Convergence Properties of a Computational Learning Model for Unknown Markov Chains

Andreas A. Malikopoulos
Department of Mechanical Engineering,
University of Michigan,
Ann Arbor, MI 48109
e-mail: amaliko@umich.edu

The increasing complexity of engineering systems has motivated continuing research on computational learning methods toward making autonomous intelligent systems that can learn how to improve their performance over time while interacting with their environment. These systems need not only to sense their environment, but also to integrate information from the environment into all decision-makings. The evolution of such systems is modeled as an unknown controlled Markov chain. In a previous research, the predictive optimal decision-making (POD) model was developed, aiming to learn in real time the unknown transition probabilities and associated costs over a varying finite time horizon. In this paper, the convergence of the POD to the stationary distribution of a Markov chain is proven, thus establishing the POD as a robust model for making autonomous intelligent systems. This paper provides the conditions that the POD can be valid, and be an interpretation of its underlying structure. [DOI: 10.1115/1.3117202]

1 Introduction

New technologies in mechatronics and actuators have induced significant enhancement in the complexity of modern engineering systems. The exact modeling of complex systems is often infeasible or expensive, and thus, deriving an optimal control policy can be intractable. This challenge has increased the need to develop computational cognitive models that will allow a system to learn how to improve its performance over time in stochastic environments. Computational intelligence, or rationality, can be achieved by modeling a system and the interaction with its environment through actions, perceptions, and associated costs (or rewards). A widely adopted paradigm for modeling this interaction is the completely observable Markov decision process.

The problem is formulated as sequential decision-making under uncertainty in which an intelligent system (decision maker), e.g., 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. This problem involves two major subproblems: (a) the system identification problem and (b) the stochastic control problem. The first is exploitation of the information acquired from the system output to identify its behavior, that is, how a state representation can be built by observing the system's state transitions. The second is assessment of the system output with respect to alternative control policies, and selecting those that optimize specified performance criteria.

Reinforcement learning (RL) [1,2] has aimed to provide simulation-based algorithms, founded on dynamic programming, for learning control policies of complex systems, where exact modeling is infeasible [3], or the analytic computation may be too high and an approximation method is necessary. Although many of these algorithms are eventually guaranteed to find suboptimal policies, their use of the accumulated data acquired over the learning process is inefficient, and they require a significant amount of experience to achieve good performance [4]. This requirement arises due to the formation of these algorithms in deriving control policies without learning the system dynamics en route; that is, they do not solve the system identification problem simultaneously.

Stochastic adaptive control provides a systematic treatment in deriving optimal control policies in systems where exact modeling is not available a priori. In this context, the evolution of the system is modeled as a countable state controlled Markov chain whose transition probability is specified up to an unknown parameter, taking values in a compact metric space; this problem has been extensively reported in literature. Mandl [5] considered an adaptive control scheme, providing a minimum contrast estimate of the unknown model of a system at each decision epoch, and then applying the optimal feedback control corresponding to this estimate. If the system satisfies a certain "identifiability condition," the sequence of parameter estimates converges almost surely to the true parameter. Borkar and Varaiya [6] removed this identifiability condition and showed that when the feasible space of the unknown parameter is finite, the maximum likelihood estimate of the parameter converges almost surely to a random variable. Borkar and Varaiya [7], and Kumar [8] examined the performance of the adaptive control scheme of Mandl [5] without the identifiability condition, but under varying degrees of generality of the state, control, and model spaces, with the attention restricted to the maximum likelihood estimate. Doshi and Shreve [9] proved that if the set of allowed control laws is generalized to include the set of randomized controls, then the cost of using this scheme will almost surely equal to the optimal cost achievable if the true parameter were known. Kumar and Becker [10] implemented a novel approach to the adaptive control problem when a set of possible models is given including a new criterion for selecting a parameter estimate. This criterion is obtained by a deliberate biasing of the maximum likelihood criterion in favor of parameters with lower optimal costs. These results were extended by assuming that a finite set of possible models is not available [11]. Sato et al. [12–14] proposed a learning controller for Markovian decision problems with unknown probabilities. The controller was designed to be asymptotically optimal, considering a conflict between estimation and control for determination of a control policy over an infinite time horizon. Kumar [15] and Varaiya [16] provided comprehensive surveys of the aforementioned research efforts.

Certainty equivalence control (CEC) is a common approach in addressing stochastic adaptive control problems. The unknown system parameter is estimated at each decision epoch while as-

Contributed by the Dynamic Systems, Measurement, and Control Division of ASME for publication in the JOURNAL OF DYNAMIC SYSTEMS, MEASUREMENT, AND CONTROL. Manuscript received March 18, 2008; final manuscript received February 4, 2009; published online May 20, 2009. Assoc. Editor: Santosh Devasia. Paper Presented at the 2008 Dynamic Systems and Controls Conference (DSC2008), Ann Arbor, MI, October 20–22, 2008.

suming that the decision maker selects a control action, as if the estimated parameter is the true one. The major drawback of this approach is that the decision maker may get locked in a false parameter when there is a conflict between learning and control. Forcing controls, different actions from those imposed by the certainty equivalence control, at some random decision epochs are often utilized to address this issue. The certainty equivalence control employing a forcing strategy is optimal in stochastic adaptive optimization problems with the average-cost-per-unit-time criterion. In these adaptive control schemes, the best possible performance depends on the on-line forcing strategy. Agrawal and Teneketzis [17] studied the rate of forcing to assess the performance of a certainty equivalence control with forcing for the multi-armed bandit problem and the adaptive control of Markov chains. Although the aforementioned research work has successfully led to asymptotically optimal adaptive control schemes when the dynamics of the system are partly known, their underlying framework imposes limitations in implementing such schemes over a varying finite time horizon.

