2011 50th IEEE Conference on Decision and Control and
European Control Conference (CDC-ECC)
Orlando, FL, USA, December 12-15, 2011

Equilibrium Control Policies for Markov Chains

Andreas A. Malikopoulos, *Member, IEEE*

***Abstract*— The average cost criterion has held great intuitive appeal and has attracted considerable attention. It is widely employed when controlling dynamic systems that evolve stochastically over time by means of formulating an optimization problem to achieve long-term goals efficiently. The average cost criterion is especially appealing when the decision-making process is long compared to other timescales involved, and there is no compelling motivation to select short-term optimization. This paper addresses the problem of controlling a Markov chain so as to minimize the average cost per unit time. Our approach treats the problem as a dual constrained optimization problem. We derive conditions guaranteeing that a saddle point exists for the new dual problem and we show that this saddle point is an equilibrium control policy for each state of the Markov chain. For practical situations with constraints consistent to those we study here, our results imply that recognition of such saddle points may be of value in deriving in real time an optimal control policy.**

I. INTRODUCTION

New technologies in mechatronics and actuators have induced significant enhancement in the complexity of modern engineering systems. Exact modeling of complex systems is often infeasible or expensive, and thus deriving an optimal control policy can be intractable. This challenge has motivated continuing research on computational learning methods towards making autonomous intelligent systems that can learn how to improve their performance over time while interacting with their environment. The problem is formulated as decision-making under uncertainly in which an intelligent system (decision maker), *e.g.*, hybrid-electric vehicle, robot, automated manufacturing system, etc, is faced with the task to select those actions in several time steps (decision epochs) to achieve long-term goals efficiently. In this paper, we focus on the system's decision-making process rather than its learning mechanism.

Decision-making problems have been the object of intense study for many decades. Blackwell's [1] influential paper provided considerable incentive in this area by utilizing the discounted cost criterion extensively. In this work, the *vanishing discount approach* was developed, *i.e.*, treating the average cost criterion as the limit of the discounted one. Derman [2] showed that *Blackwell's optimality* can be also considered in problems employing the average cost criterion. Blackwell optimal policies, however, do not necessarily exist when the state space is countable infinite or the action space is an arbitrary compact metric space. For such models, Feinberg [3] proved that under certain conditions an optimal policy may exist.

The average cost criterion for Markov chains with finite state and arbitrary action spaces has been extensively reported in the literature (see, *e.g.*, [3], [4], [5], [6] and references therein). Mathematically, the average cost criterion is prominent as being complex to analyze compared to others; while other classical criteria lead to rational complete solutions, the long-run cost does not. Although the average cost criterion in Markov chains with finite state and action spaces is well understood [7], [8], [9], [10], [11], [12], there are numerous counterexamples in which models with infinite state or action spaces do not have a nice solution. Bather [13] reviewed various techniques for a controlled Markov chain with a finite state space when there is a finite set of possible transition matrices. Feinberg [14] considered four average reward criteria on discrete time Markov decision model with a finite state space, and prove the existence of persistently nearly optimal strategies in various classes of strategies for models with complete state information.

A significant amount of research on infinite horizon, discrete-time Markov decision processes (MDPs) has focused on more general state and action spaces. Hordijk [15] extended some earlier results to countable state and action spaces by introducing the *Lyapunov function* method for controlled Markov processes. Borkar [16], [17], [18], [19], [20] presented a convex analytic approach to address this problem in a general framework with unbounded cost by treating the control problem as a constrained optimization problem on a suitably defined closed convex set of *ergodic occupation measures*. In this work, the necessary and sufficient conditions for the existence of an optimal stable stationary deterministic policy were established; moreover, Borkar provided conditions for optimality in terms of the dynamic programming when an optimal stable stationary policy is known to exist. Sennott [21] introduced conditions that guarantee an optimal control policy in problems with possibly unbounded, non-negative costs. Zhu, Guo, and Dai [22] presented another set of conditions under which an optimal stationary policy exists when both the *limit* of the *supremum* and *infimum* average criteria are employed. Cavazos-Cadena [23] considered denumerable state spaces

Manuscript received March 20, 2011. This research was supported by the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory, managed by UT-Battelle, LLC, for the U.S. Department of Energy.

A.A. Malikopoulos is with the Energy & Transportation Science Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831 USA (phone: 865-946-1529; fax: 865-946-1354; e-mail: andreas@ornl.gov) .

and stationary control policies that induce an ergodic chain; the *value iteration* scheme was utilized to construct convergent approximations of a solution to the *optimality equation* as well as a sequence of stationary policies whose limit points are optimal. Leizarowitz and Zaslavski [24] recently addressed the problem of uniqueness and stability of optimal control policies when a complete set of unicost MDPs is endowed.

Hermandez-Lerma and Lasserre [25] provided weak assumptions for MDPs on *Borel* state and action spaces with possibly unbounded costs. In this paper, it was shown that the *optimality inequality* holds everywhere and there exists a stationary policy which is optimal whenever the initial state lies in a possibly *proper* subset of the state space. Montes-de-Oca and Vega-Amaya [26], [27] considered the same problem and employed a variant of the *vanishing discount factor* that ensures the existence of a solution to the *optimality equation*. Guo and Shi [28] presented a new set of conditions under which the existence of both a solution to the optimality equations and the limiting average ϵ-optimal Markov policies is derived.

In this paper, we address the problem of controlling a Markov chain with a finite state space and compact action space so as to minimize the long run, average cost per unit time. We formulate the problem as a dual constrained optimization problem and we derive conditions that guarantee the existence of a saddle point solution. Furthermore, we show that this saddle point is an equilibrium stationary control policy for each state of the Markov chain. Equilibrium control policies may be of value in problems required to extract optimal control policies in real time, *e.g.*, powertrain systems modeled as a controlled Markov chain, as has been shown in earlier work [29].

