RATE OF CONVERGENCE OF EMPIRICAL MEASURES AND COSTS IN CONTROLLED MARKOV CHAINS AND TRANSIENT OPTIMALITY

Eitan Altman

INRIA – Sophia Antipolis

2004 Route des Lucioles

06565 Valbonne Cedex, FRANCE

Ofer Zeitouni

Electrical Engineering

Technion — Israel Institute of Technology

Haifa 32000, ISRAEL

ABSTRACT

The purpose of this paper is two-fold. First, bounds on the rate of convergence of empirical measures in Controlled Markov Chains are obtained under some recurrence conditions. These include bounds obtained through Large Deviations and Central Limit Theorem arguments. These results are then applied to optimal Control Problems. Bounds on the rate of convergence of the empirical measures that are uniform over different sets of policies are derived, resulting in bounds on the rate of convergence of the costs. Finally, new optimal control problems that involve not only average cost criteria but also measures on the transient behavior of the cost, namely the rate of convergence, are introduced and applied to a problem in telecommunications. The solution to these problems rely on the bounds introduced in previous Sections.

KEYWORDS Markov Decision Processes, CLT, Large Deviations, Rate of convergence, State action frequencies.

INTRODUCTION

In the field of Controlled Markov Chains (or Markov Decision Processes) one is usually concerned with dynamic optimization problems where some cost criterion is minimized. In the case that the long-run behavior is of interest, criteria which are of time average type are often used. This is typically the case in data communications and computer communications, where minimizing average delays or maximizing the throughput is the goal (see e.g. [5] Chap 4 and [8] Chap 6). When the near future is of greater importance (e.g. in economics) then the discounted cost criteria are used (see [8] Chap 4). Often, in such problems, the total number of states and actions taken by the controller is finite. *Throughout, we will assume that this is indeed the case.*

When average cost criteria are used, it may be of importance to know how fast the finite horizon average cost converges to the infinite horizon one. This can give an insight on the suitability of the average criteria versus the discounted criteria, and can indicate how long should one wait till the steady state behavior is achieved. Mandl [13,14] obtains Central Limit Theorems and Laws of Iterated Logarithm for the convergence of the costs under stationary deterministic policies and under other deterministic policies that behave asymptotically as stationary deterministic ones.

Note that the question of how fast the costs converge to their limits was raised not only for average cost criteria but also in optimization problems involving discounted cost criteria. A bound for the rate of convergence of the finite horizon discounted cost to the infinite horizon discounted cost which decreases geometrically in time, is given e.g. in [16] p. 36.

Depending on the definition of the average cost and on the policies used, the rate of convergence of the average costs is related to the rate of convergence of some other quantities, such as the limiting probabilities (when such exist), the state-action frequencies (known also as empirical measures or occupation measures) and the expectation of these frequencies. We therefore restrict our attention in Sections 3, 4 and 5 to studying the convergence rate of the empirical measures.

For those, we obtain in Section 3 a lower bound on the distance between the expected empirical measures and their limit points for *any time t and any policy*, and show that it converges to zero like t^{-1}. For the class of stationary policies, an upper bound with the same behavior is derived, which leads to the conclusion that non-stationary policies cannot improve substantially the rate of convergence.

In Section 4, we specify our discussion to the class U^*, the sub-class of policies for which the state frequencies and its expectation have a.s. the same limit. Note that under such policies, there may not exist a limiting probability of being in different states. U^* contains many sub-classes of policies that are of interest in the field of Markov Decision Processes, such as the stationary policies, "Policy Time Sharing" (PTS), "Action Time Sharing" (ATS) and "Asymptotically Stationary" (AS) policies [2,4]. PTS, ATS and AS policies have been used in [3] and [4] for solving constrained adaptive problems. PTS policies were used in [1] to solve optimal priority assignment in queuing networks, and were used in [2] to solve general constrained optimization problems.

In order to obtain bounds on the rate of convergence of the empirical measures, and therefore of the cost, for such policies, we first obtain a Central Limit Theorem for some associated martingale. This is done under a (relatively strong) ergodic assumption (c.f. Assumption $A1$, Sec. 2). Mandl [13,14] already obtained CLT type bounds for the cost from a CLT applied to some martingale. He considered a martingale that is related to dynamic programming, and hence obtained results for policies that behave asymptotically as the deterministic stationary policy which is optimal for some control problem. These theorems were obtained under ergodic conditions and models which are less restrictive than ours. However, in some control problems such as constrained optimization (see e.g. [12]), the optimal policies are in general not stationary deterministic, and they involve randomization. The approach of [13,14] is not applicable for such policies nor for other policies which do not behave asymptotically as a deterministic stationary policy. We thus introduce a different martingale from which we obtain bounds on the rate of convergence of frequencies, for any policy in U^*.

Information on the rate of convergence of the state frequencies in Markov Chains (without control) to the steady state probabilities has also been obtained through the theory of Large Deviations. We use these results to obtain in Section 5 bounds which are uniform in the class of stationary policies.

Using the relation between the occupation measure and the average cost, we specialize in Section 6 the bounds to the study of the rate of convergence of the costs. This relation is particularly useful for problems such as the Constrained Optimization Problem (COP), where the minimization of some functional subject to constraints on other is considered.

The bounds on (and exact behavior of) the rate of convergence of the state frequencies and of the cost functional suggest incorporating the rates of convergence as additional criteria in optimization problems which use time average. Such problems are discussed and solved in Section 7, where the relevance of this approach to a telecommunication system is demonstrated.

2. MODEL AND NOTATIONS.

Let $\{X_t\}_{t=0}^{\infty}$ be the discrete time state process, defined on the finite *state space* $\mathbf{X} = \{1, ..., J\}$; the action A_t at time t takes values in the finite *action space* $\mathcal{A}$. Without loss of generality, we assume that in any state x all actions in $\mathcal{A}$ are available. We use the space of paths $\{X_t\}$ and $\{A_t\}$ as the canonical sample space Ω, equipped with the Borel σ-field B obtained by the standard product topology.

Denote by $H_t := (X_0, A_0, X_1, A_1, ..., X_t, A_t)$ the *history* of the process up to time t. If the state at time t is x and action a is applied, then the next state will be y with probability

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

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 $\mathcal{A}$. Each policy u and initial state x induce a probability measure P_x^u on $\{\Omega, \mathcal{B}\}$. The corresponding expectation operator is denoted 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 $\mathcal{A}$, so that $p^g_{\mathcal{A}|x} = 1$; under g, X_t becomes a Markov chain with homogeneous transition probabilities, given by

$$P^g_{xy} := \sum_{a \in \mathcal{A}} p^g_{a|x} P_{xay}\,. \qquad (2.2)$$

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 $g : \mathbf{X} \to \mathcal{A}$, so that $p^g_{\cdot|x} = \delta_{g(x)}(\cdot)$ is, for each x, concentrated at the point $g(x) \in \mathcal{A}$.

Throughout the paper we often assume:

$A1$: Under any deterministic policy $g \in U(SD)$, the process X_t is a regular Markov chain (i.e. there are no transient states, and the state space consists of a single ergodic non-cyclic class).

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

For a given finite set B, denote by $M_1(B)$ the set of probability measures on B.

Next we define the state action frequencies, known also as empirical measures or occupation measures. The *state action (sa) frequency* f_{sa}^t is the (random) vector $f_{sa}^t : \Omega \to M_1(\mathbf{X} \times \mathcal{A})$ defined by $f_{sa}^t(y,a) := \frac{1}{t+1} \sum_{r=0}^t 1\{X_r = y, A_r = a\}$. The value of $f_{sa}^t(y,a;\omega)$ is the frequency at which the event of being at state y and choosing action a occurs by time t. Similarly, the *state (s) frequency* f_s^t is the (random) vector $f_s^t : \Omega \to M_1(\mathbf{X})$ defined by $f_s^t(y) := \frac{1}{t+1} \sum_{r=0}^t 1\{X_r = y\}$.

The (random) set of accumulation points of f_{sa}^t (respectively f_s^t) as $t \to \infty$, which is a mapping from Ω to subsets of $M_1(\mathbf{X} \times \mathcal{A})$ (respectively $M_1(\mathbf{X})$), is denoted by F_{sa} (respectively F_s).

