Probability in the Engineering and Informational Sciences, **14**, 2000, 511–531. Printed in the U.S.A.

ERROR BOUNDS FOR OCCUPATION MEASURE OF SINGULARLY PERTURBED MARKOV CHAINS INCLUDING TRANSIENT STATES

G. YIN
Department of Mathematics
Wayne State University
Detroit, MI 48202
E-mail: gyin@math.wayne.edu

Q. ZHANG
Department of Mathematics
University of Georgia
Athens, GA 30602
E-mail: qingz@math.uga.edu

Q. G. LIU
Department of Mathematics
Wayne State University
Detroit, MI 48202
E-mail: qgliu@math.wayne.edu

Motivated by many applications in production planning, system reliability, queueing networks, and wireless communication, this work is devoted to singularly perturbed Markov chains with finite states. Focusing on nonstationary processes with the inclusion of transient states, asymptotic error bounds of a sequence of suitably scaled occupation measures are derived. The main tools used include martingales and differential equations. The results are useful for analyzing structural properties of the underlying Markov chains and for designing nearly optimal and hierarchical controls of large-scale and complex systems.

1. INTRODUCTION

This work develops exponential error bounds for occupation measure of singularly perturbed Markov chains with finite-state space. Suppose that $(\Omega, \mathcal{F}, P)$ is the un-

© 2000 Cambridge University Press 0269-9648/00 $12.50

https://doi.org/10.1017/S0269964800144080 Published online by Cambridge University Press

derlying probability space, $\varepsilon > 0$ is a small parameter, and $\alpha^\varepsilon(t)$ is a Markov chain having finite-state space $\mathcal{M} = \{1, \ldots, m\}$ and generator $Q^\varepsilon(t)$, where

$$Q^\varepsilon(t) = \frac{\tilde{Q}(t)}{\varepsilon} + \hat{Q}(t), \tag{1.1}$$

in which both $\tilde{Q}(t)$ and $\hat{Q}(t)$ are themselves generators and $\tilde{Q}(t)$ has certain structures in accordance with classification of states. Because ε is a small parameter, the dominating part of $Q^\varepsilon(t)$ is given by the first term on the right-hand side of (1.1). It is well known that a finite-state Markov chain has at least one recurrent state; that is, it consists of either all recurrent states or also includes transient states. Associated with $\tilde{Q}(t)$, there is a Markov chain having $\tilde{Q}(t)$ as its generator. This Markov chain either has all recurrent states or it also contains transient states. This paper focuses on the case that the Markov chain corresponding to $\tilde{Q}(t)$ has several recurrent classes and transient states. That is what we meant by "inclusion of transient states." Our primary interest is on nonstationary Markov chains; the main motivations stem from various applications in manufacturing systems, queueing networks, system reliability, and so forth; see Delebecque and Quadrat [3], Pan and Başar [9], Pervozvanskii and Gaitsgori [10], Phillips and Kokotovic [11], Sethi and Zhang [12], Yin and Zhang [15], and Yong and Zhou [17], and the references therein. For motivational purpose, let us first consider a simple production planning problem that is a two-machine flowshop with machines subject to breakdown and repair. The production capacity of the machines is modeled by a finite-state Markov chain. If the machine is up, it can produce parts with a certain production rate $u(t)$. If the machine is down, it has to be repaired and the corresponding production rate becomes zero during this period. Each of the machines is either in operating condition (denoted by 1) or under repair (denoted by 0). The capacity of the flowshop is a four-state Markov chain with state space $\{(1,1),(0,1),(1,0),(0,0)\}$. Suppose that the first machine breaks down much more often than the second one. To reflect this situation, suppose that the Markov chain is generated by (1.1), with $\tilde{Q}(\cdot)$ and $\hat{Q}(\cdot)$ given by

$$\tilde{Q}(t) = \begin{pmatrix} -\lambda_1(t) & \lambda_1(t) & 0 & 0 \\ \mu_1(t) & -\mu_1(t) & 0 & 0 \\ 0 & 0 & -\lambda_1(t) & \lambda_1(t) \\ 0 & 0 & \mu_1(t) & -\mu_1(t) \end{pmatrix}$$

and

$$\hat{Q}(t) = \begin{pmatrix} -\lambda_2(t) & 0 & \lambda_2(t) & 0 \\ 0 & -\lambda_2(t) & 0 & \lambda_2(t) \\ \mu_2(t) & 0 & -\mu_2(t) & 0 \\ 0 & \mu_2(t) & 0 & -\mu_2(t) \end{pmatrix},$$

where $\lambda_i(\cdot)$ and $\mu_i(\cdot)$ are the rates of repair and breakdown of the ith machine ($i = 1,2$), respectively. Our interest is to find optimal controls of the system by minimiz-

https://doi.org/10.1017/S0269964800144080 Published online by Cambridge University Press

ing an appropriate cost function over time (subject to certain system dynamics). In this example, $\widetilde{Q}(t)$ consists of two weakly irreducible (to be defined later) classes. To bring out the distinct features of such systems, recent interest in large-scale manufacturing systems aims to use the hierarchical structure of the underlying systems to derive nearly optimal controls. It is well known that the demand in production planning problems may also be formulated as a singularly perturbed Markov chain (see [12]). In such a case, the fast changing part of the generator still takes the form (1.1), and the product demand process may fluctuate among a number of states. The process may only stay at the initial states for a short period of time (i.e., the initial demand rate of a new product may not be revisited after a short duration of time due to the increasing popularity of that product). In this case, the initial demand rates can be classified as transient states.

In the above example, we may aggregate the states in each of the two recurrent classes into a single state. Then, the aggregated process has only two states. This idea may be applied to more general systems. For a large-scale system with m states, assuming the fast-changing part of the generator has l recurrent classes, we may obtain an aggregated process with only l states. If $l \ll m$, the complexity is substantially reduced. Using the system with the aggregated states as a guidance, we may then proceed to obtain nearly optimal control of the original system. Detailed discussions on such an approach can be found in Sethi and Zhang [12]; see also Pan and Başar [9] for the corresponding optimal control problems of dynamic systems involving singularly perturbed Markov chains including transient states. Related references regarding decomposition and aggregation may also be found in Courtois [1] and Simon and Ando [13] among others. Note that when transient states are included, the aggregation is still taken within each recurrent class. To obtain any useful results for control and optimization of the underlying systems, the foremost task is to determine the asymptotic behavior (as $\varepsilon \to 0$) of the probability distribution of the underlying chains.

Continuing interest and effort have been shown to singularly perturbed Markovian models. Khasminskii, Yin, and Zhang [6] use singular perturbation techniques to find asymptotic expansion of the probability vector; singularly perturbed chains with weak and strong interactions are then treated in [7]. Denote the probability vector by

$$p^{\varepsilon}(t) = (P(\alpha^{\varepsilon}(t) = 1), \ldots, P(\alpha^{\varepsilon}(t) = m)). \tag{1.2}$$

It has been shown that as $\varepsilon \to 0$, for $t > 0$, $p^{\varepsilon}(t)$ converges to a limit measure $\bar{p}(t)$. A deterministic approach of the rate of convergence is obtained by means of asymptotic expansions in the aforementioned references. To account for the probabilistic structures, we can center a sequence of occupation measures by its mean and rescale it by $(1/\sqrt{\varepsilon})$ and show that the resulting process is dictated by the central limit theorem. Asymptotic normality for a fast-varying Markov chain is considered in Zhang and Yin [18] via a direct approach, whereas chains including both recurrent and transient classes are treated in Yin, Zhang, and Badowski [16] by using the perturbed test function method (see [8]). We may also tackle the rate of convergence

https://doi.org/10.1017/S0269964800144080 Published online by Cambridge University Press

through probabilistic error bounds of the scaled occupation measures. Singularly perturbed Markov chains with only recurrent classes are treated in [15]. The current paper aims to derive the error bounds for Markov chains including transient states. Such error bounds are needed for the asymptotic analysis of related hierarchical control problems. In addition, they can also help us to derive certain probabilistic bounds needed in the study of functional central limit theory, and they can render help for getting moments bounds for scaled occupation measures. The analysis for the inclusion of transient states is, however, more involved and difficult than that of the recurrent case.

