AGGREGATION AND TIME SCALE ANALYSIS OF PERTURBED MARKOV SYSTEMS

by

Jan Robin Rohlicek

S.B., Electrical Engineering, M.I.T.
(1981)

S.M., Electrical Engineering, M.I.T.
(1983)

SUBMITTED TO THE DEPARTMENT OF
ELECTRICAL ENGINEERING AND COMPUTER SCIENCE
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 1987

Signature of Author ____
Department of Electrical Engineering and Computer Science
January 5, 1987

Certified by ____
Alan S. Willsky
Thesis Supervisor

Accepted by ______
Arthur C. Smith
Chairman, Department Committee on Graduate Students

© Massachusetts Institute of Technology, 1987

Aggregation and Time Scale Analysis of Perturbed Markov Systems

by

Jan Robin Rohlicek

Submitted to the Department of Electrical Engineering
and Computer Science on January 5, 1987 in partial
fulfillment of the requirements for the degree of
Doctor of Philosophy

Abstract

Analysis of systems with many time scales is important in many engineering applications. This thesis addresses the approximation and decomposition of Markov processes which exhibit such multiple time scales. An algorithm is presented for the decomposition of explicitly perturbed, finite state, continuous time Markov processes. An approximation of the probability transition function which converges uniformly to zero over $t \geq 0$ is obtained. The algorithm extends previous work by providing a straightforward algorithm which has a direct probabilistic interpretation, particularly with respect to the role played by transient states. This result is then extended to consider semi-Markov and discrete time Markov processes as well. Decomposition of perturbed positive systems is also addressed. Finally, the Markov process decomposition algorithm is expressed in graphical terms and applied to a problem of determining the multiple time scale structure of a fault-tolerant system model.

Thesis Supervisor: Alan S. Willsky
Title: Professor of Electrical Engineering

Acknowledgements

I would like to thank my advisor, Alan Willsky, for his help and encouragement over the past three years. The other members of my thesis committee, George Verghese, John Tsitsiklis, and Bruce Walker, also deserve recognition for their assistance along the way. I would also like to thank Sanjoy Mitter for his optimism and confidence in my work. I am also grateful for the friendship of my fellow graduate students who have made my stay at LIDS enjoyable.

I wish to acknowledge the Air Force Office of Scientific Research for funding this project under grant AFOSR-82-0258 and also the Vinton Hayes fellowship which provided a period of my support. This work was conducted at the Laboratory for Information and Decision Systems at MIT.

Finally, special thanks to my wife Susan for her love and support (and for catching almost all of the speling errors).

to Sue,

my parents,

and our baby to be.

Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation

Many systems exhibit behavior in multiple temporal or spatial "scales". Often these different scales cause difficulty in the analysis of a system. This is either due to numerical ill-conditioning or due to excessive complexity resulting from the explicit consideration of detailed interactions within the system. One approach to the analysis of such systems is to try to isolate the various scales of behavior and analyze them separately, thereby splitting a complex problem into a set of smaller ones.

Systems which have these "multiple scale" properties arise quite frequently in practice. In many engineering techniques, a system is approximated by some simplified model which captures the behavior at the particular scale or scales of interest. An example of this is the use of "lumped" masses in the model of a mechanical system instead of explicitly considering very "stiff" coupling between the components of these masses. There are, however, many situations in which there is no general method for simplifying the problem. Much of the work presented in this thesis is aimed at providing a direct method of deriving valid simplified models for particular classes of systems.

The basic approach of decomposing the various scales of behavior of a complex system has been successfully applied to the analysis of Markov processes with rare transitions. In these systems, there is a strong relationship between coupling of sets of states through very rare transitions and multiple time scales of behavior. There is often a straightforward interpretation of the simplified models which result. States are effectively combined into "aggregate" states where the probability of occupying the aggregate is simply the sum of the probabilities of occupying the member states. A successful method of analysis of such processes has been to explicitly parameterize the rarity of the transitions between the aggregate classes of states and to consider the system as a perturbation of a system where there are no transitions between these aggregate classes.

The work to date on the multiple time scale decomposition of finite-state Markov processes has several shortcomings. As is discussed in the next section, some of the results (such as those of Courtois, [12]) are applicable to only relatively restricted classes of Markov processes. By considering restricted classes, however, the algorithms for the construction of the aggregated processes associated with various time scales are generally straightforward and involve computations with clear probabilistic interpretations. At the other extreme, Coderch [9] [11] and Delebecque [16] deal with very general classes of processes. The resulting algorithms are significantly more complex, however. Coderch *et al* [10] use results on the asymptotic approximation of singularly perturbed systems to construct a sequence of aggregate generators and to prove uniform convergence of the approximation of the probability transition function. The generality of the systems which can be considered and the guaranteed uniform convergence of the approximation are offset by the necessity of computing significantly more complex quantities which are not easily interpreted in probabilistic terms.

Despite the constraints imposed by many of the simpler decomposition techniques, explicit decomposition of Markov processes has been employed in several application. Courtois [12] has exploited the large range of service times in computer queuing networks to form decomposed system models. Many heuristic decomposition methods have also been employed in the analysis of Markov failure models with two time scales of behavior [19] [24] [43]. The use of time scale decomposition of fault-tolerant system models has also been considered [8]. Also, decomposition has been applied to determining the control laws for Markov systems [17]. It is evident that these decomposition techniques have a wide range of potential applications.

The development of a new approach and useful algorithm that addresses the general class of systems considered by Coderch should extend this set of applications still further.

The focus of this thesis is therefore threefold. First, a direct decomposition algorithm will be presented which both treats a general class of Markov processes as in [11] and maintains the simplicity of implementation and interpretation as in [12]. The goal is both to provide a useful engineering tool and to understand the reasons behind the apparent complexity involved in dealing with the general class of perturbed Markov processes. Using this procedure as a basis, the analysis is extended to deal with semi-Markov and discrete time Markov chains. The second focus of the thesis involves using the intuition gained from dealing with stochastic systems and the identification of useful features of such systems to develop a decomposition algorithm for a class of perturbed positive linear systems. Finally, use of the Markov process decomposition result is demonstrated on a potential engineering application in reliability analysis of complex systems. This application, furthermore, demonstrates a particularly useful feature of the decomposition result, namely, its graphical/connectivity interpretation which permits efficient determination of the structure of time scale decompositions by examination of a connectivity graph with links weighted by the integer orders of the corresponding Markov transition rates.

1.2 Background

An overview of several areas of related previous research are described in this section to provide a perspective on the thesis work. More specific discussion of these results appears in subsequent chapters. The related research falls into several categories. First, basic theoretical and algorithmic results related to the decomposition of singularly perturbed finite-state systems are reviewed. The engineering application of decomposition methods, particularly those related to the use of singularly perturbed finite state Markov processes, is then surveyed. For a review of the properties of continuous time and discrete time Markov processes, the reader is referred to [25] and [4] and to [21] for a treatment of semi-Markov processes.

There are several sets of results which are particularly relevant to the theoretical development in this thesis. The first set is directly related to the analysis of Markov processes with rare transitions. Second, there is extensive literature on the use of singular perturbation techniques in control theory. Some of these results have also been applied to the Markov process problem. Finally, some recent work has extended the applicability of these decomposition results to linear systems with more than two time scales and relatively unconstrained structure.

Some of the earliest works to explicitly identify the significance of weak coupling in linear dynamic systems and rare transitions in Markov processes on long-term system behavior is by Simon, Ando and Fisher [1] [42]. These results have been used and extended in more recent work by Courtois [12]. These authors addressed the validity of using certain reduced-order models to approximate the behavior of complex systems. Their motivation has included the study of economic systems where an entire sector of an economy is often represented by a single indicator for the purpose of determining long-term behavior. Their major result is a consequence of the fact that a linear, discrete-time system composed of several almost decoupled components has eigenvalues which form two groups. One group is associated with the fast dynamics of the individual components and the other group is associated with the slow inter-component interactions. This separation of the eigenvalues allows the use of different approximations at different time scales. At a fast time scale, the sectors can be treated as being totally isolated. At the slow time scales, the detailed state of each sector can be collapsed into a single indicator in order to analyze the combined behavior.

Although the work of Simon and Ando was not exclusively concerned with Markov processes, their required structure, known as *nearly complete decomposability*, has a particularly simple interpretation when the linear system defines the evolution of the transition probabilities of a Markov process. The specific formulation in the Markov process context can be expressed as follows. Note that although we will initially consider continuous time Markov processes in Chapter 2, the analogous ideas to this discrete time approach are relevant. Consider a finite state, discrete time Markov process whose state probability vector at time n is $x[n]$

which (after a suitable permutation of the states) is governed by[1]

$$x[n+1] = A(\epsilon)x[n] \tag{1.1}$$

where

$$\begin{aligned} A(\epsilon) &= A + \epsilon B && (1.2) \\ A &= \begin{bmatrix} A_1 & & & 0 \\ & A_2 & & \\ & & \ddots & \\ 0 & & & A_N \end{bmatrix} && (1.3) \end{aligned}$$

Here $A = A(0)$ is referred to as the *unperturbed generator*. Each of the A_I are irreducible (indecomposable) and therefore A is called *completely decomposable*. When the term ϵB is added, $A(\epsilon)$ is *nearly completely decomposable*. The (j, i) element of $A(\epsilon)$, $j \neq i$, is the transition probability from state i to state j. Each of the A_I generates an ergodic chain. The basic decomposition result is that there is an N-state chain which captures the slow behavior corresponding to the transitions between the ergodic classes of A. The transition probabilities of this chain are determined by the ergodic probabilities u_I of each component generated by the A_I and by the perturbation term ϵB.

$$\bar{A}_{IJ} = \sum_{j \in R_J} \sum_{i \in R_I} b_{ji} u_{I_i} \tag{1.4}$$

where R_K is the set of states in the K^{th} ergodic chain of A.

Implicit in Simon and Ando's formulation is that there are only two fundamental time scales of behavior. Courtois [12] has discussed extension of this type of decomposition in an iterative fashion to extract multiple time scale decompositions of Markov processes[2]. In order to use these results, however, the nearly complete decomposability condition must be satisfied at each intermediate time scale and this cannot be easily guaranteed in advance.

[1] Courtois' notation is changed here for consistency with the sequel. Column vectors of probabilities are used in place of row vectors.

[2] In Chapter 5, it is shown that this type of multiple time scale extension of Simon and Ando's result is not valid for more general positive systems despite the nearly completely decomposable structure.

Korolyuk and coworkers considered the decomposition of continuous time, semi-Markov processes with some very small transition probabilities, also expressed in a perturbation form [20] [31]. Although their results are more general in that the semi-Markov case is addressed, similar constraints to those of Simon and Ando are made on the ergodic structure of the system. In particular, transient states are not allowed in the unperturbed process. Furthermore, although the processes are semi-Markov, the probability distributions of the holding times do not depend on the perturbation parameter. The nature of the result is noteworthy, however, in that even though the original system is expressed in a semi-Markov form, the "slow" behavior is captured by a purely Markov process. Another interesting result is observed by Korolyuk and Turbin [33]. Although they introduce a linear perturbation term of order ϵ, due to the structure of the chain constructed, they observe "implicit time scales" for $t = O(1/\epsilon^2)$ or slower. It will be shown in Chapter 2 that such implicit time scales are in fact directly related to the complexity of decomposing an arbitrary perturbed Markov chain.

The use of singular perturbation techniques in control theory has a quite extensive literature. Although some of the basic results will be discussed here, the reader is referred to the extensive reviews by Kokotovic *et al* [28] [29] and Saksena *et al* [40]. The basic problem consider is the evolution of a singularly perturbed, continuous time, linear system

$$\dot{x}(t) = \begin{bmatrix} A_{11} & A_{12} \\ \epsilon A_{21} & \epsilon A_{22} \end{bmatrix} x(t) \tag{1.5}$$

Under suitable stability conditions[3], the "slow" system can be shown to evolve approximately according to

$$\dot{x}_s(\tau) = \left(A_{22} - A_{21}A_{11}^{-1}A_{12}\right) x_s(\tau) \ , \quad \tau = \epsilon t \tag{1.6}$$

This result has been applied to Markov processes by first performing a suitable similarity transformation to bring the generator into the proper form [17]. This expression for the slow behavior is directly related to the continuous time counterpart of Courtois' slow system. Also, the aggregated chain used by Delebecque and

[3] Stability of A_{11} and $A_{22} - A_{21}A_{11}^{-1}A_{12}$ is a sufficient condition. In the Markov process context, the semi-simple zero eigenvalue associated with the steady state can also be present.

Quadrat [17] which is essentially derived from (1.6), uses similar assumptions to those of Courtois.

Recent work has addressed the shortcomings of these available theoretical results. Coderch [9] [10] has considered the decomposition of arbitrary singularly perturbed linear systems which may have behavior at multiple time scales. The derivation draws on Kato's perturbation theory for linear operators [22]. Coderch has also considered the Markov process case and has shown that the slow time scales are associated with aggregated processes. Delebecque [16] has obtained a similar decomposition specifically in the Markov process context. Lou [36] has addressed the "gap" between Coderch's results with those previously obtained by Kokotovic *et al* and others and has developed an algebraic approach to the multiple time scale decomposition of linear systems. The specific problem formulations and results are briefly summarized below. More detailed discussion of these results is available in Section 2.1.

Coderch [10] begins with a linear system where the generator has a Taylor expansion in ϵ

$$\dot{x}(t) = A(\epsilon)x(t) \tag{1.7}$$

where

$$A(\epsilon) = A_0 + \epsilon A_1 + \epsilon^2 A_2 + \cdots \tag{1.8}$$

Under suitable conditions on $A(\epsilon)$ (which he shows are implicitly satisfied when $A(\epsilon)$ is a stochastic matrix), a new generator $\bar{A}(\epsilon)$ can be constructed which captures all the slow behavior. The construction involves identifying the eigenspace associated with all the small eigenvalues of $A(\epsilon)$ (the 0-*group*) and constructing the Taylor expansion of the associated eigenprojection $P(\epsilon)$. Following Kato [22] the new generator is then constructed as

$$\bar{A}(\epsilon) = \frac{1}{\epsilon}A(\epsilon)P(\epsilon) = \frac{1}{\epsilon}P(\epsilon)A(\epsilon)P(\epsilon) \tag{1.9}$$

which also has a similar Taylor expansion as in (1.8). The process can therefore be repeated for a finite number of steps to recover a complete set of reduced order generators[4]. An approximation of the system behavior can then be constructed

[4] Although the matrices are of the original size, the dimension of the nullspace increases and therefore the order of the systems is reduced at each stage.

from the leading order terms of these generators, $A_0, \bar{A}_0, \ldots$. This approximation can be shown to have an error which converges uniformly to 0 on $t \geq 0$ as $\epsilon \downarrow 0$.

Delebecque [16] provides a similar algorithm for the decomposition of a Markov generator though his derivation is based on randomly sampling the the original process at exponentially distributed intervals. It is not clear, however, whether the approximation he constructs has the same uniform validity as the one constructed by Coderch.

Lou [36] extends the type of analysis performed by Kokotovic to systems which have multiple time scales. His result addresses a similar class of systems as Coderch's though he provides a more direct algorithm. The procedure is based on using the Smith decomposition of the generator to construct an ϵ-independent similarity transformation such that the transformed system is in the form

$$\dot{x}(t) = \mathrm{diag}(I, \epsilon I, \epsilon^2 I, \ldots, \epsilon^m I)\bar{A}(\epsilon)x(t) \tag{1.10}$$

where $\bar{A}(0)$ has full rank. It is shown that if the system satisfies the "multiple semi-stability" (MSST) condition (i.e. all the systems describing behavior of particular time scales are semi-stable), then the ϵ-dependence of $\bar{A}(\epsilon)$ does not affect the asymptotic behavior and $\bar{A}(0)$ can therefore be used instead of $\bar{A}(\epsilon)$[5]. This allows a straightforward computation of a sequence of Schur complements to recover the complete time scale decomposition.

Although the various approaches described appear quite different, there is an underlying similarity. In particular, the aggregation result used by Courtois can be shown to be very similar to those of Kokotovic and of Coderch. Although Courtois deals with discrete time systems, the continuous time counterpart of his procedure can be expressed in the original state space in a form similar to that used by Coderch. Specifically, given a continuous time generator $(A + \epsilon B)$, the slow system would be generated by

$$\frac{1}{\epsilon}P(0)(A + \epsilon B)P(0) = P(0)BP(0) \tag{1.11}$$

where

$$P(0) = \lim_{t\to\infty} e^{At} \tag{1.12}$$

[5] Coderch [9] has shown that in the stochastic case, MSST is always satisfied.

Note that the difference here is that the projection $P(0)$ is used instead of $P(\epsilon)$ as in (1.9). The nearly complete decomposability condition in some sense guarantees that the ϵ-dependent terms in $P(\epsilon)$ are not significant. A related interpretation of the Markov decomposition algorithm developed in this thesis is provided in Section 2.1.

Similarly, a nearly completely decomposable system can be transformed into a form related to that used by Kokotovic by performing a similarity transformation. The transformed system has the form (1.5) though the blocks A_{ij} depend on ϵ. In this case, Courtois' slow system is generated by A_{22} rather than the complete expression (1.6). In this nearly completely decomposable case, the term ignored can be shown to be a regular perturbation and therefore does not affect the asymptotic behavior[6].

There are many applications which take advantage of reduced order models based on explicit identification of a perturbation term. Kokotovic *et al* [28] provide a brief survey of some of these applications. The literature on the use of reduced order models based on singularly perturbed Markov processes is more limited. There are, however, many areas where more *ad hoc* decomposition methods for Markov processes are used and where the more explicit use of singular perturbation descriptions of the systems might prove useful.

As pointed out by Courtois [12] and others, hierarchical aggregation can be of great value in analyzing complex processes. Courtois considers decomposing the Markov chain which governs the behavior of a queuing network description of a computer operating system. In this application the small rate, ϵ, is associated with the existence of an order of magnitude range in the service rates of various elements of the system. Another source of such a parameter in similar queuing applications might be some very small routing probabilities when the service rates are comparable [41].

[6] After the similarity transformation, the generator has the form

$$TA(\epsilon)T^{-1} = \begin{bmatrix} A_{11}(\epsilon) & A_{12}(\epsilon) \\ \epsilon A_{21}(\epsilon) & \epsilon A_{22}(\epsilon) \end{bmatrix}$$

where $\|A_{12}(\epsilon)\| = O(\epsilon)$. Under the nearly complete decomposability assumption, $A_{22}(0)$ is a Markov generator with one ergodic class and therefore the term $A_{21}(\epsilon)A_{11}^{-1}(\epsilon)A_{12}(\epsilon)$ which is $O(\epsilon)$ is a regular perturbation of the system.

Other analyses of queuing networks have dealt with calculation or approximation of the "zero" eigenvector corresponding to the steady-state of the system (see [7] for a review of several methods). Although these analyses are not concerned with the dynamics of the system, they apply the decomposition principle to reduce the overall complexity of the problem.

Another application area in which singularly perturbed Markov process models might prove useful is in the reliability analysis of complex systems. In this case, the small parameter is associated with some very small underlying failure rate. Although use of the perturbation methods described earlier is limited, there are many such failure analysis applications which employ heuristic methods for decomposition and order reduction. Keilson [23] [24] has shown that failure times are "asymptotically exponential" and has made heuristic arguments for partitioning the state space into "good" and "bad" sets from which the aggregate behavior is derived. Other uses of reduced order models in failure analysis are surveyed by Gertsbakh [19] though the majority of these techniques have little theoretical basis.

One area of failure analysis which has employed these decomposition techniques is in the analysis of fault tolerant systems [47] [8]. As a consequence of the fault tolerant system operating in conjunction with the underlying failure process, these models are inherently very complex and therefore are difficult to analyze using simpler methods such as those of Keilson. This application area will be discussed further in Chapter 6. Due to the complexity of the models and the presence of a physically significant small parameter (the failure probability), the decomposition algorithm developed in this thesis may provide a useful tool for the analysis of these systems.

1.3 Contributions of the Thesis

The first major contribution of this thesis is the development of multiple time scale decomposition algorithms for perturbed continuous time and discrete time Markov processes and continuous time semi-Markov processes. Several aspects of the Markov decomposition algorithm presented in Chapter 2 should be noted.

- The Markov algorithm bridges the gap between the the conceptually simple results of Simon, Ando, and Fisher and Courtois and the much more complex but general results of Coderch and Delebecque.

- Since the size of the problem is reduced at each time scale, the algorithm can be considered as a method of transforming a single large problem into a set of smaller problems which can be tackled separately. This characteristic of the algorithm should make it possible to consider much larger problems than were previously possible.

- The algorithm can be expressed in graph theoretic terms which, combined with the symbolic rather than numerical nature of the algorithm, allows various types of simplified analysis.

- In contrast to recent work aimed at approximating the steady state distribution of a perturbed Markov chain, this algorithm provides an approximation of the dynamics as well as the steady state.

The extension to continuous time semi-Markov processes presented in Chapter 3 also makes several basic contributions and demonstrates several novel features.

- The holding time probability distributions as well as the transition probabilities can be perturbed.

- Perturbation of the holding time distributions results in the slow time scale system behavior being approximated by another semi-Markov process. This is in contrast to the previously available results where only the transition probabilities are perturbed and the slow system behavior is described using an aggregated Markov process.

- The "fast" and "slow" components of a state may belong to different aggregate states at slower time scales.

The extension to discrete time Markov chains in Chapter 4 also demonstrates a novel feature.

- The slow time scale system behavior of a discrete time Markov chain can be approximated by a continuous time Markov process.

The decomposition algorithm for a class of perturbed positive systems presented in Chapter 5 makes several contributions.

- The close ties between the structure of a Markov process and a positive system are exploited to apply the Markov decomposition algorithm.
- Important differences are identified which illustrate that a direct iterative application of the Markov algorithm is not valid, even in the nearly completely decomposable case.

Finally, this thesis makes a contribution in demonstrating how some of the above results can be applied to a particular engineering problem.

- The graphical/symbolic nature of the Markov decomposition algorithm is exploited to develop a procedure for determining the critical dependence on certain small parameters in a model of a fault tolerant system.

1.4 Outline of Thesis

The remainder of this thesis is organized as follows. The algorithm for the multiple time scale decomposition of perturbed Markov processes is presented and developed in Chapter 2. This is followed by an extension to discrete time Markov chains in Chapter 3 and continuous time semi-Markov processes in Chapter 4. The algorithm is extended to consider a class of continuous time positive systems in Chapter 5. Discussion of the application of the Markov algorithm for structural decomposition with an application to a problem in fault tolerant system design in presented in Chapter 6. Conclusions and an overall discussion are provided in Chapter 7.

Chapter 2

Decomposition of Continuous Time Markov Chains

2.1 Motivation and Background

The results presented in this chapter address the decomposition of a general class of perturbed Markov processes and provide a computationally feasible algorithm for their analysis and uniform approximation. Some of the previous algorithms (such as Courtois [12] and Delebecque and Quadrat [17]) are applicable to only comparatively restricted classes of Markov processes. By considering such restricted classes, however, the algorithms for the construction of the aggregated processes associated with various time scales are generally straightforward and involve computations with clear probabilistic interpretations. At the other extreme, Coderch [9] and Delebecque [16] deal with a completely general class of perturbed Markov processes and the former also proves the uniform convergence of a decomposition-based approximation. The price, however, that is paid for this generality and the guaranteed uniform convergence are algorithms of significantly greater complexity involving the computation of complex quantities that are not easily interpreted in probabilistic terms.

The algorithm presented in this chapter, which was originally outlined in Lou *et al* [35], focuses on the gap between these two extreme sets of results. In particular an algorithm is presented for the construction of uniform multiple time scale approximations of singularly perturbed Markov processes that is as general as that of Coderch [11] and Delebecque [16] but has much the same straightforward, easily interpreted flavor as that of Courtois [12]. Indeed, when the class of systems is suitably restricted, the construction is essentially identical to that of Courtois.

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

$$\dot{X}(t) = A^{(0)}(\epsilon)X(t) \ , \quad X(0) = I \tag{2.1}$$

whose solution can be written as

$$X(t) = \mathrm{e}^{A^{(0)}(\epsilon)t} \tag{2.2}$$

The goal is to obtain an approximation of this solution which (a) explicitly displays the evolution of the process for various orders of t $(1, 1/\epsilon, 1/\epsilon^2, \ldots)$ using appropriately aggregated, ϵ-independent, Markov generators and which (b) converges uniformly over the interval $t \in [0, \infty)$ to the true probability transition function as $\epsilon \downarrow 0$. A solution to (a) and (b) is presented by Coderch *et al* [9] [11] based on associating multiple time scales with different orders of eigenvalues of $A^{(0)}(\epsilon)$. Building on Kato's [22] perturbation results for linear operators, Coderch *et al* identify the subspaces associated with these various orders of eigenvalues and devise a sequential procedure for construction of the approximation. In particular, it is shown that the solution (2.2) can be uniformly approximated using the unperturbed (ϵ-independent) "fast" evolution[2].

$$\mathrm{e}^{A^{(0)}t} \tag{2.3}$$

and a "slow" evolution

$$\mathrm{e}^{\bar{A}^{(1)}(\epsilon)t} \tag{2.4}$$

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

[2] Here $A^{(0)} = A^{(0)}(0)$ for simplicity. To avoid confusion, we will consistently write $A^{(0)}(\epsilon)$ when we are referring to the full ϵ-dependent generator as in (2.2).

where

$$\bar{A}^{(1)}(\epsilon) = \frac{1}{\epsilon} P^{(0)}(\epsilon) A^{(0)}(\epsilon) P^{(0)}(\epsilon) \tag{2.5}$$

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

The approach taken by Courtois [12] overcomes all of these drawbacks. In essence Courtois replaces the slow evolution (2.4) with

$$e^{\bar{F}^{(1)}(\epsilon)\epsilon t} \tag{2.6}$$

where

$$\bar{F}^{(1)}(\epsilon) = \frac{1}{\epsilon} P^{(0)} A^{(0)}(\epsilon) P^{(0)} \tag{2.7}$$

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

$$P^{(0)} = \lim_{t\to\infty} e^{A^{(0)}t} \tag{2.8}$$

This involves no ϵ-dependent computations. Furthermore, if $A^{(0)}$ generates N ergodic classes, the projection can be decomposed as

$$P^{(0)} = U^{(0)} V^{(0)} \tag{2.9}$$

where $U^{(0)}$ is size $n \times N$ and $V^{(0)}$ is size $N \times n$. Here $V^{(0)}$ is a "membership matrix". In the case in which there are no transient states generated by $A^{(0)}$, $V^{(0)}$ consists entirely of 0's and 1's whose rows identify which states of the process form individual ergodic classes of $A^{(0)}$ [3]. Also the columns of $U^{(0)}$ denote the ergodic probability

[3] Note that if $A^{(0)}$ generates transient states, the "membership" of a transient state may be split among several ergodic classes.

vectors, one for each ergodic class of $A^{(0)}$, and finally

$$V^{(0)}U^{(0)} = I \tag{2.10}$$

From (2.9), (2.10), the slow evolution (2.6) can be computed in an even simpler fashion

$$\begin{aligned} e^{\bar{F}^{(1)}(\epsilon)\epsilon t} &= e^{U^{(0)}A^{(1)}(\epsilon)V^{(0)}\epsilon t} && (2.11)\\ &= U^{(0)}e^{A^{(1)}(\epsilon)\epsilon t}V^{(0)} && (2.12)\end{aligned}$$

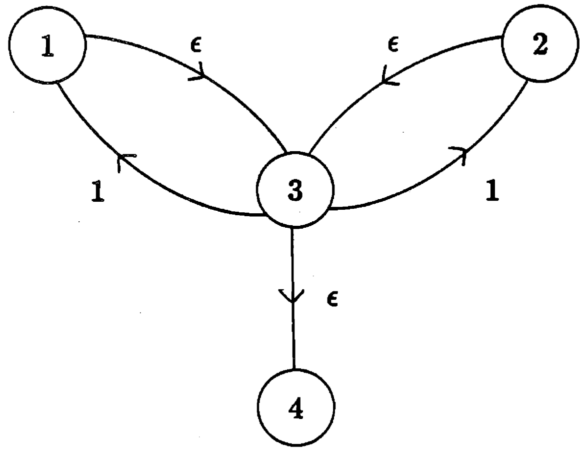
where

$$A^{(1)}(\epsilon) = \frac{1}{\epsilon}V^{(0)}A^{(0)}(\epsilon)U^{(0)} \tag{2.13}$$

is an aggregated Markov generator with one state for each ergodic class of $A^{(0)}$. Indeed (2.13) has an appealing probabilistic interpretation: the transition rate between aggregated ergodic classes of $A^{(0)}$ is computed as an "average rate", in which the rates of individual states in these classes are averaged using the ergodic probabilities of $A^{(0)}$. This is the matrix form of the summation or "averaging" expression (1.4) in Section 1.2.

While the procedure just described has a number of appealing features, it cannot be applied to arbitrary processes. In particular Courtois [12] focuses his development on the class of "nearly completely decomposable" processes introduced by Simon, Ando, and Fisher [1] [42] in which $A^{(0)}$ has no transient states and therefore the the states can be permuted to bring $A^{(0)}$ into a block diagonal form where each block is irreducible (indecomposable). While this condition can be relaxed somewhat, it is restrictive. Furthermore, while the ideas of Simon and Ando, and Courtois do allow one to consider several levels of aggregation at different time scales, iterative application of this method cannot in general be performed since the constraint of near decomposability may fail at one or more intermediate time scales.

The need for a more general algorithm can be traced to the role played by states which are transient at various time scales. To illustrate this, consider the process depicted in Figure 2.1. At $\epsilon = 0$, states 1, 2, and 4 are individual ergodic classes, while state 3 is transient, so that its steady-state probability is 0. Consequently,



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

Figure 2.1: Perturbed Markov process

application of the averaging implied by (2.13) (which uses the steady-state probabilities at $\epsilon = 0$) completely misses the possibility of transition from state 1, 2, or 3 to state 4. Thus in this case the approximation implied by (2.12) does not capture the fact that 4 is in fact a trapping state for any $\epsilon > 0$. The problem in this example is that the critical path determining long-term behavior involves a *sequence* of rare events, namely a transition from state 1 or state 2 to state 3 followed immediately by a transition to state 4.

Processes with such behavior arise in a variety of applications, and are of particular interest in analyzing the long-term reliability or availability of complex systems such as interconnected power networks (in which sequences of events lead, on infrequent occasions, to blackouts), data communication networks, and fault-prone systems possessing back-up capability. The process depicted in Figure 2.1 can in fact be thought of as an (extremely simplified) example of a system consisting of two machines, one of which acts as a backup. States 1 and 2 correspond to both machines being in working order. If a failure of one machine occurs, the process transitions to state 3 in which the machine is examined and then repaired (causing a transition to state 1) or replaced (transition to state 2). However, on rare occasions the second machine fails before the first is repaired or replaced causing a stoppage in operation (and a transition to state 4).

Although the importance of transient states has been recognized in previous work, no general approach has been developed. Korolyuk and Turbin [33] have considered a case where there is a particular ergodic structure. Recently, Bobbio and Trivedi [6] have proposed a method, similar in result to that presented in this chapter, for analyzing the effect of transient states in the two time-scale case. Multiple time-scale analysis of perturbed Markov processes with arbitrary ergodic structure is not available in these works, however, particularly with respect to the construction of a uniform asymptotic approximation.

In this chapter an algorithm for the full multiple time scale analysis of a perturbed Markov process and a proof of the uniform convergence of the approximation are presented. The key to this development is a method for handling transient states at various time scales that couple ergodic classes at slower time scales (as state 3 does between states 1 and 4 and between 2 and 4 in the example above). In general such transient states may not be transient in the full process and thus can be thought of as "almost transient" states. The way in which we accommodate the presence of such states is essentially a modification of (2.13). Specifically, recall that $V^{(0)}$ is a membership matrix indicating which states belong to which ergodic classes. When there are almost transient states it is necessary to consider an ϵ-dependent membership matrix $\tilde{V}^{(0)}(\epsilon)$ to capture the fact that states that couple ergodic classes can be thought of as being "partly" in each. Therefore, in such a case, we must identify and retain certain ϵ-dependent terms, but we can stop far short of the complete computations required by Coderch and can maintain the advantage of Courtois' approach of working directly on increasingly aggregated versions of the process.

In the next section the general algorithm is presented and illustrated on the example introduced in this section. In Section 2.3 the derivation of the procedure is provided along with the proof of the uniform convergence of the approximation. An example is presented in Section 2.4. Section 2.5 contains a discussion of several issues including computational and numerical aspects of hierarchical aggregation. Proofs of some of the supporting results are presented in the appendix, Section 2.A.

2.2 The Algorithm

In this section the general algorithm for the construction of uniform multiple time scale approximations of singularly-perturbed finite-state Markov processes is presented. For simplicity we assume that the Markov generator $A^{(0)}(\epsilon)$ has one ergodic class for $\epsilon > 0$ [4]. The basic algorithm involves the computation of a sequence of Markov generators, the k^{th} of which, $A^{(k)}(\epsilon)$, captures all behavior at time scales of order $1/\epsilon^k$ or slower. The procedure is iterative, with $A^{(k+1)}(\epsilon)$ determined directly from $A^{(k)}(\epsilon)$. There are essentially four steps (1 through 4) which are repeated at each time scale in the algorithm shown below.

Algorithm 2.1 *Begin with the generator $A^{(0)}(\epsilon)$ of a finite-state Markov process with one ergodic class for $\epsilon > 0$. Set $k \leftarrow 0$*

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

2. *For each class E_I, compute the ergodic probabilities $u_{iI}^{(k)}$, $\forall i \in E_I$ of the member states corresponding to the generator $A^{(k)}$. Also set $u_{jI}^{(k)} = 0$, $\forall j \notin E_I$.*

3. *For each transient state $j \in T$ and each class E_I, compute terms $\tilde{v}_{Ij}^{(k)}(\epsilon)$ such that*

$$\tilde{v}_{Ij}^{(k)}(\epsilon) = v_{Ij}^{(k)}(\epsilon)(1 + \mathrm{O}(\epsilon)) \tag{2.14}$$

$$\sum_{K=1}^{N} \tilde{v}_{Kj}^{(k)}(\epsilon) = 1 \tag{2.15}$$

where

$$v_{Ij}^{(k)}(\epsilon) \equiv \Pr\Big(\eta^{(k)}(\epsilon, t^*) \in E_I \;\Big|\; \eta^{(k)}(\epsilon, 0) = j,\ t^* = \inf_{t \geq 0}(t \mid \eta^{(k)}(\epsilon, t) \notin T)\Big) \tag{2.16}$$

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

[4] The generalization to more than one class is trivial, since the states of the process can be reordered such that $A^{(0)}(\epsilon)$ is block diagonal and then each block can be considered individually.

4. *Form the* $n \times N$ *and* $N \times n$ *matrices*

$$U^{(k)} = \left[u_{iJ}^{(k)}\right] \quad \text{and} \quad \tilde{V}^{(k)}(\epsilon) = \left[\tilde{v}_{Ij}^{(k)}(\epsilon)\right] \tag{2.17}$$

   Then

$$A^{(k+1)}(\epsilon) = \frac{1}{\epsilon}\tilde{V}^{(k)}(\epsilon)A^{(k)}(\epsilon)U^{(k)} \tag{2.18}$$

   Set $k \leftarrow k+1$. *Go to 1.*

5. *The overall approximation of the evolution of the transition probabilities can be written as*

$$\begin{aligned} \mathrm{e}^{A^{(0)}(\epsilon)t} &= \mathrm{e}^{A^{(0)}t} + \\ &\left(U^{(0)}\mathrm{e}^{A^{(1)}\epsilon t}V^{(0)} - U^{(0)}V^{(0)}\right) + \\ &\left(U^{(0)}U^{(1)}\mathrm{e}^{A^{(2)}\epsilon^2 t}V^{(1)}V^{(0)} - U^{(0)}U^{(1)}V^{(1)}V^{(0)}\right) + \\ &\quad \vdots \\ &\Big(U^{(0)}\ldots U^{(k-1)}\mathrm{e}^{A^{(k)}\epsilon^k t}V^{(k-1)}\ldots V^{(0)} - \\ &\qquad U^{(0)}\ldots U^{(k-1)}V^{(k-1)}\ldots V^{(0)}\Big) + \mathrm{O}(\epsilon) \end{aligned} \tag{2.19}$$

   where

$$V^{(k)} \equiv \tilde{V}^{(k)}(0) = V^{(k)}(0) \tag{2.20}$$

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

□

As indicated in the Section 2.1, this algorithm is very similar in structure to that of Courtois. In particular, compare (2.13) and (2.18). The computation in step 2 of the ergodic probabilities that form $U^{(k)}$ is identical to the corresponding step of Courtois' algorithm. The critical difference, however is the computation of the "membership matrix" $V^{(k)}(\epsilon)$. In particular, "membership", as needed here is defined in (2.16). Specifically, for each state j in the process corresponding to $A^{(k)}(\epsilon)$, the probability that the process *first enters* each ergodic class E_I of $A^{(k)}(0)$

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

is computed. If j is already a member of some E_I, then the corresponding $v_{Ij}^{(k)}(\epsilon)$ equals 1. In this case we have exactly the same membership as if we used $V^{(k)}(0)$, the quantity employed in Courtois' algorithm. Furthermore, if j is a transient state of $A^{(k)}(0)$ that does not couple transients — i.e. if j has transitions in $A^{(k)}(\epsilon)$ into only one of the E_I — the same 0–1 membership as in $V^{(k)}(0)$ is preserved. However, if j is a coupling transient state, $v_{Ij}^{(k)}(\epsilon)$ in general will be nonzero and ϵ-dependent for several values of I. While there is some ϵ-dependence to be captured here, (2.14) indicates that only the lowest-order term in each $v_{Ij}^{(k)}(\epsilon)$ needs to be matched and then higher-order terms can be picked in order to ensure that the probabilities of membership sum to 1 as in (2.15). This has important computational implications discussed in Section 2.5.

As indicated above, the only elements of $V^{(k)}(\epsilon)$ that require calculation are those which correspond to the transient state set T. The calculation of (2.16), then, is a standard problem: each ergodic class E_I of $A^{(k)}(0)$ is replaced with a single trapping state I, and all transition rates are summed together from each $j \in T$ into each E_I, forming an aggregate rate into the new state I. The probabilities in (2.16) are then simply the limiting transition probabilities as $t \to \infty$ of this simplified process. Furthermore, this is equivalent to considering the limiting probabilities of the derived discrete-time Markov chain whose transition at discrete time n corresponds to the n^{th} transition of the continuous time process. The state transition matrix $\Phi^{(k)}(\epsilon)$ of this discrete-time process (with ergodic classes of $A^{(k)}$ collapsed into trapping states) can be obtained directly from the original generator $A^{(k)}(\epsilon)$.

$$\phi_{ts}^{(k)}(\epsilon) = \frac{a_{ts}^{(k)}(\epsilon)}{-a_{ss}^{(k)}(\epsilon)}\ ,\quad \phi_{Is}^{(k)} = \sum_{i \in E_I} \frac{a_{is}^{(k)}(\epsilon)}{-a_{ss}^{(k)}(\epsilon)}\ ,\quad \phi_{tI}^{(k)}(\epsilon) = 0 \tag{2.21}$$

where $s, t \in T$ and I is a state representing the class E_I. By suitably ordering the states, $\Phi^{(k)}(\epsilon)$ can be formed as

$$\Phi^{(k)}(\epsilon) = \begin{bmatrix} \Phi_{TT}^{(k)}(\epsilon) & 0 \\ \Phi_{RT}^{(k)}(\epsilon) & I \end{bmatrix} \tag{2.22}$$

and the limit therefore becomes

$$\lim_{n\to\infty} \Phi^{(k)}(\epsilon)^n = \begin{bmatrix} 0 & 0 \\ \Phi^{(k)}_{RT}(\epsilon)\left(I - \Phi^{(k)}_{TT}(\epsilon)\right)^{-1} & I \end{bmatrix} = \begin{bmatrix} 0 \\ V^{(k)}(\epsilon) \end{bmatrix} \tag{2.23}$$

The leading order terms of $V^{(k)}(\epsilon)$ in (2.23) required in step 4 of the algorithm can be obtained in a variety of ways such as by repeated multiplication of $\Phi^{(k)}(\epsilon)$ (retaining only the leading order terms in each entry after each multiplication) or by series expansion of the inverse in (2.23) as

$$\left(I - \Phi^{(k)}_{TT}(\epsilon)\right)^{-1} = \left(I - \Phi^{(k)}_{TT}(0)\right)^{-1} \sum_{m=0}^{\infty} \epsilon^m \left(L(\epsilon)(I - \Phi^{(k)}_{TT}(0))\right)^m \tag{2.24}$$

where

$$L(\epsilon) = \frac{1}{\epsilon}\left(\Phi^{(k)}_{TT}(\epsilon) - \Phi^{(k)}_{TT}(0)\right) \tag{2.25}$$

Example 2.1 *In order to illustrate the algorithm, consider the generator*

$$A^{(0)}(\epsilon) = \begin{bmatrix} -\epsilon & 0 & 1 & 0 \\ 0 & -\epsilon & 1 & 0 \\ \epsilon & \epsilon & -2-\epsilon & 0 \\ 0 & 0 & \epsilon & 0 \end{bmatrix} \tag{2.26}$$

associated with the state transition diagram in Figure 2.1 (page 31). The ergodic classes and transient set are

$$E_1 = \{1\},\ E_2 = \{2\},\ E_3 = \{4\},\ T = \{3\} \tag{2.27}$$

The ergodic probabilities are all degenerate in this case

$$u^{(0)}_{11} = u^{(0)}_{22} = u^{(0)}_{43} = 1 \quad or \quad U^{(0)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{2.28}$$

The terms $v^{(0)}(\epsilon)$ can be determined by constructing the discrete time Markov chain with generator $\Phi^{(0)}(\epsilon)$ (with the states ordered $(3,1,2,4)$)

$$\Phi^{(0)}(\epsilon) = \begin{bmatrix} 0 & 0 \\ 1/(2+\epsilon) & \\ 1/(2+\epsilon) & I \\ \epsilon/(2+\epsilon) & \end{bmatrix} = \begin{bmatrix} \Phi^{(0)}_{TT} & 0 \\ \Phi^{(0)}_{RT} & I \end{bmatrix} \tag{2.29}$$

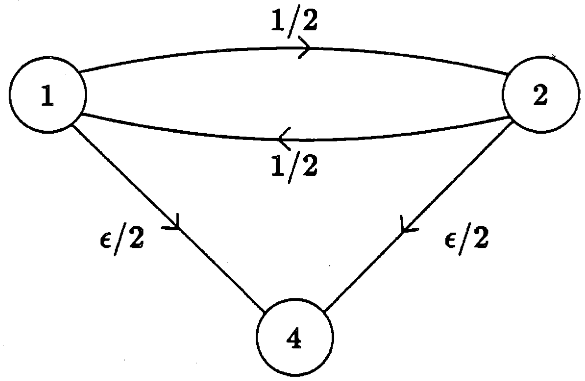


Figure 2.2: Order $1/\epsilon$ time scale model for Example 2.1

Since $\left(\Phi^{(0)}(\epsilon)\right)^2 = \Phi^{(0)}(\epsilon)$, $\lim_{n\to\infty} \Phi^{(0)}(\epsilon)^n$ *is trivial and the terms* $v^{(0)}(\epsilon)$ *can be read directly from* $\Phi^{(0)}(\epsilon)$

$$\begin{aligned} v_{13}^{(0)}(\epsilon) = v_{23}^{(0)} &= 1/(2+\epsilon) \\ v_{33}^{(0)}(\epsilon) &= \epsilon/(2+\epsilon) \end{aligned} \tag{2.30}$$

Suitable terms $\tilde{v}^{(0)}(\epsilon)$ *which satisfy (2.14-2.15) above are*

$$\begin{aligned} \tilde{v}_{13}^{(0)}(\epsilon) = \tilde{v}_{23}^{(0)}(\epsilon) &= 1/2 - \epsilon/4 \\ \tilde{v}_{33}^{(0)}(\epsilon) &= \epsilon/2 \end{aligned} \quad \text{or} \quad \tilde{V}^{(0)}(\epsilon) = \begin{bmatrix} 1 & 0 & 1/2-\epsilon/4 & 0 \\ 0 & 1 & 1/2-\epsilon/4 & 0 \\ 0 & 0 & \epsilon/2 & 1 \end{bmatrix} \tag{2.31}$$

Using these terms, $A^{(1)}(\epsilon)$ *computed using (2.18) generates the process illustrated in Figure 2.2 and given by*

$$A^{(1)}(\epsilon) = \begin{bmatrix} -1/2-\epsilon/2 & 1/2 & 0 \\ 1/2 & -1/2-\epsilon/2 & 0 \\ \epsilon/2 & \epsilon/2 & 0 \end{bmatrix} \tag{2.32}$$

This procedure is now repeated since $A^{(1)}(0)$ *generates two ergodic classes (with no transient states) with the following ergodic probabilities and membership matrices*

$$U^{(1)} = \begin{bmatrix} 1/2 & 0 \\ 1/2 & 0 \\ 0 & 1 \end{bmatrix}, \quad \tilde{V}^{(1)}(\epsilon) = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{2.33}$$

where the aggregate classes are now

$$E_1 = \{1,2\},\ E_2 = \{4\},\ T = \{\} \tag{2.34}$$

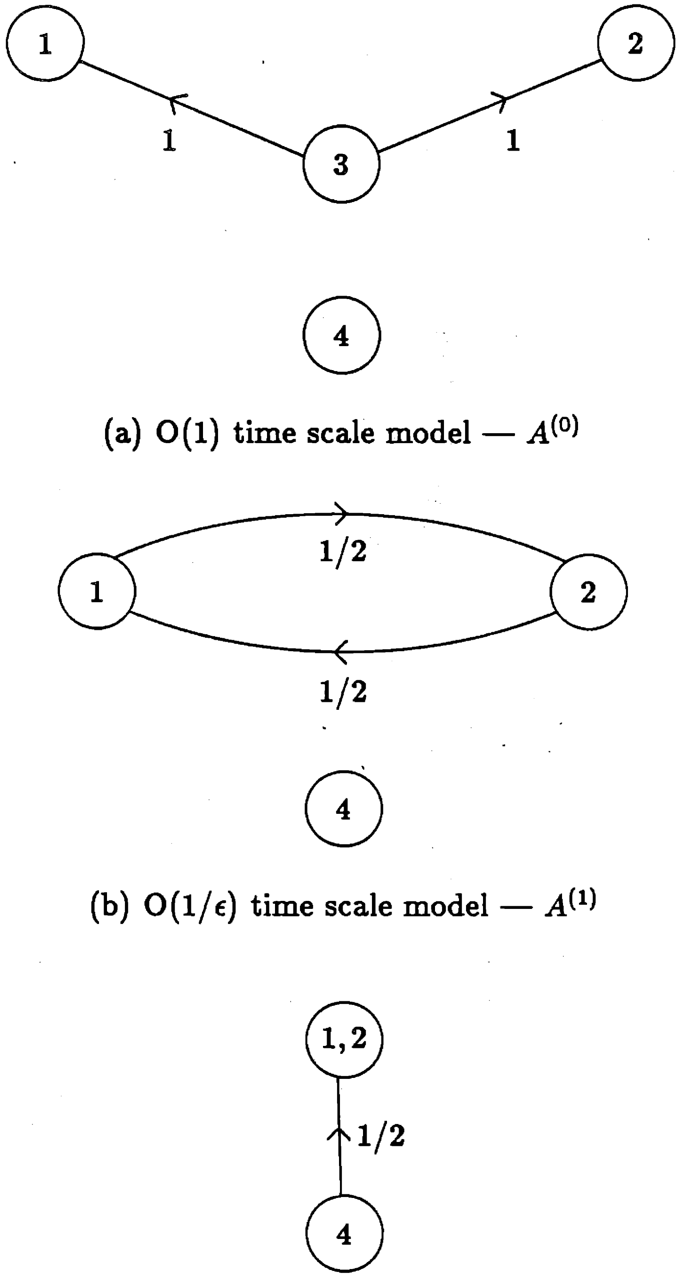
Using this, the generator $A^{(2)}(\epsilon)$ is computed.

$$A^{(2)}(\epsilon) = \begin{bmatrix} -1/2 & 0 \\ 1/2 & 0 \end{bmatrix} \tag{2.35}$$

Since $A^{(2)}(0)$ generates only one ergodic class, namely $\{4\}$, the algorithm is terminated. The set of ϵ-independent Markov models from which the approximation is derived is shown in Figure 2.3. □

Note that the process of Figure 2.1 does not have any probability transition rates of order ϵ^2 or higher. However, as seen in Figure 2.3(c), this process has time scale behavior of order $1/\epsilon^2$. The fact that there is slower behavior than is explicitly visible in the original process is directly attributable to the presence of coupling transient states or, equivalently, to critical *sequences* of rare transitions. This is precisely the case in which the ϵ-dependence of $\tilde{V}^{(k)}(\epsilon)$ is critical.

It is useful to make several comments about step 5 of the procedure which assembles an overall approximation of the transition probability matrix. The first term captures the fast, high-probability behavior at times of order 1. The next describes behavior at times of order $1/\epsilon$ by capturing transitions between ergodic classes of the fast process, and, since these transitions are sufficiently rare that the fast process can reach equilibrium between two such transitions, the probability mass within each ergodic class is distributed using the fast process ergodic probabilities. Similar interpretations can be given to subsequent terms. Such intuition is certainly present or implicit in most previous works. Indeed this idea has led researchers to develop iterative methods for computing steady-state probabilities [7] and error bounds for these computations [13]. In contrast, in the next section it is proved that the error in this approximation to the entire transition probability matrix (including the full transient behavior) goes to 0 uniformly for $0 \leq t < \infty$ as $\epsilon \downarrow 0$. Coderch [9] has a similar uniform convergence proof, but our result is stronger since we are able to work on successively aggregated versions of the process and we



(a) O(1) time scale model — $A^{(0)}$

(b) O($1/\epsilon$) time scale model — $A^{(1)}$

(c) O($1/\epsilon^2$) time scale model — $A^{(2)}$

Figure 2.3: Multiple time scale models for Example 2.1

can also discard all but the essential ϵ-dependent terms (while Coderch keeps them all). Finally, it is interesting to note that the final approximation in (2.16) uses only $\tilde{V}^{(k)}(0) = V^{(k)}(0)$, the same matrices that appear in Courtois' development. The key point here is that while $V^{(k)}(0)$ is adequate for describing the k^{th} time scale, $\tilde{V}^{(k)}(\epsilon)$ is in general needed to capture accurately all slower time scales. For example, the ϵ-dependent terms of $\tilde{V}^{(0)}(\epsilon)$ in Example 2.1 directly influence $A^{(2)}(0)$.

2.3 Derivation

The algorithm for the construction of multiple time scale decompositions of a singularly-perturbed, continuous time, finite-state Markov process is derived in this section. At the same time, the uniform convergence of the resulting approximation is established. The approach taken is as follows. First, an algorithm is derived assuming that there may be transient states at any particular time scale provided that these states cannot "couple" aggregates at slower time scales. The proof of uniform convergence in this case involves keeping track of "weak" terms in the generator which can ultimately be ignored since they do not affect the multiple time scale decomposition. This result provides a proof of the uniform validity of an approximation based on the continuous time counterpart of Courtois' procedure. Such a proof is not available in previous work. Also, this result forms the backbone for the general algorithm. Specifically, an explicit procedure for transforming a process with coupling through transient states to one without such states is provided. Finally from this, the general algorithm that involves a minimum of computation to generate the complete multiple time scale decomposition and uniform approximation is derived.

2.3.1 No coupling through transient states

The first case which will be considered involves imposing a constraint on the structure of the Markov chain for $\epsilon > 0$ (or equivalently on the structure of the generator matrix $A(\epsilon)$) based on identifying states which are transient at $\epsilon = 0$. Before specifying the specific constraints, we should note that if there are no transient states at $\epsilon = 0$, the condition is automatically satisfied. Therefore, the results in this section apply to the nearly completely decomposable case considered by Simon and Ando [42] and Courtois [12]. Furthermore, in the course of this development, a proof of the validity of the two-time-scale approximation used by Delebecque and Quadrat [17] is also obtained.

The specific condition which we will consider is that there is no coupling through transient states in the process[6].

Definition 2.1 *Consider the perturbed Markov generator $A(\epsilon) = A + B(\epsilon)$, with $\|B(\epsilon)\| = O(\epsilon)$. There is* no coupling through transient *states in this process if the following conditions hold. It is possible to partition the state set into sets R_K, each of which can further be partitioned into a single ergodic class of A, E_K, together with a possibly empty transient set T_K, so that these transient states have transitions only into the particular class with which they are associated, even with $\epsilon > 0$. That is if $m \in T_K$ then for any state $n \in R_J, J \neq K$, $a_{nm}(\epsilon) = 0$.* □

To derive and prove the validity of the general result under this condition, we need one basic result and a general recursive procedure that allows us to determine behavior at each successive time scale. The basic result, Lemma 2.1, is an adaptation of results from Lou [36], Coderch [9], and Kokotovic [30].

Definition 2.2 *Two matrices $F(\epsilon)$ and $H(\epsilon)$, analytic functions of ϵ at* 0, *are said to be* asymptotically equivalent *if*

$$\limsup_{\epsilon\downarrow 0\, t\geq 0} \left\| e^{F(\epsilon)t} - e^{H(\epsilon)t} \right\| = 0 \tag{2.36}$$

□

In this case we will write

$$e^{F(\epsilon)t} = e^{H(\epsilon)t} + O(\epsilon) \tag{2.37}$$

[6] Note that this condition could be relaxed to one of "no weak coupling through transient states" whereby any transient state at $\epsilon = 0$ must have either an $O(1)$ or identically 0 probability of first reaching any ergodic class in $O(1)$ time. Since this condition would be more complicated and the above condition is adequate for the development, the latter is used.

Definition 2.3 *A matrix $F(\epsilon)$ is said to have* well-defined time scale behavior *if there exists a constant similarity transformation S such that*

$$SF(\epsilon)S^{-1} \tag{2.38}$$

is asymptotically equivalent to a matrix of the from

$$\operatorname{diag}\left(\epsilon^{i_1}F_1, \ldots, \epsilon^{i_r}F_r\right) \tag{2.39}$$

where the F_i do not depend on ϵ. If such a similarity transformation exists, (F_1, ..., F_r; i_1, ..., i_r; S) determines a time scale decomposition of $F(\epsilon)$. □

Note that if $H(\epsilon)$ is asymptotically equivalent to $F(\epsilon)$, then a time scale decomposition for $F(\epsilon)$ then also serves as one for $H(\epsilon)$. We can now state a basic result which will be employed in the development.

