

# Model Reduction for Hidden Markov Models

by

Georgios Kotsalis

Submitted to the Department of Mechanical Engineering  
in partial fulfillment of the requirements for the degree of

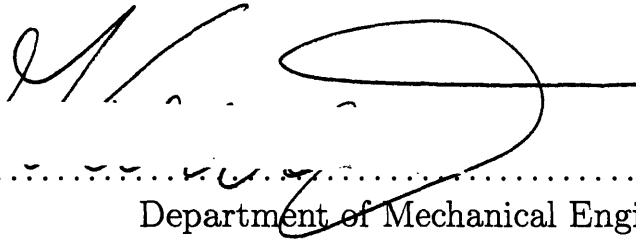
Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2006

© Massachusetts Institute of Technology 2006. All rights reserved.

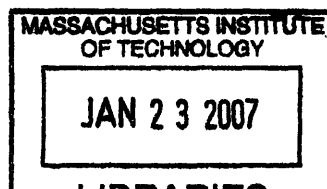


Author .....  
Department of Mechanical Engineering  
August 19, 2006

Certified by .....  
Munther A. Dahleh  
Professor of Electrical Engineering  
Thesis Supervisor

Certified by .....  
Alexandre Megretski  
Professor of Electrical Engineering  
Thesis Supervisor

Accepted by .....  
Lallit Anand  
Professor of Mechanical Engineering  
Chairman, Department Committee on Graduate Studies



ARCHIVES

# Model Reduction for Hidden Markov Models

by

Georgios Kotsalis

Submitted to the Department of Mechanical Engineering  
on August 19, 2006, in partial fulfillment of the  
requirements for the degree of  
Doctor of Philosophy

## Abstract

The contribution of this thesis is the development of tractable computational methods for reducing the complexity of two classes of dynamical systems, finite alphabet Hidden Markov Models and Jump Linear Systems with finite parameter space. The reduction algorithms employ convex optimization and numerical linear algebra tools and do not pose any structural requirements on the systems at hand.

In the Jump Linear Systems case, a distance metric based on randomization of the parametric input is introduced. The main point of the reduction algorithm lies in the formulation of two dissipation inequalities, which in conjunction with a suitably defined storage function enable the derivation of low complexity models, whose fidelity is controlled by a guaranteed upper bound on the stochastic  $L_2$  gain of the approximation error. The developed reduction procedure can be interpreted as an extension of the balanced truncation method to the broader class of Jump Linear Systems.

In the Hidden Markov Model case, Hidden Markov Models are identified with appropriate Jump Linear Systems that satisfy certain constraints on the coefficients of the linear transformation. This correspondence enables the development of a two step reduction procedure. In the first step, the image of the high dimensional Hidden Markov Model in the space of Jump Linear Systems is simplified by means of the aforementioned balanced truncation method. Subsequently, in the second step, the constraints that reflect the Hidden Markov Model structure are imposed by solving a low dimensional non convex optimization problem. Numerical simulation results provide evidence that the proposed algorithm computes accurate reduced order Hidden Markov Models, while achieving a compression of the state space by orders of magnitude.

Thesis Supervisor: Munther A. Dahleh  
Title: Professor of Electrical Engineering

Thesis Supervisor: Alexandre Megretski  
Title: Professor of Electrical Engineering



## Acknowledgments

*"I made this letter longer than usual because I lack the time to make it short."*

- Blaise Pascal.

If you're still a student and recall things like the Busta Rhymes concert at Johnson, Newbury Comics at the student center, or even the IHOP at Kenmore Square, then you can definitely relate to my epic passage at MIT. There are so many people I feel it is my duty to acknowledge, because I wouldn't be where I am today had I braved this experience alone. Each person I've crossed along the way, like Tennyson's Ulysses, *"I am part of all that I have met"*, has taught me invaluable lessons. For this, I'm deeply thankful.

I would like to start by thanking my thesis committee triumvirate: Professor Munther Dahleh, Professor Alexander "Sasha" Metgretski, and Professor Warren Seering, for the advice and support they liberally offered throughout my graduate studies at MIT. All of them have proven their personal and intellectual strength over the years, and their contribution to my growth goes beyond the academic facet of this journey. I can only hope to try and emulate the very best of them in my future endeavors.

Munther is a true clutch player and he taught me the meaning of being able to perform your very best under pressure, whether it be the few minutes before a paper deadline or when "bringing down the house" at the Palms in Las Vegas. He never loses his cool in any situation, and has the unique ability to make the seemingly impossible seem quite effortless. His remarkably deep intuition and his focus on the absolute essentials, which almost come instinctively to him, has always been a source of inspiration to me.

First thing which strikes you about Sasha is his analytical mind. He has the prodigious capacity of being able to correctly identify and decipher any problem with the slightest information available. There is no challenge he would shy away from, while always maintaining a firm yet polite demeanor. I always appreciated his generosity and self sarcasm, which I knew was meant to put others at ease - but

that didn't distract me from the fact that in most situations, while everyone else had reached their reasoning limits, Sasha had just begun his thought process.

Warren is the quintessential mentor. He wisely and naturally decides when to lead by example and when to exhibit patience, all the while being genuinely attentive to others. He can command control over every situation, but he smoothly ensures that people around him are comfortable. He understands what it takes to be successful in all aspects of life and communicates his knowledge in a witty, objective and unselfish manner.

Apart from the members of my thesis committee, I have benefited from interactions with Professors Vincent Blondel, Professor Sanjoy Mitter, and Professor Pablo Parillo during the last year of my thesis, and with Professor Jeff Shamma in 2002. Professor Dimitri Bertsekas inspired me with his art of teaching while I was a 6.041 TA.

Other members of the faculty that offered me advice at the early stages of my studies are Professors Lallit Anand, Professor Eric Feron, Professor David Miller, and Dr. Carl Blaurock.

Maria Monserrate and Leslie Regan handled every administrative issue with warmth and a smile.

I would also like to thank my Professors at ETH, Professor Mahir Behar Sayir for instilling a die-hard drive in me and teaching me to settle for nothing but the best in me, Professor Hans Peter Gehrung for inspiring by example, and Professor Christoph Wehrli for shifting my focus to academics.

I have been fortunate enough to have met some incredibly unique and talented people at MIT. Some I enjoyed playing sports with, some tagged along in the explorations of Boston and its suburbs, some made all-nighters spent in a cluster fun, and some just offered a few words that have added character and depth to my perspective. First, I would like to thank Constantinos Boussios and Nuno Martins, who could always relate to the process of doing a Ph.D. in system theory and control and its range of implications, and whom I have known since the start of my studies.

In purely alphabetical order big ups to Patrick Anquetil, Ola Ayaso, Erin Aylward,

Ioannis Bertsatos, Brian Bingham, Petros Boufounos, David Brown, Constantine Caramanis, Boris Cukalovic, Emilio Frazzoli, Bishwaroop Ganguly, Jorge Goncalvez, John Harper, Tom Harris, Shan-Yuan Ho, Tomoko Ida, Neal Jameson, Chung-Yao Kao, Mike Kim, Aleksandar Kojic, Caspar Andri Lariader, Carl Livadas, Marios Michalakis, Charles Maher, Vitaly Napadov, Reza Olfati-Saber, Peggy Paraoulaki, Georgios Papaioannou, Maria Andriana Papaioannou, Leon Patitsas, Brett Pellock, Dimitrios Rovas, Navid Sabbaghi, Sridevi Sarma, Tom Schouvenaars, Sandeep Sethuraman, Charalambos Soutatis, Zoltan Spakovszky, Deanelle Symonds, Kazutaka Takahashi, Prakash Thamburaja, Carolina Tortora, Chris Tsonis, Theodora Tzianetopoulou, Kripa Varanasi, Srikanth Vedantam, and Holly Waisanen.

Outside MIT I would like to thank my friends Michel Andenmatten, Triantafyllos Chavakis, Orestis Grigoropoulos, Ahmet Ege Karman, Orestis Servetopoulos, and Filippos Tsioros.

I would also like to thank my family for their unconditional love and support, my father Nikolaos, my mother Aggeliki, my brother Aggelos, my uncle Vasilis and my grandfather Nikolaos. Finally, I'd like to give a special thank you to my grandmother Maria for making sure that I was raised the right way.



# Contents

<b>1</b>	<b>Introduction</b>	<b>13</b>
<b>2</b>	<b>Reduction of IID Jump Linear Systems</b>	<b>17</b>
2.1	System Model . . . . .	17
2.1.1	Markov Jump Linear System . . . . .	17
2.1.2	I.I.D. Jump Linear System . . . . .	18
2.2	Stability Concepts . . . . .	18
2.3	Sensitivity Measure . . . . .	19
2.4	Error system . . . . .	22
2.5	Dissipation Inequalities . . . . .	23
2.6	Reduction by state truncation . . . . .	25
2.7	Upper bound to the approximation error . . . . .	26
2.8	Obtaining diagonal $W$ . . . . .	31
2.9	Lack of unique solutions to the dissipation inequalities . . . . .	32
2.10	A numerical example . . . . .	34
<b>3</b>	<b>Model Reduction of Hidden Markov Models</b>	<b>37</b>
3.1	Definitions . . . . .	38
3.1.1	Hidden Markov Models . . . . .	38
3.1.2	Generalized Automata . . . . .	40
3.2	Model Reduction Algorithm . . . . .	41
3.2.1	Problem Statements . . . . .	42
3.2.2	Reduction Algorithm . . . . .	42



3.3	A numerical example of the method . . . . .	46
<b>4</b>	<b>Reduction of Nearly Decomposable Hidden Markov Models</b>	<b>49</b>
4.1	Nearly Decomposable Hidden Markov Model . . . . .	49
4.2	State Aggregation . . . . .	50
4.3	Reduction Algorithm . . . . .	52

# List of Figures

2-1	Error System $\mathcal{E}$ . . . . .	23
2-2	Ratio $\frac{\lambda_{max}}{\lambda_{min}}$ of the eigenvalues of $W$ . . . . .	35
2-3	Response of $\mathcal{L}$ and $\hat{\mathcal{L}}$ to a step input . . . . .	35
3-1	Eigenvalues of $W$ , logarithmic scale on y axis . . . . .	47
3-2	Word function of error system between $\mathbf{H}$ and $\hat{\mathbf{G}}$ . . . . .	47
3-3	Word function of error system between $\mathbf{H}$ and $\hat{\mathbf{H}}$ . . . . .	48
4-1	Word function of error system between $\mathbf{H}$ and $\hat{\mathbf{H}}$ . . . . .	55

# List of Tables



# Chapter 1

## Introduction

The concept of model reduction is pervasive in all areas, where system theoretic ideas have been applied. The starting point is always some mathematical model, which exhibits high degree of complexity. The typical task is to replace the original model with a low complexity counterpart, while preserving all relevant information. What constitutes relevant information is reflected in appropriately defined error measures, which capture the fidelity of the reduced model. Model reduction algorithms are being evaluated by their ability to provide provable a priori guarantees for the degree of accuracy and the level of complexity reduction achieved by the low dimensional model, as well as the algorithmic cost associated with them.

The balanced truncation algorithm, which originated in [22] provides an example of a model reduction technique successfully employed in the context of Linear Time Invariant Systems. Theoretical justification for its use was given by the derivation of a priori bounds to the approximation error, [11]. The closely related optimal Hankel norm reduction problem was solved in [13]. Since its inception the balanced truncation algorithm has been extended to more general classes of systems. In particular multidimensional and uncertain systems, which are represented by means of Linear Fractional Transformations, are handled in [9], the case of Linear Time-Varying Systems is addressed in [14].

In this thesis a balanced truncation algorithm is presented for a class of hybrid systems that combine continuous and discrete variables, namely Jump Linear Systems, where

the parametric input varies over a finite set. Elements of that algorithm are used in a reduction method developed subsequently for discrete-time, finite state, finite alphabet Hidden Markov Models. It is remarkable that despite the widespread range of applications of Hidden Markov Models, systematic reduction methods, which do not pose any structural requirements, have been lacking so far. What follows is a brief overview of these two classes of dynamical systems.