The predictive optimal decision-making (POD) learning model [18,19] aimed to address the state estimation and system identification problem for a completely unknown system by learning in real time the system dynamics over a varying and unknown finite time horizon. It is constituted by a state-space representation that can be used to improve system performance over time in the entire state space. The POD model was employed in various applications toward making autonomous intelligent systems that can learn to improve their performance over time in stochastic environments. In the cart-pole balancing problem [19], an inverted pendulum was made capable of realizing the balancing control policy and turning into a stable system when it was released from any angle between 3 deg and −3 deg. In a vehicle cruise control implementation [19], an autonomous cruise controller was developed to learn to maintain the desired vehicle's speed at any road grade between 0 deg and 10 deg. POD has also taken steps toward development engine calibration that can capture a steady-state and transient engine operation designated by the driver's driving style [20–22]. While the engine runs the vehicle, it progressively perceives the driver's driving style and eventually learns to operate in a manner that optimizes specified performance criteria, e.g., fuel economy, emissions, or engine acceleration.

In this paper, the convergence of POD to the stationary distribution of the Markov state transitions is proven, hence, establishing POD as a robust model. The paper provides the conditions under which POD can be valid (Assumptions 3.1–3.3), and an interpretation of its underlying structure (Lemmas 4.1 and 4.2). This structure, constituting the fundamental aspect of the POD state-space representation, aims to reveal embedded properties in establishing the POD convergence (Theorem 4.1).

The remainder of this paper proceeds as follows: Section 2 presents the steps toward modeling a dynamic system incurring stochastic disturbances as a controlled Markov chain. Section 3 reviews the theory of controlled Markov chains and formulates the POD model by imposing the conditions under which it is valid. The embedded properties of the POD state-space representation and the convergence of the model are proved in Sec. 4. Conclusions are presented in Sec. 5.

2 Modeling Dynamic Systems as a Controlled Markov Chain

The stochastic system model, illustrated in Fig. 1, establishes the mathematical framework for the representation of dynamic systems that evolve stochastically over time [23,24], that is, when incurring a stochastic disturbance or noise at time k, w_k, in their portrayal. The one-dimensional model is given by an equation of the form

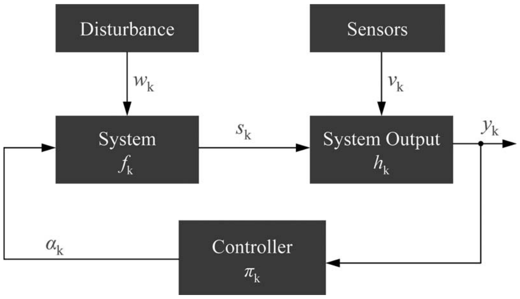

Fig. 1 Stochastic system model schematic

$$s_{k+1}=f_k(s_k,a_k,w_k), \quad k=0,1,\ldots \tag{1}$$

where s_k is the system's state that belongs to some state space $\mathcal{S}=\{1,2,\ldots,N\}$, $N\in\mathbb{N}$, f_k is a function that describes how the system's state is updated, and a_k is the input at time k; a_k represents the control action chosen by the controller from some feasible action set $A(s_k)$, which is a subset of some control space $\mathcal{A}$, namely,

$$\mathcal{A}=\cup_{s_k\in\mathcal{S}}A(s_k) \tag{2}$$

The sequence $\{w_k,k\geq 0\}$ is treated as a stochastic process, and the joint probability distribution of the random variables $w_0,w_1,\ldots,w_k$ is unknown for each k. The system output is represented by

$$y_k=h_k(s_k,v_k), \quad k=0,1,\ldots \tag{3}$$

where y_k is the observation or system's output, h_k is a function that describes how the system output is updated, and v_k is the measurement error or noise. The sequence $\{v_k,k\geq 0\}$ is also considered a stochastic process with unknown probability distribution.

We are interested in deriving a control policy so that a given performance criterion is optimized over all admissible policies π. An admissible policy consists of a sequence of functions

$$\pi=\{\mu_0,\mu_1,\ldots\} \tag{4}$$

where μ_k maps states s_k into actions $a_k=\mu_k(s_k)$, such that $\mu_k(s_k)\in A(s_k)$ and $\forall s_k\in\mathcal{S}$.

The system's state s_k depends on the input sequence $a_0,a_1,\ldots$ as well as the random variables $w_0,w_1,\ldots$, Eq. (1). Consequently, s_k is a random variable; the system output $y_k=h_k(s_k,v_k)$ is a function of the random variables $s_0,s_1,\ldots$ and $v_0,v_1,\ldots$, and thus, is also a random variable. Similarly, the sequence of control actions $a_k=\mu(s_k)$ and $\{a_k,k\geq 0\}$ constitutes a stochastic process.

Definition 2.1. *The random variables* s_0, $w_0,w_1,\ldots$, and $v_0,v_1,\ldots$, *are addressed as basic random variables, since the sequences* $\{s_k,k\geq 0\}$, $\{y_k,k\geq 0\}$, and $\{a_k,k\geq 0\}$ *are constructed from them* [23].

We explore the conditions under which the stochastic system model retains a property in imposing a condition directly on the basic random variables. That is, whether the conditional probability distribution of s_{k+1} given s_k and a_k are independent of previous values of states and control actions. Suppose the control policy $\pi=\{\mu_0,\mu_1,\ldots\}$ is employed. The corresponding stochastic processes $\{s_k^\pi,k\geq 0\}$, $\{y_k^\pi,k\geq 0\}$, and $\{a_k^\pi,k\geq 0\}$, are defined by

$$s_{k+1}^\pi=f_k(s_k^\pi,a_k^\pi,w_k), \quad s_0^\pi=s_0 \tag{5}$$

$$y_k^\pi=h_k(s_k^\pi,v_k) \tag{6}$$

and

$$a_k^{\pi} = \mu_k(s_k^{\pi}) \tag{7}$$

Suppose further that the values realized by the random variables s_k and a_k are known. These values are insufficient to determine the value of s_{k+1} since w_k is not known. The value of s_{k+1} is statistically determined by the conditional distribution of s_{k+1}, given s_k and a_k, namely,

$$\mathbb{P}^{\pi}_{s_{k+1}|s_k,a_k}(\cdot|s_k,a_k) \tag{8}$$

For any occupied state space at time $k+1$, $\mathcal{S}_{k+1}$, and from Eq. (5), we have

