Asymptotic properties of constrained Markov decision processes

Eitan Altman

UNITÉ DE RECHERCHE
INRIA-SOPHIA ANTIPOLIS

Rapports de Recherche

N°1598

Programme 1
Architectures parallèles, Bases de données,
Réseaux et Systèmes distribués

ASYMPTOTIC PROPERTIES OF CONSTRAINED MARKOV DECISION PROCESSES

Eitan ALTMAN

Février 1992

Institut National
de Recherche
en Informatique
et en Automatique

Sophia Antipolis
B.P. 109
06561 Valbonne Cedex
France
Tél.: 93 65 77 77

ASYMPTOTIC PROPERTIES OF CONSTRAINED MARKOV DECISION PROCESSES *

Eitan Altman
INRIA, Centre Sophia Antipolis
06565 Valbonne Cedex, France
Tel: 93 96 76 37.

January 1991

Abstract

We present in this paper several asymptotic properties of constrained Markov Decision Processes (MDPs) with a countable state space. We treat both the discounted and the expected average cost, with unbounded cost. We are interested in (1) the convergence of finite horizon MDPs to the infinite horizon MDP, (2) convergence of MDPs with a truncated state space to the problem with infinite state space, (3) convergence of MDPs as the discount factor goes to a limit. In all these cases we establish the convergence of optimal values and policies. Moreover, based on the optimal policy for the limiting problem, we construct policies which are almost optimal for the other (approximating) problems.

Keywords: Constrained Markov Decision Problems, countable state space, finite horizon, infinite horizon, finite approximations, asymptotic properties.

1 INTRODUCTION

In recent years growing attention was given to solving constrained MDPs (Markov Decision Problems). Such problems frequently arise in computer networks and data communications,

*This work was supported by the Chateaubriand fellowship from the french embassy in Israel

see Lazar [17], Spieksma and Hordijk [12], Nain and Ross [18] and Altman and Shwartz [1]. The theory for solving constrained MDPs was developed by Hordijk and Kallenberg [15], Kallenberg [16], Beutler and Ross [8], Altman and Shwartz [2, 4], Altman [6], Spieksma [21], Senott [19, 20] and Borkar [9].

We are interested in (1) the convergence of finite horizon MDPs to the infinite horizon MDP, (2) convergence of MDPs with a truncated state space to the problem with infinite state space, (3) convergence of MDPs as the discount factor goes to a limit. In all these cases we establish the convergence of optimal values and policies. Moreover, based on the optimal policy for the limiting problem, we construct policies which are almost optimal for the other (approximating) problems.

In the finite horizon case, the need for approximation arrizes since there are no known techniques to solve such problems, unlike the infinite horizon case. Of special importance is therefore to construct a policy that is almost optimal for the finite horizon case, based on a policy that is optimal (or almost optimal) for the infinite horizon case.

When the number of states and actions in the control problem is not finite, the question of finite approximation of the state is of interest, since the only general solution known for the constrained problem is through a LP (Linear Programming) with infinite number of decision variables (see Altman and Shwartz [2], Altman [6] and Spieksma [21]). In the finite case, however, a finite LP can be applied to obtain an optimal policy (see Hordijk and Kallenberg [15] and Altman and Shwartz [4]). The issue of approximating MDPs by other finite state MDPs was investigated in a few papers of White (e.g. [23]), Cavazos-Cadena [11], Hernandez-Lerma [13] and Thomas and Stengos [22]. Since in our case additional constraints are introduced in the control the methods used in tha above papers are not applicable, as they are all based on dynamic programming techniques. In Altman [6], a general theory for approximation for the constrained problem was developed and applied for finite approximations. The cost in Altman [6] is assumed to be non decreasing and some stochastic monotonicity assumptions on the transition probabilities are made. In this paper, in Section 5, we generalize these results to costs that are not necessarily monotone, and to transition probabilities without the stochastic monotonicity property. Yet, we restrict to transitions to "nearest neighbors", i.e., from each state we assume that it is possible to move in one step to only a finite number of states.

The third issue that we treat the convergence of MDPs as the discount factor goes to a limit. This is of special interest for the case that the discount factor tends to one. Indeed, in control problems of queueing systems without constraints, an optimal policy for the expected average cost has often been obtained as the limit of optimal policies for the discounted cost, as the discount factor goes to one. By proving the continuity of MDPs in the discount factor, we thus establish the validity of this approach to constrained MDPs as well.

We finally illustrate the above ideas by the problem of optimal priority assignment where N infinite queues compete for the attention of a single server.

Previous results on asymptotic properties of constrained MDPs have been obtained by Altman and Shwartz [4] and Altman and Gaitsgory [5] for the case of finite state space. They obtain conditions for the continuity of the optimal value and policy of MDPs in the immediate cost, transition probabilities and discount factor. These results are obtained by applying the theory of stability of finite LPs. This method is however not extendible to the case of infinite state space. An alternative method is used by Altman and Shwartz [4] that does not rely on a LP approach in order to obtain the convergence of the optimal value as the horizon tends to infinity for the case of discounted cost, and finite state space. The Theory of convergence of finite state constrained MDPs is applied in Altman and Shwartz [3, 4] to solve adaptive control problems. A characterization and a study of cases were the optimal values and policies of MDPs are not continuous is presented by Altman and Gaitsgory in [5].

The structure of the paper is as follows: after presenting the model notation and assumptions in Section 2, we cite in Section 3 key Theorems for approximation. We analyze the finite horizon problem in Section 4, and then in Section 5 we introduce and analyze the finite state approximation. The continuity of MDPs in the discounted cost is studied in Section 6. The problem of optimal priority assignment for N competing queues is finally treated in Section 7.