Lemma 2.1 *Suppose that*

$$F(\epsilon) = \begin{bmatrix} F_{11}(\epsilon) & F_{12}(\epsilon) \\ \epsilon F_{21}(\epsilon) & \epsilon F_{22}(\epsilon) \end{bmatrix} \tag{2.40}$$

where $F(\epsilon)$ has well-defined time scale behavior and $F_{11}(0)$ has eigenvalues which have strictly negative real parts.

Then $F(\epsilon)$ is asymptotically equivalent to

$$\begin{bmatrix} F_{11}(0) & 0 \\ 0 & \epsilon K(\epsilon) \end{bmatrix} \tag{2.41}$$

where

$$K(\epsilon) = F_{22}(\epsilon) - F_{21}(\epsilon)F_{11}^{-1}(\epsilon)F_{12}(\epsilon) \tag{2.42}$$

and $K(\epsilon)$ also has well-defined time scale behavior.

Proof *see [9], [30] or [36].* □

This result can be directly applied to perturbed Markov generators since (a) Coderch *et al* [10] have shown that such matrices do have well-defined time scale behavior and (b) it is straightforward to bring the generator into the form in (2.40) using an ϵ-independent similarity transformation.

Under the condition of no coupling through transient states, each transient state is uniquely associated with a single ergodic class. Thus if we assume that there is no coupling through transient states in $A^{(0)}(\epsilon)$, the states can be ordered to bring $A^{(0)}(\epsilon)$ into the block form

$$\begin{aligned} A^{(0)}(\epsilon) &= A^{(0)} + B^{(0)}(\epsilon) \qquad (2.43) \\ A^{(0)} &= \mathrm{diag}(A_1, \ldots, A_N) \qquad (2.44) \\ \left\| B^{(0)}(\epsilon) \right\| &= \mathrm{O}(\epsilon) \qquad (2.45) \end{aligned}$$

Moreover, while each A_I may now generate transient states, it must also be true that certain corresponding elements of $B^{(0)}(\epsilon)$ are identically zero to avoid coupling.[7]

In order to transform $A^{(0)}(\epsilon)$ into the form (2.40), let $U^{(0)}$ and $V^{(0)}$ denote the matrices of right and left zero eigenvectors of the unperturbed generator $A^{(0)}$ where the k^{th} column of $U^{(0)}$ and the k^{th} row of $V^{(0)}$ have nonzero entries only corresponding to states in the k^{th} set of states R_k. Specifically, the i^{th} column $U^{(0)}$ and the i^{th} row of $V^{(0)}$ are of the form

$$U_i^{(0)} = \begin{bmatrix} 0 \cdots 0 \; \pi_i \; 0 \cdots 0 \end{bmatrix}^{\mathrm{T}}, \;\; A_i \pi_i = \mathbf{0}\,, \;\; \mathbf{1}^{\mathrm{T}} \pi_i = 1 \tag{2.46}$$

and

$$V_i^{(0)} = \begin{bmatrix} 0 \cdots 0 \; \mathbf{1}^{\mathrm{T}} \; 0 \cdots 0 \end{bmatrix}, \;\; \mathbf{1}^{\mathrm{T}} A_i = \mathbf{0}^{\mathrm{T}} \tag{2.47}$$

Note that the matrices $U^{(0)}$ and $V^{(0)}$ are the terms $U^{(0)}$ and $\tilde{V}^{(0)}(\epsilon)$ computed using Algorithm 2.1 since the terms $v^{(0)}(\epsilon)$ defined in (2.16) do not depend on ϵ in this case. Also, let $Y^{(0)}$ ($Z^{(0)}$) be matrices whose columns (rows) span the right (left) eigenspace of the nonzero eigenvalues of $A^{(0)}$. Furthermore, due to the structure of $A^{(0)}$, we can clearly choose these matrices such that $A^{(0)}Y^{(0)}$ and $Z^{(0)}A^{(0)}$ are block diagonal with partitions consistent with $A^{(0)}$ and that a similarity transformation T can then be constructed using

$$T = \begin{bmatrix} Z^{(0)} \\ V^{(0)} \end{bmatrix}, \;\; T^{-1} = \begin{bmatrix} Y^{(0)} \; U^{(0)} \end{bmatrix} \tag{2.48}$$

[7] The precise structure of $B(\epsilon)$ in this case is exploited in the proof of Lemma 2.2.

Application of this similarity transformation to $A^{(0)}(\epsilon)$ results in the form given for $F(\epsilon)$ in Lemma 2.1:

$$TA^{(0)}(\epsilon)T^{-1} = \begin{bmatrix} A_{11}(\epsilon) & \epsilon A_{12}(\epsilon) \\ \epsilon A_{21}(\epsilon) & \epsilon A_{22}(\epsilon) \end{bmatrix} \tag{2.49}$$

where

$$\begin{aligned} A_{11}(\epsilon) &= Z^{(0)}A^{(0)}(\epsilon)Y^{(0)} && (2.50) \\ \epsilon A_{12}(\epsilon) &= Z^{(0)}B^{(0)}(\epsilon)U^{(0)} && (2.51) \\ \epsilon A_{21}(\epsilon) &= V^{(0)}B^{(0)}(\epsilon)Y^{(0)} && (2.52) \\ \epsilon A_{22}(\epsilon) &= V^{(0)}B^{(0)}(\epsilon)U^{(0)} && (2.53) \end{aligned}$$

Since $Z^{(0)}$ and $Y^{(0)}$ are associated with the non-zero eigenvalues of $A^{(0)}$ and since the original system has no singularities in the right half-plane, $A_{11}(0)$ has eigenvalues with negative real parts satisfying the condition of Lemma 2.1. Applying Lemma 2.1 and expressing the result in the original basis yields

$$\begin{aligned} e^{A^{(0)}(\epsilon)t} &= Y^{(0)}e^{A_{11}(0)t}Z^{(0)} + U^{(0)}e^{\epsilon G^{(1)}(\epsilon)t}V^{(0)} + O(\epsilon) && (2.54) \\ &= Qe^{A^{(0)}t} + U^{(0)}e^{\epsilon G^{(1)}(\epsilon)t}V^{(0)} + O(\epsilon) && (2.55) \\ &= e^{A^{(0)}t} + U^{(0)}e^{\epsilon G^{(1)}(\epsilon)t}V^{(0)} - U^{(0)}V^{(0)} + O(\epsilon) && (2.56) \end{aligned}$$

where

$$\begin{aligned} G^{(1)}(\epsilon) &= A_{22}(\epsilon) - \epsilon A_{21}(\epsilon)A_{11}^{-1}(\epsilon)A_{12}(\epsilon) && (2.57) \\ Q &= Y^{(0)}Z^{(0)} && (2.58) \\ P &= U^{(0)}V^{(0)} && (2.59) \\ Q &= I - P && (2.60) \end{aligned}$$

From (2.56) the problem of uniformly approximating $\exp(A^{(0)}(\epsilon)t)$ has been reduced to that of approximating $\exp(\epsilon G^{(1)}(\epsilon)t)$. In effect, one time scale has been "peeled off" leaving a lower dimension problem. However, the process is not perfectly inductive since $G^{(1)}(\epsilon)$ need not be the generator of a Markov chain (on the aggregated state space defined by $U^{(0)}$ and $V^{(0)}$). On the other hand, $G^{(1)}(\epsilon)$ is very close to being a Markov generator[8]. Specifically, a careful examination of (2.57) shows that $G^{(1)}(\epsilon)$ can be expressed as

$$G^{(1)}(\epsilon) = A^{(1)}(\epsilon) + W^{(1)}(\epsilon) \tag{2.61}$$

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

$$\begin{aligned} A^{(1)}(\epsilon) &= \frac{1}{\epsilon} V^{(0)} A^{(0)}(\epsilon) U^{(0)} && (2.62) \\ &= \frac{1}{\epsilon} V^{(0)} B^{(0)}(\epsilon) U^{(0)} && (2.63) \\ &= A^{(1)} + B^{(1)}(\epsilon) && (2.64) \end{aligned}$$

where

$$\begin{aligned} A^{(1)} &\equiv A^{(1)}(0) && (2.65) \\ \left\| B^{(1)}(\epsilon) \right\| &= \mathrm{O}(\epsilon) && (2.66) \end{aligned}$$

and

$$W^{(1)}(\epsilon) = -\frac{1}{\epsilon} V^{(0)} B^{(0)}(\epsilon) Y^{(0)} \left(Z^{(0)} A^{(0)}(\epsilon) Y^{(0)} \right)^{-1} Z^{(0)} B^{(0)}(\epsilon) U^{(0)} \tag{2.67}$$

It is not difficult to see that since $B^{(0)}(\epsilon)$ is $\mathrm{O}(\epsilon)$ and the term $Z^{(0)} A^{(0)}(0) Y^{(0)}$ is nonsingular, $W^{(1)}(\epsilon)$ is $\mathrm{O}(\epsilon)$. It will be shown that the term $W^{(1)}(\epsilon)$ can be entirely neglected. In the two time scale case, this follows from the fact that $A^{(1)}(\epsilon)$ is regularly perturbed since in this case all its nonzero eigenvalues are $\mathrm{O}(1)$. Thus $\exp(G^{(1)}(\epsilon)t)$ can then be uniformly approximated using $G^{(1)}(0) = A^{(1)}$ [10] [30][9]. This yields the two time scale result:

$$\mathrm{e}^{A^{(0)}(\epsilon)t} = \mathrm{e}^{A^{(0)}t} + U^{(0)} \mathrm{e}^{\epsilon A^{(1)} t} V^{(0)} - U^{(0)} V^{(0)} + \mathrm{O}(\epsilon) \tag{2.68}$$

If there are more than the two time scales 1 and $1/\epsilon$ in the original process, $A^{(1)}(\epsilon)$ is again singularly perturbed. $W^{(1)}(\epsilon)$ cannot therefore be ignored based

[8] Though the columns of $G(\epsilon)$ sum to zero, some of the off-diagonal elements may be small but negative.

[9] In this case, the eigenvalues of $G^{(1)}(\epsilon)$ are all strictly $\mathrm{O}(1)$ or identically zero. Since the column sums of $W^{(1)}(\epsilon)$ are zero, the zero eigenvalue can be shown to be unperturbed. The perturbation of the $\mathrm{O}(1)$ eigenvalues can then be ignored.

only on its being $O(\epsilon)$, since $O(\epsilon)$ terms may have an effect on the $O(1/\epsilon^2)$ and slower time scales. In order to show that $A^{(1)}(\epsilon)$ is asymptotically equivalent to $G^{(1)}(\epsilon)$ under the assumptions that there is no coupling through transient states, the properties of $W^{(1)}(\epsilon)$ must be considered. To do this a precise definition of "weak" terms associated with a Markov generator is given.

Definition 2.4 *Let $F(\epsilon)$ be the generator of a Markov process with one ergodic class for $\epsilon > 0$. $W(\epsilon)$ is* weak *with respect to $F(\epsilon)$ if (a) $\mathbf{1}^{\mathrm{T}}W(\epsilon) = \mathbf{0}^{\mathrm{T}}$ and (b) for any element $w_{ij}(\epsilon)$ there exists a path $S = (s_1 = j, s_2, \ldots, s_k = i)$ through the process state space such that*

$$w_{ij}(\epsilon) = \epsilon\, O\Big(f_{s_2 s_1}(\epsilon) f_{s_3 s_2}(\epsilon) \cdots f_{s_k s_{k-1}}(\epsilon)\Big) \tag{2.69}$$

□

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

Lemma 2.2 *Suppose that $A^{(0)}(\epsilon)$ is in block form (2.43) and there is no coupling through transient states, then $W^{(1)}(\epsilon)$ in (2.67) is weak with respect to $A^{(1)}(\epsilon)$ in (2.62).*

Proof *see Section 2.A.1.* □

A recursive procedure can now be defined and analyzed. Specifically, suppose that we have constructed $G^{(k)}(\epsilon) = A^{(k)}(\epsilon) + W^{(k)}(\epsilon)$, where (a) $A^{(k)}(\epsilon) = A^{(k)} + B^{(k)}(\epsilon)$ is a Markov generator with no coupling through transient states, $\left\|B^{(k)}\right\| = O(\epsilon)$ and (b) $G^{(k)}(\epsilon)$ has well-defined time scale behavior. Again applying Lemma 2.1 and stating the result in the same form as in (2.56), we obtain the following uniform approximation:

$$e^{G^{(k)}(\epsilon)t} = e^{A^{(k)}t} + U^{(k)} e^{\epsilon G^{(k+1)}(\epsilon)t} V^{(k)} - U^{(k)}V^{(k)} + O(\epsilon) \tag{2.70}$$

where

$$
\begin{aligned}
G^{(k+1)}(\epsilon) &= A^{(k+1)}(\epsilon) + W^{(k+1)}(\epsilon) && (2.71)\\
A^{(k+1)}(\epsilon) &= \frac{1}{\epsilon} V^{(k)} A^{(k)}(\epsilon) U^{(k)} && (2.72)\\
&= \frac{1}{\epsilon} V^{(k)} B^{(k)}(\epsilon) U^{(k)} && (2.73)
\end{aligned}
$$

and

$$
\begin{aligned}
W^{(k+1)}(\epsilon) &= W_1^{(k+1)}(\epsilon) + W_2^{(k+1)}(\epsilon) && (2.74)\\
W_1^{(k+1)} &= \frac{1}{\epsilon} V^{(k)} W^{(k)}(\epsilon) U^{(k)} && (2.75)\\
W_2^{(k+1)} &= -\frac{1}{\epsilon} V^{(k)} \left(B^{(k)}(\epsilon) + W^{(k)}(\epsilon)\right) Y^{(k)} \left(Z^{(k)} G^{(k)}(\epsilon) Y^{(k)}\right)^{-1} \cdot && (2.76)\\
&\qquad Z^{(k)} \left(B^{(k)}(\epsilon) + W^{(k)}(\epsilon)\right) U^{(k)}
\end{aligned}
$$

Note that for $k = 2, 3 \ldots$ the term $W^{(k)}(\epsilon)$ consists of two parts, namely the "projection" $W_1^{(k+1)}(\epsilon)$ of the preceding weak term $W^{(k)}(\epsilon)$, and a new term $W_2^{(k+1)}(\epsilon)$ defined similarly to the weak term computed previously using (2.67). We know from Lemma 2.1 that $G^{(k+1)}(\epsilon)$ has well-behaved time scales and by construction that $A^{(k+1)}(\epsilon)$ is a Markov generator. By assumption in this section, there is no coupling through transient states in $A^{(k+1)}(\epsilon)$. Thus in order to continue the recursive procedure, we need to verify the following:

Lemma 2.3 *Suppose that $G^{(k)}(\epsilon) = A^{(k)}(\epsilon) + W^{(k)}(\epsilon)$ satisfies the following*

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

Then $W^{(k+1)}(\epsilon)$ defined in (2.74) is weak with respect to $A^{(k+1)}(\epsilon)$ in (2.72)
Proof *see Section 2.A.2* □

By applying (2.56) followed by repeated use of (2.70) and finally discarding the weak terms at the last time scale (since at this point they clearly only represent a regular perturbation), the following sequence of approximations is constructed for a system exhibiting k time scales and with no coupling through transient states at any intermediate time scale:

$$\begin{aligned}
e^{A^{(0)}(\epsilon)t} &= e^{A^{(0)}t} + U^{(0)}e^{\epsilon G^{(1)}(\epsilon)t}V^{(0)} - U^{(0)}V^{(0)} + O(\epsilon)\\
e^{G^{(1)}(\epsilon)t} &= e^{A^{(1)}t} + U^{(1)}e^{\epsilon G^{(2)}(\epsilon)t}V^{(1)} - U^{(1)}V^{(1)} + O(\epsilon)\\
&\vdots \\
e^{G^{(k-2)}(\epsilon)t} &= e^{A^{(k-2)}t} + U^{(k-2)}e^{\epsilon G^{(k-1)}(\epsilon)t}V^{(k-2)} - U^{(k-2)}V^{(k-2)} + O(\epsilon)\\
e^{G^{(k-1)}(\epsilon)t} &= e^{A^{(k-1)}t} + O(\epsilon)
\end{aligned} \tag{2.77}$$

Note that determining when to stop the procedure is not a problem. Specifically, since $A^{(0)}(\epsilon)$ has one ergodic class for $\epsilon > 0$, we stop when $A^{(k-1)}$ has exactly one ergodic class. From Coderch [10], we know that since $A^{(0)}(\epsilon)$ does have well-defined time scale behavior, there is a k such that this is true, and this k is associated with the slowest time scale.

Collapsing the sum (2.77) we obtain the following result.

Theorem 2.4 *Suppose $A^{(0)}(\epsilon)$ exhibits k time scales of behavior and that there is no coupling through transient states in $A^{(j)}(\epsilon)$ for $j = 0, 1, \ldots, k-2$. Then*

$$\begin{aligned}
e^{A^{(0)}(\epsilon)t} = e^{A^{(0)}t} &+ \\
&\left(U^{(0)}e^{A^{(1)}\epsilon t}V^{(0)} - U^{(0)}V^{(0)}\right) + \\
&\left(U^{(0)}U^{(1)}e^{A^{(2)}\epsilon^2 t}V^{(1)}V^{(0)} - U^{(0)}U^{(1)}V^{(1)}V^{(0)}\right) + \\
&\vdots \\
&\Big(U^{(0)}\ldots U^{(k-2)}e^{A^{(k-1)}\epsilon^{k-1}t}V^{(k-2)}\ldots V^{(0)} - \\
&\qquad U^{(0)}\ldots U^{(k-2)}V^{(k-2)}\ldots V^{(0)}\Big) + O(\epsilon)
\end{aligned} \tag{2.78}$$

where

$$V^{(k)} \equiv \tilde{V}^{(k)}(0) = V^{(k)}(0) \tag{2.79}$$

Proof *This result follows from Lemmas 2.1 through 2.3 and the derivation above. The final approximation (2.78) is obtained by collapsing the sums in (2.77).* □

Note that in order to construct this approximation, we never need to calculate $Y^{(k)}$, $Z^{(k)}$, or any of the terms $W^{(k)}(\epsilon)$. Rather, at each time scale we begin with $A^{(k)}(\epsilon) = A^{(k)} + B^{(k)}(\epsilon)$, compute the ergodic classes and probabilities associated with $A^{(k)}$ and from these form $U^{(k)}$ and $V^{(k)}$. $A^{(k+1)}(\epsilon)$ is then calculated using (2.72). At this point, of course, we have only dealt with the case in which there is no coupling through transient states at any stage of the procedure. In the following subsection, we modify the procedure in order to remove this restriction.

Before we continue, however, let us briefly interpret the approximation (2.78). Specifically, the (i, j) element of the left-hand side of the equation is the probability that the Markov process is in state i at time t conditioned on its being in state j at time 0. If there is a sequence of order 1 rate transitions from state j to state i, then initially the behavior of this probability is captured by the first term. As t grows and this first term becomes constant ($= U^{(0)}V^{(0)}$), the probability is determined by the probability of being in the aggregate class I to which i belongs and by the relative probability of being in state i conditioned on being in the aggregate I. This behavior is determined by the generator $A^{(1)}(\epsilon)$. If there were no sequence of order 1 rate transitions in $A^{(0)}(\epsilon)$ linking state j to state i, then the probability is initially zero and the first term does not contribute. Therefore, for any element (i, j), only a subsequence of generators in (2.78) $A^{(r)}, A^{(r+1)}, \ldots, A^{(k)}$ enter into the sum. This index r can be thought of as the degree of the coupling from state j to state i. Furthermore, if a uniform approximation is required on an interval $t \in [0, T/\epsilon^s], T < \infty$, then the bracketed terms involving $A^{(s+1)}, A^{(s+2)}, \ldots, A^{(k)}$ are all O(ϵ) and can be ignored.

Finally, we note that (2.78) can be interpreted as specifying an asymptotic time scale decomposition of $A^{(0)}(\epsilon)$ in the sense given in Definition 2.3. Specifically, let

$$T^{(i)} = \begin{bmatrix} Z^{(i)} \\ V^{(i)} \end{bmatrix}, \quad T^{(i)^{-1}} = \begin{bmatrix} Y^{(i)} \; U^{(i)} \end{bmatrix} \tag{2.80}$$

The similarity transformation S can be constructed from the $T^{(i)}$.

$$S = \begin{bmatrix} I & 0 \\ 0 & T^{(k-1)} \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & T^{(k-2)} \end{bmatrix} \cdots \begin{bmatrix} I & 0 \\ 0 & T^{(1)} \end{bmatrix} \begin{bmatrix} T^{(0)} \end{bmatrix} \tag{2.81}$$

$$= \begin{bmatrix} Z^{(0)} \\ Z^{(1)}V^{(0)} \\ \vdots \\ Z^{(k-1)}V^{(k-2)}\ldots V^{(0)} \\ V^{(k-1)}V^{(k-2)}\ldots V^{(0)} \end{bmatrix} \tag{2.82}$$

$$S^{-1} = \begin{bmatrix} Y^{(0)} & U^{(0)}Y^{(1)} & \ldots & U^{(0)}\ldots U^{(k-2)}Y^{(k-1)} & U^{(0)}\ldots U^{(k-1)} \end{bmatrix} \tag{2.83}$$

2.3.2 Transient states which couple aggregates

As indicated previously, the basic approach to this general case is to reduce it to the one considered in the previous subsection. This is accomplished by constructing a new Markov process on a larger state space such that (a) the behavior of the original process can be recovered easily; and (b) the new process has the property of no coupling through transient states (Definition 2.1). Consider again a Markov generator of the form

$$A(\epsilon) = A + B(\epsilon)\ , \ \ \|B(\epsilon)\| = O(\epsilon) \tag{2.84}$$

where A generates N ergodic classes. The state space can be partitioned into $N+1$ parts E_1, $E_2, \ldots, E_N$, T where the E_K are the ergodic classes and T is the set of transient states. The basis of the construction is the observation that the set T can be "split" into N copies T_1, $T_2, \ldots, T_N$ such that each copy is associated with a unique ergodic class. The associated Markov generator $\hat{A}(\epsilon) = \hat{A} + \hat{B}(\epsilon)$ is constructed on this expanded state space such that once in a state $s \in T_K$, the next state entered that belongs to $E \equiv E_1 \cup E_2 \cup \cdots \cup E_N$ must be in E_K with probability one. By construction then, $\hat{A}(\epsilon)$ satisfies the condition in Definition 2.1. The precise nature of this construction can be stated here as a lemma. The proof of this lemma, which also gives the details of the construction, appears in the appendix. An example of this construction is provided below.

Lemma 2.5 *Let $A(\epsilon) = A + B(\epsilon)$ and let U and V be the ergodic probability and class matrices derived from the unperturbed generator A. Then there exist C, $D(\epsilon)$, $\hat{A}(\epsilon) = \hat{A} + \hat{B}(\epsilon)$ and $\hat{U}$ and $\hat{V}$ derived from $\hat{A}$ such that*

1. $e^{A(\epsilon)t} = Ce^{\hat{A}(\epsilon)t}D(\epsilon)$,
2. $\hat{A}(\epsilon)$ *does not exhibit coupling through transient states,*
3. $C\hat{U} = U$,
4. $\hat{V}D(0) = V$,
5. $D(\epsilon)U = D(0)U = \hat{U}$,
6. $C\hat{A}(\epsilon)\hat{U} = A(\epsilon)U$, *and*
7. $\mathrm{Range}(D(\epsilon))$ *is* $\hat{A}(\epsilon)$*-invariant.*

Proof *see Section 2.A.3.* □

The construction of $\hat{A}(\epsilon)$ can be described as follows. Let i, k be elements of E (i.e. recurrent states of A and $\hat{A}$). The probability transition rate from state i to state k in $\hat{A}(\epsilon)$ is then the same as that in $A(\epsilon)$. Next let $j \in T$, and let $j_1, \ldots, j_N$ denote the corresponding copies of j in the expanded process. The basic idea behind the construction is that a transition to the state j_I corresponds to a transition in the original process to state j *together with* the decision that the next ergodic class that will be entered is E_I. Consequently, the transition rates into the j_I must reflect the probability of this additional decision. Specifically, if $k \in E$, then

$$\hat{a}_{j_I k}(\epsilon) = a_{jk}(\epsilon)\, v_{Ij}(\epsilon) \tag{2.85}$$

where $v_{Ij}(\epsilon)$, defined in (2.16), is precisely the probability of that decision. Similarly, transitions out of j_I must be adjusted to reflect conditioning on knowledge of which ergodic class will be visited next. Specifically the transition rate from j_I to any state in an ergodic class other than E_I is 0, as is the rate from j to any state in T_K, $K \neq I$, i.e. to any copy of any transient state corresponding to a subsequent transition into a different ergodic class. The remaining transition rates out of j_I can be computed using Bayes' Rule and are specified as follows

$$\hat{a}_{ij_I}(\epsilon) = a_{ij}(\epsilon)\,\frac{1}{v_{Ij}(\epsilon)} \quad i \in E_I \tag{2.86}$$

$$\hat{a}_{k_I j_I}(\epsilon) = a_{kj}(\epsilon)\,\frac{v_{Ik}(\epsilon)}{v_{Ij}(\epsilon)} \quad k_I \in T_I \tag{2.87}$$

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

$$d_{j_Ij}(\epsilon) = v_{Ij}(\epsilon) \tag{2.88}$$

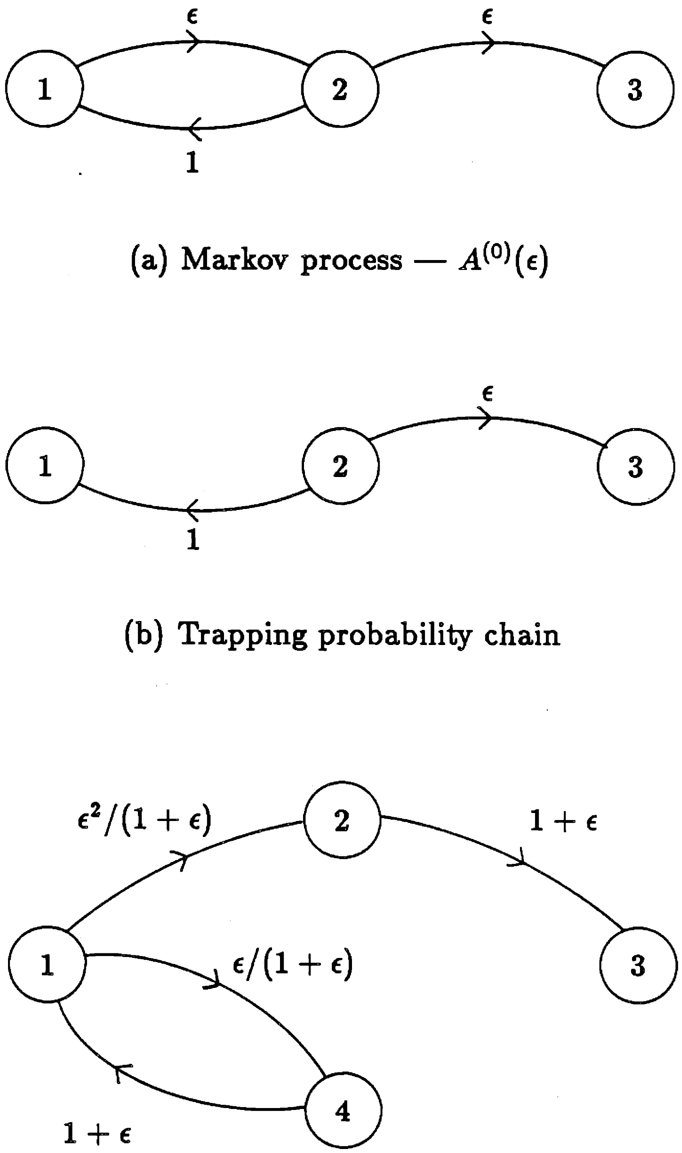
with all other elements of $D(\epsilon)$ equal to 0. The several properties in Lemma 2.5 then follow directly from the construction as shown in Section 2.A.3.

Example 2.2 *We consider the state expansion of the simple process depicted in Figure 2.4(a), for which*

$$A^{(0)}(\epsilon) = \begin{bmatrix} -\epsilon & 1 & 0 \\ \epsilon & -1-\epsilon & 0 \\ 0 & \epsilon & 0 \end{bmatrix} \tag{2.89}$$

In this case the construction in Lemma 2.5 calls for a splitting of the transient state 2. Following the procedure cited in Lemma 2.5, the key quantities are the probabilities that the perturbed process first enters each of the unperturbed ergodic classes (namely $E_1 = \{1\}$ and $E_2 = \{3\}$) given that it starts in any particular transient state. As discussed previously in Section 2.2, these can be computed as the limiting probabilities of the process illustrated in Figure 2.4(b), obtained from the chain in Figure 2.4(a) by making each unperturbed recurrent class a trapping state. The expanded state process is depicted in Figure 2.4(c) and the associated matrices are

$$\hat{C} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \quad \hat{D}(\epsilon) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{\epsilon}{1+\epsilon} & 0 \\ 0 & 0 & 1 \\ 0 & \frac{1}{1+\epsilon} & 0 \end{bmatrix} \tag{2.90}$$



(a) Markov process — $A^{(0)}(\epsilon)$

(b) Trapping probability chain

(c) Expanded Markov process — $\hat{A}^{(0)}(\epsilon)$

Figure 2.4: Markov process in Example 2.2

and

$$\hat{U} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad \hat{V} = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix} \tag{2.91}$$

Note that, as desired, states 2 and 4 in Figure 2.4(c) are transient but do not couple the ergodic classes {1} and {3}. Consequently, the derivation in Section 2.3.1 can be directly applied. □

Using properties in Lemma 2.5, the original system can be approximated as in (2.56) by considering the associated generator on this expanded state space

$$\begin{aligned} e^{A(\epsilon)t} &= Ce^{\hat{A}(\epsilon)t}D(\epsilon) && (2.92)\\ &= C\left(e^{\hat{A}t} + \hat{U}e^{\epsilon G(\epsilon)t}\hat{V} - \hat{U}\hat{V}\right)D(\epsilon) + O(\epsilon) && (2.93)\\ &= e^{At} + Ue^{\epsilon G(\epsilon)t}V(\epsilon) - UV(\epsilon) + O(\epsilon) && (2.94)\\ &= e^{At} + Ue^{\epsilon G(\epsilon)t}V - UV + O(\epsilon) && (2.95) \end{aligned}$$

where

$$V(\epsilon) \equiv \hat{V}D(\epsilon) \tag{2.96}$$

and $G(\epsilon)$ is computed from $\hat{A}(\epsilon)$.

Based on (2.95), we see how to modify the development in the preceding subsection by expanding the state space at each stage of the procedure. In order to verify that this results in a uniform approximation, we need to show the following result.

Lemma 2.6 *Suppose $G(\epsilon) = A(\epsilon) + W(\epsilon)$ where $A(\epsilon)$ is a Markov generator and $W(\epsilon)$ is weak with respect to $A(\epsilon)$. Let C, $D(\epsilon)$ and $\hat{A}(\epsilon)$ be determined as in Lemma 2.5. Then $\hat{G}(\epsilon) = \hat{A}(\epsilon) + \hat{W}(\epsilon)$ can be constructed such that*

1. $e^{G(\epsilon)t} = Ce^{\hat{G}(\epsilon)t}D(\epsilon) + O(\epsilon)$, *and*
2. $\hat{W}(\epsilon)$ *is weak with respect to* $\hat{A}(\epsilon)$

Proof *see Section 2.A.4* □

In a similar manner to (2.77), the following sequence of approximations can be constructed. $\hat{G}^{(k)}(\epsilon)$ is computed from $G^{(k)}(\epsilon)$ by expanding the state space using Lemma 2.6. $G^{(k+1)}(\epsilon)$ is then computed from the expanded $\hat{G}^{(k)}(\epsilon)$ using the construction in Lemma 2.3. By Lemmas 2.3 and 2.6,

$$G^{(k)}(\epsilon) = A^{(k)}(\epsilon) + W^{(k)}(\epsilon) \tag{2.97}$$

where $W^{(k)}(\epsilon)$ is weak with respect to $A^{(k)}(\epsilon)$. Therefore

$$\begin{aligned}
\mathrm{e}^{A^{(0)}(\epsilon)t} &= C^{(0)}\mathrm{e}^{\hat{A}^{(0)}(\epsilon)t}D^{(0)}(\epsilon) \\
\mathrm{e}^{\hat{A}^{(0)}(\epsilon)t} &= \mathrm{e}^{\hat{A}^{(0)}t} + \hat{U}^{(0)}\mathrm{e}^{\epsilon G^{(1)}(\epsilon)t}\hat{V}^{(0)} - \hat{U}^{(0)}\hat{V}^{(0)} + \mathrm{O}(\epsilon) \\
&\vdots \\
\mathrm{e}^{G^{(k-2)}(\epsilon)t} &= C^{(k-2)}\mathrm{e}^{\hat{G}^{(k-2)}(\epsilon)t}D^{(k-2)}(\epsilon) \\
\mathrm{e}^{\hat{G}^{(k-2)}(\epsilon)t} &= \mathrm{e}^{\hat{A}^{(k-2)}t} + \hat{U}^{(k-2)}\mathrm{e}^{\epsilon G^{(k-1)}(\epsilon)t}\hat{V}^{(k-2)} - \hat{U}^{(k-2)}\hat{V}^{(k-2)} + \mathrm{O}(\epsilon) \\
\mathrm{e}^{G^{(k-1)}(\epsilon)t} &= C^{(k-1)}\mathrm{e}^{\hat{G}^{(k-1)}(\epsilon)t}D^{(k-1)}(\epsilon) \\
\mathrm{e}^{\hat{G}^{(k-1)}(\epsilon)t} &= \mathrm{e}^{\hat{A}^{(k-1)}t} + \mathrm{O}(\epsilon)
\end{aligned} \tag{2.98}$$

This sequence of approximations can then be collapsed to form one overall approximation using the properties in Lemma 2.5. To illustrate this, consider combining the expression for $\exp(G^{(k)}(\epsilon)t)$ and $\exp(\hat{G}^{(k)}(\epsilon)t)$.

$$\begin{aligned}
\mathrm{e}^{G^{(k)}(\epsilon)t} &= C^{(k)}\left(\mathrm{e}^{\hat{A}^{(k)}t} + \hat{U}^{(k)}\mathrm{e}^{\epsilon\hat{G}^{(k+1)}(\epsilon)t}\hat{V}^{(k)}\right)D^{(k)}(\epsilon) + \mathrm{O}(\epsilon) && (2.99) \\
&= C^{(k)}\mathrm{e}^{\hat{A}^{(k)}t}D^{(k)}(\epsilon) + \\
&\qquad C^{(k)}\hat{U}^{(k)}\mathrm{e}^{\epsilon\hat{G}^{(k+1)}(\epsilon)t}\hat{V}^{(k)}D^{(k)}(\epsilon) + \mathrm{O}(\epsilon) && (2.100) \\
&= C^{(k)}\mathrm{e}^{A^{(k)}t}D^{(k)}(0) + \\
&\qquad C^{(k)}\hat{U}^{(k)}\mathrm{e}^{\epsilon\hat{G}^{(k+1)}(\epsilon)t}\hat{V}^{(k)}D^{(k)}(\epsilon) + \mathrm{O}(\epsilon) && (2.101) \\
&= \mathrm{e}^{A^{(k)}t} + U^{(k)}\mathrm{e}^{\epsilon\hat{G}^{(k+1)}(\epsilon)t}V^{(k)} + \mathrm{O}(\epsilon) && (2.102)
\end{aligned}$$

Combining (2.98) and (2.102) yields the following result.

Theorem 2.7 *Let $A^{(0)}(\epsilon)$ be a perturbed Markov generator with one ergodic class for $\epsilon > 0$. Then the approximation form (2.78) in Theorem 2.4 is valid where $A^{(k+1)}(\epsilon)$ is constructed from $A^{(k)}(\epsilon)$ by first forming the expanded generator $\hat{A}^{(k)}(\epsilon)$.*

Proof *This result follows directly from Lemmas 2.5 and 2.6 and the construction outlined above.* □

Let us make several comments on this result. The first is that at this point we can conclude in general that weak terms cannot affect the asymptotic approximation. More precisely, we now have the following result.

Corollary 2.8 *Let $A(\epsilon)$ be a Markov generator and $G(\epsilon)$ a matrix so that $G(\epsilon) = A(\epsilon) + W(\epsilon)$ where $W(\epsilon)$ is weak with respect to $A(\epsilon)$. Then $G(\epsilon)$ is asymptotically equivalent to $A(\epsilon)$.* □

Corollary 2.8 has the useful consequence that if one is trying to construct an approximation of a Markov process with a generator $A(\epsilon)$ which can be separated into a simpler generator $\tilde{A}(\epsilon)$ and a relatively weak part $\tilde{W}(\epsilon)$, then the weak part can safely be "pruned". A direct application of this is that only the leading order terms in ϵ of any transition rate need to be considered in the construction of the approximation.

Also as a result of Theorem 2.7 and Corollary 2.8 the behavior of a Markov process generated by $A^{(0)}(\epsilon)$ can be approximated using $A^{(0)}(0)$ and the reduced order perturbed generator $A^{(1)}(\epsilon)$ regardless of the weak term $W^{(1)}(\epsilon)$.

$$e^{A^{(0)}(\epsilon)t} = e^{A^{(0)}(0)t} + U^{(0)}e^{\epsilon A^{(1)}(\epsilon)t}V^{(0)} - U^{(0)}V^{(0)} + O(\epsilon) \qquad (2.103)$$

This result follows from the fact that all the weak terms can be discarded when the overall approximation is constructed.

Let us now consider the computation required to construct the asymptotic approximation in Theorem 2.7. Beginning with $A^{(k)}(\epsilon) = A^{(k)} + B^{(k)}(\epsilon)$ we directly compute $U^{(k)}$ and $V^{(k)}$. To continue, we must apparently expand the state space to construct $\hat{A}^{(k)}(\epsilon)$ and $D^{(k)}(\epsilon)$. The matrix $D^{(k)}(\epsilon)$ is needed to construct $V^{(k)}D^{(k)}(\epsilon)$, while $\hat{A}^{(k)}(\epsilon)$ is needed so that we can perform the aggregation required to continue to the next stage. That is,

$$A^{(k+1)}(\epsilon) = \frac{1}{\epsilon}\hat{V}^{(k)}\hat{A}^{(k)}(\epsilon)\hat{U}^{(k)} \qquad (2.104)$$

Note that in this process we first expand the state space to allow use of the previously derived procedure and then aggregate in order to move on to the next time scale. We can collapse these two steps and avoid constructing the expanded state space process completely.

Lemma 2.9 *The two step procedure for computing $\hat{A}^{(k)}(\epsilon)$ from $A^{(k)}(\epsilon)$ by Lemma 2.5 and $A^{(k+1)}(\epsilon)$ from $\hat{A}^{(k)}(\epsilon)$ by (2.104) is equivalent to*

$$\begin{aligned} A^{(k+1)}(\epsilon) &= \frac{1}{\epsilon}V^{(k)}(\epsilon)A^{(k)}(\epsilon)U^{(k)} && (2.105) \\ &= \frac{1}{\epsilon}V^{(k)}(\epsilon)B^{(k)}(\epsilon)U^{(k)} && (2.106) \end{aligned}$$

where

$$V^{(k)}(\epsilon) \equiv \hat{V}^{(k)}D^{(k)}(\epsilon) \qquad (2.107)$$

Furthermore, $V^{(k)}(\epsilon)$ is precisely the expression computed in Algorithm 2.1 (2.16).

Proof *see Section 2.A.5* □

The final step to show the general validity of Algorithm 2.1 is to show that using $\tilde{V}^{(k)}(\epsilon)$ instead of the exact $V^{(k)}(\epsilon)$ introduces only a weak term and therefore by Corollary 2.8 does not affect the asymptotic approximation.

2.4 Example

In this section, the decomposition algorithm specified above is applied to the multiple time scale example considered by Coderch [9]. This Markov process is illustrated in Figure 2.5. The generator of this process is

$$A^{(0)}(\epsilon) = \begin{bmatrix} 0 & \epsilon & 0 & 0 & 0 & 0 & 0 \\ 0 & -2-\epsilon & \epsilon & 0 & 0 & 0 & 0 \\ 0 & 0 & -\epsilon & \epsilon & 0 & \epsilon & 0 \\ 0 & 1 & 0 & -1-\epsilon & \epsilon & 0 & 0 \\ 0 & 0 & 0 & 1 & -\epsilon & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1-\epsilon & \epsilon \\ 0 & 0 & 0 & 0 & 0 & 1 & -\epsilon \end{bmatrix} \qquad (2.108)$$

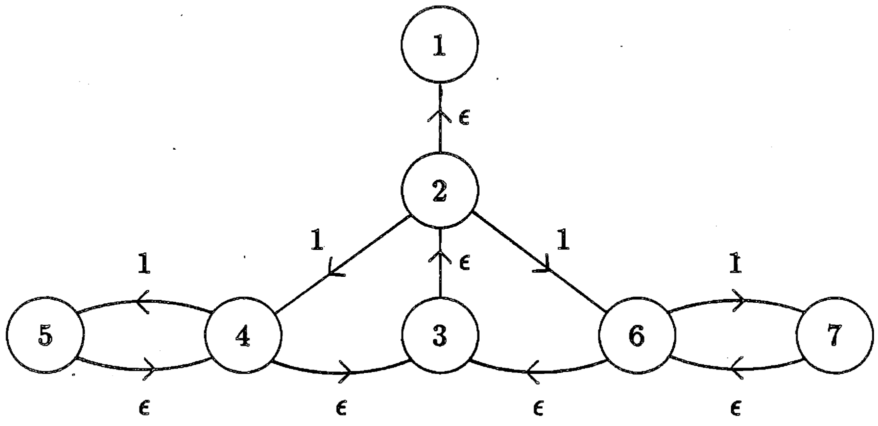


Figure 2.5: Perturbed Markov process

At $\epsilon = 0$, the ergodic classes are $\{1\}$, $\{3\}$, $\{5\}$, and $\{7\}$. All other states are transient. The ergodic probabilities of the classes are therefore given as

$$U^{(0)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{2.109}$$

$\tilde{V}^{(0)}(\epsilon)$ can be computed from the limiting probabilities of the chain shown in Figure 2.6. This yields

$$\tilde{V}^{(0)}(\epsilon) = \begin{bmatrix} 1 & \frac{\epsilon}{2} & 0 & 0 & 0 & 0 & 0 \\ 0 & \epsilon & 1 & \epsilon & 0 & \epsilon & 0 \\ 0 & \frac{1}{2} - \frac{3\epsilon}{4} & 0 & 1-\epsilon & 1 & 0 & 0 \\ 0 & \frac{1}{2} - \frac{3\epsilon}{4} & 1 & 0 & 0 & 1-\epsilon & 1 \end{bmatrix} \tag{2.110}$$

giving

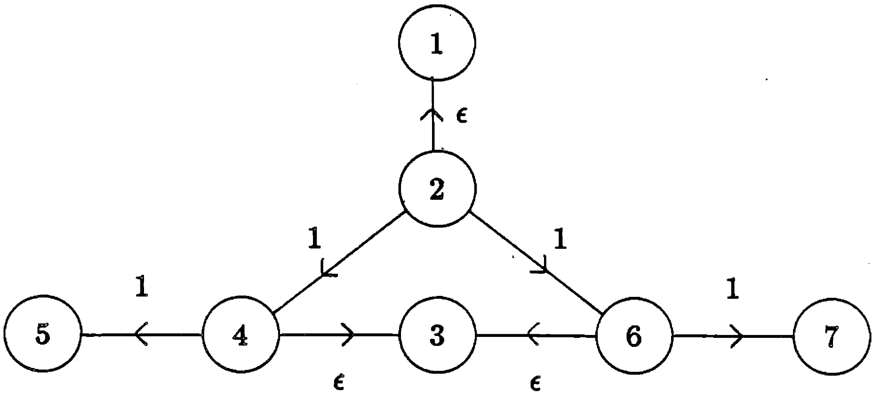


Figure 2.6: Limiting probability process

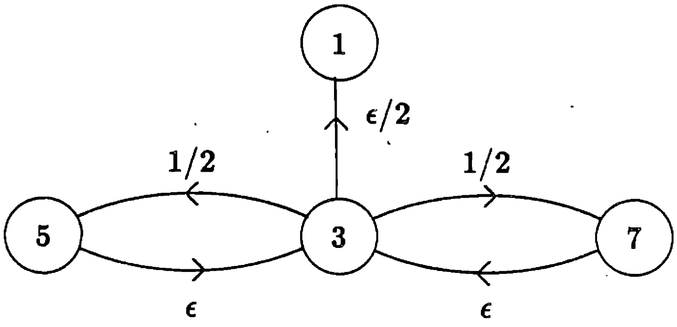


Figure 2.7: O(1/ε) time scale process

$$A^{(1)}(\epsilon) = \frac{1}{\epsilon}\tilde{V}^{(0)}(\epsilon)A^{(0)}U^{(0)} = \begin{bmatrix} 0 & \frac{\epsilon}{2} & 0 & 0 \\ 0 & -1-\epsilon & \epsilon & \epsilon \\ 0 & \frac{1}{2}-\frac{3\epsilon}{4} & -\epsilon & 0 \\ 0 & \frac{1}{2}-\frac{3\epsilon}{4} & 0 & -\epsilon \end{bmatrix} \qquad (2.111)$$

The process generated by $A^{(1)}(\epsilon)$ is shown in Figure 2.7. The recurrent classes are $\{1\}$, $\{5\}$, and $\{7\}$ and state 3 is transient. Repeating the procedure to compute

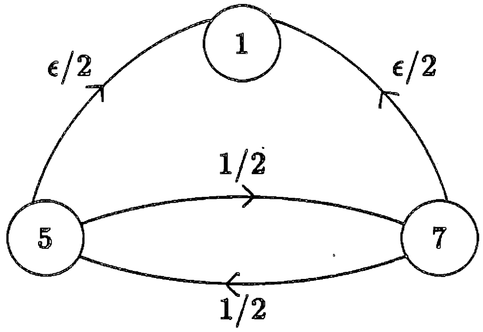


Figure 2.8: $O(1/\epsilon^2)$ time scale process

$U^{(1)}$ and $\tilde{V}^{(1)}(\epsilon)$ gives

$$U^{(1)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{2.112}$$

and

$$\tilde{V}^{(1)}(\epsilon) = \begin{bmatrix} 1 & \frac{\epsilon}{2} & 0 & 0 \\ 0 & \frac{1}{2} - \frac{\epsilon}{4} & 1 & 0 \\ 0 & \frac{1}{2} - \frac{\epsilon}{4} & 0 & 1 \end{bmatrix} \tag{2.113}$$

which yields

$$A^{(2)}(\epsilon) = \begin{bmatrix} 0 & \frac{\epsilon}{2} & \frac{\epsilon}{2} \\ 0 & -\frac{1}{2} - \frac{\epsilon}{2} & \frac{1}{2} \\ 0 & \frac{1}{2} & -\frac{1}{2} - \frac{\epsilon}{2} \end{bmatrix} \tag{2.114}$$

which is illustrated in Figure 2.8. The unperturbed process at this time scale has no transient states. Therefore we can calculate

$$A^{(3)}(\epsilon) = \frac{1}{\epsilon} V^{(2)} A^{(2)}(\epsilon) U^{(2)} = \begin{bmatrix} 0 & \frac{1}{2} \\ 0 & -\frac{1}{2} \end{bmatrix} \tag{2.115}$$

The overall approximation of the process generated by $A^{(0)}(\epsilon)$ is therefore constructed from the four unperturbed chains shown in Figure 2.9.

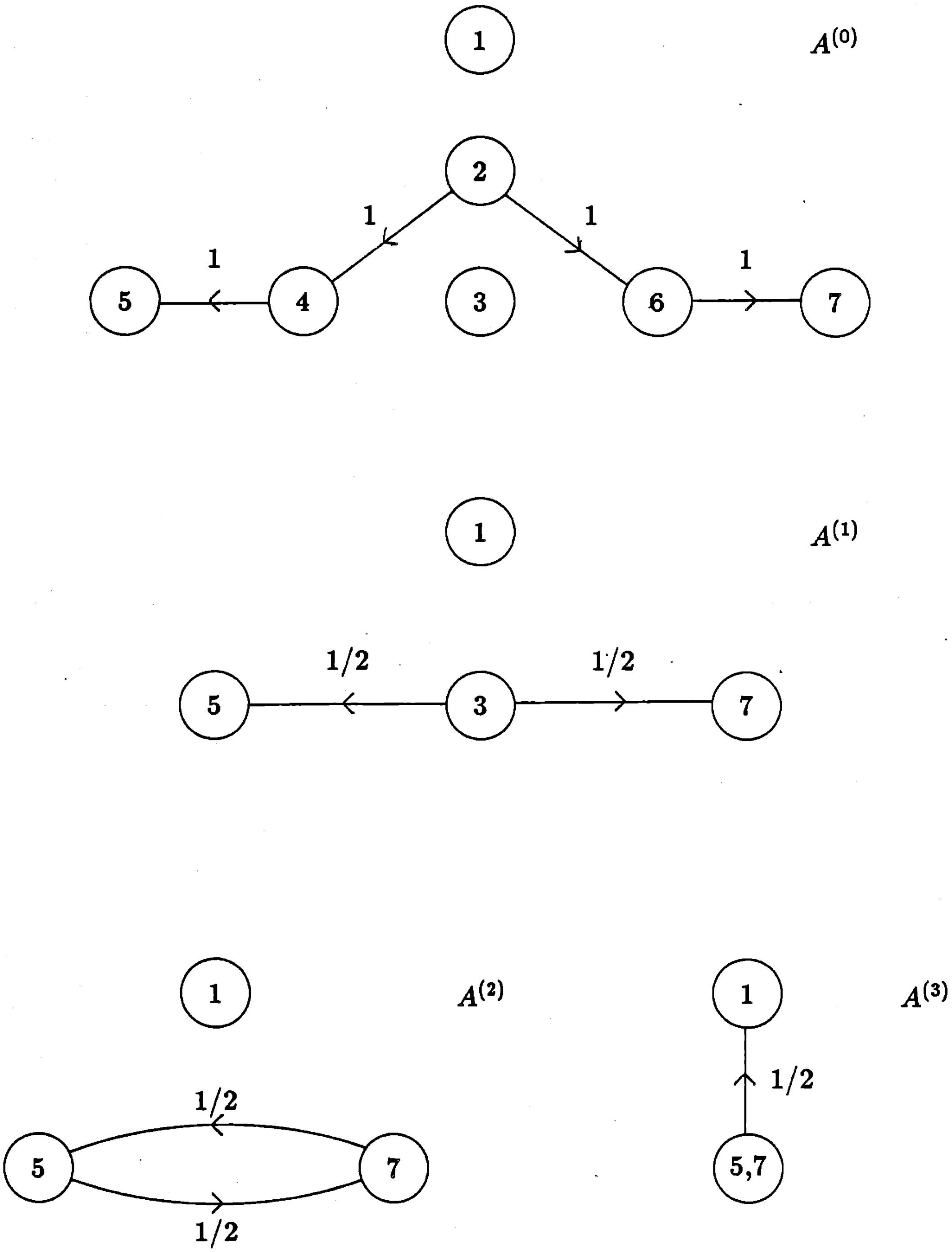


Figure 2.9: Multiple Time Scale Chains

2.5 Discussion

Several aspects of the algorithm presented in this chapter are discussed in this section. First, as was outlined in Section 2.1, there are close ties between the eigenprojection interpretations of this algorithm and those of Coderch [9] and Courtois [12]. Specifically, at each time scale, Coderch computes the *exact* terms in a Taylor expansion of the eigenprojection, $P^{(i)}(\epsilon)$, of the 0-group of eigenvalues of the generator $A^{(i)}(\epsilon)$. This is used to compute the next generator $A^{(i+1)}(\epsilon)$. By restricting the nature of the ergodic structure of $A^{(i)}(0)$ that is allowed, Courtois effectively uses the unperturbed eigenprojection $P^{(i)}(0)$ in a similar manner. The algorithm presented in this chapter has the interpretation that a projection $\tilde{P}^{(i)}(\epsilon) \equiv U^{(i)}\tilde{V}^{(i)}(\epsilon)$ is constructed which ignores certain "unimportant" terms in $P^{(i)}(\epsilon)$. One question which remains is whether a direct analysis of the terms in these eigenprojection could reduce the amount of computation still further.

Another aspect of this algorithm is that it explicitly results in a reduction in the size of the system at each time scale. Although Courtois deals with successively smaller systems at each time scale at the expense of restricting the class of systems which can be analyzed, the general algorithms presented by Coderch [9] and Delebecque [16] both manipulate systems of the original dimension. The interpretation of aggregation is introduced after all the time scales have been identified. Therefore, although at each successive time scale the rank of the generator is reduced, no advantage is taken of this fact. The algorithm presented in this chapter allows us to work with successively smaller systems as slower and slower time scales are uncovered. Hopefully, this will allow application to larger problems than was possible using previous results.

A third feature of this algorithm, which will be dealt with in Chapter 6, is that there is a very simple graphical/connectivity interpretation which is available. For example, if only the locations of the nonzero entries in the slow time scale generators are required, which effectively provides the "structure" of those systems, several simplifications are available. The identification of the ergodic classes amount to finding communicating classes in the state transition graph where only the $O(1)$ links are retained. Determining the orders of the entries of the $\tilde{V}(\epsilon)$ amounts to a shortest path problem where the length of the path is the product of the rates on each link. This can also be performed as an additive shortest path problem where the links are weighted by integer orders of the transition rates.

A fourth aspect of this algorithm which should be stressed is that is provides a simple method of approximating the *dynamics* of a Markov process. There has been much work in the area of approximating the steady state probabilities of nearly completely decomposable Markov systems (see [7] for example). Although some of these algorithms provide better approximations of the steady state through iterative refinement methods which provide an $O(\epsilon^2)$ error, they do not address the dynamic behavior of their systems. Therefore, an area in which the algorithm developed in this chapter might be applicable is in approximating the behavior of a system when it is not near its steady state. Effectively, this algorithm provides a method for approximating *transient* behavior of complex Markov processes.