$$\mathbb{P}^{\pi}_{s_{k+1}|s_k,a_k}(\mathcal{S}_{k+1}|s_k,a_k) = \mathbb{P}^{\pi}_{w_k|s_k,a_k}(\mathcal{W}_k|s_k,a_k) \tag{9}$$

where $\mathcal{W}_k := \{w|f_k(s_k,a_k,w) \in \mathcal{S}_k\}$ is the disturbance space at time k. The interpretation of Eq. (9) is that the conditional probability of reaching the state space $\mathcal{S}_{k+1}$ at time $k+1$, given s_k and a_k, is equal to the probability of being at the disturbance space $\mathcal{W}_k$ at time k. Suppose that the previous values of the random variables s_m and a_m, $m \leq k-1$ are known. Then, the conditional distribution of s_{k+1} given these values will be

$$\begin{aligned} &\mathbb{P}^{\pi}_{s_{k+1}|s_k,a_k}(\mathcal{S}_{k+1}|s_k, \ldots ,s_0,a_k, \ldots ,a_0) \\ &\quad = \mathbb{P}^{\pi}_{w_k|s_k,a_k}(\mathcal{W}_k|s_{k-1}, \ldots ,s_0,a_{k-1}, \ldots ,a_0) \end{aligned} \tag{10}$$

The conditional probability distribution of S_{k+1}, given s_k and a_k, can be independent of the previous values of states and control actions if it is guaranteed that for every control policy π, $\mathcal{W}_k$ is independent of the random variables s_m and a_m, $m \leq k-1$. Kumar and Varaiya [23] proved that this property is imposed under the following assumption.

Assumption 2.1. *The basic random variables* $s_0,w_0,w_1,\ldots$ and $v_0,v_1,\ldots$ *are all independent.*

Assumption 2.1 imposes a condition directly to the basic random variables, which eventually yields that the state s_{k+1} depends only on s_k and a_k. Moreover, the conditional probability distributions do not depend on the control policy π, and thus, the superscript π can be dropped

$$\mathbb{P}_{s_{k+1}|s_k,a_k}(s_{k+1}|s_k, \ldots ,s_0,a_k, \ldots ,a_0) = \mathbb{P}_{s_{k+1}|s_k,a_k}(s_{k+1}|s_k,a_k) \tag{11}$$

A stochastic process $\{s_k,k \geq 0\}$ satisfying the condition of Eq. (11) is called a "Markov process" and the condition is addressed as a "Markov property."

Definition 2.2. *A Markov process is a random process* $\{s_k,k \geq 0\}$, *with the property that gives the values of the process from time zero up to the current time. The conditional probability of the value of the process at any future time depends only on its value at the current time. That is, the future and past are conditionally independent given the present* [25].

Definition 2.3. *When the state of a Markov process is discrete, then the process is called a Markov chain* [26].

Consequently, under Assumption 2.1, a dynamic system incurring stochastic disturbances can be represented by a controlled Markov chain. A stochastic system is specified by the state equation f_k, $k \geq 0$, the observation equation h_k, $k \geq 0$, and the probability distribution of the basic random variables $s_0,w_0,w_1,\ldots$ and $v_0,v_1,\ldots$. A controlled Markov chain description of a stochastic system is specified by the transition probabilities $\mathbb{P}_{s_{k+1}|s_k,a_k}(\cdot|\cdot)$, the observation equation h_k, $k \geq 0$, and the probability distribution of the independent basic random variables $s_0,v_0,v_1,\ldots$. The observation function and random variables can alternatively be represented by some cost functions $R_k(s_k,a_k)$, corresponding to a system's performance criterion. These functions provide the cost associated with the state being visited by the chain at time k, $s_k = i \in \mathcal{S}$, when the control action a_k is selected.

We consider the problem of deriving an optimal control policy for a completely unknown dynamic system incurring stochastic disturbances by learning the transition probabilities and cost functions. While the system is evolving over time, the goal is to realize a control policy that optimizes a specified performance criterion, assuming the system's performance can be completely measured. The problem is formulated as a sequential decision-making problem under uncertainty. The decision-making process occurs at each sequence of decision epochs $k=0,1,2,\ldots,M$, $M \in \mathbb{N}$. At each epoch, the controller observes a system's state $s_k = i \in \mathcal{S}$, and executes an action $a_k \in A(s_k)$, from the feasible set of actions $A(s_k) \subseteq \mathcal{A}$ at this state. At the next epoch, the system transits to the state $s_{k+1} = j \in \mathcal{S}$ imposed by the conditional probabilities $\mathbb{P}(s_{k+1} = j|s_k = i,a_k)$, designated by the transition probability matrix $\mathbf{P}(\cdot|\cdot)$. The conditional probabilities of $\mathbf{P}(\cdot|\cdot)$, $\mathbb{P}: \mathcal{S} \times \mathcal{A} \rightarrow [0,1]$, satisfy the constraint

$$\sum_{j=1}^{N} \mathbb{P}(s_{k+1} = j|s_k = i,a_k) = 1 \tag{12}$$

Following this state transition, the controller receives a cost associated with the action a_k, $R(s_k = i,a_k)$, $R: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$.

A control policy π determines the probability distribution of the state process $\{s_k,k \geq 0\}$ and the control process $\{a_k,k \geq 0\}$. Different policies will lead to different probability distributions. In optimal control problems, the objective is to derive the optimal control policy that minimizes the accumulated cost incurred at each state transition per decision epoch. If a policy π is fixed, the cost incurred by π when the process starts from an initial state s_0 and up to the time horizon M is

$$J^{\pi}(s_0) = \sum_{k=0}^{M-1} R_k(s_k,a_k), \quad \forall\, s_k \in \mathcal{S}, \quad \forall\, a_k \in A(s_k) \tag{13}$$

The accumulated cost $J_{\pi}(s_0)$ is a random variable since s_k and a_k are random variables. Hence the expected accumulated cost of a control policy is given by

$$J^{\pi}(s_0) = \underset{\substack{s_k \in \mathcal{S} \\ a_k \in A(s_k)}}{E} \left\{ \sum_{k=0}^{M-1} R_k(s_k,a_k) \right\} = \underset{\substack{s_k \in \mathcal{S} \\ \mu_k \in A(s_k)}}{E} \left\{ \sum_{k=0}^{M-1} R_k(s_k,\mu_k(s_k)) \right\} \tag{14}$$

where the expectation is, with respect to the probability distribution of $\{s_k,k \geq 0\}$ and $\{a_k,k \geq 0\}$, determined by the policy π. Consequently, the control policy that minimizes Eq. (14) is defined as the optimal control policy π^*.