Often, one is not interested in the empirical measures themselves but rather in their expected value. This is particularly the case in optimal control problems where the expected average cost is considered. In anticipation of that, denote by $\bar{f}_{sa}^t(x,u)$ and $\bar{f}_s^t(x,u)$ the vectors whose components are given respectively by $\bar{f}_{sa}^t(x,u;y,a) := E^u[f_{sa}^t(y,a)|X_0 = x]$, and

$\bar{f}_s^t(x,u;y) := E^u[f_s^t(y)|X_0 = x]$, and let $\bar{F}_{sa}(x,u)$ (respectively $\bar{F}_s(x,u)$) denote the set of all accumulation points of $\bar{f}_{sa}^t(x,u)$ (respectively $\bar{f}_s^t(x,u)$) as $t \to \infty$.

For any given set of policies U', define the set of achievable expected state action frequencies $L_{sa}^x(U') := \cup_{u \in U'} \bar{F}_{sa}(x,u)$. Define similarly $L_s^x(U') := \cup_{u \in U'} \bar{F}_s(x,u)$. In particular, the set of all achievable expected frequencies is denoted by $L_{sa}^x := \cup_{u \in U} \bar{F}_{sa}(x,u)$ and $L_s^x := \cup_{u \in U} \bar{F}_s(x,u)$.

In view of the above definitions we can briefly summarize the main results of this paper. We obtain in Sections 3-5, after some preliminaries concerning the properties of L_s^x and L_{sa}^x, two different kinds of bounds on rates of convergence of frequencies:

(1) lower bounds on the distance between $\bar{f}_{sa}^t(x,u)$ (Theorem 3.2) or f_{sa}^t (Theorem 4.4) and their limits as $t \to \infty$, if such exist.

(2) upper bounds on the distance between f_s^t and the set L_s^x (Corollary 5.5).

These bounds are then applied in Sections 6 and 7 to obtain rates of convergence of costs in optimal control problems.

The following notations are used below: $\delta_a(x)$ is the Kronecker delta function. For an arbitrary set B, $1[B]$ is the indicator function of the set. When B is a finite set we denote by $|B|$ the cardinality of the set (i.e. the number of elements in B). For any function $\xi : B \to \mathbb{R}$ we define $\xi^-(y) := \max\{0, -\xi(y)\}$, $y \in B$. For two vectors or matrices ζ, P of the same dimension, the notation $\zeta \cdot P$ stands for summation over common indices (i.e., the usual scalar product when ζ, P are vectors). Given a vector of real numbers $z = \{z_1, ..., z_k\}$, denote $\|z\|_{var} := \sum_{j=1}^k |z_j|$ and $\|z\|_2 := [\sum_{j=1}^k z_j^2]^{1/2}$. Finally, for any function f, $\inf_{x \in \emptyset} f(x) = \min_{x \in \emptyset} f(x) = \infty$.

3. ACHIEVABLE FREQUENCIES AND A LOWER BOUND

In this Section we first introduce the properties of the set of limit frequencies (i.e. of the accumulation points of the frequencies) or limit expected frequencies achieved by the set of all control policies. This then enables us to investigate the rate of convergence of the expected frequencies to that set under different policies.

A particularly important role in the analysis to follow is played by the transition operator G, defined as follows. Let $G : \mathbb{R}^{|\mathbf{X}\times\mathcal{A}|} \to \mathbb{R}^{|\mathbf{X}|}$ be given by $G^y(z) := \sum_{a\in\mathcal{A}} z(y,a) - \sum_{a\in\mathcal{A}}\sum_{v\in\mathbf{X}} z(v,a)P_{vay}$. Note that for $z(v,a) = 1\{X_t = v, A_t = a\}$,

$$G^y(z) = 1\{X_t = y\} - \sum_{a\in\mathcal{A}}\sum_{w\in\mathbf{X}} 1\{X_t = w, A_t = a\}P_{way}\,, \tag{3.1}$$

and therefore for $z(v,a) = P_x^u\{X_t = v, A_t = a\}$,

$$G^y(z) = P_x^u(X_t = y) - P_x^u(X_{t+1} = y)\,, \tag{3.2}$$

which is why the operator G is regarded as a "transition operator".

In the sequel, we quote and establish some basic properties of the limit sets of the frequencies.

Lemma 3.1:

(i) Assume A1. The class of stationary policies is complete i.e. $L_{sa}^x(U(S)) = L_{sa}^x$ for every initial state x. Moreover, $L_{sa} := L_{sa}^x$ is independent of x and is equal to the convex hull of $L_{sa}^x(U(SD))$, the frequencies achieved by the stationary deterministic policies. Similarly, $L_s := L_s^x$ is independent of x and is equal to the convex hull of $L_s^x(U(SD))$, the frequencies achieved by the stationary deterministic policies.

(ii) $L_{sa}^x \subseteq \{z \in \mathbb{R}^{|\mathbf{X}|\times|\mathcal{A}|} : G(z) = 0, z \geq 0, \sum_{y\in\mathbf{X},a\in\mathcal{A}} z(y,a) = 1\}$ and $L_s^x \subseteq Y := \{z \in \mathbb{R}^{|\mathbf{X}|} : z(y) = \sum_{a\in\mathcal{A}} z'(y,a),\ y \in \mathbf{X}$ for some $z' \in L_{sa}^x\}$. Under A1, both inclusions become equality.

(iii) Assume A1. For any policy u, $F_{sa} \subset L_{sa}$ P_x^u a.s. and $F_s \subset L_s$ P_x^u a.s.

Proof: For the claims concerning state-action frequencies: for (i) and (ii) see [12] Theorem 2, and [8] Sec 7 Thm 3. For (iii) see [8] Sec 7 Thm 5.

The claims concerning the state frequencies follow from the fact that for any t and $x, y \in \mathbf{X}$ we have $f_s^t(y) = \sum_{a \in \mathbf{X}} f_{sa}^t(y,a)$ and $\bar{f}_s^t(x,u;y) = \sum_{a \in \mathbf{X}} \bar{f}_{sa}^t(x,u;y,a)$. ■

The role of the operator G is clear from Lemma 3.1, as the achievable measures belong to its null space $null(G)$. As the following theorem hints, not only does G control the location of achievable limit points but it also gives a uniform bound (in U) for the expected frequencies. With some abuse of notation, we denote by G also the matrix representation of the linear operator G. Denote $\lambda_{\max} :=$ the largest eigenvalue of the matrix $G^T G$. Clearly, under $A1$, $\lambda_{\max} > 0$. Note that for any $v \in null(G)$, and any $z \in M_1(\mathbf{X} \times \mathcal{A})$,

$$\|G(z)\|_2 = \left((z-v)^T [G^T G](z-v)\right)^{1/2} \leq \lambda_{\max}^{1/2} \|z - v\|_2 \, . \tag{3.3}$$

Consequently, for any $v \in null(G)$,

$$\{\omega : \left\|f_{sa}^t - v\right\|_2 < \delta\} \subseteq \{\omega : \lambda_{\max}^{-1/2} \left\|G(f_{sa}^t)\right\|_2 < \delta\} \, . \tag{3.4}$$

We are now ready to prove the lower bound for $\left\|\bar{f}_{sa}^t(x,u) - \zeta\right\|_2$ and $\zeta \in \bar{F}_{sa}(x,u)$. Note that the behavior of this lower bound may be considered as a measure for the rate of convergence (of $\bar{f}_{sa}^t(x,u)$) or more precisely an upper bound for the rate of convergence.

Theorem 3.2: *Assume $\lambda_{\max} > 0$ (this is satisfied, e.g., under $A1$).*

(i) For any $x \in \mathbf{X}$, $u \in U$, $\zeta \in \bar{F}_{sa}(x,u)$ and any $t > 1$,

$$\left\|\bar{f}_{sa}^t(x,u) - \zeta\right\|_2 \geq \frac{1}{(t+1)\lambda_{\max}^{1/2}} \left(\sum_{y \in \mathbf{X}} [\delta_x(y) - P_x^u(X_{t+1} = y)]^2\right)^{1/2} . \tag{3.5}$$

(ii) For any $x \in \mathbf{X}$, $u \in U$, $\zeta \in \bar{F}_{sa}(x,u)$, if a limiting probability $\pi_y^u := \lim_{t \to \infty} P_x^u(X_t = y)$ exists then

$$\varliminf_{t \to \infty} t \left\|\bar{f}_{sa}^t(x,u) - \zeta\right\|_2 \geq \lambda_{\max}^{-1/2} \left(\sum_{y \in \mathbf{X}} [\delta_x(y) - \pi_y^u]^2\right)^{1/2} .$$

Proof: Note that by (3.2),

$$G^y(\bar{f}_{sa}^t(x,u)) = \frac{1}{t+1} \sum_{s=0}^{t} [P_x^u(X_s = y) - P_x^u(X_{s+1} = y)] = \frac{1}{t+1} \left(\delta_x(y) - P_x^u(X_{t+1} = y)\right) .$$