2 Model and Assumptions

2.1 The model

Let $\{X_t\}_{t=0}^{\infty}$ be the discrete time state process, defined on the countable *state space* $\mathbf{X}$; the action A_t at time t takes values in the countable *action space* $\mathbf{A}$. However, we assume that at each state $y \in \mathbf{X}$ there is only a finite number of available actions, $\mathbf{A}(y)$. With some abuse of notation, $\mathbf{X} \times \mathbf{A}$ will denote all possible pairs $(y,a) : y \in \mathbf{X}, a \in \mathbf{A}(y)$. The *history* of the process up to time t is denoted by $\mathcal{H}_t := (X_0, A_0, X_1, A_1, ..., X_t, A_t)$. The dynamics is given by

$$\mathcal{P}_{xay} := P(X_{t+1} = y \mid X_t = x; A_t = a) = P(X_{t+1} = y \mid \mathcal{H}_{t-1} = h, X_t = x; A_t = a) \qquad (2.1)$$

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

A *stationary policy* $g \in U(S)$ is characterized by a single conditional distribution $p^g_{\bullet|x} := u(\bullet \mid X_t = x)$ 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 $P^g_{xy} := \sum_{a \in \mathbf{A}} p^g_{a|x} \mathcal{P}_{xay}$. For $g \in U(S)$, let $\Pi^g := \lim_{n \to \infty} [P^g]^n$. If all rows of Π^g are equal, we denote them by π^g.

The class of *stationary deterministic policies* $U(SD)$ is a subclass of $U(S)$ and, with some abuse of notation, every $g \in U(SD)$ is identified with a mapping $g : \mathbf{X} \to \mathbf{A}$, so that $p^g_{\cdot|x} = \delta_{g(x)}(\cdot)$ is concentrated at the point $g(x)$ in $\mathbf{A}$ for each x.

For any (finite or countable) set B, let $\mathbf{M}(B)$ denote the set of probability measures on B.

2.2 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 an initial state x. The real vector $V := \{V_k\ ,\ k = 1, ..., K\}$ is held fixed hereafter.

Call a policy u *feasible* if

$$D^k(x,u) \le V_k \ , \quad k = 1,2,\ldots,K \tag{2.2}$$

The constrained optimization problem is:

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

$C(x,u)$ and $D^k(x,u)$ will stand for either the discounted or the expected average cost, defined below. 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 assume throughout the paper that $\forall u \in U,\ x \in \mathbf{X}$, $E_x^u c(X_s,A_s)$ and $E_x^u d^k(X_s,A_s),\ k = 1,...,K$ exist. We define two versions of discounted cost functionals from $\mathbf{X} \times U$ to $\mathbb{R}$ (see Altman and Shwartz [4]).

$$\begin{aligned} C_\beta^t(x,u) &:= (\textstyle\sum_{s=0}^t \beta^s)^{-1} E_x^u \left[\sum_{s=0}^t \beta^s c(X_s,A_s)\right] \\ D_\beta^{k,t}(x,u) &:= (\textstyle\sum_{s=0}^t \beta^s)^{-1} E_x^u \left[\sum_{s=0}^t \beta^s d^k(X_s,A_s)\right] \qquad k = 1,...,K \end{aligned} \tag{2.3}$$

$$\begin{aligned} \tilde{C}_\beta^t(x,u) &:= (1-\beta) E_x^u \left[\textstyle\sum_{s=0}^t \beta^s c(X_s,A_s)\right] \\ \tilde{D}_\beta^{k,t}(x,u) &:= (1-\beta) E_x^u \left[\textstyle\sum_{s=0}^t \beta^s d^k(X_s,A_s)\right] \qquad k = 1,...,K \end{aligned} \tag{2.4}$$

$$\begin{aligned} C_\beta(x,u) &:= \overline{\lim}_{t\to\infty} C_\beta^t(x,u) \\ D_\beta^k(x,u) &:= \overline{\lim}_{t\to\infty} D_\beta^{k,t}(x,u) \qquad k = 1,...,K \end{aligned} \tag{2.5}$$

When $\beta = 1$, (2.5) reduce to the well-known definition of the expected average cost, since in that case $\sum_{s=0}^t \beta^s = t+1$. For $\beta < 1$ the definition (2.5) are a normalized version of the *standard expected discounted* costs

$$C_\beta(x,u) := (1-\beta) E_x^u \left[\sum_{s=0}^\infty \beta^s c(X_s,A_s)\right] = \lim_{t\to\infty} C_\beta^t(x,u) = \lim_{t\to\infty} \tilde{C}_\beta^t(x,u) \tag{2.6}$$

and similarly for $D_\beta(x,u)$. The addition of a normalizing factor, e.g. in (2.3) is important when one is interested in continuity of the (finite horizon) cost in the discount parameter, especially at $\beta = 1$. This is especially important in Section 6. Another advantage of using the new definition (2.5) is that it enables to obtain the same LP for both the discounted and expected average cost

for solving COP (see e.g. Altman [6]) This enables to reduce problems of continuity, sensitivity and singular perturbations of constrained MDPs in the discount parameter β as $\beta \to 1$, to the corresponding problems in Mathematical Programs for which the theory is well known (see e.g. Altman and Shwartz [4]).

Finally, for $u \in U(S)$, define $c(u) = \{c(y,u)\}$, $y \in \mathbf{X}$, where $c(y,u) = \sum_a c(y,a) p^u_{a|y}$.

