MICROCOPY RESOLUTION TEST CHART

NATIONAL BUREAU OF STANDARDS-1963-A

REPORT DOCUMENTATION PAGE

AD-A160 185	**1b. RESTRICTIVE MARKINGS**
~~ULE~~	**3. DISTRIBUTION/AVAILABILITY OF REPORT** Approved for public release; distribution unlimited.

4. PERFORMING ORGANIZATION REPORT NUMBER(S) N/A	**5. MONITORING ORGANIZATION REPORT NUMBER(S)** AFOSR-TR- 85-0833

6a. NAME OF PERFORMING ORGANIZATION Massachusets Institute of Tech.	**6b. OFFICE SYMBOL** *(If applicable)*	**7a. NAME OF MONITORING ORGANIZATION** AFOSR
6c. ADDRESS *(City, State and ZIP Code)* Laboratory for Info. and Decision Sys. Cambridge, MA 02139		**7b. ADDRESS** *(City, State and ZIP Code)* Bldg. 410 Bolling AFB, D.C. 20332-6448

8a. NAME OF FUNDING/SPONSORING ORGANIZATION AFOSR	**8b. OFFICE SYMBOL** *(If applicable)* NM	**9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER** AFOSR-82-0135

8c. ADDRESS *(City, State and ZIP Code)*
Bldg. 410
Bolling AFB, D.C. 20332-6448

10. SOURCE OF FUNDING NOS.

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT NO.
61102F	2304	A1	

11. TITLE *(Include Security Classification)*
Time Scale Decomposition: The role of scaling in linear systems & transient states in finite-space Markov processes

12. PERSONAL AUTHOR(S)
X.-C Lou, R. Rohlicek, P. G. Coxson, G. C. Verghese, A. S. Willsky

13a. TYPE OF REPORT Interim	**13b. TIME COVERED** FROM _____ TO _____	**14. DATE OF REPORT** *(Yr., Mo., Day)* 21 March 1985	**15. PAGE COUNT** 7

16. SUPPLEMENTARY NOTATION

17.		COSATI CODES	**18. SUBJECT TERMS** *(Continue on reverse if necessary and identify by block number)*
FIELD	GROUP	SUB. GR.	decomposition & Aggregation of large scale
	XXXXXXXXXX		linear systems

19. ABSTRACT *(Continue on reverse if necessary and identify by block number)*
In this paper we reports on some of our recent work on time scale decomposition and aggregation of large-scale linear systems containing weak couplings and finite-state Markov processes (FSMP's) containing rare transitions. Our work builds on that of Coderch, et. al.. The focus of the work is on the asymptotic approximation of the linear system.

DTIC
ELECTE
OCT 15 1985

DTIC FILE COPY

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. ☒ DTIC USERS ☐	**21. ABSTRACT SECURITY CLASSIFICATION** Unclassified

22a. NAME OF RESPONSIBLE INDIVIDUAL Dr. Marc Q. Jacobs	**22b. TELEPHONE NUMBER** *(Include Area Code)* (202)767-5027	**22c. OFFICE SYMBOL** NM

Title

TIME SCALE DECOMPOSITION: THE ROLE OF SCALING IN LINEAR SYSTEMS AND TRANSIENT STATES IN FINITE-STATE MARKOV PROCESSES

Principal Investigator

X.-C Lou

Laboratory for Information and Decision Systems
Massachusetts Institute of Technology
Cambridge, Mass, 02139

Time Scale Decomposition: The Role of Scaling in Linear Systems and
Transient States in Finite-State Markov Processes

X.-C. Lou*, J.R. Rohlicek*, P.G. Coxson[+], G.C. Verghese*, A.S. Willsky*

Laboratory for Information and Decision Systems
Massachusetts Institute of Technology
77 Mass. Ave., Cambridge, MA. 02139

I. Introduction

In this paper we report on some of our recent work on time scale decomposition and aggregation of large-scale linear systems containing weak couplings and finite-state Markov processes (FSMP's) containing rare transitions. Our work builds on that of Coderch, et. al. [1,2]. The focus of the work in [1] is on the asymptotic approximation of the linear system

$$\dot{x}(t) = A(\epsilon)x(t). \qquad (1.1)$$

To set the stage for our work, consider a second system

$$\dot{z}(t) = B(\epsilon)z(t). \qquad (1.2)$$

We say that (1.2) is asymptotically equivalent to (1.1) if

$$\lim_{\epsilon \to 0} \sup_{t \geq 0} \| e^{A(\epsilon)t} - e^{B(\epsilon)t} \| = 0 \qquad (1.3)$$

The focus of [1] is on the construction of a system as in (1.2) where

$$B(\epsilon) = T \, diag(A_0, \epsilon A_1 \ldots , \ldots \epsilon^r A_r) T^{-1} \qquad (1.4)$$

so that A_0 captures the order 1 time scale, ϵA_1, the $O(1/\epsilon)$ scale, etc. What is accomplished in [1] is the development of a procedure which determines if such a complete time scale decomposition is possible and, if so, computes the A_i. In our opinion, this is a very important result, but [1] left much to do, for example in "peeling back" the mathematics of [1] to allow us to obtain a far clearer and deeper understanding of time scale decompositions.

