ADAPTIVE CONTROL OF CONSTRAINED MARKOV CHAINS: CRITERIA AND POLICIES*

Eitan Altman and Adam Shwartz

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

ABSTRACT

We consider the constrained optimization of a finite-state, finite action Markov chain. In the adaptive problem, the transition probabilities are assumed to be unknown, and no prior distribution on their values is given. We consider constrained optimization problems in terms of several cost criteria which are asymptotic in nature. For these criteria we show that it is possible to achieve the same optimal cost as in the non-adaptive case.

We first formulate a constrained optimization problem under each of the cost criteria and establish the existence of optimal stationary policies.

Since the adaptive problem is inherently non-stationary, we suggest a class of "*Asymptotically Stationary*" (AS) policies, and show that, under each of the cost criteria, the costs of an AS policy depend only on it's limiting behavior. This property implies that there exist optimal AS policies. A method for generating adaptive policies is then suggested, which leads to strongly consistent estimators for the unknown transition probabilities. A way to guarantee that these policies are also optimal is to couple them with the adaptive algorithms of [3]. This leads to optimal policies for each of the adaptive constrained optimization problems under discussion.

Submitted October 1989. Revised March 1990.

---

* This work was supported in part through United States — Israel Binational Science Foundation Grant BSF 85-00306.

1. INTRODUCTION.

The problem of adaptive control of Markov chains has received considerable attention in recent years; see the survey paper by Kumar [18], Hernandez-Lerma [10] and references therein. In the setup considered there, the transition probabilities of a Markov chain are assumed to depend on some parameter, and the "true" parameter value is not known. One then tries to devise a control policy which minimizes a given cost functional, using on-line estimates of the parameter. In most of the existing work, the expected average cost is considered. Recently Schäl [22] introduced an asymptotic discounted cost criterion. Adaptive optimal policies with respect to the latter criterion were investigated by Schäl [23], and a variation of this criterion was considered by Hernandez-Lerma and Marcus [11,13,12]. Concerning the Bayesian approach to this problem, see Van Hee [25] and references therein.

Since we adopt a non-Bayesian framework it is natural, in order to formulate an adaptive problem, to consider only those cost criteria which do not depend on the finite-time behavior of the underlying controlled Markov chain. For such criteria, it may be possible to continuously improve the policy, using on-line parameter estimates, so as to obtain the same (optimal) performance as in the case where all parameters are known. We consider the well-known Expected Average cost and Sample Average (or ergodic) cost, the Schäl discounted cost, and introduce a new Asymptotic Discounted cost. In Section 2 we define these cost functionals and formulate a constrained and an adaptive constrained optimization problem for each cost structure. For example, the Average-Cost constrained problem amounts to minimizing an average cost, subject to inequality constraints in terms of other average cost functionals.

Adaptive policies for constrained problems were first introduced by Makowski and Shwartz [19,24] in the context of the expected average cost and under a single constraint. They rely on the Lagrange approach of Beutler and Ross [4], which is limited to a single constraint. The general constrained adaptive control problem of a finite state chain under the expected average cost is considered by Altman and Shwartz [3]. Using the "Action Time-Sharing" (ATS) policies, they obtain optimal adaptive policies. In Section 3 we recall the definition of these policies and show that they are useful for the sample average cost as well. We obtain some useful characterizations of the average and the expected average costs.

The adaptive control paradigm is roughly the following. At each decision epoch;

(i) Compute some estimate of the unknown parameters,

(ii) Use these values to compute a control policy,

(iii) Apply the policy of (ii) and repeat.

The "certainty equivalence" approach to the computation in step (ii) is to treat the values provided in (i) as the correct parameter values, and compute the corresponding optimal (stationary) control. There are several drawbacks to this approach; first, the complexity of computing optimal policies is very high [10]. In the constrained case, the optimal policy may be a discontinuous function of the parameter [3]. A more fundamental consideration is the following. Clearly a poor estimate may result in a bad choice of control. Such a choice may in turn suppress estimation, say by using only such control actions which do not provide new information. The use of a stationary policy may cause the algorithm to "get stuck" at an incorrect value of the parameter, hence also at a suboptimal policy.

We propose the following approach to the adaptive control problem. We first show that the costs depend only on limiting properties of the control. This makes it possible to apply "forced choices" of controls which are not optimal, but enhance estimation (see e.g. [18]). Consistent estimation is guaranteed by making these choices sufficiently often. We show that it is possible to achieve consistent estimation without changing the limiting properties. Finally, we choose a convenient implementation of a policy, incorporating the "forced choices" and possessing the correct limiting behavior to ensure optimality.

Section 3 treats the average cost and serves as a template for the development in this paper, and below we describe the analogue development in Section 3 vs. Sections 4–7. In Section 3 we first recapitulate the existence of optimal stationary policies for the (non-adaptive) constrained average cost problem. It is well known that this problem can be solved via a linear program (see e.g. [15], which extends the linear program used for the non constrained case [8 p. 80]). This suggests a method for the computation in (ii). These steps are followed for the other cost criteria in Section 4, where we show that there exists an optimal stationary policy for the constrained (non-adaptive) problem, and in Section 5 where we show that the Linear Program of [15] applies also for the asymptotic discounted cost. An extension of another Linear Program [8 p. 42] is shown to provide a method of computing stationary policies for the constrained non-adaptive problem under the discounted and the Schäl cost criteria. Unlike the non-constrained case, the optimal policies for the discounted problems depend on the initial state.

Section 3 proceeds to discuss the ATS class of policies (which includes the stationary policies) for which limits of the conditional frequencies (3.3) exist. We then show (following [3]) that average costs depend only on these limits. This provides conditions under which forced choices do not affect the final average cost. ATS policies are not applicable to the other cost criteria, since the other costs are not determined by these limits. The key concept for these cost criteria is the new and more refined limits introduced in (1.2). Policies possessing these limits are called *"Asymptotically Stationary"* (AS). In Section 6 we show that under those policies there exists a limiting distribution for the state of the process, which coincides with the invariant distribution under the corresponding stationary policy. Moreover, for each of the cost criteria, the cost of an AS policy equals the cost under the corresponding stationary policy. It follows that AS policies possess the asymptotic properties needed to achieve optimality, while retaining the flexibility which is needed to obtain consistent estimation; they are thus useful for the adaptive problems.

In Section 7 we construct an optimal adaptive policy. To do so, we first show how to compute estimates of the unknown transitions from the history of previous states and actions. We then show how to compute the (sub-optimal) stationary policy that is used till the next estimate is obtained. This stationary policy incorporates "forced choices" of actions for the sake of enhancing estimation. We finally show that this scheme yields an AS policy that is optimal for the constrained adaptive problem.

The model.

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* $\mathbf{A}$; $\mathbf{X}$ and $\mathbf{A}$ are equipped with the corresponding discrete topologies $2^{\mathbf{X}}$ and $2^{\mathbf{A}}$. Without loss of generality, we assume that in any state x all actions in $\mathbf{A}$ 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 $\mathbf{B}$ obtained by 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 for any history $h_{t-1} \in (\mathbf{X} \times \mathbf{A})^t$ of states and actions till time $t-1$, the next state will be y with probability

$$P_{xay} := P(X_{t+1} = y \mid X_t = x; A_t = a) = P(X_{t+1} = y \mid H_{t-1} = h_{t-1}, X_t = x; A_t = a) \tag{1.1}$$

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 distribution over $\mathbf{A}$. Each policy u and initial state x induce a probability measure P_x^u on $\{\Omega, \mathbf{B}\}$. The corresponding expectation operator is denoted by E_x^u.

A *Markov policy* $u \in U(M)$ is characterized by the dependence of $u_t(\cdot \mid H_{t-1}, X_t)$ on X_t only, i.e. for each t, $u_t(\cdot \mid H_{t-1}, X_t) = \tilde{u}_t(\cdot \mid X_t)$. A *stationary policy* $g \in U(S)$ is characterized by a single conditional distribution $p^g_{\cdot|x} := u(\cdot \mid X_t = x)$ over $\mathbf{A}$; under g, X_t becomes a Markov chain with stationary transition probabilities, given by $P^g_{xy} := \sum_{a \in \mathbf{A}} p^g_{a|x} P_{xay}$. 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.

Call a policy u *Asymptotically Stationary* (AS) if for some stationary policy g,

$$\lim_{t \to \infty} P^u_x(A_t = a | X_t = y, H_{t-1}) = p^g_{a|y} \quad P^u_x \ a.s. \tag{1.2}$$

for any initial state x and all a and y. Any policy satisfying (1.2) is denoted by $\hat{g}$, and $\mathbf{G}(g)$ is the set of such policies corresponding to g. The class of policies such that (1.2) holds for *some* g in $U(S)$ is denoted by $U(AS)$.

Throughout the paper we impose the following assumption:

A1: Under any stationary 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 $\pi^g(\cdot)$.

The following notation is used below: $\delta_a(x)$ is the Kronecker delta function. For an arbitrary set B, $1[B]$ is the indicator function of the set and $cl\ B$ the closure of B. For any function $\xi : B \to \mathbb{R}$ we define $\xi^-(y) := \max\{0, -\xi(y)\},\ y \in B$. When B is a finite set we denote by $|B|$ the cardinality of the set (i.e. the number of elements in B) and by $S(B)$ the $(|B|-1)$-dimensional real simplex, i.e. $S(B) := \{q : q \in \mathbb{R}^{|B|},\ \sum_{i=1}^{|B|} q_i = 1,\ 0 \le q_i \le 1 \quad \text{for all } i\}$. For vectors D and V in $\mathbb{R}^K$, the notation $D < V$ stands for $D_k < V_k\ ,\ k = 1, 2, \ldots, K.$. For two matrices ζ, P of appropriate dimensions the notation $\zeta \cdot P$ stands for summation over common indices.

2. PROBLEM FORMULATION

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

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

The constrained optimization problem is:

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

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

The *adaptive* constrained optimization problem **ACOP** is defined as follows. The values of the transition probabilities P_{xay} are unknown, except that assumption **A1** is known to hold. The objective is still to find an optimal policy for **COP**, but based on the available information, assuming that no a-priori information about the values of the $\{P_{xay}\}$ is available. In choosing the control to be used at time t, the only available information are the observed values $\{H_{t-1}, X_t\}$. The paradigm (i)–(iii) in adaptive control is then to use the observations to obtain information about the values of $\{P_{xay}\}$, leading to computation of an adaptive policy. The notation P_{xay} here stands for the *true but unknown* values of the transition probabilities, and similarly for the notation P, E and π^g.

