SENSITIVITY OF
CONSTRAINED MARKOV DECISION PROCESSES

Eitan Altman and Adam Shwartz

Electrical Engineering
Technion — Israel Institute of Technology
Haifa, 32000 Israel

ABSTRACT

We consider the optimization of finite-state, finite-action Markov Decision processes, under constraints. Costs and constraints are of the discounted or average type, and possibly finite-horizon. We investigate the sensitivity of the optimal cost and optimal policy to changes in various parameters.

We relate several optimization problems to a generic Linear Program, through which we investigate sensitivity issues. We establish conditions for the continuity of the optimal value in the discount factor. In particular, the optimal value and optimal policy for the expected average cost are obtained as limits of the discounted case, as the discount factor goes to one. This generalizes a well-known result for the unconstrained case.

We also establish the continuity in the discount factor for certain non-stationary policies. We then discuss the sensitivity of optimal policies and optimal values to small changes in the transition matrix and in the instantaneous cost functions. The importance of the last two results is related to the performance of adaptive policies for constrained MDP under various cost criteria [3, 5].

Finally, we establish the convergence of the optimal value for the discounted constrained finite horizon problem to the optimal value of the corresponding infinite horizon problem.

INTRODUCTION

Consider the standard model of a Markov Decision Process (MDP) with a finite state and action spaces. A natural generalization of the optimization problem is to include constraints. Such models arise in relation to resource-sharing systems; for example, in modern Integrated Services Digital Networks, many types of messages require bounded delays. This applies to voice messages, video, interactive computer messages, etc. On the other hand, requirements on data transfer are best modelled by imposing a cost for the delay. The resulting model is a constrained MDP [17, 19].

The analysis of Constrained MDP's started with [11]. The well-known formulation of the average-cost MDP as a Linear Program [16, 12] led naturally to the inclusion of additional constraints. Average-cost MDP's with additional constraints were also investigated by Derman [10], Kallenberg [15], and by Hordijk and Kallenberg [14] who treat the multi-chain case. Recent results by Altman and Shwartz [2] and by Borkar [7] relax the finiteness assumption. The basic results in the single-class case are the existence of an optimal stationary policy which, in the finite case, can be computed through a Linear Program. Similar results for the discounted cost are obtained by Altman and Shwartz [5] and Borkar [6].

It is of interest, from a theoretical as well as from a practical point of view, to analyse the sensitivity of the optimal value and optimal policy to small changes in the parameters. In the modeling of real systems, the values of the discount factor, transition probabilities and even costs are usually not precisely known. Moreover, infinite horizon problems are clearly an idealization of a problem with a long, but finite horizon. When we believe that the parameters are known with reasonable precision, or that the horizon is long enough, the continuity assures us that the small errors in the data lead to small errors in performance. In the unconstrained case, this continuity is easy to establish through the Dynamic Programming Equation (although Dekker [9] uses a Parametric Linear Program to discuss the sensitivity in the discount factor). However, the Dynamic Programming Equation is not applicable to constrained problems. Since the computation of optimal policies and optimal values is accomplished via Linear Programs, we can apply the powerful tool of sensitivity analysis for Linear Programs (e.g. [8]) to investigate the sensitivity of Constrained MDP's. The problem is more delicate than in the unconstrained case, since the optimal policy for the limit problem may not be feasible for the approximate problem, and so will not, in general, be ϵ-optimal.

When uncertainty is large, continuity is not sufficient and methods of adaptive control need to be applied. Here, the idea is to estimate the unknown parameters *on line*, and update the control accordingly. The first adaptive control scheme for constrained MDP's was investigated by Makowski and Shwartz [21] in the context of queueing system, and general results for finite-state MDP's under several constraints were first given by Altman and Shwartz [1, 3, 4, 5]. In order for adaptive schemes to be useful, we need to establish that as our estimates of the parameters converge to the true values, the control we compute based on these estimates becomes close to optimal. This again is an issue of sensitivity, and it is in this context that sensitivity for Constrained MDP's was first considered [3].

The purpose of this paper is to present a unified treatment of the finite-state constrained problem, and treat the sensitivity (or continuity) issues for this model. After introducing the model and the Constrained Optimization Problem (abbreviated COP henceforth) we collect in Section 2 some basic results on the linear representation of average and discounted costs as functions of "state-action frequencies" (Derman [10]). We also recall the connection between the infinite-horizon (discounted and average) COP and Linear Programs (LP). We show that both Linear Programs can be written in a single generic form. This is made possible by a suitable normalization, which facilitates the unification of discounted and average problems. This unified approach applies to the COP problem, the related LP and the "state-action frequencies". This is also the appropriate normalization for the study of continuity in the discount factor and convergence of the discounted costs to the average costs.

Section 3 is devoted to the sensitivity analysis of the generic LP. In Section 4 we apply these results to obtain continuity in the discount factor β, for $0 < \beta \leq 1$, of the optimal cost, and continuity in a weaker sense of the optimal policies. The sensitivity to the discount factor in a class of non-stationary policies is also established. In [5] the authors derive optimal adaptive controls of this class for constrained optimization problems, with asymptotic cost criteria (such as average cost). This continuity result proves that adaptive policies of this class are optimal in a stronger sense. The issue of sensitivity to other parameters, such as transition probabilities, instantaneous cost and constraints arises in connection to adaptive control. In Section 5 we derive the relevant sensitivity problem, and obtain the continuity using the results of Section 3. The optimality of the resulting adaptive policy is discussed. Finally in Section 6 we investigate the convergence of the cost of the discounted finite-horizon problem to the corresponding infinite horizon problem. The difficulty in the constrained case is that the optimal policy for the limit problem may not be feasible for the finite-time problem, and so will not, in general, be ϵ-optimal.

1. THE CONSTRAINED PROBLEM

1.1 The model.

Let $\{X_t\}_{t=0}^{\infty}$ be the discrete time state process, defined on the finite *state space* $\mathbf{X} = \{1, ..., N\}$; the action A_t at time t takes values in the finite *action space* $\mathbf{A}$. The *history* of the process up to time t is denote by $H_t := (X_0, A_0, X_1, A_1, ..., X_t, A_t)$. If the state at time t is x and action a is applied, then the next state will be y with probability

$$P_{xay} := P(X_{t+1} = y \mid X_t = x; A_t = a) = P(X_{t+1} = y \mid H_{t-1} = h, X_t = x; A_t = a)$$

A policy u in the *policy space* U is a sequence $u = \{u_0, u_1, ...\}$, where u_t is applied at time epoch t, and $u_t(\cdot \mid H_{t-1}, X_t)$ is a conditional probability measure over $\mathbf{A}$. We denote the probability measure corresponding to policy u and initial state x by P_x^u, and the expectation operator by E_x^u.

A *stationary policy* $g \in U(S)$ is characterized by a single conditional distribution $p^g_{\cdot|x} := u(\cdot \mid X_t = x)$ over $\mathbf{A}$, so that $p^g_{\mathbf{A}|x} = 1$; under g, X_t becomes a Markov chain with stationary transition probabilities, given by $P^g_{xy} := \sum_{a \in \mathbf{A}} p^g_{a|x} P_{xay}$.

Throughout the paper we impose the following assumption:

A1: Under any stationary deterministic policy g the state space of the process X_t consists of a single ergodic class.

In Section 5 we shall need the slightly stronger assumption

A1$'$: Under any stationary deterministic policy g the state space of the process X_t consists of a single ergodic non-cyclic class.