Finally, it must be pointed out that this algorithm and its development rely on the state space of the Markov process being finite. Although this is satisfactory in many engineering applications, it would be useful to extend the results to denumerable state space systems. Although the proofs presented here do not extend easily to such systems, the results seem to be applicable nevertheless. One area where such an extension would be useful is in the application of these results to the analysis of semi-Markov processes as is presented in Chapter 3. Since these results apply only to finite state Markov systems, the semi-Markov results which can be based on them can only deal with systems which have holding time distributions with rational polynomial transforms, and therefore have a finite state Markov representation. If denumerable state processes could be dealt with, the class of semi-Markov processes which could be approximated could be extended.

2.A Appendix

2.A.1 Proof of Lemma 2.2

To simplify notation in this section, we drop the superscript "$^{(0)}$" for the various quantities defined in the text. To begin, note that the term

$$(ZA(\epsilon)Y)^{-1} \tag{2.116}$$

in the expression (2.67) for $W^{(1)}(\epsilon)$ in Lemma 2.2 can be expanded as a Taylor series in ϵ. Define

$$D \equiv ZAY \tag{2.117}$$

This matrix is block-diagonal and invertible.[10] Then

$$\begin{aligned}
(ZA(\epsilon)Y)^{-1} &= (Z\,(A+B(\epsilon))\,Y)^{-1} && (2.118)\\
&= \left(I + D^{-1}ZB(\epsilon)Y\right)^{-1} D^{-1} && (2.119)\\
&= \sum_{m=0}^{\infty} \left(-D^{-1}ZB(\epsilon)Y\right)^{m} D^{-1} && (2.120)
\end{aligned}$$

where (2.120) is valid for ϵ sufficiently small since $\|B(\epsilon)\| = O(\epsilon)$. Substituting (2.120) into (2.67) gives

$$\epsilon W^{(1)}(\epsilon) = \epsilon C_1(\epsilon) + \epsilon C_2(\epsilon) + \cdots \tag{2.121}$$

where

$$\left.\begin{aligned}
\epsilon C_1(\epsilon) &= -VB(\epsilon)SB(\epsilon)U\\
\epsilon C_2(\epsilon) &= +VB(\epsilon)SB(\epsilon)SB(\epsilon)U\\
&\vdots\\
\epsilon C_m(\epsilon) &= (-1)^m VB(\epsilon)\,(SB(\epsilon))^m\, U
\end{aligned}\right. \tag{2.122}$$

and S is block diagonal

$$\begin{aligned}
S &= YD^{-1}Z && (2.123)\\
&= \mathrm{diag}(S_1, \ldots, S_N) && (2.124)
\end{aligned}$$

[10] Note that if the nonzero eigenvalues of A are distinct, then D is indeed diagonal and not simply block diagonal.

For the purpose of comparison, recall that by (2.63)

$$\epsilon A^{(1)}(\epsilon) = VB(\epsilon)U \tag{2.125}$$

The remainder of the proof involves examination of the orders of the elements of the $C_i(\epsilon)$ and $A^{(1)}(\epsilon)$. To begin, partition $B(\epsilon)$ consistently with A

$$B(\epsilon) = \begin{bmatrix} B_{11}(\epsilon) & \cdots & B_{1N}(\epsilon) \\ \vdots & \ddots & \vdots \\ B_{N1}(\epsilon) & \cdots & B_{NN}(\epsilon) \end{bmatrix} \tag{2.126}$$

Using the structure of U and V given in (2.46)-(2.47), we can write the general form of the entries of $A^{(1)}(\epsilon)$ and $C_m(\epsilon)$:

$$\epsilon a_{IJ}^{(1)} = \mathbf{1}_I^{\mathrm{T}} B_{IJ} \pi_J \tag{2.127}$$

and

$$\epsilon \left[C_m(\epsilon)\right]_{IJ} = \sum_{K_1, K_2, \ldots, K_m} \cdots \sum \mathbf{1}_I^{\mathrm{T}} B_{IK_m}(\epsilon) S_{K_m} \cdots S_{K_1} B_{K_1 J}(\epsilon) \pi_J \tag{2.128}$$

Since the S_I defined in (2.124) are all $\mathrm{O}(1)$, we have that for some sequence of aggregate states $(K_1, \ldots, K_m)$ associated with one term in the sum (2.128)

$$[\epsilon C_m(\epsilon)]_{IJ} = \mathrm{O}(\|B_{IK_m}(\epsilon)\| \cdots \|B_{K_1 J}(\epsilon)\|) \tag{2.129}$$

In fact the left hand side of (2.129) could be of higher order due to cancellation in the sum (2.128) though this possibility is not exploited.

Let us first assume that A has no transient states. Thus, because of the recurrent structure of A, all entries in the π_I are strictly positive. Since $A(\epsilon)$ is a Markov generator, the off-diagonal blocks, $B_{IJ}(\epsilon)$, $I \neq J$, have non-negative entries. Therefore

$$\begin{aligned} \|B_{IJ}(\epsilon)\| &= \mathrm{O}\Big(\mathbf{1}_I^{\mathrm{T}} B_{IJ}(\epsilon) \pi_J\Big) \quad \forall I \neq J & (2.130) \\ &= \epsilon\, \mathrm{O}\Big(a_{IJ}^{(1)}(\epsilon)\Big) & (2.131) \end{aligned}$$

Consider an off-diagonal entry of $C_m(\epsilon)$. A new sequence of aggregate states $(K_1', \ldots, K_n')$, $n \leq m$, can be formed by removing repetitions so that $K_i' \neq K_{i+1}'$. Since the $B_{IJ}(\epsilon)$ are all $\mathrm{O}(\epsilon)$

$$[\epsilon C_m(\epsilon)]_{IJ} = \epsilon^{m-n} \mathrm{O}\Big(\Big\|B_{IK_n'}(\epsilon)\Big\| \cdots \Big\|B_{K_1' J}(\epsilon)\Big\|\Big) \tag{2.132}$$

where

$$2 \leq n \leq \min(m, \dim(A)) \tag{2.133}$$

By applying (2.131) this gives

$$\epsilon\ [C_m(\epsilon)]_{IJ} = \epsilon^m \mathrm{O}\Big(a^{(1)}_{IK'_n}(\epsilon) \cdots a^{(1)}_{K'_1 J}(\epsilon)\Big) \quad I \neq J \tag{2.134}$$

Since the state space is finite, n (the number of terms in the product) is bounded as $\epsilon \downarrow 0$ and the (I, J) element $(I \neq J)$ of the series $W^{(1)}(\epsilon) = C_1(\epsilon) + C_2(\epsilon) + \cdots$ converges to a term which is weak with respect to $A^{(1)}(\epsilon)$.

The weakness of the diagonal terms of $W^{(1)}(\epsilon)$ can be established as follows. First note that since $\mathbf{1}^{\mathrm{T}} W^{(1)}(\epsilon) = \mathbf{0}^{\mathrm{T}}$ by construction

$$w^{(1)}_{ii}(\epsilon) = -\sum_{j \neq i} w^{(1)}_{ji}(\epsilon) \tag{2.135}$$

Secondly, for any i, and $j \neq i$, $w^{(1)}_{ji}(\epsilon)$ is weak and therefore there is some sequence $(k_0 = i, k_1(j), k_2(j), \ldots, k_r(j) = j)$ so that

$$w^{(1)}_{ji}(\epsilon) = \epsilon\, \mathrm{O}\Big(\, a_{k_1(j)i}(\epsilon)\, a_{k_2(j)k_1(j)}(\epsilon) \ \cdots\ a_{jk_{r-1}(j)}(\epsilon)\,\Big) \tag{2.136}$$

This in turn implies

$$w^{(1)}_{ji}(\epsilon) = \epsilon\, \mathrm{O}\Big(a_{k_1(j)i}(\epsilon)\Big) \tag{2.137}$$

Combining (2.135) and (2.137) we have

$$w^{(1)}_{ii}(\epsilon) = \epsilon\, \mathrm{O}\left(\sum_{i \neq j} a_{k_1(j)i}(\epsilon)\right) \tag{2.138}$$

Now since all the terms in the sum in (2.138) are positive, although the $k_1(j)$ need not be distinct for different values of j, we still have that

$$\sum_{j \neq j} a_{k_1(j)i} = \mathrm{O}\left(\sum_{j \neq i} a_{ji}(\epsilon)\right) = \mathrm{O}(a_{ii}(\epsilon)) \tag{2.139}$$

Combining (2.138) and (2.139) we have

$$w^{(1)}_{ii}(\epsilon) = \epsilon\, \mathrm{O}(a_{ii}(\epsilon)) \tag{2.140}$$

which proves that $W^{(1)}(\epsilon)$ is indeed weak with respect to $A^{(1)}(\epsilon)$.

Let us now consider the case where A generates transient states and satisfies the no coupling condition on transient states specified in Definition 2.1. Since there may be transient states in the unperturbed process, the probabilities π_I are not strictly positive. However, since the transient states do not couple aggregates, (2.131) still applies. Specifically, by permuting the states within each aggregate class, the blocks $B_{IJ}(\epsilon)$, $I \neq J$ and the probabilities π_J can be brought into the forms

$$B_{IJ}(\epsilon) = \left[\begin{array}{cc} B'_{IJ}(\epsilon) & \mathbf{0} \end{array} \right] \tag{2.141}$$

and

$$\pi_J = \left[\begin{array}{c} \pi'_J \\ \mathbf{0} \end{array} \right] \, , \; \pi'_J \text{ strictly positive} \tag{2.142}$$

Therefore, since no terms in the product are annihilated by the zero terms, (2.131) follows as before

$$\|B_{IJ}(\epsilon)\| = \mathrm{O}\left(\mathbf{1}_I^{\mathrm{T}} B'_{IJ}(\epsilon)\pi'_J\right) = \epsilon\,\mathrm{O}\left(a_{IJ}^{(1)}(\epsilon)\right) \quad I \neq J \tag{2.143}$$

The remainder of the proof follows as before.

2.A.2 Proof of Lemma 2.3

The first observation is that the term $W^{(k+1)}(\epsilon)$ is composed of two parts, $W_1^{(k+1)}(\epsilon)$ and $W_2^{(k+1)}(\epsilon)$. $W_1^{(k+1)}(\epsilon)$ is simply the propagation of the previous weak term $W^{(k)}(\epsilon)$. This term is weak with respect to $A^{(k+1)}(\epsilon)$ since for any term $w_{ji}^{(k)}$ there was a path from i to j using $A^{(k)}(\epsilon)$ which is of lower order. From this follow that for any term $w_{JI}^{(k+1)}$ there is an aggregate path from I to J using $A^{(k+1)}(\epsilon)$ which is of lower order.

The fact that $W_2^{(k+1)}(\epsilon)$ is weak follows from the observation that if $W^{(k)}(\epsilon)$ is weak with respect to $A^{(k)}(\epsilon) = A^{(k)} + B^{(k)}(\epsilon)$, then using the non-negativity of the off-diagonal blocks of $B^{(k)}(\epsilon)$

$$\begin{gathered} \left\|W_{JI}^{(k)}(\epsilon)\right\| = \epsilon\,\mathrm{O}\left(\left\|B_{JS_K}^{(k)}(\epsilon)\right\| \cdots \left\|B_{S_1 I}^{(k)}(\epsilon)\right\|\right) \\ \text{for some aggregate path}(I, S_1, \ldots, S_K, J) \end{gathered} \tag{2.144}$$

Using (2.144) any expression of the form $B_{IJ}^{(k)}(\epsilon) + W_{IJ}^{(k)}(\epsilon)$ can be bounded in norm as

$$\left\|B_{JI}^{(k)}(\epsilon) + W_{JI}^{(k)}(\epsilon)\right\| = \mathrm{O}\left(B_{JI}^{(k)}(\epsilon)\right) + \epsilon\, \mathrm{O}\left(\left\|B_{J S_K}^{(k)}(\epsilon)\right\| \cdots \left\|B_{S_1 I}^{(k)}(\epsilon)\right\|\right) \tag{2.145}$$

for the same aggregate path$(I, S_1, \ldots, S_K, J)$

The proof of Lemma 2.3 follows identically to the derivation (2.116)–(2.129) with $B^{(k)}(\epsilon) + W^{(k)}(\epsilon)$ used in place of $B(\epsilon)$. Finally, the right hand side of the new (2.129) can be expressed in terms of $B(\epsilon)$ alone using (2.145)

$$\epsilon\ [C_m(\epsilon)]_{IJ} = \mathrm{O}(\|B_{I S_K}(\epsilon)\| \cdots \|B_{S_1 J}(\epsilon)\|) \tag{2.146}$$

for some finite sequence of aggregate states $(S_1, S_2, \ldots, S_K)$. From (2.146), expressions analogous to (2.132)-(2.134) follow as before.

Since both components of $W^{(k+1)}(\epsilon)$ are weak, their sum must also be weak which proves the lemma.

2.A.3 Proof of Lemma 2.5

The construction of $\hat{A}(\epsilon)$, $\hat{D}(\epsilon)$, and C was specified in the text. The properties (1)-(7) in Lemma 2.5 will be shown to follow directly from this construction.

Property (2) results from each new transient state being uniquely associated with a single ergodic class and having transitions only into that class. Since the recurrent portions of the chains generated by $A(\epsilon)$ and $\hat{A}(\epsilon)$ are identical, properties (3) and (5) follow. Again from the construction, since $\hat{V}$ is composed of 0-1 entries, property (4) is true.

The constructed columns of $\hat{A}(\epsilon)$ span the same space as do the columns of $D(\epsilon)$ therefore the Range($D(\epsilon)$) forms an invariant subspace of $\hat{A}(\epsilon)$. Further, since $CD(\epsilon) = I$ by construction the columns of $D(\epsilon)$ and the rows of C are independent, matrices C^* and $D^*(\epsilon)$ can be formed such that

$$\begin{bmatrix} C \\ C^* \end{bmatrix}^{-1} = \begin{bmatrix} D(\epsilon) & D^*(\epsilon) \end{bmatrix} \tag{2.147}$$

which allows the following change of basis:

$$\begin{bmatrix} C \\ C^* \end{bmatrix} e^{\hat{A}(\epsilon)t} \begin{bmatrix} D(\epsilon) & D^*(\epsilon) \end{bmatrix} = \exp\left(\begin{bmatrix} C\hat{A}(\epsilon)D(\epsilon) & C\hat{A}(\epsilon)D^*(\epsilon) \\ C^*\hat{A}(\epsilon)D(\epsilon) & C^*\hat{A}(\epsilon)D^*(\epsilon) \end{bmatrix} t\right) \tag{2.148}$$

Since Range($\hat{A}(\epsilon)$) = Range($D(\epsilon)$) by property (7) and $C^*D(\epsilon) = 0$ by (2.147)

$$C^*\hat{A}(\epsilon)D(\epsilon) = 0 \tag{2.149}$$

Therefore the lower left block of (2.148) is zero. By construction

$$C\hat{A}(\epsilon)D(\epsilon) = A(\epsilon) \tag{2.150}$$

therefore considering the upper left block of (2.148) gives property (1)

$$Ce^{\hat{A}(\epsilon)t}D(\epsilon) = e^{A(\epsilon)t} \tag{2.151}$$

2.A.4 Proof of Lemma 2.6

Lemma 2.6 is essentially a minor extension of Lemma 2.5. The major difficulty is that $G(\epsilon) = A(\epsilon) + W(\epsilon)$ is not necessarily a Markov generator since there may be small negative terms off the main diagonal. The terms $v_{Ij}(\epsilon)$ based on $G(\epsilon)$ do not therefore have a direct probabilistic interpretation. Let $v^{(A)}(\epsilon)$ be the probabilistic terms computed using $A(\epsilon)$ alone and $v^{(G)}(\epsilon)$ be the new term computed using $G(\epsilon)$. From the respective constructions, it follows that the terms are relatively close. Specifically,

$$\left|v_{Ij}^{(G)}(\epsilon) - v_{Ij}^{(A)}(\epsilon)\right| = \epsilon \mathrm{O}\left(v_{Ij}^{(A)}(\epsilon)\right) \tag{2.152}$$

$\hat{G}(\epsilon)$ can then be computed from $G(\epsilon)$ in the same manner as in the proof of Lemma 2.5. The terms $D^{(G)}(\epsilon)$ and $C^{(G)}$ are also computed. Let $D^{(A)}(\epsilon)$ and $C^{(A)}$ be terms computed from $A(\epsilon)$ alone. $C^{(G)} = C^{(A)}$ follows from the construction and using (2.152)

$$\left\|D^{(A)}(\epsilon) - D^{(G)}(\epsilon)\right\| = \mathrm{O}(\epsilon) \tag{2.153}$$

Therefore from (2.151) and (2.153), and the uniform bound on $C^{(A)}\exp(G(\epsilon)t)$

$$\mathrm{e}^{G(\epsilon)t} = C^{(A)}\mathrm{e}^{G(\epsilon)t}D^{(A)}(\epsilon) + \mathrm{O}(\epsilon) \tag{2.154}$$

Finally, using (2.152), $\hat{G}(\epsilon)$ can be decomposed as

$$\hat{G}(\epsilon) = \hat{A}(\epsilon) + \hat{W}'(\epsilon) + \hat{W}''(\epsilon) \tag{2.155}$$

where $\hat{W}'(\epsilon)$ is obtained by construction from $W(\epsilon)$ in the same way that $\hat{A}(\epsilon)$ is obtained from $A(\epsilon)$. $\hat{W}''(\epsilon)$ results from the small difference between the term $V^{(A)}(\epsilon)$ and $V^{(G)}(\epsilon)$.

An examination of the construction of $\hat{A}(\epsilon)$ shows that if an entry of $W(\epsilon)$ satisfies

$$w_{ji}(\epsilon) = \epsilon\,\mathrm{O}(a_{js_m}(\epsilon)\cdots a_{s_1 i}(\epsilon)) \tag{2.156}$$

then the corresponding entry in $\hat{W}'(\epsilon)$ satisfies

$$\hat{w}_{ji}(\epsilon) = \epsilon\,\mathrm{O}(\hat{a}_{js_m}(\epsilon)\cdots \hat{a}_{s_1 i}(\epsilon)) \tag{2.157}$$

since both $w_{ji}(\epsilon)$ and the product in parentheses are scaled by an ϵ-dependent term which is determined by i and j and not the intermediate states s_k in the sequence. Therefore $\hat{W}'(\epsilon)$ is weak with respect to $\hat{A}(\epsilon)$.

The term $\hat{W}''(\epsilon)$ results from using $V^{(A)}(\epsilon)$ instead of $V^{(G)}(\epsilon)$. From (2.152), the elements of $\hat{W}''(\epsilon)$ satisfy

$$\hat{w}''_{ij}(\epsilon) = \epsilon\,\mathrm{O}\Big(a_{ij}(\epsilon) + w'_{ij}(\epsilon)\Big) \quad \forall i,j \tag{2.158}$$

and therefore $\hat{W}''(\epsilon)$ is weak with respect to $\hat{A}(\epsilon) + \hat{W}'(\epsilon)$. The sum $\hat{W}(\epsilon) = \hat{W}'(\epsilon) + \hat{W}''(\epsilon)$ is therefore also weak with respect to $\hat{A}(\epsilon)$

2.A.5 Proof of Lemma 2.9

The combination of the two-step procedure of computing $A^{(k+1)}(\epsilon)$ from $A^{(k)}(\epsilon)$ instead of first constructing $\hat{A}^{(k)}(\epsilon)$ can be shown to be valid using several properties in Lemma 2.5. In particular,

$$D(\epsilon)C\hat{A}(\epsilon)D(\epsilon) = \hat{A}(\epsilon)D(\epsilon) \tag{2.159}$$

$D(\epsilon)U = \hat{U}$ and by definition $V(\epsilon) = \hat{V}D(\epsilon)$. Therefore

$$\begin{aligned}
\epsilon A^{(k+1)}(\epsilon) &= \hat{V}^{(k)}\hat{A}^{(k)}(\epsilon)\hat{U}^{(k)} && (2.160)\\
&= \hat{V}^{(k)}\hat{A}^{(k)}(\epsilon)D^{(k)}(\epsilon)U^{(k)} && (2.161)\\
&= \hat{V}^{(k)}D^{(k)}(\epsilon)C^{(k)}\hat{A}(\epsilon)D^{(k)}(\epsilon)U^{(k)} && (2.162)\\
&= V^{(k)}(\epsilon)C^{(k)}\hat{A}^{(k)}(\epsilon)\hat{U}^{(k)} && (2.163)\\
&= V^{(k)}(\epsilon)A^{(k)}(\epsilon)U^{(k)} && (2.164)
\end{aligned}$$

completing the proof.

Chapter 3

Decomposition of Continuous Time, Finite State, Semi-Markov Processes

3.1 Introduction and Motivation

Analysis of the behavior of singularly perturbed Markov processes has been dealt with extensively in the literature. In Chapter 2, a completely general, straightforward algorithm for the hierarchical decomposition and uniform approximation of such processes was presented. Much less work, however, has dealt with the analysis of perturbed semi-Markov processes although such models are extremely important in numerous applications. In this chapter, using much of the machinery and many of the concepts developed in Chapter 2, we develop aggregation and decomposition methods for a very large class of perturbed semi-Markov processes. While there are strong similarities to the ideas presented in Chapter 2, there are a number of important differences and novel characteristics that arise in the semi-Markov case.

The perturbed semi-Markov processes which are considered here can be specified by a set of transition probabilities $p_{ji}(\epsilon)$ for each transition from state i to state j,[1] and holding time probability densities $h_{ji}(\epsilon, t)$ conditioned on transition from state i to state j. A basic discussion of semi-Markov processes is available in [21]. These

[1] Recall that since column vectors of probabilities are used, p_{ji} is the transition probability *from i to j*.

terms can be used to recover the evolution of the state probabilities. For example,

$$\Pr(\text{next state} = j, \text{ transition time} \leq \tau \mid \text{current state} = i) = \int_0^\tau p_{ji}(\epsilon)\, h_{ji}(\epsilon, t)\, dt \tag{3.1}$$

Such a semi-Markov process can be represented as a directed graph where the link from node i to node j is labeled with the terms $p_{ji}(\epsilon)$ and $h_{ji}(\epsilon, t)$. The state (in the sense of memory) of the semi-Markov process is composed of the observed state, a discrete variable with one of a finite number of possible values, and the time already spent in that observed state, a continuous variable. In this sense, a general semi-Markov process, even with a finite number of observed states, is an infinite state random process.

Consider the case in which the terms are analytic functions of a perturbation parameter ϵ and each $h_{ji}(\epsilon, t)$ has a Laplace transform which is the ratio of finite degree polynomials in the transform variable. This last restriction guarantees that the probabilities of occupying each of the observed states can be represented using a finite set of ordinary differential equations. An example of a semi-Markov process which cannot be represented in this way is one in which some of the holding times are deterministic delays. Such a holding time distribution does not have a rational polynomial transform and there is no finite set of ordinary differential equations which describe the evolution of the state probabilities. In the remainder of this chapter, we will use the word state to mean the observed state.

Much of the previous research dealing with perturbed semi-Markov processes has been conducted by Korolyuk and coworkers [20] [31] [32] [33]. The class of systems considered in that work allows perturbations of the transition probabilities $p_{ij}(\epsilon)$ but effectively avoids considering the effect of ϵ-dependence of the holding time distributions $h_{ij}(t)$. Furthermore, at $\epsilon = 0$ the chain is either required to have no transient states [31] [32] or to have a particular transient structure [33]. The general result derived by Korolyuk is that the slow time scale behavior can be well approximated by a purely Markov process. The work presented in this chapter considers a far wider class of systems by allowing perturbation of the holding time densities as well as the transition probabilities, and by allowing a general ergodic structure of the unperturbed ($\epsilon = 0$) semi-Markov process. As we will see, this class of processes can exhibit a variety of interesting types of behavior. In particular, the slow time scale behavior must in general be described by another semi-Markov process.

It will be demonstrated that there are essentially two characteristics of the semi-Markov model which determine the slow time scale behavior. The first is the existence of very rare state transitions which are nevertheless associated with short holding times. The second characteristic is associated with likely state transitions which have very long holding times. The former characteristic is considered in Korolyuk's work where the ergodic structure of the unperturbed chain is constrained to a very specific form. As shown in that work, this results in exponential holding time distributions at the slow time scale. The latter characteristic has not been explicitly considered in the literature. We will see that a holding time probability density, $h_{ij}(\epsilon, t)$, which is very small in the sense that $\sup_{t \geq 0} |h_{ij}(\epsilon, t)| = O(\epsilon)$, but which nevertheless has a very long tail, can result in slow behavior of this second type. The holding time distributions which result are not necessarily exponential and this accounts for the use of a semi-Markov representation of the slow time scale system.

The work presented in this chapter provides a unified framework for the decomposition and time scale approximation of semi-Markov processes which possess both types of slow behavior and arbitrary ergodic structure at $\epsilon = 0$. The decomposition technique which is developed is conceptually treated as a sequence of transformations of the representation of the original semi-Markov model. A block diagram of the procedure is shown in Figure 3.1. The major result which will be presented in this chapter is the specification of the "direct algorithm" indicated in the figure which avoids the necessity of explicitly constructing the entire sequence of representations. It is instructive, however, to examine this sequence of constructions as it provides considerable insight into the possible types of behavior of perturbed semi-Markov processes. In addition, the derivation of this sequence provides us with the proof of the validity of the direct algorithm.

The purpose of this decomposition procedure is to approximate the behavior of the probability transition function $\Phi(\epsilon, t)$ of the original semi-Markov process,

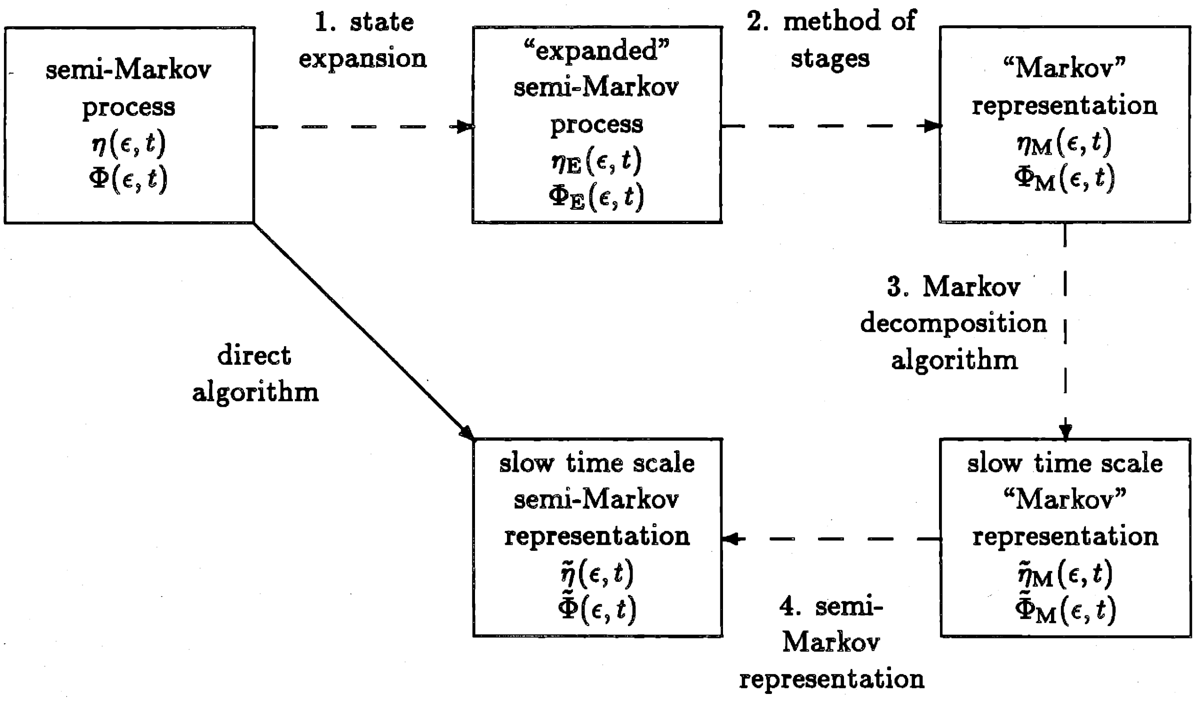

semi-Markov process
$\eta(\epsilon,t)$
$\Phi(\epsilon,t)$
1. state expansion
"expanded" semi-Markov process
$\eta_{\mathrm{E}}(\epsilon,t)$
$\Phi_{\mathrm{E}}(\epsilon,t)$
2. method of stages
"Markov" representation
$\eta_{\mathrm{M}}(\epsilon,t)$
$\Phi_{\mathrm{M}}(\epsilon,t)$
3. Markov decomposition algorithm
direct algorithm
slow time scale semi-Markov representation
$\tilde{\eta}(\epsilon,t)$
$\tilde{\Phi}(\epsilon,t)$
slow time scale "Markov" representation
$\tilde{\eta}_{\mathrm{M}}(\epsilon,t)$
$\tilde{\Phi}_{\mathrm{M}}(\epsilon,t)$
4. semi-Markov representation

Figure 3.1: Block diagram of semi-Markov decomposition algorithm

where the (i,j) element is defined as

$$\Phi_{ij}(\epsilon,t) = \Pr(\text{state at time } t \text{ is } i \mid \text{enter state } j \text{ at time } 0) \qquad (3.2)$$

This approximation is constructed using a fast time scale, ϵ-independent function of time $\Phi(0,t)$ and a slow time scale component $\tilde{\Phi}(\epsilon,t)$ which is the probability transition function of a perturbed, typically reduced-order, semi-Markov process. The direct algorithm provides a method of deriving this slow time scale model from the original semi-Markov model. The procedure can be iterated to recover unperturbed semi-Markov representations of multiple time scales from which the approximation is constructed.

The steps involved in the complete sequence of representations are summarized here. Each step is treated individually in Section 3.3.

1. *State Expansion.* The probability transition function $\Phi(\epsilon,t)$ of the original n-state semi-Markov process $\eta(\epsilon,t)$ is determined by the state transition probabilities $p_{ij}(\epsilon)$ and the holding time probability density functions $h_{ij}(\epsilon,t)$.

The function $\Phi(\epsilon,t)$ is realized using a $2n$-state semi-Markov process $\eta_{\mathrm{E}}(\epsilon,t)$ with a probability transition function $\Phi_{\mathrm{E}}(\epsilon,t)$. This "expanded" semi-Markov process is constructed by splitting each state into a "fast" and a "slow" copy and choosing new parameters appropriately. Specifically, the mean holding time in a "fast" state is $O(1)$ while in a "slow" state the probability that the holding time is less than any fixed finite time is $O(\epsilon)$. Using this construction, the function $\Phi(\epsilon,t)$ can be recovered exactly from $\Phi_{\mathrm{E}}(\epsilon,t)$.

2. *Method of Stages.* The expanded semi-Markov representation can be transformed into a "Markov" form by expanding each holding time into a sequence of exponential stages [14] [27]. The state transition function $\Phi_{\mathrm{M}}(\epsilon,t)$ associated with this representation can be thought of as the probability transition function of a Markov process in which some of the transition rates and probabilities are complex and negative. The function $\Phi_{\mathrm{E}}(\epsilon,t)$, and therefore $\Phi(\epsilon,t)$, can be recovered exactly from the Markov transition function $\Phi_{\mathrm{M}}(\epsilon,t)$. The number of states can be much greater than in the original system and therefore the actual construction of this process may not be a feasible approach in practice. However, the added states are aggregated in the steps 3 and 4 and the direct algorithm therefore avoids their explicit construction.

3. *Markov Decomposition Algorithm.* The previous step results in the specification of a system of ordinary differential equations (with complex state variables) which generates $\Phi_{\mathrm{M}}(\epsilon,t)$. Its structure is very close to that of a system which describes the evolution of the state probabilities of a finite state Markov process, although as mentioned above, some of the "transition rates" may be complex quantities. As will be discussed in Section 3.3.3, the algorithm for the decomposition of perturbed Markov generators is still applicable. Therefore $\Phi_{\mathrm{M}}(\epsilon,t)$ can be approximated using $\Phi_{\mathrm{M}}(0,t)$, and $\tilde{\Phi}_{\mathrm{M}}(\epsilon,t)$, the state transition function of a reduced-order system obtained using the method developed in Chapter 2. This reduced-order system is again almost of the form which would describe the evolution of the state probabilities of a Markov process. Also as a result of the aggregation performed in constructing $\tilde{\Phi}_{\mathrm{M}}(\epsilon,t)$, the states added in the previous step largely do not appear. In addition, a novel

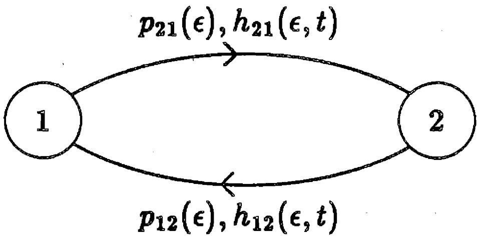

$p_{21}(\epsilon), h_{21}(\epsilon, t)$
1
2
$p_{12}(\epsilon), h_{12}(\epsilon, t)$

Figure 3.2: Two state semi-Markov process in Example 3.1

feature that can arise in the semi-Markov case (but not in the Markov case) is that the aggregation performed can result in the "fast" and "slow" copies of a particular original state belonging to *different* aggregate classes at the slow time scale.

4. *Semi-Markov Representation.* The final step involves identifying the particular terms of $\tilde{\Phi}_{\mathrm{M}}(\epsilon, t)$ which are required to approximate $\Phi(\epsilon, t)$. In particular, it will be shown that there is a valid semi-Markov process, on what is typically a reduced-order, aggregated state set, and with a probability transition function $\tilde{\Phi}(\epsilon, t)$, which captures all the necessary terms.

Example 3.1 *Consider the simple 2-state process illustrated in Figure 3.2 This process has parameters:*

$$p_{21}(\epsilon) = p_{12}(\epsilon) = 1 \tag{3.3}$$

$$h_{21}(\epsilon, t) = (1-\epsilon)\lambda_1 \mathrm{e}^{-\lambda_1 t} + \epsilon(\epsilon^N \lambda_2)\mathrm{e}^{-(\epsilon^N \lambda_2)t} \tag{3.4}$$

$$h_{12}(\epsilon, t) = \lambda_1 \mathrm{e}^{-\lambda_1 t} \tag{3.5}$$

Note that the holding time distribution in state 1, $h_{21}(\epsilon, t)$, *is made up of two components. The first is a "fast" exponential* $\lambda_1 \exp(-\lambda_1 t)$ *which is weighted by* $(1-\epsilon)$ *and the second is a "slow" exponential* $(\epsilon^N \lambda_2)\exp(\epsilon^N \lambda_2 t)$ *which is weighted by* ϵ. *An interpretation of this is that upon entering state* 1, *the holding time will either, with probability* $(1-\epsilon)$, *be exponentially distributed with a rate* λ_1, *or, with*

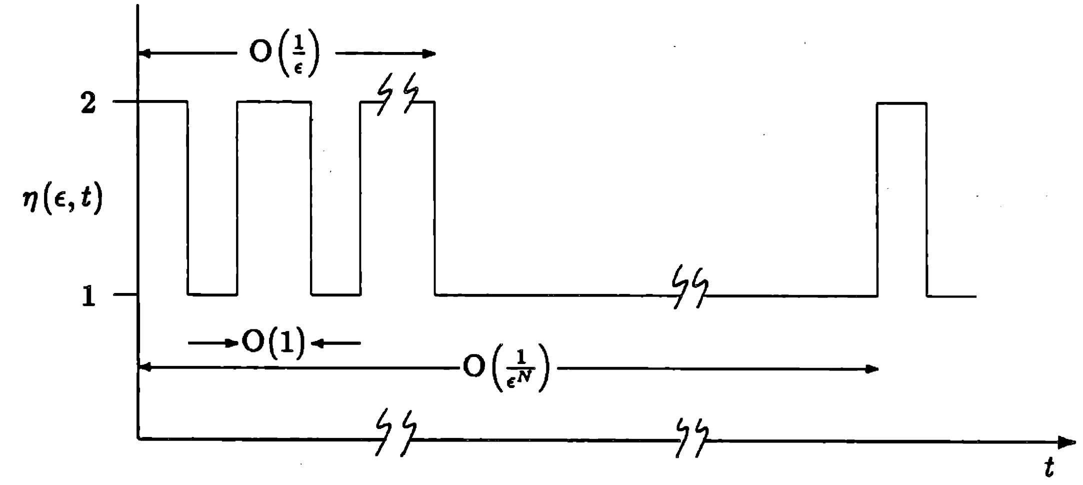


Figure 3.3: Typical semi-Markov behavior in Example 3.1

probability ϵ *be exponentially distributed with a rate* $\epsilon^N\lambda_2$. *In the former case, the mean holding time will be* $O(1)$ *while in the latter case the mean holding time will be* $O\left(1/\epsilon^N\right)$.

Based on this interpretation of the holding time in state 1, we can see that qualitatively, the behavior of this process involves rapid transitions between states 1 *and* 2 *for a period of time of length* $O(1/\epsilon)$ *followed by a period of* $O\left(1/\epsilon^N\right)$ *in state* 1, *followed again by rapid transitions for* $O(1/\epsilon)$ *time as illustrated in Figure 3.3. The slowest behavior is largely determined by the long* $O(\epsilon)$ *"tail" of the distribution* $h_{21}(\epsilon,t)$. *This simple example demonstrates how perturbation of the holding time distribution can dramatically influence the slow time-scale behavior.*

Let us examine the behavior of this process in detail and at the same time illustrate some of the steps outlined above. Specifically, consider splitting state 1 *to account explicitly for the decision concerning the holding time distribution to be used in determining when the process exits from state* 1. *We call state* f_1 *the combination of the old state* 1 *and the decision to use the "fast" component of the holding time distribution for the next transition to the old state* 2. *State* s_1 *is the combination of the old state* 2 *and the decision to use the "slow" component of the*

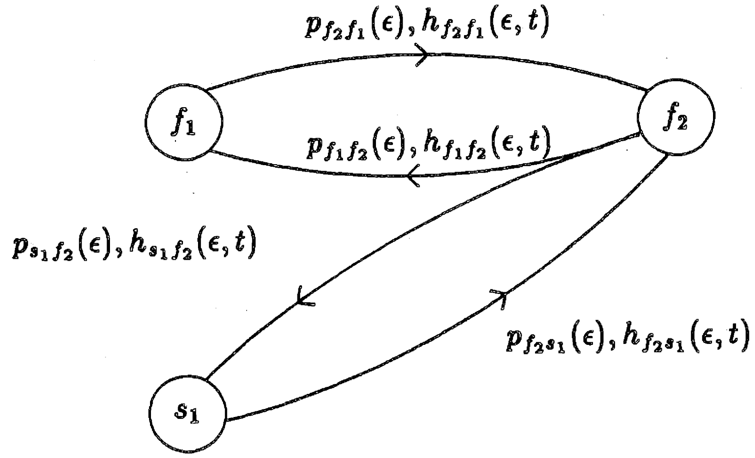

$p_{f_2f_1}(\epsilon), h_{f_2f_1}(\epsilon,t)$
f_1
f_2
$p_{f_1f_2}(\epsilon), h_{f_1f_2}(\epsilon,t)$
$p_{s_1f_2}(\epsilon), h_{s_1f_2}(\epsilon,t)$
$p_{f_2s_1}(\epsilon), h_{f_2s_1}(\epsilon,t)$
s_1

Figure 3.4: Expanded semi-Markov process $\eta_{\mathrm{E}}(\epsilon, t)$ in Example 3.1

holding time. We call the old state 2, f_2, since the holding time in state 2 is always "fast". More precisely the result of this expansion of the state space is shown in Figure 3.4 and has the following parameters.

$$p_{f_2 s_1}(\epsilon) = p_{f_2 f_1}(\epsilon) = 1 \tag{3.6}$$

$$p_{f_1 f_2}(\epsilon) = 1 - \epsilon \tag{3.7}$$

$$p_{s_1 f_2}(\epsilon) = \epsilon \tag{3.8}$$

$$h_{s_1 f_2}(\epsilon, t) = h_{f_2 f_1}(\epsilon, t) = \lambda_1 \mathrm{e}^{-\lambda_1 t} \tag{3.9}$$

$$h_{f_2 s_1}(\epsilon, t) = (\epsilon^N \lambda_2)\mathrm{e}^{-(\epsilon^N \lambda_2)t} \tag{3.10}$$

$$h_{f_2 f_1}(\epsilon, t) = \lambda_1 \mathrm{e}^{-\lambda_1 t} \tag{3.11}$$

The behavior of the original process can be recovered from that of the expanded process. The probability of being in state 1 of the original process is simply the sum of the probabilities of being in states f_1 and s_1, while the probability of being in state 2 is the probability of being in state f_2. The initial probability of being in state 1 must also be distributed among f_1 and s_2. Specifically, it will be shown that

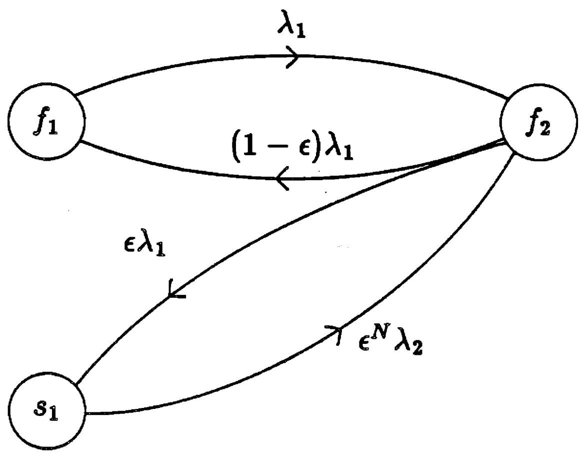


Figure 3.5: Markov representation

the behavior of the original process can be recovered exactly using

$$\Phi(\epsilon,t) = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \Phi_{\mathrm{E}}(\epsilon,t) \begin{bmatrix} 1-\epsilon & 0 \\ 0 & 1 \\ \epsilon & 0 \end{bmatrix} \tag{3.12}$$

where the states of the expanded process $\eta_{\mathrm{E}}(\epsilon,t)$ are ordered f_1, f_2, s_1.

In general the holding times would at this point be expanded using the method of stages (step 3). In this very simple example, the unconditional holding times are already exponentially distributed; therefore $\Phi_{\mathrm{M}}(\epsilon,t) = \Phi_{\mathrm{E}}(\epsilon,t)$. Note that it is not sufficient for all the holding times leaving a state to be exponentially distributed for $\Phi_{\mathrm{M}}(\epsilon,t) = \Phi_{\mathrm{E}}(\epsilon,t)$. Each state must have a single exponential constant associated with it to insure that the unconditional holding time in each state is an exponential random variable. This issue is dealt with in Section 3.3.2. The Markov representation with probability transition rates indicated is shown in Figure 3.5.

This Markov process can be decomposed into "fast" and "slow" models using the Markov decomposition algorithm presented in Chapter 2. The resulting models are shown in Figure 3.6. An interesting feature of the slow model is that the states f_1 and s_1 belong to different aggregate classes. This characteristic of the slow time scale model will be discussed more fully in Section 3.5.

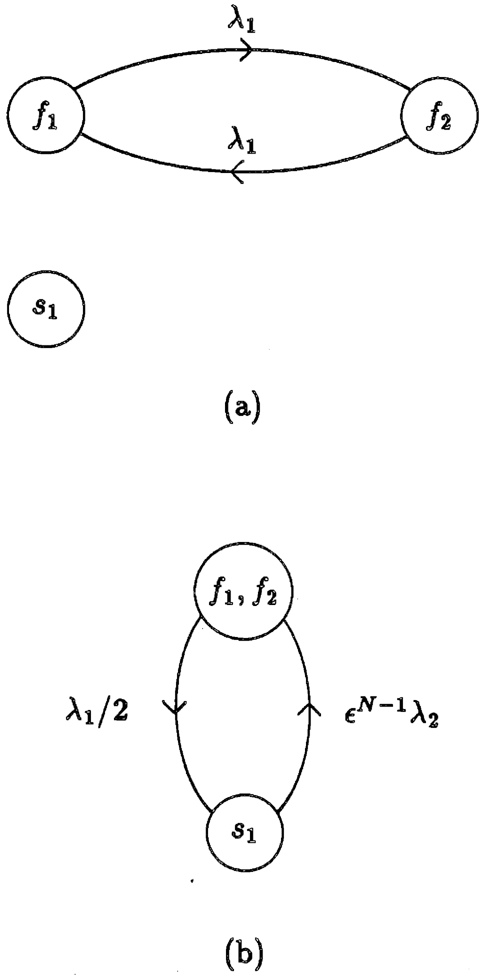


Figure 3.6: (a) "Fast" time scale $\eta_{\mathrm{M}}(0,t)$ and (b) "slow" time scale $\tilde{\eta}_{\mathrm{M}}(\epsilon,t)$ Markov processes in Example 3.1

The final step would be to represent the slow system in a semi-Markov form. Again, in this simple example, this step is not necessary; therefore $\tilde{\tilde{\Phi}}(\epsilon,t) = \tilde{\tilde{\Phi}}_{\mathrm{M}}(\epsilon,t)$. *As will be developed later, the approximation which results for this example is*

$$\Phi(\epsilon,t) = \Phi_0(\epsilon,t) + \begin{bmatrix} 1/2 & 1 \\ 1/2 & 0 \end{bmatrix} \tilde{\tilde{\Phi}}(\epsilon,\epsilon t) \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix} + \mathrm{O}(\epsilon) \qquad (3.13)$$

where the states of $\tilde{\eta}(\epsilon,t)$ *are ordered* $\{f_1, f_2\}$, $\{s_1\}$. □

The remainder of this chapter is organized as follows. In the next section, the detailed algorithm is specified. Then, in Section 3.3, the details of the derivation are provided. This development is based on first describing each of the transformations of the system outlined above and then showing how the direct algorithm provides the same result. In Section 3.4, an example is considered in detail using Algorithm 3.1. A discussion is provided in Section 3.5. Some supporting proofs and other results are provided in Section 3.A.

3.2 The Algorithm

In this section, the details of the direct algorithm are specified. We begin with the specification of a perturbed semi-Markov process, $\eta(\epsilon,t)$. The state transition function $\Phi(\epsilon,t)$, on the finite state set $\{1,2,\ldots,n\}$ is determined by the state transition probabilities

$$p_{ji}(\epsilon) = \Pr(\text{next state} = j|\text{current state} = i) \qquad (3.14)$$

and the holding time probability density functions such that

$$\int_{\tau=0}^{t} h_{ji}(\epsilon,\tau)\, d\tau = \Pr(\text{holding time in state } i \le t|\text{next state} = j) \qquad (3.15)$$

Furthermore, each holding time distribution $h(\epsilon,t)$ must have a Laplace transform

$$H_{ji}(\epsilon,s) \equiv \int \mathrm{e}^{-st} h_{ji}(\epsilon,t)\, ds \qquad (3.16)$$

which has a rational polynomial form

$$H_{ji}(\epsilon,s) = \frac{a_{ji}(\epsilon,s)}{b_{ji}^{(s)}(\epsilon,s)\, b_{ji}^{(f)}(\epsilon,s)} \qquad (3.17)$$

where a, $b^{(s)}$, and $b^{(f)}$ are finite degree polynomials in s and the roots are analytic functions of ϵ. Furthermore, all the roots of $b^{(s)}$ converge to 0 as $\epsilon \downarrow 0$ and the roots of $b^{(f)}$ converge to nonzero values.

Algorithm 3.1

1. *Construct the parameters of an "expanded" semi-Markov process, $\eta_{\mathrm{E}}(\epsilon, t)$, on a state set $\mathcal{F} \cup \mathcal{S}$ where $\mathcal{F} = \{f_1, f_2, \ldots, f_n\}$ and $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$ are the "fast" and "slow" state sets respectively.*

   Decompose each $h_{ij}(\epsilon, t)$ into the "fast" and "slow" components as

$$h(\epsilon, t) = p^{(s)}(\epsilon)h^{(s)}(\epsilon, t) + p^{(f)}(\epsilon)h^{(f)}(\epsilon, t) \tag{3.18}$$

   such that

$$H^{(s)}(\epsilon, s) = \frac{a^{(s)}(\epsilon, s)}{b^{(s)}(\epsilon, s)}\ , \quad H^{(f)}(\epsilon, s) = \frac{a^{(f)}(\epsilon, s)}{b^{(f)}(\epsilon, s)} \tag{3.19}$$

   The "probabilities" $p^{(s)}(\epsilon)$ and $p^{(f)}(\epsilon) = 1 - p^{(s)}(\epsilon)$ and the coefficients of $a^{(f)}(\epsilon, s)$ and $a^{(s)}(\epsilon, s)$ can be obtained in a variety of ways including matching the coefficients in the polynomial equation

$$a(\epsilon, s) = p^{(f)}(\epsilon)a^{(f)}(\epsilon, s)b^{(s)}(\epsilon, s) + p^{(s)}(\epsilon)a^{(s)}(\epsilon, s)b^{(f)}(\epsilon, s) \tag{3.20}$$

   If

$$\sup_{t \geq 0} h_{ij}(\epsilon, t) = O(\epsilon) \tag{3.21}$$

   then the "fast" component of the holding time distribution is not significant and we set

$$p_{ij}^{(f)}(\epsilon) = 0\ , \quad p_{ij}^{(s)}(\epsilon) = 1 \tag{3.22}$$

2. *For each state i of $\eta(\epsilon, t)$, compute*

$$\begin{aligned} p_i^{(s)}(\epsilon) &= \sum_{j \neq i} p_{ji}^{(s)}(\epsilon)p_{ji}(\epsilon) & (3.23)\\ p_i^{(f)}(\epsilon) &= 1 - p_i^{(s)}(\epsilon) & (3.24) \end{aligned}$$

The semi-Markov parameters of the expanded process $\eta_{\mathrm{E}}(\epsilon,t)$ are then computed as

$$p_{f_j f_i}(\epsilon) = p_{ji}^{(f)}(\epsilon)\, p_{ji}(\epsilon)\, \frac{p_j^{(f)}(\epsilon)}{p_i^{(f)}(\epsilon)} \tag{3.25}$$

$$p_{s_j f_i}(\epsilon) = p_{ji}^{(f)}(\epsilon)\, p_{ji}(\epsilon)\, \frac{p_j^{(s)}(\epsilon)}{p_i^{(f)}(\epsilon)} \tag{3.26}$$

$$p_{s_j s_i}(\epsilon) = p_{ji}^{(s)}(\epsilon)\, p_{ji}(\epsilon)\, \frac{p_j^{(s)}(\epsilon)}{p_i^{(s)}(\epsilon)} \tag{3.27}$$

$$p_{f_j s_i}(\epsilon) = p_{ji}^{(s)}(\epsilon)\, p_{ji}(\epsilon)\, \frac{p_j^{(f)}(\epsilon)}{p_i^{(s)}(\epsilon)} \tag{3.28}$$

and

$$h_{f_j f_i}(\epsilon,t) = h_{s_j f_i}(\epsilon,t) = h_{ji}^{(f)}(\epsilon,t) \tag{3.29}$$

$$h_{f_j s_i}(\epsilon,t) = h_{s_j s_i}(\epsilon,t) = h_{ji}^{(s)}(\epsilon,t) \tag{3.30}$$

3. *Identify the ergodic classes $E_1,\ldots,E_N$ and the transient set T of the expanded process $\eta_{\mathrm{E}}(0,t)$. Note that $T \subseteq \mathcal{F}$ and that either $E_I \subseteq \mathcal{F}$ or $E_I = \{i\}, i \in \mathcal{S}$.*

4. *Using the transition probabilities of the process $\eta_{\mathrm{E}}(\epsilon,t)$, compute terms $v_{\mathrm{E}Ij}(\epsilon)$ as in the Markov algorithm.*

$$v_{\mathrm{E}Ij}(\epsilon) \equiv \Pr\Big(\eta_{\mathrm{E}}(\epsilon,t^*) \in E_I \;\Big|\; \eta_{\mathrm{E}}(\epsilon,0) = j,\ t^* = \inf_{t\geq 0}(t \mid \eta_{\mathrm{E}}(\epsilon,t) \notin T)\Big) \tag{3.31}$$

5. *For each ergodic class $E_I \subseteq \mathcal{F}$, and each $i \in E_I$, compute $\bar{\tau}_i$, the mean holding time in state i, and $\bar{\mu}_i$, the inverse of the mean recurrence time in state i, for the process $\eta_{\mathrm{E}}(0,t)$. For each $E_I \subseteq \mathcal{F}$ also compute*

$$\Lambda_I(\epsilon) = \sum_{i\in E_I} \sum_{j\notin E_I} \bar{\mu}_{iI}\, p_{ji}(\epsilon) \tag{3.32}$$

6. *Construct a slow time scale perturbed semi-Markov process $\tilde{\eta}(\epsilon,t)$ with N states (one for each ergodic class of $\eta_{\mathrm{E}}(0,t)$) with the following parameters. If $E_I \subseteq \mathcal{F}$*

$$\tilde{h}_{JI}(\epsilon,t) = \frac{1}{\epsilon}\Lambda_I(\epsilon)\mathrm{e}^{-\Lambda_I(\epsilon)t/\epsilon} \tag{3.33}$$

and

$$\tilde{p}_{JI}(\epsilon) = \frac{1}{\Lambda_I(\epsilon)} \left(\sum_{i \in E_i} \sum_{j \in E_J} \bar{\mu}_i \, p_{ji}(\epsilon) + \sum_{i \in E_I} \sum_{t \in T} \bar{\mu}_i \, p_{ti}(\epsilon) \, v_{\mathrm{E}Jt}(\epsilon) \right) \tag{3.34}$$

If $E_I = \{i\} \in S$

$$\tilde{p}_{JI}(\epsilon) = \frac{1}{\epsilon} \left(\sum_{j \in E_J} p_{ji}(\epsilon) + \sum_{k \in T} p_{ki}(\epsilon) v_{\mathrm{E}Jk}(\epsilon) \right) \tag{3.35}$$

and

$$\begin{aligned} \tilde{h}_{JI}(\epsilon, t) \quad = \quad & \frac{1}{\tilde{p}_{JI}(\epsilon)} \frac{1}{\epsilon} \left(\sum_{j \in E_J} p_{ji}(\epsilon) h_{ji}^{(s)}(\epsilon, t/\epsilon) + \right. \\ & \left. \sum_{k \in T} p_{ki}(\epsilon) h_{ki}^{(s)}(\epsilon, t/\epsilon) v_{\mathrm{E}Jk}(\epsilon) \right) \end{aligned} \tag{3.36}$$

This slow time scale process has a probability transition function $\tilde{\Phi}(\epsilon, t)$.

