The Reduction of Perturbed Markov Generators: An Algorithm Exposing the Role of Transient States

JAN ROBIN ROHLICEK

Bolt, Beranek and Newman Laboratories, Inc., Cambridge, Massachusetts

AND

ALAN S. WILLSKY

Massachusetts Institute of Technology, Cambridge, Massachusetts

Abstract. A new algorithm for the hierarchical aggregation of singularly perturbed finite-state Markov processes is derived. The approach taken bridges the gap between conceptually simple results for a relatively restricted class of processes and the significantly more complex results for the general case. The critical role played by (almost) transient states is exposed, resulting in a straightforward algorithm for the construction of a sequence of aggregate generators associated with various time scales. These generators together provide a uniform asymptotic approximation of the original probability transition function.

Categories and Subject Descriptors: D.4.8 [**Operating Systems**]: Performance—*modeling and prediction*; *queuing theory*; *simulation*; *stochastic analysis*; D.4.5 [**Operating Systems**]: Reliability—*fault tolerance*; G.1.3 [**Numerical Analysis**]: Numerical Linear Algebra—*eigenvalues*; *conditioning*; *error analysis*; *sparse and very large systems*

General Terms: Algorithms, Performance, Reliability, Theory, Verification

Additional Key Words and Phrases: Aggregation, Markov processes, time scales

1. *Introduction*

Many systems exhibit behavior at multiple temporal or spatial "scales." Often, the existence of these different scales causes difficulty in the analysis of a system owing to either numerical ill-conditioning or excessive complexity resulting from explicit consideration of the detailed interactions within the system. A possible approach to such problems is to try to isolate the various scales of behavior and analyze them separately. This basic approach has been applied with success to the analysis of finite-state Markov processes with weakly coupled components and rare transitions. As has been shown by several authors (Coderch [5], Delebecque [9], Courtois [7], and others), processes with such structure exhibit behavior at several time

This research was conducted under the support of the Air Force Office of Scientific Research under grant AFOSR 82-0258 and the Army Research Office under grant DAAG-29-84-K-0005.

Authors' addresses: J. R. Rohlicek, Bolt, Beranek and Newman Laboratories, Inc., 10 Moulton Street, Cambridge, MA 02238; A. S. Willsky, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139.

scales. Moreover, explicit identification of the behavior at various time scales has been addressed through the construction of reduced-order aggregate processes.

The results presented in this paper address the decomposition of a general class of perturbed Markov processes and provide a computationally feasible algorithm for their analysis and uniform approximation. Some of the previous algorithms [7, 10] are applicable to only comparatively restricted classes of Markov processes. By considering such restricted classes, however, the algorithms for the construction of aggregated processes associated with various time scales are generally straightforward and involve computations with clear probabilistic interpretations. At the other extreme, Coderch [5] and Delebecque [9] deal with a completely general class of perturbed Markov processes, and the former also proves the uniform convergence of a decomposition-based approximation. The price, however, that is paid for this generality and the guaranteed uniform convergence are algorithms of significantly greater complexity involving the computation of complex quantities that are not easily interpreted in probabilistic terms.

The algorithm presented in this paper, which was originally outlined by Lou et al. [15] focuses on the gap between these two extreme sets of results. In particular, we present an algorithm for the construction of uniform multiple-time-scale approximations of singularly perturbed Markov processes that is as general as that of Coderch [5] and Delebecque [9] but has much the same straightforward, easily interpreted flavor as that of Courtois [7]. Indeed, when the class of systems is suitably restricted, the construction is essentially identical to that of Courtois [7].

The focus of this paper is on generators of continuous-time, finite-state Markov processes that are analytic functions of a small parameter ϵ, representing the presence of rare transitions between sets of states. Consider such a Markov generator, $\mathbf{A}^{(0)}(\epsilon)$[1] of size $n \times n$. The matrix probability transition function $\mathbf{X}(t)$ satisfies the dynamical equation

$$\begin{aligned} \dot{\mathbf{X}}(t) &= \mathbf{A}^{(0)}(\epsilon)\mathbf{X}(t), \\ \mathbf{X}(0) &= I, \end{aligned} \tag{1.1}$$

whose solution can be written as

$$\mathbf{X}(t) = \exp[\mathbf{A}^{(0)}(\epsilon)t]. \tag{1.2}$$

The goal is to obtain an approximation of this solution that (a) explicitly displays the evolution of the process for various orders of t $(1, 1/\epsilon, 1/\epsilon^2, \ldots)$ using appropriately aggregated, ϵ-independent, Markov generators and that (b) converges uniformly over the interval $t \in [0, \infty)$ to the true probability transition function as $\epsilon \downarrow 0$. A solution (a) and (b) is presented in Coderch [5, 6], based on associating multiple time scales with different orders of eigenvalues of $\mathbf{A}^{(0)}(\epsilon)$. Building on Kato's [12] perturbation results for linear operators, Coderch et al. identify the subspaces associated with these various orders of eigenvalues and devise a sequential procedure for construction of the approximation. In particular, it is shown that the solution (1.2) can be uniformly approximated using the unperturbed "fast" evolution[2]

$$\exp[\mathbf{A}^{(0)}t], \tag{1.3}$$

[1] The superscript $^{(0)}$ is used here to maintain a uniform notation throughout the paper. It signifies the first generator in a sequence that will be constructed in the next section.

[2] Here $\mathbf{A}^{(0)} = \mathbf{A}^{(0)}(0)$ for simplicity. To avoid confusion, we consistently write $\mathbf{A}^{(0)}(\epsilon)$ when we are referring to the full generator, as in (1.2).

and a "slow" evolution

$$\exp[\bar{\mathbf{A}}^{(1)}](\epsilon)\epsilon t, \tag{1.4a}$$

where

$$\bar{\mathbf{A}}^{(1)}(\epsilon) = \frac{1}{\epsilon}\,\mathbf{P}^{(0)}(\epsilon)\mathbf{A}^{(0)}(\epsilon)\mathbf{P}^{(0)}(\epsilon). \tag{1.4b}$$

Here $\mathbf{P}^{(0)}(\epsilon)$ is the eigenprojection associated with all the eigenvalues of order ϵ or higher. The procedure can then be iterated to produce the desired approximation, consisting of $\exp[\mathbf{A}^{(0)}t]$, $\exp[\mathbf{A}^{(1)}\epsilon t]$, $\exp[\mathbf{A}^{(2)}\epsilon^2 t]$, etc. There are, however, several drawbacks to this procedure. The first is the need to compute the entire ϵ-dependent eigenprojections, $\mathbf{P}^{(0)}(\epsilon)$, $\mathbf{P}^{(1)}(\epsilon)$, ..., and a second is the absence of a simple probabilistic interpretation of the computations being performed. Finally, although at the end of the procedure Coderch [5] provides a way in which to reorganize the approximation so that it consists of increasingly aggregated (and hence simpler) Markov models at successively slower time scales, all computations are performed on the full, unaggregated process.

The approach taken by Courtois [7] overcomes all of these drawbacks. Specifically, in essence what Courtois does is to replace the slow evolution in (1.4a) and (1.4b) by

$$\exp[\bar{\mathbf{F}}^{(1)}(\epsilon)\epsilon t] \tag{1.5a}$$

where

$$\bar{\mathbf{F}}^{(1)}(\epsilon) = \frac{1}{\epsilon}\,\mathbf{P}^{(0)}\mathbf{A}^{(0)}(\epsilon)\mathbf{P}^{(0)}. \tag{1.5b}$$

Here $\mathbf{P}^{(0)} = \mathbf{P}^{(0)}(0)$ has a simple probabilistic interpretation as the ergodic projection of the unperturbed process

$$\mathbf{P}^{(0)} = \lim_{t\to\infty} \exp[\mathbf{A}^{(0)}t]. \tag{1.6}$$

This involves no ϵ-dependent computations. Furthermore, we can always write

$$\mathbf{P}^{(0)} = \mathbf{U}^{(0)}\mathbf{V}^{(0)}. \tag{1.7}$$

Here $\mathbf{V}^{(0)}$ is a "membership matrix." In the case in which there are no transient states generated by $\mathbf{A}^{(0)}$, it consists entirely of 0's and 1's whose rows identify which states of the process form individual ergodic classes of $\mathbf{A}^{(0)}$. Also the columns $\mathbf{U}^{(0)}$ denote the ergodic probability vectors, one for each ergodic class of $\mathbf{A}^{(0)}$, and finally

$$\mathbf{V}^{(0)}\mathbf{U}^{(0)} = \mathbf{I}. \tag{1.8}$$

From (1.7) and (1.8), we see that (1.5a) can be computed in an even simpler fashion:

$$\exp[\bar{\mathbf{F}}^{(1)}(\epsilon)\epsilon t] = \mathbf{U}^{(0)}\exp[\mathbf{A}^{(1)}(\epsilon)\epsilon t]\,\mathbf{V}^{(0)}, \tag{1.9}$$

where

$$\mathbf{A}^{(1)}(\epsilon) = \frac{1}{\epsilon}\,\mathbf{V}^{(0)}\mathbf{A}^{(0)}(\epsilon)\,\mathbf{U}^{(0)} \tag{1.10}$$

is an aggregated Markov generator with one state for each ergodic class of $\mathbf{A}^{(0)}$. Indeed, (1.10) has an appealing probabilistic interpretation: We compute the

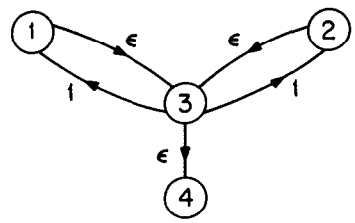


FIG. 1. Perturbed Markov process.

transition rate between aggregated ergodic classes of $\mathbf{A}^{(0)}$ as an "average rate," in which the rates of individual states in these classes are averaged using the ergodic probabilities of $\mathbf{A}^{(0)}$.

Although the procedure just described has a number of appealing features, it cannot be applied to arbitrary processes. In particular, Courtois [7] focuses his development on the class of "nearly completely decomposable" processes introduced by Simon and Ando [20] and Ando and Fisher [1], in which $\mathbf{A}^{(0)}$ has no transient states. Although this condition can be relaxed somewhat (see Section 3), it is restrictive. Furthermore, although the ideas of Simon and Ando [20] and Courtois [7] do allow one to consider several levels of aggregation at different time scales, iterative application of this method, in general, cannot be performed, since the constraint of nearly complete decomposability may fail at one or more intermediate time scales.

As the previous paragraph implies, the need for a more general algorithm can be traced to the role played by states that are transient at various time scales. To illustrate this, consider the process depicted in Figure 1. At $\epsilon = 0$, states 1, 2, and 4 are individual ergodic classes, while state 3 is transient, so that its steady-state probability is 0. Consequently, application of the averaging implied by (1.10) (which uses the steady-state probabilities at $\epsilon = 0$) completely misses the possibility of transition from states 1, 2, or 3, to state 4. Thus, in this case, the approximation implied by (1.9) and (1.10) does *not* capture the fact that 4 is really a trapping state for $\epsilon > 0$. The problem in this example is that the critical path determining long-term behavior involves a *sequence* of (in this case, two) rare events (namely, a transition from states 1 or 2 to 3, followed immediately by a transition to state 4). Processes with such behavior arise in a variety of applications, and are of particular interest for analyzing the long-term reliability or availability of complex systems such as interconnected power networks (in which sequences of events lead, on infrequent occasions, to blackouts), data communication networks, and fault-prone systems possessing backup capability. The process depicted in Figure 1 can, in fact, be thought of as an (extremely simplified) example of a system consisting of two machines, one of which acts as a backup. States 1 and 2 correspond to both machines being in working order. If a failure of one machine occurs, there is a transition of the process to state 3 at which the machine is examined and then repaired (causing a transition to state 1) or replaced (causing a transition to state 2). However, on rare occasions, the second machine fails before the first is repaired or replaced, causing a stoppage in operation (and a transition to state 4).