Hidden Markov Models (abbreviated HMM's) are one of the most basic and widespread modeling tools for discrete-time stochastic processes that take values on a finite alphabet. One of the first references that make use of the concept of a HMM is [27], where HMM's with discrete inputs were considered as models for noisy, finite state communication channels. This class of models is commonly referred to as Probabilistic Automata [24]. Another early encounter with HMM's can be found in [6]. That work provided motivation for the subsequent investigation of the stochastic realization problem. Given a finite valued, stationary process  $\{Y(t)\}$ , find necessary and sufficient conditions for it to be equivalent, in the distributional sense, with a function of a Markov chain. Those conditions were derived in [15]. Finite state stochastic processes were associated to algebraic modules and it was shown that the stochastic realization problem is essentially a question of polyhedral convexity. The work in [15] was translated in more transparent system theoretic terms in [25], however both approaches are non constructive. A remedy to that was provided in [1], where the employment of convergence results on infinite products of nonnegative matrices led to a realization algorithm, which is based on asymptotic arguments and is semi-constructive in nature. In that work it was assumed at the outset that the given process has an HMM realization of unknown order. Conditions that are essentially necessary and sufficient and are stated in terms of the given process  $\{Y(t)\}$  alone, without making the a priori assumption that  $\{Y(t)\}$  has an HMM realization, were derived recently in [31]. It is worth mentioning that the minimal stochastic realization problem is open up to this date, in fact the simpler problem of minimal realization of Linear Time Invariant Positive Systems lacks a complete solution too. A review paper on this subject is [3].

Applications of HMM's are met across the spectrum of engineering and science in fields as diverse as speech processing, computational biology and financial econometrics, see for example [26], [17] and [5] respectively. Very often the cardinality of the state space of the underlying Markov chain renders the use of a given HMM for statistical inference or decision making purposes as infeasible, motivating the investigation of possible algorithms that compress the state space without incurring much loss of information. In [21] it was suggested that the concept of approximate lumpability can be used in the context of model reduction of HMM's. Results on the approximate factorization of nonnegative matrices were used in conjunction with approximate realization of HMM's in [12].

Jump Linear Systems (abbreviated JLS's) are abstractions of hybrid systems, which combine continuous and discrete dynamics. They form an extension of Linear Time Invariant Systems, in the sense that the coefficients are functions of parameters. In this work the case where the parameter space ranges over a finite set will be considered. The transition between the different modes of operation is controlled by an exogenous parametric input. In the process of defining a distance metric between two Jump Linear Systems it will be assumed that the parametric input is a sequence of independently identically distributed random variables. The resulting system will be referred to as an IID-JLS. There is a large body of literature in the fields of econometrics and system theory pertaining to the class of JLS's, with randomly varying parameters. One of the first references is [19], which investigated the continuous-time Jump Linear Quadratic Control problem. Further results on this problem were obtained in [29] and [32]. In particular, [29] considered the finite horizon setting and a stochastic maximum principle approach is used, whereas in [32] dynamic programming is employed and the infinite horizon case is treated as well. In many applications it is reasonable to assume that the switching sequence forms a finite state Markov chain and one speaks of Markov Jump Linear System (abbreviated MJLS). In the last two decades a large collection of Linear System Theory results have been extended to this class of systems, see for example [23] and the references therein.





# Chapter 2

## Reduction of IID Jump Linear Systems

In this chapter a model reduction method for IID-JLS's will be presented. The complexity measure used is the order of the state space realization. The only assumption posed on the original system is, that it is mean square stable. The main point of the method is the formulation of two generalized dissipation inequalities, which in conjunction with a suitably defined storage function enable the derivation of reduced order models that come with a provable a priori upper bound on the stochastic  $L_2$  gain of the approximation error.

### 2.1 System Model

#### 2.1.1 Markov Jump Linear System

Consider the following stochastic system denoted by  $\mathcal{L}$  :

$$\begin{aligned}x(t+1) &= A[\theta(t)]x(t) + B[\theta(t)]f(t), \\y(t) &= C[\theta(t)]x(t), \quad t \in \mathbf{Z}_+.\end{aligned}$$

As usual  $x(t) \in \mathbf{R}^n$  is the state variable,  $f(t) \in \mathbf{R}^m$  is the control input and  $y(t) \in \mathbf{R}^p$  is the output variable, where  $n, m, p$  are positive integers. The system matrices are of conformable dimensions. The parametric input  $\theta(t)$  controls the modal transitions of the system and corresponds to the state of a Markov chain defined on a finite set  $\Theta = \{1, \dots, N\}, N \in \mathbf{Z}_+, N \geq 1$ . The associated transition probability matrix is denoted by  $P$ , and it is a row stochastic matrix. Let  $p_{ij}$  denote the entry in the  $i$ 'th row and  $j$ 'th column of  $P$ , then  $p_{ij} = \mathbf{P}[\theta(t+1) = j \mid \theta(t) = i]$ . The initial conditions of the system are given by specifying  $x(0) \in \mathbf{R}^n$  and  $\theta(0) \in \Theta$ . The system  $\mathcal{L}$  is what is called a discrete-time MJLS.

### 2.1.2 I.I.D. Jump Linear System

An IID-JLS has the same state space representation as a MJLS. The only difference lies in the transition probability matrix. In the case of an IID-JLS all the rows are equal, i.e.  $p_{ij} = p_j, \forall i, j \in \Theta$ .

## 2.2 Stability Concepts

There are several concepts of stability associated with stochastic systems, see [20] and [18].

**Definition 1.** *The system  $\mathcal{L}$  with  $f(t) = 0, \forall t \in \mathbf{Z}_+$  is called **mean square stable**, if for every initial condition  $x(0) \in \mathbf{R}^n, \theta(0) \in \Theta$ ,*

$$\mathbf{E}[|x(t)|^2] \rightarrow 0 \text{ as } t \rightarrow \infty.$$

**Definition 2.** *The system  $\mathcal{L}$  with  $f(t) = 0, \forall t \in \mathbf{Z}_+$  is called **exponentially mean square stable** if for every initial condition  $x(0) \in \mathbf{R}^n, \theta(0) \in \Theta$ , there exist real constants  $\beta > 0$  and  $\rho \in (0, 1)$  such that*

$$\mathbf{E}[|x(t)|^2] \leq \beta \rho^k |x(0)|^2, t \in \mathbf{Z}_+.$$

**Definition 3.** The system  $\mathcal{L}$  with  $f(t) = 0, \forall(t) \in \mathbf{Z}_+$  is called **stochastically stable** if for every initial condition  $x(0) \in \mathbf{R}^n, \theta(0) \in \Theta$ ,

$$\sum_{t=0}^{\infty} \mathbf{E}[|x(t)|^2] < \infty.$$

Similar to the case of Linear Time Invariant Systems, the above stability concepts are equivalent.

**Theorem 4.** [23] The following statements are equivalent :

- (a) System  $\mathcal{L}$  is mean square stable.
- (b) System  $\mathcal{L}$  is exponentially mean square stable.
- (c) System  $\mathcal{L}$  is stochastically stable.
- (d) There exists a set of  $N$  positive definite matrices  $\{G(i) > 0, G(i) \in \mathbf{R}^{n \times n}\}_{i \in \Theta}$ , such that:

$$G(i) - \sum_{j=1}^N p_{ij} A(i)' G(j) A(i) > 0, \quad i \in \Theta.$$

*Proof.* A proof of the above theorem can be found in [23]. □

In the case of iid-JLS's the set of coupled Lyapunov equations simplifies to the single equation

$$G - \sum_{j=1}^N p_j A(j)' G A(j) > 0.$$

In the following only iid-JLS's will be considered.

## 2.3 Sensitivity Measure

**Definition 5.** The **stochastic  $L_2$  gain** of the system  $\mathcal{L}$  is denoted by  $\gamma_{\mathcal{L}}$  and is defined for  $x(0) = 0$  by

$$\gamma_{\mathcal{L}}^2 = \sup_{f \in \mathbf{S}_m^2} \mathbf{E}\left[\sum_{t=0}^{\infty} |y(t)|^2\right]$$

In the above definition the set  $\mathbf{S}_m^2$  corresponds to the unit sphere in the space of square summable sequences of  $m$ -dimensional vectors.

**Theorem 6.** *Given a system  $\mathcal{L}$ , if there exists a quadratic function  $V : \mathbf{R}^n \rightarrow [0, \infty)$  with  $V(0) = 0$  satisfying :*

$$\gamma^2 |f(t)|^2 + V(x(t)) \geq \sum_{i=1}^N p_i [|C(i)x(t)|^2 + V(A(i)x(t) + B(i)f(t))], \quad (2.1)$$

$$\forall x(t) \in \mathbf{R}^n, \forall f(t) \in \mathbf{R}^m$$

then the stochastic  $L_2$  gain of  $\mathcal{L}$  does not exceed  $\gamma \geq 0$ .

*Proof.* The above relation implies

$$\gamma^2 |f(t)|^2 + \mathbf{E}[V(x(t))] \geq \mathbf{E}[|y(t)|^2] + \mathbf{E}[V(x(t+1))], \quad (2.2)$$

$$\forall f(t) \in \mathbf{R}^m, \forall t \in \mathbf{Z}_+.$$

According to the definition set  $x(0) = 0$  and sum relation (2.2) from  $t = 0$  to  $t = T$  obtaining

$$\mathbf{E}\left[\sum_{t=0}^T |y(t)|^2\right] \leq \gamma^2 \sum_{t=0}^T |f(t)|^2 - \mathbf{E}[V(x(T+1))].$$

Since  $V$  is a nonnegative valued map,  $\mathbf{E}[V(x(T+1))] \geq 0$  thus

$$\mathbf{E}\left[\sum_{t=0}^T |y(t)|^2\right] \leq \gamma^2 \sum_{t=0}^T |f(t)|^2.$$

Restricting the input signal  $f$  to be on the unit sphere  $\mathbf{S}_m^2$  gives

$$\mathbf{E}\left[\sum_{t=0}^{\infty} |y(t)|^2\right] \leq \gamma^2 \forall f \in \mathbf{S}_m^2$$

and in particular  $\gamma_{\mathcal{L}}^2 \leq \gamma^2$  completing the proof. □

The search for a quadratic storage function, that leads to finiteness of the stochastic  $L_2$  gain is guaranteed to succeed if the system is stochastically stable, in the sense defined above.

**Theorem 7.** *If the system  $\mathcal{L}$  is mean square stable, then its stochastic  $L_2$  gain is finite.*

*Proof.* Let  $Q > 0$  be an arbitrary positive definite matrix. Mean square stability guarantees the existence of a positive definite matrix  $P > 0$ , such that

$$\sum_{i=1}^N p_i A(i)' P A(i) - P = -Q < 0. \quad (2.3)$$

Define  $V(x(t)) = x(t)' \alpha P x(t)$  to be a quadratic function of the state, where  $P > 0$  and  $\alpha \geq 1$ . Using the state equations one obtains the following relation, that is equivalent to condition (2.1)

$$\begin{bmatrix} x(t)' & f(t)' \end{bmatrix} \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} \begin{bmatrix} x(t) \\ f(t) \end{bmatrix} \leq 0 \quad \forall x(t) \in \mathbf{R}^n, f(t) \in \mathbf{R}^m \quad (2.4)$$

where

$$\begin{aligned} W_{11} &= \sum_{i=1}^N p_i (A(i)' \alpha P A(i) + C(i)' C(i)) - \alpha P \\ W_{12} &= \sum_{i=1}^N p_i A(i)' \alpha P B(i) \\ W_{21} &= Q'_{12} \\ W_{22} &= \sum_{i=1}^N p_i B(i)' \alpha P B(i) - \gamma^2 I \end{aligned}$$

Using the Schur complement idea one can conclude, that a sufficient set of conditions for (2.4) to hold is

$$W_{11} < 0 \quad (2.5)$$

$$W_{22} < W_{21} W_{11}^{-1} W_{12} \quad (2.6)$$

Using (2.3), relation (2.5) can be rewritten as

$$\sum_{i=1}^N p_i C(i)' C(i) - \alpha Q < 0$$

and there is always an  $\alpha$  large enough so that it is satisfied. Setting

$$\begin{aligned} F_1 &= \sum_{i=1}^N p_i B(i)' \alpha P B(i) \\ F_2 &= W_{21} W_{11}^{-1} W_{12} \end{aligned}$$

one can rewrite (2.6) as

$$F_1 - F_2 < \gamma^2 I.$$

The above condition can always be satisfied by taking  $\gamma$  large enough. Thus, there exists an  $\alpha \geq 1$  and a  $\gamma > 0$  such that  $V(x(t)) = x(t)' \alpha P x(t)$  satisfies the dissipation inequality (2.1), leading to finiteness of the stochastic  $L_2$  gain of  $\mathcal{L}$ .  $\square$

A standing assumption in this work is that the give iid-JLS is mean square stable system.

