf{A}(x).$$

With some abuse of notation, we denote $f_{tc}(\beta, u; x) = f_{tc}(\beta, u; x, \mathsf{A}(x))$. Let

$$f_{tc}(\beta, u) := \{ f_{tc}(\beta, u; x, \cdot), x \in \mathbf{X}' \}.$$

Define $\mathcal{K}' := \{(x,a) : x \in \mathbf{X}', a \in \mathsf{A}(x)\}$, and

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

$$\mathbf{Q}_{tc}(\beta) \quad = \quad \left\{ \begin{array}{c} \rho \in M(\mathcal{K}') : \displaystyle\sum_{y \in \mathbf{X}'} \int_{\mathcal{A}(y)} \rho(y,da)(\delta_x(y) - \mathcal{P}_{yax}) = \beta(x), x \in \mathbf{X}' \\ \\ \rho(x, \mathsf{A}(x)) = 0 \text{ for } x \in \mathcal{M}, \quad \rho(x,\mathsf{A}(x)) < \infty \text{ for } x \notin \mathcal{M}, \end{array} \right\} \tag{17}$$

where $M(\mathcal{K})$ is the set of nonnegative measures over $\mathcal{K}$. We set $\mathbf{L}(\beta) = \mathbf{L}_U(\beta) \cup \mathbf{L}_{\overline{M}(U_M)}(\beta)$. For the contracting framework we define

$$\mathbf{Q}^{\mu}_{tc}(\beta) := \mathbf{Q}_{tc}(\beta) \cap \mathsf{M}^{\mu}.$$

For any sets B, B_1, B_2 in $M(\mathcal{K})$, define

- coB_1 := the convex hull of a set;
- $\min B$:= the set of minimal elements in B, i.e. $\rho \in \min B$ if there does not exist some $\rho' \in B$ such that $\rho'(y,\mathcal{A}) < \rho(y,\mathcal{A})$ for some $y \in \mathbf{X}$ and $\mathcal{A} \subset \mathsf{A}(y)$;
- $B_1 \prec B_2$ if $\forall \rho_2 \in B_2$ there exists $\rho_1 \in B_1$ such that $\rho_1 \leq \rho_2$.

Definition 5.1 *A class of policies $\overline{U}$ is called complete for the total cost criterion (for a given initial distribution β) if $\mathbf{L}_{\overline{U}}(\beta) = \mathbf{L}(\beta)$. It is called weakly complete if $\mathbf{L}_{\overline{U}}(\beta) \prec \mathbf{L}(\beta)$.*

Theorem 5.1 *(Completeness of stationary policies)*
(i) Consider transient MDPs. Then the set of stationary policies is weakly complete.
(ii) If the MDP $\mathbf{X}'$-absorbing, then the set of stationary policies is complete.

Proof: Choose a policy $u \in U$, and let w be a stationary policy satisfying

$$w_y(\mathcal{A}) = \frac{f_{tc}(\beta,u;y,\mathcal{A})}{f_{tc}(\beta,u;y)}, \ y \in \mathbf{X}', \mathcal{A} \subset \mathsf{A}(y)$$

whenever the denominator is nonzero. (When it is zero, $w_y(\cdot)$ is chosen arbitrarily). We show that $f_{tc}(\beta, w) = f_{tc}(\beta, u)$. For any $x \in \mathbf{X}$,

$$\begin{aligned} f_{tc}(\beta, u; x) &= \beta(x) + \sum_{t=2}^{\infty} p_\beta^u(t, x) \\ &= \beta(x) + \sum_{t=2}^{\infty} \int_{\mathcal{K}} p_\beta^u(t-1; d\kappa) \mathcal{P}_{\kappa x} \\ &= \beta(x) + \int_{\mathcal{K}} f_{tc}(\beta, u; d\kappa) \mathcal{P}_{\kappa x} \qquad (18) \\ &= \beta(x) + \sum_{y \in \mathbf{X}} f_{tc}(\beta, u; y) \int_{\mathsf{A}(y)} \mathcal{P}_{yax} w_y(da) \\ &= \beta(x) + \sum_{y \in \mathbf{X}} f_{tc}(\beta, u; y) P_{yx}(w) \qquad (19) \end{aligned}$$

Hence, by Lemma 3.1 (i), $f_{tc}(\beta, w; x) \leq f_{tc}(\beta, u; x)$ for all $x \in \mathbf{X}$. This implies by the definition of w that $f_{tc}(\beta, w) \leq f_{tc}(\beta, u)$, so that the set of stationary policies is weakly complete.