If transient states are included, by appropriate arrangement of the order of states, the generator $\tilde{Q}(t)$ can be put into the form

$$\tilde{Q}(t) = \begin{pmatrix} \tilde{Q}^1(t) & & & \\ & \ddots & & \\ & & \tilde{Q}^l(t) & \\ \tilde{Q}^1_*(t) & \dots & \tilde{Q}^l_*(t) & \tilde{Q}_*(t) \end{pmatrix} \tag{1.3}$$

such that for each $t \in [0,T]$ and each $k = 1,\dots,l, \tilde{Q}^k(t)$, $\tilde{Q}_*(t)$, and $\tilde{Q}^k_*(t)$ are $m_k \times m_k$, $m_* \times m_*$, and $m_* \times m_k$ matrices, respectively, and

$$m_1 + m_2 + \cdots + m_l + m_* = m.$$

For each k, $\tilde{Q}^k(t)$ is a generator of a weakly irreducible Markov chain and $\tilde{Q}_*(t)$ is a stable (to be precisely defined later) matrix representing transient behavior. Let $\mathcal{M}_k = \{s_{k1},\dots,s_{km_k}\}$ for $k = 1,\dots,l$, denote the states corresponding to recurrent states, and $\mathcal{M}_* = \{s_{*1},\dots,s_{*m_*}\}$ denote the set of transient states. Then, the state space of $\alpha^\varepsilon(t)$ can be decomposed to

$$\begin{aligned} \mathcal{M} &= \mathcal{M}_1 \cup \cdots \cup \mathcal{M}_l \cup \mathcal{M}_* \\ &= \{s_{11},\dots,s_{1m_1},\dots,s_{l1},\dots,s_{lm_l},s_{*1},\dots,s_{*m_*}\}. \end{aligned}$$

If $\hat{Q}(t) = 0$, then the underlying process will stay in one of the recurrent classes $\mathcal{M}_k$ $(k = 1,2,\dots,m)$ after entering there. If the process enters any of the transient states in $\mathcal{M}_*$, it will jump to a recurrent class in a fairly short amount of time. In view of (1.1), the use of nonzero $\hat{Q}(t)$ makes the model more interesting and complex because it allows weak interactions among these classes.

Aiming at the desired error bounds, the remainder of the paper is arranged as follows. Section 2 begins with the precise formulation of the problem and provides related preliminary results. Section 3 derives the desired exponential error bounds. As applications, Section 4 discusses the use of the error bounds for obtaining probabilistic and moment bounds of the scaled occupation measures. In addition, it discusses some ramifications. Section 5 concludes the paper with further remarks.

https://doi.org/10.1017/S0269964800144080 Published online by Cambridge University Press

2. FORMULATION AND PRELIMINARIES

We first briefly recall the notion of generator, irreducibility, and so forth. We then present certain preliminary results needed in the subsequent study. The section is concluded with the definition of scaled occupation measures.

2.1. Background

Suppose $\alpha(\cdot)$ is a continuous-time Markov chain with finite-state space $\mathcal{M} = \{1,\dots,m\}$. Let $Q(t) = (q_{ij}(t))$ denote an $m \times m$ matrix for $t \geq 0$. $Q(t)$ is called a *generator* of $\alpha(\cdot)$ if the following hold:

1. $q_{ij}(t)$ is continuous in t for all $i, j \in \mathcal{M}$.
2. $q_{ij}(t) \geq 0$ for $j \neq i$ and $q_{ii}(t) = -\sum_{j\neq i} q_{ij}(t)$, $t \geq 0$.
3. For any bounded real-valued function f defined on $\mathcal{M}$,

$$f(\alpha(t)) - \int_0^t Q(s) f(\cdot)(\alpha(s))\, ds \tag{2.1}$$

is a martingale with respect to $\mathcal{F}_t$ the σ-algebra generated by $\{\alpha(s), s \geq t\}$, where

$$Q(t)f(\cdot)(i) = \sum_{j\in\mathcal{M}} q_{ij}(t) f(j) = \sum_{j\neq i} q_{ij}(t)(f(j) - f(i)).$$

A generator $Q(t)$ is said to be *weakly irreducible* if, for each fixed $t \geq 0$, the system of equations

$$\begin{aligned} \pi(t)Q(t) &= 0, \\ \sum_{i=1}^{m} \pi_i(t) &= 1 \end{aligned} \tag{2.2}$$

has a unique solution $\pi(t) = (\pi_1(t),\dots,\pi_m(t))$ with $\pi_i(t) \geq 0$ for $i = 1,\dots,m$. Such a solution is termed a quasistationary distribution; see Davis [2] for the treatment of nonstationary Markov chains through piecewise deterministic process formulation and [15, Chap. 2] for the notion of weak irreducibility, quasistationary distributions, and so forth.

2.2. Singularly Perturbed Markov Chains

We work with a finite time horizon $t \in [0, T]$, for some $T > 0$. For a small parameter $\varepsilon > 0$, let $\alpha^\varepsilon(t)$ be a Markov chain generated by (1.1), where $\tilde{Q}(t)$ is given by (1.3). To proceed, we make the following assumptions:

(A1) For each $k = 1,\dots,l$ and $t \in [0, T]$, $\tilde{Q}^k(t)$ is weakly irreducible and $\tilde{Q}_*(t)$ is Hurwitz (i.e., all of its eigenvalues have negative real parts). In addition, $\tilde{Q}(\cdot)$ is differentiable on $[0, T]$ and its derivative is Lipschitz; $\hat{Q}(\cdot)$ is also Lipschitz on $[0, T]$.

https://doi.org/10.1017/S0269964800144080 Published online by Cambridge University Press

(A2) For $k = 1, \ldots, l$, there exist constant matrices $G_*^k \in \mathbb{R}^{m_* \times m_k}$ and $G_* \in \mathbb{R}^{m_* \times m_*}$, and matrix-valued function $B(t) \in \mathbb{R}^{m_* \times m_*}$ such that $B(\cdot)$ is Lipschitz continuous, and

$$\tilde{Q}_*^k(t) = B(t)G_*^k \quad \text{and} \quad \tilde{Q}_*(t) = B(t)G_*.$$

Remark 2.1: Note that the Hurwitzian property of $\tilde{Q}_*(t)$ implies that $B(t)$ is invertible.

For $k = 1, \ldots, l$, define

$$a_{m_k}(t) = -\tilde{Q}_*^{-1}(t)\tilde{Q}_*^k(t)\mathbb{1}_{m_k}. \tag{2.3}$$

Using assumption (A2), we have, for $k = 1, \ldots, l$,

$$\begin{aligned} a_{m_k}(t) &= -\tilde{Q}_*^{-1}(t)\tilde{Q}_*^k(t)\mathbb{1}_{m_k} \\ &= -(B(t)G_*)^{-1}(B(t)G_*^k)\mathbb{1}_{m_k} \\ &= -G_*^{-1}B^{-1}(t)B(t)G_*^k\mathbb{1}_{m_k} \\ &= -G_*^{-1}G_*^k\mathbb{1}_{m_k} \\ &= a_{m_k} \quad \text{independent of } t. \end{aligned}$$

Denote

$$\begin{aligned} A_* &= (a_{m_1}, \ldots, a_{m_l}), \\ \tilde{\mathbb{1}} &= \operatorname{diag}(\mathbb{1}_{m_1}, \ldots, \mathbb{1}_{m_l}) = \begin{pmatrix} \mathbb{1}_{m_1} & & \\ & \ddots & \\ & & \mathbb{1}_{m_l} \end{pmatrix}, \\ \tilde{\mathbb{1}}_* &= \begin{pmatrix} \tilde{\mathbb{1}} & 0_{(m-m_*)\times(m-m_*)} \\ A_* & 0_{m_*\times m_*} \end{pmatrix}, \end{aligned} \tag{2.4}$$

where $0_{j\times j}$ is a $j \times j$ matrix with all entries being zero. It can be shown by a direct calculation that

$$\tilde{Q}(t)\tilde{\mathbb{1}}_* = 0. \tag{2.5}$$

In addition, for each $t \in [0, T]$ and each $k = 1, \ldots, l$, write a_{m_k} in component form

$$a_{m_k} = (a_{m_k,1}, \ldots, a_{m_k,m_*})' \in \mathbb{R}^{m_* \times 1}.$$

Then, it can be shown (see [16]), for each $t \in [0, T]$ and $j = 1, \ldots, m_*$, that

$$a_{m_k,j} \geq 0 \quad \text{and} \quad \sum_{k=1}^{l} a_{m_k,j} = 1; \tag{2.6}$$

https://doi.org/10.1017/S0269964800144080 Published online by Cambridge University Press

that is, for each $t \in [0,T]$ and $j = 1,\dots,l$, $(a_{m_1,j},\dots,a_{m_l,j})$ can be interpreted as a probability row vector.