The remainder of the paper proceeds as follows: In Section II, we introduce our notation and develop the general framework of the long-run, average cost problem. In Section III, we reformulate the problem as minimax constrained optimization problem and provide conditions that the saddle point solution exists. In Section IV, we define the equilibrium control policies and prove that the control policy yielding the saddle point is an equilibrium control policy for all states in the Markov chain. A simple example illustrating the equilibrium control policy in a controlled Markov chain is demonstrated in Section V. Concluding remarks are presented in Section VI.

II. PROBLEM FORMULATION

A. Controlled Markov Chain

We consider a controlled Markov chain with a state space $\mathcal{S} \subset \mathbb{R}^n$ on which a controlled stochastic process evolves, and a control space $\mathcal{U} \subset \mathbb{R}^m$ from which control action are chosen. We assume that $\mathcal{S}$ and $\mathcal{U}$ are bounded and measurable. The dynamics of the system are described by a Borel measurable function $\mathcal{P}\colon \mathcal{S} \times \mathcal{S} \times \mathcal{U} \to [0,1]$. In our formulation a state-dependent constraint is incorporated; that is, for each state $i \in \mathcal{S}$, we are given a nonempty set $\mathcal{C}(i) \subset \mathcal{U}$ of admissible control actions.

Definition 2.1: We define the set of admissible state/action pairs

$$\Gamma\colon = \{(i,u) | i \in \mathcal{S} \text{ and } \mathrm{u} \in \mathcal{C}(i)\}.$$

We assume that Γ is the intersection of a closed subset of $\mathbb{R}^n \times \mathbb{R}^m$ with the set $\mathcal{S} \times \mathcal{U}$. That is, Γ is closed with respect to the induced topology on $\mathcal{S} \times \mathcal{U}$. It follows that for any $i \in \mathcal{S}$, $\mathcal{C}(i)$ is compact.

Definition 2.2: We define the set of Borel measurable functions as $\Pi_i\colon = \{\mu_i\colon \mathcal{S} \to \mathcal{U} | \mu_i \text{ is Borel measurable and } \mu_i \in \mathcal{C}(i)\}, \forall i \in \mathcal{S}$.

Let $\Pi\colon = \Pi_i^{\infty}$, $i \in \mathcal{S}$, be the set of all set of all sequences $\pi = \{\mu_1, \mu_2, ..., \mu_n\}$. Each sequence in Π is called a *stationary control policy* and operates as follows. Associated with each state $i \in \mathcal{S}$ is the Borel measurable function $\mu_i \in \mathcal{C}(i)$. If at any time the controller finds the system in state i, then the controller always chooses the action μ_i.

The evolution of the system occurs at each of a sequence of stages $t = 0, 1, ...$, and it is portrayed by the sequence of the random variables X_t and U_t corresponding to the system's state and control action. At each stage, the controller observes the system's state $X_t = i \in \mathcal{S}$, and executes an action $U_t = \mu_i$, from the feasible set of actions $\mu_i \in \mathcal{C}(i)$ at this state. At the next stage t, the system transits to the state $X_{t+1} = j \in \mathcal{S}$ imposed by the conditional probability $P(X_{t+1} = j | X_t = i, U_t = \mu_i)$, and a cost $k(X_t = i, U_t = \mu_i) = k(i, \mu_i)$ is incurred. After the transition to the next state has occurred, a new action is selected, and the process is repeated. The completed period of time over which the system is observed is called the *decision-making horizon* and is denoted by T. The horizon can be either finite or infinite; in this paper, we consider infinite-horizon decision-making problems.

B. Optimal Control Policy

We are concerned with deriving a stationary optimal control policy to minimize the long run average cost per unit time, that is

$$J(\pi) = \min_{\pi \in \Pi} \lim_{T \to \infty} \frac{1}{T+1} E\left[\sum_{0}^{T} k(X_t, U_t)\right]. \qquad (1)$$

For each policy $\pi \in \Pi$ we denote $\mathbf{P}(\pi)$ the transition probability matrix, the elements of which represent the conditional probability of moving from one state to another under the policy π, that is, $P(X_{t+1} = j | X_t = i, U_t = \mu_i)$. To guarantee that the limit in Eq. (1) exists, we assume that for each stationary control policy $\pi = \{\mu_1, \mu_2, ..., \mu_n\}$, the Markov chain $\{X_t | t = 1, 2, ...\}$ has a single ergodic class. Namely, for each stationary policy $\pi \in \Pi$, there is a unique probability distribution (row vector) $\beta(\pi) = [\beta_1(\pi), \beta_2(\pi), ..., \beta_i(\pi), ...\beta_n(\pi)], \forall i \in \mathcal{S}$, such that $\beta(\pi) = \beta(\pi) \cdot \mathbf{P}(\pi)$, with $\sum_{i \in \mathcal{S}} \beta_i(\pi) = 1$. A proof of this assertion may be found in [[30], p. 227]. Under our assumption, it is known [[31], p.175] that

$$\lim_{T \to \infty} \frac{1}{T+1} \sum_{t=0}^{T} [\mathbf{P}(\pi)]^t = \mathbf{1} \cdot \beta(\pi), \qquad (2)$$

where $\mathbf{1} = [1, 1, ..., 1]^T$ is the column vector whose elements are all unity. Substituting Eq. (2) into Eq. (1) shows that long run average cost, $J(\pi)$, does not depend on the initial state and is given more simply as

$$J(\pi) = \beta(\pi) \cdot k(\pi), \tag{3}$$

where $k(\pi) = [k(1, \mu_1), k(2, \mu_2), ..., k(i, \mu_i), ...k(n, \mu_n)]^T$ is the column vector of the cost function. Consequently, a stationary control policy is optimal if

$$J^* = J(\pi) = \inf \{J(\pi) | \pi \in \Pi\}. \tag{4}$$

Since we assume $\mathbf{P}(\pi)$ to be continuous, it follows from Eq. (2) that $\beta(\pi)$ is continuous. Since $k(\pi)$ is also assumed continuous, so is $J(\pi)$. Hence, by compactness of $\mathcal{U}$, an optimal stationary control policy exists. Our objective is to investigate the policies where the average cost is minimized.