(i) is then obtained by substituting in (3.3) $\bar{f}^t_{sa}(x,u)$ for z and ζ for v. (ii) follows immediately. ■

Clearly, there does not exist an upper bound on the left side of (3.5) that decreases to 0 as $t \to \infty$, since $\bar{f}^t_{sa}(x,u)$ may not converge at all. Yet for the case of stationary policies, the state process is in fact a Markov Chain; for Regular Markov Chains we know that an upper bound on the distance between the probabilities of visiting different states at time t and the steady-state probabilities decreases at a geometric rate. Denote by $\pi^g \in M_1(\mathbf{X})$ the steady state probability that corresponds to $g \in U(S)$. Then [4, Lemma 4.6], for any initial distribution μ on $\mathbf{X}$ and $g \in U(S)$, there exist finite positive constants $\alpha < 1$ and β such that $\|\mu[P^g]^t - \pi^g\|_2 < \beta\alpha^t$ (P^g is the one step transition probability matrix that corresponds to g). It then follows that for stationary policies, an upper bound on the left side of (3.5) exists, which decreases like t^{-1}. Indeed, let $\mu = \delta_x$. Then

$$\left\|(t+1)(\bar{f}^t_s(x,g) - \pi^g)\right\|_2 = \left\|(t+1)\left(\frac{1}{t+1}\sum_{s=0}^{t}\mu[P^g]^s - \pi^g\right)\right\|_2 < \beta\sum_{s=0}^{t}\alpha^s < \frac{\beta}{1-\alpha},$$

and hence

$$\left\|\bar{f}^t_s(x,g) - \pi^g\right\|_2 < \frac{\beta}{t+1}(1-\alpha)^{-1}.$$

Since the lower bound established in Theorem 3.2 also decreases like t^{-1}, we may conclude that the stationary policies are "good" in the sense that we cannot hope to find any policy for which the left side of (3.5) decreases at a quicker rate than t^{-1}, achieved by the stationary policies.

This result means especially that even "steering policies", that choose at each moment the action which minimizes the distance $\|f^t_{sa}(x,u) - \zeta\|_2$, cannot substantially improve the rate of convergence.

4. CLT in CONTROLLED MARKOV CHAINS

Our goal in this section is to obtain convergence bounds on the state action frequencies which are of the Central Limit Theorem (CLT) type. We identify a certain martingale (c.f. (4.1) below) and then use the martingale CLT introduced in [6]. This enables us to obtain lower bounds on the rate of convergence of the state-action frequencies to L_{sa}, uniformly in U^*, which is the class of all policies for which $\bar{F}_{sa}(x,u) = \{\bar{f}_{sa}(x,u)\}$ is a singleton, and in addition, for all z, a and for a fixed initial state x,

$$\frac{f^t_{sa}(z,a)}{\bar{f}^t_{sa}(x,u;z,a)} \to_{t\to\infty} 1 \quad (P^u_x \quad \text{a.s.}).$$

Note that under policies $u \in U^*$ there may not exist a limiting probability.

We shall assume $A1$ throughout this section. However, $A1$ turns out not to be sufficient for obtaining a CLT type lower bound. The additional assumption $A1'(y)$, defined below, is required to ensure that the transition probabilities are not "degenerate", i.e. are not all equal to zero or one, for in the latter case the process may evolve in a deterministic way and no CLT type lower bound can hold.

For any $y \in \mathbf{X}$, let

$A1'(y)$: There exists some state $z = z(y)$ such that for all $a \in \mathcal{A}$, $\eta \le P_{zay} \le 1 - \eta$ for some $\eta > 0$.

The main result, Theorem 4.4, will be meaningful if there exists some y for which $A1'(y)$ holds.

Let $\{S_t, \mathcal{F}_t\}, t = 1, 2, ...$ be a Martingale on the probability space $\{\Omega, \mathcal{F}, P\}$, with $S_0 = 0$ and $\xi_t = S_t - S_{t-1}, t = 1, 2, ...$. Let $E_r(\cdot)$ denote $E(\cdot|\mathcal{F}_r)$, and let

$\sigma_t^2 := E_{t-1}(\xi_t^2)$,

$V_t^2 := \sum_{r=1}^t \sigma_r^2$,

$s_t := \sqrt{E\ V_t^2} = \sqrt{E\ S_t^2} > 0$.

For this class of Martingales, Lindeberg's condition is said to hold if for all $\epsilon > 0$,

$$s_t^{-2} \sum_{r=1}^t E\left[\ \xi_r^2 1\{|\xi_r| \ge \epsilon s_t\}\right] \to 0 \qquad as\ t \to \infty$$

Lemma 4.1: *([6], Thm 2.) If*

$$V_t^2 s_t^{-2} \to_p 1 \qquad as\ t \to \infty$$

and Lindeberg's condition holds, then S_t/s_t converges in distribution to a $N(0,1)$ random variable.

Fix a state y, and let

$$\xi_t := 1\{X_t = y\} - E_{t-1}1\{X_t = y\} = 1\{X_t = y\} - \sum_{z,a} 1\{X_{t-1} = z, A_{t-1} = a\} P_{zay}\,. \tag{4.1}$$

With this definition of ξ_t it is easy to check that $\{S_t, \mathcal{F}_t\}$ is indeed a Martingale under any policy $u \in U$ and $x \in \mathbf{X}$, where $\mathcal{F}_t$ is the σ-field generated by the history H_t. Moreover, since under any policy $u \in U$ and $x \in \mathbf{X}$, $|\xi_t^2| \leq 1$ P_x^u *a.s.*, one concludes that $\lim_{t\to\infty} S_t/t = 0$ P_x^u *a.s.* (see e.g. the Martingale Stability Theorem [10, Theorem 2.22]). Note that

$$S_t = \sum_{s=1}^{t} \xi_s =$$

$$(t+1)\left[\sum_{z,a} f_{sa}^t(z,a)\left(\delta_y(z) - P_{zay}\right)\right] + \left[-1\{X_0 = y\} + \sum_{z,a} 1\{X_t = z, A_t = a\}P_{zay}\right]$$

$$= tG^y(f_{sa}^t) + O(1) \tag{4.2}$$

where G^y was defined before Lemma 3.1.

Lemma 4.2: *Fix $y \in \mathbf{X}$. Assume $u \in U^*$, A1 and A1$'(y)$. Then S_t/s_t converges in distribution to a $N(0,1)$ random variable.*

Proof: Some calculations yield:

$$\sigma_t^2 = \sum_{z,a} 1\{X_{t-1} = z, A_{t-1} = a\}[P_{zay} - P_{zay}^2]\,,$$

$$V_t^2 = \sum_{s=1}^{t}\sum_{z,a} 1\{X_{s-1} = z, A_{s-1} = a\}[P_{zay} - P_{zay}^2]\,, \tag{4.3}$$

$$s_t^2 = \sum_{s=1}^{t}\sum_{z,a} P_x^u\{X_{s-1} = z, A_{s-1} = a\}[P_{zay} - P_{zay}^2]\,. \tag{4.4}$$

We show below that Lindeberg's condition holds. We first note that

$$s_t^2 = \sum_{s=1}^{t}\sum_{z,a} P_x^u\{X_{s-1} = z, A_{s-1} = a\}P_{zay}(1 - P_{zay}) \geq \eta^2 \sum_{s=1}^{t} P_x^u\{X_{s-1} = z(y)\}\,.$$

Hence by showing that $\lim_{t\to\infty}\sum_{s=1}^t P_x^u\{X_s = z(y)\} = \infty$ we have $\lim_{t\to\infty} s_t^2 = \infty$, from which it follows by the boundedness of ξ_t that Lindeberg's condition holds. Indeed, by assumption $A1$ and since $U(SD)$ is finite, it follows that there exists some $\zeta > 0$ such that for any $u \in U(SD)$,

$$\underline{\lim}_{t\to\infty}\, t^{-1}\sum_{s=1}^{t} P_x^u\{X_s = z(y)\} \geq \zeta. \tag{4.5}$$

It then follows from Lemma 3.1 (i) that (4.5) holds for any $u \in U$, since L_s is the convex hull of $L_s^x(U(SD))$.