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. We shall use the following cost functions from $\mathbf{X} \times U$ to $\mathbb{R}$:

The *expected average* costs:

$$C_{ea}(x,u) := \overline{\lim_{t\to\infty}} \frac{1}{t+1} E^u \left[\sum_{s=0}^{t} c(X_s, A_s) \mid X_0 = x \right] \tag{2.1a}$$

$$D_{ea}^k(x,u) := \overline{\lim_{t\to\infty}} \frac{1}{t+1} E^u \left[\sum_{s=0}^{t} d^k(X_s, A_s) \mid X_0 = x \right] \qquad k = 1, ..., K \tag{2.1b}$$

Let $0 < \beta < 1$ be a discount factor. The *expected discounted* costs:

$$C_{ed}(x,u) := E^u \left[\sum_{t=0}^{\infty} \beta^t c(X_t, A_t) \mid X_0 = x \right] \tag{2.2a}$$

$$D_{ed}^k(x,u) := E^u \left[\sum_{t=0}^{\infty} \beta^t d^k(X_t, A_t) \mid X_0 = x \right] \qquad k = 1, ..., K \tag{2.2b}$$

As noted above, in the adaptive case, we would like to obtain the same optimal value of $C(x,u)$, despite the lack of information. Clearly, under the discounted cost criteria we cannot hope to obtain the same costs as when the $\{P_{xay}\}$ are known, since the actions taken at each finite time t will generally be sub-optimal. Therefore, the performance in terms of the cost functions C_{ed} and D_{ed} will be degraded with respect to the case where all parameters are known. Moreover, since we work in the framework of non-Bayesian adaptive control (no prior information is given on the unknown parameters), there is no natural way to define optimal adaptive policies with respect to this cost criterion (see discussion in [10] Section II.3). This motivates the following "asymptotic" definitions; consider the *N-stage expected discounted costs*:

$$C_{ed}^{N}(x,u) := E^{u}\left[\sum_{t=N}^{\infty} \beta^{t-N} c(X_t, A_t) \mid X_0 = x\right] \tag{2.3a}$$

$$D_{ed}^{N,k}(x,u) := E^{u}\left[\sum_{t=N}^{\infty} \beta^{t-N} d^k(X_t, A_t) \mid X_0 = x\right] \qquad k = 1, ..., K \tag{2.3b}$$

Define the *asymptotic expected discounted* cost

$$C_{aed}(x,u) := \overline{\lim_{N\to\infty}}\, C_{ed}^{N}(x,u) \tag{2.4}$$

with similar definitions for $D_{aed}^{k}(x,u)$, $\quad k = 1, 2, \ldots, K$.

The constrained optimization problems COP_{ea}, COP_{ed}, COP_{ed}^{N} and COP_{aed} are defined by using the appropriate definitions of the costs in the general definition of COP. We denote by, e.g., $ACOP_{ea}$ the adaptive problem that corresponds to COP_{ea}, with a similar notation for the other relevant cost criteria.

Inspired by Schäl [23], define the asymptotic constrained problem in the sense of Schäl COP_{sd} as follows. Let $g \in U(S)$ be an optimal stationary policy for COP_{ed}; we prove the existence of such a policy in Theorem 4.3. A policy u is feasible for COP_{sd} if

$$\lim_{N\to\infty} \left(D_{ed}^{N,k}(x,u) - E^{u}[D_{ed}^{k}(X_N, g) \mid X_0 = x]\right) \le 0 \ , \qquad k = 1, ..., K \tag{2.5a}$$

A policy u is optimal for COP_{sd} if it is feasible and

$$\lim_{N\to\infty} \left|C_{ed}^{N}(x,u) - E^{u}[C_{ed}(X_N, g) \mid X_0 = x]\right| = 0 \tag{2.5b}$$

This concept was first used by Schäl [23] for the unconstrained adaptive problem. The version in (2.5) is the one used by Hernandez-Lerma, Marcus et al: see [10] and references therein. The reason we prefer to use the version of [10] is that for this case there exists an optimal stationary policy for the constrained problem whenever there exists a feasible policy (see Theorem 4.9), whereas under the definition of [23] an optimal policy for the constrained problem may not exist (see Appendix).

In the unconstrained problems treated in [10,23], the optimal policy is independent of the initial state, and so $E_x^u[C_{ed}(X_N, g)]$ is the minimal cost, given the initial distribution of X_N. As is shown in Example 5.3, the optimal policy of COP_{ed} depends, in general, on the initial state. Thus, an optimal COP_{sd} policy tries to imitate, in the limit, the behavior of the optimal policy corresponding to the initial state x. Consequently, $E_x^u[C_{ed}(X_N, g)]$ is not necessarily the minimal cost, given the initial distribution of X_N since g may not be optimal for this "initial" distribution. Moreover, it is possible that if u^* is an optimal policy for COP_{sd}, then $D_{ed}^{N,k}(x, u^*) > V_k + \epsilon$ for some x and $\epsilon > 0$ and for *all* N. In COP_{aed}, in the period $0 \leq t \leq N$ we attempt to obtain the "initial" distribution $P^u(X_N \mid X_0 = x)$ which is best for obtaining good performance from $t = N$ onwards. We also require that (in the limit) the constraints are satisfied.

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

$$C_{av} := \overline{\lim_{t\to\infty}} \frac{1}{t+1} \sum_{s=0}^{t} c(X_s, A_s) \tag{2.6a}$$

$$D_{av}^k := \overline{\lim_{t\to\infty}} \frac{1}{t+1} \sum_{s=0}^{t} d^k(X_s, A_s) \ , \qquad k = 1, ..., K \tag{2.6b}$$

A policy u is feasible for COP_{av} if

$$D_{av}^k \leq V_k \quad P_x^u \text{ a.s. }, \quad x \in \mathbf{X} \ , \qquad k = 1, 2, \ldots, K \tag{2.7a}$$

A policy v is optimal for COP_{av} if it is feasible, and if for every feasible policy u and every constant M,

$$P_x^u(C_{av} \leq M) > 0 \quad \text{implies} \quad C_{av} \leq M \quad P_x^v \text{ a.s.} \tag{2.7b}$$

The main results of this paper can now be stated:

Theorem 2.1: *Under A1, if COP_{ea} (resp. COP_{ed}, COP_{ed}^N, COP_{aed}, COP_{sd} or COP_{av}) is feasible, then there exists an optimal stationary policy for COP_{ea} (resp. COP_{ed}, COP_{ed}^N, COP_{aed}, COP_{sd} or COP_{av}).*

Proof: The result for COP_{ea} is well known [8,15]. This and the other results are given in Theorems 3.1, 4.3, 4.3, 4.7, 4.9 and 4.1 respectively. ■

Theorem 2.2: *Under A1, if g in $U(S)$ is optimal for COP_{ea}, COP_{aed}, COP_{sd} or COP_{av} then any $\hat{g}$ in $\mathbf{G}(g)$ is also optimal for the respective problem.*

Proof: This claim is established in Theorem 6.4. ■

The computation of optimal stationary policies is discussed in Section 5, where a new Linear Program for COP_{ed} and COP_{sd} is obtained. We show that the well-known Linear Program which solves COP_{ea} also provides an optimal solution for COP_{av} and COP_{aed}.

In Section 7 we provide a method for modifying an adaptive algorithm so as to obtain strongly consistent estimation of the transition probabilities. The modified policy is an AS policy with the same limits, so that by Theorem 2.2 this modification does not change its optimality properties.

Explicit optimal adaptive policies can be obtained by using the estimation scheme of Section 7, combined with the methods of [3].

3. AVERAGE COSTS; frequencies and ATS policies.

In this section we investigate the average cost criteria. For the expected average cost, the constrained problem has been studied extensively. In particular, the following Theorem is proved in [16], [15] or [2, Theorem 2.9].

Theorem 3.1: *Assume A1, and that COP_{ea} is feasible. Then there exists an optimal stationary policy for COP_{ea}.*

The computation of the optimal stationary policy is presented in Section 5.1.

Remark: In the multichain case an optimal stationary policy may not exist. For that case, Hordijk and Kallenberg [15] provide a method of computing an optimal Markovian policy. Ross and Varadarajan [21] have studied a constrained problem that involves minimization of the expected average cost subject to a single constraint of the sample average type. They obtain an optimal stationary policy for the unichain case and ϵ-optimal stationary policies for the multi-chain case.

Since adaptive policies cannot usually be stationary, a larger class of policies is needed for $ACOP_{ea}$. It will be convenient to mimic, in some asymptotic sense, the behavior of stationary policies. The class of "Action Time Sharing" (ATS) policies was introduced in ([1,2]) and was used to solve $ACOP_{ea}$ in [3]. We show in Section 3.3 that this approach also solves $ACOP_{av}$.

The development of ATS policies proceeds through the following steps. Following Derman [8 p. 89] we show that the *expected state action frequencies* (see definitions below) determine the expected average cost. Hordijk and Kallenberg [16, 15] used the expected frequencies in solving COP_{ea}. Then we establish (Lemma 3.2) the measurability of the *state-action frequencies* ([8 Chapter 7]) and show below that they determine the sample average cost.

Altman and Shwartz [2] extended the analysis of the expected average case to a countable state space, and introduced a more basic quantity, the *conditional frequencies*, which determine the frequencies, and hence the sample average and the expected average costs (see also [1] and [3]). ATS policies are defined through the conditional frequencies. The analysis in Section 6 of the Asymptotically Stationary (AS) policies which are needed for other cost criteria, also relies on these results.

3.1. State-Action Frequencies.

The *state action (sa) frequency* f_{sa}^T is a random vector $f_{sa}^T : \Omega \to S(\mathbf{X} \times \mathbf{A})$ 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. We define similarly the *state (s) frequency* f_s^T as the frequency at which the event of being at state y occurs till time T. It is a random vector $f_s^T : \Omega \to S(\mathbf{X})$ defined by $f_s^T(y) := \frac{1}{T+1}\sum_{r=0}^T 1\{X_r = y\}$.

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

Similarly, the multifunction F_{sa} is a mapping from Ω whose values are subsets of $S(\mathbf{X} \times \mathbf{A})$ and is given by the (random) set of accumulation points of f_{sa}^T as $T \to \infty$. The multifunction F_s is a mapping from Ω whose values are subsets of $S(\mathbf{X})$ and is given by the set of accumulation points of the vectors f_s^T as $T \to \infty$. The multifunction (multivalued random vector) F_{sa} is said to be *Borel-measurable* ([22, Section 9] or [10, Appendix D]) if the set

$$F_{sa}^{-1}[B] := \{\omega : F_{sa}(\omega) \cap B \neq \phi\} \in \mathbf{B}$$

for every closed subset B in the standard Borel σ-field on $S(\mathbf{X} \times \mathbf{A})$. The measurability of F_s is similarly defined in terms of Borel subsets of $S(\mathbf{X})$. The measurability of F_{sa} and F_s is discussed in Lemma 3.2 below. Since the sets $\bar{F}_{sa}(x,u)$ $\bar{F}_s(x,u)$ ($F_{sa}(\omega)$ and $F_s(\omega)$) are all sets of accumulation points and are all bounded, they are all compact sets (for each ω).

Define for any given set of policies U' the set of achievable expected state action frequencies $L_x(U') := \cup_{u\in U'}\bar{F}_{sa}(x,u)$. In particular, the set of all achievable expected frequencies is denoted by $L_x := \cup_{u\in U}\bar{F}_{sa}(x,u)$.

The following Lemmas 3.2-3.3 establish some basic properties of the state action frequencies and relate them to the cost achieved by the policies.

Lemma 3.2: *Under A1;*

(i) The class of stationary policies is complete, i.e. $L_x(U(S)) = L_x$ *for every initial state* x*. Moreover,* $L := L_x$ *is independent of* x*.*

(ii) F_{sa} *and* F_s *are Borel measurable.*

(iii) Under any policy u*,* $F_{sa} \subset L(U(S))$ P_x^u *a.s.*

Proof: The first claim is proved in Derman [8, pp. 95] (a generalization to the countable state and action spaces is obtained in [2, Theorem 3.1]). We prove (ii) for F_s, as the proof for F_{sa} is the same. Pick any closed subset B in the standard Borel σ-field on $S(\mathbf{X})$. Note that

$$F_s(\omega) = \cap_{k=1}^{\infty}\ cl\left\{\cup_{t=k}^{\infty}\{f_s^t(\omega)\}\right\}$$

(see e.g. [7]). Hence

$$F_s^{-1}(B) = \{\omega : \cap_{k=1}^{\infty} = \cap_{k=1}^{\infty}\{\omega :\ cl\left\{\cup_{t=k}^{\infty}\{f_s^t(\omega)\}\right\} \cap B \neq \phi\}$$

Let $\{B_n\}_1^\infty$ be a sequence of open sets containing B, with $\cap_{n=1}^{\infty} B_n = B$. Since B is closed,

$$F_s^{-1}(B) = \cap_{k=1}^{\infty}\{\omega : \cap_{n=1}^{\infty}\left(\cup_{t=k}^{\infty} f_s^t(\omega)\cap B_n\right) \neq \phi\} = \cap_{k=1}^{\infty}\cap_{n=1}^{\infty}\cup_{t=k}^{\infty}\{\omega : \{f_s^t(\omega)\}\cap B_n \neq \phi\}\ .$$

Since f_s^t are clearly (measurable) random variables, it follows that $\{\omega : f_s^t(\omega) \in B_n\} \in \mathbf{B}$ and hence $F_s^{-1}(B) \in \mathbf{B}$, which establishes (ii). In view of (ii), (iii) follows from Derman [8, pp. 98]. ■

3.2. Costs and frequencies.

We show below that the state-action frequencies (expected frequencies) determine the cost C_{av} (respectively C_{ea}). This result obviously applies to D_{av} and D_{ea}.