(ii) Follows from Lemma 3.1 (ii) and (19). ■

Definition 5.2 *(μ-continuity)*
Consider some $\overline{U} \subset U_M$ and $Q : \overline{U} \times \mathbf{X}$. Q is said to be μ-continuous on $\overline{U}$ if for any converging sequence $u(n) \in \overline{U}$ with limit $u \in \overline{U}$

$$\lim_{n \to \infty} \sum_{y \in \mathbf{X}} |Q(u(n), y) - Q(u, y)| \mu(y) = 0.$$

Lemma 5.1 *(Continuity properties of f_{tc})*
(i) For transient MDPs, the map $f_{tc}(\beta, \cdot) : U_M \to \mathbf{L}_{U_M}$ is lower semi-continuous. The same holds for the map $f_{tc}(\beta, \cdot) : \overline{M}(U_M) \to \mathbf{L}_{U_M}$.
(ii) Consider a contracting MDP. Then the map $f_{tc}(\beta, \cdot) : U_M \to \mathbf{L}_{U_M}$ is μ-continuous. The same holds for the map $f_{tc}(\beta, \cdot) : \overline{M}(U_M) \to \mathbf{L}_{U_M}$.

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^n). Then $P_{xy}(u_t^n) \to P_{xy}(u)$ 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 $x, y \in \mathbf{X}'$. Moreover, the m step probabilities also converge, i.e. for all integers m:

$$\begin{aligned} \lim_{n\to\infty} p_\beta^{u^n}(m; y, \mathcal{A}) &= \lim_{n\to\infty} \sum_{x\in\mathbf{X}'} \beta(x)[P(u_1^n)P(u_2^n)...P(u_m^n)]_{xy} u_m^n(\mathcal{A}|y) \\ &= \sum_{x\in\mathbf{X}'} \beta(x)[P(u_1)P(u_2)...P(u_m)]_{xy} u_m(\mathcal{A}|y) = p_\beta^u(m; y, \mathcal{A}), \end{aligned} \tag{20}$$

for all $y \in \mathbf{X}'$, $\mathcal{A} \subset \mathsf{A}$. (20) is established by induction. It holds for $m = 1$. Assume it holds for arbitrary m. Consider the probability measures over $\mathbf{X}'$: $\nu(n) := p_\beta^{u^n}(m; \cdot)$ and $\nu := p_\beta^u(m; \cdot)$, and let $q_y(n)$ and q_y be the y column of $P(u_{m+1}^n)$ and $P(u_{m+1})$, respectively. Then,

$$p_\beta^{u^n}(m+1; y) = \int_{\mathbf{X}'} q_y(n) d\nu(n), \qquad p_\beta^u(m+1; y) = \int_{\mathbf{X}'} q_y d\nu.$$

The entries of q_y are bounded by 1, so by applying the generalized dominance convergence theorem ([42], Proposition 11.18) we get $\lim_{n\to\infty} p_\beta^{u^n}(m+1; y) = p_\beta^u(m+1; y)$, from which (20) follows.

Now, we fix some $y \in \mathbf{X}', \mathcal{A} \in \mathsf{A}$, and consider $p_\beta^{u^n}(m; y, \mathcal{A})$ to be a function over the integers m. This function converges pointwize to $p_\beta^u(m; y, \mathcal{A})$. Applying Fatou's Lemma ([42] Proposition 11.17) with respect to the (infinite) measure over the integers $\mu_n(m) = \mu(m) = 1$, we obtain

$$\underline{\lim}_{n\to\infty} f_{tc}(\beta, u^n; y, \mathcal{A}) \geq f_{tc}(\beta, u; y, \mathcal{A}),$$

which concludes the proof of (i) for the lower semi-continuity in 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 lower semi-continuity in in $\overline{M}(U_M)$ is established by applying again Fatou's Lemma:

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

(ii) Assume again that $u^n \to u$, where $u^n, u \in U_M$. Then by assumption,

$$\lim_{n\to\infty} \sum_{y\in\mathbf{X}'} |P_{xy}(u_1^n) - P_{xy}(u)|\mu(y) = 0$$

for all $x \in \mathbf{X}'$. Since $\sum_{y \in \mathbf{X}'} |P_{xy}(u_1^n) - P_{xy}(u)|\mu(y) \leq \mu(x)$ and $\langle \beta, \mu \rangle < \infty$, it follows from the bounded convergence theorem, that

$$\lim_{n\to\infty} \sum_{x \in \mathbf{X}'} \beta(x) \sum_{y \in \mathbf{X}'} |P_{xy}(u_1^n) - P_{xy}(y)|\mu(y) = \lim_{n\to\infty} \sum_{y \in \mathbf{X}'} |p_\beta^{u^n}(1;y) - p_\beta^u(1;y)|\mu(y) = 0. \quad (21)$$

Moreover, the m step probabilities also converge, i.e. for all integers m:

$$\lim_{n\to\infty} \sum_{y \in \mathbf{X}'} |p_\beta^{u^n}(m;y) - p_\beta^u(m;y)|\mu(y) = 0. \quad (22)$$

