The Annals of Applied Probability
2000, Vol. 10, No. 2, 549–572

ASYMPTOTIC PROPERTIES OF A SINGULARLY PERTURBED MARKOV CHAIN WITH INCLUSION OF TRANSIENT STATES

BY G. YIN,[1] Q. ZHANG[2] AND G. BADOWSKI[3]

Wayne State University, University of Georgia and Wayne State University

This work is concerned with aggregations in a singularly perturbed Markov chain having a finite state space and fast and slow motions. The state space of the underlying Markov chain can be decomposed into several groups of recurrent states and a group of transient states. The asymptotic properties are studied through sequences of unscaled and scaled occupation measures. By treating the states within each recurrent class as a single state, an aggregated process is defined and shown to be convergent to a limit Markov chain. In addition, it is shown that a sequence of suitably rescaled occupation measures converges to a switching diffusion process weakly.

1. Introduction. This paper is concerned with asymptotic properties of singularly perturbed Markov chains having fast and slow components. By considering Markov chains with finite state spaces, we obtain weak convergence results for Markov chains with state space consisting of both recurrent and transient states.

Recent advances in manufacturing systems (see [13–16] and the references therein) lead to renewed interests to the understanding of singularly perturbed systems. In the aforementioned references, to design and to control systems with stochastic capacity and random demands, models using Markov chains are used frequently. To reflect the reality of different classes of states involved, one often introduces a small parameter and models the systems under consideration as a singularly perturbed Markov chain. To illustrate, we give two examples.

EXAMPLE 1.1. Consider a two-machine flow shop with machines subject to breakdown and repair. Model the production capacity of the machines as a finite-state Markov chain. When the machine is up, it can produce parts with production rate $u(t)$; when the machine is under repair, nothing can be produced. For simplicity, suppose each of the machines is either in operating condition (denoted by 1) or under repair (denoted by 0). Then the capacity of the flowshop $\alpha(\cdot)$ is a four-state Markov chain with state space $\{(1,1),(0,1),(1,0),(0,0)\}$. Suppose the first machine breaks down much

Received January 1998; revised November 1999.

[1]Supported in part by NSF Grants DMS-95-29738 and DMS-98-77090.

[2]Supported in part by Office of Naval Research Grant N00014-96-1-0263, in part by USAF Grant F30602-99-2-0548 and in part by a University of Georgia Faculty Research grant.

[3]Supported in part by Wayne State University.

AMS 1991 *subject classifications*. 60J27, 60B10, 34E05, 60F17.

Key words and phrases. singularly perturbed Markov chain, occupation measure, aggregation, weak convergence, switching diffusion.

more often than that of the second one. For small $\varepsilon > 0$, let $\alpha(\cdot) = \alpha^\varepsilon(\cdot)$ be generated by

$$Q^\varepsilon(t) = \frac{1}{\varepsilon}\widetilde{Q}(t) + \widehat{Q}(t), \tag{1}$$

with $\widetilde{Q}(\cdot)$ and $\widehat{Q}(\cdot)$ given by

$$\widetilde{Q}(t)=\begin{pmatrix} -\lambda_1(t) & \lambda_1(t) & 0 & 0 \\ \mu_1(t) & -\mu_1(t) & 0 & 0 \\ 0 & 0 & -\lambda_1(t) & \lambda_1(t) \\ 0 & 0 & \mu_1(t) & -\mu_1(t) \end{pmatrix},$$
$$\widehat{Q}(t)=\begin{pmatrix} -\lambda_2(t) & 0 & \lambda_2(t) & 0 \\ 0 & -\lambda_2(t) & 0 & \lambda_2(t) \\ \mu_2(t) & 0 & -\mu_2(t) & 0 \\ 0 & \mu_2(t) & 0 & -\mu_2(t) \end{pmatrix},$$

where $\lambda_i(\cdot)$ and $\mu_i(\cdot)$ are the repair and breakdown rates, respectively. The matrices $\widetilde{Q}(t)$ and $\widehat{Q}(t)$ are themselves generators of Markov chains. Denote the production rates of the two machines by $u_i(\cdot)$ and assume that they are upper bounded by the machine capacities. The objective is then to choose the control within an admissible class so as to minimize an expected cost $E\int_0^T G(x(t), u(t), \alpha^\varepsilon(t))\,dt$, where $x(t)$ denotes the inventory level and $G(\cdot)$ is an appropriate function. This seemingly not so complex problem turns out to be very difficult; an analytical solution is virtually impossible to obtain. Thus one seeks approximate solutions that are asymptotically or nearly optimal.

EXAMPLE 1.2. Let $\alpha^\varepsilon(t)$ be a singularly perturbed Markov chain with a finite state space $\mathscr{M}$ and a generator $Q^\varepsilon(t)$ given by (1) with $\widetilde{Q}(t) = \operatorname{diag}(\widetilde{Q}^1, \dots, \widetilde{Q}^l(t))$ such that each $\widetilde{Q}(t)$ is a weakly irreducible generator (see Yin and Zhang [17], Chapter 2, for a definition). Consider a stochastic dynamical system with the state $x^\varepsilon(t) \in \mathbb{R}^n$, and control $u(t) \in \Gamma \subset \mathbb{R}^{n_1}$. Let $f(\cdot,\cdot,\cdot)\colon\ \mathbb{R}^n \times \mathbb{R}^{n_1} \times \mathscr{M} \mapsto \mathbb{R}^n$, $G(\cdot,\cdot,\cdot)\colon\ \mathbb{R}^n \times \mathbb{R}^{n_1} \times \mathscr{M} \mapsto \mathbb{R}$. Consider the problem

$$\begin{aligned} &\text{minimize } J^\varepsilon(u(\cdot)) = E\int_0^T G(x^\varepsilon(t), u(t), \alpha^\varepsilon(t))\,dt,\\ &\text{subject to } \frac{dx^\varepsilon(t)}{dt} = f(x^\varepsilon(t), u(t), \alpha^\varepsilon(t)), \qquad x^\varepsilon(0) = x. \end{aligned}$$

Let $\{s_{i1}, \dots, s_{im_i}\}$, $i = 1, \dots, l$, denote the states of $\alpha^\varepsilon(\cdot)$ corresponding to $\widetilde{Q}^i(t)$. Suppose that $\widetilde{m}(\cdot)$ is a relaxed control representation for $u(\cdot)$. Then the cost function and system equation can be written as

$$J^{\varepsilon}(\widetilde{m}^{\varepsilon}) = E\left(\sum_{i=1}^{l}\sum_{j=1}^{m_i}\int_0^T \int G(x^{\varepsilon}(t), v, s_{ij})\widetilde{m}_t^{\varepsilon}(v)\, dt\right),$$

$$x^{\varepsilon}(t) = x + \sum_{i=1}^{l}\sum_{j=1}^{m_i}\int_0^t \int f(x^{\varepsilon}(s), v, s_{ij})\widetilde{m}_s^{\varepsilon}(dv) I_{\{\alpha^{\varepsilon}(s)=s_{ij}\}} ds.$$

It can be shown (see [17], Chapter 9) there is an associated limit problem:

$$\text{minimize } J(m) = E\sum_{i=1}^{l}\sum_{j=1}^{m_i}\int_0^T \int G(x(t), v, s_{ij})\widetilde{m}_t(dv)\nu_j^i(t) I_{\{\bar{\alpha}(t)=i\}}\, dt,$$

$$\text{subject to } x(t) = x + \sum_{i=1}^{l}\sum_{j=1}^{m_i}\int_0^t \int f(x(s), v, s_{ij})\widetilde{m}_s(dv)\nu_j^i(s) I_{\{\bar{\alpha}(s)=i\}} ds,$$

where $(\nu_1^i(t), \dots, \nu_{m_i}^i(t))$ is the quasistationary distribution (the definition is given in what follows) corresponding to $\widetilde{Q}^i(t)$. In the limit problem, the dynamics and the cost are averaged out w.r.t. the quasistationary measures of $\alpha^{\varepsilon}(\cdot)$. One can then proceed to use the limit to design a nearly optimal control of the original problem.

The preceding examples show that, crucial to many important applications, we need to have a thorough understanding of the structure and asymptotic properties of the underlying Markov chains. Khasminskii, Yin and Zhang [8] use matched asymptotic expansion to establish the convergence of a sequence of the probability distribution vectors. Singularly perturbed Markov chains with recurrent states, naturally divisible into a number of classes are then treated in [9]. This line of research has been continued in [18, 19] with emphasis on asymptotic normality and structural properties. We have further derived asymptotic expansions for Markov chains with the inclusion of transient states and absorbing states and asymptotic distributions for Markov chains having recurrent states in [17].

This work shows our continuing effort in studying asymptotic properties of singularly perturbed Markov chains. Our main goal in this paper is to treat Markov chains with both recurrent states and transient states, where the recurrent states belong to several weakly irreducible classes. We treat nonstationary (or nonhomogeneous) Markov chains. Such a consideration is especially important when we are dealing with those control and optimization problems arising in queueing systems, system reliability and production planning, in which the generators of the underlying Markov chains are often "action" (or control) dependent. The essence of our approach is aggregation and the main focus is on unscaled and scaled occupation measures. The results to be presented include an estimate on the mean square error of a sequence of the centered occupation measures, weak convergence of an aggregated process and weak convergence of a normalized sequence of occupation measures to a switching diffusion process. We use martingale formulation and perturbed test function methods to derive the desired asymptotic properties.

The rest of the paper is arranged as follows. Precise problem formulation is given in Section 2. Section 3 proceeds with the study of unscaled occupation measures and derives a mean square estimate on a sequence of centered occupation measure. Section 4 deals with weak convergence of aggregated Markov chains and Section 5 further exploits the limit behavior of a sequence of normalized occupation measures. Finally, we close this paper with additional remarks in Section 6.