III. CONDITIONS FOR EXISTENCE OF A SADDLE POINT

A. Dual Problem Formulation

In this section, we formulate the problem of deriving an optimal control policy with the average cost criterion as a dual constrained optimization problem and we provide conditions for existence of a saddle point solution. The motivation behind this new formulation is the structure of the average cost as expressed in Eq. (3). In particular, the average cost depends on two vectors, *i.e.*, the stationary probability distribution, $\beta(\pi)$, and the vector of the cost function, $k(\pi)$. The summation of the elements of $\beta(\pi)$ equals to one, that is, $\sum_{i \in \mathcal{S}} \beta(\pi) = 1$, and the Markov chain has a single ergodic class; since we permit the single ergodic class to depend on π, different control policies will yield different probability distributions for each state $i \in \mathcal{S}$. The elements of the vector of the cost function, $k(\pi)$, on the other hand, even though depend on the control policy they are constant and known *a priori* for each state as designated by the problem. Consequently, we seek a solution ensuring that the control policy endows a stationary probability distribution of the states in the Markov chain that yields higher probability at the states with low cost, and lower probability at the states with high cost.

The implication behind the aforementioned observations is that we can formulate the problem of deriving an optimal control policy, Eqs. (3) and (4), as a dual constrained optimization problem. Namely, we try to obtain a control policy, π, that not only minimizes the cost at each state but also maximizes the probability of that state. In similar fashion, we could state that we aim at deriving a control policy that maximizes the probability of the states incurring minimum cost.

Thus, we can formulate the problem as the following dual optimization problem. Consider a function $f : \mathcal{K} \times \mathcal{B} \to \mathbb{R}$, where $\mathcal{K}$ and $\mathcal{B}$ are nonempty subsets of $\mathbb{R}^n$. We wish to either $\min_{k \in \mathcal{K}} \sup_{\beta \in \mathcal{B}} f(k, \beta) = \min_{k \in \mathcal{K}} \sup_{\beta \in \mathcal{B}} \beta \cdot k$, or $\max_{\beta \in \mathcal{B}} \inf_{k \in \mathcal{K}} f(k, \beta) = \max_{\beta \in \mathcal{B}} \inf_{k \in \mathcal{K}} \beta \cdot k$

Definition 3.1: A pair of vectors $k \in \mathcal{K}$ and $\beta \in \mathcal{B}$, where $\mathcal{K}, \mathcal{B} \subseteq \mathbb{R}^n$, is called a saddle point of the function $f(k, \beta) = \beta \cdot k$, if

$$f(k^*, \beta) \leq f(k^*, \beta^*) \leq f(k, \beta^*), \forall k \in \mathcal{K}, \forall \beta \in \mathcal{B}. \tag{5}$$

Note that (k^*, β^*) is a saddle point if and only if $k^* \in \mathcal{K}, \beta^* \in \mathcal{B}$, and

$$\sup_{\beta \in \mathcal{B}} f(k^*, \beta) = f(k^*, \beta^*) = \inf_{k \in \mathcal{K}} f(k, \beta^*). \tag{6}$$

B. Conditions of Existence

Employing the framework for duality analysis, we can derive the conditions guaranteeing that the minimax equality holds, namely, $\sup_{\beta \in \mathcal{B}} \inf_{k \in \mathcal{K}} f(k, \beta) = \inf_{k \in \mathcal{K}} \sup_{\beta \in \mathcal{B}} f(k, \beta)$.

Theorem 3.1 (Classical Saddle Point Theorem): If $\mathcal{K}$ and $\mathcal{B}$ are nonempty convex and compact subsets of $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively, and $f : \mathcal{K} \times \mathcal{B} \to \mathbb{R}$ is a function such that $f(\cdot, \beta) : \mathcal{K} \to \mathbb{R}$ is convex and closed for each $\beta \in \mathcal{B}$, and $-f(k, \cdot) : \mathcal{B} \to \mathbb{R}$ is convex and closed for each $k \in \mathcal{K}$, then the minimax equality holds and the set of saddle points of $f : \mathcal{K} \times \mathcal{B} \to \mathbb{R}$ is nonempty and compact.

Proof: See [[32], Proposition 5.5.3, p.204]. ■

In our dual constrained optimization problem formulation, the properties of the probability distribution are fixed, *i.e.*, unique probability distribution for each policy and $\sum_{i \in \mathcal{S}} \beta_i(\pi) = 1$. So the hypotheses of *Theorem 3.1* will aim to establish the desired inherent properties of the vector of the cost function. The following result inaugurates the structure of the vector cost that guarantees the existence of the saddle point in our problem formulation, that is, Eq. (6) holds.

Proposition 3.1: If $f : \mathcal{K} \times \mathcal{B} \to \mathbb{R}$ is a function such that $f(\cdot, \beta) : \mathcal{K} \to \mathbb{R}$ is convex and closed for each $\beta \in \mathcal{B}$, then the set of saddle points of $f : \mathcal{K} \times \mathcal{B} \to \mathbb{R}$ is nonempty.