This, again is established by induction. It holds for $m = 1$. Assume it holds for arbitrary m.

$$\begin{aligned}
&\sum_{y \in \mathbf{X}'} |p_\beta^{u^n}(m+1;y) - p_\beta^u(m+1;y)|\mu(y) \\
&= \sum_{y,z \in \mathbf{X}'} |p_\beta^{u^n}(m;z)P_{zy}(u_m^n) - p_\beta^u(m;z)P_{zy}(u_m)|\mu(y) \\
&\leq \sum_{y,z \in \mathbf{X}'} \left(p_\beta^{u^n}(m;z)|P_{zy}(u_m^n) - P_{zy}(u_m)| + |p_\beta^{u^n}(m;z) - p_\beta^u(m;z)|P_{zy}(u_m)\right)\mu(y) \\
&\leq \sum_{y,z \in \mathbf{X}'} p_\beta^{u^n}(m;z)|P_{zy}(u_m^n) - P_{zy}(u_m)|\mu(y) + \sum_{z \in \mathbf{X}'} |p_\beta^{u^n}(m;z) - p_\beta^u(m;z)|\xi\mu(z).
\end{aligned}$$

The first summation tends to zero as $n \to \infty$ by the same argument as in (21), since

$$\langle p_\beta^{u^n}(m;\cdot), \mu \rangle < \infty$$

by Lemma 4.1. The second summation converges to zero by the induction hypothesis. We conclude that for every m, (22) holds, and consequently $f^T(\beta, u)$ are μ-continuous on U_M. Next, we observe that for any $u \in U_M$,

$$\sum_{m=T+1}^{\infty} \sum_{y \in \mathbf{X}} p_\beta^u(m;y)\mu(y) \leq \langle \beta, \mu \rangle \sum_{m=T}^{\infty} \xi^m$$

by Lemma 4.1. We conclude that

$$\begin{aligned}
&\sum_{y \in \mathbf{X}'} |f_{tc}(\beta, u^n) - f_{tc}(\beta, u)|\mu(y) \\
&\quad \leq \sum_{y \in \mathbf{X}'} |f^T(\beta, u^n; y) - f^T(\beta, u; y)|\mu(y) + 2\langle \beta, \mu \rangle \sum_{m=T}^{\infty} \xi^m.
\end{aligned}$$

Since this holds for every T, and since for every T

$$\lim_{T\to\infty} \sum_{y\in\mathbf{X}'} |f^T(\beta,u^n;y) - f^T(\beta,u;y)|\mu(y) = 0,$$

we conclude that $f_{tc}(\beta,\cdot)$ is μ-continuous on 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 (Billingseley [12]) implies:

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

■

Lemma 5.2 *(Splitting in a state)*
Choose $w \in U_S$ and a state y. Define $w^a \in U_S$ to be the policy that chooses always action a when in state y, and otherwise behaves exactly like w. Then, there exists a probability measure γ over $\mathsf{A}(y)$ such that

$$f_{tc}(\beta,w) = \int_{\mathsf{A}(y)} \gamma(da) f_{tc}(\beta,w^a).$$

Proof: Define the stopping times $\sigma(y) \stackrel{\text{def}}{=} \inf_{r>1}\{X_r = y\}$, $y \in \mathbf{X}$, with the convention that $\inf\{\emptyset\} = \infty$. Define the total expected number of visits to state y starting from state x as:

$$W(u;x,y) = E_x^u\left(\sum_{t=2}^{\sigma(y)} 1(X_t, A_t)\right)$$

and the probability of ever reaching state y from state x:

$$\overline{p}(u;x,y) := P_x^u(\sigma(y) < \infty).$$

Define γ in the following way: for any $\mathcal{A} \subset \mathsf{A}(y)$,

$$\gamma(\mathcal{A}) \stackrel{\text{def}}{=} \int_{\mathcal{A}} \frac{w_y(da)(1-\overline{p}(w^a;y,y))}{1-\overline{p}(w;y,y)}$$

It follows from standard properties of Markov chains (see [33] Corollary 4-20) that

$$f_{tc}(x,w;y) = W(w;x,y) + \overline{p}(u;x,y) f_{tc}(y,w;y)$$

By setting $x = y$ we get

$$f_{tc}(y,w;y) = \frac{W(w;y,y)}{1-\overline{p}(w;y,y)} = \frac{\int w_y(da) W(w^a;y,y)}{1-\overline{p}(w;y,y)}$$

$$= \int \frac{w_y(da)(1-\overline{p}(w^a;y,y))}{1-\overline{p}(w;y,y)} f_{tc}(y,w^a;y) = \int_{\mathsf{A}(y)} \gamma(da) f_{tc}(y,w^a;y),$$

which establishes the proof for the case $x = y$. For general x,

$$\begin{aligned} f_{tc}(x,w;y) &= W(w;x,y) + \overline{p}(u;x,y) f_{tc}(y,w;y) \\ &= \int_{\mathsf{A}(y)} \gamma(da)[W(w;x,y) + \overline{p}(u;x,y) f_{tc}(y,w^a;y)] \\ &= \int_{\mathsf{A}(y)} \gamma(da) f_{tc}(x,w^a;y). \end{aligned}$$