## 2.4 Error system

Reduced order model candidates are denoted by  $\hat{\mathcal{L}}$ . It is required that the reduced model has the same JLS structure and that parametric input ranges over the same set  $\Theta$ . The state space equations for the reduced model are given by :

$$\begin{aligned} \hat{x}(t+1) &= \hat{A}[\theta(t)] \hat{x}(t) + \hat{B}[\theta(t)] f(t), \\ \hat{y}(t) &= \hat{C}[\theta(t)] \hat{x}(t), \quad t \in \mathbf{Z}_+, \end{aligned}$$

where  $\hat{x}(t) \in \mathbf{R}^{\hat{n}}$  and  $\hat{n} < n$ .

In order to quantify the fidelity of  $\hat{\mathcal{L}}$ , an error system  $\mathcal{E}$  is introduced, whose inputs are the common inputs  $f(t)$ ,  $\theta(t)$  of  $\mathcal{L}$  and  $\hat{\mathcal{L}}$  and whose output is the difference of

their outputs, namely  $e(t) = y(t) - \hat{y}(t)$ .

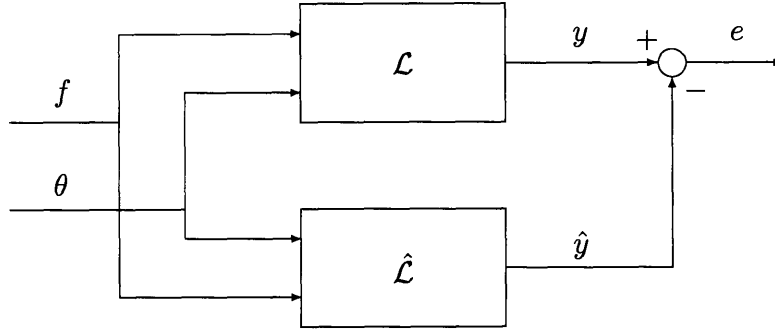


Figure 2-1: Error System  $\mathcal{E}$

The goal of the reduction process is to produce models of lower complexity  $\hat{n} \ll n$ . that, loosely speaking, satisfy  $\gamma_{\mathcal{E}} \leq \beta$ , where  $\beta$  is a “small”, a priori computable, real number. What constitutes a reasonable choice for  $\beta$  depends on how stringent are the performance requirements for a given system.

## 2.5 Dissipation Inequalities

The reduction method relies on the computation of  $P > 0$ ,  $\hat{Q} > 0$  for a given mean square stable system  $\mathcal{L}$  such that the following set of dissipation inequalities is satisfied:

$$|x(t)|_P^2 \geq \sum_{i=1}^N p_i (|A(i)x|_P^2 + |C(i)x|^2), \quad (2.7)$$

$$\forall x(t) \in \mathbf{R}^n,$$

$$|x(t)|_{\hat{Q}}^2 + |f(t)|^2 \geq \sum_{i=1}^N p_i (|A(i)x(t) + B(i)f(t)|_{\hat{Q}}^2), \quad (2.8)$$

$$\forall x(t) \in \mathbf{R}^n, \forall f(t) \in \mathbf{R}^m$$

In the above relations the notation  $|z(t)|_P^2 = z(t)'Pz(t)$  is used. There is a natural interpretation of (2.7), (2.8) in the case where  $N = 1$ , so that  $\mathcal{L}$  reduces to an LTI system. If the system matrices  $\{A(1), B(1), C(1)\}$  constitute a minimal realization of

$\mathcal{L}$ , then equation (2.7) is satisfied with equality using  $P = W_o$ , and (2.8) is satisfied with equality using  $\hat{Q} = W_c^{-1}$ , where  $W_o, W_c$  are the observability and controllability Gramians of the system respectively. They satisfy the Lyapunov equations

$$\begin{aligned} W_o &= A(1)'W_oA(1) + C(1)'C(1) \\ W_c &= A(1)W_cA(1)' + B(1)B(1)' \end{aligned}$$

In the case where  $N > 1$ , the following two lemmas provide interpretations for  $P$  and  $\hat{Q}$ .

**Lemma 8.** *Let  $T \in \mathbf{Z}_+$  and consider the unforced  $\{f(0), \dots, f(T)\} = \{0, \dots, 0\}$  response of  $\mathcal{L}$  to the initial condition  $x(0) \in \mathbf{R}^n$ . For an arbitrary  $T_0 \in \mathbf{Z}_+$ , such that  $T_0 < T$  one has*

$$\sum_{t=T_0}^T \mathbf{E}[|y(t)|^2] \leq \mathbf{E}[|x(T_0)|_P^2].$$

*Proof.* The dissipation inequality (2.7) implies in the unforced case

$$\mathbf{E}[|x(t+1)|_P^2] + \mathbf{E}[|y(t)|^2] \leq \mathbf{E}[|x(t)|_P^2],$$

Sum the above relation from  $t = T_0$  to  $t = T$  to obtain

$$\mathbf{E}[|x(T+1)|_P^2] + \sum_{k=T_0}^T \mathbf{E}[|y(k)|^2] \leq \mathbf{E}[|x(T_0)|_P^2].$$

Then, noticing that  $\mathbf{E}[|x(T+1)|_P^2] \geq 0$  leads to the desired result.  $\square$

**Lemma 9.** *Let  $T \in \mathbf{Z}_+$  and consider the evolution of  $\mathcal{L}$  that starts at rest  $x(0) = 0$ . Then, for an arbitrary input sequence  $\{f(0), \dots, f(T)\}$  one has*

$$\sum_{t=0}^T |f(t)|^2 \geq \mathbf{E}[|x(T+1)|_{\hat{Q}}^2], \quad \forall f(t) \in \mathbf{R}^m, t \in \{1 \dots T\}$$

*Proof.* The dissipation inequality (2.8) gives in this case

$$\mathbf{E}[|x(t+1)|_{\hat{Q}}^2] \leq \mathbf{E}[|x(t)|_{\hat{Q}}^2] + |f(t)|^2, \quad \forall f(t) \in \mathbf{R}^m.$$



Sum the above relation from  $t = 0$  to  $t = T$  and note that  $x(0) = 0$  to obtain the desired result.  $\square$

## 2.6 Reduction by state truncation

What follows now is a straightforward extension of the concept of state truncation, well known for LTI systems, to the JLS's case. One starts out with a state space model of  $\mathcal{L}$

$$\begin{aligned} x(t+1) &= A[\theta(t)]x(t) + B[\theta(t)]f(t), \\ y(t) &= C[\theta(t)]x(t), \quad t \in \mathbf{Z}_+, \end{aligned} \tag{2.9}$$

and applies an invertible coordinate transformation

$$x(t) = T\tilde{x}(t)$$

that puts the "most important" states in the first components of the transformed state vector  $\tilde{x}(t)$ . This transformation gives a new state space representation of  $\mathcal{L}$

$$\begin{aligned} \tilde{x}(t+1) &= \tilde{A}[\theta(t)]\tilde{x}(t) + \tilde{B}[\theta(t)]f(t), \\ y(t) &= \tilde{C}[\theta(t)]\tilde{x}(t), \quad t \in \mathbf{Z}_+. \end{aligned}$$

The state vector  $\tilde{x}(t)$  is then partitioned as

$$\tilde{x}(t) = \begin{bmatrix} \tilde{x}_1(t) \\ \tilde{x}_2(t) \end{bmatrix},$$

where the state vector  $\tilde{x}_1(t)$  corresponds to the states that are to be retained and  $\tilde{x}_2(t)$  to the states that are to be removed. With appropriate partitioning of the system

matrices the state space representation of  $\mathcal{L}$  becomes

$$\begin{aligned}\tilde{x}_1(t+1) &= \tilde{A}_{11}[\theta(t)]\tilde{x}_1(t) + \tilde{A}_{12}[\theta(t)]\tilde{x}_2(t) + \tilde{B}_1[\theta(t)]f(t), \\ \tilde{x}_2(t+1) &= \tilde{A}_{21}[\theta(t)]\tilde{x}_1(t) + \tilde{A}_{22}[\theta(t)]\tilde{x}_2(t) + \tilde{B}_2[\theta(t)]f(t), \\ y(t) &= \tilde{C}_1[\theta(t)]\tilde{x}_1(t) + \tilde{C}_2[\theta(t)]\tilde{x}_2(t), \quad t \in \mathbf{Z}_+.\end{aligned}$$

The dynamic system that one obtains by truncating the last  $r$  variables, i.e.  $\tilde{x}_2(t) \in \mathbf{R}^r$ , is equivalent to a system whose state variables are constrained in a proper subspace  $S_{n-r}$  of the original state space, where  $S_{n-r} = \{z \in \mathbf{R}^n \mid z(i) = 0, \quad n-r+1 \leq i \leq n\}$ , that is naturally isomorphic to  $\mathbf{R}^{n-r}$ . Thus the state vector  $\hat{x}(t)$  of the reduced system  $\hat{\mathcal{L}}$  will be of the form  $\hat{x}(T) = (\tilde{x}_1(t), 0)' \in S_{n-r} \subset \mathbf{R}^n$ .

## 2.7 Upper bound to the approximation error

In this section it will be shown, how to reduce the order of a given mean square stable system  $\mathcal{L}$  by means of state truncation and obtain an upper bound on the stochastic  $L_2$  gain of the resulting error system  $\mathcal{E}$ .

**Theorem 10.** *Consider a mean square stable system  $\mathcal{L}$  of order  $n$ . Consider also the positive definite matrix  $W$ , such that*

$$W = \Sigma_1 \oplus \Sigma_2,$$

where

$$\Sigma_2 = \beta_1 I_{r_1} \oplus \dots \oplus \beta_s I_{r_s}, \quad \sum_{k=1}^s r_k = r.$$

Suppose that the matrix  $P = W$  satisfies (2.7) and  $\hat{Q} = W^{-1}$  satisfies (2.8). Let  $\tilde{\mathcal{L}}$  be the reduced order model obtained by truncating the last  $r$  states of  $\mathcal{L}$ . Then, the stochastic  $L_2$  gain of the error system  $\mathcal{E}$  is bounded from above by twice the sum of the distinct entries on the diagonal of  $\Sigma_2$  :

$$\gamma_{\mathcal{E}} \leq 2(\beta_1 + \dots + \beta_s) \tag{2.10}$$

*Proof.* Using the matrix

$$E_r = \begin{bmatrix} 0 & 0 \\ 0 & I_r \end{bmatrix}$$

the state space model of  $\hat{\mathcal{L}}$  can be written as

$$\begin{aligned} \hat{x}(t+1) &= (I_n - E_r)(A[\theta(t)]\hat{x}(t) + B[\theta(t)]f(t)), \\ \hat{y}(t) &= C[\theta(t)]\hat{x}(t), \quad t \in \mathbf{Z}_+. \end{aligned} \quad (2.11)$$

The following signals will shorten the subsequent notation.

$$\begin{aligned} z(t) &= x(t) + \hat{x}(t), \\ \delta(t) &= x(t) - \hat{x}(t) \\ h[\theta(t)] &= A[\theta(t)]\hat{x}(t) + B[\theta(t)]f(t), \quad \theta(t) \in \Theta. \end{aligned}$$