Proof: In our dual problem formulation of the average cost, the set $\mathcal{B}$ of the stationary probability distribution of the Markov chain is $\mathcal{B} = [0, 1]^n$, and the set $\mathcal{K}$ of one-stage cost function is $\mathcal{K} \subset \mathbb{R}^n$. Both are convex and compact subsets of $\mathbb{R}^n$; so this portion of *Theorem 3.1* is satisfied. Furthermore, the function $-f(k, \cdot) : \mathcal{B} \to \mathbb{R}$ is convex and closed for each $k \in \mathcal{K}$ since $\sum_{i \in \mathcal{S}} \beta_i(\pi) = 1$. So, if the function $f(\cdot, \beta) : \mathcal{K} \to \mathbb{R}$ is convex and closed for each $\beta \in \mathcal{B}$, the conditions of *Theorem 3.1* are satisfied, and thus the set of saddle points of $f : \mathcal{K} \times \mathcal{B} \to \mathbb{R}$ is nonempty. ■

IV. EQUILIBRIUM CONTROL POLICY

A. Basic Definitions

In this section we define the equilibrium control policies and prove that the control policy yielding the saddle point in Eq. (6) is an equilibrium control policy for all states in the Markov chain.

Let $\mathcal{U}$ be a Borel measurement subset of the Euclidean space $\mathbb{R}^m$. We use $\boldsymbol{\mathcal{B}}(\mathcal{U})$ to denote the space of all continuous, bounded Borel measurable functions on $\mathcal{U}$. We view the Euclidean space $\mathbb{R}^m$ as a normed vector space by endowing it with the sup-norm $||\cdot||_\infty$. We also use $||\cdot||_\infty$ to denote the

sup-norm on $\boldsymbol{\mathcal{B}}(\mathcal{U})$. It is well known that $\boldsymbol{\mathcal{B}}(\mathcal{U})$ is a Banach space with respect to sup-norm [33], [34].

An operator $T\colon \boldsymbol{\mathcal{B}}(\mathcal{U}) \to \boldsymbol{\mathcal{B}}(\mathcal{U})$ is called monotone operator if $J \leqslant J'$ implies $TJ \leqslant TJ'$. Furthermore, if there exist some $k \in (0,1)$ such that $||TJ - TJ'||_\infty \leqslant k \cdot ||J - J'||_\infty$ for all $J, J' \in \boldsymbol{\mathcal{B}}(\mathcal{U})$ then T is called a *contraction operator* on $\boldsymbol{\mathcal{B}}(\mathcal{U})$ with contraction factor k.

Definition 4.1: Let $\pi \in \Pi$ be a control policy with $\pi = \{\mu_1, \mu_2, ..., \mu_i, ..., \mu_n\}$. We define the continuous function $\theta\colon \boldsymbol{\mathcal{B}}(\mathcal{U}) \to \mathbb{R}$ by

$$\theta(\pi, \pi') = \theta(\pi)\colon = \max\{0, \beta(\pi) \cdot k(\pi) - \beta(\pi') \cdot k(\pi')\}, \tag{7}$$

where $\pi' = \{\mu'_1, \mu'_2, ..., \mu'_i, ..., \mu'_n\}$ is any other control policy in Π; $\beta(\pi) = [\beta_1(\pi), \beta_2(\pi), ..., \beta_i(\pi), ...\beta_n(\pi)]$, $\beta(\pi') = [\beta_1(\pi'), \beta_2(\pi'), ..., \beta_i(\pi'), ...\beta_n(\pi')], \forall i \in \mathcal{S}$, are the stationary probability distributions (row vectors) endowed by the control policies π and π' respectively; and $k(\pi)$, $k(\pi')$ with $k(\pi) = [k(1, \mu_1), k(2, \mu_2), ..., k(i, \mu_i), ...k(n, \mu_n)]^T$, and $k(\pi') = [k(1, \mu'_1), k(2, \mu'_2), ..., k(i, \mu'_i), ...k(n, \mu'_n)]^T, \forall i \in \mathcal{S}$, are the column vectors of the cost function incurred when the control policies π and π' are employed respectively. Recall that we have assumed that for each control policy π, the Markov chain has a single ergodic class; that is, we permit the single ergodic class to depend on π.

Definition 4.2: We define the following continuous function of π by

$$\lambda(\pi) := \begin{cases} 1, & \text{if } \theta(\pi) = 0. \\ 0, & \text{if } \theta(\pi) \neq 0. \end{cases} \tag{8}$$

Definition 4.3: Let π be a control policy in Π. We define the continuous operator $T\colon \boldsymbol{\mathcal{B}}(\mathcal{U}) \to \boldsymbol{\mathcal{B}}(\mathcal{U})$ by

$$TJ(\pi, \pi') = TJ(\pi) := \frac{\lambda(\pi) \cdot J(\pi) + (1 - \lambda(\pi)) \cdot J(\pi')}{1 + \lambda(\pi) \cdot \theta(\pi)}, \tag{9}$$

where π' is any other control policy in Π.

Lemma 4.1: For any function $J\colon \boldsymbol{\mathcal{B}}(\mathcal{U}) \to \mathbb{R}$ such that $J(\pi_1) \leqslant J(\pi_2)$ and for any control policy $\pi_1, \pi_2 \in \Pi$, we have

$$TJ(\pi_1) \leqslant TJ(\pi_2). \tag{10}$$

Proof: For the following two control policies $\pi_1 = \{\mu_1, \mu_2, ..., \mu_i, ..., \mu_n\}$, and $\pi_2 = \{\mu'_1, \mu'_2, ..., \mu'_i, ..., \mu'_n\}$, let $J(\pi_1) \leqslant J(\pi_2)$. We compute the functions $\theta(\pi_1)$ and $\theta(\pi_2)$ using Eq. (7), and also the functions $\lambda(\pi_1)$ and $\lambda(\pi_2)$ using Eq. (8); that is, $\theta(\pi_1) = \max\{0, \beta(\pi_1) \cdot k(\pi_1) - \beta(\pi_2) \cdot k(\pi_2)\}$, and $\theta(\pi_2) = \max\{0, \beta(\pi_2) \cdot k(\pi_2) - \beta(\pi_1) \cdot k(\pi_1)\}$.