■

Theorem 5.2 *(Characterization of the sets of occupation measure)*
(i) For transient MDPs, $\mathbf{L}(\beta)$ *is convex, and*

$$\min \mathbf{Q}_{tc}(\beta) = \mathbf{L}_{U_S}(\beta) \prec \mathbf{L}_{U_M}(\beta) = \mathbf{L}(\beta) \subset \mathbf{Q}_{tc}(\beta).$$

(ii) For absorbing MDPs, $\mathbf{L}_{U_S}(\beta)$ *is convex and compact, and satisfies*

$$\mathbf{L}_{\mathcal{U}}(\beta) = \mathbf{L}(\beta) = \mathbf{L}_{U_S}(\beta) = co\mathbf{L}_{U_D}(\beta) = \min \mathbf{Q}_{tc}(\beta).$$

(iii) For contracting MDPs, $\mathbf{L}_{U_S}(\beta)$ *is convex and compact, and satisfies*

$$\mathbf{L}_{\mathcal{U}}(\beta) = \mathbf{L}(\beta) = \mathbf{L}_{U_S}(\beta) = co\mathbf{L}_{U_D}(\beta) = \mathbf{Q}^{\mu}_{tc}(\beta) = \min \mathbf{Q}_{tc}(\beta).$$

Proof: (i) Theorem 2.1 implies that $\mathbf{L}(\beta) = \mathbf{L}_{U_M}(\beta)$ is convex. The weakly completeness of $\mathbf{L}_{U_S}(\beta)$ was established in Theorem 5.1. That $\mathbf{L}(\beta) \subset \mathbf{Q}_{tc}(\beta)$ follows from (19). Finally, we show that $\mathbf{L}_{U_S}(\beta) = \min \mathbf{Q}_{tc}(\beta)$. For any $\rho \in \mathbf{Q}_{tc}(\beta)$, define $w(\rho)$ to be any stationary policy such that $w_y(\mathcal{A}) = \rho(y, \mathcal{A})[\rho(y, \mathsf{A}(y))]^{-1}$ whenever the denominator is nonzero. We have

$$\begin{aligned}
\rho(x, \mathsf{A}(x)) &= \beta(x) + \int_{\mathcal{K}} \rho(d\kappa)\mathcal{P}_{\kappa x} \\
&= \beta(x) + \sum_y \rho(y, \mathsf{A}(y)) \int_{\mathsf{A}(y)} \mathcal{P}_{yax} w_y(da) \\
&= \beta(x) + \sum_y \rho(y, \mathsf{A}(y)) P_{yx}(w). \qquad (23)
\end{aligned}$$

By Lemma 3.1 (i) we conclude that $f_{tc}(\beta, w(\rho); x) \le \rho(x, \mathsf{A}(x))$ for all $x \in \mathbf{X}$. By the definition of $w(\rho)$, it follows that $f_{tc}(\beta, w(\rho)) \le \rho$.

(ii) That $\mathbf{L}(\beta) = \mathbf{L}_{U_S}(\beta)$ follows from Theorem 5.1 (ii), hence $\mathbf{L}_{U_S}(\beta)$ is convex. The compactness of $\mathbf{L}_{U_M}(\beta)$ follows since by Lemma 5.1 it is the image of the compact set U_M under the continuous function $f_{tc}(\beta, \cdot)$. We show that $\mathbf{L}_{U_S}(\beta)$ is equal to the convex hull of $\mathbf{L}_{U_D}(\beta)$ (and thus of $\mathbf{L}_{\mathcal{U}}(\beta)$). Since it is compact, by the Krein-Milman theorem it is the convex hull of its extreme points. Choose some $w \in U_S$. Suppose that w is not deterministic, so that $w_y(\cdot)$ is not concentrated on a single point in $\mathsf{A}(y)$. But then by Lemma 5.2, w is not an extreme point of $\mathbf{L}_{U_S}$. The rest follow from part (i).