The proof will proceed by successive truncation of the last  $r_s, r_{s-1}, \dots, r_1$  states. Let  $\mathcal{L}_s$  denote the reduced system obtained by truncating the last  $r_s$  states and  $\mathcal{E}_s$  the corresponding error system between  $\mathcal{L}_s$  and  $\mathcal{L}$ . The state variable of  $\mathcal{L}_s$  is  $\hat{x}(t)^{(s)} \in S_{n-r_s} \subset \mathbf{R}^n$  and one can verify that the following relations hold:

$$\begin{aligned} z(t+1)^{(s)} &= A[\theta(t)]z(t)^{(s)} + 2B[\theta(t)]f(t) - E_{r_s}h[\theta(t)]^{(s)}, \\ \delta(t+1)^{(s)} &= A[\theta(t)]\delta(t)^{(s)} + E_{r_s}h[\theta(t)]^{(s)}, \\ e(t)^{(s)} &= C[\theta(t)]\delta(t)^{(s)}, \quad t \in \mathbf{Z}_+, \end{aligned}$$

where

$$\begin{aligned} z(t)^{(s)} &= x(t) + \hat{x}(t)^{(s)}, \\ \delta(t)^{(s)} &= x(t) - \hat{x}(t)^{(s)}, \\ e(t)^{(s)} &= y(t) - y(t)^{(s)}. \end{aligned}$$

In a first step it will be shown that

$$\gamma_{\varepsilon_s} \leq 2\beta_s \quad (2.12)$$

In order to prove (2.12) it is sufficient to find a storage function  $V : \mathbf{R}^n \times \mathbf{R}^n \rightarrow \mathbf{R}_+$ , such that  $V(0, 0) = 0$  and :

$$\begin{aligned} \Psi(x(t), \hat{x}(t)^{(s)}, f(t)) &\geq 0, \\ \forall x(t) \in \mathbf{R}^n, \forall \hat{x}(t)^{(s)} \in S_{n-r_s}, \forall f(t) \in \mathbf{R}^m, \end{aligned} \quad (2.13)$$

where

$$\begin{aligned} \Psi(x(t), \hat{x}(t)^{(s)}, f(t)) &= 4\beta_s^2 |f(t)|^2 - \sum_{i=1}^N p_i |C(i) \delta(t)^{(s)}|^2 - \Delta V, \\ \delta(t)^{(s)} &= x(t) - \hat{x}(t)^{(s)}, \\ \Delta V &= \sum_{i=1}^N p_i V(x(t+1), \hat{x}(t+1)_+^{(s)}) - V(x(t), \hat{x}(t)^{(s)}) \\ x(t+1) &= A(i)x(t) + B(i)f(t) \\ \hat{x}(t)_+^{(s)} &= (I_n - E_{r_s})(A(i)\hat{x}(t)^{(s)} + B(i)f(t)) \end{aligned}$$

Note that the above set of relations essentially imply

$$\begin{aligned} 0 &\leq 4\beta_s^2 |f(t)|^2 + \\ &- \mathbf{E}[|E(t)^{(s)}|^2 + V(X(t+1), \hat{X}(t+1)^{(s)}) - V(x(t), \hat{X}(t)^{(s)})], \\ &\forall f(t) \in \mathbf{R}^m \end{aligned}$$

and thus (2.12). A quadratic storage function candidate is given by :

$$V(x(t), \hat{x}(t)^{(s)}) = \beta_s^2 |z(t)^{(s)}|_{W^{-1}}^2 + |\delta(t)^{(s)}|_W^2$$

In order to verify (2.13) one needs to compute the expected increment of the storage

function along system trajectories.

$$\begin{aligned}\Delta V &= \sum_{i=1}^N p_i |A(i)\delta(t)^{(s)} + E_{r_s} h(i)^{(s)}|_W^2 + \\ &\quad \beta_s^2 \sum_{i=1}^N p_i |A(i)z(t)^{(s)} + 2B(i)f(t) - E_{r_s} h(i)^{(s)}|_{W-1}^2 + \\ &\quad -\beta_s^2 |z(t)^{(s)}|_{W-1}^2 - |\delta(t)^{(s)}|_W^2.\end{aligned}$$

Expanding the individual term in the above expressions, one obtains

$$\begin{aligned}\Delta V &= \sum_{i=1}^N p_i |A(i)\delta(t)^{(s)}|_W^2 - |\delta(t)^{(s)}|_W^2 + \tag{2.14} \\ &\quad + \beta_s^2 \sum_{i=1}^N p_i |A(i)z(t)^{(s)} + 2B(i)f(t)|_{W-1}^2 - \beta_s^2 |z(t)^{(s)}|_{W-1}^2 \\ &\quad + 2\beta_s \sum_{i=1}^N p_i |E_{r_s} h(i)^{(s)}|^2 \\ &\quad - 2\beta_s \sum_{i=1}^N p_i (E_{r_s} h(i)^{(s)})' (A(i)z(t)^{(s)} + 2B(i)f(t) - A(i)\delta(t)^{(s)}).\end{aligned}$$

Applying the dissipation inequality (2.7) on the first two terms of (2.14) gives

$$\sum_{i=1}^N p_i |A(i)\delta(t)^{(s)}|_W^2 - |\delta(t)^{(s)}|_W^2 \leq - \sum_{i=1}^N p_i |C(i)\delta(t)^{(s)}|^2.$$

Using the dissipation inequality (2.8), the second line in (2.14) becomes

$$\beta_s^2 \sum_{i=1}^N p_i |A(i)z(t)^{(s)} + 2B_i f(t)|_{W-1}^2 - \beta_s^2 |z(t)^{(s)}|_{W-1}^2 \leq 4\beta_s^2 |f(t)|^2.$$

For the last term of (2.14) note that

$$A(i)z(t)^{(s)} + 2B_i f(t) - A(i)\delta(t)^{(s)} = 2h(i)^{(s)},$$

and that  $E_{r_s}^2 = E_{r_s}$ . Using the above relations we obtain

$$\begin{aligned} \Delta V \leq & - \sum_{i=1}^N p_i |C(i)\delta(t)^{(s)}|^2 + 4\beta_s^2 |f(t)|^2 - \\ & 2\beta_s \sum_{i=1}^N p_i |E_{r_s} h(i)^{(s)}|^2. \end{aligned}$$

Substitute the above inequality in (2.14) to obtain

$$\begin{aligned} \Psi_k(x(t), \hat{x}(t)^{(s)}, f(t)) & \geq 2\beta_s \sum_{i=1}^N p_i |E_{r_s} h(i)^{(s)}|^2 \geq 0, \\ \forall \hat{x}(t)^{(s)} & \in S_{n-r_s}, \forall f \in \mathbf{R}^m, \end{aligned}$$

completing the first part of the proof. Let  $W_s$  be a submatrix of  $W$  corresponding to the retained states.

$$W_s = \Sigma_1 \oplus \beta_1 I_{r_1} \oplus \dots \oplus \beta_s I_{r_{s-1}}.$$

Note that  $W_s$  satisfies the generalized dissipation inequalities corresponding to  $\mathcal{L}_s$ , in the sense

$$\begin{aligned} \sum_{i=1}^N p_i (|A(i)\hat{x}(t)^{(s)}|_{W_s}^2 + |C(i)\hat{x}(t)^{(s)}|^2) & \leq |\hat{x}(t)^{(s)}|_{W_s}^2, \\ \forall \hat{x}(t)^{(s)} & \in S_{n-r_s}, \\ \sum_{i=1}^N p_i (|A(i)\hat{x}(t)^{(s)} + B(i)f|_{W_s^{-1}}^2) & \leq |\hat{x}(t)^{(s)}|_{W_s^{-1}}^2 + |f(t)|^2, \\ \forall \hat{x}(t)^{(s)} & \in S_{n-r_s}, \forall f \in \mathbf{R}^m. \end{aligned}$$

Thus, if the last  $r_{s-1}$  states from  $\mathcal{G}_s$  are truncated and if one denotes the resulting system  $\mathcal{G}_{s-1}$  and the corresponding error system between  $\mathcal{G}_s$ ,  $\mathcal{G}_{s-1}$  by  $\mathcal{E}_{s-1}$  then by repeating the above argument

$$\gamma_{\mathcal{E}_{s-1}} \leq 2\beta_{s-1}$$

Similarly,

$$\gamma_{\mathcal{E}_j} \leq 2\beta_j \quad j \in \{s, s-1, \dots, 1\}.$$

The desired result (2.10) is obtained by observing that

$e(t) = e(t)^{(1)} + \dots + e(t)^{(s)}$  and applying the triangle inequality on stochastic  $L_2$  gains.  $\square$

## 2.8 Obtaining diagonal $W$

The previous theorem assumes, that there exists a  $W = \Sigma_1 \oplus \Sigma_2$ ,  $\Sigma_2$  diagonal, such that  $W = P$  satisfies (2.7) and  $\hat{Q} = W^{-1}$  satisfies (2.8). In this section it will be shown that under the standing assumption of mean square stability, one can obtain in fact a diagonal matrix  $W$  with the desired properties. Mean square stability is equivalent with the existence of  $\hat{P} > 0$ , such that

$$\sum_{i=1}^N p_i A(i)' \hat{P} A(i) - \hat{P} < 0. \quad (2.15)$$

Relation (2.7) is equivalent to

$$\sum_{i=1}^N p_i A(i)' P A(i) - P \leq - \sum_{i=1}^N p_i C(i)' C(i). \quad (2.16)$$

By virtue of the above two relations, if one sets  $P = \alpha \hat{P}$  and takes  $\alpha > 0$  large enough, the dissipation inequality (2.7) can always be satisfied by some positive definite matrix  $P$ . Relation (2.8) is equivalent to

$$\begin{bmatrix} -\hat{Q} + \sum_{i=1}^N p_i A(i)' \hat{Q} A(i) & \sum_{i=1}^N p_i A(i)' \hat{Q} B(i) \\ \sum_{i=1}^N p_i B(i)' \hat{Q} A(i) & -I + \sum_{i=1}^N p_i B(i)' \hat{Q} B(i) \end{bmatrix} \leq 0 \quad (2.17)$$

Note that, if one sets  $\gamma = 1$ ,  $C(i) = 0, \forall i \in \{1, \dots, N\}$ , and  $\hat{Q} = \alpha P$  with  $\alpha > 0$ , the above relation becomes equivalent to (2.4). Let  $\hat{P}$  satisfy (2.15) and set  $\hat{Q} = \alpha \hat{P}$  with  $\alpha > 0$ . Condition (2.5) is equivalent to mean square stability and thus feasible

for all positive values of  $\alpha$ . Condition (2.6) can be rewritten as

$$\sum_{i=1}^N p_i B(i)' \hat{Q} B(i) - W_{21} W_{11}^{-1} W_{12} < I.$$

Both terms on the left hand side of the above relation scale linearly with  $\alpha$ . Thus, by taking the positive parameter  $\alpha$  to be small enough one can also satisfy (2.8) with the choice of  $\hat{Q} = \alpha \hat{P}$ . To this point one has obtained  $P > 0$  and  $\hat{Q} > 0$  such that (2.7) and (2.8) are feasible. What remains then, is to compute a transformation matrix  $T$  that diagonalizes the product  $P\hat{Q}^{-1}$ . In that case  $TP\hat{Q}^{-1}T^{-1} = W^2 > 0$  and (2.7) is satisfied by  $W$  and (2.8) by  $W^{-1}$ , justifying the assumption of the previous theorem in regards to  $W$ .