3 Finite State Controlled Markov Chains

3.1 Classification of States. The evolution of the system is modeled as a controlled Markov chain with a finite state space $\mathcal{S}$ and control action space $\mathcal{A}$. This evolution $\{s_k,k \geq 0\}$ can be seen as the motion of a notional particle, which jumps between the states $i \in \mathcal{S}$ of the state space $\mathcal{S} = \{1,2,\ldots,N\}$, $N \in \mathbb{N}$, at each decision epoch, while a certain cost incurs at each jumping.

Definition 3.1. *The chain* $\{s_k,k \geq 0\}$ is called homogeneous [27] if

$$\mathbb{P}_{ij}(s_{k+1} = j|s_k = i) = \mathbb{P}_{ij}(s_1 = j|s_0 = i), \quad \forall\, k \geq 0, \forall\, i,j \in \mathcal{S} \tag{15}$$

The classification of the states in a Markov chain aims to provide insight toward modeling appropriately the evolution of a controlled dynamic system [27].

Definition 3.2. *A Markov state* $i \in \mathcal{S}$ is called recurrent (or persistent), if

$$\mathbb{P}(s_k = i) = 1$$

for some

$$\mathbb{P}(k \geq 0|s = i) = 1 \tag{16}$$

that is, the probability of eventually returning to state i, having started from i, is 1 [28].

The first time the chain $\{s_k, k \geq 0\}$ visits a state $i \in \mathcal{S}$ is given by

$$T_1(i) := \min\{k \geq 1 : s_k = i\} \tag{17}$$

$T_1(i)$ is called the "first entrance time" or "first passage time" of state i. It may happen that $s_k \neq i$ for any $k \geq 1$. In this case, $T_1(i) = \min \varnothing$, which is taken to be ∞. Consequently, if the chain $\{s_k\}$ never visits state i for any time $k \geq 1$, $T_1(i) = \infty$. Given that the chain starts in state i, the conditional probability that the chain returns to state i in finite time is

$$f_{ii} := \mathbb{P}(T_1(i) < \infty | s_0 = i) \tag{18}$$

Consequently, for a recurrent state i, $f_{ii} = 1$. Furthermore, if the expected time for the chain to return to a recurrent state i is finite, the state is said to be a positive recurrent; otherwise, the state is said to be a null recurrent. The nth entrance time of state i is given by

$$T_n(i) := \min\{k \geq T_{n-1}(i) : s_k = i\} \tag{19}$$

DEFINITION 3.3. *The mean recurrence time* μ_i of a state i is defined as [28]

$$\mu_i := E\{T_1(i) | s_0 = i\} \tag{20}$$

DEFINITION 3.4. *The period* $d(i)$ of a state i is defined by

$$d(i) := gcd\{n : T_n(i) > 0\} \tag{21}$$

that is, the greatest common divisor of the decision epochs at which return is possible. The state i is periodic if $d(i) > 1$ and aperiodic if $d(i) = 1$ [28].

DEFINITION 3.5. *A Markov state is called ergodic, if it is a positive recurrent and aperiodic* [27].

DEFINITION 3.6. *If the chain started from state* i and visits state j, that is, $\mathbb{P}^{(n)}_{ij}(s_n = j | s_0 = i) > 0$ for some $n > 0$, it is said that i communicates with j, and it is denoted $i \to j$. It is said that i and j intercommunicate if $i \to j$ and $j \to i$, and it is denoted $i \leftrightarrow j$ [28].

DEFINITION 3.7. *A Markov chain is called irreducible if all states intercommunicate in a finite number of decision epochs, that is,* $\mathbb{P}^{(n)}_{ij}(s_n = j | s_0 = i) > 0$, $\forall i, j \in \mathcal{S}$ [26].

The behavior of a Markov chain after a long time k has elapsed is described by the stationary distributions and the limit theorem. The sequence $\{s_k, k \geq 0\}$ does not converge to some particular state $i \in \mathcal{S}$ since it enjoys the inherent random fluctuation, which is specified by the transition probability matrix. Subject to certain conditions, the distribution of $\{s_k, k \geq 0\}$ settles down to a stationary one; that is, the evolution of the Markov chain will be visiting each state with a constant probability in long term.

DEFINITION 3.8. *The vector* $\boldsymbol{\rho}$ is called a stationary distribution of the chain if $\boldsymbol{\rho}$ has entries $(\rho_i, i \in \mathcal{S})$ such that [28]

(a) $\rho_i \geq 0$ for all i, and $\Sigma_{i \in \mathcal{S}} \rho_i = 1$

(b) $\boldsymbol{\rho} = \boldsymbol{\rho} \cdot \mathbf{P}$, that is, $\rho_i = \sum_{j \in S} \rho_j \cdot \mathbb{P}_{ji}$, where $\mathbb{P}_{ji}$ is the transition probability $\mathbb{P}_{ji}(s_{k+1} = i | s_k = j)$, for all i

If the transition probability matrix of a Markov chain $\mathbf{P}(\cdot|\cdot)$ is raised to a higher power, the resulting matrix is also a transition probability matrix, and the elements in any given column start converging to the same number [29]. This property can be illustrated further in the following simple example. Let us consider a Markov chain with two states $\mathcal{S} = \{1, 2\}$, and a transition probability matrix

$$\mathbf{P} = \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix} \tag{22}$$