(iii) Since contracting MDPs are absorbing (Lemma 4.1), all the statements in (ii) hold. It remains to show that $\mathbf{Q}^{\mu}_{tc}(\beta) \subset \mathbf{L}_{U_S}(\beta)$. By applying Lemma 3.1 to (23), this statement will follow if we show that $\lim_{n\to\infty} \sum_y \rho(y, \mathsf{A}(y))[P^n(w)]_{yx}$ is zero. Indeed,

$$\sum_y \rho(y, \mathsf{A}(y))[P^n(w)]_{yx} \le \sum_y \rho(y, \mathsf{A}(y))\mu(y) \left\| [P^n(w)]_{\cdot, \mathbf{X}'} \right\|_{\mu},$$

which converges to zero by Lemma 4.1. ∎

6 Relation between cost and occupation measure

We begin by introducing different assumptions on the immediate costs, that will be used when applying either the general transient framework, or the contracting framework. When using the general transient framework we shall assume that the costs are nonnegative. We have the following properties of the total costs:

Theorem 6.1 *(Linear representation of the cost)*
(i) Assume that the MDP is contracting. Then for any instantaneous cost $c : \mathcal{K} \to \mathbb{R}$ and any β and $u \in U$

$$C_{tc}(\beta, u) = \langle c, f_{tc}(\beta, u) \rangle := \int_{\mathcal{K}} c(\kappa) f_{tc}(\beta, u; d\kappa); \tag{24}$$

the finite and infinite horizon total costs are uniformly μ-bounded over all policies:

$$\left\| C^T(\cdot, u) \right\|_\mu \leq \frac{\overline{b}}{1-\xi} \qquad\qquad \left\| C_{tc}(\cdot, u) \right\|_\mu \leq \frac{\overline{b}}{1-\xi}. \tag{25}$$

($C_{tc}(\cdot, u)$ is the vector of total cost corresponding to all initial states.)
(ii) Assume that the MDP is transient and the immediate costs are nonnegative. Then (24) holds for any $u \in U$.

Proof: (i) Fix some policy u. Since $f^T(\beta, u; y)$ converges monotonly to $f_{tc}(\beta, u; y)$ as $T \to \infty$, we have by the monotone convergence theorem

$$\lim_{T\to\infty} \langle f^T(\beta, u; y), \mu \rangle = \langle f_{tc}(\beta, u; y), \mu \rangle.$$

Since c is μ-bounded, we have by the dominated convergence theorem

$$C_{tc}(\beta, u) = \lim_{T\to\infty} C^T(\beta, u) = \lim_{T\to\infty} \langle f^T(\beta, u; y), c \rangle = \langle f_{tc}(\beta, u; y), c \rangle.$$

(25) follows from

$$\left\| C_{tc}(u) \right\|_\mu = \left\| \langle c, f_{tc}(\cdot, u) \rangle \right\|_\mu \leq \overline{b} \left\| f_{tc}(\cdot, u) \right\|_\mu \leq \frac{\overline{b}}{1-\xi}, \tag{26}$$

and

$$\left\| C^T(u) \right\|_\mu = \left\| \langle c, f^T(\cdot, u) \rangle \right\|_\mu \leq \overline{b} \left\| f_{tc}(\cdot, u) \right\|_\mu \leq \frac{\overline{b}}{1-\xi}. \tag{27}$$

(ii)

$$\begin{aligned} C_{tc}(\beta, u) &= \sum_{t=1}^{\infty} E_\beta^u c(X_t, A_t) = \sum_{t=1}^{\infty} \int_{\mathcal{K}} p_\beta^u(t; d\kappa) c(\kappa) \\ &= \int_{\mathcal{K}} \sum_{t=1}^{\infty} p_\beta^u(t; d\kappa) c(\kappa) = \langle f_{tc}(\beta, u), c \rangle, \end{aligned}$$

where the change between integration and summation follows since the integrand is non-negative (see [42] Corollary 11.14). ■

Lemma 6.1 *(The transient case: lower semi-continuity of the costs)*
Consider the transient framework, (Definition 3.1) with non-negative immediate costs. Then $C(\beta,\cdot)$ (and $D^k(\beta,\cdot), k=1,...,K$) are lower semi-continuous on U_S.

Proof: Let $w_n \to w$ be stationary. Then, by Fatou's Lemma (the arguments are as in the proof of Lemma 5.1 (i)) and Theorem 6.1, and by the lower semi-continuity of the occupation measures (Lemma 5.1 (i)), we have

$$\underline{\lim}_{n\to\infty} C_{tc}(\beta,w_n) = \underline{\lim}_{n\to\infty} \langle f_{tc}(\beta,w_n),c\rangle \geq \langle \underline{\lim}_{n\to\infty} f_{tc}(\beta,w_n),c\rangle \geq \langle f_{tc}(\beta,w),c\rangle = C_{tc}(\beta,w).$$

(The same argument holds for $D_{tc}(\beta,\cdot)$). ■

Lemma 6.2 *(Uniform convergence and continuity of the costs)*
Assume that the MDP is contracting. Then
(i) $C^T(\beta,u)$ converges to $C_{tc}(\beta,u)$ uniformly over U as $T\to\infty$.
(ii) $C_{tc}(\beta,u)$ is continuous on U_M.