From (4.3) and (4.4) we obtain

$$V_t^2/s_t^2 = \frac{\sum_{s=1}^t \sum_{z,a} 1\{X_{s-1}=z, A_{s-1}=a\}[P_{zay}-P_{zay}^2]}{\sum_{s=1}^t \sum_{z,a} P_x^u\{X_{s-1}=z, A_{s-1}=a\}[P_{zay}-P_{zay}^2]} =$$

$$\frac{\sum_{z,a} f_{sa}^{t-1}(z,a)[P_{zay}-P_{zay}^2]}{\sum_{z,a} \bar{f}_{sa}^{t-1}(x,u;z,a)[P_{zay}-P_{zay}^2]}.$$

Since $u \in U^*$ it follows that indeed

$$V_t^2 s_t^{-2} \rightarrow_{P_x^u} 1 \qquad as\ t \rightarrow \infty$$

The proof is established by applying Lemma 4.1. ■

In the following Corollary we show that the CLT for S_t is equivalent to a CLT on $G^y(f_{sa}^t)$. This, together with the results from Section 3 enable us to obtain lower bounds for the state-action frequencies, uniformly in U^*. Pick any $x \in \mathbf{X}$ and $u \in U^*$.

Denote $r(y) := \left(\sum_{z\in\mathbf{X}} \sum_{a\in\mathcal{A}} \bar{f}_{sa}(x,u;z,a)[P_{zay}-P_{zay}^2]\right)^{1/2}$.

Corollary 4.3: *Fix $y \in \mathbf{X}$. Assume A1 and $A1'(y)$. Then for every $u \in U^*$, $t^{1/2}G^y(f_{sa}^t)$ converges in law to a $N(0, r(y))$ distribution.*

Proof: Note that

$$s_t = \left(t \sum_{z\in\mathbf{X}} \sum_{a\in\mathcal{A}} \bar{f}_{sa}^{t-1}(x,u;z,a)[P_{zay}-P_{zay}^2]\right)^{1/2} \tag{4.6}$$

Denote $b^t(y) := -1\{X_0 = y\} + \sum_{z,a} 1\{X_t = z, A_t = a\}P_{zay}$, and note that, by (4.2), $S_t = (t+1)G^y(f_{sa}^t) + b^t(y)$. Hence,

$$\sqrt{t+1}G^y(f_{sa}^t) = \frac{S_t}{\sqrt{t+1}} - \frac{b^t(y)}{\sqrt{t+1}} = \frac{S_t}{s_t}\frac{s_t}{\sqrt{t+1}} - \frac{b^t(y)}{s_t}\frac{s_t}{\sqrt{t+1}}.$$

Since $\lim_{t\to\infty} b^t(y)/s_t = 0$ and $\lim_{t\to\infty} s_t t^{-1/2} = r(y)$, the proof is established. ■

Throughout, let θ denote a standard, zero mean, unit variance, Normal random variable, and let P_N denote its law. From the previous Corollary we obtain the following bounds:

Theorem 4.4: *Assume A1 . Let $Y \subset \mathbf{X}$ be the set of states for which $A1'(y)$ holds. Then (i)For any $x \in \mathbf{X}$, $u \in U^*$, $\zeta \in \bar{F}_{sa}(x,u)$*

$$\overline{\lim_{t\to\infty}} \mathbf{P}_x^u \left(t^{1/2} \left\| f_{sa}^t - \zeta \right\|_2 \le \delta \right) \le \min_{y \in Y} P_N \left(|\theta| < \delta (\lambda_{\max})^{1/2} r^{-1}(y) \right) .$$

(ii) For any $x \in \mathbf{X}$, and $u \in U^*$ we have

$$\overline{\lim_{t\to\infty}} \mathbf{P}_x^u \left(t^{1/2} d(f_{sa}^t, L_{sa}) \le \delta \right) \le \min_{y \in Y} P_N \left(|\theta| < \delta (\lambda_{\max})^{1/2} r^{-1}(y) \right) .$$

where

$$d(f_{sa}^t, L_{sa}) := \inf_{v \in L_{sa}} ||f_{sa}^t - v||_2 .$$

Proof: It follows from (3.4) that for any y,

$$\{\omega : t^{1/2} \left\| f_{sa}^t - v \right\|_2 < \delta\} \subseteq \{\omega : t^{1/2} \left\| G(f_{sa}^t) \right\|_2 < \delta \lambda_{\max}^{1/2}\} \subseteq \{\omega : |t^{1/2} G^y(f_{sa}^t)| < \delta \lambda_{\max}^{1/2}\} \quad (4.8)$$

for any v in $null(G)$. Whenever $A1'(y)$ holds, the measure of the set in the right hand side of (4.8) converges by Corollary 4.3 to $P_N \left(|\theta| < \delta (\lambda_{\max})^{1/2} r^{-1}(y) \right)$.

Recall that $L_{sa} \subset null(G)$ (Lemma 3.1). (i) is then obtained by choosing $v \in \bar{F}_{sa}(x, u)$. Since the inclusion (4.8) holds actually for any $v \in L_{sa}$ this establishes (ii). ■

Remark The bounds in Theorem 4.4 are of some interest even in the case of stationary control policies, in which case the controlled chain is Markov and the classical CLT for functionals of Markov chains apply. Note however that the limiting variance $\sigma(g, y)$ (defined below) required for the application of that classical CLT is in general hard to compute. At the cost of giving up on a precise limit, Theorem 4.4 provides an easy to compute upper bound also for this stationary case. The exact CLT when using a stationary policy g are the following: $\sqrt{t}[f_s^t(y) - \pi^g(y)]/\sigma(g, y)$ converges in distribution (P^g) to a standard Gaussian RV, where $\sigma(g, y) = \pi^g(y) - [\pi^g(y)]^2$. Similarly, $\sqrt{t}[f_{sa}^t(y, a) - \pi^g(y) p_{a|y}^g]/\sigma(g; y, a)$ converges in distribution (P^g) to a standard Gaussian RV, where $\sigma(g; y, a) = \pi^g(y) p_{a|y}^g - [\pi^g(y) p_{a|y}^g]^2$. (The expression above for $\sigma(g, y)$ follows immediately from the fact that it is the variance of the RV which has π^g as distribution. A similar argument yields the expression for $\sigma(g; y, a)$).

5. LARGE DEVIATIONS FOR CONTROLLED MARKOV CHAINS

In this section, we further restrict ourselves to discuss stationary policies only. As before, we restrict ourselves to the case when $A1$ is satisfied; we do not assume however the stronger hypothesis $A1'(y)$. Since under any stationary policy $g \in U(S)$ the state process is a Markov

Chain, it is well known that for any initial state x, $\lim_{t\to\infty} \bar{f}_s^t(x,g;\cdot) = \pi^g(\cdot)$ and $\lim_{t\to\infty} f_s^t(\cdot) = \pi^g(\cdot), \quad P^g$ *a.s.*.

We compute the exponential rate of convergence of the state frequencies using standard Large Deviations theory and apply them to yield an upper bound for the state frequencies, which is uniform in $u \in U(S)$.

Let W be the class of bounded functions $c : \mathbf{X} \to [1,\infty)$. Define the function $J_g : M_1(\mathbf{X}) \to \mathbb{R}$, $g \in U(S)$ by $J_g(\nu) := \sup_{c\in W}\{-\int \log\left(c^{-1}[P^g c]\right) d\nu\}$.

Lemma 5.1: *([7], Section 3.1.2) Assume A1. Then for every $g \in U(S)$, the Large Deviations Principle holds with rate function $J_g(\nu)$, i.e. for any Borel measurable set $B \subset M_1(\mathbf{X})$,*

$$-\inf_{\nu\in B^o} J_g(\nu) \le \varliminf_{t\to\infty} \frac{1}{t}\log \mathbf{P}_x^g(f_s^t \in B) \le \varlimsup_{t\to\infty} \frac{1}{t}\log \mathbf{P}_x^g(f_s^t \in B) \le -\inf_{\nu\in\bar{B}} J_g(\nu) \tag{5.1}$$

where B^o (and $\bar{B}$) are the interior (respectively the closure) of B.