In [3] we presented some of our first results on an algebraic approach to the problem of time scale decomposition of (1.1) based on viewing $A(\epsilon)$ as a matrix over the ring W of functions of ϵ that are analytic at $\epsilon=0$. This work allowed us to relate the general result of [1] to earlier work as in [4] on special cases for which the form of the time scale decomposition is intuitively clear.

Furthermore, by making clear the role of invariant factors in time scale decompositions, we were able to formulate and solve a "time scale control" problem. This algebraic approach also allows us to consider and solve several other important problems which we report on in Section II. In particular, we have been able to obtain a complete characterization of the relationship between the eigenvalues of $A(\epsilon)$, its invariant factors, and a condition introduced in [1] in a complicated way but to which we can now give a far clearer interpretation. This characterization is then used to develop (a) a procedure for computing the invariant factors (and thus determining the number of modes at each time scale) from the gcd's of principal minors of $A(\epsilon)$, and (b) a method for scaling the system (1.1) when it does not have a uniform time scale approximation to obtain a system that does.

The work in [2] applies the method of [1] to FSMP's with rare transitions (as parametrized by ϵ). In such a case $x(t)$ in (1.1) is a vector of state probabilities and $A(\epsilon)$ is a stochastic matrix (offdiagonal elements ≥ 0 and column sums $= 0$). What is done in [2] is to interpret the results of [1] as defining a succession of stochastically discontinuous processes representing the evolution of x at successively slower time scales ($t, t/\epsilon,\ldots$) so that at each stage transitions that occur at a faster rate appear to occur instantaneously. This led naturally to an aggregation at each stage that had the effect of removing these discontinuous transitions. While this is an extremely important result, it does have some drawbacks. In particular, the direct application of the results of [1] involves a procedure whose probabilistic interpretation is, at best, obscure. Furthermore, the computational feasibility of this approach is dubious. This is in marked contrast to other work in this area, such as [8], in which an intuitively appealing approach to aggregation is described for the special class of models devoid of transient states at any time scale: to obtain an aggregate description of the FSMP at the slower time scale, we lump together the states in each separate ergodic class at the faster time scale (i.e. with $\epsilon = 0$) and compute an average transition rate between these ergodic classes to be used to describe evolution at the slower time scale.

While the intuition provided in this method is desireable, the limitations of methods as in [8] are both the absence of proofs of uniform asymptotic equivalence of the approximations produced and their inability to handle transient states. In our recent work, described in Section III, we have been working to bridge the gap

between the methods of [1] and [8]. In particular, we have developed an understanding of the role of transient states and, in particular, of what we call "splitting transient states." This understanding has allowed us to formulate an approach to aggregation and asymptotic approximation that in essence follows the procedure of [1] but does so by modifying the FSMP at each stage so that the computations involved are essentially those of the extension of the methods of [8] to allow for transient (but non-splitting) states. The procedure we describe and illustrate in Section III is computationally feasible for the analysis of very complex systems.

II. Algebraic Methods for Time Scale Analysis

The perturbed matrix $A(\varepsilon)$ from (1.1) can be expressed in the Smith-decomposed form $A(\varepsilon) = P(\varepsilon)D(\varepsilon)Q(\varepsilon)$, where $P(\varepsilon)$ and $Q(\varepsilon)$ are unimodular (i.e. $|P(0)|\neq0$ and $|Q(0)|\neq0$) and

$$D(\varepsilon) = \text{block diagonal } [\varepsilon^{j_1}I_{n_1},\ldots,\varepsilon^{j_m}I_{n_m}],$$

$$0 \leq j_1 < \ldots < j_m .$$

The diagonal elements of $D(\varepsilon)$ are the invariant factors of $A(\varepsilon)$. Since $P(\varepsilon)$ and its inverse are well defined in a neighborhood of $\varepsilon = 0$, we can make the change of variables $x(t) = P(\varepsilon)z(t)$ in (1.1), which results in a description that we call explicit form:

$$\dot{z}(t) = D(\varepsilon)Q(\varepsilon)P(\varepsilon)z(t) = D(\varepsilon)\bar{A}(\varepsilon)z(t),$$

$$\bar{A}(\varepsilon) \text{ unimodular.} \qquad (2.1)$$

What we term a reduced explicit form for (1.1) is then obtained by replacing the unimodular matrix $\bar{A}(\varepsilon)=Q(\varepsilon)P(\varepsilon)$ by the constant matrix $\bar{A}(0)=Q(0)P(0)$, which we shall from now on simply denote by A, to form the system