Now, we return to the slowly varying part of the generator. Put $\hat{Q}(t)$ in a partitioned form corresponding to the division of recurrent and transient states:

$$\hat{Q}(t) = \begin{pmatrix} \hat{Q}^{11}(t) & \hat{Q}^{12}(t) \\ \hat{Q}^{21}(t) & \hat{Q}^{22}(t) \end{pmatrix}, \tag{2.7}$$

where the block matrices

$$\hat{Q}^{11}(t) \in \mathbb{R}^{(m-m_*)\times(m-m_*)}, \qquad \hat{Q}^{12}(t) \in \mathbb{R}^{(m-m_*)\times m_*},$$
$$\hat{Q}^{21}(t) \in \mathbb{R}^{m_*\times(m-m_*)}, \qquad \hat{Q}^{22}(t) \in \mathbb{R}^{m_*\times m_*}.$$

The "average" of $\hat{Q}(t)$ with respect to the quasistationary measure is defined by

$$\begin{aligned} \bar{Q}(t) &= \operatorname{diag}(\nu^1(t),\dots,\nu^l(t))(\hat{Q}^{11}(t)\tilde{\mathbb{1}} + \hat{Q}^{12}(t)A_*), \\ \bar{Q}_*(t) &= \operatorname{diag}(\bar{Q}(t),0_{m_*\times m_*}). \end{aligned} \tag{2.8}$$

Denote

$$\begin{aligned} \nu(t) &= \operatorname{diag}(\nu^1(t),\dots,\nu^l(t)), \\ \nu_*(t) &= \operatorname{diag}(\nu(t),0_{m_*\times m_*}). \end{aligned} \tag{2.9}$$

Using $\nu_*(t)$, we can write

$$\bar{Q}_*(t) = \nu_*(t)\hat{Q}(t)\tilde{\mathbb{1}}_*.$$

It is readily verified that both $\bar{Q}(t)$ and $\bar{Q}_*(t)$ are generators. Let U be a random variable uniformly distributed on $[0,1]$ that is independent of $\alpha^\varepsilon(\cdot)$. For each $j = 1,\dots,m_*$, define an integer-valued random variable U_j by

$$U_j = I_{\{0\le U\le a_{m_1,j}\}} + 2I_{\{a_{m_1,j}<U\le a_{m_1,j}+a_{m_2,j}\}} + \cdots + lI_{\{a_{m_1,j}+\cdots+a_{m_{l-1},j}<U\le 1\}}.$$

Using the definition of U_j, we can now define an aggregated process. The idea is to treat all m_i states in the recurrent class $\mathcal{M}_i$ as one state. This results in the definition of the aggregated process:

$$\bar{\alpha}^\varepsilon(t) = \begin{cases} i & \text{if } \alpha^\varepsilon(t) \in \mathcal{M}_i \\ U_j & \text{if } \alpha^\varepsilon(t) = s_{*j}. \end{cases} \tag{2.10}$$

Note that the state space of $\bar{\alpha}^\varepsilon(t)$ is $\bar{\mathcal{M}} = \{1,\dots,l\}$. It has been proven in [16] that although $\bar{\alpha}^\varepsilon(\cdot)$ is not necessarily Markovian, it converges weakly to a Markov chain $\bar{\alpha}(\cdot)$ that is generated by $\bar{Q}(t)$. Clearly, the limiting Markov chain has a reduced state space. Such a result provides guidelines for many control and optimization problems and allows us to treat a limit problem in lieu of the more complicated original problem so that a substantial reduction of complexity is achieved (i.e., a particular important case is that $l \ll m$, where l and m are the number of states in the aggregated process and the original process, respectively). The current paper examines the as-

https://doi.org/10.1017/S0269964800144080 Published online by Cambridge University Press

ymptotic properties of the singularly perturbed Markov chains through the derivation of error bounds of exponential type.

We will use the notation of the partitioned vector

$$I^{\varepsilon,i}(t) = (I_{\{\alpha^\varepsilon(t)=s_{i1}\}},\dots,I_{\{\alpha^\varepsilon(t)=s_{im_i}\}}) \in \mathbb{R}^{1\times m_i} \quad \text{for } i = 1,\dots,l,$$

$$I^{\varepsilon,*}(t) = (I_{\{\alpha^\varepsilon(t)=s_{*1}\}},\dots,I_{\{\alpha^\varepsilon(t)=s_{*m_*}\}}) \in \mathbb{R}^{1\times m_*}, \tag{2.11}$$

and denote the jth components of $I^{\varepsilon,i}(t)$ and $I^{\varepsilon,*}(t)$ by $I_j^{\varepsilon,i}(t)$ and $I_j^{\varepsilon,*}(t)$, respectively. Define

$$I^\varepsilon(t) = (I^{\varepsilon,1}(t),\dots,I^{\varepsilon,l}(t),I^{\varepsilon,*}(t)). \tag{2.12}$$

To proceed, we state a lemma. Its proof can be obtained from [15, Sect. 6.4]. We omit the details here.

LEMMA 2.2: *Assume conditions* (A1) *and* (A2) *are satisfied. Then, the following assertions hold*:

(A) *For the probability distribution vector*,

$$p^\varepsilon(t) = (\theta(t)\,\mathrm{diag}(\nu^1(t),\dots,\nu^l(t)),0_{m_*}) + O(\varepsilon + \exp(-\kappa_0 t/\varepsilon)),$$

where $0_{m_*} \in \mathbb{R}^{1\times m_*}$ *and* $\theta(t) = (\theta_1(t),\dots,\theta_l(t)) \in \mathbb{R}^{1\times l}$, *satisfying*

$$\frac{d\theta(t)}{dt} = \theta(t)\bar{Q}(t), \qquad \theta_i(0) = p^{\varepsilon,i}(0)\mathbb{1}_{m_i} + p^{\varepsilon,*}(0)a_{m_i}$$

and $p^\varepsilon(0) = (p^{\varepsilon,1}(0),\dots,p^{\varepsilon,l}(0),p^{\varepsilon,*}(0))$ *with* $p^{\varepsilon,i}(0) \in \mathbb{R}^{1\times m_i}$ *and* $p^{\varepsilon,*}(0) \in \mathbb{R}^{1\times m_*}$.