Proof: (i) For any policy u,

$$\begin{aligned}
&|C_{tc}(\beta,u) - C^T(\beta,u)| \\
&\leq \sum_{t=T+1}^{\infty} E^u_\beta |c(X_t,A_t)| = \sum_{t=T+1}^{\infty} \int_{\mathcal{K}} p^u_\beta(t;d\kappa)|c(\kappa)| \\
&\leq \overline{b} \sum_{t=T+1}^{\infty} \sum_{y\in\mathbf{X}'} p^u_\beta(t;y)\mu(y) \leq \frac{\overline{b}\langle\beta,\mu\rangle\xi^T}{1-\xi}
\end{aligned}$$

which converges to 0 as $T\to\infty$. The last inequality follows from Lemma 4.1.
(ii) Consider any $u,u'\in U_M$. It follows from Theorem (6.1) that

$$\begin{aligned}
|C_{tc}(\beta,u) - C_{tc}(\beta,u')| &= \langle c, f_{tc}(\beta,u)\rangle - \langle c, f_{tc}(\beta,u')\rangle \leq \langle c, f_{tc}(\beta,u) - f_{tc}(\beta,u')\rangle \\
&\leq \overline{b} \sum_{y\in\mathbf{X}'} |c, f_{tc}(\beta,u) - f_{tc}(\beta,u')|\mu(y).
\end{aligned}$$

Since f_{tc} are μ-continuous (Lemma 5.1 (ii). ■

Lemma 6.3 *(Extension to $\mathcal{U}$ and $\overline{M}(U_M)$)*
The results of Lemmas 6.1 and 6.2 as well as Theorem 6.1 holds also for $\overline{M}(U_M)$ (and thus, in particular, for $\mathcal{U}$).

Proof: The extension of Lemma 6.1: 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 lower semi-continuity in in $\overline{M}(U_M)$ is established by applying again Fatou's Lemma:

$$\underline{\lim}_{n\to\infty} C_{tc}(\beta,\hat{\gamma}^n) = \underline{\lim}_{n\to\infty} \langle \gamma^n, C_{tc}(\beta,\cdot)\rangle \geq \langle \underline{\lim}_{n\to\infty} \gamma^n, C_{tc}(\beta,\cdot)\rangle = C_{tc}(\beta,\hat{\gamma}).$$

The extension of Lemma 6.2: we thus consider the contracting framework.

$$\begin{aligned}
&\sup_{\hat{q}\in\overline{M}(U_M)} |C^T(\beta,\hat{q}) - C_{tc}(\beta,\hat{q})| \\
&= \sup_{q\in M_1(U_M)} \left| \int_{U_M} C^T(\beta,u)q(du) - \int_{U_M} C_{tc}(\beta,u)q(du)\right| \\
&\leq \sup_{q\in M_1(U_M)} \int_{U_M} |C^T(\beta,u) - C_{tc}(\beta,u)|q(du) \\
&= \sup_{u\in U_M} \int_{U_M} |C^T(\beta,u) - C_{tc}(\beta,u)|
\end{aligned}$$

We conclude that Lemma 6.2 (i) holds for $\overline{M}(U_M)$.

Let $q^n, n = 1,2,...$ and q be probability measures over U_M, and let $\hat{q}^n$ and $\hat{q}$ be the corresponding policies in $\overline{M}(U_M)$. Assume that $\hat{q}^n$ converges to $\hat{q}$ (by which we mean that q^n converges to q weakly). Then

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

Indeed, this follows (see Billingseley [12]) since, by Theorem 6.1 (i), $C_{tc}(\beta,u)$, are bounded on U_M, and since, by Lemma 6.2, $C_{tc}(\beta,u)$ are continuous on U_M This establishes Lemma 6.2 (ii) for $\overline{M}(U_M)$.

The extension of Theorem 6.1 to $\overline{M}(U_M)$ is straightforward. ■

7 Dominating classes of policies

Theorem 7.1 *(Dominating policies)*
(i) Consider the transient framework, (Definition 3.1) with non-negative immediate costs. Then U_S *and* $\mathcal{U}$ *are dominating class of policies.*
(ii) Consider the contracting framework, (Definition 4.1). Then any complete class of policies (definition 5.1) is a dominating class of policies.
(iii) Under the assumptions of (i) or of (ii), if **COP** *is feasible, then there exist an optimal policy in* U_S *and in* $\mathcal{U}$.

Proof: (i) follows from the linear representation of the cost (Theorem 6.1) as well as the weak completeness of the of stationary policies (Theorem 5.1). The proof for $\mathcal{U}$ can be found in [4].

(ii) follows from similar arguments.

