October, 1982

LIDS-P-1198
Revised

HIERARCHICAL AGGREGATION OF SINGULARLY PERTURBED FINITE STATE MARKOV PROCESSES

by

M. Coderch, A.S. Willsky, S.S. Sastry and D.A. Castanon*

LABORATORY FOR INFORMATION AND DECISIONS SYSTEMS

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

CAMBRIDGE, MA. 02139 USA

Abstract

In this paper we study the asymptotic behavior of Finite State Markov Processes with rare transitions. We show how to construct a sequence of increasingly simplified models of a singularly perturbed FSMP and how to combine these aggregated models to produce an asymptotic approximation of the original process uniformly valid over $[0,\infty)$

*Research supported in part by the DOE under grant ET-76-C-01-2295. The first author also acknowledges thankfully the continuing support of the Fundación ITP Madrid-Spain. Dr. D.A. Castanon is presently at Alphatech, Inc. 3 New England Executive Park, Burlington, Mass. 01803

Section 1 Introduction and Motivating Example

1.1 Introduction

In this paper we study the asymptotic behavior of continuous time Finite State Markov Processes (FSMP's) with rare transitions. Let $\eta^{\varepsilon}(t)$ be a FSMP with transition probability matrix $P^{\varepsilon}(t) = \exp\{A_0(\varepsilon)t\}$, where

$$A_0(\varepsilon) = \sum_{p=0}^{\infty} \varepsilon^p A_{0p} \tag{1.1}$$

is the matrix of transition rates, and $\varepsilon \in [0,\varepsilon_0]$ is a small parameter modelling rare transitions in $\eta^{\varepsilon}(t)$. We establish that, if the perturbation in (1.1) is singular (in the sense that the number of ergodic classes of $\eta^{\varepsilon}(t)$ changes at $\varepsilon = 0$), then

(i) the limits

$$\lim_{\varepsilon \downarrow 0} P^{\varepsilon}(t/\varepsilon^k) \triangleq P_k(t) \qquad k = 1,2,\ldots m$$

are well defined and determine a finite sequence of (in general stochastically discontinuous) FSMP's $\eta_k(t)$, $k = 1,2,\ldots,m$, with transition probability matrix $P_k(t)$;

ii) the limit processes $\eta_k(t)$ can be aggregated to produce a hierarchy of simplified, approximate models of $\eta^{\varepsilon}(t)$ each of which is a FSMP valid at a certain time scale t/ε^k describing changes in $\eta^{\varepsilon}(t)$ at a distinct level of detail; and

iii) the collection of aggregated models $\hat{\eta}_k(t)$, $k = 1,2,\ldots,m$, can then be combined to construct an asymptotic approximation to $\eta^{\varepsilon}(t)$ uniformly valid for $t \geq 0$.

The idea of using aggregated models to describe gross features of the evolution of Markov processes with rare transitions (i.e., with time scale separation) has been explored by several authors (see [1] - [10]). With the exception of [7] - [10], on which we will comment later, the works referred to above deal with the nearly-decomposable case, i.e., the matrix $A(\varepsilon)$ is assumed to be decomposed into $A(\varepsilon) = A_0 + \varepsilon B$ with A_0 block-diagonal. For this simple case it was proven in [2] that the rare transitions among weakly interacting groups of states (which take place for times of order t/ε) can be modeled by a Markov process with one state for each block in A_0. In [7] the authors allow the chain A_0 to have transient states (i.e. A_0 not block-diagonal) but because only the time scale t/ε was considered, the presence of such states did not modify the basic result in [2]. In [8] the case where transitions between weakly interacting groups do not take place until times of order t/ε^k for some $k \geq 0$ was also included and the same ideas were shown to be applicable to more general Markov processes but, as in previous work, only one aggregated model of a given process was considered. The notion of a hierarchy of aggregated models of a process, each associated with a certain time scale t/ε^k, was discussed in [9] and subsequently in [10]. This latter paper, however, differs in substantial ways from the results and approach taken here. In particular, in [10] the focus is essentially entirely in the frequency domain, i.e. on the resolvent of $A(\varepsilon)$. While a set of aggregated models is developed in this way (using a set of recursive calculations somewhat different than ours), reference [10] does not make a rigorous connection between the aggregated models and the construction of an asymptotic approximation to the original process which is valid on the semi-infinite interval $[0,\infty)$. Such a connection is made in the present paper, using recent results [13] on singularly perturbed linear systems, thus providing for the first time a precise statement of how the set of aggregated models should be

interpreted. Furthermore in our development we have exposed the critical role played by stochastically discontinuous processes in the analysis and aggregation of singularly perturbed FSMP's. Together these results represent significant extensions of previous work and, more importantly, provide a new, concise, and complete picture of the relationships between rare transitions, singular perturbations, stochastic discontinuity, multiple time scale behavior, aggregated modelling and asymptotic approximations for FSMP's.

The paper is organized as follows: In Section 1.2 we use a simple example to motivate the subsequent development. Section 2 is devoted to the introduction of basic definitions, notation and results that we use to establish our main contributions in Section 4. The interpretation of limiting results and aggregation is greatly facilitated by the introduction of stochastically discontinuous FSMP's, (i.e. processes with instantaneous transitions), whose properties have not received much attention in the literature. We study them in Section 3. An example is given in Section 5 and some conclusions are drawn in Section 6.

1.2 A Motivating Example.

Consider the process $\eta^{\varepsilon}(t)$ portrayed in Figure 1 and suppose that $\eta^{\varepsilon}(0) \in \{e_1, e_2\}$. For $\varepsilon > 0$, the process will spend a random amount of time switching between e_1 and e_2 and eventually it will get trapped in e_3. It is clear that we can identify phenomena occuring at two time scales. At the "fast" time scale only transitions between e_1 and e_2 occur and $\eta^{0}(t)$ (in which there is no possibility of transition to e_3) is a good model for that. At the "slow" time scale the important phenomena is a transition to state e_3. Suppose we are interested only in the phenomena occuring at the slow time scale and for ε very small. It is then logical to study the process $\eta^{\varepsilon}(t/\varepsilon)$ in the limit as $\varepsilon \downarrow 0$.

Figure 2 shows a typical sample function of $\eta^{\varepsilon}(t/\varepsilon)$. Each sojourn in states e_1 and e_2 has an average duration of order ε, and on the order of $1/\varepsilon$

such sojourns take place before absorption. In the limit as $\varepsilon \downarrow 0$ the sample functions of $\eta^{\varepsilon}(t/\varepsilon)$ approach functions with an infinite number of discontinuities on finite time intervals. In fact, as we will see in Sections 3 and 4, the finite dimensional distributions of $\eta^{\varepsilon}(t/\varepsilon)$ converge to those of a stochastically discontinuous Markov process $\eta_1(t)$ with sample functions of the type shown in Figure 3. Furthermore, the time to absorption, τ^{ε}, is the sum of a geometrically distributed number of i.i.d. positive random variables and it has mean of order 1. Using results in [11] we can conclude that it converges to an exponentially distributed random variable τ. Therefore, if we define a new process $\hat{\eta}_1(t)$ by

$$\hat{\eta}_1(t) = \begin{cases} \hat{e}_1 & \text{if } \eta_1(t) \in \{e_1, e_2\} \\ \hat{e}_2 & \text{if } \eta_1(t) = e_3 \end{cases}$$

it is clear that $\hat{\eta}_1(t)$ is the Markov process shown in Figure 4 which can be thought of as an approximate, aggregated model for the slow behavior of $\eta^{\varepsilon}(t)$.

This example indicates the need to deal with stochastically discontinuous processes when analyzing the multiple time scale behavior of singularly perturbed FSMP's. Stochastic discontinuity reflects the fact that when a specific time scale is selected, each transition that is likely to occur at a faster time scale (if any) appears to occur instantaneously upon entering some state. Aggregation in this context is simply the avoidance of stochastic discontinuity by discarding the details modeled by these faster, asymptotically discontinuous transitions.

The rest of the paper is devoted to making these ideas precise, and to show that they generalize to arbitrary FSMP's with generator of the form (1.1) and phenomena occurring at several different time scales.

Section 2 Preliminaries

In this section we introduce several definitions and notation and we state some basic results on matrix theory and on linear differential equations possesing multiple time scales. The reader is referred to [12], [13] and [21] for proofs and further development of this preliminary material.

2.1 Matrix Theory. Definitions and Results

Let T denote an (nxn) matrix. The function:

$$R(\xi,T) \triangleq (T-\xi I)^{-1} \qquad \xi \in C$$

is called the resolvent of T and it is analytic with singularities at λ_k, $k = 1,\ldots,s$, the eigenvalues of T. The spectral representation of T is given by

$$T = \sum_{k=1}^{s} (\lambda_k P_k + D_k)$$

where

$$P_k = -\frac{1}{2\pi i}\int_{\Gamma_k} R(\xi,T)\, d\xi$$

and

$$D_k = -\frac{1}{2\pi i}\int_{\Gamma_k} (\xi-\lambda_k)\, R(\xi,T)\, d\xi$$

are respectively the eigenprojection and the eigennilpotent for the eigenvalue λ_k (Γ_k is a positively oriented contour enclosing λ_k but no other eigenvalue of T).

The integer

$$m_k = \dim \mathcal{R}(P_k)$$

is called the algebraic multiplicity of λ_k. An eigenvalue λ_k is said to be semisimple if the associated eigennilpotent is zero and simple if in addition $m_k = 1$. We will say that a matrix T has semisimple null structure (SSNS) if zero is a semisimple eigenvalue of T and we will say that it is semistable if it has SSNS and in addition all its non-zero eigenvalues have negative real parts.

Proposition 2.1

Let T have SSNS and denote by P_0 the eigenprojection for the zero eigenvalue of T. Then:

i) $\mathbb{R}^n = R(T) \oplus N(T)$

ii) P_0 is the projection on $N(T)$ along $R(T)$

iii) $(T + P_0)$ is invertible.

iv) $T^{\#} = (T + P_0)^{-1} - P_0$ satisfies $TT^{\#}T = T$, $T^{\#}TT^{\#} = T^{\#}$ and $T^{\#}T = TT^{\#}$.

It is therefore the group generalized inverse of T.

