PH.D. THESIS

Modeling and Reduction with Applications to Semiconductor Processing

by Andrew J. Newman
Advisor: P.S. Krishnaprasad

CDCSS Ph.D. 99-2
(ISR Ph.D. 99-5)

The Center for Dynamics and Control of Smart Structures (CDCSS) is a joint Harvard University, Boston University, University of Maryland center, supported by the Army Research Office under the ODDR&E MURI97 Program Grant No. DAAG55-97-1-0114 (through Harvard University). This document is a technical report in the CDCSS series originating at the University of Maryland.

Web site http://www.isr.umd.edu/CDCSS/cdcss.html

ABSTRACT

Title of Dissertation: MODELING AND REDUCTION WITH APPLICATIONS TO SEMICONDUCTOR PROCESSING

Andrew J. Newman, Doctor of Philosophy, 1999

Dissertation directed by: Professor P. S. Krishnaprasad
Department of Electrical and Computer Engineering

This thesis consists of several somewhat distinct but connected parts, with an underlying motivation in problems pertaining to control and optimization of semiconductor processing. The first part (Chapters 3 and 4) addresses problems in model reduction for nonlinear state-space control systems. In 1993, Scherpen generalized the balanced truncation method to the nonlinear setting. However, the Scherpen procedure is not easily computable and has not yet been applied in practice. We offer a method for computing a working approximation to the controllability energy function, one of the main objects involved in the method. Moreover, we show that for a class of second-order mechanical systems with dissipation, under certain conditions related to the dissipation, an exact formula for

the controllability function can be derived. We then present an algorithm for a numerical implementation of the Morse-Palais lemma, which produces a local coordinate transformation under which a real-valued function with a non-degenerate critical point is quadratic on a neighborhood of the critical point. Application of the algorithm to the controllabilty function plays a key role in computing the balanced representation. We then apply our methods and algorithms to derive balanced realizations for nonlinear state-space models of two example mechanical systems: a simple pendulum and a double pendulum.

The second part (Chapter 5) deals with modeling of rapid thermal chemical vapor deposition (RTCVD) for growth of silicon thin films, via first-principles and empirical analysis. We develop detailed process-equipment models and study the factors that influence deposition uniformity, such as temperature, pressure, and precursor gas flow rates, through analysis of experimental and simulation results. We demonstrate that temperature uniformity does not guarantee deposition thickness uniformity in a particular commercial RTCVD reactor of interest. In the third part (Chapter 6) we continue the modeling effort, specializing to a control system for RTCVD heat transfer. We then develop and apply ad-hoc versions of prominent model reduction approaches to derive reduced models and perform a comparative study.

MODELING AND REDUCTION WITH APPLICATIONS TO SEMICONDUCTOR PROCESSING

by

Andrew J. Newman

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
1999

Advisory Committee:

Professor P. S. Krishnaprasad, Chairman
Professor Carlos A. Berenstein
Professor William S. Levine
Professor Steven I. Marcus
Professor Shihab A. Shamma

PREFACE

"What have you done for science?"

—P. S. Krishnaprasad

This question has been posed to me repeatedly over the past five years. With this thesis, I finally offer an answer.

The contents herein take some small steps toward making nonlinear balancing, a new model reduction method introduced by other researchers, accessible for use in practical applications. I hope that this work is followed by improvements, and leads to additional interesting and useful results, and, ultimately, to practical implementation. I intend to pursue this goal and would be pleased if other researchers find fruitful points of departure. I also hope that readers who have worked with standard versions of the methods described in this thesis will gain a better understanding of when and how to use them.

It has been my good fortune, through the skillful and tireless efforts of my advisor, to have my research funded by various grants and a joint project with an industrial partner, Northrop Grumman Corporation (Electronic Sensors and Systems Sector) of Linthicum, Maryland. This project has afforded me the invaluable experience of performing research toward solving "real-world" manufacturing problems from which I have benefited greatly. It is worthwhile to mention that in such a situation the objectives of the parties may not perfectly coincide, e.g.,

enhanced fundamental understanding of physical-chemical processes versus manufacturing support leading to immediately tangible cost reductions. I can only hope that I have balanced the competing pressures in a way that is somewhat satisfying to all of the involved parties.

During the more than seven years I spent here in College Park, the campus and the city have changed and improved signficantly in many ways. I consider myself lucky to have worked here during the presidency of Dr. William E. Kirwan. Moreover, the ISR, under the leadership of Dr. Steven Marcus and Dr. Gary Rubloff, and the ECE Department, under the leadership of Dr. Nariman Farvardin, have been wonderful and stimulating places to work over this period. I have noticed the results of their efforts on a daily basis. Moreover, it has been a pleasure to work with Dr. David Bader and Dr. Raadhakrishnan Poovendran toward the establishment of a thriving ECE graduate student association. Finally, it is a good thing that College Park has an establishment dedicated to *"smoothies,"* as they kept my energy level high without resorting to coffee (the beverage of choice, it seems, for graduate students).

My advisor gave me all sorts of advice and guidance over the past several years on a variety of subjects. Yet there was a common theme that is characterized in a statement that he repeated often. It is one that I will remember and try to apply in the future.

"Be relentless in the pursuit of knowledge."

—P. S. Krishnaprasad

Andrew Joseph Newman
College Park, Maryland
December, 1999

DEDICATION

To Mom and Dad

ACKNOWLEDGEMENTS

> "The life of a graduate student is a lonely one. Deep thinking requires solitude. Original ideas are, by definition, from within oneself."
>
> —P. S. Krishnaprasad

These observations partially describe my experience in this journey. I am pleased to have this opportunity to write about the other part.

It has been my privilege to study under the direction of my advisor, Prof. P. S. Krishnaprasad. It is through his dedication to teaching and producing new knowledge, insistence on meeting all challenges forthrightly, and granting of freedom to explore ideas, that I have developed the skill, perspective, and discipline necessary to complete this thesis. I thank him for providing copious amounts of thoughtful guidance, constructive criticism, encouragement, and expertise on every subject that arose, in addition to generous financial support, and the opportunity to travel and learn from others.

There are a number of other individuals who provided crucial assistance toward the research in this thesis. I am extremely grateful to:

- Prof. Jacqueline Scherpen of Delft University, The Netherlands, and Prof. W. Steven Gray of Old Dominion University for reviewing my work during their visits to give inspiring talks at the University of Maryland. Their expertise in the subject of nonlinear balancing is unique and invaluable;

- Prof. Howard Stone of Harvard University, for discussions about issues of scaling and power laws in the analysis of gas flows and the relationship between growth rate and flow rates in the Epsilon-1 RTCVD reactor;

- Prof. Steven Marcus and Prof. David Elliott, who provided advice and expertise on a variety of topics. In particular, this thesis greatly profited from their deep knowledge and vast experience in the area of stochastically excited systems. I am also indebted to Prof. Marcus for his suggestions and encouragement toward making my job search a successful one.

- Prof. Raymond Adomaitis, who served as my co-advisor from the summer of 1994 until the summer of 1996, for sparking and nurturing my interest in model reduction and semiconductor process modeling. His expertise and perspective in these areas had an immense and positive influence on my work.

- Mr. Sam Ponczak, Mr. Paul Brabant, Dr. Thomas Knight, and Dr. Michael O'Loughlin, who were our colleagues and industrial partners at Northrop Grumman ESSS. Each of these individuals provided a unique perspective, motivation, ideas, analysis, and commentary throughout the two year project. Mr. Brabant worked with me to perform the growth and lamp heating experiments using the Epsilon-1 reactor, and supplied a great deal of information about the reactor and its operation including the results of previous experiments and production runs. His experience and involvement with the project were indispensable. Finally, I am indebted to Mr. Ponczak for his primary role in securing funding for my research on the project.

- Dr. Doug Meyer, Mr. Tony Komasa, Mr. Rod Smith, and co-workers at ASM America, Inc., Phoenix, AZ, the manufacturers of the Epsilon-1 reactor, who

provided us with invaluable tours, demonstrations, and detailed information. In addition, I thank Dr. Meyer for his analysis and comments regarding process and equipment details and physics.

I would like to thank Prof. Carlos Berenstein, Prof. William Levine, Prof. Steven Marcus, and Prof. Shihab Shamma for serving on my Ph. D. examining committee and providing many useful comments and improvements to this thesis. I also offer thanks to Dr. Eric Justh for reading my thesis in advance and communicating many useful comments and improvements.

I am grateful to all of my teachers. In particular, I would like to thank Prof. Stuart Antman for his fascinating lectures, Prof. A. Yavuz Oruc and Prof. William Levine for offering many words of encouragement and taking a personal interest in my studies (in what was often an all too impersonal environment), and Prof. David Meyer and Prof. Doug Costa who taught me the basics when I was a beginning graduate student at the University of Virginia.

I have been fortunate to work with many talented, interesting, and humorous colleagues, all of whom I consider to be my friends. I would like to thank Dr. George Kantor, Dr. Ram Venkataraman, Dr. Eric Justh, Dr. Vikram Manikonda, Dr. Herbert Struemper, Mr. Jinqiu Shao, Mr. Babak Azimi Sadjadi, Mr. Tharmarajah Kugaraja, and Mr. Vadim Polyakov for teaching me about all sorts of subjects and contributing in many ways to my experience here. I offer special thanks to George for being a great lab co-manager, housemate, and friend.

I would also like to thank the administration and staff of the ISR and the ECE Department who assisted me in a wide variety of tasks. Special thanks are due to Ms. Pamela White, Mr. Prasad Dharma, Ms. Kathy Penn, and Mr. Amar Vadlamudi.

This research was supported in part by a grant from the National Science Foundation's Engineering Research Centers Program: NSFD CDR 8803012 and by the NSF Grant EEC-952757, and by the Army Research Office under the ODDR&E MURI97 Program Grant No. DAAG55-97-1-0114 to the Center for Dynamics and Control of Smart Structures (through Harvard University), and by the Electronic Sensors and Systems Sector of Northrop Grumman.

It has been my fabulous *mazel* to be surrounded by a wonderful family during this endeavor. I have been blessed to receive the constant love, support, encouragement, and good humor of my wife Cindy, my parents Malcolm and Lilly, my sister Kathryn, my grandmother Lillian and late grandparents Max, Margit, and Julius, my new family members Merle, Richard, and Jennifer, and my siblings in spirit Bruce and Shari. This accomplishment is theirs to share with me. I am especially thankful to Mom and Dad for being the best parents in every way, and to Cindy and Kathryn for inspiring me with their courage, conviction, determination, unique talents, vision, and resourcefulness. My admiration for them constantly grows.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

This thesis consists of several somewhat distinct but connected parts, with an underlying motivation in problems pertaining to control and optimization of semiconductor processing. The first part (Chapters 3 and 4) addresses problems in model reduction for nonlinear state-space control systems. The problems are motivated by a thorough discussion and analysis of prominent state-of-the-art approaches. We then offer solutions via methods, tools, and algorithms for computation of balanced realizations, both in general and motivated by specific applications. The second part (Chapter 5) deals with modeling of rapid thermal chemical vapor deposition (RTCVD) for growth of silicon thin films, via first-principles and empirical analysis. We present detailed process-equipment models and study the factors that influence deposition uniformity through analysis of experimental and simulation results. In the third part (Chapter 6) we continue the modeling effort, specializing to a control system for RTCVD heat transfer. We then develop and apply ad-hoc versions of prominent model reduction approaches to derive reduced models and perform a comparative study.

In this introductory chapter we provide background material, an overview of the

scope and contributions of this thesis, and a guide to its organization by chapters.

1.1 Background

The modeling of complex dynamical systems is one of the most important subjects in science and engineering. For control engineers, the subject is crucial, since control law design requires the formulation of a suitable mathematical model for the system of interest. For many systems, the underlying physics is known and physics-based models exist whose predictive capability has been demonstrated experimentally and is well established. For example, the Navier-Stokes equations (see, e.g., [61]) together with appropriate boundary conditions and initial conditions provide a reliable mathematical description for the flow of a Newtonian fluid.

It is often the case that a model is too complicated to be useful for its intended application. Highly complex models cause difficulties in controller synthesis and may place excessive computational burdens on software and hardware used for simulation and control. For example, the difficulties involved with the design of control algorithms using a nonlinear partial differential equation (PDE) model such as Navier-Stokes are well known (e.g., [70, 99]). One remedy is to make approximations to the model, based on physical considerations and mathematical analysis, in order to derive a simpler model from the original complicated one. This is what is generally referred to as model reduction.

The distinction between modeling and model reduction is blurry and varies among different authors. Verriest [160] defines modeling as the process whereby an abstract mathematical model is matched to the physical reality, and model reduction the process whereby a simpler mathematical model is derived from an existing mathematical model. This notion is intuitive but we mention the following

exceptions. We interpret simplifications based solely on physical considerations, such as eliminating terms describing conductive effects in a heat transfer problem where radiative effects are dominant, as falling within the realm of modeling, and take the resulting simpler (but still complicated) model as the new starting point for model reduction. Also, there are situations in control engineering and physics in which the model contains redundancies that can be eliminated through mathematical analysis, yielding simplified models which make exactly the same, rather than approximate, predictions as the original. Examples include non-minimal linear state-space models (see e.g., [72]) and certain conservative systems with symmetries (e.g., [100]). Again, we take the model with redundancies already eliminated as the starting point for reduction.

1.1.1 Model Reduction

Two basic attributes of a mathematical model are its fidelity and complexity. Generally speaking, the fidelity of a model, also called correctness, refers to its capability to predict the behavior of the system being modeled. It can also be thought of as the degree to which characteristics of the physical system are reflected by the model. The complexity of a model is given, roughly, by the number of unknowns that must be determined in order to characterize the system behavior. There is a natural trade-off between these two attributes. Approximations resulting in a complexity reduction necessarily degrade fidelity and vice-versa (otherwise, as stated earlier, the original model is an unacceptable starting point).

The advantages of low complexity are clear. It allows for an easier understanding of model dynamics and simplification of controller synthesis. The reduced computational burden of low complexity models leads to faster and easier computer

simulation, faster control algorithms, and more reliable controller implementations (whether in hardware or software) since there are fewer sources of potential failure. Again referring to the control of fluid flows, initial successes have recently been shown toward using low complexity models, derived from Navier-Stokes, in the development of control algorithms for the wall region of a turbulent boundary layer [99].

The model reduction problem, then, is one of finding a systematic methodology within a given mathematical framework to produce an efficient or optimal trade-off of fidelity versus complexity. By efficient and optimal, respectively, we mean relatively small and the smallest possible degradation in fidelity for a given complexity reduction. Thus, the procedure should quantify the effect of a given approximation on fidelity in some meaningful way. Guaranteed error bounds are desirable. It is also useful for the procedure to guarantee the preservation of properties such as open-loop and closed-loop stability. It is then up to the designer to choose the degree of reduction based on considerations for the particular application of interest.

This thesis deals with model reduction within the framework of continuous-time state-space control systems. By control system we mean a dynamical system with exogenous inputs (e.g., controls, disturbances) and outputs (e.g., measurements, variables of interest). The dimension of a state-space model, also known as the model order, is the number, possibly infinite, of independent variables needed to characterize the "state" of the system, which, roughly speaking, represents the memory that the system has of its past. These variables are called state variables, or state components, and an ordered collection of all of them is called the system state. The set of allowable values for the state is called the state-space, also

known as the phase-space. Definitions and a mathematical set-up are presented in Section 2.1.

If there are a finite number of state variables, then we call the system finite-dimensional. Otherwise, by convention the system order is set to infinity and we call the system infinite-dimensional, also known as a distributed parameter system. In the state-space context, complexity is equivalent to model order. Thus, it is clear that model reduction is essential in the infinite-dimensional setting. The original (physics-based) model is called the full-order model, while approximations are called reduced-order models.

One measure of fidelity, i.e., the quality of approximation, is given by

$$\sup_{u \in \mathcal{U}} \quad \frac{\|y - y_r\|}{\|u\|} \tag{1.1}$$

where y represents the full-order system output, y_r represents the reduced-order system output, u represents the input belonging to the admissible class $\mathcal{U}$, and $\|\cdot\|$ denotes an appropriate norm. This amounts to measuring the worst-case error between the outputs of the original and reduced models over all admissible input signals. Other measures of fidelity can also be used depending upon the situation.

The general methodology for state-space model reduction involves coordinate transformation followed by component truncation. The procedure is illustrated in Figure 1.1. The state can be expressed in terms of coordinates, i.e., as a linear combination of basis elements for the state-space. For reduction we find a coordinate system in which each state component is ranked according to its contribution, or importance, to the relevant (e.g., input-to-output) system behavior. Then, the system evolution equation is expressed in terms of the new coordinates, and state components with relatively little importance are deleted from the model. Integration of the reduced evolution equation gives the trajectory of the reduced-order

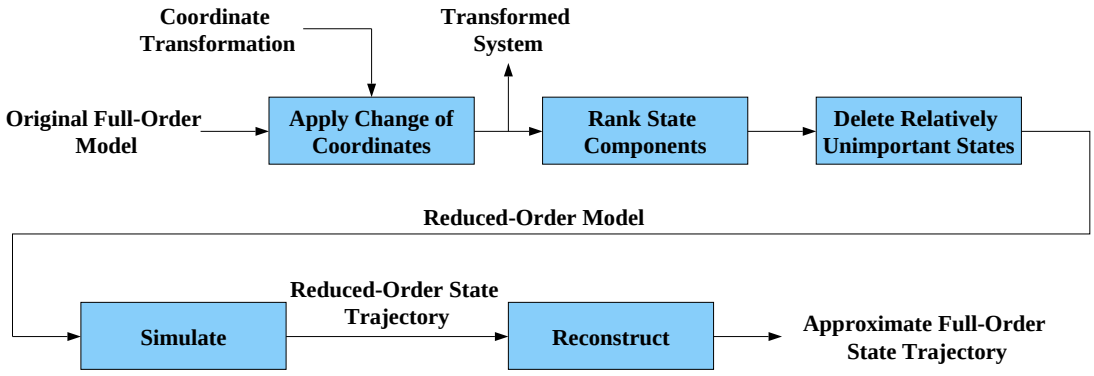


Figure 1.1: General state-space model reduction methodology.

state. Finally, an approximation to the original full-order state is reconstructed from the reduced-order state. The choice of coordinate transformation is what generally distinguishes different methods and is the key to achieving efficiency.

In earlier practice, model approximation has been largely based on heuristics and ad-hoc trial-and-error methods. However, over the past few decades, model reduction on a solid mathematical basis has been the subject of extensive research from a broad range of viewpoints and over a large number of application areas. This research has resulted in a variety of reduction tools.

One tool that has found much recent application in simplification of models for fluid flow, especially in the area of turbulence, is the proper orthogonal decomposition (POD) (see, e.g., [15, 67, 96, 147]), also known as the Karhunen-Loeve expansion (see, e.g., [126, 166]) of a second-order stochastic process. The POD can be described, roughly, as a procedure for extracting a basis for an orthogonal decomposition of the state-space from an ensemble of signals. In the context of model reduction for dynamical systems, the ensemble must capture, or represent, the relevant system behavior. The procedure is attractive for several reasons. It is applicable to linear and nonlinear models, and to models of finite and infinite dimension. It provides a meaningful ranking of state components from an energy

contribution viewpoint, and enjoys properties such as optimality (see e.g., [15]) in the sense of data compression and error minimization. The POD is a tool of major importance to this thesis and is described in detail in Section 3.2.

The POD has been independently rediscovered and analyzed from different points of view several times since the 1940's (see [15] for a brief history). It has been applied in a variety of areas including image processing (e.g., [146]), modeling and control of chemical processes (e.g., [25, 52]), and turbulence modeling (e.g., [11, 12, 149]). While the properties of the POD are well known and useful from the point of view of reducing the dimension of a single ordinary or partial differential equation model, the control viewpoint introduces new issues. By letting controls take values in a suitable function space, a family of ordinary or partial differential equations is obtained. From the empirical perspective, it becomes unclear how to generate a representative data ensemble, since the system response depends strongly on the chosen input signal. The ranking of states, and properties such as optimality, lose their precise meaning. Furthermore, the relationship between states and outputs is ignored in determining the POD basis.

Nevertheless, during the 1990's, the POD has been prominent as a tool for model reduction of state-space control systems, particularly in the area of temperature control for rapid thermal processing (RTP) (e.g., [1, 5, 6, 13, 120, 157]), a process used for several functions involved in manufacturing semiconductor devices (see Section 1.1.2). In control system applications of POD, mathematical rigor is replaced by ad-hoc procedures. For example, in [13] the authors generate several reduced models for an RTP system, each corresponding to a different operating point, and switch among them via heuristic rules according to the state trajectory. Thus, the method can be effective as part of an overall ad-hoc procedure, but not

necessarily satisfying from a control-theoretic point of view.

The basis elements generated by the POD procedure are often referred to as principal components (see Section 2.3) or empirically determined eigenfunctions, because they are extracted from an empirically generated ensemble of signals. There are other model reduction procedures that use a fixed, rather than empirical, basis for an orthogonal decomposition of the state-space. For example, wavelet bases (see, e.g., [35]) have recently been used in model approximation for the control of heat diffusion and vibration damping in a visco-thermoelastic rod [20]. However, these methods do not take advantage of existing physical or empirical knowledge of the system in choosing the basis. Consequently, there is little that can be said about their effectiveness in general situations.

Considerable work has been devoted to model reduction for finite-dimensional linear time-invariant (LTI) control systems, dating back to the 1960s (see [53] for a complete list of references through 1976). These efforts generally fall into the categories of polynomial approximations in the frequency-domain, state-space transformation and component truncation in the time-domain, and parametric optimization techniques (see [48] for a complete overview). Some of these are motivated by and designed for a particular application. For example, modal analysis (see, e.g., [103]) is mainly used as a tool for reducing the complexity of linear lightly damped mechanical systems (e.g., [26]).

An LTI state-space method of general importance and applicability is balancing, introduced by Moore [109] in 1981 for stable, minimal, finite-dimensional LTI systems. In this method, a system is transformed to balanced form, which means that it is "equally controllable and observable." The states of a balanced realization can be ranked according to their influence on the input-to-output behavior of

the system, as measured by its input-to-output gain, or Hankel norm. For LTI systems, balancing is strongly related to the POD, in the sense that the basis for the balancing coordinate transformation can be derived using principal components generated via injection of impulsive inputs. We elaborate on balancing for LTI systems in Section 3.3.

During the 1980s and 1990s, various versions of balancing and other Hankel-norm based methods (e.g., [34, 54, 113]) were developed for finite- and infinite-dimensional LTI systems. Concurrently, there was a substantial effort toward development of algorithms and computational tools (e.g., [91, 136]) for practical implementation of linear balancing, resulting in its wide application to produce low-order models for LTI control systems. The main objects of importance in linear balancing are the controllability and observability Gramian matrices. Thus, the computational tools are based mainly on well known and efficient algorithms for matrix algebra problems.

Scherpen [140, 141] extended the balancing approach, introducing in 1993 a general theory and procedure of balancing for a class of stable, affine, smooth, finite-dimensional nonlinear systems. The main objects of importance in nonlinear balancing are the controllability and observability energy functions. The balancing coordinate transformation is local to a neighborhood of the origin and determined by application of the Morse-Palais lemma, which gives a canonical form for functions in the neighborhood of a non-degenerate critical point. Scherpen's nonlinear balancing procedure forms a major part of the foundation for this thesis, and is detailed throughout Chapter 4.

In contrast to the linear case, the nonlinear balancing procedure is not immediately amenable to computational implementation. For example, the controllability

energy function corresponds to the value function for a nonlinear optimal control problem. Also, the Morse-Palais lemma guarantees the existence of a transformation to a canonical form for the controllability energy function, but provides no constructive procedure for obtaining it. Thus, tools have not yet appeared for computing balanced realizations for nonlinear systems, and the procedure has not yet been applied as a tool for model reduction.

In this thesis we offer methods for computing balanced realizations for nonlinear systems. We rely heavily on the theory of stochastically excited dynamical systems, i.e., control systems with Gaussian white noise injected at the input terminals. The state of such a system is a stochastic process with an associated density function. The evolution of the state is governed by a stochastic differential equation, and the evolution of the density function is governed by a pair of hypoelliptic diffusion equations. In the case of a linear system, the covariance matrix of the steady-state density is equal to the controllability Gramian matrix of the corresponding deterministic system, a fact which motivates useful generalizations to the nonlinear setting. Mathematical preliminaries for dealing with stochastically excited systems are presented in Section 2.6.

The model reduction tools that we develop have general applicability but become impractical for systems of sufficiently high dimension. However, for certain specific types of systems, we can obtain computable results, e.g., an exact formula for the controllability function. We study the class of second-order mechanical systems characterized by a Hamiltonian (conservative) structure and perturbed by dissipation and forcing. Under certain conditions, steady-state densities can be obtained for these systems from which the controllability energy function can be derived.

1.1.2 Modeling: Rapid Thermal CVD for Silicon Growth

In recent years, the semiconductor industry has begun to employ mathematical and computational models as tools to aid in equipment design (e.g., [29, 49, 155]), simulation (e.g., [84, 98]), process optimization (e.g., [135]), and model-based process control (e.g., [30, 139, 63]). These modeling efforts have been motivated primarily by gains in manufacturing cost effectiveness. In particular, the design of new equipment and manufacturing processes is typically performed via costly trial-and-error procedures. Modeling and simulation are used to reduce the number of required experiments, thus reducing the associated cost and development cycle time. Furthermore, as device sizes shrink, and as wafer sizes grow, process control becomes more challenging and specifications become tighter. In many cases, model-based process control is needed to accomplish the task.

One semiconductor manufacturing process that has been the subject of much recent modeling activity from various points of view is chemical vapor deposition (CVD) (see e.g., [75, 133, 144]), a common way of depositing thin layers of conducting and insulating films over the surface of a semiconductor wafer. In CVD, the material is deposited from a gaseous precursor to the substrate via chemical reactions that are activated by heat energy. In this thesis, we are concerned with dynamic and steady-state models of CVD for growth of silicon thin films on a silicon wafer. In particular, we focus on rapid thermal CVD, i.e., CVD processes that employ RTP technology for wafer heating. The process is illustrated in Figure 1.2.

RTP (see e.g., [18, 132]) is a technology for rapidly heating and cooling a single semiconductor wafer, allowing manufacturing processes to achieve high temperatures for short (e.g., 5 seconds), well-controlled periods of time. The wafer is usually heated by energy radiated from specially designed high-power lamps. A tempo-

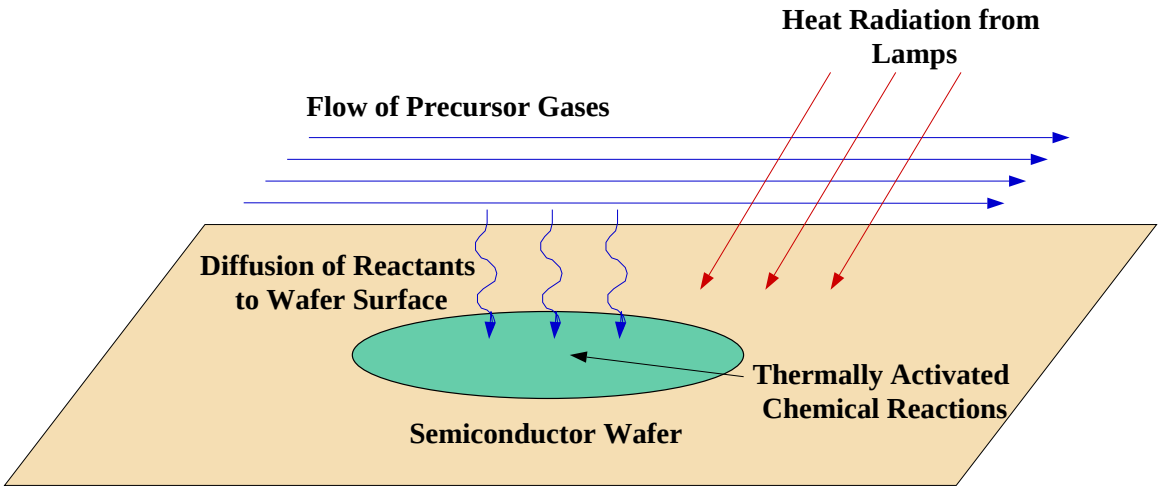


Figure 1.2: Simplified illustration of rapid thermal CVD.

rally varying temperature profile is programmed to achieve the desired processing step. RTP technology is versatile; its capabilities have been used in several wafer processing functions including annealing, CVD, oxidation, nitridation, and contact sintering. Several different designs exist for RTCVD equipment (e.g., [27, 80, 155]).

Precise control of the temperature distribution across the wafer surface during RTP is critical to achieving a uniform film thickness across the wafer surface, ensuring reproducibility from wafer to wafer, and minimizing thermal stress on the wafer during processing. Even small temperature variations (non-uniformities) can cause large thickness variations, resulting in costly reductions in process yield. Various control strategies have been tried recently, including Run-to-Run (RtR) control (e.g., [68, 86, 168]) and model-based feedback control (e.g., [63, 62, 39]).

The modeling and analysis of CVD equipment and processes presented in this thesis are mainly the result of a joint project [116, 117, 115] between the Institute for Systems Research (ISR) of the University of Maryland, College Park, and the Electronic Sensors and Systems Sector of Northrop Grumman Corporation (NG-ESSS), Baltimore, MD, undertaken during 1997 and 1998. The overall objective

of the project was to improve manufacturing effectiveness for epitaxial growth of silicon and silicon-germanium (Si-Ge) thin films on a silicon wafer. Epitaxial growth, or epitaxy, (see, e.g., [134, 162]) refers to the deposition of a thin layer of material onto the surface of a single-crystal substrate in such a manner that the layer is also single-crystal and has a fixed and predetermined crystallographic orientation with respect to the substrate.

The equipment used at NG-ESSS to deposit the thin films (and currently a production tool in use for various processes) was the Epsilon-1 RTCVD reactor, manufactured by ASM America, Phoenix, AZ. NG-ESSS uses the Epsilon-1 to deposit both poly-crystalline and epitaxial (single-crystal) layers of silicon, depending on the application. Silicon epitaxy provides flexibility for a device designer to tailor or optimize the device performance by allowing for greater control of doping concentration and profile in deposited layers. Si-Ge films are always epitaxial. Chapter 5 contains details regarding the CVD equipment and processes involved in our modeling effort.

One important modeling objective is the prediction of deposition rates and thickness uniformity given operating conditions such as flow rates of process gases, wafer temperature, and chamber pressure. Therefore, the process-equipment model for silicon growth in the Epsilon-1 describes the flow of process gases through the chamber, heat transfer in the gas phase and within and among the various solids in the chamber including the wafer, the transport of chemical species within a multicomponent gas, and chemical mechanisms for gas phase and surface reactions. It takes the form of a set of coupled nonlinear PDEs and associated boundary conditions together with chemical kinetics equations. Moreover, we note that due to the asymmetrical design of the Epsilon-1 deposition chamber and heat lamp ap-

paratus, models incorporating three spatial dimensions are required to sufficiently describe the various phenomena (approximations incorporating one or two spatial dimensions via symmetries are useful in certain situations). Discretization (e.g., finite-volumes, finite-elements) at a suitable resolution results in a model with thousands of states.

Due to the scale and scope of the overall process-equipment model, it is advantageous to develop a separate model that focuses specifically on heat transfer among the various solid surfaces in the RTP chamber including the semiconductor wafer. Such models are often used in model-based control strategies for achieving temperature uniformity across the wafer surface. However, as we demonstrate in Chapter 5, temperature uniformity does not necessarily ensure deposition thickness uniformity.

Nevertheless, in Chapter 6, we develop such a RTP heat transfer model pertaining specifically to the Epsilon-1. Physics-based models for RTP heat transfer give a distributed parameter control system with a radiative (4-th power) nonlinearity in the governing equations and boundary conditions. Low-order approximations to the model are desired in order to facilitate control law design, model-based feedback control implementation, and computer simulation. We derive low-order models for the RTP heat transfer control system using the reduction approaches described in this thesis, and compare their relative merits and drawbacks.

1.2 Scope and Contributions

We list here the main contributions of this thesis.

- We develop useful methods, tools, and algorithms to compute the energy functions and coordinate transformations involved in the Scherpen theory and procedure for nonlinear balancing. We apply our approach to derive, for the first time, balanced representations of nonlinear state-space models.

- We offer a new method involving stochastic excitation for approximating the controllability energy function of a nonlinear system.

- We determine conditions under which an exact formula can be written for the controllability energy function of a nonlinear Hamiltonian system perturbed by dissipation and forcing. We apply our result to provide an expression for the controllability function of a 4-state nonlinear mechanical system.

- We present an algorithm for a numerical implementation of the Morse-Palais lemma, i.e., computation of a local coordinate transformation under which a real-valued function with a non-degenerate critical point is quadratic on a neighborhood of the critical point.

- We develop a collection of programs and utilities in a standard programming language to facilitate the practical application of our methods and algorithms.

- We develop a high-fidelity process-equipment model for deposition of silicon thin films in a commercial rapid thermal CVD reactor. The model allows us to simulate growth experiments under a broad range of process conditions while taking account of the various physical and chemical phenomena involved in CVD of silicon from a multi-component gas.

- We investigate the factors that influence deposition rate and uniformity in a commercial reactor including the effects of temperature, precursor gas flow rates, and chamber pressure. We determine relationships between the various factors and growth rate that can be used to predict their effects in a particular situation.

- We demonstrate through anecdotal evidence, simulation results, and experimental data that achieving deposition thickness uniformity requires a certain degree of temperature non-uniformity across the wafer surface.

- We apply view factor methods to develop a radiative heat transfer model for the heating of a semiconductor wafer via tungsten-halogen lamps in a commercial RTP reactor. The model incorporates a non-symmetric 3-dimensional chamber and lamp array geometry, a feature not commonly found in the literature. Furthermore, the model is partially validated through an ad-hoc experimental procedure.

- We formulate ad-hoc procedures for applying standard reduction methodologies to physics-based models for RTP heat transfer. We apply the procedures to a high-order control system model for heat transfer in a commercial RTP chamber to derive low-order model approximations that faithfully reproduce the relevant input-to-output behavior of the original model.

- We provide a guide to the use of standard and ad-hoc model reduction approaches that does not sacrifice rigor while serving as a practical tool that emphasizes computational issues and potential hazards. In the process we illuminate important connections between prominent methods.

We list here some contributions of this thesis that either support the main body of work or are necessary for completeness of the exposition but are not of primary interest or importance.

- We present a new proof of a theorem by Scherpen that appeals to the connections between the result and optimal control theory.
- We provide new and more general conditions on the output map of a nonlinear system such that the observability energy function exists, i.e., is finite.
- We bring to light and illustrate through examples issues of non-uniqueness regarding the Morse coordinate transformation and balancing transformation.
- We demonstrate that the consumption of process gases in the Epsilon-1 RTCVD reactor can be reduced by decreasing the purge gas flow.

1.3 Thesis Outline

The material presented in this thesis is reasonably self contained. It is organized by chapters as follows.

Chapter 2 We provide the mathematical preliminaries necessary for working with the main topics of this thesis.

Chapter 3 We introduce two prominent model reduction approaches: POD and balanced truncation. We describe the current state-of-the-art, including the underlying theory, computational issues, advantages and shortcomings, and selected applications. The material in this chapter motivates the research

in Chapter 4 and explains the methods and computational tools used in Chapter 6.