2.3 Assumptions and Notation

Let $\mu : \mathbf{X} \to \mathbb{R}$ be some function. Following Dekker and Hordijk [10] and Spieksma [21], define the μ-norm of vectors $\pi \in \mathbf{X}$ and of matrices $P \in \mathbf{X} \times \mathbf{X}$ as

$$\begin{aligned} \|\pi\|_\mu &= \sup_{x \in \mathbf{X}} \mu_x^{-1} |\pi_x| \\ \|P\|_\mu &= \sup_{x \in \mathbf{X}} \mu_x^{-1} \sum_{y \in \mathbf{X}} |P_{xy}| \mu_y \end{aligned} \tag{2.7}$$

For a subset $M \subset \mathbf{X}$, let ${}_M P$ be the taboo matrix corresponding to P, i.e.

$${}_M P_{xy} = \begin{cases} P_{xy}, & y \notin M \\ 0, & y \in M \end{cases} \tag{2.8}$$

A MDP is said to be μ-uniform geometric recurrent (μ-UGR), if a finite set M and a $\alpha < 1$ exist, such that for any $u \in U(SD)$,

$$\|{}_M P^u\|_\mu \leq \alpha. \tag{2.9}$$

The following assumptions are used frequently in the paper:

B1: Under any $g \in U(SD)$, $\mathbf{X}$ consists of a single ergodic aperiodic class, (with no transient states).

B2: there exists some policy u (not necessarily an optimal policy) such that

$$D^k(x,u) < V_k \ , \quad k = 1, 2, \ldots, K \tag{2.10}$$

B3: The family of stationary probabilities $\{\pi^u(\cdot)\}$, $\pi^u \in \mathbf{M}(\mathbf{X})$, corresponding to policies $u \in U(S)$, is tight.

B4(i): There exists some function $\mu \in \mathbf{X} \to \mathbb{R}$ such that if $\beta = 1$ then $\mu_y \geq 1$ and the MDP corresponding to the dynamics $\{\mathcal{P}_{xay}\}$ is μ-UGR; if $\beta < 1$ then $\exists \alpha$, $0 < \alpha < 1$ such that for all $u \in U(SD)$, $\|\beta P\|_\mu \leq \alpha$.

B4(ii): both $c(\cdot,\cdot)$ $d^k(\cdot,\cdot)$, $k = 1, ..., K$ are bounded below, and $\exists R < \infty$ such that $\|\max_a c(\bullet, a)\|_\mu \leq R$, $\left\|\max_a d^k(y,a)\right\|_\mu \leq R$, $k = 1, ..., K$.

B5: For any x, a, $\mathcal{P}_{xa\bullet}$ has finite support (that may depend on x, a).

We shall often need the following result, extending B4(i) to $U(S)$.

Lemma 2.1 *Assume B4(i). Then for any $u \in U(S)$,*

$$\begin{aligned} \|_M P^u\|_\mu &\leq \alpha, \quad \text{if } \beta = 1 \\ \|\beta P^u\|_\mu &\leq \alpha, \quad \text{if } \beta < 1 \end{aligned} \tag{2.11}$$

Proof: For any $u \in U(S)$,

$$\|P^u\|_\mu = \sup_{x \in \mathbf{X}} \mu_x^{-1} \sum_{y \in \mathbf{X}} |\sum_a P_{xay} p^u_{a|x}| \mu_y \leq \sup_{x \in \mathbf{X}} \mu_x^{-1} \sum_{y \in \mathbf{X}} |\max_a P_{xay} p^u_{a|x}| \mu_y = \|P^w\|_\mu \tag{2.12}$$

where $w \in U(SD)$ is a policy that achieves the max in 2.12. ■

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, $|B|$ the cardinality of this set (if B is finite then $|B|$ is 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, $\zeta \cdot Q$ stands for summation over common indices (scalar product). Q^T denotes the transposed of the matrix Q.

3 Key Theorems for Approximation

In this Section we quote some results on approximations of constrained Markov Problems. Assume that for any initial state x and policy u we have a sequence of approximations $C_n(x,u)$ and $D_n^k(x,u)$, $n = 1,2,...$ of the costs $C(x,u)$ and $D^k(x,u)$ with $k = 1,2,...,K$. Below $C(x,u)$ and $D^k(x,u)$ will stand for either the expected average cost $\beta = 1$ or for the discounted cost $\beta < 1$, where as $C_n(x,u)$ and $D_n^k(x,u)$, may be **arbitrary functions** from $\mathbf{X} \times U$ to $\mathbb{R}^+$.

Consider the following sequence of problems:

$\mathbf{COP_n}$: Find $C_n(x)$ which is given by:

$C_n(x) := \inf_{u \in U}\{C_n(x,u); D_n^k(x,u) \leq V_k,\ k = 1,2,...,K\}$.

Let $C(x)$ be the optimal value of COP. Note that COP_n may not have any optimal policy even if it is feasible. Moreover, there may not exist any ϵ-optimal stationary policy for COP_n. In Theorem 3.1 below,

Theorem 3.1 *Assume*

(1) B2 and that $c(\cdot,\cdot)$ and $d^k(\cdot,\cdot)$ are bounded below.

(2) $\lim_{n\to\infty} C_n(x,u) = C(x,u)$ and $\lim_{n\to\infty} D_n^k(x,u) = D^k(x,u)$ for every stationary policy u and a given initial state x, uniformly in $u \in U(S)$.

(3) $\beta < 1$, or $\{\beta = 1$ and B1 holds$\}$.

Then (i) $\lim_{n\to\infty} C_n(x) = C(x)$.

(ii) Choose a sequence $\xi_n \to 0$. Let $r(n)$ be a ξ_n-optimal policy for COP_n if COP_n is feasible, otherwise let it be an arbitrary stationary policy. Let $w(n)$ be a stationary policy that satisfies

$$C_\beta(x,w(n)) \leq C_\beta(x,r(n)), \qquad D_\beta^k(x,w(n)) \leq D_\beta^k(x,r(n)), \quad k = 1,2,...,K. \tag{3.1}$$

Let w be an arbitrary accumulation point of $w(n)$, $n = 1,2,...$ i.e. there exists a subsequence $\{n_i\}_{i=1}^\infty$ such that for all $x \in \mathbf{X}, a \in \mathbf{A}$,

$$\lim_{i\to\infty} p_{a|y}^{w(n_i)} = p_{a|y}^{w}. \tag{3.2}$$

Then w is optimal for COP.