Suppose that $\alpha(\cdot)$ is a continuous-time Markov chain. In this paper, all we need is the following modified definitions (see [17], Chapter 2). Let $\mathscr{M} = \{1, \dots, m\}$. For $i,\ j \in \mathscr{M}$, denote $Q(t) = (q_{ij}(t))$, for $t \geq 0$. For any real-valued function f on $\mathscr{M}$ and $i \in \mathscr{M}$, write

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

The matrix $Q(t)$, $t \geq 0$, is a *generator* of $\alpha(\cdot)$ if $q_{ij}(t)$ is continuous for all $i, j \in \mathscr{M}$ and $t \geq 0$, $q_{ij}(t) \geq 0$ for $j \neq i$ and $q_{ii}(t) = -\sum_{j\neq i} q_{ij}(t)$, $t \geq 0$, and for all bounded real-valued functions f defined on $\mathscr{M}$

$$f(\alpha(t)) - \int_0^t Q(\varsigma)f(\cdot)(\alpha(\varsigma))\,d\varsigma \tag{2}$$

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

$$\begin{aligned} \nu(t)Q(t) &= 0, \\ \sum_{i=1}^{m} \nu_i(t) &= 1 \end{aligned} \tag{3}$$

has a unique solution $\nu(t) = (\nu_1(t), \dots, \nu_m(t))$ and $\nu(t) \geq 0$. (Here and hereafter, for a vector v, $v \geq 0$ means that each of its components $v_i \geq 0$.) The nonnegative solution of (3) is termed a quasistationary distribution. For more general approach to nonstationary Markov chains, we refer the reader to Davis [1].

Let $\alpha(t)$, $t \geq 0$, be a Markov chain generated by $Q(t)$. Then it is well known that the probability distribution vector

$$p(t) = (P(\alpha(t) = 1), \dots, P(\alpha(t) = m)) \in \mathbb{R}^{1\times m}$$

satisfies the forward equation

$$\dot{p}(t) = p(t)Q(t), \qquad p(0) = p_0 \tag{4}$$

such that $p_0 = (p_{0,\ 1}, \dots, p_{0,\ m})$ satisfying $p_{0,\ i} \geq 0$ for all $i = 1, \dots, m$ and $p_0 \mathbb{1} = \sum_{i=1}^{m} p_{0,\ i} = 1$, where $\mathbb{1} = (1, \dots, 1)' \in \mathbb{R}^{m\times 1}$.

2. Problem formulation. We arrange this section in three parts. The first part presents the formulation, the second part gives some preliminaries and the third part concentrates on aggregation of the underlying processes.

2.1. *Formulation.* Let $\varepsilon > 0$ be a small parameter. For some $T > 0$ and all $t \in [0,\ T]$, let $\alpha^\varepsilon(t)$ be a Markov chain generated by (1), where $\widetilde{Q}(t)/\varepsilon$ governs the rapidly changing part and $\widehat{Q}(t)$ describes the slowly changing components. Assume

$$\widetilde{Q}(t) = \begin{pmatrix} \widetilde{Q}^1(t) & & & \\ & \ddots & & \\ & & \widetilde{Q}^l(t) & \\ \widetilde{Q}^1_*(t) & \cdots & \widetilde{Q}^l_*(t) & \widetilde{Q}_*(t) \end{pmatrix} \tag{5}$$

such that, for each $t \in [0,\ T]$ and each $k = 1, \ldots, l$, $\widetilde{Q}^k(t)$ is a generator with dimension $m_k \times m_k$, $\widetilde{Q}_*(t)$ is an $m_* \times m_*$ matrix, $\widetilde{Q}^k_*(t) \in \mathbb{R}^{m_* \times m_k}$ and

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

The state space of the underlying Markov chain is given by

$$\mathscr{M} = \mathscr{M}_1 \cup \cdots \cup \mathscr{M}_l \cup \mathscr{M}_*$$

where $\mathscr{M}_k = \{s_{k1}, \ldots, s_{km_k}\}$ are recurrent classes that are weakly irreducible and $\mathscr{M}_* = \{s_{*1}, \ldots, s_{*m_*}\}$ includes the transient states. In what follows, for a vector $y \in \mathbb{R}^{1\times m}$, we often use a partitioned form $y = (y^1, \ldots, y^l, y^*)$, where $y^i \in \mathbb{R}^{1\times m_i}$ for each $i = 1, \ldots, l$ and $i = *$.

Our formulation is inspired by the work of Phillips and Kokotovic [14] and Delebecque and Quadrat [2]; see also the recent work of Pan and Başar [13], in which the authors treated time-invariant $\widetilde{Q}$ matrix of a similar form.

In the literature, many people have studied singular perturbed systems. Notably, Khasminskii [7] considered two-time-scale singularly perturbed diffusions of the form

$$\begin{aligned} dx^\varepsilon(t) &= A_1(x^\varepsilon(t), y^\varepsilon(t))\,dt + \sigma_1(x^\varepsilon(t), y^\varepsilon(t))\,dw_1(t),\\ dy^\varepsilon(t) &= \frac{1}{\varepsilon}A_2(x^\varepsilon(t), y^\varepsilon(t))\,dt + \frac{1}{\sqrt{\varepsilon}}\sigma_2(x^\varepsilon(t), y^\varepsilon(t))\,dw_2(t), \end{aligned}$$

for independent Brownian motions w_1 and w_2. In [6], Chapter 6, Friedlin and Wentzell considered the singularly perturbed dynamic system

$$dx^\varepsilon(t) = b(x^\varepsilon(t))\,dt + \varepsilon\,dw_t$$

and the associated Markov chains with state space being the equivalent classes generated by the unperturbed ($\varepsilon = 0$) dynamic system. Ethier and Kurtz [4], Chapter 12, and Kushner [12] treated systems of the form

$$\dot{x}^\varepsilon(t) = f\left(x^\varepsilon(t), y\left(\frac{t}{\varepsilon^2}\right)\right) + \frac{1}{\varepsilon}g\left(x^\varepsilon(t), y\left(\frac{t}{\varepsilon^2}\right)\right)$$

and the like. The driving noise $y(\cdot)$ is known as a wideband noise; see [12] for various applications of controlled dynamic systems. The problems considered in this paper bare the spirit of singular perturbation, but are very different from the aforementioned references.

2.2. *Conditions and preliminaries.* We assume the following conditions.

(A1) For all $t \in [0, T]$ and $k = 1, \dots, l$, $\widetilde{Q}^k(t)$ are weakly irreducible, and $\widetilde{Q}_*(t)$ is Hurwitz (i.e., all of its eigenvalues have negative real parts).

(A2) The matrix $\widetilde{Q}(\cdot)$ is differentiable on $[0, T]$ and its derivative is Lipschitz. Moreover, $\widehat{Q}(\cdot)$ is Lipschitz continuous on $[0, T]$.

Use the partition

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

where $\widehat{Q}^{11}(t) \in \mathbb{R}^{(m-m_*)\times(m-m_*)}$, $\widehat{Q}^{12}(t) \in \mathbb{R}^{(m-m_*)\times m_*}$, $\widehat{Q}^{21}(t) \in \mathbb{R}^{m_*\times(m-m_*)}$ and $\widehat{Q}^{22}(t) \in \mathbb{R}^{m_*\times m_*}$, and write

$$\begin{aligned} \overline{Q}_*(t) &= \operatorname{diag}(\nu^1(t), \dots, \nu^l(t))\Big(\widehat{Q}^{11}(t)\tilde{\mathbb{1}} + \widehat{Q}^{12}(t)(a_{m_1}(t), \dots, a_{m_l}(t))\Big), \\ \overline{Q}(t) &= \operatorname{diag}(\overline{Q}_*(t), 0_{m_*\times m_*}), \end{aligned} \tag{6}$$

where

$$\tilde{\mathbb{1}} = \operatorname{diag}(\mathbb{1}_{m_1}, \dots, \mathbb{1}_{m_l}), \qquad \mathbb{1}_{m_j} = (1, \dots, 1)' \in \mathbb{R}^{m_j\times 1}$$

and

$$a_{m_i}(t) = -\widetilde{Q}_*^{-1}(t)\widetilde{Q}_*^i(t)\mathbb{1}_{m_i} \quad \text{for } i = 1, \dots, l. \tag{7}$$

In what follows, if $a_{m_i}(t)$ is time independent, we will simply write it as a_{m_i}. It will be shown that $\overline{Q}_*(t)$ is a generator of a Markov chain corresponding to the "limit" of the aggregated process.

REMARK. The requirement on $\widetilde{Q}_*(t)$ in (A1) implies that the corresponding states are transient. The Hurwitzian property also has the following interesting implication: For each $t \in [0, T]$, and each $i = 1, \dots, l$, $a_{m_i}(t) = (a_{m_i,1}(t), \dots, a_{m_i,m_*}(t))' \in \mathbb{R}^{m_*\times 1}$. Then

$$a_{m_i,j}(t) \geq 0 \quad \text{and} \quad \sum_{i=1}^{l} a_{m_i,j}(t) = 1 \tag{8}$$

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

To see this, note that, for each $t \in [0,\ T]$,

$$\int_0^\infty \exp(\widetilde{Q}_*(t)s)\,ds = -\widetilde{Q}_*^{-1}(t),$$

which has nonnegative components (see [17], page 147). From the definition it follows that $a_{m_i}(t) \geq 0$. Furthermore,

$$\sum_{i=1}^{l} a_{m_i}(t) = -\widetilde{Q}_*^{-1}(t)\sum_{i=1}^{l}\widetilde{Q}_*^{i}(t)\mathbb{1}_{m_i} = \Big(-\widetilde{Q}_*^{-1}(t)\Big)\Big(-\widetilde{Q}_*(t)\Big)\mathbb{1}_{m_*} = \mathbb{1}_{m_*}.$$

Thus (8) follows.

Denote the transition probability of $\alpha^\varepsilon(\cdot)$ by $P^\varepsilon(t, t_0)$, that is, $P^\varepsilon(t, t_0) = (P(\alpha^\varepsilon(t) = j|\alpha^\varepsilon(t_0) = i))$. The following results concern the asymptotic expansion of the probability distribution and the associated transition probabilities. The proof of the first statement is in [17], Section 6.4, and the proof of the second statement is a slight modification of the first. We thus omit the proofs.