Lemma 3.3: *Under A1, for every policy* $u \in U$ *and any instantaneous cost function* c*,*

(i) there exists some $\bar\zeta \in \bar F_{sa}(x,u)$ *for which the cost* $C_{ea}(x,u)$ *obeys*

$$C_{ea}(x,u) = \sum_y \sum_a c(y,a)\bar\zeta(y,a) \tag{3.1}$$

(ii) there exists some random variable ζ^* *with* $\zeta^*(\omega) \in F_{sa}(\omega)$ *for which the cost* C_{av} *obeys*

$$C_{av} = \sum_y \sum_a c(y,a)\zeta^*(y,a). \tag{3.2}$$

Remark: The main issue in (ii) is to show that ζ^* can be chosen so that it is a *measurable* mapping from $\{\Omega, \mathbf{B}\}$ to $S(\mathbf{X} \times \mathbf{A})$.

Proof: The first claim is in Derman [8, pp. 89]. To prove (ii), define the function $v : S(\mathbf{X}\times\mathbf{A}) \to \mathbb{R}$ by $v(\zeta) := c \cdot \zeta := \sum_{x,a} c(x,a)\zeta(x,a)$. Then

$$C_{av} = \overline{\lim_{t\to\infty}} \frac{1}{t+1} \sum_{s=0}^{t} c(X_s, A_s) = \overline{\lim_{t\to\infty}}\, c \cdot f_{sa}^t$$

Since for each ω, any accumulation point of $c \cdot f_{sa}^t$ can be expressed as $c \cdot \zeta$ with $\zeta(\omega) \in F_{sa}(\omega)$, we obtain

$$C_{av} = \sup_{\zeta\in F_{sa}} c \cdot \zeta = \max_{\zeta\in F_{sa}} v(\zeta)$$

where the second equality follows from the compactness (for each ω) of the range of F_{sa}. Since v is continuous and since F_{sa} is Borel measurable with compact range, it follows from a standard *Measurable Selection Theorem* (see [10, Appendix D] or [22, Section 9]) that there exists a random variable (called a *Selector*) ζ^* with $\zeta^*(\omega) \in F_{sa}(\omega)$ such that $v(\zeta^*) = c \cdot \zeta^* = \max_{\zeta\in F_{sa}} v(\zeta)$. ■

3.3. Conditional frequencies and ATS policies.

The *conditional frequency* f_c^t is a collection $\{f_c^t(a|y),\ \mathit{all}\ a, y\}$, so that the components of the random vector $f_c^t : \Omega \to S(\mathbf{A})^J$ (with $J = |\mathbf{X}|$) are given by

$$f_c^t(a|y) := \frac{\sum_{s=0}^t 1\{X_s = y, A_s = a\}}{\sum_{s=0}^t 1\{X_s = y\}} \tag{3.3}$$

if $\sum_{s=0}^t \{X_s = y\} = 0$ set $f_c^t(a|y) := \frac{1}{|\mathbf{A}|}$. This quantity represents the frequency that action a is used conditioned on state y being visited.

The set of limit points of f_c^t as $t \to \infty$ is denotes by F_c. If only one limit exists then it is denoted by f_c. Lemma 3.5 below states that the conditional frequencies determine the state-action frequencies and the expected frequencies, hence they determine the cost. The conditional frequencies are the key quantities for the controller, since they can be directly steered to the desired values. For instance, if we need to obtain $f(a|y) = 0.2$ for some a, y then one possibility for achieving this is to use action a every fifth visit to state y. That this is possible, is a consequence of the following Lemma, whose proof is available in [1, Lemma 4.1] or [2, Corollary 5.3]. Note that there is no such direct way to control the (expected) state-action frequencies.

Lemma 3.4: *Under A1, for any policy u, each state $y \in \mathbf{X}$ is visited infinitely often P_x^u a.s.*

A policy u is called an ATS policy corresponding to some stationary policy g if F_c is a singleton, with $f_c(a|y) = p^g_{a|y} \quad P_x^u$ a.s.

Lemma 3.5: *Assume A1, and fix some stationary policy g. If under a policy u, $f_c(a|y) = p^g_{a|y} \quad P_x^u$ a.s. Then*

(i) F_{sa} and $\bar{F}_{sa}$ are singletons, and for all a, y, $f_{sa}(y,a) = \bar{f}_{sa}(x,u;y,a) = p^g_{a|y}\pi^g_y \quad P_x^u$ a.s.

(ii) For any initial state x, $C_{ea}(x,g) = C_{ea}(x,u) = C_{av} \quad P_x^u$ a.s. and $C_{ea}(x,g) = C_{av} \quad P_x^g$ a.s.

Proof: The first claim is proved in [2, eq. (4.5)] (see also [5, p. 969]). The second claim then follows from Lemma 3.3. ■

From the definition it is clear that these conditional frequencies are not sensitive to the use of "bad" (non-optimal) controls at some initial finite time interval, nor are they affected by the use of non-optimal controls provided that the frequency at which they are used decreases to zero. The cost functions C_{ea} and C_{av} also have these properties due to their definition as time-averages. Thus it is to be expected that only Cesaro properties of the control should influence the costs. This property makes ATS policies attractive for $ACOP_{ea}$, since we do not need to know the optimal stationary policy g at any finite time; a (strongly) consistent estimator of $p^g_{a|y}$ suffices for optimality. On the other hand, ATS policies facilitate estimation through probing, i.e. by testing non-optimal actions without affecting the cost. Such policies were indeed used in [3] to solve $ACOP_{ea}$.

It is shown in Theorem 4.1 that the stationary policies are optimal for COP_{av}. Moreover, the stationary policy g which is optimal for COP_{ea} is also optimal for COP_{av}. It then follows from Lemma 3.5 (ii) that the ATS policy that solves $ACOP_{ea}$ also solves $ACOP_{av}$. Therefore the method described in [3] to solve $ACOP_{ea}$ is optimal for $ACOP_{av}$ as well.

ATS policies are not adequate for the other cost criteria introduced in Section 2, since the conclusion (ii) of Lemma 3.5 does not hold for cost criteria (even asymptotic) which are based on discounting rather than time average. We therefore have to develop other policies in order to obtain an adaptive method that can be applied to all ACOP. In the next section we show that stationary policies are optimal for all cost criteria under consideration. We then develop, in Section 6, the Asymptotically Stationary policies.

4. STATIONARY POLICIES: OPTIMALITY

The optimality of stationary policies for COP_{ea} is discussed in the beginning of Section 3 (see Theorem 3.1). In this section we prove the optimality of stationary policies for the other cost criteria under consideration.

4.1. The sample average cost

Theorem 4.1: *Assume A1. Then*

(i) COP_{av} is feasible if and only if COP_{ea} is feasible.

(ii) The stationary policies are optimal for COP_{av}.

(iii) A stationary policy is optimal for COP_{ea} if and only if it is optimal for COP_{av}.

Proof: Since under a stationary policy g, $D_{av} = D_{ea}(g)$ P_x^g a.s. and similarly for C, (iii) follows from (i) and (ii).

If COP_{ea} is feasible, then there exists a feasible stationary policy. But under a stationary policy g, $D_{av} = D_{ea}(g)$ P_x^g a.s., so COP_{av} is feasible.

Now assume COP_{av} is feasible. By Fatou's Lemma, since clearly $\frac{1}{t+1}\sum_{s=0}^{t} d^k(X_s, A_s)$ is bounded,

$$\overline{\lim_{t\to\infty}}\, E_x^u \frac{1}{t+1}\sum_{s=0}^{t} d^k(X_s, A_s) \le E_x^u\, \overline{\lim_{t\to\infty}}\, \frac{1}{t+1}\sum_{s=0}^{t} d^k(X_s, A_s) = E_x^u D_{av}^k \le V_k$$

and (i) follows.

Let v be a stationary optimal policy for COP_{ea} and fix some policy u and initial state x. Let $Y := \{\omega : C_{av} < C_{ea}(x,v),\ D_{av} \le V\}$. Fix a selector $\xi \in F_{sa}$ that satisfies $C_{av} = \xi \cdot c$. Fix an arbitrary $\omega_0 \in Y$ and denote $\zeta := \xi(\omega_0)$. Let t_n be an increasing subsequence of times, such that $\zeta := \lim_{n\to\infty} f_{sa}^{t_n}(\omega_0)$.

Claim: $\zeta \notin L$. To establish the claim by contradiction, assume $\zeta \in L$. Then there exists a stationary policy g such that $\bar{F}_{sa}(x,g) = \{\zeta\}$. By definition of Y and ζ it follows that

$$\zeta \cdot c = C_{ea}(x,g) < C_{ea}(x,v)$$

$$D_{ea}^k(x,g) = \zeta \cdot d^k \le \overline{\lim_{t\to\infty}}\, f_{sa}^t(\omega_0) \cdot d^k \le V_k$$

This contradicts the optimality of v, which establishes the claim.

From this it follows that $P_x^u(F_{sa} \not\subset L_x(S)) \geq P_x^u(Y)$. However by Lemma 3.2 (see Derman [8 p. 98]), for any policy u, $P_x^u(F_{sa} \not\subset L_x(S)) = 0$, so that necessarily $P_x^u(Y) = 0$. Since u is arbitrary and $C_{ea}(x,v) = C_{av}$ and $D_{ea}(x,v) = D_{av}$ P_x^v *a.s.*, (ii) holds. ∎

4.2. The expected discounted cost

We begin by defining some basic quantities that play an important role in the expected discounted cost. Define the matrix $\{\bar{\phi}_{sa}(x,u;y,a)\}_{y,a}$ by

$$\bar{\phi}_{sa}(x,u;y,a) := \sum_{t=0}^{\infty} \beta^t P^u(X_t = y, A_t = a | X_0 = x) \tag{4.1}$$

These quantities determine the cost in the following way: for each instantaneous cost $c(y,a)$, $y \in \mathbf{X}, a \in \mathbf{A}$, the overall cost has the representation (see [6]):

$$C_{ed}(x,u) = \sum_{y \in \mathbf{X}, a \in \mathbf{A}} c(y,a)\bar{\phi}_{sa}(x,u;y,a) \tag{4.2}$$

and similarly for D_{ed}.

This representation was suggested and investigated by Borkar [6] who developed a similar representation also for finite time cost criteria and for the exit problem. In his paper Borkar considers a countable state space and a compact action space. He calls the quantity $\bar{\phi}_{sa}$ "occupation measure". Borkar's approach is somewhat different, in that he defines the occupation measure through the cost representation and not explicitly through (4.1).

Let L_x^β denote the set of matrices $\{\bar{\phi}_{sa}(x,u;y,a)\}_{y,a}$ achieved by all policies in U, and $L_x^\beta(S)$ the set of matrices $\{\bar{\phi}_{sa}(x,u;y,a)\}_{y,a}$ achieved by all policies in $U(S)$.