Though the importance of transient states has been recognized in previous work, no general approach has been developed. Korolyuk and Turbin [13a] have considered a case in which there is a particular ergodic structure. Recently, Bobbio and Trivedi [3] have proposed a method, similar to our own, for analyzing the effect of transient states in the two-time-scale case. Multiple-time-scale analysis of perturbed Markov processes with arbitrary ergodic structure is not available in these

works, however, particularly with respect to the construction of a uniform asymptotic approximation.

In this paper we perform this full multiple-time-scale analysis and prove uniform convergence. The key to our development is a method for handling transient states at various time scales (state 3 in the example) that couple ergodic classes at slower time scales (as state 3 does between states 1 and 4 and between states 2 and 4). In general, such transient states may not be transient in the full process and thus can be thought of as *almost transient* states. The way in which we accommodate the presence of such states is essentially a modification of (1.10). Specifically, recall that $\mathbf{V}^{(0)}$ is a membership matrix indicating which states are in which ergodic classes. When there are transient states, it is necessary to consider an ϵ-dependent membership matrix $\tilde{\mathbf{V}}^{(0)}(\epsilon)$ to capture the fact that states that couple ergodic classes can be thought of as being "partly" in each. Therefore, in such a case, we must identify and retain certain ϵ-dependent terms, but we can stop far short of the complete computations required by Coderch [5] and maintain the advantage of Courtois's approach of working directly on increasingly aggregated versions of the process.

In the next section we present our algorithm and illustrate it using the example introduced in this section, and in Section 3 we outline the derivation of the procedure and prove uniform convergence. Section 4 contains a discussion of several issues, including computational and numerical aspects of hierarchical aggregation.

2. *The Algorithm*

In this section we present and apply our algorithm for the construction of uniform multiple-time-scale approximations of singularly perturbed finite-state Markov processes. For simplicity, we assume that we begin with a Markov generator $\mathbf{A}^{(0)}(\epsilon)$ that has one ergodic class for $\epsilon > 0$.[3] The basic algorithm involves the computation of a sequence of generators, the kth of which, $\mathbf{A}^{(k)}(\epsilon)$, captures all behavior at time scales of order $1/\epsilon^k$ or slower. The procedure is iterative, with $\mathbf{A}^{(k+1)}(\epsilon)$ determined directly from $\mathbf{A}^{(k)}(\epsilon)$. There are essentially four steps involved at each step of this algorithm as shown below.

Algorithm

(0) Set $k \leftarrow 0$.
Begin with the generator $\mathbf{A}^{(0)}(\epsilon)$ of a finite-state Markov process.

(1) Partition the state set into the communicating classes $E_1, E_2, \ldots, E_N$ and the transient set T generated by $\mathbf{A}^{(k)}(0)$. If there is only a single class ($N = 1$), go to (5).

(2) For each class E_I, compute the ergodic probabilities of the member states at $\epsilon = 0$, $u_{i,I}^{(k)}$, $\forall i \in E_I$.

(3) For each transient state $j \in T$ and each class E_I, compute a term $\tilde{v}_{I,j}^{(k)}(\epsilon)$ such that[4]

$$\tilde{v}_{I,j}^{(k)}(\epsilon) = v_{I,j}^{(k)}(\epsilon)(1 + O(\epsilon)), \tag{2.1a}$$

$$\sum_{K=1}^{N} \tilde{v}_{K,j}^{(k)}(\epsilon) = 1, \tag{2.1b}$$

where

$$v_{I,j}^{(k)}(\epsilon) \equiv \Pr(\eta^{(k)}(\epsilon, t^*) \in E_I \mid \eta^{(k)}(\epsilon, 0) = j,\ t^* = \inf_{t \geq 0}(t \mid \eta^{(k)}(\epsilon, t) \notin T) \tag{2.2}$$

and $\eta^{(k)}(\epsilon, t)$ is a sample path of the Markov process generated by $A^{(k)}(\epsilon)$.

[3] The generalization to more than one class is trivial, since we can reorder the states of the process so that $\mathbf{A}^{(0)}(\epsilon)$ is block diagonal and then consider each block individually.

[4] Here $O(\epsilon^k)$ denotes a quantity of order ϵ^k.

(4) Form the matrices

$$\mathbf{U}^{(k)} = [u_{i,J}^{(k)}], \qquad \text{where} \quad u_{(i,j)}^{(k)} = 0 \quad \text{if} \quad i \notin E_J, \tag{2.3}$$

$$\tilde{\mathbf{V}}^{(k)}(\epsilon) = [\tilde{v}_{I,j}^{(k)}(\epsilon)]. \tag{2.4}$$

Then

$$\mathbf{A}^{(k+1)}(\epsilon) = \frac{1}{\epsilon}\,\tilde{\mathbf{V}}^{(k)}(\epsilon)\mathbf{A}^{(k)}(\epsilon)\mathbf{U}^{(k)}. \tag{2.5}$$

Set $k \leftarrow k+1$ and go to (1).

(5) The overall approximation of the evolution of the transition probabilities can be written as

$$\begin{aligned}\exp[\mathbf{A}^{(0)}(\epsilon)t] = {} & \exp[\mathbf{A}^{(0)}t] \\ & + (\mathbf{U}^{(0)}\exp[\mathbf{A}^{(1)}\epsilon t]\mathbf{V}^{(0)} - \mathbf{U}^{(0)}\mathbf{V}^{(0)}) \\ & + (\mathbf{U}^{(0)}\mathbf{U}^{(1)}\exp[\mathbf{A}^{(2)}\epsilon^2 t]\mathbf{V}^{(1)}\mathbf{V}^{(0)} - \mathbf{U}^{(0)}\mathbf{U}^{(1)}\mathbf{V}^{(1)}\mathbf{V}^{(0)}) \\ & \quad \vdots \\ & + (\mathbf{U}^{(0)} \cdots \mathbf{U}^{(k-1)}\exp[\mathbf{A}^{(k)}\epsilon^k t]\mathbf{V}^{(k-1)} \cdots \mathbf{V}^{(0)} \\ & \quad - \mathbf{U}^{(0)} \cdots \mathbf{U}^{(k-1)}\mathbf{V}^{(k-1)} \cdots \mathbf{V}^{(0)}) + O(\epsilon),\end{aligned} \tag{2.6}$$

where $\mathbf{V}^{(k)} \equiv \tilde{\mathbf{V}}^{(k)}(0) = \mathbf{V}^{(k)}(0)$.

This approximation is uniformly valid for $t \geq 0$.[5]

As indicated in the previous section, (1)–(4) of the algorithm are very similar in structure to the algorithm of Courtois. In particular, compare (1.10) and (2.5). The computation in step (2) of the ergodic probabilities that form $\mathbf{U}^{(k)}$ is identical to the corresponding step of Courtois's algorithm. The critical difference, however, is the computation of the "membership matrix" $\mathbf{V}^{(k)}(\epsilon)$. In particular, "membership," as needed here is defined in (2.2). Specifically, for each state j in the process corresponding to $\mathbf{A}^{(k)}(\epsilon)$, we compute the probability that the process *first enters* each ergodic class E_I of $\mathbf{A}^{(k)}(0)$. If j is already a member of some E_I, then the corresponding $v_{I,j}^{(k)}(\epsilon)$ equals 1, that is, in this case we have exactly the same membership as if we used $V^{(k)}(0)$, the quantity employed in Courtois's algorithm. Furthermore, if j is a transient state of $\mathbf{A}^{(k)}(0)$ that does not couple ergodic classes (i.e., if j has transitions in $\mathbf{A}^{(k)}(\epsilon)$ into only one of the E_I), we still have the same 0–1 membership as in $\mathbf{V}^{(k)}(0)$. However, if j *is* a coupling transient state, $v_{I,j}^{(k)}(\epsilon)$ in general will be nonzero and ϵ-dependent for several values of I. Although there is some ϵ-dependence to be captured here, (2.1a) indicates that we actually only need to match the lowest order term in each $v_{I,j}^{(k)}(\epsilon)$ and then can pick higher order terms as we like in order to ensure that the probabilities of membership sum to 1 (eq. (2.1b)). This has important computational implications, as we discuss in Section 4.

As indicated above, the only elements of $\mathbf{V}^{(k)}(\epsilon)$ that require calculation are those that correspond to the transient state set T. The calculation of (2.2), then, is a standard problem: We replace each ergodic class E_I of $\mathbf{A}^{(k)}(0)$ with a single trapping state I and sum together all transition rates from each $j \in T$ into each E_I, forming an aggregate rate into the new state I; the probabilities in (2.2) are then simply the limiting transition probabilities as $t \to \infty$ of this simplified process. Furthermore, this is equivalent to considering the limiting possibilities of the derived discrete-time Markov chain whose transition at discrete time n corresponds to the nth transition of the continuous time process. The state transition matrix $\mathbf{P}(\epsilon)$ of this discrete-time process (with ergodic classes of $\mathbf{A}^{(k)}$ collapsed into trapping states)

[5] Specifically, $O(\epsilon)$ is some (matrix) function $\mathbf{F}(\epsilon, t)$ such that $\lim_{\epsilon \downarrow 0} \sup_{t \geq 0} \| [\mathbf{F}(\epsilon, t)/\epsilon] \| = \mu < \infty$.

can be obtained directly from the original generator $\mathbf{A}^{(k)}(\epsilon)$.

$$p_{k,j}(\epsilon) = \frac{a_{k,j}(\epsilon)}{-a_{j,j}(\epsilon)}, \qquad p_{I,j}(\epsilon) = \sum_{i \in E_I} \frac{a_{i,j}(\epsilon)}{-a_{j,j}(\epsilon)}, \qquad p_{j,I}(\epsilon) = 0, \tag{2.7}$$

where $j, k \in T, j \neq k$, and I is a state representing a class E_I. By suitably ordering the states, $\mathbf{P}(\epsilon)$ can be formed as

$$\mathbf{P}(\epsilon) = \begin{bmatrix} \mathbf{P}_{TT}(\epsilon) & 0 \\ \mathbf{P}_{TR}(\epsilon) & I \end{bmatrix}, \tag{2.8}$$

and the limit therefore becomes

$$\lim_{n \to \infty} \mathbf{P}(\epsilon)^n = \begin{bmatrix} 0 & 0 \\ \mathbf{P}_{TR}(\epsilon)(I - \mathbf{P}_{TT}(\epsilon))^{-1} & I \end{bmatrix} \equiv \begin{bmatrix} 0 \\ \mathbf{V}(\epsilon) \end{bmatrix}. \tag{2.9}$$