(iii) Recall that the set U_S of stationary policies and $\mathcal{U}$ are compact. Under the assumptions of (i) or (ii), the costs are lower semi-continuous on U_S and on $\mathcal{U}$ (Lemma 6.2, 6.1 and 6.3). This implies that the feasible set of stationary policies $\Pi_V := \{u : u \in U_S, D_{tc}(\beta, u) \leq V\}$ is compact, since it is obtained as the intersection of the compact set U_S and the inverse map of the closed sets $(-\infty, V_k]$. Finally, by the lower semi-continuity of $C_{tc}(\beta, u)$ on Π_V we conclude that $C_{tc}(\beta, u)$ achieves its minimum on $\Pi_V(\beta, u)$, i.e. there exists an optimal stationary policy for COP. Similarly, it follows that there exists an optimal policy within $\mathcal{U}$. ■

8 Equivalent linear program

We show below that **COP** is equivalent to a LP with countable number of decision variables and a countable number of constraints. Such equivalence was obtained for the total cost criterion for finite states and actions by Kallenberg [32]. The LP formulation constitutes an important method for computing stationary optimal policies.

Consider the following LP, that will correspond to the transient case:
$\mathbf{LP_1}(\beta)$: Find the infimum $\mathcal{C}^*$ of $\mathcal{C}(\rho) := \langle c, \rho \rangle$ subject to:

$$\mathcal{D}^k(z) := \langle d^k, \rho \rangle \leq V_k, k = 1, ..., K, \qquad \rho \in \mathbf{Q}_{tc}(\beta) \tag{28}$$

where $\mathbf{Q}_{tc}(\beta)$ was defined in (17).

We similarly define the LP for the contracting case:
$\mathbf{LP}_{\mathbf{1}}^{\mu}(\beta)$: Find the infimum $\mathcal{C}^*$ of $\mathcal{C}(\rho) := \langle c, \rho \rangle$ subject to:

$$\mathcal{D}^k(z) := \langle d^k, \rho \rangle \leq V_k, k = 1, ..., K, \qquad \rho \in \mathbf{Q}_{tc}^{\mu}(\beta). \tag{29}$$

We show that there is a one to correspondence between feasible (and optimal) solutions to the LP, and the feasible (and optimal) solutions to **COP**.

Theorem 8.1 *(Equivalence between COP and LP, the transient case)*
Assume that the MDP is transient and the immediate costs are nonnegative. Then
(i) $\mathcal{C}^* = C_{tc}(\beta)$.
(ii) For any $u \in U$, $\rho(u) := f_{tc}(\beta, u) \in \mathbf{Q}_{tc}(\beta)$, $C_{tc}(\beta, u) = \mathcal{C}(\rho(u))$ *and* $D_{tc}(\beta, u) = \mathcal{D}(\rho(u))$*; conversely, for any* $\rho \in \mathbf{Q}_{tc}(\beta)$*, the stationary policy* $w(\rho)$ *(defined above (23)) satisfies* $C_{tc}(\beta, w(\rho)) \leq \mathcal{C}(\rho)$ *and* $D_{tc}(\beta, w(\rho)) \leq \mathcal{D}(\rho)$.
(iii) $\mathbf{LP}_{\mathbf{1}}(\beta)$ *is feasible if and only if* **COP** *is. Assume that* **COP** *is feasible. Then there exists an optimal solution* ρ^* *for* $\mathbf{LP}_{\mathbf{1}}(\beta)$*, and the stationary policy* $w(\rho^*)$ *is optimal for* **COP**.

Proof: We start from (ii). The first claim follows from eq. (18). The claims on the costs follow from Theorem 6.1 and Theorem 5.1.
(i) and (iii) now follows from (ii) and Theorem 7.1. ■

For the contracting case we get similarly:

Theorem 8.2 *(Equivalence between COP and LP, the contracting case)*
Assume that the MDP is contracting. Then
(i) $\mathcal{C}^* = C_{tc}(\beta)$.
(ii) For any $u \in U$, $\rho(u) := f_{tc}(\beta, u) \in \mathbf{Q}_{tc}^{\mu}(\beta)$, $C_{tc}(\beta, u) = \mathcal{C}(\rho(u))$ *and* $D_{tc}(\beta, u) = \mathcal{D}(\rho(u))$*; conversely, for any* $\rho \in \mathbf{Q}_{tc}^{\mu}(\beta)$*, the stationary policy* $w(\rho)$ *(defined above (23) satisfies* $C_{tc}(\beta, w(\rho)) = \mathcal{C}(\rho)$ *and* $D_{tc}(\beta, w(\rho)) = \mathcal{D}(\rho)$.
(iii) $\mathbf{LP}_{\mathbf{1}}^{\mu}(\beta)$ *is feasible if and only if* **COP** *is. Assume that* **COP** *is feasible. Then there exists an optimal solution* ρ^* *for* $\mathbf{LP}_{\mathbf{1}}^{\mu}(\beta)$*, and the stationary policy* $w(\rho^*)$ *is optimal for* **COP**.