Proof: The Proof of the Theorem is given in Altman [6] under slightly different conditions for $\beta = 1$. It follows from Proposition 5.1 (vi) p. 97 of Flos [21] and Lemma 2.1 that B3 holds. The condition A4 there follows from B1, and condition A3(i) is implied by B1 and B3, see Hordijk [14] Section 10.

■

Next we construct a policy which is almost optimal for COP_n for all n large enough. Let u^* be a policy for which (1). the following limit exists

$$\bar{f}^{\beta}(x,u;y,a) := \lim_{t\to\infty}[\sum_{s=0}^{t}\beta^s]^{-1}\sum_{s=0}^{t}\beta^s P_x^u(X_s = y, A_s = a) \tag{3.3}$$

(2)

$$C(x,u) = c\cdot\bar{f}(x,u), \qquad D^k(x,u) = d^k\cdot\bar{f}(x,u), \quad k = 1,...,K, \tag{3.4}$$

and (3) u^* is ϵ-optimal for COP. Let $u(\epsilon)$ be the policy that satisfies (3.3), the linear representation (3.4) and such that

$$\bar{f}(x,u(\epsilon)) = (1-\epsilon)\bar{f}(x,u^*) + \epsilon\bar{f}(x,v) \tag{3.5}$$

(The existence of these policies under the conditions of Theorem 3.1 is established in Altman [6]).

Theorem 3.2 *Let $-M$, $M \geq 0$ be a lower bound on the immediate costs c. Under the conditions of Theorem 3.1, $u(\epsilon)$ defined in (3.5) is $\hat{\epsilon}$-optimal for COP_n for all n large enough, where $\hat{\epsilon} = \epsilon[C(x,v) + M + 3]$.*

4 The finite horizon case

Let COP_n denote the COP with finite horizon costs $C_\beta^n(x,u)$ and $D_\beta^n(x,u)$. Let $C_\beta^n(x)$ be the optimal value for COP_n.

Theorem 4.1 *Assume B2, B4, and either B1 or $\beta < 1$. Then*
(1) $\lim_{n\to\infty} C_\beta^n(x) = C_\beta(x)$.
(2) Choose some $\epsilon > 0$. Let u^* be an ϵ-optimal (or optimal) policy for COP that satisfies (3.4) (e.g. any ϵ-optimal stationary policy). There exists some $N(\epsilon)$ such that for all $n \geq N(\epsilon)$, the policy $u(\epsilon)$ satisfying (3.4) and (3.5) is $\hat{\epsilon}$-optimal for COP_n, where $\hat{\epsilon}$ is given in Theorem 3.2.
(3) Choose a sequence $\xi_n \to 0$. Let $r(n)$ be a ξ_n-optimal policy for COP_n if COP_n is feasible, otherwise let it be an arbitrary stationary policy. The stationary policy w obtained by applying the limiting procedure in Theorem 3.1 (ii) to the policies $r(n)$ is optimal for COP.

Proof: Assume first $\beta < 1$. Then,

$$\left\|\tilde{C}_\beta^n(\bullet,u) - C_\beta(\bullet,u)\right\|_\mu \leq (1-\beta)\beta^{n+1}\left\|\sum_{k=0}^{\infty}\beta^k[P^u]^k\right\|_\mu \|c(u)\|_\mu \leq \frac{\beta^{n+1}(1-\beta)R}{1-\alpha} \tag{4.1}$$

$$C_\beta^n(\bullet,u) - \tilde{C}_\beta^n(\bullet,u) = (1-\beta)C_\beta^n(\bullet,u)\left([1-\beta]^{-1} - \sum_{k=0}^{n}\beta^l\right) = (1-\beta)\beta^{n+1}C_\beta^n(\bullet,u) \tag{4.2}$$

and hence

$$\left\|C_\beta^n(\bullet,u) - \tilde{C}_\beta^n(\bullet,u)\right\|_\mu \leq \frac{\beta^{n+1}(1-\beta)R}{1-\alpha} \tag{4.3}$$

Combining (4.1) and (4.3) we get

$$\left\|C_\beta^n(\bullet,u) - C_\beta(\bullet,u)\right\|_\mu \leq \frac{2\beta^{n+1}(1-\beta)R}{1-\alpha} \tag{4.4}$$

and similarly for $D_\beta^n(\bullet,u)$. Hence the conditions of Theorem 3.1 and 3.2 are satisfied, which establishes the proof for $\beta < 1$.

Next, let $\beta = 1$. B4 implies that $\exists r > 0$ s.t. $\|\Pi^u\|_\mu \leq r$, $\|[P^u]^n\|_\mu \leq r$, uniformly in $n \in \mathbb{N}$ and $u \in U(S)$. This follows from Proposition 5.1 p. 97 in Spieksma [21] for $u \in U(SD)$ and generalizes readily to $U(S)$ due to Lemma 2.1. The MDP satisfies μ-geometric ergodicity, i.e. $\exists \rho > 0$, $\alpha < 1$, s.t.

$$\begin{gathered}\|[P^u]^n - \Pi^u\|_\mu \leq \rho\alpha^n, \ \forall n \in \mathbb{N}, \\ \|P^u\|_\mu \leq \rho\end{gathered} \tag{4.5}$$

for all $u \in U(SD)$ (see Key Theorem I, p. 24 in Spieksma [21]). This easily extends to all $u \in U(S)$ by Lemma 2.1. Hence for all $u \in U(S)$ and $n \in \mathbb{N}$,

$$\|C_1^n(\bullet,u) - C_1(\bullet,u)\|_\mu \leq \sum_{k=0}^{n} \frac{\left\|[P^u]^k - \Pi^u\right\|_\mu}{n+1} \|c(u)\|_\mu \leq \frac{\rho R \sum_{k=0}^{n} \alpha^k}{n+1} \leq \frac{\rho R}{(n+1)(1-\alpha)}, \quad (4.6)$$

and similarly with $D_1(\bullet,u)$. Hence the conditions of Theorem 3.1 and 3.2 are satisfied, which establishes the proof. ■