For the sake of completeness, and in anticipation of future use, we bring here an alternative description of $J_g(\nu)$, using the pair empirical measure, $(t+1)^{-1}\sum_{s=0}^t 1_{\{X_s,X_{s+1}\}}(\cdot,\cdot)$. For $\tilde{\mu} \in M_1(\mathbf{X^2})$, which is stationary (i.e., the marginal law on the first coordinate is equal to the marginal on the second), let $\mu(x) = \sum_{y\in\mathbf{X}} \tilde{\mu}(x,y)$ and let $\mu(x|y) = \tilde{\mu}(x,y)[\mu(y)]^{-1}$, where $\mu(x|y) = 0$ if $\mu(y) = 0$. With some abuse of notations, we identify P^g with the conditional measure that it induces P_{xy}^g. We have then (c.f. [9, (4.4.46)]):

Lemma 5.2:

$$J_g(\nu) = \inf_{\tilde{\nu}\in\Theta} \sum_{y\in\mathbf{X}} \nu(y) H(\nu(\bullet|y)|P_{\bullet y}^g)$$

where

$$\Theta = \left\{ \tilde{\nu}(x,y) \in M_1(\mathbf{X^2}) : \sum_{y\in\mathbf{X}} \tilde{\nu}(x,y) = \sum_{y\in\mathbf{X}} \tilde{\nu}(y,x) = \nu(x) \right\}$$

and

$$H(\nu(\bullet|y)|P_{\bullet y}^g) = \sum_{x\in\mathbf{X}} \nu(x|y)\log\frac{\nu(x|y)}{P_{xy}^g}\,.$$

Remark: Lemma 5.2 is actually Varadhan's characterization of the spectral radius of positive matrices. A transparent proof of it is obtained by applying the contraction principle of large deviation theory to the pair empirical measure.

Lemma 5.1 in principle provides one with the exact rate of convergence of $f_s^t(\cdot)$ for each $g \in U(S)$ used. It is now possible to derive a lower bound on $J_g(\nu)$ which is shown to be proportional to the square of the variation distance between ν and π^g, uniformly in $g \in U(S)$. As a first step, we consider the case where P_{xy}^g is uniformly bounded below. The general case is postponed to Theorem 5.4.

Theorem 5.3: *Assume A1. Assume moreover that $P_{xy}^g > \epsilon$ for some $\epsilon > 0$ and all $x, y \in \mathbf{X}$ and $g \in U(S)$. Then for every policy $g \in U(S)$ the following inequality holds:*

$$J_g(\nu) \geq 0.5\epsilon^2 \parallel \nu - \pi^g \parallel_{var}^2 . \tag{5.3}$$

Proof: For any ν and $g \in U(S)$, we have by an argument of Donsker and Varadhan (c.f. [9, Lemma 4.1.45]),

$$J_g(\nu) \geq 0.5 \parallel \nu - \nu P^g \parallel_{var}^2 . \tag{5.4}$$

To evaluate the right hand side of (5.4), we note that

$$\|\nu P^g - \pi^g\|_{var} \leq (1-\epsilon)\|\nu - \pi^g\|_{var} , \tag{5.5}$$

see [4, Lemma 4.5]. (5.3) then follows from (5.4) and (5.5) since

$$\|\nu - \nu P^g\|_{var} \geq |\|\nu P^g - \pi^g\|_{var} - \|\nu - \pi^g\|_{var}| \geq \epsilon\|\nu - \pi^g\|_{var} .$$

■

We note that the ϵ in the assumption in Theorem 5.3 depends heavily on the dimension of the state space X since it is upper bounded by $1/|\mathbf{X}|$. This fact therefore prevents a direct generalization of the results to more general problems, and in particular to the countable state space case. Nevertheless, in the finite state space case, a sufficient condition for the assumption in Theorem 5.3, i.e. that $P_{xy}^g > \epsilon$ for some $\epsilon > 0$ and all $x, y \in \mathbf{X}$ and $g \in U(S)$, is that $P_{xay} > 0$ for all $x, y \in \mathbf{X}$ and $a \in \mathcal{A}$. Moreover, in the finite state space case, one actually may dispense of the condition involving ϵ, as the next theorem shows:

Theorem 5.4: *Assume A1. Let $\delta > 0$ be given and assume that $||\nu - \pi^g||_{var} > \delta$. Then there exists some positive real number $\Delta(\delta) > 0$ independent of the specific ν chosen such that*

$$J_g(\nu) \geq \Delta(\delta) ,$$

uniformly in $g \in U(S)$.

Proof: Note that, as soon as $||\nu - \pi^g||_{\mathrm{var}} \geq \delta$, one has that $||\nu - \nu P^g||_{\mathrm{var}} > 0$. Since the latter is continuous in ν and g, and since the set $\Theta = \{\nu : ||\nu - \pi^g||_{\mathrm{var}} \geq \delta\}$ is compact, it follows that

$$\inf_{\nu \in \Theta} ||\nu - \nu P^g||_{\mathrm{var}} > \Delta(\delta)$$

for some $\Delta(\delta) > 0$. The theorem follows by combining this with (5.4). ■

Remark An explicit estimate of $\Delta(\delta)$ may be provided as follows. Under $A1$ there exists some positive ϵ and an integer M such that $[P^M]^g_{xy} > \epsilon$ for all $x, y \in \mathbf{X}$ and $g \in U(S)$ (see the proof of Lemma 4.6 in [4]). It follows immediately that $\Delta(\delta) = \epsilon^2\delta^2/M$. In the sequel, we use the assumption that $P^g_{xy} > \epsilon$, recalling that by using M as above one may always transform the general situation to this one.

In the following Corollary we obtain a uniform upper bound on the distance between the state frequencies and the set $L_s(U(S))$.

Corollary 5.5: *Assume $A1$, and assume that all entries of P^g, $g \in U(S)$ are greater than some $\epsilon > 0$. Let B be any closed set such that $B \subset \{\nu \in M_1(\mathbf{X}), \inf_{\theta \in L_s} \parallel \nu - \theta \parallel^2_{var} \geq \delta\}$. Then $\overline{\lim}_{t \to \infty} t^{-1} \log \mathbf{P}^g_x(f^t_s \in B) \leq -0.5\epsilon^2\delta$ uniformly in $g \in U(S)$.*

Proof: Follows immediately from Lemmas 5.1 and Theorem 5.3. ■

It follows from assumption $A1$ and Lemma 3.1 (i)-(ii) that for every $y \in \mathbf{X}$ there exist positive constants $\Delta_1(y) > 0$ and $\Delta_2(y) < 1$ such that for every $x \in \mathbf{X}$, $u \in U$ and $z \in \bar{F}_s(x, u; y)$ we have $\Delta_2(y) \geq z \geq \Delta_1(y)$. In particular, under any $g \in U(S)$ the stationary probability π^g satisfies $\Delta_2(y) \geq \pi^g(y) \geq \Delta_1(y)$. From Corollary 5.5 we immediately obtain the following Corollary:

Corollary 5.6: *Assume $A1$, and assume that all entries of P^g, $g \in U(S)$ are greater than some $\epsilon > 0$. Then for any $x, y \in \mathbf{X}$, $g \in U(S)$ and $\delta > 0$ we have:*

$$\overline{\lim_{t \to \infty}} \frac{1}{t} \log \mathbf{P}^g_x(f^t_s(y) \leq \Delta_1(y) - \delta^{1/2}) \leq -0.5\epsilon^2\delta$$

and similarly

$$\overline{\lim_{t \to \infty}} \frac{1}{t} \log \mathbf{P}^g_x(f^t_s(y) \geq \Delta_2(y) + \delta^{1/2}) \leq -0.5\epsilon^2\delta\,.$$

6. OPTIMAL CONTROL AND RATE OF CONVERGENCE

In this Section we introduce optimal control problems (known as Markov Decision Processes); after relating the cost to the frequencies in Subsection 6.1, we obtain in Subsection 6.2 bounds on the rate of convergence of the frequencies that correspond to special classes of control policies. In Subsection 6.3 we finally obtain bounds on the rate of the convergence of the finite-horizon costs to the expected average costs.

6.1 Optimal control problems and relation between cost and frequencies.

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

The *expected average* costs:

$$C_{ea}^t(x,u) := \frac{1}{t+1} E^u \left[\sum_{s=0}^{t} c(X_s,A_s) \mid X_0 = x \right] ,$$

$$D_{ea}^{t,k}(x,u) := \frac{1}{t+1} E^u \left[\sum_{s=0}^{t} d^k(X_s,A_s) \mid X_0 = x \right] , \qquad k=1,...,K ,$$

$$C_{ea}(x,u) := \overline{\lim_{t\to\infty}}\, C_{ea}^t(x,u) , \qquad D_{ea}^k(x,u) := \overline{\lim_{t\to\infty}}\, D_{ea}^{t,k} , \quad k=1,...,K .$$

The *sample average* costs are the random variables, given by

$$C_{av}^t := \frac{1}{t+1} \sum_{s=0}^{t} c(X_s,A_s) , \qquad D_{av}^{t,k} := \frac{1}{t+1} \sum_{s=0}^{t} d^k(X_s,A_s) , \quad k=1,...,K ,$$

$$C_{av} := \overline{\lim_{t\to\infty}}\, C_{av}^t , \qquad D_{av}^k := \overline{\lim_{t\to\infty}}\, D_{av}^{t,k} , \quad k=1,...,K .$$

Given real numbers $\{V_k\ ,\ k=1,...,K\}$ define the following control problems:

CP (Constrained Problem) Find a policy that satisfies

$$D_{ea}^k(x,u) \le V_k , \qquad 1 \le k \le K . \tag{6.1}$$

Denote all the policies that satisfy (6.1) by U_{cp}.