Lemma 4.2: *Assume that under any policy in $U(SD)$, the state space includes a single recurrent class, where the structure is independent of the policy. Then $L_x^\beta = L_x^\beta(S)$, and is closed and convex.*

Proof: See Borkar [6]. ∎

Theorem 4.3: *Under A1 the stationary policies are optimal for COP_{ed} and COP_{ed}^N.*

Proof: The claim for COP_{ed} follows immediately from (4.2) and Lemma 4.2; in fact, this holds even if we relax Assumption A1 to allow some fixed transient states. In order to prove the second claim consider the process $\hat{X}_t$ defined on the enlarged state space $\{\mathbf{X} \times \{1,2,\ldots,N\}\ \}$ and the same action space $\mathbf{A}$. Let the transition probability be given by: $\hat{P}_{\{x,j\}a\{y,l\}} = P_{xay}$ for $l = j+1$ or $j = l = N$, and 0 otherwise. Let $\hat{c}(\{x,j\},a) = 0$ for $j \neq N$ and $\hat{c}(\{x,N\},a) = \beta^{-N} c(x,a)$ with the analogue definition for $\hat{d}^k$. Then clearly for any policy u and initial state x, $C^N_{ed}(x,u) = \hat{C}_{ed}(\{x,1\},u)$ and $D^{N,k}_{ed}(x,u) = \hat{D}^k_{ed}(\{x,1\},u)$. This process satisfies the hypotheses of Lemma 4.2, so that the proof for COP^N_{ed} follows immediately from the proof for COP_{ed}. ■

4.3 The asymptotic expected discounted cost

Theorem 4.4: *Under A1, for any stationary policy g, $C_{aed}(x,g) = [1-\beta]^{-1} C_{ea}(x,g)$, and $D^k_{aed}(x,g) = [1-\beta]^{-1} D^k_{ea}(x,g)$, $k = 1,2,\ldots,K$.*

Proof: For any policy u, note that $C^N_{ed}(x,u)$ has the representation:

$$\begin{aligned} C^N_{ed}(x,u) &= E^u[\sum_{t=N}^{\infty} \beta^{t-N} c(X_t,A_t) | X_0 = x] \\ &= \sum_{t=N}^{\infty} \beta^{t-N} \sum_{y,a} c(y,a) P^u(X_t = y, A_t = a | X_0 = x) \end{aligned} \tag{4.3}$$

Since for any stationary policy the limit $\eta^g(y,a) := \lim_{t\to\infty} P^g_x(X_t = y, A_t = a)$ exists,

$$\lim_{t\to\infty} \sum_{y,a} c(y,a) P^g(X_t = y, A_t = a | X_0 = x) = \sum_{y,a} c(y,a) \pi^g(y) p^g_{a|y} = C_{ea}(g,x) \tag{4.4}$$

(see Lemma 3.5). It then follows from (4.3) that

$$C_{aed}(x,g) := \lim_{N\to\infty} C^N_{ed}(x,g) = [1-\beta]^{-1} \sum_{y,a} c(y,a) \pi^g(y) p^g_{a|y} = [1-\beta]^{-1} C_{ea}(x,g) \tag{4.5}$$

The derivation for D^k_{aed} is identical. ■

In order to establish the existence of an optimal stationary policy for COP_{aed} we need the following Lemmas. For any real valued vector $z \in \mathbb{R}^{|\mathbf{X}|}$, let $\|z\|_1 := \sum_{y\in\mathbf{X}} |z(y)|$. Let $\zeta(\cdot)$ be any distribution on $\mathbf{X}$, and recall the convention $(\zeta \cdot P)(y) = \sum_x \zeta(x) P_{xy}$.

Lemma 4.5: *Let $P = \{P_{xy}\}$ be the transitions of a regular Markov chain on $\mathbf{X}$, with invariant measure π. If $P_{xy} \geq \epsilon' > 0$ for all $x,y \in \mathbf{X}$, then $\|\zeta \cdot P - \pi\|_1 \leq (1-\epsilon')\|\zeta - \pi\|_1$.*

Proof:

$$\|\zeta \cdot P - \pi\|_1 = \|(\zeta - \pi) \cdot P\|_1 = \sum_y |\sum_x [\zeta(x) - \pi(x)] P_{xy}| \tag{4.6}$$

$$\|\zeta - \pi\|_1 = \sum_x |(\zeta(x) - \pi(x)| = \sum_x |\zeta(x) - \pi(x)| \sum_y P_{xy} = \sum_x \sum_y |\zeta(x) - \pi(x)| P_{xy} \tag{4.7}$$

Since $\sum_x [\zeta(x) - \pi(x)] = 0$, we have $\sum_x [\zeta(x) - \pi(x)]^+ = \sum_x [\zeta(x) - \pi(x)]^-$ and we obtain after some algebra

$$\|\zeta - \pi\|_1 - \|\zeta \cdot P - \pi\|_1 = \sum_y \sum_x 2[\zeta(x) - \pi(x)]^- \cdot P_{xy} \geq 2\epsilon' \sum_x [\zeta(x) - \pi(x)]^- = \epsilon' \|\zeta - \pi\|_1 \tag{4.8}$$

The Lemma now follows from (4.8). ∎

Lemma 4.6: *Under A1, there exists a single constant $\alpha < 1$, independent of x and of $g \in U(S)$, and an integer M such that $\|P^g(X_t = \cdot | X_0 = x) - \pi^g\|_1 \leq \alpha^{t-M}$ for all $t > M$.*

Proof: A1 implies that under any stationary deterministic g, $\{X_t\}$ is a regular Markov chain. Hence (see [17]) for some integer $M(g)$, $[P^g]^{M(g)}$ has all components nonzero. Furthermore, since $U(SD)$ is a finite set, then there is an integer M such that $[P^g]^M$ has all components strictly positive for any $g \in U(SD)$, and is a transition matrix for a regular Markov chain with invariant measure π^g. Since $U(SD)$ is a finite set, $\epsilon'' := \min\{\ \{[P^g]^M\}_{yz} : g \in U(SD),\ y, z \in \mathbf{X}\} > 0$. Now enumerate the stationary deterministic policies as $\{g_1, \ldots, g_{|U(SD)|}\}$. Fix g in $U(S)$; then $P^g = \sum_{i=1}^{|U(SD)|} q_i P^{g_i}$ where $q_i \geq 0$ and $\sum_i q_i = 1$. Therefore

$$\{[P^g]^M\}_{yz} \geq \sum_{i=1}^{|U(SD)|} q_i^M \{[P^{g_i}]^M\}_{yz} \geq |U(SD)|^{-M} \epsilon'' := \epsilon'$$

By Lemma 4.5 applied to the transition matrix $[P^g]^M$ it follows that

$$\|\zeta \cdot [P^g]^{(n+1)M} - \pi^g\|_1 \leq (1 - \epsilon') \|\zeta \cdot [P^g]^{nM} - \pi^g\|_1$$

for any initial distribution $\zeta(\cdot)$ and any $g \in U(S)$. Hence

$$\|\zeta \cdot [P^g]^{nM} - \pi^g\|_1 \leq (1 - \epsilon')^n$$

uniformly in $g \in U(S)$ and in ζ. Thus for $t = nM + l$ where $0 \leq l < M$ we have

$$\|\zeta \cdot [P^g]^t - \pi^g\|_1 = \|\ (\zeta \cdot [P^g]^l)[P^g]^{nM} - \pi^g\|_1 \leq (1 - \epsilon')^n \leq (1 - \epsilon')^{(t/M)-1}$$

and the result follows. ■

Theorem 4.7: *Under A1, if COP_{aed} is feasible then there exists an optimal stationary policy.*

Proof: Let g_n be an optimal stationary policy for COP^n_{ed} and let u be any (not necessarily stationary) feasible policy for COP_{aed}, i.e. $D_{aed}(x,u) \le V$. By definition,

$$C_{aed}(x,u) = \overline{\lim_{N\to\infty}}\, C^N_{ed}(x,u)$$

Thus there exists some increasing sequence of integers n_l such that

$$C_{aed}(x,u) = \lim_{l\to\infty} C^{n_l}_{ed}(x,u) \tag{4.9}$$

Since A is finite, there is some subsequence n_m of n_l such that for all y, $g_{n_m}(\cdot;y)$ converges to some distribution $g(\cdot;y)$. By Theorem 4.4 we have:

$$C_{aed}(x,g) = [1-\beta]^{-1}C_{ea}(x,g) =$$

$$\lim_{m\to\infty}[1-\beta]^{-1}C_{ea}(x,g_{n_m}) = \lim_{m\to\infty} C_{aed}(x,g_{n_m}) \tag{4.11}$$

where the second equality follows from the representation (4.4) since $\pi^{g'}$ is continuous in g' (see e.g. [14]). Now by (4.3),

$$\mid C_{aed}(x,g_{n_m}) - C^{n_m}_{ed}(x,g_{n_m}) \mid \le \sup_{v\in U(S)} \mid C_{aed}(x,v) - C^{n_m}_{ed}(x,v) \mid$$

$$\le \sup_{v\in U(S)} \sum_{t=n_m}^{\infty} \beta^{t-n_m} \sum_{y,a} |c(y,a)| \cdot p^v_{a|y} \mid P^v_x(X_t = y) - \pi^v(y) \mid$$

However, by Lemma 4.6 $\mid P^v_x(X_t=y) - \pi^v(y) \mid \to 0$ as $t\to\infty$, *uniformly in* v. Therefore

$$C_{aed}(x,g) = [1-\beta]^{-1}C_{ea}(x,g) = \lim_{m\to\infty} C^{n_m}_{ed}(x,g_{n_m}) \tag{4.12}$$

Similarly for D_{aed} we obtain:

$$D_{aed}(x,g) = \lim_{m\to\infty} D^{n_m}_{ed}(x,g_{n_m}) \tag{4.13}$$

On the other hand we have from (4.9)

$$C_{aed}(x,u) = \lim_{m\to\infty} C_{ed}^{n_m}(x,u) \tag{4.14}$$

Clearly, by definition of D_{aed},

$$D_{aed}(x,u) \geq \overline{\lim_{m\to\infty}} D_{ed}^{n_m}(x,u) \tag{4.15}$$

Since $C_{ed}^{n_m}(x, g_{n_m}) \leq C_{ed}^{n_m}(x,u)$ we obtain from (4.12) and (4.14) that $C_{aed}(x,g) \leq C_{aed}(x,u)$. Similarly, (4.13) and (4.15) imply $D_{aed}(x,g) \leq D_{aed}(x,u)$. Thus, for every policy there is a stationary policy with better performance. To conclude note that the continuity of $g \to C_{aed}(g)$ established in (4.11) together with the compactness of the space of stationary policies imply the existence of an optimal stationary policy. ■

The connection between optimal policies for COP_{ea}, COP_{av} and COP_{aed} is now summarized.

Lemma 4.8: *Consider problems COP_{ea} and COP_{av} with constraints given by V, and problem COP_{aed} with constraints given by $[1-\beta]^{-1} \cdot V$. Under A1, if a stationary policy v is optimal for one of the problems, then it is optimal for all of them.*

Proof: Recall that there exist stationary optimal policies for COP_{ea}, COP_{av} and COP_{aed} (Theorem 3.1, 4.1 and 4.7 resp.). The relation between COP_{ea} and COP_{av} is established in Theorem 4.1. The relation between COP_{aed} and COP_{ea} is immediate from Theorem 4.4. ■

4.4 Schäl optimality

Theorem 4.9: *Assume A1. If COP_{sd} is feasible, there exists an optimal stationary policy for COP_{sd}. If COP_{ed} is feasible and g is an optimal stationary policy for COP_{ed}, than g is optimal for COP_{sd}.*

Proof: If COP_{sd} is feasible, then by definition so is COP_{ed}. Thus the first claim follows from the second. If g is optimal for COP_{ed} then by definition it is optimal for COP_{sd}. ■

Note that while there is a stationary policy which is optimal for COP_{ea}, COP_{av} and COP_{aed} uniformly in the initial conditions, this is not the case for COP_{sd}. As is shown in Example 5.3, for each initial condition the optimal stationary policy may be different.