This matrix represents the one-step transition probabilities of the states. Consequently, if the chain is at state 1, there is a probability of 0.7 that it will remain there and 0.3 that it will transit to state 2. Similarly, if the chain is at state 2, there is a probability of 0.4 that it will transit to state 1 and 0.6 that it will remain at state 2. If this matrix is raised to the second order, the resulting matrix yields the two-step transition probabilities

$$\mathbf{P}^2 = \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix} \cdot \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix} = \begin{bmatrix} 0.61 & 0.39 \\ 0.52 & 0.48 \end{bmatrix} \tag{23}$$

The elements of the two-step transition probability matrix essentially return the conditional probability that the chain will transit to a particular state within two decision epochs. Consequently, the value $\mathbb{P}^2_{12}(s_{k+1} = 2 | s_k = 1) = 0.39$ in the above matrix is the conditional probability that the chain will go from state 1 to state 2 in two decision epochs. If the one-step transition probability matrix is raised to the eighth power, it is noticed that the elements in any given column start converging to 0.57 and 0.43, respectively, namely,

$$\mathbf{P}^8 = \begin{bmatrix} 0.5715 & 0.4285 \\ 0.5714 & 0.4286 \end{bmatrix} \tag{24}$$

These numbers constitute the stationary distribution of the chain, vector $\boldsymbol{\rho}$, that is,

$$\boldsymbol{\rho} = \begin{bmatrix} \rho_1 \\ \rho_2 \end{bmatrix} = \begin{bmatrix} 0.57 \\ 0.43 \end{bmatrix} \tag{25}$$

The limit theorem states that if a chain is irreducible with positive recurrent states, the following limit exists:

$$\rho_j = \lim_{n \to \infty} \mathbb{P}^n_{ij}(s_{k+1} = j | s_k = i) = \mathbb{P}(s_n = j) \tag{26}$$

THEOREM 3.1 ("LIMIT THEOREM"). *An irreducible Markov chain has a stationary distribution* $\boldsymbol{\rho}$ if and only if all the states are positive recurrent. Furthermore, $\boldsymbol{\rho}$ is the unique stationary distribution and is given by $\rho_i = \mu_i^{-1}$ for each $i \in \mathcal{S}$, where μ_i is the mean recurrence time of state i [28].

Stationary distributions have the following property:

$$\boldsymbol{\rho} = \boldsymbol{\rho} \cdot \mathbf{P}^n, \quad \forall\, n \geq 0 \tag{27}$$

The accumulated cost $J_\pi(s_0)$, Eq. (14), can be readily evaluated in terms of the stationary probability distributions as follows:

$$J^\pi(s_0) = \sum_{k=0}^{M} \rho_i \cdot R_k(s_k = i, a_k)$$

$$\forall\, i \in \mathcal{S}, \quad \forall\, a_k \in A(s_k) \tag{28}$$

where ρ_i is the stationary probability of the visiting state i.

3.2 Formulation of the Predictive Optimal Decision-Making Model.

The POD learning model consists of a new state-space system representation. This representation accumulates gradually enhanced knowledge of the system's transition from each state to another in conjunction with actions taken for each state. While the system interacts with its environment, the POD model learns the transition probabilities of the Markov state transitions and associated cost functions. This realization determines the stationary distribution of the Markov chain that can then be used in deriving the optimal control policy through Eq. (28).

The model considers systems that their evolution can be modeled as a controlled Markov chain under the following assumptions:

ASSUMPTION 3.1. *The Markov chain is homogeneous.*

ASSUMPTION 3.2. *The Markov chain is ergodic, that is, the states are positive recurrent and aperiodic.*

ASSUMPTION 3.3. *The Markov chain is irreducible. Consequently, each state i* of the Markov chain intercommunicates with each other $i \leftrightarrow j$, $\forall i$, and $j \in \mathcal{S}$, that is, each system's state can be reached with a positive probability from any other state in finite decision epochs.

The new state-space representation defines the POD domain $\tilde{\mathcal{S}}$, which is implemented by a mapping $\mathcal{H}$ from the Cartesian product of the finite state space and action space of the Markov chain $\{s_k, k \geq 0\}$

$$\mathcal{H}: \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \tilde{\mathcal{S}} \tag{29}$$

where $\mathcal{S}=\{1,2,\ldots,N\}$, $N \in \mathbb{N}$ denotes the Markov state space, and $\mathcal{A}=\cup_{s_k \in \mathcal{S}} A(s_k)$, $\forall s_k=i \in \mathcal{S}$ stands for the finite action space. Each state of the POD domain represents a Markov state transition from $s_k=i \in \mathcal{S}$ to $s_{k+1}=j \in \mathcal{S}$ for all $k \geq 0$, that is,

$$\tilde{\mathcal{S}} := \left\{ \tilde{s}_{k+1}^{ij} \middle| \tilde{s}_{k+1}^{ij} \equiv s_k = i \underset{\mu(s_k) \in A(s_k)}{\longrightarrow} s_{k+1} = j, \sum_{j=1}^{N} p(s_{k+1} = j | s_k = i, a_k) \right.$$
$$\left. = 1, N = |\mathcal{S}| \right\}$$
$$\forall i,j \in \mathcal{S}, \forall\, \mu(s_k) \in A(s_k) \tag{30}$$

Definition 3.9. *The mapping* $\mathcal{H}$ generates an indexed family of subsets, $\tilde{\mathcal{S}}_i$, for each Markov state $s_k=i \in \mathcal{S}$, defined as predictive representation nodes (PRNs). Each PRN is constituted by the set of POD states $\tilde{s}_{k+1}^{ij} \in \tilde{\mathcal{S}}_i$ representing the state transitions from the state $s_k=i \in \mathcal{S}$ to all other Markov states

$$\tilde{\mathcal{S}}_i := \{ \tilde{s}_{k+1}^{ij} | s_k = i \underset{\mu(s_k) \in A(s_k)}{\longrightarrow} s_{k+1} = j, \forall\, j \in \mathcal{S} \} \tag{31}$$

PRNs partition the POD domain insofar as the POD underlying structure captures the state transitions in the Markov domain, namely,

$$\tilde{\mathcal{S}} = \cup_{\tilde{s}_k^{ij} \in \tilde{\mathcal{S}}_i} \tilde{\mathcal{S}}_i \tag{32}$$

with

$$\cap_{\tilde{s}_k^{ij} \in \tilde{\mathcal{S}}_i} \tilde{\mathcal{S}}_i = \emptyset\ . \tag{33}$$

PRNs, constituting the fundamental aspect of the POD state representation, provide an assessment of the Markov state transitions along with the actions executed at each state. This assessment aims to establish a necessary embedded property of the new state representation so as to consider the stationary distribution in long term.