PROPOSITION 2.1. *Assume conditions* (A1) *and* (A2) *are satisfied. Then the following assertions hold*:

A. *For the probability distribution vector*

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

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*

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

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* $t \geq t_0 \geq 0$,

$$P^\varepsilon(t, t_0) = P^0(t, t_0) + O\left(\varepsilon + \exp(-\kappa_0(t - t_0)/\varepsilon)\right), \tag{9}$$

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

$$P^0(t, t_0) = \begin{pmatrix} \mathbb{1}_{m_1}\nu^1(t)\theta_{11}(t, t_0) & \cdots & \mathbb{1}_{m_1}\nu^l(t)\theta_{1l}(t, t_0) & 0_{m_*} \\ & \vdots & & \\ \mathbb{1}_{m_l}\nu^1(t)\theta_{l1}(t, t_0) & \cdots & \mathbb{1}_{m_l}\nu^l(t)\theta_{ll}(t, t_0) & 0_{m_*} \\ \nu^1(t)\theta^*_{11}(t, t_0) & \cdots & \nu^l(t)\theta^*_{1l}(t, t_0) & 0_{m_*} \\ & \vdots & & \\ \nu^1(t)\theta^*_{m_*1}(t, t_0) & \cdots & \nu^l(t)\theta^*_{m_*l}(t, t_0) & 0_{m_*} \end{pmatrix},$$

and $\Theta(t, t_0)$ *satisfies the differential equation*

$$\frac{\partial\Theta(t, t_0)}{\partial t} = \Theta(t, t_0)\overline{Q}_*(t), \tag{10}$$

with $\big(\theta_{ij}(t_0,\ t_0)\big) = I \in \mathbb{R}^{l\times l}$ *and* $(\theta^*_{ij}(t_0,\ t_0)) = (a_{m_1}(t_0), \dots, a_{m_l}(t_0)) \in \mathbb{R}^{m_*\times l}$.

2.3. *Aggregation.* Looking at the form of the generator (5), naturally, we can aggregate the states in $\mathscr{M}_i$ (for $i = 1, \dots, l$) as one state. This aggregation leads to the definition of an aggregated process $\overline{\alpha}^\varepsilon(\cdot)$ with $\overline{\alpha}^\varepsilon(t) = i$ if $\alpha^\varepsilon(t) \in \mathscr{M}_i$ together with appropriate definition for $\alpha^\varepsilon(\cdot)$ on $\mathscr{M}_*$. Much of the rest of the paper is concerned with this aggregated process. Then we study further the properties of a scaled sequence of the occupation measures. Accompanying the results of asymptotic expansion, for each $j = 1, \dots, m_i$, we define a sequence of centered occupation measures by

$$O^\varepsilon_{ij}(t) = \begin{cases} \displaystyle\int_0^t \Big(I_{\{\alpha^\varepsilon(s)=s_{ij}\}} - \nu^i_j(s)I_{\{\alpha^\varepsilon(s)\in\mathscr{M}_i\}}\Big)\, ds, & \text{for } i = 1, \dots, l,\\ \displaystyle\int_0^t I_{\{\alpha^\varepsilon(s)=s_{*j}\}}\, ds, & \text{for } i = *, \end{cases} \tag{11}$$

where I_A is the indicator of the set A, and $\nu^i_j(t)$ denotes the jth component of the quasistationary distribution of $\nu^i(t)$. For notational convenience, here and hereafter "$*$" is treated as an index corresponding to those transient states, which we purposely choose to be distinct from the indices of the recurrent states. Note that by Proposition 2.1, the probability distribution corresponding to the states in $\mathscr{M}_*$ goes to 0 as $\varepsilon \to 0$. To give motivation of the use of the aggregated process, we present an example below.

EXAMPLE 2.1. Let $\alpha^\varepsilon(\cdot)$ denote a Markov chain generated by a time-independent generator $Q^\varepsilon = \widetilde{Q}/\varepsilon + \widehat{Q}$ such that $\widetilde{Q}$ has the structure as in (5) and $\widehat{Q}$ is another generator. Consider a class of hybrid linear quadratic Gaussian (LQG) systems consisting of a set of diffusions coupled by $\alpha^\varepsilon(\cdot)$:

$$\begin{aligned} dx(t) &= [A(\alpha^\varepsilon(t))x(t) + B(\alpha^\varepsilon(t))u(t)]\, dt + \sigma\, dw(t),\\ x(s) &= x, \qquad \text{for } s \le t \le T, \end{aligned} \tag{12}$$

where $x(t) \in \mathbb{R}^{n_1}$ is the state; $u(t) \in \mathbb{R}^{n_2}$ is the control; $A(i) \in \mathbb{R}^{n_1\times n_1}$ and $B(i) \in \mathbb{R}^{n_1\times n_2}$ are well defined and have finite values for $i \in \mathscr{M}$; and $w(\cdot)$ is a standard Brownian motion. Our objective is to find the optimal control $u(\cdot)$ so that the expected quadratic cost function

$$\begin{aligned} &J(s, x, \alpha, u(\cdot))\\ &\quad = E\bigg\{\int_s^T [x'(t)M(\alpha^\varepsilon(t))x(t) + u'(t)N(\alpha^\varepsilon(t))u(t)]\, dt + x'(T)Dx(T)\bigg\} \end{aligned} \tag{13}$$

is minimized, where E is the expectation given $\alpha^\varepsilon(s) = \alpha$ and $x(s) = x$, $M(i)$, $i = 1, \dots, m$, are symmetric nonnegative definite matrices, and $N(i)$, $i = 1, \dots, m$, and D are symmetric positive definite matrices.

By means of an optimal control approach ([5], Chapter 5), using the solution of the Riccati equation,

$$\begin{aligned} \dot{K}^\varepsilon(s,i) &= -K^\varepsilon(s,i)A(i) - A'(i)K^\varepsilon(s,i) - M(i) \\ &\quad + K^\varepsilon(s,i)B(i)N^{-1}(i)B'(i)K^\varepsilon(s,i) - Q^\varepsilon K^\varepsilon(s,\cdot)(i), \end{aligned} \tag{14}$$

with $K^\varepsilon(T,i) = D$, the optimal control $u^{\varepsilon,*}$ can be written as

$$u^{\varepsilon,*}(s,i,x) = -N^{-1}(i)B'(i)K^\varepsilon(s,i)x. \tag{15}$$

Using the classical approach, to find the optimal control, one has to solve the Riccati equations. However, in many problems in manufacturing, such solutions are very difficult to obtain due to the large dimensionality. Therefore, one has to resort to approximation schemes.

It is shown in [20] that, as $\varepsilon \to 0$, $K^\varepsilon(s, s_{kj}) \to \overline{K}(s,k)$ for $k = 1, \dots, l$, $j = 1, \dots, m_k$, $K^\varepsilon(s, s_{*j}) \to \overline{K}_*(s,j)$ for $j = 1, \dots, m_*$, uniformly on $[0, T]$, where

$$\overline{K}_*(s,j) = a_{m_1,j}\overline{K}(s,1) + \cdots + a_{m_l,j}\overline{K}(s,l),$$

and $\overline{K}(s,k)$ is the unique solution of

$$\begin{aligned} \dot{\overline{K}}(s,k) &= -\overline{K}(s,k)\overline{A}(k) - \overline{A}'(k)\overline{K}(s,k) - \overline{M}(k) \\ &\quad + \overline{K}(s,k)\overline{BN^{-1}B'}(k)\overline{K}(s,k) - \overline{Q}_*\,\overline{K}(s,\cdot)(k), \end{aligned} \tag{16}$$

with $\overline{K}(T,k) = D$ for $k = 1, \dots, l$. In the above, $\overline{A}$, $\overline{M}$ and $\overline{BN^{-1}B'}$ denote the averages w.r.t. the corresponding quasistationary distributions.

Note that the dimension of the limit Riccati equation is much smaller than that of the original one so the complexity is much reduced. Replacing K^ε by $\overline{K}$ in (15), we obtain a limit control $\overline{u}(s,\alpha,x)$. The asymptotic results of this paper yields that such a control is nearly optimal, that is,

$$\lim_{\varepsilon\to 0} |J^\varepsilon(s,\alpha,x,\overline{u}(\cdot)) - v^\varepsilon(s,\alpha,x)| = 0.$$

3. Occupation measures. This section is devoted to the study of a sequence of occupation measures. The main result is a mean square estimate that is stated in the following theorem.

THEOREM 3.1. *Assume* (A1) *and* (A2). *Then for each* $j = 1, \dots, m_i$,

$$E|O^\varepsilon_{ij}(t)|^2 = \begin{cases} O(\varepsilon), & \text{for } i = 1, \dots, l, \\ O(\varepsilon^2), & \text{for } i = *, \end{cases} \tag{17}$$

uniformly in $t \in [0, T]$.

REMARK. It is worthwhile to note the order of the estimates in (17). Although $E|O^{\varepsilon}_{ij}(s)|^2$ goes to zero for all i, j, it diminishes an order of magnitude faster for the states in the transient class. This property will be used when evaluating the limit covariance of a sequence of scaled occupation measures.

PROOF. For $i = 1, \dots, l$, $j = 1, \dots, m_i$ and $s, r \in [0, T]$, define

$$\Psi^{\varepsilon}_{1,ij}(s,r) = P\big(\alpha^{\varepsilon}(s) = s_{ij}, \alpha^{\varepsilon}(r) = s_{ij}\big) - \nu^i_j(s)P\big(\alpha^{\varepsilon}(s) \in \mathscr{M}_i, \alpha^{\varepsilon}(r) = s_{ij}\big),$$

$$\begin{aligned}\Psi^{\varepsilon}_{2,ij}(s,r) = {} & \nu^i_j(r)\nu^i_j(s)P\big(\alpha^{\varepsilon}(s) \in \mathscr{M}_i, \alpha^{\varepsilon}(r) \in \mathscr{M}_i\big)\\ & -\nu^i_j(r)P\big(\alpha^{\varepsilon}(s) = s_{ij}, \alpha^{\varepsilon}(r) \in \mathscr{M}_i\big).\end{aligned}$$

Then for $i = 1, \dots, l$,

$$\begin{aligned} E|O^{\varepsilon}_{ij}(t)|^2 &= \int_0^t \int_0^t \Big(\Psi^{\varepsilon}_{1,ij}(s,r) + \Psi^{\varepsilon}_{2,ij}(s,r)\Big)\, dr\, ds \\ &= \int_0^t \int_0^s (\Psi^{\varepsilon}_{1,ij}(s,r) + \Psi^{\varepsilon}_{2,ij}(s,r))\, dr\, ds \\ &\quad + \int_0^t \int_0^r \Big(\Psi^{\varepsilon}_{1,ij}(s,r) + \Psi^{\varepsilon}_{2,ij}(s,r)\Big)\, ds\, dr. \end{aligned} \tag{18}$$

By using the form of $\Psi^{\varepsilon}_{1,ij}(\cdot)$ and $\Psi^{\varepsilon}_{2,ij}(\cdot)$, it is easily seen that

$$E|O^{\varepsilon}_{ij}(t)|^2 = 2\int_0^t \int_0^s \Big(\Psi^{\varepsilon}_{1,ij}(s,r) + \Psi^{\varepsilon}_{2,ij}(s,r)\Big)\, dr\, ds.$$