The leading order terms of $\mathbf{V}(\epsilon)$ in (2.9) required in step (3) of the algorithm can be obtained in a variety of ways such as by repeated multiplication of $\mathbf{P}(\epsilon)$ (retaining only the leading order terms after each multiplication) or by series expansion of the inverse in (2.9) as

$$(I - \mathbf{P}_{TT}(\epsilon))^{-1} = (I - \mathbf{P}_{TT}(0))^{-1} \sum_{m=0}^{\infty} \epsilon^m (\mathbf{L}(\epsilon)(I - \mathbf{P}_{TT}(0)))^m,$$

where

$$\mathbf{L}(\epsilon) \equiv \frac{1}{\epsilon} (\mathbf{P}_{TT}(\epsilon) - \mathbf{P}_{TT}(0)).$$

Example 1. In order to illustrate the algorithm, consider the generator $\mathbf{A}^{(0)}(\epsilon)$ associated with the state transition diagram in Figure 1. The communicating classes and transient set are

$$E_1 = \{1\}, \qquad E_2 = \{2\}, \qquad E_3 = \{4\}, \qquad T = \{3\}.$$

The ergodic probabilities are all degenerate in this case;

$$u_{1,1} = u_{2,2} = u_{4,3} = 1 \qquad \text{or} \qquad \mathbf{U}^{(0)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Suitable terms $\tilde{v}(\epsilon)$ that satisfy (2.1) above are

$$\tilde{v}_{1,3}(\epsilon) = \tilde{v}_{2,3}(\epsilon) = \frac{1}{2} - \frac{\epsilon}{4}$$

or

$$\tilde{\mathbf{V}}^{(0)}(\epsilon) = \begin{bmatrix} 1 & 0 & \frac{1}{2} - \frac{\epsilon}{4} & 0 \\ 0 & 1 & \frac{1}{2} - \frac{\epsilon}{4} & 0 \\ 0 & 0 & \frac{\epsilon}{2} & 1 \end{bmatrix},$$

$$\tilde{v}_{3,3}(\epsilon) = \frac{\epsilon}{2}.$$

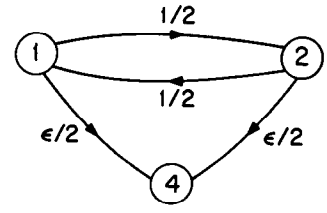


FIG. 2. $O(1/\epsilon)$ time scale.

Using these terms, $\mathbf{A}^{(1)}(\epsilon)$ computed using (2.5) generates the process illustrated in Figure 2:

$$\mathbf{A}^{(1)}(\epsilon) = \begin{bmatrix} -\frac{1}{2} - \frac{\epsilon}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & -\frac{1}{2} - \frac{\epsilon}{2} & 0 \\ \frac{\epsilon}{2} & \frac{\epsilon}{2} & 0 \end{bmatrix}.$$

This procedure is now repeated since $\mathbf{A}^{(1)}(0)$ generates two ergodic classes with the following ergodic probabilities:

$$E_1 = \{1, 2\}, \qquad E_2 = \{4\}, \qquad T = \phi,$$

$$\mathbf{U}^{(1)} = \begin{bmatrix} \frac{1}{2} & 0 \\ \frac{1}{2} & 0 \\ 0 & 1 \end{bmatrix}, \qquad \tilde{\mathbf{V}}^{(1)}(\epsilon) = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Using this, the generator $\mathbf{A}^{(2)}(\epsilon)$ is computed:

$$\mathbf{A}^{(2)}(\epsilon) = \begin{bmatrix} -\frac{1}{2} & 0 \\ \frac{1}{2} & 0 \end{bmatrix}.$$

Since $\mathbf{A}^{(2)}(0)$ generates only one ergodic class, namely $\{4\}$, the algorithm is terminated. The set of ϵ-independent Markov models from which the approximation is derived is shown in Figure 3.

Note that the process of Figure 1 has explicitly only order ϵ rates. However, as seen in Figure 3c, this process has time-scale behavior of order $1/\epsilon^2$. The fact that there is slower behavior than is explicitly visible in the original process is directly attributable to the presence of coupling transient states or, equivalently, to critical *sequences* of rare transitions. This is precisely the case in which the ϵ-dependence of $\tilde{\mathbf{V}}^{(k)}(\epsilon)$ is critical.

It is useful to make several comments about step (5) of the procedure that assembles an overall approximation of the transition probability matrix. The first term captures the fast, high-probability behavior at times of order 1. The next describes behavior at times of order $1/\epsilon$ by capturing transitions between ergodic classes of the fast process, and, since these transitions are sufficiently rare that the fast process can reach equilibrium between two such transitions, the probability mass within each ergodic class is distributed using the fast-process ergodic probabilities. Similar interpretations can be given to subsequent terms. Such intuition is

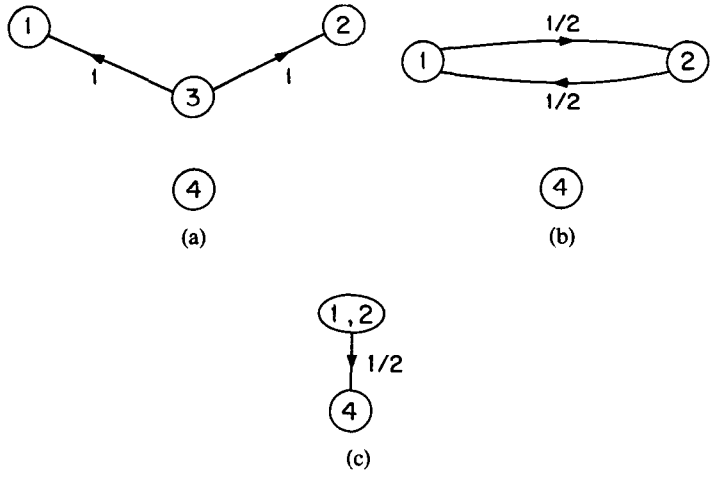


FIG. 3. (a) $O(1)$ time scale. (b) $O(1/\epsilon)$ time scale. (c) $O(1/\epsilon^2)$ time scale.

certainly present or implicit in most previous works. Indeed this idea has led researchers to develop iterative methods for computing steady-state probabilities [4] and error bounds for these computations [8]. In contrast, what we prove in the next section is that the error in this approximation to the entire transition probability matrix (including the full transient behavior) goes to 0 *uniformly* for $0 \leq t < \infty$ as $\epsilon \downarrow 0$. Coderch [6] has a similar uniform convergence proof, but our result is stronger since we are able to work on successively aggregated versions of the process, and we can also discard all but the essential ϵ-dependent terms (whereas Coderch [6] keeps them all). Finally, it is interestiong to note that the final approximation in (2.6) uses only $\tilde{\mathbf{V}}^{(k)}(0) = \mathbf{V}^{(k)}(0)$, the same matrices that appear in Courtois's development. The key point here is that while $\mathbf{V}^{(k)}(0)$ is adequate for describing the kth time scale, $\tilde{\mathbf{V}}^{(k)}(\epsilon)$ is in general needed to capture accurately all slower time scales. For example, the ϵ-dependent terms of $\tilde{\mathbf{V}}^{(0)}(\epsilon)$ in Example 1 directly influence $\tilde{\mathbf{V}}^{(1)}(0)$.

3. *The Derivation*

The algorithm for the construction of multiple-time-scale decompositions of a singularly perturbed, finite-state Markov process is derived in this section. At the same time, the uniform convergence of the associated approximation (2.6) is established. The approach taken is as follows. We first derive the algorithm assuming that there may be transient states at any particular time scale provided that these states cannot "couple" aggregates (i.e., aggregated ergodic classes) at slower time scales. The proof of uniform convergence in this case involves keeping track of "weak" terms in the generator that can ultimately be ignored since they do not affect the multiple-time-scale decomposition. The uniform convergence result for this case is stronger than that of Coderch [5] in that the continuous time analog of Courtois's multiple-time-scale procedure (using ergodic projections rather than the full perturbed eigenprojections) can be shown to provide a uniform approximation. Also, this result forms the backbone for our general algorithm in which we minimize the number of ϵ-dependent terms that must be computed in order to generate the complete multiple-time-scale decomposition and uniform approximation when there are transient states that couple aggregates. The

generalization to this case is proved in Section 3.2 by showing that it is equivalent to first constructing a process with an expanded state that does not have coupling transient states and then recovering the probability transition function of the original model after the procedure of Section 3.1 is applied to the expanded process.

3.1 No Coupling through Transient States. We first consider the case in which any almost transient state has transitions into a single ergodic class for $\epsilon > 0$. In this case we show that the "Courtois/Simon–Ando" approach is valid in that transient states have no effect on multiple-time-scale behavior. As mentioned in Section 2, in this case the $\tilde{\mathbf{V}}^{(k)}(\epsilon)$ are composed of entries which are either 0 or 1 since each transient state is associated with a unique ergodic class. In order to analyze the more restrictive case, the following result is useful:

Lemma 1. *Suppose*

$$\mathbf{F}(\epsilon) = \begin{bmatrix} \mathbf{F}_{11}(\epsilon) & \epsilon\mathbf{F}_{12}(\epsilon) \\ \epsilon\mathbf{F}_{21}(\epsilon) & \epsilon\mathbf{F}_{22}(\epsilon) \end{bmatrix}, \tag{3.1}$$

where

(1) $\mathbf{F}_{11}(\epsilon)$ *has eigenvalues with strictly negative real parts for all* $\epsilon \in [0, \epsilon_0)$ *for some* $\epsilon_0 > 0$,

(2) $\mathbf{F}(\epsilon)$ *has "well-defined multiple-time-scale behavior" in the sense defined in* [6].

Then

$$\mathbf{F}(\epsilon) \quad \textit{and} \quad \mathbf{H}(\epsilon) = \begin{bmatrix} \mathbf{F}_{11}(0) & 0 \\ 0 & \epsilon\mathbf{K}(\epsilon) \end{bmatrix},$$

where

$$\mathbf{K}(\epsilon) \equiv \mathbf{F}_{22}(\epsilon) - \epsilon\mathbf{F}_{21}(\epsilon)\mathbf{F}_{11}^{-1}(\epsilon)\mathbf{F}_{12}(\epsilon),$$

are "asymptotically equivalent" in the sense that

$$\sup_{t \geq 0} \| \exp[\mathbf{F}(\epsilon)t] - \exp[\mathbf{H}(\epsilon)t] \| = O(\epsilon).$$

Proof. This result is an adaptation of the basic perturbation result used by several authors. See, for example, Lou [16], Coderch et al. [6], or Kokotovic [13]. □

This result is applicable to perturbed Markov generators since (a) Coderch et al. [6] have shown that such matrices do have well-defined time-scale behavior and (b) it is straightforward to bring the generator into the form in (3.1) by an ϵ-independent similarity transformation. Before doing this, let us introduce the following definition:

Definition 1. Consider the perturbed Markov generator $\mathbf{A}(\epsilon) = \mathbf{A} + \mathbf{B}(\epsilon)$, with $\| \mathbf{B}(\epsilon) \| = O(\epsilon)$. There is *no coupling through transient states* in this process if the following conditions hold. It is possible to partition the state set into sets R_K, each of which consists, at $\epsilon = 0$, of a single ergodic class E_K together, perhaps, with some transient states T_K so that these transient states have transitions only into the particular class with which they are associated, even with $\epsilon > 0$. That is, if $m \in T_K$, then for any state $n \notin R_K$, $a_{n,m}(\epsilon) = 0$.