References

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

[2] E. Altman, "Denumerable constrained Markov Decision Processes and finite approximations", *Math. of Operations Research*, **19**, No. 1, pp. 169-191, 1994.

[3] E. Altman, "Constrained Markov decision processes with total cost criteria: Lagrange approach and dual LP", submitted, 1996.

[4] E. Altman, *Constrained Markov Decision Processes*, INRIA Report, May 1995.

[5] E. Altman and A. Shwartz, "Optimal priority assignment: a time sharing approach", *IEEE Transactions on Automatic Control* Vol. AC-34 No. 10, pp. 1089-1102, 1989.

[6] E. Altman and A. Shwartz, "Markov decision problems and state-action frequencies," *SIAM J. Control and Optimization.* **29**, No. 4, pp. 786-809, 1991

[7] E. Altman and A. Shwartz, "Sensitivity of constrained Markov Decision Problems", *Annals of Operations Research*, **32**, pp. 1-22, 1991.

[8] R. J. Aumann, "Mixed and behavior strategies in infinite extensive games", *Advances in Game Theory*, Ann. Math. Study. **52**, pp. 627-650, 1964.

[9] P. Bernhard, "Information and strategies in dynamic games", *SIAM J. Cont. and Opt.* **30**, pp. 212-228, 1992.

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

[11] F. J. Beutler and K. W. Ross, "Time-Average Optimal Constrained Semi-Markov Decision Processes", *Advances of Applied Probability* **18**, No. 2, pp. 341-359, 1986.

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

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

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

[15] V. S. Borkar, "Ergodic control of Markov Chains with constraints – the general case", *SIAM J. Control and Optimization.* **32**, No. 1, pp. 176-186, 1994.

[16] R. Dekker and A. Hordijk, “Average, sensitive and Blackwell optimal policies in denumerable Markov decision chains with unbounded rewards", *Mathematics of Operations Research*, **13**, pp. 395-421, 1988.

[17] C. Derman, *Finite State Markovian Decision Processes,* Academic Press, 1970.

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

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

[20] C. Derman and R. E. Strauch, “A note on memoryless rules for controlling sequential control processes", *Ann. Math. Stat.* **37**, pp. 276-278, 1966.

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

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

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

[24] E. A. Feinberg and I. Sonin, “Notes on equivalent stationary policies in Markov decision processes with total rewards", submitted to *ZOR - Methods and Models in Operations Research*, 1995.

[25] E. A. Feinberg E. and A. Shwartz Constrained discounted dynamic programming, to appear in *Math. of Operations Research*, 1995.

[26] M. Haviv, “On constrained Markov decision processes", submitted to *OR letters*, 1995.

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

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

[29] A. Hordijk and L. C. M. Kallenberg, “Constrained undiscounted stochastic dynamic programming", *Mathematics of Operations Research,* **9**, No. 2, pp. 276-289, 1984.

[30] A. Hordijk and F. Spieksma, "Constrained admission control to a queuing system" *Advances of Applied Probability* Vol. 21, pp. 409-431, 1989.

[31] D. Kadelka, "On randomized policies and mixtures of deterministic policies", manuscript.

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

[33] J. G. Kemeney, J. L. Snell and A. W. Knapp, *Denumerable Markov Chains*, Springer-Verlag, 1976.

[34] N. Krylov, "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, 1985.

[35] H. W. Kuhn, "Extensive games and the problem of information", *Ann. Math. Stud.* **28**, pp. 193-216, 1953.

[36] A. Lazar, "Optimal flow control of a class of queuing networks in equilibrium", *IEEE Transactions on Automatic Control*, Vol. 28 no. 11, pp. 1001-1007, 1983.

[37] P. Nain and K. W. Ross, "Optimal Priority Assignment with hard Constraint", *Transactions on Automatic Control,* **31**, pp. 883-888, 1986.

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

[39] K. W. Ross, "Randomized and past-dependent policies for Markov decision processes with multiple constraints", *Operations Research* **37**, No. 3, pp. 474-477, 1989.

[40] K. W. Ross and B. Chen, "Optimal scheduling of interactive and non interactive traffic in telecommunication systems", *IEEE Trans. on Auto. Control,* Vol. 33 No. 3 pp. 261-267, 1988.

[41] K. W. Ross and R. Varadarajan, "Markov Decision Processes with Sample path constraints: the communicating case", *Operations Research*, **37**, No. 5, pp. 780-790, 1989.

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

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

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

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

[46] J. Van Der Wal, *Stochastic Dynamic Programming*, Mathematisch Centrum, Amsterdam, 1990.