Proof. See [13]

From the spectral representation of $\exp\{Tt\}$ it follows immediately that the limit of $\exp\{Tt\}$ as $t \to \infty$ exists if and only T is semistable. In this case,

$$\lim_{t\to\infty} \exp\{Tt\} = P_0$$

If T is the matrix of transition rates of a FSMP then the rows of P_0 are the different ergodic probability vectors of the process, and

$$-T^{\#} = (T + P_0)^{-1} - P_0 = \int_0^{\infty} (\exp\{T\,t\} - P_0)\, dt$$

is referred to as the potential matrix.

2.2 Matrix Preturbation Theory

Suppose now that $T(\varepsilon)$, $\varepsilon \in [0,\varepsilon_0]$ is a matrix valued function with an absolutely convergent series of the form:

$$T(\varepsilon) = \sum_{p=0}^{\infty} \varepsilon^p T_p \tag{2.1}$$

An important problem is the nature of the ε-dependence of the eigenvalues, eigenprojections and eigennilpotents of $T(\varepsilon)$ as $\varepsilon \downarrow 0$. For a detailed account the reader is referred to [12], here we briefly state several results that we use later on.

The number of distinct eigenvalues of $T(\varepsilon)$ is constant except at some isolated values of ε. Without loss of generality let $\varepsilon = 0$ be the only exceptional point in $[0,\varepsilon_0]$. The eigenvalues of $T(\varepsilon)$ are continuous functions of ε and at $\varepsilon = 0$ several of them may collapse into a single eigenvalue of $T(0)$. Suppose that rank $T(\varepsilon)$ > rank $T(0)$ and let Γ_0 be a positively oriented contour enclosing zero but no other eigenvalue of $T(0)$. For ε small enough, all eigenvalues of $T(\varepsilon)$ that collapse into the origin as $\varepsilon \to 0$ (referred to as the zero-group of eigenvalues), are inside Γ_0. The matrix

$$P_0(\varepsilon) = -\frac{1}{2\pi i}\int_{\Gamma_0} R(\xi,T(\varepsilon))\, d\xi$$

is therefore equal to the sum of the eigenprojections for eigenvalues of the zero group and it is called the total projection for the zero group of $T(\varepsilon)$. In the subsequent sections we will need the following result.

Proposition 2.2

Let $T(\varepsilon)$ be as in (2.1) and assume that $T_0 = T(0)$ has SSNS then,

$$\frac{T(\varepsilon)P_0(\varepsilon)}{\varepsilon} = \frac{P_0(\varepsilon)T(\varepsilon)}{\varepsilon} = \sum_{n=0}^{\infty} \varepsilon^n \tilde{T}_n \tag{2.2}$$

where

$$\tilde{T}_n = - \sum_{p=1}^{n+1} \sum_{\substack{v_1+\ldots+v_p=n+1 \\ k_1+\ldots+k_{p+1}=p-1 \\ v_i \geq 1,\ k_j \geq 0}} S^{(k_1)} T_{v_1} S^{(k_2)} \ldots S^{(k_p)} T_{v_p} S^{(k_{p+1})} \tag{2.3}$$

where $S^{(0)} = - P_0(0) \triangleq - P_0$ and $S^{(k)} = (T_0^{\#})^k$.

Proof: See [13]

2.3 Asymptotic Approximation of $\exp\{A_0(\varepsilon)t\}$.

Let $A_o(\varepsilon)$ be an nxn matrix having an (absolutely convergent) expansion of the form:

$$A_0(\varepsilon) = \sum_{p=0}^{\infty} \varepsilon^p A_{0p} \tag{2.4}$$

Construct a sequence of matrices $A_k(\varepsilon)$, k = 0,1,...,m, as follows. Let $P_0(\varepsilon)$ denote the total projection for the zero group of eigenvalues of $A_0(\varepsilon)$ and define:

$$A_1(\varepsilon) \triangleq \frac{P_0(\varepsilon)A_0(\varepsilon)}{\varepsilon} = \frac{A_0(\varepsilon)P_0(\varepsilon)}{\varepsilon} = \frac{P_0(\varepsilon)A_0(\varepsilon)P_0(\varepsilon)}{\varepsilon}$$

If A_{00} has SSNS then $A_1(\varepsilon)$ has a series expansion of the form

$$A_1(\varepsilon) = \sum_{p=0}^{\infty} \varepsilon^p A_{1p}$$

If A_{10} also has SSNS then

$$A_2(\varepsilon) \triangleq \frac{P_1(\varepsilon)A_1(\varepsilon)}{\varepsilon} = \frac{P_1(\varepsilon)P_0(\varepsilon)A_0(\varepsilon)}{\varepsilon}$$

also has a series expansion

$$A_2(\varepsilon) = \sum_{p=0}^{\infty} \varepsilon^p A_{2p}$$

where $P_1(\varepsilon)$ is the total projection for the zero group of eigenvalues of $A_1(\varepsilon)$. This recursion can continue if A_{20} also has SSNS. The sequence ends at step m, i.e., at

$$A_m(\varepsilon) = \frac{P_{m-1}(\varepsilon)A_{m-1}(\varepsilon)}{\varepsilon} = \frac{P_{m-1}(\varepsilon)\ldots P_1(\varepsilon)P_0(\varepsilon)A_0(\varepsilon)}{\varepsilon}$$

$$= \sum_{p=0}^{\infty} \varepsilon^p A_{mp}$$

if A_{mo} does not have SSNS, or if

$$\sum_{k=0}^{m} \text{rank } A_{k0} = \text{rank } A_0(\varepsilon) \qquad \text{for } \varepsilon > 0 \tag{2.5}$$

It is not difficult to prove that this stopping condition always occurs at some finite m (see [13]). Of special interest in the sequel are the leading terms in the matrices $A_k(\varepsilon)$, $P_k(\varepsilon)$ and $Q_k(\varepsilon) \triangleq I-P_k(\varepsilon)$, which for convenience we denote as

$$A_k \triangleq \lim_{\varepsilon \downarrow 0} A_k(\varepsilon) = A_{k0} \tag{2.6}$$

$$P_k \triangleq \lim_{\varepsilon \downarrow 0} P_k(\varepsilon) \tag{2.7}$$

$$Q_k = I - P_k$$

To construct a uniform asymptotic approximation of $\exp\{A_0(\varepsilon)t\}$ we require the following condition.

Definition 2. An analytic matrix function $A_0(\varepsilon)$ of ε satisfies the multiple semistability (MSST) condition if the sequence of matrices A_k, k=0,1,...,m, defined in (2.6) are all semistable and satisfy (2.5).

The following is a central result in the asymptotic approximation of $\exp\{A_0(\varepsilon)t\}$.

Theorem 2.3

Let $A_0(\varepsilon)$ be as in (2.4) and let A_k, k=0,1,...,m, be the sequence of matrices constructed above.

i) If A_k, k=0,1,...ℓ-1, are all semistable then $\forall T < \infty$,

$$\lim_{\varepsilon \downarrow 0} \sup_{t \in [0, T/\varepsilon^{\ell}]} ||\exp\{A_0(\varepsilon)t\} - \phi_{\ell}(t,\varepsilon)|| = 0 \tag{2.8}$$

where $\phi_{\ell}(t,\varepsilon)$ is any of the following expressions

$$\phi_{\ell}(t,\varepsilon) = \sum_{k=0}^{\ell} Q_k \exp\{A_k \varepsilon^k t\} + P_0 P_1 \ldots P_{\ell} \tag{2.9}$$

$$= \sum_{k=0}^{\ell} \exp\{A_k \varepsilon^k t\} - \ell I \tag{2.10}$$

$$= \prod_{k=0}^{\ell} \exp\{A_k \varepsilon^k t\} \tag{2.11}$$

$$= \exp\{\sum_{k=0}^{\ell} A_k \varepsilon^k t\} \tag{2.12}$$

and

$$\mathbb{R}^n = R(A_0) \oplus R(A_1) \oplus \ldots \oplus R(A_{\ell}) \oplus (\bigcap_{k=0}^{\ell} N(A_k)) \tag{2.13}$$

ii) If $A_0(\varepsilon)$ satisfies the MSST condition then

$$\lim_{\varepsilon \downarrow 0} \sup_{t \geq 0} ||\exp\{A_0(\varepsilon)t\} - \phi_m(t,\varepsilon)|| = 0 \tag{2.14}$$

where $\phi_m(t,\varepsilon)$ is as in (2.9) - (2.12), and (2.13) is satisfied for $\ell = m$.

Proof: See [13]

Remarks:

1) Theorem 2.3 asymptotically decomposes $\exp\{A_0(\varepsilon)t\}$ into several evolutions, each taking place at a different time scale.

In section 4.3 we interpret the matrices A_K and P_K, $K=0,1,\ldots,m$ in the context of FSMP's as determing a hierarchical sequence of reduced order models of a process $\eta^{\varepsilon}(t)$ with transition probability matrix given by $A_0(\varepsilon)$.

2) If $A_0(\varepsilon)$ satisfies the MSS condition then, for $k=0,1,\ldots,m$, P_k is the projection on $N(A_k)$ along $R(A_k)$ and it follows from (2.13) that:

i) $P_i P_j = P_j P_i \qquad i,j = 0,1,\ldots,m$

ii) $Q_i Q_j = 0 \qquad i\neq j \quad i,j=0,1,\ldots,m$

iii) $A_i A_j = 0 \qquad i\neq j \quad i,j=0,1,\ldots,m$

iv) $$P_j A_i = A_i P_j = \begin{cases} 0 & i=j \\ A_i & i\neq j \end{cases}$$

3. Stochastically Discontinuous Finite State Markov Processes

We consider here continuous-time, stationary, finite-state Markov processes $\{\eta(t),\ t \geq 0\}$ that may undergo an infinite number of transitions in finite time intervals. Such processes violate the continuity condition:

$$\lim_{t\downarrow 0} \Pr\{\eta(t) = \eta(0)\} = 1$$

and, accordingly, are referred to as stochastically discontinous [14]. They were first analyzed in [15] and [16] but were considered pathological from an applications view point and since then stochastic continuity has been a standard assumption in the literature (see for example [17] and [18]). As we have

indicated in Section 1.2, however, stochastically discontinuous processes are obtained as limits of Markov processes with transition rates of different orders of magnitude and the stochastic discontinuity property has a natural and important interpretation in this context.

In this section we carry out an analysis of FSMP's along the same lines usually followed for stochastically continuous processes (as in [19], for example), but for the general, stochastically discontinuous case.