$$\frac{d}{dt}\begin{bmatrix} y_1(t) \\ \vdots \\ \dot{y}_m(t) \end{bmatrix} = \begin{bmatrix} \varepsilon^{j_1}I_{n_1} & & 0 \\ & \ddots & \\ 0 & & \varepsilon^{j_m}I_{n_m} \end{bmatrix}\begin{bmatrix} \bar{A}_{11} \cdots \bar{A}_{1m} \\ \vdots \quad \ddots \quad \vdots \\ \bar{A}_{m1} \cdots \bar{A}_{mm} \end{bmatrix}\begin{bmatrix} y_1(t) \\ \vdots \\ y_m(t) \end{bmatrix} \qquad (2.2)$$

The partitioning in (2.2) is that induced by the block sizes in $D(\varepsilon)$.

A key role in our theorems is played by a set of matrices derived from $\bar{A}$ in (2.2). To obtain this set, we first write

$$A^{(1)} = \bar{A} \text{ and } \bar{A}_{11} = \bar{A}_{11} . \qquad (2.3a)$$

Now let $A^{(2)}$ denote the Schur complement of $\bar{A}_{11}$ in $A^{(1)}$ and A_{22} denote the $n_2 \times n_2$ leading principal submatrix of this Schur complement, so that

$$\bar{A}_{22} = \bar{A}_{22} - \bar{A}_{21}(\bar{A}_{11})^{-1}\bar{A}_{12} . \qquad (2.3b)$$

Thus $\bar{A}_{22}$ is defined iff $\bar{A}_{11}$ is nonsingular. Continuing this, we define $\bar{A}_{ii}$, $i = 1$ to m, as the $n_i \times n_i$ leading principal submatrix of the Schur complement, $A^{(i)}$, of $\bar{A}_{i-1,i-1}$ in $A^{(i-1)}$; again, it is defined iff $\bar{A}_{i-1,i-1}$ is nonsingular.

A. Connections to Results of [1,5]

It was shown in [5] that the necessary and sufficient condition for (1.1) to have a complete time-scale decomposition is that $A(\varepsilon)$ satisfy a so-called "multiple semi-stability" (MSST) condi-

tion. Another condition on $A(\varepsilon)$ that will be of considerable interest to us here, though it plays only a subsidiary role in [5] and [1], is the multiple semi-simple null structure" (MSSNS) condition.

The proofs of all the results that follow are in (or may be readily deduced from) [5,6]

Theorem 1: The following are equivalent:
(a) $A(\varepsilon)$ in (1.1) satisfies the MSSNS condition of [1].
(b) The orders of the eigenvalues of $A(\varepsilon)$ are identical to the orders of its invariant factors.
(c) $D(\varepsilon)\bar{A}$ in (2.2) satisfies MSSNS.
(d) $\bar{A}_{ii}$, $i = 1$ to m, are defined and nonsingular.

Although our focus here is on MSSNS, we note that the statements in Theorem 1 have analogs valid for MSST. If we replace MSSNS by MSST and replace nonsingular by Hurwitz in 1d, then the following implications hold: (a) $<=>$ (c)$<=>$ (d).

Theorem 2: If $A(\varepsilon)$ satisfied MSSNS, the eigenvalues of $A(\varepsilon)$ and $D(\varepsilon)\bar{A}$ are clustered in m groups, with those in the k-th group lying within $O(\varepsilon^{j_k+1})$ of the eigenvalues of $\varepsilon^{j_k} \bar{A}_{kk}$.

It is evident from Theorem 1(b) and Theorem 2 that the MSSNS condition is of value in frequency-scale approximation, a topic traditionally studied in the context of root loci, see for example [7]. The frequency scales of (1.1) are equal to its invariant factors precisely under the MSSNS condition, which is directly checked via Theorem 1(d). The eigenvalues of $A(\varepsilon)$ at the different frequency scales are then approximated, according to Theorem 2, via the $\bar{A}_{ii}$.

In general, the invariant factors of $A(\varepsilon)$ are obtained from the gcd $\sigma(i)$ of all ixi minors for each i, while the eigenvalues of $A(\varepsilon)$ are determined solely by its principal minors. By (b) of Theorem 1, it must be true that, when $A(\varepsilon)$ satisfies MSSNS, the invariant factors are also determined by the principal minors alone -- in fact, by the gcd's of the ixi principal minors for each i, as described in the next theorem. For the statement of the theorem, denote the orders of the gcd's of the ixi principal minors by $p(i)$, $i = 1$ to n, and define $p(0) = 0$. Now let $b(i)$ be the slopes of the line segments forming the lower (boundary of the convex) hull in the graph of $p(i)$ versus i.