(B) *We have, for* $t \ge t_0 \ge 0$,

$$P^\varepsilon(t,t_0) = P^{(0)}(t,t_0) + O(\varepsilon + \exp(-\kappa_0(t-t_0)/\varepsilon)) \tag{2.13}$$

for some $\kappa_0 > 0$, *where*

$$P^{(0)}(t,t_0) = \tilde{\mathbb{1}}_*\Theta_*(t,t_0)\nu_*(t),$$
$$\Theta_*(t,t_0) = \mathrm{diag}(\Theta(t,t_0),I_{m_*\times m_*}) \tag{2.14}$$

and $\Theta(t,s) = (\theta_{ij}(t,s)) \in \mathbb{R}^{l\times l}$ *satisfies the differential equation*

$$\frac{\partial\Theta(t,s)}{\partial t} = \Theta(t,s)\bar{Q}(t), \qquad \Theta(s,s) = I. \tag{2.15}$$

https://doi.org/10.1017/S0269964800144080 Published online by Cambridge University Press

Remark 2.3: In view of the definitions of $\nu_*(t)$ and $\Theta_*(t,s)$, we can write

$$P^{(0)}(t,t_0) = \begin{pmatrix} \tilde{1\!\!1}\Theta(t,t_0)\nu(t) & 0_{(m-m_*)\times(m-m_*)} \\ A_*\Theta(t,t_0)\nu(t) & 0_{m_*\times m_*} \end{pmatrix}.$$

Moreover, using (2.15), the $(l+m_*)\times(l+m_*)$ matrix-valued function $\Theta_*(t,s)$ also satisfies

$$\frac{\partial\Theta_*(t,s)}{\partial t} = \Theta_*(t,s)\bar{Q}_*(t), \qquad \Theta_*(s,s) = I. \tag{2.16}$$

2.3. Occupation Measures

Let $\beta_{ij}(t)$ be a bounded and Borel-measurable deterministic function for each $i = 1,\dots,l,*$, $j = 1,\dots,m_i$ and $t \geq 0$. Consider the normalized occupation measures

$$n^\varepsilon(t) = (n^\varepsilon_{11}(t),\dots,n^\varepsilon_{1m_1}(t),\dots,n^\varepsilon_{l1}(t),\dots,n^\varepsilon_{lm_l}(t),n^\varepsilon_{*1}(t),\dots,n^\varepsilon_{*m_*}(t)),$$

where

$$n^\varepsilon_{ij}(t) = \frac{1}{\sqrt{\varepsilon}}\int_0^t \tilde{n}^\varepsilon_{ij}(s,\alpha^\varepsilon(s))\,ds, \tag{2.17}$$

with

$$\tilde{n}^\varepsilon_{ij}(t,\alpha) = \begin{cases} (I_{\{\alpha=s_{ij}\}} - \nu^i_j(t)I_{\{\alpha\in\mathcal{M}_i\}})\,\beta_{ij}(t) & \text{for } i = 1,\dots,l, \\ I_{\{\alpha=s_{*j}\}}\beta_{*j}(t) & \text{for } i = *. \end{cases} \tag{2.18}$$

Remark 2.4: The centered and suitably scaled processes defined in (2.17) represent approximation of the original occupation measures by that of aggregated ones weighted by the quasistationary distributions. They are termed occupation measures because they indicate how much time the processes spend in the given sets. They give much needed information in various applications. The scale $1/\sqrt{\varepsilon}$ is suggested by central-limit-type results; for the related study in asymptotic normality, see [18] (see also [15,16]). The insertion of the functions $\beta_{ij}(t)$ make the definition of the scaled occupation measures more general since $\beta_{ij}(\cdot)$ can be chosen from a rather large class with only boundedness and measurability being assumed. The inclusion of $\beta_{ij}(\cdot)$ are primarily motivated by applications in the context of hierarchical controls; see, for example, [12] for more discussion.

3. ERROR BOUNDS

Our effort in what follows is to obtain the exponential error bounds for the sequence of the occupation measures $n^\varepsilon(\cdot)$. According to Lemma 2.2, there exists a constant $\kappa_0 > 0$ such that

$$|P^\varepsilon(t,s) - P^{(0)}(t,s)| = O\left(\varepsilon + \exp\left(-\frac{\kappa_0(t-s)}{\varepsilon}\right)\right). \tag{3.1}$$

https://doi.org/10.1017/S0269964800144080 Published online by Cambridge University Press

For any fixed $T > 0$, let

$$K_T = \max\left\{1, \sup_{0\le s\le t\le T}\left(\frac{|P^\varepsilon(t,s) - P^{(0)}(t,s)|}{\varepsilon + \exp(-\kappa_0(t-s)/\varepsilon)}\right)\right\}. \tag{3.2}$$

Use $O_1(y)$ to denote a function of y such that $|O_1(y)|/|y| \le 1$. Then, we may write (3.1) as

$$|P^\varepsilon(t,s) - P^{(0)}(t,s)| = K_T O_1\left(\varepsilon + \exp\left(-\frac{\kappa_0(t-s)}{\varepsilon}\right)\right). \tag{3.3}$$

Denoting

$$\beta(t) = \operatorname{diag}(\beta_{11}(t),\ldots,\beta_{1m_1}(t),\ldots,\beta_{l1}(t),\ldots,\beta_{lm_l}(t),\beta_{*1}(t),\ldots,\beta_{*m_*}(t)),$$

$n^\varepsilon(t)$ may be written as

$$n^\varepsilon(t) = \frac{1}{\sqrt{\varepsilon}}\int_0^t (I^\varepsilon(s) - I^\varepsilon(s)\tilde{\mathbb{1}}_*\nu_*(s))\beta(s)\,ds.$$

To proceed, the desired error bounds are provided in the following theorem. The proof is a combined use of stochastic analysis based on stochastic differential equations with the driving noise being sequences of square integrable martingales and the theory of differential equations and dynamic systems.

THEOREM 3.1: *Assume* (A1) *and* (A2). *Then, there exist* $\varepsilon_0 > 0$ *and* $K > 0$ *such that for* $0 < \varepsilon \le \varepsilon_0$, $T \ge 0$, *and for any bounded, Borel-measurable, and deterministic process* $\beta_{ij}(\cdot)$,

$$E\exp\left(\frac{\theta_T}{(T+1)^3}\sup_{0\le t\le T}|n^\varepsilon(t)|\right) \le K,$$

where θ_T *is a constant satisfying*

$$0 \le \theta_T \le \frac{\min\{1,\kappa_0\}}{K_T|\beta|_T(1+|\hat{Q}|_T)}, \tag{3.4}$$

where

$$|\beta|_T = \max_{i,j}\sup_{0\le t\le T}|\beta_{ij}(t)|,$$

$$|\hat{Q}|_T = \max_{i,j}\sup_{0\le t\le T}|\hat{Q}_{ij}(t)|. \tag{3.5}$$

PROOF: The proof is divided into four steps. We detail these steps as follows.

Step 1: Relationship between $M^\varepsilon(t)$ *and* $I^\varepsilon(t)$. Consider the partitioned vectors defined in (2.11) and (2.12). Then, $M^\varepsilon(\cdot)$ given by

$$M^\varepsilon(t) = I^\varepsilon(t) - I^\varepsilon(0) - \int_0^t I^\varepsilon(s)Q^\varepsilon(s)\,ds \tag{3.6}$$

https://doi.org/10.1017/S0269964800144080 Published online by Cambridge University Press

is a martingale with respect to the σ-algebra $\mathcal{F}_t^\varepsilon = \{\alpha^\varepsilon(s), s \le t\}$. Denote $M^\varepsilon(t) = (M^{\varepsilon,r}(t), M^{\varepsilon,*}(t))$, where for each $t \in [0,T]$, $M^{\varepsilon,r}(t) \in \mathbb{R}^{1\times(m-m_*)}$ and $M^{\varepsilon,*}(t) \in \mathbb{R}^{1\times m_*}$. Equation (3.6) can be written as a linear stochastic differential equation driven by a martingale (see Elliott [4])

$$dI^\varepsilon(t) = I^\varepsilon(t)Q^\varepsilon(t)\,dt + dM^\varepsilon(t). \tag{3.7}$$

Denote by $P^\varepsilon(t,s)$, the principal matrix solution of

$$\frac{dP^\varepsilon(t,s)}{dt} = P^\varepsilon(t,s)Q^\varepsilon(t), \qquad P^\varepsilon(s,s) = I. \tag{3.8}$$

The solution to the stochastic differential equation (3.7) is

$$I^\varepsilon(t) = I^\varepsilon(0)P^\varepsilon(t,0) + \int_0^t (dM^\varepsilon(s))P^\varepsilon(t,s).$$