By virtue of the asymptotic expansion (see Proposition 2.1),

$$P\big(\alpha^{\varepsilon}(s) = s_{ij} \big| \alpha^{\varepsilon}(r) = s_{ij}\big) = \nu^i_j(s)\theta_{ii}(s,r) + O\big(\varepsilon + \exp(-\kappa_0(s-r)/\varepsilon)\big) \tag{19}$$

and

$$\begin{aligned} & P\big(\alpha^{\varepsilon}(s) \in \mathscr{M}_i \big| \alpha^{\varepsilon}(r) = s_{ij}\big) \\ &\quad = \sum_{k=1}^{m_i} \nu^i_k(s)\theta_{ii}(s,r) + O\big(\varepsilon + \exp(-\kappa_0(s-r)/\varepsilon)\big) \\ &\quad = \theta_{ii}(s,r) + O\big(\varepsilon + \exp(-\kappa_0(s-r)/\varepsilon)\big). \end{aligned} \tag{20}$$

Using Proposition 2.1 together with (19) and (20),

$$\Psi^{\varepsilon}_{1,ij}(s,r) = O\big(\varepsilon + \exp(-\kappa_0(s-r)/\varepsilon)\big)$$

and similarly

$$\Psi^{\varepsilon}_{2,ij}(s,r) = O\big(\varepsilon + \exp(-\kappa_0(s-r)/\varepsilon)\big).$$

Moreover the estimates hold uniformly in $t \in [0, T]$.

Using the asymptotic expansion for the transient states,

$$\begin{aligned}
E&\left(\int_0^t I_{\{\alpha^\varepsilon(s)=s_{*j}\}}\,ds\right)^2\\
&=2\int_0^t\int_0^s P\left(\alpha^\varepsilon(s)=s_{*j},\alpha^\varepsilon(r)=s_{*j}\right)\,dr\,ds\\
&=\int_0^t\int_0^s O(\varepsilon+\exp(-\kappa_0(s-r)/\varepsilon))O(\varepsilon+\exp(-\kappa_0 r/\varepsilon))\,dr\,ds\\
&=\int_0^t\int_0^s O(\varepsilon^2+\varepsilon\exp(-\kappa_0 r/\varepsilon)+\varepsilon\exp(-\kappa_0(s-r))+\exp(-\kappa_0 s/\varepsilon))\,dr\,ds\\
&=O\left(\varepsilon^2+\int_0^t s\exp(-\kappa_0 s/\varepsilon)\,ds\right)\\
&=O(\varepsilon^2),
\end{aligned}$$

uniformly in $t\in[0,\ T]$, where we used the estimate

$$s\ \exp(-\kappa_0 s/\varepsilon)\le K\varepsilon\ \exp(-\kappa_0 s/(2\varepsilon))$$

with $K=\sup_{u\in[0,\ \infty)} ue^{-\kappa_0 u/2}<\infty$. The desired order estimate then follows. □

4. Weak convergence of aggregate process. This section is devoted to the weak convergence of the process $\overline{\alpha}^\varepsilon(\cdot)$. To proceed, we modify the assumption on the transient part in the generator $\widetilde{Q}(t)$ as follows.

(A3) Assume $\widetilde{Q}_*^i(t)=\widetilde{Q}_*^i$ and $\widetilde{Q}_*(t)=\widetilde{Q}_*$; that is, they are independent of t.

Owing to (A3), $a_{m_i}(t)=a_{m_i}$. Let ξ be a random variable uniformly distributed on $[0,\ 1]$ that is independent of $\alpha^\varepsilon(\cdot)$. For each $j=1,\ldots,m_*$, define an integer-valued random variable ξ_j by

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

Now define the aggregated process $\overline{\alpha}^\varepsilon(\cdot)$ by

$$\overline{\alpha}^\varepsilon(t)=\begin{cases} i, & \text{if } \alpha^\varepsilon(t)\in\mathscr{M}_i,\\ \xi_j, & \text{if } \alpha^\varepsilon(t)=s_{*j}.\end{cases}\tag{21}$$

Note that the state space of $\overline{\alpha}^\varepsilon(t)$ is $\overline{\mathscr{M}}=\{1,\ldots,l\}$ and that $\overline{\alpha}^\varepsilon(\cdot)\in D[0,\ T]$. In addition,

$$P(\overline{\alpha}^\varepsilon(t)=i|\alpha^\varepsilon(t)=s_{*j})=a_{m_i,j}.\tag{22}$$

The aggregated process $\overline{\alpha}^\varepsilon(\cdot)$ is non-Markovian in general; we show that the limit process $\overline{\alpha}(\cdot)$ is a Markov chain, however. To do so, we first prove that if there is a weak convergence, the weak limit must be the solution of an appropriate martingale problem; we then demonstrate that the weak convergence indeed takes place.

LEMMA 4.1. *Under* (A1)–(A3)*, if* $\overline{\alpha}^\varepsilon(\cdot)$ *converges to* $\overline{\alpha}(\cdot)$ *weakly, then the weak limit* $\overline{\alpha}(\cdot)$ *is a solution of the martingale problem with generator* $\overline{Q}_*(\cdot)$*, that is,*

$$f(\overline{\alpha}(t)) - \int_0^t \overline{Q}_*(s) f(\overline{\alpha}(s))\,ds$$

is a martingale for each bounded and Borel-measurable real-valued function $f(\cdot)$.

PROOF. For any bounded and Borel-measurable real-valued function $f(\cdot)$, define

$$\begin{aligned}\overline{f}(\alpha) &= \sum_{k=1}^{l} f(k) I_{\{\alpha\in\mathscr{M}_k\}} + \sum_{k=1}^{l} f(k) a_{m_k,j} I_{\{\alpha=s_{*j}\}}\\ &= \begin{cases} f(i), & \text{if } \alpha\in\mathscr{M}_i, \text{ if } i=1,\dots,l,\\ \displaystyle\sum_{k=1}^{l} f(k) a_{m_k,j}, & \text{if } \alpha = s_{*j},\end{cases}\end{aligned} \tag{23}$$

where a_{m_k} is given by (7).

From the definition it follows immediately that $\widetilde{Q}(t)\overline{f}(\cdot)(\alpha)=0$. Since $Q^\varepsilon(\cdot)$ is a generator of $\alpha^\varepsilon(\cdot)$,

$$\overline{f}(\alpha^\varepsilon(t)) - \int_0^t Q^\varepsilon(s)\overline{f}(\alpha^\varepsilon(s))\,ds$$

is a martingale. We further have

$$\begin{aligned}\int_0^t Q^\varepsilon(s)\overline{f}(\alpha^\varepsilon(s))\,ds &= \int_0^t \widehat{Q}(s)\overline{f}(\alpha^\varepsilon(s))\,ds\\ &= \int_0^t \sum_{i=1}^{l}\sum_{j=1}^{m_i} I_{\{\alpha^\varepsilon(s)=s_{ij}\}} \widehat{Q}(s)\overline{f}(\cdot)(s_{ij})\,ds\\ &\quad + \int_0^t \sum_{j=1}^{m_*} I_{\{\alpha^\varepsilon(s)=s_{*j}\}} \widehat{Q}(s)\overline{f}(\cdot)(s_{*j})\,ds\\ &= \int_0^t \sum_{i=1}^{l}\sum_{j=1}^{m_i} \Big(I_{\{\alpha^\varepsilon(s)=s_{ij}\}} - \nu_j^i(s) I_{\{\alpha^\varepsilon(s)\in\mathscr{M}_i\}}\Big) \widehat{Q}(s)\overline{f}(\cdot)(s_{ij})\,ds\\ &\quad + \int_0^t \sum_{i=1}^{l}\sum_{j=1}^{m_i} \nu_j^i(s) I_{\{\alpha^\varepsilon(s)\in\mathscr{M}_i\}} \widehat{Q}(s)\overline{f}(\cdot)(s_{ij})\,ds\\ &\quad + \int_0^t \sum_{j=1}^{m_*} I_{\{\alpha^\varepsilon(s)=s_{*j}\}} \widehat{Q}(s)\overline{f}(\cdot)(s_{*j})\,ds.\end{aligned} \tag{24}$$

By virtue of Theorem 3.1, the first term on the third equality sign in (24) goes to 0 in mean square. Similarly, the last term in (24) also goes to 0. We then concentrate on the fifth line of (24). Recall the definition of $\overline{Q}_*(s)$ in (6). Note that

$$\sum_{i=1}^{l}\sum_{j=1}^{m_i} \nu^i_j(s) I_{\{\alpha^\varepsilon(s)\in\mathscr{M}_i\}} \widehat{Q}(s)\overline{f}(\cdot)(s_{ij}) = \sum_{i=1}^{l} I_{\{\alpha^\varepsilon(s)\in\mathscr{M}_i\}} \overline{Q}_*(s) f(\cdot)(i).$$

As a result,

$$\sup_{t\in[0,\ T]} E\left|\int_0^t \widehat{Q}(s)\overline{f}(\cdot)(\alpha^\varepsilon(s)) - \int_0^t \overline{Q}_*(s) f(\cdot)(\overline{\alpha}^\varepsilon(s))\, ds\right| \to 0, \tag{25}$$

as $\varepsilon \to 0$.