Chapter 4 We address the problem of computability pertaining to the Scherpen theory and procedure for balancing of nonlinear systems. We offer methods and algorithms toward balancing stable affine nonlinear control systems, with some emphasis on computation of the controllability energy function and the Morse coordinate transformation of a function around a non-degenerate critical point.

Chapter 5 We develop high-fidelity physical-chemical models for predicting the behavior and output of a commercial RTCVD reactor used for depositing thin films of Si and Si-Ge on silicon wafers in a manufacturing environment. We present the results of simulations and growth experiments and use them to study the factors that influence deposition rate and uniformity in the reactor.

Chapter 6 We formulate a physical model describing heat transfer in a commercial RTCVD reactor. We then derive low-order models for the resulting RTP heat transfer control system, using ad-hoc versions of methods described in Chapter 3.

Chapter 7 We present concluding remarks and comments on future research opportunities.

Appendices The Appendices contain supporting material that is essential for completeness but that would be disruptive within the main exposition.

Chapter 2

Preliminaries

This thesis makes use of tools, and draws concepts and ideas, from several different areas of science and mathematics. Here we collect the basic definitions and results so that they may be used without any detailed explanation later in the thesis. Topics are covered in additional depth in the listed references.

2.1 State-Space Control Systems

This thesis deals with model reduction of continuous-time state-space control systems. We focus on methods and algorithms for reduction of finite-dimensional models. The mathematical framework for these models is that of ordinary differential equations (ODEs) for the state evolving on a smooth manifold and represented in terms of a local coordinate system. The necessary machinery for working with manifolds and local coordinates is set up in Appendix B. The reduction approaches that we consider require some elements of system theory (continuous-time finite-dimensional) which appear in Section 2.2.

Some of the modeling and reduction methods are directly applicable in the

infinite-dimensional setting. However, for purposes of this thesis, we generally assume that an infinite-dimensional model has been suitably discretized and work with the finite-dimensional approximation.

In the finite-dimensional case, the state-space control system model is given by a pair of equations, the state equation and the output equation, respectively, describing the evolution of the state on a smooth manifold given a specified input, and the relationship between the state and output. Here we present these equations in their most general form, followed by some important specializations. An important component of the state-space model reduction procedure is coordinate transformation. For this reason, we describe what happens to the evolution equations under diffeomorphic change of coordinates.

The material contained in this section is standard. Our treatment is based on texts by Khalil [79], Nijmeijer and van der Schaft [121], and Isidori [69]. For proofs we refer to the literature.

System Equations

The state-space is assumed to be an n-dimensional smooth manifold M. The finite-dimensional control system evolving on M is given by the equations

$$\dot{x}(t) = f(t, x(t), u(t)) \tag{2.1}$$

$$y(t) = h(t, x(t), u(t)) \tag{2.2}$$

where $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ denotes local coordinates for the *state*, $u = (u_1, \ldots, u_m) \in U \subset \mathbb{R}^m$ denotes the *input (control)*, and $y = (y_1, \ldots, y_p) \in \mathbb{R}^p$ denotes the *output*. The maps f and h are to be interpreted as their respective corresponding local representatives. The map $f : \mathbb{R}^+ \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ is called the *system map*. The map $h : \mathbb{R}^+ \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^p$ is called the *output map*.

We often make assumptions about the regularity of f and h, e.g., that they are of class C^k or possibly smooth. The ordinary differential equation (2.1) is called the *state equation* and governs the time evolution of the state given a specified input and initial state. Equation (2.2) is called the *output equation*. These equations are written in vector notation, shorthand for

$$\begin{aligned} \dot{x}_1(t) &= f_1(t, x_1(t), \ldots, x_n(t), u_1(t), \ldots, u_m(t)) \\ &\vdots \qquad\quad \vdots \\ \dot{x}_n(t) &= f_n(t, x_1(t), \ldots, x_n(t), u_1(t), \ldots, u_m(t)) \end{aligned} \tag{2.3}$$

and

$$\begin{aligned} y_1(t) &= h_1(t, x_1(t), \ldots, x_n(t), u_1(t), \ldots, u_m(t)) \\ &\vdots \qquad\quad \vdots \\ y_p(t) &= h_p(t, x_1(t), \ldots, x_n(t), u_1(t), \ldots, u_m(t)) \end{aligned} \tag{2.4}$$

Each input signal u belongs to the class $\mathcal{U}$ of *admissible controls*, which we take as the set

$$\mathcal{U} = \left\{ u : \mathbb{R}^+ \to U \subset \mathbb{R}^m : u \in C^\infty \right\} \tag{2.5}$$

or sometimes more generally

$$\mathcal{U} = \left\{ u : \mathbb{R}^+ \to U \subset \mathbb{R}^m : u \text{ is piecewise continuous from the right} \right\} \tag{2.6}$$

Given a specified input $\bar{u}$ and initial state x_0, the time evolution of the state is given by the initial value problem

$$\dot{x}(t) = \bar{f}(t, x(t)) \qquad x(t_0) = x_0 \tag{2.7}$$

where

$$\bar{f}\left(t, x(t)\right) \triangleq f\left(t, x(t), \bar{u}(t)\right)$$

In order for (2.7) to predict the state trajectory, it must have a unique solution. This can be guaranteed under conditions given by the following result.

Theorem 2.1.1 (Local Existence and Uniqueness) *Let $\bar{f}(t,x)$ be piecewise continuous in t and satisfy the Lipschitz condition*

$$\left\| \bar{f}(t,x) - \bar{f}(t,y) \right\| \leq L \, \| x - y \| \tag{2.8}$$

for all $x, y \in B = \{x \in \mathbb{R}^n : \| x - x_0 \| \leq r\}$ and for all $t \in [t_0, t_1]$. Then, there exists some $\delta > 0$ such that the initial value problem (2.7) has a unique solution over $[t_0, t_0 + \delta]$. □

Remark 2.1.2 *The unique solution of (2.7) on $[t_0, t_0 + \delta]$, if it exists, is given by*

$$x(t) = x_0 + \int_{t_0}^{t} \bar{f}(s, x(s)) \, ds \qquad t \in [t_0, t_0 + \delta] \tag{2.9}$$

It is referred to as the state trajectory *and sometimes denoted $x(t, x_0, t_0, \bar{u})$ to explicitly indicate the initial state, initial time, and specified input. The corresponding output $y(t)$ is referred to as the* output trajectory. □

Remark 2.1.3 *The Lipschitz property is weaker than continuous differentiability. In this thesis, we usually work with f and u that are smooth in their respective arguments. Thus, we generally assume local existence and uniqueness of solutions for the systems that we consider, unless specified otherwise.* □

A point x for which $f(\cdot, x, 0) \equiv 0$ is called an *equilibrium*. If a system has an equilibrium, then, without loss of generality, we assume that it is at $x = 0$, i.e., $f(\cdot, 0, 0) = 0$, unless otherwise specified. We also assume that $h(\cdot, 0, 0) = 0$ so that the output is zero whenever the state is zero.

We focus our attention on state-space models for which the functions f and h do not depend explicitly on t, i.e.,

$$\dot{x}(t) = f(x(t), u(t)) \tag{2.10}$$

$$y(t) = h(x(t), u(t)) \tag{2.11}$$

With a specified input u, these models are referred to as *autonomous* or *time-invariant*.

A particular class of systems that we consider in this thesis is the class of autonomous *affine systems*

$$\dot{x}(t) = f(x(t)) + \sum_{i=1}^{m} g_i(x(t))\, u_i(t) \tag{2.12}$$

$$y(t) = h(x(t)) \tag{2.13}$$

in which the input enters the state equation in an affine way, and there is no direct feedthrough of the input to the output. The maps $f : \mathbb{R}^n \to \mathbb{R}^n$, $g_i : \mathbb{R}^n \to \mathbb{R}^n$, $i \in \underline{m}$, and $h : \mathbb{R}^n \to \mathbb{R}^p$ are to be interpreted as local representatives. As before, we assume, without loss of generality, the existence of an equilibrium at $x = 0$, i.e., $f(0) = 0$, as well as $h(0) = 0$.

Remark 2.1.4 *Sometimes we use the notation g_{ij} which refers to the i-th component of the j-th input function, i.e., $g_{ij} = (g_j)_i$.* □

Coordinate Transformations

Model reduction involves smooth coordinate transformations of the state-space. Let $\{e_1, \ldots, e_n\}$ denote the standard basis for $\mathbb{R}^n$, i.e., e_i is a vector with a 1 in the i-th position and a 0 in every other position. We assume that the functions f, g, and h have been formulated with respect to the standard coordinate system,

i.e., the basis for the original local coordinate system is the standard basis, so that in terms of local coordinates the state vector can be written

$$x = \sum_{i=1}^{n} x_i \, e_i$$

We now consider what happens to Equations (2.12) and (2.13) under coordinate transformation. Let U and V be subsets of $\mathbb{R}^n$ containing 0. Let $S : U \to V$ be a diffeomorphism such that $S(0) = 0$ (to preserve the equilibrium at 0). The fact that S is a diffeomorphism allows for reversing the transformation and recovering the original state, and guarantees that the system in the new coordinates is still smooth. We call S a smooth local coordinate transformation about the origin.

Under the smooth local coordinate transformation

$$z \mapsto x = S(z) \tag{2.14}$$

the control system (2.12)-(2.13) transforms to

$$\dot{z}(t) = \widehat{f}\,(z(t)) + \sum_{i=1}^{m} \widehat{g}_i\,(z(t))\; u_i(t) \tag{2.15}$$

$$y(t) = \widehat{h}\,(z(t)) \tag{2.16}$$

where

$$\begin{aligned} \widehat{f}(z) &= [DS(z)]^{-1} f\,(S(z)) \\ \widehat{g}_i(z) &= [DS(z)]^{-1} g_i\,(S(z)) \qquad i \in \underline{m} \\ \widehat{h}(z) &= h\,(S(z)) \end{aligned}$$

Linear Time-Invariant Systems

It will be useful on occasion to consider the special case of a linear time-invariant (LTI) system. The LTI specialization of (2.12)-(2.13) takes the form

$$\dot{x} = Ax + Bu \tag{2.17}$$

$$y \;=\; Cx \tag{2.18}$$

where A is $n \times n$, B is $n \times m$, and C is $p \times n$.

In the LTI case, the state-space manifold M is equal to $\mathbb{R}^n$. The unique global solution of (2.17) with initial state $x(0) = x_0$ is given by the *variation of constants* formula

$$x\,(t) = \exp\,(A\,t)\;x_0 + \int_0^t \exp\,(A\,(t-s))\;B\,u(s)\,ds \tag{2.19}$$

A coordinate transformation is global and linear, represented by an invertible *transformation matrix*. Let S be the transformation matrix. Under the linear change of coordinates

$$z \mapsto x = S\,z \tag{2.20}$$

the LTI control system (2.17)-(2.18) transforms to

$$\dot{z} \;=\; \widehat{A}\,z + \widehat{B}u \tag{2.21}$$

$$y \;=\; \widehat{C}\,z \tag{2.22}$$

where

$$\widehat{A} = S^{-1}\,AS\,, \quad \widehat{B} = S^{-1}\,B\,, \quad \widehat{C} = C\,S \tag{2.23}$$

2.2 Some Elements of System Theory

In this thesis certain key elements of systems theory appear frequently. The notions of stability, controllability, and observability of a control system are essential. They are used in their most general form as well as specializations for particular situations. For example, in Section 4.6 we will need to show that certain example systems are locally accessible, locally observable, and asymptotically stable.

The material contained in this section is drawn mainly from texts by Khalil [79], Nijmeijer and van der Schaft [121], Vidyasagar [161], and class notes in Geometric Control presented by Dayawansa [37] at the University of Maryland. For proofs we refer to the literature.

Stability

Here we present some standard definitions and results on the local stability of a time-invariant system without inputs, i.e.,

$$\dot{x}(t) = f(x(t)) \tag{2.24}$$

where $f : D \subset \mathbb{R}^n \to \mathbb{R}^n$ is locally Lipschitz so that there exists a unique solution on an interval $[0, \delta]$.

We are concerned with the stability of equilibrium points. Without loss of generality we can assume that the system has an equilibrium at 0, i.e., $f(0) = 0$.

Definition 2.2.1 (Stability of Equilibrium) *The equilibrium point $x = 0$ of system (2.24) is said to be* stable *if for any neighborhood U of 0 there exits a neighborhood V of 0 such that if $x(0) \in V$ then the solution $x(t, 0, x(0))$ belongs to U for all $t \geq 0$.* □

Remark 2.2.2 *The equilibrium point $x = 0$ of (2.24) is said to be* unstable *if it is not stable.* □

Definition 2.2.3 (Asymptotic Stability of Equilibrium) *The equilibrium point $x = 0$ of (2.24) is said to be* asymptotically stable *if it is stable and there exists a neighborhood W such that if $x(0) \in W$ then*

$$\lim_{t \to \infty} x(t, 0, x(0)) = 0 \tag{2.25}$$

□

Definition 2.2.4 (Region of Attraction) *Let $x = 0$ be asymptotically stable for the system (2.24). The* region of attraction *is defined as the set*

$$\left\{ x_0 \in \mathbb{R}^n : \lim_{t \to \infty} x(t, 0, x_0) = 0 \right\} \tag{2.26}$$

i.e., the set of points from which the trajectory approaches the origin as $t \to \infty$. □

Definition 2.2.5 (Exponential Stability of Equilibrium) *The equilibrium point $x = 0$ of (2.24) is said to be* exponentially stable *if there exist constants $k > 0$ and $\gamma > 0$, such that*

$$\| x(t) \| \leq k \, x(0) \exp(-\gamma \, t) \qquad t \geq 0 \tag{2.27}$$

□

Remark 2.2.6 *Depending on the situation, stability (asymptotic stability) of (2.24) can be verified via the direct method of Lyapunov, indirect method of Lyapunov, or the invariance principle of LaSalle. Instability can be verified via a theorem of Cetaev. Since we do not explicitly use these results in this thesis, we refer the reader to the literature.* □

Controllability

Here we present some standard definitions and results regarding the controllability and reachability of a nonlinear system, some of which pertain specifically to the affine system (2.12). In addition, we introduce the notion of asymptotic reachability, which appears in Chapter 4. See [121] for background on the notion of a Lie algebra.

Definition 2.2.7 (Controllable System) *The system (2.12) is said to be* controllable *if for any two points x_1, x_2 in M there exists a finite time T and an admissible input $u : [0, T] \to U$ such that $x(T, 0, x_1, u) = x_2$.* □

Definition 2.2.8 (Reachable Set) *The* reachable set $R^V(x_0, T)$ *from x_0 at time $T > 0$ following trajectories which remain for $t \leq T$ in the neighborhood V of x_0 is defined as the set of all points $x \in M$ for which there exists $u : [0, t] \to U$ such that $x(t, x_0, 0, u) \in V$, $t \in [0, T]$ and $x(T) = x$. We also denote*

$$R_T^V(x_0) = \cup_{\tau \leq T} R^V(x_0, \tau) \tag{2.28}$$

□

Definition 2.2.9 (Locally Accessible System) *The system (2.12) is said to be*

(i) locally accessible from x_0 *if $R_T^V(x_0)$ contains a non-empty open set of M for all neighborhoods V of x_0 and all $T > 0$;*

(ii) locally accessible *if the condition in (i) holds for every $x_0 \in M$;*

(iii) locally strongly accessible from x_0 *if, for any neighborhood V of x_0, the set $R^V(x_0, T)$ contains a non-empty open set for any $T > 0$ sufficiently small;*

(iv) locally strongly accessible *if the condition in (iii) holds for every $x_0 \in M$.*

□

Definition 2.2.10 (Accessibility Algebra, Accessibility Distribution) *For the system (2.12)-(2.13), the*

(i) accessibility algebra $\mathcal{C}$ *is the smallest subalgebra of the Lie algebra of vector fields on M that contains $\{f, g_1, \ldots, g_n\}$;*

(ii) strong accessibility algebra $\mathcal{C}_0$ *is the smallest subalgebra that contains* $g_1, \ldots, g_m$ *and satisfies* $[f, X] \in \mathcal{C}_0$ *for all* $X \in C_0$*;*

(iii) accessibility distribution $C(x)$ *at* $x \in M$ *is the distribution generated by the accessibility algebra, i.e.,*

$$C(x) = \text{span}\,\{X(x) : X \text{is a vector field in } \mathcal{C}\}$$

(iv) strong accessibility distribution $C_0(x)$ *at* $x \in M$ *is the distribution generated by the strong accessibility algebra, i.e.,*

$$C_0(x) = \text{span}\,\{X(x) : X \text{is a vector field in } \mathcal{C}_0\}$$

□

Theorem 2.2.11 *For the system (2.12)-(2.13), if*

(i) $\dim(C(x_0)) = n$ *then the system is locally accessible from* x_0*;*

(ii) $\dim(C(x)) = n$ *for all* $x \in M$ *then the system is locally accessible;*

(iii) $\dim(C_0(x_0)) = n$ *then the system is locally strongly accessible from* x_0*;*

(iv) $\dim(C_0(x)) = n$ *for all* $x \in M$ *then the system is locally strongly accessible.*

□

Definition 2.2.12 (Asymptotically Reachable System) *The system (2.12) is said to be* asymptotically reachable *from* x_0 *on a neighborhood* W *of* x_0 *if for each* $x \in W$ *there exists a* $u \in \mathcal{U}$ *such that* $x(t, 0, x_0, u) \in W$ *for all* $t \geq 0$ *and*

$$\lim_{t \to \infty} x(t, 0, x_0, u) = x$$

□

Observability

Here we present some standard definitions and results regarding the observability of a nonlinear system, some of which pertain specifically to the affine system (2.12)-(2.13).

Definition 2.2.13 (Indistinguishable States) *Two states $x_1, x_2 \in M$ are said to be* V-indistinguishable *(denoted $x_1 \, I^V \, x_2$ for system (2.12)-(2.13) if for each admissible input $u : [0,t] \to U$, $T > 0$, with the property that $x(t, 0, x_1, u)$ and $x(t, 0, x_2, u)$ both remain in V for $t \leq T$, the output function $t \mapsto y(t, 0, x_1, u)$ for $t \geq 0$ and initial state $x(0) = x_1$ and the output function $t \mapsto y(t, 0, x_2, u)$ for $t \geq 0$ and initial state $x(0) = x_2$ are identical on their common domain of definition.* □

Definition 2.2.14 (Observable System) *The system (2.12)-(2.13) is said to be* observable *if $x_1 \, I^M \, x_2$ implies that $x_1 = x_2$.* □

Definition 2.2.15 (Locally Observable System) *The system (2.12)-(2.13) is said to be*

(i) locally observable at x_0 *if there exists a neighborhood W of x_0 such that for every neighborhood $V \subset W$ of x_0 the relation $x_0 \, I^V x_1$ implies that $x_1 = x_0$, i.e., indistinguishability implies equality;*

(ii) locally observable *if it is locally observable at each x_0.*

□

Definition 2.2.16 (Zero-state Observable System) *The system (2.12)-(2.13) is said to be*

(i) locally zero-state observable *if there exists a neighborhood W of 0 such that for each $x \in W$, if $y(t, 0, x, 0) = 0$ for $t \geq 0$ then $x(t, 0, x, 0) = 0$ for $t \geq 0$;*

(ii) zero-state observable *if the above holds for all* $x \in M$.

□

Definition 2.2.17 (Observation Space, Observability Codistribution)

For the system (2.12)-(2.13), the

(i) observation space $\mathcal{O}$ *is the linear space of functions on* M *containing* $h_1, \ldots, h_p$ *and all repeated Lie derivatives* $L_{X_1} L_{X_2} \cdots L_{X_k} h_j$ *for* $j \in \underline{p}$ *and with* X_i, $i = 1, 2, \ldots$ *in the set* $\{f, g_1, \ldots, g_n\}$;

(ii) observability codistribution $d\mathcal{O}$ *at* $x \in M$ *is defined by*

$$d\mathcal{O}(x) = \operatorname{span}\{dH(x) : H \in \mathcal{O}\}$$

□

Theorem 2.2.18 *For the system (2.12)-(2.13), if*

(i) $\dim(d\mathcal{O}(x_0)) = n$ *then the system is locally observable at* x_0;

(ii) $\dim(d\mathcal{O}(x)) = n$ *for all* $x \in M$ *then the system is locally observable.*

□

2.3 Principal Component Analysis

Principal component analysis (PCA) refers to a particular type of orthogonal decomposition for a matrix-valued signal $F(t)$ as described below. The signal is represented by a piecewise continuous map $F : \mathbb{R}^+ \to \mathbb{R}^{n \times m}$. We use (and adapt somewhat) the terminology presented by Moore [109].

The Gramian matrix is an object of primary interest.

Definition 2.3.1 (Gramian Matrix) *Given a piecewise continuous map* $F : [t_1, t_2] \to \mathbb{R}^{n \times m}$*, the* Gramian matrix $W^2 [t_1, t_2]$ *for* F *is defined by*

$$W^2 [t_1, t_2] = \int_{t_1}^{t_2} F(t)\ F^{\mathsf{T}}(t)\, dt \tag{2.29}$$

□

We usually deal with signals on the interval $[0, \infty)$ (infinite-time-horizon) and will use the term Gramian matrix and notation W^2 to mean $W^2[0, \infty)$ unless otherwise noted. Gramians with finite time horizons are useful in situations where the signals of interest grow unbounded, e.g., unstable systems.

The Gramian matrix W^2 is a non-negative definite matrix. Therefore, it has n non-negative real eigenvalues $\sigma_1^2 \geq \ldots \geq \sigma_n^2 \geq 0$ and n corresponding mutually orthogonal unit eigenvectors $v_1, \ldots, v_n$ (we ignore the case where W^2 has repeated eigenvalues and Jordan blocks of order 2 or higher).

The standard Fourier analysis tells us that any signal $F : [t_1, t_2] \to \mathbb{R}^{n \times m}$ can be represented as the linear combination of dyads

$$F(t) = \sum_{i=1}^{n} v_i\, a_i{}^{\mathsf{T}}(t) \tag{2.30}$$

where

$$a_i{}^{\mathsf{T}}(t) = v_i{}^{\mathsf{T}}\, F(t)\,, \quad i \in \underline{n} \tag{2.31}$$

correspond to the Fourier coefficients.

Remark 2.3.2 *PCA refers to an orthogonal decomposition (2.30) of signal* $F(t)$ *in which the basis vectors* v_i*,* $i \in \underline{n}$ *are the unit eigenvectors of* W^2*.* □

Remark 2.3.3 *Regarding (2.30), we use the following standard terminology, referring to the i-th, respectively, principal component* $v_i\, a_i{}^{\mathsf{T}}(t)$*, component vector* v_i*, component magnitude* σ_i*, and component function vector* $a_i(t)$*.* □

The PCA enjoys some useful properties.

Proposition 2.3.4 (Moore [109]) *For $F : \mathbb{R}^+ \to \mathbb{R}^{n\times m}$ with PCA given by (2.30) the following relationships hold:*

$$\int_{t_1}^{t_2} a_i^{\mathsf{T}}(t)\, a_j(t)\, dt = 0\,, \quad i \neq j \tag{2.32}$$

$$\int_{t_1}^{t_2} \| a_i(t) \|^2\, dt = \sigma_i^2 \tag{2.33}$$

$$\int_{t_1}^{t_2} \| F(t) \|_F^2\, dt = \sum_{i=1}^{n} \sigma_i^2 \tag{2.34}$$

where $\| \cdot \|_F$ denotes the Frobenius norm. □

The efficiency of the PCA as an orthogonal decomposition is due to the following result.

Proposition 2.3.5 (Moore [109]) *Let S_F denote the space*

$$S_F = \{v : v \in \mathrm{Im}(F(t)),\, t \in [t_1, t_2]\}$$

Let k be a fixed integer, $1 \leq k \leq n$. Over the class of piecewise continuous $F_A(t)$ satisfying $\dim(S_{F_A}) = k$, the residuals

$$J_F = \int_{t_1}^{t_2} \| F(t) - F_A(t) \|_F^2\, dt \tag{2.35}$$

$$J_S = \max_{\| v \|=1} \int_{t_1}^{t_2} \left\| v^{\mathsf{T}}\, (F(t) - F_A(t)) \right\|^2 dt \tag{2.36}$$

are minimized by

$$F_A(t) = F_k(t) \stackrel{\triangle}{=} \sum_{i=1}^{k} v_i\, a_i^{\mathsf{T}}(t) \tag{2.37}$$

with error residuals

$$J_F = \sum_{i=k+1}^{n} \sigma_i^2 \tag{2.38}$$

$$J_S = \sigma_{k+1}^2 \tag{2.39}$$

□

Remark 2.3.6 *The set S_F is spanned by those component vectors of W^2 corresponding to non-zero component magnitudes.* □

Remark 2.3.7 *Proposition 2.3.5 says that the most efficient k-th order approximation to F is given by the PCA.* □

2.4 Hilbert Spaces

The notion of a Hilbert space appears prominently in this thesis, particularly in the context of second-order stochastic processes and stochastically excited systems. We work with several examples of Hilbert spaces and frequently use the concepts of orthogonality, basis, and separability. The material contained in this section is standard. It is drawn mainly from texts by Akhiezer and Glazman [3] and Gohberg and Goldberg [56]. We refer to the literature for the proofs.

Definition 2.4.1 (Hilbert Space) *A* Hilbert space $\mathcal{H}$ *is a vector space over* $\mathbb{R}$ *or* $\mathbb{C}$ *together with an inner product* $\langle \cdot, \cdot \rangle$ *and which is complete as a metric space.* □

Remark 2.4.2 *The norm is defined as* $\| \phi \| = \sqrt{\langle \phi, \phi \rangle}$ *for* $\phi \in \mathcal{H}$ *and the metric is defined as* $d(\phi, \psi) = \| \phi - \psi \|$ *for* $\phi, \psi \in \mathcal{H}$. *The members of a Hilbert space are called* elements *or* vectors. *In this thesis, we consider only Hilbert spaces over* $\mathbb{R}$. □

The concepts of orthogonality and orthonormal sets will be crucial.

Definition 2.4.3 (Orthogonal Vectors) *Two distinct vectors ϕ and ψ in a Hilbert space $\mathcal{H}$ are said to be* orthogonal *if*

$$\langle \phi, \psi \rangle = 0 \tag{2.40}$$

Definition 2.4.4 (Orthonormal Set) *A countable collection of vectors* $\Phi = \{\phi_1, \phi_2, \ldots\}$ *in a Hilbert space* $\mathcal{H}$ *is said to be an* orthonormal set *if any two distinct vectors* $\phi_i, \phi_j \in \Phi$, $i \neq j$ *are orthogonal and* $\| \phi_i \| = 1$ *for all* $i \geq 1$. □

Definition 2.4.5 (Complete Orthonormal Set) *An orthonormal set* $\Phi = \{\phi_1, \phi_2, \ldots\}$ *in a Hilbert space* $\mathcal{H}$ *is said to be* complete *in* $\mathcal{H}$ *if there exists no vector in* $\mathcal{H}$*, except the zero vector, that is orthogonal to every vector in* Φ. □

When a Hilbert space contains an orthonormal set, every element of the Hilbert space can be represented as a convergent series expansion.

Proposition 2.4.6 (Series Expansion Representation) *If* $\Phi = \{\phi_1, \phi_2, \ldots\}$ *is an orthonormal set in* $\mathcal{H}$ *then for each* $y \in \mathcal{H}$*, the series*

$$\sum_{k=1}^{\infty} \langle\, y, \phi_k \,\rangle\, \phi_k \tag{2.41}$$

converges. Conversely, if

$$y = \sum_{k} \alpha_k \, \phi_k \tag{2.42}$$

then $\alpha_k = \langle\, y, \phi_k \,\rangle$. □

We wish to establish conditions under which every vector in the Hilbert space is guaranteed to have the stated expansion.

Definition 2.4.7 (Orthonormal Basis) *A countable orthonormal set* $\Phi = \{\phi_1, \phi_2, \ldots\}$ *is said to be an* orthonormal basis *for* $\mathcal{H}$ *if for each* $y \in \mathcal{H}$ *and for some* $\alpha_1, \alpha_2, \ldots \in \mathbb{R}$

$$y = \sum_{i} \alpha_i \, \phi_i \tag{2.43}$$

□

Remark 2.4.8 *By the previous result we know that* $\alpha_i = \langle y, \phi_i \rangle$. *Each* $\langle y, \phi_i \rangle$ *is called a* Fourier coefficient *of* $y \in \mathcal{H}$. □

Proposition 2.4.9 *The orthonormal set* $\Phi = \{\phi_1, \phi_2, \ldots\}$ *is an orthonormal basis for the Hilbert space* $\mathcal{H}$ *if and only if* Φ *is complete in* $\mathcal{H}$. □

Remark 2.4.10 *Thus, if there exists a complete orthonormal set in the Hilbert space, then every element of the Hilbert space can be expanded in terms of the basis vectors and Fourier coefficients.* □

It is logical to now ask under what conditions a Hilbert space will contain such a complete orthonormal set.

Definition 2.4.11 (Separable Hilbert Space) *A Hilbert space* $\mathcal{H}$ *is* separable *if* $\mathcal{H}$ *contains a countable set which is dense in* $\mathcal{H}$. □

Proposition 2.4.12 *A Hilbert space contains an orthonormal basis if and only if it is separable.* □

Remark 2.4.13 *In summary, we find that those Hilbert spaces that are separable contain a countable, complete, orthonormal set of vectors, i.e. an orthonormal basis for the Hilbert space, in which every vector in the Hilbert space can be expanded.* □

Finally, we note the following result which states that if an orthonormal basis exists, it is not unique.

Proposition 2.4.14 (Non-uniqueness of Orthonormal Basis) *Given a complete orthonormal set of vectors* $\{\phi_i,\ i = 1, 2, \ldots\}$, *the set* $\{\psi_i,\ i = 1, 2, \ldots\}$ *where*

$$\psi_i = \sum_j \alpha_{ij}\, \phi_j \tag{2.44}$$

for coefficients satisfying

$$\sum_k \alpha_{ik}\,\alpha_{jk} = \delta_{ij} \tag{2.45}$$

is also a complete orthonormal set of vectors. □

Now we present some important examples of Hilbert spaces that we will use in this thesis.

Example 2.4.15 ($\mathbb{R}^n$) *The space of n-tuples $(x_1, \ldots, x_n)$ of real numbers for which*

$$\sum_{i=1}^{n} |x_i|^2 < \infty \tag{2.46}$$

is denoted $\mathbb{R}^n$ and is an n-dimensional Hilbert space. □

Example 2.4.16 (ℓ_2) *The space of infinite sequences $(x_1, x_2, \ldots)$ of real numbers for which*

$$\sum_{i=1}^{\infty} |x_i|^2 < \infty \tag{2.47}$$

is denoted ℓ_2 and is an infinite-dimensional Hilbert space. □

Example 2.4.17 ($\mathcal{L}_2(\mathcal{D})$) *The space of real-valued Lebesgue-measurable square-integrable functions f on a domain $\mathcal{D}$ such that*

$$\int_{\mathcal{D}} |f(x)|^2\, dx < \infty \tag{2.48}$$

is denoted $\mathcal{L}_2(\mathcal{D})$ and generally is an infinite-dimensional Hilbert space. (It is actually a Hilbert space of equivalence classes of functions but we can treat it as a space of functions by identifying functions which are equal almost everywhere.) The inner product on $\mathcal{L}_2$ is given by

$$\langle f, g \rangle_{\mathcal{L}_2} = \int_{\mathcal{D}} f(x)\; g(x)\, dx \tag{2.49}$$

□

Remark 2.4.18 *In this thesis we often deal with the spaces* $\mathcal{L}_2[a,b]$, $\mathcal{L}_2[0,\infty)$, *and* $\mathcal{L}_2(-\infty,0]$. □

We are especially concerned with the following property of the above Hilbert spaces.

Fact 2.4.19 *All of the above Hilbert spaces are separable, i.e. contain a countable orthonormal basis.* □

Moreover, we have the following result.

Proposition 2.4.20 *Any two separable infinite-dimensional Hilbert spaces are isomorphic.* □

Remark 2.4.21 *Actually, we can make the stronger statement that any two separable Hilbert spaces are linearly isometric. Hence,* ℓ_2 *and* $\mathcal{L}_2$ *are indistinguishable as Hilbert spaces.* □

2.5 Stochastic Processes

This thesis relies heavily on the theory of continuous-parameter stochastic processes. In this section, we set up the mathematical framework. The material contained in this section is standard. It is drawn mainly from texts by Ash and Gardner [8], Astrom [10], Davis [36], Papoulis [126], and Wong [166], and class notes in Random Processes presented by Narayan [114] at the University of Maryland. Some basic elements of probability theory are needed, including probability spaces, measurable functions, and expectation. These subjects are covered in the aforementioned texts. We refer to the literature for all proofs.

In what follows, we use the notation $(\Omega, \mathcal{A}, \mathcal{P})$ to denote a *probability space* with *sample space* Ω, associated *σ-algebra* $\mathcal{A}$, and *probability measure* $\mathcal{P}$. The terminology *random variable* refers to an $\mathcal{A}$-measurable $\mathbb{R}^n$-valued function X defined on Ω. Such a function is often called a *random vector* for the case where $n > 1$. However, we use the term random variable regardless of the value of the positive integer n. We will not deal with complex-valued random variables. We usually suppress the dependence of X on $\omega \in \Omega$ and write X as shorthand for $X(\omega)$. The *expected value* of a random variable X is defined by

$$E[X] = \int_{\Omega} X(\omega)\, \mathcal{P}(d\omega) \tag{2.50}$$

Definition 2.5.1 (Stochastic Process) *A* stochastic process $\{X_t,\, t \in T\}$ *is a family of $\mathbb{R}^n$-valued random variables indexed by a real parameter t and defined on a common probability space $(\Omega, \mathcal{A}, \mathcal{P})$.* □

Remark 2.5.2 *The parameter set T is usually taken to be an interval $[a, b]$ where $a < b$. In the cases where $a = -\infty$, $b = \infty$, or both, the interval, respectively, is $(-\infty, b]$, $[a, \infty)$, or $(-\infty, \infty)$. The parameter t represents time unless otherwise specified.* □

Remark 2.5.3 *The ω-dependence is suppressed in the notation $\{X_t,\, t \in T\}$ which is shorthand for $\{X(\omega, t),\, \omega \in \Omega,\, t \in T\}$.* □

Remark 2.5.4 *Similarly, we can define a stochastic process with multiple parameters, e.g., a two-parameter stochastic process $\{X_{t,x},\, t \in T,\, x \in \mathcal{D}\}$ indexed by two real parameters t and x with respective index sets T and $\mathcal{D}$. In this example, the parameters t and x, respectively, typically represent time and a spatial variable.* □

Definition 2.5.5 (Sample Path) *For each $\omega \in \Omega$, $\{X_t,\, t \in T\}$ is a $\mathbb{R}^n$-valued function defined on T and is called a* sample function *or* sample path *of the process.* □

Remark 2.5.6 *In addition, we also note that, by definition, for each $t \in T$, the function$X_t : \Omega \to \mathbb{R}^n$ is a random variable.* □

Transition Properties

When working with stochastically excited systems we will encounter processes that possess the Markov property.