## 2.9 Lack of unique solutions to the dissipation inequalities

In general LMI's may have multiple solutions, and thus there is no unique solution to (2.7) and (2.8). However the dissipation inequality (2.7) and its equivalent form (2.16) possess a unique minimal solution, which can be computed by solving the linear algebraic equation :

$$\sum_{i=1}^N p_i (A(i)' P A(i) + C(i)' C(i)) = P$$

When it comes to relation (2.8) or its equivalent form (2.17) the situation is different. For  $N = 1$ , the inverse of the controllability grammian corresponds to a maximal solution of (2.17). In the case of a JLS, where  $N > 1$ , there is no maximal solution though. For example, let  $N = 2$  and

$$\hat{Q}_i = \arg \max_{\hat{Q} > 0} \text{trace}(R(i)\hat{Q}), \quad i \in \{1, 2\} \quad (2.18)$$



subject to (2.17), where  $R(i) \geq 0, i \in \{1, 2\}$ . If there was a maximal solution to (2.17), then one should have

$$\hat{Q}_1 = \hat{Q}_2 \tag{2.19}$$

The following system shows that (2.19) is not satisfied. Let

$$\begin{aligned} R(1) &= \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, & A(1) &= \begin{bmatrix} 0.2 & 0.0 \\ 0.3 & 0.5 \end{bmatrix}, \\ R(2) &= \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, & A(2) &= \begin{bmatrix} 0.3 & 0.3 \\ 0.2 & 0.2 \end{bmatrix}, \\ B(1) &= B(2) = \begin{bmatrix} 0.4 \\ 0.2 \end{bmatrix}, & q_1 &= q_2 = \frac{1}{2}. \end{aligned}$$

Solving the optimization problem (2.18) subject to (2.17) gives

$$\hat{Q}(1) = \begin{bmatrix} 17.4 & -16.7 \\ -16.7 & 21.3 \end{bmatrix}, \quad \hat{Q}(2) = \begin{bmatrix} 14.9 & -16.0 \\ -16.0 & 24.9 \end{bmatrix}.$$

The lack of a maximal solution to (2.17), is to some extent unfortunate, since the diagonal entries of  $\Sigma_2$  that appear in (2.10) are monotonic in  $P$  and  $\hat{Q}^{-1}$ . A reasonable remedy is to compute a positive definite matrix  $\hat{Q}$  such that  $\text{trace}(P^{-1} \hat{Q})$  is maximized subject to the constraint (2.17). The motivation for this objective function comes from the fact that

$$\text{trace}(P^{-1} \hat{Q}) = \sum_{i=1}^N \frac{1}{\beta_i^2}$$

and thus the smaller eigenvalues of  $W$  are more heavily penalized in this optimization criterion, which is desirable given the nature of the error bound (2.10).

## 2.10 A numerical example

The reduction method will be demonstrated on a simple example that involves a system  $\mathcal{L}$  with 2 modes, 2 states, 1 input and 1 output having the system matrices :

$$A(1) = \begin{bmatrix} \beta + \alpha & 0 \\ 0 & \beta - \alpha \end{bmatrix}, \quad A(2) = \begin{bmatrix} \beta - \alpha & 0 \\ 0 & \beta + \alpha \end{bmatrix}$$

where  $\beta$  and  $\alpha$  are positive parameters.

$$B(1) = B(2) = B = \begin{bmatrix} 1 \\ 1 \end{bmatrix} = C(1)' = C(2)' = C'.$$

Note that the above system is worst case stable if and only if

$$\beta + \alpha < 1.$$

The parametric input is randomized by setting  $q_1 = q_2 = \frac{1}{2}$ . System  $\mathcal{L}$  is mean square stable if and only if

$$\beta^2 + \alpha^2 < 1.$$

As expected the requirement of stochastic stability relaxes the constraints on the parameters  $\alpha, \beta$ . Let

$$W = \begin{bmatrix} \lambda_{max} & 0 \\ 0 & \lambda_{min} \end{bmatrix},$$

and set  $\beta = 0.7$ . The ratio  $\frac{\lambda_{max}}{\lambda_{min}}$  is depicted in the following figure as a function of  $\alpha$ .

Given the nature of the error bound, one can expect, that the larger the eigenvalue ratio of  $W$  the better the quality of the reduction. Note that as  $\alpha$  converges to 0,  $\mathcal{L}$  converges to a first order linear time-invariant system. Truncating one state from  $\mathcal{L}$  leads to a reduced system  $\hat{\mathcal{L}}$ , that turns out to be a linear time invariant system with a single pole at  $\beta$ . The response of the two systems to a step input for a particular realization of the parametric input is depicted in the following figure for  $\beta = 0.7$  and

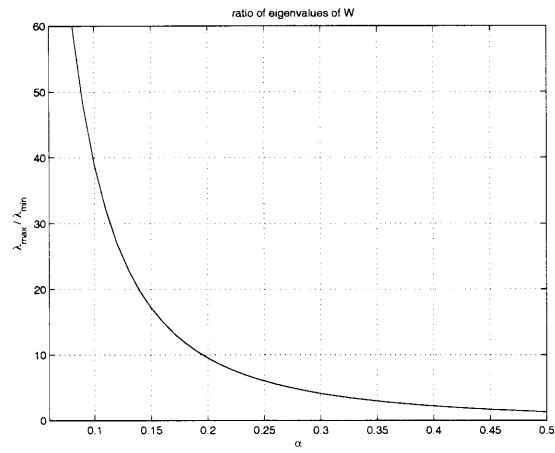


Figure 2-2: Ratio  $\frac{\lambda_{max}}{\lambda_{min}}$  of the eigenvalues of  $W$ .

$\alpha = 0.1$ .

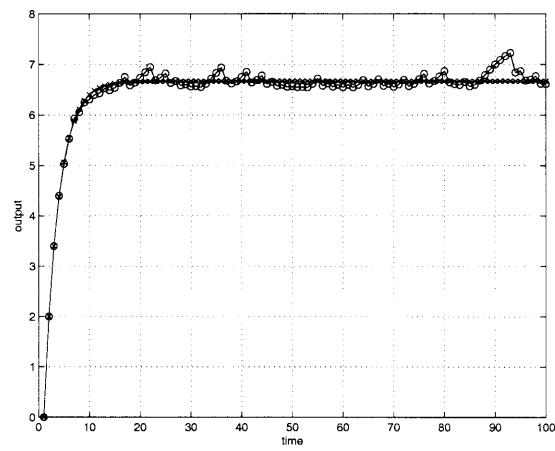


Figure 2-3: Response of  $\mathcal{L}$  and  $\hat{\mathcal{L}}$  to a step input



## Chapter 3

# Model Reduction of Hidden Markov Models

This chapter presents a two step model reduction algorithm for discrete-time, finite state, finite alphabet Hidden Markov Models. The complexity measure used is the cardinality of the state space of the underlying Markov chain. In the first step, Hidden Markov Models are associated with a certain class of stochastic Jump Linear Systems, namely the ones where the parametric input is a sequence of independent identically distributed random variables. The image of the high dimensional Hidden Markov Model in this class of stochastic Jump Linear Systems is simplified by means of a balanced truncation algorithm, which was developed in the previous chapter. Subsequently, the reduced order stochastic Jump Linear System is modified, so that it meets the constraints of an image of a Hidden Markov Model of the same order. This is achieved by solving a low dimensional non convex optimization problem. Numerical simulation results provide evidence that the proposed algorithm computes accurate reduced order Hidden Markov Models, while achieving a compression of the state space by orders of magnitude.

## 3.1 Definitions

Let  $\mathbf{Y}$  be a nonempty, finite set, which will be called the alphabet. The elements of  $\mathbf{Y}$  will be referred to as letters. A monoid  $\mathbf{Y}^*$ , is formed, referred to as the language, consisting of all finite sequences of elements of  $\mathbf{Y}$ , as well as the empty set, which is denoted by  $\emptyset$ . The finite sequences of letters will be called words or strings. The required law of composition is the concatenation operation between strings and the identity element is the empty set  $\emptyset$ . Let  $v$  be a word, its length will be denoted by  $|v|$ , the empty string  $\emptyset$  has length 0. The set of all strings of length  $k \in \mathbf{Z}_+$  is denoted by  $\mathbf{Y}^k$ . The concatenation of  $v$  and  $u$  is written as  $vu$ , and  $|vu| = |v| + |u|$ . Strings are read from right to left, in the sense that in the expression  $vu$ ,  $u$  is followed by  $v$ . This convention will lead to a less cluttered notation in subsequent parts.

### 3.1.1 Hidden Markov Models

Hidden Markov Models can be defined in many equivalent ways. Throughout the paper the definition introduced in the context of realization theory of HMM's will be used, see for instance [25],[1],[31]. Consider  $\{Y(t)\}$  a discrete-time, stationary stochastic process over some fixed probability space  $\{\Omega, \mathcal{F}, \mathbf{P}\}$ , with values on a finite set  $\mathbf{Y} = \{1, \dots, N\}$ ,  $N \in \mathbf{Z}_+$ ,  $N \geq 2$ . The “future” of the process after time  $t$  is denoted by  $Y_t^+ = \{\dots, Y(t+2), Y(t+1)\}$ . At time  $t$  the “past and present” of the process are denoted by  $Y_t^- = \{Y(t), Y(t-1), \dots\}$ . Let  $v = v_k \dots v_1 \in \mathbf{Y}^*$  the notation  $Y_t^+ = v$  stands for the event  $\{\omega \in \Omega | Y(t+k) = v_k, \dots, Y(t+1) = v_1\}$ . Consider now another discrete-time, stationary stochastic process  $\{X(t)\}$  with values on  $\mathbf{X} = \{s_1, \dots, s_n\}$ , where  $n \in \mathbf{Z}_+$ ,  $n \geq 2$ . The joint process  $\{X(t), Y(t)\}$  is a discrete-time, stationary, finite state, finite alphabet HMM of order  $n$  if  $\{X(t), Y(t)\}$  is a Markov process and given arbitrary strings  $\sigma \in \mathbf{X}^*$  and  $v \in \mathbf{Y}^*$ , the following “splitting property” holds

$$\begin{aligned} \mathbf{P}(X_t^+ = \sigma, Y_t^+ = v | X_t^-, Y_t^-) = \\ \mathbf{P}(X_t^+ = \sigma, Y_t^+ = v | X(t)). \end{aligned}$$

The above definition insures that  $\{X(t)\}$  is by itself a Markov chain of order  $n$ , meaning

$$\mathbf{P}(X_t^+ = \sigma | X_t^-) = \mathbf{P}(X_t^+ = \sigma | X(t)).$$

It also insures that  $\{Y(t)\}$  is a probabilistic function of the Markov chain  $\{X(t)\}$  in the sense that

$$\mathbf{P}(Y_t^+ = v | X_t^-, Y_t^-) = \mathbf{P}(Y_t^+ = v | X(t)).$$

The process  $\{X(t)\}$  will be referred to as the state process, which is hidden from the observer and the observable process  $\{Y(t)\}$  will be referred to as the output process. The output process is characterized by the concept of the probability function, which is not to be confused with a probability measure. The probability function  $p : \mathbf{Y}^* \rightarrow \mathbf{R}$  is defined as

$$p(v) = \mathbf{P}(Y_t^+ = v), \quad \forall v \in \mathbf{Y}^*, \forall t \in \mathbf{Z}.$$

Note that since the process is stationary, the value of  $p(v)$  in the above definition does not depend on  $t$ . It can be readily verified, that the probability function satisfies the properties:

$$p(\emptyset) = 1 \tag{3.1}$$

$$p(v) \in [0, 1], \quad \forall v \in \mathbf{Y}^*, \tag{3.2}$$

$$p(v) = \sum_{u \in \mathbf{Y}^k} p(vu), \quad \forall v \in \mathbf{Y}^*, k \in \mathbf{Z}_+. \tag{3.3}$$

A HMM  $\mathbf{H} \in \mathcal{H}$  will be identified with an ordered quadruple  $(\mathbf{X}, e_n, M, \pi)$ . The first quantity is the finite state space. The next three quantities  $e_n \in \mathbf{R}^{1 \times n}$ ,  $M : \mathbf{Y} \rightarrow \mathbf{R}_+^{n \times n}$ ,  $\pi \in \mathbf{R}_+^{n \times 1}$  encode the statistical description of  $\mathbf{H}$ . In particular the row vector of  $n$  one's is denoted by  $e_n$ . The map  $M$  has the probabilistic interpretation

$$M(k)_{ij} = \mathbf{P}(X(t+1) = i, Y(t+1) = k | X(t) = j) \\ \forall i, j \in \mathbf{X}, \forall k \in \mathbf{Y}, \forall t \in \mathbf{Z}_+.$$

Let  $\Pi$  denote the state transition matrix of the Markov process  $\{X(t)\}$ , then clearly

$$\Pi = \sum_{k \in \mathbf{Y}} M(k).$$

The stationary distribution of  $X(t)$  is denoted by  $\pi \in \mathbf{R}_+^n$  satisfying  $\Pi\pi = \pi$ . Using the notation above one can derive that for an arbitrary string  $v = v_k v_{k-1} \dots v_1$ , where  $k \in \mathbf{Z}_+, k \geq 1$ , the following recursive relation holds

$$p_{\mathbf{H}}(v) = e_n M(v_k) \dots M(v_1) \pi.$$

One can think of the map  $M$  being extended by means of a homomorphism from the output alphabet  $\mathbf{Y}$  to the whole language  $\mathbf{Y}^*$  and write  $M(v) = M(v_k) \dots M(v_1)$ . The elements of the matrix  $M(v)$  have the probabilistic interpretation

$$\begin{aligned} M(v)_{ij} &= \mathbf{P}(X(t+|v|) = i, Y_t^+ = v | X(t) = j) \\ \forall i, j \in \mathbf{X}, \forall v \in \mathbf{Y}^*, \forall t \in \mathbf{Z}_+. \end{aligned}$$