Since $J(\pi_1) \leqslant J(\pi_2)$ we have that $\theta(\pi_1) = 0$, and $\theta(\pi_2) \neq 0$ yielding $\lambda(\pi_1) = 1$ and $\lambda(\pi_2) = 0$. Substituting the above functions into Eq. (9) for each control policy $\pi_1, \pi_2 \in \Pi$ separately we have:

$$TJ(\pi_1) = \frac{\lambda(\pi_1) \cdot J(\pi_1) + (1 - \lambda(\pi_1)) \cdot J(\pi_2)}{1 + \lambda(\pi_1) \cdot \theta(\pi_1)} = J(\pi_1). \tag{11}$$

Following the same procedure when the operator is applied to $J(\pi_2)$ we have:

$$TJ(\pi_2) = \frac{\lambda(\pi_2) \cdot J(\pi_2) + (1 - \lambda(\pi_2)) \cdot J(\pi_1)}{1 + \lambda(\pi_2) \cdot \theta(\pi_2)} = J(\pi_1). \tag{12}$$

That is, for any control policy $\pi_1, \pi_2 \in \Pi$, if $J(\pi_1) \leqslant J(\pi_2)$ then $TJ(\pi_1) \leqslant TJ(\pi_2)$. ■

Lemma 4.2: For any function $J\colon \boldsymbol{\mathcal{B}}(\mathcal{U}) \to \mathbb{R}$, a stationary policy $\pi \in \Pi$, and a scalar $c \in \mathbb{R}$ we have

$$T(J(\pi) + c) \leqslant TJ(\pi) + \frac{c}{1 + \lambda(\pi) \cdot \theta(\pi)}. \tag{13}$$

Proof: We apply the operator defined in Eq. (9) to $J(\pi) + c$. So, $\forall \pi' \in \Pi$ we have

$$\begin{aligned} T(J(\pi) + c) &= \frac{\lambda(\pi) \cdot (J(\pi) + c) + (1 - \lambda(\pi)) \cdot J(\pi')}{1 + \lambda(\pi) \cdot \theta(\pi)} \\ &= TJ(\pi) + \frac{c}{1 + \lambda(\pi) \cdot \theta(\pi)}. \end{aligned} \tag{14}$$

■

Proposition 4.1: For any two bounded functions $J\colon \boldsymbol{\mathcal{B}}(\mathcal{U}) \to \mathbb{R}$ there holds

$$||TJ(\pi) - TJ(\pi')||_\infty \leqslant k \cdot \ ||J(\pi) - J(\pi')||_\infty, k \in (0,1). \tag{15}$$

Proof: Denote $c = ||J(\pi) - J(\pi')||_\infty, \forall \pi \in \Pi$. Then we have $J(\pi) - c \leqslant J(\pi') \leqslant J(\pi) + c, \forall \pi \in \Pi$. Applying the operator $T\colon \boldsymbol{\mathcal{B}}(\mathcal{U}) \to \boldsymbol{\mathcal{B}}(\mathcal{U})$ using *Lemma 4.1* and then *Lemma 4.2*, we obtain

$$T(J(\pi) - c) \leqslant TJ(\pi') \leqslant T(J(\pi) + c), \tag{16}$$

Since $c = ||J(\pi) - J(\pi')||_\infty < 1 + \theta(\pi)$, it follows that $||TJ(\pi) - TJ(\pi')||_\infty \leqslant k \cdot ||J(\pi) - J(\pi')||_\infty$, where $0 \leqslant k = \frac{c}{1+\theta(\pi)} < 1$. ■

B. Equilibrium Control Policy

Definition 4.4 : A control policy $\pi = \{\mu_1, \mu_2, ..., \mu_i, ..., \mu_N\}$ is an equilibrium control policy if and only if the policy yields the saddle point of the product of the stationary probability distribution $\beta(\pi)$ and cost function $k(\pi)$, that is

$$\beta^*(\pi) \cdot k^*(\pi) = \beta(\pi) \cdot k(\pi) \leqslant \beta(\pi') \cdot k(\pi'), \tag{17}$$

where $\pi' = \{\mu'_1, \mu'_2, ..., \mu'_i, ..., \mu'_n\}$ is any other control policy in Π.

Thus an equilibrium control policy yields those probability distribution and cost function vectors that minimize the average cost of the controlled Markov chain.

Theorem 4.1: The control policy $\pi = \{\mu_1, \mu_2, ..., \mu_i, ..., \mu_N\}$ that yields the saddle point at each state of the Markov chain is an equilibrium control policy.