We next show that

$$\lim_{\varepsilon\to 0} \sup_{s\in[0,\ T]} E\left|\overline{f}(\alpha^\varepsilon(s)) - f(\overline{\alpha}^\varepsilon(s))\right| = 0. \tag{26}$$

Since, for $\alpha^\varepsilon(s) \notin \mathscr{M}_*$, $\overline{f}(\alpha^\varepsilon(s)) = f(\overline{\alpha}^\varepsilon(s))$ and, for $\alpha^\varepsilon(s) \in \mathscr{M}_*$,

$$|\overline{f}(\alpha^\varepsilon(s)) - f(\overline{\alpha}^\varepsilon(s))| \le K$$

for some $K > 0$ by the boundedness of $f(\cdot)$, we have

$$\begin{aligned} E|\overline{f}(\alpha^\varepsilon(s)) - f(\overline{\alpha}^\varepsilon(s))| &= E|\overline{f}(\alpha^\varepsilon(s)) - f(\overline{\alpha}^\varepsilon(s))| I_{\{\alpha^\varepsilon(s)\notin\mathscr{M}_*\}} \\ &\quad + E|\overline{f}(\alpha^\varepsilon(s)) - f(\overline{\alpha}^\varepsilon(s))| I_{\{\alpha^\varepsilon(s)\in\mathscr{M}_*\}} \\ &\le KP(\alpha^\varepsilon(s) \in \mathscr{M}_*) \\ &= O(\varepsilon + \exp(-\kappa_0 s/\varepsilon)) \to 0 \quad \text{as } \varepsilon \to 0. \end{aligned}$$

Recall that

$$\overline{f}(\alpha^\varepsilon(t)) - \int_0^t \widehat{Q}(u)\overline{f}(\alpha^\varepsilon(u))\, du$$

is a martingale. By virtue of the definition of $\overline{\alpha}^\varepsilon(\cdot)$ and the martingale property mentioned above for any bounded and measurable function $h(\cdot)$, for any $0 \le t \le t+s$ and for arbitrary k and $t_i \le t$ for $i = 1, \ldots, k$,

$$\begin{aligned} \lim_{\varepsilon\to 0} Eh(\overline{\alpha}^\varepsilon(t_i),\ i \le k)\Big(&\overline{f}(\alpha^\varepsilon(t+s) - \overline{f}(\alpha^\varepsilon(t)) \\ &- \int_t^{t+s} \widehat{Q}(u)\overline{f}(\alpha^\varepsilon(u))\, du\Big) = 0. \end{aligned} \tag{27}$$

Owing to (25), (26) and the convergence of $\overline{\alpha}^\varepsilon(\cdot)$ to $\overline{\alpha}(\cdot)$, (27) can be rewritten in terms of the limit process. That is,

$$Eh(\overline{\alpha}(t_i),\ i \le k)\left(f(\overline{\alpha}(t+s) - f(\overline{\alpha}(t)) - \int_t^{t+s} \overline{Q}_*(u) f(\overline{\alpha}(u))\, du\right) = 0.$$

This establishes the lemma. □

THEOREM 4.2. *Under conditions* (A1)–(A3), $\overline{\alpha}^\varepsilon(\cdot)$ *converges weakly to* $\overline{\alpha}(\cdot)$, *a Markov chain generated by* $\overline{Q}_*(\cdot)$.

PROOF. Let f be a function defined on $\{1, \dots, l\}$, and let

$$\psi^\varepsilon(t, s, \alpha) = E[f(\overline{\alpha}^\varepsilon(t+s)) - f(\overline{\alpha}^\varepsilon(t))|\alpha^\varepsilon(t) = \alpha],$$

for $t, s \geq 0$ and $\alpha \in \mathscr{M}$. To complete the proof, we use the following lemma.

LEMMA 4.3. *Under the conditions of the theorem,*

$$|\psi^\varepsilon(t, s, \alpha)| \leq K(s + \varepsilon + e^{-\kappa_0 s/\varepsilon}), \tag{28}$$

for some positive constants K *and* κ_0.

Suppose the lemma holds for the moment. Then as $\varepsilon \to 0$,

$$\left|\frac{1}{\sqrt{\varepsilon}} \int_0^{\sqrt{\varepsilon}} \psi^\varepsilon(t, s, \alpha^\varepsilon(t))\, ds\right| \leq \frac{K}{\sqrt{\varepsilon}} \int_0^{\sqrt{\varepsilon}} (s + \varepsilon + e^{-\kappa_0 s/\varepsilon})\, ds = O(\sqrt{\varepsilon}) \to 0$$

and

$$\left|\frac{1}{\sqrt{\varepsilon}} \psi^\varepsilon(t, \sqrt{\varepsilon}, \alpha^\varepsilon(t))\right| \leq \frac{K}{\sqrt{\varepsilon}} \left(\sqrt{\varepsilon} + \varepsilon + e^{-\kappa_0/\sqrt{\varepsilon}}\right) \leq C < \infty.$$

It follows that

$$\lim_{\varepsilon \to 0} E \sup_{0 \leq t \leq T} \left|\frac{1}{\sqrt{\varepsilon}} \int_0^{\sqrt{\varepsilon}} \psi^\varepsilon(t, s, \alpha^\varepsilon(t))\, ds\right| = 0,$$
$$\sup_\varepsilon E \left(\int_0^T \left(\frac{1}{\sqrt{\varepsilon}} \psi^\varepsilon(t, \sqrt{\varepsilon}, \alpha^\varepsilon(t))\right)^2 dt\right)^{1/2} < \infty.$$

By virtue of Corollary 4.8.6 in Ethier and Kurtz [4], $\overline{\alpha}^\varepsilon(\cdot)$ converges to $\overline{\alpha}(\cdot)$ weakly. Then Lemma 4.1 implies that the weak limit $\overline{\alpha}(\cdot)$ is a Markov chain generated by $\overline{Q}_*(\cdot)$.

It remains to prove Lemma 4.3. For $\alpha = s_{ij}$ and $i \neq *$, we have

$$\begin{aligned}
\psi^\varepsilon(t, s, \alpha) &= E\big[f(\overline{\alpha}^\varepsilon(t+s)) - f(i)|\alpha^\varepsilon(t) = s_{ij}\big] \\
&= \sum_{k=1}^{l} E\big[(f(k) - f(i)) I_{\{\alpha^\varepsilon(t+s) \in \mathscr{M}_k\}}|\alpha^\varepsilon(t) = s_{ij}\big] \\
&\quad + \sum_{j_1=1}^{m_*} E\big[(f(\xi_{j_1}) - f(i)) I_{\{\alpha^\varepsilon(t+s) = s_{*j_1}\}}|\alpha^\varepsilon(t) = s_{ij}\big]. \\
&= \sum_{k \neq i} (f(k) - f(i)) P\big(\alpha^\varepsilon(t+s) \in \mathscr{M}_k|\alpha^\varepsilon(t) = s_{ij}\big) + O(\varepsilon + e^{-\kappa_0 s/\varepsilon}),
\end{aligned}$$

because

$$\begin{aligned}
&\big|E\big[(f(\xi_{j_1}) - f(i)) I_{\{\alpha^\varepsilon(t+s) = s_{*j_1}\}}|\alpha^\varepsilon(t) = s_{ij}\big]\big| \\
&\qquad \leq KP\big(\alpha^\varepsilon(t+s) = s_{*j_1}|\alpha^\varepsilon(t) = s_{ij}\big) = O(\varepsilon + e^{-\kappa_0 s/\varepsilon}).
\end{aligned}$$

Moreover,

$$\begin{aligned}
&\sum_{k\neq i}(f(k)-f(i))P\big(\alpha^{\varepsilon}(t+s)\in \mathscr{M}_k|\alpha^{\varepsilon}(t)=s_{ij}\big)\\
&\quad\le K\sum_{k\neq i}P\big(\alpha^{\varepsilon}(t+s)\in \mathscr{M}_k|\alpha^{\varepsilon}(t)=s_{ij}\big)\\
&\quad=\sum_{j_1=1}^{m_k}P\big(\alpha^{\varepsilon}(t+s)=s_{kj_1}|\alpha^{\varepsilon}(t)=s_{ij}\big)\\
&\quad=\theta_{ik}(t+s,t)+O(\varepsilon+e^{-\kappa_0 s/\varepsilon})\quad\big(\text{ with } k\neq i\big)\\
&\quad=O(s)+O(\varepsilon+e^{-\kappa_0 s/\varepsilon})=O(s+\varepsilon+e^{-\kappa_0 s/\varepsilon}),
\end{aligned}$$

because $\Theta(t,t_0)=(\theta_{ij}(t,t_0))$ is the solution to (10) with initial data $\Theta(t_0,\ t_0)=I\in\mathbb{R}^{l\times l}$.

If $\alpha=s_{*j}\in\mathscr{M}_*$, then we have

$$\begin{aligned}
\psi^{\varepsilon}(t,s,\alpha)&=E\big[f(\overline{\alpha}^{\varepsilon}(t+s))-f(\xi_j)|\alpha^{\varepsilon}(t)=s_{*j}\big]\\
&=\sum_{k=1}^{l}f(k)P\big(\alpha^{\varepsilon}(t+s)\in\mathscr{M}_k|\alpha^{\varepsilon}(t)=s_{*j}\big)\\
&\qquad-\sum_{k=1}^{l}f(k)a_{m_k,j}+O(\varepsilon+e^{-\kappa_0 s/\varepsilon})\\
&=\sum_{k=1}^{l}f(k)\theta^{*}_{jk}(t+s,t)-\sum_{k=1}^{l}f(k)a_{m_k,j}+O(\varepsilon+e^{-\kappa_0 s/\varepsilon})\\
&=\sum_{k=1}^{l}f(k)(a_{m_k,j}+O(s))-\sum_{k=1}^{l}f(k)a_{m_k,j}+O(\varepsilon+e^{-\kappa_0 s/\varepsilon})\\
&=O(s+\varepsilon+e^{-\kappa_0 s/\varepsilon}).
\end{aligned}$$

Thus, (28) follows. This completes the proof. □

REMARK. When $\widetilde{Q}^i_*(t)$, for $i=1,\dots,l$, and $\widetilde{Q}_*(t)$ are time dependent, define the aggregated process $\overline{\alpha}^{\varepsilon}(\cdot)$ in the same way. If the underlying process belongs to the space $D[0,\ T]$, all the results obtained so far carry over.

Note that, in the limit, the transient states become isolated states because the corresponding generator $\overline{Q}(t)$ defined in (6) is diagonal with the last block being zero. Therefore the transient states are “asymptotically” not important provided that the initial state is nontransient.

The proof of Theorem 4.2 is based on the results in [4]. A perturbed test function method may also be used to establish the desired results.

5. Asymptotic normality. This section is devoted to the asymptotic properties of a sequence of scaled occupation measures. One natural question to ask is how fast is the convergence taking place for the occupation measures and what is the appropriate scaling.

For $t \geq 0$ and $\alpha \in \mathscr{M}$, let $\beta_{ij}(t)$ be bounded Borel-measurable deterministic functions, and let

$$W_{ij}(t,\alpha) = \begin{cases} \big(I_{\{\alpha=s_{ij}\}} - \nu^i_j(t) I_{\{\alpha\in\mathscr{M}_i\}}\big)\beta_{ij}(t), & \text{for } i=1,\dots,l,\ j=1,\dots,m_i, \\ I_{\{\alpha=s_{*j}\}}\beta_{ij}(t), & \text{for } i=*,\ j=1,\dots,m_*. \end{cases} \tag{29}$$

Consider the normalized occupation measure

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

where

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

Let $\mathscr{F}^\varepsilon_t = \sigma\{\alpha^\varepsilon(s) \colon 0 \leq s \leq t\}$ denote the filtration generated by $\alpha^\varepsilon(\cdot)$.