If we assume that $\mathbf{A}^{(0)}(\epsilon)$ has no coupling through transient states, we can order the states of the process so that

$$\mathbf{A}^{(0)}(\epsilon) = \mathbf{A}^{(0)} + \mathbf{B}^{(0)}(\epsilon),$$

where $\| \mathbf{B}^{(0)}(\epsilon) \| = O(\epsilon)$ and

$$\mathbf{A}^{(0)} = \operatorname{diag}(A_1, A_2, \ldots, A_N).$$

Each A_I corresponds to a process with a single ergodic class and possibly some transient states that are uniquely associated with that class. If such states are present, then the no coupling assumption implies that certain corresponding elements of $\mathbf{B}^{(0)}(\epsilon)$ are identically zero.

In order to transform $\mathbf{A}^{(0)}(\epsilon)$ into the form (3.1), let $\mathbf{U}^{(0)}$ and $\mathbf{V}^{(0)}$ denote the matrices of right and left zero eigenvectors of the unperturbed generator $\mathbf{A}^{(0)}$, where the kth column of $\mathbf{U}^{(0)}$ and the kth row of $\mathbf{V}^{(0)}$ have nonzero entries only corresponding to the states in the set R_k. Note that the matrices $\mathbf{U}^{(0)}$ and $\mathbf{V}^{(0)}$ correspond to $\mathbf{U}^{(k)}$ and $\tilde{\mathbf{V}}^{(k)}(\epsilon)$ constructed in the Algorithm since there is no ϵ-dependence in this case. Also, let $\mathbf{Y}^{(0)}(\mathbf{Z}^{(0)})$ be matrices whose columns (rows) span the right (left) eigenspace of the nonzero eigenvalues of $\mathbf{A}^{(0)}$. Furthermore, owing to the structure of $\mathbf{A}^{(0)}$, we can clearly choose these matrices such that $\mathbf{A}^{(0)}\mathbf{Y}^{(0)}$ and $\mathbf{Z}^{(0)}\mathbf{A}^{(0)}$ are block diagonal matrices with partitions consistent with $\mathbf{A}^{(0)}$ and that a similarity transformation T can then be constructed as

$$\mathbf{T} = \begin{bmatrix} \mathbf{Z}^{(0)} \\ \mathbf{V}^{(0)} \end{bmatrix}, \qquad \mathbf{T}^{-1} = [\mathbf{Y}^{(0)}\mathbf{U}^{(0)}].$$

Application of this similarity transformation $\mathbf{A}^{(0)}(\epsilon)$ results in the form (3.1) given for $\mathbf{F}(\epsilon)$ in Lemma 1:

$$\mathbf{T}\mathbf{A}^{(0)}(\epsilon)\mathbf{T}^{-1} = \begin{bmatrix} \mathbf{A}_{11}(\epsilon) & \epsilon\mathbf{A}_{12}(\epsilon) \\ \epsilon\mathbf{A}_{21}(\epsilon) & \epsilon\mathbf{A}_{22}(\epsilon) \end{bmatrix},$$

where

$$\begin{aligned} \mathbf{A}_{11}(\epsilon) &= \mathbf{Z}^{(0)}\mathbf{A}^{(0)}(\epsilon)\mathbf{Y}^{(0)}, \\ \epsilon\mathbf{A}_{12}(\epsilon) &= \mathbf{Z}^{(0)}\mathbf{B}^{(0)}(\epsilon)\mathbf{U}^{(0)}, \\ \epsilon\mathbf{A}_{21}(\epsilon) &= \mathbf{V}^{(0)}\mathbf{B}^{(0)}(\epsilon)\mathbf{Y}^{(0)}, \\ \epsilon\mathbf{A}_{22}(\epsilon) &= \mathbf{V}^{(0)}\mathbf{B}^{(0)}(\epsilon)\mathbf{U}^{(0)}. \end{aligned}$$

Since $\mathbf{Z}^{(0)}$ and $\mathbf{Y}^{(0)}$ are associated with the nonzero eigenvalues of $\mathbf{A}^{(0)}$ and since the original system has no eigenvalues in the right half-plane, $\mathbf{A}_{11}(\epsilon)$ satisfies the conditions of Lemma 1. Applying Lemma 1 and expressing the result in the original basis yields the following uniform asymptotic approximation

$$\begin{aligned} \exp[\mathbf{A}^{(0)}(\epsilon)t] &= \mathbf{Y}^{(0)}\exp[\mathbf{A}_{11}(0)t]\mathbf{Z}^{(0)} + \mathbf{U}^{(0)}\exp[\mathbf{G}^{(1)}(\epsilon)\epsilon t]\mathbf{V}^{(0)} + O(\epsilon) \\ &= \exp[\mathbf{A}^{(0)}t] + \mathbf{U}^{(0)}\exp[\mathbf{G}^{(1)}(\epsilon)\epsilon t]\mathbf{V}^{(0)} - \mathbf{U}^{(0)}\mathbf{V}^{(0)} + O(\epsilon), \end{aligned} \tag{3.2a}$$

where

$$\mathbf{G}^{(1)}(\epsilon) \equiv \mathbf{A}_{22}(\epsilon) - \epsilon\mathbf{A}_{21}(\epsilon)\mathbf{A}_{11}^{-1}(\epsilon)\mathbf{A}_{12}(\epsilon). \tag{3.2b}$$

From (3.2a) we see that the problem of uniformly approximating $\exp[\mathbf{A}^{(0)}(\epsilon)t]$ has been reduced to that of approximating $\exp[\epsilon\mathbf{G}^{(1)}(\epsilon)t]$; one time scale has been "peeled off" leaving a lower dimension problem. However, the procedure is not perfectly inductive since $\mathbf{G}^{(1)}(\epsilon)$ need not be the generator of a Markov chain. On

the other hand, it is very close to being one.[6] Specifically, a careful examination of (3.2b) shows that $\mathbf{G}^{(1)}(\epsilon)$ can be expressed as

$$\mathbf{G}^{(1)}(\epsilon) = \mathbf{A}^{(1)}(\epsilon) + \mathbf{W}^{(1)}(\epsilon),$$

where $\mathbf{A}^{(1)}(\epsilon)$ is a Markov generator given by

$$\begin{aligned}\mathbf{A}^{(1)}(\epsilon) &= \frac{1}{\epsilon}\mathbf{V}^{(0)}\mathbf{A}^{(0)}(\epsilon)\mathbf{U}^{(0)} \\ &= \frac{1}{\epsilon}\mathbf{V}^{(0)}\mathbf{B}^{(0)}(\epsilon)\mathbf{U}^{(0)} \equiv \mathbf{A}^{(1)} + \mathbf{B}^{(1)}(\epsilon),\end{aligned} \tag{3.3}$$

where

$$\mathbf{A}^{(1)} \equiv \mathbf{A}^{(1)}(0), \qquad \|\mathbf{B}^{(1)}(\epsilon)\| = O(\epsilon),$$

and

$$\mathbf{W}^{(1)}(\epsilon) = -\frac{1}{\epsilon}\mathbf{V}^{(0)}\mathbf{B}^{(0)}(\epsilon)\mathbf{Y}^{(0)}(\mathbf{Z}^{(0)}\mathbf{A}^{(0)}(\epsilon)\mathbf{Y}^{(0)})^{-1}\mathbf{Z}^{(0)}\mathbf{B}^{(0)}(\epsilon)\mathbf{U}^{(0)}, \tag{3.4}$$

where it is straightforward to show that $\|\mathbf{W}^{(1)}(\epsilon)\| = O(\epsilon)$.

What will be shown is that the term $\mathbf{W}^{(1)}(\epsilon)$ can be entirely neglected. In the two-time-scale case, this follows from the fact that $\mathbf{A}^{(1)}(\epsilon)$ is regularly perturbed, since all the nonzero eigenvalues of $\mathbf{A}^{(1)}(\epsilon)$ are $O(1)$, and from the fact that $\mathbf{W}^{(1)}(0) = 0$. $\mathbf{G}^{(1)}(\epsilon)$ can then be uniformly approximated using $\mathbf{G}^{(1)}(0) = \mathbf{A}^{(1)}$. This yields the two-time-scale result

$$\exp[\mathbf{A}^{(0)}(\epsilon)t] = \exp[\mathbf{A}^{(0)}t] + \mathbf{U}^{(0)}\exp[\mathbf{A}^{(1)}\epsilon t]\mathbf{V}^{(0)} - \mathbf{U}^{(0)}\mathbf{V}^{(0)} + O(\epsilon).$$

If there are more than two time scales 1 and $1/\epsilon$ in the original process, $\mathbf{A}^{(1)}(\epsilon)$ is again a singularly perturbed Markov generator. $\mathbf{W}^{(1)}(\epsilon)$, therefore, cannot be ignored only on the basis of its being $O(\epsilon)$ when considering the order $1/\epsilon^2$ and slower time scales, as was done above. In particular, discarding an arbitrary $O(\epsilon)$ term can lead to errors in subsequent time-scale approximations. Thus, in order to show that we *can* discard $\mathbf{W}^{(1)}(\epsilon)$, we must determine some special property that it possesses that guarantees that $\mathbf{W}^{(1)}(\epsilon)$ has *no* effect on slower time-scale approximations. To do this, let us first give a precise definition of what we mean by "weak" terms associated with a Markov generator.

Definition 2. Let $\mathbf{F}(\epsilon)$ be the generator of a Markov process with one ergodic class for $\epsilon > 0$. $\mathbf{W}(\epsilon)$ is *weak* with respect to $\mathbf{F}(\epsilon)$ if (a) $1^{\mathrm{T}}\mathbf{W}(\epsilon) = 0$ and (b) for any element $w_{i,j}(\epsilon)$ there exists a path $\mathbf{S} = (s_1 = j, s_2 \cdots s_k = i)$ through the process state space such that

$$w_{i,j}(\epsilon) = \epsilon O(f_{s_2,s_1}f_{s_3,s_2}\cdots f_{s_k,s_{k-1}}). \tag{3.5}$$

Condition (a) is necessary to avoid perturbation of the zero eigenvalue of $\mathbf{F}(\epsilon)$, which is associated with the sum of the probabilities being identically 1. In the derivations presented, however, this condition is satisfied by construction; therefore, we concentrate on property (b). Roughly what this property means is that if we think of $w_{i,j}(\epsilon)$ as a "transition rate" from state j to state i (although it may be negative), we can find a product of rates in the generator $\mathbf{F}(\epsilon)$ leading from j to i that is of lower order in ϵ and therefore represents a significantly more likely sequence of events.

[6] Though the columns of $\mathbf{G}^{(1)}(\epsilon)$ sum to zero, some of the off-diagonal elements may be small but negative.

In the Appendix we provide a proof of the following:

LEMMA 2. *Suppose that* $\mathbf{A}^{(0)}(\epsilon)$ *is as in* (1.3) *and* (1.4) *and there is no coupling through transient states; then* $\mathbf{W}^{(1)}(\epsilon)$ (3.4) *is weak with respect to* $\mathbf{A}^{(1)}(\epsilon)$ *in* (3.3).