COP (Constrained Optimization Problem)

Find a policy $u \in U_{cp}$ that minimizes $C_{ea}(x,u)$.

Denote the set of all policies that are optimal for COP by U_{cop}.

In the following Lemma we show that information on costs may be obtained from information on the (expected or not) empirical measures.

Lemma 6.1: *For every policy $u \in U$ and any instantaneous cost function c,*
(i)

$$C^t_{ea}(x,u) = \sum_y \sum_a c(y,a) \bar{f}^t_{sa}(x,u;y,a)\,,$$

$$C^t_{av} = \sum_y \sum_a c(y,a) f^t_{sa}(y,a)\,.$$

(ii) For any $\zeta \in \bar{F}_{sa}(x,u)$,

$$C_{ea}(x,u) \geq \sum_y \sum_a c(y,a)\zeta(y,a)\,, \tag{6.2}$$

and there exists some $\zeta \in \bar{F}_{sa}(x,u)$ for which equality is obtained.

Proof: (i) follows immediately from the definitions of the frequencies and expected frequencies. (ii) is proved in [8] pp. 89. ■

Conclusion: For any policy for which $\bar{F}_{sa}(x,u)$ is a singleton, the convergence of $\bar{f}^t_{sa}(x,u)$ to its limit implies the convergence $\lim_{t\to\infty} C^t_{ea}(x,g) = C_{ea}(x,g)$. If moreover u is a policy such that f^t_{sa} converges a.s. to the limit of $\bar{f}^t_{sa}(x,u)$ then $\lim_{t\to\infty} C^t_{av} = C_{ea}(x,g) \quad P^g$ *a.s.*.

6.2 Rate of convergence of frequencies in optimal control problems

Next we restrict results obtained in previous sections to the limit frequencies obtained by the set of policies which are feasible for CP or for COP, i.e. U_{cp} and U_{cop}. We thus present the analog results to Theorem 4.4 and Corollary 5.5. Recall the definition of the class of policies U^* from Section 4.

Corollary 6.2: *Assume A1. Let $Y \subset \mathbf{X}$ be the set of states for which $A1'(y)$ holds. Then for any $x \in \mathbf{X}$, any class of policies $\tilde{U}$ and $u \in \tilde{U} \cap U^*$ we have*

$$\overline{\lim_{t\to\infty}}\, P^u_x \left(t^{1/2} d(f^t_{sa}, L^x_{sa}(\tilde{U})) \leq \delta \right) \leq \min_{y\in Y} P_N \left(|\theta| < \delta(\lambda_{\max})^{1/2} r^{-1}(y) \right)\,, \tag{6.3}$$

where

$$d(f^t_{sa}, L_{sa}(\tilde{U})) := \inf_{v \in L_{sa}(\tilde{U})} ||f^t_{sa} - v||_2.$$

In particular, (6.3) holds for $\tilde{U} = U^*, U_{cp}, U_{cop}$ *respectively.*

Proof: We use the same arguments as in the proof of Theorem 4.4 (ii). Since the inclusion (4.8) holds actually for any $v \in L_{sa}$ and thus in particular for $v \in L^x_{sa}(\tilde{U})$ this establishes the proof. ■

Next we adapt the Large Deviation type bound to the policies U_{cp} and U_{cop}, to obtain the equivalent of Corollary 5.5. We obtain a uniform upper bound for the distance between the state frequencies and the sets $L_s(U_{cp})$ and $L_s(U_{cop})$.

Corollary 6.3: *Assume A1, and assume that all entries of* P^g, $g \in U(S)$ *are greater than* ϵ. *Let* B_{cp} *be any closed set such that* $B_{cp} \subset \{\nu \in M_1(\mathbf{X}), \inf_{\theta \in L_s(U_{cp})} \parallel \nu - \theta \parallel^2_{var} \geq \delta\}$, *and let* B_{cop} *be any closed set such that* $B_{cop} \subset \{\nu \in M_1(\mathbf{X}), \inf_{\theta \in L_s(U_{cop})} \parallel \nu - \theta \parallel^2_{var} \geq \delta\}$. *Then*

$\overline{\lim}_{t\to\infty} t^{-1} \log \mathbf{P}^g_x(f^t_s \in B_{cp}) \leq -0.5\epsilon^2\delta$ *uniformly in* $g \in U_{cp}$,

$\overline{\lim}_{t\to\infty} t^{-1} \log \mathbf{P}^g_x(f^t_s \in B_{cop}) \leq -0.5\epsilon^2\delta$ *uniformly in* $g \in U_{cop}$.

Proof: Follows immediately from Lemmas 5.1 and 5.3. ■

Remark: The assumption on the positivity of all entries of P^g can be weakened in Corollary 6.3 following the proof of Lemma 4.6 in [4].

6.3 Large Deviation Principle for the cost.

Having computed a lower bound on the exponential rate of convergence of the state frequencies in Section 5 which is uniform in $u \in U(S)$, this is then used in this subsection to obtain, via the contraction principle, a lower bound on the rate of convergence of the costs C^t_{av} to C_{ea}.

We assume below that the cost does not depend on the action, i.e. $c(y,a) = c(y)$, $\forall y, a$.

Lemma 6.4: *Assume A1. Then for every* $g \in U(S)$, *the Large Deviations Principle holds with rate function* $I_g(\xi)$, *i.e. for any Borel measurable set* $B \subset \mathbb{R}$,

$$-\inf_{\xi \in B^o} I_g(\xi) \leq \underline{\lim}_{t\to\infty} \frac{1}{t} \log \mathbf{P}^g_x(C^t_{av} \in B) \leq \overline{\lim}_{t\to\infty} \frac{1}{t} \log \mathbf{P}^g_x(C^t_{av} \in B) \leq -\inf_{\xi \in \bar{B}} I_g(\xi) \qquad (6.4)$$

where $I_g(\xi)$ is given by

$$I_g(\xi) = \inf\left\{ J_g(\nu')|\nu' \in M_1(\mathbf{X}), \xi = \sum_{y\in X} c(y)\nu'(y) \right\};$$

B^o (and $\bar{B}$) are the interior (respectively the closure) of B. Moreover, (6.4) holds with equality for closed sets which are the closure of open sets in $\mathbb{R}$.

Proof: The first claim is obtained by applying the contraction principle (c.f. [9], Lemma 2.1.4). The second claim follows since $I_g(\xi)$ is a convex and lower semi continuous function on $\mathbb{R}$, and hence continuous in its (open) domain. ■

Corollary 6.5: *Assume A1, and assume that $P^g_{xy} > \epsilon$ for some $\epsilon > 0$ and all $x, y \in \mathbf{X}$ and $g \in U(S)$. Then for every $\Delta > 0$ and policy $g \in U(S)$*

$$\overline{\lim_{t\to\infty}} \frac{1}{t} \log \mathbf{P}^g_x \left(|C^t_{av} - C_{ea}(x,g)| \geq \Delta \right) \leq -\frac{\epsilon^2\Delta^2}{2\max_y |c(y)|}, \tag{6.5}$$

and for any real number V:

$$\overline{\lim_{t\to\infty}} \frac{1}{t} \log \mathbf{P}^g_x \left(C^t_{av} \geq V \right) \leq -\frac{\epsilon^2 (C_{ea}(x,g) - V)^2}{2\max_y |c(y)|}. \tag{6.6}$$

Proof: Let $B' := \{w \in M_1(\mathbf{X}); |\sum_y c(y)(w(y) - \pi^g(y))| \geq \Delta\}$ and
$B'' := \{w \in M_1(\mathbf{X}); \max_y |c(y)| \cdot \|w - \pi^g\|_{var} \geq \Delta\}$.
Since $|\sum_y c(y)(w(y) - \pi^g(y))| \leq \max_y |c(y)| \cdot \|w - \pi^g\|_{var}$, it follows that $B' \subset B''$. To evaluate the right side of (6.4), which will yield the upper bound sought in the statement of the corollary, note that

$$\inf_{\left\{\xi : \left|\xi - \sum_{y\in\mathbf{X}} c(y)\pi^g(y)\right| \geq \Delta\right\}} I_g(\xi) = \inf\{J_g(\nu')|\nu' \in B'\} \geq \inf\{J_g(\nu')|\nu' \in B''\} \geq \frac{\epsilon^2\Delta^2}{2\max_y |c(y)|}$$

where the last equality follows from Theorem 5.3. This establishes (6.5). (6.6) is obtained similarly. ■

Remark Again, one may dispense of the assumption that $P^g_{xy} > \epsilon$ by using Theorem 5.4 instead of Theorem 5.3. In this case, ϵ^2 in (6.5) and (6.6) need to be replaced by $\Delta(\delta)$.