4 Convergence of POD Model

While the system interacts with its environment, the POD model learns the system dynamics in terms of the Markov state transitions. The POD state representation attempts to provide a process in realizing the sequences of state transitions that occurred in the Markov domain, as infused in PRNs. The different sequences of the Markov state transitions are captured by the POD states. We show that this realization determines the stationary distribution of the Markov chain.

Definition 4.1. *Given a set* $C \subset \mathbb{R}$ and a variable x, the indicator function, denoted by $I_C(x)$, is defined by

$$I_C(x) := \begin{cases} 1, x \in C \\ 0, x \notin C \end{cases} \tag{34}$$

Lemma 4.1. *Each PRN is irreducible, that is,* $\tilde{\mathcal{S}}_i \leftrightarrow \tilde{\mathcal{S}}_j$, $\forall i$, $j \in \mathcal{S}$.

Proof. At the decision epoch k, the state transition from i to j corresponds to the $\tilde{s}_k^{ij}$ inside the PRN $\overline{\mathcal{S}}_i$. The next state transition will occur from the state j to any other Markov state. Consequently, by Definition 3.9, the next state transition will occur in $\overline{\mathcal{S}}_j$. By Assumption 3.3, all states intercommunicate with each other, that is, $i \leftrightarrow j$, $\forall i$, $j \in \mathcal{S}$. So PRNs intercommunicate and thus they are irreducible. The lemma is proved. □

The number of visits of the chain to the state $j \in \mathcal{S}$ between two successive visits to state $i \in \mathcal{S}$ at the decision epoch $k=M$, that is, the number of visits of the POD state $\tilde{s}_M^{ij} \in \tilde{\mathcal{S}}$, is given by

$$V(\tilde{s}_M^{ij}) := \sum_{k=1}^{M} I_{\{s_k=j\} \cap \{T_1(i) \geq k\}}(s_k) \tag{35}$$

where $T_1(i)$ is the time of the first return to state $i \in \mathcal{S}$.

Definition 4.2. *The mean number of visits of the chain to the state* $j \in \mathcal{S}$ between two successive visits to state $i \in \mathcal{S}$ is

$$\bar{V}(\tilde{s}_M^{ij}) := E\{V(\tilde{s}_M^{ij}) | s_k = i\}$$

or

$$\bar{V}(\tilde{s}_M^{ij}) := \sum_{k=1}^{M} \mathbb{P}(s_k = j, T_1(i) \geq k | s_0 = i) \tag{36}$$

Definition 4.3. *The mean recurrence time* $\mu_{\tilde{\mathcal{S}}_i}$ that the chain spends at the PRN $\tilde{\mathcal{S}}_i$ is

$$\mu_{\tilde{\mathcal{S}}_i} := \sum_{j \in S} \bar{V}(\tilde{s}_M^{ij}) = \sum_{j \in S} \sum_{k=1}^{M} \mathbb{P}(s_k = j, T_1(i) \geq k | s_0 = i) \tag{37}$$

Lemma 4.2. *The mean recurrence time of each PRN* $\tilde{\mathcal{S}}_i$, $\mu_{\tilde{\mathcal{S}}_i}$, *is equal to the mean recurrence time of state* $i \in \mathcal{S}$, μ_i.

Proof. It was shown (Lemma 3.1) that each time the Markov chain transits from one state $i \in \mathcal{S}$ to a state $j \in \mathcal{S}$, there is a corresponding transition from the PRN $\overline{\mathcal{S}}_i$ to $\overline{\mathcal{S}}_j$. Consequently, the number of visits of the chain to the state $i \in \mathcal{S}$ is equal to the number of visits to the PRN $\overline{\mathcal{S}}_i$. Taken the expectation of this number yields the mean recurrence time, by Definition 4.3. The lemma is proved. □

Proposition 4.1. *If A, B, and C are some events and*

$$\mathbb{P}(A | B \cap C) = \mathbb{P}(A | B) \tag{38}$$

then

$$\mathbb{P}(A \cap C | B) = \mathbb{P}(A | B) \cdot \mathbb{P}(C | B) \tag{39}$$

Proof.

$$\mathbb{P}(A \cap C | B) = \frac{\mathbb{P}(A \cap B \cap C)}{\mathbb{P}(B)} \tag{40}$$

using the identity $\mathbb{P}(A|B) \cdot \mathbb{P}(B) = \mathbb{P}(A \cap B)$, Eq. (40) yields

$$\frac{\mathbb{P}(A | C \cap B) \cdot \mathbb{P}(C \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A|B) \cdot \mathbb{P}(C \cap B)}{\mathbb{P}(B)}$$

by using Eq. (38)

$$= \frac{\mathbb{P}(A|B) \cdot \mathbb{P}(C \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A|B) \cdot \mathbb{P}(C|B) \cdot \mathbb{P}(B)}{\mathbb{P}(B)} = \mathbb{P}(A|B) \cdot \mathbb{P}(C|B)$$

□

It remains to present the main result of the POD learning model, namely, that the realization of the sequences of state transitions that occurred in the Markov domain as infused by the PRNs determines the stationary distribution of the Markov chain.

Theorem 4.1. *The POD state representation generates the stationary distribution* $\boldsymbol{\rho}$ of the Markov chain. Moreover, the stationary probability is given by the mean recurrence time of each PRN $\tilde{\mathcal{S}}_i$, $\rho_i = \mu_{\tilde{\mathcal{S}}_i}^{-1}$.