To use the estimate of $|P^\varepsilon(t,s) - P^{(0)}(t,s)|$ in Lemma 2.2, we rewrite the solution of (3.7) as

$$\begin{aligned} I^\varepsilon(t) &= I^\varepsilon(0)P^{(0)}(t,0) \\ &\quad + \int_0^t (dM^\varepsilon(s))P^{(0)}(t,s) \\ &\quad + I^\varepsilon(0)(P^\varepsilon(t,0) - P^{(0)}(t,0)) \\ &\quad + \int_0^t (dM^\varepsilon(s))(P^\varepsilon(t,s) - P^{(0)}(t,s)). \end{aligned} \tag{3.9}$$

To proceed, multiply both sides of (3.6) by $\tilde{\mathbb{1}}_*$:

$$M^\varepsilon(t)\tilde{\mathbb{1}}_* = I^\varepsilon(t)\tilde{\mathbb{1}}_* - I^\varepsilon(0)\tilde{\mathbb{1}}_* - \int_0^t I^\varepsilon(s)Q^\varepsilon(s)\tilde{\mathbb{1}}_*\,ds.$$

Note that $M^\varepsilon(t)\tilde{\mathbb{1}}_*$ is a square-integrable martingale. Based on (2.5), $Q^\varepsilon(t)\tilde{\mathbb{1}}_* = \hat{Q}(t)\tilde{\mathbb{1}}_*$. Thus, the above equation can be written as

$$M^\varepsilon(t)\tilde{\mathbb{1}}_* = I^\varepsilon(t)\tilde{\mathbb{1}}_* - I^\varepsilon(0)\tilde{\mathbb{1}}_* - \int_0^t I^\varepsilon(s)\hat{Q}(s)\tilde{\mathbb{1}}_*\,ds. \tag{3.10}$$

Step 2: "Aggregation" and "Averaging." Introduce

$$\begin{aligned} \tilde{I}^\varepsilon(t) &= I^\varepsilon(t)\tilde{\mathbb{1}}_*, \\ \bar{I}^\varepsilon(t) &= \tilde{I}^\varepsilon(t)\nu_*(t). \end{aligned} \tag{3.11}$$

Note that

$$\begin{aligned} \tilde{I}^\varepsilon(t) &= (I^{\varepsilon,1}(t)\,\mathbb{1}_{m_1} + I^{\varepsilon,*}(t)a_{m_1}, \ldots, I^{\varepsilon,l}(t)\mathbb{1}_{m_l} + I^{\varepsilon,*}(t)a_{m_l}, 0_{m_*}), \\ \bar{I}^\varepsilon(t) &= ((I^{\varepsilon,1}(t)\mathbb{1}_{m_1} + I^{\varepsilon,*}(t)a_{m_1})\nu^1(t), \ldots, (I^{\varepsilon,l}(t)\mathbb{1}_{m_l} + I^{\varepsilon,*}(t)a_{m_l})\nu^l(t), 0_{m_*}). \end{aligned}$$

https://doi.org/10.1017/S0269964800144080 Published online by Cambridge University Press

Effectively, $\tilde{I}^\varepsilon(t)$ aggregates the occupation measures in each recurrent substate space, and $\bar{I}^\varepsilon(t)$ takes an "average" with respect to the quasistationary distributions.

Then, we have

$$\begin{aligned} I^\varepsilon(t)\hat{Q}(t)\tilde{\mathbb{1}}_* &= \tilde{I}^\varepsilon(t)\bar{Q}_*(t) + (I^\varepsilon(t)\hat{Q}(t)\tilde{\mathbb{1}}_* - \tilde{I}^\varepsilon(t)\bar{Q}_*(t)) \\ &= \tilde{I}^\varepsilon(t)\bar{Q}_*(t) + (I^\varepsilon(t) - \bar{I}^\varepsilon(t))\hat{Q}(t)\tilde{\mathbb{1}}_*. \end{aligned}$$

Thus, we can rewrite (3.10) as

$$\begin{aligned} M^\varepsilon(t)\tilde{\mathbb{1}}_* = \tilde{I}^\varepsilon(t) - \tilde{I}^\varepsilon(0) - \int_0^t \tilde{I}^\varepsilon(s)\bar{Q}_*(s)\,ds \\ + \int_0^t (I^\varepsilon(s) - \bar{I}^\varepsilon(s))\hat{Q}(s)\tilde{\mathbb{1}}_*\,ds. \end{aligned} \tag{3.12}$$

This is a stochastic differential equation driven by a square-integrable martingale

$$d\tilde{I}^\varepsilon = (\tilde{I}^\varepsilon(t)\bar{Q}_*(t) + (I^\varepsilon(t) - \bar{I}^\varepsilon(t))\hat{Q}(t)\tilde{\mathbb{1}}_*)\,dt + dM^\varepsilon(t)\tilde{\mathbb{1}}_*$$

with given $\tilde{I}^\varepsilon(0)$. Denote the principal matrix solution to (2.16) by $\Theta_*(t,s)$. The solution of (3.12) can then be written as

$$\begin{aligned} \tilde{I}^\varepsilon(t) = \tilde{I}^\varepsilon(0)\Theta_*(t,0) + \int_0^t (dM^\varepsilon(s)\tilde{\mathbb{1}}_*)\Theta_*(t,s) \\ + \int_0^t (I^\varepsilon(s) - \bar{I}^\varepsilon(s))\hat{Q}(s)\tilde{\mathbb{1}}_*\Theta_*(t,s)\,ds. \end{aligned} \tag{3.13}$$

Postmultiplying $\nu_*(t)$ to both sides of the above equality, we obtain

$$\begin{aligned} \tilde{I}^\varepsilon(t)\nu_*(t) = \tilde{I}^\varepsilon(0)\Theta_*(t,0)\nu_*(t) + \int_0^t (dM^\varepsilon(s)\tilde{\mathbb{1}}_*)\Theta_*(t,s)\nu_*(t) \\ + \int_0^t (I^\varepsilon(s) - \bar{I}^\varepsilon(s))\hat{Q}(s)\tilde{\mathbb{1}}_*\Theta_*(t,s)\nu_*(t)\,ds. \end{aligned} \tag{3.14}$$

By the definitions of $\bar{I}^\varepsilon(t)$, $\tilde{I}^\varepsilon(t)$, and $P^{(0)}(t,s)$, the above expression can be written as

$$\begin{aligned} \bar{I}^\varepsilon(t) = I^\varepsilon(0)P^{(0)}(t,0) + \int_0^t dM^\varepsilon(s)P^{(0)}(t,s) \\ + \int_0^t (I^\varepsilon(s) - \bar{I}^\varepsilon(s))\hat{Q}(s)P^{(0)}(t,s)\,ds. \end{aligned}$$

https://doi.org/10.1017/S0269964800144080 Published online by Cambridge University Press

Comparing the above equation with (3.9), we have

$$\begin{aligned} I^\varepsilon(t) - \bar{I}^\varepsilon(t) &= I^\varepsilon(0)(P^\varepsilon(t,0) - P^{(0)}(t,0)) \\ &\quad + \int_0^t dM^\varepsilon(s)(P^\varepsilon(t,s) - P^{(0)}(t,s)) \\ &\quad - \int_0^t (I^\varepsilon(s) - \bar{I}^\varepsilon(s))\hat{Q}(s)P^{(0)}(t,s)\,ds. \end{aligned} \tag{3.15}$$

Denote

$$\begin{aligned} g^\varepsilon(t) &= I^\varepsilon(0)(P^\varepsilon(t,0) - P^{(0)}(t,0)) \\ &\quad + \int_0^t dM^\varepsilon(s)(P^\varepsilon(t,s) - P^{(0)}(t,s)) \\ &\quad - \int_0^t (I^\varepsilon(s) - \bar{I}^\varepsilon(s))\hat{Q}(s)(P^{(0)}(t,s) - P^\varepsilon(t,s))\,ds. \end{aligned} \tag{3.16}$$

Then, (3.15) can be written as

$$I^\varepsilon(t) - \bar{I}^\varepsilon(t) - g^\varepsilon(t) + \int_0^t (I^\varepsilon(s) - \bar{I}^\varepsilon(s))\hat{Q}(s)P^\varepsilon(t,s)\,ds = 0. \tag{3.17}$$