Theorem 3: (a) If $A(\varepsilon)$ satisfies MSSNS, then the orders of its invariant factors are equal to $b(i)$, $i = 1$ to n.
(b) If $A(\varepsilon)$ has invariant factor orders equal to the $b(i)$ of the explicit form, then $A(\varepsilon)$ satisfies MSSNS.

Such "Newton polygon" constructions are to be expected in the context of the present problem, cf. [7]. However, we have not encountered a statement as simple as that of the above theorem in the literature. The result will be useful in the discussion of scaling (II.B). Another use is illustrated in example 1.

Example 1: Suppose n=4, and suppose $p(1)=3$, $p(2)=2$, $p(3)=2$ and $p(4)=3$. Then Figure 2.1 shows that the $b(i)$ are 2/3, 2/3, 2/3 and 1. Since invariant factors of $A(\varepsilon)$ cannot be of fractional order, the only possible conclusion is that $A(\varepsilon)$ does not satisfy MSSNS.

B. Scaling

A matrix $A(\epsilon)$ that does not satisfy MSSNS -- and that therefore has eigenvalue orders different from invariant factor orders, see Theorem 1(b) -- can often be transformed by non-unimodular similarity transformations to a matrix that does satisfy MSSNS. An important reason for trying to induce MSSNS like this is to enable the application of decomposition results such as Theorem 2 to estimating the natural frequencies of (1.1).

We restrict ourselves to ϵ-dependent scaling of variables, i.e. to non-unimodular diagonal similarity transformations. This enables us to build directly on Theorem 3(b), because such transformations leave both eigenvalues and principal minors unchanged, while they still permit some modification of invariant factors. Before stating the procedure in a general way, we consider an example.

Example 2: $A(\epsilon) = \begin{bmatrix} -\epsilon & 1 \\ 0 & -\epsilon \end{bmatrix} = \begin{bmatrix} 1-\epsilon & 1 \\ -\epsilon & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & \epsilon^2 \end{bmatrix}\begin{bmatrix} -\epsilon & 1+\epsilon \\ -\epsilon & 1 \end{bmatrix}$

The explicit form and reduced explicit form of $A(\epsilon)$ are then

$$A_e(\epsilon) = \begin{bmatrix} 1 & 0 \\ 0 & \epsilon^2 \end{bmatrix}\begin{bmatrix} -2\epsilon & 1 \\ -1 & 0 \end{bmatrix}, \quad A_r(\epsilon) = \begin{bmatrix} 1 & 0 \\ 0 & \epsilon^2 \end{bmatrix}\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

Since $\bar{\lambda}_{11} = 0$, it is evident from Theorem 1(d) that $A(\epsilon)$ does not satisfy MSSNS. If we let $S(\epsilon) = $ diagonal $[\epsilon, 1]$, and transform the explicit form to $S(\epsilon)D(\epsilon)A(\epsilon)S^{-1}(\epsilon)$ the resulting matrix satisfies both MSSNS and MSST. Below, we outline a systematic approach for generating appropriate scaling matrices $S(\epsilon)$.

The first step of our general scaling procedure, again driven by Theorem 3(b), is to transform $A(\epsilon)$ to its explicit form, $A_e(\epsilon) = D(\epsilon)A(\epsilon)$, see (2.1). The second step then involves marking what we term a skeleton in the explicit form: precisely one element from each row and column of $A_e(\epsilon)$, with the additional constraint that no other element in a row have lower order (in ϵ) than the skeleton element. Since $\bar{\lambda}(0)$ is nonsingular, the skeleton element in the i-th row has order equal to the order of the i-th entry of $D(\epsilon)$. (The choice of skeleton may not be unique, but see Remark 1 below.)

Now identify the skeleton above with the nxn permutation matrix that has 1's at the locations of the skeleton elements and 0's elsewhere. Recall that any permutation can be uniquely expressed as a product of disjoint cycles. It follows that, perhaps after some re-ordering of the variables associated with our system, the elements of the skeleton can be brought to the positions occupied by 1's in a block diagonal canonical circulant matrix, whose diagonal blocks take the form:

$$\begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 \\ & & \dots & & \\ 0 & 0 & & \dots & 0 & 1 \\ 1 & 0 & & \dots & & 0 \end{bmatrix} \text{ (or simply 1 for a scalar block) .}$$

We shall restrict ourselves here to the case of only a single block; the extension of the following results to the multiple block case is described in [6]. Note that the required re-ordering of variables corresponds to using a permutation matrix for similarity transformation of the explicit form,

and the result is still in explicit form. This re-ordering of variables is the third step of our procedure.