3.1 Characterization of Finite State Markov Processes

A stationary Markov process $\{\eta(t),\ t \geq 0\}$ taking values in a finite state space $E = \{e_1, e_2, \ldots, e_n\}$ is completely described by its transition probability matrix $P(t)$ whose elements are the transition probabilities:

$$p_{ij}(t) = \Pr\{\eta(t) = j \mid \eta(0) = i\} \quad i,j \in E,\ t \quad 0$$

An (nxn) matrix-valued function $P(t)$, $t \geq 0$ is a transition probability matrix of some FSMP if and only if it satisfies the following conditions:

$$\text{i)} \quad P(0) = I \tag{3.1}$$

$$\text{ii)} \quad P(t) \geq 0, \quad \forall t \geq 0 \tag{3.2}$$

$$\text{iii)} \quad P(t) \cdot \mathbb{1} = \mathbb{1}^* \tag{3.3}$$

$$\text{iv)} \quad P(t)\,P(\tau) = P(t+\tau) \quad , \forall t, \tau \geq 0 \tag{3.4}$$

In addition, it is known (see [15], [16]) that if $P(t)$ is the transition probability matrix of a FSMP then it is continuous for $t > 0$ and the limit

$$\lim_{t \downarrow 0} P(t) = \Pi \tag{3.5}$$

always exists. It follows from (3.2) - (3.4) and the continuity of $P(t)$ that Π satisfies:

$$\Pi > 0, \quad \Pi \cdot \mathbb{1} = \mathbb{1}, \quad \Pi^2 = \Pi \tag{3.6}$$

*$\mathbb{1} = [1, 1, \ldots, 1]^T$

and also

$$\Pi P(t) = P(t)\Pi = P(t) \tag{3.7}$$

If Π is the identity matrix then the process $\eta(t)$ with transition probability matrix $P(t)$ is called stochastically continuous, otherwise it is called stochastically discontinuous.

Theorem 3.1

If $P(t)$ is the transition probability matrix of a FSMP then,

$$P(t) = \Pi \exp\{At\} \quad t > 0 \tag{3.8}$$

for a pair of matrices Π, A satisfying:

$$\text{(i)} \quad \Pi \geq 0,\ \Pi \cdot \mathbb{1} = \mathbb{1},\ \Pi^2 = \Pi; \tag{3.9}$$

$$\text{(ii)} \quad \Pi A = A\Pi = A; \tag{3.10}$$

$$\text{(iii)} \quad A \cdot \mathbb{1} = 0; \tag{3.11}$$

$$\text{(iv)} \quad A + c\Pi \geq 0 \quad \text{for some } c \geq 0. \tag{3.12}$$

Conversely, any pair of matrices A, Π satisfying (i) - (iv) uniquely determine a FSMP with transition probability matrix given by (3.8).

Proof: The proof of (3.8) given here adapts a more general result on semigroups in [20] to the context of FSMP's. By the continuity properties of $P(t)$ we have:

$$\lim_{h \downarrow 0} \frac{1}{h} \int_t^{t+h} P(\tau)\,d\tau = \begin{cases} P(t) & \text{if } t > 0 \\ \Pi & \text{if } t = 0 \end{cases}$$

for some Π satisfying (3.6) and (3.7).

It follows from (3.4) and (3.7) that $\forall t > 0$,

$$\int_0^t P(h+\tau)\, d\tau - \int_0^t P(\tau) d\tau = (P(h) - \Pi) \int_0^t P(\tau)\, d\tau$$

which gives

$$\frac{1}{h}\int_t^{t+h} P(\tau) d\tau - \frac{1}{h}\int_0^h P(\tau) d\tau = \frac{1}{h}(P(h) - \Pi)\int_0^t P(\tau) d\tau \qquad (3.12)$$

As $h \downarrow 0$ the left-hand side converges to $P(t) - \Pi$ and therefore

$$\lim_{h\downarrow 0} \frac{P(h) - \Pi}{h} \triangleq A \qquad (3.13)$$

exists. Taking limits as $h \downarrow 0$ in (3.12) we get

$$P(t) = \Pi + A\int_0^t P(\tau) d\tau \qquad \forall t > 0$$

establishing (3.8). Definition (3.13) together with (3.6) and (3.7) give (3.10), and (3.11) follows immediately from (3.8) and the fact that $\Pi \cdot \Pi = \Pi$. The positivity of $P(t)$, i.e.,

$$\frac{1}{h} P(h) = \frac{1}{h}\Pi \exp\{At\} = \frac{1}{h}\Pi + A + \frac{o(h)}{h} \geq 0 \quad \text{for } h \geq 0$$

implies that for h small enough $A + \Pi/h \geq 0$ establishing (3.12). To prove the converse suppose now that Π and A satisfy (3.9) - (3.12). Then, $P(t) = \Pi \exp\{At\}$ clearly satisfies (3.3) and (3.4) and the positivity condition

follows from (3.12) as indicated below:

$$\Pi \exp\{At\} = \Pi e^{-ct} \exp\{(A + cI)t\}$$

$$= e^{-ct} \Pi \sum_{n=o}^{\infty} \frac{(A\Pi + c\Pi)^n}{n!} t^n \geq 0$$

We shall refer to the projection $\Pi = \lim_{t \downarrow 0} P(t)$ as the ergodic projection at zero and to the matrix

$$A = \lim_{h \downarrow 0} \frac{P(h) - \Pi}{h} \tag{3.14}$$

as the infinitesimal generator of $P(t)$.

Remarks

1) It follows from (3.9) that Π is the matrix of ergodic probabilities of a Markov chain and as such it determines a partition of E in terms of ergodic clases, E_i^o, i=1,...s, and transient states, E_T^o,

$$E = (\bigcup_{i=1}^{s} E_i^o) \cup E_T^o$$

that we will refer to as the ergodic partition at zero. As we will see later, this partition corresponds to a classification of states into different types. While the process is in absorbing state (i.e. in an ergodic class E_i^o with a single element), the process behaves as a stochastically continuous FSMP. Instantaneous transitions occur between states belonging to the same ergodic class at zero, and transient states are visited only during transitions between ergodic classes, with no time spent in them.

2) For stochasticaaly continuous processes $\Pi = I$ and conditions (i) - (iv) only require that the rows of A add up to zero and that all its off-diagonal entries be non-negative. In the general case some off-diagonal entries of A can be negative provided the corresponding entry in Π is non-zero (see Example 3.2 below). The usual interpretation of a_{ij} as the rate of transitions from state i to j is thus no longer valid in the stochastically discontinuous case. To interpret these entries it is first necessary to perform an aggregation as discussed in Section 3.3.

Example 3.2

The following is a stochastically discontinuous transition probability matrix:

$$P(t) = \begin{bmatrix} p_1 e^{-\lambda t} & p_2 e^{-\lambda t} & 1-e^{-\lambda t} \\ p_1 e^{-\lambda t} & p_2 e^{-\lambda t} & 1-e^{-\lambda t} \\ 0 & 0 & 1 \end{bmatrix}, \quad p_1 + p_2 = 1, \quad t > 0$$

with initial projection and infinitesimal generator given by:

$$\Pi = \begin{bmatrix} p_1 & p_2 & 0 \\ p_1 & p_2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad ; \quad A = \begin{bmatrix} -p_1\lambda & -p_2\lambda & \lambda \\ -p_1\lambda & -p_2\lambda & \lambda \\ 0 & 0 & 0 \end{bmatrix}$$

For $p_1 = p_2 = \lambda = 1/2$ this is the stochastically discontinuous limit process $\eta_1(t)$ described in Section 1.2.

3.2 Implications of Stochastic Discontinuity

If we consider a separable version of a stochastically continuous FSMP then its sample functions are easily visualized as piecewise continuous functions taking values in E [19]. The evolution of the process can be thought of as succesion of stays in different states of E, each being of random duration and exponentially distributed. The sequence of states visited follows a Markov chain law with one-step transition probabilities determined by the entries of the generator A. On the contrary, the sample functions of a stochastically discontinuous process are much more irregular. As we will now prove, these processes have instantaneous states, i.e., states in which the process spends no time with probability one. Furthermore, in general, a stochastically discontinuous process spends a non-zero amount of time switching among instantaneous states. The sample functions are therefore nowhere continuous on certain time intervals.

Consider a separable version of a FSMP $\eta(t)$ with initial projection Π and generator A, and let Λ be a separating set. For $t > 0$ and $n = 0,1,\ldots,$ take

$$0 = t_{0n} < t_{1n} < \ldots < t_{nn} = t$$

in such a way that the sets

$$\Lambda_n = \{t_{0n}, t_{1n}, \ldots, t_{nn}\}$$

increase monotonically and $\bigcup \Lambda_n = \Lambda \cap [0,t]$. Then we have:

$$\Pr\{\eta(\tau) = i,\ \forall \tau \in [0,t] \mid \eta(0) = i\} =$$

$$\Pr\{\eta(\tau) = i,\ \forall \tau \in [0,t] \cap \Lambda \mid \eta(0) = i\} =$$

$$\lim_{n\to\infty} \Pr\{\eta(\tau) = i,\ \forall \tau \in [0,t] \cap \Lambda_n \mid \eta(0) = i\} =$$

$$\lim_{n\to\infty} \prod_{k=0}^{n-1} p_{ii}\,(t_{k+1,n} - t_{k,n}) =$$

$$\exp\left\{\lim_{n\to\infty} \sum_{k=0}^{n-1} \log p_{ii}(t_{k+1,n} - t_{k,n})\right\} \tag{3.15}$$

where $p_{ii}(t)$ are the diagonal elements of $P(t) = \Pi \exp\{At\}$. Equation (3.15) facilitates a clasification of the states of $\eta(t)$ according to the diagonal entires, π_{ii}, of Π. If $\pi_{ii} = 0$ then $p_{ii}(h) \to 0$ as $h \to 0$ and therefore (3.15) gives:

$$\Pr\{\eta(\tau) = i,\ \forall \tau \in [0,t] \mid \eta(0) = i\} = 0 \qquad \forall t > 0$$

If, on the other hand, $0 < \pi_{ii} \leq 1$ use (3.14) to write:

$$\frac{p_{ii}(h)}{\pi_{ii}} = 1 + \frac{a_{ii}}{\pi_{ii}} h + o(h)$$

or

$$\log p_{ii}(h) = \log \pi_{ii} + \frac{a_{ii}}{\pi_{ii}} h + o(h)$$

and it follows from (3.15) that

$$\Pr\{\eta(\tau) = i, \forall\tau \in [0,t] \,|\, \eta(0) = i\} = \begin{cases} 0 & \text{if } \pi_{ii} < 1 \\ \exp\{a_{ii}t\} & \text{if } \pi_{ii} = 1 \end{cases} \tag{4.16}$$

Definition 3.3

A state i will be called instantaneous if $\pi_{ii} < 1$ and regular if $\pi_{ii} = 1$. An instantaneous state j will be called evanescent if $\pi_{jj} = 0$.