Step 3: Estimates via Dynamic Systems and Differential Equations. Note that

$$P^\varepsilon(t,s) = P^\varepsilon(0,s)P^\varepsilon(t,0) \quad \text{and} \quad P^\varepsilon(t,0)P^\varepsilon(0,t) = I.$$

We have

$$(I^\varepsilon(t) - \bar{I}^\varepsilon(t) - g^\varepsilon(t))P^\varepsilon(0,t) + \int_0^t (I^\varepsilon(s) - \bar{I}^\varepsilon(s))\hat{Q}(s)P^\varepsilon(0,s)\,ds = 0.$$

Let

$$h^\varepsilon(t) = (I^\varepsilon(t) - \bar{I}^\varepsilon(t) - g^\varepsilon(t))P^\varepsilon(0,t).$$

Then,

$$h^\varepsilon(t) + \int_0^t h^\varepsilon(s)P^\varepsilon(s,0)\hat{Q}(s)P^\varepsilon(0,s)\,ds + \int_0^t g^\varepsilon(s)\hat{Q}(s)P^\varepsilon(0,s)\,ds = 0.$$

Clearly, $h^\varepsilon(0) = 0$ and the above equation is equivalent to

$$\frac{dh^\varepsilon(t)}{dt} = -h^\varepsilon(t)P^\varepsilon(t,0)\hat{Q}(t)P^\varepsilon(0,t) - g^\varepsilon(t)\hat{Q}(t)P^\varepsilon(0,t).$$

Solving this equation leads to

$$h^\varepsilon(t) = -\int_0^t g^\varepsilon(s)\hat{Q}(s)P^\varepsilon(0,s)\Theta_1(t,s)\,ds,$$

https://doi.org/10.1017/S0269964800144080 Published online by Cambridge University Press

where $\Theta_1(t,s)$ is the solution of the initial-value problem

$$\frac{d\Theta_1(t,s)}{dt} = -\Theta_1(t,s)P^\varepsilon(t,0)\hat{Q}(t)P^\varepsilon(0,t),$$
$$\Theta_1(s,s) = I.$$

Denote $\Theta_2(t,s) = P^\varepsilon(0,s)\Theta_1(t,s)P^\varepsilon(t,0)$. Then,

$$I^\varepsilon(t) - \bar{I}^\varepsilon(t) = g^\varepsilon(t) - \int_0^t g^\varepsilon(s)\hat{Q}(s)\Theta_2(t,s)\,ds. \tag{3.18}$$

By using $P^\varepsilon(0,t)P^\varepsilon(t,0) = I$, we have

$$\begin{aligned}\frac{d\Theta_2(t,s)}{dt} &= P^\varepsilon(0,s)\left(\frac{d\Theta_1(t,s)}{dt}P^\varepsilon(t,0) + \Theta_1(t,s)\frac{dP^\varepsilon(t,0)}{dt}\right)\\ &= P^\varepsilon(0,s)\Bigg(-\Theta_1(t,s)P^\varepsilon(t,0)\hat{Q}(t)P^\varepsilon(0,t)P^\varepsilon(t,0)\\ &\qquad + \Theta_1(t,s)P^\varepsilon(t,0)\left(\frac{1}{\varepsilon}\tilde{Q}(t) + \hat{Q}(t)\right)\Bigg)\\ &= P^\varepsilon(0,s)\Theta_1(t,s)P^\varepsilon(t,0)\left(\frac{1}{\varepsilon}\tilde{Q}(t)\right)\\ &= \Theta_2(t,s)\left(\frac{1}{\varepsilon}\tilde{Q}(t)\right)\\ \Theta_2(s,s) &= I.\end{aligned} \tag{3.19}$$

Equation (3.19) is a forward equation since $\tilde{Q}(t)/\varepsilon$ is a generator. The solution $\Theta_2(t,s)$ is the transition probability matrix corresponding to $\tilde{Q}(t)/\varepsilon$. Therefore,

$$\sup_{0\le s\le t\le T}|\Theta_2(t,s)| \le 1 \quad \text{and hence} \quad |\Theta_2|_T \le 1. \tag{3.20}$$

From (3.18), we have, for $t \in [0,T]$,

$$\begin{aligned}&\int_0^t (I^\varepsilon(u) - \bar{I}^\varepsilon(u))\beta(u)\,du\\ &\qquad = \int_0^t g^\varepsilon(u)\beta(u)\,du - \int_0^t\int_0^u g^\varepsilon(s)\hat{Q}(s)\Theta_2(u,s)\beta(u)\,ds\,du.\end{aligned}$$

Since

$$\int_0^t\int_0^u g^\varepsilon(s)\hat{Q}(s)\Theta_2(u,s)\beta(u)\,ds\,du = \int_0^t g^\varepsilon(s)\left(\int_s^t \hat{Q}(s)\Theta_2(u,s)\beta(u)\,du\right)ds,$$

https://doi.org/10.1017/S0269964800144080 Published online by Cambridge University Press

we obtain

$$\int_0^t (I^\varepsilon(u) - \bar{I}^\varepsilon(u))\beta(u)\,du = \int_0^t g^\varepsilon(u)\tilde{\beta}(t,u)\,du,$$

where

$$\tilde{\beta}(t,u) = \beta(u) - \int_u^t \hat{Q}(u)\Theta_2(r,u)\beta(r)\,dr.$$

Using (3.20) in the defining equation of $\tilde{\beta}(t)$,

$$|\tilde{\beta}|_T \le (1+T)|\beta|_T(1+|\hat{Q}|_T). \tag{3.21}$$

Denote

$$g_1^\varepsilon(t) = I^\varepsilon(0)(P^\varepsilon(t,0) - P^{(0)}(t,0)) + \int_0^t dM^\varepsilon(s)(P^\varepsilon(t,s) - P^{(0)}(t,s)),$$

$$g_2^\varepsilon(t) = \int_0^t (I^\varepsilon(s) - \bar{I}^\varepsilon(s))\hat{Q}(s)(P^{(0)}(t,s) - P^\varepsilon(t,s))\,ds.$$

Step 4: Completion of the Proof. In view of (3.16) and (3.17),

$$\frac{1}{\sqrt{\varepsilon}}\left|\int_0^t (I^\varepsilon(u) - \bar{I}^\varepsilon(u))\beta(u)\,du\right|$$

$$\le \frac{1}{\sqrt{\varepsilon}}\left|\int_0^t g_1^\varepsilon(u)\tilde{\beta}(t,u)\,du\right| + \frac{1}{\sqrt{\varepsilon}}\left|\int_0^t g_2^\varepsilon(u)\tilde{\beta}(t,u)\,du\right|.$$

By virtue of the definition of $g_2(\cdot)$,

$$|g_2^\varepsilon(t)| = \int_0^t O(\varepsilon + \exp(-\kappa_0(t-s)/\varepsilon))\,ds = O(\varepsilon(t+1)) \quad \text{for all } 0 \le t \le T.$$

Consequently,

$$\left|\int_0^t g_2^\varepsilon(u)\tilde{\beta}(t,u)\,du\right| \le |\tilde{\beta}|_T \sup_{0\le t\le T}\int_0^t O(\varepsilon(u+1))\,du$$

$$\le K|\tilde{\beta}|_T(T^2+T)O(\varepsilon). \tag{3.22}$$

The estimate (3.21) then yields

$$\left|\int_0^t g_2^\varepsilon(u)\tilde{\beta}(t,u)\,du\right| \le (1+T)^3|\beta|_T(1+|\hat{Q}|_T)O(\varepsilon).$$

https://doi.org/10.1017/S0269964800144080 Published online by Cambridge University Press

Therefore, there exists an $\varepsilon_0 > 0$, such that for all $0 < \varepsilon \le \varepsilon_0$,

$$\begin{aligned}\exp\left(\frac{\theta_T}{(T+1)^3}\sup_{0\le t\le T}\left|\int_0^t \frac{g_2^\varepsilon(s)\tilde{\beta}(s)}{\sqrt{\varepsilon}}\,ds\right|\right) &\le \exp(O(\sqrt{\varepsilon})\min(1,\kappa_0)/K_T)\\ &\le \exp(O(\sqrt{\varepsilon_0})\min(1,\kappa_0))\\ &\le K.\end{aligned}$$

Similarly, we can show that

$$\exp\left(\frac{\theta_T}{(T+1)^3}\sup_{0\le t\le T}\left|\int_0^t \frac{g_1^\varepsilon(s)\tilde{\beta}(s)}{\sqrt{\varepsilon}}\,ds\right|\right) \le K.$$

This completes the proof of Theorem 3.1. ■

Remark 3.2: If a Markov chain contains only recurrent states, there are no transient states in the definition of $n^\varepsilon(t)$. Therefore, the result of Theorem 3.1 is consistent with what was obtained for a Markov chain containing only the recurrent states.

