**Proceedings of DSCC2008**
**2008 ASME Dynamic Systems and Control Conference**
**October 20-22, 2008, Ann Arbor, Michigan, USA**

**DSCC2008-2174**

# CONVERGENCE PROPERTIES OF A COMPUTATIONAL LEARNING MODEL FOR UNKNOWN MARKOV CHAINS

**Andreas A. Malikopoulos***
Department of Mechanical Engineering
University of Michigan
Ann Arbor, MI 48109-2133
amaliko@umich.edu

## ABSTRACT

*The increasing complexity of engineering systems 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. These systems need not only to be able to sense their environment, but should also integrate information from the environment into all decision making. The evolution of such systems is modeled as an unknown controlled Markov chain. In previous research, the predictive optimal decision-making (POD) model was developed that aims to learn in real time the unknown transition probabilities and associated costs over a varying finite time horizon. In this paper, the convergence of POD to the stationary distribution of a Markov chain is proven, thus establishing POD as a robust model for making autonomous intelligent systems. The paper provides the conditions that POD can be valid, and an interpretation of its underlying structure.*

## 1. 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 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 uncertainly 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 sub-problems: (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 sub-optimal 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

* Currently appointed as a Researcher in the Powertrain Systems Research Lab (PSR) at the General Motors Research & Development Center, Phone: (586) 986-4564, Fax: (586) 986-0176, e-mail: andreas.malikopoulos@gm.com

 

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 the 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 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, Abe, and Takeda [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] have 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 assuming 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. 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 [17, 18] was 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 has been employed in various applications towards making autonomous intelligent systems that can learn to improve their performance over time in stochastic environments. In the cart-pole balancing problem [18], 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° and -3°. In a vehicle cruise control implementation [18], an autonomous cruise controller was developed to learn to maintain the desired vehicle's speed at any road grade between 0° and 10°. POD has also taken steps toward development engine calibration that can capture steady-state and transient engine operation designated by the driver's driving style [19-21]. While the engine is running 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 the paper proceeds as follows: Section 2 presents the steps towards 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 POD state-space representation and the convergence of the model are proved in Section 4. Conclusions are presented in Section 5.

## 2. MODELING DYNAMIC SYSTEMS AS A CONTROLLED MARKOV CHAIN

The stochastic system model, illustrated in Figure 1, establishes the mathematical framework for the representation of dynamic systems that evolve stochastically over time [22, 23], 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

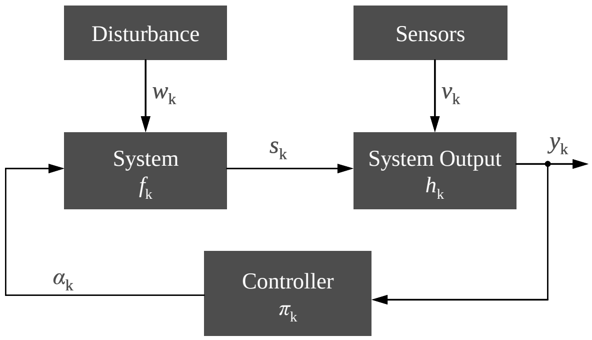


FIGURE 1. STOCHASTIC SYSTEM MODEL SCHEMATIC.

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

where $s_k$ is system's state that belongs to some state space $\mathcal{S} = \{1, 2, ..., N\}$, $N \in \mathbb{N}$, $f_k$ 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,

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

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

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

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 $\Pi$. An admissible policy consists of a sequence of functions

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

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

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

*Definition 2.1* [22]: The random variables $s_0, w_0, w_1, ...,$ $v_0, v_1, ...,$ 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.

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 $a_k$ is independent of previous values of states and control actions. Suppose the control policy $\pi = \{\mu_0, \mu_1, ...\}$ 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

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

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

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

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,

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

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

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

where $\mathcal{W}_k := \{w \mid f_k(s_k, a_k, w) \in \mathcal{S}_k\}$ 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

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

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 $\pi$, $w_k$ is independent of the random variables $s_m$ and $a_m$, $m \leq k-1$. Kumar and Varaiya [22] proved that this property is imposed under the following assumption.

*Assumption 2.1*: The basic random variables $s_0, w_0, w_1, ...,$ $v_0, v_1, ...,$ 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 $\pi$, and thus, the superscript $\pi$ can be dropped

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

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*.