Definition 2.5.7 (Markov Process) *A process $\{X_t,\, t \in T\}$ is said to be a* Markov process *if for any increasing collection $t_1, t_2, \ldots, t_n \in T$*

$$\mathcal{P}(X_{t_n} < x_n | X_{t_\nu} = x_\nu,\, \nu = 1, \ldots, n-1) = \mathcal{P}(X_{t_n} < x_n | X_{t_{n-1}} = x_{n-1}) \tag{2.51}$$

□

Definition 2.5.8 (Transition Density) *Let $\{X_t,\, t \in T\}$ be a Markov process. The* transition function *of the process is defined by*

$$P\left(x, t; y, s\right) = \mathcal{P}\left(X_t < x; X_s = y\right) \tag{2.52}$$

If there is a function $p\left(x, t; y, s\right)$ such that

$$P\left(x, t; y, s\right) = \int_{-\infty}^{x} p\left(u, t; y, s\right) du \tag{2.53}$$

then we call $p\left(x, t; y, s\right)$ the transition density function. □

Remark 2.5.9 *The transition density function $p\left(x, t; y, s\right)$ represents the probability density of being in state x at time t given that the process is in state y at time s.* □

Time Independence and Averaging Properties

Time independence and averaging properties are of importance for our purposes. We wish to be able to take time averages in lieu of ensemble averages, since explicit probability measures may not be available. This requires the property of ergodicity, which in turn requires the property of stationarity.

Definition 2.5.10 (Stationary Process) *A process* $\{X_t,\, t \in \mathbb{R}\}$ *is said to be* stationary *if for any* $(t_1, \ldots, t_n)$ *the joint distribution of* $\{X_{t_1+t_0}, X_{t_2+t_0}, \ldots, X_{t_n+t_0}\}$ *does not depend on* t_0. □

A rigorous definition of what it means for a stationary process to be ergodic requires additional machinery (see, e.g., [166]) which we do not provide here. Instead, we state the following property of an ergodic process which is of primary interest for our purposes.

Proposition 2.5.11 *Let* $\{X_t,\, t \in \mathbb{R}\}$ *be a separable and measurable ergodic process. Let* f *be any Borel function such that* $E\left[\,|f(X_0)|\,\right] < \infty$. *Then*

$$E\left[\,f(X_0)\,\right] = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} f(X_t)\, dt \qquad \textit{almost surely} \tag{2.54}$$

Conversely, if (2.54) holds for every such f, *then* $\{X_t,\, t \in \mathbb{R}\}$ *is ergodic.* □

Remark 2.5.12 *By stationarity,* $E\left[\,f(X_t)\,\right] = E\left[\,f(X_0)\,\right]$ *for all* t. *We interpret Proposition 2.5.11 as saying that, for an ergodic process,* time average equals ensemble average. □

Second-Order Processes

A class of stochastic processes of great importance is the class of second-order processes. In this thesis we make extensive use of the covariance function of a

stochastic process. In order for a process to have a covariance function, it must be a second-order process.

Definition 2.5.13 (Second-Order Random Variable) *A random variable X is said to be a* second-order random variable *if $E[\|X\|^2] < \infty$ where $\|\cdot\|$ denotes the usual Euclidean norm.* □

Definition 2.5.14 (Second-Order Stochastic Process) *A process $\{X_t,\, t \in T\}$ is said to be a* second-order stochastic process *if for each fixed $t \in T$, X_t is a second order random variable.* □

Thus, a second-order stochastic process is a parameterized family of second-order random variables. It has at least a first moment, second moment, and second central moment, called its mean, correlation, and covariance.

Definition 2.5.15 (Mean, Correlation, Covariance) *Let $\{X_t,\, t \in T\}$ be a second-order process. The* mean function $\mu : T \to \mathbb{R}^n$, correlation function $\mathcal{R} : T \times T \to \mathbb{R}^{n \times n}$, *and* covariance function $R : T \times T \to \mathbb{R}^{n \times n}$ *are defined, respectively, as*

$$\mu(t) = E[X_t] \tag{2.55}$$

$$\mathcal{R}(t,s) = E[X_t\, X_s^{\mathsf{T}}] \tag{2.56}$$

$$R(t,s) = E[(X_t - \mu(t))(X_s - \mu(s))^{\mathsf{T}}] \tag{2.57}$$

□

Remark 2.5.16 *The correlation and covariance functions are sometimes referred to, respectively, as the* autocorrelation *and* autocovariance *functions.* □

Remark 2.5.17 *For each $t \in T$, the mean function $\mu(t)$ is an n-vector. For each $t, s \in T$, the correlation and covariance functions, respectively, $\mathcal{R}(t,s)$ and $R(t,s)$, are $n \times n$-matrices.* □

Remark 2.5.18 *For a process with zero mean, $R(t,s) = \mathcal{R}(t,s)$ and the covariance and correlation functions can be used interchangeably.* □

A covariance function satisfies a number of important properties. Two of importance for our purposes are as follows.

Proposition 2.5.19 (Symmetry of Covariance) *Let $\{X_t,\, t \in T\}$ be a second-order process. Its covariance function is symmetric, i.e.,*

$$R(t,s) = R(s,t) \qquad t, s \in T \tag{2.58}$$

□

Proposition 2.5.20 (Non-negativity of Covariance) *Let $\{X_t,\, t \in T\}$ be a second-order process. Its covariance function is non-negative definite, i.e., for any finite collection $t_1, \ldots, t_n$ and real constants $\alpha_1, \ldots, \alpha_n$*

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \, \alpha_j \, R(t_i, t_j) \geq 0 \tag{2.59}$$

□

We are interested in working with covariance functions that are time independent. This can be ensured by assuming that the second-order process is stationary. However, stationarity is stronger than we actually need.

Definition 2.5.21 (Wide-Sense Stationary Process) *A second-order process $\{X_t,\, t \in \mathbb{R}\}$ is said to be* wide-sense stationary *if its covariance function $R(t,s)$ is*

a function only of the difference $t - s$, i.e.,

$$R(t,s) = R(t-s) \qquad t,s \in T \tag{2.60}$$

□

Remark 2.5.22 *A stationary second-order process is wide-sense stationary, but the converse is not necessarily true. Also, note that for a wide-sense stationary process, the mean function must be a constant, i.e., $\mu(t) \equiv \mu$.* □

One object of importance that is associated with a wide-sense stationary second-order process is its spectral density function, which will appear in the description of white noise.

Definition 2.5.23 (Spectral Density Function) *The* spectral density function *for a wide-sense stationary second-order process $\{X_t,\, t \in T\}$ with covariance $R(\tau)$ is defined as*

$$S(\nu) = \int_{-\infty}^{\infty} \exp(-i2\pi\nu\tau)\; R(\tau)\, d\tau \qquad \nu \in \mathbb{R} \tag{2.61}$$

□

Remark 2.5.24 *The inversion integral is*

$$R(\tau) = \int_{-\infty}^{\infty} \exp(i2\pi\nu\tau)\; S(\nu)\, d\nu \qquad \tau \in \mathbb{R} \tag{2.62}$$

□

We will be working with series expansions and stochastic differential equations. Thus, we need to take limits and derivatives. We use the following notions of limits and continuity when dealing with a second-order process.

Definition 2.5.25 (Quadratic Mean Convergence) *A sequence of random variables* $\left\{X^{(n)}\right\}$ *is said to* converge in quadratic mean *to* X *if*

$$\lim_{n\to\infty} E\left[\left\| X^{(n)} - X \right\|^2\right] = 0 \tag{2.63}$$

We call X *the* limit in quadratic mean (q.m. limit) *of* $\left\{X^{(n)}\right\}$ *and use the notation*

$$X = \lim_{n\to\infty} \text{in q.m. } X^{(n)} \tag{2.64}$$

Definition 2.5.26 (Quadratic Mean Continuous Process) *A second-order process* $\{X_t,\, t \in T\}$ *is said to be* continuous in quadratic mean (q.m. continuous) *at* t *if*

$$\lim_{h\to 0} E\left[\left\| X_{t+h} - X_t \right\|^2\right] = 0 \tag{2.65}$$

A process that is q.m. continuous at every $t \in T$ *is said to be a* q.m. continuous process. □

An important fact about q.m. continuous processes that we will use is

Proposition 2.5.27 (Continuity of Covariance) *If* $\{X_t,\, t \in T\}$ *is a second-order q.m. continuous process then its covariance function* $R(\cdot,\cdot)$ *is continuous at every point on the square* $T \times T$. □

Two-Parameter Second-Order Processes

We will find it useful to redefine some of the above notions in the context of two-parameter stochastic processes. Consider a two-parameter second-order process $\{X_{t,x},\, t \in T,\, x \in \mathcal{D}\}$. The mean function, correlation function, and covariance function, respectively, are given by

$$\mu(t,x) = E[X_{t,x}] \tag{2.66}$$

$$\mathcal{R}(t,x,s,y) = E[X_{t,x}\, X_{s,y}^{\mathsf{T}}] \tag{2.67}$$

$$R(t,x,s,y) = E[(X_{t,x} - \mu(t,x))(X_{s,y} - \mu(s,y))^{\mathsf{T}}] \tag{2.68}$$

It is often useful to fix one of the two parameters. When the parameter t represents time and the parameter x represents space, we use the following terminology. The functions

$$\begin{aligned}
\mathcal{R}(t,x,t,y) &= E[X_{t,x}\, X_{t,y}^{\mathsf{T}}] && (2.69)\\
R(t,x,t,y) &= E[(X_{t,x}-\mu(t,x))(X_{t,y}-\mu(t,y))^{\mathsf{T}}] && (2.70)\\
\mathcal{R}(t,x,s,x) &= E[X_{t,x}\, X_{s,x}^{\mathsf{T}}] && (2.71)\\
R(t,x,s,x) &= E[(X_{t,x}-\mu(t,x))(X_{s,x}-\mu(s,x))^{\mathsf{T}}] && (2.72)
\end{aligned}$$

are called, respectively, the *spatial correlation*, *spatial covariance*, *temporal correlation*, and *temporal covariance*. Sometimes, the term *two-point* precedes the object name, e.g., two-point spatial covariance. The symbols $\mathcal{R}$ and R are used to denote, respectively, the correlation and covariance functions, in all cases.

If a two-parameter second-order process is wide-sense stationary with respect to time, then the spatial correlation and spatial covariance, respectively, can be written in terms of the spatial parameters only, i.e.,

$$\begin{aligned}
\mathcal{R}(t,x,t,y) &= \mathcal{R}(x,y)\\
R(t,x,t,y) &= R(x,y)
\end{aligned}$$

If a two-parameter second-order process is wide-sense stationary with respect to the spatial variable, then the temporal correlation and temporal covariance, respectively, can be written in terms of the temporal parameters only, i.e.,

$$\begin{aligned}
\mathcal{R}(t,x,s,x) &= \mathcal{R}(t,s)\\
R(t,x,s,x) &= R(t,s)
\end{aligned}$$

Finally, the notion of q.m. continuity must be considered with respect to one particular parameter, i.e., a two-parameter second-order process

$\{X_{t,x},\, t \in T,\, x \in \mathcal{D}\}$ is q.m. continuous with respect to $t \in T$ if for each $x \in \mathcal{D}$

$$\lim_{h \to 0} E\left[\| X_{t+h,x} - X_{t,x} \|^2 \right] = 0 \tag{2.73}$$

and similarly for q.m. continuity with respect to $x \in \mathcal{D}$.

Hilbert Space Properties

It will be necessary to collect random variables as members of a Hilbert space.

Definition 2.5.28 (Linear Operation on a Second-Order Process) *Let* $\{X_t,\, t \in T\}$ *be a second-order process. A random variable* Y *is said to be* derived from a linear operation *on* $\{X_t,\, t \in T\}$ *if either of the following are true:*

(i) For some integer N *and times* $\{t_1, \ldots, t_N\}$

$$Y = \sum_{i=1}^{N} \alpha_i X_{t_i} \tag{2.74}$$

(ii) Y *is the q.m. limit of a sequence of such finite linear combinations.*

□

Definition 2.5.29 ($\mathcal{H}_X$) *The collection of all random variables derived from linear operations on a process* $\{X_t,\, t \in T\}$ *is denoted* $\mathcal{H}_X$. □

Remark 2.5.30 *The set* $\mathcal{H}_X$ *is generally an infinite-dimensional Hilbert space. It is separable, and so is linearly isometric with* ℓ_2 *(see Section 2.4). The inner product on* $\mathcal{H}_X$ *is given by*

$$\langle Y, Z \rangle_{\mathcal{H}_X} = E\left[Y\, Z^{\mathsf{T}} \right] \tag{2.75}$$

□

Gaussian Processes

A special case of a second-order stochastic process of great importance for our purposes is a Gaussian process. We first need to define what we mean by a Gaussian random variable.

Definition 2.5.31 (Gaussian Random Variable) *A second-order random variable Z with $\mu = E[Z]$ and $\sigma^2 = E[(Z-\mu)^2]$ is said to be* Gaussian *if $\sigma^2 = 0$, in which case $Z = \mu$ with probability* 1*, or*

$$\Pr(Z < a) = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\frac{(z-\mu)^2}{\sigma^2}\right] dz \tag{2.76}$$

□

Remark 2.5.32 *A random variable that is Gaussian is also referred to as* normal*.* □

Remark 2.5.33 *A Gaussian random n-vector has a density function determined only by parameters μ and R, given by*

$$p_Z(z) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2}(z-\mu)^{\mathsf{T}} R^{-1} (z-\mu)\right] \tag{2.77}$$

□

Definition 2.5.34 (Gaussian Process) *A second-order stochastic process $\{X_t,\, t \in T\}$ is said to be a* Gaussian process *if for some integer N and times $\{t_1, \ldots, t_N\}$, every finite linear combination of the form*

$$Z = \sum_{i=1}^{N} \alpha_i X_{t_i} \tag{2.78}$$

is a Gaussian random variable. □

The Wiener Process (Brownian Motion)

One important special case of a Gaussian process is the Wiener process, which is the mathematical abstraction of the physical phenomena known as Brownian motion. We use the terms Wiener process and Brownian motion process interchangeably. The Wiener process is a zero mean process with certain properties, defined as follows.

Definition 2.5.35 (Orthogonal Increments) *A process* $\{X_t,\, t \in T\}$ *is said to have* orthogonal increments *if for any non-overlapping intervals* (s,t) *and* (s',t')

$$E\left[\,(X_{t'} - X_{s'})(X_t - X_s)^{\mathsf{T}}\,\right] = 0 \tag{2.79}$$

□

Remark 2.5.36 *A process that has orthogonal increments is said to be an* orthogonal increments process. □

Definition 2.5.37 (Independent Increments) *A process* $\{X_t,\, t \in T\}$ *is said to have* independent increments *if for any two non-overlapping intervals* (s,t) *and* (s',t')*, the random variables* $(X_{t'} - X_{s'})$ *and* $(X_t - X_s)$ *are independent.* □

Remark 2.5.38 *A process that has independent increments is said to be an* independent increments process. □

Remark 2.5.39 *Clearly, the class of independent increments processes is a subclass of the class of orthogonal increments processes.* □

Definition 2.5.40 (Stationary Increments) *A process* $\{X_t,\, t \in T\}$ *has* stationary increments *if the variance of the increment* $(X_t - X_s)$ *depends only on the distance* $|t-s|$*, i.e.,*

$$E[(X_t - X_s)^2] = E[(X_{t+r} - X_{s+r})^2] \qquad r, s, t \in T \tag{2.80}$$

Remark 2.5.41 *A process that has stationary increments is said to be a* stationary increments process. □

Proposition 2.5.42 *A q.m. continuous process* $\{X_t,\, t \in T\}$ *is said to have stationary orthogonal increments if and only if its covariance function is*

$$R(t,s) = \sigma^2 \min(t,s) \tag{2.81}$$

□

Now we define the Wiener process. It is defined for positive time, usually on the interval $[0,\infty)$.

Definition 2.5.43 (Wiener Process) *A process* $\left\{W_t,\, t \in \mathbb{R}^+\right\}$ *is said to be a* Wiener process *or a* Brownian motion process *if it has zero mean, i.e.,* $E\left[W_t\right] = 0$ *for* $t \geq 0$*, and it has stationary independent Gaussian increments.* □

Remark 2.5.44 *By a* standard Wiener process *we mean that* $W_0 = 0$ *and* $E[W_1^2] = 1$*. Thus, a standard Wiener process is a Gaussian process with mean* $\mu(t) \equiv 0$ *and covariance function* $R(t,s) = \min(t,s)$*.* □

Remark 2.5.45 *Some important properties of a Wiener process are*

(i) Any sample path $W(t)$ *is continuous everywhere with probability 1, differentiable nowhere, and of infinite length.*

(ii) The increment $(W_{t+h} - W_t)$ *is of order* $O(\sqrt{h})$*, i.e.,* dW_t *is proportional to* $\sqrt{dt}$*.*

□

Even though the sample path of a Wiener process is nowhere differentiable, it is often written, formally, that there is a process $\left\{\zeta_t,\, t \in \mathbb{R}^+\right\}$ such that

$$\zeta_t = \frac{d}{dt} W_t \tag{2.82}$$

and the converse

$$W_t = \int_0^t \zeta_s \, ds \tag{2.83}$$

are true in some useful sense. Such a process $\left\{\zeta_t,\, t \in \mathbb{R}^+\right\}$ is called *white noise*. A process with the above relationship to Brownian motion, i.e., its formal derivative, is extremely useful in applications, even though it does not exist in the traditional sense. We elaborate and justify its relationship with a Wiener process as follows.

White Noise

A white noise is usually described (and sometimes defined) as a wide-sense stationary process with a spectral density function that is constant over all frequencies, i.e.,

$$S(\nu) = S_0 \qquad \nu \in \mathbb{R} \tag{2.84}$$

This description has the following implications:

(i) $R(0) = \infty$

(ii) $R(\tau) = \delta(\tau)\, S_0$

Thus, a process with the above description is not a second-order process and does not have a well-defined spectral density. In fact, a white noise is not well-defined as a stochastic process. Rather, it is a generalized process. We will not proceed with a digression on generalized processes here. See Arnold [7] for an exposition on this subject. For completeness, we include the definition of a white noise process.

Definition 2.5.46 (White Noise) *A generalized Gaussian stochastic process* Φ_ζ *is said to be a* Gaussian white noise *process if it has mean functional* $E[\Phi_\zeta] = 0$ *and covariance functional*

$$C_\zeta(\phi, \psi) = \int_{-\infty}^{\infty} \phi(t)\,\psi(t)\,dt \tag{2.85}$$

Remark 2.5.47 *When the Wiener process is considered as a generalized process, the covariance function of its derivative is given by*

$$R(t, s) = \delta(t - s)\,S_0 \tag{2.86}$$

which is the covariance function of a white noise process. Thus, white noise $\left\{\zeta_t,\, t \in \mathbb{R}^+\right\}$ *is the derivative of the Wiener process* $\left\{W_t,\, t \in \mathbb{R}^+\right\}$ *when both processes are considered as generalized processes. This justifies the relationships (2.82) and (2.83).* □

It suffices for our purposes to use the formal description of a white noise process, justified by noting that $\left\{\zeta_t,\, t \in \mathbb{R}^+\right\}$ is never used outside of an integral. In particular, an expression of the form

$$\int_a^b \zeta_t\,\phi(t)\,dt \tag{2.87}$$

is said to be a *white noise integral*. Expression (2.87) is merely formal; there is no stochastic process $\left\{\zeta_t,\, t \in \mathbb{R}^+\right\}$ for which such an integral exits. Rather, it is to be interpreted as a stochastic integral,

$$\int_a^b \phi(t)\,dW_t \tag{2.88}$$

which is defined in Section 2.6.

2.6 Stochastically Excited Dynamical Systems

A stochastically excited dynamical system is a control system with white noise injected at the input terminals. These systems play a crucial role in the model reduction approach presented in this thesis. The time evolution of the state of such a system is not governed by differential equations using the ordinary Stieltjes calculus. Instead, it is necessary to work with the stochastic calculus, including stochastic differential equations (SDEs). Moreover, the state is a stochastic process, with an associated probability density function, the evolution of which is governed by a pair of diffusion equations.

The material contained in this section is based on that presented in texts by Arnold [7], Astrom [10], Davis [36], and Wong [166], papers by Brockett [22] and Fuller [51], and class notes in Stochastic Control presented by Marcus [101] at the University of Maryland. It relies heavily on the material presented in Section 2.5. We refer to the literature for all proofs.

2.6.1 State Equations

Recall the form of the state equation for an affine control system

$$\dot{x}\,(t) = f\,(t, x(t)) + \sum_{i=1}^{m} g_i\,(t, x(t))\; u_i(t) \tag{2.89}$$

where for purposes of generality we include the possibility of explicit time dependence. By a *stochastically excited dynamical system*, we mean an affine control system for which the m components of the input, u_i, $i \in \underline{m}$, have been replaced by the sample paths of m Gaussian white noises, $\left\{(\zeta_t)_i,\, t \in \mathbb{R}^+\right\}$, $i \in \underline{m}$.

The evolution equation for the state process $\left\{X_t,\, t \in \mathbb{R}^+\right\}$ takes the form

$$\frac{d}{dt}\, X_t = f(t, X_t) + \sum_{i=1}^{m} g_i(t, X_t)\,(\zeta_t)_i \tag{2.90}$$

The meaning of (2.90) is given in terms of the stochastic integral. In particular, given a function $\phi : \Omega \times \mathbb{R} \to \mathbb{R}^{n\times m}$, a *stochastic integral* is a quantity of the form

$$I(\phi) = \int_a^b \phi(\omega, t)\, dW(\omega, t) \tag{2.91}$$

Because $\left\{W_t,\, t \in \mathbb{R}^+\right\}$ is neither differentiable nor of bounded variation, (2.91) does not have a well-defined interpretation as an integral in the ordinary sense. Therefore, it is necessary to define what we mean by (2.91).

We use the following norm for functions $\phi : \Omega \times \mathbb{R} \to \mathbb{R}^{n\times m}$

$$\| \phi \| = \int_a^b \sum_{i=1}^{n} \sum_{j=1}^{m} E\left[|\phi_{ij}(\cdot, t)|^2 \right] dt \tag{2.92}$$

The stochastic integral is defined in terms of a step function or a sequence of step functions.

Definition 2.6.1 ((ω, t)-Step Function) *Let ϕ be jointly measurable in (ω, t) and such that $\| \phi \| < \infty$. If there exist times $t_0, \ldots, t_n$, independent of ω as functions, such that $a < t_0 < \cdots < t_n < b$ and*

$$\phi(\omega, t) = \phi_\nu(t) \qquad t_\nu \le t < t_{\nu+1} \qquad \nu = 1, \ldots, n-1 \tag{2.93}$$

then ϕ is said to be an (ω, t)-step function. □

Definition 2.6.2 (Stochastic Integral) *Let ϕ be jointly measurable in (ω, t) and such that $\| \phi \| < \infty$. The quantity*

$$I(\phi) = \int_a^b \phi(\omega, t)\, dW(\omega, t) \tag{2.94}$$

is said to be a stochastic integral *defined as follows:*

(i) If ϕ is an (ω, t)-step function then

$$\int_a^b \phi(\omega, t)\, dW(\omega, t) = \sum_{\nu=1}^{n-1} \phi_\nu(\omega)\, [W(\omega, t_{\nu+1}) - W(\omega, t_\nu)] \tag{2.95}$$

(ii) Otherwise,

$$\int_a^b \phi(\omega,t)\,dW(\omega,t) = \lim_{n\to\infty} \text{in q.m.} \int_a^b \phi_n(\omega,t)\,dW(\omega,t) \tag{2.96}$$

where $\{\phi_n\}$ *is a sequence of* (ω,t)*-step functions satisfying*

$$\lim_{n\to\infty} \| \phi - \phi_n \|^2 = 0 \tag{2.97}$$

□

Remark 2.6.3 *The existence of the convergent sequence in (2.97) is guaranteed (see, e.g., [166] Chap. 4 Prop. 2.1).* □

Remark 2.6.4 *In what follows, and throughout this thesis, we take the limits of integration as* $a = 0$ *and* $b = \infty$ *unless specified otherwise. In this case, the stochastic integral is defined via the limit in q.m. as* $b \to \infty$*.* □

The mathematical model for a stochastically excited dynamical system uses the notion of a stochastic differential equation and a white noise driven differential equation, both of which are interpreted precisely via the stochastic integral.

Definition 2.6.5 (Stochastic Differential Equation) *Given functions* $f : \mathbb{R}^+ \times \mathbb{R}^n \to \mathbb{R}^n$ *and* $g_i : \mathbb{R}^+ \times \mathbb{R}^n \to \mathbb{R}^n$*,* $i \in \underline{m}$*, a* stochastic differential equation *(SDE) is an equation of the form*

$$dX_t = f(t,X_t)\,dt + \sum_{i=1}^{m} g_i(t,X_t)\,(dW_t)_i \tag{2.98}$$

□

Definition 2.6.6 (Solution of SDE) *A process* $\left\{X_t,\, t \in \mathbb{R}^+\right\}$ *is said to* satisfy the SDE (2.98) *with initial condition* $X_0 = X$ *if*

(i) The quantities

$$\int_0^t g_i(s, X_s)\,(dW_s)_i \qquad i \in \underline{n}$$

are capable of being interpreted as stochastic integrals.

(ii) For each t, X_t is almost surely equal to the random variable defined by

$$X + \int_0^t f(s, X_s)\,ds + \int_0^t \sum_{i=1}^m g_i(s, X_s)\,(dW_s)_i$$

where the first integral is of ordinary type and the second is a stochastic integral.

□

Remark 2.6.7 *Thus, the SDE (2.98) is an expression that means*

$$X_t = X_0 + \int_0^t f(s, X_s)\,ds + \int_0^t \sum_{i=1}^m g_i(s, X_s)\,(dW_s)_i \tag{2.99}$$

where the first integral is of ordinary type and the second is a stochastic integral.
□

The existence and uniqueness of a solution $\{X_t,\, t \in [0, b]\}$ of SDE (2.98) is guaranteed under certain regularity conditions on f and g. Furthermore, under those conditions, the unique solution is a Markov process.

Proposition 2.6.8 (Properties of Solutions) *Let $\{W_t,\, t \in [0, b]\}$ be a Wiener process and X be a second-order random variable. Let $f(t, x)$ and $g(t, x)$, $t \in [0, b]$, $x \in \mathbb{R}$, be measurable in (t, x). Suppose that f and g satisfy the following conditions:*

$$|f(t, x) - f(t, y)| + |g(t, x) - g(t, y)| \leq K\,|x - y| \tag{2.100}$$

$$|f(t, x)| + |g(t, x)| \leq K\,\sqrt{1 + x^2} \tag{2.101}$$

Then there exists a process $\{X_t,\, t \in [0, b]\}$ such that

(i) $\{X_t,\, t \in [0,b]\}$ *satisfies the SDE (2.98) with initial condition* $X_0 = X$.

(ii) $\{X_t,\, t \in [0,b]\}$ *is unique with probability* 1.

(iii) $\{X_t,\, t \in [0,b]\}$ *is a Markov process.*

(iv) $\{X_t,\, t \in [0,b]\}$ *has continuous sample paths with probability* 1.

□

Remark 2.6.9 *The condition (2.100) is called the* uniform Lipschitz condition *and the condition (2.101) is called the* restriction on growth condition. *The constants K can be the same. If the restriction on growth condition is violated, we get the effect of an "explosion" of the solution, i.e., a finite escape time.* □

The connection between a stochastically excited system and SDEs is made using the notion of a white noise driven differential equation.

Definition 2.6.10 (White Noise Driven Differential Equation) *Given functions $f : \mathbb{R}^+ \times \mathbb{R}^n \to \mathbb{R}^n$ and $g_i : \mathbb{R}^+ \times \mathbb{R}^n \to \mathbb{R}^n$, $i \in \underline{m}$, a* white noise driven differential equation *is an equation of the form (2.90) (repeated below)*

$$\frac{d}{dt} X_t = f(t, X_t) + \sum_{i=1}^{m} g_i(t, X_t)\,(\zeta_t)_i$$

where for each i, $\left\{(\zeta_t)_i,\, t \in \mathbb{R}^+\right\}$ is a Gaussian white noise. □

The interpretation of (2.90) is that of a sequence of SDEs

$$\frac{d}{dt} X_t^{(n)} = f(t, X_t^{(n)}) + \sum_{i=1}^{m} g_i(t, X_t^{(n)})\,(\zeta_t^{(n)})_i \tag{2.102}$$

where $\left\{\zeta_t^{(n)},\, t \in \mathbb{R}^+\right\}$ represents a sequence of Gaussian processes that converges in some suitable sense to a white noise, yet for each n, $\left\{\zeta_t^{(n)},\, t \in \mathbb{R}^+\right\}$ has well-behaved sample paths. If the sequence of processes $\left\{X_t^{(n)},\, t \in \mathbb{R}^+\right\}$ converges (say, in q.m.) to a process $\left\{X_t,\, t \in \mathbb{R}^+\right\}$ then we interpret X_t as the solution of (2.90).

It was shown by Wong and Zakai [164, 165] that, given the above interpretation, the precise mathematical meaning of (2.90) is that of an SDE (given elementwise)

$$(dX_t)_i = \bar{f}_i(t, X_t)\ dt + \sum_{i=1}^{m} g_i(t, X_t)\ (dW_t)_i \qquad i \in \underline{n} \tag{2.103}$$

where

$$\bar{f}_i(t, X_t) = f_i(t, X_t) + \frac{1}{2} \sum_{j=1}^{n} \sum_{k=1}^{m} \frac{\partial g_{ik}}{\partial X_j}(t, X_t)\ g_{jk}(t, X_t) \qquad i \in \underline{n} \tag{2.104}$$

and $g_{ij} = (g_j)_i$, $i \in \underline{n}, j \in \underline{m}$.

Definition 2.6.11 (Correction Term) *The second term on the right side of (2.104) is called the* correction term *(sometimes referred to as the* Ito-Stratonovich correction term*). Its appearance is due to the fact that dW_t is proportional to $\sqrt{dt}$ (see Remark 2.5.45).* □

Remark 2.6.12 *To summarize, a stochastically excited system is an affine control system with Gaussian white noise injected at the input terminals. It is modeled by a white noise driven differential equation (2.90), which is interpreted as an SDE of the form (2.103). The SDE (2.103) is defined in terms of the stochastic integral, i.e., by an integral equation of the form (2.99), but with f replaced by $\bar{f}$, i.e., the sum of f and the correction term.* □

Remark 2.6.13 *When the functions g_i, $i \in \underline{m}$ are independent of x, i.e., $g_i(\cdot, x) = g_i(\cdot)$, then the correction term vanishes.* □

Remark 2.6.14 (Simulation) *Care must be taken in order to correctly implement a numerical simulation of SDE (2.103). In particular, the continuous-time Wiener process must be approximated by a sequence of Gaussian random variables. The statistics of these random variables must be chosen in a manner consistent with the approximation scheme. Details are provided in Appendix C.* □

2.6.2 Diffusion Equations

We have presented the mathematical framework regarding the time evolution of the state for a stochastically excited system. The state, which is represented by a Markov process, has an associated transition probability, which also evolves in time. The transition probability of a process satisfying an SDE can be obtained by solving either of a pair of parabolic PDEs. These equations are called the backward and forward equations of Kolmogorov, or diffusion equations.

Let the process $\{X_t,\, t \in T\}$ be the unique solution of the SDE (2.98) and have transition function $P(x,t;y,s)$ and transition density function $p(x,t;y,s)$ (see Definition 2.5.8). The forward time evolution of $p(x,t;y,s)$ is governed by the *forward equation of Kolmogorov*, also known as the *Fokker-Planck equation*, given by

$$\begin{aligned}\frac{\partial p}{\partial t}(x,t;y,s) = \mathcal{L}\, p = \\ -\sum_{i=1}^{n} \frac{\partial}{\partial x_i}\left(f_i(t,x)\, p(x,t;y,s)\right) \\ +\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial^2}{\partial x_i\, \partial x_j}\left(b_{ij}(t,x)\, p(x,t;y,s)\right)\end{aligned} \tag{2.105}$$

with initial condition

$$p(x,s;y,s) = \delta(x-y)$$

and where

$$b_{ij}(t,x) = \sum_{k=1}^{m} g_{ik}(t,x)\; g_{jk}(t,x) = \left[\left[g(t,x)\right]\, \left[g(t,x)\right]^{\mathsf{T}}\right]_{ij}$$

The reverse time evolution of $P(x,t;y,s)$ is governed by the *backward equation of Kolmogorov*, given by

$$\frac{\partial P}{\partial s}(x,t;y,s) = \mathcal{L}^*\, P =$$

$$\sum_{i=1}^{n} f_i(s,y)\,\frac{\partial}{\partial y_i}\left(P(x,t;y,s)\right) + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} b_{ij}(s,y)\,\frac{\partial^2}{\partial y_i\,\partial y_j}\left(P(x,t;y,s)\right) \tag{2.106}$$

with terminal condition

$$\lim_{t\uparrow s} P(x,t;y,s) = \begin{cases} 1 & x > y \\ 0 & x < y \end{cases}$$

Detailed derivations of Equations (2.105) and (2.106) are presented in [7, 166]. For a history and derivation of the Fokker-Planck equation from a physical point of view see [51].

Remark 2.6.15 *The operators $\mathcal{L}$ and $\mathcal{L}^*$ are linear and adjoints.* □

Remark 2.6.16 *When working with a white noise driven equation, it is important to keep in mind that the functions f_i, $i \in \underline{n}$, must incorporate the correction term.* □

Existence and uniqueness of solutions to Equations (2.105) and (2.106) can be shown under suitable regularity conditions (see, e.g., [7, 166]). However, the following result by Elliott [42, 43] (and elaborated upon by Brockett [21, 22]) is more useful for our purposes (because it appeals to our control theoretic viewpoint).

Theorem 2.6.17 (Elliott [42, 43]) *Suppose that*

(i) the Lie algebra of vector fields generated by $\{f, g_1, \ldots, g_m\}$ consists of complete vector fields on a manifold M; and

(ii) the smallest Lie algebra which contains $g_1, \ldots, g_n$ and which is closed under bracketing with f spans the tangent space of M at each point.

Then the corresponding SDE (2.103) defines smooth transition densities on M. □

Remark 2.6.18 *By completeness of a vector field f, we mean that the solution of $\dot{x} = f(x), \quad x(0) = x_0$ is defined for all time, i.e., no finite escape times in forward or backward time.* □

Remark 2.6.19 *Theorem 2.6.17 states that strong local accessibility of the corresponding affine control system, together with a completeness condition on vector fields, guarantees the existence of smooth transition densities, i.e., smooth solutions of the Fokker-Planck equation, for all times t. We apply this result in Section 4.3.2.* □

In many applications, the functions f_i, $i \in \underline{n}$, and g_{ij}, $i \in \underline{n}$, $j \in \underline{m}$, (and hence the b_{ij}) are time-independent. In such cases we are often interested in the steady-state probability density (if any) which p approaches as t becomes large. In the steady-state, $\dfrac{\partial p}{\partial t}$ vanishes, i.e., the probability density is stationary, and Equation (2.105) simplifies to the *stationary Fokker-Planck equation*

$$0 = -\sum_{i=1}^{n} \frac{\partial}{\partial x_i} \left(f_i(x)\, p_\infty(x) \right) + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial^2}{\partial x_i\, \partial x_j} \left(b_{ij}(x)\, p_\infty(x) \right) \tag{2.107}$$

where $p_\infty(x)$ denotes the stationary probability density (if it exists).