5. COMPUTATION OF OPTIMAL POLICIES

In this Section we show that optimal policies for COP_{ea}, COP_{av}, COP_{aed} and for COP_{ed}, COP_{sd} (which may be found in $U(S)$, according to the results of Section 4) can be obtained by solving the appropriate Linear Programs. A similar solution for COP_{ed}^N is obtained through the embedding described in the proof of Theorem 4.3.

5.1 Optimal policies for $\mathbf{COP_{ea}}$, $\mathbf{COP_{av}}$, and $\mathbf{COP_{aed}}$

Since by Lemma 4.8, if a stationary policy v is optimal for COP_{ea} or COP_{av} or COP_{aed} then it is optimal for all of them, it suffices to describe the LP that yield an optimal (stationary) policy for COP_{ea}. This LP is well known [8]; it was recently extended to the multi-chain case in [16] and [15]. An extension to the countable state and action case was introduced in [2].

LP1: Find $\{z^*(y,a)\}_{y,a}$ that minimizes $c \cdot z := \sum_{y,a} c(y,a)z(y,a)$ subject to:

$$\sum_{y,a} z(y,a)\left[P_{yav} - \delta_v(y)\right] = 0 \qquad v \in \mathbf{X} \tag{5.1a}$$

$$\sum_{y,a} d^k(y,a)z(y,a) \le V_k \qquad 1 \le k \le K \tag{5.1b}$$

$$\sum_{y,a} z(y,a) = 1 \qquad z(y,a) \ge 0 \tag{5.1c}$$

Theorem 5.1: *Under A1,*

(i) If the stationary policy g is feasible for COP_{ea}, then a feasible solution z to (5.1) is given by

$$z(y,a) = \pi_y^g \cdot p_{a|y}^g \tag{5.2}$$

(ii) If g is an optimal stationary policy for COP_{ea} then (5.2) defines an optimal solution for LP1.

(iii) Conversely, let $z(y,a)$ satisfy (5.1). Then a feasible policy g for COP_{ea} is defined through

$$p_{a|y}^g = \frac{z(y,a)}{\sum_{a' \in \mathbf{A}} z(y,a')} \tag{5.3}$$

(iv) If z is an optimal solution of LP1, then the stationary policy g defined by (5.3) is optimal for COP_{ea}.

Proof: See [16]. ■

5.2 Optimal policies for $\mathbf{COP_{ed}}$, $\mathbf{COP^N_{ed}}$ and $\mathbf{COP_{sd}}$

We find below optimal stationary policies for COP_{ed} through Linear Program. The same technique can be used to solve COP^N_{ed} by using the embedding described in the proof of Theorem 4.3. By Theorem 4.9, any stationary policy which is optimal for COP_{ed} is also optimal for COP_{sd}.

Define the following LP:

LP2: Find $\{z^*(y,a)\}_{y,a}$ that minimizes $C(z) := \sum_{y,a} c(y,a)z(y,a)$ subject to:

$$\sum_{y,a} z(y,a)\left[\delta_v(y) - \beta P_{yav}\right] = \delta_x(v) \qquad v \in \mathbf{X} \tag{5.4a}$$

$$\sum_{y,a} d^k(y,a)z(y,a) \le V_k \qquad 1 \le k \le K \tag{5.4b}$$

$$z(y,a) \ge 0 \tag{5.4c}$$

Note that due to (5.4a), each initial condition x leads to a distinct Linear Program.

Theorem 5.2: *Assume A1.*

(i) If the stationary policy w is feasible for COP_{ed}, then the matrix $\bar{\phi}_{sa}(x,w)$, defined in (4.1) satisfies (5.4). For any stationary policy w, $c \cdot \bar{\phi}_{sa}(x,w) = C_{ed}(x,w)$.

(ii) If g is an optimal stationary policy for COP_{ed} then there exists an optimal solution for LP2 satisfying

$$z^*(y,a) = \bar{\phi}_{sa}(x,g;y,a) \tag{5.5a}$$

(iii) Conversely, let $z(y,a)$ satisfy (5.4). Then the policy w given by

$$p^w_{a|y} = \frac{z(y,a)}{\sum_{a' \in \mathbf{A}} z(y,a')} \tag{5.5b}$$

is feasible for COP_{ed}, and $C_{ed}(x,w) = C(z)$.

(iv) If z^* solves LP2, then the stationary policy g defined by

$$p^g_{a|y} = \frac{z^*(y,a)}{\sum_{a' \in \mathbf{A}} z^*(y,a')} \tag{5.5c}$$

is optimal for COP_{ed}.

Proof: The Lemma is a generalization of [8, p. 42] where the non-constrained expected discounted case is considered. To prove (i) assume that the stationary policy w is feasible for COP_{ed}. Then (5.4c) is clearly satisfied, and (5.4b) is satisfied by (4.2). (5.4a) is satisfied since

$$\sum_{y,a} \bar{\phi}_{sa}(x,w;y,a)\delta_v(y) = \sum_a \bar{\phi}_{sa}(x,w;v,a) =$$

$$= \delta_x(v) + \sum_{t=1}^{\infty} \beta^t P^w(X_t = v | X_0 = x)$$

Since

$$P^w(X_t = v \mid X_0 = x) = \sum_{a,y} P^w(X_{t-1} = y, A_{t-1} = a \mid X_0 = x) P_{yav} \tag{5.6}$$

we obtain

$$\sum_{y,a} \bar{\phi}_{sa}(x,w;y,a)\delta_v(y) = \delta_x(v) + \beta \sum_{y,a} \sum_{t=0}^{\infty} \beta^t P^w(X_t = y, A_t = a | X_0 = x) P_{yav}$$

$$= \delta_x(v) + \beta \sum_{y,a} \bar{\phi}_{sa}(x,w;y,a) P_{yav}$$

which proves the first claim in (i). The second claim follows from (4.2).

To prove (iii) let $z(y,a)$ satisfy (5.4). Then $\sum_a z(y,a) = \sum_a \bar{\phi}_{sa}(x,w;y,a)$ due to (5.4a) and (5.4c) (see proof in [8, p. 43]), where w is given in the Theorem. Since $P^w(X_t = y, A_t = a \mid X_0 = x) = P^w(X_t = y \mid X_0 = x) \cdot p^w_{a|y}$, we have $\bar{\phi}_{sa}(x,w;y,a) = p^w_{a|y} \sum_a \bar{\phi}_{sa}(x,w;y,a)$, it then follows from (5.5b) that $z(y,a) = \bar{\phi}_{sa}(x,w;y,a)$. Thus by (4.2) it follows that w is feasible for COP_{ed} and $C_{ed}(x,w) = C(z)$, which establishes (iii). Thus $p^w_{a|y} = z(y,a)[\sum_{a \in \mathbf{A}} z(y,a)]^{-1}$ is a one-to-one mapping of the feasible solutions of LP2 onto the stationary policies that are feasible for COP_{ed}. This, together with (4.2) establish (ii) and (iv). ■

Since to the best of our knowledge there are no previous studies on the constrained problem with expected discounted cost, we point out at some of its' properties and some important differences from the non-constrained case.

(i) In the unconstrained case it is known that an optimal stationary deterministic policy can be found. In the constrained case we do not have in general optimal stationary deterministic policies, but do have optimal stationary randomized policies. It can easily be shown (as in the expected average case [20]) that an optimal stationary policy can be computed, that has at most K states in which randomization is needed.

(ii) Unlike the unconstrained case, the optimal stationary policy depends on the initial state (or initial distribution). Moreover, the optimality principle does not hold in the constrained case. The following example with one constraint exhibits these points.

Example 5.3: Consider COP_{ed} with $\mathbf{X} = \{1, 2\}$, $\mathbf{A} = \{a, b\}$, discount factor $\beta = 0.1$,

$$P_{1a1} = P_{1b1} = P_{2a1} = P_{2b1} = 0.1 \ , \qquad P_{1a2} = P_{1b2} = P_{2a2} = P_{2b2} = 0.9$$

$$c(1, a) = c(1, b) = 0, \qquad c(2, a) = 1, \quad c(2, b) = 0$$

$$d(1, a) = d(1, b) = 1 \ , \qquad d(2, a) = 0, \quad d(2, b) = 0.1$$

Note that the transitions do not depend on the control. Let g_1 be the policy which chooses always action a and set $V = D_{ed}(1, g_1)$. Any feasible policy for the problem starting at 1 must always choose a at 2, hence g_1 is optimal for that initial condition. However, if g_2 chooses a at 1 and b at 2, then clearly g_2 achieves the minimal cost. If the initial condition is 2 then

$$D_{ed}(2, g_2) < 0.1 + \sum_{n=1}^{\infty} \beta^n < 1 < V \tag{5.7}$$

so that g_2 is feasible, and hence optimal for the problem starting at 2. Hence the optimal policy depends on the initial state.

Suppose that at any time s, $X_s = z$. Then the optimality principle states that one should use the policy which is optimal for the optimization problem that starts with $X_0 := z$. Clearly in the example this principle does not hold, since when $X_s = 2$ we must not use g_2 if the initial state was $X_0 = 1$.

6. ASYMPTOTICALLY STATIONARY POLICIES

From the definition (1.2) of asymptotically stationary policies, for each $u \in \mathbf{G}(g)$ there exists $\epsilon(t) = \epsilon(t, \omega)$ such that (outside a set in Ω of probability P_x^u zero),

$$\epsilon(t) \downarrow 0 \text{ as } t \to \infty, \quad \text{and} \quad \left| P_x^u(A_t = a | H_{t-1}, X_t = y) - p^g_{a|y} \right| < \epsilon(t) \tag{6.1}$$

for all x, y and a. We show that these *limiting conditional distributions* of a policy determine the cost achieved by asymptotically stationary policies, for all the cost criteria under consideration.

This enables the application of AS policies to the solution of all the relevant adaptive constrained problems.

Given a policy u, denote

$$\eta_x^u(y,a) := \lim_{t\to\infty} P_x^u(X_t = y, A_t = a) \tag{6.2}$$

whenever this limit exists. Note that under A1, for any stationary policy g, η_x^g exists and $\eta_x^g(y,a) = \eta^g(y,a) = \pi^g(y)\cdot g(a|y)$ is independent of x. The main result of this Section is that the conditional distributions which enter the definition (1.2) of an AS policy $\hat{g}$ have a role which is similar to the conditional frequencies (3.3) in the sense that they determine the limiting probabilities (6.2) (Theorem 6.1), which in turn determine the costs (see proof of Theorem 6.4).