Under assumption A1, each stationary policy g induces a unique stationary steady state distribution on $\mathbf{X}$, denoted by $\pi^g(\cdot)$.

The following notation is used below: $\delta_a(x)$ is the Kronecker delta function. For any set B, $1[B]$ is the indicator function of the set, and if B is finite we denote by $|B|$ the cardinality of this set (i.e. the number of elements in B). For vectors D and V in $\mathbb{R}^K$, the notation $D < V$ stands for $D_k < V_k$, $k = 1, 2, \ldots, K$, with a similar convention for matrices. For two matrices ζ, Q of appropriate dimensions the notation $\zeta \cdot Q$ stands for summation over common indices (scalar product).

1.2 The constrained problem

Let $C(x,u)$ and $D(x,u) := \{D^k(x,u)\ ,\ 1 \le k \le K\}$ be cost functions associated with each policy u and initial state x. The precise definitions of the cost functions of interest are given below. The real vector $V := \{V_k\ ,\ k = 1, ..., K\}$ is held fixed thereafter. Call a policy u *feasible* if

$$D^k(x,u) \le V_k\ , \quad k = 1, 2, \ldots, K \tag{1.1}$$

The constrained optimization problem is:

(**COP**) Find a feasible $v \in U$ that minimizes $C(x,u)$

The unconstrained problem (where $K = 0$) is denoted by **OP**. We shall frequently assume

A2: COP is feasible when replacing the inequalities in (1.1) by strict inequalities, i.e.

$$D^k(x,u) < V_k \qquad 1 \le k \le K$$

Let $c(x,a)$, $d(x,a) := \{d^k(x,a)\ ,\ k = 1, ..., K\}$ be real ($\mathbb{R}^K$) valued instantaneous cost functions, i.e. costs per state-action pair. Let $0 < \beta \le 1$ be a discount factor. We shall use the following *normalized discounted* cost functionals from $U \times \mathbf{X}$ to $\mathbb{R}$.

$$C^t_\beta(x,u) := (\sum_{s=0}^{t} \beta^s)^{-1} E^u_x \left[\sum_{s=0}^{t} \beta^s c(X_s, A_s) \right] \tag{1.2a}$$

$$D^{k,t}_\beta(x,u) := (\sum_{s=0}^{t} \beta^s)^{-1} E^u_x \left[\sum_{s=0}^{t} \beta^s d^k(X_s, A_s) \right] \qquad k = 1, ..., K \tag{1.2b}$$

$$C_\beta(x,u) := \overline{\lim_{t\to\infty}}\ C^t_\beta(x,u) \tag{1.3a}$$

$$D^k_\beta(x,u) := \overline{\lim_{t\to\infty}}\ D^{k,t}_\beta(x,u) \qquad k = 1, ..., K \tag{1.3b}$$

When $\beta = 1$, (1.2-1.3) reduce to the well-known definition of the expected average cost, since in that case $\sum_{s=0}^{t} \beta^s = t + 1$.

As we deal with finite state and action spaces, it follows immediately that for $\beta < 1$ (1.3) can be written as the *standard expected discounted* costs with a normalization constant:

$$C_\beta(x,u) := (1-\beta) E^u_x \left[\sum_{s=0}^{\infty} \beta^s c(X_s, A_s) \right] \tag{1.4a}$$

$$D^k_\beta(x,u) := (1-\beta) E^u_x \left[\sum_{s=0}^{\infty} \beta^s d^k(X_s, A_s) \right] \qquad k = 1, ..., K \tag{1.4b}$$

The constrained optimization problems COP^t_β and COP_β are defined by using the the costs (1.2) and (1.3) respectively in the general definition of COP. Denote by $C_\beta(x)$ ($C^t_\beta(x)$) the optimal value of COP_β (COP^t_β respectively) and by $U_\beta(x)$ the set of optimal policies of COP_β.

We are interested in the following problems:

(i) The continuity of $C_\beta(x)$ in β, especially near $\beta = 1$, and the "continuity" of $U_\beta(x)$ in β.

(ii) The continuity of $U_\beta(x)$ and $C_\beta(x)$ in the transition probabilities and in the one-step costs.

(iii) The convergence of $C^t_\beta(x)$ to $C_\beta(x)$ as $t \to \infty$.

These problems are related to the Sensitivity of LPs. In particular we shall reduce (i) and (ii) to questions of convergence of the optimal sets and optimal values of a sequence of Linear Programs (LPs) when the coefficients of the LPs converge. In Sections 4 and 5 we show that it suffices to consider the following generic sequence of LPs denoted by LP_n, and the limit LP, denoted LP_∞.

$\mathbf{LP_n}$: Find z^*_n that minimizes $c_n \cdot z$ subject to:

$$\psi^n(z) = 0 \tag{LPa}$$

$$z \geq 0 \tag{LPb}$$

$$d^k_n \cdot z \leq V^k \qquad 1 \leq k \leq K \tag{LPc}$$

where z, $c := \{c(y,a)\}_{ya}$ and $d := \{d(y,a)\}_{ya}$ are considered matrices of dimension $|\mathbf{X}| \times |\mathbf{A}|$, ψ^n are affine, $\psi^n \to \psi$ pointwise, $d^k_n \to d^k$ and $c_n \to c$.

2. CONSTRAINED OPTIMIZATION AND LINEAR PROGRAMS

2.1 The linear representation of the cost.

We begin by defining some basic quantities that play an important role in the expected discounted and in expected average cost. Given a discount factor $0 < \beta \leq 1$ define the collections $\{\bar f^{t,\beta}_{sa}(x,u;y,a)\}_{y,a}$ and $\{\bar f^{\beta}_{sa}(x,u;y,a)\}_{y,a}$ by

$$\bar f^{t,\beta}_{sa}(x,u;y,a) := [\sum_{s=0}^{t} \beta^s]^{-1} \sum_{s=0}^{t} \beta^s P^u_x(X_s = y, A_s = a) \tag{2.1a}$$

$$\bar f^{t,\beta}_{s}(x,u;y) := [\sum_{s=0}^{t} \beta^s]^{-1} \sum_{s=0}^{t} \beta^s P^u_x(X_s = y) \tag{2.1b}$$

By $\bar{f}^{\beta}_{sa}(x,u) := \{\bar{f}^{\beta}_{sa}(x,u;y,a)\}_{ya}$ we denote a generic accumulation point of $\bar{f}^{t,\beta}_{sa}(x,u)$. Similarly, we define $\bar{f}^{\beta}_{s}(x,u) := \{\bar{f}^{\beta}_{s}(x,u;y)\}_{ya}$ to be a generic accumulation point of $\bar{f}^{t,\beta}_{s}(x,u)$. Note that for $\beta < 1$ there is a single accumulation point

$$\bar{f}^{\beta}_{sa}(x,u;y,a) = [1-\beta]\sum_{t=0}^{\infty}\beta^t P^u_x(X_t = y, A_t = a) \tag{2.2a}$$

$$\bar{f}^{\beta}_{s}(x,u;y) = [1-\beta]\sum_{t=0}^{\infty}\beta^t P^u_x(X_t = y) \tag{2.2b}$$

For each β, t, x and u, $\bar{f}^{t,\beta}_{sa}(x,u)$ can be considered as a probability measure over $\mathbf{X}\times\mathbf{A}$. Since $\mathbf{X}$ and $\mathbf{A}$ are finite, the set of probability measures $\{\bar{f}^{t,\beta}_{sa}(x,u)\}_t$ is tight. It follows that all limit points are also probability measures, and we obtain $\sum_{y,a}\bar{f}^{\beta}_{sa}(x,u;y,a) = 1$. Similarly, $\sum_y \bar{f}^{\beta}_{s}(x,u;y) = 1$.