Remark 2.6.20 *The dependence of the transition density on y and s vanishes in the steady-state case.* □

A solution $p_\infty(x)$, if it is to represent a probability density function, must also satisfy

$$p_\infty(x) \geq 0 \qquad x \in \mathbb{R}^n \tag{2.108}$$

and

$$\int_{\mathbb{R}^n} p_\infty(x)\, dx = 1 \tag{2.109}$$

Definition 2.6.21 (Stationary Solution) *A solution of (2.107) which also satisfies (2.108) and (2.109) is called a* stationary solution *of the Fokker-Planck equation.* □

Boundary conditions for (2.107) are assigned as

$$\lim_{x\to\infty} p_\infty(x) = 0 \tag{2.110}$$

and as a consequence

$$\lim_{x\to\infty} \frac{\partial p_\infty}{\partial x}(x) = 0 \tag{2.111}$$

We can use Theorem 2.6.17 in order to establish the existence of a smooth invariant density. In addition, Zakai [169] shows how to prove existence via a Lyapunov criterion. Moreover, Fuller [51] argues that we can heuristically assume the existence of a unique stationary asymptotically stable transition density if

(i) No part of the system is completely isolated from the effects of the white noise.

(ii) The system has restoring forces which prevent the ensemble from dispersing to infinity.

We shall make the standing assumption that the stochastically excited systems we work with yield a unique stationary asymptotically stable transition density, unless noted otherwise. We call this transition density the *steady-state density*, *stationary density*, or *invariant density*.

Closed form solutions of (2.107) exist in certain special cases such as for systems with a conservative (Hamiltonian) part and a dissipative part [51, 171]. These solutions will be exploited in Chapter 4.

2.7 Mechanical Systems

Mechanical systems whose dynamics can be described by the *Euler-Lagrange* or *Hamilton* equations of motion form a subclass of nonlinear control systems that is of major importance in this thesis. In particular, we will consider mechanical systems consisting of open chained rigid links. In this section we briefly present the mathematical framework for working with this subclass of systems. The material is standard and drawn mainly from texts by Murray, Li, and Sastry [112] and Nijmeijer and van der Schaft [121] and a paper by Fuller [51]. We refer to the literature for all proofs.

The motion of a mechanical system can be described by a set of variables that completely determines the configuration of the system. We refer to such a set of variables as *generalized coordinates*, denoted in vector form as $q = (q_1, \ldots, q_n) \in \mathbb{R}^n$, where n denotes the number of *degrees of freedom* (DOF) of the system. For a mechanical system consisting of rigid links, the generalized coordinates are almost always chosen to be the angles of the joints. We also refer to the q_i as the generalized positions and $\dot{q}_i$ as the generalized velocities.

We express the external forces applied to the system in terms of components along the generalized coordinates. These forces are referred to as *generalized forces*, denoted in vector form as $F = (F_1, \ldots, F_n) \in \mathbb{R}^n$. For the rigid link system with joint angles acting as generalized coordinates, the generalized forces are the torques applied about the joint axes.

The kinetic energy K of the system is a function of the generalized positions and velocities, i.e., $K = K(q, \dot{q})$. For a system of rigid links it is usually written as the sum of a translational component and a rotational component. The potential energy U is a function of position only, i.e., $U = U(q)$. It is usually written as

the sum of stored energies due to gravity and mechanical stiffness. The dissipation energy R (also called the *Rayleigh dissipation function*) is generally a function of position and velocity, i.e., $R = R(q, \dot{q})$. It contains terms reflecting generalized mechanical damping.

We define the *Lagrangian* as the difference between the kinetic and potential energies of the system, i.e.,

$$L(q, \dot{q}) = K(q, \dot{q}) - U(q) \tag{2.112}$$

The equations of motion for the system can be derived from the Lagrangian L and the Rayleigh dissipation function R via the *Euler-Lagrange equations of motion*.

Theorem 2.7.1 (Euler-Lagrange Equations of Motion) *The equations of motion for a mechanical system with generalized coordinates $q \in \mathbb{R}^n$, generalized forces $F \in \mathbb{R}^n$, and Lagrangian L are given by*

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{q}_i} - \frac{\partial L}{\partial q_i} = F_i - \frac{\partial R}{\partial \dot{q}_i} \qquad i \in \underline{n} \tag{2.113}$$

□

A control system model in standard state-space form is obtained from the Euler-Lagrange equations of motion by interpreting the external forces as the control inputs and expressing the kinetic energy as

$$K(q, \dot{q}) = \frac{1}{2}\dot{q}^{\mathsf{T}} M(q)\, \dot{q} \tag{2.114}$$

where $M(q)$ is a positive-definite matrix called the *inertia matrix* or *mass matrix*. The equations of motion can be written

$$M(q)\, \ddot{q} + C(q, \dot{q}) + N(q, \dot{q}) = F \tag{2.115}$$

where $C(q, \dot{q})$ represents Coriolis and centrifugal force terms and $N(q, \dot{q})$ includes gravity and other forces which act at the joints (e.g., torsional damping, stiffness). The state-space model is given by

$$\frac{d}{dt}\begin{bmatrix} q \\ \dot{q} \end{bmatrix} = \begin{bmatrix} \dot{q} \\ -M^{-1}(q)\ (C(q, \dot{q}) + N(q, \dot{q})) \end{bmatrix} + \begin{bmatrix} 0 \\ M^{-1}(q) \end{bmatrix} F \tag{2.116}$$

We define the *generalized momenta* $p = (p_1, \ldots, p_n) \in \mathbb{R}^n$ in terms of the generalized coordinates q and Lagrangian L via the *Legendre transformation*

$$p_i = \frac{\partial L}{\partial \dot{q}_i} \qquad i \in \underline{n} \tag{2.117}$$

We define the *Hamiltonian*, in terms of the generalized positions and momenta, as the sum of the kinetic and potential energies of the system, i.e.,

$$H(q, p) = K(q, p) + U(q) \tag{2.118}$$

The Hamiltonian H and Lagrangian L are related by

$$H(q, p) = \langle p, \dot{q} \rangle - L(q, \dot{q}) \tag{2.119}$$

The equations of motion can be restated in coordinates (q, p), in terms of the Hamiltonian H, mass matrix M, and Rayleigh dissipation function R (reformulated in terms of q and p), in the obvious vector notation as

$$\frac{d}{dt}\begin{bmatrix} q \\ p \end{bmatrix} = \begin{bmatrix} \dfrac{\partial H}{\partial p}(q, p) \\ -\dfrac{\partial H}{\partial q}(q, p) - M(q)\ \dfrac{\partial R}{\partial p}(q, p) \end{bmatrix} + \begin{bmatrix} 0 \\ \mathbb{I} \end{bmatrix} F \tag{2.120}$$

Remark 2.7.2 *An advantage of this formulation is that the equations of motion immediately constitute a control system in standard state-space form.* □

Remark 2.7.3 *Using either formulation, the equations of motion yield a state-space model of dimension* $2n$*, i.e., two state variables per DOF. We refer to such a model as a* second-order system*.* □

Remark 2.7.4 *A system (2.120) for which there is no dissipation or forcing is referred to as* Hamiltonian *or* conservative. *This terminology reflects the fact that the total energy of the system remains constant, i.e.,* $\dot{H} = 0$. *We refer to the system (2.120) as a Hamiltonian system perturbed by dissipation and forcing.* □

The *Poisson bracket* is generally a bilinear map from $C^\infty(M) \times C^\infty(M)$ into $C^\infty(M)$, where M is a manifold, satisfying the properties of skew-symmetry, Jacobi identity, and the Leibniz rule (see [121]). We will use a special case of the Poisson bracket where $M = \mathbb{R}^n \times \mathbb{R}^n$, i.e., represents the space of generalized positions and momenta, defined as follows.

Definition 2.7.5 (Poisson Bracket) *Let* $F : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ *and* $G : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ *be smooth functions. The* Poisson bracket *of* F *and* G *is the bilinear map defined by*

$$\{F, G\}(q, p) = \sum_{i=1}^{n} \left(\frac{\partial F}{\partial p_i} \frac{\partial G}{\partial q_i} - \frac{\partial F}{\partial q_i} \frac{\partial G}{\partial p_i} \right)(q, p) \qquad (q, p) \in \mathbb{R}^n \times \mathbb{R}^n \tag{2.121}$$

We will use the following lemma in Chapter 4.

Lemma 2.7.6 *Let* $F : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ *be a functional of the Hamiltonian* H*, i.e.,* $F = F(H(q, p)) \in \mathbb{R}$. *Then* $\{F, H\} = 0$.

Proof

$$\begin{aligned}
\{F, H\} &= \sum_{i=1}^{n} \left[\frac{\partial F}{\partial p_i} \frac{\partial H}{\partial q_i} - \frac{\partial F}{\partial q_i} \frac{\partial H}{\partial p_i} \right] && (2.122)\\
&= \sum_{i=1}^{n} \left[\frac{\partial F}{\partial H} \frac{\partial H}{\partial p_i} \frac{\partial H}{\partial q_i} - \frac{\partial F}{\partial H} \frac{\partial H}{\partial q_i} \frac{\partial H}{\partial p_i} \right] && (2.123)\\
&= \sum_{i=1}^{n} \left[\frac{\partial F}{\partial H} \left(\frac{\partial H}{\partial p_i} \frac{\partial H}{\partial q_i} - \frac{\partial F}{\partial q_i} \frac{\partial H}{\partial p_i} \right) \right] && (2.124)\\
&= 0 && (2.125)
\end{aligned}$$

■

Chapter 3

Standard and Ad-Hoc Approaches to Model Reduction

3.1 Introduction

This chapter introduces two prominent and related approaches for deriving low-order approximations to high-order nonlinear system models, referred to generically as the POD and balanced truncation. Basic versions of these methods have become standard model reduction tools and have been used in a variety of application areas during the past two decades.

Although the POD and balanced truncation are founded in rigorous mathematical results, application of these standard tools to model reduction for nonlinear control systems requires ad-hoc assumptions and procedures. Perhaps the most obvious ad-hoc procedure is linearization, whereby it is assumed that the original nonlinear model can be approximated (locally) by a linear system derived from a Taylor series expansion. The literature regarding applied balanced truncation to this date is concerned only with model reduction for linear systems. The POD, on

the other hand, can be applied to nonlinear control system models, but assumptions about locality, and situational procedures, are still needed. The POD has been applied recently in situations where the input is pre-determined or the system is expected to evolve close to a pre-specified trajectory.

The purpose of this chapter is to

- provide an overview of the state-of-the-art, including important aspects of the underlying theory, computational issues, advantages and shortcomings, and selected applications;
- motivate the research presented in Chapter 4; and
- explain the methods and computational tools used in Chapter 6.

The POD and balancing methods for determining a suitable coordinate transformation are presented, respectively, in Sections 3.2 and 3.3. The general procedure for component truncation is outlined in Section 3.4. We summarize and make some additional remarks in Section 3.5. The subject matter relies heavily on concepts introduced in Sections 2.1-2.5.

3.2 Proper Orthogonal Decomposition

The proper orthogonal decomposition (POD) of a second-order stochastic process is one member of the class of representations known as orthogonal expansions (the Fourier series, or harmonic decomposition, is another example). Its usefulness in the area of model reduction stems from its mathematical properties pertaining to its efficiency in terms of representing an ensemble of signals. The POD is also known as the Karhunen-Loeve expansion (named after two [76, 95] of the several

scientists who are credited with its independent discovery; see [15]), and in certain contexts as principal component analysis (PCA).

An orthogonal expansion of a second-order stochastic process $\{X_t,\, t \in T\}$ is an expression of the form

$$X_t = \sum_{i=1}^{\infty} \sigma_i(t)\, Z_i \qquad t \in T \tag{3.1}$$

where the set $\{Z_1, Z_2, \ldots\}$ is an orthonormal basis for $\mathcal{H}_X$ (see Definition 2.5.29) and the coefficient functions $\{\sigma_i(t) = \langle\, X_t, Z_i \,\rangle\,,\ i = 1, 2, \ldots\}$ are completely deterministic and square-integrable on T. Representations of this form permit the family of random variables $\{X_t,\, t \in T\}$ to be expressed as a linear combination of a countable number of orthonormal random variables $\{Z_1, Z_2, \ldots\}$.

The separable Hilbert space $\mathcal{H}_X$ contains an infinite number of possible orthonormal basis sets $\{Z_1, Z_2, \ldots\}$ (see Proposition 2.4.14). We shall see that the basis derived via the POD is an advantageous choice. In particular, the coefficient functions $\{\sigma_1, \sigma_2, \ldots\}$ form an orthonormal set in $\mathcal{L}_2(T)$, the span of which is capable of representing all members of the ensemble, and the individual terms in the series (3.1) can be ranked according their respective relative contributions to the energy, on average, contained in members of the ensemble. This ranking allows for an efficient representation via truncation of (3.1) at a suitably low index. The POD and its properties are derived mainly using the spectral theory of compact, self-adjoint, integral operators, as described in the following sections.

The exposition contained in this section is based on that from several sources, but we believe that it is original in its treatment, in particular toward illuminating the applied aspects in a rigorous way. Rigorous mathematical treatments are given, from a purely theoretical standpoint in [8, 96, 166], and with a view toward applications to model reduction in [15, 67, 118, 119, 147, 149]. For simplicity we

introduce the main results in the context of one-parameter processes. Later, results are interpreted in terms of two-parameter processes for purposes of application.

3.2.1 Derivation

Consider an ensemble of signals

$$\{X(\omega,\cdot) : [a,b] \to \mathbb{R}^n,\, \omega \in \Omega\} \tag{3.2}$$

Each member of the ensemble (i.e., for each fixed ω) is a function in $\mathcal{L}_2[a,b]$. We assign to the ensemble a probabilistic structure including an associated averaging operation $E[\,\cdot\,]$. The nature of the randomness is not important for the sake of this discussion. It could be due, e.g., to strong dependence on unpredictable initial conditions. We assume that the stochastic process $\{X_t,\, t \in [a,b]\}$ is second-order, and without loss of generality, has zero mean, i.e., $E[\,X_t\,] \equiv 0$.

Now, consider the problem of determining which single deterministic function $\phi \in \mathcal{L}_2[a,b]$, is most similar, on average, to the members of the ensemble, i.e., find ϕ such that

$$\max_{\psi} \frac{E\left[\,\langle\, X_t, \psi \,\rangle_{\mathcal{L}_2}\,\right]}{\|\,\psi\,\|^2_{\mathcal{L}_2}} = \frac{E\left[\,\langle\, X_t, \phi \,\rangle_{\mathcal{L}_2}\,\right]}{\|\,\phi\,\|^2_{\mathcal{L}_2}} \tag{3.3}$$

Remark 3.2.1 *The function ϕ is most nearly parallel to signals in the ensemble, on average, in the function space $\mathcal{L}_2[a,b]$.* □

The maximization problem (3.3) is a classical problem in the calculus of variations. A necessary condition for (3.3) to hold is that ϕ be an eigenfunction of the integral operator with kernel given by the two-point covariance function $R(t,s) = E\left[\,X_t\,X_s^{\mathsf{T}}\,\right]$, i.e.,

$$\int_a^b R(t,s)\,\phi(s)\,ds = \lambda\,\phi(t) \qquad t \in [a,b] \tag{3.4}$$

Remark 3.2.2 *The spectral theory of compact, self-adjoint, integral operators (see, e.g., [150]) ensures that the maximum in (3.3) is achieved and corresponds to the largest eigenvalue* $\lambda_{\max}$ *of the integral operator in (3.4). Furthermore, under the condition that* $[a, b]$ *is bounded, Hilbert-Schmidt theory (see, e.g., [56]) guarantees the existence of a countably infinite number of solutions* $\{\phi_1, \phi_2, \ldots\}$ *of (3.4).* □

The key result is Mercer's theorem (see, e.g., [56]), which gives the *spectral decomposition* of an integral operator with continuous, self-adjoint, non-negative definite kernel.

Theorem 3.2.3 (Mercer) *Let* $k(\cdot, \cdot)$ *be a continuous, Hermitian symmetric, non-negative definite function on* $[a, b] \times [a, b]$*. If* $\{\phi_1, \phi_2, \ldots\}$ *are the orthonormal eigenvectors corresponding to the non-zero eigenvalues* $\{\lambda_1, \lambda_2, \ldots\}$ *of the integral operator with kernel* $k(\cdot, \cdot)$ *then for all* $t, s \in [a, b]$

$$k(t, s) = \sum_{i=1}^{\infty} \lambda_i \, \phi_i(t) \, \phi_i(s) \tag{3.5}$$

The series converges absolutely and uniformly on $[a, b] \times [a, b]$*.* □

Remark 3.2.4 *The spectral decomposition of the covariance is given by*

$$R(t, s) = \sum_{i=1}^{\infty} \lambda_i \, \phi_i(t) \, \phi_i(s) \tag{3.6}$$

where $\{\phi_1, \phi_2, \ldots\}$ *are solutions of the integral equation (3.4). It follows from non-negative definiteness of* R *that the eigenvalues* $\{\lambda_1, \lambda_2, \ldots\}$ *are non-negative. By convention they are ordered such that* $\lambda_i \geq \lambda_{i+1}$*.* □

The POD is a direct consequence of Mercer's theorem.

Theorem 3.2.5 (Proper Orthogonal Decomposition) *Let* $\{X_t,\, t \in [a,b]\}$ *be a zero-mean q.m. continuous second-order stochastic process with covariance function* $R(t,s)$*. The process* X_t *has an orthogonal decomposition*

$$X(\omega,t) = \lim_{N\to\infty} \text{in q.m.} \sum_{i=1}^{N} \sqrt{\lambda_i}\, a_i(\omega)\, \phi_i(t) \qquad t \in [a,b] \tag{3.7}$$

with

$$E\,[\, a_i\, a_j\,] = \delta_{ij} \tag{3.8}$$

and

$$\langle\, \phi_i, \phi_j\, \rangle_{\mathcal{L}_2[a,b]} = \int_a^b \phi_i(t)\, \phi_j(t)\, dt = \delta_{ij} \tag{3.9}$$

if and only if the $\{\phi_1, \phi_2, \ldots\}$ *are the orthonormal eigenfunctions and the* $\{\lambda_1, \lambda_2, \ldots\}$ *are the corresponding eigenvalues of the integral operator with kernel* $R(\cdot,\cdot)$*, i.e.,*

$$\int_a^b R(t,s)\, \phi_i(s)\, ds = \lambda_i\, \phi_i(t) \qquad t \in [a,b] \qquad i = 1,2,\ldots \tag{3.10}$$

In that case, the series (3.7) converges uniformly on $[a,b]$*.*

Proof See Appendix D and [8, 96, 166]. ■

Remark 3.2.6 *The coefficient functions* ϕ_i *that correspond to non-zero eigenvalues* λ_i *(and hence that contribute to the convergent series in (3.5) and (3.7)) are called the* empirical eigenfunctions *of the ensemble. They form an orthonormal basis for the subspace of* $\mathcal{L}_2\,[a,b]$ *to which all members of the ensemble belong (except for a set of measure zero; see Fact 3.2.24).* □

Remark 3.2.7 *The uncorrelated random variables* $\{a_1, a_2, \ldots\}$ *are given by*

$$a_i(\omega) = \left(\sqrt{\lambda_i}\right)^{-1} \int_a^b \phi_i(t)\, X(\omega,t)\, dt \qquad i = 1,2,\ldots \tag{3.11}$$

and form the desired orthonormal basis for $\mathcal{H}_X$*.* □

Remark 3.2.8 *Each eigenvalue λ_i can be interpreted as the mean energy of the signals in the ensemble projected onto the ϕ_i-axis in function space $\mathcal{L}_2[a,b]$ (see, e.g., [118] for the calculation), i.e.,*

$$\lambda_i = E\left[\,|\,\langle\,\phi_i, X_t\,\rangle_{\mathcal{L}_2[a,b]}\,|^2\,\right] \qquad i = 1, 2, \ldots \tag{3.12}$$

This interpretation justifies the ranking of terms in the series (3.7) by relative energy contribution. □

Two Parameter Processes

To model the two-parameter case, consider the ensemble of signals

$$\{X(\omega, \cdot, \cdot) : [0,\infty) \times \mathcal{D} \to \mathbb{R}^n,\, \omega \in \Omega\} \tag{3.13}$$

where $\mathcal{D}$ is bounded. Each member of the ensemble is a function in, respectively, $\mathcal{L}_2[0,\infty)$ and $\mathcal{L}_2(\mathcal{D})$, for fixed x and fixed t. As before, we assign a probabilistic structure and assume that the process $\{X_{t,x},\, t \in [0,\infty),\, x \in \mathcal{D}\}$ is second-order and has zero mean, i.e., $E[\,X_{t,x}\,] \equiv 0$. We also assume that the process is wide-sense stationary and ergodic with respect to t. It then has a spatial covariance function given by

$$R(x,y) = E\left[\,X_{t,x}\,X_{t,y}^{\mathsf{T}}\,\right] = \lim_{T\to\infty} \frac{1}{T} \int_0^T X_{t,x}\,X_{t,y}^{\mathsf{T}}\,dt \tag{3.14}$$

The spectral decomposition and *spatial POD*, respectively, are given by

$$R(x,y) = \sum_{i=1}^{\infty} \lambda_i\,\phi_i(x)\,\phi_i(y) \tag{3.15}$$

where convergence is absolute and uniform on $\mathcal{D} \times \mathcal{D}$, and

$$X(\omega,t,x) = \lim_{N\to\infty} \text{in q.m.} \sum_{i=1}^{N} \sqrt{\lambda_i}\,a_i(\omega,t)\,\phi_i(x) \qquad t \in [0,\infty),\, x \in \mathcal{D} \tag{3.16}$$

where convergence is uniform on $[0,\infty) \times \mathcal{D}$.

Remark 3.2.9 *The coefficient functions ϕ_i that correspond to non-zero eigenvalues λ_i are called the* spatial empirical eigenfunctions *of the ensemble. They are the unit magnitude solutions of the family of integral equations*

$$\int_{\mathcal{D}} R(x,y)\,\phi_i(y)\,dy = \lambda_i\,\phi_i(x) \qquad x \in \mathcal{D} \qquad i = 1,2,\ldots \tag{3.17}$$

and form an orthonormal basis for the subspace of $\mathcal{L}_2(\mathcal{D})$ to which all members of the ensemble belong for each fixed t (except for a set of measure zero). □

Remark 3.2.10 *The random functions $\{a_1, a_2, \ldots\}$ are given by*

$$a_i(\omega,t) = (\sqrt{\lambda_i})^{-1} \int_{\mathcal{D}} \phi_i(x)\,X(\omega,t,x)\,dx \qquad i = 1,2,\ldots \tag{3.18}$$

They are stochastic processes, inherit ergodicity from $X_{t,x}$, and are uncorrelated in the sense that

$$E\left[\,a_i(t)\,a_j(t)\,\right] = \lim_{T\to\infty} \int_0^T a_i(t)\,a_j(t)\,dt = \delta_{ij} \tag{3.19}$$

□

Remark 3.2.11 *Sirovich [147] refers to the spatial empirical eigenfunctions $\{\phi_1, \phi_2, \ldots\}$ as* coherent structures. *This terminology stems from the interpretation of the signals as realizations of a physical flow in time and space (e.g., fluid momentum, heat). The empirical eigenfunctions then correspond to physically manifested, coherent spatial structures in the flow.* □

Remark 3.2.12 *Each eigenvalue λ_i can be interpreted as the mean energy of the signals in the ensemble projected onto the ϕ_i-axis in function space $\mathcal{L}_2(\mathcal{D})$, and equivalently, by ergodicity with respect to time, as the average relative time spent by signals in the ensemble along the ϕ_i-axis, i.e.,*

$$\lambda_i = E\left[\,|\,\langle\,\phi_i, X_{t,x}\,\rangle_{\mathcal{L}_2(\mathcal{D})}\,|^2\,\right] = \lim_{T\to\infty} \frac{1}{T} \int_0^T \left|\langle\,\phi_i, X_{t,x}\,\rangle_{\mathcal{L}_2(\mathcal{D})}\right|^2 dt \qquad i = 1,2,\ldots \tag{3.20}$$

Remark 3.2.13 *We note that since the time domain is unbounded (which typically models the evolution of a dynamical system), there is no "temporal POD" for this two-parameter case. This is due to the fact that the integral operator with kernel $R(t,s)$ is compact if and only if its domain is bounded, and hence has no spectral decomposition in the case of unbounded domain.* □

Sampled Data Processes

In most practical applications, we work with processes that are sampled in time or space or both. Consider the ensemble of sampled signals

$$\{X(\omega,\cdot) : \{1,2,\ldots\} \to \mathbb{R}^n,\, \omega \in \Omega\} \tag{3.21}$$

Each member of the ensemble is a vector in ℓ_2. As usual, we assume that the process $\{X_k,\, k = 1,2,\ldots\}$ is second-order, and without loss of generality, has zero mean, i.e., $E[\,X_k\,] \equiv 0$. The discrete covariance function is given by

$$R(j,k) = E\left[\,X_j\,{X_k}^{\mathsf{T}}\,\right] \tag{3.22}$$

which, in the case that X_k is scalar valued, can be written as a matrix $R = [R]_{jk} = [R(j,k)]$ (if X_k is vector valued then R can be written as a fourth-order tensor). The matrix (tensor) R is real, symmetric, and non-negative definite.

The spectral theorem (see, e.g., [154]) states that every real symmetric matrix R can be diagonalized by an orthogonal matrix, i.e., there exists an orthogonal matrix Φ and a real diagonal matrix $\Lambda = \mathrm{diag}\,(\lambda_1,\ldots,\lambda_n)$ such that $R = \Phi\,\Lambda\,\Phi^{\mathsf{T}}$. We write the spectral decomposition of R in (3.22) as

$$R = \sum_{i=1}^{\infty} \lambda_i\,\phi_i\,{\phi_i}^{\mathsf{T}} \tag{3.23}$$

where each vector ϕ_i is the i-th column of Φ.

Remark 3.2.14 *The operator given by multiplication by the covariance matrix is always compact, even with an infinite number of rows and columns. We need not worry about boundedness of the time domain here (i.e., even though $k = 1, 2, \ldots$, the spectral decomposition exists).* □

Remark 3.2.15 *Because the covariance matrix R is non-negative definite, the eigenvalues $\{\lambda_1, \lambda_2, \ldots\}$ are non-negative. By convention they are ordered such that $\lambda_i \geq \lambda_{i+1}$.* □

The *sampled data POD* is a direct consequence of the spectral theorem.

Theorem 3.2.16 (Sampled Data POD) *Let $\{X_k,\ k = 0, 1, \ldots\}$ be a zero-mean scalar-valued discrete-parameter second-order stochastic process with covariance matrix R. The process X_k has an orthogonal decomposition*

$$X(\omega, k) = \lim_{N \to \infty} \text{in q.m.} \sum_{i=1}^{N} \sqrt{\lambda_i}\, a_i(\omega)\, (\phi_i)_k \qquad k = 0, 1, \ldots \tag{3.24}$$

with

$$E\,[\, a_i\, a_j\,] = \delta_{ij} \tag{3.25}$$

and

$$\langle\, \phi_i, \phi_j\, \rangle = \phi_i{}^{\mathsf{T}}\, \phi_j = \delta_{ij} \tag{3.26}$$

if and only if the $\{\phi_1, \phi_2, \ldots\}$ are the orthonormal eigenvectors and the $\{\lambda_1, \lambda_2, \ldots\}$ are the corresponding eigenvalues of the matrix R, i.e.,

$$R\,\phi_i = \lambda_i\,\phi_i \qquad i = 1, 2, \ldots \tag{3.27}$$

Proof See Appendix D. ■

Remark 3.2.17 *The vectors ϕ_i corresponding to non-zero eigenvalues λ_i are called the* empirical eigenvectors *of the ensemble. They form an orthonormal basis for*

the subspace of ℓ_2 to which all members of the ensemble belong (except for a set of measure zero). □

Remark 3.2.18 *The uncorrelated random variables $\{a_1, a_2, \ldots\}$ are given by*

$$a_i(\omega) = \left(\sqrt{\lambda_i}\right)^{-1} \sum_{k=1}^{\infty} (\phi_i)_k \, X(\omega, k) \tag{3.28}$$

and form the desired orthonormal basis for $\mathcal{H}_X$. □

Remark 3.2.19 *It is often convenient to express the scalar-valued sampled data process $\{X_k,\, k = 0, 1, \ldots\}$ as a random vector $X = [X_1, X_2, \ldots]^{\mathsf{T}}$. In this case the covariance matrix is given by $R = E\left[\, X\, X^{\mathsf{T}} \,\right]$ and the sampled data POD (3.24) is written compactly as*

$$X = \Phi\, \Lambda^{1/2}\, a = \Phi\, b \tag{3.29}$$

where $\Phi = [\Phi]_{ki} = [(\phi_i)_k]$ is an orthogonal matrix whose columns are the empirical eigenvectors, $\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \ldots)$, and

$$a = [a_1, a_2, \ldots]^{\mathsf{T}} = \Lambda^{-1/2}\, \Phi^{\mathsf{T}}\, X \tag{3.30}$$

and $b = \Lambda^{1/2}\, a$ are random vectors. The mean energy interpretation of the eigenvalues is expressed by

$$\lambda_i = E\left[\, |\phi_i{}^{\mathsf{T}}\, X|^2 \,\right] \qquad i = 1, 2, \ldots \tag{3.31}$$

□

Two Parameter Sampled Data Processes

The typical situation that arises in dynamical systems applications is that of an ensemble of signals that evolves on a time continuum but whose spatial domain has been discretized. Moreover it is usually the case that the spatial domain is

bounded, resulting in a finite number of samples in the spatial parameter (e.g., time evolution of the temperature field in a solid body discretized via finite-elements).

Consider an ensemble of such signals

$$\{X(\omega,\cdot,\cdot) : [0,\infty) \times \{1,\ldots,n\} \to \mathbb{R}^n,\, \omega \in \Omega\} \tag{3.32}$$

We assume as before that the process

$$\{X_{t,k},\, t \in [0,\infty),\, k = 1,\ldots,n\}$$

is second-order and has zero mean, i.e., $E[\,X_{t,k}\,] \equiv 0$. We also assume that the process is wide-sense stationary and ergodic with respect to t. The discrete spatial covariance function is given by

$$R(j,k) = E[\,X_{t,j}\,X_{t,k}\,] = \lim_{T\to\infty} \int_0^T X_{t,j}\,X_{t,k}\,dt \tag{3.33}$$

which as before can be expressed as a matrix $R = [R]_{jk} = [R(j,k)]$ with spectral decomposition $R = \sum_{i=1}^{n} \lambda_i\,\phi_i\,\phi_i^{\mathsf{T}}$.

The *sampled-data spatial POD* is given by

$$X(\omega,t,k) = \sum_{i=1}^{n} \sqrt{\lambda_i}\,a_i(\omega,t)\,(\phi_i)_k \qquad k = 1,\ldots,n \tag{3.34}$$

Remark 3.2.20 *The vectors ϕ_i corresponding to non-zero eigenvalues λ_i are called the* spatial empirical eigenvectors *of the ensemble. They are the unit length solutions of (3.27) (i.e., unit eigenvectors of the matrix R) and form an orthonormal basis for the subspace of $\mathbb{R}^n$ to which all members of the ensemble belong for each fixed t (except for a set of measure zero).* □

Remark 3.2.21 *The random functions $\{a_1,\ldots,a_n\}$ are given by*

$$a_i(\omega,t) = \left(\sqrt{\lambda_i}\right)^{-1} \sum_{k=1}^{n} (\phi_i)_k\,X(\omega,t,k) \tag{3.35}$$

They are stochastic processes, inherit ergodicity from $X_{t,k}$, and are uncorrelated in the sense of (3.19). □

Remark 3.2.22 *In applications, the continuous-time scalar-valued sampled spatial data process*

$$\{X_{t,k},\, t \in [0,\infty),\, k = 1,\ldots,n\}$$

is often expressed as a one-parameter vector process $X_t = [X_{t,1},\ldots,X_{t,n}]^{\mathsf{T}}$. *In this case the spatial covariance matrix is given by*

$$R = E\left[\, X_t\, X_t^{\mathsf{T}}\,\right] = \lim_{T\to\infty} \int_0^T X_t\, X_t^{\mathsf{T}}\, dt \tag{3.36}$$

and the sampled data spatial POD (3.34) is written compactly as

$$X_t = \Phi\, \Lambda^{1/2}\, a(t) = \Phi\, b(t) \tag{3.37}$$

where $\Phi = [\Phi]_{ki} = [(\phi_i)_k]$ *is an orthogonal* $n \times n$*-matrix whose columns are the spatial empirical eigenvectors,* $\Lambda = \operatorname{diag}(\lambda_1,\ldots,\lambda_n)$, *and*

$$a(t) = [a_1(t),\ldots,a_n(t)]^{\mathsf{T}} = \Lambda^{-1/2}\, \Phi^{\mathsf{T}}\, X_t \tag{3.38}$$

and $b(t) = \Lambda^{1/2}\, a(t)$ *are vector processes. The mean energy and average time duration interpretations of the eigenvalues are expressed, respectively, by*

$$\lambda_i = E\left[\, \left|\phi_i{}^{\mathsf{T}}\, X_t\right|^2 \,\right] = \lim_{T\to\infty} \int_0^T \left|\phi_i{}^{\mathsf{T}}\, X_t\right|^2 dt \qquad i = 1,\ldots,n \tag{3.39}$$

□

Remark 3.2.23 *The relationship between the POD of a second-order stochastic process and the PCA of a matrix valued signal (see Section 2.3) becomes apparent by observing that the infinite time horizon Gramian matrix* W^2 *given by (2.29) for signal* $X(t)$ *corresponds to the two-point spatial covariance matrix* R. *The sampled-data spatial POD (3.34) is then the same as the PCA (2.30) where the square root of the* λ_i *have been subsumed into the coefficient functions* $a_i(t)$. □

3.2.2 Properties

There are two properties of the POD in which we are most interested here. The first says that the POD does in fact produce a representation that is capable of describing all of the observed phenomena from which it was derived. The other is the main property of interest in the context of model reduction. It says that the POD is optimal, or most efficient, in terms of modeling the signal set with the fewest number of modes.

Recall the ensemble of signals

$$\{X(\omega, \cdot) : [a, b] \to \mathbb{R}^n,\, \omega \in \Omega\} \tag{3.40}$$

and suppose that the associated POD orthonormal basis set is $\{\phi_1, \phi_2, \ldots\}$. We define the *span of the empirical basis* as the collection of functions that can be represented by a convergent sequence of a linear combination of empirical eigenfunctions, i.e.,

$$S_\phi = \left\{ \sum_{i=1}^{\infty} \alpha_i \, \phi_i : \sum_{i=1}^{\infty} \alpha_i < \infty \right\} \tag{3.41}$$

Similarly, the span of all members of the ensemble is given by

$$S_X = \left\{ \sum_{i=1}^{\infty} \beta_i \, X(\omega_i, t) : \sum_{i=1}^{\infty} \beta_i < \infty \right\} \tag{3.42}$$

It is shown in [15] that

Fact 3.2.24 *The sets S_ϕ and S_X are equivalent with the exception of a set of measure zero.* □

Remark 3.2.25 *Thus, every member of the ensemble that generated the empirical eigenfunctions, and linear combinations thereof, can be represented by a convergent series of a linear combination of the empirical eigenfunctions.* □

The optimality of the POD in terms of modeling a signal from among an ensemble of signals is expressed by the following result (see, e.g., [15, 147]).