7. OPTIMAL RATE CONTROL AND AN APPLICATION TO TELECOMMUNICATION SYSTEMS

In this section, we suggest a method of incorporating a performance measure on the convergence rates in average type control problems into the problem's formulation. In such problems, the optimal control may not to be unique but may rather belong to some convex closed set. It seems natural therefore, when one is faced with the problem of choosing one of the optimal controls, to use the one which achieves as fast as possible the optimal performance. This can be done in two ways: one may, among all the optimal control, choose the one which achieves the fastest rate of convergence (in a Large Deviations sense, say). We refer to this as "lexicographic optimization", for which the second criterion is applied on the solutions of the first optimization problem. An alternative, which we describe in some details in this section, would be to impose some minimal rate of convergence on the solution of the control problem and to seek an optimal control which achieves this rate. To be specific, we consider the following prototype problem of optimal scheduling of interactive and noninteractive traffic in telecommunication systems, which serves as motivation for studying a class of optimal control problems where the objectives include rates of convergence. Consider a discrete time system; packets of information of K different types, such as file transfers, video and voice signals, compete for access of some shared resource. Each type of arriving packets waits in a buffer of finite capacity till it gets access to the resource. If an arriving packet finds its buffer full then it is lost. 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. We assume that with positive probability, each packet may be removed from the queue before starting service.

The traffic types are partitioned into two groups: interactive and noninteractive. The average delay of each of the interactive types is required to be below some given thresholds. The problem COP' is to find a scheduling policy that minimizes a linear combination of the average delays for the noninteractive types subject to the constraint above. Similar problems were solved under different assumptions on the arrival and service statistics by Ross and Chen in [15] and by Altman and Shwartz [1] and references therein. However, Ross and Chen point out in [15] that the model is inappropriate for voice interactive signals due to the choice of average cost criteria: "We remark that the model studied here would not be appropriate if real-time voice packets were also competing for the resource. This is because [COP] imposes constraints on the average delay ... and not on the actual delay."

We thus reformulate the model above as a constrained control problem (COP^*) that is more sensitive to the actual delay of voice signal than COP', by introducing constraints on the large deviation rate function. We shall focus on the events that the actual average delays of speech signals till time t are larger than some given thresholds; We then replace the original constraints in COP' that relate to the voice traffic with constraints on the asymptotical rate at which the probability of such events decays to zero (as t goes to infinity). Although this kind of constraints is also asymptotic, it seems to be more sensitive to variability in the actual delay of voice signals than the previous average type constraints. In fact it enables us to perform another more refined selection among different policies that are optimal for COP'. Note that it is not possible to impose constraints on the actual delay, since due to noise, the service may fail again and again, so that the actual delay can not be bounded uniformly in the chance parameter ω.

Formulation of the problem

We shall model the problem as the following discrete-time queuing system. A generic element of the state is given by $x = \{x^1, x^2, ..., x^K\}$, $\ 0 \leq x^k \leq N \ \ k = 1, ..., K$ and it represents a K dimensional vector of the different queues' size. An action $a \in \mathcal{A}$ where $\mathcal{A} = \{1, 2, ..., K\}$ represents the type of traffic to which service is assigned at each time unit. Assume that the arrivals of different types are independent; moreover, the arrivals of each type are independent from slot to slot. We assume that there is a positive probability for any vector of arrivals at each time unit, and therefore, under any action, all transition probabilities are larger than some $\epsilon > 0$.

Let L_1 and L_2 be given nonnegative integers with $L_1 \leq L_2 \leq K$. Let $1, 2, ..., L_1$ be the set of queues that are assigned for voice traffic, $L_1 + 1, L_1 + 2, ..., L_2$ the set of queues that are assigned for other interactive traffic, and $L_2 + 1, ..., K$ the set of queues assigned to the remaining noninteractive traffic.

Consider the following linear instantaneous cost functions: $d^k(x, a) = x^k$ for $1 \leq k \leq L_2$, and $c(x, a) := \sum_{k=1}^{K-L_2} c_k x^{k+L_2}$, where c_k, $k = 1, 2, ..., K - L_2$ are non-negative constants.

Note that with these definitions of the costs, $C_{ea}(x, u)$ has the interpretation of a linear combination of the expected average length of queues $L_2 + 1, L_2 + 2, ..., K$, and $D^k_{ea}(x, u)$ is the expected average length of queue k, where $k = 1, 2, ..., L_2$. By Little's law these quantities are proportional to the respective waiting times in the different queues. However, the rest of this

Section does not rely on this specific choice of the immediate costs $c(\cdot,\cdot)$ and $d^k(\cdot,\cdot)$, and in fact the results hold for general nonnegative immediate costs provided that they do not depend on the action.

Let x be a given initial state and let V be a given vector of L_2 positive real numbers. Similarly to the definition in Section 6, problem COP' is to find a policy $u \in U(S)$ that minimizes $C_{ea}(x,u)$ s.t. $D_{ea}(x,u) \le V$. We shall also mention problem CP' which is to find a policy $u \in U(S)$ s.t. $D_{ea}(x,u) \le V$.

Let r be a given vector of L_1 positive real numbers. In view of the discussion above concerning the modeling for voice signals, consider the following problems

CP* : Find a policy $u \in U(S)$ s.t.

$$\overline{\lim}_{t\to\infty} \frac{1}{t} \log P_x^u(D_{av}^{k,t} \ge V_k) \le -r_k \,, \qquad 1 \le k \le L_1 \,, \tag{7.1a}$$

$$D_{ea}^k(x,u) \le V_k \,, \qquad L_1 < k \le L_2 \,. \tag{7.1b}$$

COP* : Find a policy $u \in U(S)$ that minimizes $C_{ea}(x,u)$ s.t.

$$\overline{\lim}_{t\to\infty} \frac{1}{t} \log P_x^u(D_{av}^{k,t} \ge V_k) \le -r_k \,, \qquad 1 \le k \le L_1 \,, \tag{7.2a}$$

$$D_{ea}^k(x,u) \le V_k \,, \qquad L_1 < k \le L_2 \,. \tag{7.2b}$$

Note that the definition of problem COP^* assumes the existence of an optimal solution. In Theorem 7.5 we provide conditions for the existence of such a solution.

Solving COP^* and CP^*

We need first the following two Lemmas that apply in general to COP and CP (and not only to COP' and CP').

Given some $z \in M_1(\mathbf{X} \times \mathcal{A})$, let $u(z)$ be the stationary policy that satisfies

$$\pi_{a|y}^{u(z)} := z(y,a)[\sum_{a'} z(y,a')]^{-1} \tag{7.3}$$

whenever this is well defined. Let Y be as defined in Lemma 3.1.

Lemma 7.1: *Assume A1 and pick any initial state x. Then (i) for any $z \in Y$, $u(z)$ is well defined and $\bar{f}_{sa}(x,u(z)) = z$. (ii) $C_{ea}(x,u(z)) = c \cdot z$ and $D_{ea}^k(x,u(z)) = d^k \cdot z, \quad k = 1,...,L_2$.*

Proof: see e.g. [12] ■

Let $\Phi_v^k := \{z : z \cdot d^k \leq V_k\}$, let $\Phi_v = \cap_k \Phi_v^k$ and define $Y_v := Y \cap \Phi_v$ Note that Y_v is a closed convex polytope.

Lemma 7.2: *Assume A1. Then*

(i) $L_{sa}(U_{cp}) = Y_v$.

(ii) $L_s(U_{cp}) \subset L_s$ *is a (closed convex) polytope and equals the set*

$\{z \in \mathbb{R}^{|\mathbf{X}|} : z = \sum_{a \in \mathcal{A}} z'(y,a)$ *for some* $z'(y,a) \in L_{sa}(U_{cp})\}$.

(iii) $L_{sa}(U_{cop}) \subset L_{sa}(U_{cp})$ *and* $L_s(U_{cop}) \subset L_s(U_{cp})$ *are (closed convex) polytopes.*

Proof: It follows from Lemma 7.1 that $L_{sa}(U(S) \cap U_{cp})$ is closed and convex. Now,

$$L_{sa}(U(S) \cap U_{cp}) = L_{sa}(U(S)) \cap L_{sa}(U_{cp}) = L_{sa} \cap L_{sa}(U_{cp}) = L_{sa}(U_{cp}),$$

which establishes the proof of (i).