Finally there is a minor technical assumption posed to all the HMM's under consideration in this work. It is assumed that  $\exists k \in \mathbf{Z}_+ : p(v) < 1 \forall v \in \mathbf{Y}^k$ . This assumption excludes a possible deterministic evolution of the HMM at hand and guarantees also existence of a positive definite matrix  $P > 0$ , which satisfies the Linear Matrix Inequality (abbreviated LMI)

$$\sum_{y \in \mathbf{Y}} M'(y) P M(y) - P < 0. \quad (3.4)$$

### 3.1.2 Generalized Automata

The concept of a Generalized Automaton was introduced in [30]. A Generalized Automaton over the finite alphabet  $\mathbf{Y}$  of order  $n$  is defined as an ordered quadruple  $(\mathbf{X}, c, A, b)$ , where  $c \in \mathbf{R}^{1 \times n}$ ,  $A : \mathbf{Y} \rightarrow \mathbf{R}^{n \times n}$ ,  $b \in \mathbf{R}^{n \times 1}$  and  $\mathbf{X} = \{s_1, \dots, s_n\}$  is a finite set of states. Let  $v = v_k \dots v_1 \in \mathbf{Y}^*$ , where  $k \in \mathbf{Z}_+, k \geq 2$ , the domain of  $A$  is



extended from  $\mathbf{Y}$  to  $\mathbf{Y}^*$  by defining

$$\begin{aligned} A(\emptyset) &= I_n \\ A(v_k \dots v_1) &= A(v_k) \dots A(v_1) \end{aligned}$$

where  $I_n$  denotes the identity matrix in  $\mathbf{R}^{n \times n}$ .

As already mentioned Generalized Automata (abbreviated GA) are equivalent to recognizable FPS in  $|\mathbf{Y}|$  noncommuting indeterminates with real coefficients, a concept that has been frequently used in the study of formal languages in theoretical computer science, see for instance [4], [10]. In the context of system theory, they have appeared in connection with realization problems of multi-linear, state-affine and uncertain systems, see [16], [28], [2] respectively.

Associated to every Generalized Automaton  $\mathbf{G}$  is a map from the language to the real numbers, called the word function  $q_{\mathbf{G}} : \mathbf{Y}^* \rightarrow \mathbf{R}$  where

$$q_{\mathbf{G}}(v) = cA(v)b, \forall v \in \mathbf{Y}^*.$$

The set of all possible values of the map  $q_{\mathbf{G}}(\mathbf{Y}^*)$  will be referred to also as the coefficients of  $\mathbf{G}$ .

## 3.2 Model Reduction Algorithm

Given is a HMM  $\mathbf{H} = (\mathbf{X}, e_n, M, \pi)$  over the alphabet  $\mathbf{Y}$  of order  $n \in \mathbf{Z}_+, n \geq 3$ , which also satisfies the technical assumption (3.4). The model reduction algorithm envisioned for the class of HMM's under consideration consists of the solution to the following two problems.

### 3.2.1 Problem Statements

#### Problem A

Given is a Generalized Automaton  $\mathbf{G} = (\mathbf{X}, c, A, b)$  over the alphabet  $\mathbf{Y}$  of order  $n \in \mathbf{Z}_+, n \geq 3$ , such that  $\forall Q > 0, \exists P > 0$  where  $P$  and  $Q \in \mathbf{R}^{n \times n}$  satisfying

$$\sum_{y \in \mathbf{Y}} A'(y)PA(y) - P = Q. \quad (3.5)$$

Required is an algorithm that produces for any  $\hat{n} \in \mathbf{Z}_+$ , where  $2 \leq \hat{n} < n$ , a Generalized Automaton  $\hat{\mathbf{G}} = (\hat{\mathbf{X}}, \hat{c}, \hat{A}, \hat{b})$  over the alphabet  $\mathbf{Y}$  of order  $\hat{n}$  that satisfies

$$\sum_{v \in \mathbf{Y}^*} (q_{\mathbf{G}}(v) - q_{\hat{\mathbf{G}}}(v))^2 \leq \epsilon_{\hat{n}},$$

where  $\epsilon_{\hat{n}}$  is an a priori computable error bound.

Note that the above statement is relevant since,  $\mathcal{H} \subset \mathcal{G}$ , so that every HMM can be identified with a Generalized Automaton of the same order. Condition (3.5) is a translation of (3.4) in terms of the structural parameters of a Generalized Automaton.

#### Problem B

Given the Generalized Automaton  $\hat{\mathbf{G}}$  of the above problem statement find an algorithm that computes a HMM  $\hat{\mathbf{H}}$  of order  $\hat{n}$  that minimizes the distance measure

$$\sum_{v \in \mathbf{Y}^*} (p_{\hat{\mathbf{H}}}(v) - q_{\hat{\mathbf{G}}}(v))^2$$

### 3.2.2 Reduction Algorithm

#### Solution to Problem A

The following correspondence between  $\mathcal{G}$  and  $\mathcal{L}$ , where  $\Theta = \mathbf{Y}$ , is established. The Generalized Automaton  $\mathbf{G} = (\mathbf{X}, c, A, b)$  that satisfies (3.5) is mapped to the iid-JLS

$\mathbf{L}$  that has the state space realization

$$\begin{aligned}x(t+1) &= \sqrt{N}A[\theta(t)]x(t) + bf(t), \\y(t) &= cx(t), \quad t \in \mathbf{Z}_+, \end{aligned}$$

and

$$\mathbf{P}(\theta(t) = i) = q_i = \frac{1}{N}, \quad \forall i \in \Theta.$$

Note that  $\mathbf{L}$  is mean square stable, by virtue of the assumption (3.5). By applying the aforementioned balanced truncation algorithm for iid-JLS's one gets the reduced system  $\hat{\mathbf{L}}$  of order  $\hat{n}$  with the state space realization

$$\begin{aligned}\hat{x}(t+1) &= \sqrt{N}\hat{A}[\theta(t)]\hat{x}(t) + \hat{b}f(t), \\ \hat{y}(t) &= \hat{c}\hat{x}(t), \quad t \in \mathbf{Z}_+. \end{aligned}$$

The distance in terms of the stochastic  $L_2$  gain between the two systems satisfies the bound

$$d_{\mathcal{L}}(\mathbf{L}, \hat{\mathbf{L}}) \leq 2(\beta_1 + \dots + \beta_s) = \sqrt{\epsilon_{\hat{n}}}$$

The reduced order iid-JLS  $\hat{\mathbf{L}}$  is mapped back to the Generalized Automaton  $\hat{\mathbf{G}} = (\hat{\mathbf{X}}, \hat{c}, \hat{A}, \hat{b})$  of order  $\hat{n}$ . A topological equivalence is established between  $\mathcal{G}$  and  $\mathcal{L}$  by means of the above correspondence and by inducing the following metric on  $\mathcal{G}$

$$d_{\mathcal{G}}(\mathbf{G}, \hat{\mathbf{G}}) := d_{\mathcal{L}}(\mathbf{L}, \hat{\mathbf{L}}).$$

Automatically one gets then

$$d_{\mathcal{G}}(\mathbf{G}, \hat{\mathbf{G}}) \leq \sqrt{\epsilon_{\hat{n}}}$$

This is a stronger bound than the one required in problem statement A. Let  $x(0) = 0$  and apply the input  $f = (1, 0, 0, \dots)$  to the error system  $\mathcal{E}$  between  $\mathbf{L}$  and  $\hat{\mathbf{L}}$ . One obtains then

$$\sum_{v \in \mathbf{Y}^*} (q_{\mathbf{G}}(v) - q_{\hat{\mathbf{G}}}(v))^2 = \mathbf{E}\left[\sum_{t=0}^{\infty} |e(t)|^2\right].$$

The above equation and the definition of the stochastic  $L_2$  gain lead directly to the inequality

$$\sum_{v \in \mathbf{Y}^*} (q_{\mathbf{G}}(v) - q_{\hat{\mathbf{G}}}(v))^2 \leq d_G(\mathbf{G}, \hat{\mathbf{G}})^2,$$

and consequently to

$$\sum_{v \in \mathbf{Y}^*} (q_{\mathbf{G}}(v) - q_{\hat{\mathbf{G}}}(v))^2 \leq \epsilon_{\hat{n}}$$

By virtue of the relation above, problem statement A is considered solved.

### Solution to problem B

The word function of the reduced order Generalized Automaton  $\hat{\mathbf{G}}$  approximates the probability function of the original HMM  $\mathbf{H}$  within an a priori computable error bound. However,  $q_{\hat{\mathbf{G}}}$  does not necessarily satisfy relations (3.1)-(3.3). In fact, even in the special case when the structural parameters of a Generalized Automaton have integer entries, checking nonnegativity of  $q_{\hat{\mathbf{G}}}$  is an undecidable problem [7]. For the case of a single letter alphabet NP-hardness complexity results may be found in [8].

These circumstances motivate the search for a HMM  $\hat{\mathbf{H}} = (\hat{\mathbf{X}}, e_{\hat{n}}, \hat{M}, \hat{\pi})$  over the alphabet  $\mathbf{Y}$  of order  $\hat{n}$  that minimizes the objective

$$\sum_{v \in \mathbf{Y}^*} (p_{\hat{\mathbf{H}}}(v) - q_{\hat{\mathbf{G}}}(v))^2. \quad (3.6)$$

The constraints that are satisfied by the structural parameters of  $\hat{\mathbf{H}}$  are:

$$\hat{\pi}_i \geq 0 \quad \forall i \in \hat{\mathbf{X}}, \quad (3.7)$$

$$\hat{M}(k)_{ij} \geq 0 \quad \forall i, j \in \hat{\mathbf{X}}, \forall k \in \mathbf{Y}, \quad (3.8)$$

$$e_{\hat{n}} = e_{\hat{n}} \sum_{k \in \mathbf{Y}} [M(k)], \quad (3.9)$$

$$\hat{\pi} = \sum_{k \in \mathbf{Y}} [M(k)] \hat{\pi}. \quad (3.10)$$

One can introduce the error system between  $\hat{\mathbf{G}}$  and  $\hat{\mathbf{H}}$ , which is a Generalized Au-

tomaton  $\tilde{\mathbf{G}}$  of order  $\tilde{n} = 2\hat{n}$ , where

$$\tilde{A}(k) = \begin{bmatrix} \hat{M}(k) & 0 \\ 0 & \hat{A}(k) \end{bmatrix}, \forall k \in \mathbf{Y}$$

$$\tilde{b} = \begin{bmatrix} e_{\hat{n}} \\ \hat{b} \end{bmatrix}, \tilde{c} = \begin{bmatrix} \hat{\pi} & -\hat{c} \end{bmatrix}$$

The objective function can then be written as

$$\sum_{v \in \mathbf{Y}^*} (p_{\hat{\mathbf{H}}}(v) - q_{\tilde{\mathbf{G}}}(v))^2 = \tilde{b}' W_c \tilde{b}$$

where  $W_c$  satisfies the Lyapunov like equation

$$W_c = \tilde{c}' \tilde{c} + \sum_{k \in \mathbf{Y}} [\tilde{A}'(k) W_c \tilde{A}(k)].$$

The above problem is a non convex optimization problem that is being solved by means of a gradient flow algorithm. As it is typical with non convex formulations there are no guarantees of convergence to the global minimum, however numerical simulation results have been encouraging as far as the approximation error incurred in this step is concerned. It is worth mentioning that one can consider stronger metrics than (3.6). For instance let  $W_b$  be defined as

$$W_b = \tilde{b} \tilde{b}' + \sum_{k \in \mathbf{Y}} [\tilde{A}(k) W_b \tilde{A}'(k)],$$

then one could attempt to minimize the objective

$$\text{trace}[W_c W_b],$$

instead of (3.6). This direction is worth exploring, especially the tradeoff associated with the increase of the computational cost, due to the solution of an additional Lyapunov like equation at each iteration step.