Remarks

(1) Notice that this classification is based on the ergodic partition at zero.

(2) We have just seen that the sojourn time in instantaneous states is zero w.p.1.

(3) Also, the sojourn time in regular states is exponentially distributed. All states of a stochastically continuous process are regular.

(4) In Example 3.2, states $\{1,2\}$ are instantaneous, non-evanescent states while 3 is regular.

(5) Even though the duration of stays in a given instantaneous state is zero w.p.1, there is, in general, a non-zero probability of finding the process in an instantaneous state at any given time (as in states $\{1,2\}$ of Example 3.2).

(6) The probability of finding the process in an evanescent state at any given time is zero. This follows from the fact that $\pi_{ii} = 0$ implies $\pi_{ji} = 0$ $j=1,\ldots,n$ (i.e. evanescent states are transient states of the chain Π) and

because:

$$P(t) = \Pi\exp\{At\} = \Pi\exp\{At\}\ \Pi \quad \forall t > 0$$

we have

$$p_{ji}(t) = 0, \quad \forall t > 0, \quad j=1,\ldots,n \quad .$$

The evanescent states can thus be neglected in the sense that there exists a version of the process $\eta(t)$ with the same finite dimensional distributions which does not take values in the set of evanescent states.

As we will see in the next section, the evolution of a stochastically continuous FSMP can be thought of as follows: While in a regular state, it behaves as a stochastically continuous process. Upon entering a state belonging to, an ergodic class at zero with more than one state, say, E_k°, the process starts switching instantaneously among states in E_k°. The amount of time spent in E_k° is exponentially distributed and after a random stay in E_k° the process jumps to some state in $E - E_k^\circ$. Evanescent states may be visited during transitions between ergodic classes at zero but, as we said, they can be pruned without affecting the finite dimensional distributions of $\eta(t)$.

3.4 Aggregation of Stochastically Discontinuous FSMP

We prove now that all probabilistic properties of a stochastically discontinuous process can be derived from its ergodic projection at zero plus an aggregated version of the process that is stochastically continuous.

The following well known proposition establishes notation that we use in the sequel.

Proposition 3.4

Let Π be the ergodic projection at zero of a FSMP then, by an adequate ordering,

$$\Pi = \begin{bmatrix} \Pi_{11} & 0 & \cdots & 0 \\ 0 & \Pi_{22} & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & \Pi_{ss} & 0 \\ \Pi_{1,s+1} & \cdots & \Pi_{s,s+1} & 0 \end{bmatrix} \tag{3.19}$$

with $\Pi_{kk} = \mathbb{1} \cdot \mu_k^T$, $k = 1,\ldots,s$, for some vector $\mu_k > 0$ such that $\mu_k^T \cdot \mathbb{1} = \mathbb{1}$; and $\Pi_{k,s+1} = \delta_k \cdot \mu_k^T$, $k=1,\ldots,s$ for a set of vectors $\delta_k \geq 0$ such that $\sum_{k=1}^{s} \delta_k = \mathbb{1}$.

Furthermore, define the (nxs) matrix V and the (sxn) matrix U as follows:

$$V = \begin{bmatrix} \mathbb{1} & 0 & \cdots & 0 \\ 0 & \mathbb{1} & \cdots & 0 \\ 0 & 0 & \cdots & \mathbb{1} \\ \delta_1 & \delta_2 & \cdots & \delta_s \end{bmatrix} \tag{3.20}$$

$$U = \begin{bmatrix} \mu_1^T & 0 & \cdots & 0 & 0 \\ 0 & \mu_2^T & \cdots & 0 & 0 \\ 0 & 0 & \cdots & \mu_s^T & 0 \end{bmatrix}$$

then,

$$V \cdot U = \Pi \tag{3.21}$$

$$U \cdot V = I$$

Proof: Follows from the fact that Π is the matrix of ergodic probabilities of a Markov chain. The vector μ_k is the vector of ergodic probabilities of a

chain with state space E_k° and transition matrix Π_{kk}. The vectors δ_k are the trapping probabilities from transient states to ergodic classes [19]. □

Remark Notice that the structure of (3.19) makes explicit the ergodic partition at zero.

We shall refer to (3.21) as the canonical product decomposition of Π. Notice that U and V satisfy

$$U \cdot \mathbb{1} = \mathbb{1} \tag{3.23}$$

$$V \cdot \mathbb{1} = \mathbb{1} \tag{3.24}$$

$$U \cdot \Pi = U \tag{3.25}$$

$$\Pi \cdot V = V \tag{3.26}$$

Theorem 3.5

Let $P(t) = \Pi \exp\{At\}$ be the transition probability matrix of a FSMP $\eta(t)$ taking values in $E = \{e_1,\ldots,e_n\}$ and let s be the number of ergodic classes at zero. Let $\Pi = V \cdot U$ be the canonical product decomposition of Π. Then:

$$\hat{P}(t) \triangleq U\, P(t)\, V = \exp\{UAVt\} \quad \forall t > 0 \tag{3.27}$$

is the transition probability matrix of a FSMP taking values in $\hat{E} = \{\hat{e}_1,\ldots,\hat{e}_s\}$ and

$$P_{(t)} = V\ \hat{P}(t)\ U \qquad \forall t > 0 \tag{3.28}$$

Proof: $\hat{P}(t) \geq 0$, $\hat{P}(t) \cdot \mathbb{1} = \mathbb{1}$ and $\hat{P}(t)\ \hat{P}(t) = \hat{P}(t+\tau)$ follow immediately from positivity of U and V, from (3.23) and (3.24), and from (3.7) and (3.21),

respectively. Use now $I = U\Pi V$ and $A = VUA$ to write

$$\hat{P}(t) = U\Pi V + \sum_{k=1}^{\infty} \frac{t^k}{k!} UA\ (VUA)^{k-1}\ V$$

$$= I + \sum_{k=1}^{\infty} \frac{t^k}{k!} (UAV)^k = \exp\{UAV\ t\} \tag{3.29}$$

To prove (3.28) notice that

$$V\hat{\hat{P}}(t)U = VUP(t)VU = \Pi P(t)\Pi = P(t) \tag{3.30}$$

□

Equation (3.27) can be interpreted as performing an aggregation operation that masks the stochastically discontinuous nature of $P(t)$. Define the aggregated process $\hat{\eta}(t)$ taking values in $\hat{E} = \{\hat{e}_1,\ldots,\hat{e}_s\}$ as follows:

$$\hat{\eta}(t) = \hat{e}_i \qquad \text{if } \eta(t) = E_i^o \quad i=1,\ldots,s \tag{3.31}$$

Assuming that we deal with a version of $\eta(t)$ which does not take values in E_T^o, $\hat{\eta}(t)$ is well defined for $t \geq 0$.

Corollary 3.6

The aggregated process $\hat{\eta}(t)$ is a stochastically continuous FSMP with transition probability matrix $\hat{P}(t)$, i.e.,

$$\Pr\{\eta(t) \in E_j^o | \eta(0) = e_k \in E_i^o\} = \Pr\{\hat{\eta}(t) = \hat{e}_j | \hat{\eta}(0) = \hat{e}_i\} = \hat{p}_{ij}(t) \quad \forall t > 0$$

Proof: Follows directly from (3.31) and the structure of the matrices U and V. □

Note also that (3.28) can be interpreted as follows:

$$\Pr\{\eta(t) = e_i | \eta(0) = e_j\} = \mu_\ell^i \cdot \Pr\{\hat{\eta}(t) = \hat{e}_\ell | \hat{\eta}(0) = \hat{e}_p\} \qquad (3.32)$$

$$e_j \in E_p^\circ,\ e_i \in E_\ell^\circ$$

where μ_ℓ^i is the component of the ergodic probability vector μ_ℓ corresponding to e_i. That is, the transitions between the ergodic classes E_i° are governed by the aggregated process while, once in one of the classes E_i°, the ergodic probabilities μ_ℓ^i are immediately established due to the instantaneous nature of the transitions.

Example 3.7

Consider the process $\eta(t)$ in Example 3.2. Its ergodic partition at zero is

$$E_1^\circ = \{3\} \quad ; \quad E_2^\circ = \{1,2\}$$

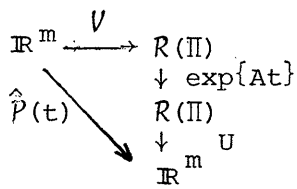
Its aggregated version $\hat{\eta}(t)$ has two states one of which corresponds to the consolidation of states 1 and 2 of $\eta(t)$. The canonical product decomposition of Π is

$$\Pi = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} p_1 & p_2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

and matrix of transition rates for the aggregated process $\hat{\eta}(t)$ is

$$A = U \cdot \begin{bmatrix} -p_1\lambda & -p_2\lambda & \lambda \\ -p_1\lambda & -p_2\lambda & \lambda \\ 0 & 0 & 0 \end{bmatrix} \cdot V = \begin{bmatrix} -\lambda & \lambda \\ 0 & 0 \end{bmatrix}$$

In view of the interpretations described above, we shall refer to (3.27) as the aggregation operation and to (3.28) as the disaggregation operation. These operations can also be interpreted from a geometrical point of view. Notice that the stochastically discontinuous transition probability matrix $P(t) = \Pi \exp\{At\}$ defines a transition matrix on $\mathcal{R}(\Pi)$ which is continuous at zero. By construction, the matrix V maps $\mathbb{R}^s$ into $\mathcal{R}(\Pi)$ on a one-to-one basis and U maps $\mathcal{R}(\Pi)$ back into $\mathbb{R}^s$ also one-to-one. We thus have the following diagram:



From this point of view the aggregation operation is interpreted as a restriction to the range of Π of the domain of definition of the generator A.