Proposition 3.2.26 (Optimality of the POD) *Consider the ensemble*

$$\{X(\omega, \cdot) : [a, b] \to \mathbb{R}^n, \, \omega \in \Omega\}$$

and let $\{\phi_1, \phi_2, \ldots\}$ *be the empirical eigenfunctions with corresponding eigenvalues* $\{\lambda_1, \lambda_2, \ldots\}$. *Let*

$$\{X(\bar{\omega}, t), \, \bar{\omega} \in \Omega, \, t \in [a, b]\}$$

be a member of the ensemble with POD

$$X(\bar{\omega}, t) = \sum_{i=1}^{\infty} b_i(\bar{\omega}) \, \phi_i(t) \tag{3.43}$$

where the eigenvalues have been subsumed into the random coefficients, i.e., for each i, $b_i(\omega) = \sqrt{\lambda_i} \, a_i(\omega)$. *Let* $\{\psi_1, \psi_2, \ldots\}$ *be an arbitrary orthonormal set such that for some random variables* $\{c_1, c_2, \ldots\}$

$$X(\bar{\omega}, t) = \sum_{i=1}^{\infty} c_i(\bar{\omega}) \, \psi_i(t) \tag{3.44}$$

Then for each truncation index N

$$\sum_{i=1}^{N} E\left[\, | \, \langle \, \phi_i, X_t \, \rangle \, | \, \right] = \sum_{i=1}^{N} \lambda_i \geq \sum_{i=1}^{N} E\left[\, | \, \langle \, \psi_i, X_t \, \rangle \, | \, \right] \tag{3.45}$$

and equivalently

$$\sum_{i=1}^{N} E\left[\, \| \, b_i \, \|^2 \, \right] = \sum_{i=1}^{N} \lambda_i \geq \sum_{i=1}^{N} E\left[\, \| \, c_i \, \|^2 \, \right] \tag{3.46}$$

□

Remark 3.2.27 *Thus, for any given number of modes* N, *the projection of any member of the ensemble onto the subspace spanned by the most energetic* N *members of the empirical basis will contain more energy, on average, than the projection onto the subspace spanned by* N *members of any other orthonormal basis.* □

Remark 3.2.28 *The optimality property may also be interpreted as a minimization of the error, on average, between members of the ensemble and the truncated orthogonal expansion. To see this, observe that the minimum of the mean-squared error*

$$E\left[\left\| X_t - \sum_{i=1}^{N} c_i(\omega)\,\psi_i(t)\right\|^2\right] = E\left[\,\| X_t \|^2\,\right] + \sum_{i=1}^{N} E\left[\,\| c_i\,\psi_i \|^2\,\right] - 2\sum_{i=1}^{N} E\left[\,\langle\, c_i\,\psi_i, X_t \,\rangle\,\right] \tag{3.47}$$

is achieved when $\sum_{i=1}^{N} E\left[\,\langle\, c_i\,\psi_i, X_t \,\rangle\,\right]$ *is maximized. Equation (3.45) gives the empirical basis as the maximizing orthonormal set.* □

Remark 3.2.29 *There exists no explicit error bound, e.g., corresponding to (1.1), in terms of the eigenvalues or otherwise.* □

Remark 3.2.30 *In the two-parameter case, the optimality properties (3.45) and (3.46) are expressed, respectively, as*

$$\sum_{i=1}^{N} E\left[\,|\,\langle\,\phi_i(x), X_{t,x}\,\rangle\,|\,\right] = \sum_{i=1}^{N} \lambda_i \geq \sum_{i=1}^{N} E\left[\,|\,\langle\,\psi_i(x), X_{t,x}\,\rangle\,|\,\right] \tag{3.48}$$

and equivalently

$$\sum_{i=1}^{N} E\left[\,\| b_i(t)\,\|^2\,\right] = \sum_{i=1}^{N} \lambda_i \geq \sum_{i=1}^{N} E\left[\,\| c_i(t)\,\|^2\,\right] \tag{3.49}$$

for each N*, where*

$$X(\bar{\omega}, t, x) = \sum_{i=1}^{\infty} b_i(\bar{\omega}, t)\,\phi_i(x) = \sum_{i=1}^{\infty} c_i(\bar{\omega}, t)\,\psi_i(x) \tag{3.50}$$

□

3.2.3 Computation

The computational aspects of the POD are crucial to analyzing its advantages and shortcomings as a model reduction methodology. We consider the issues of centering, practical computation of the empirical basis, and practical derivation of low-order models for nonlinear control systems.

Centering and Zero-Mean Processes

Consider the ensemble,

$$\{Z(\omega, \cdot) : T \to \mathbb{R}^n, \, \omega \in \Omega\} \tag{3.51}$$

It is often the case that members of the ensemble are very similar to each other in the sense of (3.3) because their respective differences are small in magnitude compared with the magnitude of the signals themselves. This centering problem is addressed simply by subtracting out the average signal $\mu(t) = E[\, Z_t \,]$ from each ensemble member, i.e.,

$$X_t = Z_t - \mu(t) \tag{3.52}$$

to give the zero-mean process X_t.

Remark 3.2.31 *The process X_t represents the deviation or fluctuation from the mean signal. The recentering minimizes the effect of noise and numerical error in computations and generates a POD basis set and corresponding ranking of modes that more accurately reflects the differences in energy content between signals. This justifies our emphasis on zero-mean processes in Section 3.2.1.* □

Remark 3.2.32 *The centering is similarly applied to the two-parameter and sampled data cases, i.e.,*

$$\begin{aligned} X_{t,x} &= Z_{t,x} - \mu(t,x) & \mu(t,x) &= E\left[\, Z_{t,x} \,\right] \\ X_{k} &= Z_{k} - \mu(k) & \mu(k) &= E\left[\, Z_{k} \,\right] \\ X_{t,k} &= Z_{t,k} - \mu(t,k) & \mu(t,k) &= E\left[\, Z_{t,k} \,\right] \end{aligned} \tag{3.53}$$

□

Computing the Empirical Basis

Here we present standard methods for computing the empirical basis. Consider the typical dynamical systems application where $X_t = [X_{t,1}, \ldots, X_{t,n}]^{\mathsf{T}}$ is a zero-mean vector process representing the fluctuation from the mean of a physical flow in time (continuous) and space (discretized), or possibly some other multi-variable state evolution in continuous time. We make the standard assumptions as usual.

We need to compute an approximation to the spatial covariance matrix

$$R = E\left[\, X_t \, X_t^{\mathsf{T}} \,\right] = \lim_{T \to \infty} \frac{1}{T} \int_0^T X_t \, X_t^{\mathsf{T}} \, dt \tag{3.54}$$

This is accomplished by sampling the flow X_t at times $\{t_1, t_2, \ldots\}$, i.e., capturing "snapshots"

$$\{X(t_1), X(t_2), \ldots\}$$

The times can be equally spaced by a fixed interval τ, i.e., $t_\nu = (\nu - 1)\,\tau$ (mimicking the action of a strobe). We define the approximation

$$\widehat{R} = \lim_{M \to \infty} \frac{1}{M} \sum_{\nu=1}^{M} X(t_\nu) \, X^{\mathsf{T}}(t_\nu) \tag{3.55}$$

For all practical purposes, we must make another approximation and use only a fixed finite number M of samples, i.e.,

$$\widehat{R}_M = \frac{1}{M} \sum_{\nu=1}^{M} X(t_\nu) \, X^{\mathsf{T}}(t_\nu) \tag{3.56}$$

Remark 3.2.33 *The approximation improves as the number of samples M increases. The actual number of samples that are captured and used will depend on practical considerations. However, it is reasonable to assume that $M \ll n$.* □

Remark 3.2.34 *In the applied literature, one often sees equation (3.56) written as*

$$\widehat{R}_M = \frac{1}{M} \Xi \, \Xi^{\mathsf{T}} \tag{3.57}$$

where

$$\Xi = [\Xi]_{k\nu} \stackrel{\triangle}{=} [X_k(t_\nu)] \tag{3.58}$$

is a $n \times M$ matrix called the data matrix. *Actually, it is common for authors to ignore the $1/M$ factor and the fact that an approximation is being made, and to write $R = \Xi \, \Xi^{\mathsf{T}}$. Also, this procedure is commonly mistaken for the "method of snapshots," which refers to something somewhat different, to be described shortly.* □

Remark 3.2.35 *The approximate spatial covariance has only M non-zero eigenvalues and hence M approximate empirical eigenvectors. Thus, the span of the empirical basis is at most an M-dimensional subspace of $\mathbb{R}^n$.* □

Now we can use standard matrix algebra algorithms to compute the spectral decomposition

$$\widehat{R}_M = \Phi \, \Lambda \, \Phi^{\mathsf{T}} \tag{3.59}$$

The approximate empirical eigenvectors are given by the first M columns $\{\phi_1, \ldots, \phi_M\}$ of the $n \times n$ orthogonal matrix Φ. They correspond to the M non-zero eigenvalues $\{\lambda_1, \ldots, \lambda_M\}$, i.e., the non-zero diagonal entries of Λ (in decreasing order).

Remark 3.2.36 *In the applied literature, the spectral decomposition of* $\widehat{R}_M$ *is usually accomplished by computing the singular value decomposition (SVD) of the data matrix* Ξ*, i.e.,*

$$\Xi = \Phi \, \Sigma \, \Psi^{\mathsf{T}} \tag{3.60}$$

where Φ *is the* $n \times n$ *orthogonal matrix containing the empirical eigenvectors,* Σ *is* $n \times M$ *given by*

$$\Sigma = \begin{bmatrix} \operatorname{diag}(\sigma_1, \ldots, \sigma_M) \\ \\ 0 \end{bmatrix} = \begin{bmatrix} (M \operatorname{diag}(\lambda_1, \ldots, \lambda_M))^{1/2} \\ \\ 0 \end{bmatrix} \tag{3.61}$$

and Ψ *is* $M \times M$ *and orthogonal. Thus, the empirical eigenvectors are computed correctly via the SVD.*

The ordering of the σ_i *remains the same as that of the* λ_i *(and hence also the ranking of modes). However, the precise meaning via relative energy resides with the eigenvalues. Sometimes authors have ignored the square root operation and mistakenly based their truncation analysis on the singular values. Moreover, it is common for authors to claim that the POD and the SVD are equivalent procedures, which ignores the underlying theory, assumptions, and approximations pertaining to the POD. It is more accurate to think of the SVD as a tool that can be used in practical computation of the POD.* □

Computational difficulties occur when the dimension n of X_t is large, e.g., due to a high resolution spatial discretization. In particular, this forces the spectral decomposition of a large matrix, possibly requiring very high computational expense, and possibly leading to the accumulation of large numerical errors. Sirovich [147] introduced a method to deal with this situation, coining the names *"method of snapshots"* and *"method of strobes"*. It is a method for computing the empirical

eigenvectors without forming and decomposing the very large two-point spatial covariance matrix.

Assume that the snapshots $X(t_\nu)$ are linearly independent vectors in $\mathbb{R}^n$, i.e., the data matrix Ξ is has full column rank, and define the $M \times M$ matrix

$$\widehat{C}_M = \frac{1}{M} \Xi^{\mathsf{T}} \Xi \tag{3.62}$$

Let $\{\psi_1, \ldots, \psi_M\}$ and $\{\mu_1, \ldots, \mu_M\}$ be, respectively, the eigenvectors and eigenvalues of $\widehat{C}_M$, i.e.,

$$\mu_i \psi_i = \widehat{C}_M \psi_i = \frac{1}{M} \Xi^{\mathsf{T}} \Xi \psi_i \qquad i \in \underline{M} \tag{3.63}$$

A simple calculation reveals that if we define

$$\phi_i = \Xi \psi_i \qquad i \in \underline{M} \tag{3.64}$$

then

$$\mu_i \phi_i = \frac{1}{M} \Xi \Xi^{\mathsf{T}} \phi_i = \widehat{R}_M \phi_i \qquad i \in \underline{M} \tag{3.65}$$

Remark 3.2.37 *Thus, the eigenvalues μ_i of the $M \times M$ matrix $\widehat{C}_M$ correspond exactly to the non-zero eigenvalues λ_i of the $n \times n$ matrix $\widehat{R}_M$. The relationship between the empirical eigenvectors ϕ_i and the eigenvectors ψ_i of $\widehat{C}_M$ is given by (3.64), i.e., the empirical eigenvectors are linear combinations of the snapshots $X(t_\nu)$.* □

Remark 3.2.38 *The advantage of the method of snapshots is that we compute the empirical eigenvectors via spectral decomposition of the $M \times M$ matrix $\widehat{C}_M$ instead of the much larger $n \times n$ matrix $\widehat{R}_M$.* □

Remark 3.2.39 *In the literature, sometimes $\widehat{C}_M$ is referred to as the covariance matrix. It is actually a temporal covariance. Time sampling produces a compact*

operator which admits a spectral decomposition yielding temporal empirical eigenvectors, which are related to the spatial empirical eigenvectors by (3.64). □

Deriving the Reduced-Order Model

Here we describe the typical procedure for deriving a reduced-order model from the full-order model. Our presentation is in the context of finite-dimensional control systems, although it generalizes easily to the infinite-dimensional setting (see, e.g., [118, 119]).

We assume that the full-order state-space model has been derived via first principles or empirical analysis or both, with autonomous state equation and output equation, respectively,

$$\dot{x} = f(x, u) \tag{3.66}$$

$$y = h(x) \tag{3.67}$$

where $x \in \mathbb{R}^n$ represents local coordinates for the full-order state. It is also possible that the actual physical system being modeled is available for experimentation.

An ensemble of signals is needed to compute the empirical basis. In order for the POD to yield an efficient basis for an orthogonal expansion, the signals in the ensemble should represent or capture the essential system behavior. This means that the modeler must choose one or more sets of inputs and initial conditions to produce what he deems to be the representative system response. An ensemble of state trajectories is then generated via numerical simulation of the state equation (3.66), or possibly via experimentation if practical. A suitable set of sampling times must be chosen to determine the snapshot data which forms the data matrix.

Given the data matrix, the empirical eigenvectors $\{\phi_1, \ldots, \phi_M\}$ corresponding to the non-zero eigenvalues $\{\lambda_1, \ldots, \lambda_M\}$ are computed via the POD, whether by

direct SVD or the Sirovich method of snapshots, and arranged into the matrix

$$\Phi = [\phi_1, \ldots, \phi_M] \in \mathbb{R}^{n \times M} \tag{3.68}$$

The relative magnitudes of the eigenvalues are then analyzed in order to choose a truncation index $k < M \ll n$ such that, from the viewpoint of the modeler, the resulting model order is sufficiently low while retaining sufficiently high signal energy on average, i.e., provides a favorable tradeoff between fidelity and complexity. This analysis yields a truncated transformation matrix

$$\Phi_k = [\phi_1, \ldots, \phi_k] \in \mathbb{R}^{n \times k} \tag{3.69}$$

The reduced-order state $z \in \mathbb{R}^k$ and the approximate reconstruction of the original full-order state $\widehat{x} \in \mathbb{R}^n$, respectively, are defined by

$$z \triangleq \Phi_k{}^{\mathsf{T}} x \tag{3.70}$$

$$\widehat{x} \triangleq \Phi_k z \tag{3.71}$$

The reduced state equation is computed via Galerkin projection

$$\dot{z} = \Phi_k{}^{\mathsf{T}} \dot{x} = \Phi_k{}^{\mathsf{T}} f(x, u) \approx \Phi_k{}^{\mathsf{T}} f(\Phi_k z, u) \tag{3.72}$$

yielding the reduced-order control system

$$\dot{z} = \widehat{f}(z, u) \tag{3.73}$$

$$\widehat{y} = \widehat{h}(z) \tag{3.74}$$

where the reduced system map $\widehat{f} : \mathbb{R}^k \times \mathcal{U} \to \mathbb{R}^k$ and reduced output map $\widehat{h} : \mathbb{R}^k \to \mathbb{R}^p$ are defined, respectively, by

$$\widehat{f}(z, u) \triangleq \Phi_k{}^{\mathsf{T}} f(\Phi_k z, u) \qquad \widehat{h}(z) \triangleq h(\Phi_k z) \tag{3.75}$$

and $\widehat{y} \in \mathbb{R}^p$ is the approximate reconstruction of the output.

The reduced state equation (3.73) is numerically integrated using initial reduced-order state $z(0) = {\Phi_k}^{\mathsf{T}} x(0)$ where $x(0)$ corresponds to the desired initial full-order state. This produces the reduced-order state trajectory $z(t)$, from which an approximation to the full-order trajectory is reconstructed via $\widehat{x}(t) = {\Phi_k}^{\mathsf{T}} z(t)$.

Remarks

We conclude this section with some remarks on the advantages and drawbacks of the POD as a tool for computing a coordinate transformation for state-space model reduction.

- The modeler has a great deal of discretion in determining the reduced-order model. He chooses the ensemble of signals (via choice of inputs and initial conditions), sampling times, and truncation index. While the POD is a natural tool for efficiently representing signals in an ensemble, it is merely part of an ad-hoc procedure for reducing the order of a dynamical system. It does not work directly with the system map f.

- The POD is optimal for efficiently representing signals that belong to an ensemble, e.g., state trajectories generated via simulation of the ODE (3.66) for a chosen set of inputs. However, because the admissible controls constitute a much larger set, the resulting family of ODEs will produce trajectories that do not belong to the ensemble. The efficiency of the POD in terms of representing all possible state trajectory signals is unknown.

- There have been no results of which we are aware regarding a rigorous or systematic methodology for generating a representative ensemble of signals to characterize the state response of a nonlinear control system.

- Snapshot data may fail to capture dynamical effects occurring at widely differing time scales.

- The state-to-output relationship $y = h(x)$ is not used in determining the empirical basis. Since the output consists of variables of particular interest, it would appear that ignoring this relationship is to the method's detriment.

3.2.4 Applications

The POD has become prominent as a tool for complexity reduction during the 1980s and 1990s, finding application in a wide variety of areas. Here we describe some examples in order to illustrate its capabilities for efficient representation and the ad-hoc nature of the procedure as applied to model reduction for nonlinear dynamical systems. We emphasize applications to order reduction for RTP models in order to provide background and motivation for subsequent material.

The capabilities of the POD become apparent in the context of data compression for image processing, for which the POD is a natural tool. For example, in [146], the authors apply the POD to compress the amount of data needed to reconstruct pixelized images of human faces ($2^7 \times 2^7$ pixels with 2^8 gray levels). In their study, linear combinations of 40 dominant "empirical eigenfaces" are capable of representing face images, both within and outside of the original population (115 faces), to within 3% error, thus reducing the dimension of the representation space from 2^{22} to 40.

Various studies have demonstrated the effectiveness of the POD as a tool for deriving low-order characterizations of a spatially distributed flow $v(t, x)$ representing the time evolution of some physical phenomenon. Examples include applications to pipe flow in a wall region, [11], Rayleigh-Benard convective flow [40], turbulent

channel flow [12], and vibration of a thin membrane in a stadium [19]. In these studies, the ensemble of signals $\{v(t,x)\}$ is a collection of realizations for the time-varying flow field, i.e., the unique solutions of a family of initial value problems. In each study it is shown that the pre-computed flow can be represented with high accuracy using a number of empirical eigenfunctions that is relatively small compared with the discretization resolution used to simulate the original evolution equations ($O\left(10^3\right)$ reduction is consistently achieved). Moreover, the structural aspect of the eigenfunctions is observed, e.g., as rolls and shearing motions in [12].

However, it is important to note that, in these examples, low-order approximations to the original evolution equations are not derived. Rather, the focus is on deriving low-order representations of pre-computed flow fields, a task naturally suited for the POD. Thus, these applications do not necessarily fall within the realm of what we consider to be model reduction. The subject of deriving low-order evolution equations for spatially distributed flows using the POD basis is covered in, e.g., [148], but examples are not offered and computational issues such as those pointed out in Section 3.2.3 are not addressed.

During the 1990's, the POD has appeared in various ad-hoc methodologies related to the control of state-space models for dynamical systems. Strategies for control of turbulent flows are proposed in [99], where the authors use low-dimensional models and knowledge of the extracted coherent structures to determine how, when, and where to interfere with the flow in the boundary layer. In [25], a nonlinear feedback law is constructed which requires information only about the dominant empirical eigenfunctions without using the original nonlinear model.

There has been much recent activity in the development of models to be used for control of the temperature distribution on a semiconductor wafer in a RTP

chamber. Dynamic models of heat transfer in a generic RTP system with 5 lamp banks are presented in [5, 6]. The full-order model has 116 states, each representing the temperature at a physical location in the chamber. In [6], an ensemble of representative trajectories is generated by simulating the state equation using a collection of control inputs consisting of a nominal optimal control (a time-varying power setting for each lamp bank) and several perturbations via pseudo-random binary sequences. It is then shown that a reduced model derived via Galerkin projection onto the span of 30 POD basis vectors is sufficient to reproduce temperature dynamics to within one-third of a degree. In [5], an ensemble is generated by simulations using the nominal optimal control perturbed by a uniform 5%-10%. After truncation of all but the most energetic 40 POD modes, further reductions to a 15 state system were obtained by various procedures, including the "slaving" of modes, i.e., setting the time derivatives of a pre-determined number of "slave" modes to zero, resulting in a "steady manifold" and a set of differential-algebraic equations.

Similar studies of model reduction for heat transfer in a RTP chamber with 3 lamp banks are presented in [1, 157]. In these studies, an ensemble is generated via three simulations: one with each of the lamp banks set to 100% power and the other two turned off. A reduced model with, respectively, 4 and 5 states, is derived from the original models of order 100 and 76, respectively, using the POD and an orthogonal collocation discretization scheme. The reduced model is then used to compute a control for tracking a desired temperature trajectory at several points on the wafer surface.

Another RTP heat transfer application, for which the order reduction is of considerably greater magnitude, is presented in [13]. A finite-element discretiza-

tion of a generic RTP chamber results in a dynamic model with 5060 unknowns. There are 3 operating points of interest, each corresponding to a particular uniform steady-state temperature across the wafer surface. Consequently, the authors derive three reduced models of order 10, each corresponding to one of the operating points, and each capable of reproducing temperature trajectories in a vicinity of the operating point to within 1% error. The overall strategy then involves switching among the three reduced models according to a pre-determined set of rules. Various computational issues regarding the numerical integration of a switched set of state equations are addressed.

Remark 3.2.40 *The ad-hoc nature of the POD applied to model reduction for state-space control systems is made clear in the examples presented above. The discretion of the modeler in choosing the ensemble and designing the overall procedure is evident in all cases. Moreover, the state-to-output relationship is consistently ignored in determining the reduced model.* □

3.3 Balanced Truncation for Linear Systems

One technique that is used often in dealing with nonlinear state-space models is linearization, in which the model is approximated, locally about an equilibrium, by a linear system derived from a Taylor series expansion of the system map. We then work with the resulting linear model for purposes of, e.g., control. In addition, there are various system identification techniques that directly yield a linear model. One method for reducing the order of a linear system is balanced truncation. It is well established and widely applied, mainly due to its simplicity, computability, and good performance.

The balanced realization is one of the infinitely many different state-space realizations for a given LTI system. Its properties have made it very useful in control engineering and signal processing. Mullis and Roberts [111] first introduced the balanced realization in 1976 to study roundoff noise in digital filters. In 1981, Moore [109] proposed the balanced truncation method for reducing the order of a stable finite-dimensional LTI system. The terminology "balanced" reflects the characterization of the realization as one that is equally controllable and observable, a notion that is made precise in Section 3.3.1. When a system is in balanced form, the importance of an individual state component to the input-to-output behavior of the system is indicated by the relative magnitude of its corresponding Hankel singular value. This provides for a meaningful ranking of state components and a guide for truncating those with relatively small contribution.

The method of balanced truncation, or balancing, has been extended in several directions, including to infinite-dimensional linear systems [34, 55], unstable linear systems [104, 122], closed-loop linear systems incorporating various types of controller structures (e.g., LQG [71], H_∞ [113]), and conservative mechanical systems [158]. To accommodate many practical applications, the method has been modified to incorporate frequency weighting [4, 44]. Glover [54] showed that the balanced truncation method is not optimal with respect to the Hankel norm and introduced a closely related method that achieves optimality. We will not work with these extensions or modifications in this thesis. However, generalizations of the method to the nonlinear setting are of paramount importance here and are introduced in Section 3.3.5.

The material contained in this section is well known. Our exposition is mainly based on the presentations in [54, 109, 141, 170], and some results drawn from [127].

We refer to the literature for the proofs.

3.3.1 Derivation

Consider the LTI system

$$\dot{x} = A\,x + B\,u \qquad y = C\,x \tag{3.76}$$

where A is $n \times n$, B is $n \times m$, and C is $p \times n$. The transfer function matrix for (3.76) is given by

$$G(s) = C\,(s\,\mathbb{1} - A)^{-1}\,B \tag{3.77}$$

We say that (A, B, C) is a *realization* for $G(s)$. We say that A is stable if spec $(A) \subset \mathbb{C}^-$.

Definition 3.3.1 (Controllability and Observability Gramians) *Consider the realization (A, B, C) and let A be stable. The $n \times n$ symmetric non-negative definite matrices*

$$W_c = \int_0^\infty \exp(A\,t)\; B\,B^{\mathsf{T}} \exp\left(A^{\mathsf{T}}\,t\right) dt \tag{3.78}$$

$$W_o = \int_0^\infty \exp\left(A^{\mathsf{T}}\,t\right)\; C^{\mathsf{T}}\,C \exp(A\,t)\,dt \tag{3.79}$$

exist and are called, respectively, the controllability Gramian *and the* observability Gramian. □

Interpretations of the Gramians are important to understanding their use in model reduction. Consider the following interpretations from the energy point of view. The minimum control energy required to reach state x_0 from 0 in infinite time is given by $x_0\,{W_c}^{-1}\,x_0$. Hence, ${W_c}^{-1}$ can be used as an indicator of the amount of input energy needed to reach a given state. Similarly, the output energy generated

by releasing the system from initial state x_0 with the input turned off is given by $x_0\, W_o\, x_0$. Hence, W_o can be used as an indicator of the effect that a given initial state has on the output.

Another important interpretation is what Moore referred to as a signal injection view of Kalman's minimal realization theory. The controllability Gramian appears in the PCA (see Section 2.3) of the matrix-valued signal $X(t) = \exp(A\,t)\; B$. This signal corresponds to a collection of state responses to unit impulses injected at the input terminals. Similarly, the observability Gramian appears in the PCA of the matrix-valued signal $Y^{\mathsf{T}}(t) = \exp\left(A^{\mathsf{T}}\,t\right)\; C^{\mathsf{T}}$. The signal $Y(t)$ corresponds to a collection of output responses to unit impulses injected as disturbances at the output terminals of the input filter. In his derivation, Moore used the PCA of these signals to characterize the controllable subspace and the orthogonal complement of the unobservable subspace, and to find a coordinate system in which those subspaces are spanned by the same PCA component vectors. This point of view illustrates the connections between the POD, PCA, and balanced realizations for linear systems.

Example 3.3.2 *The controllability or observability Gramian alone cannot give an accurate indication of the dominance of the state components pertaining to the input-to-output behavior. Consider the following example from [170]. For the transfer matrix*

$$G(s) = \frac{3\,s + 18}{s^2 + 3\,s + 18} \tag{3.80}$$

we have the family of realizations

$$A = \begin{bmatrix} -1 & -4/\alpha \\ 4\,\alpha & -2 \end{bmatrix} \qquad B = \begin{bmatrix} 1 \\ 2\,\alpha \end{bmatrix} \qquad C = \begin{bmatrix} -1 & 2/\alpha \end{bmatrix} \tag{3.81}$$

with Gramians

$$W_c = \begin{bmatrix} 1/2 & 0 \\ 0 & \alpha^2 \end{bmatrix} \qquad W_o = \begin{bmatrix} 1/2 & 0 \\ 0 & 1/\alpha^2 \end{bmatrix} \tag{3.82}$$

We see that the degree to which the second state component is controllable or observable can be made arbitrarily high or low, but not independently. □

The following properties of the Gramians are important for our purposes.

Theorem 3.3.3 (Lyapunov Equations) *Consider the realization* (A, B, C) *and let* A *be stable. The controllability Gramian* W_c *and the observability Gramian* W_o *are the unique solutions of the matrix Lyapunov equations, respectively,*

$$A\,W_c + W_c\,A^{\mathsf{T}} + B\,B^{\mathsf{T}} = 0 \tag{3.83}$$

$$A^{\mathsf{T}}\,W_o + W_o\,A + C^{\mathsf{T}}\,C = 0 \tag{3.84}$$

□

Theorem 3.3.4 *Consider the realization* (A, B, C) *and let* A *be stable. The controllability Gramian* W_c *is positive definite if and only if* (A, B) *is controllable. Likewise, the observability Gramian* W_o *is positive definite if and only if* (C, A) *is observable.* □

A meaningful ranking of states is provided by the Hankel singular values.

Definition 3.3.5 (Hankel Singular Values) *Consider the realization* (A, B, C) *with* A *stable, controllability Gramian* W_c*, and observability Gramian* W_o*. The* Hankel singular values *of the system are defined as the positive square roots of the eigenvalues of the product* $W_c\,W_o$*, i.e.,*

$$\sigma_i = (\lambda_i(W_c\,W_o))^{1/2} \qquad i \in \underline{n} \tag{3.85}$$

where by convention they are ordered such that $\sigma_i \geq \sigma_{i+1}$. □

Remark 3.3.6 *Under similarity transformation $x = S\,\widehat{x}$, the Gramians transform to, respectively,*

$$\widehat{W}_c = S^{-1}\,W_c\left(S^{\mathsf{T}}\right)^{-1} \qquad \widehat{W}_o = S^{\mathsf{T}}\,W_o\,S \tag{3.86}$$

and the product $W_c\,W_o$ transforms to $S^{-1}\,W_c\,W_o\,S$. Thus, the Hankel singular values are invariant under similarity transformation. □

Remark 3.3.7 *If the realization (A, B, C) is minimal then the Hankel singular values are strictly positive. We will work only with minimal realizations.* □

Fact 3.3.8 *Let $G(s)$ be the transfer function matrix for the minimal realization (A, B, C) with A stable. The largest Hankel singular value of the system is equal to the Hankel norm of the system, i.e.,*

$$\|\,G\,\|_H^2 = \sigma_1^2 \tag{3.87}$$

The other Hankel singular values may be characterized inductively in a similar way. □

Example 3.3.9 *The Hankel singular values in Example 3.3.2 are* 1 *and* 0.5. □

We now define the balanced realization in terms of the Gramians and note its relationship with the Hankel singular values.

Definition 3.3.10 (Balanced Realization) *A minimal realization (A, B, C) with A stable, controllability Gramian W_c, and observability Gramian W_o is said to be* balanced *if*

$$W_c = \Sigma = W_o \tag{3.88}$$

where $\Sigma = \mathrm{diag}\,(\sigma_1, \ldots, \sigma_n) > 0$. □

Remark 3.3.11 *Such a realization is also referred to as* internally balanced *[109] and* diagonal balanced *[66].* □

Remark 3.3.12 *The diagonal entries in* Σ *correspond to the Hankel singular values of the system.* □

Remark 3.3.13 *The origin of the terminology "balanced" is apparent, i.e., the system is equally controllable and observable, as indicated by the equality of the Gramians.* □

Theorem 3.3.14 (Existence and Uniqueness of a Balanced Realization) *Let the realization* (A, B, C) *be minimal and let* A *be stable. Then there exists a coordinate transformation* $x = S_{\text{bal}}\,\widehat{x}$ *such that*

$$\widehat{W}_c \stackrel{\triangle}{=} {S_{\text{bal}}}^{-1}\, W_c \Big({S_{\text{bal}}}^{\mathsf{T}}\Big)^{-1} = \Sigma = {S_{\text{bal}}}^{\mathsf{T}}\, W_o\, S_{\text{bal}} \stackrel{\triangle}{=} \widehat{W}_o \tag{3.89}$$

where $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_n) > 0$. *It is unique up to an arbitrary orthogonal transformation* T *such that* $T\,\Sigma = \Sigma\,T$. □

Remark 3.3.15 *We refer to* S_{bal} *as the* balancing coordinate transformation. □

There are two other special forms that are related to the balanced realization and that will be important for computing balanced realizations in the nonlinear setting.

Definition 3.3.16 (Input-Normal, Output-Normal) *A minimal realization* (A, B, C) *with* A *stable, controllability Gramian* W_c*, and observability Gramian* W_o *is said to be* input-normal *if*

$$W_c = \mathbb{I} \qquad W_o = \Sigma^2 \tag{3.90}$$

where Σ *is the diagonal matrix of Hankel singular values. Furthermore, it is said to be* output-normal *if*

$$W_c = \Sigma^2 \qquad W_o = \mathbb{I} \tag{3.91}$$

□

Remark 3.3.17 *The input-normal and output-normal realizations are easily obtained from the balanced realization by scalings on the states, respectively,* $x = \Sigma^{1/2}\,\widehat{x}$ *and* $x = \Sigma^{-1/2}\,\widehat{x}$. □

3.3.2 Properties

We now justify the use of the balanced realization as a tool for model reduction. Let A be stable and (A, B, C) be a minimal realization in balanced form so that $W_c = \Sigma = W_o$. Then the magnitude of the Hankel singular value σ_i, relative to the others, is an indication of the degree to which the i-th state component is, simultaneously, controllable and observable, relative to the others. Equivalently, a small σ_i means that it is relatively difficult both to reach and to observe the state $(0, \ldots, 0, x_i, 0, \ldots, 0)$, and visa-versa. Finally, from an energy point of view, σ_i is interpreted as indicating the contribution of the i-th state component to the input-to-output energy gain of the system, as measured by the Hankel norm.

Suppose that $\sigma_k \gg \sigma_{k+1}$ for some k. Then those states corresponding to $\sigma_{k+1}, \ldots, \sigma_n$ are considerably less important than states corresponding to $\sigma_1, \ldots, \sigma_k$. Consequently, truncation of the less important states causes little degradation to the predictive capability of the model, as pertaining to the input-to-output behavior. These ideas are made more precise in the following discussion.

We derive the reduced-order model via partitioning and truncation. Consider

the partition of the balanced Gramian

$$\Sigma = \left[\begin{array}{c|c} \Sigma_1 & 0 \\ \hline 0 & \Sigma_2 \end{array}\right] \tag{3.92}$$

where $\Sigma_1 = \mathrm{diag}\,(\sigma_1, \ldots, \sigma_k)$ and $\Sigma_2 = \mathrm{diag}\,(\sigma_{k+1}, \ldots, \sigma_n)$. In addition, we partition (A, B, C) accordingly as

$$A = \left[\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array}\right] \qquad B = \left[\begin{array}{c} B_1 \\ \hline B_2 \end{array}\right] \qquad C = \left[\begin{array}{c|c} C_1 & C_2 \end{array}\right] \tag{3.93}$$

We refer to the subsystem (A_{11}, B_1, C_1) as the *truncated system*. It can be used as a k-th-order reduced model to approximate the n-th order full model (A, B, C).