4. APPLICATIONS AND SOME RAMIFICATIONS

4.1. Applications

To study the asymptotic behavior of $\{n^\varepsilon(\cdot)\}$, an important step is to verify its tightness. In [16], we obtained such a result by using a tightness criterion (see [5] and [8]). In deriving such tightness, one often needs to verify that $n^\varepsilon(t)$ is bounded in probability. As an application of Theorem 3.1, we show in what follows that such a probabilistic boundedness can be obtained.

THEOREM 4.1: *Under the conditions of Theorem* 3.1, *for any* $\eta > 0$, *there exists a* $K_\eta > 0$ *such that*

$$\sup_\varepsilon P(|n^\varepsilon(t)| > K_\eta) < \eta. \tag{4.1}$$

PROOF: Denote $c_T = \theta_T/(T+1)^3$. By using the results of Theorem 3.1 together with the familiar Markov inequality, we have

$$\begin{aligned}P(|n^\varepsilon(t)| > K_\eta) &= P(c_T|n^\varepsilon(t)| > c_T K_\eta)\\ &\le P\left(c_T \sup_{0\le t\le T}|n^\varepsilon(t)| > c_T K_\eta\right)\\ &\le P\left(\exp\left(c_T \sup_{0\le t\le T}|n^\varepsilon(t)|\right) > \exp(c_T K_\eta)\right)\\ &\le \exp(-c_T K_\eta) E\exp\left(c_T \sup_{0\le t\le T}|n^\varepsilon(t)|\right)\\ &\le K\exp(-c_T K_\eta).\end{aligned}$$

https://doi.org/10.1017/S0269964800144080 Published online by Cambridge University Press

Note that the right-hand side is independent of ε. Thus, we also have

$$\sup_{\varepsilon} P(|n^{\varepsilon}(t)| > K_{\eta}) \leq K \exp(-c_T K_{\eta}). \tag{4.2}$$

To make the right-hand side of (4.2) to be less than η, it suffices to have $K_{\eta} > (1/c_T)\log(K/\eta)$. Such a choice then leads to (4.1). Hence, the theorem is established. ■

Remark 4.2: What has been proved in fact is stronger than (4.1). We have showed that the bound is uniform in $t \in [0, T]$.

Note that in view of the exponential bounds obtained, any finite moment of $|n^{\varepsilon}(t)|$ exists. Another consequence of Theorem 3.1 is the following moment bounds.

THEOREM 4.3: *Assume the conditions of Theorem* 3.1. *Then, for any* $0 < k < \infty$,

$$E\Big(\sup_{0\leq t\leq T} |n^{\varepsilon}(t)|\Big)^k \leq \frac{Kk!}{(c_T)^k}, \tag{4.3}$$

where K is given in Theorem 3.1.

PROOF: Observe that for any $z \in \mathbb{R}$,

$$\exp(z) = \sum_{k=0}^{\infty} \frac{z^k}{k!}.$$

Thus,

$$\Big(c_T \sup_{0\leq t\leq T} |n^{\varepsilon}(t)|\Big)^k \leq k! \exp\Big(c_T \sup_{0\leq t\leq T} |n^{\varepsilon}(t)|\Big) \quad \text{w.p.1.}$$

Taking the expectation and using Theorem 3.1,

$$E\Big(c_T \sup_{0\leq t\leq T} |n^{\varepsilon}(t)|\Big)^k \leq k! E \exp\Big(c_T \sup_{0\leq t\leq T} |n^{\varepsilon}(t)|\Big) \leq Kk!.$$

Thus, (4.3) follows. ■

4.2. Ramification

The upper bounds obtained in Theorem 3.1 can be improved if the generators are time invariant. In fact, if $\tilde{Q}(t) = \tilde{Q}$ and $\hat{Q}(t) = Q$, the quasistationary distributions of $\tilde{Q}^i$ for $i \leq l$ are also independent of t, so $\nu_*(t) = \nu_*$. Let

$$Q_0 = \hat{Q}\tilde{\mathbb{1}}_*\nu_*.$$

By induction, it is readily seen that

$$\hat{Q}\tilde{\mathbb{1}}_*(\bar{Q}_*)^k\nu_* = Q_0^{k+1} \quad \text{for } k \geq 0. \tag{4.4}$$

https://doi.org/10.1017/S0269964800144080 Published online by Cambridge University Press

In view of Remark 2.3, rewriting $P^{(0)}(t,s)$ in terms of $\Theta_*(t,s)$ via the use of (2.16) and using (4.4),

$$\begin{aligned}\hat{Q}P^{(0)}(t,s) &= \hat{Q}\tilde{\mathbb{1}}_* \exp(\bar{Q}_*(t-s))\nu_* \\ &= \hat{Q}\tilde{\mathbb{1}}_* \sum_{k=0}^{\infty} \frac{(t-s)^k}{k!}(\bar{Q}_*)^k\nu_* \\ &= \sum_{k=0}^{\infty} \frac{(t-s)^k}{k!}\hat{Q}\tilde{\mathbb{1}}_*(\bar{Q}_*)^k\nu_* \\ &= \sum_{k=0}^{\infty} \frac{(t-s)^k}{k!}(Q_0)^{k+1} \\ &= Q_0 \exp(Q_0(t-s)).\end{aligned}$$

Let $\phi^\varepsilon(t) = (I^\varepsilon(t) - \bar{I}^\varepsilon(t)) - g^\varepsilon(t)$. Then, $\phi^\varepsilon(\cdot)$ satisfies

$$\phi^\varepsilon(t) = -\int_0^t (\phi^\varepsilon(s) + g^\varepsilon(s))Q_0 \exp(Q_0(t-s))\,ds.$$

The solution is given by

$$\phi^\varepsilon(t) = -\int_0^t g^\varepsilon(s)Q_0\,ds.$$

Write $I^\varepsilon(t) - \bar{I}^\varepsilon(t)$ in terms of $\phi^\varepsilon(t)$ and $g^\varepsilon(t)$. The rest of the detail is similar to that of Theorem 3.1. We thus obtain the following result.

THEOREM 4.4: *In addition to the conditions of Theorem* 3.1, *assume* $\tilde{Q}(t) = \tilde{Q}$ *and* $\hat{Q}(t) = \hat{Q}$. *Then, there exist positive constants* ε_0 *and* K *such that for* $0 < \varepsilon \le \varepsilon_0$ *and* $T \ge 0$,

$$E\exp\left(\frac{\theta_T}{(T+1)^{3/2}} \sup_{0\le t\le T} |n^\varepsilon(t)|\right) \le K.$$

5. FURTHER REMARKS

The error bounds obtained can be used to develop asymptotic or nearly optimal controls of systems involving singularly perturbed Markov chains. They are suitable for infinite horizon control problems under discounted cost and/or risk-sensitive control problems.

https://doi.org/10.1017/S0269964800144080 Published online by Cambridge University Press

In lieu of (1.3), we may consider

$$\tilde{Q}(t) = \begin{pmatrix} \tilde{Q}^1(t) & & & & \\ & \ddots & & & \\ & & \tilde{Q}^l(t) & & \\ & & & 0_{m_a \times m_a} & \\ \tilde{Q}^1_*(t) & \dots & \tilde{Q}^l_*(t) & \tilde{Q}^a_*(t) & \tilde{Q}_*(t) \end{pmatrix}. \tag{5.1}$$

Not only are recurrent and transient states included, but absorbing states are also in the collection.