To conclude this section notice that all relevant information about a stochastically discontinuous process is contained in its aggregated version and its ergodic projection at zero. Therefore, the analysis of such processes can be reduced, using Theorem 3.5, to the well known stochastically continuous case. As an example of this reduction consider the following corollary to Theorem 3.5.

Corollary 3.8

If $P(t) = \Pi \exp\{At\}$ is the transition probability matrix of a FSMP then the limit

$$\lim_{t\to\infty} P(t) \triangleq \Pi^{\infty} \tag{3.33}$$

always exists and satisfies

i) $\Pi^\infty \geq 0, \quad \Pi^\infty \cdot \mathbb{1} = \mathbb{1}, \quad (\Pi^\infty)^2 = \Pi^\infty$

ii) $\Pi^\infty P(t) = P(t)\, \Pi^\infty = \Pi^\infty$

iii) $A\Pi^\infty = \Pi^\infty A = 0$

iv) $\Pi^\infty \Pi = \Pi \cdot \Pi^\infty = \Pi^\infty$

Proof: Follows from (3.30) and the fact that (3.33) always exists for stochastically continuous processes [19]. □

In the sequel we shall refer to Π^∞ as the ergodic projection at ∞.

For future reference it is important to notice that because

$$P(t) = \Pi \exp\{At\} = \exp\{At\} - I + \Pi$$

equation (3.33) implies that generators of FSMP's are semistable matrices.

Section 4. Singularly Perturbed Finite State Markov Processes

4.1 Regular and Singular Perturbations

Consider now a stochastically continuous FSMP $\eta^\varepsilon(t)$ that takes values in $E_o = \{e_1,\ldots,e_{n_0}\}$ with infinitesimal generator of the form:

$$A_0(\varepsilon) = \sum_{p=0}^{\infty} \varepsilon^p A_{0p} \qquad \varepsilon \in [0,\varepsilon_0] \tag{4.1}$$

The small parameter ε models rare transitions in $\eta^\varepsilon(t)$ and we shall refer to $\eta^\varepsilon(t)$ for $\varepsilon > 0$ as a perturbed version of the process $\eta^0(t)$. Let $P^\varepsilon(t)$ and $P^0(t)$ denote the transition probability matrices of $\eta^\varepsilon(t)$ and $\eta^0(t)$ respectively. Our objective is to analyze the behavior of $\eta^\varepsilon(t)$ (or equivalently, that of $P^\varepsilon(t)$) as $\varepsilon \downarrow 0$ on the time interval $[0,\infty)$.

First, it is straightforward to verify that on any interval of the form $[0,T]$, $\eta^\varepsilon(t)$ can be approximated by $\eta^0(t)$. Precisely,

$$\lim_{\varepsilon \downarrow 0} \sup_{0 \leq t \leq T} ||P^{\varepsilon}(t) - P^{0}(t)|| = 0 \quad \forall T < \infty \tag{4.2}$$

i.e., the finite dimensional distributions of $\eta^{\varepsilon}(t)$ converge to those of $\eta^{0}(t)$ uniformly on $[0,T]$. However, as the example in Section 1.2 illustrates, the behavior of $\eta^{\varepsilon}(t)$ on the infinite time interval $[0,\infty)$ may differ markedly from that of $\eta^{0}(t)$. We shall say that $\eta^{\varepsilon}(t)$ is regularly perturbed if

$$\lim_{\varepsilon \downarrow 0} \sup_{t \geq 0} ||P^{\varepsilon}(t) - P^{0}(t)|| = 0 \tag{4.3}$$

otherwise, we will say that the perturbation is singular. In what follows we focus on the singularly perturbed case, since failure of (4.3) is symptomatic of the existence of distinct behavior at different time scales.

Definition 4.1

We will say that $\eta^{\varepsilon}(t)$ has well defined behavior at time scale t/ε^{k}, $k > 0$, if there exists a continuous, time-dependent matrix $Y_{k}(t)$ such that for any $\delta > 0$, $T < \infty$,

$$\lim_{\varepsilon \downarrow 0} \sup_{\delta \leq t \leq T} ||P^{\varepsilon}(t/\varepsilon^{k}) - Y_{k}(t)|| = 0 \tag{4.4}$$

Remarks:

1) It is readily verified that the limit matrix $Y_{k}(t)$ in (4.4) must be the transition probability matrix of some FSMP $\eta_{k}(t)$ taking values in E_{0}. Thus (4.4) is equivalent to say that $\eta^{\varepsilon}(t/\varepsilon^{k})$ converges to some FSMP $\eta_{k}(t)$ as $\varepsilon \downarrow 0$ in the sense of finite dimensional distributions.

2) As we will see in Section 4.3, $\lim_{\varepsilon \downarrow 0} P^{\varepsilon}(t/\alpha(\varepsilon))$ exists for any order function $\alpha(\varepsilon)$ (α: $[0,\varepsilon_{0}] \to \mathbb{R}^{+}$, $\alpha(0) = 0$ and $\alpha(\cdot)$ continuous and monotone increasing). It turns out, however, that only the limits $Y_{k}(t)$ for a finite

number of positive integers $k = 0,1,\ldots,m$ are required to construct an asymptotic approximation to $P^{\varepsilon}(t)$ uniformly valid for $t \geq 0$. We shall call $t, t/\varepsilon,\ldots, t/\varepsilon^m$ the fundamental or natural time scales of the process $\eta^{\varepsilon}(t)$.

3) Regularly perturbed processes have trivial time scale behavior. For any order function $\alpha(\varepsilon)$

$$\lim_{\substack{\varepsilon \downarrow 0 \\ t>0}} \sup \left|\left|P^{\varepsilon}(t/\alpha(\varepsilon)) - \Pi_0^{\infty}\right|\right| = 0 \tag{4.5}$$

where Π_0^{∞} is the ergodic projection at ∞ of the unperturbed process $\eta^0(t)$.

Proposition 4.2

The process $\eta^{\varepsilon}(t)$ is singularly perturbed if and only if the number of ergodic classes at ∞ of the perturbed process $\eta^{\varepsilon}(t)$ is different than that of $\eta^0(t)$ or, equivalently, if rank $A_0(\varepsilon) \neq$ rank A_{00} for $\varepsilon > 0$.

Proof. See [13] for a proof of the proposition in terms of the rank condition. The statement in terms of the number of ergodic classes at ∞ follows from the fact that this number equals nul $A_0(\varepsilon)$. □

4.2 Multiple Time Scale Behavior and Aggregation

In Section 2.3 we indicated that if a matrix $A_0(\varepsilon)$ satisfies the MSST property then $\exp\{A_0(\varepsilon)t\}$ has an asymptotic approximation that clearly displays its multiple time scale behavior (eqs. (2.9) - (2.11)). We now prove that generators of FSMP's always satisfy the MSST condition and we construct a uniform asymptotic approximation of singularly perturbed FSMP's based on a hierarchy of aggregated models.

The basic result is the following:

Theorem 4.3

Let $\eta^{\varepsilon}(t)$ be a singularly perturbed stochastically continuous FSMP $\eta^{\varepsilon}(t)$

taking values in $E_0 = \{1,2,\ldots,n_0\}$ with transition probability matrix $P^\varepsilon(t) = \exp\{A_0(\varepsilon)t\}$ and infinitesimal generator $A_0(\varepsilon)$ of the form (4.1). Denote by A_k, P_k, $k=0,\ldots,m$, the sequence of matrices constructed from $A_0(\varepsilon)$ as indicated in Section 2.2. Then,

i) A_k and $\Pi_k \triangleq P_0 P_1 \ldots P_{k-1}$ are respectively the infinitesimal generator and the ergodic projection at zero of some FSMP $\eta_k(t)$ taking values in E_0, and

$$\lim_{\varepsilon \downarrow 0} \sup_{t \in [\delta,T]} \| P^\varepsilon(t/\varepsilon^k) - \Pi_k \exp\{A_k t\} \| = 0 \tag{4.6}$$

for $\forall \delta > 0$, $T < \infty$ and $k = 1,2,\ldots,m$ (T can be taken equal to ∞ for $k = m$). Furthermore, let $\Pi_k = V_k \cdot U_k$ be the canonical product decomposition of Π_k. Then,

ii) $$P^\varepsilon(t) = \exp\{A_0(\varepsilon)t\}$$

$$= \sum_{k=0}^{m} \exp\{A_k \varepsilon^k t\} - mI + o(1) \tag{4.7}$$

$$= \prod_{k=0}^{m} \exp\{A_k \varepsilon^k t\} + o(1) \tag{4.8}$$

$$= \exp\{A_0 t\} + \sum_{k=1}^{m} (V_k \exp\{\hat{A}_k \varepsilon^k t\} U_k - \Pi_k) + o(1) \tag{4.9}$$

uniformly for $t \geq 0$, where $\hat{A}_k \triangleq U_k A_k V_k$ is the infitesimal generator of a stochastically continuous FSMP $\hat{\eta}_k(t)$ taking values in $E_k = \{1,2,\ldots,n_k\}$ and

$$n_k = n_0 - \sum_{p=0}^{k-1} \operatorname{rank} A_p \tag{4.10}$$

Proof: The first step is to prove that $A_0(\varepsilon)$ satisfies the MSST property. We use induction. Suppose that $A_0, A_1, \ldots, A_\ell$ are semistable. Then, by Theorem 2.3,

the limit

$$\lim_{\varepsilon \downarrow 0} P^{\varepsilon}(t/\varepsilon^{\ell+1}) = P_0 P_1 \ldots P_{\ell} \exp\{A_{\ell+1} t\} \triangleq \mathcal{P}_{\ell+1}(t) \quad \forall t > 0 \tag{4.11}$$

is well defined. Clearly, $\mathcal{P}_{\ell+1}(t) \cdot \mathbb{1} = \mathbb{1}$, $\mathcal{P}_{\ell+1}(t) \geq 0$ and $\mathcal{P}_{\ell+1}(t) \cdot \mathcal{P}_{\ell+1}(\tau) = \mathcal{P}_{\ell+1}(t+\tau)$ (remember that the projections P_j, $j = 0,1,\ldots,\ell$ conmute with each other and with $A_{\ell+1}$). Therefore $\mathcal{P}_{\ell+1}(t)$ is the transition probability matrix of a FSMP and it follows from Corollary 3.8 that $A_{\ell+1}$ must be semistable. Because A_0 is semistable, MSST is proven, and this together with Theorem 2.3 gives i). The second part of the theorem follows from (2.14) and the fact that, by Theorem 3.5,

$$\exp\{A_k t\} = I - \Pi_k + V_k \exp\{\hat{A}_k t\} U_k$$

with $\hat{A}_k = U_k A_k V_k$ being the infinitesimal generator of a stochastically continuous FSMP with $n_k = \text{rank } \Pi_k$ states. Equation (4.10) then follows from:

$$\text{rank } \Pi_k = \dim \mathcal{R}(P_0 P_1 \ldots P_{k-1}) = n_0 - \sum_{p=0}^{k-1} \text{rank } A_p \qquad \square$$