The following description now takes $A_e(\epsilon)$ to have a skeleton corresponding to a single canonical circulant block. We denote the order of the skeleton element in the i-th row by $a(i)$ -- note that these are just the orders of the invariant factors. We make three further assumptions. The first of these is that the orders of the diagonal entries in the matrix are in nondecreasing order; this assumption is also lifted in [6]. We have, however, been unable to relax the remaining two assumptions:

Assumption 1: $b(i) \geq a(j)$ for $i,j = 1$ to $n-1$. To visualize what the assumption states, plot both $p(i)$ and $q(i)$ versus i, as in Figure 2.2. Then the slopes of the (solid) line segments making up the lower hull of the $p(i)$ curve are assumed to be not less than the slopes of those making up the $q(i)$ curve (the dotted lines), except at the last step (from n-1 to n).

Assumption 2: In any principal submatrix of $A_e(\epsilon)$, the order of any term formed by taking the product of precisely one element from every row and column of the submatrix is not less than the order of the corresponding principal minor.

With all the above assumptions, the following scaling can be shown to transform the matrix to one that satisfies MSSNS:

$$S(\epsilon) = \text{diagonal } [\epsilon^{s_1}, \epsilon^{s_2}, \dots, \epsilon^{s_{n-1}}, 1], \quad (2.5a)$$

where

$$s_i = s_{i+1} + b(i) - a(i), \quad s_n = 0. \quad (2.5b)$$

The arguments, even for the special case we are considering, are rather intricate, and are presented in [6]. They show that, under Assumptions 1 and 2, the above scaling produces invariant factor orders equal to $b(i)$.

The following example illustrates the process.

Example 3: Suppose

$$A(\epsilon) = \begin{bmatrix} \epsilon^3 & \epsilon^4 & \epsilon^5 & 1 \\ \epsilon & \epsilon^3 & \epsilon^3 & \epsilon \\ \epsilon^3 & \epsilon & \epsilon^2 & \epsilon^7 \\ \epsilon^6 & \epsilon^8 & \epsilon^6 & \epsilon^7 \end{bmatrix}, \text{ where the circled elements constitute a skeleton.}$$

It is easy to see that $A(\epsilon)$ is already in explicit form. Similarity transformation by a permutation matrix, corresponding to a re-ordering of variables, brings it to the form

$$A_e(\epsilon) = \begin{bmatrix} \epsilon^2 & \epsilon & \epsilon^3 & \epsilon^7 \\ \epsilon^3 & \epsilon^3 & \epsilon & \epsilon \\ \epsilon^5 & \epsilon^4 & \epsilon^3 & 1 \\ \epsilon^6 & \epsilon^8 & \epsilon^6 & \epsilon^7 \end{bmatrix}.$$

It is evident that $a(1)=1$, $a(2)=1$, $a(3)=0$, $a(4)=6$, while some computation shows that $b(i)=2$ for $i=1$ to 4. This information can be visualized via Figure 2.3. From (2.5b) we have $s_3=2$, $s_2=3$ and $s_1=4$. Similarity transforming $A_e(\epsilon)$ by $S(\epsilon)$ defined in (2.5a), we get the matrix

$$\begin{bmatrix} \epsilon^2 & \epsilon^2 & \epsilon^5 & \epsilon^{11} \\ \epsilon^2 & \epsilon^3 & \epsilon^2 & \epsilon^4 \\ \epsilon^3 & \epsilon^3 & \epsilon^3 & \epsilon^2 \\ \epsilon^2 & \epsilon^5 & \epsilon^4 & \epsilon^7 \end{bmatrix}.$$

It is easy to check that this matrix satisfies MSSNS.

Remark 1: While the validity of Assumptions 1 and 2 is independent of which particular skeleton is chosen (when the choice is not unique), it may be that one choice leads to a simpler procedure than another choice.

Remark 2: Our scaling procedure involves (in $S^{-1}(\epsilon)$) and may produce matrices whose entries are outside the ring W. However, by changing the time scale (i.e. redefining ϵ), one can always bring the result of the scaling to a form that our theorems apply to.

Remark 3: See [6] for application of this scaling procedure to cases treated in [10].

Remark 4: Other scalings are possible, even when our assumptions are violated, and further work in this direction will be worthwhile.

III. Aggregation and Time-scale Decomposition of Finite-state Markov Processes

In this section we describe our recent work on aggregation of finite-state Markov processes. In order to provide some perspective on the key ideas underlying our approach, we begin by reviewing the case of "nearly completely decomposable systems" and contrasting what the approaches of [1] and [8] have to say in this case.

A. Nearly Completely Decomposable Systems

Consider a FSMP whose probabilistic evolution is described by (1.1) with $A(\epsilon) = A + \epsilon B$, where A describes a FSMP with several ergodic classes and no transient states. The precise structure of the transitions between these classes for $\epsilon > 0$ is specified by the matrix B. Such a process can be shown to have only two fundamental time scales (t and t/ϵ) due to the combination of the irreducible structure of A and the restriction to linear perturbations of the form ϵB.