The following results say that the truncated system is balanced and stable.

Theorem 3.3.18 (Pernebo and Silverman [127]) *Let (A, B, C) be balanced with Gramian Σ and partitions (3.92) and (3.93). Then both subsystems (A_{11}, B_1, C_1) and (A_{22}, B_2, C_2) are balanced and their controllability and observability Gramians are equal to, respectively, Σ_1 and Σ_2.* □

Theorem 3.3.19 (Pernebo and Silverman [127]) *Let (A, B, C) be balanced with Gramian Σ and partitions (3.92) and (3.93). If $\sigma_k > \sigma_{k+1}$ (i.e., $\sigma_k \neq \sigma_{k+1}$) then both subsystems (A_{11}, B_1, C_1) and (A_{22}, B_2, C_2) are asymptotically stable.* □

Remark 3.3.20 *Truncation amounts to setting*

$$x_{k+1} = \cdots = x_n = 0 \tag{3.94}$$

There is another method [46] for generating a reduced model in which we set

$$\dot{x}_{k+1} = \cdots = \dot{x}_n = 0 \tag{3.95}$$

Using the latter method, the full-order and reduced-order models have matching DC gains (steady-state response). The former method typically produces a better approximation to the full-order model over a range of frequencies, but the DC gains are not guaranteed to match. Using either method, the reduced model is stable and balanced. □

The performance of balanced truncation as a model reduction method is characterized by the following error bound.

Theorem 3.3.21 (Glover [54]) *Let (A, B, C) be balanced with Gramian Σ, partitions (3.92) and (3.93), and transfer function matrix $G(s)$. Let the truncated system (A_{11}, B_1, C_1) have transfer function matrix $G_1(s)$. Then*

$$\| G - G_1 \|_H \leq \| G - G_1 \|_\infty \leq 2 \sum_{i=k+1}^{n} \sigma_i \tag{3.96}$$

□

Remark 3.3.22 *Thus, if $\sigma_{k+1}, \ldots, \sigma_n$ are small then the error is small and the truncated system is a good approximation in terms of the Hankel norm, i.e., the error bound can be used as a measure of model fidelity corresponding to (1.1).* □

Remark 3.3.23 *The upper bound for the error (3.96) is not optimal, but is close to optimal. See Glover [54] for a characterization of all optimal Hankel norm approximations to (A, B, C) and associated error bounds.* □

3.3.3 Computation

Given a minimal realization (A, B, C) with A stable, a balanced realization can be obtained efficiently through the standard algorithms of Laub [90], Moore [109], and the more elegant algorithm of Laub, et.al. [91]. The latter algorithm is currently,

and has been since 1994, the standard balancing algorithm used in MATLAB [102]. These algorithms use standard, efficient, and stable matrix algebra and decomposition routines such as for SVD and Cholesky decomposition. There is little or no trouble in computing balanced realizations for LTI systems in most situations.

Numerical difficulties arise in using these algorithms when the condition number of the product matrix $W_c W_o$ is high. This corresponds to the situation where some states are nearly uncontrollable or nearly unobservable, i.e., the realization is nearly non-minimal. Safonov and Chiang [136] introduced a method to deal with this situation, which they called a "Schur method" in reference to the fact that it is computed using the Schur decomposition of the product matrix $W_c W_o$.

The idea is to avoid balancing altogether by computing orthonormal bases, via the ordered Schur form, for the left and right eigenspaces associated with the "big" eigenvalues of $W_c W_o$. The reduced k-th order model is not necessarily balanced, but has exactly the same transfer function as would any k-th order balanced realization. Thus, the reduced unbalanced model enjoys the same error bound as a reduced balanced model. The algorithm is stable and effective without regard to nearness to unobservability or uncontrollability.

Finally, we note that Helmke and Moore [66] have presented gradient flows on the class of positive definite matrices that converge to the unique symmetric positive definite balancing transformation matrix, as well as to all balancing transformation matrices for a given realization (A, B, C). The gradient flows converge exponentially fast to the balanced realization. The gradient flow method has not yet become prominent for applications due to issues of practical implementation.

The details of the various algorithms discussed here are presented in Appendix E.

3.3.4 Applications

Balanced truncation has become prominent as a tool for model reduction during the 1980s and 1990s mainly due to its simplicity, computability, and good performance. It has been applied in a wide variety of areas, including the several that we briefly discuss here. We note that in these and most practical situations, the balancing procedure is modified somewhat to suit the intended application. However, the role which ad-hoc techniques play in balanced model reduction is small compared with that for POD methods.

One drawback of balancing is that the state variables in the balanced realization lose their physical meaning. Blelloch, Mingori, and Wei [17] addressed this problem in their application of balanced truncation to linear models for lightly damped mechanical systems with gyroscopic and small circulatory forces such as large flexible space structures. Specifically, they used the result that the modal representation for certain systems becomes asymptotically balanced as the damping approaches zero. Thus, a lightly damped structure in modal form, which retains the physical meaning of the state variables, is approximately balanced. After deriving an approximate balanced realization for the model of a dual-spin spacecraft, they take advantage of its properties to reduce the model order.

Friswell, Penny, and Garvey [50] conducted a comparative study of reduction methods for models of high degree-of-freedom mechanical structures with local nonlinearities. The nonlinear term (nonlinear forces) was ignored in computing the linear coordinate transformations, which were then applied to the full nonlinear model. The authors applied the reduction methods to models of a cantilever beam system and a pinned beam system. For the simulations they conducted, reduced-order models derived via balanced truncation predicted the response of the full-

order nonlinear model more accurately than the other methods, including modal coordinates.

Ramirez and Maciejowski [130] directly formulated a balanced model for a stirred tank chemical reactor from pulse response data. They used the canonical form of Ober [122, 123] and a related system identification algorithm to produce a high-order balanced realization. Truncation yielded a 3-state linear model. The input-to-output response of the low-order balanced model was compared with that of a 3-state nonlinear physical-chemical model and its linearization. For the simulations they conducted, the balanced model captured the nonlinear dynamics of the system more accurately than did the linearization. It was then used in the design of an LQG optimal control law and Kalman filter for regulation in the presence of large disturbances.

Finally, we note that balanced truncation has been studied [116, 120], and used in at least one commercial software package [129], for order reduction of nonlinear models for heat transfer in RTP chambers. In these applications, linearized versions of the models are used to compute the balancing coordinate transformation, which is then applied to the original nonlinear model. Truncation yields low-order models for use in simulation, control, and optimization.

3.3.5 Nonlinear Generalizations

There have been two recent independent attempts (of which we are aware) to generalize the method of balanced truncation to the nonlinear setting. In 1993, Scherpen [140, 141] introduced a general theory and procedure of balancing for a class of stable, affine, smooth, finite-dimensional nonlinear systems. The approach is inherently nonlinear; it produces a nonlinear balancing coordinate transforma-

tion that is local to a neighborhood of the origin. In 1999, Lall, Marsden, and Glavaski [137] introduced a method that is inherently linear; it produces a linear change of coordinates by constructing and decomposing matrices that serve as generalized Gramians for the nonlinear system. We introduce and remark on these methods below to provide background for Chapter 4.

Scherpen Nonlinear Balancing

The Scherpen methodology departs from the signal injection viewpoint of Moore while remaining consistent with its results in the LTI case. The main objects of importance are the controllability and observability energy functions. These functions serve the role that the controllability and observability Gramian matrices do in linear balancing, i.e., they provide a well-defined measure of the degree to which a system is, respectively, controllable or observable. However, unlike the Gramians, they are not easily computable.

In the LTI case, the energy functions specialize to quadratic forms involving the Gramian matrices. Thus, it is natural that, in the nonlinear setting, the first step in the balancing procedure is to determine a nonlinear change of coordinates in a neighborhood of the origin under which the controllability function is locally quadratic. This is accomplished by application of the Morse-Palais lemma, which gives a quadratic canonical form for functions in the neighborhood of a nondegenerate critical point. Again, there exist no practical methods for computing the desired change of coordinates.

Further nonlinear coordinate transformations take the system to special forms that are analogous to input-normal, output-normal, and balanced. When the system is in balanced form, state components can be ranked and deleted according

to their respective contributions to the input-to-output energy of the system, as indicated by the respective relative magnitudes of the singular value functions, which are generalizations of the Hankel singular values.

Remark 3.3.24 *The method is consistent with the LTI balancing procedure in the following sense. Suppose the nonlinear system is realized with* (f, g, h) *which has linearization* (A, B, C) *about* 0*. Let* Ψ *be the nonlinear balancing transformation for* (f, g, h) *about* 0 *and* $\left(\widehat{f}, \widehat{g}, \widehat{h}\right)$ *be the balanced realization. Let* S *and* $\left(\widehat{A}, \widehat{B}, \widehat{C}\right)$ *be the linearizations, respectively, of* Ψ *and* $\left(\widehat{f}, \widehat{g}, \widehat{h}\right)$ *about* 0*. Let* T *be the balancing transformation matrix for* (A, B, C)*. Then* $S = T$ *and* $\left(\widehat{A}, \widehat{B}, \widehat{C}\right) = (T^{-1} A T, T^{-1} B, C T)$*.* □

Remark 3.3.25 *In contrast to the linear case, the nonlinear balancing procedure is not immediately amenable to computational implementation. For example, the controllability energy function corresponds to the value function for a nonlinear optimal control problem. Also, the Morse-Palais lemma guarantees the existence of a transformation to the desired canonical form, but provides no constructive procedure for obtaining it. Thus, tools have not yet appeared for computing balanced realizations for nonlinear systems, and the Scherpen procedure has not yet been applied as a tool for model reduction.* □

Scherpen Pseudo-Balancing

We note that, in 1994, Scherpen [141] also introduced a method for balancing in the special case of a nonlinear Hamiltonian system. A special technique is necessary because such a system is conservative and therefore not asymptotically stable. The method is a nonlinear generalization of the pseudo-balancing approach of van der Schaft and Oeloff [158]. The idea is to derive an associated "gradient system"

of dimension n, from the original Hamiltonian system of dimension $2n$, that is asymptotically stable and thus can be balanced.

Remark 3.3.26 *To date, the only "balanced" realizations for nonlinear systems that have appeared in the literature have been derived via pseudo-balancing [141, 159], i.e., are pseudo-balanced Hamiltonian systems.* □

LMG Nonlinear "Balancing"

The method of Lall, Marsden, and Glavaski (LMG) adopts the signal injection viewpoint of Moore and extends it by expanding the class of allowable impulsive test signals to include rigid rotations and uniform scalings of the original impulsive vector signals. In a procedure analogous to that of Moore, application of PCA to the resulting collection of system response signals produces matrices, called the empirical controllability Gramian and the empirical observability Gramian, that serve the role that their respective counterparts do in the LTI case. Thus, again like that of Moore, the approach is strongly related to the POD.

Remark 3.3.27 *As with the Scherpen method, the LMG method is consistent with the LTI balancing procedure. It is shown that the use of the expanded input signal space has no effect when applied in the LTI case, i.e., the LMG method specializes to the usual balanced truncation for LTI systems.* □

Remark 3.3.28 *The computational framework of the LMG method is the same as that of Moore, i.e., it involves matrix algebra and decompositions. Therefore, it retains the computational ease and efficiency of LTI balancing.* □

Remark 3.3.29 *The nomenclature "empirical Gramian" used to distinguish the new objects from the familiar Gramians is misleading, since the empirical Gramians*

are no more nor less "empirical" than the familiar Gramians. This is made clear by Moore's signal injection viewpoint. The new objects are better described as extended Gramians, to indicate that they are constructed by taking account of the broader class of impulsive test signals. Moreover, the properties of these Gramians pertaining to controllability and observability in the nonlinear setting are unclear and not discussed. □

The authors claim that the method provides a balanced truncation via ranking of subspaces, but the authors never actually define what they mean by a "balanced nonlinear system." The proposed change of coordinates balances the empirical Gramians, i.e., makes them equal to the same positive-definite diagonal matrix. While the consequences of such a change of coordinates are known in the LTI case, the implications on the nonlinear system realization are not known and never discussed. Moreover, no quantification of the importance of a particular subspace is offered. Thus, it is unclear as to how exactly the LMG method is an extension of balancing to nonlinear systems.

The LMG method does provide an organized framework for injecting test signals with additional degrees of freedom beyond that which Moore described, while remaining compatible with the linear theory, and reinforcing the importance and special properties of impulsive signals as inputs. In particular, the rigid rotations and scalings of the impulsive inputs are chosen such that they excite the nonlinear system in some appropriate manner. The authors suggest that experience and experimental data may be useful in choosing parameters. Thus, the method suffers from some of the shortcomings associated with the POD, albeit with the choice of input signals parameterized in an organized fashion.

Remark 3.3.30 *Generalizations of the Gramian matrices to the nonlinear setting that are compatible with the Scherpen theory for nonlinear balancing have recently appeared [59]. These generalizations have the advantage that their properties pertaining to controllability and observability are well-defined.* □

3.4 Component Truncation

Recall the general methodology for state-space model reduction outlined in Section 1.1.1 and illustrated by Figure 1.1. Once a coordinate transformation has been selected, it is applied to the model equations to yield the transformed system. Then, state components are ranked, and some are deleted. We will refer to this process as component truncation. We have already presented component truncation procedures for the special cases of

- a linear orthogonal coordinate transformation Φ for a nonlinear model (POD - via projection Φ_k); and
- a linear coordinate transformation S_{bal} for a LTI model (balanced truncation - via partition and truncation).

Here we present the general procedure and make some remarks about practical computation.

Let the full-order state-space model be given by the autonomous state equation and output equation, respectively,

$$\dot{x} = f(x,u) \tag{3.97}$$

$$y = h(x) \tag{3.98}$$

where $x \in \mathbb{R}^n$ represents local coordinates for the full-order state. Suppose that we apply the diffeomorphic change of coordinates

$$z \mapsto x = S(z) \tag{3.99}$$

under which the system map and output map transform to (see Section 2.1), respectively,

$$\widehat{f}(z,u) = [DS(z)]^{-1} f\left(S(z),u\right) \tag{3.100}$$

$$\widehat{h}(z) = h\left(S(z)\right) \tag{3.101}$$

We reduce the system order, or truncate state components, by setting

$$z_{k+1} = \cdots = z_n = 0 \tag{3.102}$$

so that $\widehat{f}_{k+1}, \ldots, \widehat{f}_n$ are no longer relevant to the model. We define the truncated state

$$z^1 \triangleq [z_1, \ldots, z_k]^{\mathsf{T}} \tag{3.103}$$

and the truncated system map and output map, respectively,

$$\widehat{f}^1\left(z^1,u\right) \triangleq \begin{bmatrix} \widehat{f}_1(z_1,\ldots,z_k,0,\ldots,0,u) \\ \vdots \\ \widehat{f}_k(z_1,\ldots,z_k,0,\ldots,0,u) \end{bmatrix} \qquad \widehat{h}^1\left(z^1\right) \triangleq \begin{bmatrix} \widehat{h}_1(z_1,\ldots,z_k,0,\ldots,0) \\ \vdots \\ \widehat{h}_p(z_1,\ldots,z_k,0,\ldots,0) \end{bmatrix} \tag{3.104}$$

The reduced-order model and approximate state reconstruction are given by

$$\dot{z}^1 = \widehat{f}^1\left(z^1,u\right) \tag{3.105}$$

$$\widehat{y} = \widehat{h}^1\left(z^1\right) \tag{3.106}$$

and

$$\widehat{x} = S(z_1,\ldots,z_k,0,\ldots,0) \tag{3.107}$$

The reduction in dimensionality of the state-space manifold from n to k is not necessarily reflected in practical computations. Suppose that we are given (f, g, h) and compute a suitable local transformation S. Consider a numerical integration of the reduced state equation (3.105). This computation requires, for each time step, evaluation of the k real-valued functions in the reduced system map $\widehat{f}^1 = \left(\widehat{f}_1, \ldots, \widehat{f}_k\right)$. However, practically speaking, these function evaluations must be performed using the known original f and known diffeomorphism S via (3.100). Thus, it is actually necessary to evaluate the n real-valued functions in the full state map $f = (f_1, \ldots, f_n)$, partially defeating the purpose of the reduction. This dilemma has been alluded to in the applied POD literature [5, 6] without further explanation.

Remark 3.4.1 *It appears desirable to have a computational procedure in which the transformed system map $\widehat{f}$ can be evaluated without needing to evaluate the full system map f.* □

Remark 3.4.2 *This issue does not enter into the LTI case. The linear structure allows for the elimination of a subsystem that is completely irrelevant to computations for the reduced model.* □

3.5 Remarks

We have studied two prominent methodologies for model reduction of linear and nonlinear systems. The POD approach is applicable to linear and nonlinear models, and to models of finite and infinite dimension. It produces a set of basis vectors for a linear orthogonal change of coordinates. The basis vectors are called empirical eigenvectors, and in certain contexts they correspond to physically manifested

spatial structures in a spatially distributed flow. The procedure can be characterized as statistical and empirical in nature, in that the basis is derived via spectral decomposition of the covariance associated with an ensemble of empirically generated signals. Its effectiveness relies on the ability of the ensemble to capture the essential system behavior. The POD coordinate transformation is not derived directly from the model or its intrinsic properties. Furthermore, it completely ignores the state-to-output relationship.

The empirical eigenvectors are computed easily using SVD. The corresponding eigenvalues provide a meaningful ranking of state components in terms of signal energy as captured by the ensemble. However, there exist no explicit error bounds in terms of the eigenvalues or otherwise. Coordinate transformation and component truncation occur simultaneously via linear projection onto the low-dimensional subspace. Versions of the method have been applied within overall ad-hoc procedures for particular situations.

The balancing approach produces a linear change of coordinates for an LTI system. The balanced realization is derived directly from the model parameters (A, B, C) through decompositions of the Gramians W_c and W_o. The procedure can be characterized as control-theoretic in nature, in that it derives from controllability and observability properties of the system, although the signal injection view of Moore reveals connections to the POD.

The balancing transformation is easily computable using efficient matrix algebra algorithms. The Hankel singular values provide a well-defined and meaningful ranking of state components in terms of contribution to the input-to-output behavior. Furthermore, the norm of the error between the full and reduced models is explicitly computed in terms of the discarded Hankel singular values. Component

truncation is performed via partition and subsystem elimination. The method has been applied to linear models for various physical systems.

The theory and procedure of Scherpen generalizes the established linear method to the nonlinear setting. The balancing transformation is nonlinear and local to a neighborhood of the origin. It retains some of the appealing features of the linear method, e.g., the balancing transformation is derived directly from the model parameters f, g, and h, and emphasizes state components that are both strongly controllable and strongly observable, so that state components which are least likely to influence the measurements are truncated. However, the procedure is not easily computable and its performance has not yet been observed in applications. We address these issues in Chapter 4.

Chapter 4

Computing Balanced Realizations for Nonlinear Systems

4.1 Introduction

This chapter addresses the problem of computability pertaining to the Scherpen theory and procedure for balancing of nonlinear systems. We offer methods and algorithms toward computing balanced realizations for stable affine nonlinear control systems, i.e., state-space models of the form

$$\dot{x}(t) = f(x(t)) + \sum_{i=1}^{m} g_i(x(t))\, u_i(t) \tag{4.1}$$

$$y(t) = h(x(t)) \tag{4.2}$$

where $u = (u_1, \ldots, u_m) \in U \subset \mathbb{R}^m$, $y = (y_1, \ldots, y_p) \in \mathbb{R}^p$, and $x = (x_1, \ldots, x_n)$ are local coordinates for a smooth state-space manifold M. The maps $f, g_1, \ldots, g_m$ are smooth and we assume that $f(0) = 0$ and $h(0) = 0$.

We say that f is stable (asymptotically stable) if 0 is a stable (asymptotically stable) equilibrium for $\dot{x} = f(x)$, and normally assume asymptotic stability of f.

We refer to the triple (f, g, h) as a realization of the nonlinear system.

In Sections 4.2 and 4.3 we consider the problem of computing the controllability energy function without solving the family of optimal control problems implied in its definition. Stochastically excited systems (see Section 2.6) play a major role in our methodology. We present a stochastic method for computing an estimate of the controllability function, and show that in certain situations the method provides an exact solution. The procedure is tested on applications via Monte-Carlo experiments.

The crucial step in the balancing procedure is finding a local coordinate transformation under which the controllability function is quadratic in a neighborhood of 0. The Morse-Palais lemma ensures the existence of a quadratic canonical form for a function with a non-degenerate critical point at 0, such as the controllability and observability energy functions. However, it provides no constructive procedure for obtaining the desired local change of coordinates. In Section 4.4 we develop an algorithm for computing the desired nonlinear transformation.

In Section 4.5 we present the overall procedure for computing the balancing transformation and algorithms for performing the required computations. We apply the methods developed in this chapter to two example problems. The results are presented in Section 4.6. We compute a balanced realization for a forced damped pendulum system, and make progress toward balancing a forced damped double pendulum system. Additional remarks and a summary are presented in Section 4.7.

4.2 Energy Functions

As stated earlier, a balanced realization is one that is equally controllable and observable. In order to make such a statement meaningful, we must provide a measure of the degree to which a system realization, and its state components, are, respectively, controllable and observable. These properties can be quantified in a precise way via, respectively, the controllability and observability energy functions of the system.

Definition 4.2.1 (Controllability Function) *The controllability function, $L_c : \mathbb{R}^n \to \mathbb{R}$, for system (4.1)-(4.2) is defined by*

$$L_c(x_0) = \min_{\substack{u \in \mathcal{L}_2(-\infty, 0) \\ x(-\infty) = 0\,,\; x(0) = x_0}} \frac{1}{2} \int_{-\infty}^{0} \| u(t) \|^2 \, dt \tag{4.3}$$

□

Definition 4.2.2 (Observability Function) *The observability function, $L_o : \mathbb{R}^n \to \mathbb{R}$, for system (4.1)-(4.2) is defined by*

$$L_o(x_0) = \frac{1}{2} \int_{0}^{\infty} \| y(t) \|^2 \, dt \qquad x(0) = x_0 \qquad u(t) \equiv 0,\, t \geq 0 \tag{4.4}$$

□

Remark 4.2.3 *The value of L_c at state x_0 is the minimum amount of control energy required to reach the state x_0 from 0. The value of L_o at state x_0 is the amount of output energy generated by the system's natural response to initial state x_0.* □

Remark 4.2.4 *There are other definitions of L_o for which the input u plays a direct role. These can be considered closed loop generalizations of (4.4), which*

then corresponds to an open loop observability function. For details and results see [58, 142]. In this thesis we use only the L_o given by (4.4). □

Fact 4.2.5 *In the case of an LTI system (3.76) the controllability and observability functions, respectively, specialize to the quadratic functions*

$$L_c(x_0) = \frac{1}{2}\, x_0{}^{\mathsf{T}}\, W_c{}^{-1}\, x_0 \tag{4.5}$$

$$L_o(x_0) = \frac{1}{2}\, x_0{}^{\mathsf{T}}\, W_o\, x_0 \tag{4.6}$$

where the symmetric positive-definite matrices W_c and W_o are, respectively, the familiar controllability and observability Gramian matrices, given, respectively, by (3.78) and (3.79) (see [140, 141]). □

4.2.1 Properties

There are several properties of the controllability and observability energy functions that we will use in our computational effort. It is clear that both of these functions are non-negative (vanish only at 0). However, they are not necessarily finite everywhere in a neighborhood of the origin, nor is the minimum at 0 necessarily non-degenerate, i.e., isolated. We need both L_c and L_o to be finite (i.e., to exist) and non-degenerate in a neighborhood of 0 in order to perform the balancing computations. We now discuss conditions under which the energy functions enjoy those properties.

Theorem 4.2.6 (Scherpen and Gray [142]) *Suppose that f is asymptotically stable on a neighborhood W of 0. Then $L_c(x)$ is smooth, finite, and satisfies $L_c(x) > 0$ for $x \in W$, $x \neq 0$ if and only if the system (4.1)-(4.2) is asymptotically reachable from 0 on W.* □

Remark 4.2.7 *It makes intuitive sense that a reachability property determines whether L_c is finite, since if we have a state x_0 that is not reachable, the minimum in (4.3) will fail to exist at x_0. In that case, by convention, we take $L_c(x_0) = \infty$.* □

Remark 4.2.8 *Non-degeneracy of L_c is guaranteed by asymptotic stability of f. Intuitively, some positive control energy must be required to steer away from the asymptotically stable origin to states in some open neighborhood.* □

Interestingly, but not surprisingly, we have a dual situation regarding L_o. In a reversal of the situation for L_c, non-degeneracy (rather than finiteness) of L_o is determined by an observability property, and finiteness (rather than non-degeneracy) of L_o is determined by stability properties. We first state the condition for non-degeneracy of L_o.

Theorem 4.2.9 (Scherpen [140]) *Suppose that f is asymptotically stable on a neighborhood W of 0. If the system (4.1)-(4.2) is zero-state observable on W, then $L_o(x) > 0$ for $x \in W$, $x \neq 0$.* □

To ensure that L_o is finite, we use stability conditions on f together with additional conditions on h. It turns out that the typical situation in which the origin is exponentially stable and h is smooth with $h(0) = 0$ is sufficient to guarantee that L_o is finite on a neighborhood W of 0. In fact, we derive somewhat weaker sufficient conditions, presented in the following.

Proposition 4.2.10 *Suppose that f is asymptotically stable with region of attraction W. Let $x_0 \in W$. Let $x(t)$ be the unique solution to the initial value problem*

$$\dot{x} = f(x) \qquad x(0) = x_0 \tag{4.7}$$

Suppose that there exists a time $t_1 \geq 0$ *and positive constants* α *and* β *such that for all* $t \geq t_1$

$$\| x(t) \| \leq \alpha \exp\left(-\beta(t - t_1)\right) \tag{4.8}$$

Suppose that h *is locally Lipschitz on* W *with* $h(0) = 0$. *Define* $y(t) = h\left(x(t)\right)$ *for* $t \geq 0$. *Then* $y \in \mathcal{L}_2[0, \infty)$.

Proof Since h is locally Lipschitz on W, there exists a constant $L > 0$ such that

$$\| h(x) - h(0) \| \leq L \| x - 0 \| \tag{4.9}$$

in some compact neighborhood $V \subset W$ of 0. By $h(0) = 0$ we have $\| h(x) \| \leq L \| x \|$ for all $x \in V$. Let $U = V \cap \mathrm{B}(0, \alpha)$. Let $t_2 = \min\{t \geq 0 : x(t) \in U\}$. Then for $t \geq t_2$

$$\| y(t) \| = \| h(x(t)) \| \leq L\, \alpha \exp\left(-\beta\left(t - t_1\right)\right) \tag{4.10}$$

and

$$\begin{aligned} \int_{t_2}^{\infty} \| y(t) \|^2 \, dt &\leq L^2\, \alpha^2 \exp\left(2\, \beta\, t_1\right) \int_{t_2}^{\infty} \exp\left(-2\, \beta\, t\right) dt \\ &= \frac{L^2\, \alpha^2}{2\, \beta} \exp\left(2\, \beta\left(t_1 - t_2\right)\right) \end{aligned} \tag{4.11}$$

Now, by asymptotic stability, $x(t)$ is finite for all $t \geq 0$ for arbitrary $x_0 \in W$. Furthermore, h is continuous on V (by Lipschitz property and compactness of V). Therefore $y(t) = h\left(x(t)\right) < \infty$ for $0 \leq t \leq t_2$. This implies that for some $\gamma > 0$

$$\int_0^{t_2} \| y(t) \|^2 \, dt = \gamma < \infty \tag{4.12}$$

Thus,

$$\int_0^{\infty} \| y(t) \|^2 \, dt \leq \gamma + \frac{L^2\, \alpha^2}{2\, \beta} \exp\left(2\, \beta\left(t_1 - t_2\right)\right), \tag{4.13}$$

i.e., $y \in \mathcal{L}_2[0, \infty)$. ■

Remark 4.2.11 *The proposition asserts that L_o is finite on W whenever all of the following hold:*

(i) The equilibrium 0 is asymptotically stable with region of attraction containing W.

(ii) There exists a neighborhood of 0, possibly smaller, in which 0 is exponentially stable.

(iii) The function h is at least locally Lipschitz on W.

□

Remark 4.2.12 *Scherpen [141] considers only the more general case where h is smooth and states the more conservative condition that the linearization $A = \left[\frac{\partial f}{\partial x}(0)\right]$ be asymptotically stable, i.e., that 0 be exponentially stable.* □

Remark 4.2.13 *Given a Lipschitz h, mere asymptotic stability of 0 without a neighborhood of exponential stability is insufficient to guarantee a finite L_o. Systems whose linearizations yield eigenvalues on the imaginary axis are candidates for belonging to such a class of systems. For example, consider the system $\dot{x} = -x^3$ with $h(x) = x$ (smooth) and $x(0) = 0$. The output trajectory is $y(t) = \left(1/\sqrt{2}\right) t^{-1/2}$ so that $\| y(t) \|^2 \overset{t\to\infty}{\longrightarrow} 0$ as t^{-1}, but y does not belong to $\mathcal{L}_2[0,\infty)$. If L_o fails to exist at x_0 then by convention we take $L_o(x_0) = \infty$.* □

Remark 4.2.14 *The systems we work with generally have smooth h with $h(0) = 0$. Moreover, many physical systems, including mechanical systems with damping, enjoy exponential stability. Throughout this thesis, we usually can assume that the functions L_c and L_o are finite, smooth, and non-degenerate everywhere in their domain, without further explanation.* □

We now present an important theoretical result that we use later to analyze the results of our computational methods. The functions L_c and L_o each satisfy a familiar nonlinear PDE, respectively, a Hamilton-Jacobi-Bellman (HJB) equation associated with an optimal control problem, and a nonlinear type of Lyapunov equation.

Theorem 4.2.15 (Scherpen [140, 141])

(i) Suppose that 0 *is an asymptotically stable equilibrium of* $-\left(f + g\, g^{\mathsf{T}} \left[\frac{\partial L_c}{\partial x}\right]^{\mathsf{T}}\right)$ *on a neighborhood* V *of* 0*. Then, for all* $x \in V$*,* L_c *is the unique smooth solution of*

$$0 = \frac{\partial L_c}{\partial x}(x)\; f(x) + \frac{1}{2}\frac{\partial L_c}{\partial x}(x)\; g(x)\; g^{\mathsf{T}}(x)\; \left[\frac{\partial L_c}{\partial x}(x)\right]^{\mathsf{T}} \qquad L_c(0) = 0 \quad (4.14)$$

under the assumption that (4.14) has a smooth solution on V*.*

(ii) Suppose that 0 *is an asymptotically stable equilibrium of* $f(x)$ *on a neighborhood* W *of* 0*. Then, for all* $x \in W$*,* $L_o(x)$ *is the unique smooth solution of*

$$0 = \frac{\partial L_o}{\partial x}(x)\; f(x) + \frac{1}{2} h^{\mathsf{T}}(x)\; h(x) \qquad L_o(0) = 0 \qquad (4.15)$$

under the assumption that (4.15) has a smooth solution on W*.*

Proof See Remark 4.2.16 and Appendix F. ■

Remark 4.2.16 *Scherpen [140, 141] proves Theorem 4.2.15 via a completing the square argument and straightforward manipulation of Equations (4.14) and (4.15) and Definitions 4.3 and 4.4. We offer a different proof in Appendix F that appeals to the connections between Equations (4.14) and (4.15) and optimal control theory.* □

Remark 4.2.17 *Smooth solutions of (4.14) and (4.15) exist locally about* 0 *if the matrix* $A = \left[\frac{\partial f}{\partial x}(0)\right]$ *is Hurwitz, so we will assume that the theorem is generally applicable.* □

Remark 4.2.18 *For the case of a linear system, Equations (4.14) and (4.15) specialize to the matrix Lyapunov equations (3.83) and (3.84) (repeated below)*

$$\begin{aligned} A\,W_c + W_c\,A^{\mathsf{T}} + B\,B^{\mathsf{T}} &= 0 \\ A^{\mathsf{T}}\,W_o + W_o\,A + C^{\mathsf{T}}\,C &= 0 \end{aligned}$$

□

4.2.2 Remarks on Computation and Applications

Here we briefly point out some of the difficulties involved with computing the controllability and observability energy functions, and discuss some straightforward but likely impractical approaches.

Suppose that we have suitably discretized the state-space in such a way that there are p evenly spaced grid points along each of the n dimensions. This means that there are p^n total grid points at which the energy functions are to be evaluated. We denote the set of discrete grid points by $\mathcal{X} \subset \mathbb{R}^n$.

A direct computation of L_c from Definition 4.3 requires the numerical solution of p^n optimal control problems for the nonlinear system (4.1), one for each initial state $x_0 \in \mathcal{X}$. In particular, to use a standard solution approach, the minimization problem in (4.3) should be restated so that the optimal control problem corresponds to signals in positive time, i.e., controls in $\mathcal{L}_2(0,\infty)$. We make the changes of variables

$$\tau \stackrel{\triangle}{=} -t \qquad z(\tau) \stackrel{\triangle}{=} x(-\tau) \qquad v(\tau) \stackrel{\triangle}{=} u(-\tau) \tag{4.16}$$

for $t \leq 0$ and $\tau \geq 0$ so that (4.1) and (4.3), respectively, transform to

$$\dot{z}\,(t) = -f\,(z(t)) - g\,(z(t))\; v(t) \tag{4.17}$$

and

$$L_c\,(x_0) = \min_{\substack{v \in \mathcal{L}_2(0,\infty) \\ z(0) = x_0\,,\; z(\infty) = 0}} \frac{1}{2} \int_0^\infty \| v\,(\tau) \|^2 \, d\tau \tag{4.18}$$

Remark 4.2.19 *For the system (4.17),* 0 *is an unstable equilibrium. Therefore, the minimum energy control which takes the state from* $x_0 \neq 0$ *to* 0 *cannot be* $u \equiv 0$*, again demonstrating the non-degeneracy of* L_c*.* □

There exist computational methods and at least one software toolbox (RIOTS [143]) for solving broad classes of optimal control problems such as (4.18). However, regardless of the computational complexity of the solution algorithms, the overall computational complexity is at least $o(p^n)$, the number of optimal control problems we need to solve. Even for very low-order systems, the computational expense is prohibitive.

Remark 4.2.20 *Similarly, the overall computational complexity for a direct computation of* L_o *from Definition 4.4 is at least* $o(p^n)$*. However, for each* $x \in \mathcal{X}$*, rather than a numerical solution of an optimal control problem, we merely need a numerical integration of the system equations. Although still impractical in general, it is feasible for low-order systems.* □

Another computational approach that immediately comes to mind is numerical solution of the nonlinear PDE (4.14) for the value function L_c. Equation (4.14) is of Hamilton-Jacobi type (see [47, 93] for an extensive analysis of this type of

equation, including existence and uniqueness results, and properties of solutions) and of the general form

$$H(x, u, Du) = 0,\ x \in \Omega \qquad u = \phi,\ x \in \partial\Omega \tag{4.19}$$

where the function H is called the Hamiltonian, u is the unknown function, Du denotes the gradient of u, Ω is the domain of definition for u, and ϕ is a prescribed boundary condition. In the case of (4.14) we have $u = L_c$ and $\Omega = \mathbb{R}^n$. Rather than a boundary condition we have the supplemental condition $L_c(0) = 0$.