LEMMA 5.1. *Assume* (A1) *and* (A2). *Then, for all* $0 \leq s \leq t \leq T$, *the following assertions hold:*

(a) $\sup_{s\leq t\leq T} E[(n^\varepsilon(t) - n^\varepsilon(s)) | \mathscr{F}^\varepsilon_s] = O(\sqrt{\varepsilon})$;
(b) $E[(n^\varepsilon(t) - n^\varepsilon(s))^2 | \mathscr{F}^\varepsilon_s] = O(t-s)$.

PROOF. For fixed i, j,

$$E[(n^\varepsilon_{ij}(t) - n^\varepsilon_{ij}(s)) | \mathscr{F}^\varepsilon_s] = \frac{1}{\sqrt{\varepsilon}} \int_s^t E[W_{ij}(r, \alpha^\varepsilon(r)) | \mathscr{F}^\varepsilon_s]\, dr$$

Then, by the Markov property, for $0 \leq s \leq r$, $i = 1, \dots, l$ and $j = 1, \dots, m_i$,

$$E\big[W_{ij}(r, \alpha^\varepsilon(r)) | \mathscr{F}^\varepsilon_s\big] = E\big[W_{ij}(r, \alpha^\varepsilon(r)) | \alpha^\varepsilon(s)\big]. \tag{30}$$

Consider $E[W_{ij}(r, \alpha^\varepsilon(r)) | \alpha^\varepsilon(s) = \alpha]$. If $\alpha = s_{*j} \in \mathscr{M}_*$, then we have

$$\begin{aligned} E\big[&\big(I_{\{\alpha^\varepsilon(r)=s_{ij}\}} - \nu^i_j(r) I_{\{\alpha^\varepsilon(r)\in\mathscr{M}_i\}}\big) | \alpha^\varepsilon(s) = s_{*j_1}\big] \\ &= \nu^i_j(r)\theta^*_{j_1 i}(r,s) - \nu^i_j(r)\theta^*_{j_1 i}(r,s) + O\big(\varepsilon + \exp(-\kappa_0(r-s)/\varepsilon)\big) \\ &= O\big(\varepsilon + \exp(-\kappa_0(r-s)/\varepsilon)\big). \end{aligned}$$

If $\alpha \notin \mathscr{M}_*$, then using the argument of [17], Section 7.4, the third line above is of the order $O(\varepsilon + \exp(-\kappa_0(r-s)/\varepsilon))$. Thus,

$$E\big[W_{ij}(r, \alpha^\varepsilon(r)) | \alpha^\varepsilon(s)\big] = O\big(\varepsilon + \exp(-\kappa_0(r-s)/\varepsilon)\big);$$

so is $E[W_{ij}(r, \alpha^\varepsilon(r)) | \mathscr{F}^\varepsilon_s]$.

Now, for $i = *$ and $j = 1, \dots, m_*$, using the Markovian property, we have

$$\begin{aligned} E\big[W_{*j}(r, \alpha^\varepsilon(r))|\mathscr{F}_s^\varepsilon\big]\beta_{ij}(r) &= E\big[I_{\{\alpha^\varepsilon(r)=s_{*j}\}}|\mathscr{F}_s^\varepsilon\big]\beta_{ij}(r) \\ &= E\big[I_{\{\alpha^\varepsilon(r)=s_{*j}\}}|\alpha^\varepsilon(s)\big]\beta_{ij}(r) \\ &= P\big(\alpha^\varepsilon(r) = s_{*j}|\alpha^\varepsilon(s)\big)\beta_{ij}(r) \\ &= O\big(\varepsilon + \exp(-\kappa_0(r-s)/\varepsilon)\big). \end{aligned}$$

Note that

$$\frac{1}{\sqrt{\varepsilon}}\int_s^t O\big(\varepsilon + \exp\left(\frac{-\kappa_0(r-s)}{\varepsilon}\right) dr = O(\sqrt{\varepsilon}).$$

Thus (a) follows.

To verify (b), fix i, j and define

$$\eta^\varepsilon(t) = E\left[\left(\int_s^t W_{ij}(r, \alpha^\varepsilon(r))\,dr\right)^2\Big|\mathscr{F}_s^\varepsilon\right].$$

Then, by definition of $n_{ij}^\varepsilon(\cdot)$,

$$E\big[(n_{ij}^\varepsilon(t) - n_{ij}^\varepsilon(s))^2|\mathscr{F}_s^\varepsilon\big] = \frac{\eta^\varepsilon(t)}{\varepsilon}.$$

Denote, for $i = 1, \dots, l$ and $j = 1, \dots, m_i$,

$$\begin{aligned} \Psi_{1,ij}^\varepsilon(t,r) &= I_{\{\alpha^\varepsilon(r)=s_{ij}\}}I_{\{\alpha^\varepsilon(t)=s_{ij}\}} - \nu_j^i(t)I_{\{\alpha^\varepsilon(r)=s_{ij}\}}I_{\{\alpha^\varepsilon(t)\in\mathscr{M}_i\}}, \\ \Psi_{2,ij}^\varepsilon(t,r) &= -\nu_j^i(r)I_{\{\alpha^\varepsilon(r)\in\mathscr{M}_i\}}I_{\{\alpha^\varepsilon(t)=s_{ij}\}} + \nu_j^i(r)\nu_j^i(t)I_{\{\alpha^\varepsilon(r)\in\mathscr{M}_i\}}I_{\{\alpha^\varepsilon(t)\in\mathscr{M}_i\}} \end{aligned}$$

and for $i = *$ and $j = 1, \dots, m_*$,

$$\begin{aligned} \Psi_{1,ij}^\varepsilon(t,r) &= I_{\{\alpha^\varepsilon(r)=s_{*j}\}}I_{\{\alpha^\varepsilon(t)=s_{*j}\}}, \\ \Psi_{2,ij}^\varepsilon(t,r) &= 0. \end{aligned}$$

Then

$$\frac{d\eta^\varepsilon(t)}{dt} = 2\int_s^t E\big[\big(\Psi_{1,ij}^\varepsilon(t,r) + \Psi_{2,ij}^\varepsilon(t,r)\big)|\mathscr{F}_s^\varepsilon\big]\beta_{ij}(r)\beta_{ij}(t)\,dr$$

and

$$E[\Psi_{1,ij}^\varepsilon(t,r)|\mathscr{F}_s^\varepsilon] = E[\Psi_{1,ij}^\varepsilon(t,r)|\alpha^\varepsilon(s)].$$

By considering $E[\Psi_{1,ij}^\varepsilon(t,r)|\alpha^\varepsilon(s) = \alpha]$ with $\alpha \in \mathscr{M}_*$ and $\alpha \notin \mathscr{M}_*$, respectively, we can show

$$E[\Psi_{1,ij}^\varepsilon(t,r)|\alpha^\varepsilon(s)] = O(\varepsilon + \exp(-\kappa_0(t-r)/\varepsilon)).$$

Similarly

$$E[\Psi^{\varepsilon}_{2,ij}(t,r)|\mathscr{F}^{\varepsilon}_s] = O\big(\varepsilon + \exp(-\kappa_0(t-r)/\varepsilon)\big).$$

As a consequence,

$$\frac{d\eta^{\varepsilon}(t)}{dt} = O(\varepsilon) \quad \text{and} \quad \eta^{\varepsilon}(s) = 0.$$

Integrating both sides over $[s,t]$ yields

$$\frac{\eta^{\varepsilon}(t)}{\varepsilon} = O(t-s).$$

This completes the proof of the lemma. □

LEMMA 5.2. *Assume* (A1)–(A3). *Then* $\{(n^{\varepsilon}(\cdot), \overline{\alpha}^{\varepsilon}(\cdot))\}$ *is tight in* $D^{m+1}[0,\ T]$.

PROOF. Owing to Theorem 4.2, it suffices to prove the tightness of $\{n^{\varepsilon}(\cdot)\}$. In view of Theorem 3.1, for each $\delta > 0$ and each rational $t \geq 0$, there exists a $K_{t,\delta} \geq \sqrt{KT/\delta}$ such that

$$\inf_{\varepsilon} P(|n^{\varepsilon}(t)| \leq K_{t,\delta}) \geq \inf_{\varepsilon} \big(1 - E|n^{\varepsilon}(t)|^2/K^2_{t,\delta}\big) \geq 1 - Kt/K^2_{t,\delta} \geq 1 - \delta.$$

Using Lemma 5.1, for any $i = 1, \dots, l, *$ and $j = 1, \dots, m_j$,

$$\lim_{\Delta\downarrow 0} \limsup_{\varepsilon\downarrow 0} \sup_{0\leq s\leq \Delta} E\left(E^{\varepsilon}_t\big[n^{\varepsilon}_{ij}(t+s) - n^{\varepsilon}_{ij}(t)\big]^2\right) = 0,$$

where E^{ε}_t denotes the conditional expectation w.r.t. the σ-algebra generated by $\{n^{\varepsilon}(u);\ 0 \leq u \leq t\}$. Thus by virtue of [10], Theorem 2.7 on page 10, the desired result follows. □

In what follows, we often need to separate $W^{\varepsilon}_{ij}(\cdot)$ and a function of $n^{\varepsilon}(\cdot)$. The following lemma that is a generalization of Theorem 3.1 deals with this case. The main idea is to divide the underlying interval into small subintervals so that on each of the small subintervals the separation can be achieved. The proof of the lemma is similar to that of [17], Lemma 7.14, and is thus omitted.

LEMMA 5.3. *Let* $\xi(t,x)$ *be a real-valued function that is Lipschitz in* $(t,x) \in \mathbb{R}^{m+1}$. *Then*

$$\sup_{0\leq\zeta\leq T} E\left|\int_0^{\zeta} W_{ij}(s, \alpha^{\varepsilon}(s))\xi(s, n^{\varepsilon}(s))\,ds\right|^2 \to 0,$$

where $W_{ij}(t,\alpha)$ *was defined in* (29).