Thanks to this lemma, an iterative procedure can now be defined and analyzed. Specifically, suppose that we have constructed $\mathbf{G}^{(k)}(\epsilon) = \mathbf{A}^{(k)}(\epsilon) + \mathbf{W}^{(k)}(\epsilon)$, where (a) $\mathbf{A}^{(k)}(\epsilon) = \mathbf{A}^{(k)} + \mathbf{B}^{(k)}(\epsilon)$ is a Markov generator with no coupling through transient states, $\| \mathbf{B}^{(k)}\epsilon \| = O(\epsilon)$, and (b) $\mathbf{G}^{(k)}(\epsilon)$ has well-defined time-scale behavior. Applying Lemma 1 and stating the result as in (3.2), we obtain the following uniform approximation:

$$\exp[\mathbf{G}^{(k)}(\epsilon)t] = \exp[\mathbf{A}^{(k)}t] + \mathbf{U}^{(k)}\exp[\mathbf{G}^{(k+1)}(\epsilon)\epsilon t]\mathbf{V}^{(k)} - \mathbf{U}^{(k)}\mathbf{V}^{(k)} + O(\epsilon),$$

where

$$\begin{aligned} \mathbf{G}^{(k+1)}(\epsilon) &= \mathbf{A}^{(k+1)}(\epsilon) + \mathbf{W}^{(k+1)}(\epsilon), \\ \mathbf{A}^{(k+1)}(\epsilon) &= \frac{1}{\epsilon}\mathbf{V}^{(k)}\mathbf{A}^{(k)}(\epsilon)\mathbf{U}^{(k)} = \frac{1}{\epsilon}\mathbf{V}^{(k)}\mathbf{B}^{(k)}(\epsilon)\mathbf{U}^{(k)}, \end{aligned} \tag{3.6}$$

and

$$\begin{aligned} \mathbf{W}^{(k+1)}(\epsilon) &= \mathbf{W}_1^{(k+1)}(\epsilon) + \mathbf{W}_2^{(k+1)}(\epsilon), \\ \mathbf{W}_1^{(k+1)}(\epsilon) &= \frac{1}{\epsilon} V^{(k)}\mathbf{W}^{(k)}(\epsilon)\mathbf{U}^{(k)}, \\ \mathbf{W}_2^{(k+1)}(\epsilon) &= -\frac{1}{\epsilon}\mathbf{V}^{(k)}(\mathbf{B}^{(k)}(\epsilon) + \mathbf{W}^{(k)}(\epsilon))\mathbf{Y}^{(k)}(\mathbf{Z}^{(k)}\mathbf{G}^{(k)}(\epsilon)\mathbf{Y}^{(k)})^{-1}\mathbf{Z}^{(k)} \\ &\quad \times (\mathbf{B}^{(k)}(\epsilon) + \mathbf{W}^{(k)}(\epsilon))U^{(k)}. \end{aligned}$$

Note that for $k = 2, 3, \ldots$ the term $\mathbf{W}^{(k)}(\epsilon)$ consists of two parts, namely, the "projection" $\mathbf{W}_1^{(k)}(\epsilon)$ of the preceding weak term $\mathbf{W}^{(k+1)}(\epsilon)$, and a new term $\mathbf{W}_2^{(k)}(\epsilon)$ defined similarly to the weak term computed previously in (3.4). We know from Lemma 1 that under the conditions stated above $\mathbf{G}^{(k+1)}(\epsilon)$ has well-defined time scales and by construction that $\mathbf{A}^{(k+1)}(\epsilon)$ is a Markov generator. By assumption in this section, there is no coupling through transient states in $\mathbf{A}^{(k+1)}(\epsilon)$. What we must show, however, is that the property of "weakness" is preserved; that is, we must show that *both* $\mathbf{W}_1^{(k+1)}(\epsilon)$ and $\mathbf{W}_2^{(k+1)}(\epsilon)$ are weak. Thus, in order to continue the iterative procedure, we need to verify the following, which is done in the Appendix.

LEMMA 3. *Suppose that* $\mathbf{G}^{(k)}(\epsilon) = \mathbf{A}^{(k)}(\epsilon) + \mathbf{W}^{(k)}(\epsilon)$ *satisfies the following*:

(1) $\mathbf{G}^{(k)}(\epsilon)$ *has well-defined time-scale behavior*;
(2) $\mathbf{A}^{(k)}(\epsilon) = \mathbf{A}^{(k)} + \mathbf{B}^{(k)}(\epsilon)$ *is a Markov generator with no coupling through transient states,* $\| \mathbf{B}^{(k)}(\epsilon) \| = O(\epsilon)$;
(3) $\mathbf{W}^{(k)}(\epsilon)$ *is weak with respect to* $\mathbf{A}^{(k)}(\epsilon)$.

Then

$\mathbf{G}^{(k+1)}(\epsilon)$ *has well-defined time-scale behavior, and*
$\mathbf{W}^{(k+1)}(\epsilon)$ *is weak with respect to* $\mathbf{A}^{(k+1)}(\epsilon)$ *in* (3.6).

What we now have is the following: We proceed by first applying (3.2), followed by the iterative application of (3.6). At each stage we accumulate weak terms. However, at the *last* time scale, we *know* that we can discard the weak terms, since there are no further time scales to be perturbed. Consequently, we can actually discard these weak terms *as we proceed*, since we know that they only produce

weak terms at slower time scales. Thus, the following sequence of approximations is constructed for a system exhibiting k time scales and no coupling transient states at any intermediate time scale:

$$\begin{aligned}
&\exp[\mathbf{A}^{(0)}(\epsilon)t] = \exp[\mathbf{A}^{(0)}t] + \mathbf{U}^{(0)}\exp[\mathbf{G}^{(1)}(\epsilon)\epsilon t]\mathbf{V}^{(0)} - \mathbf{U}^{(0)}\mathbf{V}^{(0)} + O(\epsilon),\\
&\exp[\mathbf{G}^{(1)}(\epsilon)t] = \exp[\mathbf{A}^{(1)}t] + \mathbf{U}^{(1)}\exp[\mathbf{G}^{(2)}(\epsilon)\epsilon t]\mathbf{V}^{(1)} - \mathbf{U}^{(1)}\mathbf{V}^{(1)} + O(\epsilon),\\
&\qquad\vdots\\
&\exp[\mathbf{G}^{(k-2)}(\epsilon)t] = \exp[\mathbf{A}^{(k-2)}t] + \mathbf{U}^{(k-2)}\exp[\mathbf{G}^{(k-1)}(\epsilon)\epsilon t]\mathbf{V}^{(k-2)}\\
&\qquad\qquad - \mathbf{U}^{(k-2)}\mathbf{V}^{(k-2)} + O(\epsilon),\\
&\exp[\mathbf{G}^{(k-1)}(\epsilon)t] = \exp[\mathbf{A}^{(k-1)}t] + O(\epsilon).
\end{aligned} \tag{3.7}$$

Note that there is no problem here in determining when to stop the procedure. Stop when $\mathbf{A}^{(k-1)}$ has exactly one ergodic class. From Coderch [5] we know that, since $\mathbf{A}^{(0)}(\epsilon)$ does have well-defined time-scale behavior, there is a k such that this is true, and this k is associated with the slowest time scale. The approximation (2.6) follows directly by collapsing the equations in (3.7).

Note also that in order to construct this approximation, we never need to calculate $\mathbf{Y}^{(k)}$, $\mathbf{Z}^{(k)}$, or any of the terms $\mathbf{W}^{(k)}(\epsilon)$. Rather, at each time scale we begin with $\mathbf{A}^{(k)}(\epsilon) = \mathbf{A}^{(k)} + \mathbf{B}^{(k)}(\epsilon)$ and compute the ergodic classes and probabilities associated with $\mathbf{A}^{(k)}$ to form $\mathbf{U}^{(k)}$ and $\mathbf{V}^{(k)}$. $\mathbf{A}^{(k+1)}(\epsilon)$ is then calculated using (3.6). At this point, of course, we have only dealt with the case in which there is no coupling through transient states at any stage of the procedure. We now modify the procedure in order to remove this restriction.

3.2 TRANSIENT STATES THAT COUPLE AGGREGATES. Our basic approach to this general case is to reduce it to the one considered in the previous subsection by expanding the state space, when necessary, by defining an associated generator that satisfies the no-coupling constraint. Specifically, consider a generator $\mathbf{A}(\epsilon) = \mathbf{A} + \mathbf{B}(\epsilon)$ where A generates N ergodic classes. The state space can be partitioned into $N + 1$ parts $E_1, E_2, \ldots, E_N, T$ where the E_K are the ergodic classes of A and T is the set of transient states. The set T is then "split" into N copies $T_1, T_2, \ldots, T_N$ such that each copy is associated with a unique ergodic class. Specifically an associated generator $\hat{\mathbf{A}}(\epsilon) = \hat{\mathbf{A}} + \hat{\mathbf{B}}(\epsilon)$ is constructed on this expanded state space such that once the process is in a state $s \in T_k$, the next state entered that belongs to $E \equiv E_1 \cup E_2 \cup \cdots \cup E_N$ *must* be in E_k. By construction, then, $\hat{\mathbf{A}}(\epsilon)$ satisfies Definition 1. The precise nature of this construction can be stated as follows:

LEMMA 4. *Let* $\mathbf{A}(\epsilon) = \mathbf{A} + \mathbf{B}(\epsilon)$ *and let* $\mathbf{U}$ *and* $\mathbf{V}$ *be the ergodic probability and membership matrices for the unperturbed generator* $\mathbf{A}$. *Then there exist* $\mathbf{C}$, $\mathbf{D}(\epsilon)$, $\hat{\mathbf{A}}(\epsilon) = \hat{\mathbf{A}} + \hat{\mathbf{B}}(\epsilon)$, *and* $\hat{\mathbf{U}}$ *and* $\hat{\mathbf{V}}$ *similarly derived from* $\hat{\mathbf{A}}$ *such that*

(1) $\exp[\mathbf{A}(\epsilon)t] = \mathbf{C}\exp[\hat{\mathbf{A}}(\epsilon)t]D(\epsilon)$,
(2) $\hat{\mathbf{A}}(\epsilon)$ *does not exhibit coupling through transient states,*
(3) $\mathbf{C}\hat{\mathbf{U}} = \mathbf{U}$,
(4) $\hat{\mathbf{V}}\mathbf{D}(0) = \mathbf{V}$,
(5) $\mathbf{D}(\epsilon)\mathbf{U} = \mathbf{D}(0)\mathbf{U} = \hat{\mathbf{U}}$,
(6) $\mathbf{C}\hat{\mathbf{A}}(\epsilon)\hat{\mathbf{U}} = \mathbf{A}(\epsilon)\mathbf{U}$,
(7) *The range of* $\mathbf{D}(\epsilon)$ *is* $\hat{\mathbf{A}}(\epsilon)$*-invariant.*

The construction of $\hat{\mathbf{A}}(\epsilon)$ can be described as follows. Let i, k be elements of E (i.e., recurrent states of $\mathbf{A}(\epsilon)$ and $\hat{\mathbf{A}}(\epsilon)$). Then, the transition probability from i to k in $\hat{\mathbf{A}}(\epsilon)$ is the same as in $\mathbf{A}(\epsilon)$. Next, let $j \in T$, and let $j_1, \ldots, j_N$ denote the