Proof. Since the chain is ergodic with irreducible states, it is guaranteed that the chain has a unique stationary distribution, and for each state $i \in \mathcal{S}$ the stationary probability is equal to $\rho_i = \mu_i^{-1}$ (Theorem 3.1)

$$\rho_i \cdot \mu_i =$$

$$= \rho_i \cdot \mu_{\tilde{S}_i}$$

by Lemma 4.2

$$= \sum_{j \in S} \sum_{k=1}^{M} \mathbb{P}(s_k = j, T_1(i) \geq k | s_0 = i) \cdot \mathbb{P}(s_0 = i) \tag{41}$$

$$= \sum_{j \in S} \sum_{k=1}^{M} \mathbb{P}(s_k = j, T_1(i) \geq k, s_0 = i) \tag{42}$$

by using the identity $\mathbb{P}(A|B) \cdot \mathbb{P}(B) = \mathbb{P}(A \cap B)$.

For $k=1$, Eq. (42) yields

$$\sum_{j \in S} \mathbb{P}(s_k = j, T_1(i) \geq 1, s_0 = i) = 1 \tag{43}$$

For $k \geq 2$, Eq. (41) yields

$$\sum_{j \in S} \sum_{k=1}^{M} \mathbb{P}(s_k = j, T_1(i) \geq k | s_0 = i) \cdot \mathbb{P}(s_0 = i)$$

$$= \sum_{j \in S} \sum_{k=1}^{M} \mathbb{P}(s_k = j, s_m \neq i \text{ for } 1 \leq m \leq k-1, s_0 = i) \tag{44}$$

Using Proposition 4.1 and since $\mathbb{P}(s_k = j | s_m \neq i \text{ for } 1 \leq m \leq k-1, s_0 = i) = \mathbb{P}(s_k = j | s_0 = i)$, Eq. (44) becomes

$$\begin{aligned}
&\sum_{j \in S} \sum_{k=1}^{M} \mathbb{P}(s_k = j | s_0 = i) \cdot \mathbb{P}(s_m \neq i \text{ for } 1 \leq m \leq k-1 | s_0 = i) \cdot \mathbb{P}(s_0 \\
&= i) = \sum_{j \in S} \sum_{k=1}^{M} \mathbb{P}(s_k = j | s_0 = i) \cdot \mathbb{P}(s_m \neq i \text{ for } 1 \leq m \leq k \\
&- 1, s_0 = i) = \sum_{k=1}^{M} \left(\sum_{j \in S} \mathbb{P}(s_k = j | s_0 = i) \right) \cdot \mathbb{P}(s_m \\
&\neq i \text{ for } 1 \leq m \leq k-1, s_0 = i) = \sum_{k=1}^{M} \mathbb{P}(s_m \neq i \text{ for } 1 \\
&\leq m \leq k-1, s_0 = i)
\end{aligned} \tag{45}$$

by using the identity $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$, Eq. (45) becomes

$$\begin{aligned}
&\sum_{k=1}^{M} \mathbb{P}(s_0 = i) + \mathbb{P}(s_m \neq i \text{ for } 1 \leq m \leq k-1) - \mathbb{P}(s_m \neq i \text{ for } 0 \\
&\leq m \leq k-1)
\end{aligned}$$

Since the Markov chain is homogeneous (Assumption 3.1)

$$\begin{aligned}
&= \sum_{k=1}^{M} \{\mathbb{P}(s_0 = i) + \mathbb{P}(s_0 \neq i) + \mathbb{P}(s_m \neq i \text{ for } 0 \leq m \leq k-3) - \mathbb{P}(s_m \\
&\neq i \text{ for } 0 \leq m \leq k-1)\}
\end{aligned}$$

$$\begin{aligned}
&= \sum_{k=1}^{M} \{\mathbb{P}(s_0 = i) + \mathbb{P}(s_0 \neq i)\} + \lim_{k \to \infty} (\mathbb{P}(s_m \neq i \text{ for } 0 \leq m \leq k-3)) \\
&- \lim_{k \to \infty} (\mathbb{P}(s_m \neq i \text{ for } 0 \leq m \leq k-1))
\end{aligned} \tag{46}$$

since the Markov states are irreducible (Assumption 3.3)

$$\lim_{k \to \infty} (\mathbb{P}(s_m \neq i \text{ for } 0 \leq m \leq k-3)) = 0$$

and

$$\lim_{k \to \infty} (\mathbb{P}(s_m \neq i \text{ for } 0 \leq m \leq k-1)) = 0$$

Equation (46) becomes

$$\sum_{k=1}^{M} \{\mathbb{P}(s_0 = i) + \mathbb{P}(s_0 \neq i)\} = \sum_{k=1}^{M} \{1\} = 1$$

We have shown that

$$\rho_i \cdot \mu_i = \rho_i \cdot \mu_{\tilde{S}_i} = 1$$

Consequently, the stationary distribution is given by the mean recurrence time of each PRN $\tilde{\mathcal{S}}_i$, $\mu_{\tilde{S}_i}$

$$\rho_i = \frac{1}{\mu_{\tilde{S}_i}} \tag{47}$$

□

5 Concluding Remarks

The POD model aimed to address the state estimation and system identification problem for a completely unknown system by learning in real time the system dynamics when the system's performance can be measured. The model possesses a structure that enables a convergent behavior of the conditional probabilities infused by the POD state-space representation to the stationary distribution. This behavior is desirable in the effort toward making autonomous intelligent systems that can learn to improve their performance over time in stochastic environments. The implementation of the POD model along with a lookahead control algorithm in various applications to date cited in the Introduction support these theoretical results.

The major advantage of the POD model, compared with the stochastic adaptive control approaches, is that it can solve the state estimation and system identification problem over a varying and unknown finite time horizon. This property arises due to the structure of the POD model in addressing the system identification problem separately from the stochastic one. Under the assumption that the basic random variables are all independent, the transition probabilities do not depend on the control policy. Consequently, system identification can be independent of the control policy imposed by the controller, and be addressed separately.