To characterize the limit of $(n^{\varepsilon}(\cdot), \overline{\alpha}^{\varepsilon}(\cdot))$, consider the martingale problem associated with $(n^{\varepsilon}(\cdot), \overline{\alpha}^{\varepsilon}(\cdot))$. Note that

$$\frac{dn^{\varepsilon}(t)}{dt} = \frac{1}{\sqrt{\varepsilon}} W(t, \alpha^{\varepsilon}(t)) \quad \text{and} \quad n^{\varepsilon}(0) = 0,$$

where

$$W(t,\alpha)=\big(W_{11}(t,\alpha),\ldots,W_{1m_1}(t,\alpha),\ldots,W_{*1}(t,\alpha),\ldots,W_{*m_*}(t,\alpha)\big).$$

Let $G^\varepsilon(t)$ denote the generator

$$G^\varepsilon(t)f(t,x,\alpha)=\frac{\partial}{\partial t}f(t,x,\alpha)+\frac{1}{\sqrt{\varepsilon}}\langle W(t,\alpha),\nabla_x f(t,x,\alpha)\rangle+Q^\varepsilon(t)f(t,x,\cdot)(\alpha),$$

for all $f(\cdot,\cdot,\alpha)\in C^1$, where $\langle\cdot,\cdot\rangle$ denotes the usual inner product in a Euclidean space and ∇_x is the gradient with respect to x. It is well known (see [1]) that

$$f(t,n^\varepsilon(t),\alpha^\varepsilon(t))-\int_0^t G^\varepsilon(s)f(s,n^\varepsilon(s),\alpha^\varepsilon(s))\,ds \tag{31}$$

is a martingale.

For any real-valued function $f^0(\cdot,i)\in C^2$, with bounded second partial derivatives, define

$$\begin{aligned}\overline{f}(x,\alpha)&=\sum_{k=1}^{l}f^0(x,k)I_{\{\alpha\in\mathscr{M}_k\}}+\sum_{k=1}^{l}a_{m_k,j}f^0(x,k)I_{\{\alpha=s_{*j}\}}\\&=\begin{cases}f^0(x,i), & \text{if } \alpha\in\mathscr{M}_i \text{ for} i=\ldots,l,\\ \displaystyle\sum_{k=1}^{l}f^0(x,k)a_{m_k,j}, & \text{if } \alpha=s_{*j}.\end{cases}\end{aligned}$$

Consider the function

$$f(t,x,\alpha)=\overline{f}(x,\alpha)+\sqrt{\varepsilon}h(t,x,\alpha),$$

where $h(t,x,\alpha)$ is to be specified later.

In view of the definition of $\overline{f}(x,\alpha)$, it is readily seen that $\widetilde{Q}(t)\overline{f}(x,\cdot)(\alpha)=0$. Using the function $f(\cdot)$ in (28) results in

$$\begin{aligned}&\overline{f}(n^\varepsilon(t),\alpha^\varepsilon(t))+\sqrt{\varepsilon}h(t,n^\varepsilon(t),\alpha^\varepsilon(t))\\&\quad-\int_0^t\bigg\{\frac{1}{\sqrt{\varepsilon}}\langle W(s,\alpha^\varepsilon(s)),\nabla_x\overline{f}(n^\varepsilon(s),\alpha^\varepsilon(s))+\sqrt{\varepsilon}\,\nabla_x h(s,n^\varepsilon(s),\alpha^\varepsilon(s))\rangle\\&\qquad+\sqrt{\varepsilon}\frac{\partial}{\partial s}h(s,n^\varepsilon(s),\alpha^\varepsilon(s))\\&\qquad+\frac{1}{\sqrt{\varepsilon}}\widetilde{Q}(s)h(s,n^\varepsilon(s),(\cdot))(\alpha^\varepsilon(s))\\&\qquad+\widehat{Q}(s)(\overline{f}(n^\varepsilon(s),(\cdot))+\sqrt{\varepsilon}h(s,n^\varepsilon(s),(\cdot))(\alpha^\varepsilon(s))\bigg\}\,ds\end{aligned}$$

defining a martingale.

The methods of perturbation of generators (see Ethier and Kurtz [4], Section 1.7) and the perturbed test functions (see, e.g., Kushner [11], Chapter 5 and the references therein) provide useful machinery to carry out the analysis. To apply their methods, we need to choose the function $h(\cdot)$ so that it cancels the "bad" terms of the order $1/\sqrt{\varepsilon}$:

$$\widetilde{Q}(s)h(s, x, \cdot)(\alpha) = -\langle W(s, \alpha), \nabla_x \overline{f}(x, \alpha)\rangle.$$

The matrix $\widetilde{Q}(t)$ has rank $m - l$. Therefore, as discussed in [17], page 194, $h(\cdot)$ can be chosen to satisfy the following properties, assuming $\beta_{ij}(\cdot)$ to be Lipschitz on $[0, \ T]$:

(1) $h(t, x, \alpha)$ is continuously differentiable in x;
(2) $(\partial/\partial t)h(t, x, \alpha)$ exists a.e. in (t, x, α);
(3) $\left|h(t, x, \alpha)\right|$, $\left|\nabla_x h(t, x, \alpha)\right|$ and $|(\partial/\partial t)h(t, x, \alpha)|$ are bounded by $K(1 + |x|)$.

Such an $h(\cdot)$ leads to

$$\begin{aligned}
&\overline{f}(n^\varepsilon(t), \alpha^\varepsilon(t)) + \sqrt{\varepsilon} h(t, n^\varepsilon(t), \alpha^\varepsilon(t)) \\
&\quad - \int_0^t \Bigg\{ \langle W(s, \alpha^\varepsilon(s)), \nabla_x h(s, n^\varepsilon(s), \alpha^\varepsilon(s))\rangle \\
&\qquad\qquad + \sqrt{\varepsilon} \frac{\partial}{\partial s} h(s, n^\varepsilon(s), \alpha^\varepsilon(s)) + \widehat{Q}(s)\overline{f}(n^\varepsilon(s), \cdot)(\alpha^\varepsilon(s)) \\
&\qquad\qquad\qquad + \sqrt{\varepsilon}\widehat{Q}(s) h(s, n^\varepsilon(s), \cdot)(\alpha^\varepsilon(s)) \Bigg\}\, ds
\end{aligned}$$

being a martingale.

For each s, x, α, define

$$g(s, x, \alpha) = \langle W(s, \alpha), \nabla_x h(s, x, \alpha)\rangle.$$

Using $g(s, x, \alpha)$ defined above, we obtain

$$\begin{aligned}
&\int_0^t \langle W(s, \alpha^\varepsilon(s)), \nabla_x h(s, n^\varepsilon(s), \alpha^\varepsilon(s))\rangle\, ds \\
&\quad = \int_0^t g(s, n^\varepsilon(s), \alpha^\varepsilon(s))\, ds \\
&\quad = \int_0^t \sum_{i=1}^{l} \sum_{j=1}^{m_i} I_{\{\alpha^\varepsilon(s) = s_{ij}\}} g(s, n^\varepsilon(s), s_{ij})\, ds \\
&\qquad + \int_0^t \sum_{j=1}^{m_*} I_{\{\alpha^\varepsilon(s) = s_{*j}\}} g(s, n^\varepsilon(s), s_{*j})\, ds \\
&\quad = \int_0^t \sum_{i=1}^{l} \sum_{j=1}^{m_i} \left(I_{\{\alpha^\varepsilon(s) = s_{ij}\}} - \nu_j^i(s) I_{\{\overline{\alpha}^\varepsilon(s) \in \mathscr{M}_i\}} \right) g(s, n^\varepsilon(s), s_{ij})\, ds \\
&\qquad + \int_0^t \sum_{i=1}^{l} \sum_{j=1}^{m_i} I_{\{\alpha^\varepsilon(s) \in \mathscr{M}_i\}} \nu_j^i(s) g(s, n^\varepsilon(s), s_{ij})\, ds \\
&\qquad + \int_0^t \sum_{j=1}^{m_*} I_{\{\alpha^\varepsilon = s_{*j}\}} g(s, n^\varepsilon(s), s_{*j})\, ds.
\end{aligned}$$

In view of Lemma 5.3, the terms in the fifth and seventh lines above go to zero in mean square uniformly in $t \in [0,\ T]$. Let

$$\overline{g}(s, x, i) = \sum_{j=1}^{m_i} \nu_j^i(s) g(s, x, s_{ij}). \tag{32}$$

It is an average with respect to the quasistationary distribution corresponding to the states in $\mathscr{M}_i$.

It follows that

$$\begin{aligned}\int_0^t \sum_{i=1}^{l} \sum_{j=1}^{m_i} & I_{\{\overline{\alpha}^\varepsilon(s)\in\mathscr{M}_i\}} \nu_j^i(s) g(s, n^\varepsilon(s), s_{ij})\, ds \\ &= \int_0^t \sum_{i=1}^{l} I_{\{\alpha^\varepsilon(s)\in\mathscr{M}_i\}} \overline{g}(s, n^\varepsilon(s), i)\, ds \\ &= \int_0^t \overline{g}(s, n^\varepsilon(s), \overline{\alpha}^\varepsilon(s))\, ds.\end{aligned}$$

Therefore, as $\varepsilon \to 0$,

$$E\left|\int_0^t \langle W(s, \alpha^\varepsilon(s)), \nabla_x h(s, n^\varepsilon(s), \alpha^\varepsilon(s))\rangle\, ds - \int_0^t \overline{g}(s, n^\varepsilon(s), \overline{\alpha}^\varepsilon(s))\, ds\right|^2 \to 0$$

uniformly in $t \in [0,\ T]$.

Furthermore, similarly to the proof of Lemma 4.1, it follows that as $\varepsilon \to 0$,

$$E\left|\int_0^t \widehat{Q}(s)\overline{f}(n^\varepsilon(s), \cdot)(\alpha^\varepsilon(s))\, ds - \int_0^t \overline{Q}_*(s) f^0(n^\varepsilon(s), \cdot)(\overline{\alpha}^\varepsilon(s))\, ds\right| \to 0$$

uniformly in $t \in [0,\ T]$, where $\overline{Q}_*(s)$ is given by (6).