It immediately follows from (4.1) that

Corollary 4.1 *Let $\beta < 1$. Then Theorem 4.1 holds also for the finite horizon discounted cost $\tilde{C}_\beta^n(x,u)$ replacing $C_\beta^n(x,u)$, and similarly with $\tilde{D}_\beta^n(x,u)$.*

5 Finite approximations

In this Section we consider the problem of approximating the Controlled Markov Chain (CMC) which is characterized by the dynamics (i.e. the transition probabilities) $\mathcal{P}_{xay}$ by a sequence of Controlled Markov Chains CMC_m governed by the dynamics $\{\mathcal{P}_{xay}(m)\}$, $m = 1, 2, ...$ We denote by $C_\beta(m;x,u)$ and $D_\beta(m;x,u)$ the costs (given in (2.5)) under policy u and initial state x corresponding to CMC_m and discount factor β $(0 < \beta \leq 1)$. Define similarly $C_\beta^n(m;x,u)$ and $D_\beta^n(m;x,u)$ the finite horizon costs given in (2.3). Let $C_\beta(m,x)$ denote the optimal value for COP_m.

We shall construct the CMC_m such that for all $x, y \in \mathbf{X}$, $a \in \mathbf{A}$ $\lim_{m\to\infty} \mathcal{P}_{xay}(m) = \mathcal{P}_{xay}$. We then show that the construction ensures that $\lim_{m\to\infty} C_m^\beta(x,u) = C^\beta(x,u)$ and $\lim_{m\to\infty} D_m^{k,\beta}(x,u) = D^{k,\beta}(x,u)$, $k = 1, ..., K$ uniformly in $u \in U(S)$ and hence by Theorem 3.1 $\lim_{n\to\infty} C_n^\beta(x) = C^\beta(x)$.

Introduce the following approximation scheme **FA**:
(i) For each $m = 0, ...$ the state space is decomposed in two disjoint classes of states: E^m, which contains a finite number of states, and T^m.

(ii) Under any stationary policy u, E^m is a recurrent class, T^m is a transient class, and absorption into the positive recurrent class takes place in finite expected time from any initial state.
(iii) $E_m \subset E_{m+1}$, $m = 1, ...$; $E_0 := \{\emptyset\}$; $E_\infty = \mathbf{X}$.
(iv) There is some partial order on $\mathbf{X}$; CMC is μ-UGR and μ is non decreasing in $\mathbf{X}$ w.r.t. the partial order.
(v) The following holds:

$$\mathcal{P}_{xay}(m) \begin{cases} = \mathcal{P}_{xay} & x \in E_m,\ y \in E_{m-1} \\ \geq \mathcal{P}_{xay} & x \in E_m,\ y \in E_m \backslash E_{m-1} \\ = 0 & x \in E_m,\ y \notin E_m \\ = 1\{y = 1\} & x \notin E_m \end{cases} \tag{5.1}$$

where 1 is some arbitrary state such that $1 \in \cap_m E_m$. Moreover, for every $m > 0$ and each $y \in E_m \backslash E_{m-1}$ and $x \in E_{m-1}$, we have $x \leq y$ w.r.t. the partial order in (iv), if x and y are comparable.

For $u \in U(S)$, denote $P^u_{xy}(m) := \sum_a \mathcal{P}_{xay}(m) p^u_{a|x}$

Theorem 5.1 *Consider a sequence of finite approximations COP_n obtained by applying* **FA**. *Assume B2, B4, B5 and {B1 or $\beta < 1$}. Then*
(1) $\lim_{m\to\infty} C_\beta(m; x) = C_\beta(x)$.
(2) Choose some $\epsilon > 0$. Let u^* be an ϵ-optimal (or optimal) policy for COP that satisfies (3.4) (e.g. any ϵ-optimal stationary policy). There exists some $N(\epsilon)$ such that for all $m \geq N(\epsilon)$, the policy $u(\epsilon)$ satisfying (3.4) and (3.5) is $\hat{\epsilon}$-optimal for COP_m, where $\hat{\epsilon}$ is given in Theorem 3.2.
(3) Choose a sequence $\xi_m \to 0$. Let $r(m)$ be a ξ_m-optimal policy for COP_m if COP_m is feasible, otherwise let it be an arbitrary stationary policy. The stationary policy w obtained by applying the limiting procedure in Theorem 3.1 (ii) to the policies $r(m)$ is optimal for COP.

Proof:

$$\begin{aligned} |C_\beta(m; x, u) - C_\beta(x, u)| \ \leq\ & \left|C_\beta(m; x, u) - C^n_\beta(m; x, u)\right| \\ & + \left|C^n_\beta(m; x, u) - C^n_\beta(x, u)\right| \\ & + \left|C^n_\beta(x, u) - C_\beta(x, u)\right| \end{aligned}$$

By (4.4),(4.6), $\left|C_\beta^n(x,u) - C_\beta(x,u)\right|$ converges to zero as $n \to \infty$ uniformly in $u \in U(S)$.