Proof: If the control policy is fixed under the operator T, then $\theta(\pi) = 0$, and so $\lambda(\pi) = 1$, since

$$\beta^*(\pi) \cdot k^*(\pi) \leqslant \beta(\pi') \cdot k(\pi'). \tag{18}$$

This means there is no control policy that can do any better. Conversely, if $\pi = \{\mu_1, \mu_2, ..., \mu_i, ..., \mu_N\}$ is an equilibrium policy, it is immediate that $\theta(\pi) = 0$, and so $\lambda(\pi) = 1$, making the control policy $\pi \in \Pi$ a fixed point of the operator $T\colon \boldsymbol{\mathcal{B}}(\mathcal{U}) \rightarrow \boldsymbol{\mathcal{B}}(\mathcal{U})$. Since $\boldsymbol{\mathcal{B}}(\mathcal{U})$ is a Banach space, the operator $T\colon \boldsymbol{\mathcal{B}}(\mathcal{U}) \rightarrow \boldsymbol{\mathcal{B}}(\mathcal{U})$ is a *contraction operator* on $\boldsymbol{\mathcal{B}}(\mathcal{U})$. According to Banach fixed-point theorem the operator $T\colon \boldsymbol{\mathcal{B}}(\mathcal{U}) \rightarrow \boldsymbol{\mathcal{B}}(\mathcal{U})$ has unique fixed point, which must be the equilibrium point. ■

V. ILLUSTRATIVE EXAMPLE

A. Problem Formulation

We consider a controlled Markov chain with a state space $\mathcal{S}$ consisting of two states numbered 1 and 2, $\mathcal{S} = \{1, 2\}$, and a control space consisting of two control actions - also numbered 1 and 2 - for each state, namely, $\mathcal{C}(1) = \mathcal{C}(2) = \mathcal{U} = \{1, 2\}$. The transition probability matrices associated with the control actions 1 and 2 respectively are: $\mathbf{P}_1 = \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix}$, and $\mathbf{P}_2 = \begin{bmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{bmatrix}$. The transition cost matrices associated with the control actions 1 and 2 respectively are: $\mathbf{R}_1 = \begin{bmatrix} 6 & 5 \\ 7 & 12 \end{bmatrix}$, and $\mathbf{R}_2 = \begin{bmatrix} 10 & 17 \\ 14 & 13 \end{bmatrix}$.

In this problem there are 4 possible control policies that can be employed to control the Markov chain, namely, $\pi_1 = \{1, 1\}, \pi_2 = \{1, 2\}, \pi_3 = \{2, 1\}$ and $\pi_4 = \{2, 2\}$. We seek to derive the optimal control policy that minimizes the average cost per unit time. The transition probability and cost matrices for each control policy are: $\mathbf{P}_{\pi_1} = \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix}$, $\mathbf{P}_{\pi_2} = \begin{bmatrix} 0.7 & 0.3 \\ 0.2 & 0.8 \end{bmatrix}$, $\mathbf{P}_{\pi_3} = \begin{bmatrix} 0.9 & 0.1 \\ 0.4 & 0.6 \end{bmatrix}$, $\mathbf{P}_{\pi_4} = \begin{bmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{bmatrix}$, $\mathbf{R}_{\pi_1} = \begin{bmatrix} 6 & 5 \\ 7 & 12 \end{bmatrix}$, $\mathbf{R}_{\pi_2} = \begin{bmatrix} 6 & 5 \\ 14 & 13 \end{bmatrix}$, $\mathbf{R}_{\pi_3} = \begin{bmatrix} 10 & 17 \\ 7 & 12 \end{bmatrix}$, and $\mathbf{R}_{\pi_4} = \begin{bmatrix} 10 & 17 \\ 14 & 13 \end{bmatrix}$.

B. Average Cost

The stationary probability distributions endowed by the control policies π_1, π_2, π_3 and π_4 can be computed by Eq. (2) employing the transition probability matrices corresponding to each policy; namely, $\beta(\pi_1) = [0.57, 0.43], \beta(\pi_2) = [0.40, 0.60]$, and $\beta(\pi_3) = [0.80, 0.20], \beta(\pi_4) = [0.67, 0.33]$.

The column vectors, $k(\pi)$, of the cost function for each control policy is computed as follows:

$$\begin{aligned} k(\pi_1) &= [k(1,1), k(2,1)]^T \\ &= [\mathbf{P}_{\pi_1}(1,1) \cdot \mathbf{R}_{\pi_1}(1,1) + \mathbf{P}_{\pi_1}(1,2) \cdot \mathbf{R}_{\pi_1}(1,2), \\ &\quad \mathbf{P}_{\pi_1}(2,1) \cdot \mathbf{R}_{\pi_1}(2,1) + \mathbf{P}_{\pi_1}(2,2) \cdot \mathbf{R}_{\pi_1}(2,2)]^T \\ &= [5.7, 10]^T. \end{aligned} \tag{19}$$

The cost functions for the control policies π_2, π_3, and π_4 are computed in a similar fashion yielding: $k(\pi_2) = [k(1,1), k(2,2)]^T = [5.7, 13.2]^T$, $k(\pi_3) = [k(1,2), k(2,1)]^T = [10.7, 10]^T$, and $k(\pi_4) = [k(1,2), k(2,2)]^T = [10.7, 13.2]^T$.

The average cost per unit time as imposed by each control policy is computed by Eq. (3) and are equal to: $J(\pi_1) = \beta(\pi_1) \cdot k(\pi_1) = 7.55$, $J(\pi_2) = \beta(\pi_2) \cdot k(\pi_2) = 10.20$, $J(\pi_3) = \beta(\pi_3) \cdot k(\pi_3) = 10.56$, and $J(\pi_4) = \beta(\pi_4) \cdot k(\pi_4) = 11.53$. The optimal control policy can be derived by Eq. (4); namely, $J^* = \inf\{J(\pi_1), J(\pi_2), J(\pi_3), J(\pi_4) | \pi_1, \pi_2, \pi_3, \pi_4 \in \Pi\} = J(\pi_1)$. That is, the control policy π_1 is the optimal control policy.