If we follow the approach of [1], [2], the "slow" dynamics of the FSMP are captured by the generator $B(\epsilon) = P(\epsilon)A(\epsilon)P(\epsilon)$ where $P(\epsilon)$ is the oblique projection onto the eigenspace of $A(\epsilon)$ of all the $o(1)$ eigenvalues along the space of $O(1)$ eigenvalues. More precisely, the fast dynamics, represented by $A(0)$ (which capture the behavior of (1.1) over intervals of the form $[0,T_1]$, $T_1 < \infty$) and the slow dynamics $\epsilon D(0)$, where

$$D(0) = \lim_{\epsilon \to 0} \frac{P(\epsilon)A(\epsilon)P(\epsilon)}{\epsilon} \qquad (3.1)$$

(which capture (1.1) over intervals of the form $(\frac{T_2}{\epsilon}, \infty)$, $T_2 > 0$) together provide a uniform approximation to the original FSMP.

An issue here is whether it is really necessary to calculate $P(\epsilon)$. In particular, interpretation of the reduction performed by Courtois [8] on nearly completely decomposable chains is that $C(\epsilon) = P(0)A(\epsilon)P(0) = P(0)BP(0)$ generates a Markov process whose probability transition function is an approximation of the original function with bounded error on an interval $t \geq T(\epsilon)$ for some $T(\epsilon)$ that grows without bound as $\epsilon \to 0$. What

we conjecture is a good bit stronger. Specifically we claim that $A(0)$ and $P(0)BP(0)$ together provide a uniform approximation to the original FSMP that is that $A(\epsilon) = A + \epsilon B$ can be replaced by $\hat{A}(\epsilon) = A + \epsilon P(0)BP(0)$. Note that in this approach all that is required is $P(0)$, which is nothing more than the ergodic projection matrix associated with $A(0) = A$. An example illustrating this is shown in Figure 3.1. The original process is shown in (a), for this process.

$$P(0) = \begin{bmatrix} 1/2 & 1/2 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 2/3 & 2/3 \\ 0 & 0 & 1/3 & 1/3 \end{bmatrix} \qquad (3.2)$$

The process corresponding to $\hat{A}(\epsilon)$ is shown in (b). While this process may appear to be more complex, what we have in fact done is to maintain the equilibrium of the fast dynamics after rare transitions. Further, since A and $P(0)BP(0)$ commute

$$e^{\hat{A}(\epsilon)t} = e^{At} e^{P(0)BP(0)\epsilon t} \qquad (3.3)$$

As in [2], we can write $P(0)$ as

$$P(0) = \begin{bmatrix} 1/2 & 0 \\ 1/2 & 0 \\ 0 & 2/3 \\ 0 & 1/3 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \overset{\Delta}{=} UV \qquad (3.4)$$

where the columns of U represent the two possible ergodic probability vectors of A, and V lumps the states in each ergodic class.

Combining (3.3) and (3.4) we have that

$$e^{\hat{A}(\epsilon)t} = e^{At} U e^{B'\epsilon t} V \qquad (3.5)$$

where

$$B' = VBU = \begin{bmatrix} -1/2 & 0 \\ 1/2 & 0 \end{bmatrix} \qquad (3.6)$$

Figure 3.1(c) provides the interpretation of (3.5): the matrix B' describes the evolution of the slow, aggregated process, while the matrix A specifies the faster evolution within either of the two aggregate classes.

This procedure can be extended in a straightforward fashion to systems exhibiting multiple time scales when there are no transient states at the first time scale. In such cases the generator has the form $A(\epsilon) = A + \epsilon B(\epsilon)$ where A generates no transient states. In this case, we again conjecture that $A(\epsilon)$ and $\hat{A}(\epsilon) = A + \epsilon P(0)B(\epsilon)P(0)$ are asymptotically equivalent. One can then proceed as in the example to aggregate $B(\epsilon)$; one may then repeat this procedure several times as in the procedures of [2] and [9].

A natural question that arises at this point concerns the role and effect of transient states. This topic is taken up in the next subsection.

B. Transient States

Though indecomposable structure of $A(0)$ is a sufficient condition for using the simplified procedure described above, a less restrictive condition is available. In particular, we can allow $A(0)$ to possess "non-splitting transient states," i.e. transient states that may have $O(1)$ transitions

into more than one A(0)-ergodic class but do not have direct transitions into other such classes with states of any higher order. If such splitting transient rates are present, then the FSMP may exhibit implicit time scales that can't be captured directly by our simplified procedure. As an example, consider the FSMP shown in Figure 3.2. Though this chain has only linear perturbation terms, the eigenvalues are 0, $O(1)$, and $O(\epsilon^2)$. The generator $P(0)A(\epsilon)P(0) = 0$, obviously does not capture the t/ϵ^2 time scale behavior. An intuitive explanation of this is that using this reduction process implicitly assumes that the "fast" components equilibrate between rare $(O(\epsilon)$ rate) transitions. In this example, the t/ϵ^2 behavior is associated with a sequence of two consecutive rare transitions (state 1 to 2 followed by 2 to 3). Beginning in state 2, there is an $O(1)$ probability of entering state 1 next and an $O(\epsilon)$ probability of entering 3. Effectively, this $O(\epsilon)$ probability is lost in the reduction procedure.