Another possible problem formulation is to eliminate the first reduction step and minimize directly the objective

$$\sum_{v \in \mathbf{Y}^*} (p_{\hat{\mathbf{H}}}(v) - p_{\mathbf{H}}(v))^2.$$

Note that in such a situation the error system between  $\mathbf{H}$  and  $\hat{\mathbf{H}}$  is a Generalized Automaton  $\tilde{\mathbf{G}}$  of order  $\tilde{n} = \hat{n} + n$ . As  $n$  increases the solution of such a non convex optimization problem becomes intractable. The main advantage of the first reduction step is that in typical applications very accurate approximations of the original HMM in the space of GA can be found with  $\hat{n} \ll n$ .

### 3.3 A numerical example of the method

Denote by  $S_n$  the set of all possible permutations of the set  $\{1, \dots, n\}$ ,  $n \in \mathbf{Z}_+$ . Let  $\sigma \in S_n$  and denote by  $\Gamma_\sigma$  the corresponding permutation matrix. Set  $n = 100$  and pick a  $\sigma \in S_n$  at random. Define the matrix  $\Sigma = \Gamma_\sigma + \epsilon \Delta$ , where  $\epsilon = 0.02$  and  $\Delta \in \mathbf{R}^{n \times n}$ . Each entry of  $\Delta$  is drawn from a uniform distribution on the unit interval,  $[0, 1]$ . Let  $w = e_n \Sigma$  and form the diagonal matrix  $D$  with  $D_{ii} = w_i^{-1}$ ,  $i \in \{1, \dots, n\}$ . One can verify that the matrix  $\Pi = \Sigma D$  is column stochastic, thus it can be considered as the transition matrix of a Markov chain evolving on  $\mathbf{X} = \{s_1, \dots, s_n\}$ . Note also that  $\mathbf{P}(\Pi_{ij} > 0) = 1, \forall i, j \in \{1, \dots, n\}$ , due to the way that  $\Pi$  was generated, thus the stationary distribution corresponding to  $\Pi$  is unique. Consider the stationary Markov process  $\{X(t)\}$  on  $\mathbf{X}$  with transition matrix  $\Pi$  and obtain a stationary HMM  $\mathbf{H}$  over  $\mathbf{Y} = \{0, 1\}$  by defining the output process as a deterministic function of the state process. In particular  $Y(t) = 0$  if  $X(t) \in \{s_1, \dots, s_{50}\}$  and  $Y(t) = 1$  if  $X(t) \in \{s_{51}, \dots, s_{100}\}$ . The HMM  $\mathbf{H}$  generated following the procedure described above is used to demonstrate the reduction algorithm. The first figure depicts the eigenvalues of the matrix  $W$ , that controls the error bound between  $\mathbf{H}$  and  $\hat{\mathbf{G}}$ . Overall there is an evident decay in the eigenvalues of  $W$ . The HMM  $\mathbf{H}$  is truncated to a Generalized Automaton  $\hat{\mathbf{G}}$  that has  $\hat{n} = 5$  states. The language  $\mathbf{Y}^*$  is countable

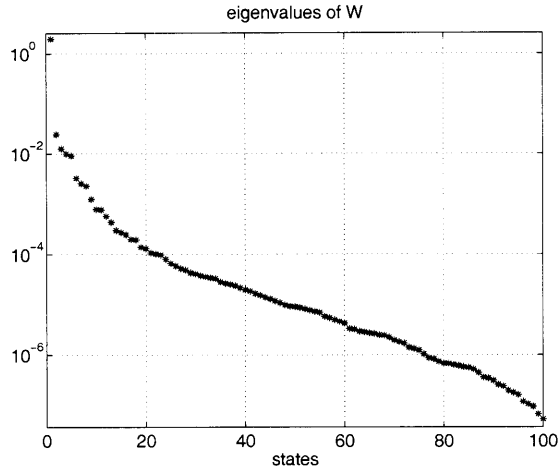


Figure 3-1: Eigenvalues of  $W$ , logarithmic scale on y axis

and one can impose for example the first-lexical order on it,  $fl : \mathbf{Z}_+ \rightarrow \mathbf{Y}^*$ , where  $fl(1) = \emptyset, fl(2) = 0, fl(3) = 1, fl(4) = 00$ , and so on. The next figure depicts the word-function of the error system between  $\mathbf{H}$  and  $\hat{\mathbf{G}}$ , i.e. the function  $p_{\mathbf{H}}(v) - q_{\hat{\mathbf{G}}}(v)$  for all strings up to length 10. The strings in the x-axis of the above graph are arranged

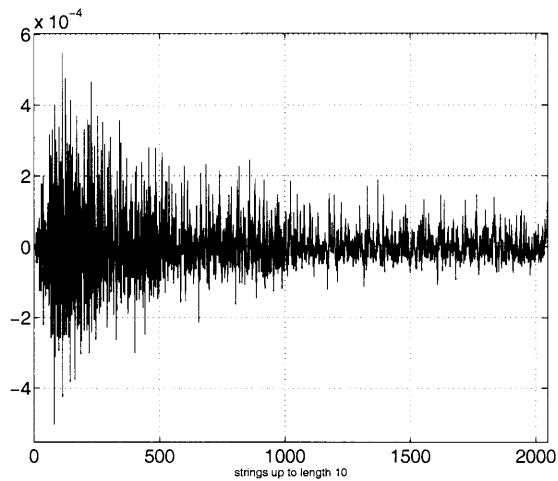


Figure 3-2: Word function of error system between  $\mathbf{H}$  and  $\hat{\mathbf{G}}$

according to the aforementioned first-lexical order. The calculated a posteriori bound  $\sup_{v \in \mathbf{Y}^*} |q_{\hat{\mathbf{G}}} - p_{\mathbf{H}}| < 5.5 \times 10^{-4}$  confirms that the word-function of  $q_{\hat{\mathbf{G}}}$  provides a good approximation of  $p_{\mathbf{H}}$ . Now  $\hat{\mathbf{G}}$  is turned into a HMM  $\hat{\mathbf{H}}$  by solving the low dimensional

non convex optimization problem B. The last figure depicts the word-function of the error system between the high dimensional HMM  $\mathbf{H}$  of order 100 and the low dimensional HMM  $\hat{\mathbf{H}}$  of order 5, for all strings up to length 10, again the first-lexical order was used in the x-axis. The a posteriori bound  $\sup_{v \in \mathcal{Y}^*} |p_{\hat{\mathbf{H}}} - p_{\mathbf{H}}| < 9.3 \times 10^{-3}$  shows that the reduction algorithm performed adequately in this particular example, achieving a compression by a factor of 20.

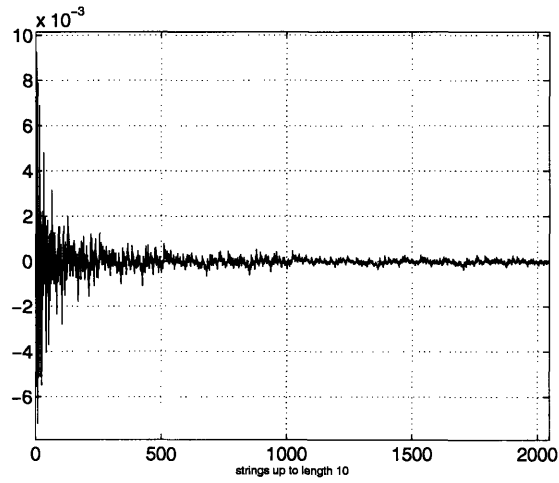


Figure 3-3: Word function of error system between  $\mathbf{H}$  and  $\hat{\mathbf{H}}$



# Chapter 4

## Reduction of Nearly Decomposable Hidden Markov Models

In this chapter the concept of aggregation, clustering, of the state space is investigated as means of reducing the complexity of HMM's. One of the main uses of clustering has been in the computation of the asymptotic distribution of subsets of the states of large scale Markov chains. Existing ideas will be extended to the HMM case.

### 4.1 Nearly Decomposable Hidden Markov Model

As in the previous chapter the point of departure is a stationary discrete-time, finite state, finite alphabet HMM  $\mathbf{H} = (\mathbf{X}, e_n, M, \pi)$  over the alphabet  $\mathbf{Y}$  of order  $n \in \mathbf{Z}_+, n \geq 3$ . However, in the nearly decomposable case, each transition matrix has almost block diagonal structure.

$$M(y) = M^*(y) + \epsilon\Delta(y),$$
$$M^*(y) = \begin{bmatrix} M_1(y) & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & M_K(y) \end{bmatrix}, y \in \mathbf{Y},$$

where  $\epsilon$  denotes a coupling factor that appears once the entries of  $\Delta$  are normalized  $|\Delta_{ij}(y)| \leq 1 \quad \forall i, j \in \{1, \dots, n\}, y \in \mathbf{Y}$  and  $K$  denotes the number of clusters.

The concept of a nearly decomposable HMM is compatible with the notion of a nearly decomposable Markov chain, which are Markov chains with almost block diagonal transition matrix. Nearly decomposable Markov chains were introduced as means for modeling stochastic systems whose dynamics exhibit multiple scales. The behavior of such systems can be analyzed in stages. In a first stage one considers each cluster that corresponds to a respective block independently until a partial equilibrium is achieved. In a second stage each cluster is considered as a single aggregated entity and interactions between aggregates lead to the steady state distribution for the system as a whole. The reduction procedure for nearly decomposable HMM's employs aggregation of the state space, otherwise known as clustering.

## 4.2 State Aggregation

The state space  $\mathbf{X}$  is partitioned in  $\hat{n} = K$  disjoint clusters. The new state space of aggregated states is  $\hat{\mathbf{X}} = \{\hat{S}_1, \hat{S}_2, \hat{S}_3, \dots, \hat{S}_{\hat{n}}\}$  where

1.  $\hat{S}_i \subset \mathbf{X} \quad i \in \{1, \dots, \hat{n}\},$
2.  $\hat{S}_i \cap \hat{S}_j = \emptyset$  if  $i \neq j$
3.  $\bigcup_{i=1}^{\hat{n}} \hat{S}_i = \mathbf{X}.$

Next the aggregation operator  $L : \mathbf{R}^n \rightarrow \mathbf{R}^{\hat{n}}$  where

$$L_{ij} = \begin{cases} 1 & \text{if } s_j \in \hat{S}_i, \\ 0 & \text{otherwise} \end{cases}$$

is introduced. Let

$$\pi(t) = \left[ \mathbf{P}[X(t) = s_1] \quad \dots \quad \mathbf{P}[X(t) = s_n] \right]'$$

denote the instantaneous probability distribution of the underlying Markov chain of the original model and respectively

$$\hat{\pi}(t) = \left[ \mathbf{P}[\hat{X}(t) = \hat{S}_1] \quad \dots \quad \mathbf{P}[\hat{X}(t) = \hat{S}_{\hat{n}}] \right]'$$

the corresponding quantity for the aggregated model, then it holds

$$\hat{\pi}(t) = L \pi(t) = \left[ \sum_{s_i \in \hat{S}_1} \mathbf{P}[X(t) = s_i], \quad \dots, \quad \sum_{s_i \in \hat{S}_{\hat{n}}} \mathbf{P}[X(t) = s_i] \right]'$$

Let  $\lambda_{jm}(t)$  denote the conditional probability of the state  $s_j$  in cluster  $\hat{S}_m$  at time instant  $t$  i.e.,

$$\lambda_{jm}(t) = \frac{\pi_j(t)}{\hat{\pi}_m(t)}$$

the entries of  $\hat{\Pi}(t)$ , the aggregated transition probability matrix at instant  $t$  are given by

$$\hat{\Pi}_{mk}(t) = \sum_{j : s_j \in \hat{S}_m} \lambda_{jm}(t) \sum_{i : s_i \in \hat{S}_k} \Pi_{ji} \quad (4.1)$$

the evolution of the probability distribution for the aggregated system is given by

$$\hat{\pi}(t+1) = \hat{\Pi}(t) \hat{\pi}(t) \quad (4.2)$$

Note that the Markovian property is preserved, the chain is time inhomogenous though. Computation of the exact value of  $\hat{\Pi}(t)$  requires at each instant a disaggregation step in order to obtain the values of the conditional distributions  $\lambda_{jm}(t)$  in every cluster  $\hat{S}_m$ . From a computational standpoint this is equivalent with working with the original system, thus exact calculation of the aggregation matrix does not bear any benefit. For a given aggregation operator  $L$  we define the compact set of stochastic matrices  $\Pi_L$  where

$$\Pi_L = \{ \hat{\Pi} : \hat{\Pi}_{mk} = \sum_{j : s_j \in \hat{S}_m} \lambda_{jm} \sum_{i : s_i \in \hat{S}_k} \Pi_{ji} ; \lambda_{jm} \in [0, 1] \}$$

An equivalent way to equation (4.2) for describing the evolution of the probability distribution of the aggregated system is given by:

$$\hat{\pi}(t+1) = \hat{\Pi}(t)\hat{\Pi}(t-1)\dots\hat{\Pi}(1)\hat{\pi}(0) \quad (4.3)$$

where  $\hat{\Pi}(t) \in \Pi_L$ . A low order approximation of the original system by a homogenous Markov chain requires a selection of a fixed matrix  $\hat{\Pi}$  where  $\hat{\Pi} \in \Pi_L$ . This is equivalent with fixing the values of conditional probabilities in each cluster  $\forall t \in \mathbf{Z}_+$ . The approximate dynamics on the aggregated state space will be described by:

$$\hat{\pi}^*(t+1) = \hat{\Pi}\hat{\pi}^*(t)$$

The vector  $\hat{\pi}^*(t)$  is regarded as an approximation to the exact probability distribution  $\hat{\pi}(t)$  on the aggregated state space.