Lemma 2.1: *For each instantaneous cost $c(y,a)$, $y\in\mathbf{X}, a\in\mathbf{A}$,*

$$C^t_\beta(x,u) = \sum_{y\in\mathbf{X},a\in\mathbf{A}} c(y,a)\bar{f}^{t,\beta}_{sa}(x,u;y,a) \tag{2.3a}$$

$$C_\beta(x,u) = \overline{\lim_{t\to\infty}}\sum_{y\in\mathbf{X},a\in\mathbf{A}} c(y,a)\bar{f}^{t,\beta}_{sa}(x,u;y,a) = \sum_{y\in\mathbf{X},a\in\mathbf{A}} c(y,a)\bar{f}^{\beta}_{sa}(x,u;y,a) \tag{2.3b}$$

for some accumulation point $\bar{f}^{\beta}_{sa}(x,u)$. In particular, if $\beta < 1$ or if $u\in U(S)$ then $\bar{f}^{\beta}_{sa}(x,u)$ is unique and

$$C_\beta(x,u) = \sum_{y\in\mathbf{X},a\in\mathbf{A}} c(y,a)\bar{f}^{\beta}_{sa}(x,u;y,a) \tag{2.3c}$$

Similar expressions apply to D_β.

The representation (2.3b) is well-known in the expected average case ($\beta = 1$); Derman [10] calls $\bar{f}^{t,\beta}_{sa}(x,u;y,a)$ "state action frequencies". Kallenberg [15] and Hordijk and Kallenberg [14] used these frequencies in relation with constrained optimization with expected average costs. This representation is extended to countable spaces in Altman and Shwartz [2]. Derman [10] and Borkar [6] use a similar representation in the case $\beta < 1$ without the normalization, but the derivation is identical. Borkar [6] extends the representation to a countable state space and a compact action space. He calls the quantity $\bar{f}^{\beta}_{sa}$ "occupation measure". Borkar's approach is somewhat different, in that he defines the occupation measure through the cost and not explicitly through (2.2).

Let L_x^β denote the space of all state-action frequency matrices achieved by all policies in U, i.e. $L_x^\beta := \{\bar{f}_{sa}^\beta(x,u) : u \in U\}$ and $L_x^\beta(S)$ the space achieved by all policies in $U(S)$.

Lemma 2.2: *Under A1, $L_x^\beta = L_x^\beta(S)$ and is closed and convex.*

Proof: See [10] for $\beta = 1$. For $\beta < 1$ the proof can be obtained from Borkar [6]. ■

Theorem 2.3: *Under A1 the stationary policies are optimal for COP_β for each β.*

Proof: For $\beta = 1$ see [14]. For $\beta < 1$ it follows immediately from (2.3d) and Lemma 2.2. ■

The linear programs

Optimal policies for COP_β (which may be found in $U(S)$, according to Theorem 2.3) can be obtained by solving the appropriate Linear Programs. The following LP applies to both the discounted and average case (i.e. for any $0 < \beta \le 1$).

LP(β) : Find $z^* := \{z^*(y,a)\}_{y,a}$ that minimizes $\mathcal{C}(z) := \sum_{y,a} c(y,a)z(y,a)$ subject to:

$$\sum_{y,a} z(y,a)\left[\delta_v(y) - \beta P_{yav}\right] = [1-\beta]\delta_x(v) \qquad v \in \mathbf{X} \tag{2.4a}$$

$$\mathcal{D}^k(z) := \sum_{y,a} d^k(y,a)z(y,a) \le V_k \qquad 1 \le k \le K \tag{2.4b}$$

$$z(y,a) \ge 0 \qquad \sum_{y,a} z(y,a) = 1 \tag{2.4c}$$

Note that due to (2.4a), each initial condition x leads to a distinct Linear Program, unless $\beta = 1$. Given a set of nonnegative real numbers $z = \{z(y,a)\}_{y,a}$ define $\gamma(z) = \{\gamma_y^a(z)\}_{y,a}$ by $\gamma_y^a(z) := z(y,a)[\sum_a z(y,a)]^{-1}$. If the denominator is zero then γ is defined arbitrarily but such that $\gamma_y^a(z) \ge 0$ and $\sum_a \gamma_y^a(z) = 1$. Define $g(z)$ to be the stationary policy given by $p_{a|y} = \gamma_y^a(z)$.

Theorem 2.4: *Under A1,*

(i) For every policy u, any accumulation point $\zeta := \bar{f}_{sa}^\beta(x,u)$ satisfies (2.4a) and (2.4c).

(ii) If $\beta < 1$ or u is stationary, then $\mathcal{C}(\zeta) = C_\beta(x,u)$ and $\mathcal{D}^k(\zeta) = D_\beta^k(x,u)$ $\quad 1 \le k \le K$. Consequently, if u is feasible for COP_β then ζ satisfies (2.4b).

Proof: For $\beta = 1$ this result is well-known (see, e.g. [10]). When $\beta < 1$, (i) follows from [10, Thm. 3 p. 42] provided u is stationary. For general u, it follows from Lemma 2.2 that there exists a stationary policy g with $\bar{f}^{\beta}_{sa}(x,u) = \bar{f}^{\beta}_{sa}(x,g)$, so (i) holds. But Lemma 2.1 now implies (ii). ∎

Theorem 2.5: *Assume A1. Pick any ζ that satisfies (2.4a) and (2.4c).*

(i) ζ has, for each state y, at least one non-zero component $\zeta(y,a)$. Moreover, $\bar{f}^{\beta}_{sa}(x, g(\zeta)) = \zeta$.

(ii) $\mathcal{C}(\zeta) = C_\beta(x, g(\zeta))$ and $\mathcal{D}^k(\zeta) = D^k_\beta(x, g(\zeta)) \quad 1 \le k \le K$.

(iii) Assume that ζ satisfies also (2.4b). Then $g(\zeta)$ is a feasible policy for COP_β.

Proof: For $\beta = 1$ this result is well-known (see, e.g. [10]). When $\beta < 1$, it follows from [10, Thm. 3 p. 42] that $\bar{f}^{\beta}_{sa}(x, g(\zeta)) = \zeta$. Under A1,

$$\sum_a \bar{f}^{\beta}_{sa}(x, g(\zeta); y, a) = \bar{f}^{\beta}_{s}(x, g(\zeta); y) > 0$$

for each y, and (i) is established. Lemma 2.1 now implies (ii) and (iii). ∎

We are now in a position to establish the useful relationship between COP_β and $LP(\beta)$.

Theorem 2.6: *Assume A1. Then $C_\beta(x) = \mathcal{C}(z^*)$. Moreover,*

(i) let u be an optimal policy for COP_β and let $\bar{f}^{\beta}_{sa}(x,u)$ be a limit point such that 2.3b holds. Then $z := \bar{f}^{\beta}_{sa}(x,u)$ is optimal for $LP(\beta)$.

(ii) Conversely, let z^* be optimal for $LP(\beta)$. Then $g(z^*)$ is optimal for COP_β.