7. *The original transition function* $\Phi(\epsilon, t)$ *can be approximated*

$$\Phi(\epsilon, t) = \Phi(0, t) + U\tilde{\Phi}(\epsilon, \epsilon t)V - UV + \mathrm{O}(\epsilon) \tag{3.37}$$

where

$$U = [u_{iJ}] \ , \ \ u_{iJ} = \begin{cases} \bar{\mu}_{f_i} \bar{\tau}_{f_i} & f_i \in E_J \\ 1 & \{s_i\} = E_J \\ 0 & \textit{otherwise} \end{cases} \tag{3.38}$$

and

$$V = [v_{Ij}] \ , \ \ v_{Ij} = p_j^{(f)}(0) v_{EIf_j} + p_j^{(s)}(0) v_{EIs_j} \tag{3.39}$$

8. *The above step can be iterated to provide a complete decomposition of* $\Phi^{(0)}(\epsilon, t)$. *Specifically, begin with* $\Phi(\epsilon, t) = \Phi^{(0)}(\epsilon, t)$. *The computed* $\tilde{\Phi}(\epsilon, t)$ *is then* $\Phi^{(1)}(\epsilon, t)$ *which serves as* $\Phi(\epsilon, t)$ *for the next iteration. This is continued until* $\eta^{(k)}(0, t)$ *has only one ergodic class. The process* $\eta^{(0)}(\epsilon, t)$ *therefore exhibits*

$k+1$ time scales and the overall approximation can then be constructed as

$$\begin{aligned}\Phi^{(0)}(\epsilon,t) &= \Phi^{(0)}(0,t) + \\ &\left(U^{(0)}\Phi^{(1)}(0,\epsilon t)V^{(0)} - U^{(0)}V^{(0)}\right) + \\ &\left(U^{(0)}U^{(1)}\Phi^{(2)}(0,\epsilon^2 t)V^{(1)}V^{(0)} - U^{(0)}U^{(1)}V^{(1)}V^{(0)}\right) + \\ &\vdots \\ &\left(U^{(0)}\cdots U^{(k-1)}\Phi^{(k)}(0,\epsilon^k t)V^{(k-1)}\cdots V^{(0)} - \right. \\ &\left. \qquad U^{(0)}\cdots U^{(k-1)}V^{(k-1)}\cdots V^{(0)}\right) + \mathrm{O}(\epsilon)\end{aligned} \tag{3.40}$$

□

Several comments can be made about the computations required in the various steps of this algorithm.

- In step 1, the probabilities $p^{(f)}(\epsilon)$ and $p^{(s)}(\epsilon)$ and the coefficients of $a^{(f)}(\epsilon,s)$ and $b^{(f)}(\epsilon,s)$ can be determined easily by solving a set of linear equations. The derivation of these equations is shown in Section 3.A.1.

- Step 3 requires identifying the ergodic classes and transient states of the process $\eta_{\mathrm{E}}(0,t)$. This step can in fact be performed using the original process $\eta(0,t)$. The ergodic classes of $\eta_{\mathrm{E}}(0,t)$ are of two basic types. First, each slow state $i \in S$ must necessarily be a ergodic class with exactly one state. This follows from the fact that as $\epsilon \downarrow 0$, the probability that the holding time is less than any time $T = \mathrm{O}(1)$ converges to zero. Therefore, at $\epsilon = 0$, the state is a trapping state and consequently is a *degenerate* ergodic class. Second, if a set of states $\{i, j, \ldots\}$ of the process $\eta(0,t)$ forms an ergodic class, then necessarily, the set $\{f_i, f_j, \ldots\}$ forms an ergodic class of $\eta_{\mathrm{E}}(0,t)$. This isomorphism of the ergodic classes of $\eta(0,t)$ and $\eta_{\mathrm{E}}(0,t)$ also simplifies the computation of the terms $\bar{\mu}_{iJ}$ in step 5.

- In step 4, the terms $v_{\mathrm{E}Ij}(\epsilon)$ depend only on the transition probabilities $p_{ij}(\epsilon)$ and not explicitly on the holding time distributions (other than to identify the transient and recurrent states). Therefore, computation of these trapping probabilities can be performed in much the same manner as in the Markov case. Note that in the continuous time Markov algorithm, a discrete time chain was constructed from the state transition rates. In this semi-Markov case, the transition probabilities taken alone already defines a discrete time Markov chain. After making the recurrent states trapping, the limiting probabilities can be computed directly from these transition probabilities.

- The computation of $\bar{\mu}$ for any nondegenerate ergodic class of $\eta_{\mathrm{E}}(0,t)$ can be carried out using the statistics of the original process $\eta(0,t)$. Suppose that a set $\{i,j,\ldots\}$ forms an ergodic class of $\eta(0,t)$. If ρ_i is the steady state probability of entering state i on the next transition and $\bar{\tau}_i$ is the mean holding time in state i, then

  $$\bar{\mu}_i = \frac{\rho_i}{\sum_j \rho_j \bar{\tau}_j} \tag{3.41}$$

  Therefore, since the associated ergodic class of $\eta_{\mathrm{E}}(0,t)$ shares the same statistics,

  $$\bar{\mu}_{If_i} = \bar{\mu}_i \tag{3.42}$$

  where $I = \{f_i, f_j, \ldots\}$.

The steps described in this algorithm, as well as the steps involved in the explicit transformation are described in the next section. Simple examples are provided in the text to demonstrate specific aspects of the derivation. A complete example is dealt with in Section 3.4.

3.3 Development

In this section, each of the transformations described in Section 3.1 is dealt with in detail. Then, the direct algorithm is shown to follow from this sequence of transformations. Simple examples are given in this development. Section 3.4 follows with a complete example of the transformations and the direct algorithm.

3.3.1 State expansion

The first transformation involves explicit representation of each holding time probability distribution as a sum of "fast" and "slow" components. This type of representation is then introduced into the entire model to construct the "expanded" system.

Consider a holding time probability density function $h(\epsilon, t)$ such that its Laplace transform $H(\epsilon, s)$ has the form

$$H(\epsilon, s) = \frac{a(\epsilon, s)}{b(\epsilon, s)} = \frac{a(\epsilon, s)}{b^{(f)}(\epsilon, s)\, b^{(s)}(\epsilon, s)} \tag{3.43}$$

where all the terms are analytic functions of ϵ. We assume that there is zero probability of an instantaneous transition, therefore the numerator degree of $H(\epsilon, s)$ must be lower than the denominator degree. The denominator polynomial $b(\epsilon, s)$ is factored into a product such that all the roots of $b^{(s)}(\epsilon, s)$ approach 0 as $\epsilon \downarrow 0$ and all the roots of $b^{(f)}(\epsilon, s)$ remain bounded away from zero. As shown in the following lemma, the transform $H(\epsilon, s)$ can then be written as a sum of two terms. Furthermore, if we require that $\sup_{t \geq 0} h(\epsilon, t) = \mathrm{O}(1)$, then this representation also has a direct probabilistic interpretation (i.e. "probabilities"$\in [0, 1]$).

Lemma 3.1 *Suppose that $h(\epsilon, t)$ is a probability density function, ($h(\epsilon, t) \geq 0$, $\int_0^\infty h(\epsilon, t)\, dt = 1$) with $H(\epsilon, s)$ of the form shown in (3.43). Then $h(\epsilon, t)$ can be written as*

$$h(\epsilon, t) = p^{(f)}(\epsilon) h^{(f)}(\epsilon, t) + p^{(s)}(\epsilon) h^{(s)}(\epsilon, t) \tag{3.44}$$

where

- $p^{(f)}(\epsilon) + p^{(s)}(\epsilon) = 1$,
- $p^{(s)}(\epsilon) \geq 0$,
- *$h^{(s)}(\epsilon, t)$ is a valid probability density function,*
- *$H^{(f)}(\epsilon, s)$ has poles s_k such that $\Re(s_k) = \mathrm{O}(1)$, and*
- $\sup_{t \geq 0} h^{(s)}(\epsilon, t) = \mathrm{O}(\epsilon)$.

Furthermore, if $\sup_{t \geq 0} |h(\epsilon, t)| = \mathrm{O}(1)$,

- *$p^{(f)}(\epsilon) = \mathrm{O}(1) > 0$, and*
- *$h^{(f)}(\epsilon, t)$ is a valid probability density function.*

Proof *See Section 3.A.1 for a proof and a construction of a set of linear equations for $p^{(f)}(\epsilon)$ and $p^{(s)}(\epsilon)$, and the coefficients of $a^{(f)}(\epsilon, s)$ and $a^{(s)}(\epsilon, s)$.* □

An important fact to note is that regardless of the condition on $\sup_{t\geq 0} |h(\epsilon, t)|$, this decomposition can be performed. However, if $\sup_{t\geq 0} |h(\epsilon, t)| = O(\epsilon)$, then the "fast" probability $p^{(f)}(\epsilon)$ is $O(\epsilon)$ and we cannot guarantee that it is positive, and therefore the probabilistic interpretation of the decomposition is not necessarily valid. Furthermore, as will be seen below, if $p^{(f)}(\epsilon) = O(\epsilon)$, then even though there is a fast component of the holding time distribution, it is not associated with fast time scale behavior due to its small probability. It will be seen that such a low probability fast component can in fact be entirely discarded without affecting the validity of the overall approximation.

Using the decomposition in Lemma 3.1, each state i is "split" into two copies f_i and s_i such that the mean holding time in f_i is $O(1)$ and in s_i is $O(1/\epsilon^q)$ for some $q \geq 1$. As will be demonstrated below, this expansion simplifies the analysis in the further steps.

Example 3.2 *Consider the following holding time transform*

$$H(\epsilon, s) = \frac{(1-\epsilon)\lambda_1}{s+\lambda_1} + \frac{\epsilon^{N+1}\lambda_2}{s+\epsilon^N\lambda_2} \tag{3.45}$$

$$= \frac{s((1-\epsilon)\lambda_1 + \epsilon^{N+1}\lambda_2) + \epsilon^N\lambda_1\lambda_2}{(s+\lambda_1)(s+\epsilon^N\lambda_2)} \tag{3.46}$$

Note that a rational polynomial transform of a holding time distribution must satisfy two properties. First, since $\int_0^\infty h(\epsilon, t)\, dt = 1$, $H(\epsilon, 0) = 1$. Also, since there is zero probability of an instantaneous transition, we know that the numerator degree of $H(\epsilon, s)$ must be less than the denominator degree. This must also be true for $H^{(s)}(\epsilon, s)$ and $H^{(f)}(\epsilon, s)$. Consequently, $a^{(f)}(\epsilon, s)$ and $a^{(s)}(\epsilon, s)$ have degree zero in the example. Using $H^{(f)}(\epsilon, 0) = 1$ and $H^{(s)}(\epsilon, 0) = 1$, the coefficients of s^0 of $a^{(f)}(\epsilon, s)$ and $a^{(s)}(\epsilon, s)$ are found to be

$$a^{(f)}(\epsilon, s) = \lambda_1\ , \quad a^{(s)}(\epsilon, s) = \epsilon^N\lambda_2 \tag{3.47}$$

and therefore

$$s((1-\epsilon)\lambda_1 + \epsilon^{N+1}\lambda_2) + \epsilon^N\lambda_1\lambda_2 = p^{(f)}(\epsilon)\lambda_1(s+\epsilon^N\lambda_2) + p^{(s)}(\epsilon)\epsilon^N\lambda_2(s+\lambda_1) \tag{3.48}$$

Matching the coefficients in this polynomial equation gives

$$p^{(f)}(\epsilon) = 1 - \epsilon\ , \ \ p^{(s)}(\epsilon) = \epsilon \tag{3.49}$$

The decomposition is therefore

$$H(\epsilon, s) = (1-\epsilon)\frac{\lambda_1}{s+\lambda_1} + \epsilon\frac{\epsilon^N \lambda_2}{s+\epsilon^N \lambda_2} \tag{3.50}$$

which has the time domain form

$$h(\epsilon, t) = (1-\epsilon)\lambda_1 \mathrm{e}^{-\lambda_1 t} + \epsilon(\epsilon^N\lambda_2)\mathrm{e}^{-\epsilon^N\lambda_2 t} \tag{3.51}$$

which was used in Example 3.1.

This decomposition could have been performed easily in this case by considering the time domain form (as was done in Example 3.1). However, this general transform approach will work in cases where there is no obvious answer available by inspection. □

Example 3.3 *In order to illustrate a case where* $p^{(f)}(\epsilon) = \mathrm{O}(\epsilon) < 0$, *consider the holding time transform*

$$H(\epsilon, s) = \frac{a(\epsilon, s)}{b^{(f)}(\epsilon, s)\, b^{(s)}(\epsilon, s)} = \frac{s^2((\epsilon+\epsilon^2)\lambda) + s(2(\epsilon+\epsilon^2)\lambda - \epsilon) + \epsilon\lambda}{(s+1)^2(s+\epsilon\lambda)} \tag{3.52}$$

Solving for the decomposed form gives

$$H(\epsilon, s) = -\epsilon\frac{1}{(s+1)^2} + (1+\epsilon)\frac{\epsilon\lambda}{(s+\epsilon\lambda)} \tag{3.53}$$

which has a time domain form

$$h(\epsilon, t) = -\epsilon(\mathrm{e}^{-t} * \mathrm{e}^{-t}) + (1+\epsilon)(\epsilon\lambda)\mathrm{e}^{-\epsilon\lambda t} \tag{3.54}$$

Note that the "fast probability" $p^{(f)}(\epsilon) = -\epsilon$ *is negative even though* $h(\epsilon, t) > 0$.

It will be shown in Section 3.3.4, that in this case, since $p^{(f)}(\epsilon) = \mathrm{O}(\epsilon)$, *that* $h(\epsilon, t)$ *can be approximated by* $h^{(s)}(\epsilon, t) = (\epsilon\lambda)\exp(-\epsilon\lambda t)$ *without affecting the overall approximation.* □

Applying this type of expansion of each of the holding time distributions as in Algorithm 3.1, step 2, to the entire process $\eta(\epsilon,t)$ results in the expanded process $\eta_{\mathrm{E}}(\epsilon,t)$. The basic interpretation of equations (3.25)-(3.28) which determine the transition probabilities of $\eta_{\mathrm{E}}(\epsilon,t)$ is as follows. Consider a state transition from state i to state j in $\eta(\epsilon,t)$. Essentially, four state transitions are introduced in the process $\eta_{\mathrm{E}}(\epsilon,t)$, namely the transition probabilities from state f_i to state f_j in $\eta_{\mathrm{E}}(\epsilon,t)$ is made up of two basic terms. First, the conditional probability of making an i to j transition, conditioned on that transition being "fast", must be computed. This is then multiplied by the unconditional probability of the holding time in state j being "fast". The unconditional probability of a fast holding time in state i is simply

$$\begin{aligned} p_i^{(f)}(\epsilon) &\equiv \Pr(\text{"fast" in } i) && (3.55)\\ &= \sum_{k\neq i} \Pr(\text{"fast" in } i \mid i \to k \text{ transition}) \Pr(i \to k \text{ transition}) && (3.56)\\ &= \sum_{k\neq i} p_{ki}^{(f)}(\epsilon)\, p_{ki}(\epsilon) && (3.57)\end{aligned}$$

which is (3.23). The probability of making an $i \to j$ transition conditioned on having a fast holding time in i can be expressed using Bayes' Rule as

$$\begin{aligned} \Pr(i \to j \mid \text{fast in } i) &= \frac{\Pr(\text{fast in } i \mid i \to j)\Pr(i \to j)}{\Pr(\text{fast in } i)} && (3.58)\\ &= \frac{p_{ji}^{(f)}(\epsilon)\, p_{ji}(\epsilon)}{p_i^{(f)}(\epsilon)} && (3.59)\end{aligned}$$

Finally, since the holding time in state j is independent of the previous state,

$$\begin{aligned} p_{f_j f_i}(\epsilon) &= \Pr(i \to j, \text{fast in } j \mid \text{fast in } i) && (3.60)\\ &= \Pr(i \to j \mid \text{fast in } i)\Pr(\text{fast in } j) && (3.61)\\ &= p_{ji}^{(f)}(\epsilon)\, p_{ji}(\epsilon)\, \frac{p_j^{(f)}(\epsilon)}{p_i^{(f)}(\epsilon)} && (3.62)\end{aligned}$$

which is (3.25). The remaining equations for $p_{s_j f_i}(\epsilon)$, $p_{f_j s_i}(\epsilon)$ and $p_{s_j s_i}(\epsilon)$ follow similarly. Finally, the holding time in f_i conditioned on moving to either f_j or s_j is distributed as $h_{ji}^{(f)}(\epsilon,t)$ and in s_i as $h_{ji}^{(s)}(\epsilon,t)$.

The behavior of the original process can be recovered easily. The process $\eta(\epsilon,t)$ is in state i if $\eta_{\mathrm{E}}(\epsilon,t)$ is either in the fast copy, state f_i, or in the slow copy, state s_i. If $\eta(\epsilon,t)$ initially enters state i, then with probability $p_i^{(f)}(\epsilon)$ it will make a "fast" transition and with probability $p_i^{(s)}(\epsilon)$ it will make a "slow" transition. These two observations can be combined to form the following identity

$$\Phi(\epsilon,t) = L_{\mathrm{E}}\,\Phi_{\mathrm{E}}(\epsilon,t)\,R_{\mathrm{E}} \tag{3.63}$$

where

$$L_{\mathrm{E}} = \begin{bmatrix} 1 & \cdots & 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1 & 0 & \cdots & 1 \end{bmatrix} \tag{3.64}$$

and

$$R_{\mathrm{E}} = \begin{bmatrix} p_1^{(f)}(\epsilon) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & p_n^{(f)}(\epsilon) \\ p_1^{(s)}(\epsilon) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & p_n^{(s)}(\epsilon) \end{bmatrix} \tag{3.65}$$

and the states of $\eta_{\mathrm{E}}(\epsilon,t)$ are ordered $f_1,\ldots,f_n,s_1,\ldots,s_n$.

3.3.2 Markov representation

In this section we consider the Markov representation of a semi-Markov process with rational polynomial holding time transforms. An arbitrary semi-Markov process on a state set $\{1,2,\ldots,n\}$ is considered.[2] The essential step in representing the semi-Markov process as a Markov process on an expanded state space involves constructing a sequence of stages with exponential holding times for each of the holding time distributions, $h(t)$, using the "method of stages" [14] [27]. The basis

[2] For notational simplicity, we assume in this section that we are dealing with a process with a state set that has been relabeled as $1,\ldots,n$. In the full sequential procedure we would first generate the semi-Markov process with states $f_1,\ldots,f_n$ and $s_1,\ldots,s_n$ and then apply the procedure described in this section to that process.

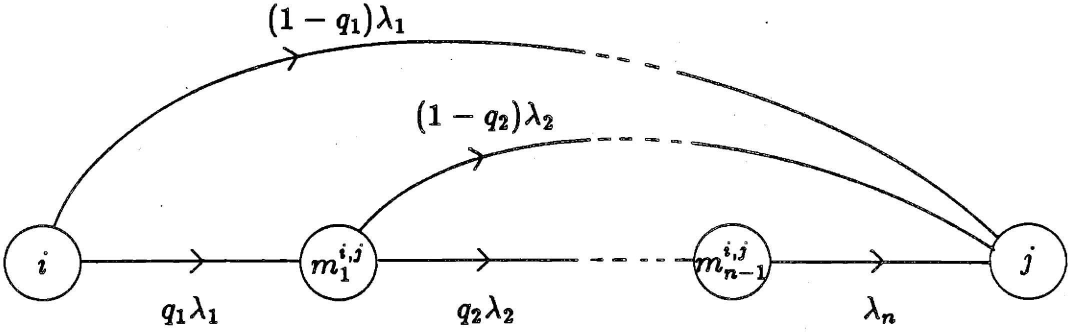


Figure 3.7: Method of stages expansion

of this approach is the observation that a rational polynomial transform with a denominator of degree n can be written as

$$\begin{aligned} H(s) = \frac{a(s)}{b(s)} &= \frac{a(s)}{\prod_{i=1}^{n}(s+\lambda_i)} && (3.66) \\ &= (1-q_1)\frac{\lambda_1}{(s+\lambda_1)} + q_1(1-q_2)\frac{\lambda_1}{(s+\lambda_1)}\frac{\lambda_2}{(s+\lambda_2)} + \cdots && (3.67) \\ &\quad + q_1 q_2 \cdots q_{n-1} \prod_{i=1}^{n} \left(\frac{\lambda_i}{s+\lambda_i}\right) \end{aligned}$$

which has the "Markov" stage form shown in Figure 3.7. The states added between the old states i and j are labeled $m_1^{i,j}, m_2^{i,j}, \ldots$. Also define the sets $\mathcal{M}(i,j) \equiv \{m_k^{i,j}, k = 1, 2, \ldots\}$ and $\mathcal{M}(i) \equiv \{i\} + \bigcup_j \mathcal{M}(i,j)$. It is well known, however, that the λ_i are not necessarily real. Cox [14] has shown how such complex probabilities can be manipulated by formally applying results for purely real rates. It will be shown in Section 3.A.2 that such complex probabilities in fact do not cause difficulties in applying the Markov decomposition algorithm to this expanded system.

Example 3.4 *Consider a holding time transform*

$$H(s) = \frac{2s^2 + 11s + 24}{4(s+1)(s+2)(s+3)} \qquad (3.68)$$

The method of stages form is

$$H(s) = \frac{1}{2}\frac{1}{s+1} + \frac{1}{2}\frac{1}{4}\frac{1}{s+1}\frac{2}{s+2} + \frac{1}{2}\frac{3}{4}\frac{1}{s+1}\frac{2}{s+2}\frac{3}{s+3} \qquad (3.69)$$

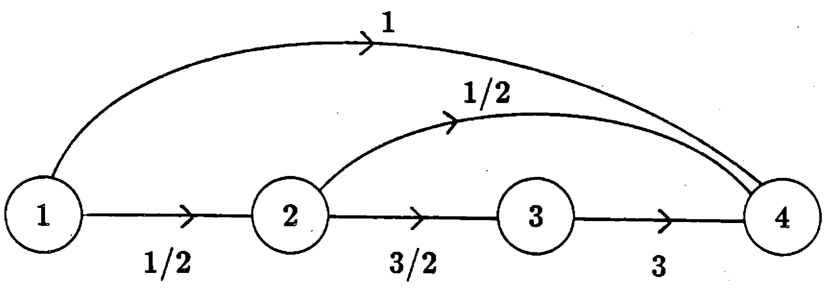


Figure 3.8: Method of stages in Example 3.4

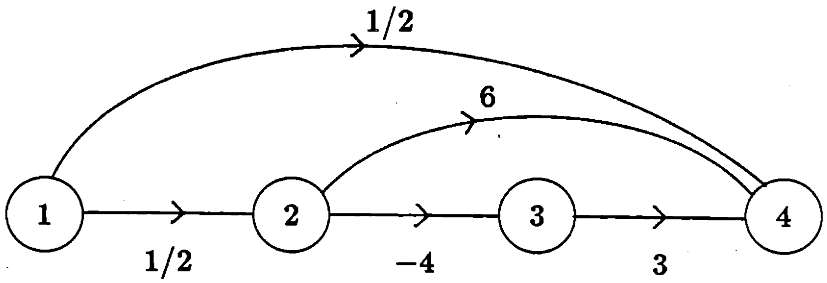


Figure 3.9: Method of stages in Example 3.5

which is illustrated in Figure 3.8. □

Example 3.5 *To illustrate that the probabilities in this type expansion are not necessarily positive (even when the poles of $H(s)$ are real), consider*

$$h(t) = \frac{1}{2}e^{-t} + \frac{1}{2}(2e^{-2t} * 3e^{-3t}) \tag{3.70}$$

$$H(s) = \frac{s^2 + 11s + 12}{2(s+1)(s+2)(s+3)} \tag{3.71}$$

The method of stages expansion of $H(s)$ is

$$H(s) = \frac{1}{2}\frac{1}{s+1} + \frac{1}{2}3\frac{1}{s+1}\frac{2}{s+2} + \frac{1}{2}(-2)\frac{1}{s+1}\frac{2}{s+2}\frac{3}{s+3} \tag{3.72}$$

which is shown in Figure 3.9. Note that $q_2 = -2 < 0$ is not a valid probability. □

Once each of the holding time distributions is represented as a sequence of exponential stages, these representations must be incorporated into a single, overall model. In order to do this, the first stage of each transition leaving a particular state must have a common exponential rate. Then, the rates leaving the initial state are weighted by the original state transition probabilities. For example, consider the situation illustrated in Figure 3.10(a) where

$$h_1(\epsilon,t) = \lambda_1 e^{-\lambda_1 t} \ , \ \ h_2(\epsilon,t) = \lambda_2 e^{-\lambda_2 t} \tag{3.73}$$

Each of the two holding time distributions leaving state i are exponentially distributed, but they have different rate constants. Therefore the unconditional holding time in state i is not exponentially distributed. Consequently, we do not have a Markovian representation. In order to obtain such a representation, we introduce a common root λ^* into the transforms of each of the two holding time transforms

$$H_1(s) = \frac{(s+\lambda^*)}{(s+\lambda^*)}\frac{\lambda_1}{(s+\lambda_1)} \ , \ \ H_2(s) = \frac{(s+\lambda^*)}{(s+\lambda^*)}\frac{\lambda_2}{(s+\lambda_2)} \tag{3.74}$$

This allows expansion of each of the holding times as in (3.67) as shown in Figure 3.10(b).

$$H_1(s) \ = \ (1-q_{11})\frac{\lambda^*}{(s+\lambda^*)} + q_{11}\frac{\lambda^*}{(s+\lambda^*)}\frac{\lambda_1}{(s+\lambda_1)} \ , \ \ q_{11} = 1 - \frac{\lambda_1}{\lambda^*} \tag{3.75}$$

$$H_2(s) \ = \ (1-q_{21})\frac{\lambda^*}{(s+\lambda^*)} + q_{21}\frac{\lambda^*}{(s+\lambda^*)}\frac{\lambda_2}{(s+\lambda_2)} \ , \ \ q_{21} = 1 - \frac{\lambda_2}{\lambda^*} \tag{3.76}$$

These two expansions can now be combined into the Markov representation by adding the effect of the transition probabilities p_{ji} and p_{ki} shown in Figure 3.10(c).

Though the term λ^* is arbitrarily chosen, it will be seen that its choice does not affect the final approximation result when expressed in semi-Markov form. As we will see, the basic reason that λ^* does not enter into the overall approximation is that the statistic needed in computing the parameters in the aggregate model at the next time scale is the mean time between entries into state i and not the steady state probability of occupying state i. The rate λ^* which was introduced only directly affects the holding time in the state and not the recurrence statistics.

Another feature of this expansion which will be used deals with the case in which $\sup_{t\geq 0}|h(\epsilon,t)| = O(\epsilon)$ but where there are $O(1)$ poles of $H(\epsilon,s)$. An argument which

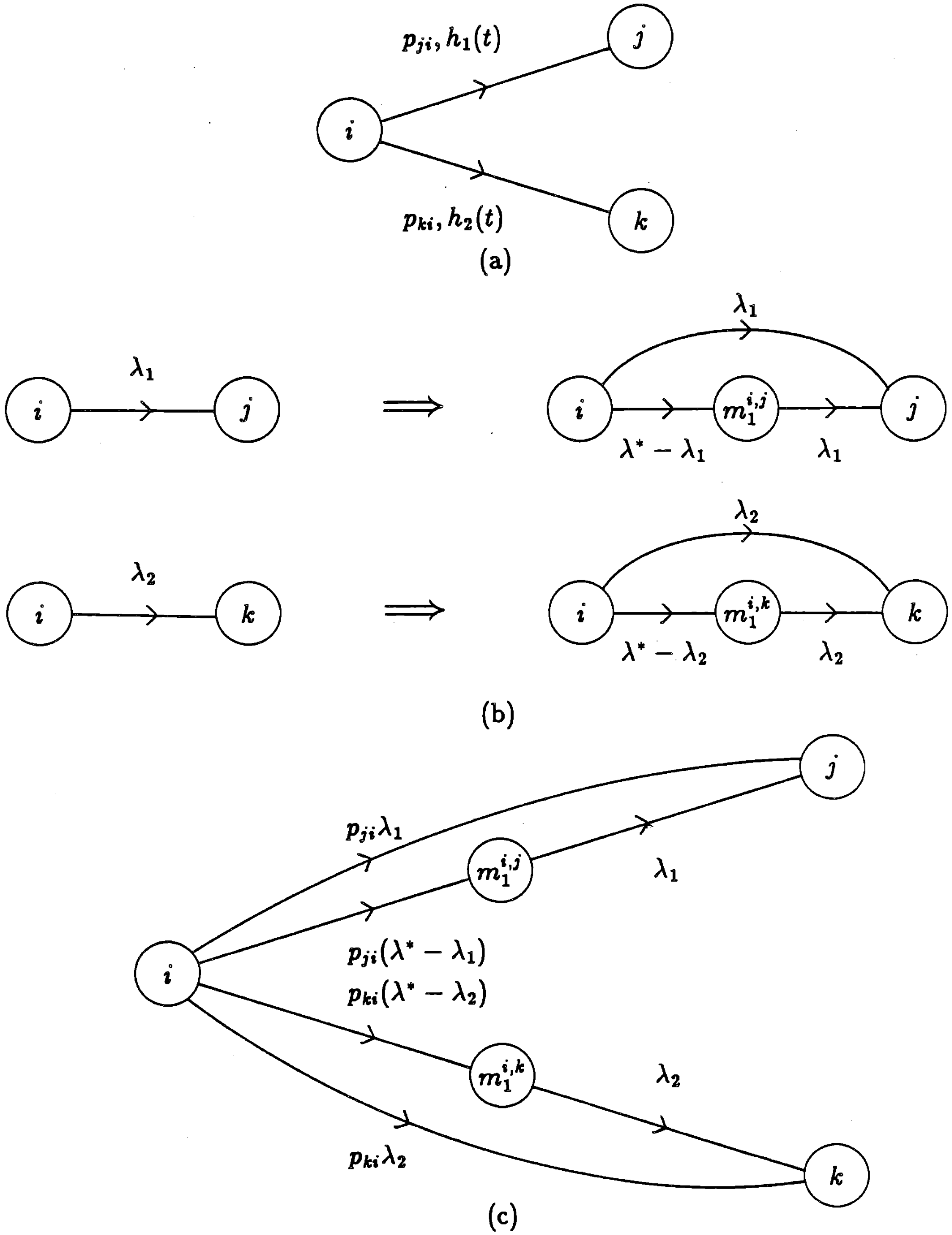


Figure 3.10: Combined method of stages

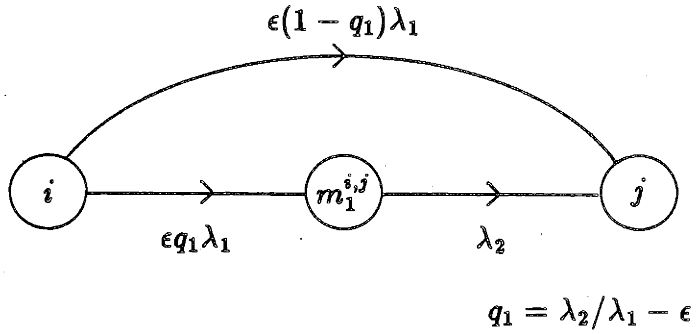

$\epsilon(1-q_1)\lambda_1$
i
$m_1^{i,j}$
j
$\epsilon q_1\lambda_1$
λ_2
$q_1 = \lambda_2/\lambda_1 - \epsilon$

Figure 3.11: Markov representation in Example 3.6

will be employed below is that if $\sup_{t\geq 0} |h(\epsilon, t| = \mathrm{O}(\epsilon)$ and if some of the denominator poles are $\mathrm{O}(1)$, then if the roots of $b(\epsilon, s)$ are ordered such that $\lambda_1, \ldots, \lambda_k$ are $\mathrm{O}(\epsilon)$ and $\lambda_{k+1}, \ldots, \lambda_n$ are $\mathrm{O}(1)$, then the states $m_k^{i,j}, m_{k+1}^{i,j}, \ldots, m_{n-1}^{i,j}$ introduced between states i and j are all transient at $\epsilon = 0$.

Example 3.6 *Consider the Markov representation of*

$$H_{ji}(\epsilon, s) = (1+\epsilon)\frac{\epsilon\lambda_1}{s+\epsilon\lambda_1} - \epsilon\frac{\lambda_2}{s+\lambda_2} \tag{3.77}$$

which is illustrated in Figure 3.11. Note that the state $m_1^{i,j}$ is necessarily transient at $\epsilon = 0$ since all the rates entering it are $\mathrm{O}(\epsilon)$ and there is a $\lambda_2 = \mathrm{O}(1)$ rate leaving it. □

Finally, let us explicitly state how to recover the behavior of the process $\eta_{\mathrm{E}}(\epsilon, t)$ from the Markov representation $\eta_{\mathrm{M}}(\epsilon, t)$. First, $\eta_{\mathrm{E}}(\epsilon, t)$ is in state i only if $\eta_{\mathrm{M}}(\epsilon, t)$ is in *any* of the states in the set $\mathcal{M}(i)$. Second, if $\eta_{\mathrm{E}}(\epsilon, t)$ has just entered state i, this exactly corresponds to the process $\eta_{\mathrm{M}}(\epsilon, t)$ entering state i as well. Using these two observations, we can write

$$\Phi_{\mathrm{E}}(\epsilon, t) = L_{\mathrm{M}}\, \Phi_{\mathrm{M}}(\epsilon, t)\, R_{\mathrm{M}} \tag{3.78}$$

where

$$R_{\mathrm{M}} = \begin{bmatrix} I \\ \mathbf{0} \end{bmatrix} \tag{3.79}$$

and

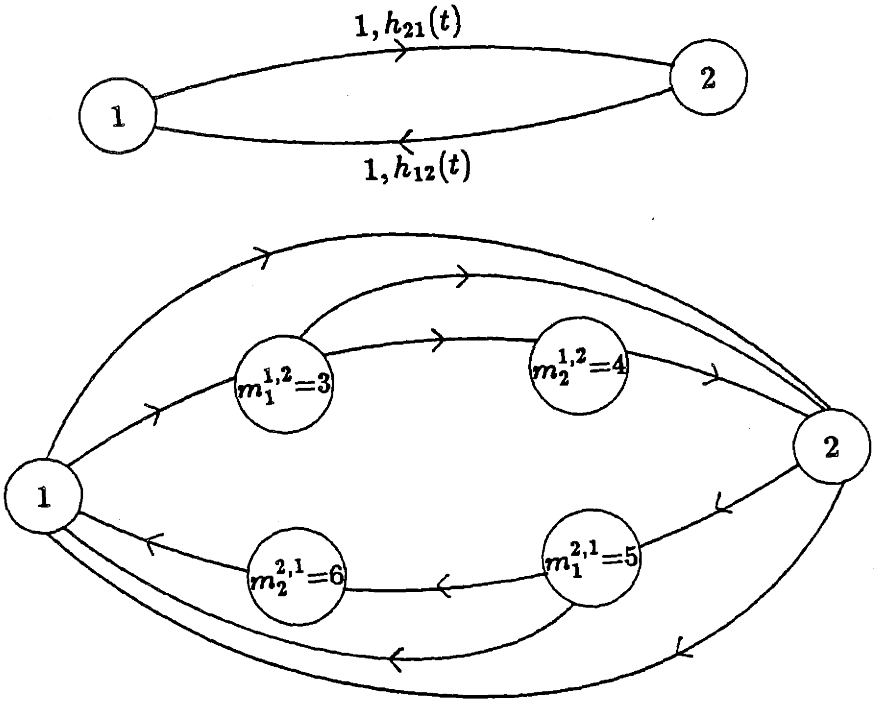


Figure 3.12: Semi-Markov and Markov representation in Example 3.7

$$L_{\mathrm{M}} = \begin{bmatrix} 1 & \cdots & 0 & \mathbf{1}^{\mathrm{T}} & \cdots & \mathbf{0}^{\mathrm{T}} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1 & \mathbf{0}^{\mathrm{T}} & \cdots & \mathbf{1}^{\mathrm{T}} \end{bmatrix} \tag{3.80}$$

and the states of $\eta_{\mathrm{M}}(\epsilon, t)$ are ordered $1, 2, \ldots, n, m_1^{1,2}, \ldots, m_{k_{1,2}}^{1,2}, \ldots, m_{k_{n,n-1}}^{n,n-1}$

Example 3.7 *In order to illustrate this reconstruction of a semi-Markov process from the Markov representation, consider the ϵ-independent semi-Markov process and its Markov representation shown in Figure 3.12. (Although the specific Markov rates are not indicated, it is assumed that $H_{12}(s)$ and $H_{21}(s)$ have three poles each and therefore have three stages each in the Markov representation). For this process,* *the behavior of $\eta(t)$ can be recovered from that of $\eta_{\mathrm{M}}(t)$ using*

$$\Phi(t) = \begin{bmatrix} 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 \end{bmatrix} \Phi_{\mathrm{M}}(t) \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \tag{3.81}$$

□

3.3.3 Decomposition of the Markov representation

In this section, the slow time scale approximation of the Markov representation is addressed. First, the general properties of the Markov process $\eta_{\mathrm{M}}(\epsilon, t)$ are discussed. Then, the construction of the slow time scale Markov process, $\tilde{\eta}_{\mathrm{M}}(\epsilon, t)$, is addressed under the assumption that the process $\eta_{\mathrm{M}}(\epsilon, t)$ is a valid Markov process, i.e. all the rates and probabilities are real and positive. In Section 3.A.2, the validity of the procedure is demonstrated in the general case in which the probabilities and transition rates resulting from the Markov expansion are complex.

Example 3.8 *In order to demonstrate some of the features of the Markov representation, consider the processes shown in Figure 3.13. The expanded semi-Markov process, $\eta_{\mathrm{E}}(\epsilon, t)$ is shown in Figure 3.13(a). The states are partitioned into the fast set $\mathcal{F} = \{1, 2\}$ and the slow set $\mathcal{S} = \{3\}$. The Markov realization is shown in Figure 3.13(b) and the slow time scale Markov model derived using the algorithm presented in Chapter 2 is shown in Figure 3.13(c).*

The holding time transforms for the semi-Markov process are

$$\begin{aligned} H_{21}(\epsilon, s) &= \frac{(1-\epsilon)s + 1}{(s+1)^2} && (3.82) \\ H_{12}(\epsilon, s) &= H_{32}(\epsilon, s) = \frac{1}{(s+1)^2} && (3.83) \\ H_{32}(\epsilon, s) &= \frac{\epsilon^2}{(s+\epsilon)(s+1)} && (3.84) \end{aligned}$$

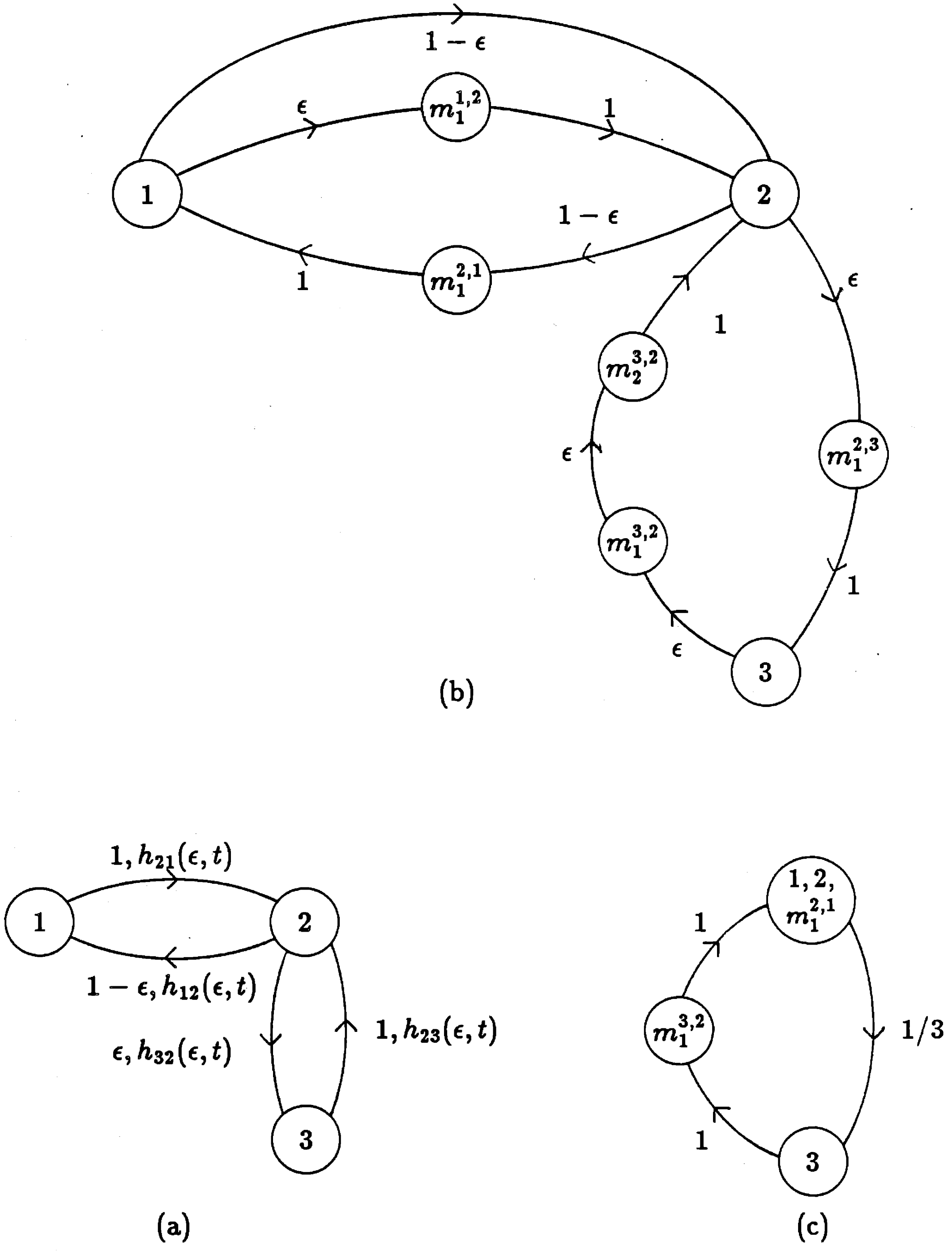


Figure 3.13: (a) Semi-Markov, $\eta_{\mathrm{E}}(\epsilon, t)$, (b) Markov, $\eta_{\mathrm{M}}(\epsilon, t)$), and (c) slow time scale Markov, $\tilde{\eta}_{\mathrm{M}}(\epsilon, t)$ processes in Example 3.8

The ergodic classes of $\eta_{\mathrm{E}}(0,t)$ are $\{1,2\} \subseteq \mathcal{F}$ and $\{3\} \subseteq \mathcal{S}$. Consider the class $\{1,2\}$. The set of states $\mathcal{M}(1) \cup \mathcal{M}(2) = \{1, m_1^{1,2}, 2, m_1^{2,1}, m_1^{2,3}\}$ contains one ergodic class $\{1,2,m_1^{2,1}\}$ and transient states $m_1^{1,2}$ and $m_1^{2,3}$. The ergodic class $\{3\}$ of $\eta_{\mathrm{E}}(0,t)$ results in two ergodic classes of $\eta_{\mathrm{M}}(0,t)$, namely $\{3\}$ and $\{m_1^{3,2}\}$, and one transient state $m_2^{3,2}$. The slow time scale, aggregated model associated with $\tilde{\eta}_{\mathrm{M}}(\epsilon,t)$ is shown in Figure 3.13(c). □

In order to discuss the slow time scale approximation of the Markov representation, it is useful to explicitly identify the various types of states which can occur. This characterization of the states of $\eta_{\mathrm{M}}(0,t)$ will simplify the discussion below. We assume that the semi-Markov process resulted from the expansion step described in Section 3.3.1 and that the states are partitioned into the fast set $\mathcal{F}$ and the slow set $\mathcal{S}$. In the discussion below, an ergodic class of $\eta_{\mathrm{M}}(0,t)$ is denoted by E and is associated with the class E' of $\eta_{\mathrm{E}}(0,t)$. Specifically, if $E' = \{i,j,\ldots\}$ then $E \subseteq \mathcal{M}(i) \cup \mathcal{M}(j) \cup \cdots$. This association is discussed further below.

- If a set of states $\{i,j,\ldots\} = E' \subseteq \mathcal{F}$ is an ergodic class of $\eta_{\mathrm{E}}(0,t)$, then the set $\mathcal{M}(i) \cup \mathcal{M}(j) \cup \cdots$ contains exactly one ergodic class E of $\eta_{\mathrm{M}}(0,t)$ and perhaps some transient states. In Example 3.8, the set $E' = \{1,2\}$ results in the ergodic class $E = \{1,2,m_1^{2,1}\} \subseteq \mathcal{M}(1) \cup \mathcal{M}(2)$ and the transient states $\{m_1^{1,2}, m_1^{2,3}\} \subseteq \mathcal{M}(1) \cup \mathcal{M}(2)$. An important feature of the associated ergodic classes E and E' is that the statistics of the class E' of $\eta_{\mathrm{E}}(0,t)$ can be derived from $\eta_{\mathrm{M}}(0,t)$. Specifically, the probability of occupying a state $i \in E'$ is the sum of the probabilities of occupying the states in $\mathcal{M}(i)$. Also, the inter-arrival time distribution of a state $i \in E'$ is the same as that for $i \in E$ in $\eta_{\mathrm{M}}(0,t)$.

- If state $i \in E_I \subseteq \mathcal{F}$ is recurrent and for some $j \notin E_I$, $p_{ji}(\epsilon) \neq 0$, then all the states in the set $\mathcal{M}(i,j)$ are transient. In the example above, state $m_1^{2,3} \in \mathcal{M}(2,3)$ is such a transient state. Note, however, that the terms in $V(\epsilon)$ associated with these transient states are particularly simple since the transient states are all uniquely associated with the ergodic class, as is state $m_1^{2,3}$ with class $\{3\}$ in the example.

- If state $i \in S$, then necessarily, $\{i\}$ is an ergodic class of $\eta_{\mathrm{M}}(0,t)$. The remaining states in the set $\mathcal{M}(i)$ are either also degenerate ergodic classes or are transient. In the example, $\{m_1^{3,2}\}$ forms such an ergodic class and the state $m_2^{3,2}$ is such a transient state. Recall that these transient states result from a holding time distribution which has a fast time scale component with very small probability associated with it.

- If a state $i \in \mathcal{F}$ is transient in $\eta_{\mathrm{E}}(0,t)$, then necessarily all the states in the set $\mathcal{M}(i)$ are also transient in the process $\eta_{\mathrm{M}}(0,t)$. Although this is not illustrated in the example, it follows from the argument that any state in $\mathcal{M}(i)$ can be visited no more frequently than state i and that state i has the same arrival statistics in $\eta_{\mathrm{M}}(0,t)$ as in $\eta_{\mathrm{E}}(0,t)$.

Consider now the terms required to compute the aggregated transition rates of the process $\tilde{\eta}_{\mathrm{M}}(\epsilon,t)$. Specifically, if the sets E_I and E_J are ergodic classes of $\eta_{\mathrm{M}}(0,t)$, and T is the transient set of $\eta_{\mathrm{M}}(0,t)$, then directly applying the Markov algorithm gives

$$\tilde{\lambda}_{JI}(\epsilon) = \frac{1}{\epsilon} \sum_{i\in E_I} \sum_{j\in E_J} u_{iI}\, \lambda_{ji}(\epsilon) + \frac{1}{\epsilon} \sum_{i\in E_I} \sum_{t\in T} u_{iI}\, \lambda_{ti}(\epsilon)\, v_{Jt}(\epsilon) \tag{3.85}$$

This expression is trivial to evaluate in the case when the class E_I consists of only a single state. We will therefore concentrate on the the case in which it is a larger subset of the fast states $\mathcal{F}$. In this case $E_I \subseteq \bigcup_{i\in E_I'} \mathcal{M}(i)$. If $E_J \neq E_I$ then from the construction, the nonzero terms $\lambda_{ji}(\epsilon)$ originate from the original states $i, j, \ldots$ and not from any of the states in a set of the form $\mathcal{M}(i,j)$. In the example, this means that any transition leaving the class $\{1, 2, m_1^{2,1}\}$ must leave through either state 1 or state 2 and *not* the intervening state $m_1^{2,1}$ introduced in the Markov representation.

In order to evaluate the sum (3.85), the ergodic probabilities u_{iI} are required. We first compute one such term in the example and then express the computation required in general. Consider state 2 in the example which is a member of the ergodic class $E_1 = \{1, 2, m_1^{2,1}\}$ of $\eta_{\mathrm{M}}(0,t)$. By considering this process at $\epsilon = 0$, it is evident that $u_{21} = 1/3$. One method of computing this quantity is as the quotient of the mean holding time in state 2 and the mean time between arrivals in that state. The mean holding time in state 2 is 1 and the mean time between arrivals is the sum of the mean holding times in states 2, $m_1^{2,1}$, and 1, which is 3. The fact to note is that the recurrence time for state 2 could have been computed using the statistics of $\eta_{\mathrm{E}}(0,t)$ as the sum of the mean holding time in state 1 and in state 2, which is also 3.

The general form of this computation can be expressed as follows. Consider an ergodic class E' of $\eta_{\mathrm{E}}(0,t)$. If ρ_i is the steady state probability of entering a state $i \in E'$ *on the next transition*, and $\bar{\tau}_i$ is the mean holding time in that state not conditioned on the next state to be visited, then the mean recurrence frequency can be computed as [21]

$$\bar{\mu}_i = \frac{\rho_i}{\sum_j \rho_j \bar{\tau}_j} \tag{3.86}$$

where $\bar{\tau}_j$ can be computed in a variety of ways including

$$\bar{\tau}_j \equiv \int_0^\infty t\, h_j(t)\, dt = (-\frac{d}{ds} H_j(s))_{s=0} \tag{3.87}$$

Now consider the state $i \in E'$ in the Markov process $\eta_{\mathrm{M}}(0,t)$. If we define the total exponential rate leaving state i in $\eta_{\mathrm{M}}(0,t)$ as λ_i^*, then the ergodic probability of the Markov state i can be expressed as

$$u_{iI} = \frac{\bar{\mu}_i}{\lambda_i^*} \tag{3.88}$$

Therefore, the terms u_{iI} required in the sum (3.85) can be evaluated using only simple terms from the semi-Markov process, the $\bar{\mu}_i$, and the exponential rates, λ_i^*, leaving that state in the Markov representation.

The sum (3.85) can therefore be written as

$$\tilde{\lambda}_{JI}(\epsilon) = \frac{1}{\epsilon} \sum_{i \in E_I} u_{iI}\, \lambda_i^* \left(\sum_{j \in E_J} \frac{\lambda_{ji}(\epsilon)}{\lambda_i^*} + \sum_{t \in T} \frac{\lambda_{ti}(\epsilon)}{\lambda_i^*} v_{Jt}(\epsilon) \right) \tag{3.89}$$

Recall that $u_{iI}\, \lambda_i^* = \bar{\mu}_{iI}$ which was computed above. Note that the terms $\lambda_{ji}(\epsilon)/\lambda_i^*$ can be interpreted as probabilities. In fact by construction,

$$\frac{\lambda_{ji}(\epsilon)}{\lambda_i^*} + \sum_{t \in \mathcal{M}(i,j)} \frac{\lambda_{ti}(\epsilon)}{\lambda_i^*} = p_{ji}(\epsilon) \tag{3.90}$$

where $p_{ji}(\epsilon)$ is the transition probability of the semi-Markov process $\eta_{\mathrm{E}}(\epsilon, t)$. Therefore the aggregated exponential rate can be completely computed in terms of quantities related to the semi-Markov process,

$$\tilde{\lambda}_{JI}(\epsilon) = \frac{1}{\epsilon} \sum_{i \in E'_I} \bar{\mu}_{iI} \left(\sum_{j \in E'_J} p_{ji}(\epsilon) + \sum_{t \in T'} p_{ti}(\epsilon) v_{Jt}(\epsilon) \right) \tag{3.91}$$

Note that as mentioned previously, λ_i^* no longer appears in the expression. Note that the term $\Lambda_I(\epsilon)$ defined in (3.32) is the total rate leaving the aggregate class I

$$\Lambda_I(\epsilon) = \sum_J \tilde{\lambda}_{JI}(\epsilon) \tag{3.92}$$

and the probabilities $\tilde{p}_{JI}(\epsilon)$ in (3.34) can be written as

$$\tilde{p}_{JI}(\epsilon) = \frac{\tilde{\lambda}_{JI}(\epsilon)}{\Lambda_I(\epsilon)} \tag{3.93}$$

Therefore the aggregated set E_I has an exponential holding time with a total rate $\Lambda_I(\epsilon)$ and transition probabilities $\tilde{p}_{JI}(\epsilon)$.

In terms of Example 3.8, consider the aggregated transition rate from class $E_1 = \{1, 2, m_1^{2,1}\}$ to class $E_2 = \{3\}$. The term $\tilde{\lambda}_{21}(\epsilon)$ can be computed using the Markov representation as

$$\tilde{\lambda}_{21} = \frac{1}{\epsilon} \sum_{i \in E_1} u_{i1} \left(\sum_{j \in E_2} \lambda_{ji}(\epsilon) + \sum_{t \in T} \lambda_{ti}(\epsilon) v_{2j}(\epsilon) \right) \tag{3.94}$$

which has only one nonzero term, $i = 2$, $t = m_1^{2,3}$.

$$\tilde{\lambda}_{21}(\epsilon) = \frac{1}{\epsilon} \frac{1}{3} \epsilon = \frac{1}{3} \tag{3.95}$$

This could have been expressed in terms of the set $E'_1 = \{1,2\}$, $E'_2 = \{3\}$, and $T' = \{\}$ of the process $\eta_{\mathrm{E}}(\epsilon, t)$ as

$$\tilde{\lambda}_{21}(\epsilon) = \frac{1}{\epsilon} \sum_{i \in E'_1} \bar{\mu}_{i1} \left(\sum_{j \in E'_2} p_{ji}(\epsilon) + \sum_{t \in T'} p_{ti}(\epsilon) v_{2j}(\epsilon) \right) \tag{3.96}$$

This again has only one nonzero term, $i = 2$, $j = 3$, which gives the same value for $\tilde{\lambda}_{21}(\epsilon)$ as above.