Remarks

1) The matrices $\Pi_k = V_k \cdot U_k$ and A_k satisfy:

$$\text{a) } \Pi_k \Pi_{\ell} = \Pi_{\ell} \Pi_k = \Pi_k \qquad k \geq \ell \tag{4.12}$$

$$\text{b) } U_k \Pi_{\ell} = U_k \qquad k \geq \ell \tag{4.13}$$

$$\Pi_{\ell} V_k = V_k \qquad k \geq \ell$$

$$\text{c)}\quad \Pi_k A_\ell = A_\ell \Pi_k = \begin{cases} A_\ell & k \le \ell \\ 0 & k > \ell \end{cases} \tag{4.14}$$

2) Part i) implies that the finite dimensional distributions of $\eta^\varepsilon(t/\varepsilon^k)$ converge to those of $\eta_k(t)$ which is, in general, a stochastically discontinuous process.

3) As shown in Section 3.3, each of the ergodic projections at zero Π_k, $k = 1,2,\ldots,m$, determines an aggregation operation performed by collapsing all states that belong to a given ergodic class of Π_k into a single state. If $\eta^\varepsilon(t)$ is aggregated according to the partition specified by Π_k, we get:

$$U_k\, P^\varepsilon(t)\, V_k = \exp\{\hat{A}_k \varepsilon^k t\} + o(1) \tag{4.15}$$

uniformly on $[0,T/\varepsilon^k]$. (This follows using (4.13) and (4.14) in (4.9)). Thus, the aggregation partition specified at stage k isolates transitions between groups of states that are likely to occur over time intervals of order T/ε^k but not over shorter time intervals. In addition, and to first approximation, these transitions follow a markovian law with rates specified by $\hat{A}_k$. It is in this sense that we refer to $\hat{\eta}_k(t)$ as an <u>aggregated model of $\eta^\varepsilon(t)$ valid at time scale t/ε^k</u>. If such aggregation is not performed, the approximate model for time scale t/ε^k, $\eta_k(t)$, is stochastically discontinuous because transitions that occur at slower time scales look as instantaneous in the limit as $\varepsilon \downarrow 0$.

4) The sequence of aggregated models $\hat{A}_k$, $k=1,\ldots,m$, is a hierarchy. We have already seen that the sequence n_k is non-increasing. Furthermore, (4.12) implies that if two rows of Π_ℓ are equal, the corresponding two rows in Π_k are also equal

$\forall k \geq \ell$ and therefore if two states are aggregated together at a certain stage then they are also aggregated together at all stages thereafter.

4.4 Computation of Aggregated Models

The asymptotic approximation of singularly perturbed FSMP's in terms of aggregated models developed in previous sections is based on the matrices A_k, k=1,...,m which are the leading terms in the series expansion of:

$$A_k(\varepsilon) = \frac{P_{k-1}(\varepsilon) \ \ldots \ P_1(\varepsilon) \ P_o(\varepsilon) \ A_o(\varepsilon)}{\varepsilon^k} = \sum_{p=0}^{\infty} \varepsilon^p \ A_{kp}$$

We consider now the computations involved in calculating the matrices U_k, V_k and $\hat{A}_k$ used in (4.9).

It is convenient to visualize the matrices A_{ij} ordered in an array as indicated in Table I. The $(i+1)^{th}$ row can be computed from the i^{th} row using formula (2.3) and the array can be grown triangularly so that computation of A_{k0} requires only the computation of A_{ij} for i=0,...,k-1 and j=0,...,k-i. The following proposition gives explicit expressions for the first three aggregated models $\hat{A}_k$.

Proposition 4.5

The matrices $\hat{A}_k$ for k=0,1,2 and 3 are given by:

$$\hat{A}_0 = A_0 = A_{00}$$

$$\hat{A}_1 = U_1 A_1 V_1 = U_1 A_{01} V_1$$

$$\hat{A}_2 = U_2 A_2 V_2 = U_2 \ (A_{02} - A_{01} A_0^{\#} A_{01}) \ V_2$$

$$\begin{aligned} \hat{A}_3 = U_3 A_3 V_3 = U_3 \ (&A_{03} - A_{01}A_0^{\#}A_{02} - A_{02}A_0^{\#}A_{01} + A_{01}A_0^{\#}A_{01}A_0^{\#}A_{01} \\ &- A_{02}A_1^{\#}A_{02} + A_{02}A_1^{\#}A_{01}A_0A_{01} - A_{01}A_0^{\#}A_{01}A_1^{\#}A_{02} \\ &- A_{01}A_0^{\#}A_{01}A_1^{\#}A_{02} - A_{01}A_0^{\#}A_{01}A_1^{\#}A_{01}A_0^{\#}A_{01}) \ V_3 \end{aligned}$$

Proof: Use (2.3) recursively. □

Remarks

1) In addition to the matrix multiplications and additions indicated in Proposition 4.5, each aggregation stage involves computing a new projection Π_{k-1} and a new generalized inverse $A^{\#}_{k-2}$. Such calculations can be carried out with matrices of increasingly smaller dimension. At the first stage we need to compute Π_1 and $A^{\#}_0$ and even though these are $(n_0 x n_0)$ matrices the computation can be decomposed into a set of smaller problems (essentially one per ergodic class of the unperturbed process). At the second stage, Π_2 can be computed as,

$$\Pi_2 = \lim_{t\to\infty} \Pi_1 \exp\{A_1 t\} = V_1 \left(\lim_{t\to\infty} \exp\{\hat{A}_1 t\}\right) U_1$$

and therefore only the ergodic projection of the aggregated model $\hat{A}_1$ needs to be calculated. Similarly with the generalized inverse $A^{\#}_1$,

$$A^{\#}_1 = -\int_0^\infty (e^{A_1 t} - P_1)dt = \int_0^\infty (V_1 e^{\hat{A}_1 t} U_1 + I - \Pi_1 - P_1)dt$$

$$= -\int_0^\infty V_1 (e^{\hat{A}_1 t} - \hat{P}_1) U_1 dt = V_1 \hat{A}^{\#}_1 U_1$$

which requires only $\hat{A}^{\#}_1 = (\hat{A}_1 + \hat{P}_1)^{-1} - \hat{P}_1$, where $\hat{P}_1$ is the ergodic projection of the aggregated process $\hat{A}_1$. The canonical product decomposition of Π_2 can also be computed from that of Π_1 and that of $\hat{P}_1 = \hat{V}_1 \hat{U}_1$ as follows:

$$\Pi_2 = V_1 \hat{P}_1 U_1 = \underbrace{V_1 \hat{V}_1}_{V_2} \cdot \underbrace{\hat{U}_1 U_1}_{U_2}$$

In summary, the complexity of the computations required decreases with the number of states of the succesive aggregated models and they can be implemented in a recursive fashion. We illustrate this procedure with an example in the next section.

2) In [10] Delebecque gives a recursive algorithm to compute the aggregated models $\hat{A}_k$ by constructing an array analogous to Table 1 but using a somewhat simplified version of formula (2.3) (essentially eliminating terms that are cancelled at subsequent stages of the recursion).

Section 5. An Example

Consider the process $\eta^{\varepsilon}(t)$ in Figure 5. A quick look at the unperturbed version in Figure 6 will convince the reader of the singular nature of the perturbation. The ergodic projection at ∞ of $\eta^{0}(t)$ determines four ergodic classes $E_1 = \{1,2\}^*$, $E_2 = \{3\}$, $E_3 = \{4,5\}$ and $E_4 = \{7\}$ and a transient state $E_T = \{6\}$. The aggregated model $\hat{\eta}_1(t)$ valid at time scale t/ε is portrayed in Figure 7 and it has the following infinitesimal generator:

$$\hat{A}_1 = U_1 B V_1 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1/2 & -1 & 1/2 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Notice that the aggregation operation in addition to collapsing $\{1,2\}$ and $\{4,5\}$ into two states also prunes the evanescent state $\{6\}$. At the next stage we get

$$\hat{P}_1 = \lim_{t\to\infty} e^{\hat{A}_1 t} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

which leads to the following aggregation partition:

$$E_1' = \{1,2\}\ ,\ E_2' = \{4,5\}\ ,\ E_3' = \{7\} \text{ and } E_T' = \{3,6\}\ .$$

The corresponding aggregated model valid at time scale t/ε^2, $\hat{\eta}_2(t)$ has generator:

*Notice that if a transient state communicates with only one ergodic class, as states 2 and 4 do in this example, it can be included in that ergodic class for aggregation purposes.

$$\hat{A}_2 = - U_2 B A_0^{\#} B V_2 = \begin{bmatrix} -1/2 & 1/2 & 0 \\ 1/2 & -1/2 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

and it is represented in Figure 8. Finally,

$$\hat{P}_2 = \lim_{t \to \infty} e^{\hat{A}_2 t} = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 1/2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

leads to the next aggregation partition: $E_1'' = \{1,2,3,4,5,6\}$, $E_2'' = \{7\}$. The aggregated model valid at t/ε^3 has rates

$$\hat{A}_3 = - U_3 B A_0^{\#} B V_1 \hat{A}_1^{\#} U_1 B A_0^{\#} B V_3 = \begin{bmatrix} -1/2 & 1/2 \\ 0 & 0 \end{bmatrix}$$

and it is portrayed in Fig. 9. The hierarchy of models ends here because

$$\text{rank } A_0 + \text{rank } \hat{A}_1 + \text{rank } \hat{A}_2 + \text{rank } \hat{A}_3 = \text{rank } A_0(\varepsilon) = 6 \quad \varepsilon > 0$$

This example illustrates how a comparatively complex singularly perturbed FSMP can be asymptotically approximated by a collection of very simple FSMP.