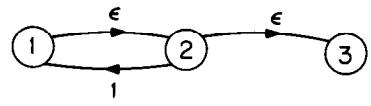


FIG. 4. Example 2.

corresponding copies of j in the expanded process. The basic idea behind the construction is that a transition to the state j_I corresponds to a transition in the original process to state j, *together with* the decision that the next ergodic class that will be entered is E_I. Consequently, the transition rates into the j_I must reflect the probability of this additional decisions. Specifically, if $k \in E$, then

$$\hat{a}_{j_I,k}(\epsilon) = a_{j,k}(\epsilon)\nu_{I,j}(\epsilon), \tag{3.8}$$

where $\nu_{I,j}(\epsilon)$, defined in (2.2), is precisely the probability of that decision. Similarly, transitions out of j_I must be adjusted to reflect conditioning on knowledge of which ergodic class will be visited next. Specifically, the transition rate from j_I to any state in an ergodic class other than E_I is 0, as is the rate from j_I to any state in T_K, $K \neq I$, that is, to any copy of any transient state corresponding to a subsequent transition into a different ergodic class. The remaining transition rates out of j_I are specified as follows:

$$\begin{aligned} \hat{a}_{i,j_I}(\epsilon) &= a_{i,j}(\epsilon)\frac{1}{\nu_{I,j}(\epsilon)}, \qquad i \in E_I, \\ \hat{a}_{k_I,j_I}(\epsilon) &= a_{k,j}(\epsilon)\frac{\nu_{I,k}(\epsilon)}{\nu_{I,j}(\epsilon)}, \qquad k_I \in T_I. \end{aligned} \tag{3.9}$$

The construction of $\mathbf{C}$ is quite simple: The various copies of each transient state are collapsed by summing their probabilities. Specifically, for each $i \in E$, $c_{i,i} = 1$, and $c_{j,j_I} = 1$ for each $j \in T$ and all its copies $j_1, \ldots, j_N$. All other elements of $\mathbf{C}$ are 0. In the case of $\mathbf{D}(\epsilon)$ the initial probability of each transient state j must be split by again making a decision concerning which E_I is visited first. Thus, for each $i \in E$, $[\mathbf{D}(\epsilon)]_{i,i} = 1$, while for $j \in I$

$$d_{j_I,j}(\epsilon) = \nu_{I,j}(\epsilon), \tag{3.10}$$

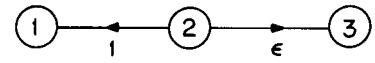
with all other elements of $\mathbf{D}(\epsilon)$ equal to 0. The several properties (1)–(7) in the lemma then follow directly from the construction (see Rohlicek [18] for detailed verification).

Example 2. We illustrate the state expansion construction on the simple process depicted in Figure 4 for which

$$\mathbf{A}^{(0)}(\epsilon) = \begin{bmatrix} -\epsilon & 1 & 0 \\ \epsilon & -1-\epsilon & 0 \\ 0 & \epsilon & 0 \end{bmatrix}.$$

In this case the construction of Lemma 4 calls for a splitting of the transient state 2. Following the procedure cited in Lemma 4, the key quantities are the probabilities that the perturbed process first enters each of the unperturbed recurrent classes (viz., $E_1 = \{1\}$ and $E_2 = \{3\}$), given that it starts in any particular transient state. These can be compared as the limiting probabilities of the process illustrated in Figure 5, which is obtained from the chain in Figure 4 by making each unperturbed recurrent class a trapping state. The expanded state process is depicted in Figure 6

FIG. 5. Modified process.



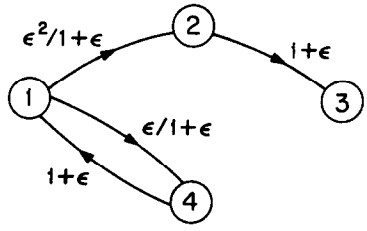


FIG. 6. Expanded state process.

and the associated matrices are

$$\hat{\mathbf{C}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \qquad \hat{\mathbf{D}}(\epsilon) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \epsilon/(1+\epsilon) & 0 \\ 0 & 0 & 1 \\ 0 & \epsilon/(1+\epsilon) & 0 \end{bmatrix},$$

$$\hat{\mathbf{U}} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}, \qquad \hat{\mathbf{V}} = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}.$$

Note that, as desired, states 2 and 4 in Figure 6 are transient but do *not* couple the ergodic classes {1} and {3}. Consequently, the procedure of Section 3.1 can be directly applied.

A similar expansion of the state set can be performed using the generator $\mathbf{G}^{(k)}(\epsilon)$ defined in (3.6). In this case, the following properties also follow:

LEMMA 5. *Suppose* $\mathbf{G}(\epsilon) = \mathbf{A}(\epsilon) + \mathbf{W}(\epsilon)$ *where* $\mathbf{A}(\epsilon)$ *is a Markov generator and* $\mathbf{W}(\epsilon)$ *is weak with respect to* $A(\epsilon)$. *Let* $\mathbf{C}$, $\mathbf{D}(\epsilon)$, *and* $\mathbf{A}$ *be determined as in Lemma 4 from* $\mathbf{A}(\epsilon)$. *Then* $\hat{\mathbf{G}}(\epsilon) = \hat{\mathbf{A}}(\epsilon) + \hat{\mathbf{W}}(\epsilon)$ *can be constructed such that*

(1) $exp[\mathbf{G}(\epsilon)t] = \mathbf{C}\,exp[\hat{\mathbf{G}}(\epsilon)t]\mathbf{D}(\epsilon) + O(\epsilon)$ *and*
(2) $\hat{\mathbf{W}}(\epsilon)$ *is weak with respect to* $\hat{\mathbf{A}}(\epsilon)$.

Lemma 5 is essentially a minor extension of Lemma 4, and we limit ourselves here to a brief sketch of the proof. We refer the reader to Rohlicek [18] for a complete proof. The only complication here is that $\mathbf{G}(\epsilon)$ is not necessarily a Markov generator. Nevertheless, we can follow the same construction for $\mathbf{G}(\epsilon)$ as that for $\mathbf{A}(\epsilon)$, where in this case the $v_{J,i}(\epsilon)$, computed as in (2.7)–(2.9) with $g_{i,j}(\epsilon)$ in place of $a_{i,j}(\epsilon)$, are not given direct probabilistic interpretations. This construction yields the same $\mathbf{C}$ matrix as that produced from $\mathbf{A}(\epsilon)$ and a slightly different set of $v_{J,i}(\epsilon)$, which show up in both $\hat{\mathbf{G}}(\epsilon)$ and the corresponding $\mathbf{D}(\epsilon)$ matrix. The weakness of $\mathbf{W}(\epsilon)$, however, implies that the difference in the $v_{J,i}(\epsilon)$ values is higher order, from which we can immediately conclude that we can replace the $\mathbf{D}(\epsilon)$ computed from $\mathbf{G}(\epsilon)$ with that constructed from $\mathbf{A}(\epsilon)$ and incur only an $O(\epsilon)$ error. Finally, we can write $\hat{\mathbf{G}}(\epsilon) = \hat{\mathbf{A}}(\epsilon) + \mathbf{W}'(\epsilon) + \mathbf{W}''(\epsilon)$. Here $\mathbf{W}'(\epsilon)$ results directly from the construction, that is, it is obtained from $\mathbf{W}(\epsilon)$ in the same way $\hat{\mathbf{A}}(\epsilon)$ is obtained from $\mathbf{A}(\epsilon)$ (see (3.8)–(3.9)). The weakness of $\mathbf{W}(\epsilon)$ allows us to conclude that $\mathbf{W}'(\epsilon)$

is weak. The term $\mathbf{W}''(\epsilon)$ captures the fact that the $\hat{\mathbf{G}}(\epsilon)$ used slightly different $v_{I,j}(\epsilon)$ values than those used in $\hat{\mathbf{A}}(\epsilon)$. The fact that this difference is higher order allows us to conclude that $\mathbf{W}''(\epsilon)$ is also weak.

We now can piece together a complete algorithm: At any stage k we begin with $G^{(k)}(\epsilon) = \mathbf{A}^{(k)}(\epsilon) + \mathbf{W}^{(k)}(\epsilon)$ (starting with $\mathbf{G}^{(0)}(\epsilon) = \mathbf{A}^{(0)}(\epsilon)$); we first expand the state space, thereby eliminating all coupling transient states, and then perform the aggregation step described in Section 3.1 to produce $\mathbf{G}^{(k+1)}(\epsilon)$. This yields the following uniform approximations:

$$\begin{aligned}
&\exp[\mathbf{G}^{(k)}(\epsilon)t] \\
&\quad = \mathbf{C}^{(k)}\exp[\hat{\mathbf{G}}^{(k)}(\epsilon)t]\mathbf{D}^{(k)}(\epsilon) + O(\epsilon) \\
&\quad = \mathbf{C}^{(k)}(\exp[\hat{\mathbf{A}}^{(k)}t] + \hat{\mathbf{U}}^{(k)}\exp[\mathbf{G}^{(k+1)}(\epsilon)\epsilon t]\hat{\mathbf{V}}^{(k)} - \hat{\mathbf{U}}^{(k)}\hat{\mathbf{V}}^{(k)})\mathbf{D}^{(k)}(\epsilon) + O(\epsilon) \\
&\quad = \exp[\mathbf{A}^{(k)}t] + U^{(k)}\exp[\mathbf{G}^{(k+1)}(\epsilon)\epsilon t]\mathbf{V}^{(k)}(\epsilon) + O(\epsilon). \qquad (3.11)
\end{aligned}$$

Here $\hat{\mathbf{U}}^{(k)}$, $\hat{\mathbf{V}}^{(k)}$ are the ergodic probability and membership matrices corresponding to $\hat{\mathbf{A}}^{(k)}$, and

$$\mathbf{G}^{(k+1)}(\epsilon) = \frac{1}{\epsilon}\hat{\mathbf{V}}^{(k)}\hat{\mathbf{G}}^{(k)}(\epsilon)\hat{\mathbf{U}}^{(k)} = \mathbf{A}^{(k+1)}(\epsilon) + \mathbf{W}^{(k+1)}(\epsilon). \qquad (3.12)$$

Also, with $\mathbf{C}^{(k)}$ and $\mathbf{D}^{(k)}(\epsilon)$ constructed from $\hat{\mathbf{A}}^{(k)}(\epsilon)$ as in Lemma 4, we obtain the final form in (3.11), where it is straightforward to check that

$$\mathbf{V}^{(k)}(\epsilon) = \hat{\mathbf{V}}^{(k)}\mathbf{D}^{(k)}(\epsilon), \qquad (3.13)$$

where $\mathbf{V}^{(k)}(\epsilon)$ is defined in (2.4) using the actual $v_{I,j}(\epsilon)$.

Combining Lemmas 2–5 shows that this procedure yields the sequences of matrices $\mathbf{U}^{(i)}$, $\mathbf{V}^{(i)}$, $\mathbf{A}^{(i)}$, $i = 0, 1, \ldots$ and the uniform approximation (2.6). However, we can take this several steps farther. Specifically, although we have used state expansion in order to prove that we can construct a uniform approximation, we do not actually need to perform this expansion to obtain the approximation. Indeed, although (3.11) implies a two-step procedure for computing $\mathbf{A}^{(k+1)}(\epsilon)$, it is a straightforward consequence of Lemma 4 that we can compute $\mathbf{A}^{(k+1)}(\epsilon)$ directly from $\mathbf{A}^{(k)}(\epsilon)$:

$$\mathbf{A}^{(k+1)}(\epsilon) = \frac{1}{\epsilon}\mathbf{V}^{(k)}(\epsilon)\mathbf{A}^{(k)}(\epsilon)\mathbf{U}^{(k)} \qquad (3.14)$$

(see Rohlicek [18] for the demonstration of the validity of (3.11)).