Another feature of the calculation of the slow time scale, aggregated transition rates involves the "slow transient" states such as state $m_2^{3,2}$ in Example 3.8. Specifically, we are concerned with transient states $m_k^{i,j}$ for which $i \in S$. By the construction outlined in the preceding section, we can assume that state $m_{k-1}^{i,j}$, (or state i if $k = 1$), is not transient and that all the states $m_{k'}^{i,j}$, $k' \geq k$ are also transient. It will be shown here that these transient states can effectively be "bypassed". The effect of considering this modified Markov process is identical to simply retaining the "slow" component of the holding time distribution. For example, the holding time distribution in Example 3.6

$$H_{ji}(\epsilon, s) = (1+\epsilon)\frac{\epsilon\lambda_1}{s+\epsilon\lambda_1} - \epsilon\frac{\lambda_2}{s+\lambda_2} \tag{3.97}$$

could be replaced by

$$H_{ji}(\epsilon, s) = \frac{\epsilon\lambda_1}{s+\epsilon\lambda_1} \tag{3.98}$$

In Example 3.8, the argument is that the Markov representation for the transition between state 3 and 2 can either be represented exactly as in Figure 3.14(a) or by bypassing the transient state $m_2^{3,2}$ as in (b). The holding time distribution of the stages in (a) is $h_{32}(\epsilon, t)$ while in (b), it is the slow component of $h_{32}(\epsilon, t)$, $h_{32}^{(s)}(\epsilon, t)$. By considering the Markov decomposition algorithm, we see that the behavior of the system using the bypassed form (b) is identical to that which has the original staged form (a). This justifies ignoring a fast component of a holding time distribution if $p^{(f)}(\epsilon) = O(\epsilon)$ as discussed in Section 3.3.1. The slow time scale Markov representation, $\tilde{\eta}_M(\epsilon, t)$ can now be constructed from the Markov representation. For the example above, this is shown in Figure 3.13(c).

The uniform asymptotic approximation of $\Phi_M(\epsilon, t)$ obtained using the Markov decomposition algorithm is

$$\Phi_M(\epsilon, t) = \Phi_M(0, t) + U_M\tilde{\Phi}_M(\epsilon, \epsilon t)V_M - U_M V_M + O(\epsilon) \tag{3.99}$$

In the next section, we will see that (3.99) can be expressed using a slow time scale *semi-Markov* process $\tilde{\eta}(\epsilon, t)$ instead of explicitly using $\tilde{\eta}_M(\epsilon, t)$ and that $\Phi_M(0, t)$ and that U_M and V_M do not have to be computed explicitly as well.

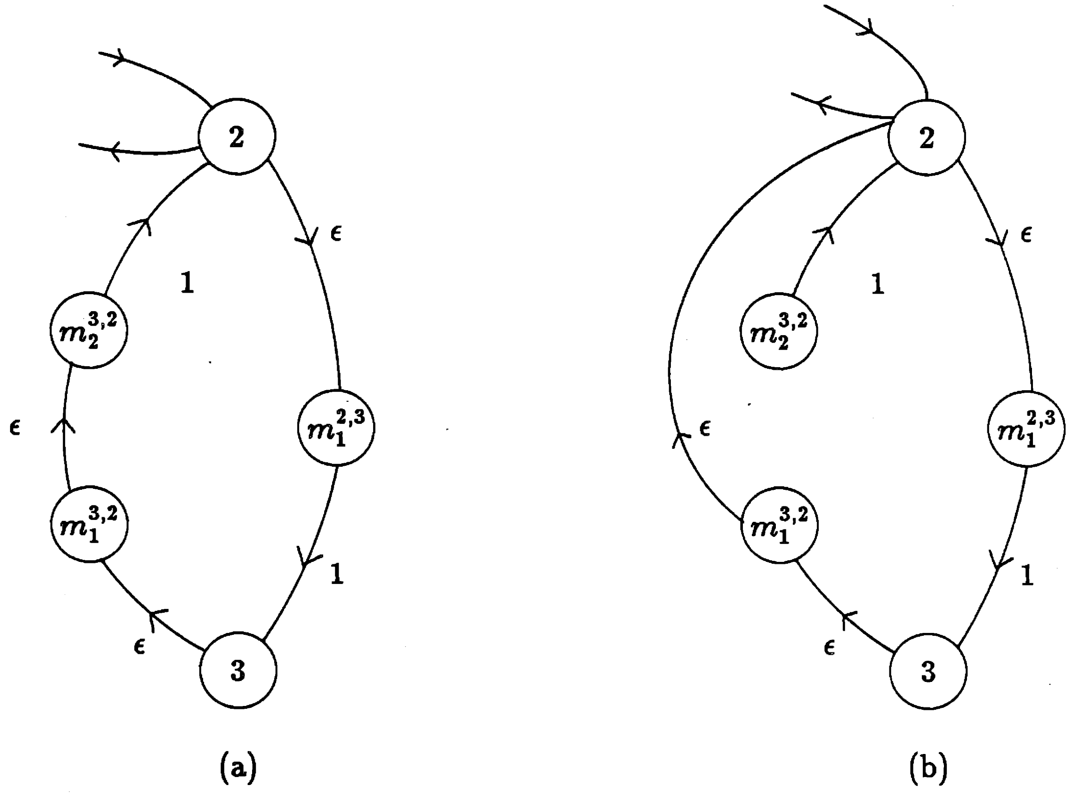


Figure 3.14: Bypassing transient states

3.3.4 Semi-Markov representation

To this point in the development, we have performed three transformations and approximations of the original semi-Markov process:

1. State expansion of the original semi-Markov process to form $\eta_{\mathrm{E}}(\epsilon, t)$. Associated with this expansion is the identity (3.63).
2. Markov representation using a method of stages expansion to form $\eta_{\mathrm{M}}(\epsilon, t)$. Associated with this representation is the identity (3.78).
3. Slow time scale approximation using the Markov algorithm presented in Chapter 2. This results in the asymptotic approximation (3.99).

These identities and approximations can be combined to approximate $\Phi(\epsilon, t)$

$$\begin{aligned}\Phi(\epsilon, t) &= L_{\mathrm{E}} L_{\mathrm{M}} \Phi_{\mathrm{M}}(0, t) R_{\mathrm{M}} R_{\mathrm{E}} \\ &\quad + L_{\mathrm{E}} L_{\mathrm{M}} U_{\mathrm{M}} \tilde{\Phi}_{\mathrm{M}}(\epsilon, \epsilon t) V_{\mathrm{M}} R_{\mathrm{M}} R_{\mathrm{E}} \\ &\quad - L_{\mathrm{E}} L_{\mathrm{M}} U_{\mathrm{M}} V_{\mathrm{M}} R_{\mathrm{M}} R_{\mathrm{E}} + \mathrm{O}(\epsilon)\end{aligned} \tag{3.100}$$

Consider the first term. By following the construction, it follows simply that

$$L_{\mathrm{E}} L_{\mathrm{M}} \Phi_{\mathrm{M}}(0, t) R_{\mathrm{M}} R_{\mathrm{E}} = \Phi(0, t) \tag{3.101}$$

By considering $\lim_{t\to\infty} \Phi_{\mathrm{M}}(0, t) = U_{\mathrm{M}} V_{\mathrm{M}}$ where U_{M} and V_{M} are the ergodic probability and membership matrices computed from the original semi-Markov process, the third term becomes

$$- L_{\mathrm{E}} L_{\mathrm{M}} U_{\mathrm{M}} V_{\mathrm{M}} R_{\mathrm{M}} R_{\mathrm{E}} = - \lim_{t\to\infty} \Phi(0, t) = -P \tag{3.102}$$

where P is the ergodic projection of $\eta(0, t)$.

The second term

$$L_{\mathrm{E}} \left(L_{\mathrm{M}} U_{\mathrm{M}} \tilde{\Phi}_{\mathrm{M}}(\epsilon, \epsilon t) V_{\mathrm{M}} R_{\mathrm{M}} \right) R_{\mathrm{E}} \tag{3.103}$$

can also be expressed in terms of the behavior of a semi-Markov process. We will first consider the parenthesized term in (3.103). Recall from the discussion above, that the states of $\tilde{\eta}_{\mathrm{M}}(\epsilon, t)$ are either made up of a single "slow" state $j \in \mathcal{M}(i)$ of

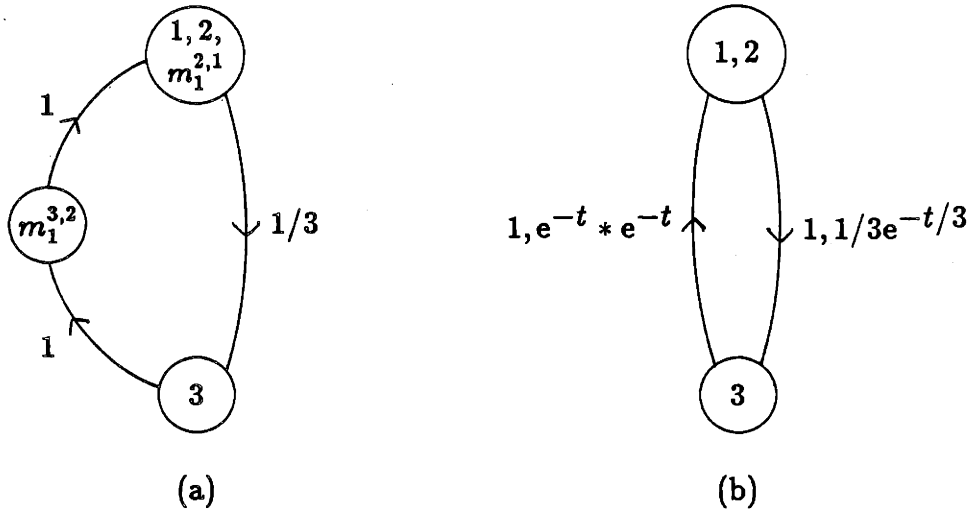


Figure 3.15: Slow time scale Markov and semi-Markov models in Example 3.8

$\eta_{\mathrm{M}}(\epsilon, t)$ where $i \in S$, or the aggregation of an ergodic class of $\eta_{\mathrm{M}}(0, t)$ made up of "fast" states. Due to the structure of R_{M}, the only terms in $\tilde{\Phi}_{\mathrm{M}}(\epsilon, \epsilon t)$ which enter into the product in parentheses in (3.103) correspond to initially being in either such an aggregated class, or in a slow state i which is also a state of the semi-Markov process $\eta_{\mathrm{E}}(\epsilon, t)$. Similarly, due to the structure of L_{M}, we are only interested in the probability of being in some set $\mathcal{M}(i)$ and not in any particular state of $\tilde{\eta}_{\mathrm{M}}(\epsilon, \epsilon t)$.

Consider the slow time scale Markov model for Example 3.8. First, we know that we are not interested in the transition probabilities from state $m_1^{3,2}$ since interaggregate transitions must leave through either state 1 or state 2. Second, the probabilities of being in states 3 and $m_1^{3,2}$ are always combined. In this example, it is easy to demonstrate that the terms of interest could be recovered from the semi-Markov process shown in Figure 3.15.

This observation that the terms in $\tilde{\Phi}_{\mathrm{M}}(\epsilon, t)$ which affect the product (3.103) can be computed using $\tilde{\Phi}(\epsilon, t)$, the probability transition function of a reduced order semi-Markov process, is in fact quite general. Although a complete proof would be tedious, the generality of this step can be understood as follows. A slow holding time distribution for the transition from state i to state j in the process $\eta_{\mathrm{E}}(\epsilon, t)$ is expanded using the method of stages to produce a portion of the Markov model of $\eta_{\mathrm{M}}(\epsilon,t)$. States $i, m_1^{i,j}, m_2^{i,j}, \ldots$ are all trapping states of the process $\eta_{\mathrm{M}}(0,t)$ and therefore each becomes a separate aggregate class in the slow time scale process $\tilde{\eta}_{\mathrm{M}}(\epsilon,t)$. In the procedure to compute the transition rates of $\tilde{\eta}_{\mathrm{M}}(\epsilon,t)$, all the Markov transition rates are scaled by $1/\epsilon$. This results in the Markov representation still having the same form but scaled rates. This set of stages is again the staged form of a semi-Markov holding time distribution in which time is scaled by $1/\epsilon$ or a holding time transform where s is scaled by ϵ. Consequently, the collapsing of these stages back to a semi-Markov form is immediate.

Finally, the term which is needed can be expressed as

$$L_{\mathrm{E}}U_{\mathrm{E}}\tilde{\Phi}(\epsilon,\epsilon t)V_{\mathrm{E}}R_{\mathrm{E}} \tag{3.104}$$

Define $U \equiv L_{\mathrm{E}}U_{\mathrm{E}}$ and $V \equiv V_{\mathrm{E}}R_{\mathrm{E}}$. These are precisely the terms U and V which are constructed in the algorithm. The ergodic projection of $\eta(0,t)$ can then be expressed as $P = UV$ and therefore the overall approximation in the algorithm is uniformly valid.

Therefore, by identifying the slow time scale semi-Markov process, the series of transformations can be avoided. The algorithm provides a direct method of obtaining the parameters of this process. A complete example is considered in the next section in order to clarify the steps in Algorithm 3.1 which avoid explicit construction of the representations discussed above.

3.4 Example

In this section, Algorithm 3.1 is applied to a simple semi-Markov process to illustrate the steps involved. The initial process, illustrated in Figure 3.16, has three states and the following parameters:

$$p_{21}(\epsilon) = 1\,, \quad h_{21}(\epsilon,t) = \mathrm{e}^{-t} * 2\mathrm{e}^{-2t} \tag{3.105}$$

$$p_{12}(\epsilon) = 1-\epsilon\,, \quad h_{12}(\epsilon,t) = (1-\epsilon)\mathrm{e}^{-t} + \epsilon^2\mathrm{e}^{-\epsilon t} \tag{3.106}$$

$$p_{32}(\epsilon) = \epsilon\,, \quad h_{32}(\epsilon,t) = \epsilon\mathrm{e}^{-t} + (1-\epsilon)\mathrm{e}^{-\epsilon t} \tag{3.107}$$

$$p_{23}(\epsilon) = 1\,, \quad h_{23}(\epsilon,t) = \epsilon^2\mathrm{e}^{-\epsilon t} * \mathrm{e}^{-\epsilon t} \tag{3.108}$$

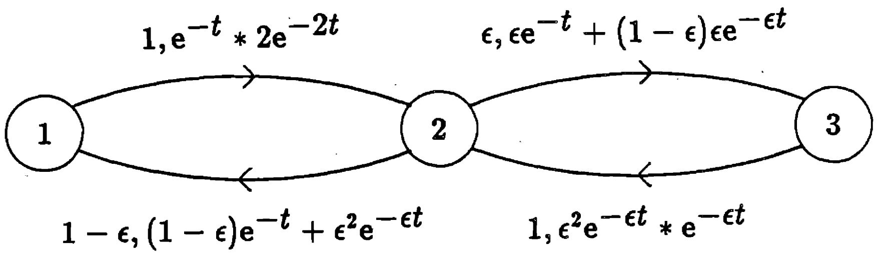

$1, e^{-t} * 2e^{-2t}$
$\epsilon, \epsilon e^{-t} + (1-\epsilon)\epsilon e^{-\epsilon t}$
1
2
3
$1-\epsilon, (1-\epsilon)e^{-t} + \epsilon^2 e^{-\epsilon t}$
$1, \epsilon^2 e^{-\epsilon t} * e^{-\epsilon t}$

Figure 3.16: Perturbed semi-Markov process

Step 1 of Algorithm 3.1 is to decompose each holding time distribution into its fast and slow components as in (3.18). The result of this procedure is:

$$p_{21}^{(f)}(\epsilon) = 1\ , \quad h_{21}^{(f)}(\epsilon, t) = e^{-t} * 2e^{-2t} \tag{3.109}$$

$$p_{21}^{(s)}(\epsilon) = 0 \tag{3.110}$$

$$p_{12}^{(f)}(\epsilon) = (1-\epsilon)\ , \quad h_{12}^{(f)}(\epsilon, t) = e^{-t} \tag{3.111}$$

$$p_{12}^{(s)}(\epsilon) = \epsilon\ , \quad h_{12}^{(s)}(\epsilon, t) = \epsilon e^{-\epsilon t} \tag{3.112}$$

$$p_{32}^{(f)}(\epsilon) = 0 \tag{3.113}$$

$$p_{32}^{(s)}(\epsilon) = 1\ , \quad h_{32}^{(s)}(\epsilon, t) = \epsilon e^{-\epsilon t} \tag{3.114}$$

$$p_{23}^{(f)}(\epsilon) = 0 \tag{3.115}$$

$$p_{23}^{(s)}(\epsilon) = 1\ , \quad h_{23}^{(s)}(\epsilon, t) = \epsilon^2 e^{-\epsilon t} * e^{-\epsilon t} \tag{3.116}$$

Note that the holding time from state 2 to 3 has on $O(\epsilon)$ fast component therefore $p_{32}^{(f)}(\epsilon)$ can be set to zero. The unconditional probability of making fast and slow transitions out of each state are then computed using (3.23)-(3.24).

$$p_1^{(f)}(\epsilon) = 1\ , \quad p_1^{(s)}(\epsilon) = 0 \tag{3.117}$$

$$p_2^{(f)}(\epsilon) = (1-\epsilon)^2\ , \quad p_2^{(s)}(\epsilon) = 2\epsilon - \epsilon^2 \tag{3.118}$$

$$p_3^{(f)}(\epsilon) = 0\ , \quad p_3^{(s)}(\epsilon) = 1 \tag{3.119}$$

This allow us to compute the transition probabilities and holding time distributions of the expanded process $\eta_E(\epsilon, t)$. The model of this process is illustrated in Figure 3.17. The nonzero transition probabilities computed using (3.25)-(3.28) are:

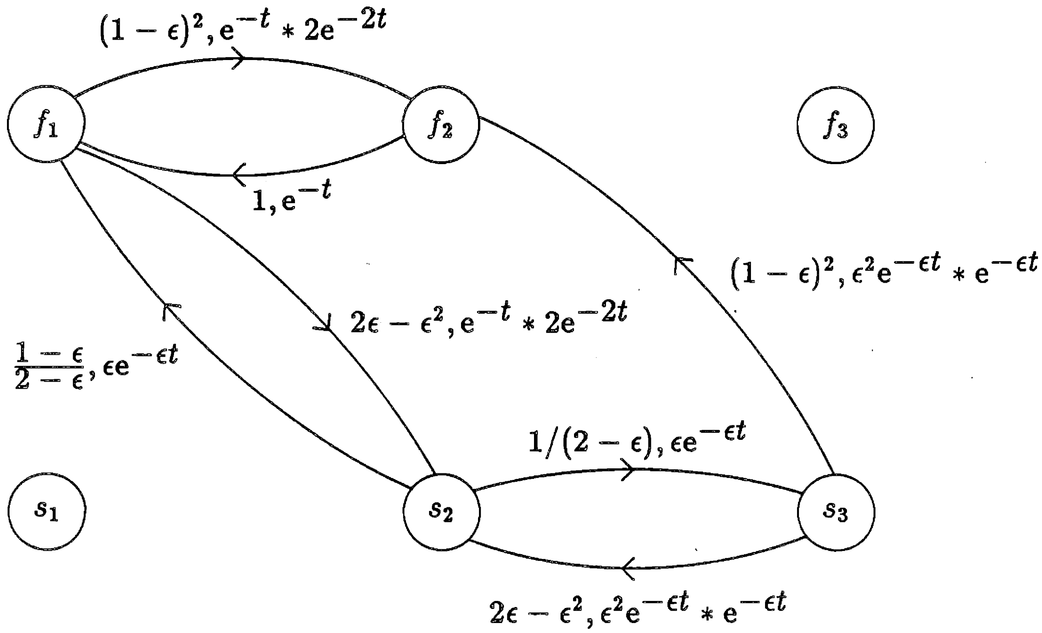

$(1-\epsilon)^2, \mathrm{e}^{-t} * 2\mathrm{e}^{-2t}$
f_1
f_2
f_3
$1, \mathrm{e}^{-t}$
$(1-\epsilon)^2, \epsilon^2\mathrm{e}^{-\epsilon t} * \mathrm{e}^{-\epsilon t}$
$2\epsilon-\epsilon^2, \mathrm{e}^{-t} * 2\mathrm{e}^{-2t}$
$\frac{1-\epsilon}{2-\epsilon}, \epsilon\mathrm{e}^{-\epsilon t}$
$1/(2-\epsilon), \epsilon\mathrm{e}^{-\epsilon t}$
s_1
s_2
s_3
$2\epsilon-\epsilon^2, \epsilon^2\mathrm{e}^{-\epsilon t} * \mathrm{e}^{-\epsilon t}$

Figure 3.17: Expanded semi-Markov representation

$$\begin{aligned}
p_{f_2 f_1}(\epsilon) &= (1-\epsilon)^2 && (3.120)\\
p_{s_2 f_1}(\epsilon) &= 2\epsilon - \epsilon^2 && (3.121)\\
p_{f_1 f_2}(\epsilon) &= 1 && (3.122)\\
p_{f_1 s_2}(\epsilon) &= (1-\epsilon)/(2-\epsilon) && (3.123)\\
p_{s_3 s_2}(\epsilon) &= 1/(2-\epsilon) && (3.124)\\
p_{f_2 s_3}(\epsilon) &= (1-\epsilon)^2 && (3.125)\\
p_{s_2 s_3}(\epsilon) &= 2\epsilon - \epsilon^2 && (3.126)
\end{aligned}$$

and the holding time distributions are determined using (3.29)-(3.30).

$$\begin{aligned}
h_{f_2 f_1}(\epsilon,t) = h_{s_1 f_1}(\epsilon,t) &= \mathrm{e}^{-t} * 2\mathrm{e}^{-2t} && (3.127)\\
h_{f_1 f_2}(\epsilon,t) &= \mathrm{e}^{-t} && (3.128)\\
h_{f_1 s_2}(\epsilon,t) &= \epsilon \mathrm{e}^{-\epsilon t} && (3.129)\\
h_{s_3 s_2}(\epsilon,t) &= \epsilon \mathrm{e}^{-t} + (1-\epsilon)\epsilon \mathrm{e}^{-\epsilon t} && (3.130)\\
h_{f_2 s_3}(\epsilon,t) = h_{s_2 s_3}(\epsilon,t) &= \epsilon^2 \mathrm{e}^{-\epsilon t} * \mathrm{e}^{-\epsilon t} && (3.131)
\end{aligned}$$

The ergodic classes and transient states of the expanded process $\eta_{\mathrm{E}}(0,t)$ can now be identified using the transition probabilities computed above. The ergodic classes are

$$\begin{aligned}
E_1 &= \{f_1, f_2\} && (3.132)\\
E_2 &= \{s_2\} && (3.133)\\
E_3 &= \{s_3\} && (3.134)\\
E_4 &= \{s_1\} && (3.135)\\
E_5 &= \{f_3\} && (3.136)
\end{aligned}$$

and there are no transient states. Note that the classes E_4 and E_5 form completely decoupled chains, even for $\epsilon > 0$, and therefore they can be completely ignored as they have no influence on the aggregation of the remainder of the process (although they could be carried through the entire algorithm). Since there are no transient states, the computation of the trapping probabilities v_{E} in step 4 is straightforward

$$v_{\mathrm{E}Ij} = 1 \iff j \in E_I \qquad (3.137)$$

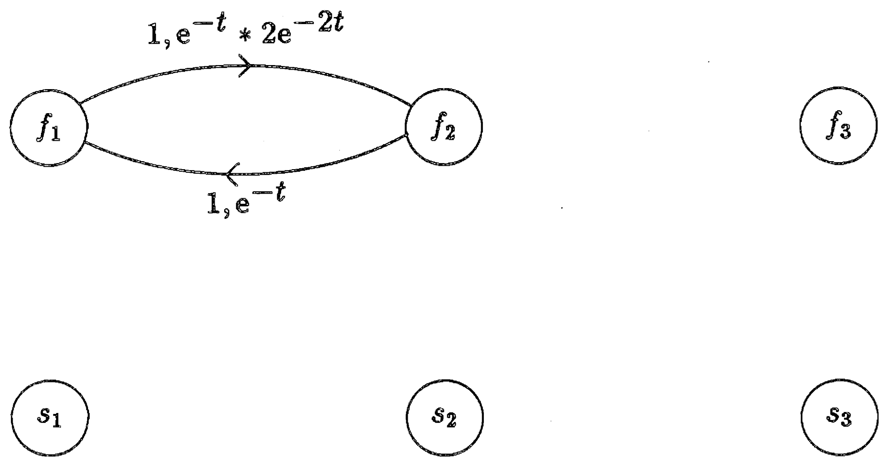


Figure 3.18: Expanded semi-Markov representation, $\epsilon = 0$

Only the class E_1 is composed of a set of "fast" states, therefore in step 5, only Λ_1 needs to be computed. By considering the process $\eta_{\mathrm{E}}(0,t)$, shown in Figure 3.18, the mean holding times in these states are

$$\bar{\tau}_{f_1} = 3 \tag{3.138}$$

$$\bar{\tau}_{f_2} = 1 \tag{3.139}$$

and the inverse of the mean time between arrivals in these states are

$$\bar{\mu}_{f_1} = 1/4 \tag{3.140}$$

$$\bar{\mu}_{f_2} = 1/4 \tag{3.141}$$

It should be noted that these parameters of a "fast" ergodic class could just as easily have been computed using the original process $\eta(0,t)$ since the class $\{1,2\}$ of that process must have the same statistics as the class $\{f_1, f_2\}$ of the expanded process $\eta_{\mathrm{E}}(0,t)$. The exponential rate leaving E_1 can then be computed using (3.32) as

$$\Lambda_1(\epsilon) = \bar{\mu}_{f_1} p_{s_2 f_1}(\epsilon) = \epsilon/2 - \epsilon^2/4 \tag{3.142}$$

The slow time scale semi-Markov process can now be constructed as in step 6 of the algorithm. This process is illustrated in Figure 3.19. The holding time leaving

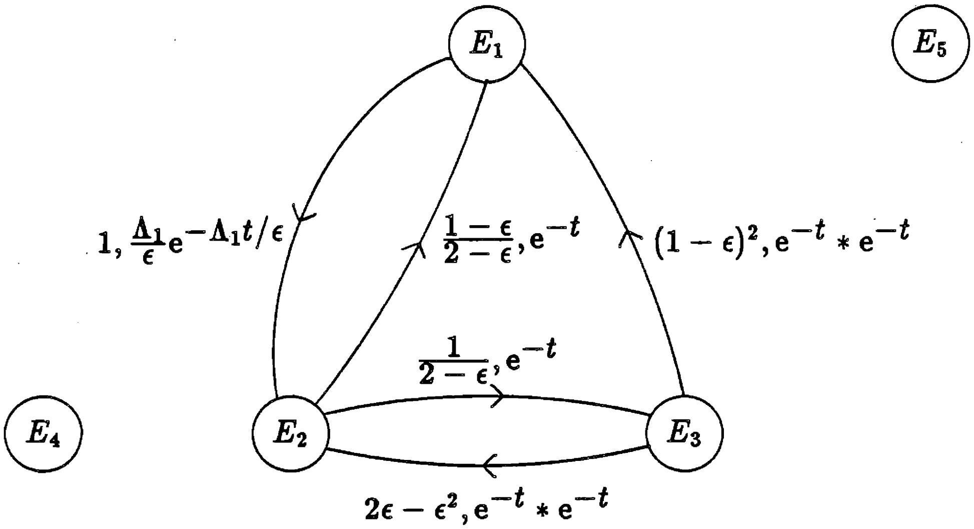

E_1
E_5
$1, \frac{\Lambda_1}{\epsilon} e^{-\Lambda_1 t/\epsilon}$
$\frac{1-\epsilon}{2-\epsilon}, e^{-t}$
$(1-\epsilon)^2, e^{-t} * e^{-t}$
$\frac{1}{2-\epsilon}, e^{-t}$
E_4
E_2
E_3
$2\epsilon - \epsilon^2, e^{-t} * e^{-t}$

Figure 3.19: $O(1/\epsilon)$ time scale semi-Markov process

the aggregate class E_1 is exponentially distributed with the form (3.33)

$$\tilde{h}_{21}(\epsilon, t) = \frac{1}{\epsilon}\Lambda_1(\epsilon)e^{-\Lambda_1(\epsilon)t/\epsilon} = (1/2 - \epsilon/4)e^{-(1/2 - \epsilon/4)t} \tag{3.143}$$

while the distributions leaving the other classes are directly derived from the distributions in $\eta(\epsilon, t)$

$$\begin{aligned} \tilde{h}_{12}(\epsilon, t) &= e^{-t} & (3.144)\\ \tilde{h}_{32}(\epsilon, t) &= e^{-t} & (3.145)\\ \tilde{h}_{13}(\epsilon, t) &= e^{-t} * e^{-t} & (3.146)\\ \tilde{h}_{23}(\epsilon, t) &= e^{-t} * e^{-t} & (3.147) \end{aligned}$$

The transition probabilities of this process are

$$\begin{aligned} \tilde{p}_{21}(\epsilon) &= 1 & (3.148)\\ \tilde{p}_{12}(\epsilon) &= (1-\epsilon)/(2-\epsilon) & (3.149)\\ \tilde{p}_{32}(\epsilon) &= 1/(2-\epsilon) & (3.150)\\ \tilde{p}_{13}(\epsilon) &= (1-\epsilon)^2 & (3.151)\\ \tilde{p}_{23}(\epsilon) &= 2\epsilon - \epsilon^2 & (3.152) \end{aligned}$$

The matrices U and V can be constructed as in step 7

$$U = \begin{bmatrix} 3/4 & 0 & 0 \\ 1/4 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad V = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3.153}$$

The procedure is terminated after one iteration since $\tilde{\eta}(0,t)$ has only one ergodic class. The overall approximation therefore has the form

$$\begin{aligned} \Phi(\epsilon,t) &= \Phi(0,t) + \begin{bmatrix} 3/4 & 0 & 0 \\ 1/4 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tilde{\bar{\Phi}}(0,\epsilon t) \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \\ &\quad \begin{bmatrix} 3/4 & 3/4 & 0 \\ 1/4 & 1/4 & 0 \\ 0 & 0 & 1 \end{bmatrix} + O(\epsilon) \end{aligned} \tag{3.154}$$

The processes $\eta(0,t)$ and $\tilde{\eta}(0,t)$ are illustrated in Figure 3.20.

In order to demonstrate that the same result could be obtained by explicitly constructing the Markov representation and applying Algorithm 2.1, consider the process $\eta_{\mathrm{M}}(\epsilon,t)$ illustrated in Figure 3.21. Applying Algorithm 2.1 results in the process $\tilde{\eta}_{\mathrm{M}}(\epsilon,t)$ illustrated in Figure 3.22. Note that the same process would result if the Markov representation of $\tilde{\eta}_{\mathrm{E}}(\epsilon,t)$ were constructed as shown in Figure 3.19.

3.5 Discussion

The decomposition algorithm presented in this chapter makes several contributions. First, although the details are relatively complicated compared to the purely Markov case, the algorithm provides a computationally feasible and easily interpreted method of decomposing semi-Markov processes. In the case where the holding time distributions do not depend on ϵ, the result is very similar to that provided by Korolyuk [31]. The more general case with perturbed holding time distributions shows that in general, a semi-Markov process is needed to recover the slow time scale behavior of the original system.

A novel feature of this algorithm not found in the Markov case is that in the slow time scale model, the fast and slow copies of a state may be combined into

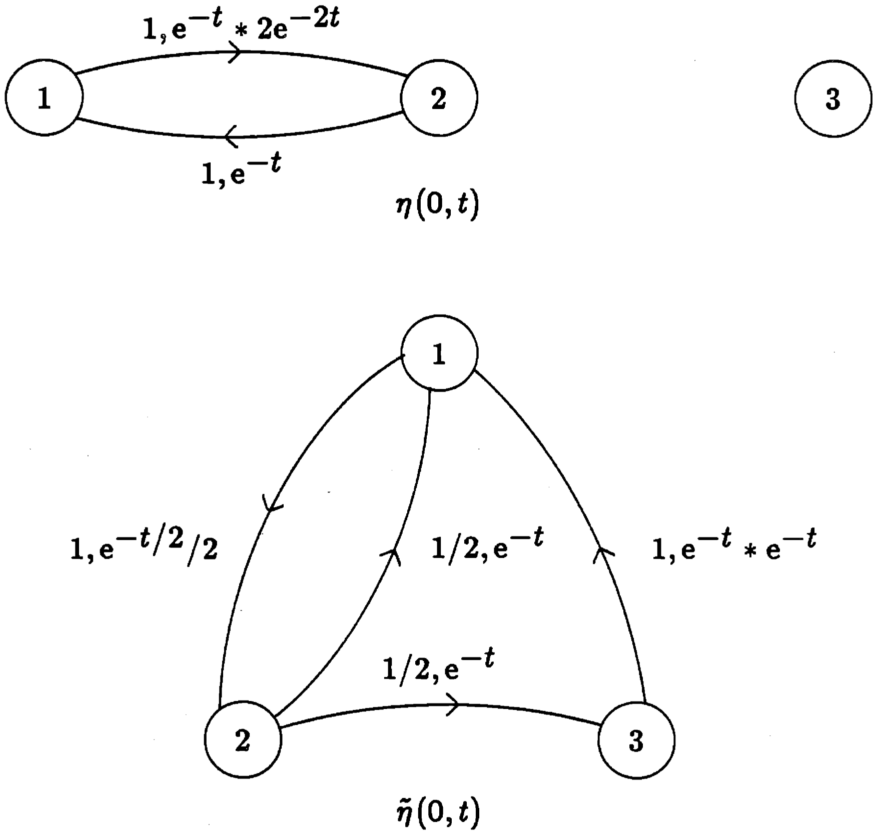


Figure 3.20: Multiple time scale semi-Markov models

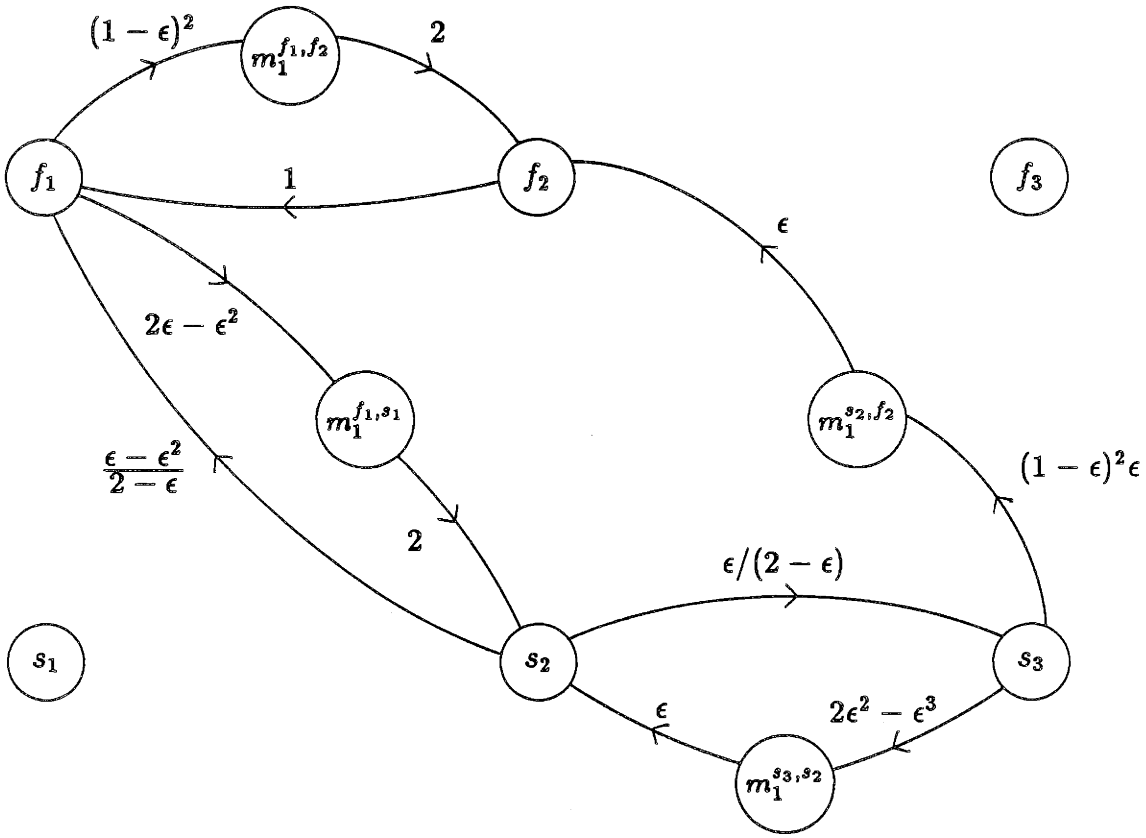


Figure 3.21: Markov representation

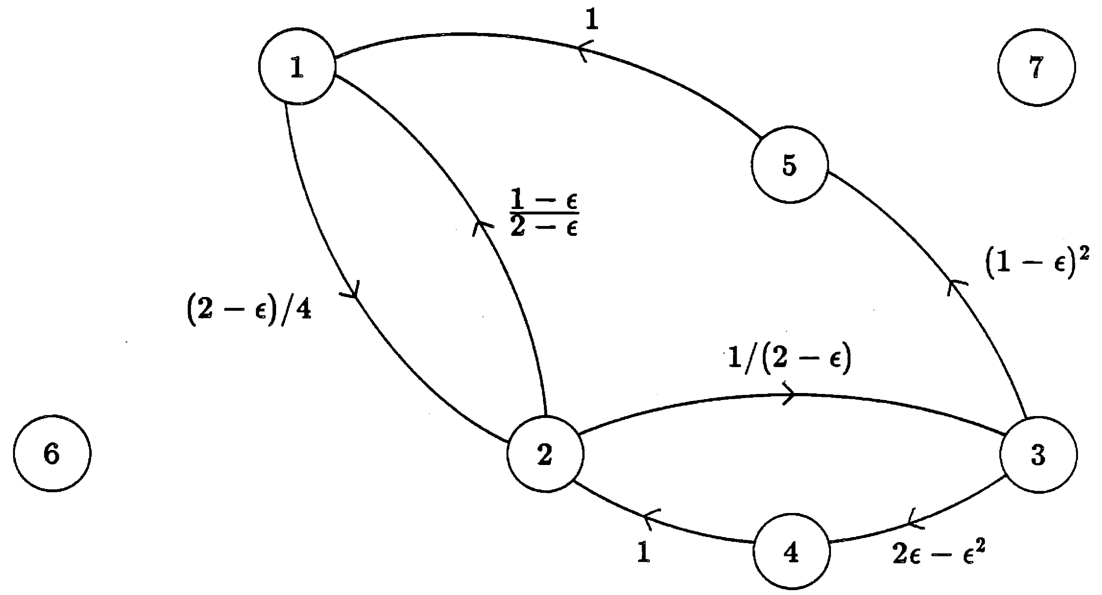


Figure 3.22: $O(1/\epsilon)$ time scale Markov representation

different aggregate classes. Although this may seem paradoxical, one interpretation of this results is as follows. Recall that the state (in the sense of memory) of a semi-Markov process is the observed state currently occupied and the time already spent in that state. Aggregation in some sense partitions this infinite state space. However, the process is again represented in terms of the finite number of observable states which can be visited. In this light, it is not as surprising that the parts of an observed state can be aggregated into different classes.

A restriction imposed by the method of proof — deriving the finite state Markov equivalent and applying the previously derived algorithm — is that the class of semi-Markov processes which can be considered must be restricted to those with rational polynomial holding time transforms. This restriction seems to be a limitation of the proof technique and not of the basic algorithm. An arbitrary holding time distribution can be decomposed into components which are fast and slow. The definitions of the "slow component" would have to be changed. For instance, a slow distribution is one which uniformly approaches 0 over any finite interval. Essentially, the limit has all the probability at infinity. An extension of the algorithm to this more general case would be useful in practice where a rational polynomial form is not sufficient. For instance, holding times with uniform distributions cannot be represented in this way.

3.A Appendix

3.A.1 Proof of Lemma 3.1

The decomposition of $h(\epsilon, t)$ into "fast" and "slow" terms follows from a partial fraction expansion of the Laplace transform.

$$H(\epsilon, s) = \frac{a(\epsilon, s)}{b(\epsilon, s)} = \frac{a(\epsilon, s)}{b^{(f)}(\epsilon, s)\, b^{(s)}(\epsilon, s)} = \frac{c^{(f)}(\epsilon, s)}{b^{(f)}(\epsilon, s)} + \frac{c^{(s)}(\epsilon, s)}{b^{(s)}(\epsilon, s)} \tag{3.155}$$

Let $b(\epsilon, s)$ have degree $n+1$. For $h(\epsilon, t)$ to be a valid probability density function, $a(\epsilon, s)$ has a maximum degree n. Let $b^{(f)}(\epsilon, s)$ have degree $n-m+1$ and $b^{(s)}(\epsilon, s)$ have degree m. The coefficients (expressed as vectors) of the polynomials $c^{(f)}(\epsilon, s)$ and $c^{(s)}(\epsilon, s)$ are of degree at most $n-m$ and $m-1$ respectively and satisfy the polynomial equation $a(\epsilon, s) = b^{(s)}(\epsilon, s)c^{(f)}(\epsilon, s) + b^{(f)}(\epsilon, s)c^{(s)}(\epsilon, s)$. Matching coefficients in these polynomials, this can be expressed as a set of linear equations where a polynomial $x_n s^n + \cdots + x_1 s + x_0$ is represented as a vector $\vec{x} = (x_0, \ldots, x_n)^{\mathrm{T}}$

$$\vec{a}(\epsilon) = B(\epsilon) \begin{bmatrix} \vec{c}^{(f)}(\epsilon) \\ \vec{c}^{(s)}(\epsilon) \end{bmatrix} \tag{3.156}$$

where

$$B(\epsilon) = \begin{bmatrix} b_0^{(s)}(\epsilon) & 0 & \cdots & 0 & b_0^{(f)}(\epsilon) & 0 & \cdots & 0 \\ b_1^{(s)}(\epsilon) & b_0^{(s)}(\epsilon) & \ddots & 0 & b_1^{(f)}(\epsilon) & b_0^{(f)}(\epsilon) & \ddots & 0 \\ \vdots & & & \vdots & \vdots & & & \vdots \\ 0 & 0 & & b_{m-2}^{(s)}(\epsilon) & 0 & 0 & & b_{n-m-1}^{(f)} \\ 0 & 0 & \cdots & b_{m-1}^{(s)}(\epsilon) & 0 & 0 & \cdots & b_{n-m}^{(f)} \end{bmatrix} \tag{3.157}$$

Since the roots of $b^{(f)}(\epsilon, s)$ and $b^{(s)}(\epsilon, s)$ remain separated as $\epsilon \downarrow 0$, this set of equations has a solution (which may not be unique) for all $\epsilon \in [0, \epsilon_0)$. The coefficients of $c^{(f)}(\epsilon, s)$ and $c^{(s)}(\epsilon, s)$ are therefore bounded and real.

Having solved for $c^{(f)}(\epsilon, s)$ and $c^{(s)}(\epsilon, s)$, $p^{(f)}(\epsilon)$ and $p^{(s)}(\epsilon)$ can be determined as

$$p^{(f)}(\epsilon) = \frac{c^{(f)}(\epsilon, 0)}{b^{(f)}(\epsilon, 0)}, \quad p^{(s)}(\epsilon) = \frac{c^{(s)}(\epsilon, 0)}{b^{(s)}(\epsilon, 0)} \tag{3.158}$$

and

$$H^{(f)}(\epsilon,s)=\frac{1}{p^{(f)}(\epsilon)}\frac{c^{(f)}(\epsilon,s)}{b^{(f)}(\epsilon,s)}\ ,\quad H^{(s)}(\epsilon,s)=\frac{1}{p^{(s)}(\epsilon)}\frac{c^{(s)}(\epsilon,s)}{b^{(s)}(\epsilon,s)} \tag{3.159}$$

Several properties of this decomposition follow directly:

- By construction, $p^{(f)}(\epsilon)+p^{(s)}(\epsilon)=1$.

- The positivity of $p^{(s)}(\epsilon)$ can be argued by considering $h(\epsilon,t)$ for $t\gg 1/\epsilon$. For such time, $p^{(f)}(\epsilon)h^{(f)}(\epsilon,t)$ is $\mathrm{O}(\exp(-1/\epsilon))$ and can be ignored. Since $h(\epsilon,t)$ must remain positive, both $p^{(s)}(\epsilon)$ and $h^{(s)}(\epsilon,t)$ must also be positive for $t\gg 1/\epsilon$. Since $h^{(s)}(\epsilon,t)$ is slowly varying (i.e. $\frac{d}{dt}h^{(s)}(\epsilon,t)=\mathrm{O}(\epsilon)$) $h^{(s)}(\epsilon,t)\geq 0$ for all $t\geq 0$.

- Since the derivative of $h^{(s)}(\epsilon,t)$ is $\mathrm{O}(\epsilon)$, this also implies that $\sup_{t\geq 0}h^{(s)}(\epsilon,t)=\mathrm{O}(\epsilon)$.

- In the case that $\sup_{t\geq 0}|h(\epsilon,t)|=\mathrm{O}(1)$, in order to guarantee the positivity of $h(\epsilon,t)$ for $t=\mathrm{O}(1)$, $p^{(f)}(\epsilon)$ must be nonnegative since $h^{(s)}(\epsilon,t)$ is only $\mathrm{O}(\epsilon)$.

3.A.2 The effect of complex "probabilities"

In Section 3.3.3, Algorithm 2.1 was applied to the Markov representation $\eta_{\mathrm{M}}(\epsilon,t)$ of the semi-Markov process $\eta_{\mathrm{E}}(\epsilon,t)$. The derivation in that section assumed that all the transition probabilities and rates in the Markov representation were real and positive. However, as discussed in Section 3.3.2, these quantities can be negative or complex in general. In this section, it will be argued that although many terms in the Markov representation may be complex, application of Algorithm 2.1 is still valid. The essential aspect of that algorithm which must be demonstrated is that there can be no cancellation of the lowest order term in the sum (3.85). This can be shown by carefully considering the construction of the Markov representation.

A useful aspect of the method of stages representation of an arbitrary holding time distribution $h(t)$ is that since $h(0)$ is necessarily nonnegative and real, the term $(1-q_1)\lambda_1$ in the staged form shown in Figure 3.7 (page 94) is also nonnegative and real. This can be shown by considering the Initial Value Theorem applied to the Laplace transform $H(s)$ of $h(t)$. Furthermore, without loss of generality, we can assume that the first exponential rate λ_1 is real. This can always be guaranteed by introducing another stage with rate λ^* as discussed in Section 3.3.2.

There are three cases which must be dealt with separately to show that there is no cancellation in the sum (3.85).

1. The initial class E_I is an aggregation of "fast" states of $\eta_{\mathrm{M}}(\epsilon, t)$,
2. E_I is composed of a single state $m_n^{i,k}$ where $i \in S$, the set of slow states of $\eta_{\mathrm{E}}(\epsilon, t)$, and
3. E_I is composed of a single state $i \in S$.

The first case is straightforward. In Section 3.3.3 the sum (3.85) is expressed in terms of positive, real quantities related to the semi-Markov process $\eta_{\mathrm{E}}(\epsilon, t)$ as (3.91). There cannot be any cancellation due to the positivity of the terms.

The non-cancellation of the sum in the second case can be argued from the fact that there can only be either be one or two terms in the sum. If state $m_{n+1}^{i,k}$ is not a transient state of $\eta_{\mathrm{M}}(0, t)$, then there is only one term in the sum. The only possible classes E_J to which transitions can occur are $\{m_{n+1}^{i,k}\}$ or the class to which state k belongs. If state $m_{n+1}^{i,k}$ is transient, then by the construction $v_{J\, m_{n+1}^{i,k}} = 1$ if $k \in E_J$ and 0 otherwise. The two terms in the sum can be expressed as a single term and therefore there can be no cancellation in this case as well.

The third case relies on the observation outlined previously that the term $(1 - q_1)\lambda_1$ is always nonnegative and real. If the next class is of the form $\{m_1^{i,k}\}$ then there is only one term in the sum as in the previous case. If the next class is not of this form then we know that the transition rates must all be real since all the transition rates $\lambda_{ji}(\epsilon)$ in the sum are of the form $(1 - q_1)\lambda_1$ which are real and nonnegative.

Chapter 4

Decomposition of Discrete Time Markov Chains

4.1 Introduction

In this chapter, the decomposition of discrete time, finite state Markov chains is addressed. Recall that in Chapter 2, the behavior of a continuous time Markov chain was approximated using a fast time scale, ϵ-independent, continuous time process and a reduced order perturbed process. In the discrete time case presented in this chapter, the basic approximation which is derived has a "hybrid" form. In this form, the fast time scale behavior is approximated using an ϵ-independent, discrete time Markov chain, and the slow behavior is captured by a perturbed, continuous time process. This extension to discrete time chains bridges the gap between previous multiple time scale decomposition results, which have dealt exclusively with either continuous time or discrete time processes, and provides a uniform framework for the analysis of both types of systems.

Consider the state probabilities, $x[t]$, of a discrete time Markov chain which satisfy the difference equation

$$x[t+1] = \Phi^{(0)}(\epsilon)x[t] \ , \ \ t \in \mathrm{N}_0 \tag{4.1}$$

where $\phi_{ji}^{(0)}(\epsilon)$ is the one-step transition probability from state i to state j. Note that all the entries of $\Phi^{(0)}(\epsilon)$ are nonnegative and that $\mathbf{1}^{\mathrm{T}}\Phi^{(0)}(\epsilon) = \mathbf{1}^{\mathrm{T}}$. The solution of this difference equation has the form

$$x[t] = \Phi^{(0)}(\epsilon)^t x[0] \tag{4.2}$$

which is analogous to the matrix exponential form which results in the continuous time case in which the system is governed by a differential equation. As in the previous chapters, it is assumed that $\Phi^{(0)}(\epsilon)$ is an analytic function of a small parameter ϵ.

The basic result which will be demonstrated is that the exact system behavior can be approximated using an ϵ-independent discrete time chain, and a smaller, perturbed continuous time chain. Specifically, the form of the approximation which will be derived is

$$\Phi^{(0)}(\epsilon)^t = \Phi^{(0)}(0)^t + U^{(0)} e^{\epsilon A^{(1)}(\epsilon) t} V^{(0)} - U^{(0)} V^{(0)} + O(\epsilon) \tag{4.3}$$

where $A^{(1)}(\epsilon)$ is the generator of a continuous time Markov chain and $O(\epsilon)$ is a function of ϵ and t which converges uniformly to zero over the interval $t \geq 0$ as $\epsilon \downarrow 0$. The matrices $U^{(0)}$ and $V^{(0)}$ are computed from $\Phi^{(0)}(0)$ and have similar interpretations to the corresponding quantities in the continuous time case presented in Chapter 2. The term $\Phi^{(0)}(0)^t$ captures the behavior of the fast time scale and $\exp(\epsilon A^{(1)}(\epsilon) t)$ captures the slow time scale, aggregate behavior.

Note that the form of the approximation (4.3) is identical to the continuous time case (2.103) except that the "fast" term, $\exp(A^{(0)}(0)t)$, is replaced with $\Phi^{(0)}(0)^t$. Although use of this type of hybrid form of approximation has been suggested in the past in the more limited context of a two time scale approximation [5], the uniform validity of (4.3) has not been addressed. The importance of this uniform validity is that it allows us to apply the continuous time algorithm of Chapter 2 to approximate *all* the slow time scale behavior generated by $A^{(1)}(\epsilon)$. Therefore, the key to defining a decomposition algorithm and overall approximation for discrete time Markov chains lies in proving the uniform validity of (4.3). This result extends currently available techniques in two major ways. First, as is true in the continuous time algorithm, there is essentially no restriction on the ergodic structure of the initial Markov process. Therefore, a complete time scale decomposition of any singularly perturbed Markov chain can be obtained. Second, although discrete time chains have been studied by many other authors, (see [12] [42] for example), there has been little connection with other approaches to decomposition of continuous time processes. In particular, the notion of considering a scaled time variable has not been stressed. Furthermore, although the use of a differential equation to describe the slow behavior of a difference equation is not a new idea (see [5] for example), there are few results related to the explicit construction of such a continuous time system in the Markov process context.

One restriction on the structure of $\Phi^{(0)}(\epsilon)$ not present in the continuous time case is that $\Phi^{(0)}(0)$ must be aperiodic and therefore the only eigenvalues of $\Phi^{(0)}(0)$ on the unit circle are at the point $\lambda = 1$. A consequence of periodicity at the fast time scale is that $\lim_{t\to\infty} \Phi^{(0)}(0)^t$ does not exist. This limit must exist for there to exist a multiple time scale decomposition of the process. It will be shown that the slow time scales are aperiodic since they can be approximated by a continuous time Markov process.

The remainder of this chapter is organized as follows. The detailed algorithm is presented in the next section. A proof of the validity of this algorithm is then provided in Section 4.3, followed by a simple example in Section 4.4. A discussion is presented in Section 4.5. Proofs of supporting results are presented in Section 4.A.

4.2 The Algorithm

The decomposition algorithm is based on the approximation (4.3) and Algorithm 2.1.

Algorithm 4.1 *Begin with the generator, $\Phi^{(0)}(\epsilon)$, of a discrete time Markov process with one ergodic class for $\epsilon > 0$ and which is aperiodic at $\epsilon = 0$.*

1. *Define*

$$A^{(0)}(\epsilon) \equiv \Phi^{(0)}(\epsilon) - I \tag{4.4}$$

   $A^{(0)}(\epsilon)$ is the generator of a continuous time Markov process.