Transient states which do not exhibit such $O(\epsilon)$ probabilities of entering various recurrent classes do not cause this problem. Consider the related FSMP shown in Figure 3.3. Both states 2a and 2b are transient at $\epsilon=0$, but neither of them splits as state 2 does in the previous example. The $O(\epsilon^2)$ rate is explicit and $P(0)A(\epsilon)P(0)$ successfully captures the t/ϵ^2 behavior. This chain is derived from the first by "splitting" the transient state 2 into the nonsplitting transient states 2a and 2b, depending on the first recurrent class entered. If we imagine having an observation mechanism for this process that yields the value 1 or 3 if the FSMP is in state 1 or 3 respectively and the value 2 if the FSMP is in 2a or 2b, then the transition rates between these observation values are exactly those given in Figure 3.2. An approximation of the process shown in Figure 3.3 can then be used to construct an approximation of the original process.

C. The General Procedure

An interative procedure can be derived to construct a sequence of aggregate perturbed generators such as $B'(\epsilon)$ at each successive time scale. The steps involved in computing $B'(\epsilon)$ from $A(\epsilon)$ consist of (i) identifying the recurrent classes and transient states at $\epsilon=0$, (ii) calculating the invariant probabilities of the recurrent classes at $\epsilon=0$ (together (i) and (ii) determine $P(0)$), (iii) calculating the ϵ-dependence of the trapping probabilities starting in any state of the transient class (so that we may split any splitting transient class), and finally (iv) computing the aggregate rates $B'(\epsilon)$ from these quantities. From $A(0)$, $B'(0)$, etc., an approximation can be calculated which we conjecture is the same uniform approximation derived in [2].

This procedure can further be simplified by identifying modifications of a perturbed chain which preserve its time scale behavior. For example, it is conjectured that only the leading order term in ϵ of any transition rate affects asymptotic behavior. Also, if there is an indirect sequence of $O(1)$ rate transitions from one state to another, then any direct $O(\epsilon)$ rate between these states can be safely "pruned."

Though the reduction algorithm as outlined above has produced the same uniform approximation as the methods of [2] and [9] in all the examples we have considered, the proof that this is necessarily true has not yet been established. Explicitly, two fundamental conjectures form the basis of the result.

1) $A(\epsilon)=A(0)+\epsilon B(\epsilon)$: Markov generator with $A(0)$ irreducible

$F(\epsilon)=P(0)A(\epsilon)P(0) = P(0)B(\epsilon)P(0)$

conjecture

$$\lim_{\epsilon\downarrow 0} \sup_{\epsilon t \geq \delta > 0} \| \exp(A(\epsilon)t) - \exp(F(\epsilon)t) \| = 0 \qquad (3-7)$$

$$\lim_{\epsilon\downarrow 0} \sup_{t \geq 0} \| \exp(A(\epsilon)t) - \exp(A(0)t)\,\exp(B(\epsilon)t) \| = 0 \qquad (3-8)$$

2) Conjecture 1 is also true if $A(0)$ has no splitting transient states, that is under the following condition. Let T denote the set of transient states of $A(0)$ and let $R_1,\ldots,R_m$ denote its ergodic classes. Let $\rho(t)$ denote the sample path of the FSMP (with ϵ included). Then for all $x_0 \epsilon T$ and all $i=1,\ldots,m$

$$\Pr(\rho(t_1)\epsilon R_i \mid t_1 = \inf(t:\rho(t)\notin T), \rho(0)=x_0) = 0 \text{ or } O(1) \qquad (3-9)$$

Proving this second result allows us to consider arbitrary generators $A(\epsilon)$ by conceptually splitting the transient class T into m copies, $T_1,\ldots,T_m$ associated with the classes $R_1,\ldots,R_m$.

REFERENCES

1. M. Coderch, A.S. Willsky, S.S. Sastry, and D.A. Castanon, "Hierarchical Aggregation of Linear Systems with Multiple Time Scales," IEEE Trans. Aut. Control, Vol. AC-28, No. 11, Nov. 1983, pp. 1017-1030.

2. M. Coderch, A.S. Willsky, S.S. Sastry, and D. A. Castanon, "Hierarchical Aggregation of Singularly Perturbed Finite State Markov Processes," Stochastics, Vol. 8, 1983, pp. 259-289.

3. X.-C. Lou, G.C. Verghese, A.S. Willsky, and M. Vidyasagar, "An Algebraic Approach to Analysis and Control of Time-Scales," Proc. 1984 American Control Conf., June 1984, pp. 1368-1372.