These types of equations are in general nonlinear first-order problems for which there is no hope to find classical solutions (i.e., a solution of class C^1 at least). Instead, one must deal with suitable generalized solutions (i.e., locally Lipschitz on Ω, continuous on $\bar{\Omega}$, and almost everywhere differentiable). The correct class of generalized solutions was established by Crandall and Lions in [33, 93]. There they introduced the notion of the viscosity solution of nonlinear first-order PDEs which are the generalized solutions of primary interest in many areas of application including this one. Briefly, under certain hypotheses, for $\epsilon > 0$, the solution u^ϵ of

$$H(x, u^\epsilon, Du^\epsilon) - \epsilon \triangle u^\epsilon = 0,\ x \in \Omega \qquad u^\epsilon = \phi,\ x \in \partial\Omega \tag{4.20}$$

approximates the viscosity solution of (4.19) with error estimate

$$|u^\epsilon(x) - u(x)| \leq c\sqrt{\epsilon} \tag{4.21}$$

for some constant c.

Crandall and Lions [32] give finite-difference schemes for approximating these viscosity solutions along with error estimates. Souganidis [153] establishes results concerning the convergence of explicit and implicit finite-difference schemes to viscosity solutions. However, except when dealing with low-dimensional state-

spaces, the finite-difference schemes become impractical as the number of grid points becomes prohibitively large.

Remark 4.2.21 *Similarly, finite-difference schemes for solution of the nonlinear PDE (4.15) become impractical in higher dimensions.* □

4.3 Stochastic Methods for Computation

We seek a method for computing the controllability energy function without solving the family of optimal control problems implied in its definition, or solving the associated HJB equation. In this section we offer an approach, based primarily on the theory of stochastically excited dynamical systems, for computing an estimate of the controllability function. We show that in certain situations the method provides an exact solution.

4.3.1 Stationary Densities and the Controllability Function

In Section 2.6 we introduced the notion of a stochastically excited dynamical system, i.e., a control system for which the m components of the input, u_i, $i \in \underline{m}$, have been replaced by the sample paths of m Gaussian white noises, $\left\{(\zeta_t)_i,\, t \in \mathbb{R}^+\right\}$, $i \in \underline{m}$. The state equation is given by

$$\frac{d}{dt} X_t = f(X_t) + \sum_{i=1}^{m} g_i(X_t)\,(\zeta_t)_i \tag{4.22}$$

Recall that the white noise driven system (4.22) is interpreted correctly via the SDE (given elementwise)

$$(dX_t)_i = \bar{f}_i\,(X_t)\; dt + \sum_{i=1}^{m} g_i\,(X_t)\;(dW_t)_i \qquad i \in \underline{n} \tag{4.23}$$

where

$$\bar{f}_i(X_t) = \left[f_i(X_t) + \frac{1}{2} \sum_{j=1}^{n} \sum_{k=1}^{m} \frac{\partial g_{ik}}{\partial X_j}(X_t)\ g_{jk}(X_t) \right] \qquad i \in \underline{n} \tag{4.24}$$

and $g_{ij} = (g_j)_i$, $i \in \underline{n}$, $j \in \underline{m}$, and where (4.23) is defined in terms of a corresponding stochastic integral.

Recall also that the state X_t is a Markov process with transition probability density $p(x,t;y,s)$. Time evolution of $p(x,t;y,s)$ is governed by the Fokker-Planck equation, given by

$$\begin{aligned} \frac{\partial p}{\partial t}(x,t;y,s) = & \\ - & \sum_{i=1}^{n} \frac{\partial}{\partial x_i} \left(\bar{f}_i(x)\, p(x,t;y,s) \right) \\ + & \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial^2}{\partial x_i\, \partial x_j} \left(b_{ij}(x)\, p(x,t;y,s) \right) \end{aligned} \tag{4.25}$$

with initial condition

$$p(x,s;y,s) = \delta(x-y)$$

and where

$$b_{ij}(x) = \sum_{k=1}^{m} g_{ik}(x)\ g_{jk}(x) = \left[[g(x)]\ [g(x)]^{\mathsf{T}} \right]_{ij}$$

Finally, recall that in the steady-state, the probability density is stationary, and Equation (4.25) simplifies to the stationary Fokker-Planck equation

$$0 = -\sum_{i=1}^{n} \frac{\partial}{\partial x_i} \left(\bar{f}_i(x)\, p_\infty(x) \right) + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial^2}{\partial x_i\, \partial x_j} \left(b_{ij}(x)\, p_\infty(x) \right) \tag{4.26}$$

where $p_\infty(x)$ denotes the stationary probability density (if it exists).

In this section, we propose a method for computing the controllability function that relies heavily on the above framework. We are motivated initially by observations concerning the relationship between the controllability function of a linear system and the stationary density of the corresponding linear stochastically excited system.

Linear Case

The solution of a linear stochastic initial value problem is given by a variation of constants formula for the state process. Moreover, the time evolutions of its mean and its covariance are governed by a pair of ODEs. These results are summarized in the following theorem (see, e.g., [36]).

Theorem 4.3.1 *Let $\{W_t\}$ be an m-vector process with stationary orthogonal increments, x_0 an n-vector random variable orthogonal to $\mathcal{H}_W$ with*

$$\mu_0 = E[x_0] \tag{4.27}$$

$$R_0 = E\left[(x_0 - \mu_0)(x_0 - \mu_0)^{\mathsf{T}}\right] \tag{4.28}$$

and $A(\cdot)$, $B(\cdot)$ matrices of dimension $n \times n$ and $n \times m$, respectively, whose elements are piecewise continuous real-valued functions. Then the stochastic initial value problem

$$dX_t = A(t)\,X_t\,dt + B(t)\,dW_t \qquad X_0 = x_0 \tag{4.29}$$

has the unique solution

$$X_t = \Phi(t,0)\,x_0 + \int_0^t \Phi(t,s)\,B(s)\,dW_s \tag{4.30}$$

where $\Phi(\cdot,\cdot)$ is the transition matrix corresponding to $A(\cdot)$. The moments $\mu(t) = E[X_t]$ and $R(t) = E\left[(X_t - \mu(t))(X_t - \mu(t))^{\mathsf{T}}\right]$ are the unique solutions of the initial value problems

$$\dot{\mu}(t) = A(t)\,\mu(t) \qquad \mu(0) = \mu_0 \tag{4.31}$$

$$\dot{R}(t) = A(t)\,R(t) + R(t)\,A^{\mathsf{T}}(t) + B(t)\,B^{\mathsf{T}}(t) \qquad R(0) = R_0 \tag{4.32}$$

□

Remark 4.3.2 *It is well known that the response of a linear system to a Gaussian stochastic input such as the white noise process ζ_t is also a Gaussian process (see, e.g., [126]). Thus, Theorem 4.3.1 implies that the state process X_t, given by (4.30), is Gaussian, with mean $\mu(t)$ and covariance $R(t)$ evolving according to (4.31) and (4.32), respectively. Additional moments are not required to characterize the density.* □

We can immediately apply this result to the case of an asymptotically stable LTI system.

Corollary 4.3.3 *Consider the LTI system realization (A, B, C) with A stable and controllability Gramian matrix W_c. Suppose that we replace the input signal $u(t)$ with the sample function of a Gaussian white noise process $\{\zeta_t\}$. Suppose that the evolution of the white noise driven system generates the state process X_t with mean $\mu(t)$ and covariance $R(t)$. Then the mean and covariance satisfy*

$$\lim_{t\to\infty} \mu(t) = 0 \tag{4.33}$$

$$\lim_{t\to\infty} R(t) = W_c \tag{4.34}$$

Proof The key observation is that (4.32) is also satisfied if we replace $R(t)$ with the finite-time-horizon Gramian matrix

$$W(t) = \int_0^t \exp(As)\, B\, B^{\mathsf{T}} \exp\left(A^{\mathsf{T}} s\right) ds \tag{4.35}$$

and let $W(0) = 0$ (see, e.g., [72]). By asymptotic stability, $R(t) = W(t) \to W_c$ as $t \to \infty$. In addition, asymptotic stability together with Equation (4.31) implies that $\mu(t) \to 0$ as $t \to \infty$. ■

Thus, for a stochastically excited LTI system, the transition density function $p(x, t; y, s)$ describing the random properties of the state process X_t is Gaussian.

The stationary density $p_\infty(x)$ has zero mean and covariance equal to the controllability Gramian matrix W_c, i.e.,

$$p_\infty(x) = \left[(2\pi)^n \det(W_c)\right]^{-1/2} \exp\left(-\frac{1}{2}\, x^\mathsf{T} {W_c}^{-1}\, x\right) \tag{4.36}$$

Recall that in the LTI case, the controllability function L_c is given by the quadratic form in Equation (4.5). Thus, in the LTI case, the controllability function L_c and the stationary density p_∞ are related exactly by

$$p_\infty(x) = \left[(2\pi)^n \det(W_c)\right]^{-1/2} \exp\left(-L_c(x)\right) \tag{4.37}$$

and

$$L_c(x) = -\log\left(p_\infty(x)\right) + \log\left(\left[(2\pi)^n \det(W_c)\right]^{-1/2}\right) \tag{4.38}$$

Nonlinear Setting

In the nonlinear setting, the density $p(x,t;y,s)$, and in particular the stationary density $p_\infty(x)$, are not, in general, Gaussian, nor determined completely by their mean and covariance, i.e., higher order moments are involved. However, because the balancing coordinate transformation is local to a neighborhood of the origin, we are mainly interested in capturing a local characterization of the controllability function. In light of this, Equation (4.38) suggests that a useful approximation of L_c is defined by

$$L_c'(x) \stackrel{\triangle}{=} -\log\left(p_\infty(x)\right) + C \tag{4.39}$$

where C is a normalizing constant, dependent on the particular system, such that $L_c'(0) = 0$. By Equation (4.38), L_c' specializes to the exact L_c in the LTI case.

Remark 4.3.4 *The approximation L_c' captures the nonlinearity intrinsic to the realization (f,g), which is manifested in the stationary density p_∞ through the*

evolution of the nonlinear stochastically excited system. It provides a useful working approximation, i.e., a reasonably accurate measure of the degree to which state components are controllable in a neighborhood of the origin. Substitution of the approximation L_c' for L_c into the Scherpen balancing procedure results in a realization that is not balanced, but nearly balanced. The property of equal controllability and observability of state components is satisfied so closely that the attractive properties of such a realization in terms of model reduction are still enjoyed. □

There exist certain nonlinear systems for which Equation (4.39) provides an exact, rather than approximate, formula for the controllability function. As one simple example, consider the process X_t governed by the first-order SDE

$$dX_t = -\nabla \phi(X_t) + dW_t \tag{4.40}$$

where $\phi : \mathbb{R}^n \to \mathbb{R}^n$ is a C^1 map such that $-\nabla \phi$ is asymptotically stable.

Remark 4.3.5 *In general, a process modeled by a first-order SDE is referred to as a* Langevin process *or an* Ornstein-Uhlenbeck process. *The terminology originates with the equation of Langevin [89], which describes the motion of a free particle in a viscous fluid, where the random noise models the impulsive forces due to collisions between the fluid molecules and the free particle.* □

The stationary Fokker-Planck equation for the steady-state transition density p_∞ of the Langevin process governed by (4.40) is given by

$$0 = \frac{1}{2} \sum_{i=1}^{n} \frac{\partial^2 p_\infty}{\partial {x_i}^2}(x) + \sum_{i=1}^{n} \frac{\partial^2 \phi}{\partial {x_i}^2}(x)\ p_\infty(x) + \sum_{i=1}^{n} \frac{\partial \phi}{\partial x_i}(x)\ \frac{\partial p_\infty}{\partial x_i}(x), \tag{4.41}$$

i.e.,

$$0 = \frac{1}{2} \triangle p_\infty(x) + p_\infty(x)\ \triangle \phi(x) + (\nabla p_\infty(x))^{\mathsf{T}} \nabla \phi(x) \tag{4.42}$$

Proposition 4.3.6 *The density function*

$$p_{\infty}^{\mathrm{MB}}(x) = C \exp\left(-2\,\phi(x)\right) \tag{4.43}$$

satisfies the stationary Fokker-Planck equation (4.42) where C is a constant such that $\int p_{\infty}^{\mathrm{MB}} = 1$.

Proof Equation (4.42) follows directly from

$$\begin{aligned} \nabla\, p_{\infty}^{\mathrm{MB}}\,(x) &= -2\,C \exp\left(-2\,\phi\,(x)\right)\,\nabla\,\phi\,(x) \\ &= -2\,p_{\infty}^{\mathrm{MB}}\,(x)\,\nabla\,\phi\,(x) \end{aligned} \tag{4.44}$$

and

$$\begin{aligned} \triangle p_{\infty}^{\mathrm{MB}}\,(x) &= -2\,\nabla\,\cdot\,\left(p_{\infty}^{\mathrm{MB}}\,(x)\,\nabla\,\phi\,(x)\right) \\ &= -2\left[p_{\infty}^{\mathrm{MB}}\,(x)\,\triangle\phi\,(x) + \left(\nabla\, p_{\infty}^{\mathrm{MB}}\,(x)\right)^{\mathsf{T}}\nabla\,\phi\,(x)\right] \end{aligned} \tag{4.45}$$

■

Remark 4.3.7 *A density of the form (4.43) is referred to as a Maxwell-Boltzmann density. It originally appeared in the work of Maxwell and Boltzmann on modeling heat in a medium as the random motion of the constituent molecules, where ϕ represents the total energy in the system. The steady-state density describing the random properties of molecule positions and velocities is of the form (4.43).* □

Now, using (4.39), define

$$L_c^{\mathrm{MB}}\,(x) = -\log\left(p_{\infty}^{\mathrm{MB}}(x)\right) + \log\,(C) = 2\,\phi(x) \tag{4.46}$$

Proposition 4.3.8 *The function L_c^{MB} satisfies the HJB equation (4.14) and thus is the unique controllability energy function for the Langevin system, i.e., stable affine nonlinear system with $f(x) = -\nabla\,\phi(x)$ and $g(x) = \mathbb{1}$.*

Proof The statement follows from straightforward substitution. ∎

Systems modeled by the first-order SDE (4.40) do not comprise a sufficiently general class to be useful in many situations of interest. In particular, if $g(x) \neq \mathbb{I}$, the density p_∞^{MB} does not generally satisfy the stationary Fokker-Planck equation and the function L_c^{MB} does not generally satisfy the HJB equation. In the next section, we seek conditions under which a broader class of systems admits an exact relationship between the stationary density and the controllability function.

4.3.2 Second-Order Mechanical Systems

In this section we determine conditions under which the controllability function for a second-order mechanical system can be expressed exactly in terms of the stationary density for the corresponding stochastically excited system. We adopt and modify somewhat the notation and framework of Fuller [51] and Zhu and Yang [171]. These authors have presented conditions under which exact solutions of the stationary Fokker-Planck equation can be derived. We show that in certain cases, the same conditions are sufficient for expressing the controllability function in terms of the stationary density, while in other cases, additional conditions are required.

Hamiltonian System Perturbed by Dissipation and Forcing

We consider a forced, dissipatively perturbed, n-DOF Hamiltonian system as described in Section 2.7. Let $q = (q_1, \ldots, q_n) \in \mathbb{R}^n$ and $p = (p_1, \ldots, p_n) \in \mathbb{R}^n$ denote, respectively, the generalized displacements and momenta. Let the Hamiltonian $H' = H'(q, p)$, i.e., the sum of the kinetic and potential energies of the system, be C^2. Let $c'_{ij} = c'_{ij}(q, p)$ for $i, j \in \underline{n}$ be C^1 functions representing gener-

alized nonlinear dissipation coefficients. Let $d_{ij} = d_{ij}(q,p)$ for $i,j \in \underline{n}$ be C^2. The system that we consider is governed by the equations of motion, for $i \in \underline{n}$

$$\dot{q}_i = \frac{\partial H'}{\partial p_i} \tag{4.47}$$

$$\dot{p}_i = -\frac{\partial H'}{\partial q_i} - \sum_{j=1}^{n} c'_{ij} \frac{\partial H'}{\partial p_j} + \sum_{k=1}^{m} d_{ik}\, u \tag{4.48}$$

The system is realized in standard state-space form with coordinates $x = (q,p) \in \mathbb{R}^{2n}$ and

$$\begin{aligned} f_i &= \frac{\partial H'}{\partial p_i} \qquad i = 1,\ldots,n \\ f_i &= -\frac{\partial H'}{\partial q_i} - \sum_{j=1}^{n} c'_{ij} \frac{\partial H'}{\partial p_j} \qquad i = n+1,\ldots,2n \end{aligned} \tag{4.49}$$

$$\begin{aligned} (g_k)_i &= 0 \qquad i = 1,\ldots,n; \quad k = 1,\ldots,m \\ (g_k)_i &= d_{ik} \qquad i = n+1,\ldots,2n; \quad k = 1,\ldots,m \end{aligned} \tag{4.50}$$

The output map h is irrelevant for purposes of the discussion here.

Stochastically Excited System

The corresponding stochastically excited system is governed by the SDEs, for $i \in \underline{n}$

$$dQ_i = \frac{\partial H'}{\partial P_i}\, dt \tag{4.51}$$

$$\begin{aligned} dP_i = &-\left(\frac{\partial H'}{\partial Q_i} + \sum_{j=1}^{n} c'_{ij} \frac{\partial H'}{\partial P_j} + \frac{1}{2} \sum_{j=1}^{n} \sum_{k=1}^{m} \frac{\partial d_{ik}}{\partial P_j}\, d_{jk} \right) dt \\ &+ \sum_{k=1}^{m} d_{ik}\, (dW_t)_i \end{aligned} \tag{4.52}$$

where we have adopted the usual notation by substituting Q for q and P for p when dealing with the corresponding random variables.

It is usually the case that the correction terms $\frac{1}{2} \sum_{j=1}^{n} \sum_{k=1}^{m} \frac{\partial d_{ik}}{\partial P_j}\, d_{jk}$, $i \in \underline{n}$, can be split into two parts: one which modifies the conservative forces and the other that

modifies the damping forces (see [171]). We assume that this can be accomplished, and

(i) combine the first part with $-\dfrac{\partial H'}{\partial Q_i}$ to form effective conservative force terms $\dfrac{\partial H}{\partial Q_i}$ with a new Hamiltonian $H = H(Q,P)$ such that $\dfrac{\partial H}{\partial P_i} = \dfrac{\partial H'}{\partial P_i}$;

(ii) combine the second part with $\displaystyle\sum_{j=1}^{n} c'_{ij}\,\frac{\partial H'}{\partial P_j}$ to form effective dissipative force terms $\displaystyle\sum_{j=1}^{n} c_{ij}\,\frac{\partial H}{\partial P_j}$ with new damping coefficients $c_{ij} = c_{ij}(Q,P)$.

Equations (4.51) and (4.52) can be rewritten, for $i \in \underline{n}$

$$dQ_i = \frac{\partial H}{\partial P_i}\,dt \tag{4.53}$$

$$dP_i = -\left(\frac{\partial H}{\partial Q_i} + \sum_{j=1}^{n} c_{ij}\,\frac{\partial H}{\partial P_j}\right)dt + \sum_{k=1}^{m} d_{ik}\,(dW_t)_i \tag{4.54}$$

Stationary Fokker-Planck Equation

The stationary Fokker-Planck equation governing the stationary transition density $p_\infty = p_\infty(q,p)$ associated with the SDEs (4.53) and (4.54) is given by

$$\begin{aligned} 0 &= \sum_{i=1}^{n}\left[-\frac{\partial}{\partial Q_i}\left(\frac{\partial H}{\partial P_i}\,p_\infty\right) + \frac{\partial}{\partial P_i}\left(\frac{\partial H}{\partial Q_i}\,p_\infty\right)\right] \\ &+ \sum_{i=1}^{n}\left[\frac{\partial}{\partial P_i}\left(\sum_{j=1}^{n} c_{ij}\,\frac{\partial H}{\partial P_j}\,p_\infty\right) + \frac{1}{2}\sum_{j=1}^{n}\frac{\partial^2}{\partial P_i\,\partial P_j}\left(b_{ij}\,p_\infty\right)\right] \end{aligned} \tag{4.55}$$

where

$$b_{ij} = \sum_{k=1}^{m} d_{ik}\,d_{jk} = \left[[d]\;[d]^{\mathsf{T}}\right]_{ij}$$

and subject to boundary conditions (vanishing probability flow)

$$\lim_{\|(q,p)\|\to\infty} \frac{\partial H}{\partial P_i}\,p_\infty = 0 \tag{4.56}$$

and

$$\lim_{\|(q,p)\|\to\infty} \left(\frac{\partial H}{\partial Q_i} + \sum_{j=1}^{n} c_{ij}\,\frac{\partial H}{\partial P_j}\right) p_\infty + \frac{1}{2}\sum_{j=1}^{n}\frac{\partial}{\partial P_j}\left(b_{ij}\,p_\infty\right) = 0 \tag{4.57}$$

Observe that the first summation term on the right-hand-side of (4.55) is equal to the Poisson bracket of p_∞ and H, i.e.,

$$\begin{aligned}\{p_\infty, H\} &= \sum_{i=1}^{n}\left[-\frac{\partial H}{\partial P_i}\frac{\partial p_\infty}{\partial Q_i}+\frac{\partial H}{\partial Q_i}\frac{\partial p_\infty}{\partial P_i}\right] \\ &= \sum_{i=1}^{n}\left[-\frac{\partial}{\partial Q_i}\left(\frac{\partial H}{\partial P_i}\,p_\infty\right)+\frac{\partial}{\partial P_i}\left(\frac{\partial H}{\partial Q_i}\,p_\infty\right)\right] \end{aligned} \tag{4.58}$$

Thus, we can rewrite the stationary Fokker-Planck equation (4.55) as

$$0=\{p_\infty, H\}+\sum_{i=1}^{n}\left[\frac{\partial}{\partial P_i}\sum_{j=1}^{n}\left(c_{ij}\,\frac{\partial H}{\partial P_j}\,p_\infty\right)+\frac{1}{2}\sum_{j=1}^{n}\frac{\partial^2}{\partial P_i\,\partial P_j}\left(b_{ij}\,p_\infty\right)\right] \tag{4.59}$$

Existence of Smooth Stationary Densities

Before proceeding, we must establish the existence of a smooth stationary density in this special case of a stochastically excited and dissipatively perturbed Hamiltonian system. We appeal to Theorem 2.6.17. First we show that the corresponding deterministic system has the required property of local strong accessibility.

A system of the form (4.47)-(4.48) can be derived from the equations of motion (see Section 2.7)

$$M\left(q\right)\,\ddot{q}+C\left(q,\dot{q}\right)+N\left(q,\dot{q}\right)=F \tag{4.60}$$

Let q_d and $\dot{q}_d$ represent desired position and velocity trajectories. Define $e=q-q_d$ as the error between the actual and desired trajectories. Consider the control law

$$F=M\left(q\right)\left(\ddot{q}_d-K_v\,\dot{e}-K_p\,e\right)+C\left(q,\dot{q}\right)+N\left(q,\dot{q}\right)+w \tag{4.61}$$

where K_v and K_p are constant gain matrices and w is an exogenous input vector. The resulting error dynamics can be written as

$$M\left(q\right)\,\left(\ddot{e}+K_v\,\dot{e}+K_p\,e\right)=w \tag{4.62}$$

Since M is positive definite for all q, we can write

$$\ddot{e}+K_v\,\dot{e}+K_p\,e=M^{-1}\left(q\right)\,w \tag{4.63}$$

We can choose K_v and K_p so that this linear ODE governing the error yields a stable, controllable LTI system. Moreover, for the LTI system, controllability is equivalent to local strong accessibility. Thus, using the above feedback transformation, we may conclude that the forced, dissipatively perturbed, Hamiltonian system with equations of motion (2.115) is locally strongly accessible.

Remark 4.3.9 *The control law (4.61) is often referred to as the* computed torque *control law in the robotics literature (see, e.g., [112]).* □

It is clear that the completeness property is satisfied in the case that (4.60) yields a linear system. We now argue, formally, that there are many interesting cases in the nonlinear setting for which the vector fields in the Lie algebra generated by $\{f, g_1, \ldots, g_n\}$ are complete.

Observe that the Hamiltonian is typically of the form

$$H(q,p) = p^{\mathsf{T}} M(q)^{-1} p + U(q) \tag{4.64}$$

By the conservation law $\dot{H} = 0$ for the system with zero dissipation, we have that q lies within the interior a n-dimensional torus (a compact set) for all time (with zero dissipation q lies on the torus; otherwise within the interior). Furthermore, p lies within the interior of a compact set formed by taking a union of ellipsoids parameterized by q for all time. Thus, there exist no exploding solutions of $\dot{x} = f(x)$.

Now, consider the case where the d_{ik} are constant vector fields. This situation is a common one, e.g., torque inputs at the joints of a serial manipulator. Clearly, such vector fields produce no exploding solutions. On the other hand, it is possible, due to the nonlinear mass matrix $M(q)$, for brackets of the form $[f, g_i]$ to produce vector fields that are not complete. For purposes of this discussion, we take the

position, without justification, that such brackets will often produce complete vector fields, and assume that smooth transition densities exist for the systems with which we work.

Constant Parameter Case

We first consider the case where the parameters c'_{ij} and d_{ik} are independent of q and p, i.e., are constants. In this situation, the correction term vanishes, $H = H'$, and $c_{ij} = c'_{ij}$. The following is modified from Fuller [51].

Theorem 4.3.10 (Fuller [51]) *Consider the stochastically excited system corresponding to the forced, dissipatively perturbed, n-DOF Hamiltonian system governed by the SDEs (4.53)-(4.54) where H is the Hamiltonian. Suppose that the coefficients c_{ij} and d_{ik} are independent of q and p. Furthermore, suppose that the following constant ratio holds for all $i, j \in \underline{n}$:*

$$\frac{c_{ij}}{b_{ij}} = \ell = \text{constant} \tag{4.65}$$

Then the unique stationary density p_∞ that satisfies Equation (4.59) is

$$p_\infty(q,p) = C \exp\left(-2\,\ell\, H(q,p)\right) \tag{4.66}$$

where C is a constant such that $\int p_\infty = 1$.

Proof Observe that p_∞ is a functional of H, i.e., $p_\infty = p_\infty(H(q,p))$, which implies by Lemma 2.7.6 that $\{p_\infty, H\} = 0$. Observe also that

$$\frac{\partial p_\infty}{\partial P_j} = -2\,\ell\,\frac{\partial H}{\partial P_j}\,p_\infty \tag{4.67}$$

Since the c_{ij} and d_{ij} are constants, the stationary Fokker-Planck equation (4.59) can be written

$$0 = \sum_{i=1}^{n} \frac{\partial}{\partial P_i}\left[\sum_{j=1}^{n}\left(c_{ij}\,\frac{\partial H}{\partial P_j}\,p_\infty + \frac{1}{2}\,b_{ij}\,\frac{\partial p_\infty}{\partial P_j}\right)\right] \tag{4.68}$$

By Equations (4.65) and (4.67) we have, for $i, j \in \underline{n}$

$$c_{ij} \frac{\partial H}{\partial P_j} p_\infty + \frac{1}{2} b_{ij} \frac{\partial p_\infty}{\partial P_j} = c_{ij} \frac{\partial H}{\partial P_j} p_\infty - \ell\, b_{ij} \frac{\partial H}{\partial P_j} p_\infty \tag{4.69}$$

$$= (c_{ij} - \ell\, b_{ij}) \frac{\partial H}{\partial p_j} \tag{4.70}$$

$$= 0 \tag{4.71}$$

Finally, we observe that p_∞ satisfies the boundary conditions (4.56) and (4.57). ■

Remark 4.3.11 *The density (4.66) is of Maxwell-Boltzmann form (see Remark 4.3.7).* □

Remark 4.3.12 *The condition (4.65) is referred to as the* equipartition of energy *condition. The terminology derives from the situation in statistical mechanics where each DOF of a multi-particle system is associated with the same mean energy.* □

Remark 4.3.13 *The equipartition of energy condition imposes a severe restriction on the class of systems for which (4.66) is the stationary density.* □

The controllability function L_c uniquely satisfies the HJB equation (4.14), which, for realization (f, g) given in Equations (4.49) and (4.50), takes the form

$$0 = \{L_c, H\} + \sum_{i=1}^{n} \left[\frac{\partial L_c}{\partial p_i} \sum_{j=1}^{n} \left(-c_{ij} \frac{\partial H}{\partial p_j} + \frac{1}{2} b_{ij} \frac{\partial L_c}{\partial p_j} \right) \right] \tag{4.72}$$

The relationship between the stationary density and the controllability function, and an exact formula for the latter in terms of the Hamiltonian, are given in the following result.

Theorem 4.3.14 *Consider the forced, dissipatively perturbed, n-DOF Hamiltonian control system, governed by the evolution equations (4.47) and (4.48), and*

realized by (f, g) given in Equations (4.49) and (4.50). Under the conditions stated in Theorem 4.3.10, the unique controllability energy function for the system is given by

$$\begin{aligned} L_c(q,p) &= -\log\left(p_\infty(q,p)\right) + C' \\ &= 2\,\ell\,H(q,p) + C' \end{aligned} \tag{4.73}$$

where p_∞ is the stationary density of the corresponding stochastically excited system and C' is a constant such that $L_c(0,0) = 0$.

Proof It is necessary and sufficient to show that L_c satisfies Equation (4.72). Since L_c is a functional of H, Lemma 2.7.6 implies that $\{L_c, H\} = 0$. Furthermore, $\dfrac{\partial L_c}{\partial p_j} = 2\,\ell\,\dfrac{\partial H}{\partial p_j}$, so that Equation (4.72) becomes

$$0 = \sum_{i=1}^{n}\left[2\,\ell\,\frac{\partial H}{\partial p_i}\sum_{j=1}^{n}\left(-c_{ij}\,\frac{\partial H}{\partial p_j} + \ell\,b_{ij}\,\frac{\partial H}{\partial p_j}\right)\right] \tag{4.74}$$

$$= \sum_{i=1}^{n}\left[2\,\ell\,\frac{\partial H}{\partial p_i}\sum_{j=1}^{n}\frac{\partial H}{\partial p_j}\left(\ell\,b_{ij} - c_{ij}\right)\right] \tag{4.75}$$

which is clearly satisfied given the equipartition of energy condition (4.65). ■

General Setting

We now consider the more general situation where the parameters c'_{ij} and d_{ik} are permitted to be functions of q and p. The following is modified from Zhu and Yang [171].

Theorem 4.3.15 (Zhu and Yang [171]) *Consider the stochastically excited system corresponding to the forced, dissipatively perturbed, n-DOF Hamiltonian system governed by the SDEs (4.53)-(4.54) where H is the Hamiltonian. Suppose that*

the following ratio holds for all $i \in \underline{n}$ *and for some functional* h *of* H*:*

$$\frac{\sum_{j=1}^{n} \left(2\, c_{ij} \dfrac{\partial H}{\partial P_j} + \dfrac{\partial b_{ij}}{\partial P_j} \right)}{\sum_{j=1}^{n} b_{ij} \dfrac{\partial H}{\partial P_j}} = h\left(H\right) \tag{4.76}$$

Then the unique stationary density p_∞ *that satisfies Equation (4.59) is*

$$p_\infty\left(q,p\right) = C \exp\left(- \int_0^{H(q,p)} h(u)\, du \right) \tag{4.77}$$

where C *is a constant such that* $\int p_\infty = 1$*.*

Proof Assume that there exists $\phi\left(H\right)$ such that

$$p_\infty\left(q,p\right) = C \exp\left(-\phi\left(H(q,p)\right)\right) \tag{4.78}$$

Then as before $\{p_\infty, H\} = 0$ and $\dfrac{\partial p_\infty}{\partial P_j} = -\dfrac{\partial \phi}{\partial H} \dfrac{\partial H}{\partial P_j} p_\infty$. The stationary Fokker-Planck equation (4.59) can then be written

$$0 = \sum_{i=1}^{n} \left[\sum_{j=1}^{n} \left(2\, c_{ij} \frac{\partial H}{\partial P_j} + \frac{\partial b_{ij}}{\partial P_j} - b_{ij} \frac{\partial H}{\partial P_j} \frac{\partial \phi}{\partial H} \right) \right] \tag{4.79}$$

which is clearly satisfied if we assign $\dfrac{\partial \phi}{\partial H} = h\left(H\right)$ where $h\left(H\right)$ is defined by Equation (4.76). Thus, the desired functional ϕ is obtained through integration yielding Equation (4.77). ■

Remark 4.3.16 *The density (4.77) is of Maxwell-Boltzmann form.* □

Remark 4.3.17 *The condition (4.76) is analogous to an equipartition of energy condition, again imposing a severe restriction on the class of systems for which (4.77) is the stationary density.* □

The relationship between the stationary density and the controllability function, and an exact formula for the latter in terms of the Hamiltonian, are given in the following result.

Theorem 4.3.18 *Consider the forced, dissipatively perturbed, n-DOF Hamiltonian control system, governed by the evolution equations (4.47) and (4.48), and realized by (f, g) given in Equations (4.49) and (4.50). Suppose that the following ratio holds for all $i \in \underline{n}$ and for some functional r of H:*

$$\frac{\sum_{j=1}^{n} c_{ij} \dfrac{\partial H}{\partial p_j}}{\sum_{j=1}^{n} b_{ij} \dfrac{\partial H}{\partial p_j}} = r(H) \tag{4.80}$$

Then the unique controllability energy function for the system is given by

$$L_c(q,p) = 2 \int_0^{H(q,p)} r(u)\, du + C' \tag{4.81}$$

where C' is a constant such that $L_c(0,0) = 0$. Furthermore, if the b_{ij} are independent of p then

$$L_c(q,p) = -\log\left(p_\infty(q,p)\right) + C' \tag{4.82}$$

where p_∞ is the stationary density of the corresponding stochastically excited system.

Proof Assume that $L_c(q,p) = \phi(H(q,p))$ for some functional ϕ of H. Then $\{L_c, H\} = 0$. The HJB equation (4.72) can be written

$$0 = \sum_{i=1}^{n} \frac{\partial H}{\partial p_i} \frac{\partial \phi}{\partial H} \left(b_{ij} \frac{\partial H}{\partial p_j} \frac{\partial \phi}{\partial H} - 2\, c_{ij} \frac{\partial H}{\partial p_j} \right) \tag{4.83}$$

which is clearly satisfied if we assign $\dfrac{\partial \phi}{\partial H} = 2\, r(H)$ where $r(H)$ is defined by Equation (4.80). Thus, the desired functional ϕ is obtained through integration yielding Equation (4.81). Moreover, if the b_{ij} are independent of p then $2\, r(H) = h(H)$ where $h(H)$ is defined in Equation (4.76). In that case Equation (4.82) holds. ∎

4.3.3 Monte-Carlo Experiments

In situations where we do not have an exact formula for the controllability function, we wish to use the approximation given by Equation (4.39) in the nonlinear balancing procedure. This requires determining the stationary density $p_\infty(x)$, or a suitable estimate. Approximating $p_\infty(x)$ via Monte-Carlo experiments is a natural approach.