2. *Construct the terms $U^{(k)}$, $V^{(k)}$, $k = 0, 1, 2, \ldots, K$, and $A^{(k)}$, $k = 1, 2, \ldots, K$ using steps 1-4 of Algorithm 2.1.*

3. *The overall approximation of the evolution of the transition probabilities can be written as*

$$\begin{aligned}\Phi^{(0)}(\epsilon)^t &= \Phi^{(0)}{}^t + \\ &\Big(U^{(0)}e^{A^{(1)}\epsilon t}V^{(0)} - U^{(0)}V^{(0)}\Big) + \\ &\Big(U^{(0)}U^{(1)}e^{A^{(2)}\epsilon^2 t}V^{(1)}V^{(0)} - U^{(0)}U^{(1)}V^{(1)}V^{(0)}\Big) + \\ &\vdots \\ &\Big(U^{(0)}\ldots U^{(k-1)}e^{A^{(k)}\epsilon^k t}V^{(k-1)}\ldots V^{(0)} - \\ &\qquad U^{(0)}\ldots U^{(k-1)}V^{(k-1)}\ldots V^{(0)}\Big) + \mathrm{O}(\epsilon)\end{aligned} \tag{4.5}$$

where $\mathrm{O}(\epsilon)$ *is a function of* ϵ *and* t *which converges uniformly to zero over* $t \geq 0$.

□

Several features of the algorithm should be noted. First, note that $\Phi^{(0)}(0)$ must be aperiodic. As stated in the previous section, this is necessary for there to exist a multiple time scale decomposition of the process. This fact will be demonstrated in the next section where the derivation of the algorithm relies on this characteristic. The necessity of aperiodicity will also be discussed more fully in Section 4.5.

Note also that in step 1, $A^{(0)}(\epsilon)$ is in fact a generator of a continuous time Markov process. Though the fast behavior of the continuous time process generated by $A^{(0)}(\epsilon)$ is very different from that of the discrete time process generated by $\Phi^{(0)}(\epsilon)$, it will be shown that their *slow time scale* behaviors are approximately equal. This fact forms the basis of the argument for using $A^{(1)}(\epsilon)$ to approximate the slow behavior of the original discrete time process.

4.3 Derivation

The derivation of the above approximation is composed of three distinct parts. First, the "fast" and "slow" components of $\Phi^{(0)}(\epsilon)^t$ are identified. The next two parts consist of approximating these components separately. The superscript "$^{(0)}$" is omitted in the derivation to simplify the notation.

4.3.1 Separation of "fast" and "slow" components

The behavior of $\Phi(\epsilon)^t$ can be separated into "fast" and "slow" components. The "slow" component is associated with eigenvalues which converge to 1 as $\epsilon \downarrow 0$ while the "fast" component is associated with those eigenvalue which converge to points within the unit circle. The approach taken here is based of Kato's perturbation results for linear operators [22] and parallels Coderch's approach to separation of time scales in the continuous time, general linear system case [9].

The generator $\Phi(\epsilon)$ can be expressed as the spectral sum

$$\Phi(\epsilon) = \sum_i \lambda_i(\epsilon) P_i(\epsilon) + D_i(\epsilon) \tag{4.6}$$

where $P_i(\epsilon)$ is the eigenprojection and $D_i(\epsilon)$ is the eigennilpotent associated with the eigenvalue $\lambda_i(\epsilon)$. Note that in general, these projections and nilpotents are not analytic functions of ϵ even if $\Phi(\epsilon)$ is.

Example 4.1 *In order demonstrate that the projections and nilpotents are not necessarily analytic, consider the matrix (which is not a Markov generator)*

$$A(\epsilon) = \begin{bmatrix} 1-\epsilon & 1 \\ 0 & 1-\epsilon^2 \end{bmatrix} \tag{4.7}$$

At $\epsilon = 0$, the decomposition sum has one term:

$$A(\epsilon) = 1 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \tag{4.8}$$

For $\epsilon > 0$, the eigenvalues are distinct and therefore there are two terms in the sum and no eigennilpotents:

$$A(\epsilon) = (1-\epsilon) \begin{bmatrix} 1 & \frac{-1}{\epsilon-\epsilon^2} \\ 0 & 0 \end{bmatrix} + (1-\epsilon^2) \begin{bmatrix} 0 & \frac{1}{\epsilon-\epsilon^2} \\ 0 & 1 \end{bmatrix} \tag{4.9}$$

□

These eigenprojections and nilpotents have the properties that

$$P_i(\epsilon)P_j(\epsilon) = \begin{cases} P_j(\epsilon) & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \tag{4.10}$$

$$P_i(\epsilon)D_j(\epsilon) = \begin{cases} D_j(\epsilon) & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \tag{4.11}$$

The *total projection of the* 1*-group* can be formed as

$$P(\epsilon) \equiv \sum_{i:\,\lambda_i(\epsilon)\to 1} P_i(\epsilon) \tag{4.12}$$

Although as stated above, the individual projections and nilpotents are not necessarily analytic functions of ϵ, Kato [22] shows that any *total projection* of an eigengroup of a perturbed matrix is analytic. An *eigengroup* is a set of eigenvalues which converge to a common point. Therefore since $\Phi(\epsilon)$ is an analytic function of ϵ, $P(\epsilon)$ is analytic at $\epsilon = 0$.

The generator $\Phi(\epsilon)$ can therefore be decomposed into the sum of two parts

$$\Phi(\epsilon) = P(\epsilon)\Phi(\epsilon) + Q(\epsilon)\Phi(\epsilon) \tag{4.13}$$

where

$$Q(\epsilon) \equiv I - P(\epsilon) \tag{4.14}$$

Recall that by assumption, $\Phi(0)$ is aperiodic. Therefore, all the eigenvalues on the unit circle are in fact concentrated at $\lambda = 1$. The eigenvalues of $Q(\epsilon)\Phi(\epsilon)$ therefore converge to points strictly within the unit circle.

Using this decomposition of $\Phi(\epsilon)$ and the properties of the eigenprojections stated above, the following decomposition is possible

$$\begin{aligned} \Phi(\epsilon)^t &= \Big(P(\epsilon)\Phi(\epsilon) + Q(\epsilon)\Phi(\epsilon)\Big)^t && (4.15)\\ &= P(\epsilon)\Big(P(\epsilon)\Phi(\epsilon)\Big)^t + Q(\epsilon)\Big(Q(\epsilon)\Phi(\epsilon)\Big)^t && (4.16) \end{aligned}$$

In order to prove that the approximation (4.3) is valid, the two terms in the sum (4.16) will be treated separately.

4.3.2 Approximation of the fast behavior

Using the decomposition (4.16), the fast behavior, which is determined by $Q(\epsilon)\Phi(\epsilon)$ can be easily approximated since the ϵ-dependence is a regular perturbation of $Q(0)\Phi(0)$. The goal is to show that

$$\Big(Q(\epsilon)\Phi(\epsilon)\Big)^t - \Big(Q(0)\Phi(0)\Big)^t = O(\epsilon) \tag{4.17}$$

from which follows that

$$
\begin{aligned}
Q(\epsilon)\big(Q(\epsilon)\Phi(\epsilon)\big)^t &= Q(0)\big(Q(0)\Phi(0)\big)^t + O(\epsilon) && (4.18)\\
&= Q(0)\Phi(0)^t + O(\epsilon) && (4.19)\\
&= \Phi(0)^t - P(0) + O(\epsilon) && (4.20)\\
&= \Phi(0)^t - UV + O(\epsilon) && (4.21)
\end{aligned}
$$

The validity of (4.17) can be argued from the fact that the eigenvalues of $Q(0)\Phi(0)$ are all strictly inside the unit circle. From Kato [22], the Z-transform, given by

$$T(\epsilon, z) \equiv \big(I - z^{-1}Q(\epsilon)\Phi(\epsilon)\big)^{-1} \qquad (4.22)$$

converges uniformly away from the singularities of $T(0, z)$.[1] The difference in (4.17) can be written as

$$
\begin{aligned}
\Delta(\epsilon, t) &\equiv \big(Q(\epsilon)\Phi(\epsilon)\big)^t - \big(Q(0)\Phi(0)\big)^t && (4.23)\\
&= \frac{1}{2\pi\imath}\oint_\Gamma z^{t-1}\big(T(\epsilon, z) - T(0, z)\big)dz && (4.24)
\end{aligned}
$$

where Γ is a positively oriented contour of length $|\Gamma|$ contained inside the unit circle. Since on the contour $|z^t| \leq 1$ for $t \geq 0$

$$\|\Delta(\epsilon, t)\| \leq \frac{1}{2\pi}|\Gamma| \sup_{z\in\Gamma} \left\| \frac{1}{z}\big(T(\epsilon, z) - T(0, z)\big) \right\| = O(\epsilon) \qquad (4.25)$$

and therefore (4.17)–(4.21) are true.

4.3.3 Approximation of the slow behavior

The approximation of the slow behavior determined by $P(\epsilon)\Phi(\epsilon)$ is based on its further separation into components that evolve at various time scales. Within each time scale, we employ the matrix equivalent of the scalar approximation

$$(1 + \epsilon\lambda)^t = e^{\epsilon\lambda t} + O(\epsilon) \qquad (4.26)$$

[1] Kato states these results in terms of the resolvent $R(\varsigma, A(\epsilon)) \equiv (A(\epsilon) - \varsigma I)^{-1}$. The Z-transform is more commonly used in the context of Markov chains.

whenever $\Re(\lambda) < 0, \epsilon \in [0, \epsilon_0)$, a fact that can easily be verified by series expansion of the terms. Note also that (4.26) is obviously true for $\lambda = 0$. The ultimate goal in this section is to show that

$$P(\epsilon)\big(P(\epsilon)\Phi(\epsilon)\big)^t - P(\epsilon)\mathrm{e}^{P(\epsilon)(\Phi(\epsilon) - I)t} = \mathrm{O}(\epsilon) \tag{4.27}$$

Before continuing with the general development, we should note that the proof of the validity of the approximation (4.27) is particularly simple in the special situation when the eigenvalues of $P(\epsilon)\Phi(\epsilon)$ are semi-simple over an interval $\epsilon \in [0, \epsilon_0)$. Specifically, since the eigenvalues are semi-simple, there are no eigennilpotents and therefore

$$\begin{aligned} P(\epsilon)\Phi(\epsilon) &= \sum_i \lambda_i(\epsilon)P_i(\epsilon) && (4.28)\\ (P(\epsilon)\Phi(\epsilon))^t &= \sum_i \lambda_i(\epsilon)^t P_i(\epsilon) && (4.29)\\ \mathrm{e}^{P(\epsilon)(\Phi(\epsilon) - I)t} &= \sum_i \mathrm{e}^{(\lambda_i(\epsilon) - 1)t} P_i(\epsilon) && (4.30) \end{aligned}$$

By matching the terms in these sums and applying the scalar result (4.26), the approximation (4.27) follows directly.

When the eigenvalues of $P(\epsilon)\Phi(\epsilon)$ are not semi-simple on $\epsilon \in [0, \epsilon_0)$, the approximation (4.27) can still be shown to be valid but the proof is not as straightforward. The basic result which will be employed is the matrix form of the scalar approximation (4.26) above.

Lemma 4.1 *Consider a matrix A with semi-simple null structure and such that all the nonzero eigenvalues have negative real parts. Then*

$$(I + \epsilon A)^t - \mathrm{e}^{\epsilon A t} = \mathrm{O}(\epsilon) \tag{4.31}$$

Proof *The case where the eigenvalues are semi-simple has been discussed in the text above. A general proof is provided in Section 4.A.1.* □

The key to application of this lemma lies in isolating the various time scales and applying the result to each separately. As in Algorithm 4.1, we define the continuous time Markov generator

$$A^{(0)}(\epsilon) \equiv \Phi^{(0)}(\epsilon) - I \tag{4.32}$$

$P(\epsilon)A^{(0)}(\epsilon)$ can be decomposed into terms

$$P(\epsilon)A^{(0)}(\epsilon) = \sum_{i=1}^{K} \epsilon^i B^{(i)}(\epsilon) \tag{4.33}$$

where

$$\epsilon^i B^{(i)}(\epsilon) = R^{(i)}(\epsilon)A^{(0)}(\epsilon) \tag{4.34}$$

and

$$R^{(i)}(\epsilon) \equiv \sum_{i:\, \lambda_i(\epsilon)-1=O\left(\epsilon^i\right)} P_i(\epsilon) \tag{4.35}$$

All the eigenprojections $R^{(i)}(\epsilon)$ exist and are analytic at $\epsilon = 0$. This fact follows since $A^{(0)}(\epsilon)$ is the generator of a continuous time Markov process and therefore satisfies the "Multiple Semi-Simple Null Structure" condition which in turn guarantees that all these eigenprojections exist at $\epsilon = 0$ [9]. A more detailed discussion of this is provided in Section 4.A.2. Essentially, $\epsilon^i B^{(i)}(\epsilon)$ captures all the eigenvalues of $\Phi^{(0)}(\epsilon) - I$ which are strictly $O\left(\epsilon^i\right)$.

Since the eigenvalues of $B^{(i)}(\epsilon)$ are all identically zero or have strictly negative $O(1)$ real parts, the ϵ-dependence is a regular perturbation

$$\begin{aligned} e^{B^{(i)}(\epsilon)t} &= e^{B^{(i)}(0)t} + O(\epsilon) && (4.36) \\ (I + \epsilon^i B^{(i)}(\epsilon))^t &= (I + \epsilon^i B^{(i)}(0))^t + O(\epsilon) && (4.37) \end{aligned}$$

Therefore applying Lemma 4.1 to the right hand sides of the above equations gives

$$(I + \epsilon^i B^{(i)}(\epsilon))^t = e^{\epsilon^i B^{(i)}(\epsilon)t} + O(\epsilon) \tag{4.38}$$

By decomposing the terms in (4.27)

$$\begin{aligned} P(\epsilon)\left(P(\epsilon)\Phi^{(0)}(\epsilon)\right)^t &= \sum_{i=1}^{k} R^{(i)}(\epsilon)\left(I + \epsilon^i B^{(i)}(\epsilon)\right)^t && (4.39) \\ P(\epsilon)e^{P(\epsilon)(\Phi^{(0)}(\epsilon) - I)t} &= \sum_{i=1}^{k} R^{(i)}(\epsilon)e^{\epsilon^i B^{(i)}(\epsilon)t} && (4.40) \end{aligned}$$

and matching the finite number of terms in these sums proves that the approximation (4.27) is indeed valid.

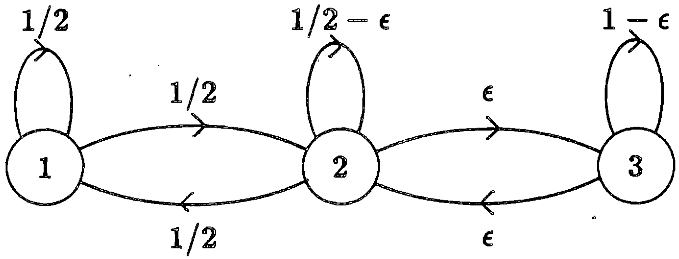


Figure 4.1: Discrete time perturbed Markov process

Finally, the term

$$P(\epsilon)\mathrm{e}^{P(\epsilon)(\Phi^{(0)}(\epsilon)-I)t} = P(\epsilon)\mathrm{e}^{P(\epsilon)A^{(0)}(\epsilon)t} \tag{4.41}$$

is identically the term for the slow behavior of the continuous time process generated by $A^{(0)}(\epsilon) = \Phi^{(0)}(\epsilon) - I$. Using the results of Chapter 2, this can be written as

$$P(\epsilon)\mathrm{e}^{P(\epsilon)A^{(0)}(\epsilon)t} = U^{(0)}\mathrm{e}^{\epsilon A^{(1)}(\epsilon)t}V^{(0)} + \mathrm{O}(\epsilon) \tag{4.42}$$

where $A^{(1)}(\epsilon)$ is a reduced order Markov generator and $U^{(0)}$ and $V^{(0)}$ are the ergodic probability and membership matrices determined from $A^{(0)}(0)$.

The approximation (4.3) needed to prove the validity of Algorithm 4.1 therefore follows from the approximation of the fast component (4.21) and (4.42). The remainder of the approximation (4.5) follows from the approximation of $\exp(A^{(1)}(\epsilon)\tau)$, $\tau = t/\epsilon$, which is available using Algorithm 2.1.

4.4 Example

In this section, a simple two time scale discrete time Markov chain is decomposed. Consider the process with the transition probability graph illustrated in Figure 4.1 and with generator

$$\Phi^{(0)}(\epsilon) = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 1/2-\epsilon & \epsilon \\ 0 & \epsilon & 1-\epsilon \end{bmatrix} \tag{4.43}$$

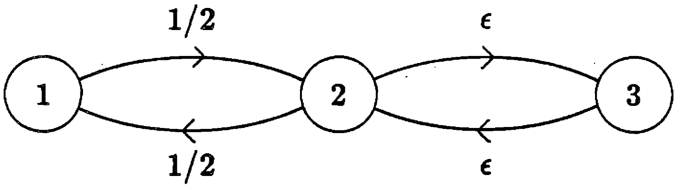


Figure 4.2: Associated continuous time process

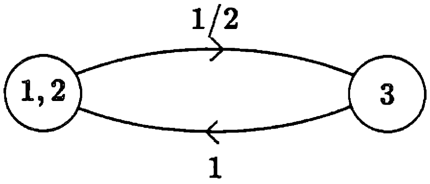


Figure 4.3: O$(1/\epsilon)$ time scale continuous time process

The transition rates of the continuous time process generated by $A^{(0)}(\epsilon) \equiv \Phi^{(0)}(\epsilon) - I$ are shown in Figure 4.2. The slow time scale process obtained using the Markov algorithm is shown in Figure 4.3 and has a generator

$$A^{(1)}(\epsilon) = \begin{bmatrix} -1/2 & 1 \\ 1/2 & -1 \end{bmatrix} + \mathrm{O}(\epsilon) \tag{4.44}$$

The combined approximation is therefore

$$\Phi^{(0)}(\epsilon)^t = \Phi^{(0)}(0)^t + U^{(0)} e^{\epsilon A^{(1)}(0) t} V^{(0)} - U^{(0)} V^{(0)} + \mathrm{O}(\epsilon) \tag{4.45}$$

$$= \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 1/2 & 0 \\ 0 & 0 & 1 \end{bmatrix}^t + \begin{bmatrix} 1/2 & 0 \\ 1/2 & 0 \\ 0 & 1 \end{bmatrix} \exp\left(\begin{bmatrix} -1/2 & 1 \\ 1/2 & -1 \end{bmatrix} \epsilon t \right) \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{4.46}$$

$$- \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 1/2 & 0 \\ 0 & 0 & 1 \end{bmatrix} + \mathrm{O}(\epsilon)$$

4.5 Discussion

The decomposition result presented in this chapter, together with the continuous time algorithm in Chapter 2, provide us with an efficient algorithm for aggregation of discrete time, finite state, Markov processes. One novel feature encountered in the discrete time case is that the fast time scale is required to be aperiodic. In continuous time systems, there is no comparable issue since such processes cannot be periodic. The restriction to systems with aperiodic fast time scales seems fundamental and not simply a consequence of the derivation employed. If the fast time scale is periodic, then the limit

$$\lim_{t\to\infty} \Phi^{(0)}(0)^t \tag{4.47}$$

does not exist. The existence of such a limit is prerequisite for there to exist a complete multiple time scale decomposition. It may, however, be possible to construct a uniform approximation on an interval of the form $[\delta/\epsilon^q, \infty)$, though this extension was not investigated.

Another feature of the discrete time result is related to the results for semi-Markov processes presented in Chapter 3. A discrete time Markov process can be considered as a semi-Markov process where all the holding times are deterministic and equal. Such a semi-Markov process does not satisfy the restriction to rational polynomial holding time transforms. However, a minor modification of the semi-Markov algorithm would give the same result as is derived in this chapter. This suggests that the semi-Markov algorithm presented in Chapter 3 may be in fact more general than can be justified using the derivation presented.

4.A Appendix

4.A.1 Proof of Lemma 4.1

By performing a similarity transformation we can assume, without loss of generality, that the matrix A is in the form

$$A = \begin{bmatrix} \Lambda + D & 0 \\ 0 & 0 \end{bmatrix} \tag{4.48}$$

where Λ is a diagonal matrix with strictly negative entries, and D is a nilpotent matrix ($D^m = 0$) which commutes with Λ. We will consider the case where $A = \Lambda + D$ since the result for the more general form for A follows directly.

Since Λ and D commute,

$$\begin{aligned} e^{\epsilon A t} &= e^{\epsilon \Lambda t} e^{\epsilon D t} && (4.49) \\ &= e^{\epsilon \Lambda t}\left(I + \epsilon D t + \cdots + \frac{1}{(m-1)!}(\epsilon D t)^{m-1}\right) && (4.50) \end{aligned}$$

The term $(I + \epsilon A)^t$ can be similarly expanded

$$\begin{aligned} (I+\epsilon A)^t &= \Big((I+\epsilon\Lambda) + \epsilon D\Big)^t && (4.51) \\ &= (I+\epsilon\Lambda)^t + t(I+\epsilon\Lambda)^{t-1}(\epsilon D) + \cdots + && (4.52) \\ &\qquad \frac{t!}{(t-m+1)!(m-1)!}(I+\epsilon\Lambda)^{t-m+1}(\epsilon D)^{m-1} \end{aligned}$$

The terms in these sums can be matched term by term. The first is essentially the scalar result

$$\left\| e^{\epsilon \Lambda t} - (I+\epsilon\Lambda)^t \right\| = O(\epsilon) \tag{4.53}$$

since the diagonal elements of Λ have negative real parts by assumption.

The difference in the $(j+1)^{\text{th}}$ terms in the sums, $1 < j < m-1$, is

$$\left(\frac{t^j}{j!} e^{\epsilon \Lambda t} (\epsilon D)^j\right) - \left(\frac{t!}{(t-j)!}\frac{1}{j!}(I+\epsilon\Lambda)^{t-j}(\epsilon D)^j\right) \tag{4.54}$$

which can be written as

$$\left(e^{\epsilon \Lambda t}(\epsilon t)^j - \frac{t!}{(t-j)!}\epsilon^j (I+\epsilon\Lambda)^{t-j}\right)\frac{D^j}{j!} \tag{4.55}$$

Note that the first term in the parenthesis in (4.55) is diagonal and that the term $D^j/j!$ is $\mathrm{O}(1)$. Using

$$\frac{t!}{(t-j)!} = t(t-1)\cdots(t-j+1) = t^j + \mathrm{O}\left(t^{j-1}\right) \tag{4.56}$$

a diagonal element of the first term can be written as

$$\mathrm{e}^{\epsilon\lambda t}(\epsilon t)^j - (\epsilon t)^j(1+\epsilon\lambda)^{t-j} + \mathrm{O}\left(\epsilon^j t^{j-1}(1+\epsilon\lambda)^{t-j}\right) \tag{4.57}$$

The maximum of this expression occurs with $t = \mathrm{O}(1/\epsilon)$ at which point the entire expression is $\mathrm{O}(\epsilon)$. Therefore, since each of the finite number of terms in the expansions are $\mathrm{O}(\epsilon)$, the lemma follows directly.

4.A.2 Analytic eigenprojections, $R^{(i)}(\epsilon)$

The analytic nature of the eigenprojections $R^{(i)}(\epsilon)$ follows from the fact that each of these projections is the total projection of the zero-group of eigenvalues of a slow time scale operator. Essentially, this is the implicit fact used by Coderch [9]. In that work, we begin with $A^{(0)}(\epsilon)$, a continuous time Markov generator and construct $P^{(0)}(\epsilon)$, the total projection of the zero-group. Then $A^{(1)}(\epsilon) = P^{(0)}(\epsilon)A^{(0)}(\epsilon)/\epsilon$ which again has a zero-group projection $P^{(1)}(\epsilon)$.

Coderch has shown that all these projections $P^{(i)}(\epsilon)$ are analytic functions of ϵ. The terms $R^{(i)}(\epsilon)$ used in the derivation in Section 4.3 can be defined in terms of the $P^{(i)}(\epsilon)$ in [9] as

$$R^{(i)} \equiv \left(I - P^{(i)}(\epsilon)\right) P^{(i-1)}(\epsilon) \cdots P^{(0)}(\epsilon) \tag{4.58}$$

Chapter 5

Decomposition of Autonomous Positive Systems

5.1 Introduction and Background

5.1.1 Motivation

The results presented in Chapter 2 provide a straightforward algorithm for the decomposition of Markov processes with rare transitions. The development of this algorithm uses the probabilistic nature of the systems both in the proof of its validity and in its interpretation. It is not immediately clear which characteristics of stochastic matrices[1] are exploited in the Markov context which are not available or not used in the more general linear system algorithms presented in [9] and [36].

In this chapter, we will see that a wider class of linear systems than those associated with perturbed stochastic matrices can in fact be analyzed using a decomposition algorithm which is substantially based on the algorithm developed in the Markov context. However, it is also true that direct application of the Markov process algorithm is not in general valid. Certain linear systems, called *positive systems*, in which a nonnegative initial state results in a nonnegative state at any future time share many of the important properties with linear systems describing the evolution of state occupancy probabilities of continuous time Markov

[1] We will use the term "stochastic matrix" to refer to the generator of a continuous time Markov process, i.e. A is stochastic if $a_{ij} \geq 0 \, \forall i \neq j$ and $\mathbf{1}^{\mathrm{T}} A = \mathbf{0}^{\mathrm{T}}$.

chains. Although there are strong ties with the results in the stochastic case, the positive system algorithm is more complex. As will be discussed below, the major additional computation which must be carried out is the identification of the ϵ-dependent dominant eigenvector associated with the generator matrix.

It should be noted that some of the earliest work on decomposition of Markov chains by Simon, Ando, and Fisher [1] [42] was in fact an application of their results on nearly completely decomposable ***positive systems***. This work was motivated by problems in econometrics. The application of their initial results has since then been almost exclusively in a probabilistic context. The work presented in this chapter goes beyond their results in that they only considered a class of systems where the previously discussed Markov algorithm would be directly applicable. The algorithm presented in this chapter addresses a larger class of positive systems.

In this chapter, perturbed, positive, linear systems of the form

$$\dot{x}(t) = A(\epsilon)x(t) \tag{5.1}$$

are considered. It is required that for any $\epsilon \in [0, \epsilon_0)$, if $x(t_0)$ has nonnegative components, $x(t)$ also has nonnegative components for all $t \geq t_0$. Also all the nonzero eigenvalues of $A(\epsilon)$ must have negative real parts and the zero eigenvalue must be semi-simple. The major result which will be demonstrated is that under certain restrictions, the algorithm presented in Chapter 2 can be applied, although it does not share the probabilistic interpretation.

Analysis of perturbed positive systems may be useful in several areas. One application involves ***compartmental models*** [38] in which there may be some very small flow rates. Small flow rates are analogous to small probability transition rates which are associated with very slow time scales of behavior in the probabilistic case. Also, the analysis of slow behavior of positive systems may be useful in the area of ***chemical kinetics*** where the use of various types of aggregate modeling has already proved useful [15].

5.1.2 Positive Systems

The class of positive linear systems is quite well understood. Several properties which will be useful in the development are summarized in this section in order to demonstrate the similarities and to provide some insight into the meaning of these results. The basic material in this section follows closely the presentations in [3] and [37, Chapter 6]. The material relating to the eigenvectors of reducible positive matrices is also treated in [39]. The following notation will be employed with respect to inequalities involving vectors and matrices.

$$x \geq \mathbf{0} \iff x_i \geq 0 \, \forall i \tag{5.2}$$

$$x > \mathbf{0} \iff x \geq \mathbf{0} \text{ and } x_i > 0 \text{ for some } i \tag{5.3}$$

$$x \gg \mathbf{0} \iff x_i > 0 \, \forall i \tag{5.4}$$

These relationships will be termed *nonnegative*, *positive*, and *strictly positive*, respectively.[2] Using this notation, a *positive system* is defined as

$$\dot{x}(t) = Ax(t) \tag{5.5}$$

such that

$$x(t_0) \geq \mathbf{0} \implies x(t) \geq \mathbf{0} \, \forall t \geq t_0 \tag{5.6}$$

In order for this condition to be satisfied, the matrix A must be a *Metzler matrix*, i.e. $a_{ij} \geq 0$ for all $i \neq j$. This guarantees that $\dot{x}_i(t) \geq 0$ when $x_i(t) = 0$ and $x(t) \geq \mathbf{0}$. The state vector cannot therefore leave the positive orthant and the vector $x(t)$ must remain nonnegative [37, Chapter 6].

Irreducibility

In the stochastic context, the ergodicity of an underlying Markov chain is relied upon to guarantee the strict positivity of the limiting probabilities. In the positive system case, we cannot rely on the probabilistic interpretation of the state vector. The analogous concept in the positive system context is irreducibility of the generator matrix.

Definition 5.1 *A square matrix A is* irreducible *if there exists no permutation matrix P such that*

$$PAP^{\mathrm{T}} = \begin{bmatrix} B & C \\ 0 & D \end{bmatrix} \tag{5.7}$$

[2] Note that for a matrix $A > 0$ has the meaning analogous to (5.3) and does *not* mean that A is positive definite.

where B and D are square matrices. □

Note that if A is a stochastic matrix for a Markov chain with one ergodic class but which also has transient states, the matrix B would be a stochastic matrix which generates one ergodic subchain, and the states corresponding to D would be transient. If the chain has two ergodic classes and no transient states, then B and D would each generate a single ergodic class and $C = 0$.

Example 5.1 *Consider the stochastic matrix A*

$$A = \begin{bmatrix} -1 & 0 & 2 & 0 \\ 0 & -1 & 0 & 2 \\ 1 & 0 & -3 & 0 \\ 0 & 1 & 1 & -2 \end{bmatrix} \tag{5.8}$$

which can be permuted as follows

$$PAP^{\mathrm{T}} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} -1 & 0 & 2 & 0 \\ 0 & -1 & 0 & 2 \\ 1 & 0 & -3 & 0 \\ 0 & 1 & 1 & -2 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix} \tag{5.9}$$

$$= \begin{bmatrix} -1 & 2 & 0 & 0 \\ 1 & -2 & 0 & 1 \\ 0 & 0 & -1 & 2 \\ 0 & 0 & 1 & -3 \end{bmatrix} = \begin{bmatrix} B & C \\ 0 & D \end{bmatrix} \tag{5.10}$$

Since A is stochastic, the state transition diagrams for A and B can be drawn as in Figure 5.1. Note that this chain has one ergodic class, described by B. □

From the definition of irreducibility, the following lemma dealing with nonnegative matrices is available.

Lemma 5.1 *[3, Theorem 2.2.1] A nonnegative matrix A is irreducible if and only if for every pair (i, j) there exists a natural number q such that*

$$a_{ij}^{(q)} > 0 \tag{5.11}$$

where $a_{ij}^{(q)}$ denotes the (i, j) element of A^q. □

$$A = \begin{bmatrix} -1 & 0 & 2 & 0 \\ 0 & -1 & 0 & 2 \\ 1 & 0 & -3 & 0 \\ 0 & 1 & 1 & -2 \end{bmatrix}$$

$$B = \begin{bmatrix} -1 & 2 \\ 1 & -2 \end{bmatrix}$$

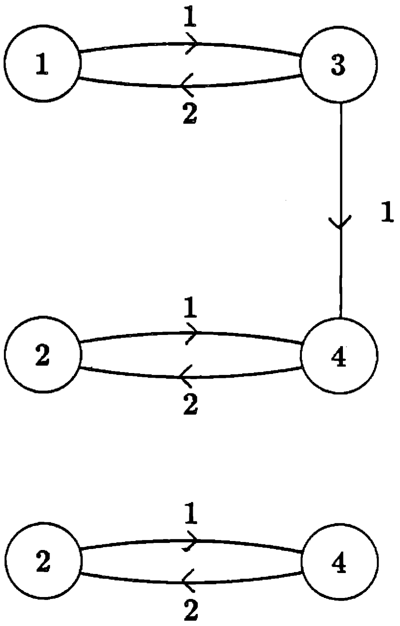
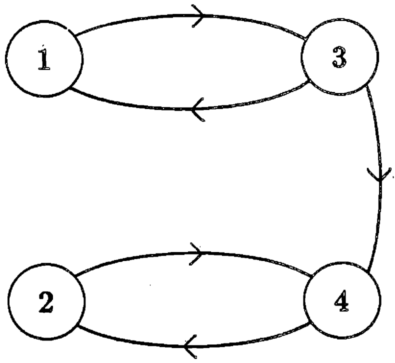


Figure 5.1: Reducible stochastic matrix in Example 5.1



$$A = \begin{bmatrix} -1 & 0 & 2 & 0 \\ 0 & -1 & 0 & 2 \\ 1 & 0 & -3 & 0 \\ 0 & 1 & 1 & -2 \end{bmatrix}$$

Figure 5.2: Graph G(A) where A is stochastic

Graphical analysis

From the above characterization of irreducibility follows a simple graphical interpretation.

Definition 5.2 *The* associated graph G(A) $= (\mathcal{V}, \mathcal{E})$ *of an* $n \times n$ *matrix* A *consists of vertices* $\mathcal{V} = \{1, \ldots, n\}$ *and directed edges* $\mathcal{E} = \{(i, j) \mid a_{ji} \neq 0\}$.[3] *We say that* i *has* access *to* j *if there exists a directed path from* i *to* j. *The graph* G(A) *is* strongly connected *if for any pair* (i, j), i *has access to* j *and* j *has access to* i. □

The graph of the matrix A in Example 5.1 is illustrated in Figure 5.2. Note that this graph is directly related to the state transition diagram in Figure 5.1 with the exception that the arcs are not labeled with the entries in the matrix A. The irreducibility of the matrix A can be determined from the graph G(A) in the same manner as the ergodicity of a Markov chain can be identified from its state transition diagram.

[3] Note that the direction of the edge is reversed from the convention used in [3]. This is done to provide a more intuitive connection with the Markov process case where column vectors of probabilities are used and the (i, j) element of a generator matrix is associated with a $j \rightarrow i$ transition. The theorem statements are therefore slightly different from those cited.

Theorem 5.2 *A Metzler matrix A is irreducible if and only if the graph $G(A)$ is strongly connected.*

Proof *This follows directly from the "i, j, q" characterization in Lemma 5.1 and the fact that if a positive matrix B is irreducible, then so is any $A = B - sI$, $s \geq 0$.*

□

Applying this theorem to the matrix A in Figure 5.2, we see that it is not irreducible since there exists a path from 3 to 4 but not from 4 to 3.

Finally, the graph $G(A)$ can be partitioned into strongly connected sets α_1, $\alpha_2, \ldots$. These sets are called *classes* if they do not belong to any larger strongly connected sets, i.e. they are maximal strongly connected sets. For example $\alpha_1 = \{1,3\}$ and $\alpha_2 = \{2,4\}$ are classes of the graph in Figure 5.2. Note that a class α is an equivalence class in that if $i \in \alpha$ has access to k, then so do all other $j \in \alpha$. Note in the case of stochastic matrices, these classes would correspond to individual ergodic sub-chains and to a disjoint partition of the transient states into subsets so that any transient state in one of these classes can be reached from any other state in that class.

M-matrices

Requiring that A is irreducible and that all the non-zero eigenvalues have negative real parts furthermore implies that $-A$ is an *M-matrix* [3, Theorems 6.4.6 and 6.4.16]. Among the many properties of M-matrices is that the matrix $-A$ can be written as

$$-A = sI - B \tag{5.12}$$

where $B > 0$ is irreducible, $\rho(B)$ is the spectral radius (magnitude of the largest eigenvalue) of B and $s \geq \rho(B)$ [2]. In the special case when $s = \rho(B)$, $-A$ is a *singular M-matrix.* The negatives of the stochastic matrices considered in Chapter 2 belong to this class of singular M-matrices. For example, a typical stochastic matrix A can be written as

$$-A = \begin{bmatrix} 1 & -2 \\ -1 & 2 \end{bmatrix} = 2\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 1 & 2 \\ 1 & 0 \end{bmatrix} = \rho(B)I - B \tag{5.13}$$

The following results are extensions of the Perron-Frobenius theorem to irreducible nonnegative matrices.[4]

Theorem 5.3 *[37, Theorem 6.2] An irreducible matrix $B > 0$ has a simple eigenvalue $\rho(B)$. Associated with this eigenvalue are strictly positive right and left eigenvectors.* □

Corollary 5.4 *An irreducible singular M-matrix A has a simple zero eigenvalue with strictly positive right and left eigenvectors.*

Proof *The matrix A can be written as $A = \rho(B)I - B$, where $B > 0$ and B is irreducible. Therefore by Theorem 5.3, B has strictly positive eigenvectors which it shares with A. Since $\rho(B)$ is a simple eigenvalue of B, 0 must be a simple eigenvalue of A.* □

Reducible positive matrices

Although the case when A is not irreducible is more complicated, the properties of reducible positive matrices are directly related to the applicability of the Markov process algorithm to positive systems. We again consider matrices A of the form

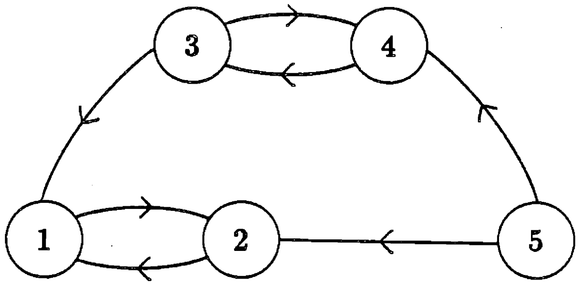
$$A = B - sI\ , \quad B > 0\ , \quad s \geq \rho(B) \tag{5.14}$$

Since A is a Metzler matrix, it can always be written as $B - sI$ for some $B > 0$, $s \geq 0$. The condition that $s \geq \rho(B)$ is required so that there are no eigenvalues of A with positive real parts. The case where A is singular is of most interest in the multiple time scale analysis to be performed. We therefore consider here only the singular case where $s = \rho(B)$.

Since A is reducible, there exists some permutation matrix P such that

$$PAP^{\mathrm{T}} = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1S} \\ 0 & A_{22} & & A_{2S} \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & A_{SS} \end{bmatrix} \tag{5.15}$$

[4] Note that the Perron-Frobenius theorem itself is applicable only to strictly positive matrices.



$$A = \begin{bmatrix} 1/4 & 1/4 & 1/2 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 1 \\ 0 & 0 & 1/2 & 1/2 & 0 \\ 0 & 0 & 1/2 & 1/2 & 1 \\ 0 & 0 & 0 & 0 & 1/2 \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ 0 & A_{22} & A_{23} \\ 0 & 0 & A_{33} \end{bmatrix}$$

Figure 5.3: Graph G(A) in Example 5.2

where each of the A_{jj} are irreducible. By considering these blocks separately, it follows from Theorem 5.2 that the nodes of the graph G(A) can be partitioned into classes $\alpha_1, \alpha_2, \ldots, \alpha_S$ where the class α_i is associated with the block A_{ii}. Based on connectivity of the graph G(A), several properties of the dominant eigenvalue and its generalized eigenvectors can be established. The following definitions are useful in this exposition.

Definition 5.3 *A class α of $A > 0$ is* basic *if $\rho(A[\alpha]) = \rho(A)$ where $A[\alpha]$ is the submatrix of A based on the indices of α. A class α is* singular *if $A[\alpha]$ is singular. A class is* final *if it has access to no other class and it is* initial *if no other class has access to it.*

A class chain *of G(A) is a sequence of classes $\alpha_{k_1}, \alpha_{k_2}, \ldots, \alpha_{k_m}$ such that α_{k_i} has access to $\alpha_{k_{i+1}}$ for $i = 1, 2, \ldots, m-1$. The* length *of such a chain is the number of basic classes in the sequence.* □

Example 5.2 *In order to illustrate the concepts of basic and singular classes consider the matrix A and its associated graph shown in Figure 5.3. The classes of A*

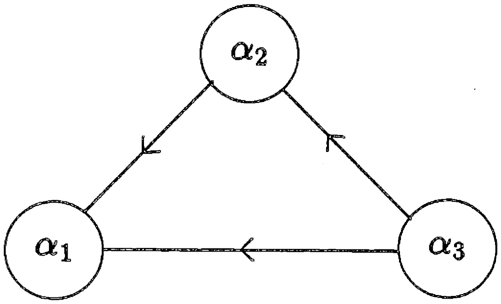


Figure 5.4: Class chains for Example 5.2

are

$$\alpha_1 = \{1,2\}\ ,\ \ \alpha_2 = \{3,4\}\ ,\ \ \alpha_3 = \{5\} \tag{5.16}$$

By considering the submatrices $A[\alpha]$, the basic classes are determined

$$\rho(A[\alpha_1]) \ = \ \rho(A_{11}) = 3/4 \tag{5.17}$$
$$\rho(A[\alpha_2]) \ = \ \rho(A_{22}) = 1 \tag{5.18}$$
$$\rho(A[\alpha_3]) \ = \ \rho(A_{33}) = 1/2 \tag{5.19}$$

Since $\rho(A[\alpha_2]) = \rho(A) = 1$, the class α_2 is basic. The class chains for the graph in Figure 5.3 are shown in Figure 5.4. The chain $(\alpha_1, \alpha_2, \alpha_3)$ is length 1 while the chain (α_1, α_3) is length 0. Class α_1 is final and α_3 is initial. Class α_2 is neither initial nor final. □

Definition 5.3 refers to positive matrices. As described previously, we are interested in singular matrices of the form

$$A = B - \rho(B)I\ ,\ \ B > 0 \tag{5.20}$$

Note that the graphs G(A) and G(B) are identical and therefore the classes of A and B are identical. Note also that $A[\alpha]$ is singular if and only if $\rho(B[\alpha]) = \rho(B)$ therefore the basic classes of B are identically the singular classes of A. To provide some intuition about these concepts, let us examine what they mean for generators of Markov processes. For such a process, the singular classes are exactly the ergodic

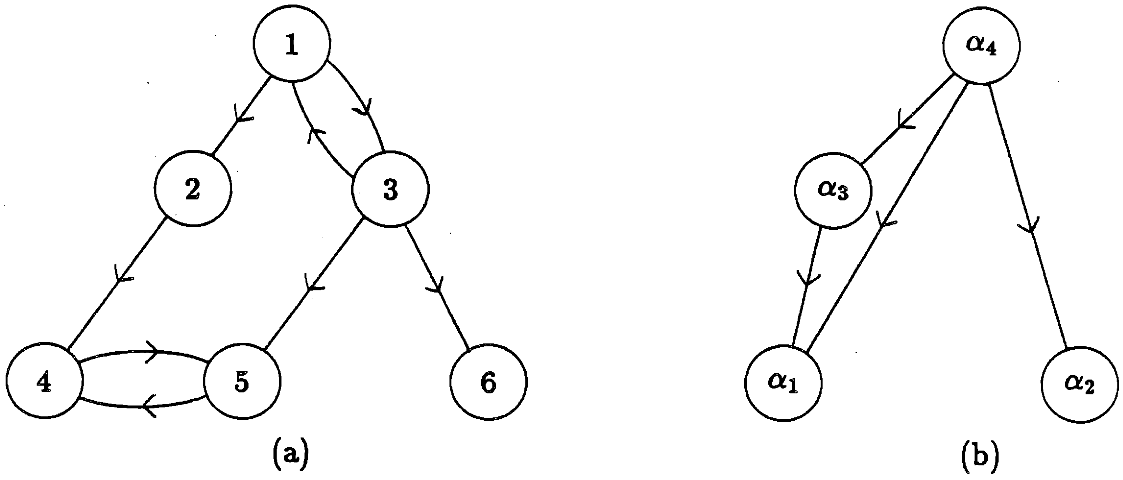


Figure 5.5: G(A) and class chains in Example 5.3

classes of the process and they are all final since by definition there are no transitions leaving an ergodic class [39].

Theorem 5.5 *[3, Theorem 2.3.30] Let $B > 0$ have spectral radius $\rho(B)$ and m basic classes $\alpha_1, \alpha_2, \dots, \alpha_m$. A positive basis $(x^{(1)^{\mathrm{T}}}, x^{(2)^{\mathrm{T}}}, \dots, x^{(m)^{\mathrm{T}}})$ for the left algebraic eigenspace associated with the eigenvalue $\rho(B)$ can be formed such that*

$$x_i^{(j)} > 0 \Longleftrightarrow i \text{ has access to } \alpha_j \tag{5.21}$$

□

This theorem can be applied to $A = B - \rho(B)I$ since the zero eigenvalue of A shares the eigenvectors $x^{(i)}$ above. In the case where $A = B - \rho(B)I$ is a stochastic matrix, this theorem can be interpreted as saying that a basis of the left eigenspace associated with the zero eigenvalue can be constructed such that there is one basis vector for each ergodic class and that the nonzero components of that vector correspond to those states which can reach that class with nonzero probability.

Example 5.3 *Consider the the stochastic matrix A and its associated graph and class chains shown in Figure 5.5. The classes of* G(A) *are $\alpha_1 = \{4,5\}$, $\alpha_2 = \{6\}$,*

$\alpha_3 = \{2\}$ and $\alpha_4 = \{1,3\}$. The classes α_1 and α_2 correspond to ergodic classes of the Markov process and are both singular and final. The non-singular classes α_3 and α_4 correspond to the transient states $1, 2$ and 3.

By applying Theorem 5.5, we can determine the nonzero entries in the right and left eigenvectors of the zero eigenvalue of A. Consider the class $\alpha_1 = \{4,5\}$. The states $1,2,3,4$ and 5 have access to α_1 and therefore one of the left eigenvectors will be of the form $(*,*,*,*,*,0)$ where $*$ represents some strictly positive number. Similarly the left eigenvector associated with α_2 has the form $(*,0,*,0,0,*)$. Also, by applying Theorem 5.5 to A^{T}, the right eigenvectors have the form $(0,0,0,*,*,0)^{\mathrm{T}}$ and $(0,0,0,0,0,*)^{\mathrm{T}}$.

By considering the stochastic nature of the matrix, this structure is also available by considering the transient/recurrent structure of the process. The canonical set of eigenvectors obtained in this way are $(1,1,1,1,1,0)$, $(1,0,1,0,0,1)$, and $(0,0,0,\pi_4,\pi_5,0)^{\mathrm{T}}$, $(0,0,0,0,0,1)^{\mathrm{T}}$, where π_4 and π_5 are the ergodic probabilities of the class of states $\{4,5\}$. □

The zero eigenvalue of a singular matrix $A = B - \rho(B)I$ associated with any time scale must be semi-simple in order that the system has a well-defined multiple time scale decomposition. Note that the zero eigenvalue is semi-simple if and only if $\mathcal{N}(A^2) = \mathcal{N}(A)$ where $\mathcal{N}(C)$ denotes the nullspace of C. This condition can be expressed in terms of the associated graph.

Definition 5.4 *The* index *(or* degree*), $\nu(B)$, of a matrix B is the smallest natural number q such that*

$$\mathcal{N}((B - \rho(B)I)^q) = \mathcal{N}((B - \rho(B)I)^{q+1}) \tag{5.22}$$

□

Note that the semi-simplicity of the zero eigenvalue of $B - \rho(B)I$ corresponds to $\nu(B) \leq 1$.

Example 5.4 *Consider the matrix B*

$$B = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \tag{5.23}$$

where $\rho(B) = 1$. $\mathcal{N}((B-\rho(B)I)^2) \neq \mathcal{N}(B-\rho(B)I)$ *but* $\mathcal{N}((B-\rho(B)I)^{q+1}) = \mathcal{N}((B-\rho(B)I)^q)$ *for* $q \geq 2$. *Therefore* $\nu(B) = 2$. □

Theorem 5.6 *[39, Theorem 3.1]* $\nu(B) \leq 1$ *if and only if the class chains in* G(B) *have length at most one.* □

Note that the matrix B in Example 5.4 has degree two and the class chain ({3}, {2}, {1}) has length two since both classes {1} and {3} are basic.

Example 5.3 (continued) *There are two singular classes in the chain shown in Figure 5.5, namely* α_1 *and* α_2. *The zero eigenvalue therefore has multiplicity two. Since there is no class chain containing both* α_1 *and* α_2, *the zero eigenvalue is semisimple. There are therefore two independent eigenvectors associated with the zero eigenvalue.* □

Singular positive systems

The multiple time scale analysis of continuous time positive systems involves analysis of the unperturbed system at any particular time scale. Since the slow behavior of the system is governed by the small eigenvalues of the generator matrix, the unperturbed matrix must be singular for there to be any nontrivial slow time scale behavior. Therefore, the development of the algorithm considers the properties of the system

$$\dot{x} = Ax\ , \quad A = B - \rho(B)I\ , \quad B > 0 \tag{5.24}$$

In order for the slow time scale behavior to be well-defined, the limit

$$\lim_{t \to \infty} \mathrm{e}^{At} \tag{5.25}$$

must exist. For this to be true, the zero eigenvalue must be semi-simple. This is equivalent to the degree $\nu(B) \leq 1$. This has the graphical interpretation that any class chain of A must contain at most one singular class. Note that the dominant eigenvalue (in this case zero) of a positive system must be real. If zero is not semi-simple in this case, the state $x(t)$ must grow without bound as $t \to \infty$.

Example 5.5 *In order to demonstrate the difficulty encountered when this condition is not satisfied, consider*

$$A(\epsilon) = A + \epsilon\bar{A} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} -\epsilon & 0 \\ 0 & -\epsilon \end{bmatrix} \tag{5.26}$$

The graph G(*A*) *has two singular classes,* $\alpha_1 = \{1\}$ *and* $\alpha_2 = \{2\}$. *The chain* (α_1, α_2) *has two singular classes and therefore the zero eigenvalue of* A *is not semi-simple. The limit*

$$\lim_{t\to\infty} e^{At} = \lim_{t\to\infty} \begin{bmatrix} 0 & t \\ 0 & 0 \end{bmatrix} \tag{5.27}$$

does not exist. □

5.1.3 Chapter Outline

Having established basic properties of positive matrices and systems which will be useful in the development, the remainder of this chapter is organized as follows. In Section 5.2 the decomposition of a class of perturbed positive systems is addressed. Two simple examples are presented in Section 5.3. Then, the issues involved in extending these results to a large number of positive systems and a discussion of the results that are available are presented in Section 5.4.

5.2 Decomposition Algorithm

In this section, the decomposition algorithm for Markov systems is applied to a class of positive systems. The main result is that positive systems which can be analyzed in this way must essentially have a multiple time scale "structure" which has the same characteristics as that associated with Markov systems. The precise characterization of these systems is presented below.

The main step in applying the previously developed algorithm is the identification of a "scaling" of the state variables which results in either a stochastic matrix, or more generally a substochastic matrix where the column sums are negative. In the substochastic case, an additional state variable is added to make the new generator matrix stochastic. The Markov algorithm is then applied and the result is expressed using the original, unscaled state variables.

In order to demonstrate the difficulty in applying essentially the algorithm of Simon and Ando to positive systems, consider the simple generator

$$A(\epsilon) = \begin{bmatrix} -1 & 1-\epsilon \\ 1+\epsilon & 1 \end{bmatrix} \tag{5.28}$$

which has eigenvalues which are approximately -2 and $-\epsilon^2/2$. The matrix $A(0)$ forms a single irreducible block and therefore we can attempt to apply the decomposition algorithm used in Chapter 2 to this system. Computing the slow time scale generator as

$$\tilde{A}(\epsilon) = \frac{1}{\epsilon} V A(\epsilon) U = \frac{1}{\epsilon} \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} -1 & 1-\epsilon \\ 1+\epsilon & 1 \end{bmatrix} \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix} = \begin{bmatrix} 0 \end{bmatrix} \tag{5.29}$$

does not, however, capture the correct slow time scale behavior associated with the eigenvalue $\mathrm{O}(\epsilon^2)$. Even when the matrix $A(0)$ is composed of irreducible blocks, the slow time scale behavior (at time scales $t = \mathrm{O}(1/\epsilon^2)$ and slower) is not approximated correctly using this straightforward approach.

5.2.1 Scaling state variables

Consider the behavior of the system

$$\dot{x}(t) = A(\epsilon)x(t) \tag{5.30}$$

where $A(\epsilon)$ is a Metzler matrix and where there also exists some vector $\alpha^{\mathrm{T}}(\epsilon)$ such that

$$\alpha^{\mathrm{T}}(\epsilon)A(\epsilon) \leq \mathbf{0}^{\mathrm{T}}\ , \quad \epsilon \in [0, \epsilon_0) \tag{5.31}$$

and

$$\alpha(0) \gg \mathbf{0} \tag{5.32}$$

If such a vector exists, then we can perform the similarity transformation

$$B(\epsilon) = T(\epsilon)A(\epsilon)T^{-1}(\epsilon) \tag{5.33}$$

where

$$T(\epsilon) \equiv \operatorname{diag}(\alpha_1(\epsilon), \ldots, \alpha_n(\epsilon)) \tag{5.34}$$

The result of this transformation is that $B(\epsilon)$ is in general a substochastic matrix for $\epsilon \in [0, \epsilon_0)$, since

$$\sum_{j=1}^{n} b_{ij}(\epsilon) = \frac{1}{\alpha_i(\epsilon)} \sum_{j=1}^{n} \alpha_j(\epsilon) a_{ij}(\epsilon) \leq 0 \tag{5.35}$$

and

$$b_{ij}(\epsilon) = \frac{\alpha_i(\epsilon)}{\alpha_j(\epsilon)} a_{ij}(\epsilon) \geq 0\ , \quad \forall i \neq j \tag{5.36}$$

If $B(0)$ is substochastic then there exists only one time scale since all the eigenvalues are O(1). We therefore consider the case in which $B(0)$ is stochastic.

The state variables can be augmented resulting in the purely stochastic matrix

$$C(\epsilon) = \begin{bmatrix} B(\epsilon) & 0 \\ c_{n+1}(\epsilon) & 0 \end{bmatrix} \tag{5.37}$$

where

$$c_{n+1}(\epsilon) \equiv -\alpha^{\mathrm{T}}(\epsilon) A(\epsilon) T^{-1}(\epsilon) \tag{5.38}$$

Since $T^{-1}(0)$ exists, we know that the entries of $C(\epsilon)$ are analytic functions of ϵ. Furthermore, since $C(\epsilon)$ is a stochastic matrix, we know that it generates a system with well defined multiple time scale behavior.