(ii) follows from the fact that for any t and $x, y \in \mathbf{X}$ we have

$\bar{f}_s^t(x,u;y) = \sum_{a \in \mathbf{X}} \bar{f}_{sa}^t(x,u;y,a)$.

Finally (iii) follows from the fact ([12]) that $L(U_{cop})$ is given by the optimal set of the linear programming problem: minimize $c \cdot z$ subject to $z \in Y_v$; the rest of the derivation follows the same line as (i) and (ii). ■

We present below a mathematical program that is equivalent to CP^*, in the sense that any feasible solution for it provides a feasible solution for CP^* and vice versa. Since the computation of feasible solutions for MP below may be quite complex, we use bounds from Section 5 to obtain a Linear Program, whose set of feasible solutions are a subset of the feasible set of the Mathematical Program. Hence any feasible solution for it provides a feasible solution for CP^*. A similar approach is then used to solve COP^* by a Linear Program. Define

$$I_g^k(\xi) = \inf\left\{ J_g(\nu') | \nu' \in M(\mathbf{X}), \xi = \sum_{y \in X} d^k(y)\nu'(y) \right\}, \qquad k = 1, 2, ..., L_1.$$

By Lemma 5.2 an alternative expression for $I_g^k(\xi)$ is

$$I_g^k(\xi) = \inf\left\{ \sum_{y \in \mathbf{X}} \nu(y) H(\nu(\bullet|y)|P_{\bullet y}^g) : \tilde{\nu}(x,y) \in \Psi_s \right\}$$

where

$$\Psi_s := \left\{ \tilde\nu(x,y) : \tilde\nu(x,y) \in M_1(\mathbf{X^2}), \sum_{x\in\mathbf{X}} \tilde\nu(x,y) = \sum_{x\in\mathbf{X}} \tilde\nu(y,x), \xi = \sum_{x,y\in\mathbf{X^2}} d^k(y)\tilde\nu(x,y) \right\}.$$

Consider the following problem:

MP: Find $z \in Y$ (defined in Lemma 3.1 (ii)) s.t.

$$\inf \left\{ I^k_{u(z)}(\xi) : \ \xi \geq V_k \ \right\} \geq r_k\,, \quad k = 1,2,...,L_1\,, \tag{7.4a}$$

$$\sum_{y,a} z(y,a) d^k(y) \leq V_k\,, \qquad k = L_1+1, L_1+2, ..., L_2. \tag{7.4b}$$

We remark that (7.4a) is equivalent to the minimization

$$\inf \left\{ \sum_{x,y\in\mathbf{X^2}} \nu(y) H(\nu(\bullet|y)|P^g_{\bullet y}) : \tilde\nu(x,y) \in \Psi_s, \sum_{y\in\mathbf{X}} d^k(y)\tilde\nu(x,y) \geq V_k \right\} \geq r_k, \quad k = 1,2,...,L_1.$$

Theorem 7.3: *Assume A1. Then*

(i) for any solution z^* of MP, $u(z^*)$ (defined in (7.3)) is feasible for CP^*,

(ii) for any $g \in U(S)$ that solves CP^*, z^* is a solution of MP, where z^* is given by

$$z^*(y,a) := \pi^g(y) p^g_{a|y}\,.$$

Proof: (i) Assume that z solves MP. Then (7.1b) holds from (7.4b) and following Lemma 7.1 (ii). (7.1a) holds due to (6.4) and (7.4a).

(ii) Assume that g solves CP^*. For any $1 \leq k \leq K_2$, $D^k(x,g) = z^* \cdot d^k$ and hence (7.1b) implies (7.4b). Note that since z^* is achieved using $P^g_{a|y}$, it follows that $z^* \in Y$ (defined in Lemma 3.1). Next, note that $u(z^*) = g$ (where $u(z^*)$ is defined in (7.3)), hence (7.1a) together with (6.4) and the fact that the equality is indeed achieved in the large deviations bound (due to the second claim of Lemma 6.4) implies (7.4a), which establishes the proof. ■

Consider the set Z of $\{z \in \mathbb{R}^{|\mathbf{X}|\times|\mathcal{A}|}\}$ given by the following linear inequalities:

$$z \cdot d^k \leq V_k - \left[\frac{2\max_y d^k(y)}{\epsilon^2 r_k} \right]^{1/2}, \qquad k = 1,2,...,L_1\,, \tag{7.5a}$$

$$\sum_{y,a} z(y,a) d^k(y) \le V_k\,, \qquad k = L_1+1, L_1+2, ..., L_2\,, \tag{7.5b}$$

$$G(z) = 0, \qquad z \ge 0, \qquad \sum_{y\in\mathbf{X}, a\in\mathcal{A}} z(y,a) = 1\,.$$

Theorem 7.4: *Assume A1. If $z \in Z$ then $u(z)$ (defined in (7.3)) is feasible for CP^*.*

Proof: Let $z \in Z$. By Lemma 7.1 (ii), (7.1b) holds. By (6.6)

$$\overline{\lim_{t\to\infty}} \frac{1}{t} \log P^g\left(D_{av}^{k,t} \ge V_k\right) \le -\frac{\epsilon^2 (D_{ea}^k(x,g) - V_k)^2}{2 \max_y d^k(y)}\,.$$

By Lemma 7.1 (ii) and (7.5a),

$$(V_k - D_{ea}(x, u(z)))^2 \ge \frac{2\max_y d^k(y)}{\epsilon^2 r_k}\,,$$

and hence (7.1a) holds. ■

Consider next the following Linear Program

LP: Find $z \in Z$ that minimizes $c \cdot z$.

In the following Theorem we show that bounds on rate of convergence may be used for a further selection between policies that are optimal for COP'. Let B^* be the set of optimal solutions for LP, and and let B be the set of optimal solutions for LP when omitting the constraint (7.5a).

Theorem 7.5: *Assume A1. Assume that $B^* \subset B$ and B^* is nonempty. Then for any $z \in B^*$, $u(z)$ (defined in (7.3)) is optimal for COP^*.*

Proof: By [12], for any $z \in B^*$, $u(z)$ is optimal for COP^* when omitting the constraint (7.2a). By Theorem 7.4, for any $z \in B^*$, $u(z)$ is feasible for COP^*. Combining these arguments, it follows that $u(z)$ is optimal for COP^*. ■

Acknowledgments We thank the referees for a careful reading of the original manuscript and for helpful remarks.

REFERENCES

[1] Altman E. and A. Shwartz, "Optimal priority assignment: a time sharing approach", *IEEE Trans. on Automatic Control* Vol. AC-34 No. 9, pp. 1098-1102, 1989.

[2] — , "Non-stationary policies for controlled Markov Chains", EE. PUB. No. 633, Department of Electrical Engineering, Technion, June 1987, to appear, Operations Research.

[3] — , "Adaptive control of constrained Markov chains", *IEEE Trans. Auto. Control.* **36** No. 4, pp. 454-462, 1991.

[4] — , "Adaptive control of constrained Markov chains: criteria and policies", *Annals of Oper. Res.*, **28**, special issue on Markov Decision Processes, pp. 101-134, 1991.

[5] Bertsekas D. and Gallager R., *Data Networks*, Prentice Hall, 1987.

[6] Brown B. M., "Martingale Central Limit Theorems", *Annals of Math. Stat.*, Vol. 42, No. 1, pp. 59-66, 1971.

[7] Dembo A. and Zeitouni O., *Large Deviations Techniques and Applications,* Jones and Bartlett, 1993.

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

[9] Deuschel J. D. and Stroock D., *Large Deviations*, Academic Press, 1989.

[10] Hall P. and C. C. Heyde, *Martingale Limit Theory and its Applications,* John Wiley, New York, 1980.

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

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

[13] Mandl P, "Estimation and Control in Markov Chains", *Adv. Appl. Probability,* **4**, pp. 279-290, 1974.

[14] Mandl P, "On the Adaptive Control of Countable Markov Chains", *Probability Theory* Banach Center Publications, Vol. 5, PWN – Polish Scientific Publishers, pp. 159-173, 1979.

[15] Ross K. W. and B. Chen, "Optimal Scheduling of interactive and noninteractive traffic in Telecommunication Systems", *IEEE Trans. Auto. Cont.,* Vol. 33, No. 3, pp. 261-267, March 1988.

[16] Ross S., *Introduction to Stochastic Dynamic Programming*, Academic Press, 1984.