Finally, when this procedure reaches the last time scale, we can discard *all* of the accumulated weak terms, since at this point they are a regular perturbation. Furthermore, we can also replace the $\mathbf{V}^{(k)}(\epsilon)$ with the $\tilde{\mathbf{V}}^{(k)}(\epsilon)$ introduced in the algorithm in Section 2, since the difference between these is of higher order and is consequently weak. This then yields the following:

THEOREM 1. *The iterative algorithm given in Section* 2 *(eqs.* (2.1)–(2.6)*) yields the uniform multiple-time-scale approximation* (2.6).

There is another extremely important consequence of the derivation we have just sketched. We state it in the following:

COROLLARY 1. *Let* $\mathbf{F}(\epsilon)$ *and* $\mathbf{G}(\epsilon)$ *be two Markov generators so that* $\mathbf{F}(\epsilon) = \mathbf{G}(\epsilon) + \mathbf{W}(\epsilon)$, *where* $\mathbf{W}(\epsilon)$ *is weak with respect to* $\mathbf{G}(\epsilon)$. *Then* $\mathbf{F}(\epsilon)$ *is asymptotically equivalent (defined in Lemma* 1*) to* $\mathbf{G}(\epsilon)$.

This corollary has the useful consequence that, if one is trying to construct an approximation of a Markov process with a generator $\mathbf{A}(\epsilon)$ that can be separated into a simpler generator $\bar{\mathbf{A}}(\epsilon)$ and a relatively weak part $\mathbf{W}(\epsilon)$, then the weak part can safely be "pruned." A direct application of this is that only the leading order terms in ϵ of any transition rate need to be considered in the construction of the approximation. This corollary not only implies that we can use $\tilde{\mathbf{V}}^{(k)}(\epsilon)$ rather than $\mathbf{V}^{(k)}(\epsilon)$ but also has significant additional computational implications elaborated on in the next section.

4. *Conclusion*

In this paper we have developed a new procedure for the hierarchical, multiple-time-scale approximation of singularly perturbed, finite-state Markov processes. Our results bridge the gap between conceptually simple results such as those of Courtois [7] and the significantly more complex results of Coderch [5] and Delebecque [9]. In addition to providing a general algorithm, our work also provides additional insight into the nature of multiple-time-scale approximations and the role played by almost transient states. In particular, if we write out the expression for a single element of $\mathbf{A}^{(k+1)}(\epsilon)$ in (2.5), we obtain

$$a_{J,I}^{(k+1)}(\epsilon) = \frac{1}{\epsilon} \sum_{j \in E_J} \sum_{i \in E_I} u_{i,I}^{(k)} a_{j,i}^{(k)}(\epsilon) + \frac{1}{\epsilon} \sum_{j \in T} \sum_{i \in E_I} u_{i,I}^{(k)} a_{j,i}^{(k)}(\epsilon) \tilde{v}_{J,i}^{(k)}(\epsilon). \tag{4.1}$$

The first term corresponds to the usual average rate between aggregates used by Courtois. The second term, on the other hand, involves transient states, and the additional weighting, captured by $\tilde{v}_{J,i}$, reflects the critical "split membership" of transient states that couple ergodic classes.

Another insight our work provides concerns Coderch's eigenprojection interpretation. In particular, as we have seen, the key to Coderch's approach is the eigenprojection $\mathbf{P}(\epsilon)$ of a Markov generator $\mathbf{A}(\epsilon)$. When there are no-coupling transient states, we can approximate $\mathbf{P}(\epsilon)$ by $\mathbf{P}(0)$, which has an easily computed factorization $\mathbf{UV}$ that can be exploited to construct an aggregated process at the next time scale. When there are such coupling states, this approach fails, but what our results show is that we can approximate $\mathbf{P}(\epsilon)$ by the factored approximation $\mathbf{U}\tilde{\mathbf{V}}(\epsilon)$, which can again be calculated in a straightforward manner and exploited to construct an aggregate approximation.

Application of our decomposition to the area of reliability analysis seems natural [17, 21]. If faults occur at rate $O(\epsilon)$ and are repaired at rate $O(1)$, then at $\epsilon = 0$, there exist many transient states. Furthermore, the goal of fault-tolerant design in general is to create an apparent failure rate that is orders of magnitude smaller than the natural failure rate. In the context of this paper, this corresponds to reaching some implied state only at time scales of order $1/\epsilon^2$ or slower. As we have seen, such implicit time-scale behavior requires the presence of coupling transient states.

Other applications may be found in engineering techniques based on very large Markov models. For example, such models have been used as the basis of estimation algorithms in speech recognition [2] and electrocardiogram analysis [11]. In applications, computational requirements grow quickly as more ambitious analysis tasks are undertaken. Use of multiple-time-scale decomposition of the underlying model may suggest possible hierarchical approximation methods that are computationally feasible.

Finally, let us comment on numerical and computational aspects of hierarchical, multiple-time-scale approximation algorithms in general and our procedure in particular. First of all, an assumption common to our work and previous treatments is the use of ϵ-*independent* Markov generators in the approximation at each time scale. This raises an important point that provides insight into why one would seek this type of approximation. Specifically, there is an implicit assumption in this and previous work that the ϵ-dependent perturbation terms in $A^{(0)}(\epsilon)$ capture all rare events and ill-conditioning in the original Markov process. To the extent that it is true, all of the $O(1)$ computations in our or any other procedure are well-conditioned. Thus, by using these ϵ-independent generators for each time scale, the approximation of (1.2) becomes a numerically stable problem, since the effect of the small parameter ϵ is isolated from the approximation at any particular time scale.

In our case the critical quantities to be calculated in each step of our algorithm are the ergodic probabilities that comprise $\mathbf{U}^{(k)}$ and the leading order terms of the trapping probabilities $v_{I,j}(\epsilon)$ for each transient state j. Since ϵ-dependence is completely absent in the $\mathbf{U}^{(k)}$ calculation, the terms of interest are guaranteed to be $O(1)$. The calculation of the leading order coefficient of $v_{I,j}(\epsilon)$ is also an $O(1)$ computation. In fact, referring to (2.7)–(2.8) and the accompanying discussion, we see that this computation consists of a clearly stable symbolic part—identifying the lowest power of ϵ in the various elements of $\mathbf{P}(\epsilon)^n$ and an $O(1)$ computation corresponding to the multiplication of the coefficients of these leading order terms as we compute the successive powers of $\mathbf{P}(\epsilon)$.

To illustrate what can happen if we allow ϵ-dependencies in the generators used at each time scale, consider again the process depicted in Figure 4. Suppose we initially group states 1 and 2 together as one ergodic class at the fastest time scale and state 3 as the other. In doing this, we keep the ϵ rate from state 2 to state 1 as part of our fast time-scale model (and in essence are then treating it in the same manner as the $O(1)$ terms), while the ϵ rate from 2 to 3 is viewed as a perturbation. With this grouping, there are no transient states, and thus we can directly apply Courtois's procedure. In doing this, we find that the "fast" ergodic probability vector for the $\{1, 2\}$ class is

$$u(\epsilon) = \begin{bmatrix} \dfrac{1}{1+\epsilon} \\ \dfrac{\epsilon}{1+\epsilon} \end{bmatrix},$$

which, as expected, contains a small value for the probability of being in state 2. This is the source of the difficulty with this approach. First of all, it becomes necessary to know ahead of time which small terms should be thought of as small and which should not. Also, since these probabilities are used as weights in computing the aggregate behavior at the next time scale, it is actually necessary to know the $O(\epsilon)$ component of $u(\epsilon)$ to within $O(\epsilon^2)$ in order to extract a uniformly valid approximation. As the next paragraph makes clear, this is a far more stringent numerical requirement than is needed in our procedure. Furthermore, if this approach is used for the model presented in Section 1, states 1, 2, and 3 must all be grouped together. Not only must the small ergodic probability be calculated, but what was a set of two small (degenerate) problems has become larger; the advantage of decomposition is partially lost.

Finally, let us comment on the significant computational implications of Corollary 1. Specifically, this states that it is only the leading order terms in all transition rates that matter at any stage of our procedure. Consequently, errors of order ϵ in the computation of $\mathbf{U}^{(k)}$ or $\tilde{\mathbf{V}}^{(k)}(\epsilon)$ *have no effect* on the asymptotic approximation, since errors that are introduced into the approximation by such perturbations in our calculations are at worst of the same order as the accuracy of the overall computation. This lemma also has another important implication. Specifically, thanks to this lemma, using only knowledge of the (integer) orders of the elements of $\mathbf{A}^{(k)}(\epsilon)$ we can determine the location of the nonzero entries in $\mathbf{U}^{(k)}$ and the orders of magnitudes of the entries in $\mathbf{V}^{(k)}(\epsilon)$. Therefore, the orders of magnitude of the transition rates in $\mathbf{A}^{(k+1)}(\epsilon)$ can be determined. Consequently, the problem of determining the structure of the full set of time-scale models (i.e., what states are aggregated at what stage and the orders of the transition rates between these aggregates) involves only connectivity calculations on the state transition graph where transitions are labeled with their orders of magnitude. Such analysis is then essentially an extension of the type of analysis method used by Siljak [19] for large-scale systems. This structural property suggests an interesting problem, namely, the effect that a change in the order of one or more transition rates has on the overall time-scale structure. Rohlicek [17] presents an example of this applied to the problem of determining the effect on overall system reliability of adjustments in component failure rates and the rates at which faults are detected or incorrectly indicated.