The key element of this procedure is therefore identifying a suitable vector $\alpha^{\mathrm{T}}(\epsilon)$. If $A(\epsilon)$ is already a stochastic matrix, this is particularly simple. Since the sum of the state probabilities is a constant, $\alpha^{\mathrm{T}}(\epsilon) = \mathbf{1}^{\mathrm{T}}$ is suitable. The general problem of finding $\alpha^{\mathrm{T}}(\epsilon)$ and identifying systems for which such a vector does not exist is discussed in Section 5.2.2.

Having constructed $C(\epsilon)$, Algorithm 2.1 can be applied. Specifically,

$$\mathrm{e}^{C(\epsilon)t} = \mathrm{e}^{C(0)t} + U_C(0)\mathrm{e}^{\epsilon\tilde{C}(\epsilon)t} V_C(0) - U_C(0)V_C(0) + \mathrm{O}(\epsilon) \tag{5.39}$$

where

$$\tilde{C}(\epsilon) \equiv \frac{1}{\epsilon} V_C(\epsilon) C(\epsilon) U_C \tag{5.40}$$

as in Algorithm 2.1.

This approximation can be expressed in terms of the matrix $B(\epsilon) = T(\epsilon)A(\epsilon)T^{-1}(\epsilon)$. Specifically, since $B(0)$ is stochastic, we can form matrices U_B and $V_B(\epsilon)$ such that

$$U_C = \left[\begin{array}{cc} U_B & 0 \\ 0^{\mathrm{T}} & 1 \end{array} \right], \quad V_C(\epsilon) = \left[\begin{array}{cc} V_B(\epsilon) & 0 \\ 0^{\mathrm{T}} & 1 \end{array} \right] \tag{5.41}$$

Therefore, the matrix $\tilde{C}(\epsilon)$ can be written as

$$\tilde{C}(\epsilon) = \frac{1}{\epsilon} \left[\begin{array}{cc} V_B(\epsilon)T(\epsilon)A(\epsilon)T^{-1}(\epsilon)U_B & 0 \\ -\alpha^{\mathrm{T}}(\epsilon)A(\epsilon)T^{-1}(\epsilon)U_B & 0 \end{array} \right] \tag{5.42}$$

If the terms $U(\epsilon)$ and $V(\epsilon)$ are defined as

$$U(\epsilon) \equiv T^{-1}(\epsilon)U_B \,, \quad V(\epsilon) \equiv V_B(\epsilon)T(\epsilon) \tag{5.43}$$

then the approximation of the original system can be written as

$$e^{A(\epsilon)t} = e^{A(0)t} + U(0)e^{\epsilon\tilde{A}(\epsilon)t}V(0) - U(0)V(0) + O(\epsilon) \tag{5.44}$$

where

$$\tilde{A}(\epsilon) = \frac{1}{\epsilon}V(\epsilon)A(\epsilon)U(\epsilon) \tag{5.45}$$

The computation of the slow time scale generator $\tilde{A}(\epsilon)$ can be expressed in the original, unaggregated, basis as

$$A^{(1)}(\epsilon) = \frac{1}{\epsilon}\hat{P}(\epsilon)A(\epsilon)\hat{P}(\epsilon) \tag{5.46}$$

where

$$\hat{P}(\epsilon) \equiv U(\epsilon)V(\epsilon) = T^{-1}(\epsilon)U_B V_B(\epsilon)T(\epsilon) \tag{5.47}$$

is an "approximation" of the true eigenprojection $P(\epsilon)$ of the zero group of eigenvalues. In some sense, $\hat{P}(\epsilon)$ retains only the "important" terms of $P(\epsilon)$. This is similar to the interpretation of Algorithm 2.1 where $\tilde{P}(\epsilon) = U\tilde{V}(\epsilon)$ captures the important terms of $P(\epsilon)$.

5.2.2 Existence of $\alpha^{\mathrm{T}}(\epsilon)$

The key to the construction presented in the previous section is the identification of the vector $\alpha^{\mathrm{T}}(\epsilon)$. Several aspects of the structure of the matrix $A(\epsilon)$ are related to the existence of a suitable vector.

First, using Theorem 5.5, if $A(\epsilon)$ is irreducible for $\epsilon \in (0, \epsilon_0)$, then there is a strictly positive eigenvector $\alpha_0^{\mathrm{T}}(\epsilon)$ associated with the dominant eigenvector $\lambda_0(\epsilon)$. Since by assumption the dominant eigenvalue is nonpositive,

$$\alpha_0^{\mathrm{T}}(\epsilon)A(\epsilon) = \lambda_0(\epsilon)\alpha_0^{\mathrm{T}}(\epsilon) \leq \mathbf{0}^{\mathrm{T}} \tag{5.48}$$

However, strict positivity for $\epsilon > 0$ is not sufficient since it is still possible for $\alpha(0) \not\gg \mathbf{0}$. If any of the components of $\alpha^{\mathrm{T}}(\epsilon)$ are not $\mathrm{O}(1)$ then $T^{-1}(0)$ does not exist and $T^{-1}(\epsilon)$ is not an analytic function of ϵ.

Consider for example

$$A(\epsilon) = \begin{bmatrix} -1 & 1-\epsilon & \epsilon^2 \\ 1+\epsilon & -1 & 0 \\ 0 & \epsilon & -\epsilon-\epsilon^2 \end{bmatrix} \tag{5.49}$$

The dominant eigenvalue is identically zero and the associated eigenvector is

$$\alpha^{\mathrm{T}}(\epsilon) = (1+\epsilon, 1, \epsilon) \tag{5.50}$$

which is strictly positive for $\epsilon > 0$ but not for $\epsilon = 0$.

The condition that $\lim_{\epsilon \downarrow 0} \alpha(\epsilon) \gg \mathbf{0}$ is related to the graph structure of the sequence of aggregated generator matrices. Specifically, if a multiple time scale decomposition of a generator $A^{(0)}(\epsilon)$ is performed to obtain $A^{(0)}(0), \ldots A^{(k)}(0)$, then each graph $\mathrm{G}(A^{(i)}(0))$, $i = 0, \ldots, k-1$ must have "singular/final" structure. That is, the singular classes of $\mathrm{G}(A^{(i)}(0))$ must all be final as well. Note that by Theorem 5.6, this implies that there can not be a class chain with two singular classes and therefore the zero eigenvalue is semi-simple. Since each of the $A^{(i)}$ satisfy this condition, the original system must have well defined multiple time scale behavior (see Definition 2.3, page 42). If the condition is not satisfied for some $A^{(i)}(0)$, then the dominant left eigenvector of $A^{(i)}(0)$ has components which are identically zero. Although a complete proof is not provided here, it can be shown that such zero components are associated with components of the dominant eigenvector of $A^{(0)}(\epsilon)$, $\alpha^{\mathrm{T}}(\epsilon)$ which converge to zero as $\epsilon \downarrow 0$. To illustrate this, consider $A(\epsilon)$ in (5.49). At the second time scale, states 1 and 2 are aggregated and state 3 forms another aggregate. At that time scale, the aggregate made up of the original states 1 and 2 forms a singular class that is not final. The net result is that the dominant eigenvector has a component which converges to zero.

Computation of the dominant eigenvector, $\alpha^{\mathrm{T}}(\epsilon)$ is not in general trivial. Consider the singular case in which there is only a single zero eigenvalue of $A(\epsilon)$ for $\epsilon > 0$. The vector $\alpha^{\mathrm{T}}(\epsilon)$ must be found such that

$$\alpha^{\mathrm{T}}(\epsilon)A(\epsilon) = 0 \tag{5.51}$$

This can be expanded as a Taylor series as

$$(\alpha_0 + \epsilon\alpha_1 + \cdots)(A_0 + \epsilon A_1 + \cdots) = 0 \tag{5.52}$$

By matching powers of ϵ, a sequence of linear equations can be formed.

$$\begin{aligned} \alpha_0 A_0 &= 0 \\ \epsilon(\alpha_0 A_1 + \alpha_1 A_0) &= 0 \\ \epsilon^2(\alpha_2 A_0 + \alpha_1 A_1 + \alpha_0 A_2) &= 0 \\ &\vdots \end{aligned} \tag{5.53}$$

The matrix A_0 is not invertible and therefore these equations do not have unique solutions. However, if A_0 has the zero eigenvalue with multiplicity m, then there are m degrees of freedom in each of the equations. This could be exploited to compute the overall $\alpha^{\mathrm{T}}(\epsilon)$.

An alternative to the direct solution of these equations is to use an iterative approach which is an extension of that proposed by Vantilborgh [45] for the computation of the steady state eigenvector of a nearly completely decomposable Markov chain. In that approach, the eigenvectors of each block are first computed in isolation. Then, the aggregated model is constructed and its eigenvector is computed. The eigenvector associated with each block is then updated to take into account the weak interactions with the other aggregates resulting in an $\mathrm{O}(\epsilon^2)$ approximation.

In the nearly completely decomposable, two time scale case, this procedure can be iterated to provide an $O\left(\epsilon^k\right)$ approximation. Extension of this type of iterative approach may be feasible for the computation of $\alpha^{\mathrm{T}}(\epsilon)$.

5.3 Example

In order to demonstrate the decomposition algorithm for perturbed positive systems, two examples are considered. The first has a singular generator matrix which results in the transformed generator being stochastic. The second example has a nonsingular generator.

Consider the system generated by

$$A(\epsilon) = \begin{bmatrix} -1 & 1-\epsilon & 2\epsilon \\ 1+\epsilon & -1 & 0 \\ 0 & \epsilon^2/2 & -\epsilon-\epsilon^2 \end{bmatrix} \tag{5.54}$$

The dominant left eigenvector associated with the zero eigenvalue is

$$\alpha^{\mathrm{T}}(\epsilon) = (1+\epsilon, 1, 2) \tag{5.55}$$

The similarity transformation is therefore

$$T(\epsilon) = \begin{bmatrix} 1+\epsilon & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix} \tag{5.56}$$

resulting in the stochastic matrix

$$B(\epsilon) = T(\epsilon)A(\epsilon)T^{-1}(\epsilon) = \begin{bmatrix} -1 & 1-\epsilon^2 & \epsilon+\epsilon^2 \\ 1+\epsilon & -1 & 0 \\ 0 & \epsilon^2 & -\epsilon-\epsilon^2 \end{bmatrix} \tag{5.57}$$

The behavior of the system generated by $B(\epsilon)$ can be approximated using Algorithm 2.1.

$$e^{B(\epsilon)t} = e^{B(0)t} + U_B e^{\epsilon\tilde{B}(\epsilon)t} V_B(0) - U_B V_B(0) + O(\epsilon) \tag{5.58}$$

where

$$\tilde{B}(\epsilon) = \frac{1}{\epsilon} V_B(\epsilon) B(\epsilon) U_B = \begin{bmatrix} -\epsilon & 1+\epsilon \\ \epsilon & -1-\epsilon \end{bmatrix} \tag{5.59}$$

and

$$V_B(\epsilon) = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad U_B = \begin{bmatrix} 1/2 & 0 \\ 1/2 & 0 \\ 0 & 1 \end{bmatrix} \tag{5.60}$$

Expressed using the original variables, this approximation is

$$\mathrm{e}^{A(\epsilon)t} = \mathrm{e}^{A(0)t} + U(0)\mathrm{e}^{\epsilon \tilde{A}(\epsilon)t} V(0) - U(0)V(0) + \mathrm{O}(\epsilon) \tag{5.61}$$

where in this case $\tilde{A}(\epsilon) = \tilde{B}(\epsilon)$ and

$$U(\epsilon) = T^{-1}(\epsilon) U_B = \begin{bmatrix} 1/(2+2\epsilon) & 0 \\ 1/2 & 0 \\ 0 & 1/2 \end{bmatrix} \tag{5.62}$$

and

$$V(\epsilon) = V_B(\epsilon) T(\epsilon) = \begin{bmatrix} 1+\epsilon & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix} \tag{5.63}$$

Next, consider the matrix $A(\epsilon)$ which is nonsingular for $\epsilon > 0$.

$$A(\epsilon) = \begin{bmatrix} -1 & 1-2\epsilon & \epsilon^2 \\ 1+2\epsilon & -1 & 0 \\ 0 & \epsilon^2 & -\epsilon^2 \end{bmatrix} \tag{5.64}$$

A suitable vector $\alpha^{\mathrm{T}}(\epsilon)$ is

$$\alpha^{\mathrm{T}}(\epsilon) = (1+2\epsilon, 1, 2) \tag{5.65}$$

Note that when $A(\epsilon)$ is singular, there is a unique vector $\alpha^{\mathrm{T}}(\epsilon)$ (within a scale factor). When $A(\epsilon)$ is non-singular, the suitable vector $\alpha^{\mathrm{T}}(\epsilon)$ is not unique and therefore is somewhat easier to obtain since small multiples of the other eigenvectors can be added. After similarity transformation, we obtain the substochastic matrix

$$B(\epsilon) = \begin{bmatrix} -1 & 1-4\epsilon^2 & \epsilon^2/2+\epsilon^3 \\ 1 & -1 & 0 \\ 0 & 2\epsilon^2 & -\epsilon^2 \end{bmatrix} \tag{5.66}$$

which can be augmented to form the stochastic matrix $C(\epsilon)$

$$C(\epsilon) = \begin{bmatrix} -1 & 1-4\epsilon^2 & \epsilon^2/2+\epsilon^3 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 2\epsilon^2 & -\epsilon^2 & 0 \\ 0 & 2\epsilon^2 & \epsilon^2/2-\epsilon^3 & 0 \end{bmatrix} \tag{5.67}$$

Proceeding directly from $B(\epsilon)$ we can compute

$$U_B = \begin{bmatrix} 1/2 & 0 \\ 1/2 & 0 \\ 0 & 1 \end{bmatrix}, \quad V_B(\epsilon) = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{5.68}$$

from which are obtained

$$U(\epsilon) = T^{-1}(\epsilon)U_B = \begin{bmatrix} 1/(2+4\epsilon) & 0 \\ 1/2 & 0 \\ 0 & 1/2 \end{bmatrix}, \quad V(\epsilon) = V_B(\epsilon)T(\epsilon) = \begin{bmatrix} 1+2\epsilon & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix} \tag{5.69}$$

Finally, the slow time scale generator $\tilde{A}(\epsilon)$ can be calculated directly using $U(\epsilon)$, $V(\epsilon)$, and the original generator $A(\epsilon)$

$$\tilde{A}(\epsilon) = \frac{1}{\epsilon}V(\epsilon)A(\epsilon)U(\epsilon) = \begin{bmatrix} -2\epsilon & \epsilon+2\epsilon^2 \\ \epsilon & -2\epsilon \end{bmatrix} \tag{5.70}$$

5.4 Discussion and Conclusion

The decomposition approach presented in the chapter extends previous multiple time scale results by providing a direct algorithm for constructing a multiple time scale approximation for a class of positive systems. This class is characterized by having a structure that is very similar to that of systems which describe the evolution of state probabilities of a Markov process. The only additional necessary computation beyond the Markov decomposition algorithm is the computation of the dominant eigenvector of the generator of the system.

Consideration of positive systems points out several aspects of the Markov decomposition algorithm which are significant. The first is that Markov processes and the positive system which can be analyzed have a very specific structure. At each time scale the singular classes are final. It is clear that any more general algorithm dealing with positive systems will have to deal with the more general structure which can occur in arbitrary positive systems. The second aspect is that in a conservative system, such as in the Markov case, the left eigenvector is trivial to obtain. It is now clear that part of the ease with which Markov systems can be analyzed is due to knowing the form of the dominant left eigenvector.

It would be desirable to compute the slow time scale generator matrix $\tilde{A}(\epsilon)$ directly from $A(\epsilon)$ without explicitly computing the stochastic form of the generator. This type of result might be available through extension of the concept of "weak" terms in a generator. In the Markov context, it was required that the off-diagonal weak terms be of higher order than the product of several remaining terms. Also, a weak matrix $W(\epsilon)$ had to satisfy $\mathbf{1}^{\mathrm{T}}W(\epsilon) = \mathbf{0}^{\mathrm{T}}$. If this definition is extended to the class of systems which can be decomposed through the stochastic representation presented in this chapter, a weak matrix must satisfy $\alpha^{\mathrm{T}}(\epsilon)W(\epsilon) = \mathbf{0}^{\mathrm{T}}$.

It is now clear that direct application of the Markov algorithm to positive systems is not in general valid. It should be pointed out that this is true even in the nearly completely decomposable case originally considered by Simon and Ando. Their algorithm only gives valid approximations for systems with two time scales of behavior. It is therefore evident that more ϵ-dependent computation is in general necessary. This can take the form of computation of the dominant left eigenvector. Another feasible approach may be to identify which ϵ-dependent terms in the eigenprojection of the zero-group, $P(\epsilon)$, are "needed" in order to obtain a valid decomposition. Once the dominant eigenvector is found, for example, we have already seen that $\hat{P}(\epsilon) = U(\epsilon)V(\epsilon)$ in (5.47) is a sufficient approximation.

Finally, in terms of applications, it may seem that finding $\alpha^{\mathrm{T}}(\epsilon)$ is too difficult. However, there are many systems for which this is straightforward. One major class includes models of compartmental systems [38]. For these systems, $\alpha^{\mathrm{T}}(\epsilon) = \mathbf{1}^{\mathrm{T}}$ is again sufficient and the Markov algorithm can be applied almost directly.

Chapter 6

Structural Decomposition and Fault-Tolerant Systems

6.1 Introduction and Motivation

In this chapter, analysis of the multiple time scale *structure* of a perturbed Markov process is addressed. Structure of a perturbed Markov process refers to the complete multiple time scale decomposition of the type performed in Chapter 2. In this case, only the position of the nonzero transition rates of each of the unperturbed, aggregated time scale models, and the sets of states which constitute the aggregate classes, are determined. Although the detailed behavior of the system cannot be recovered from these descriptions, much useful information is retained. It will be shown that this subset of information about the decomposition can be obtained using very simple graph-theoretic algorithms which are implicit in Algorithm 2.1.

Use of these structural aspects of a Markov chain can have many applications. For instance, an algorithm similar to that presented in this chapter has been applied to the analysis of the behavior of "simulated annealing" optimization methods [44]. Other applications include computing order of magnitudes for the time of events in systems with rare transitions. Specific applications include analysis of the failure time of a fault-tolerant system or the error rate of a communication protocol.

The nature of the decomposition algorithms presented in this chapter are very different from those presented in Chapter 2. In the latter algorithms, numerical calculations of dominant eigenvectors of unperturbed systems, and of ϵ-dependent "trapping" probabilities were required. The algorithms in this chapter basically consist of computing various types of connectivity in labeled graphs. Since the computations involve only integer quantities, issues of numerical stability are not relevant. Also, by introducing the graph-theoretic formalism, it is possible to employ standard graphical algorithms for computing quantities such as shortest paths between vertices of a graph.

This chapter is organized as follows. In the next section, the basic structural decomposition algorithm is presented along with the necessary graph theoretic formalism. An example is analyzed in detail to demonstrate various aspects of the algorithm. Some possible extensions of the graphical algorithm are also discussed. In Section 6.3, an algorithm which can be used to analyze a system with several unknown orders of magnitude of rare transitions is presented. This algorithm is applied to the analysis of a simple fault-tolerant system. The final section contains a discussion of the structural decomposition results.

6.2 Structural Decomposition Algorithm

In Chapter 2, an algorithm for approximating the transition probability function $\Phi^{(0)}(\epsilon,t)$ of a perturbed Markov process $\eta^{(0)}(\epsilon,t)$ was presented. The resulting uniform approximation (2.19) is expressed as a combination of the behavior of a set of ϵ-independent, aggregated Markov processes with generators $A^{(0)}(0),\ldots,A^{(k)}(0)$ and some associated matrices (the $U^{(i)}$ and $V^{(i)}$) which characterize the recurrent and transient states of these processes. The algorithm can be greatly simplified if only the constituent states of the aggregate classes and the allowable (i.e. nonzero rate) state transitions are desired. In this section, the structure of the graphical representation is presented. Then, the graphical algorithm for the derivation of this entire structural decomposition is described.

6.2.1 Graphical Structure

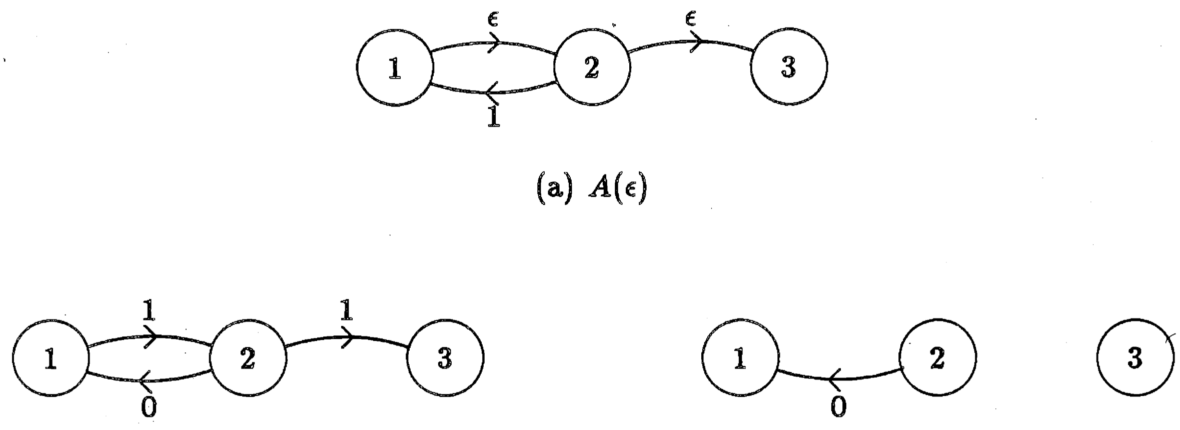
Given a perturbed Markov generator $A^{(k)}(\epsilon)$ of the type considered in Chapter 2, an associated graph $G^{(k)} = (\mathcal{N}^{(k)}, \mathcal{E}^{(k)})$ can be constructed as follows. If $A^{(k)}(\epsilon)$ generates an n state process, then the set of vertices is $\mathcal{N}^{(k)} = \{1, 2, \ldots, n\}$. Each edge is a triple, $e = (i, j, w) \in \mathcal{E}^{(k)}$ of the initial vertex, i, the final vertex, j, and a nonnegative integer weight, w. For each nonzero entry $a_{ji}^{(k)}(\epsilon), j \neq i$, a directed, weighted link from vertex i to vertex j is introduced. If $a_{ji}^{(k)}(\epsilon)$ is strictly $O(\epsilon^w)$, the weight of the edge is w. It is also useful to construct the "zero weight" graph $G_0^{(k)} = (\mathcal{N}^{(k)}, \mathcal{E}_0^{(k)})$ from the generator $A^{(k)}(0)$. The set of vertices is unchanged, and the new set of edges, $\mathcal{E}_0^{(k)} \subseteq \mathcal{E}^{(k)}$, corresponds to the subset of zero-weight edges in $\mathcal{E}^{(k)}$.

Example 6.1 *Consider the generator*

$$A(\epsilon) = \begin{bmatrix} -\epsilon & 1 & 0 \\ \epsilon & -1-\epsilon & 0 \\ 0 & \epsilon & 0 \end{bmatrix} \tag{6.1}$$

of the process illustrated in Figure 6.1(a). The associated graph G is shown in Figure 6.1(b) and has edges $(1,2,1)$, $(2,1,0)$, *and* $(2,3,1)$. *The graph G_0 is shown in Figure 6.1(c) and has only the single edge* $(2,1,0)$. □

Several properties of the graph $G = (\mathcal{N}, \mathcal{E})$ will be used. First, define the distance between two vertices $d(i,j)$ as the minimum sum of weights along directed paths from i to j. If there is no such path, then $d(i,j) = \infty$. Using terminology that is consistent with that used in Chapter 5, define a *communicating class* C as a maximal set of states such that $i, j \in C$ if and only if $d(i,j) < \infty$. A class C is *final* if there is no vertex $j \notin C$ such that $d(i,j) < \infty$ for any $i \in C$. Also define the distance $\tilde{d}_A(i, C)$ as the minimum weight path from i to the class C with *no* intermediate vertices in the set A.



(a) $A(\epsilon)$

(b) G

(b) G_0

Figure 6.1: Markov generator $A(\epsilon)$ and graphs G and G_0 for Example 6.1

As discussed in Chapter 5, there is a direct association between the Markov process generated by A and the graph G. Most important, an ergodic class of A exactly corresponds to a final class of G. Also, a transient state of A corresponds to a vertex of G in some non-final class.

6.2.2 The Algorithm

The basic structural decomposition algorithm is specified in this section. We begin with a graph $G^{(0)}$ constructed using the Markov generator $A^{(0)}(\epsilon)$. Then a sequence of graphs associated with successive time scales, $G_0^{(0)}, G_0^{(1)}, \ldots, G_0^{(K)}$ is constructed such that $G_0^{(i)}$ is associated with the generator $A^{(i)}(0)$ constructed using Algorithm 2.1. Also, the constituent states in an aggregate class associated with each of the vertices in the graph are computed. The set $\mathcal{U}_i^{(k)}$ is the set of states of $\eta^{(0)}(\epsilon, t)$ associated with the aggregate class represented by vertex i in the graph $G_0^{(k)}$. The set $\mathcal{V}_i^{(k)}$ is the set of states which have an O(1) probability of entering the aggregate class $\mathcal{U}_i^{(k)}$ by the k^{th} time scale. The roles of the sets $\mathcal{U}$ and $\mathcal{V}$ are discussed in Section 6.2.4.

Algorithm 6.1

1. *Begin with the graph $G^{(0)} = (\mathcal{N}^{(0)}, \mathcal{E}^{(0)})$ constructed from the Markov generator $A^{(0)}(\epsilon)$ of the n state process $\eta^{(0)}(\epsilon, t)$.*
   Define $\mathcal{U}_i^{(0)} = \{i\}, \mathcal{V}_i^{(0)} = \{i\}$, $i = 1, 2, \ldots, n$.
   Set $k \leftarrow 0$.

2. *Construct the graph $G_0^{(k)} = (\mathcal{N}^{(k)}, \mathcal{E}_0^{(k)})$ using the zero-weight subset of edges $\mathcal{E}_0^{(k)} \subseteq \mathcal{E}^{(k)}$.*

3. *Identify the final communicating classes of $G_0^{(k)}$, $E_I^{(k)}$, $I = 1, 2, \ldots, N$. The complementary set*

   $$T^{(k)} \equiv \mathcal{N}^{(k)} - E^{(k)} \ , \quad E^{(k)} \equiv \bigcup_I E_I^{(k)} \tag{6.2}$$

   is the set of vertices in non-final classes. Also define

   $$\tilde{E}_I^{(k)} = \{i \mid \tilde{d}_{E^{(k)}}(i, E_I^{(k)}) = 0\} \tag{6.3}$$

   If there is only one final class $(N = 1)$, then the decomposition is completed.

4. *Form the new aggregate sets*

   $$\mathcal{U}_I^{(k+1)} = \bigcup_{i \in E_I^{(k)}} \mathcal{U}_i^{(k)} \tag{6.4}$$

   and

   $$\mathcal{V}_I^{(k+1)} = \bigcup_{i \in \tilde{E}_I^{(k)}} \mathcal{V}_i^{(k)} \tag{6.5}$$

5. *Construct a new graph $\bar{G}^{(k+1)} = (\mathcal{N}^{(k+1)}, \bar{\mathcal{E}}^{(k+1)})$, $\mathcal{N}^{(k+1)} = \{1, \ldots, N\}$. For each pair of vertices (I, J) of $\bar{G}^{(k+1)}$, $I \neq J$, the following set of edges comprise $\bar{\mathcal{E}}^{(k+1)}$.*

   (a) *For each edge, $e = (v_1, v_2, w) \in \mathcal{E}^{(k)}$, where $v_1 \in E_I^{(k)}$, $v_2 \in E_J^{(k)}$, an edge $(I, J, w - 1)$ is added to the set $\bar{\mathcal{E}}^{(k+1)}$.*

   (b) *For each edge, $e = (v_1, t, w) \in \mathcal{E}^{(k)}$, where $v_1 \in E_I^{(k)}$, $t \in T^{(k)}$, such that $\tilde{d}_{E^{(k)}}(t, E_J^{(k)}) < \infty$, the edge $(I, J, w - 1 + \tilde{d}_{E^{(k)}}(t, E_J^{(k)}))$ is added to the set $\bar{\mathcal{E}}^{(k+1)}$.*

6. *The graph $\bar{G}^{(k+1)}$ is "reduced" by retaining only the lowest weight link between any pair of vertices producing the graph $G^{(k+1)} = (\mathcal{N}^{(k+1)}, \mathcal{E}^{(k+1)})$. Specifically,*
    (a) *Set $\mathcal{E}^{(k+1)} \leftarrow \bar{\mathcal{E}}^{(k+1)}$.*
    (b) *For each edge $e = (v_1, v_2, w) \in \mathcal{E}^{(k+1)}$, remove edge e if there is some other edge $e' = (v_1, v_2, w'), w' \leq w$ remaining in $\mathcal{E}^{(k+1)}$*

    Set $k \leftarrow k+1$. Go to step 2

□

The major computational requirement of this algorithm involves determining the shortest paths, the $\tilde{d}$ terms. This can be accomplished using a variety of standard graphical algorithms. Since all shortest paths from non-final vertices to all final classes are required, an algorithm such as the Floyd algorithm can be employed. This algorithm computes all shortest path pairs for a graph as follows.

Algorithm 6.2 *(Floyd [18])*
Begin with a graph $G = (\mathcal{N}; \mathcal{E})$, $\mathcal{N} = \{1, \ldots, N\}$, $\mathcal{E} = \{(i, j, d_{ij}) \mid i, j \in \mathcal{N}\}$.

1. *Set $D_{ij}^{(0)} = \min(d_{ij} \mid (i, j, d_{ij}) \in \mathcal{E})$. $D_{ij}^{(0)} = \infty$ if there is no edge from i to j.*
2. *For $n = 0, \ldots, N-1$*

$$D_{ij}^{(n+1)} = \min\left(D_{ij}^{(n)}, D_{in}^{(n)} + D_{nj}^{(n)}\right) \quad \forall i \neq j$$

3. *$D_{ij}^{(N)}$ is then the shortest path from i to j.*

The interpretation of $D_{ij}^{(k)}$, $k < N$, is the shortest part from i to j with all intermediate vertices in the set $\{1, \ldots, k\}$. Note that the complexity of this algorithm is $O(N^3)$. □

This algorithm is easily extended to compute the distances $\tilde{d}_E$ required in Algorithm 6.1. Since $\tilde{d}_E(i,j)$ is defined to be the shortest path from i to j with no intermediate vertices in the set E, step 2 can be modified to introduce only those n such that $n \notin E$.

Algorithm 6.3

Begin with a graph $G = (\mathcal{N}, \mathcal{E})$, $\mathcal{N} = \{1, \ldots, N\}$, $\mathcal{E} = \{(i, j, d_{ij}) \mid i, j \in \mathcal{N}\}$, and a set of vertices E which may not be used as intermediated vertices.

1. *Set $D_{ij}^{(0)} = \min(d_{ij} \mid (i, j, d_{ij}) \in \mathcal{E})$. $D_{ij}^{(0)} = \infty$ if there is no edge from i to j.*
2. *For $n = 0, \ldots, N-1$*
   if $n \in E$ then $D^{(n+1)} = D^{(n)}$ otherwise
   $$D_{ij}^{(n+1)} = \min\left(D_{ij}^{(n)}, D_{in}^{(n)} + D_{nj}^{(n)}\right) \quad \forall i \neq j$$
3. *$D_{ij}^{(N)}$ is then $\tilde{d}_E(i, j)$*

□

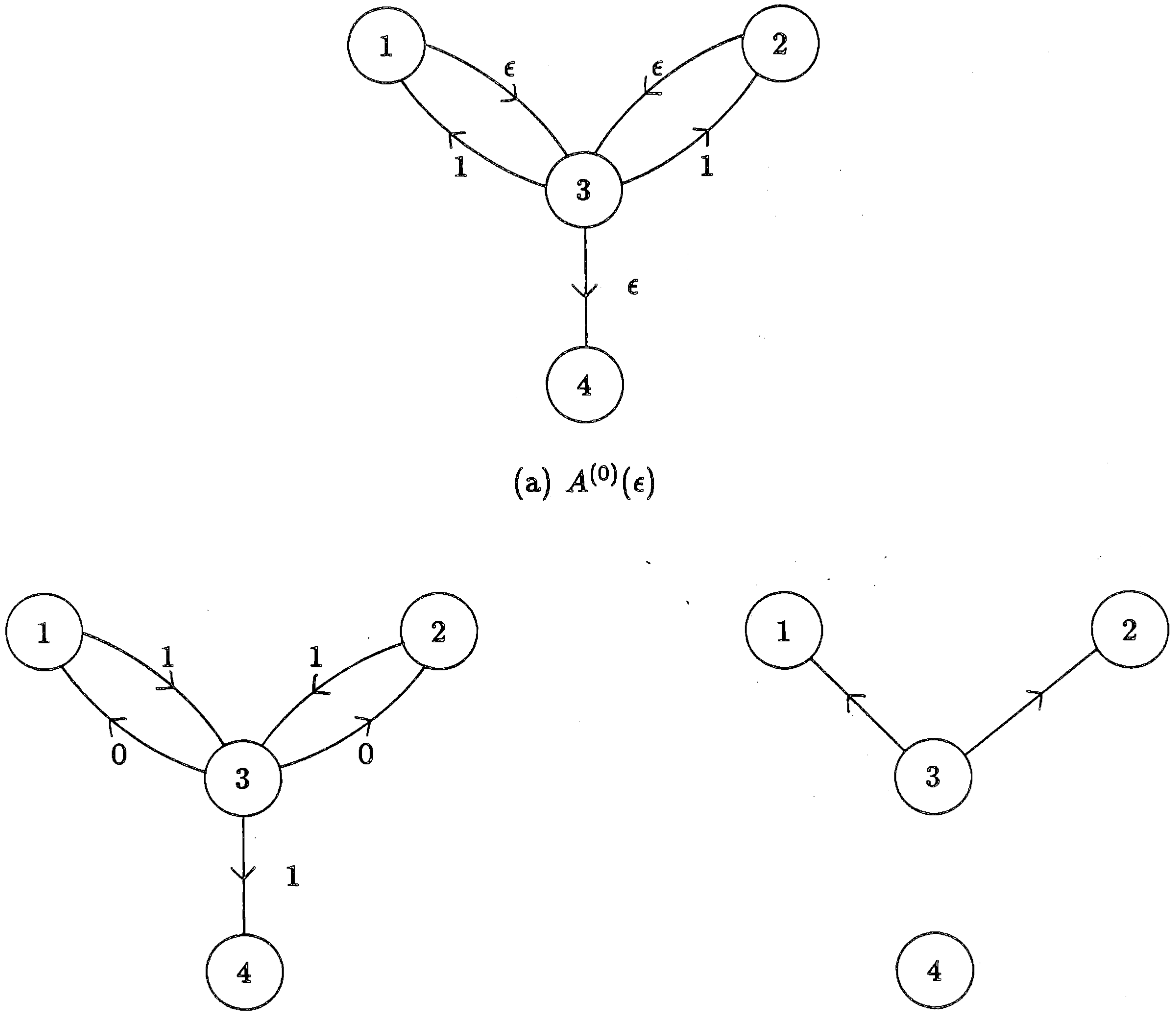
6.2.3 Examples

A simple example is considered using Algorithm 6.1 in order to clarify the approach. The Markov process considered appears in Example 2.1 (page 36) where Algorithm 2.1 is demonstrated.

The generator $A^{(0)}(\epsilon)$ and the associated graph $G^{(0)}$ and $G_0^{(0)}$ are shown in Figure 6.2. The graph $G_0^{(0)}$ has only the edges $(3, 1, 0)$ and $(3, 2, 0)$. The final classes of the graph $G_0^{(0)}$ are $E_1^{(0)} = \{1\}$, $E_2^{(0)} = \{2\}$, and $E_3^{(0)} = \{4\}$ and the non-final set is $T^{(0)} = \{3\}$. Using (6.3) $\tilde{E}_1^{(0)} = \{1, 3\}$, $\tilde{E}_2^{(0)} = \{2, 3\}$, and $\tilde{E}_3^{(0)} = \{4\}$. The distances in the graph $G^{(0)}$ to the final classes of $G_0^{(0)}$ are $\tilde{d}_{E^{(0)}}(3, 1) = \tilde{d}_{E^{(0)}}(3, 2) = 0$ and $\tilde{d}_{E^{(0)}}(3, 4) = 1$. The sets $\mathcal{U}^{(1)}$ and $\mathcal{V}^{(1)}$ are $\mathcal{U}_1^{(1)} = \{1\}$, $\mathcal{U}_2^{(1)} = \{2\}$, and $\mathcal{U}_3^{(1)} = \{4\}$, and $\mathcal{V}_1^{(1)} = \{1, 3\}$, $\mathcal{V}_2^{(1)} = \{2, 3\}$, and $\mathcal{V}_3^{(1)} = \{4\}$.

The graph $\tilde{G}^{(1)}$, which in this simple example is already reduced and therefore also the graph $G^{(1)}$, is shown in Figure 6.3. The sets constructed in this iteration on the algorithm ($k = 1$) are $E_1^{(1)} = \tilde{E}_1^{(1)} = \{1, 2\}$, $E_2^{(1)} = \tilde{E}_2^{(1)} = \{3\}$, $T^{(1)} = \{\}$, $\mathcal{U}_1^{(2)} = \{1, 2\}$, $\mathcal{V}_1^{(2)} = \{1, 2, 3\}$, and $\mathcal{U}_2^{(2)} = \mathcal{V}_2^{(2)} = \{4\}$.

Finally, the graph $\tilde{G}^{(2)}$ is constructed as shown in Figure 6.4(a). The reduction in this case corresponds to removing one of the duplicated links from vertex 1 to 2. The sets constructed in this iteration are $E_1^{(2)} = \{2\}$, $\tilde{E}_1^{(2)} = \{1, 2\}$, and $T^{(2)} = \{1\}$. The set of graphs $G_0^{(i)}$ and the sets $\mathcal{U}^{(i)}$ and $\mathcal{V}^{(i)}$ are illustrated in Figure 6.5. This

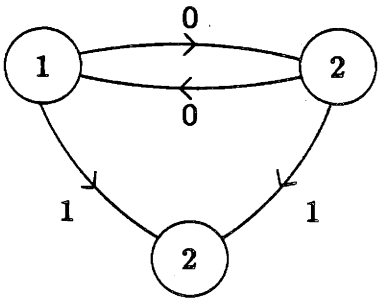


(a) $A^{(0)}(\epsilon)$

(b) $G^{(0)}$

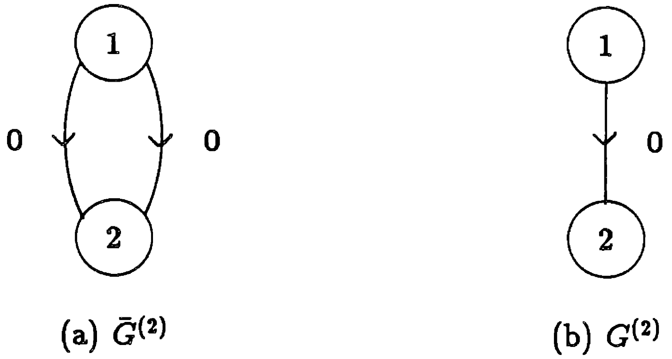
(c) $G_0^{(0)}$

Figure 6.2: Perturbed Markov process and associated graphs



$\bar{G}^{(1)} = G^{(1)}$

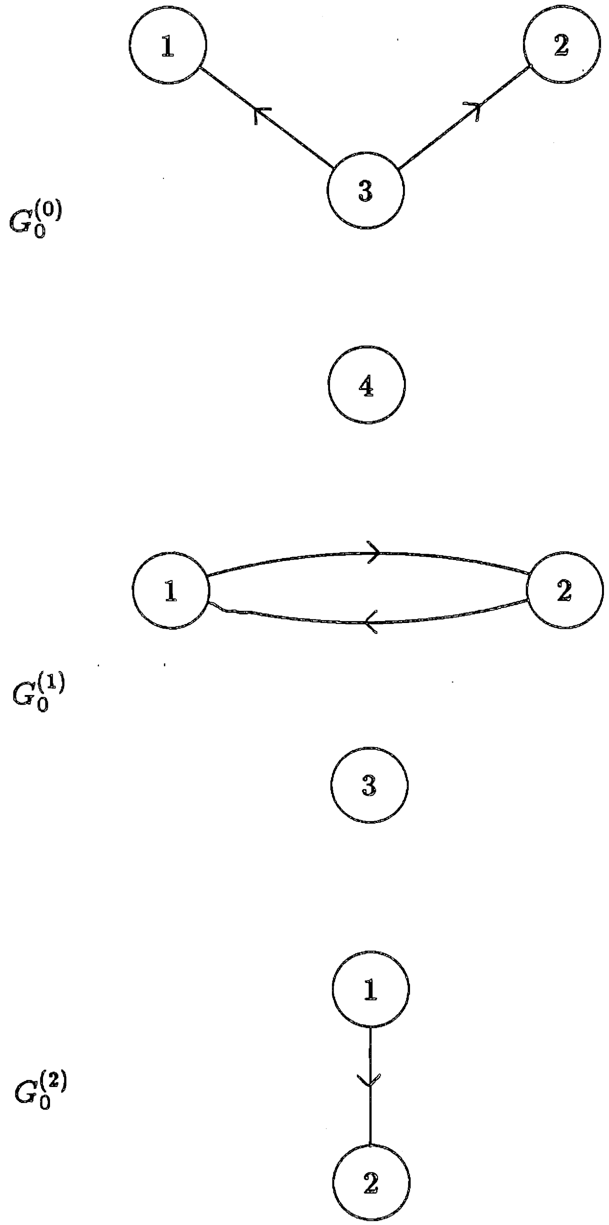
Figure 6.3: O($1/\epsilon$) time scale graph



(a) $\bar{G}^{(2)}$

(b) $G^{(2)}$

Figure 6.4: O($1/\epsilon^2$) time scale graphs



$$\begin{array}{ll} \mathcal{U}_1^{(1)} = \{1\} & \mathcal{V}_1^{(1)} = \{1,3\} \\ \mathcal{U}_2^{(1)} = \{2\} & \mathcal{V}_2^{(1)} = \{2,3\} \\ \mathcal{U}_3^{(1)} = \{4\} & \mathcal{V}_3^{(1)} = \{4\} \end{array}$$

$$\begin{array}{ll} \mathcal{U}_1^{(2)} = \{1,2\} & \mathcal{V}_1^{(2)} = \{1,2,3\} \\ \mathcal{U}_2^{(2)} = \{4\} & \mathcal{V}_2^{(2)} = \{4\} \end{array}$$

Figure 6.5: Multiple time scale graphs

should be compared to the set of Markov processes illustrated in Figure 2.3 (page 39) which is obtained using Algorithm 2.1.

6.2.4 Relationship to Algorithm 2.1

In Algorithm 2.1 (page 33), a sequence of Markov generators, $A^{(k)}(0), k = 0, 1, \dots, K$ describing slower and slower time scales was constructed. Associated with each of these generators are matrices $U^{(k)}$ and $V^{(k)}$ which can be obtained from $A^{(k)}(0)$. The overall approximation (2.19) constructed consists of a sum of terms of the general form

$$U^{(0)}U^{(1)}\dots U^{(k-1)}e^{A^{(k)}(0)\epsilon^k t}V^{(k-1)}\dots V^{(1)}V^{(0)} \tag{6.6}$$

and

$$U^{(0)}U^{(1)}\dots U^{(k-1)}V^{(k-1)}\dots V^{(1)}V^{(0)} \tag{6.7}$$

The relationship which is shown below is that the graph $G_0^{(k)}$ constructed using Algorithm 6.1 is associated with $A^{(k)}(0)$ in the manner described above. Furthermore, the nonzero entries in the products $U^{(0)}\dots U^{(k-1)}$ and $V^{(k-1)}\dots V^{(0)}$ are determined by the sets $\mathcal{U}^{(k)}$ and $\mathcal{V}^{(k)}$ respectively.

Specific relationships between the terms computed in the two algorithms can be summarized as follows. These properties will be demonstrated on an example following the theorem statement.

Theorem 6.1 *For $k = 0, 1, \dots, K$*

1. *$a_{ji}^{(k)}(0)$ is nonzero if and only if there is a link $(i, j, 0)$ in the graph $G_0^{(k)}$ (or equivalently in the graph $G^{(k)}$).*
2. *$u_{iJ}^{(k)}$ is nonzero if and only if $i \in E_J^{(k)}$.*
3. *$v_{Ij}^{(k)}(0)$ is nonzero if and only if $j \in \tilde{E}_I^{(k)}$.*
4. *The (i, J) element of the product $U^{(0)}\dots U^{(k-1)}$ is nonzero if and only if $i \in \mathcal{U}_J^{(k)}$.*
5. *The (I, j) element of the product $V^{(k-1)}\dots V^{(0)}$ is nonzero if and only if $j \in \mathcal{V}_I^{(k)}$.*

Proof *(outline)*

In order to demonstrate that these properties are true, consider the following inductive argument. Assume that these properties are satisfied for k. We will see that they must also be satisfied for $k+1$ as well.

- *Given $A^{(k)}(0)$, the ergodic classes and transient states must be identified in Algorithm 2.1. An ergodic class can be determined by considering the nonzero entries in the matrix $A^{(k)}(0)$. Specifically, if states i and j are recurrent and belong to the same ergodic class, then there is some sequence of state transitions from i to j and back, and there is no state k for which there is a sequence of transitions from i to k but none from k to i. This identification of recurrent classes is precisely what is performed in determining the final classes of the graph $G_0^{(k)}$. Therefore, the set of vertices, $E_I^{(k)}$, exactly corresponds to the I^{th} recurrent class generated by $A^{(k)}(0)$.*
- *Since the set of vertices $E_I^{(k)}$ exactly corresponds to the ergodic class generated by $A^{(k)}(0)$, the (i,J) entry of $U^{(k)}$ is nonzero if and only if $i \in E_J^{(k)}$.*
- *From the definition of $v_{Ij}^{(k)}(0)$ as the probability that the process $\eta^{(k)}(0,t)$ enters class I from state j before any other class, it is evident that this term is nonzero if and only if $j \in \tilde{E}_I^{(k)}$. This follows since there must be some sequence of nonzero probability transition from state j to the I^{th} ergodic class for the probability to be nonzero, and if there is no such sequence, then the probability must be identically zero.*
- *Consider the term $v_{Ij}^{(k)}(\epsilon)$. The $\mathrm{O}(1)$ terms are identified by the sets $\tilde{E}^{(k)}$. From the definition of $v_{Ij}^{(k)}(\epsilon)$ as the probability that the process $\eta^{(k)}(\epsilon,t)$ enters class I from state j before any other class, this term can be interpreted as the sum of the probabilities of all paths through transient states from j to class I. The order of the sum for all paths must be the same as the order of the sum for all* direct *paths. The order of this latter sum is simply the distance $\tilde{d}_{E^{(k)}}(j, E_I^{(k)})$ in the graph $G^{(k)}$ since the weight w of each link (i,j,w), $i \in T^{(k)}$, is simply the order of the transition probability from state i to state j in the process $\eta^{(k)}(\epsilon,t)$. Therefore $v_{Ij}^{(k)}(\epsilon) = \mathrm{O}\left(\epsilon^{\tilde{d}_{E^{(k)}}(j,E_I^{(k)})}\right)$.*

- *The order of the term $a_{JI}^{(k+1)}(\epsilon)$ can be determined by considering the sum*

  $$a_{JI}^{(k+1)}(\epsilon) = \frac{1}{\epsilon} \sum_{i \in E_I^{(k)}} \sum_{j \in E_J^{(k)}} u_{iI}^{(k)} a_{ji}^{(k)}(\epsilon) + \frac{1}{\epsilon} \sum_{i \in E_I^{(k)}} \sum_{t \in T^{(k)}} u_{iI}^{(k)} a_{ti}^{(k)}(\epsilon) v_{Jt}^{(k)}(\epsilon) \tag{6.8}$$

  Each term in the sums is positive, therefore the order of $a_{JI}^{(k)}(\epsilon)$ is the minimum order of the individual terms summed. Each term exactly corresponds to an edge of the graph $\bar{G}^{(k+1)}$. The graph $G^{(k+1)}$ preserves the lowest weight edge between all pairs of vertices and therefore, $G^{(k+1)}$ is associated with the orders of the entries in $A^{(k+1)}(\epsilon)$.

- *From the assumption that the (i, j) entry of $U^{(0)} \cdots U^{(k-1)}$ is nonzero if and only if $i \in \mathcal{U}_j^{(k)}$ and from the association of the sets $E_J^{(k)}$ and $U_J^{(k)}$ shown above, the (i, J) entry of $U^{(0)} \cdots U^{(k)}$ is nonzero if and only if $i \in \mathcal{U}_J^{(k+1)}$.*

- *Similarly, from the association of $\tilde{E}_I^{(k)}$ and $V_I^{(k)}(0)$, the (I, j) entry in $V^{(k)} \cdots V^{(0)}$ is nonzero if and only if $j \in \mathcal{V}_I^{(k+1)}$.*

- *Finally, these properties follow simply for $k = 0, 1$. By induction, they follow for all $1 < k \leq K$ as well.*

□

The above properties can be demonstrated on the example presented in Section 6.2.3. Consider the graph $G^{(1)}$ and the sets $U^{(1)}$ and $V^{(1)}$. Using the above properties, the structures of $A^{(1)}(0)$, $U^{(1)}$ and $V^{(1)}(0)$ are

$$A^{(1)}(0) = \begin{bmatrix} * & * & 0 \\ * & * & 0 \\ 0 & 0 & * \end{bmatrix} \tag{6.9}$$

and

$$U^{(1)} = \begin{bmatrix} * & 0 \\ * & 0 \\ 0 & * \end{bmatrix}, \quad V^{(1)}(0) = \begin{bmatrix} * & * & 0 \\ 0 & 0 & * \end{bmatrix} \tag{6.10}$$

where $*$ is some non-zero entry. Similarly

$$U^{(2)} = \begin{bmatrix} 0 \\ * \end{bmatrix}, \quad V^{(2)}(0) = \begin{bmatrix} * & * \end{bmatrix} \tag{6.11}$$

Consider also the product

$$U^{(0)}U^{(1)} = \begin{bmatrix} * & 0 & 0 \\ 0 & * & 0 \\ 0 & 0 & 0 \\ 0 & 0 & * \end{bmatrix} \begin{bmatrix} * & 0 \\ * & 0 \\ 0 & * \end{bmatrix} \tag{6.12}$$

$$= \begin{bmatrix} * & 0 \\ * & 0 \\ 0 & 0 \\ 0 & * \end{bmatrix} \tag{6.13}$$

It is straightforward to verify that this structure is consistent with the sets $\mathcal{U}_1^{(2)}$ and $\mathcal{U}_2^{(2)}$.

Another relationship to the development in Chapter 2 is that the reduction of a graph $\bar{G}^{(k)}$ to produce $G^{(k)}$ can be extended to yield an even simpler graph which still produces that same sequence of graphs of zero-weight links. By applying Corollary 2.8, links in $G^{(k)}$ can be removed if they are associated with "weak" terms in $A^{(k)}(\epsilon)$. This type of "pruning" of the graph $G^{(k)}$, as well as some more extensive methods are discussed in the next subsection.

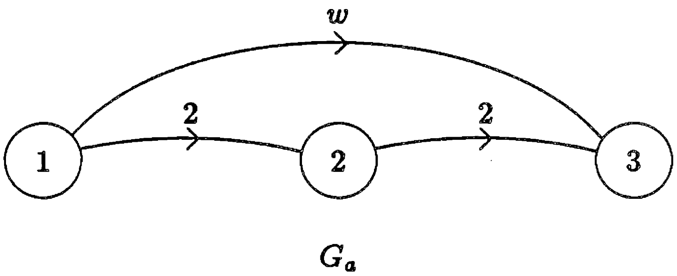
6.2.5 Extensions of the Graphical Algorithm

The reduction step in Algorithm 6.1 which removes unnecessary edges in a graph $\bar{G}^{(k)}$ to produce $G^{(k)}$ can be improved to produce an even simpler graph with fewer edges. The first improvement involves applying Corollary 2.8 which states that any "weak" transition rate can be removed. A further improvement will be considered so that more edges can be removed from the graph $G^{(k)}$ without affecting the sequence of zero-weight graphs $G_0^{(k)}$ produced. The nature of the graph-theoretic formalism provides a more straightforward method of showing the validity of this improved pruning method than is possible by considering the details of Algorithm 2.1.

Corollary 2.8 states that if a generator $\bar{A}(\epsilon)$ can be written as $\bar{A}(\epsilon) = A(\epsilon) + W(\epsilon)$ where $W(\epsilon)$ is weak with respect to $A(\epsilon)$, then $\bar{A}(\epsilon)$ and $A(\epsilon)$ are asymptotically equivalent. That is, the multiple time scale decomposition of $\bar{A}(\epsilon)$ and $A(\epsilon)$ are identical. This corollary can be stated in terms of the associated graph. Let $\bar{A}(\epsilon)$ be associated with the graph $\bar{G} = (\mathcal{N}, \bar{\mathcal{E}})$ and $A(\epsilon)$ with the graph $G = (\mathcal{N}, \mathcal{E})$. Since the $\bar{A}(\epsilon)$ and $A(\epsilon)$, are asymptotically equivalent, the sequence of zero-weight graphs produced using Algorithm 6.1, beginning with either $\bar{G}$ or G, are identical.

By the definition of a weak term, for any nonzero entry $w_{ji}(\epsilon), j \neq i$ there is a sequence of state transitions in $A(\epsilon)$ with a product of transition rates of lower order than $w_{ji}(\epsilon)$. By the construction of the associated graph, the order of a product of transition rates along a path is simply the sum of the weights along that path in the associated graph. Therefore, for any weak term $w_{ji}(\epsilon) = O(\epsilon^w)$, there is a link (i, j, w) such that $w > d(i, j)$ where $d(i, j)$ is the shortest path from i to j in the graph G. The further reduction of the graph G therefore justified by Corollary 2.8 corresponds simply to the removal of any edge $e = (i, j, w)$ which has a weight greater than the shortest path from i to j. The computational requirement of this type of "pruning" consists of computing the shortest path between each pair of vertices. This can be accomplished using an algorithm such as the Floyd algorithm discussed previously.