Proof: (i) Since u is feasible for COP_β and since $\bar{f}^{\beta}_{sa}(x,u)$ is an accumulation point,

$$V_k \ge D^k_\beta(x,u) := \overline{\lim_{t\to\infty}} \sum_{y\in\mathbf{X}, a\in\mathbf{A}} d^k(y,a) \bar{f}^{t,\beta}_{sa}(x,u;y,a) \ge \sum_{y\in\mathbf{X}, a\in\mathbf{A}} d^k(y,a) \bar{f}^{\beta}_{sa}(x,u;y,a) := \mathcal{D}^k(z) \tag{2.5}$$

so that z is feasible for $LP(\beta)$. If z is not optimal, then there exists a solution z' of $LP(\beta)$ with $z' \cdot c < z \cdot c$. But by Theorem 2.5(ii), $g(z')$ is feasible for COP_β and $C_\beta(x, g(z')) < C_\beta(x,u)$, which contradicts the optimality of u.

(ii) By Theorem 2.5(ii), $g(z^*)$ is feasible for COP_β and $z^* \cdot c = C_\beta(x, g(z^*))$. If u' is feasible and satisfies $C_\beta(x,u') < C_\beta(x, g(z^*))$ then by (i) z^* is not optimal for $LP(\beta)$. Thus $g(z^*)$ is optimal for COP_β. The first claim now follows from 2.3b and the definition of the cost in $LP(\beta)$. ∎

Since the COP problems are related to LP problems, in order to prove the continuity results for a COP problem it suffices to study the generic sequence of Linear Programs LP_n. This is done in the next Section.

Since the constrained problem with $\beta < 1$ is quite new, we point out some of its' properties and some important differences from the non-constrained case.

(i) In the unconstrained case it is well-known that an optimal stationary deterministic policy can be found (for any $0 < \beta \leq 1$). In the constrained case we do not in general have optimal stationary deterministic policies, but rather optimal stationary randomized policies. It can easily be shown (as for the expected average case [18]) that there exists an optimal stationary policy which applies randomization in at most K states.

(ii) Unlike the unconstrained case, or the constrained case with $\beta = 1$, the optimal stationary policy depends on the initial state (or initial distribution). Consequently, the optimality principle does not apply to the discounted constrained case. An example with one constraint that exhibits these points can be found in [5].

3. SENSITIVITY OF LPs

In this Section we discuss the sensitivity of the Linear Programs $LP(\beta)$ which were derived in previous the Section. To this end, consider the sequence of LPs

$\mathbf{LP_n}$: Find $\{z_n^*(y,a)\}_{y,a}$ that minimizes $\mathcal{C}_n(z) := \sum_{y,a} c_n(y,a)z(y,a)$ subject to:

$$\sum_{y,a} z(y,a)\left[\delta_v(y) - \beta_n P^n_{yav}\right] = [1-\beta_n]\delta_x(v) \qquad v \in \mathbf{X} \tag{3.1a}$$

$$\mathcal{D}^k_n(z) := \sum_{y,a} d^k_n(y,a)z(y,a) \leq V^n_k \qquad 1 \leq k \leq K \tag{3.1b}$$

$$z(y,a) \geq 0 \qquad \sum_{y,a} z(y,a) = 1 \tag{3.1c}$$

To simplify the notation, we represent all equality constraints as $\psi^n(z) = 0$, and all inequality constraints as $\phi^n(z) \leq 0$. Thus ϕ^n and ψ^n are affine functions from $\mathbb{R}^l$ to $\mathbb{R}^m$ and $\mathbb{R}^{m'}$ respectively, where l is the number of decision variables.

Concerning LP_n, we assume throughout this Section:

$\phi^n \to \phi, \psi^n \to \psi$ pointwise and $c_n \to c$ (so that $\mathcal{C}_n \to \mathcal{C}$).

For each n, $0 < \beta_n \le 1$ and $\{P^n_{yav}\}_{yav}$ is the conditional transition matrix of some Markov chain.

With this notation, LP_n takes the form

$\mathbf{LP_n}$: Find $\{z^*_n(y,a)\}_{y,a}$ that minimizes $\mathcal{C}_n(z) := \sum_{y,a} c_n(y,a)z(y,a)$ subject to:

$$\psi^n(z) = 0 \tag{3.2a}$$

$$\phi^n(z) \le 0 \tag{3.2b}$$

The notation of Section 1 for this generic LP_n, which displays the inequality constraints explicitly is also used below. The limiting LP is $LP(\beta)$ of Section 2, which is also denoted LP_∞.

The results of this Section concern the convergence of optimal sets and optimal value of the generic sequence of LPs LP_n to those of the limiting LP-LP_∞. Specific applications are dealt with in subsequent Markov chain. With this notation, LP_n takes the form (cf. Section 1) Sections.

We begin by presenting a Lemma obtained by Dantzig, Folkman and Shapiro [8]. Let H and H_n be some sets. We need the following notion of convergence;

$$\overline{\lim_n} H_n = \{x \mid x = \lim_i x_{n_i} \text{ for some subsequence } x_{n_i}, \text{ with } x_{n_i} \in H_{n_i}\}$$

$$\underline{\lim_n} H_n = \{x \mid x = \lim_n x_n, \text{ with } x_n \in H_n \text{ except for a finite number of } n\}$$

We write $H_n \to H$ if $\underline{\lim}_n H_n = \overline{\lim}_n H_n = H$. Given any function σ from $\mathbb{R}^l$ to $\mathbb{R}$ and a set $H \in R^l$, define $M(\sigma|H)$ to be the subset of H where σ achieves its minimum, i.e.

$$M(\sigma|H) := \{x \in H : \sigma(x) = \inf\{\sigma(y)|y \in H\}\}$$

Denote $H(\phi,\psi) := \{x \in \mathbb{R}^l \mid \phi(x) \le 0, \text{ and } \psi(x) = 0\}$; this is the feasible set of the Linear Program associated with ϕ, ψ.

Lemma 3.1: *Assume that $lim_{n\to\infty} H(\phi^n, \psi^n) = H$ for some set H. Let σ and $\{\sigma_n\}$ be linear functions such that $\sigma_n \to \sigma$ pointwise. Then*

$$\underline{\lim}_{n\to\infty} M(\sigma_n|H(\phi^n,\psi^n)) \subset M(\sigma|H)$$

Proof: See [8] Theorem I.2.2. ■

Thus the convergence of the set of optimal solutions of a Linear Program depends on the convergence of the feasible set. Next we quote another result from [8] that gives conditions for this convergence. Denote

$I := \{i \mid 1 \le i \le m,\ \phi_i(x) = 0 \text{ for all } x \in H(\phi, \psi)\}$

This set depends only on the limiting LP, and identifies those *inequality* constraints which are, in fact, equalities for *all feasible points.* Let $\phi_I := \{\phi_i\ ,\ \ i \in I\}$ and denote by $\text{rank}(\phi_I, \psi)$ the rank of the matrix whose $|I| + m'$ rows are the coefficients of the linear part of $\phi_i, i \in I$, and $\psi_j, 1 \le j \le m'$.

Lemma 3.2: ([8] Cor. II.3.4) *Assume that* $\overline{\lim}_{n\to\infty} \text{rank}(\phi_{\rm I}^{\rm n}, \psi^{\rm n}) \le \text{rank}(\phi_{\rm I}, \psi)$ *and that* $H(\phi, \psi)$ *is non empty. Then either* $\lim_{n\to\infty} H(\phi^n, \psi^n) = H(\phi, \psi)$ *or* $H(\phi^n, \psi^n)$ *is empty for infinitely many* n*.*

In order to apply the previous Lemmas to LP_n we first show that A1 and A2 imply the hypotheses of Lemma 3.2. We then show that under A1-A2 the feasible set of LP_n is empty only finitely often. It is then possible to apply the results of Lemma 3.1.