Next we show that this also holds for the first term. Without loss of generality, assume that $1 \in M$ (1 is the state defined in **FA** (v), and M is defined in (2.9)). We have for any $u \in U(S)$ for $\beta = 1$

$$\begin{aligned}
\|_M P^u(m)\|_\mu &= \sup_{x \in E_m} \mu_x^{-1} \sum_{y \notin M} |P_{xy}^u| \mu_y \\
&= \max_{x \in E_m} \mu_x^{-1} \left[\sum_{y \in E_{m-1} \setminus M} P_{xy}^u \mu_y + \sum_{y \in E_m \setminus E_{m-1} \setminus M} P_{xy}^u(m) \mu_y \right] \\
&\leq \max_{x \in E_m} \mu_x^{-1} \left[\sum_{y \in E_{m-1} \setminus M} P_{xy}^u \mu_y + \sum_{y \in E_m \setminus E_{m-1} \setminus M} P_{xy}^u \mu_y \right] \\
&\leq \max_{x \in E_m} \mu_x^{-1} \sum_{y \in \mathbf{X} \setminus M} P_{xy}^u \mu_y \\
&\leq \|_M P^u\|_\mu
\end{aligned}$$

For $\beta < 1$ we get similarly $\|\beta P^u(n)\|_\mu \leq \|\beta P^u(n)\|_\mu$. We show that this implies that the first term in (5.2) converges to 0 as $n \to \infty$ uniformly in $U(S)$ and in m. Consider a new MDP^* which is identical to the original MDP, except that in each state an extra action is added that allows to reach state 1 in one step w.p.1. Let $U(S^*)$ and $U(SD^*)$ be the set of stationary policies and stationary deterministic policies in MDP^*. Let $\bar{P}^u$ and $\bar{C}_\beta(x,u)$ correspond to the transition probabilities and cost associated with the new MDP. For each policy $u \in U(S)$ and m in the original MDP, there exists $w \in U(S^*)$ such that $P^u(m) = \bar{P}^w$, and thus $C_\beta^n(m;x,u) = \bar{C}_\beta(m,x,w)$. It is easily seen that the new MDP is also μ-UGR with the same M and α. The conditions of Theorem 4.1 therefore hold for MDP^*. It thus follows from (4.1),(4.6), that $\left|C_\beta^n(m;x,u) - C_\beta(m;x,u)\right|$ converges to zero as $n \to \infty$ uniformly in $u \in U(S)$.

Choose some ϵ. Since both the first and the last term converge to zero uniformly in $u \in U(S)$, it follows that $\exists n(\epsilon)$ such that

$$\left|C_\beta(m;x,u) - C_\beta^{n(\epsilon)}(m;x,u)\right| < \epsilon/2, \qquad \left|C_\beta^{n(\epsilon)}(x,u) - C_\beta(x,u)\right| < \epsilon/2 \tag{5.2}$$

B5 implies that for any n and x, $\exists L(n,x)$ s.t. $\forall m > L(n,x), \forall u \in U(S)$,

$$\left| C_\beta^n(m;x,u) - C_\beta^n(x,u) \right| = 0. \tag{5.3}$$

Combining this with (5.2) we obtain $\forall u \in U(S), \forall \epsilon > 0$,

$$|C(m;x,u) - C(x,u)| < \epsilon \tag{5.4}$$

for all $m \geq L(n(\epsilon),x)$, and hence as $m \to \infty$, $C(m;x,u)$ converges to zero uniformly in $u \in U(S)$. We obtain similarly the uniform convergence of $D(m;x,u)$. Thus the conditions of Theorem 3.1 and 3.2 are satisfied, which establishes the proof. ■

6 Convergence in the discount factor

Let COP_β denote the COP with cost $C_\beta(x,u)$ and $D_\beta(x,u)$, where $9 < \beta \leq 1$. Let $C_\beta(x)$ be the optimal value for COP_β.

Theorem 6.1 *Let $\gamma \to \beta$, $0 < \gamma, \beta \leq 1$. Assume B2, B4 and {B1 or $\beta < 1$}. Then*
(1) $\lim_{\gamma \to \beta} C_\gamma(x) = C_\beta(x)$.
(2) Choose some $\epsilon > 0$. Let u^* be an ϵ-optimal (or optimal) policy for COP_β that satisfies (3.4) (e.g. any ϵ-optimal stationary policy). There exists some $\Delta(\epsilon)$ such that for all γ such that $|\gamma - \beta| < \Delta(\epsilon)$, the policy $u(\epsilon)$ satisfying (3.4) and (3.5) is $\hat{\epsilon}$-optimal for COP_γ, where $\hat{\epsilon}$ is given in Theorem 3.2.
(3) Let $\gamma(m)$ be a subsequence converging to β. Choose a sequence $\xi_m \to 0$. Let $r(m)$ be a ξ_m-optimal policy for COP_β if $COP_{\gamma(m)}$ is feasible, otherwise let it be an arbitrary stationary policy. The stationary policy w obtained by applying the limiting procedure in Theorem 3.1 (ii) to the policies $r(m)$ is optimal for $COP - \beta$.

Proof: For any γ and β such that $0 < \gamma, \beta \leq 1$,

$$\begin{aligned} &\|C_\gamma(x,u) - C_\beta(x,u)\|_\mu \leq \\ &\left\|C_\gamma(x,u) - C_\gamma^n(x,u)\right\|_\mu + \left\|C_\gamma^n(x,u) - C_\beta^n(x,u)\right\|_\mu + \left\|C_\beta^n(x,u) - C_\beta(x,u)\right\|_\mu \end{aligned} \tag{6.1}$$