Acknowledgment

This research was partially supported by the Automotive Research Center (ARC), a U.S. Army Center of Excellence in modeling and simulation of ground vehicles at the University of Michigan. This support is gratefully acknowledged.

References

[1] Bertsekas, D. P., and Tsitsiklis, J. N., 1996, *Neuro-Dynamic Programming* (Optimization and Neural Computation Series Vol. 3), 1st ed., Athena Scientific, Nashua, NH.

[2] Sutton, R. S., and Barto, A. G., 1998, *Reinforcement Learning: An Introduction (Adaptive Computation and Machine Learning)*, MIT, Cambridge, MA.

[3] Borkar, V. S., 2000, "A Learning Algorithm for Discrete-Time Stochastic Control," Probability in the Engineering and Informational Sciences, **14**, pp. 243–258.

[4] Kaelbling, L. P., Littman, M. L., and Moore, A. W., 1996, "Reinforcement Learning: A Survey," J. Artif. Intell. Res., **4**, pp. 237–285.

[5] Mandl, P., 1974, "Estimation and Control in Markov Chains," Adv. Appl. Probab., **6**, pp. 40–60.

[6] Borkar, V., and Varaiya, P., 1979, "Adaptive Control of Markov Chains. I. Finite Parameter Set," IEEE Trans. Autom. Control, **AC-24**, pp. 953–957.

[7] Borkar, V., and Varaiya, P., 1982, "Identification and Adaptive Control of Markov Chains," SIAM J. Control Optim., **20**, pp. 470–489.

[8] Kumar, P. R., 1982, "Adaptive Control With a Compact Parameter Set," SIAM J. Control Optim., **20**, pp. 9–13.

[9] Doshi, B., and Shreve, S. E., 1980, "Strong Consistency of a Modified Maximum Likelihood Estimator for Controlled Markov Chains," J. Appl. Probab., **17**, pp. 726–734.

[10] Kumar, P. R., and Becker, A., 1982, "A New Family of Optimal Adaptive Controllers for Markov Chains," IEEE Trans. Autom. Control, **AC-27**, pp. 137–146.

[11] Kumar, P. R., and Lin, W., 1982, “Optimal Adaptive Controllers for Unknown Markov Chains,” IEEE Trans. Autom. Control, **AC-27**, pp. 765–774.

[12] Sato, M., Abe, K., and Takeda, H., 1982, “Learning Control of Finite Markov Chains With Unknown Transition Probabilities,” IEEE Trans. Autom. Control, **AC-27**, pp. 502–505.

[13] Sato, M., Abe, K., and Takeda, H., 1985, “An Asymptotically Optimal Learning Controller for Finite Markov Chains With Unknown Transition Probabilities,” IEEE Trans. Autom. Control, **AC-30**, pp. 1147–1149.

[14] Sato, M., Abe, K., and Takeda, H., 1988, “Learning Control of Finite Markov Chains With an Explicit Trade-Off Between Estimation and Control,” IEEE Trans. Syst. Man Cybern., **18**, pp. 677–684.

[15] Kumar, P. R., 1985, “A Survey of Some Results in Stochastic Adaptive Control,” SIAM J. Control Optim., **23**, pp. 329–380.

[16] Varaiya, P., 1982, “Adaptive Control of Markov Chains: A Survey,” Proceedings of the IFAC Symposium, New Delhi, India, pp. 89–93.

[17] Agrawal, R., and Teneketzis, D., 1989, “Certainty Equivalence Control With Forcing: Revisited,” Proceedings of the IEEE Conference on Decision and Control Including the Symposium on Adaptive Processes, Tampa, FL, p. 2107.

[18] Malikopoulos, A. A., 2008, “Real-Time, Self-Learning Identification and Stochastic Optimal Control of Advanced Powertrain Systems,” Ph.D. thesis, Department of Mechanical Engineering, University of Michigan, Ann Arbor, MI.

[19] Malikopoulos, A. A., Papalambros, P. Y., and Assanis, D. N., 2007, “A State-Space Representation Model and Learning Algorithm for Real-Time Decision-Making Under Uncertainty,” Proceedings of the 2007 ASME International Mechanical Engineering Congress and Exposition, Seattle, WA, Nov. 11–15.

[20] Malikopoulos, A. A., Papalambros, P. Y., and Assanis, D. N., 2007, “A Learning Algorithm for Optimal Internal Combustion Engine Calibration in Real Time,” Proceedings of the ASME 2007 International Design Engineering Technical Conferences Computers and Information in Engineering Conference, Las Vegas, NV, Sept. 4–7.

[21] Malikopoulos, A. A., Assanis, D. N., and Papalambros, P. Y., 2007, “Real-Time, Self-Learning Optimization of Diesel Engine Calibration,” Proceedings of the 2007 Fall Technical Conference of the ASME Internal Combustion Engine Division, Charleston, SC, Oct. 14–17.

[22] Malikopoulos, A. A., Assanis, D. N., and Papalambros, P. Y., 2008, “Optimal Engine Calibration for Individual Driving Styles,” Proceedings of the SAE 2008 World Congress and Exhibition, Detroit, MI, Apr. 14–17, SAE Paper No. 2008-01-1367.

[23] Kumar, P. R., and Varaiya, P., 1986, *Stochastic Systems*, Prentice-Hall, Englewood Cliffs, NJ.

[24] Bertsekas, D. P., 2001, *Dynamic Programming and Optimal Control (Volumes 1 and 2)* (Optimization and Neural Computation Series), 1st ed., Athena Scientific, Nashua, NH.

[25] Kemeny, J. G., and Snell, J. L., 1983, *Finite Markov Chains*, 1st ed., Springer, New York.

[26] Krishnan, V., 2006, *Probability and Random Processes*, 1st ed., Wiley, New York.

[27] Gubner, J. A., 2006, Probability and Random Processes for Electrical and Computer Engineers, 1st ed., Cambridge University Press, Cambridge.

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

[29] Gosavi, A., 2003, *Simulation-Based Optimization: Parametric Optimization Techniques and Reinforcement Learning*, 1st ed., Springer, New York.