Theorem 6.1: *Under A1, given any stationary policy g, if $u \in \mathbf{G}(g)$ then for any state y and initial state x, $\lim_{t\to\infty} P^u(X_t = y|X_0 = x) = \pi^g(y)$, and η_x^u exists and is equal to η^g.*

In order to prove Theorem 6.1 we need the following Lemma. Fix a stationary policy g and a policy u in $\mathbf{G}(g)$. Define the matrix $P(t)$ by

$$\{P(t)\}_{yz} := P^u(X_t = z|X_{t-1} = y, X_0 = x) \tag{6.3}$$

Lemma 6.2: *Assume A1 and fix some arbitrary $\tilde{\epsilon} > 0$, $g \in U(S)$ and $u \in \mathbf{G}(g)$. If $P_x^u(X_{t_n} = y) > \tilde{\epsilon}$ for some increasing sequence $\{t_n\}_1^\infty$, then $\lim_{n\to\infty}\{P(t_n)\}_{yz} = \{P^g\}_{yz}$ for each z.*

Proof of Lemma 6.2: By (6.1), for any $\Delta > 0$ there exists a $T_\Delta > 0$ such that $P_x^u(\epsilon(T_\Delta) > \Delta) < \Delta$. For $t > T_\Delta$,

$$\begin{aligned} |P_x^u(A_t = a|X_t = y) - p^g_{a|y}| \le \Big| E_x^u \Big(& 1\{\epsilon(t) \le \Delta\} \Big[P_x^u(A_t = a \mid H_{t-1}, X_t = y) - p^g_{a|y}\Big] \\ & +1\{\epsilon(t) > \Delta\} \Big[P_x^u(A_t = a \mid H_{t-1}, X_t = y) - p^g_{a|y}\Big] \mid X_t = y\Big)\Big| \\ \le \Delta + E_x^u\big(& 1\{\epsilon(t) > \Delta\} \mid X_t = y\big) \end{aligned} \tag{6.4}$$

For any $t_n > T_\Delta$, $E_x^u\big(1\{\epsilon(t_n) > \Delta\} \mid X_{t_n} = y\big) \le \frac{\Delta}{\tilde{\epsilon}}$. Since Δ is arbitrary, $\lim_{n\to\infty} P^u(A_{t_n} = a|X_{t_n} = y, X_0 = x) = p^g_{a|y}$. The Lemma then follows by noting that $P_{yz}(t) = \sum_a P_{yaz}\cdot P^u(A_{t-1} = a|X_{t-1} = y, X_0 = x)$. ■

Proof of Theorem 6.1: Fix x, a and $u \in \mathbf{G}(g)$. Under g, $\{X_t\}$ is a regular Markov chain. Hence for some integer M (as in the proof of Lemma 4.6), $[P^g]^M$ has all components nonzero (in fact it can easily be shown that one such M is the smallest integer divisible by $2, 3, \ldots, J$).

Let ρ_j be a row vector with 1 in the j-th component and 0 otherwise. Denote $Q_x^t := P^u(X_t | X_0 = x)$. Observe that $Q_x^t = \rho_x \cdot \prod_{s=1}^t P(s)$, where $P(s)$ is defined in (6.3). This follows by iterating

$$\begin{aligned} Q_x^t(y) &= P^u(X_t = y | X_0 = x) \\ &= \sum_z P^u(X_{t-1} = z | X_0 = x) \cdot P^u(X_t = y | X_{t-1} = z, X_0 = x) = \sum_z Q_x^{t-1}(z) P_{zy}^u(t) \end{aligned} \tag{6.5}$$

where the second equality is obtained from Bayes rule. In matrix notation the last equation reads $Q_x^t = Q_x^{t-1} \cdot P^u(t)$. Given some $\tilde{\epsilon} > 0$ let $P^*(t)$ be the matrix whose elements are given by

$$\{P^*(t)\}_{yz} := \{P(t)\}_{yz} + \{P^g - P(t)\}_{yz} \cdot 1\{Q_x^{t-1}(y) \le \tilde{\epsilon}\} \tag{6.6}$$

for all y, z in $\mathbf{X}$. Denote

$$R(t) := \prod_{s=t+1}^{t+M} P(t) \ , \quad R^*(t) := \prod_{s=t+1}^{t+M} P^*(t) \tag{6.7}$$

Next, we evaluate $Q_x^{t+M} - Q_x^t \cdot R^*(t)$.

$$\begin{aligned} \left| \left[Q_x^{t+1} - Q_x^t P^*(t+1) \right] (y) \right| &= |Q_x^{t+1}(y) - \sum_z Q_x^t(z) P_{zy}^*(t+1)| \\ &= |\sum_z Q_x^t(z)[P_{zy}(t+1) - P_{zy}^*(t+1)]| \le \tilde{\epsilon} \cdot M \end{aligned} \tag{6.8}$$

Similarly, for an arbitrary stochastic matrix $\tilde{P}$,

$$\left| \left(\left[Q_x^{t+1} - Q_x^t P^*(t+1) \right] \tilde{P} \right) (y) \right| \le \tilde{\epsilon} \cdot M^2 \tag{6.9}$$

Using (6.7)-(6.9) we obtain

$$\left| \left[Q_x^{t+M} - Q_x^t R^*(t) \right] (y) \right| = \left| \left[Q_x^t \left(R(t) - R^*(t) \right) \right] (y) \right| \tag{6.10a}$$

$$\le \left| \left[Q_x^t P(t+1) \left(\prod_{s=t+2}^{t+M} P(t) - \prod_{s=t+2}^{t+M} P^*(t) \right) \right] (y) \right|$$

$$+ \left| \left[Q_x^t \left(P(t+1) - P^*(t+1) \right) \prod_{s=t+2}^{t+M} P^*(t) \right] (y) \right| \tag{6.10b}$$

$$\le \left| \left[Q_x^{t+1} \left(\prod_{s=t+2}^{t+M} P(t) - \prod_{s=t+2}^{t+M} P^*(t) \right) \right] (y) \right| + \tilde{\epsilon} M^2 \tag{6.10c}$$

The first term in (6.10c) is now handled as in (6.10a)-(6.10c), to obtain

$$\left|\left[Q_x^{t+M} - Q_x^t R^*(t)\right](y)\right| \le \tilde{\epsilon} M^3 \tag{6.11}$$

Denote $\epsilon_1 := \tilde{\epsilon} \cdot M^4$. By Lemma 6.2 and the definition (6.6), $P^*(t)$ converges to P^g. The continuity of the mapping $\{P^*(s)\}_{s=t+1}^{t+M} \to \prod_{s=t+1}^{t+M} P^*(s)$ now implies $R^*(t) \to [P^g]^M \quad P_x^u$ a.s. Thus there exists some t_0 such that for all $t > t_0$, all components of $R^*(t)$ are greater than some positive ϵ', and ϵ' depends only on P^g. Denote $\alpha := 1 - \epsilon'$.

Let π^t denote the stationary probability of a Markov chain whose transition probabilities are given by the matrix $R^*(t)$. Since all components of $R^*(t)$ are strictly positive, π^t exists and is unique for all $t > t_0$. The convergence $R^*(t) \to [P^g]^M$ now implies (see e.g. [14]) $\lim_{t\to\infty} \pi^t = \pi^g$ since π^g is the unique invariant distribution of $[P^g]^M$. Let $t_1 > t_0$ be such that $\|\pi^t - \pi^g\|_1 \le \epsilon_1$ for all $t > t_1$. For $t \ge t_1$:

$$\begin{aligned}
\|Q_x^{t+nM} - \pi^g\|_1 &\le \|Q_x^{t+nM} - \pi^{t+(n-1)M}\|_1 + \|\pi^{t+(n-1)M} - \pi^g\|_1 && (6.12a)\\
&= \|Q_x^{t+(n-1)M} \cdot R(t+(n-1)M) - \pi^{t+(n-1)M}\|_1 + \|\pi^{t+(n-1)M} - \pi^g\|_1 \\
&\le \|Q_x^{t+(n-1)M} \cdot R^*(t+(n-1)M) - \pi^{t+(n-1)M}\|_1 \\
&\quad + \|Q_x^{t+(n-1)M} \cdot (R(t+(n-1)M) - R^*(t+(n-1)M))\|_1 + \|\pi^{t+(n-1)M} - \pi^g\|_1 \\
&\le \alpha \|Q_x^{t+(n-1)M} - \pi^{t+(n-1)M}\|_1 + \epsilon_1 + \epsilon_1 && (6.12b)
\end{aligned}$$

where (6.12b) follows from Lemma 4.5, (6.10) and (6.11). Iterating (6.12) we obtain:

$$\|Q_x^{t+nM} - \pi^g\|_1 \le \alpha^n \|Q_x^t - \pi^g\|_1 + \frac{2\epsilon_1}{1-\alpha} \le 2\alpha^n + \frac{2\epsilon_1}{1-\alpha} \tag{6.13}$$

Since $\tilde{\epsilon}$ and hence ϵ_1 can be chosen arbitrarily small, α is independent of this choice and since α^n converges to 0, it follows that

$$\lim_{t\to\infty, n\to\infty} \|Q_x^{t+nM} - \pi^g\|_1 = 0$$

from which it follows

$$\lim_{t\to\infty} P^u(X_t = \cdot \,|X_0 = x) = \lim_{t\to\infty} Q_x^t = \pi^g \tag{6.14}$$

This proves the first claim.

Now observe that from (6.14), if $\tilde{\epsilon} < \pi^g(y)$ then $P_x^u(X_t = y) < \tilde{\epsilon}$ only a finite number of times. By A1, $\pi^g(y) > 0$ for all y, so Lemma 6.2 and the argument following (6.4) imply that

$$\lim_{t\to\infty} P^u(A_t = a|X_t = y, X_0 = x) = p_{a|y}^g$$

Since $P^u(X_t = y, A_t = a|X_0 = x) = P^u(X_t = y|X_0 = x) \cdot P^u(A_t = a|X_t = y, X_0 = x)$, η^u exists and is equal to η^g. ■

In fact, (6.14) allows to restate Lemma 6.2 in the following stronger form.

Lemma 6.2′: *Assume A1 and fix some $g \in U(S)$ and $u \in \mathbf{G}(g)$. Then $\lim_{t\to\infty} P(t) = P^g$.*

Asymptotically stationary policies turn out to be a special case of ATS policies, as the next Lemma shows. Denote $\mathcal{F}_t := \sigma\{H_{t-1}, X_t\}$, fix y in $\mathbf{X}$ and define

$$\tau(y;n) := \text{ the time of the n}th\text{ visit } \textit{strictly after } 0 \text{ to state } y.$$

Lemma 6.3: *Under A1, if $u \in \mathbf{G}(g)$ then $F_c = \{p^g_{a|y}\}$ P^u_x a.s.*

Proof: Fix a and y and define

$Y_s := 1\{X_s = x, A_s = a\} - E^u_x[1\{X_s = x, A_s = a\} \mid \mathcal{F}_s] \qquad s \geq 0$

$S_t := \sum_{s=0}^t Y_s$

Y_t and S_t are $\mathcal{F}_{t+1}$ measurable. Moreover, $\{S_t, \mathcal{F}_{t+1},\ t \geq 0\}$ is a P^u_x Martingale. Note that the term in expectation in Y_s can be expressed as:

$$E^u_x[1\{A_s = a, X_s = y\} \mid \mathcal{F}_s] = 1\{X_s = y\} \cdot E^u_x[1\{A_s = a\} \mid \mathcal{F}_s] = 1\{X_s = y\} \cdot u_s(a \mid \mathcal{F}_s)$$

$$= 1\{X_s = y\} \cdot u_s(a \mid H_{s-1}, X_s = y) \tag{6.16}$$

Note that $U_t := \sum_{s=0}^t 1\{X_s = y\}$ is nondecreasing, $\mathcal{F}_t$ measurable and, by Lemma 3.4, converges to infinity P^u_x a.s. To apply a version of the Martingale Stability Theorem [9 Theorem 2.18 p. 35], observe that by (6.16),

$$\sum_{s=0}^{\infty} U_s^{-2} E\left(|Y_s|^2 \mid \mathcal{F}_s\right) \leq \sum_{s=0}^{\infty} U_s^{-2} 1\{X_s = y\}. \tag{6.17}$$

With $\tau(y;n)$ as in (6.15), $X_t = y$ if and only if $t = \tau(y;n)$ for some n, and $U_{\tau(y;n)} = n$. Thus (6.17) implies

$$\sum_{s=0}^{\infty} U_s^{-2} E\left(|Y_s|^2 \mid \mathcal{F}_s\right) \leq \sum_{n=0}^{\infty} n^{-2} < \infty \tag{6.18}$$

The Martingale Stability Theorem now implies that

$$\lim_{t\to\infty} \frac{S_t}{\sum_{s=0}^{t} 1\{X_s = y\}} = 0 \qquad P_x^u \ a.s.$$

from which we obtain:

$$\lim_{t\to\infty} \left[f_c^t(a|y) - \frac{\sum_{s=0}^{t} 1\{X_s = y\} \cdot u_s(a \mid H_{s-1}, X_s = y)}{\sum_{s=0}^{t} 1\{X_s = y\}} \right] = 0 \qquad P_x^u \ a.s.$$

Since $\lim_{t\to\infty} u_t(a \mid H_{t-1}, X_t = y) = p_{a|y}^g \quad P_x^u \ a.s.$ and $\sum_{s=0}^{t} 1\{X_s = y\} \to \infty$ it follows that

$$\lim_{t\to\infty} f_c^t(a|y) = p_{a|y}^g \qquad P_x^u \ a.s.$$

which proves that $F_c = \{p_{a|y}^g\}$ P_x^u a.s. ∎

We are now in a position to prove Theorem 2.2. We show that asymptotically stationary policies are equivalent to the respective stationary policies, in that the costs associated with them are identical. Therefore, by Theorem 6.4, whenever $g \in U(S)$ is optimal under one of the cost criteria, any $\hat{g} \in \mathbf{G}(g)$ is optimal as well. In Section 7 we construct optimal AS policies for ACOP.