Appendix

A1. PROOF OF LEMMA 2.[7] First note that the term $[ZA(\epsilon)Y]^{-1}$ in (3.4) can be expressed as an infinite series

$$\begin{aligned}(\mathbf{ZA}(\epsilon)\mathbf{Y})^{-1} &= (\mathbf{Z}(\mathbf{A}+\mathbf{B}(\epsilon))\mathbf{Y})^{-1} \\ &= (\mathbf{I}+\mathbf{D}^{-1}\mathbf{ZB}(\epsilon)\mathbf{Y})^{-1}\mathbf{D}^{-1} \\ &= \sum_{m=0}^{\infty}(-\mathbf{D}^{-1}\mathbf{ZB}(\epsilon)\mathbf{Y})^{m}\mathbf{D}^{-1}, \qquad \text{where} \quad \mathbf{D}=\mathbf{ZAY}.\end{aligned}$$

Since $\mathbf{Z}$ and $\mathbf{Y}$ are associated with the nonzero eigenvalues of $\mathbf{A}$, $\mathbf{D}^{-1}$ exists. Substituting this expression into (3.4) gives

$$\begin{aligned}\epsilon\mathbf{W}^{(1)}(\epsilon) &= \mathbf{VB}(\epsilon)\mathbf{SB}(\epsilon)\mathbf{U}+\mathbf{VB}(\epsilon)\mathbf{SB}(\epsilon)\mathbf{SB}(\epsilon)\mathbf{U}+\cdots+ \\ &\equiv \epsilon\mathbf{C}_1(\epsilon)-\epsilon\mathbf{C}_2(\epsilon)+\cdots+ \\ &\qquad\qquad \text{where} \quad \mathbf{S}\equiv\mathbf{YD}^{-1}\mathbf{Z}=\operatorname{diag}(S_1,\ldots,S_N).\end{aligned}$$

Without loss of generality, we assume that the states of each block are ordered with any transient states at the end, so that the ergodic probability vectors can be written as

$$u_I=\begin{pmatrix}\pi_I\\0\end{pmatrix} \qquad \text{where} \quad \pi_I>0, \quad \mathbf{A}_I u_I=0.$$

If $\mathbf{B}(\epsilon)$ is partitioned consistently with $\mathbf{A}(\epsilon)$, then from the no-coupling assumption, the (I, J) block must have the form

$$\mathbf{B}_{I,J}(\epsilon)=\begin{bmatrix}\mathbf{B}_{I,J}^{(R)}(\epsilon)\\0\end{bmatrix}, \qquad \mathbf{B}_{I,J}^{(R)}\geq 0 \qquad \text{for} \quad I\neq J.$$

[7] We drop the superscript $^{(0)}$ in this proof to simplify the notation.

The (I, J) elements of $\mathbf{A}^{(1)}(\epsilon)$ and $\mathbf{C}_m(\epsilon)$ can now be expressed as

$$\epsilon a^{(1)}_{I,J}(\epsilon) = 1^{\mathrm{T}}\mathbf{B}_{I,J}(\epsilon)u_J,$$

and

$$[\epsilon\mathbf{C}_m(\epsilon)]_{I,J} = \sum_{K_1,K_2,\ldots,K_m} 1^{\mathrm{T}}\mathbf{B}_{I,K_m}(\epsilon)S_{K_m}\cdots S_{K_1}B_{K_1,J}(\epsilon)u_J.$$

There must therefore exist a sequence of aggregate states $(K_1, K_2, \ldots, K_m)$ such that

$$[\epsilon\mathbf{C}_m(\epsilon)]_{I,J} = O(\|\mathbf{B}_{I,K_m}(\epsilon)\|\cdots\|\mathbf{B}_{K_1,J}(\epsilon)\|). \tag{A1}$$

From the structure of $\mathbf{B}_{I,J}(\epsilon)$ shown above, and the positivity of π_J, it follows that for an $I \neq J$,

$$\|\mathbf{B}_{I,J}(\epsilon)\| = O(\|1^T\mathbf{B}^{(R)}_{I,J}(\epsilon)\pi_J\|) = O(a^{(1)}_{I,J}(\epsilon)).$$

In (A1) the path $(J, K_1, K_2, \ldots, K_m, I)$ may have cycles. A new path $(J, K'_1, K'_2, \ldots, K'_{m'}, I)$, $m' \leq m$ can be constructed by removing the cycles in the original path. Since the number of states is bounded by the dimension of $\mathbf{A}(\epsilon)$, $m' \leq \dim(\mathbf{A}(\epsilon))$. Using this new path, the entry of $\mathbf{C}_m(\epsilon)$ in (A1) can be bounded:

$$[\epsilon\mathbf{C}_m(\epsilon)]_{I,J} = \epsilon^{m-m'}O(\|\mathbf{B}_{I,K'_{m'}}(\epsilon)\|\cdots\|\mathbf{B}_{K'_1,J}(\epsilon)\|),$$

from which it follows that

$$[\epsilon\mathbf{C}_m(\epsilon)]_{I,J} = \epsilon^{m-m'}O(a^{(1)}_{I,K'_{m'}}(\epsilon)\cdots a^{(1)}_{K'_1,J}(\epsilon)).$$

Introducing this bound on the (I, J) elements of $\mathbf{C}_1(\epsilon)$, $\mathbf{C}_2(\epsilon)$, ... allows bounding the infinite series for $w^{(1)}_{I,J}(\epsilon)$ as

$$w^{(1)}_{I,J}(\epsilon) = \epsilon O(a^{(1)}_{I,K''_{m''}}(\epsilon)\cdots a^{(1)}_{K''_1,J}(\epsilon)),$$

where the path $(J, K''_1, \ldots, K''_{m''}, I)$ corresponds to the path used in the bound for $C_m(\epsilon)$, for the m that results in the term of lowest order. □

A2. Proof of Lemma 3. First, $\mathbf{G}^{(k+1)}(\epsilon)$ has well-defined time-scale behavior by Lemma 1. The proof that $\mathbf{W}^{(k+1)}(\epsilon)$ is weak with respect to $\mathbf{A}^{(k+1)}(\epsilon)$ follows the proof of Lemma 2 with the following exception. We first write $\mathbf{G}^{(k)}(\epsilon)$ as

$$\mathbf{G}^{(k)}(\epsilon) = \mathbf{A}^{(k)} + (\mathbf{B}^{(k)}(\epsilon) + \mathbf{W}^{(k)}(\epsilon)) \equiv \mathbf{A}^{(k)} + \overline{\mathbf{B}}^{(k)}(\epsilon).$$

Using the nonnegativity of the off-diagonal blocks of $\mathbf{B}^{(k)}(\epsilon)$,

$$w^{(k)}_{J,I}(\epsilon) = \epsilon O(\|\mathbf{B}_{J,S_n}(\epsilon)\|\cdots\|\mathbf{B}_{S_1,I}(\epsilon)\|), \qquad I \neq j$$

for some path $(I, S_1, \ldots, S_n, J)$, $S_1 \neq I$, $S_i \neq S_{i+1}$, $S_n \neq J$; therefore

$$\|\overline{\mathbf{B}}_{J,I}(\epsilon)\| = O(\|\mathbf{B}_{J,I}(\epsilon)\|) + \epsilon O(\|\mathbf{B}_{J,S_n}(\epsilon)\|\cdots\|\mathbf{B}_{S_1,I}(\epsilon)\|).$$

This expression can be substituted into (A1) where $\|\mathbf{B}_{J,I}(\epsilon)\|$ appears. Equation (A1) is therefore valid for some new path $(K_1, \ldots, K_m)$, and the remainder of the proof follows as in Lemma 2. □

ACKNOWLEDGMENTS. The authors wish to acknowledge the contributions of Pamela Coxson during her stay as a visiting scientist at the Laboratory for Information and Decision Systems at M.I.T. Also, the many helpful suggestions of the reviewers have contributed to a much improved version of the paper.

REFERENCES

1. ANDO, A., AND FISHER, F. M. Near-decomposability, partition and aggregation. In *Essays on the Structure of Social Science Models*, A. Ando, and F. H. Fisher, Eds. MIT Press, Cambridge, Mass., 1963, pp. 92–106.
2. BAHL, L. R., JELINEK, F., AND MERCER, R. L. A maximum likelihood approach to continuous speech recognition. *IEEE Trans. Pattern Anal. Mach. Intell. 5*, 2 (1983), 179–190.
3. BOBBIO, A., AND TRIVEDI, K. S. An aggregation technique for the analysis of stiff Markov chains. *IEEE Trans. Comput.*, to appear.
4. CAO, W.-L., AND STEWART, W. J. Iterative aggregation/disaggregation techniques for nearly uncoupled Markov chains. *J. ACM 32*, 3 (July 1985), 702–719.
5. CODERCH, M., WILLSKY, A. S., SASTRY, S. S., AND CASTANON, D. A. Hierarchical aggregation of singularly perturbed finite state Markov processes. *Stochastics 8* (1983), 259–289.
6. CODERCH, M., WILLSKY, A. S., SASTRY, S. S., AND CASTANON, D. A. Hierarchical aggregation of linear systems with multiple time scales. *IEEE Trans. Autom. Contr. AC-28*, 11 (1983), 1017–1030.
7. COURTOIS, P. J. *Decomposability: Queuing and Computer System Applications.* Academic Press, Orlando, Fla., 1977.
8. COURTOIS, P. J., AND SEMAL, P. Error bounds for the analysis by decomposition of non-negative matrices. In *Mathematical Computer Performance and Reliability*, G. Iazeolla, P. J. Courtois, and A. Hordijk, Eds. North-Holland, Amsterdam, 1984, pp. 209–224.
9. DELEBECQUE, F. A reduction process for perturbed Markov chains. *SIAM J. Appl. Math. 43*, 2 (1983), pp. 325–330.
10. DELEBECQUE, F., AND QUADRAT, J.-P. Optimal Control of Markov chains admitting strong and weak interactions. *Automatica 17* (1981), 281–296.
11. DOERSCHUK, P. C. A Markov chain approach to electrocardiogram modeling and analysis. Ph.D. dissertation. Massachusetts Institute of Technology, Cambridge, Mass., 1985.
12. KATO, T. *Perturbation Theory for Linear Operators.* Springer-Verlag, Berlin, 1966.
13. KOKOTOVIC, P. V. Singular perturbations and iterative separation of timescales. *Automatica 16* (1980), 23–24.

13a. KOROLYUK, V. S., AND TURBIN, A. F. On the asymptotic behavior of the occupancy time of a semi-Markov process in a reducible subset of states. *Theor. Probl. Math. Stat. 2* (1974), 133–143.

14. KOROLYUK, V. S., AND TURBIN, A. F. Limit theorems for Markov random evolution in the scheme of asymptotic state lumping. In Lecture Notes in Mathematics, vol. 1021. Springer-Verlag, Berlin, 1983.
15. LOU, X.-C., ROHLICEK, J. R., COXSON, P. G., VERGHESE, G. C., AND WILLSKY, A. S. Time scale decomposition: The role of scaling in linear systems and transient states in finite-state Markov processes. In *Proceedings of the 1985 American Control Conference* (June). American Automatic Control Council, Green Valley, Ariz., 1985, pp. 1408–1413.
16. LOU, X.-C., VERGHESE, G., WILLSKY, A. S., AND VIDYASAGAR, M. An algebraic approach to analysis and control of timescales. In *Proceedings of the 1984 American Control Conference* (June). American Automatic Control Council, Green Valley, Ariz., 1984, pp. 1365–1372.
17. ROHLICEK, J. R. Aggregation and time scale analysis of perturbed Markov systems. Ph.D. dissertation. Massachusetts Institute of Technology, Cambridge, Mass., 1987.
18. ROHLICEK, J. R., AND WILLSKY, A. S. The reduction of perturbed Markov generators: An algorithm exposing the role of transient states. M.I.T. Report LIDS-P-1492. Massachusetts Institute of Technology, Cambridge, Mass., Sept. 1985.
19. SILJAK, D. D. *Large-Scale Dynamic Systems: Stability and Structure.* North-Holland, Amsterdam, 1978.
20. SIMON, H., AND ANDO, A. Aggregation of variables in dynamic systems. *Econometrica 29* (1963), 111–139.
21. WALKER, B. K. A semi-Markov approach to quantifying fault-tolerant system performance. Ph.D. dissertation. Massachusetts Institute of Technology, Cambridge, Mass., 1980.

RECEIVED SEPTEMBER 1985; REVISED SEPTEMBER 1986, AUGUST 1987, NOVEMBER 1987; ACCEPTED NOVEMBER 1987