Let J_n denote the number of linearly independent equations among the $|X| + 1$ equality constraints in LP_n, $n = 1, 2, \ldots, \infty$.

Lemma 3.3: $J_n \le |\mathbf{X}|$, *and under A1* $J_\infty = |\mathbf{X}|$.

Proof: It is easy to check (by separating the cases $\beta < 1$ and $\beta = 1$) that $J_n \le |X|$. Under A1, suppose $J_\infty < |\mathbf{X}|$. Then there exists some z satisfying (3.1a) and (3.1c) with at most J_∞ non-zero components. But this contradicts Theorem 2.5 (i). ∎

Lemma 3.4: *Under A1, A2 is equivalent to* $H(\phi, \psi) \ne \emptyset$ *together with* $I = \emptyset$*, and implies* $\text{rank}(\phi_{\rm I}^{\rm n}, \psi^{\rm n}) \le \text{rank}(\phi_{\rm I}, \psi)$ *for all* n*.*

Proof: Since for all n $\text{rank}\psi^{\rm n} \le |{\rm X}|$, Lemma 3.3 and $I = \emptyset$ imply

$$\text{rank}\psi^{\rm n} = \text{rank}(\phi_{\rm I}^{\rm n}, \psi^{\rm n}) \le \text{rank}(\phi_{\rm I}, \psi) = |\mathbf{X}| \quad \text{for all n}.$$

Since the feasible set of LP_∞ is convex, $H(\phi, \psi) \ne \emptyset$ together with $I = \emptyset$ is equivalent to the requirement that there exists some z that satisfies $\psi(z) = 0$ and $\phi(z) < 0$. By Theorem 2.5 (ii) this implies A2. On the other hand, we conclude from Theorem 2.4 (ii) that A2 implies the existence of some stationary policy v such that $z(y, a) := \pi^v(y) \cdot p^v_{a|y}$ satisfy

$$d^k \cdot z < V_k \qquad 1 \le k \le K \qquad \psi(z) = 0 \tag{3.3}$$

We shall construct some z' satisfying both (3.3) and $z' > 0$.

Pick any stationary policy w that satisfies $p^w_{a|y} > 0, \quad a \in \mathbf{A}, y \in \mathbf{X}$. Let $z''(y,a) = \pi^w(y) \cdot p^w_{a|y}$. By assumption A1 and Theorem 2.4 (i) z'' satisfies $z'' > 0$ and $\psi(z'') = 0$. Let $z' := (1-\delta)z + \delta z''$, where δ is some positive constant. Since ψ is affine, $\psi(z') = 0$, and clearly $z' > 0$. By choosing δ small enough, z' satisfies (3.3), so $H(\phi, \psi) \neq \emptyset$ and $I = \emptyset$. ■

Lemma 3.5: *Assume A1,A2. Assume that for all but a finite number of n, there exists some point z_n that satisfies $\psi^n(z_n) = 0$ and $z_n \geq 0$. Then $H(\phi^n, \psi^n)$ is at most finitely often empty.*

Proof: Consider the auxiliary Linear Program

$\mathbf{LP2_n}$: Find z and η that minimize η, subject to

$$\psi^n(z) = 0 \tag{3.4a}$$

$$z \geq 0 \tag{3.4b}$$

$$d^j_n \cdot z \leq V^j + \eta \tag{3.4c}$$

Denote the optimal η by η^n. Let H'_n denote the set of (z, η) satisfying (3.4a–3.4c). Let $LP2_\infty$ denote the Linear Program obtained by replacing ψ^n by ψ and d^j_n with d^j in $LP2_n$.

By A2 and Theorem 2.4 (i) there exists a stationary policy u for which $\zeta := \bar{f}^\beta_{sa}(x, u)$ satisfies (3.4a)–(3.4c) for $n = \infty$ with $\eta < 0$. Therefore $\eta^\infty < 0$.

Since the state and action spaces are finite, there exists some positive constant L such that $D^k_x(u) \leq L$ for all policies u and all k. Thus by choosing $\eta' = L + \max_k\{| \ V^k \ |\}$ we have $(z_n, \eta') \in H'_n$, so H'_n is nonempty except for a finite number of n. Define $\tilde{\phi}^n$ as the respective inequality constraints in $LP2_n$. Note that ψ^n is exactly the function defining LP_n.

From Lemma 3.5 and (3.4c) it easily follows that

$$I := \{i \mid 1 \leq i \leq m, \ \tilde{\phi}_i(x) = 0 \text{ for all } x \in H'_\infty(\tilde{\phi}, \psi)\} = \emptyset$$

As in Lemma 3.5 we conclude $\mathrm{rank}(\tilde{\phi}^{\mathrm{n}}_{\mathrm{I}}, \psi^{\mathrm{n}}) \leq \mathrm{rank}(\tilde{\phi}_{\mathrm{I}}, \psi)$ for all n. It then follows from Lemma 3.2 that the feasible sets $H'(\tilde{\phi}^n, \psi^n)$ converge to $H'_\infty(\tilde{\phi}, \psi)$. We claim that $\eta^n \leq 0$ for all large enough n. To prove the claim, assume the converse. Then, since H'_n is nonempty for all large n, there exists a subsequence n_i such that $0 \leq \eta^{n_i} \leq \eta'$ and $(z^*_n, \eta^{n_i}) \to (z^*, \eta)$, where $\eta \geq 0$. But by Lemma 3.1 applied to the subsequence n_i, $\eta^n \to \eta^\infty < 0$ and the claim follows. Thus $LP2_n$ is feasible for all n large, so that $H(\phi^n, \psi^n)$ is at most finitely often empty. ■

Theorem 3.6: *Assume A1,A2 and consider LP_n, as defined at the beginning of this Section. Then*

(i) $\underline{\lim}_{n\to\infty} M(\mathcal{C}_n | H(\phi^n, \psi^n)) \subset M(\mathcal{C}|H)$

(ii) $\overline{\lim}_{n\to\infty} M(\mathcal{C}_n | H(\phi^n, \psi^n)) \subset M(\mathcal{C}|H)$

(iii) If moreover LP_∞ possesses a unique solution z^*, then any sequence $z_n^* \in M(\mathcal{C}_n|H(\phi^n, \psi^n))$ converges to z_∞^*.

(iv) The optimal values of LP_n converge to $\mathcal{C}(z^*)$.

Proof: Fix an arbitrary stationary deterministic policy g, and consider the transition matrices $\{P_{xy}^{n;g}\}$. Recall that the ergodic structure of a finite Markov chain is determined by the location of the zero-entries in the transition matrix. By the hypotheses, $\{P_{xy}^{n;g}\} \to \{P_{xy}^{g}\}$. Thus, for n large enough, $\{P_{xy}^{n;g}\}$ is the transition matrix of a Markov chain with a single recurrent class. Theorem 2.4(i) now implies that there is a point z_n such that $\psi^n(z_n) = 0$ and $z \geq 0$ for all large n. From Lemma 3.4 we conclude that H_n is at most finitely often empty, and Lemma 3.1 implies (i).