6.1 Conclusions

We have presented a methodology for isolating different events in a singularly perturbed FSMP according to their level of rareness. This methodology leads to a hierarchy of reduced-order models for such processes, each describing the evolution of the process with a different degree of detail and being adequate at a certain time scale. The complete (finite) collection of models obtained in

this way can then be combined to produce an approximation valid on the infinite time interval $[0,\infty)$. We refer the reader to [21] for the more general case of singularly perturbed linear dynamical systems and for some filtering applications based on the hierarchical description of FSMP's.

Acknowledgments

We acknowledge gratefully many helpful discussions with Profs. G. Verghese and J.C. Willems.

References

1. H.A. Simon and A. Ando, "Aggregation of Variables in Dynamic Systems", Econometrica, 29, (1961), 111-138.

2. V.S. Korolyuk, L.I. Polishchuk and A.A. Tomusyak, "A Limit Theorem for Semi-Markov Processes", Kibernetika, Vol. 5, No. 4 (1969) pp. 144-145.

3. V.G. Gaitsgori and A.A. Pervozvanskii, "Aggregation of States in a Markov Chain with Weak Interactions", Kibernetica, 3, (1975) pp. 91-98.

4. P.V. Kokotovic, "Subsystems, Time Scales and Multimodeling", Invited Paper, 2nd IFAC Symposium on Large Scale Systems: Theory and Applications, Toulouse, France (1980).

5. D. Teneketzis, H. Javid and B. Shridhar, "Control of Weakly-Coupled Markov Chains", Proc. CDC Conference, Albuquerque, (1980).

6. P.J. Curtois, "Decomposability: Queuing and Computer Systems Applications", Academic Press, New York (1977).

7. F. Delebecque and J.P. Quadrat, "Optimal Control of Markov Chains Admitting Strong and Weak Interactions," Automatica, 17, (1981) pp. 281-296.

8. V.S. Korolyuk and A.F. Turbin, "Mathematical Foundations of Phase Consolidation for Complex Systems", Naukova Dumka, Kiev, (1978), (In Russian).

9. D.A. Castanon, M. Coderch, B.C. Levy and A.S. Willsky, "Asymptotic Analysis, Approximation and Aggregation Methods for Stochastic Hybrid Systems", Proceedings 1980 JACC, San Diego, California.

10. F. Delebecque, "A Reduction Process for Perturbed Markov Chains", submitted to SIAM J. Appl. Math.

11. J. Keilson, "Markov Chain Models - Rarity and Exponentiality", Springer Verlag, New York, (1978), pp. 132-133.

12. T. Kato, Perturbation Theory for Linear Operators, Springer Verlag, Berlin, (1966).

13. M. Coderch, A.S. Willsky, S.S. Sastry and D.A. Castanon, "Hierarchical Aggregation of Linear Systems with Multiple Time Scales", to appear in IEEE Trans. on AC.

14. A.A. Dynkin, Markov Processes, Springer Verlag, Berlin, 1965.

15. W. Doeblin, "Sur l'Equation Matricielle A(t+s) = A(t) A(s) et ses Applications aux Probabilités en Chaine", Bull. Sci. Math., 62, (1938), pp. 21-32.

16. J.L. Doob, "Topics in the Theory of Markoff Chains", Trans. Am. Math. Soc., 52, (1942), pp. 37-64.

17. A.V. Skorokhod, "Constructive Methods of Specifying Stochastic Processes", Russ. Math. Surreys, 20, (1965) p. 70.

18. D. Williams, Diffusions, Markov Processes and Martingales, Vol. I: Foundations, Wiley, New York, (1979).

19. J.L. Doob, Stochastic Processes, Wiley, New York, 1953, Chapter VI.

20. E. Hille, R.S. Phillips, Functional Analysis and Semigroups, Am. Math. Soc., Providence, 1957.

21. M. Coderch, "Multiple Time Scale Approach to Hierarchical Aggregation of Linear Systems and Finite State Markov Processes", Ph.D Thesis, MIT-LIDS, August 1982.

Figure Captions

Fig. 1 The process $\eta^{\varepsilon}(t)$.

Fig. 2 A typical sample function of $\eta^{\varepsilon}(t/\varepsilon)$.

Fig. 3 A typical sample function of $\eta_1(t) = \lim_{\varepsilon \downarrow 0} \eta^{\varepsilon}(t/\varepsilon)$.

Fig. 4 The process $\eta^{\varepsilon}(t)$ and its approximate, aggregated model.

Fig. 5 The perturbed process $\eta^{\varepsilon}(t)$.

Fig. 6 The unperturbed process $\eta^{\circ}(t)$.

Fig. 7 Aggregated model valid at time scale t/ε.

Fig. 8 Aggregated model valid at time scale t/ε^2.

Fig. 9 Aggregated model valid at time scale t/ε^3.

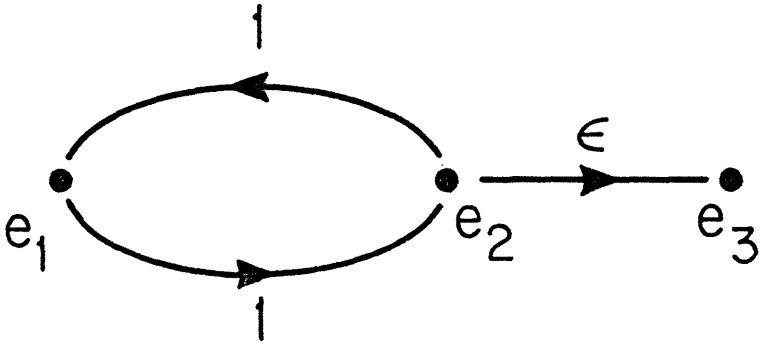


Fig. 1 The process $\eta^{\varepsilon}(t)$.

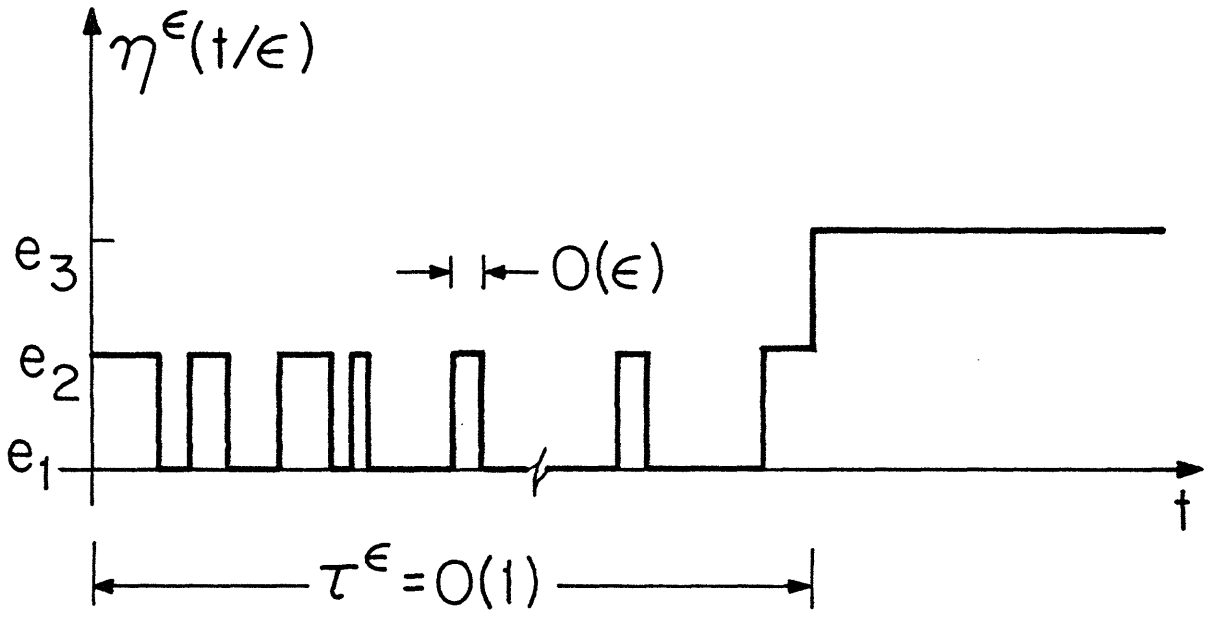


Fig. 2 A typical sample function of $\eta^{\varepsilon}(t/\varepsilon)$.

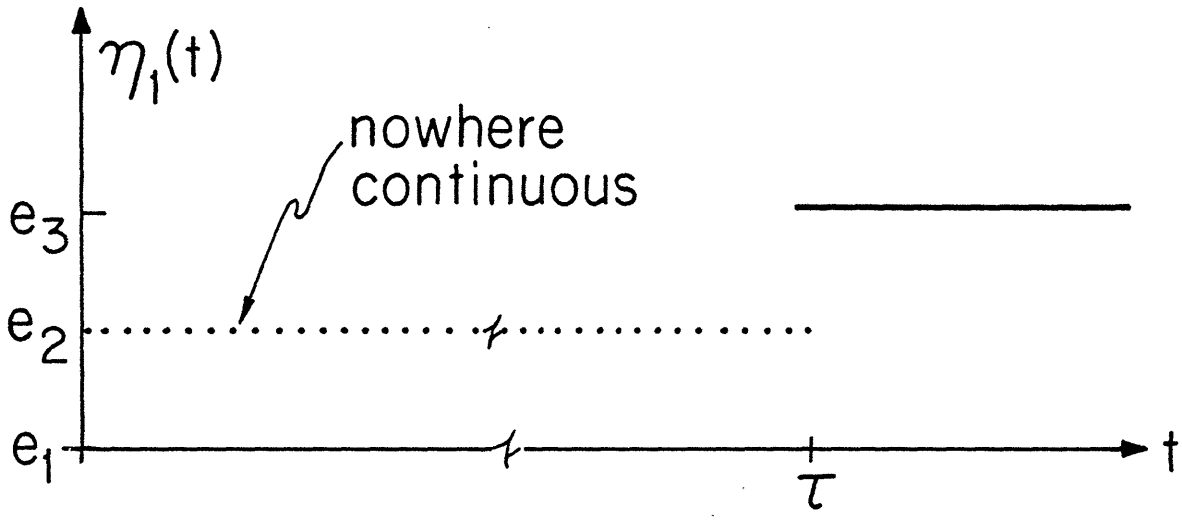


Fig. 3 A typical sample function of $\eta_1(t) = \lim_{\varepsilon \downarrow 0} \eta^{\varepsilon}(t/\varepsilon)$.