Theorem 6.4: *Assume A1. Then for any stationary $g \in U(S)$, any $\hat{g} \in \mathbf{G}(g)$ and any x,*

(i) $C_{ea}(x, \hat{g}) = C_{ea}(x, g)$

(ii) $C_{av} = C_{ea}(x, g)$ $P_x^{\hat{g}}$ *a.s. and* P_x^g *a.s.*

(iii) $C_{aed}(x, \hat{g}) = C_{aed}(x, g)$

The results (i)-(iii) hold also for the costs associated with $d^k(\cdot,\cdot)$, $k = 1, 2, \ldots, K$.

(iv)

$$\lim_{N\to\infty} |C_{ed}^N(x, \hat{g}) - E^{\hat{g}}[C_{ed}(X_N, g)|X_0 = x]| = 0 \tag{6.19}$$

$$\lim_{N\to\infty} |D_{ed}^{N,k}(x, \hat{g}) - E^{\hat{g}}[D_{ed}^k(X_N, g)|X_0 = x]| = 0 \ , \quad 1 \le k \le K \tag{6.20}$$

Consequently, whenever g above is optimal for COP_{ea}, COP_{av}, COP_{aed} or COP_{sd}, so is any $\hat{g}$.

Proof: (i) From Lemma 6.1 we have $\eta^{\hat{g}} = \eta^g$ and hence

$$\lim_{t\to\infty} \frac{1}{t+1} \sum_{s=0}^{t} P^u(X_s = y, A_s = a \mid X_0 = x) = \eta^g(y, a)$$

Hence $\bar{F}_{sa}(x,u) = \{\eta^g\} = \bar{F}_{sa}(x,g)$. The result then follows from (3.1).

(ii) By Lemma 6.3, $F_c = \{p^g_{a|y}\}$ $P^{\hat{g}}_x$ a.s. The result (ii) then follows from Lemma 3.5.

(iii) Since by Lemma 6.1 $\eta^{\hat{g}}(y,a)$ exists and is equal to $\eta^g(y,a)$, the derivation in Theorem 4.4 applies verbatim to yield

$$C_{aed}(x,\hat{g}) = [1-\beta]^{-1} C_{ea}(x,g) = C_{aed}(x,g) \; . \tag{6.21}$$

These arguments obviously hold also for $d^k(\cdot,\cdot)$, $k = 1, 2, \ldots, K$.

(iv) From (4.3) it follows that

$$E^{\hat{g}}[C_{ed}(X_N,g)|X_0 = x] = \sum_{t=N}^{\infty} \beta^{t-N} \sum_{y,a} \sum_{z} c(y,a) P^g(X_t = y, A_t = a | X_N = z) P^{\hat{g}}(X_N = z | X_0 = x) \tag{6.22}$$

Let u^N be the policy that uses $\hat{g}$ at $t \le N$ and then uses g. Then

$$\lim_{N\to\infty} |C^N_{ed}(x,\hat{g}) - E^{\hat{g}}[C_{ed}(X_N,g)|X_0 = x]|$$
$$= \sum_{y,a} c(y,a) \left[\sum_{t=N}^{\infty} \beta^{t-N} \Big(P^{\hat{g}}(X_t = y, A_t = a | X_0 = x) - P^{u^N}(X_t = y, A_t = a | X_0 = x) \Big) \right] \tag{6.23}$$

Since $P^{\hat{g}}(X_t = y, A_t = a | X_0 = x) = P^{\hat{g}}(X_t = y | X_0 = x) \cdot P^{\hat{g}}(A_t = a | X_t = y, X_0 = x)$ it follows by Lemma 6.1 and by definition of AS policies that the right hand side of (6.23) converges to zero as $N \to \infty$, which establishes (6.19). The argument for (6.20) is identical. The last claim of the Lemma follows then from the definitions (2.5a-b). ■

7. ESTIMATION AND CONTROL

In this section we introduce an optimal adaptive policy, based on "probing" and on the method of [3]. Recall that we assume no prior information about the transition probabilities, except that A1 holds.

7.1 Estimation of the transition probabilities.

Define

$$\hat{P}^t_{zay} := \frac{\sum_{s=2}^{t} 1\{X_{s-1} = z, A_{s-1} = a, X_s = y\}}{\sum_{s=2}^{t} 1\{X_{s-1} = z, A_{s-1} = a\}}$$

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

$$\lim_{t\to\infty} \hat{P}^t_{zay} = P_{zay} \quad P^u_x \ a.s. \tag{7.1}$$

for all states $z, y \in \mathbf{X}$ and $a \in \mathbf{A}$, it is sufficient that each state is visited infinitely often and moreover, that at each state, each action is used infinitely often. If this condition is met, then (7.1) holds by the strong law of large numbers, as we show in (7.12)–(7.13).

By Lemma 3.4, each state y is indeed visited infinitely often P^u_x a.s. under any policy u. By an appropriate "probing" (described below) we shall obtain $\sum_{s=2}^{\infty} 1\{X_{s-1} = z, A_{s-1} = a\} = \infty \quad P^u_x \ a.s.$ for all $z \in \mathbf{X},\ a \in \mathbf{A}$, implying consistent estimation.

The adaptive policies below define actions at stopping times — when particular states are visited. The connection between this definition and the definition of AS policies is given in the following Lemma. Recall the definition (6.15) of $\tau(y;n)$.

Lemma 7.1: *Assume A1. Then $u \in \mathbf{G}(g)$ iff for every a, y we have P^u_x a.s.*

$$\lim_{n\to\infty} P^u_x(A_{\tau(y;n)} = a | \mathcal{F}_{\tau(y;n)}) = p^g_{a|y} \tag{7.2}$$

($\tau(y;n)$ was defined below Lemma 6.2′).

Proof: Assume $u \in \mathbf{G}(g)$. Since $\tau(y;n) \geq n$ and is finite P^u_x *a.s.* (Lemma 3.4) we have

$$P^u(A_{\tau(y;n)} = a | \mathcal{F}_{\tau(y;n)}) - p^g_{a|y} = \sum_{t=n}^{\infty} 1\{\tau(y;n) = t\} \left[P^u(A_t = a | \mathcal{F}_{\tau(y;n)}) - p^g_{a|y} \right] \tag{7.3}$$

To proceed, we need to show that P^u_x a.s.,

$$1\{\tau(y;n) = t\} P^u_x(A_t = a | \mathcal{F}_{\tau(y;n)}) = 1\{\tau(y;n) = t\} P^u_x(A_t = a | H_{t-1}, X_t = y) \tag{7.4}$$

Let $Y := 1\{A_{\tau(y;n)} = a\} 1\{\tau(y;n) = t\}$. Clearly $E^u_x(Y|\mathcal{F}_{\tau(y;n)}) = 1\{\tau(y;n) = t\} E^u_x(Y|\mathcal{F}_{\tau(y;n)})$, so that for each set $B \in \mathcal{F}_t$

$$\begin{aligned} E^u_x\left[1_B \cdot E^u_x(Y|\mathcal{F}_{\tau(y;n)})\right] &= E^u_x\left[1_B 1\{\tau(y;n) = t\} \cdot E^u_x(Y|\mathcal{F}_{\tau(y;n)})\right] \\ &= E^u_x\left[E^u_x(1_B 1\{\tau(y;n) = t\} \cdot Y | \mathcal{F}_{\tau(y;n)})\right] = E^u_x[1_B \cdot Y] \end{aligned} \tag{7.5}$$

where the second equality follows since by definition of $\mathcal{F}_{\tau(y;n)}$, for every $\mathcal{F}_t$ measurable random variable Z, $Z \cdot 1\{\tau(y;n) = t\}$ is $\mathcal{F}_{\tau(y;n)}$ measurable. By a similar argument, $1\{\tau(y;n) = t\} E^u_x(Y|\mathcal{F}_{\tau(y;n)})$ is $\mathcal{F}_t$ measurable, so that $E^u_x(Y|\mathcal{F}_{\tau(y;n)})$ is $\mathcal{F}_t$ measurable and by (7.5), $E^u_x(Y|\mathcal{F}_{\tau(y;n)}) = E^u_x(Y|\mathcal{F}_t)$. Now

$$
\begin{aligned}
E^u_x(Y|\mathcal{F}_t) &= 1\{\tau(y;n) = t\} \sum_z P^u_x(A_t = a|H_{t-1}, X_t = z) \cdot 1\{X_t = z\} \\
&= 1\{\tau(y;n) = t\} P^u_x(A_t = a|H_{t-1}, X_t = y)
\end{aligned} \tag{7.6}
$$

which completes the proof of (7.4). Using the definition (6.1), equations (7.3)-(7.4) yield

$$
\begin{aligned}
P^u_x(A_{\tau(y;n)} = a|\mathcal{F}_{\tau(y;n)}) - p^g_{a|y} &= \sum_{t=n}^{\infty} 1\{\tau(y;n) = t\} \left[P^u_x(A_t = a|H_{t-1}, X_t = y) - p^g_{a|y} \right] \\
&\leq \epsilon(n) \sum_{t=n}^{\infty} 1\{\tau(y;n) = t\} = \epsilon(n)
\end{aligned} \tag{7.7}
$$

and (7.2) follows.