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 $s_0, w_0, w_1, ..., v_0, v_1, ...$ 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 $s_0, v_0, v_1, ...$ 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 of a sequence of decision epochs $k = 0,1,2,...,M$, $M \in \mathbb{N}$. 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 $\mathbb{P}(s_{k+1} = j \mid s_k = i, a_k)$, designated by the transition probability matrix $\mathbf{P}(\cdot|\cdot)$. The conditional probabilities of $\mathbf{P}(\cdot|\cdot)$, $\mathbb{P}: \mathcal{S} \times \mathcal{A} \to [0,1]$, satisfy the constraint

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

Following this state transition, the controller receives a cost associated with the action $a_k$, $R(s_k = i, a_k)$, $R: \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. A control policy $\pi$ determines the probability distribution of 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 $\pi$ is fixed, the cost incurred by $\pi$ when the process starts from an initial state $s_0$ and up to the time horizon *M* is

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

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

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

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 $\pi$. Consequently, the control policy that minimizes Eq.(14) is defined as the optimal control policy $\pi^*$.

## 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 $\mathcal{S} = \{1, 2, ..., N\}$, $N \in \mathbb{N}$, at each decision epoch, while a certain cost incurs at each jumping.

*Definition 3.1* [24]: The chain $\{s_k, k \geq 0\}$ is called homogeneous if

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

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

*Definition 3.2* [25]: A Markov state $i \in \mathcal{S}$ is called recurrent (or persistent), if

$$\mathbb{P}(s_k = i \text{ for some } k \geq 0 \mid s_0 = i) = 1, \tag{16}$$

that is, the probability of eventually return to state $i$, having started from $i$, is one.

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

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

$T_1(i)$ 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 $\infty$. 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

$$f_{ii} := \mathbb{P}(T_1(i) < \infty \mid s_0 = i). \quad (18)$$

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 positive recurrent; otherwise, the state is said to be null recurrent. The *n*th entrance time of state $i$ is given by

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

*Definition 3.3* [25]: The mean recurrence time $\mu_i$ of a state $i$ is defined as

$$\mu_i := E\{T_1(i) \mid s_0 = i\}. \quad (20)$$

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 3.4* [25]: 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:

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

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

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

$$\rho_j = \lim_{n \to \infty} \mathbb{P}_{ij}^n (s_{k+1} = j \mid s_k = i) = \mathbb{P}(s_n = j). \quad (21)$$

*Theorem 3.1 ("Limit Theorem")* [25]: 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 $\mu_i$ is the mean recurrence time of state $i$.

Stationary distributions have the following property

$$\boldsymbol{\rho} = \boldsymbol{\rho} \cdot \mathbf{P}^n, \forall n \geq 0 \quad (22)$$

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

$$J^\pi(s_0) = \sum_{k=0}^{M} \rho_i \cdot R_k(s_k = i, a_k), \ \forall i \in \mathcal{S}, \forall a_k \in A(s_k). \quad (23)$$

where $\rho_i$ is the stationary probability of visiting the state $i$.

### 3.2 Formulation of the Predictive Optimal Decision-Making Model

The POD learning model [17, 18] 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 than can be then used in deriving the optimal control policy through Eq. (24).

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 $i \leftrightarrow j, \forall i, j \in \mathcal{S}$, 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 $\{s_k, k \geq 0\}$

$$\mathcal{H} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \tilde{\mathcal{S}}, \quad (24)$$

where $\mathcal{S} = \{1, 2, ..., N\}$, $N \in \mathbb{N}$ denotes the Markov state space, and $\mathcal{A} = \bigcup_{s_k \in \mathcal{S}} A(s_k), \forall s_k = i \in \mathcal{S}$ 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

$$\tilde{\mathcal{S}} := \left\{ \tilde{s}_{k+1}^{ij} \mid \tilde{s}_{k+1}^{ij} \equiv s_k = i \underset{\mu(s_k) \in A(s_k)}{\rightarrow} s_{k+1} = j \right\}, \quad (25)$$

where
$\sum_{j=1}^{N} p(s_{k+1} = j \mid s_k = i, a_k) = 1, N = |\mathcal{S}|, \ \forall i, j \in \mathcal{S}, \ \forall \mu(s_k) \in A(s_k).$

*Definition 3.5*: 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

$$\tilde{\mathcal{S}}_i := \left\{ \tilde{s}_{k+1}^{ij} \mid s_k = i \underset{\mu(s_k)\in A(s_k)}{\rightarrow} s_{k+1} = j, \forall j \in \mathcal{S} \right\}. \quad (26)$$

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

$$\tilde{\mathcal{S}} = \bigcup_{\tilde{s}_k^{ij} \in \tilde{\mathcal{S}}_i} \tilde{\mathcal{S}}_i, \text{ with} \quad (27)$$

$$\bigcap_{\tilde{s}_k^{ij} \in \tilde{\mathcal{S}}_i} \tilde{\mathcal{S}}_i = \varnothing. \quad (28)$$