An issue which remains is whether a more "aggressive" pruning of the edges of a graph G can be performed to simplify the computations required without changing the sequence of zero weight graphs produced. Consider the graphs G_a and G_b shown in Figure 6.6. Assume that the shortest path from 1 to 3 is indeed 4. Using the arguments above, the edge of weight w from 1 to 3 can be removed in both graphs if w is greater than the shortest path, i.e. $w \geq 5$. In the graph G_a, we can see that since there is no edge from 2 back to 1, the direct edge from 1 to 3 can be removed if $w \geq 3$. Therefore, we see that by considering the detailed structure of the graph rather than simply the shortest paths, a stronger pruning condition could be specified. Formalizing this type of intuition may reduce the amount of computation involved. Furthermore, this type of observation could be used to reduce the computation in the full algorithm presented in Chapter 2.



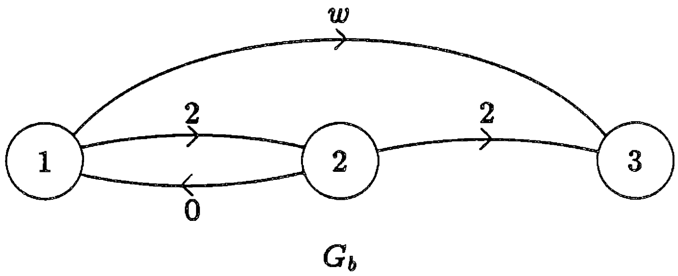


Figure 6.6: Pruning unnecessary links in G

6.3 Application to fault-tolerant systems

6.3.1 Introduction

In this section, structural decomposition of perturbed Markov processes is applied the analysis of fault-tolerant system models. Stochastic models of fault tolerant systems are difficult to analyze due to both their inherent complexity and the presence of very small probabilities related to the fault rates of physical components. The algorithm presented in the previous section, and an extension of it dealing with multiple unknown parameters, provides a useful tool in analyzing the behavior of such systems.

A *fault tolerant system* is designed to function in spite of failures of certain components which make up the system. This is typically accomplished through the use of redundant critical components and the introduction of a system to manage that redundancy. The goal of the design of such a system can be viewed as trying to create a system with an overall failure rate which is much smaller than that of the individual components which make up the system. The complexity of a model describing such a system results from the interaction of the redundancy management system and the behavior of the components themselves. In the design of such a system, some parameters cannot be changed easily. For example, failure rates of physical components might be viewed as constants. Other parameters, particularly those related to the redundancy management system, may be easily changed. For example, a tradeoff may be made between the detection rate and false alarm rate of a fault detection system. An important part of the design procedure is determining a proper set of these parameters.

One approach to the analysis of such systems which has been employed is to construct detailed Markov and semi-Markov representations of the entire system and to study their behavior. In order to deal with the difficulties introduced by the presence of the small failure rate, asymptotic analyses have been employed. A review of several such asymptotic methods is available in [19]. Some of these methods are based on previously available results related to the *two time scale* decomposition of perturbed Markov and semi-Markov processes discussed in the previous chapters.

Such approximation methods have been applied to complex fault tolerant systems [8] [46].

In order to address the complexity as well as the very large range in transition rates inherent in fault tolerant system systems, the remainder of this section addresses the ***structural decomposition*** of such Markov models of fault tolerant systems. The goal here is not to approximate the detailed behavior but rather to determine the structure of the various time scale models. In this way, it may be possible to identify the various operating or failure "modes" of the system. This goal of determining the multiple time scale structure is extended to deal with the presence of several unknown small parameters. It will be shown that using a minor modification of Algorithm 6.1, several structural decompositions can be obtained such that each is associated with a particular set of constraints on the relationship of these unknown parameters. Using this information, it may be possible to design a system which has a desirable multiple time scale structure. Also, by identifying the underlying constraints on the unknown parameters, some sort of "optimal" design procedure could be developed.

In the next subsection, background to the fault tolerant system problem is presented. In Section 6.3.3, the specific problem is formulated and the extended algorithm is presented. Results of applying this algorithm to a simple fault tolerant system model are presented in Section 6.3.4.

6.3.2 Fault-tolerant systems

Several aspects of fault tolerant systems complicate the analysis of these systems. These aspects, however, make the problem particularly suited to an analysis based on multiple time scale structure.

First, due to the very small probabilities associated with the failure rates of various components, there is a natural parameterization of the system in a perturbation form. It is straightforward to consider the failure events as a perturbation of a nominal system where there are no failures. There may also be other sources of "small" parameters. For example, the time until a failure is detected may be short compared to the time between failures but long compared to other events in the system such as normal operating dynamics or the time to fix or replace a failed component. Therefore, there are generally two, and possibly several, parameters of such a system with widely separated magnitudes. Such a wide range of parameters can be associated with overall system dynamics evolving at several time scales. Such a separation of time scales generally poses problems if one uses other analysis methods. For instance, numerical solution of a linear system model might be difficult due to the inherent ill-conditioning of the system. Also, if a simulation approach is taken, a very long time interval must be generated in order to observe the behavior at all the time scales. Furthermore, traditional combinatorics-based reliability analysis methods do not capture the interaction of the dynamics of the physical system and the fault management algorithm. Given these problems, an approximate multiple time scale decomposition would be useful.

Another relevant aspect of fault tolerant systems is the complexity of the overall system model. This complexity results from the interaction of several components, each with its own dynamic characteristics. For example, in a simple system where there are several redundant sensors and a system managing this redundancy through some sort of fault detection and isolation algorithm, the combined system can have a very large state space. The state of the entire system is the combined state of all the sensors and the state of the detection algorithm. As more sophisticated algorithms are employed, and more complex component models are incorporated, there is a combinatorial growth in the complexity of the overall model. There must be a tradeoff in the accuracy of the model describing the overall system and the "accuracy" of the analysis method. The two extremes are not particularly useful; an analytic solution to an oversimplified model is no more useful than an intractable but very detailed model. The structural decomposition algorithm described in this chapter is in some sense a compromise between accuracy of the model and accuracy of the solution.

Third, an interpretation of the the goal of increasing the time until failure of the overall system is that the time scale at which the total failure occurs should be increased over its rate if no redundancy management system were used. For example, consider a system with two redundant components where once an initial failure occurs, the backup unit is relied upon. Although the time until failure is increased over that obtained through the use of a single component, the time scale at which the overall failure occurs has not changed. If, however, a failed component is repaired once its failure is detected, and the repair time is an order of magnitude faster than the failure time of a single component, there is an order of magnitude increase in the total failure time of the overall system. Overall failure only occurs when the backup component fails before the other component is repaired. The inclusion of a repair or replacement policy in a fault tolerant system allows one to design systems which have an overall failure time which is orders of magnitude longer than that of any of the individual components. Including this type of repair policy in the overall model allows one to include such things as periodic maintenance and retirement of components after a certain age. It is these types of complex models which are most interesting for multiple time scale analysis.

Finally, despite the use of a fault management system, there is inevitably some sequence of events which results in total failure. Therefore, in some sense, any fault tolerant system model is composed predominantly of transient states. Only the unrecoverable failure states are trapping. The goal in analysis of a fault tolerant system is to identify the time until some such trapping state is reached, and the sequence of events, and the timing of those events, which lead to that state. This observation implies that any sort of steady-state analysis of a fault-tolerant system model is necessarily inadequate. Furthermore, if we consider the model assuming that no failures can occur (the unperturbed model), then any state which represents a partial failure which can be repaired or replaced will be transient. Only the completely functioning and the completely failed states will then be trapping states. Both these aspects point out the necessity of considering the transient characteristics of fault tolerant system models.

6.3.3 Problem formulation

In this thesis, the analysis of fault tolerant systems is restricted to the identification of the relationship of several unknown or undetermined parameters in the system model and the resulting multiple time scale structural decomposition. The problem formulation is as follows. We begin with a Markov model of the system. As in the previous development, there are some very small transition rates. In this case, however, there are several perturbation terms, $\epsilon_1, \epsilon_2, \ldots, \epsilon_m$. In order to analyze the slow time scale behavior of the system, using the algorithms presented, the relative magnitudes of these terms must be specified. One alternative is to simply assume that each of these small terms is a known analytic function of a single term ϵ. This basic type of reasoning has been used in [26] and [34]. The previously developed algorithms can then be applied. This alternative of expressing each of the small parameters in terms of a single quantity has several shortcomings. First the long term behavior of the system may be quite sensitive to this parameterization. Second, as will be discussed below, there may be some freedom in choosing this parameterization in the design of a system.

Another approach, which will be explored in this section, is to assume that each of the parameters is indeed some analytic function of a single term, ϵ, but that function is unknown. In particular, we assume that ϵ_i is strictly $O(\epsilon^{x_i})$ [1], where x_i is a nonnegative integer, but otherwise unconstrained. For any particular set of values for $x_1, x_2, \ldots$, a set of slow time scale graphs results from applying Algorithm 6.1. Instead of exhaustive enumeration of the decompositions associated with each of the possible sets of these values, an algorithm is developed which provides a small set of decompositions, each associated with a particular set of linear inequalities which the parameters x_i must satisfy. In this way, the relationship of the values of the unknown parameters and the overall structural decomposition is available to be used as an analysis or design tool.

The goal of the analysis algorithm is therefore to provide a small set of decompositions where each decomposition consists of a sequence of graphs $G_0^{(i)}$ and the associated sets $\mathcal{U}^{(i)}$ and $\mathcal{V}^{(i)}$, *and* a set of linear inequalities on the values x_k which must be satisfied for the decomposition to be valid.

Example 6.2 *In order to illustrate this goal, consider first the graph, $G^{(0)}$, in Figure 6.7. Applying Algorithm 6.1 gives the remaining sequence of graphs shown in Figure 6.7.*

Note first that there are only two graphs which have any zero weight edges, $G^{(1)}$

[1] That is, $\lim_{\epsilon \downarrow 0} \epsilon_i / \epsilon^{x_i} = \mu_i$, $0 < \mu_i < \infty$.

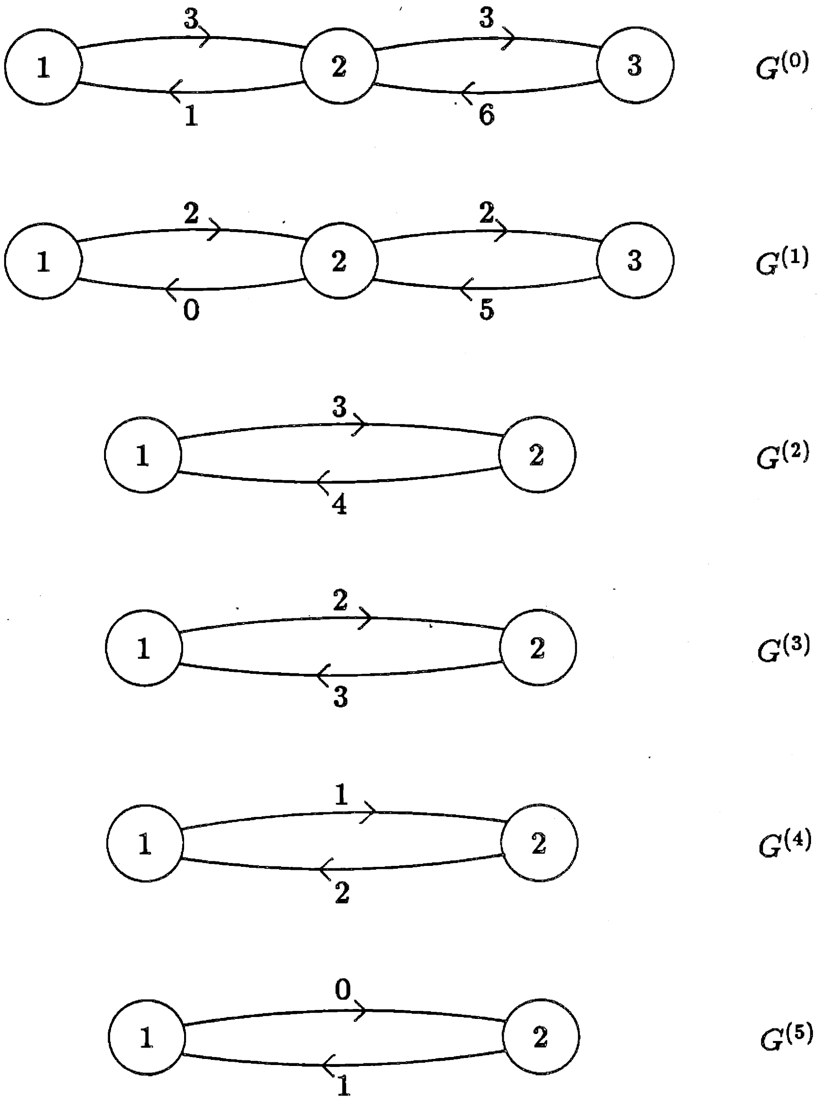


Figure 6.7: Graphs $G^{(i)}$ for Example 6.2

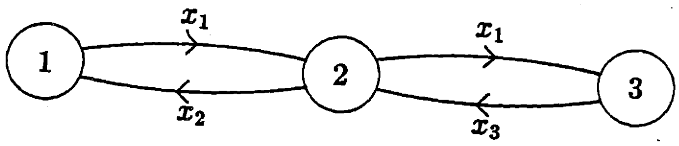


Figure 6.8: Graph $G^{(0)}$ for Example 6.3

and $G^{(5)}$. This corresponds to there being only two time scales which exhibit any dynamic behavior, $t = O(1/\epsilon)$ and $t = O(1/\epsilon^5)$. Therefore, the graphs $G_0^{(1)}$ and $G_0^{(5)}$ are sufficient to reconstruct all the graphs $G_0^{(0)}, \ldots, G_0^{(5)}$.

A second feature to note is that the particular weight of the $3 \to 2$ edge does not affect the decomposition as long as it is ≥ 6. There is therefore an infinite set of weights for this graph which result in the same decomposition. □

The type of decomposition used in Example 6.2 using Algorithm 6.1 can also be performed using the unknown order of magnitude formulation presented above.

Example 6.3 *Consider the graph in Figure 6.8. Note that the graph in the previous example is an instance of this graph with $x_1 = 3, x_2 = 1, x_3 = 6$. Following the same type of procedure, consider the case where* a priori *we know that*

$$x_i \geq 0,\ i = 1,2,3 \quad \text{and} \quad x_3 > 2x_1 - x_2 > x_2 \tag{6.14}$$

(which the values in the previous example satisfy). The graphs in the sequence constructed are therefore as shown in Figure 6.9. The zero weight graphs of interest in this case are $G_0^{(x_2)}$ and $G_0^{(2x_1-x_2)}$ as shown in Figure 6.10. Note that the decomposition depends on the particular inequalities which must be satisfied. For instance, if the linear constraints were

$$x_i \geq 0,\ i = 1,2,3 \quad \text{and} \quad 2x_1 - x_2 > x_3 > x_2 \tag{6.15}$$

then the graphs of interest would be $G_0^{(x_2)}$ and $G_0^{(x_3)}$ as shown in Figure 6.11. □

The basic algorithm is structured as a tree search. Each node of the tree has associated with it a graph $G^{(k)}$ and a set of linear inequalities on the unknowns x_i.

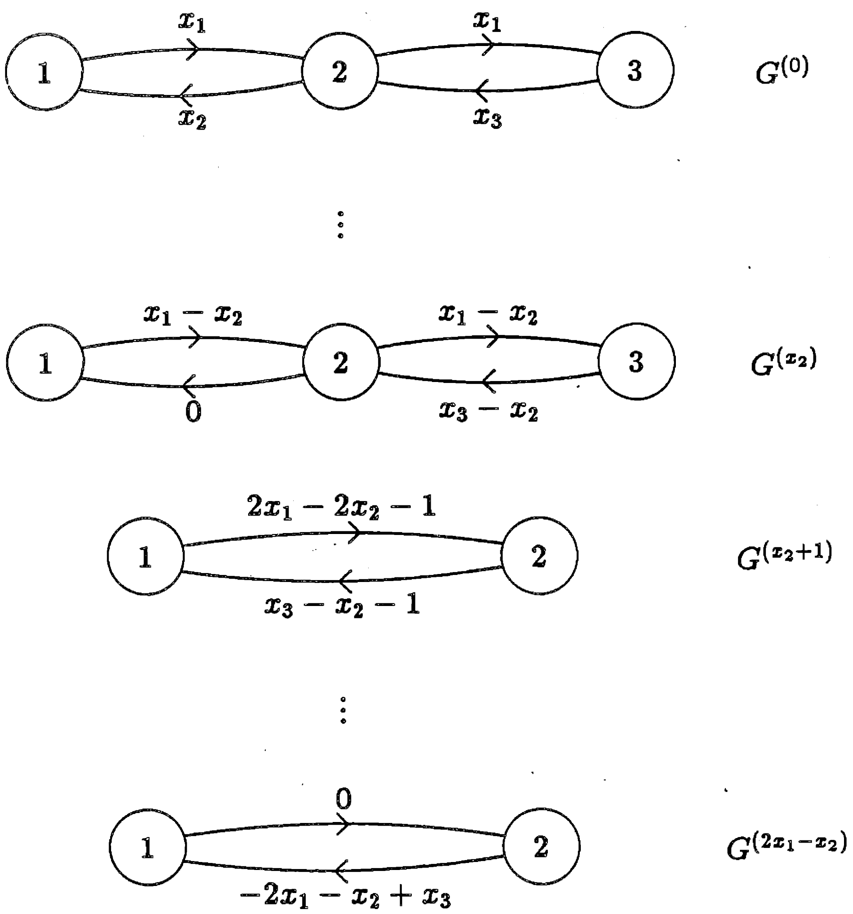


Figure 6.9: Graphs $G^{(i)}$ for Example 6.3

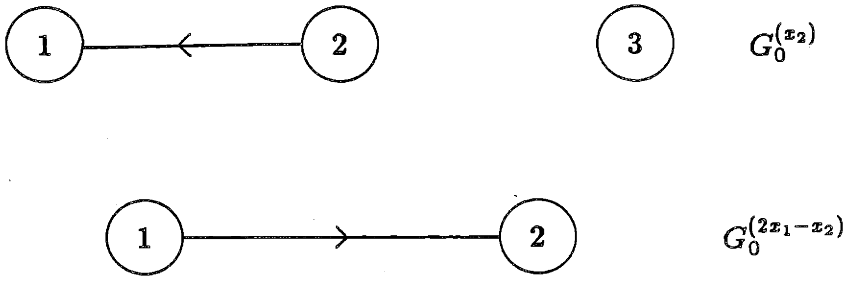


Figure 6.10: Graphs $G_0^{(i)}$ for Example 6.3

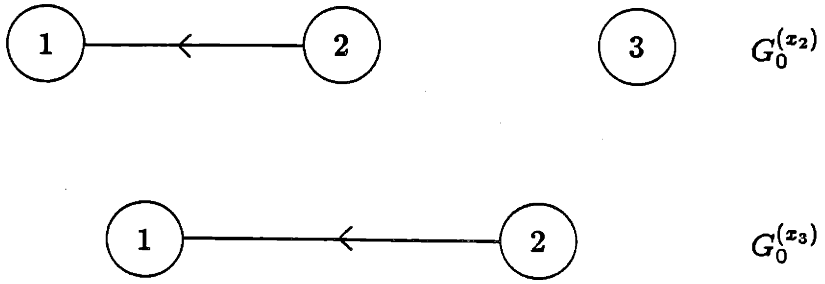


Figure 6.11: Graphs $G_0^{(i)}$ for Example 6.3 using alternative constrains

The weights of the edges of the graph are no longer constants but rather are linear combinations of the unknowns. Due to the constraints on the unknowns, each of the weights is guaranteed to be a nonnegative integer for any feasible set $\{x_i\}$.

The recursive step of the algorithm involves extending the branches from such a node. Conceptually, each subset of edges is considered in turn to see whether it is feasible for all the edges in that set to have the minimum weight in the graph, (i.e. whether there exist integer values for all the x_i satisfying all constraints and so that the particular subset of edges under consideration all have the same minimum weight). If a set of edges has such a feasible set of weights, then a branch is added in the tree search. The inequalities associated with the new node in the tree are those of the parent and the new inequalities introduced with the assumption that the edges precisely form the minimum weight set. The new graph is constructed in two steps. First, the weights of all the edges in the graph are modified by subtracting the minimum weight from all the edges. This leaves the edges in the minimum weight set with weight identically zero. The remaining weights in the graph must be strictly positive for any feasible set of unknowns. This graph is then decomposed using essentially one iteration of Algorithm 6.1 (extended to deal with the unknown values of the positive weights).

The major difficulty encountered in extending Algorithm 6.1 to multiple unknowns is the computation of the shortest paths $\tilde{d}$. To deal with the possibility that different paths may be the shortest for different feasible values of the unknowns, the weight of the shortest path must be expressed as the minimum of several terms. There are some simplifications which can be performed. For instance,

$$\min(x, 2x, y) = \min(x, y) \text{ if } x \geq 0 \tag{6.16}$$

The generalization is that given a set $\mathcal{X}$ of feasible values for the unknown parameters, an expression of the form

$$\min(f_1(x), \ldots, f_n(x)) \tag{6.17}$$

can be simplified by eliminating any $f_k(x)$ which for each $x' \in \mathcal{X}$ there exists some other $f_j(x)$, $j \neq k$, such that $f_j(x') < f_k(x')$.

If there are no remaining edges in the graph, then a complete decomposition has been produced. The set of inequalities which must be satisfied for this decomposition to be valid is associated with the leaf node of the tree. The graphs associated with the decomposition can be found by following the tree to the root and computing the zero weight graphs at each node.

Example 6.4 *In order to illustrate this procedure, consider the graph shown in Figure 6.12.*

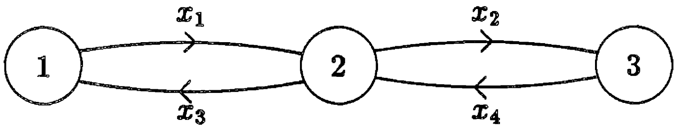


Figure 6.12: Initial graph $G^{(0)}$ for Example 6.4

Initially, assume that there are no inequalities imposed except

$$x_i \geq 0\ ,\ i = 1, 2, 3, 4 \tag{6.18}$$

Consider one possibility for the set $\mathcal{L}$ of lowest weight edges

$$\mathcal{L} = \{(2, 1, x_3)\} \tag{6.19}$$

The additional inequalities introduced with this choice are

$$x_3 < x_1\ ,\ x_3 < x_2\ ,\ x_3 < x_4 \tag{6.20}$$

which combined with the initial inequalities is feasible, i.e. there exist values for x_i, i=1,2,3,4 *consistent with all these inequalities. The "scaled" graph after subtracting the minimum weight is therefore as illustrated in Figure 6.13.*

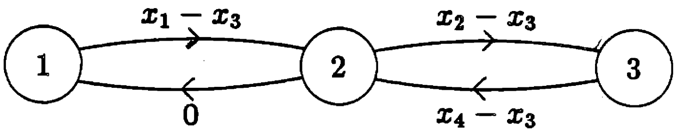


Figure 6.13: Scaled graph for Example 6.4

The result of applying one iteration of Algorithm 6.1 to this graph is shown in Figure 6.14.

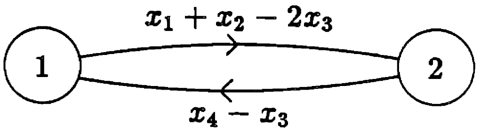


Figure 6.14: Graph after one iteration of Algorithm 6.1 for Example 6.4

The sets associated with this graph are

$$\mathcal{U}_1 = \{1\}\quad ,\quad \mathcal{U}_2 = \{3\} \tag{6.21}$$

$$\mathcal{V}_1 = \{1,2\}\quad ,\quad \mathcal{V}_2 = \{3\} \tag{6.22}$$

There are now three possible sets of lowest weight edges which can be considered:

$$\mathcal{L}_1 = \{(1,2,x_1+x_2-2x_3)\} \tag{6.23}$$

$$\mathcal{L}_2 = \{(2,1,x_4-x_3)\} \tag{6.24}$$

$$\mathcal{L}_3 = \{(1,2,x_1+x_2-2x_3),(2,1,x_4-x_3)\} \tag{6.25}$$

Let us consider extending the node associated with the set $\mathcal{L}_1$. The inequality introduced with this choice is

$$x_1 + x_2 - 2x_3 < x_4 - x_3 \tag{6.26}$$

which is feasible. The graph after subtracting the minimum weight is shown in Figure 6.15.

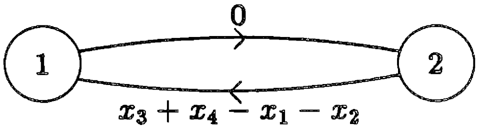


Figure 6.15: Graph after subtracting minimum weight for Example 6.4

The complete algorithm would recursively identify all the leaves associated with this sort of decomposition. □

6.3.4 Results

In order to demonstrate the utility of the modified algorithm outlined above, consider a simple system with three failure prone components. For example, these components may be redundant sensors in a nuclear reactor which are required for proper operation. Components are either *in service* if they are believed to be functioning properly or *out of service* if the fault detection algorithm has identified that component as a probable failure. *Total failure* occurs when more than half of the components which are in service have failed. A failure detection algorithm is employed to identify the component failures. In addition to detecting failures, this algorithm generates false alarms. Once a real failure has been detected, that component is repaired or replaced and returned to service. False alarms are returned to service after they have been identified as false alarms. The basic state transition rates in the system model are therefore

- Failure rate of individual components in service, $O(\epsilon^{x_1})$. (Components out of service do not fail).

- False alarm rate of the failure detection algorithm, $O(\epsilon^{x_2})$.

- Detection rate of real failure, $O(\epsilon^{x_3})$.

- Repair rate of a failed (and detected) component, $O(\epsilon^{x_4})$.

- Rate of recognition of a false alarm, $O(\epsilon^{x_5})$.

Several features of such a system are of interest. First, if no failure detection algorithm were used to identify and then repair failed components, then total failure would occur at a rate $O(\epsilon^{x_1})$. Although redundancy is available through the use of several independent components, the order of magnitude of the time to failure is unchanged from that of a single component. The basic purpose of the failure detection algorithm is to make the system tolerant of these failures. A desirable effect is to increase the order of magnitude of the expected time until total failure occurs. Another aspect which should be noted is that it may not be possible to independently set the values of the x_i in practice. For instance, there may be little control over the order of the basic failure rate, x_1. It may also not be possible to increase the fault detection rate without also increasing the false alarm rate. Finally, it may be possible to reduce the false alarm rate but at a large cost. It would be beneficial to determine whether such a reduction in false alarm rate would indeed modify the order of magnitude of the total failure rate. For example, it is quite possible that the false alarm rate does not have a strong influence on the overall slow time scale behavior and therefore does not warrant the cost of reducing it.

The states of the particular 7 state system which is considered are enumerated below. The graph $G^{(0)}$ for this system is shown in Figure 6.16.

State 1 All 3 components are functioning and are in service.

State 2 One functioning unit is out of service due to a false alarm. Two other functioning units remain in service.

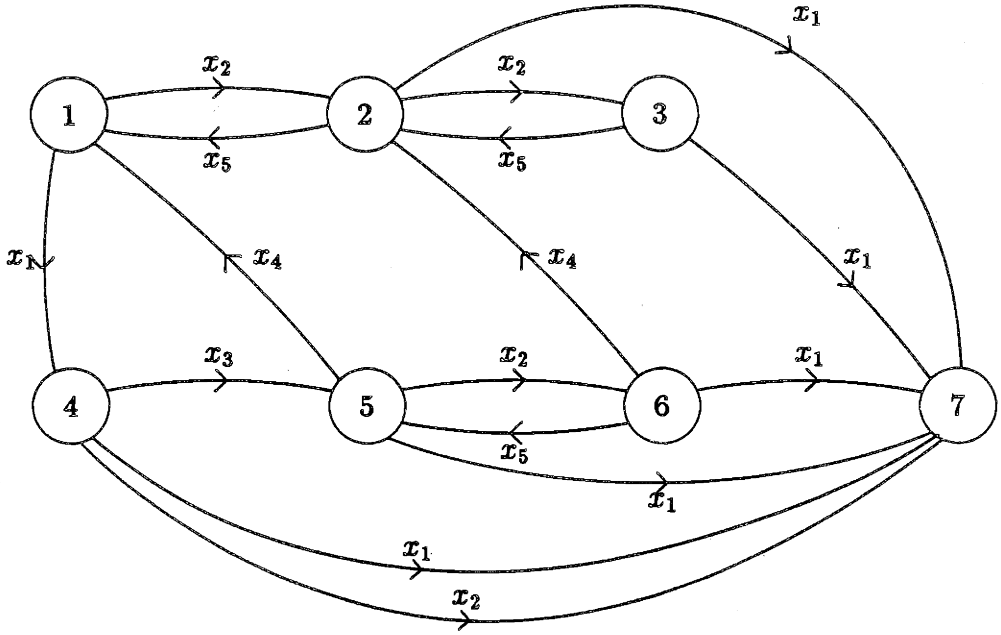


Figure 6.16: Graph $G^{(0)}$ for the fault tolerant system model

State 3 One functioning unit is in service. Two other functioning units are out of service due to false alarms.

State 4 All 3 components are in service. Two are functioning and one has failed.

State 5 One failed unit is out of service. Two functioning units are in service.

State 6 One functioning unit is in service. One failed and one functioning unit are out of service.

State 7 The total failure state. This state is trapping.

The recursive algorithm described above was performed using this system together with the following *a priori* constraints on the unknowns

$$x_i \geq 0 \quad , \quad i = 1,2,3,4,5 \tag{6.27}$$

$$x_1 > x_i \quad , \quad i = 3,4,5 \tag{6.28}$$

$$x_2 \geq x_3 \tag{6.29}$$

The first constraint is always required; the unknowns must be nonnegative. The second constraint corresponds to the natural failure rate, $O(\epsilon^{x_1})$, to be at least an order of magnitude slower that failure detection rate, $O(\epsilon^{x_3})$, the repair rate, $O(\epsilon^{x_4})$, and the detection of false alarm rate, $O(\epsilon^{x_5})$. The final constraint is that the failure detection rate, $O(\epsilon^{x_3})$, must be the same order of magnitude or faster than the false alarm rate, $O(\epsilon^{x_2})$. These constraints are chosen as a typical example of the types of constraints which can be introduced.

The resulting decompositions of interest are those for which the time scale at which total failure occurs is slower that $O(1/\epsilon^{x_1})$, the natural failure time scale of each of the components. By restricting our attention to decompositions with this property, we find that there are only four slower time scales at which total failure can occur. These time scales, and the constraints on the unknowns which must be satisfied are

$$O\left(1/\epsilon^{2x_1 - x_3}\right) \implies x_4 < x_3 < x_5 \; , \; x_1 + x_5 < x_2 + x_3 \text{ or} \tag{6.30}$$

$$x_5 \leq x_4 < x_3 < x_1 \leq x_2 \text{ or} \tag{6.31}$$

$$x_4 < x_5 \leq x_3 < x_1 \leq x_2 \tag{6.32}$$

$$O\left(1/\epsilon^{2x_1 - x_4}\right) \implies x_3 < x_4 \leq x_4 \; , \; x_1 + x_5 < x_2 + x_4 \text{ or} \tag{6.33}$$

$$x_3 \leq x_4 \leq x_5 \; , \; x_1 + x_5 < x_2 + x_4 \text{ or} \tag{6.34}$$

$$x_5 \leq x_3 \leq x_4 \; , \; x_1 + x_3 < x_2 + x_4 \tag{6.35}$$

$$O\left(1/\epsilon^{x_1 + x_2 - x_3}\right) \implies x_5 \leq x_4 < x_3 < x_2 < x_1 \text{ or} \tag{6.36}$$

$$x_4 < x_5 \leq x_3 < x_2 < x_1 \; , \; x_5 < x_2 \text{ or} \tag{6.37}$$

$$x_5 \leq x_3 \leq x_4 \; , \; x_3 < x_2 \; , \; x_2 + x_4 \leq x_1 + x_3 \tag{6.38}$$

$$O\left(1/\epsilon^{x_1 + x_2 - x_5}\right) \implies x_4 < x_3 < x_5 < x_2 \; , \; x_2 + x_3 \leq x_1 + x_5 \text{ or} \tag{6.39}$$

$$x_3 \leq x_4 < x_5 < x_2 \; , \; x_2 + x_4 \leq x_1 + x_5 \text{ or} \tag{6.40}$$

$$x_3 < x_5 \leq x_4 \; , \; x_5 < x_2 \; , \; x_2 + x_4 \leq x_1 + x_5 \tag{6.41}$$

In order to illustrate the type of decomposition which is associated with each of these sets of constraints, consider that associated with (6.30). The set of zero weight

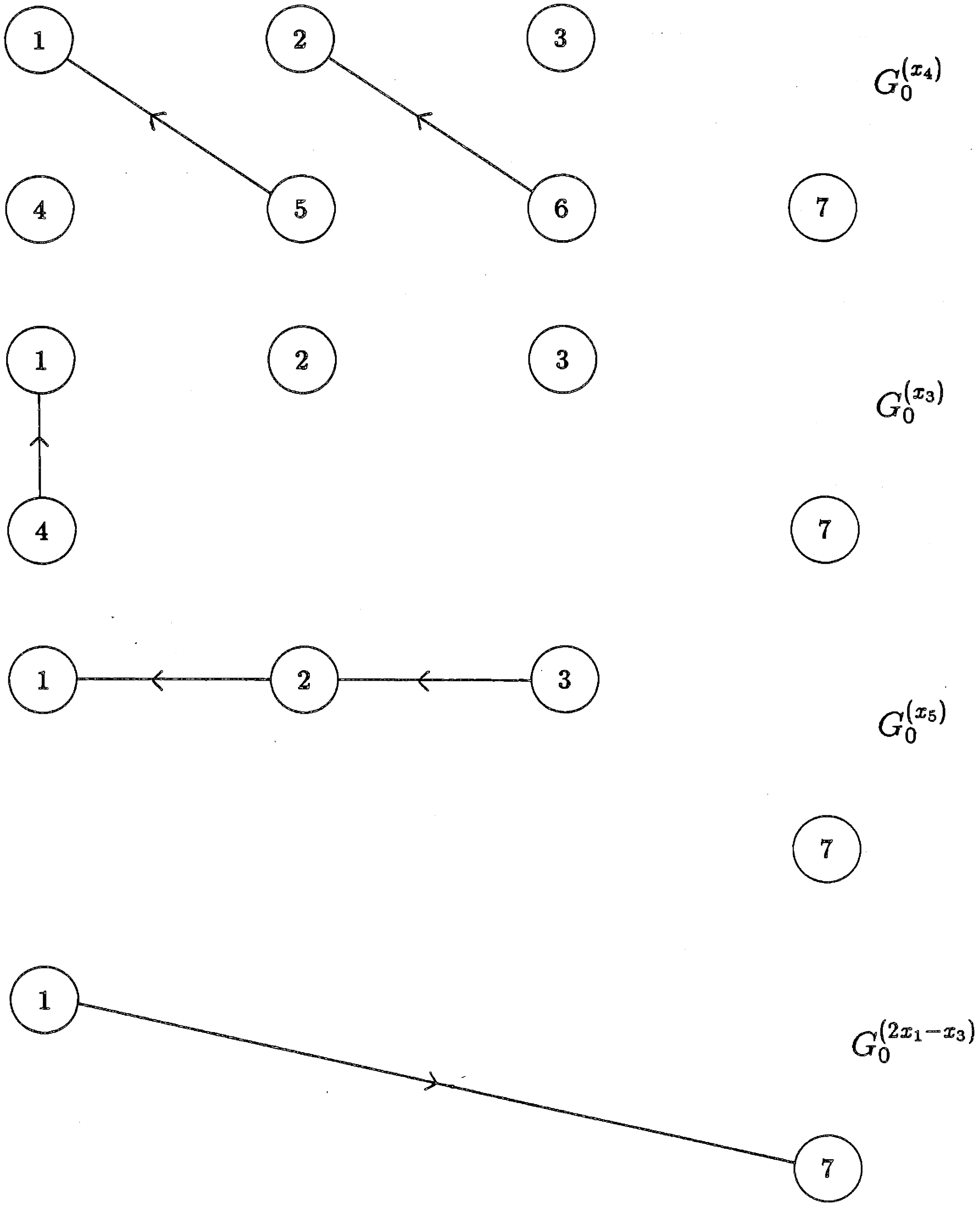


Figure 6.17: Graphs $G_0^{(i)}$ for the fault tolerant system model

graphs associated with these constraints are shown in Figure 6.17. A possible set of values for the unknown parameters is

$$x_1 = 3,\ x_2 = 6,\ x_3 = 1,\ x_4 = 0,\ x_5 = 2 \tag{6.42}$$

This means that failures occur at rate $O(\epsilon^3)$, false alarms at rate $O(\epsilon^6)$, detection of failures at rate $O(\epsilon)$, repairs at rate $O(1)$, and recognition of false alarms at rate $O(\epsilon^2)$. The slowest behavior of the system therefore occurs at a time scale $t = O(1/\epsilon^5)$. By considering the graphs in Figure 6.17, we can see that the fastest behavior occurs at a time scale $t = O(1)$ and corresponds to repair of a failed and detected component. The next slower behavior occurs at a time scale $t = O(1/\epsilon)$ and corresponds to recovery from a real failure. Next, recovery from false alarms occurs at the time scale $t = O(1/\epsilon^2)$. Finally, total failure occurs at the time scale $t = O(1/\epsilon^5)$. This failure corresponds to the failure of a second component before the first failure is detected.

In order to illustrate the utility of this analysis, we can consider the largest false alarm rate possible while preserving the same structural decomposition. The constraint on the parameter x_2, the order of magnitude of the failure rate, is

$$x_2 > x_1 + x_5 - x_3 \tag{6.43}$$

which for the particular values of the other parameters used above corresponds to the constraint $x_2 > 4$. As long as this inequality, as well as the *a priori* constraints, are satisfied, then this decomposition shown in Figure 6.17 is valid. Therefore false alarm rates could occur at a rate $O(\epsilon^5)$ (corresponding to $x_2 = 5$) while preserving the structure.

Although the system described here is artificially simple, the approach can be applied to much more complicated systems. More components could be added and controlled using a similar algorithm. It is possible, for example, to determine how the order of magnitude of the total failure time varies as more components are added. Also, alternative maintenance policies could be investigated.

6.4 Discussion

The two major contributions of this chapter are first, the restatement of the structural aspects of Algorithm 2.1 in simple graph-theoretic terms, and second, a method of using this graphical result to deal with a problem involving multiple perturbation parameters.

Algorithm 6.1 isolates the issues related to the detailed behavior of a system at successive time scales leaving only the necessary aspects related to the structure of the decomposition. This graphical formulation shows that the determination of the structure of the decomposition has a very simple interpretation and implementation.

The algorithm for analysis of systems with multiple perturbation parameters is significant in two ways. First, it shows the utility of the graphical formulation of the decomposition algorithm presented earlier in the chapter. An algorithm dealing with this multiple perturbation parameter case would be very much more tedious if the details of each time scale were desired as well. In terms of a methodology for dealing with multiple perturbation terms in general, this algorithm extends the class of systems which could be considered.

Previous work by Ladde and Siljak [34] and Khalil [26] considered a case where there are two sets of perturbation parameters, $\epsilon_1, \epsilon_2, \ldots, \epsilon_r$ and $\mu_1, \mu_2, \ldots, \mu_s$. These terms satisfy

$$\underline{\epsilon} \le \frac{\epsilon_i}{\epsilon_j} \le \bar{\epsilon} \ \ \forall i, j \tag{6.44}$$

and

$$\underline{\mu} \le \frac{\mu_i}{\mu_j} \le \bar{\mu} \ \ \forall i, j \tag{6.45}$$

Furthermore, by assumption,

$$\mu/\epsilon \rightarrow 0 \text{ as } \epsilon \downarrow 0 \tag{6.46}$$

where ϵ and μ are defined to be the geometric means $(\Pi_i \epsilon_i)^{1/r}$ and $(\Pi_i \mu_i)^{1/s}$ respectively. Further restrictions on the structure of the system result in exactly three time scales of behavior, $t = O(1)$, $t = O(1/\epsilon)$, and $t = O(1/\mu)$.

In contrast to previous work, the decomposition algorithm presented in this chapter can deal with multiple time scale systems. The interpretation of this is that a system may have behavior at a time scale $t = O(1/(\epsilon^m \mu^n))$ for many different values of m and n. Essentially, the results of Siljak and Khalil are extensions of the two time scale results for single perturbation parameter singular systems. The result in this chapter attempts to provide a method for dealing with the multiparameter version of multiple time scale systems.

Chapter 7

Conclusions

Multiple time scale analysis of Markov systems is an important tool in many engineering applications in which there are events which happen with probabilities of different orders of magnitude. In this thesis, we have considered the decomposition of explicitly perturbed, finite state Markov and semi-Markov processes. The orientation of the thesis has been to exploit the particular structure of Markov processes to develop practical analysis tools which have a firm theoretical basis.

7.1 Contributions of thesis

Specific contributions of the thesis can be summarized as follows.

- A straightforward, multiple time scale decomposition algorithm for perturbed, finite state, continuous time and discrete time Markov processes is presented in Chapters 2 and 4. The continuous time algorithm addresses the same general class of systems as do the algorithms of Coderch [11] and Delebecque [16]. However, the algorithm presented here requires much less computation than those algorithms. Furthermore, the approach has a simple probabilistic interpretation and in fact provides a bridge between the simpler algorithms of Simon and Ando, and Courtois which deal with restricted classes of Markov systems and the more complex algorithms of Coderch and Delebecque which deal with more general Markov systems. The approximation which is constructed based on the decomposition has the property that the approximation error converges uniformly to zero over the entire interval $t \geq 0$. The algorithms take the form of recursively applied aggregation steps. That is, at each time scale, an aggregated slow time scale system is derived. That aggregated system is then analyzed using the same algorithm. This continues until the final system analyzed has only one time scale of behavior. Finally, the discrete time algorithm shows that the approximation of a discrete time Markov chain takes a "hybrid" form where the fast time scale is approximated by a discrete time chain and all slower time scales are approximated by a *continuous time* Markov process.

- An algorithm for the multiple time scale decomposition of perturbed semi-Markov processes is presented in Chapter 3. There are two major contributions of this result which extend currently available results in the literature. First, systems with more than two time scales and arbitrary ergodic structure at each time scale can be analyzed. Second, the holding time probability distributions as well as the state transition probabilities in the system can be functions of the explicit perturbation parameter. In previous work where only the transition probabilities are perturbed, the slow time scale system is described by a Markov model. In the more general case considered in this thesis, the slow time scale is described by another semi-Markov system. Another consequence of this more general form of perturbation is that the aggregation step to form a slow time scale semi-Markov model can split a state such that parts of the state at the fast time scale are in *different* aggregates at the slow time scales.

- A decomposition algorithm for a class of perturbed positive systems is presented in Chapter 5. This result demonstrates the role played by the structure of a positive system. While the Markov decomposition result is not directly applicable to positive systems, a modification involving computation of the dominant left eigenvector allows application of a similar algorithm to a class of positive systems which share similar structure with the class of Markov systems. Potential applications of this result include analysis of compartmental models.

- The Markov decomposition algorithm developed in Chapter 2 is expressed in graphic-theoretic terms to provide a straightforward graphical algorithm which provides the structure of the multiple time scale decomposition. The algorithm is based on simple concepts of connectivity of sets of nodes in the graph and shortest paths between sets of states where the path lengths are nonnegative integers.

- The graphical decomposition result is extended in Chapter 6 to address the analysis of a fault-tolerant system model. This application demonstrated that the computation of the multiple time scale structure of a system with several types of rare events can be useful in providing intuition into the overall behavior of the system and in particular into the way in which this behavior changes as the relative orders of magnitude of the various transition rates are altered.

7.2 Future work

The work presented in this thesis suggests many areas for future work. In many ways, the relatively simple results for the decomposition of Markov systems suggests that more general algorithms for the multiple time scale decomposition of linear systems might benefit from detailed consideration of the structure of the system being considered. Several potential areas of future work are listed below.

- The proof of the validity of the Markov decomposition result presented in Chapter 2 relies on the finiteness of the state space. It is not clear, however, that such a restriction is really necessary. For example, extension to systems on a denumerable state set should be feasible. An example of such a system would be an infinite birth/death process. Not only would such an extension increase the set of systems which can be analyzed directly, extension to infinite state sets would also allow application of the Markov result to the decomposition of semi-Markov processes without restriction to holding time transforms which are rational.

- A direct approach to the decomposition of semi-Markov processes would be desirable. As discussed above, removing the restriction to rational polynomial holding time transforms seems possible. This may possibly be accomplished by direct spectral analysis of the probability transition function without realization in an intermediate Markov form.

- Although a decomposition result is obtained for a particular class of positive systems, a more general result would be desirable. Such a result will likely have to take advantage of the particular structure of positive systems which is surveyed in Chapter 5. Once a general result is available, direct consideration of more general cone-invariant (K-positive [3]) systems may also be possible.

- In Chapter 6, a graph-theoretic decomposition result is presented. An open question which remains concerns the investigation of graph simplification (pruning) procedures that can be performed while maintaining the multiple time scale structure. This type of result would have implications on the computation complexity of both the structural decomposition algorithm presented in Chapter 6 and the full algorithm presented in Chapter 2.

- Computational aspects of the algorithms presented in the thesis have not been explored. Though these algorithms seems less complex than those previously available, a more detailed analysis of computational complexity and the development of numerically robust procedures would be desirable. Of interest here is the use of the structural algorithm as a first step followed by the numerical computation at each time scale. Such an algorithm should avoid the numerically sensitive problem of finding the dimension of the approximation at each time scale.

- Possible applications of the decomposition algorithms should be investigated. For instance, an interesting application which has not been explored is the development of estimation algorithms which exploit the multiple time scale structure of a system. For example, "asymptotically optimal" estimation of a state trajectory based on noisy observations may be possible.

- An interpretation of many of the decomposition results presented in this thesis is that through the particular constructions derived, approximations of the true eigenprojection of the zero-group, $P(\epsilon)$, are effectively specified. This is in contrast to simpler but relatively restrictive results which assume equilibration of the "fast" dynamics before the aggregates interact, i.e. results which effectively use $P(0)$ as their approximation. An open question is how can the "important" terms of $P(\epsilon)$ be identified in a computationally efficient manner. This could be used to provide a direct, computationally feasible, decomposition algorithm which may be applicable to general linear systems.

Bibliography

[1] A. Ando, F. M. Fisher, and H. A. Simon. Near-decomposability, partition and aggregation and the relevance of stability discussions. In A. Ando and F. M. Fisher, editors, *Essays on the Structure of Social Science Models*, MIT Press, Cambridge, Massachusetts, 1963.

[2] T. Ando. Inequalities for M-matrices. *Linear and Multilinear Algebra*, 8(4):291–316, 1980.

[3] Abraham Berman and Robert J. Plemmons. *Nonnegative Matrices in the Mathematical Sciences.* Academic Press, New York, 1979.

[4] A. Bharucha-Reid. *Elements in the Theory of Markov Processes and Their Applications.* 1960.

[5] G. Blankenship. Singularly perturbed difference equations in optimal control problems. *IEEE Transactions on Automatic Control*, AC-26(4):911–917, August 1981.

[6] A. Bobbio and K. S. Trivedi. *An Aggregation Technique for the Transient Analysis of Stiff Markov Chains.* Computer Science Memo CS-1984-25, Duke University, 1985. (to appear in *IEEE Transactions on Computers*).

[7] W.-L. Cao and W. J. Stewart. Iterative aggregation/disaggregation techniques for nearly uncoupled Markov chains. *Journal of the ACM*, 32(3):702–719, July 1985.

[8] S. K. Chu. *Approximate behavior of generalized Markovian models of fault-tolerant systems.* Master's thesis, MIT, 1986. Department of Aeronautics and Astronautics.

[9] M. Coderch. *Multiple Timescale Approach to Hierarchical Aggregation of Linear Systems.* PhD thesis, MIT, 1982.

[10] M. Coderch, A. S. Willsky, S. S. Sastry, and D. A. Castinon. Hierarchical aggregation of linear systems with multiple time scales. *IEEE Transaction on Automatic Control*, AC-28(11), 1983.

[11] M. Coderch, A. S. Willsky, S. S. Sastry, and D. A. Castinon. Hierarchical aggregation of singularly perturbed finite state Markov processes. *Stochastics*, 8:259–289, 1983.

[12] P. J. Courtois. *Decomposability: Queuing and Computer Systems Applications*. Academic Press, New York, 1977.

[13] P. J. Courtois and P. Semal. Error bounds for the analysis by decomposition of non-negative matrices. In G. Iazeolla and P. J. Courtois, editors, *Mathematical Computer Performance and Reliability*, pages 209–224, North Holland, Amsterdam, 1984.

[14] D. R. Cox. A use of complex probabilities in the theory of stochastic processes. *Proceedings of the Cambridge Philosophical Society*, 51:313–319, 1955.

[15] P. G. Coxson. *Model Reduction: Identifying Partitions from Structured Aggregates*. Technical Report LIDS-P-1400, MIT, 1984.

[16] F. Delebecque. A reduction process for perturbed Markov chains. *SIAM Journal of Applied Mathematics*, 43(2), 1983.

[17] F. Delebecque and J.-P. Quadrat. The optimal cost expansion of finite controls, finite state Markov chains. In *Analysis and Optimization of Systems*, pages 322–337, Springer-Verlag, 1980. Lecture notes in Control and Information Science no. 28.

[18] R. W. Floyd. Algorithm 97 — shortest path. *Communications of ACM*, 5:345, 1962.

[19] I. B. Gertsbakh. Asymptotic methods in reliability theory: a review. *Advances in Applied Probability*, 16:147–175, 1984.

[20] D. V. Gusak and V. S. Korolyuk. The asymptotic behavior of semi-Markov processes with decomposable sets of states. *Theory of Probability and Mathematical Statistics*, 5, 1975.

[21] R. Howard. *Dynamic Probabilistic Systems. Volume 2.* Wiley, 1971.

[22] T. Kato. *Perturbation Theory for Linear Operators*. Springer-Verlag, Berlin, 1966.

[23] J. Keilson. *Markov Chain Models: Rarity and Exponentiality*. Springer-Verlag, New York, 1978.

[24] J. Keilson. *Robustness of The Failure Time Distribution for Systems of Independent Components.* Working Paper QM 8426, University of Rochester, 1984.

[25] J. G. Kemeny and J. L. Snell. *Finite Markov Chains.* Van Nostrand, New York, 1960.

[26] H. K. Khalil and P. V. Kokotovic. Control of linear systems with multiparameter singular perturbation. *Automatica*, 15:197, 1979.

[27] L. Kleinrock. *Queuing Systems. Volume I: Theory.* Wiley, New York, 1975.

[28] P. Kokotovic, R. O'Malley, and P. Sannuni. Singular perturbation and order reduction in control theory — an overview. *Automatica*, 12:123–132, 1976.

[29] P. V. Kokotovic. Applications of singular perturbation techniques to control problems. *SIAM Review*, 26(4):501–550, 1984.

[30] P. V. Kokotovic. Singular perturbations and iterative separation of timescales. *Automatica*, 16:23–24, 1980.

[31] V. S. Korolyuk, L. I. Polishchuk, and Tomusak. On a limit theorem for semi-Markov processes. *Cibernetics*, 5(4):524–526, 1969.

[32] V. S. Korolyuk and A. F. Turbin. Asymptotic enlarging of semi-Markov processes with an arbitrary state space. In A. Dold and B. Eckman, editors, *Proceedings Third Japan-USSR Symposium on Probability Theory*, pages 297–315, Springer-Verlag, Berlin, 1972. Lecture Notes in Math no. 550.

[33] V. S. Korolyuk and A. F. Turbin. On the asymptotic behavior of the occupancy time of semi-Markov processes with an arbitrary state space. *Theoretical Probability and Mathematical Statistics*, 2:133–143, 1974.

[34] G. S. Ladde and D. D. Siljak. Multiparameter singular perturbations of linear systems with multiple time scales. *Automatica*, 19(4):385–394, 1983.

[35] X.-C. Lou, J. R. Rohlicek, P. G. Coxson, G. C. Verghese, and A. S. Willsky. Time scale decomposition: the role of scaling in linear systems and transient states in finite state Markov processes. In *Proceedings of the American Control Conference*, June 1985.

[36] Xi-Cheng Lou. *An algebraic approach to time scale analysis and control.* PhD thesis, MIT, October 1985.

[37] David G. Luenberger. *Introduction to Dynamic Systems.* Wiley, New York, 1979.

[38] Y. Ohta, H. Maeda, S. Kodama, and T. Itoh. Fault diagnosis of compartmental systems. *Electronics and Communication in Japan*, 67-A(8):833–840, August 1984.

[39] Uriel G. Rothblum. Algebraic eigenspace of nonnegative matrices. *Linear Algebra and its Applications*, 12:281–292, 1975.

[40] V. R. Saksena, J. O'Reilly, and P. V. Kokotovic. Linear perturbation and time scale methods in control theory. Survey: 1976–1983. *Automatica*, 20(3):273–293, 1984.

[41] C. Sauer and K. M. Chandy. *Computer Performance Modelling.* Prentice-Hall, New Jersey, 1981.

[42] H. Simon and A. Ando. Aggregation of variables in dynamic systems. *Econometrica*, 29(2):111–139, April 1961.

[43] K. S. Trivedi and R. M. Geist. Decomposition in reliability analysis of fault-tolerant systems. *IEEE Transactions on Reliability*, R-32(5), 1983.

[44] J. N. Tsitsiklis. *Markov Chains with Rare Transitions and Simulated Annealing.* Technical Report LIDS-P-1497, MIT, September 1985.

[45] H. T. Vantilborgh. Aggregation with an error $O(\epsilon^2)$. *Journal of the ACM*, 32(1):162–190, 1985.

[46] B. K. Walker. *A Semi-Markov Approach to Quantifying Fault-Tolerant System Performance.* PhD thesis, MIT, 1980. Department of Aeronautics and Astronautics.

[47] B. K. Walker and D. K. Gerber. Evaluation of fault-tolerant system performance by approximate techniques. In *Proceedings of* 7th *IFAC Symposium on Identification and System Parameter Estimation*, July 1985.