LEMMA 5.4. *Assume* (A1) *and* (A2). *If* $(n^\varepsilon(\cdot), \bar{\alpha}^\varepsilon(\cdot))$ *converges weakly to* $(n(\cdot), \overline{\alpha}(\cdot))$, *then for* $f^0(\cdot, i) \in C^2$ *with bounded second partial derivatives*

$$f^0(n(t), \overline{\alpha}(t)) - \int_0^t (\overline{g}(s, n(s), \overline{\alpha}(s) + \overline{Q}_*(s) f^0(n(s), \cdot)(\overline{\alpha}(s)))\, ds$$

is a martingale.

PROOF. First, note that $\overline{f}(n^\varepsilon(t), \alpha^\varepsilon(t)) = f^0(n^\varepsilon(t), \overline{\alpha}^\varepsilon(t))$ for $\alpha^\varepsilon(t) \notin \mathscr{M}_*$, and for $\alpha^\varepsilon(t) \in \mathscr{M}^*$,

$$\left|\overline{f}(n^\varepsilon(t), \alpha^\varepsilon(t)) - f^0(n^\varepsilon(t), \overline{\alpha}^\varepsilon(t))\right| \le K$$

for some $K > 0$ by the boundedness of $f^0(\cdot)$. Therefore, by similar reasoning as in the proof of Lemma 4.1, we can show that

$$\lim_{\varepsilon\to 0} \sup_{t\in[0,\ T]} E\left|\overline{f}(n^\varepsilon(t), \alpha^\varepsilon(t)) - f^0(n^\varepsilon(t), \overline{\alpha}^\varepsilon(t))\right| = 0,$$

and then the rest of the proof is the same as in the proof of [17], Lemma 7.17. □

Now, let us return to the function $\overline{g}(\cdot)$. For $i = 1, \dots, l$, as in [17] (7.52), $\overline{g}(s, x, i)$ can be written as

$$\overline{g}(s, x, i) = \frac{1}{2} \sum_{i_1, i_2 = 1}^{m_i} a_{i_1 i_2}(s, i)\, \partial^2_{i, i_1, i_2} f_0(x, i),$$

for some continuous function $a_{i_1 i_2}$. It follows that

$$\overline{g}(s, x, s_{*j}) = 0. \tag{33}$$

Here $a_{i_1 i_2}(s, i)$ can be computed as in [17], page 203.

LEMMA 5.5. *Let $\mathscr{L}_s$ denote the generator given by*

$$\mathscr{L}_s f^0(x, i) = \overline{g}(s, x, i) + \overline{Q}_*(s) f(x, \cdot)(i).$$

The martingale problem with generator $\mathscr{L}_s$ has a unique solution.

PROOF. The structure of $\overline{Q}(\cdot)$ implies that $\overline{Q}(s) f(x, \cdot)(s_{*j}) = 0$ for any bounded and measurable function $f(\cdot)$. This and (33) then imply the transient states have no contribution to the generator of the limit martingale problem. As for $i = 1, \dots, l$, the uniqueness can be proved as in [17], Lemma 7.18. □

Similar to [17] Lemma 7.19, the first assertion in what follows can be established, whereas the second assertion is a consequence of (32) and (33).

LEMMA 5.6. *The following assertions hold:*

(a) *for $s \in [0, T]$ and $i = 1, \dots, l$, $A(s, i) = (a_{i_1 i_2}(s, i))$ is symmetric and non-negative definite*;

(b) $A(s, s_{*j}) = 0$ *for* $j = 1, \dots, m_*$.

Hence, we have shown that $(n^\varepsilon(\cdot), \overline{\alpha}^\varepsilon(\cdot))$ converges weakly to $(n(\cdot), \overline{\alpha}(\cdot))$, where $n(\cdot)$ is a switching diffusion, whose covariance matrix has the following form:

$$\sigma(s, i) = \operatorname{diag}(0_{m_1 \times m_1}, \dots, \sigma^0(s, i), \dots, 0_{m_l \times m_l}, 0_{m_* \times m_*}), \tag{34}$$

where $\sigma^0(s, i)$ is an $m_i \times m_i$ matrix such that

$$\sigma^0(s, i) \sigma^{0,\prime}(s, i) = \tfrac{1}{2} A(s, i) \quad \text{for } i = 1, \dots, l.$$

We summarize this in the following theorem.

THEOREM 5.7. *Assume* (A1) *and* (A3)*, and suppose $\widetilde{Q}(\cdot)$ is twice differentiable with Lipschitz continuous second derivative. Moreover, $\widehat{Q}(\cdot)$ is differentiable with Lipschitz continuous derivative. Let $\beta_{ij}(\cdot)$ (for $i = 1, \dots, l$,*

$j = 1, \ldots, m_i$) be a bounded and Lipschitz continuous deterministic function. Then $n^\varepsilon(\cdot)$ converges weakly to a switching diffusion $n(\cdot)$, where

$$n(t) = \int_0^t \sigma(s, \overline{\alpha}(s))\, dw(s) \tag{35}$$

and $w(\cdot)$ is a standard m-dimensional Brownian motion.

Our result includes an arbitrary function $\beta(\cdot)$ for the reason of more generality. Such a function often arises in various applications in manufacturing systems [15], Chapter 5. A simplest case is $\beta_{ij}(t) = 1$ for all i and all j.

One of the conditions used in this paper is the smoothness of the generator. When this condition is missing, although it is possible to obtain convergence of $P(\alpha^\varepsilon(t) = s_{ij})$, one cannot expect to be able to obtain the asymptotic expansion. The asymptotic expansion, however, is crucial in calculating the covariance of the limit switching diffusion.

6. Concluding remark. This paper concentrates on the aggregation aspect of a singularly perturbed Markov chain with inclusion of transient states. The results obtained will be of help in treating control and optimization problems involving such a Markov chain so that the complexity of the models can be reduced. Further study can be carried out in deriving exponential-type upper bounds for the scaled sequence of occupation measures.

Condition (A3) can be much relaxed and time-varying functions can be dealt with. Assume (A3) with the following modifications: For each k, $\widetilde{Q}^k(t) = B(t)\widetilde{Q}^k_{*,c}$, $\widetilde{Q}_*(t) = B(t)\widetilde{Q}_{*,c}$, where $B(t) \in \mathbb{R}^{m_* \times m_*}$ and $\widetilde{Q}^k_{*,c} \in \mathbb{R}^{m_* \times m_k}$ and $\widetilde{Q}_{*,c} \in \mathbb{R}^{m_* \times m_*}$ are time-independent matrices. It is readily seen that $B(t)$ is invertible for each $t \in [0, T]$ and for each k, $a_{m_k}(t)$ remains to be a time-independent vector a_{m_k}. Using exactly the same arguments, Lemmas 4.1, 4.3, 5.2 and Theorems 4.2 and 5.7 continue to hold.

Acknowledgment. We gratefully acknowledge the insightful comments and detailed suggestions provided by the reviewer, which led to much improvement of the paper.

REFERENCES

[1] DAVIS, M. H. A. (1993). *Markov Models and Optimization*. Chapman and Hall, London.

[2] DELEBECQUE, F. and QUADRAT, J. (1981). Optimal control for Markov chains admitting strong and weak interactions. *Automatica* **17** 281–296.

[3] ELLIOTT, R. J., AGGOUN, L. and MOORE, J. (1995). *Hidden Markov Models: Estimation and Control*. Springer, New York.

[4] ETHIER, S. N. and KURTZ T. G. (1986). *Markov Processes: Characterization and Convergence*. Wiley, New York.

[5] FLEMING, W. H. and RISHEL, R. W. (1975). *Deterministic and Stochastic Optimal Control*. Springer, New York.

[6] FRIEDLIN, M. I. and WENTZELL, A. D. (1984). *Random Perturbations of Dynamical Systems*. Springer, New York.

[7] KHASMINSKII, R. Z. (1968). On an averaging principle for Itô stochastic differential equations. *Kybernetika* **4** 260–279.

[8] KHASMINSKII, R. Z., YIN, G. and ZHANG, Q. (1996). Asymptotic expansions of singularly perturbed systems involving rapidly fluctuating Markov chains. *SIAM J. Appl. Math.* **56** 277–293.

[9] KHASMINSKII, R. Z., YIN G. and ZHANG, Q. (1997). Constructing asymptotic series for probability distribution of Markov chains with weak and strong interactions. *Quart. Appl. Math.* **55** 177–200.

[10] KURTZ, T. G. (1981). *Approximation of Population Processes*. SIAM, Philadelphia.

[11] KUSHNER, H. J. (1984). *Approximation and Weak Convergence Methods for Random Processes, with Applications to Stochastic Systems Theory*. MIT Press.

[12] KUSHNER, H. J. (1990). *Weak Convergence Methods and Singularly Perturbed Stochastic Control and Filtering Problems*. Birkhäuser, Boston.

[13] PAN, Z. G. and BAŞAR, T. (1995). H^∞-control of Markovian jump linear systems and solutions to associated piecewise-deterministic differential games. In *New Trends in Dynamic Games and Applications* (G. J. Olsder, ed.) 61–94. Birkhäuser, Boston.

[14] PHILLIPS, R. G. and KOKOTOVIC, P. V. (1981). A singular perturbation approach to modelling and control of Markov chains. *IEEE Trans. Automat. Control* **26** 1087–1094.

[15] SETHI, S. P. and ZHANG, Q. (1994). *Hierarchical Decision Making in Stochastic Manufacturing Systems*. Birkhäuser, Boston.

[16] YIN, G. and ZHANG, Q. (1997). *Mathematics of Stochastic Manufacturing Systems*. Amer. Math. Soc., Providence, RI.

[17] YIN, G. and ZHANG, Q. (1998). *Continuous-Time Markov Chains and Applications: A Singular Perturbations Approach*. Springer, New York.

[18] ZHANG, Q. and YIN, G. (1996). A central limit theorem for singularly perturbed nonstationary finite state Markov chains. *Ann. Appl. Probab.* **6** 650–670.

[19] ZHANG, Q. and YIN, G. (1997). Structural properties of Markov chains with weak and strong interactions. *Stochastic Process. Appl.* **70** 181–197.

[20] ZHANG, Q. and YIN, G. (1999). On nearly optimal controls of hybrid LQG problems. *IEEE Trans. Automat. Control* **44** 2271–2282.

G. YIN
G. BADOWSKI
DEPARTMENT OF MATHEMATICS
WAYNE STATE UNIVERSITY
DETROIT, MICHIGAN 48202
E-MAIL: gyin@math.wayne.edu
grazyna@math.wayne.edu

Q. ZHANG
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF GEORGIA
ATHENS, GEORGIA 30602
E-MAIL: qingz@math.uga.edu