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. It is shown 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

$$I_C(x) := \begin{cases} 1, x \in C \\ 0, x \notin C \end{cases} \quad (29)$$

*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 $\tilde{\mathcal{S}}_i$. The next state transition will occur from the state $j$ to any other Markov state. Consequently, by Definition 3.5, the next state transition will occur in $\tilde{\mathcal{S}}_j$. 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

$$V(\tilde{s}_M^{ij}) := \sum_{k=1}^{M} I_{\{s_k = j\} \cap \{T_1(i) \geq k\}}(s_k), \quad (30)$$

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

$$\begin{aligned} \bar{V}(\tilde{s}_M^{ij}) &:= E\{V(\tilde{s}_M^{ij}) \mid s_k = i\} = \\ &= \sum_{k=1}^{M} \mathbb{P}(s_k = j, T_1(i) \geq k \mid s_0 = i). \end{aligned} \quad (31)$$

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

$$\mu_{\tilde{\mathcal{S}}_i} := \sum_{j \in \mathcal{S}} \bar{V}(\tilde{s}_M^{ij}) = \sum_{j \in \mathcal{S}} \sum_{k=1}^{M} \mathbb{P}(s_k = j, T_1(i) \geq k \mid s_0 = i). \quad (32)$$

*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 $i \in \mathcal{S}$, $\mu_i$.

*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 $\tilde{\mathcal{S}}_i$ to $\tilde{\mathcal{S}}_j$. 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 $\tilde{\mathcal{S}}_i$. 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

$$\mathbb{P}(A \mid B \cap C) = \mathbb{P}(A \mid B), \text{ then} \quad (33)$$

$$\mathbb{P}(A \cap C \mid B) = \mathbb{P}(A \mid B) \cdot \mathbb{P}(C \mid B) \quad (34)$$

*Proof*:

$$\mathbb{P}(A \cap C \mid B) = \frac{\mathbb{P}(A \cap B \cap C)}{\mathbb{P}(B)} \quad (35)$$

using the identity $\mathbb{P}(A \mid B) \cdot \mathbb{P}(B) = \mathbb{P}(A \cap B)$, Eq.(35) yields

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

by using Eq. (33)

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

$$= \mathbb{P}(A \mid B) \cdot \mathbb{P}(C \mid B).$$

□

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).

$$\rho_i \cdot \mu_i =$$

$$= \rho_i \cdot \mu_{\tilde{S}_i} \text{ by Lemma 4.2}$$

$$= \sum_{j \in S} \sum_{k=1}^{M} \mathbb{P}(s_k = j, T_1(i) \geq k \mid s_0 = i) \cdot \mathbb{P}(s_0 = i) \quad (36)$$

$$= \sum_{j \in S} \sum_{k=1}^{M} \mathbb{P}(s_k = j, T_1(i) \geq k, s_0 = i) \quad (37)$$

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

For $k = 1$, Eq. (37) yields

$$\sum_{j \in S} \mathbb{P}(s_k = j, T_1(i) \geq 1, s_0 = i) = 1. \quad (38)$$

For $k \geq 2$, Eq. (36) yields

$$\sum_{j \in S} \sum_{k=1}^{M} \mathbb{P}(s_k = j, s_m \neq i \text{ for } 1 \leq m \leq k-1, s_0 = i). \quad (39)$$

Using Proposition 4.1 and since

$\mathbb{P}(s_k = j \mid s_m \neq i \text{ for } 1 \leq m \leq k-1, s_0 = i) = \mathbb{P}(s_k = j \mid s_0 = i)$, Eq.(39) becomes

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

$$= \sum_{k=1}^{M} \mathbb{P}(s_m \neq i \text{ for } 1 \leq m \leq k-1, s_0 = i) \quad (40)$$

by using the identity $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$, Eq. (40) becomes

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

Since the Markov chain is homogeneous (Assumption 3.1)

$$= \sum_{k=1}^{M} \{\mathbb{P}(s_0 = i) + \mathbb{P}(s_0 \neq i)\} + \lim_{k \to \infty} \left( \mathbb{P}(s_m \neq i \text{ for } 0 \leq m \leq k-3) \right) -$$

$$- \lim_{k \to \infty} \left( \mathbb{P}(s_m \neq i \text{ for } 0 \leq m \leq k-1) \right), \quad (41)$$

since the Markov states are irreducible (Assumption 3.3)

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

$\lim_{k \to \infty} \left( \mathbb{P}(s_m \neq i \text{ for } 0 \leq m \leq k-1) \right) = 0$, and Eq. (41) 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 $\tilde{\mathcal{S}}_i$, $\mu_{\tilde{\mathcal{S}}_i}$.