Pick any sequence $z_n \in M(\mathcal{C}_n|H(\phi^n, \psi^n))$ and let $n(i)$ be an increasing subsequence along which $z_{n(i)}$ converges. (ii) is then obtained by applying (i) to that subsequence. (iii) is immediate from the last argument and (ii). Since $\mathcal{C}_n$ are linear and converge to $\mathcal{C}$ and since for all large n the feasible set $H(\phi^n, \psi^n)$ is compact, non-empty and uniformly bounded, it follows that $M(\mathcal{C}_n|H(\phi^n, \psi^n))$ is non-empty for all large n, and uniformly bounded. But this together with (i) implies (iv), since by definition $\mathcal{C}(z)$ is constant on $M(\mathcal{C}|H)$. ■

4. SENSITIVITY IN THE DISCOUNT FACTOR

As an immediate Corollary to Theorem 3.6 we obtain the continuity in the discount factor.

Theorem 4.1: *Assume A1,A2. Then $C_\beta(x)$ is continuous in β. If g^n is a sequence of stationary policies which are optimal for COP_{β_n} and if $\beta_n \to \beta$, then any limit point g of g_n is optimal for COP_β.*

Remark: it is easy to establish, by a direct argument, that under A2, if the stationary policy g is optimal for COP_β then $C_{\beta_n}(x, g) \to C_\beta(x, g)$. But, in contrast with the unconstrained case, it is not true in general that g is ϵ-optimal for COP_{beta_n} since g may not be feasible for COP_{β_n}. However, it is straightforward to construct, using the arguments of Theorem 6.1, Section 6, a series of modifications g_n of g so that $g_n \to g$ and $C_{\beta_n}(x, g_n) - C_{\beta_n}(x) \to 0$.

Next we treat a sensitivity problem for non-stationary policies. In the study of adaptive problems the class of Asymptotically Stationary policies [5] has proved quite useful. The structure and main ideas of the adaptive problem are discussed in the next Section. We refer the reader to [5] for the precise definitions, and only remark that under $A1'$, for any Asymptotically Stationary policy u, the limit $\eta(y,a) := \lim_{t\to\infty} P_x^u(X_t = y, A_t = a)$ exists for all y, a. We show by a direct argument that as the discount factor goes to one, the costs under Asymptotically Stationary policies converge to the expected average costs. This result implies that under A2, methods used to obtain optimal policies for the adaptive COP_1 can be used to obtain ϵ-optimal policies for the adaptive COP_β for β sufficiently close to one. This provides a new framework for adaptive problems with discounting — see Section 5. Previous results for discounted adaptive problems assume a Bayesian framework, whereas for the non-Bayesian framework, there are results only for the case where the discounted cost criterion is replaced by asymptotic cost criteria (see [20, 13,5]).

Theorem 4.2: *Under A1, if u is such that $\eta(y,a) := \lim_{t\to\infty} P_x^u(X_t = y, A_t = a)$ exists for all y, a, then*

$$\lim_{\beta\uparrow 1} C_\beta(x,u) = C_1(x,u)$$

Proof: For any integer N we have

$$\begin{aligned} C_\beta(x,u) =& (1-\beta)E_x^u \sum_{t=0}^{\infty} \beta^t c(X_t, A_t) \\ =& (1-\beta)E_x^u \sum_{t=0}^{N-1} \beta^t c(X_t, A_t) + \beta^N (1-\beta) E_x^u \sum_{t=N}^{\infty} \beta^{t-N} c(X_t, A_t) \end{aligned}$$

Let $c_{max} := \max_{y,a} |c(y,a)|$. For every N and ϵ there exists exists some $\beta_1(N,\epsilon)$ such that for every $\beta_1(N,\epsilon) < \beta < 1$

$$|(1-\beta)E_x^u \sum_{t=0}^{N-1} \beta^t c(X_t, A_t)| < \epsilon \tag{4.1}$$

since $|(1-\beta)\sum_{t=0}^{N-1} \beta^t c(X_t, A_t)| \le c_{max}(1-\beta^N)$. Note that $C_1(x,u) = \eta \cdot c$, so that the hypotheses imply

$$E_x^u c(X_t, A_t) = \sum_{y,a} c(y,a) P_x^u(X_t = y, A_t = a) \to \eta \cdot c$$

Therefor, for every $\epsilon > 0$ there exists some $N = N(\epsilon)$ such that

$$|1 - \beta^N| < \epsilon \tag{4.2}$$

$$|(1-\beta)E_x^u \sum_{t=N}^{\infty} \beta^{t-N} c(X_t, A_t) - C_1(x,u)| < \epsilon \tag{4.3}$$

uniformly in $0 < \beta < 1$. From (4.1)–(4.3) we have

$$|C_\beta(x,u) - C_1(x,u)| < \epsilon + (1 + c_{max})\epsilon$$

provided $\beta_1(N(\epsilon), \epsilon) < \beta < 1$ and the Theorem is established. ■

The policy u arises in relation to adaptive control, and in Section 5 we construct such a policy and discuss the importance of Theorem 4.2.

5. SENSITIVITY AND THE ADAPTIVE PROBLEM

In this Section we discuss the stability of optimal policies to small changes in the transition matrix and in the instantaneous cost functions. The motivation for this question comes from the adaptive constrained optimization problem **ACOP**, where the following assumptions on the cost structure and the transitions are made:

(i) The instantaneous costs $c(x,a)$, $d(x,a) := \{d^k(x,a)\ ,\ k = 1, ..., K\}$ are replaced by random costs. We denote by c_t and d_t the random costs obtained at time t. Given that $X_t = y, A_t = a$, we assume these costs are statistically independent of the history of the process (including past costs) and are generated from the probability distributions $Q_c(y,a)$ and $Q_d(y,a)$ respectively. We assume that these distributions are unknown. $C_\beta(x,u)$ and $D_\beta^k(x,u)$ are defined as before. Note that due to the definition of the average costs, they depend only on the means, which we denote as $c(y,a) := \int c_t dQ_c(y,a)$ and $d(y,a) := \int d_t dQ_d(y,a)$.

(ii) The values of the transition probabilities P_{xay} are unknown, except that assumption $A1'$ is known to hold.

The objective is still to find an optimal policy for $\mathbf{COP_1}$, but based on the available information, and without using any a-priori information about the values of the $\{P_{xay}\}$ and the means of the costs. In choosing the control to be used at time t, the only available information are the observed values $\{H_{t-1}, X_t\}$, where in this case $H_t := \{X_s, A_s, c(X_s, A_s), d(X_s, A_s),\ 0 \le s \le t\}$. The paradigm in adaptive control is then to use the observations to obtain information about the values of $\{P_{xay}\}$ and about the expected values of the costs, leading to computation of an adaptive policy. The notation P_{xay} here stands for the *true but unknown* values of the transition probabilities, and similarly for the notation P, E and π^g.

The estimator of the transition probabilities at time t is defined by:

$$\hat{P}^t_{zay} := \frac{\sum_{s=1}^t 1\{X_{s-1} = z, A_{s-1} = a, X_s = y\}}{\sum_{s=1}^t 1\{X_{s-1} = z, A_{s-1} = a\}}$$

If the denominator is zero then $\hat{P}^t$ is chosen arbitrarily, but such that for every z and a, $\hat{P}^t_{zay}$ is a probability distribution.