In addition to the notation used earlier, denote by $\mathcal{M}_a = \{s_{a1}, \dots, s_{am_a}\}$ the subspace corresponding to the absorbing states. Then, we can write the entire state space as a union of

$$\mathcal{M} = \mathcal{M}_1 \cup \mathcal{M}_2 \cup \cdots \cup \mathcal{M}_l \cup \mathcal{M}_a \cup \mathcal{M}_*.$$

We will use the following convention: If $\mathcal{M}_a = \varnothing$ and $\mathcal{M}_* = \varnothing$, then (5.1) reduces to the case that all states are recurrent. If only one of them is empty, then we get the case of inclusion of either transient states or absorbing states.

Note that an absorbing state can be considered to be a recurrent state. There are m_a states corresponding to the zero matrix 0_a. To proceed, one may rewrite $0_{m_a \times m_a} = \operatorname{diag}(0,0,\dots,0) \in \mathbb{R}^{m_a \times m_a}$, where each diagonal element in the matrix is a scalar 0. Thus, in addition to the l groups of recurrent states, we have m_a groups of additional recurrent states (each of which has only one state). From an application point of view, under various scenarios (e.g., in reliability analysis and competing risk modeling [14]), one wants to consider the inclusion of absorbing states separately.

We then replace $\hat{\mathbb{1}}_*(t)$ *by*

$$\hat{\mathbb{1}}(t) = \begin{pmatrix} \mathbb{1}_{m_1} & & & & \\ & \ddots & & & \\ & & \mathbb{1}_{m_l} & & \\ & & & I_{m_a} & \\ a_{m_1} & \dots & a_{m_l} & a_{m_a} & 0_{m_* \times m_*} \end{pmatrix},$$

where a_{m_i} for $i = 1, \dots, l$ are as earlier, and I_{m_a} denotes an $m_a \times m_a$ identity matrix, and use the partitioned matrix

$$\hat{Q}(t) = \begin{pmatrix} \hat{Q}^{1r}(t) & \hat{Q}^{1a}(t) & \hat{Q}^{1*}(t) \\ \hat{Q}^{2r}(t) & \hat{Q}^{2a}(t) & \hat{Q}^{2*}(t) \\ \hat{Q}^{3r}(t) & \hat{Q}^{3a}(t) & \hat{Q}^{3*}(t) \end{pmatrix},$$

https://doi.org/10.1017/S0269964800144080 Published online by Cambridge University Press

where $\hat{Q}^{1r}(t) \in \mathbb{R}^{(m_1+\cdots+m_l)\times(m_1+\cdots+m_l)}$, $\hat{Q}^{2a}(t) \in \mathbb{R}^{m_a\times m_a}$, $\hat{Q}^{3*}(t) \in \mathbb{R}^{m_*\times m_*}$, and the rest of $\hat{Q}^{ij}(t)$ have compatible dimensions. Define

$$\bar{Q}_0(t) = \operatorname{diag}(\nu^1(t),\ldots,\nu^l(t), I_{m_a\times m_a}) \times \begin{pmatrix} \hat{Q}^{1r}(t)\tilde{\mathbb{1}} + \hat{Q}^{1*}(t)A_* & \hat{Q}^{1a}(t) + \hat{Q}^{1*}(t)a_{m_a} \\ \hat{Q}^{2r}(t)\tilde{\mathbb{1}} + \hat{Q}^{2*}(t)A_* & \hat{Q}^{2a}(t) + \hat{Q}^{2*}(t)a_{m_a} \end{pmatrix},$$

$$\bar{Q}_u(t) = \operatorname{diag}(\bar{Q}_0(t), 0_{m_*\times m_*}).$$

Then, the limit transition matrix takes the form

$$P(t,t_0) = \hat{\mathbb{1}}(t)\,\hat{\Theta}(t,t_0)\hat{\nu}(t),$$

where

$$\hat{\Theta}(t,t_0) = \operatorname{diag}(\Theta_0(t,t_0), I_{m_*\times m_*}), \qquad \hat{\nu}(t) = \operatorname{diag}(\nu(t), I_{m_a\times m_a}),$$

and $\Theta_0(t,t_0)$ satisfies

$$\frac{d\Theta_0(t,t_0)}{dt} = \Theta_0(t,t_0)\bar{Q}_0(t),$$
$$\Theta_0(t_0,t_0) = I_{(l+m_a)\times(l+m_a)}.$$

Assuming the same conditions and using essentially the same argument as presented in the previous sections, we can show that the desired exponential bounds on the errors continue to hold.

Acknowledgments

The research of G.Y. was supported in part by the National Science Foundation (grant DMS-9877090). The research of Q.Z. was supported in part by the Office of Naval Research (grant N000014-96-1-0263).

References

1. Courtois, P.J. (1977). *Decomposability: Queueing and computer system applications.* New York: Academic Press.
2. Davis, M.H.A. (1993). *Markov models and optimization.* London: Chapman & Hall.
3. Delebecque, F. & Quadrat, J. (1981). Optimal control for Markov chains admitting strong and weak interactions. *Automatica* 17: 281–296.
4. Elliott, R.J. (1982). *Stochastic calculus and applications.* New York: Springer-Verlag.
5. Ethier, S.N. & Kurtz, T.G. (1986). *Markov processes: Characterization and convergence.* New York: Wiley.
6. Khasminskii, R.Z., Yin, G., & Zhang, Q. (1996). Asymptotic expansions of singularly perturbed systems involving rapidly fluctuating Markov chains. *SIAM Journal of Applied Mathematics* 56: 277–293.
7. Khasminskii, R.Z., Yin, G., & Zhang, Q. (1997). Constructing asymptotic series for probability distribution of Markov chains with weak and strong interactions. *Quarterly Applied Mathematics* 55: 177–200.

https://doi.org/10.1017/S0269964800144080 Published online by Cambridge University Press

8. Kushner, H.J. (1984). *Approximation and weak convergence methods for random processes, with applications to stochastic systems theory.* Cambridge, MA: MIT Press.
9. Pan, Z.G. & Başar, T (1995). H^{∞}-control of Markovian jump linear systems and solutions to associated piecewise-deterministic differential games. In G.J. Olsder (ed.), *New trends in dynamic games and applications.* Boston: Birkhäuser, pp. 61–94.
10. Pervozvanskii, A.A. & Gaitsgori, V.G. (1988). *Theory of suboptimal decisions: Decomposition and aggregation.* Dordrecht: Kluwer.
11. Phillips, R.G. & Kokotovic, P.V. (1981). A singular perturbation approach to modelling and control of Markov chains. *IEEE Transactions on Automated Control* AC-26: 1087–1094.
12. Sethi, S.P. & Zhang, Q. (1994). *Hierarchical decision making in stochastic manufacturing systems.* Boston: Birkhäuser.
13. Simon, H.A. & Ando, A. (1961). Aggregation of variables in dynamic systems. *Econometrica* 29: 111–138.
14. Thompson, W.A., Jr. (1988). *Point process models with applications to safety and reliability.* New York: Chapman & Hall.
15. Yin, G. & Zhang, Q. (1998). *Continuous-time Markov chains and applications: A singular perturbation approach.* New York: Springer-Verlag.
16. Yin, G., Zhang, Q., & Badowski, G. (2000). Asymptotic properties of a singularly perturbed Markov chain with inclusion of transient states. *Annals of Applied Probability*, in press.
17. Yong, J. & Zhou, X.Y. (1999). *Stochastic controls: Hamiltonian systems and HJB equations.* New York: Springer-Verlag.
18. Zhang, Q. & Yin, G. (1996). A central limit theorem for singularly perturbed nonstationary finite state Markov chains. *Annals of Applied Probability* 6: 650–670.

https://doi.org/10.1017/S0269964800144080 Published online by Cambridge University Press