According to Theorem 4.1, the first and last term converge to zero as $n \to \infty$ uniformly in $U(S)$. For any ϵ, $\exists\ N(\epsilon)$ such that for $n = N(\epsilon)$, the first term and the last term in (6.1) are smaller than $\epsilon/3$ for any γ such that $|\gamma - \beta| < \epsilon,\ \gamma \leq 1$. (This follows from the fact that the term on the right hand side of (4.4) is continuous in β in (0,1), and converges to $2R(1-\alpha)^{-1}(n+1)^{-1}$ as $\beta \to 1$). Next we examine the second term. For any $u \in U(S)$ and $n = N(\epsilon)$,

$$\left\| C_\gamma^n(\bullet, u) - C_\beta^n(x,u) \right\|_\mu = \left\| \sum_{j=0}^{n} \left[\frac{\gamma^j}{\sum_{s=0}^n \gamma^s} - \frac{\beta^j}{\sum_{s=0}^n \beta^s} \right] P^j c(u) \right\|_\mu$$

$$\leq \begin{cases} \displaystyle\sum_{j=0}^{n} \left[\frac{\gamma^j}{\sum_{s=0}^n \gamma^s} - \frac{\beta^j}{\sum_{s=0}^n \beta^s} \right] rR & \text{if } \beta = 1 \\ \displaystyle\sum_{r=0}^{n} \left[\frac{\gamma^r/\beta^r}{\sum_{s=0}^n \gamma^s} - \frac{1}{\sum_{s=0}^n \beta^s} \right] \alpha^j R^j & \text{if } \beta < 1 \end{cases} \tag{6.2}$$

Let $\delta(\epsilon)$ be such that for all γ that satisfy $|\gamma - \beta| < \delta(\epsilon)$, the second term in (6.1) is less than $\epsilon/3$. It then follows that $\|C_\gamma(x,u) - C_\beta(x,u)\|_\mu < \epsilon$, $\forall \gamma$ such that $|\gamma - \beta| < \min\{\epsilon, \delta(\epsilon)\}$. Hence $C_\gamma(x,u)$ converges to $C_\beta(x,u)$ uniformly in $U(S)$, and similarly with $D_\gamma(x,u)$. Thus the conditions of Theorem 3.1 and 3.2 are satisfied, which establishes the proof. ■

7 Application to a queueing model

Consider the following discrete time system (Altman and Shwartz [1], [2] Section 6, Nain and Ross [18], Spieksma [21]). Packets of information of N different types, such as data files, video and voice signals, compete for access to some shared resource. Each type of arriving packets waits in a buffer till it gets access to the resource. At the beginning of each time slot, priority is given to one of the traffic types according to some prespecified decision rule, and the packet is served for one unit of time. Service problems and errors due to noises are modeled by allowing the service to fail with positive (class dependent) probability. If the service is successful, the packet disappears from the system; otherwise, it remains in the queue. The problem COP_{queues} is to find a scheduling policy that minimizes a linear combination of the average delays of

some types of traffic (typically, of the noninteractive types) subject to constraints on (linear combination of) average delays of other types (typically the interactive traffic).

At time t, M_t^i customers arrive to queue i, $1 \leq i \leq N$. Arrival vectors $M_t = \{M_t^1, ..., M_t^N\}$ are independent from slot to slot and form a renewal sequence with finite means λ_i. During a time slot (t,t+1) a customer from any class i, $1 \leq i \leq N$ may be served, according to some policy, which is a prespecified dynamic priority assignment. If served, with probability μ_i it completes its service and leaves the system; otherwise it remains in its queue. (μ_i should not be confused with the μ-norm introduced in Section 2. We use the same notation for both since both are widely used in Literature with this notation). A generic element of the state is given by $x = \{x^1, x^2, ..., x^N\}$ and it represents an N dimensional vector of the different queues' sizes. Throughout we restrict to non-idling policies.

Assume $\rho := \sum_{i=1}^N \lambda_i/\mu_i < 1$. Consider the linear cost function $c(x,a) := \sum_{i=1}^N c_i x^i$ and $d^k(x,a) = \sum_{i=1}^N d_i^k x^i$ for $1 \leq k \leq K$, where c_i and d_i^k are non-negative constants. Thus the costs $C(x,u)$ and $D^k(x,u)$ are related to linear combinations of expected average length of the different queues, and COP_{queues} has the form: find $u \in U$ that minimizes $C(x,u)$ s.t. $D^k(x,u) \leq V_k,\ k = 1, ..., K$, where V_k are given constants. Consider the expected average cost. By Little's law these quantities are proportional to the respective waiting times in the different queues.

Let $\mathbf{G} = \{g_j\}$ be the set of all strict priority policies, i.e. each type of customer has an index, and a customer of a given type is served only if there are no customers with lower priority in the system, and if it is the first in its buffer. Let $|\mathcal{G}| = L$. For the unconstrained control problem, there exists an optimal policy within $\mathbf{G}$; it is the well known "μc" rule, for which the priorities are set according to increasing order of the $\mu_i c_i$ (see Baras et al. [7]). Thus, the queue for which $\mu_i c_i$ is the largest has the highest priority, and so on. Optimal policies for COP_{queues} are obtained by time multiplexing between the different g_j's. More specifically, define an L dimensional vector parameter $\alpha = \{\alpha_1, \alpha_2, ..., \alpha_L\}$, where α is a probability measure. Define a "cycle" as the time between two consecutive instants that the system is empty. During any cycle, a fixed g_j is used. A PTS policy $\hat{\alpha}$ is defined as a policy that chooses different policies g_j in such a way that the relative average number of cycles during which g_j was used is equal to

α_j, as t goes to infinity. (The precise definition can be found in Altman and Shwartz [1]). It is shown in Altman and Shwartz [1] that

$$\bar{f}^1(x,\hat{\alpha}) = \sum_{j=1}^{L} \alpha_j \bar{f}^1(x,g_j) \tag{7.1}$$

and

$$C_1(x,\hat{\alpha}) = c \cdot \bar{f}^1(x,\hat{\alpha}) = \sum_{j=1}^{L} \alpha_j c \cdot \bar{f}^1(x,g_j) = \sum_{j=1}^{L} \alpha_j C_1(x,g_j). \tag{7.2}$$

where $\bar{f}$ is defined in (3.3). For a given $\delta > 0$, consider the following **LP**: find $\alpha \in \mathbf{M}(\{1,...,L\})$ that

$$\begin{aligned} &\text{minimize } \sum_{j=1}^{L} \alpha_j C_1(x,g_j) \\ &\text{subject to } \sum_{j=1}^{L} \alpha_j D_1^k(x,g_j) \le V_k - \delta, \qquad k = 1,...,K \end{aligned} \tag{7.3}$$