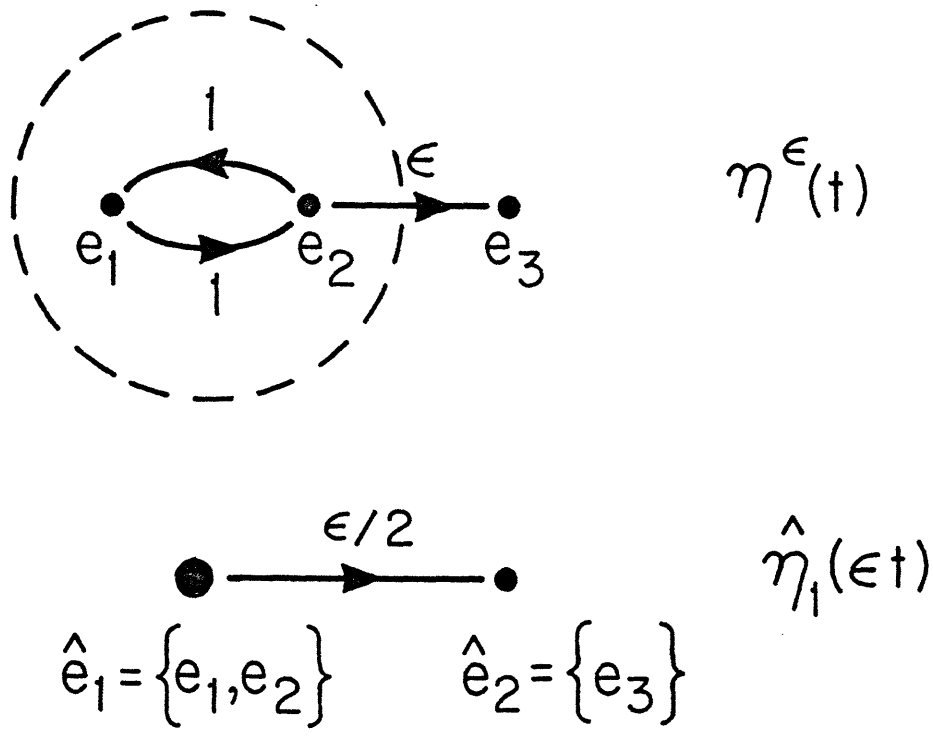


Fig. 4. The process $\eta^{\varepsilon}(t)$ and its approximate, aggregated model.

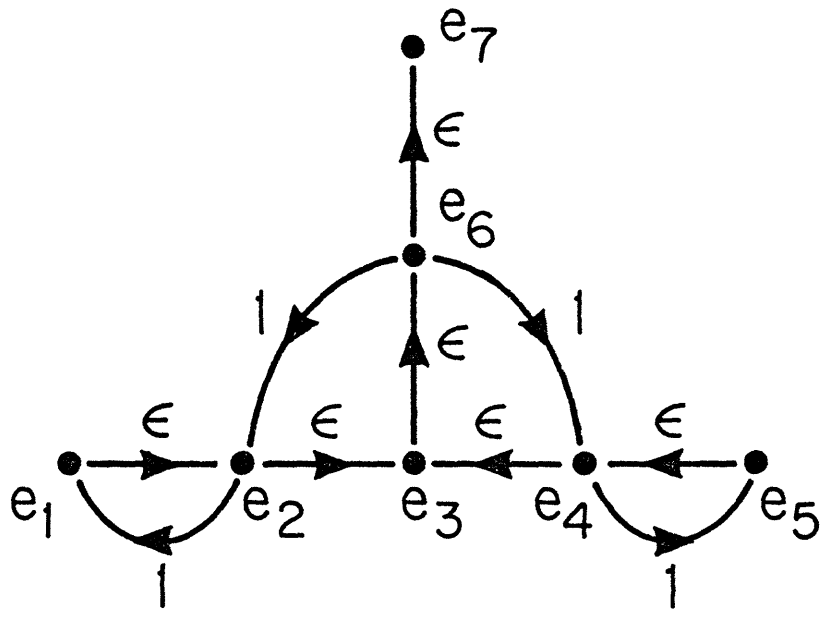


Fig. 5 The perturbed process $\eta^{\varepsilon}(t)$.

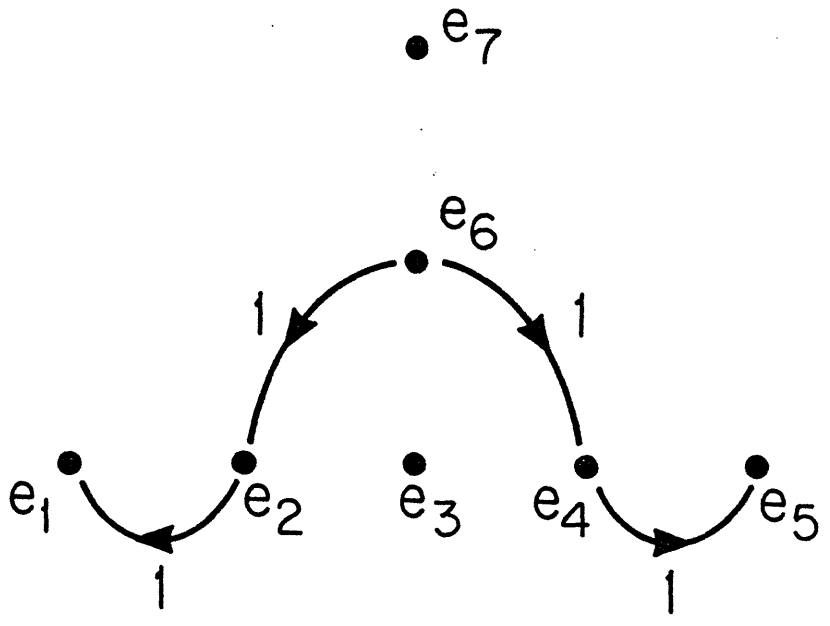


Fig. 6 The unperturbed process $\eta^{\circ}(t)$

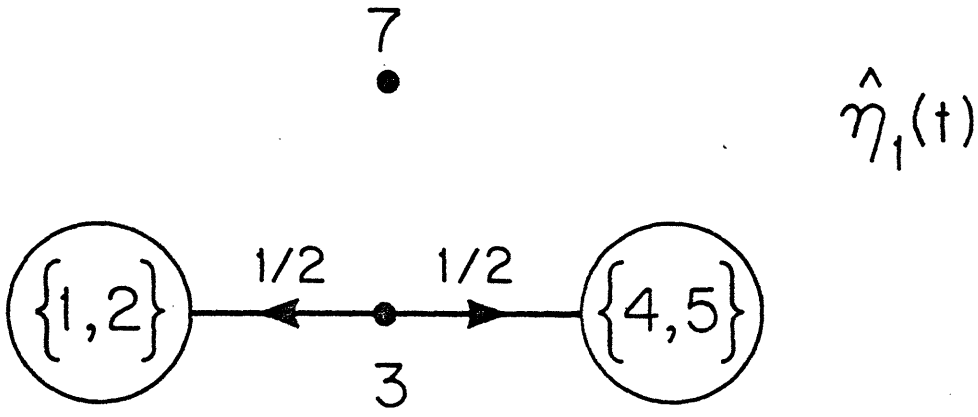


Fig. 7 Aggregated model valid at time scale t/ε.

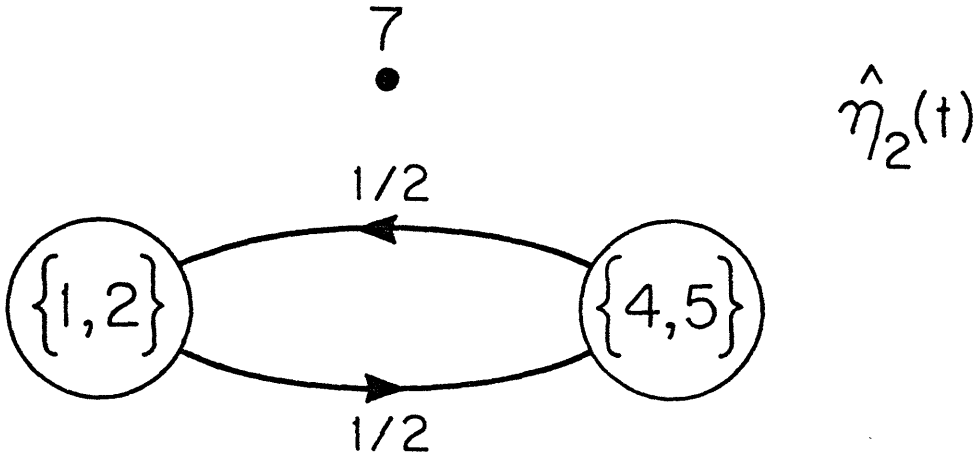


Fig. 8 Aggregated model valid at time scale t/ε^2.

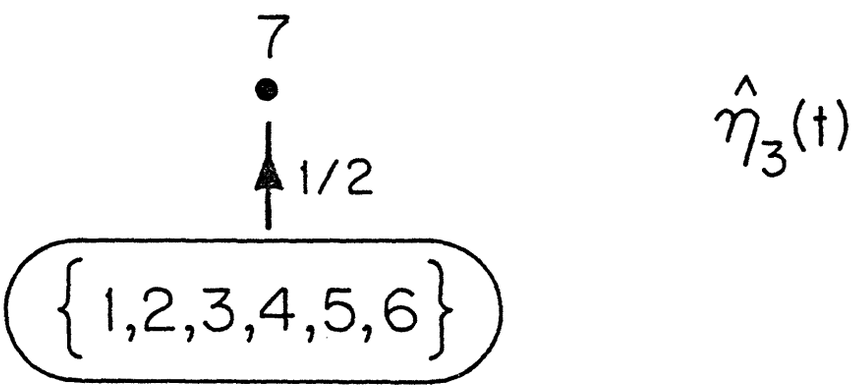


Fig. 9 Aggregated model valid at time scale t/ε^3.