C. Equilibrium Control Policy

To demonstrate the equilibrium control policy in this problem we compute both $\sup_{\beta \in \mathcal{B}} \inf_{k \in \mathcal{K}} \beta(\pi) \cdot k(\pi)$ and $\inf_{k \in \mathcal{K}} \sup_{\beta \in \mathcal{B}} \beta(\pi) \cdot k(\pi)$ among the control policies π_1, π_2, π_3, and π_4 for each state. The infimum over the product of the stationary probability distribution $\beta(\pi)$ and cost function $k(\pi)$ is

$$\begin{aligned} \inf_{k \in \mathcal{K}} \beta(\pi) \cdot k(\pi) &= \inf_{k \in \mathcal{K}} \begin{pmatrix} \beta(\pi_1) \cdot k(\pi_1) \\ \beta(\pi_2) \cdot k(\pi_2) \\ \beta(\pi_3) \cdot k(\pi_3) \\ \beta(\pi_4) \cdot k(\pi_4) \end{pmatrix} \\ &= \begin{pmatrix} [0.57, 0.43] \\ [0.40, 0.60] \\ [0.80, 0.20] \\ [0.67, 0.33] \end{pmatrix} \cdot \inf_{k \in \mathcal{K}} \begin{pmatrix} [5.7, 10]^T \\ [5.7, 13.2]^T \\ [10.7, 10]^T \\ [10.7, 13.2]^T \end{pmatrix} \\ &= \begin{pmatrix} [0.57, 0.43] \\ [0.40, 0.60] \\ [0.80, 0.20] \\ [0.67, 0.33] \end{pmatrix} \cdot [5.7, 10]^T. \end{aligned} \tag{20}$$

Taking the supremum over the Eq. (20), we have

$$\begin{aligned} \sup_{\beta \in \mathcal{B}} \inf_{k \in \mathcal{K}} \beta(\pi) \cdot k(\pi) &= \sup_{\beta \in \mathcal{B}} \begin{pmatrix} [0.57, 0.43] \\ [0.40, 0.60] \\ [0.80, 0.20] \\ [0.67, 0.33] \end{pmatrix} \cdot [5.7, 10]^T \\ &= [0.57, 0.43] \cdot [5.7, 10]^T = 7.55. \end{aligned} \tag{21}$$

since the supremum of the probability distribution is the one which is closest to 0.5 in this simple case with two states.

To compute $\inf_{k \in \mathcal{K}} \sup_{\beta \in \mathcal{B}} \beta(\pi) \cdot k(\pi)$ we first compute the supremum of the product of the stationary probability distribution $\beta(\pi)$ and cost function $k(\pi)$ is

$$\begin{aligned} \sup_{\beta \in \mathcal{B}} \beta(\pi) \cdot k(\pi) &= \sup_{\beta \in \mathcal{B}} \begin{pmatrix} \beta(\pi_1) \cdot k(\pi_1) \\ \beta(\pi_2) \cdot k(\pi_2) \\ \beta(\pi_3) \cdot k(\pi_3) \\ \beta(\pi_4) \cdot k(\pi_4) \end{pmatrix} \\ &= \sup_{\beta \in \mathcal{B}} \begin{pmatrix} [0.57, 0.43] \\ [0.40, 0.60] \\ [0.80, 0.20] \\ [0.67, 0.33] \end{pmatrix} \cdot \begin{pmatrix} [5.7, 10]^T \\ [5.7, 13.2]^T \\ [10.7, 10]^T \\ [10.7, 13.2]^T \end{pmatrix} \\ &= [0.57, 0.43] \cdot \begin{pmatrix} [5.7, 10]^T \\ [5.7, 13.2]^T \\ [10.7, 10]^T \\ [10.7, 13.2]^T \end{pmatrix}. \end{aligned} \tag{22}$$

Taking the infimum over the Eq. (22), we have

$$\inf_{k\in\mathcal{K}}\sup_{\beta\in\mathcal{B}}\beta(\pi)\cdot k(\pi) = [0.57, 0.43]\cdot \inf_{k\in\mathcal{K}}\begin{pmatrix}[5.7,10]^T\\ [5.7,13.2]^T\\ [10.7,10]^T\\ [10.7,13.2]^T\end{pmatrix} = [0.57, 0.43]\cdot[5.7,10]^T = 7.55. \tag{23}$$

In this problem, the hypotheses of *Proposition 3.1* hold, and thus, the saddle point exists. By applying the equilibrium control policy we achieve the optimal average cost per unit time.

VI. CONCLUDING REMARKS

The results presented here address the problem of minimizing the average cost per unit time for a controlled Markov chain. The problem is essentially formulated as a dual constrained optimization problem on nonempty convex and compact subsets of $\mathbb{R}^n$. Conceptually, we seek a solution ensuring that the control policy endows a stationary probability distribution yielding higher probability at the states with low cost and lower probability at the states with high cost. The control policy that yields the saddle point solution of this optimization problem is an equilibrium control policy. Recognition of such saddle points may be of value in practical situations with constraints consistent to those studied here when deriving an optimal control policy in real time is required. Solving the original stochastic control problem is computational expensive and real-time implementation may be prohibitive; alternatively, we can design a controller with the aim to achieve higher probability for the states with low cost and lower probability for the states with high cost.

VII. ACKNOWLEDGMENTS

The author gratefully acknowledge the support of the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory, managed by UT-Battelle, LLC, for the U.S. Department of Energy. The author would also like to acknowledge the reviewers' comments as well as Seddik Djouadi and Mohammed Olama for their remarks and suggestions.

REFERENCES

[1] D. Blackwell, "Discrete dynamic programming," *Ann. Math. Statist.*, vol. 33, no. 33, pp. 719–726, 1962.