To prove the converse assume (7.2) holds. Then there exists $\tilde{\epsilon}(n) = \tilde{\epsilon}(n, \omega)$ such that (outside a set in Ω of probability P^u_x zero), $\tilde{\epsilon}(n) \downarrow 0$ as $n \to \infty$, and $\left| P^u_x(A_{\tau(y;n)} = a|\mathcal{F}_{\tau(y;n)}) - p^g_{a|y} \right| < \tilde{\epsilon}(n) \quad P^u_x$ *a.s.* Since $n \leq \tau(y;n)$ we obtain from (7.4):

$$
\begin{aligned}
P^u_x(A_t = a|H_{t-1}, X_t = y) - p^g_{a|y} &= \sum_{n=1}^{t} 1\{\tau(y;n) = t\} \cdot \left[P^u_x(A_t = a|H_{t-1}, X_t = y) - p^g_{a|y} \right] \\
&= \sum_{n=1}^{t} 1\{\tau(y;n) = t\} \left[P^u_x(A_{\tau(y;n)} = a|\mathcal{F}_{\tau(y;n)}) - p^g_{a|y} \right]
\end{aligned} \tag{7.8}
$$

Since $\tau(y;n)$ is finite P^u_x a.s., (7.2) implies for each $s < t$

$$
\lim_{t \to \infty} \sum_{n=1}^{s} 1\{\tau(y;n) = t\} \left[P^u_x(A_{\tau(y;n)} = a|\mathcal{F}_{\tau(y;n)}) - p^g_{a|y} \right] = 0 \quad P^u_x \text{a.s.} \tag{7.9}
$$

On the other hand,

$$
\sum_{n=s+1}^{t} 1\{\tau(y;n) = t\} \left[P^u_x(A_{\tau(y;n)} = a|\mathcal{F}_{\tau(y;n)}) - p^g_{a|y} \right] \leq \tilde{\epsilon}(s) \sum_{n=1}^{t} 1\{\tau(y;n) = t\} = \tilde{\epsilon}(s) \tag{7.10}
$$

which can be made arbitrarily small by choosing s large enough. Combining (7.8)-(7.10), it follows that $\lim_{t \to \infty} P^u_x(A_t = a|H_{t-1}, X_t = y) - p^g_{a|y} = 0 \ P^u_x$ *a.s.* which concludes the proof. ■

7.2 The adaptive policy

Define $B :=$ the space of matrices $\{\beta^a_y\}$ such that for each state y, $\{\beta^a_y\ ,\ a \in \mathbf{A}\}$ is a probability distribution on $\mathbf{A}$. Let g be some optimal stationary policy for COP. Let α be the stationary policy defined by $\alpha_{a|y} = |\mathbf{A}|^{-1}$ for each a, y. Fix a decreasing sequence $\epsilon_r \downarrow 0$ such that $\sum_1^\infty \epsilon_r = \infty$, and an increasing sequence of times T_r with $T_1 = 0$. T_r are the times at which we update the estimate. Suppose we are given a sequence $\Phi = \{\Phi^r\ ,\ r = 1, 2, ...\}$ where each $\Phi^r \in B$ depends on the estimates of the transition probabilities $\hat{P}_{zay},\ z, y \in \mathbf{X},\ a \in \mathbf{A}$ and is an approximation of g. Such control laws are introduced in [3], and are based on sensitivity analysis of Linear Programs (e.g. [7]). In the following Algorithm we construct a policy u which is shown later to be optimal for ACOP.

Algorithm 7.2: If $X_t = y$ and y has been visited n times, and $T_r \le t < T_{r+1}$ then u_t is defined through:

$$u_t(\cdot|H_{t-1}, X_t) = \epsilon_n \cdot \alpha + (1 - \epsilon_n) \cdot \Phi^r(\hat{P}^{T_r}) \tag{7.11}$$

Remark: A way to obtain each Φ^r from the estimate $\hat{P}^{T_r}$ is given in [3], and it involves solving two Linear Programs. One can choose $T_r = q \cdot r$ where q is some positive integer that is proportional to the computation time of Φ^r.

Theorem 7.3: *Assume A1 and assume that COP is feasible. Assume moreover that $\hat{P}^{T_r} \to P$ implies that $\Phi^r(P^{T_r})$ converges to g. Then the policy u obtained through Algorithm 7.2 satisfies $u \in \mathbf{G}(g)$ and is optimal for ACOP.*

Proof: Pick some state y and an action a. By Lemma 3.4 y is visited infinitely often P^u_x a.s. under any policy u. Recall the definition of $\tau(y; n)$ in (6.15). According to (7.11) and the construction of $\{\epsilon_n\}$ we have $\sum_{n\to\infty} P(A_{\tau(y;n)} = a) = \infty$. Hence by Borel Cantelli Lemma, at state y action a is used infinitely often. Let $\sigma(y, a, m)$ be the mth time that $X_t = y, A_t = a$. It follows that $\sigma(y, a, m)$ are finite P^u a.s. since the event $\{X_t = y, A_t = a\}$ occurs i.o. P^u a.s. It can be shown that $1\{X_{\sigma(y,a,m)+1} = z\},\ m = 1, 2, ...$ are i.i.d. variables. It then follows by the Strong Law of Large Numbers that

$$\lim_{m\to\infty} \frac{\sum_{j=1}^m 1\{X_{\sigma(y,a,j)+1} = z\}}{m} = P_{yaz} \quad P^u\ a.s. \tag{7.12}$$

But since T_r and $\sigma(y, a, m)$ are finite P^u a.s. we have

$$\lim_{r\to\infty} \hat{P}^{T_r}_{yaz} = \lim_{r\to\infty} \frac{\sum_{t=2}^{T_r} 1\{X_{t-1} = y, A_{t-1} = a, X_t = z\}}{1\{X_{t-1} = y, A_{t-1} = a\}} =$$

$$= \lim_{m\to\infty} \frac{\sum_{j=1}^{m} 1\{X_{\sigma(y,a,j)+1} = z\}}{m} = P_{yaz} \quad P^u \ a.s. \tag{7.13}$$

Since this holds for any y and a it follows that $\hat{P}^{T_r} \to P$ w.p.1. Hence by hypothesis $\Phi^r(P^{T_r})$ converges to g, which implies by (7.11) that $P_x^u(A_{\tau(y;n)} = a|\mathcal{F}_{\tau(y;n)}) \to p^g_{a|y}$. Hence $u \in \mathbf{G}(\mathbf{g})$ is an asymptotically stationary policy by Lemma 7.1. By Theorem 6.4 it is optimal for the case of expected average costs, average costs, asymptotic expected discounted costs, and for Schäl's criterion. ■

8. APPENDIX

In this Section we generalize Schäl's original criterion to the constrained problem and show, using Example 5.3 that there need not exist an optimal (or even an ϵ-optimal) policy for this problem.

Call a policy u feasible with respect to COP_{Schal} if

$$\overline{\lim_{n\to\infty}}\, E_x^u \left[\sum_{t=n}^{\infty} \beta^{t-n} d^k(X_t, A_t) - V_k \right]^+ = 0 \qquad k = 1, ..., K$$

where $[z]^+ := \max\{z, 0\}$.

Let $C^*_{ed}(x)$ be the optimal value of COP_{ed} starting at $X_0 = x$. A policy u is said to be optimal with respect to COP_{Schal} if it is feasible and

$$\lim_{n\to\infty} E_x^u \left| \sum_{t=n}^{\infty} \beta^{t-n} c(X_t, A_t) - C^*_{ed}(X_n) \right| = 0$$

Consider Example 5.3, and note that for $t \geq 2$ we have $P(X_t = 1) = 0.1$ under any policy and for any initial state. For any policy u we have:

$$E_x^u \left[\sum_{t=n}^{\infty} \beta^{t-n} d(X_t, A_t) - V \right]^+ \geq P_x^u(X_n = 1) E_x^u \left(\left[\sum_{t=n}^{\infty} \beta^{t-n} d(X_t, A_t) - V \right]^+ \middle| X_n = 1 \right)$$

$$\geq 0.1 E_x^u \left(\left[\sum_{t=n}^{\infty} \beta^{t-n} d(X_t, A_t) - V \right] \middle| X_n = 1 \right)$$

where the last inequality follows from Jensen's inequality and the fact that the last expression is always non-negative. For it to converge to zero, it is clearly necessary that

$$\lim_{n\to\infty} P_x^u\left(A_{n+1} = a \mid X_{n+1} = 2\right) = 1 \tag{8.1}$$

But note that $C^*(x) = 1\{x = 1\}C^*(1) + 1\{x = 2\}C^*(2)$ where

$$\begin{aligned} C^*(1) &= \sum_{t=1}^{\infty} \left[0.1^t \cdot 1 \cdot P_1^{g_1}(X_t = 2)\right] = 0.9\left(\frac{1}{1-0.1} - 1\right) = 0.1 \\ C^*(2) &= 0 \end{aligned}$$

But using Jensen's inequality,

$$\begin{aligned} E_x^u \left| \sum_{t=n}^{\infty} \beta^{t-n} c(X_t, A_t) - C_{ed}^*(X_n) \right| &\geq P_x^u(X_n = 2) E_x^u \left(\left| \sum_{t=n}^{\infty} [\beta^{t-n} c(X_t, A_t) - C_{ed}^*(X_n)] \right| |X_n = 2 \right) \\ &\geq P_x^u(X_n = 2) \left| E_x^u \left[\sum_{t=n}^{\infty} \beta^{t-n} c(X_t, A_t) - 0 \middle| X_n = 2 \right] \right| \end{aligned}$$

and therefore for u to be optimal it should follow asymptotically g_2, i.e.

$$\lim_{n\to\infty} P_x^u\left(A_{n+1} = a \mid X_{n+1} = 2\right) = 0$$

which contradicts (8.1). Therefore there does not exist any optimal policy, and moreover, there exists an $\epsilon > 0$ such that any feasible policy is not even ϵ-optimal (in the obvious sense).

REFERENCES

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

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

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

[4] Beutler F. J. and K. W. Ross, "Optimal policies for controlled Markov chains with a constraint", *Math. Anal. Appl.* Vol. 112, pp. 236-252, 1985.

[5] Borkar V. S., "On minimum cost per unit time control of Markov chains", *SIAM J. Control Opt.*, Vol. 22 No. 6, pp. 965-978, November 1984.

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

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

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

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

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

[11] Hernandez-Lerma O. and S. I. Marcus, "Adaptive Control of Discounted Markov Decision Chains", *J. Opt. Theory Appl.* Vol. 46 No. 2, pp. 27-235, June 1985.

[12] Hernandez-Lerma O. and S. I. Marcus, "Discretization procedures for adaptive Markov control processes", *J. Math. Anal . Appl.,* to appear.

[13] Hernandez-Lerma O. and S. I. Marcus, "Adaptive policies for discrete-time stochastic systems with unknown disturbance distribution" *System Control Letters* Vol. 9, pp. 307-315, 1987.

[14] Hordijk A., *Dynamic Programming and Markov potential theory* , Mathematical Center Tracts No. 51, Amsterdam, 1974.

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

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

[17] Kemeny J. G. and J. L. Snell, *Finite Markov Chains* , D. Van Nostrand Company, 1960.

[18] Kumar P. R., "A survey of some results in stochastic adaptive control", *SIAM J. Control Opt.*, Vol. 23 No. 3, pp. 329-380, May 1985.

[19] Makowski A. M. and A. Shwartz, "Implementation issues for Markov decision processes", *Proc. of a workshop on Stochastic Differential Systems, Stochastic Control Theory and Applications,* Inst. Math. Appl., Univ. of Minnesota, Springer-Verlag Lecture Notes, Control and Inf. Sci., Edited by W. Fleming and P.-L. Lions, 1986.

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

[21] Ross K. W. and A R. Varadarajan, "Markov decision processes with sample path constraints: Unichain, communicating and deterministic cases", Submitted to *Operations Research,* July 1986.

[22] Schäl M., "Conditions for optimality in dynamic programming and for the limit of n-stage optimal policies to be optimal", *Z. Wahrscheinlichkeitstheorie und verw. Geb.* Vol. 32, pp. 179-196, 1975.

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

[24] Shwartz A. and A. M. Makowski, "An Optimal Adaptive Scheme for Two Competing Queues with Constraints", *Analysis and Opt. of Systems,* edited by A. Bensoussan and J. L. Lions, Springer Verlag lecture notes in Control and Info. Scie. No. 83, 1986.

[25] Van Hee, K. M. *Bayesian Control of Markov Chains*, Mathematical Centre Tracts 95, Amsterdam, 1978.