$C_1(x,g_j)$ and $D_1^k(x,g_j)$ can be obtained as in Nain and Ross [18]. Let $\alpha^*(\delta)$ be the solution of LP with a given δ. Altman and Shwartz [1] show that $\hat{\alpha}^*(0)$ is an optimal policy for COP_{queues}. Under B2, it can be shown that there exists some $\delta^* > 0$ such that $\hat{\alpha}^*(\delta)$ is feasible for COP_{queues} (this follows from the fact that the PTS policies are sufficient for COP_{queues}, see Altman and Shwartz [1]).

In the following Theorem we consider (1) a sequence of problems COP^n_{queues} for the systems with buffers of sizes $\mathbf{R}^n = \{R_1^n,...,R_L^n\}$, $1 \le n \le \infty$, where COP^∞_{queues} is the one with all buffers infinite, and $\mathbf{R}^n \subset \mathbf{R}^{n+1}$ (where the inclusion is strict). Assume without loss of generality that the initial state x satisfies $x \in \cap_{n=1}^{\infty} \mathbf{R}^n$. (2) The problem with finite horizon and (3) convergence in the discount factor.

Theorem 7.1 . *Assume B2. Then*
(i) $\lim_{n\to\infty} C_n(x) = C(x)$, $\lim_{m\to\infty} C_\beta(m;x) = C_\beta(x)$, $\lim_{\gamma\to\beta} C_\gamma(x) = C_\beta(x)$.
(ii) Let $\beta = 1$. *Choose some* $0 < \epsilon < 1$. *Let* $u(\epsilon)$ *be the PTS policy with*

$$\alpha = \epsilon\alpha^*(\delta^*) + (1-\epsilon)\alpha^*(0). \tag{7.4}$$

Then for all n large enough $u(\epsilon)$ is $\hat{\epsilon}$-optimal for COP^n_{queues}; for all m large enough $u(\epsilon)$ is $\hat{\epsilon}$-optimal for COP_{queues} with finite horizon m; for all β close enough to 1, $u(\epsilon)$ is $\hat{\epsilon}$-optimal for COP^β_{queues}. $\hat{\epsilon} = \epsilon[C(x, \hat{\alpha}^*(\delta^*)) + 3]$.

(iii) In case that there are no constraints, for all n large enough the μc rule is ϵ-optimal for COP^n_{queues}; for all m large enough it is $\hat{\epsilon}$-optimal for COP_{queues} with finite horizon m; for all β close enough to 1, it is $\hat{\epsilon}$-optimal for COP^β_{queues}.

Proof: B4(i) and B4(ii) are satisfied by Spieksma [21] Theorem 9.1. Moreover, the function μ has the form

$$\mu(x) = \prod_{i=1}^{N} (1 + z_i)^{x^i} \tag{7.5}$$

where z are some numbers with $z_i > 0, \forall i$. It is easily seen that this implies (iv) in the finite approximation scheme **FA**, and that the other features of **FA** are satisfied as well. B1 is satisfied (see e.g. Altman and SHwartz [2]) and B5 clearly holds too. The Theorem then follows by applying Theorems 4.1, 5.1 and 6.1. ■

References

[1] 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.

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

[3] E. Altman and A. Shwartz, "Adaptive control of constrained Markov chains", *IEEE Transactions on Automatic Control*, **36**, No. 4, pp. 454-462, 1991.

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

[5] E. Altman and V. A. Gaitsgory, "Stability and Singular Perturbations in Constrained Markov Decision Problems", submitted to *IEEE Transactions on Automatic Control*, 1990.

[6] E. Altman, "Denumerable constrained Markov Decision Problems and finite approximations", under revision, 1991.

[7] J. S. Baras, D. -J. Ma, and A. M. Makowski, "K competing queues with geometric service requirements and linear costs: the μc rule is always optimal," *Systems and Control Letters,* **6** No. 3 pp. 173-180, August 1985.

[8] 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.

[9] V. S. Borkar, "Ergodic control of Markov Chains with constraints – the general case", manuscript.

[10] 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.

[11] R. Cavazos-Cadena, "Finite-state approximations for denumerable state discounted Markov Decision Processes", *J. Applied Mathematics and Optimization* **14** pp. 27-47, 1986.

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

[13] O. Hernandez-Lerma, "Finite state approximations for denumerable multidimensional - state discounted Markov decision processes", *J. Mathematical Analysis and Applications* **113** pp. 382-389, 1986.

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

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

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

[17] 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.

[18] Nain P. and K. W. Ross, "Optimal Priority Assignment with hard Constraint," *Transactions on Automatic Control,* Vol. 31 No. 10, pp. 883-888, October 1986.

[19] L. I. Sennott, "Constrained discounted Markov decision chains", submitted, 1990.

[20] L. I. Sennott, "Constrained average cost Markov decision chains", submitted, 1990.

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

[22] L. C. Thomas and D. Stengos, "Finite State Approximation Algorithms for Average Cost Denumerable State Markov Decision Processes", *OR Spectrum*, **7**, pp. 27-37, 1985.

[23] D. J. White, "Finite State Approximations for Denumerable State Infinite Horizon Discounted Markov Decision Processes with Unbounded Rewards", *J. Mathematical Analysis and Applications* **86**, pp. 292-306, 1982.