[2] C. Derman, "On sequential decisions and Markov chains," *Management Sci.*, vol. 9, no. 1, pp. 16–24, 1962.

[3] E. A. Feinberg, "On controlled finite state Markov processes with compact control sets," *Theor. Probab. Appl.*, vol. 20, pp. 856–861, 1975.

[4] A. Arapostathis, V. Borkar, E. Fernandez-Gaucherand, M. K. Ghosh, and S. I. Marcus, "Discrete-time controlled Markov processes with average cost criterion: a survey," *SIAM Journal on Control and Optimization*, vol. 31, no. 2, pp. 282–344, 1993.

[5] R. Y. Chitashvili, "A controlled finite Markov chain with an arbitrary set of decisions," *Theory of Probability and Its Applications*, vol. 20, no. 4, pp. 839–847, 1975.

[6] E. A. Feinberg, "The existence of a stationary ϵ -optimal policy for a fintie Markov chain," *Theory of Probability and Its Applications*, vol. 23, no. 2, pp. 297–313, 1978.

[7] P. Varaiya, "Optimal and suboptimal stationary controls for Markov chains," *IEEE Transactions on Automatic Control*, vol. AC-23, no. 3, pp. 388–394, 1978.

[8] D. P. Bertsekas and S. E. Shreve, *Stochastic Optimal Control: The Discrete-Time Case*, 1st ed. Athena Scientific, February 2007.

[9] H. J. Kushner, *Introduction to Stochastic Control*. Holt, Rinehart and Winston, 1971.

[10] P. R. Kumar and P. Varaiya, *Stochastic systems*. Prentice Hall, June 1986.

[11] R. A. Howard, *Dynamic Programming and Markov Processes*. The MIT Press, June 1960.

[12] J. L. Doob, *Stochastic Processes*. Wiley-Interscience, January 1990.

[13] J. Bather, "Optimal decision procedures for finite Markov chains. I. Examples," *Advances in Applied Probability*, vol. 5, no. 2, pp. 328–339, 1973.

[14] E. A. Feinberg, "Finite state Markov decision models with average reward criteria," *Stochastic Processes and their Applications*, vol. 49, no. 1, pp. 159–177, 1994.

[15] A. Hordjik, "Dynamic Programming and Markov Potential Theory," *Mathematical Centre Tracts*, vol. 51, 1974.

[16] V. S. Borkar, "Controlled Markov chains and stochastic networks," *SIAM Journal on Control and Optimization*, vol. 21, no. 4, pp. 652–665, 1983.

[17] ——, "On minimum cost per unit time control of Markov chains," *SIAM Journal on Control and Optimization*, vol. 22, no. 6, pp. 965–978, 1984.

[18] ——, "Control of Markov chains with long-run average cost criterion," *Stochastic Differential Systems, Stochastic Control Theory and Applications*, pp. 57–77, 1988.

[19] ——, "Control of Markov chains with long-run average cost criterion: the dynamic programming equations," *SIAM Journal on Control and Optimization*, vol. 27, no. 3, pp. 642–657, 1989.

[20] ——, "Controlled Markov chains with constraints," *Sadhana - Academy Proceedings in Engineering Sciences*, vol. 15, no. pt 4-5, p. 405, 1990.

[21] L. I. Sennott, "Another set of conditions for average optimality in Markov control processes," *Systems and Control Letters*, vol. 24, no. 2, pp. 147–151, 1995.

[22] Q. Zhu, X. Guo, and Y. Dai, "Unbounded cost Markov decision processes with limsup and liminf average criteria: new conditions," *Math. Methods Oper. Res.*, vol. 61, no. 3, pp. 469–482, 2005.

[23] R. Cavazos-Cadena, "Value iteration in a class of communicating Markov decision chains with the average cost criterion," *SIAM Journal on Control and Optimization*, vol. 34, no. 6, pp. 1848–73, 1996.

[24] A. Leizarowitz and A. J. Zaslavski, "Uniqueness and stability of optimal policies of finite state Markov decision processes," *Mathematics of Operations Research*, vol. 32, no. 1, pp. 156–167, 2007.

[25] O. Hernandez-Lerma and J. Lasserre, "Weak conditions for average optimality in Markov control processes," *Syst. Control Lett.*, vol. 22, no. 4, pp. 287–91, 1994.

[26] R. Montes-de Oca, "The average cost optimality equation for Markov control processes on Borel spaces," *Syst. Control Lett.*, vol. 22, no. 5, pp. 351–7, 1994.

[27] O. Vega-Amaya and R. Montes-de Oca, "Application of average dynamic programming to inventory systems," *Math. Methods Oper. Res.*, vol. 47, no. 3, pp. 451–71, 1998.

[28] X. Guo and P. Shi, "Limiting average criteria for nonstationary Markov decision processes," *SIAM Journal on Optimization*, vol. 11, no. 4, pp. 1037–1053, 2001.

[29] A. A. Malikopoulos, *Real-Time, Self-Learning Identification and Stochastic Optimal Control of Advanced Powertrain Systems*. Ph.D. Dissertation, Department of Mechanical Engineering, The University of Michigan, Ann Arbor, USA, 2008.

[30] G. R. Grimmett and D. R. Stirzaker, *Probability and Random Processes*, 3rd ed. Oxford University Press, August 2001.

[31] S. M. Ross, *Stochastic Processes*, 2nd ed. Wiley, January 1995.

[32] D. P. Bertsekas, *Convex Optimization Theory*. Athena Scientific, June 2009.

[33] R. B. Ash, *Measure, Integration and Functional Analysis*. Academic Press, 1972.

[34] M. A. Khamsi and W. A. Kirk, *An Introduction to Metric Spaces and Fixed Point Theory*. Wiley-Interscience, 2001.