□

## 5. CONCUDING 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 towards 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 to 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 from the control policy imposed by the controller, and be addressed separately.

## ACKNOWLEDGMENTS

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., *Neuro-Dynamic Programming (Optimization and Neural Computation Series, 3)*, 1st edition, Athena Scientific, May 1996.

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

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

[4] Kaelbling, L. P., Littman, M. L., and Moore, A. W., "Reinforcement Learning: a Survey," *Journal of Artificial Intelligence Research*, vol. 4, 1996.

[5] Mandl, P., "Estimation and Control in Markov Chains," *Advances in Applied Probability*, vol. 6, pp. 40-60, 1974.

[6] Borkar, V. and Varaiya, P., "Adaptive Control of Markov Chains. I. Finite Parameter Set," *IEEE Transactions on Automatic Control*, vol. AC-24, pp. 953-7, 1979.

[7] Borkar, V. and Varaiya, P., "Identification and Adaptive Control of Markov Chains," *SIAM Journal on Control and Optimization*, vol. 20, pp. 470-89, 1982.

[8] Kumar, P. R., "Adaptive Control With a Compact Parameter Set," *SIAM Journal on Control and Optimization*, vol. 20, pp. 9-13, 1982.

[9] Doshi, B. and Shreve, S. E., "Strong Consistency of a Modified Maximum Likelihood Estimator for Controlled Markov Chains," *Journal of Applied Probability*, vol. 17, pp. 726-34, 1980.

[10] Kumar, P. R. and Becker, A., "A new Family of Optimal Adaptive Controllers for Markov Chains," *IEEE Transactions on Automatic Control*, vol. AC-27, pp. 137-46, 1982.

[11] Kumar, P. R. and Lin, W., "Optimal Adaptive Controllers for Unknown Markov Chains," *IEEE Transactions on Automatic Control*, vol. AC-27, pp. 765-74, 1982.

[12] Sato, M., Abe, K., and Takeda, H., "Learning Control of Finite Markov Chains with Unknown Transition Probabilities," *IEEE Transactions on Automatic Control*, vol. AC-27, pp. 502-5, 1982.

[13] Sato, M., Abe, K., and Takeda, H., "An Asymptotically Optimal Learning Controller for Finite Markov Chains with Unknown Transition Probabilities," *IEEE Transactions on Automatic Control*, vol. AC-30, pp. 1147-9, 1985.

[14] Sato, M., Abe, K., and Takeda, H., "Learning Control of Finite Markov Chains with an Explicit Trade-off Between Estimation and Control," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 18, pp. 677-84, 1988.

[15] Kumar, P. R., "A Survey of Some Results in Stochastic Adaptive Control," *SIAM Journal on Control and Optimization*, vol. 23, pp. 329-80, 1985.

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

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

[18] Malikopoulos, A. A., Papalambros, P. Y., and Assanis, D. N., "A State-Space Representation Model and Learning Algorithm for Real-Time Decision-Making Under Uncertainty," in Proceedings of the 2007 ASME International Mechanical Engineering Congress and Exposition, Seattle, Washington, November 11-15, 2007.

[19] Malikopoulos, A. A., Papalambros, P. Y., and Assanis, D. N., "A Learning Algorithm for Optimal Internal Combustion Engine Calibration in Real Time," in Proceedings of the ASME 2007 International Design Engineering Technical Conferences Computers and Information in Engineering Conference, Las Vegas, Nevada, September 4-7, 2007.

[20] Malikopoulos, A. A., Assanis, D. N., and Papalambros, P. Y., "Real-Time, Self-Learning Optimization of Diesel Engine Calibration," in Proceedings of the 2007 Fall Technical Conference of the ASME Internal Combustion Engine Division, Charleston, South Carolina, October 14-17, 2007.

[21] Malikopoulos, A. A., Assanis, D. N., and Papalambros, P. Y., "Optimal Engine Calibration for Individual Driving Styles," in Proceedings of the SAE 2008 World Congress and Exhibition, Detroit, Michigan, April 14-17, 2008, SAE 2008-01-1367.

[22] Kumar, P. R. and Varaiya, P., *Stochastic Systems*, Prentice Hall, June 1986.

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

[24] Gubner, J. A., *Probability and Random Processes for Electrical and Computer Engineers*, 1st edition, Cambridge University Press, June 5, 2006.

[25] Grimmett, G. R. and Stirzaker, D. R., *Probability and Random Processes*, 3rd edition, Oxford University Press, July 16, 2001.