### 4.3 Reduction Algorithm

When it comes to reducing the dimensionality of a nearly decomposable HMM with  $n$  states one can readily obtain a reduced order model by considering an aggregated state space induced by the block diagonal structure of the transition matrices. The conditional probabilities in each cluster are fixed to the values corresponding to the asymptotic distribution of the original Markov chain. The following example demonstrates the method. Consider the HMM  $\mathbf{H}$  that has 8 states and a binary output alphabet  $\mathbf{Y} = \{0, 1\}$ . The transition matrix of the underlying Markov chain has nearly decomposable structure :

$$\Pi = \begin{bmatrix} \Pi_{11} & \Pi_{12} & \Pi_{13} \\ \Pi_{21} & \Pi_{22} & \Pi_{23} \\ \Pi_{31} & \Pi_{32} & \Pi_{33} \end{bmatrix}$$

$$\begin{aligned} \Pi_{11} &= \begin{bmatrix} 0.85 & 0 & 0.149 \\ 0.1 & 0.65 & 0.249 \\ 0.1 & 0.8 & 0.0996 \end{bmatrix} \\ \Pi_{12} &= \begin{bmatrix} 0.0009 & 0 \\ 0 & 0.0009 \\ 0.0003 & 0 \end{bmatrix} \\ \Pi_{13} &= \begin{bmatrix} 5 \cdot 10^{-5} & 0 & 5 \cdot 10^{-5} \\ 5 \cdot 10^{-5} & 0 & 5 \cdot 10^{-5} \\ 0 & 4 \cdot 10^{-4} & 0 \end{bmatrix} \\ \Pi_{21} &= \begin{bmatrix} 0 & 4 \cdot 10^{-4} & 0 \\ 5 \cdot 10^{-4} & 0 & 4 \cdot 10^{-4} \end{bmatrix} \\ \Pi_{22} &= \begin{bmatrix} 0.7 & 0.2995 \\ 0.399 & 0.6 \end{bmatrix} \\ \Pi_{23} &= \begin{bmatrix} 0 & 0.0001 & 0 \\ 0.0001 & 0 & 0 \end{bmatrix} \\ \Pi_{31} &= \begin{bmatrix} 0 & 5 \cdot 10^{-5} & 0 \\ 3 \cdot 10^{-5} & 0 & 3 \cdot 10^{-5} \\ 0 & 5 \cdot 10^{-5} & 0 \end{bmatrix} \\ \Pi_{32} &= \begin{bmatrix} 0 & 5 \cdot 10^{-5} \\ 4 \cdot 10^{-5} & 0 \\ 0 & 5 \cdot 10^{-5} \end{bmatrix} \\ \Pi_{33} &= \begin{bmatrix} 0.6 & 0.2499 & 0.15 \\ 0.1 & 0.8 & 0.0999 \\ 0.1999 & 0.25 & 0.55 \end{bmatrix} \end{aligned}$$

The above stochastic matrix  $\Pi$  is irreducible and nearly decomposable, thus it can be written in the form  $\Pi = \Pi^* + \epsilon\Delta$  with  $\epsilon = 0.001$  and

$$P^* = \begin{bmatrix} P_1^* & 0 & 0 \\ 0 & P_2^* & 0 \\ 0 & 0 & P_3^* \end{bmatrix}$$

$$\Pi_1^* = \begin{bmatrix} 0.85 & 0 & 0.15 \\ 0.10 & 0.65 & 0.25 \\ 0.10 & 0.80 & 0.10 \end{bmatrix}$$

$$\Pi_2^* = \begin{bmatrix} 0.70 & 0.30 \\ 0.40 & 0.60 \end{bmatrix}$$

$$\Pi_3^* = \begin{bmatrix} 0.60 & 0.25 & 0.15 \\ 0.10 & 0.80 & 0.10 \\ 0.20 & 0.25 & 0.55 \end{bmatrix}$$

The matrix  $\Delta$  is omitted since it is not relevant to the subsequent calculations. The observations are related to the states by means of a deterministic output function. In particular  $\mathbf{P}[Y(t+1) = 0 | X(t) = i] = 1, i \in \{s_1, s_2, s_3\}$  and  $\mathbf{P}[Y(t+1) = 0 | X(t) = i] = 1, i \in \{s_4, s_5, s_6, s_7, s_8\}$ . In accordance to the block structure of  $P^*$  a suitable partition of the state space consists of 2 clusters,  $\hat{S} = \{\hat{S}_1, \hat{S}_2\}$  with  $\hat{S}_1 = \{1, 2, 3\}$ ,  $\hat{S}_2 = \{4, 5, 6, 7, 8\}$ . The following figure depicts the word function of the error system for all strings up to length 10.

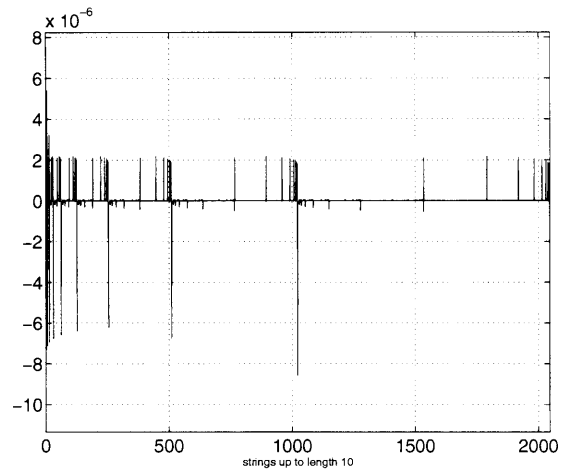


Figure 4-1: Word function of error system between  $\mathbf{H}$  and  $\hat{\mathbf{H}}$





# Bibliography

- [1] B.D.O. Anderson. The realization problem for hidden markov models. *Mathematics of Control, Signals, and Systems*, 12(1):80–122, 1999.
- [2] C. Beck. On formal power series representations for uncertain systems. *IEEE Transactions on Automatic Control*, 46(2):314–319, February 2001.
- [3] L. Benvenuti and L. Farina. A tutorial on the positive realization problem. *IEEE Transactions on Automatic Control*, 49(5):651–664, 2004.
- [4] J. Berstel and C. Reutenauer. *Rational series and their languages*. Springer-Verlag, 1988.
- [5] R. Bhar and S. Hamori. *Hidden Markov Models : applications to financial economics*. Kluwer Academic Publishers, 2004.
- [6] D. Blackwell and L. Koopmans. On the identifiability problem for functions of finite markov chains. *Annals of Mathematical Statistics*, 28(4):1011–1015, 1957.
- [7] V.D. Blondel. *private communication*. 2006.
- [8] V.D. Blondel and N. Portier. The presence of a zero in an integer linear recurrent sequence is np-hard to decide. *Linear Algebra and its Applications*, 351:91–118, August 2002.
- [9] J. Doyle C. L. Beck and K. Glover. Model reduction of multidimensional and uncertain systems. *IEEE Transactions on Automatic Control*, 41(10):1466–1477, October 1996.

- [10] S. Eilenberg. *Automata, languages, and machines*. Academic Press, 1974.
- [11] D. Enns. Model reduction with balanced realizations: An error bound and a frequency weighted generalization. In *Proceedings of the 23rd IEEE Conference on Decision and Control*, pages 127 – 132, Las Vegas, USA, December 1984.
- [12] L. Finesso and P. Spreij. Approximate realization of finite hidden markov chains. In *Proceedings of the IEEE Information Theory Workshop*, pages 90 – 93, 2002.
- [13] K. Glover. All optimal hankel-norm approximations of linear-multivariable systems and their  $l^\infty$  - error bounds. *International Journal of Control*, 39(6):1115–1193, 1984.
- [14] A. Rantzer H. Sandberg. Balanced truncation of linear time-varying systems. *IEEE Transactions on Automatic Control*, 49(02):217 – 229, February 2004.
- [15] A. Heller. On stochastic processes derived from markov chains. *Annals of Mathematical Statistics*, 36(4):1286–1291, August 1965.
- [16] A. Isidori. Direct construction of minimal bilinear realizations from nonlinear input-output maps. *IEEE Transactions on Automatic Control*, 18(6):626–631, December 1973.
- [17] T. Koski. *Hidden Markov Models for Bioinformatics*. Kluwer Academic Publishers, 2001.
- [18] F. Kozin. A survey of stability of stochastic systems. *Automatica*, 5(1):95–112, 1969.
- [19] N.N. Krasovskii and E.A. Lidskii. Analytical design of controllers in systems with random attributes i,ii,iii. *Automation and Remote Control*, 22:1021–1025, 1141–1146, 1289–1294, 1961.
- [20] H. Kushner. *Stochastic stability and control*. Academic Press, 1967.

- [21] R. Mahony L. B. White and G. D. Brushe. Lumpable hidden markov models - model reduction and reduced complexity filtering. *IEEE Transactions on Automatic Control*, 45(12):2297–2306, 2000.
- [22] B.C. Moore. Principal component analysis in linear systems, controllability, observability, and model reduction. *IEEE Transactions on Automatic Control*, 26(1):17–32, January 1981.
- [23] M.D. Fragoso O.L.V. Costa and R.P. Marques. *Discrete-time Markov jump linear systems*. Springer, 2005.
- [24] A. Paz. *Introduction to probabilistic automata*. Academic Press, 1971.
- [25] G. Picci. On the internal structure of finite state stochastic processes. In *Recent Developments in Variable Structure Systems*, volume 162 of *Lecture Notes in Economics and Mathematical Systems*, pages 288–304. Springer, 1978.
- [26] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257 – 286, February 1989.
- [27] C. E. Shannon. A mathematical theory of communication. *The Bell Systems Technical Journal*, 27:379–423, July 1948.
- [28] E. Sontag. Realization theory of discrete-time nonlinear systems: Part i-the bounded case. *IEEE Transactions on Circuits and Systems*, 26(5):342–356, May 1979.
- [29] D. Sworder. Feedback control of a class of linear systems with jump parameters. *TAC*, 14(1):9–14, 1969.
- [30] P. Turakainen. Generalized automata and stochastic languages. *Proceedings of the American Mathematical Society*, 21(2):303 – 309, May 1969.
- [31] M. Vidyasagar. The realization problem for hidden markov models: The complete realization problem. In *Proceedings of the 44th IEEE Conference on Decision and Control*, pages 6632 – 6637, Seville, Spain, December 2005.

- [32] W. M. Wonham. Random differential equations in control theory. In A. T. Bharucha-Reid, editor, *Probabilistic Methods in Applied Mathematics*, volume 2, pages 131–212. Academic Press, 1970.