The expectation of the costs $c(\cdot,\cdot)$ and $d(\cdot,\cdot)$ with respect to Q_c and Q_d respectively is sufficient for computing an optimal adaptive policy. The estimate of this expectation at time t is given by

$$\hat{c}_t(y,a) := \frac{\sum_{s=1}^t c_s 1\{X_s = y, A_s = a\}}{\sum_{s=1}^t 1\{X_s = y, A_s = a\}}$$

$$\hat{d}^k_t(y,a) := \frac{\sum_{s=1}^t d^k_s 1\{X_s = y, A_s = a\}}{\sum_{s=1}^t 1\{X_s = y, A_s = a\}}$$

In order that

$$\lim_{t\to\infty} \hat{P}^t_{zay} = P_{zay} \quad P^u_x \ a.s. \tag{5.1a}$$

$$\lim_{t\to\infty} \hat{c}_t(y,a) = c(y,a) \quad P^u_x \ a.s. \tag{5.1b}$$

$$\lim_{t\to\infty} \hat{d}_t(y,a) = d(y,a) \quad P^u_x \ a.s. \tag{5.1c}$$

for all states $z, y \in \mathbf{X}$ and all $a \in \mathbf{A}$, it is sufficient that each state is visited infinitely often and moreover, that at each state, each action is used infinitely often P^u_x a.s. If this condition is met, then (5.1) holds by the strong law of large numbers, see [3].

By [2, Cor. 5.3] or [4, Lemma 3.2], each state y is indeed visited infinitely often P^u_x a.s. under any policy u. By an appropriate "probing" (described below) we shall obtain $\sum_{s=1}^\infty 1\{X_{s-1} = z, A_{s-1} = a\} = \infty \quad P^u_x \ a.s.$ for all $z \in \mathbf{X}$, $a \in \mathbf{A}$, implying consistent estimation.

Based on these estimations, two different algorithms are presented in [3] and [5] to compute optimal adaptive policies for the case that the instantaneous cost are not random. These algorithms can be generalized to our case. They involve solving a sequence of LP. We present below a generalization of these LPs that takes into account the fact that the instantaneous costs are unknown. Fix an increasing sequence of times T_n with $T_1 = 0$. T_n are the times at which we update the estimate.

$\mathbf{LP3_n}$: Find some $\{z_n^*(y,a)\}_{y,a}$ that minimizes $\mathcal{C}_n(z) := \sum_{y,a} \hat{c}_{T_n}(y,a) z(y,a)$ subject to:

$$\sum_{y,a} z(y,a) \left[\delta_v(y) - \hat{P}_{yav}^{T_n} \right] = 0$$

$$\mathcal{D}_n^k(z) := \sum_{y,a} \hat{d}_{T_n}^k(y,a) z(y,a) \leq V_k \qquad 1 \leq k \leq K$$

$$z(y,a) \geq 0 \qquad \sum_{y,a} z(y,a) = 1$$

In order to obtain an optimal adaptive policy, we need to verify that

(i) The estimators of the parameters entering $LP3_n$ are consistent,

(ii) Convergence of the estimators implies the convergence of z_n^* to z^*, the optimal solution of $LP(1)$, and

(iii) The policy u constructed from z_n^* is indeed optimal.

That the following algorithm indeed satisfies these requirements is the subject of Theorem 5.3 below. Item (ii) is exactly a sensitivity problem, to which Theorem 3.6 applies.

Fix a decreasing sequence $\epsilon_r \downarrow 0$ such that $\sum_1^\infty \epsilon_r = \infty$. Let $\tau(y;r)$ denote the time of the rth visit to state y. In the following Algorithm we construct an adaptive policy u for ACOP. For the case that all instantaneous costs are known, it was shown in [5,4] that u is optimal for ACOP.

Algorithm 5.1: If $t = \tau(y;r)$ and $T_n \leq t < T_{n+1}$ then set for every action $a \in \mathbf{A}$:

$$u_t(a|H_{t-1}, X_t) = \frac{\epsilon_n}{|\mathbf{A}|} + (1 - \epsilon_n) \cdot \gamma_{X_t}^a(z_n^*) \tag{5.2}$$

Remark: the first term in the right side ensures that each action is chosen infinitely often, so that the estimation is consistent. The second term is an approximation of the optimal policy.

We need below the following assumption:

A3: The non-adaptive problem COP has a single optimal stationary policy (or equivalently, there is only one optimal solution of LP(1)).

Theorem 5.2: *Assume A1′, A2 and A3. If* $\hat{P}^{T_r} \to P$ *w.p.1. implies that* z_n^* *converges to* z^**, then the policy* u *obtained through Algorithm 5.1 is asymptotically stationary with limit* $\eta = \bar{f}_{sa}(x, g(z^*))$ *and is optimal for ACOP.*

Proof: This is a slight modification of the results in [5, 3]. ■

From this and Theorem 3.6 we obtain

Theorem 5.3: *Under A1′, A2 and A3 Algorithm 5.1 is optimal for ACOP.*

Finally we obtain the following Corollary to Theorem 4.2.

Corollary 5.4: *Assume that the conditions of Theorem 5.2 hold, and consider the Asymptotically Stationary policy* u *generated by Algorithm 5.1. Then for every* $\epsilon > 0$ *there exists some* $\beta'(u, \epsilon) < 1$ *such that* u *is* ϵ*-optimal for the adaptive* COP_β *for every* $\beta'(u, \epsilon) < \beta < 1$.

See also the remark preceding Theorem 4.2.

6. CONVERGENCE OF FINITE HORIZON PROBLEMS

In the discounted unconstrained case, it is immediate that the optimal cost of the finite-time problem converges to that of the infinite-horizon problem. The presence of constraints complicates the issue, since the optimal control of the infinite-horizon problem may not be feasible for the finite case, and vice versa.

Theorem 6.1: *Assume A1, A2. Then for any* $0 < \beta < 1$, $\lim_{t\to\infty} C_\beta^t(x) = C_\beta(x)$.

Proof: Assume that $c(\cdot,\cdot)$ and $d^k(\cdot,\cdot)$ are all nonnegative. If this is not the case then the proof will follow by considering the problem with all immediate costs and all constraints shifted by a large enough real number. We first show that

$$\varliminf_{t\to\infty} C_\beta^t(x) \geq C_\beta(x) \tag{6.1}$$

Assume (6.1) does not hold. Then there exists some $\epsilon > 0$ such for every $T > 0$ there exists some $t > T$ and a policy u^t where u^t is feasible for COP_β^t and

$$C_\beta^t(x, u^t) < C_\beta(x) - \epsilon \tag{6.2}$$

It follows from A2 that there exists a policy v and some positive real number η such that $D^k(\beta, v) < V_k - \eta$, for all $1 \leq k \leq K$. Pick some $0 < \alpha < 0.5$ such that $\alpha C_\beta(x, v) < \epsilon/4$, and $\delta > 0$ that satisfies $0 < \delta < min\{\frac{\epsilon}{2}\ , \eta\alpha\}$. Fix T such that $\beta^T < \delta$ and

$$\beta^T max_{a,y,k}\{d^k(y,a)\}|\mathbf{A}||\mathbf{X}| < \delta \tag{6.3}$$

$$\beta^T max_{a,y}\{c(y,a)\}|\mathbf{A}||\mathbf{X}| < \delta \tag{6.4}$$

Let $t > T$ be such that (6.2) holds. For any policy u and every y, a we have

$$0 \leq \frac{1}{1-\beta}\bar{f}^\beta_{sa}(x,u;y,a) - [\sum_{s=0}^{t}\beta^s]\bar{f}^{t,\beta}_{sa}(x,u;y,a) \leq \frac{\beta^t}{1-\beta} \tag{6.5}$$

and hence

$$\bar{f}^\beta_{sa}(x,u;y,a) < (1-\beta)[\sum_{s=0}^{t}\beta^s]\bar{f}^{t,\beta}_{sa}(x,u;y,a) + \beta^t \leq \bar{f}^{t,\beta}_{sa}(x,u;y,a) + \beta^t \tag{6.6}$$