4. P.V. Kokotovic, R.E. O'Malley, Jr., and P. Sannuti, "Singular Perturbations and Order Reduction in Control Theory--An Overview," Automatica, Vol. 12, 1976, pp. 123-132.

5. M. Coderch, "Multiple Time Scale Approach to Hierarchical Aggregation of Linear Systems and Finite State Markov Processes," Inst. Technol., Dept. Elec. Eng. Comput. Sci., Massachusetts Inst. Technol., Cambridge, August 1982.

6. X.-C. Lou, "An Algebraic Approach to the Analysis and Control of Time Scales," Ph.D. Dissertation, Dept. Elec. Eng. Comput. Sci., Massachusetts Inst. Technol., Cambridge, June 1985.

7. C.I. Byrnes and P.K. Stevens, "The McMillan and Newton Polygons of a Feedback Systems and the Construction of Root Loci," *Int. J. Control*, Vol. 35, No. 1, 1982, pp. 29-53.

8. P.J. Courtois, *Decomposability: Queueing and Computer System Applications*, Academic Press, New York, 1977.

9. F. Delebecque, "A Reduction Process for Perturbed Markov Chains," *SIAM J. Appl. Math.*, Vol. 43, No. 2, April 1983.

10. P. Sannuti, "Direct Singular Perturbation Analysis of High Gain and Cheap Control Problems," *Automatica*, Vol. 19, No. 1, 1983, pp. 41-51.

FIGURES
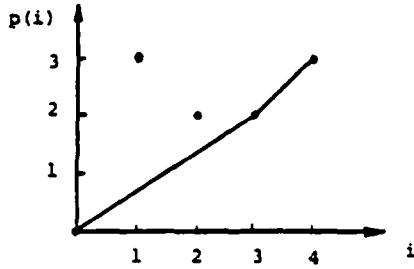
Figure 2.1
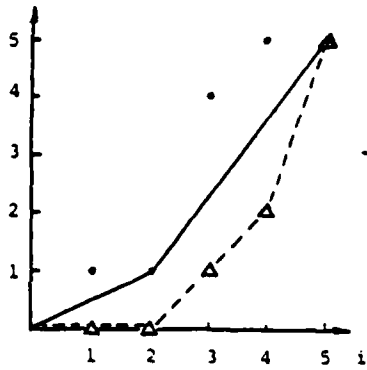
Figure 2.2

Figure 2.3

(a) $A(\epsilon) = A + \epsilon B$

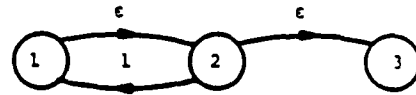(b) $\hat{A}(\epsilon) = A + \epsilon P(0) B P(0)$
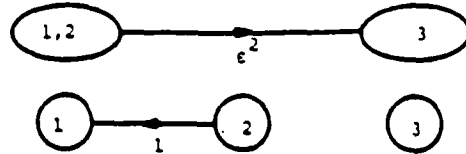
(c)
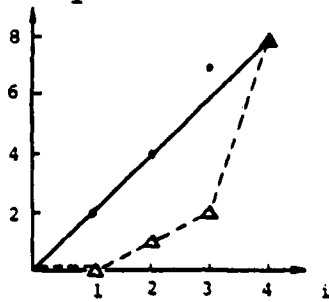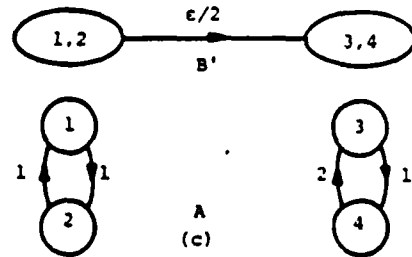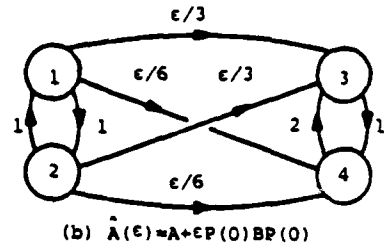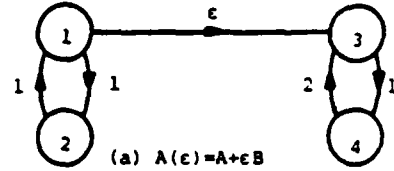
Figure 3.1

Figure 3.2
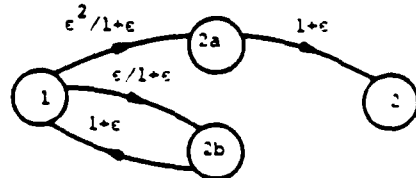
Figure 3.3
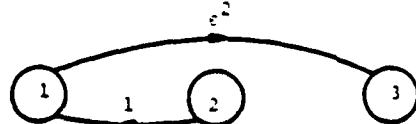
Figure 3.4

Figure 3.5

END

FILMED

11-85

DTIC