From (6.3) and (6.6) we get

$$D^k_\beta(x,u^t) = \bar{f}^\beta_{sa}(x,u^t)\cdot d^k \leq \bar{f}^{t,\beta}_{sa}(x,u^t)\cdot d^k + \beta^t\sum_{y,a}d^k(y,a) < D^{k,t}_\beta(x,u^t) + \delta$$

and similarly from (6.4) and (6.6)

$$C_\beta(x,u^t) = \bar{f}^\beta_{sa}(x,u^t)\cdot c \leq \bar{f}^{t,\beta}_{sa}(x,u^t) + \beta^t\sum_{y,a}c(y,a) < C^t_\beta(x,u^t) + \delta$$

We then obtain

$$C_\beta(x,u^t) \leq C_\beta(x) + \delta - \epsilon \tag{6.7}$$

$$D^k_\beta(x,u^t) \leq D^{k,t}_\beta(x,u^t) + \delta \tag{6.8}$$

From Lemma 2.2 it follows that the set of frequencies that are achievable by the stationary policies is closed, convex and contains L_x. Therefore there exists some stationary policy u'' such that $\bar{f}^\beta_{sa}(x,u^t) = \bar{f}^\beta_{sa}(x,u'')$ and hence a stationary policy u' such that

$$\bar{f}^\beta_{sa}(x,u') = (1-\alpha)\bar{f}^\beta_{sa}(x,u^t) + \alpha\bar{f}^\beta_{sa}(x,v) \tag{6.9}$$

It follows from (6.8) and (6.9) that u' is feasible since

$$D^k_\beta(x,u') = \bar{f}^\beta_{sa}(x,u')\cdot d^k \leq (1-\alpha)(V_k+\delta) + \alpha(V_k-\eta)$$

$$= V_k + (1-\alpha)\delta - \alpha\eta < V_k + \delta - \alpha\eta < V_k$$

From (6.7) and (6.9) we obtain

$$C_\beta(x, u') = \bar{f}^\beta_{sa}(x, u') \cdot c = (1-\alpha)C_\beta(x, u^t) + \alpha C_\beta(x, v)$$

$$< (1-\alpha)(C_\beta(x) + \delta - \epsilon) + \epsilon/4 \leq C_\beta(x) + (\delta - \epsilon)/2 + \epsilon/4 < C_\beta(x)$$

But this contradicts the definition of $C_\beta(x)$, which proves (6.1). Next we prove that

$$\overline{\lim_{t\to\infty}}\, C^t_\beta(x) \leq C_\beta(x) \tag{6.10}$$

Let u^* be any policy which is optimal for COP_β. Pick a stationary policy u_ϵ such that

$$\bar{f}^\beta_{sa}(x, u_\epsilon) = (1-\epsilon)\bar{f}^\beta_{sa}(x, u^*) + \epsilon \bar{f}^\beta_{sa}(x, v) \tag{6.11}$$

It follows from (6.11) and the linear representation of the cost (see 2.3d) that

$$\overline{\lim_{t\to\infty}}\, C^t_\beta(x, u_\epsilon) = C_\beta(x, u_\epsilon)$$

$$\overline{\lim_{t\to\infty}}\, D^{k,t}_\beta(x, u_\epsilon) = D^k_\beta(x, u_\epsilon) \leq V_k - \epsilon\eta$$

for all $1 \leq k \leq K$. Therefore for every ϵ, u_ϵ is feasible for COP^t_β for all t large enough. Since u_ϵ may not be optimal for COP^t_β, we clearly have

$$\overline{\lim_{t\to\infty}}\, C^t_\beta(x) \leq \overline{\lim_{t\to\infty}}\, C^t_\beta(x, u_\epsilon) \leq (1-\epsilon)C_\beta(x) + \epsilon \sum_{y,a} c(y, a)$$

where the last inequality follows again from (6.11) and the linear representation of the cost (see 2.3c). Since this holds for every ϵ, (6.10) follows. This completes the proof. ∎

Note that in addition to proving the convergence of the optimal cost, the proof of Theorem 6.1 provides a way to construct ϵ-optimal policies for the finite problem using the optimal policy to the infinite-horizon problem and vice versa. The main point is that the policies need to be modified in order to guarantee feasibility.

REFERENCES

[1] Altman E. and A. Shwartz, "Non-stationary policies for controlled Markov Chains", EE PUB. 633, Technion, June 1987, under revision.

[2] Altman E. and A. Shwartz, "Markov decision problems and state-action frequencies", EE. PUB. No. 693, Technion, November 1988, under revision.

[3] Altman E. and A. Shwartz, "Adaptive control of constrained Markov chains", EE. PUB. No. 713, Technion, March 1989, under revision.

[4] Altman E. and A. Shwartz, "Adaptive Control of constrained Markov chains", Trans. of the 14th Symposium on Operations Research, Ulm, Germany, September 1989.

[5] Altman E. and A. Shwartz, "Adaptive Control of constrained Markov chains: Criteria and Policies", submitted to Annals of Oper. Res., 1989.

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

[7] Borkar V. S., "Controlled Markov chains with constraints", preprint (revised), 1989.

[8] Dantzig G. B., J. Folkman and N. Shapiro, "On the continuity of the minimum set of a continuous function", *J. Math. Anal. and Applications,* Vol. 17, pp. 519-548, 1967.

[9] Dekker R., *Denumerable Markov decision chains: optimal policies for small interest rates*, Thesis, Inst. for Applied Math. and Comp. Science, University of Leiden, 1984.

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

[11] C. Derman M. Klein, "Some remarks on finite horizon Markovian decision models", Oper. Res. Vol. 13, pp. 272-278, 1965.

[12] Heilmann W.-R. "Solving Stochastic Dynamic Programming problems by Linear Programming — an annotated bibliography", Zeit. Oper. Res. 22 pp. 43-53, Physica-Verlag, 1978.

[13] Hernandez-Lerma O., *Adaptive Control of Markov Processes*, Springer Verlag, 1989.

[14] Hordijk A. and L. C. M. Kallenberg, "Constrained undiscounted stochastic dynamic programming", *Mathematics of Operations Research,* Vol. 9, No. 2, May 1984.

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

[16] Manne A. S. "Linear Programming and Sequential decisions", MS 6, pp. 259-267, 1960.

[17] Nain P. and K. W. Ross, "Optimal Priority Assignment with hard Constraint," *IEEE Trans. on Automatic Control,* Vol. 31 No. 10, October 1986.

[18] Ross K. W., "Randomized and past-dependent policies for Markov decision processes with

multiple constraints", *Operations Research,* Vol. 37, No. 3, May 1989.

[19] Ross K. W. 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, March 1988.

[20] Schäl M., "Estimation and control in discounted dynamic programming," *Stochastics* **20**, pp. 51-71, 1987.

[21] Shwartz A. and A. M. Makowski, "An optimal adaptive scheme for two competing queues with constraints", in *Analysis and optimization of systems* pp. 515-532, Edited by A. Bensoussan, J. L. Lions, Springer Verlag Lect. Notes Control and Info. Sci., 1986.