Each experiment corresponds to a numerical simulation of the white noise driven system (4.22), with zero initial state and input corresponding to a discretized approximation of a sample path for Gaussian white noise. The numerical schemes that we used for approximating a white noise signal and integrating the SDEs are detailed in Appendix C. The state response trajectory X_t for each experiment is sampled and recorded. An approximation of the steady-state, i.e., $\lim_{t\to\infty} X_t$, is generated by simulating the system over a sufficiently large time period, e.g., several multiples of its largest time constant.

We approximate the time evolution of the density function $p(x, t; 0, 0)$ by histogramming the collection of trajectories at fixed values of t. Likewise, we approximate the stationary density $p_\infty(x)$ by histogramming the collection of steady-state responses. Naturally, the approximations improve as the number of experiments in the collection increases. Moreover, a larger set of experiments allows for histogramming at a higher resolution. Statistics of the density such as $\mu(t)$, μ_∞, $R(t)$, and R_∞ can be computed and analyzed to confirm the correctness of the data.

The results of Monte-Carlo experiments to approximate the controllability functions for an example problem are presented in Section 4.6.

4.4 Computing the Morse Coordinate Transformation

Recall that for an LTI system, the energy functions L_c and L_o globally take the form of quadratic functions given, respectively, by (4.5) and (4.6). We wish to generalize the linear balancing procedure to the nonlinear setting, but the functions L_c and L_o are not, in general, quadratic. However, we can appeal to some important results from critical point theory (see, e.g., [108]) in order to find a change of coordinates under which a smooth function takes a quadratic form locally around a non-degenerate critical point. The key result is the Morse lemma [110], which guarantees the existence of the desired canonical form for functions with a non-degenerate critical point defined on a finite-dimensional manifold, and an analogous result of Palais [125], which generalizes the notion to functions defined on a Hilbert space. The established results are presented from various points of view in [16, 57, 60, 88, 108].

4.4.1 The Morse-Palais Lemma

The functions L_c and L_o are smooth real-valued mappings defined on local coordinates $x \in \mathbb{R}^n$ for n-dimensional manifold M. Thus, we can use the fact that the local behavior of a smooth real-valued function on a manifold is known at almost every point up to diffeomorphism. To see this, we introduce the following terminology (see [57, 60]).

Definition 4.4.1 (Critical Point) *A point p is said to be a* critical point *of the smooth real-valued function f if the partial derivatives with respect to local coordi-*

nates $\{x_1, \ldots, x_n\}$ *satisfy*

$$\frac{\partial f}{\partial x_i}(p) = 0 \qquad i \in \underline{n} \tag{4.84}$$

Otherwise, the point p *is said to be a* regular point *of* f. □

Remark 4.4.2 *If a point* p *is regular, then we can invoke the implicit function theorem and choose a coordinate system so that* f *is simply the first coordinate function in a neighborhood of* p*. Thus the local behavior of* f *around regular points is completely characterized.* □

The functions L_c and L_o each have a critical point at 0. We now focus on characterizing the local behavior of a function around critical points.

Definition 4.4.3 (Non-degenerate Critical Point) *A critical point* p *of the smooth real-valued function* f *is called* non-degenerate *if the Hessian matrix of second partials at* p

$$D^2 f(p) = \left[\frac{\partial^2 f}{\partial x_i \, \partial x_j}(p)\right] \tag{4.85}$$

is nonsingular. Otherwise p *is called* degenerate. □

Definition 4.4.4 (Morse Function) *A smooth real-valued function* f *with a non-degenerate critical point at* p *is said to be a* Morse function *at* p. □

Remark 4.4.5 *Under conditions outlined in Section 4.2, the functions* L_c *and* L_o *are Morse functions at* 0. □

Definition 4.4.6 (Index, Nullity) *The* index *of a bilinear functional* $\mathcal{A}$ *on* $\mathbb{R}^n$ *is defined to be the maximal dimension of a subspace of* $\mathbb{R}^n$ *on which* $\mathcal{A}$ *is negative definite. The* nullity *of* $\mathcal{A}$ *is defined as the dimension of the nullspace of* $\mathcal{A}$*. The index and nullity of a critical point* p *of function* f *are, respectively, the index and nullity of the bilinear functional* $\mathcal{A}(x, y) = \langle\, D^2 f(p)\, x, y \,\rangle$. □

There is a canonical form for a Morse function f in the neighborhood of its non-degenerate critical point p, completely described by the index of p. This idea is made precise in finite-dimensions in a theorem by Morse [110] and generalized to Hilbert spaces in a theorem by Palais [125]. The theorems state that there exists a local change of coordinates under which a Morse function is quadratic on some neighborhood of its non-degenerate critical point. We refer to this result as the Morse-Palais lemma, for which we present a version based on that in Milnor [108] and Lang [88]. Henceforth we assume without loss of generality that $p = 0$.

Theorem 4.4.7 (Morse-Palais) *Let f be a smooth real-valued function defined on an open neighborhood $\mathcal{O}$ of 0 in the Hilbert space $\mathcal{E}$. Assume that $f(0) = 0$ and that 0 is a non-degenerate critical point of f. Then there exists a neighborhood $U \subset \mathcal{O}$ of 0, a local change of coordinates ϕ on U, and an invertible symmetric operator $\mathcal{A}$ such that*

$$f(x) = \langle \mathcal{A}\,\phi(x), \phi(x) \rangle_{\mathcal{E}} \qquad x \in U \tag{4.86}$$

□

We defer the proof momentarily and present some related and supporting results.

Corollary 4.4.8 *Let f be a smooth real-valued function defined on an open neighborhood $\mathcal{O}$ of 0 in the Hilbert space $\mathcal{E}$. Assume that $f(0) = 0$ and that 0 is a non-degenerate critical point of f. Then there exists a neighborhood $U \subset \mathcal{O}$ of 0, a local change of coordinates $z = \xi(x)$ on U, and an orthogonal decomposition $\mathcal{E} = \mathcal{F} + \mathcal{F}^{\perp}$ such that if we write $z = \xi(x) = u + v$ with $u \in \mathcal{F}$ and $v \in \mathcal{F}^{\perp}$ then*

$$f(z) = f(\xi(x)) = \langle u, u \rangle_{\mathcal{E}} - \langle v, v \rangle_{\mathcal{E}} \qquad x \in U \tag{4.87}$$

□

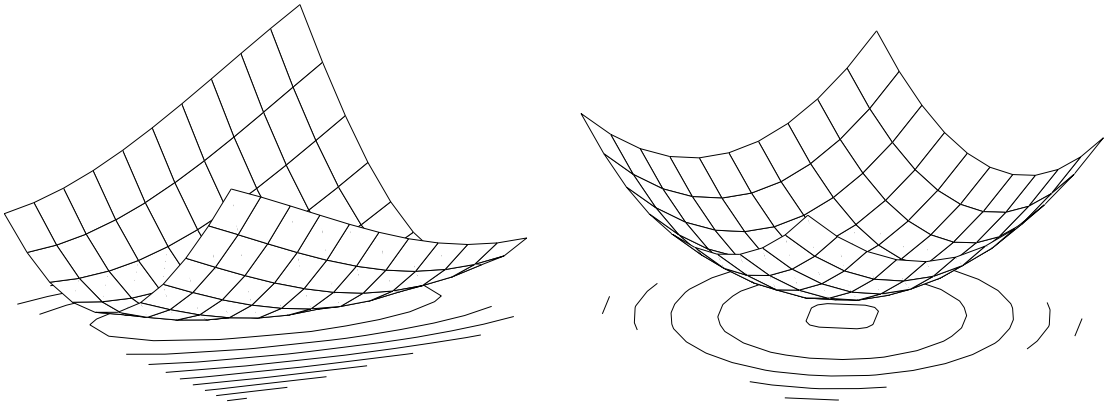


Figure 4.1: An example of a Morse function on $\mathbb{R}^2$ (with level contours) before and after transformation to spherical quadratic form.

Remark 4.4.9 *Consider the special case where $\mathcal{E} = \mathbb{R}^n$ and critical point 0 has index r. Define $z = \xi(x)$ for $x \in U$ and $\psi = \xi^{-1}$ on $\xi(U)$. Then Corollary 4.4.8 implies that*

$$f(x) = f(\psi(z)) = -\sum_{i=1}^{r} z_i^2 + \sum_{i=r+1}^{n} z_i^2 \tag{4.88}$$

In the new coordinates, the function f is said to be in spherical quadratic form. *The transformation is illustrated in Figure 4.1.* □

Definition 4.4.10 (Morse Coordinate Transformation) *A change of coordinates ψ satisfying (4.88) is said to be a* Morse coordinate transformation *for f around* 0. □

The original proof by Morse uses the Gram-Schmidt orthogonalization process which is essentially a coordinate-by-coordinate induction argument. The generalization by Palais is proved without a coordinate-wise procedure, which, as we demonstrate later, is advantageous for purposes of computation. Moreover, certain decompositions of smooth functions with non-degenerate critical points are

integral to the proofs. We merely outline the proofs, emphasizing the decompositions since they are essential to computing the desired transformation. For a concise presentation of the Morse lemma and proof see Milnor [108]. The Palais version is found in [88, 125].

The following lemma provides a decomposition of any smooth real-valued function defined on a finite-dimensional manifold. It is a simple application of the first fundamental theorem of integral calculus.

Lemma 4.4.11 *Let f be a smooth real-valued function defined in an open convex neighborhood $\mathcal{O}$ of 0 in n-dimensional manifold M. Then there exist smooth functions g_i, $i \in \underline{n}$ on $\mathcal{O}$ such that*

$$f(x) = \sum_{i=1}^{n} g_i(x)\, x_i \qquad i \in \underline{n}, \quad x \in \mathcal{O} \tag{4.89}$$

Furthermore, if 0 is a critical point of f then

$$g_i(0) = \frac{\partial f}{\partial x_i}(0) \qquad i \in \underline{n} \tag{4.90}$$

□

The proof is instructive in that it shows us how to compute one such collection of functions g_i, $i \in \underline{n}$.

Proof By the fundamental theorem of calculus

$$f(x) - f(0) = \int_0^1 \frac{df}{dt}(tx)\, dt = \int_0^1 \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(tx)\; x_i\, dt \tag{4.91}$$

Define

$$g_i(x) = \int_0^1 \frac{\partial f}{\partial x_i}(tx)\, dt \tag{4.92}$$

to yield (4.89). ■

Applying Lemma 4.4.11 twice to f around a critical point at 0 results in the following decomposition.

Lemma 4.4.12 *Let f be a smooth real-valued function defined in an open convex neighborhood $\mathcal{O}$ of 0 in n-dimensional manifold M. Let 0 be a critical point of f. Then there exist smooth functions h_{ij}, $i, j \in \underline{n}$ on $\mathcal{O}$ such that*

$$f(x) = \sum_{i=1}^{n} \sum_{j=1}^{n} h_{ij}(x)\, x_i\, x_j \qquad i, j \in \underline{n}, \quad x \in \mathcal{O} \tag{4.93}$$

Moreover, the symmetry property

$$h_{ij}(x) = h_{ji}(x) \qquad i, j \in \underline{n}, \quad x \in \mathcal{O} \tag{4.94}$$

holds, and it is true that

$$h_{ij}(0) = \frac{1}{2} \frac{\partial^2 f}{\partial x_i\, \partial x_j}(0) = \frac{1}{2} D^2 f(0) \tag{4.95}$$

□

We now return to the proof of Theorem (4.4.7). The argument follows from decomposition (4.93). We denote $H(x) = [H(x)]_{ij} = [h_{ij}(x)]$. Some details are omitted here (see [88]).

Morse-Palais Lemma (Theorem 4.4.7)

Proof Non-degeneracy of critical point 0 ensures that $H(0)$ is nonsingular. Equation (4.86) is satisfied if, for all $x \in \mathcal{O}$, we define $\mathcal{A}$ by

$$\mathcal{A}(x) = H(0)\, x \tag{4.96}$$

and define ϕ by

$$\phi(x) = \mathcal{C}(x)\, x \tag{4.97}$$

where

$$\left(\mathcal{C}(x)^{*}\, \mathcal{A}(x)\, \mathcal{C}(x)\right)(x) = H(x)\, x \tag{4.98}$$

and $\mathcal{C}(x)^*$ denotes the adjoint operator. The desired operator-valued map $\mathcal{C}$ is given, for $x \in \mathcal{O}$, by

$$\mathcal{C}(x) = \mathcal{B}(x)^{1/2} \tag{4.99}$$

where $\mathcal{B}$ is defined for $x \in \mathcal{O}$ by

$$\mathcal{B}(x)\ x = H(0)^{-1} H(x)\ x \tag{4.100}$$

The operator square root in (4.99) is guaranteed to exist for x in some neighborhood $U \subset \mathcal{O}$ of 0 because $\mathcal{B}(x)$ is close to the identity operator $\mathbb{I}$ on a neighborhood of 0, and the square root function has a convergent power series expansion near $\mathbb{I}$. ■

Remark 4.4.13 *Corollary 4.4.8 then results directly from the fact that operator $\mathcal{A}$ is symmetric, positive definite on F, and negative definite on $F^{\perp}$. This allows for the change of coordinates $z = \mathcal{A}^{1/2} x$ on F and $z = -\mathcal{A}^{1/2} x$ on $F^{\perp}$ to yield (4.88).* □

4.4.2 Properties

A Morse coordinate transformation ψ for f around 0 is not unique. This can be argued as a consequence of the non-uniqueness of the functions g_i in the decomposition (4.89) and the functions h_{ij} in the decomposition (4.93). Consider the isotropy transformation $T : \mathbb{R}^n \to \mathbb{R}^n$ such that

$$T(x)\, x = x \qquad x \in \mathbb{R}^n \tag{4.101}$$

At each point x, the isotropy $T(x)$ is a pure rotation about an axis passing through the origin and x. Let $G(x) = (g_1(x), \ldots, g_n(x))$. Then Equation (4.89) implies that

$$f(x) = G(x)\, x = G(x)\, T(x)\, x = \widetilde{G}(x)\, x \tag{4.102}$$

where $\widetilde{G}(x) = G(x)\,T(x)$ comprises another set of g_i satisfying (4.89). The non-uniqueness of the functions h_{ij} in (4.93) follows as an immediate consequence.

Remark 4.4.14 *Even though the functions g_i and h_{ij} are not unique, their values at the origin, i.e., $g_i(0)$ and $h_{ij}(0)$, are invariants of the function f and given by first and second derivatives, respectively.* □

The non-uniqueness of the Morse coordinate transformation can also be shown from the following viewpoint. Consider f with index r, and rewrite (4.88) as

$$f\left(\psi(z)\right) = z^{\mathsf{T}}\,E\,z \tag{4.103}$$

where E is the block diagonal matrix

$$E = \left[\begin{array}{c|c} -\mathbb{I}_r & 0 \\ \hline 0 & \mathbb{I}_{n-r} \end{array}\right] \tag{4.104}$$

Let $\Theta_1 \in O(r)$ and $\Theta_2 \in O(n-r)$ and define the block diagonal matrix

$$\Theta = \left[\begin{array}{c|c} \Theta_1 & 0 \\ \hline 0 & \Theta_2 \end{array}\right] \tag{4.105}$$

and the change of coordinates

$$\widehat{\psi}\left(z\right) = \psi\left(\Theta\,z\right) \qquad z \in \psi^{-1}\left(U\right) \tag{4.106}$$

Then

$$f\left(\widehat{\psi}(z)\right) = \left(\Theta z\right)^{\mathsf{T}}\,E\,\left(\Theta z\right) = z^{\mathsf{T}}\,\left(\Theta^{\mathsf{T}}\,E\,\Theta\right)\,z = z^{\mathsf{T}}\,E\,z \tag{4.107}$$

Thus, $\widehat{\psi}$ is also a Morse coordinate transformation for f around 0.

Remark 4.4.15 *A function f with non-degenerate critical point 0 admits a family of Morse coordinate transformations, parameterized by the spaces of orthogonal matrices $O(r)$ and $O(n-r)$. It is not clear, however, if this family exhausts the entire collection of Morse transformations for f around 0.* □

Example 4.4.16 *To illustrate the nonuniqueness property, consider the polynomial function f on $\mathbb{R}^2$*

$$f(x) = 3x_1^3 - x_1^2 x_2 - x_1 x_2^2 + 2x_2^3 + 5x_1^2 - 2x_1 x_2 + 2x_2^2 \tag{4.108}$$

which has a non-degenerate critical point at $(0,0)$. Applying the decomposition (4.93), i.e., $f(x) = x^{\mathsf{T}} H(x)\, x$, yields the invariant

$$H(0) = \begin{bmatrix} 5 & -1 \\ -1 & 2 \end{bmatrix} \tag{4.109}$$

One valid choice for $H(x)$, computed via repeated application of (4.92), is given by

$$H(x) = \begin{bmatrix} 5 + 3x_1 - \frac{1}{3}x_2 & -1 - \frac{1}{3}x_1 - \frac{1}{3}x_2 \\ -1 - \frac{1}{3}x_1 - \frac{1}{3}x_2 & 2 - \frac{1}{3}x_1 + 2x_2 \end{bmatrix} \tag{4.110}$$

Two other valid choices are

$$H(x) = \begin{bmatrix} 5 + 3x_1 & -1 - 0.5x_1 - 0.5x_2 \\ -1 - 0.5x_1 - 0.5x_2 & 2 + 2x_2 \end{bmatrix} \tag{4.111}$$

and

$$H(x) = \begin{bmatrix} 5 + 3x_1 - x_2 & -1 \\ -1 & 2 + 2x_2 - x_1 \end{bmatrix} \tag{4.112}$$

Each different choice of $H(x)$ will result in a different Morse coordinate transformation ψ via (4.100). □

Remark 4.4.17 *We were able to perform the above calculations for a given (contrived) polynomial f. However, in general, there exist no closed form expressions for functions g_i, h_{ij}, or ψ, even if we have a formula for f. Moreover, for our purposes, the Morse function f may be known only at discrete points on a grid. Thus, we always apply Equation (4.92) to our computations.* □

It is worth noting that not every function ψ satisfying (4.88) is smooth, a condition required for a function to be a valid change of coordinates. Consider the following example.

Example 4.4.18 *Let ψ be a function on U such that*

$$z = \psi^{-1}(x) = \left(\frac{\sqrt{f(x)}}{\| x \|} \right) x \qquad x \in U \tag{4.113}$$

Then ψ satisfies (4.103) with $E = \mathbb{I}_n$. However, the first derivative of ψ^{-1} does not exist at 0. The function ψ is not smooth and thus is not a valid change of coordinates. □

Remark 4.4.19 *Observe that the proof of Theorem 4.4.7 also uses a square root operation, but avoids any associated pitfalls by ensuring that the square root operand remains close to the identity.* □

4.4.3 Algorithm

We present here, somewhat loosely, an algorithm for numerical implementation of Theorem 4.4.7 and Corollary 4.4.8. The algorithm is presented more rigorously in Section 4.5.4 once the computational framework has been introduced. The algorithm takes a Morse function f and returns a neighborhood U, Morse coordinate transformation ϕ, and invertible symmetric matrix A under which f takes the desired form (4.86) on U. An additional algorithm takes ϕ, A, and U and returns a coordinate transformation ξ under which f takes the spherical quadratic form (4.87).

The algorithms are based primarily on calculations appearing in the proofs in Section 4.4.1. The main building blocks are as follows.

smooth function decomposition Given smooth real-valued function f, return smooth functions g_i, $i \in \underline{n}$, such that Equation (4.89) holds. This is accomplished via the integration in Equation (4.92).

1. Compute approximate partial derivatives $\dfrac{\partial f}{\partial x_i}$, $i \in \underline{n}$.
2. For each point x in the domain of definition of f, compute approximate integrals $g_i(x) = \displaystyle\int_0^1 \frac{\partial f}{\partial x_i}(tx)\, dt$, $i \in \underline{n}$.

Morse function decomposition Given Morse function f, i.e., f has non-degenerate critical point at 0, return smooth functions h_{ij}, $i, j \in \underline{n}$ such that Equation (4.93) holds. This is accomplished via $n+1$ smooth function decompositions.

1. Apply the smooth function decomposition to f yielding g_i, $i \in \underline{n}$.
2. Apply the smooth function decomposition to each of the g_i yielding h_{ij}, $i, j \in \underline{n}$.

matrix square root Given matrix B close to the identity, return its square root C, i.e., $B = C^2$. The matrix B must satisfy

$$\| \mathbb{I} - B \| < 1 \tag{4.114}$$

In that case, the following algorithm converges to a fixed point corresponding to the desired matrix $C = B^{1/2}$.

$$\begin{aligned} C_{k+1} &= C_k + \frac{1}{2}\left(B - C_k^2\right) \qquad k = 0, 1, \ldots \\ C_0 &= \mathbb{I} \end{aligned} \tag{4.115}$$

The convergence of the sequence $\{C_k\}$ to the fixed point $B^{1/2}$ can be shown to be a consequence of the contraction mapping principle.

Morse-Palais transformation Given Morse function f, return neighborhood U, coordinate transformation ϕ, and invertible symmetric matrix A such that Equation (4.86) holds.

1. Apply the Morse function decomposition to f yielding h_{ij}, $i, j \in \underline{n}$. Let $H(x) = [h_{ij}(x)]$ and $A = H(0)$.
2. For each point x in the domain of definition of f:
   (a) Compute the solution B of the matrix equation $A\,B = H(x)$.
   (b) If $\| \mathbb{1} - B \| < 1$ then:
      i. Apply the matrix square root algorithm to compute $C = B^{1/2}$.
      ii. Let $\phi(x) = C\,x$.
      iii. Include the point x in the neighborhood U.
   (c) Otherwise, the point x is not in the neighborhood U and no further calculations apply.

This procedure provides an estimate of the neighborhood U for which the function can be transformed to the canonical quadratic form. It is possible that the maximal neighborhood is larger.

spherical transformation Given transformation ϕ and invertible symmetric matrix A such that Equation (4.86) holds, return index r and coordinate transformation ψ such that Equation (4.88) holds.

1. Compute the spectral decomposition of matrix A, i.e., $A = V\,\Lambda\,V^{\mathsf{T}}$.
2. Let $E = \mathrm{diag}(|\lambda_1|, \ldots, |\lambda_n|)$.
3. Let r equal the number of λ_i such that $\lambda_i < 0$.
4. Let $R = E\,V^{\mathsf{T}}$.

5. For each point x in the domain of definition of f, let $\psi(x) = R\,\phi(x)$.

Remark 4.4.20 *The terminology "Morse function decomposition" is somewhat misleading since the decomposition (4.93) merely requires a critical point that is not necessarily non-degenerate. However, we adopt the terminology for lack of a better name and because we are applying the decomposition to Morse functions.* □

4.5 Computing the Balancing Transformation

A realization (f, g, h) for a nonlinear system is transformed to balanced form in the Scherpen procedure by composing several local coordinate transformations as illustrated in Figure 4.2. The transformations are, in general, nonlinear, and result from manipulations on the controllability and observability energy functions. Each transformation is a local generalization of a corresponding linear transformation in the procedure for balancing LTI systems.

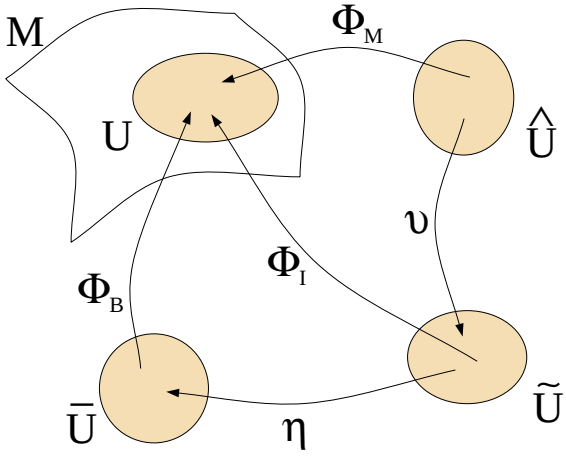


Figure 4.2: Overview of coordinate transformations for nonlinear balancing.

We use the following terminology and notation in describing the required transformations. Applying the Morse-Palais lemma to the controllability function gives

the Morse coordinate transformation, denoted $\Phi_M : \widehat{U} \to U$, where U and $\widehat{U}$ are neighborhoods of 0. We also define local transformations which bring the realization to forms analogous to input-normal and balanced, denoted $\nu : \widehat{U} \to \widetilde{U}$ and $\eta : \widetilde{U} \to \bar{U}$, respectively, where $\widetilde{U}$ and $\bar{U}$ are neighborhoods of 0. The transformations are written

$$x = \Phi_M(\widehat{x}) \qquad x \in U, \quad \widehat{x} \in \widehat{U} \tag{4.116}$$

$$\widetilde{x} = \nu(\widehat{x}) \qquad \widehat{x} \in \widehat{U}, \quad \widetilde{x} \in \widetilde{U} \tag{4.117}$$

$$\bar{x} = \eta(\widetilde{x}) \qquad \widetilde{x} \in \widetilde{U}, \quad \bar{x} \in \bar{U} \tag{4.118}$$

Composing Φ_M with ν^{-1} results in the input-normal coordinate transformation Φ_I, i.e.,

$$x = \Phi_I(\widetilde{x}) = \Phi_M(\nu^{-1}(\widetilde{x})) \qquad x \in U, \quad \widetilde{x} \in \widetilde{U} \tag{4.119}$$

Composing Φ_M with ν^{-1} and η^{-1} results in the balancing transformation Φ_B, i.e.,

$$x = \Phi_B(\bar{x}) = \Phi_M(\nu^{-1}(\eta^{-1}(\bar{x})) \qquad x \in U, \quad \bar{x} \in \bar{U} \tag{4.120}$$

The balanced realization $\left(\bar{f}, \bar{g}, \bar{h}\right)$ is given by

$$\begin{aligned} \bar{f}(\bar{x}) &= [D\Phi_B(z)]^{-1} f\left(\Phi_B(z)\right) \\ \bar{g}_i(\bar{x}) &= [D\Phi_B(z)]^{-1} g_i\left(\Phi_B(z)\right) \qquad i \in \underline{m} \\ \bar{h}(\bar{x}) &= h\left(\Phi_B(z)\right) \end{aligned}$$

Remark 4.5.1 *There is an equivalent dual procedure (see [140]), in which the Morse-Palais lemma is applied to the observability energy function, and the intermediate step takes the realization to output-normal (instead of input-normal) form. Its use in an appropriately revised balancing procedure results in an equivalent balanced realization. The computational methods are essentially identical.* □

In this section, we provide the mathematical and computational frameworks for performing the individual steps and combining them into the overall balancing procedure.

4.5.1 Morse-Palais Form

The Morse-Palais lemma is applied in the Scherpen balancing procedure by observing that the controllability function L_c has a non-degenerate critical point at 0. Therefore, there exists a Morse coordinate transformation under which L_c is quadratic on a neighborhood of 0. Equation 4.88 leads directly to the following result.

Corollary 4.5.2 (Morse Coordinate Transformation for L_c Around 0) *There exist neighborhoods U and $\widehat{U}$ of 0 and a local coordinate transformation*

$$\Phi_M : \widehat{U} \to U\,, \quad \widehat{x} \to x = \Phi_M(\widehat{x})\,, \quad \Phi_M(0) = 0 \tag{4.121}$$

such that

$$\widehat{L}_c(\widehat{x}) \triangleq L_c(\Phi_M(\widehat{x})) = \frac{1}{2}\,\widehat{x}^{\mathsf{T}}\,\widehat{x} \qquad \widehat{x} \in \widehat{U} \tag{4.122}$$

where $U = \Phi_M(\widehat{U})$. □

Example 4.5.3 *Consider the case of a LTI system with controllability Gramian W_c and controllability function $L_c(x) = \frac{1}{2}\,x^{\mathsf{T}}\,{W_c}^{-1}\,x$. Let $W_c = L\,L^{\mathsf{T}}$ be the Cholesky decomposition for the symmetric positive-definite matrix W_c. Then the Morse coordinate transformation for L_c around 0 is given by*

$$x = \Phi_M\,(\widehat{x}) = L\,\widehat{x} \tag{4.123}$$

resulting in $\widehat{L}_c(x) = \frac{1}{2}\,\widehat{x}^{\mathsf{T}}\,\widehat{x}$. □

4.5.2 Input-Normal Form

Recall the input-normal form for a stable minimal linear system realization (A, B, C), i.e., Gramians take the form $W_c = \mathbb{I}$ and $W_o = \Sigma^2 = \mathrm{diag}\,(\sigma_1^2, \ldots, \sigma_n^2)$. Consequently, the corresponding energy functions are given by

$$L_c\,(x) \;=\; \frac{1}{2}\,x^{\mathsf{T}}\,x \tag{4.124}$$

$$L_o\,(x) \;=\; \frac{1}{2}\,x^{\mathsf{T}}\,\Sigma^2\,x \tag{4.125}$$

We seek a local coordinate transformation under which the nonlinear system realization (f, g, h) and corresponding energy functions L_c and L_o take an analogous form in a neighborhood of 0.

Assume that we already have applied the Morse coordinate transformation Φ_M guaranteed by Corollary 4.5.2. Let the energy functions in the new coordinates be denoted $\widehat{L}_c$ and $\widehat{L}_o$. The transformed controllability function $\widehat{L}_c$ is of the desired form (4.124). We need an additional change of coordinates under which $\widehat{L}_o$ takes a form analogous to (4.125) while preserving the form of $\widehat{L}_c$. The following is a direct result of Lemma 4.4.12 applied to $\widehat{L}_o$.

Corollary 4.5.4 *For each $\widehat{x} \in \widehat{U}$ there exists a matrix $M(\widehat{x})$ such that*

$$\widehat{L}_o(\widehat{x}) \stackrel{\triangle}{=} L_o(\Phi_M(\widehat{x})) = \frac{1}{2}\,\widehat{x}^T\,M(\widehat{x})\,\widehat{x}. \tag{4.126}$$

Moreover, $M(\widehat{x})$ is symmetric everywhere on $\widehat{U}$ and

$$M\,(0) = D^2\widehat{L}_o\,(0) \tag{4.127}$$

□

To complete the input-normal analogy we must diagonalize M throughout $\widehat{U}$ while preserving the form of $\widehat{L}_c$. The following result from Kato [77] provides conditions for the existence of such a transformation.

Lemma 4.5.5 (Kato [77]) *If the number of distinct eigenvalues of $M(\widehat{x})$ is constant on a neighborhood $\widehat{U}$ of 0, then the eigenvalues and eigenvectors of $M(\widehat{x})$ are smooth functions of $\widehat{x} \in \widehat{U}$.* □

Henceforth, we assume that $M(\widehat{x})$ always has a constant number of distinct eigenvalues on $\widehat{U}$. Since $M(0)$ is symmetric and positive-definite, it is diagonalizable. Thus, our assumption together with Lemma 4.5.5 implies that $M(\widehat{x})$ is smoothly diagonalizable throughout $\widehat{U}$, i.e., there exist smooth matrix-valued functions T and Λ such that

$$M(\widehat{x}) = T(\widehat{x})\,\Lambda(\widehat{x})\,T(\widehat{x})^{\mathsf{T}} \qquad \widehat{x} \in \widehat{U} \tag{4.128}$$

where $T(\widehat{x})$ is orthogonal for each $\widehat{x} \in \widehat{U}$ and $\Lambda(\widehat{x})$ takes the form

$$\Lambda(\widehat{x}) = \mathrm{diag}\,(\lambda_1(\widehat{x}), \ldots \lambda_n(\widehat{x})) \qquad \widehat{x} \in \widehat{U} \tag{4.129}$$

with $\lambda_1(\widehat{x}) \geq \cdots \geq \lambda_n(\widehat{x}) \geq 0$ by convention.

To construct the input-normal coordinate transformation, we define the change of coordinates

$$\nu : \widehat{U} \to \widetilde{U}\,, \quad \widehat{x} \to \widetilde{x} = \nu(\widehat{x})\,, \quad \nu(0) = 0 \tag{4.130}$$

by

$$\widetilde{x} = \nu(\widehat{x}) \stackrel{\triangle}{=} T(\widehat{x})^{\mathsf{T}}\,\widehat{x} \qquad \widehat{x} \in \widehat{U} \tag{4.131}$$

where $\widetilde{U} = \nu(\widehat{U})$. Observe that ν is linear and orthogonal for each fixed $\widehat{x}$. Composing ν^{-1} with Φ_M yields the input-normal coordinate transformation

$$x = \Phi_I(\widetilde{x}) \stackrel{\triangle}{=} \Phi_M(\nu^{-1}(\widetilde{x})) \qquad \widetilde{x} \in \widetilde{U} \tag{4.132}$$

where $U = \Phi_I(\widetilde{U})$. This is summarized in the following.

Lemma 4.5.6 (Input-Normal Form) *There exist neighborhoods U and $\widetilde{U}$ of 0 and a local change of coordinates*

$$\Phi_I : \widetilde{U} \to U\,, \quad \widetilde{x} \to x = \Phi_I(\widetilde{x})\,, \quad \Phi_I(0) = 0 \tag{4.133}$$

such that

$$\widetilde{L}_c(\widetilde{x}) \stackrel{\triangle}{=} L_c(\Phi_I(\widetilde{x})) \;=\; \frac{1}{2}\,\widetilde{x}^T\,\widetilde{x} \tag{4.134}$$

$$\widetilde{L}_o(\widetilde{x}) \stackrel{\triangle}{=} L_o(\Phi_I(\widetilde{x})) \;=\; \frac{1}{2}\,\widetilde{x}^T\,\widetilde{W}(\widetilde{x})\,\widetilde{x} \tag{4.135}$$

where

$$\widetilde{W}(\widetilde{x}) \;\stackrel{\triangle}{=}\; \mathrm{diag}\,(\mu_1(\widetilde{x}), \ldots, \mu_n(\widetilde{x})) \tag{4.136}$$

$$\mu_i(\widetilde{x}) \;\stackrel{\triangle}{=}\; \lambda_i(\nu^{-1}(\widetilde{x})) \qquad i \in \underline{n} \tag{4.137}$$

Example 4.5.7 *We continue with the LTI system of Example 4.5.3. Let the system have observability Gramian W_o and observability function $L_o(x) = \frac{1}{2}\,x^{\mathsf{T}}\,W_o\,x$. Let $W_o = T\Sigma^2 T^{\mathsf{T}}$ be the spectral decomposition for the symmetric positive-definite matrix W_o. Then the input-normal transformation is given by*

$$\widetilde{x} \;=\; \nu(\widehat{x}) = T^{\mathsf{T}}\,\widehat{x} \tag{4.138}$$

$$x \;=\; \Phi_I\,(\widetilde{x}) = L\,T\,\widetilde{x} \tag{4.139}$$

resulting in $\widetilde{L}_c(\widetilde{x}) = \frac{1}{2}\,\widetilde{x}^{\mathsf{T}}\,\widetilde{x}$ and $\widetilde{L}_o(\widetilde{x}) = \frac{1}{2}\,\widetilde{x}^{\mathsf{T}}\,\Sigma^2\,\widetilde{x}$. □

4.5.3 Balanced Form

Recall the balanced form for a stable minimal linear system realization (A, B, C), i.e., Gramians are such that $\widehat{W}_c = \Sigma = \widehat{W}_o$. Consequently, the corresponding energy functions are given by

$$L_c\,(x) \;=\; \frac{1}{2}\,x^{\mathsf{T}}\,\Sigma^{-1}\,x \tag{4.140}$$

$$\widehat{L}_o(\widehat{x}) \;=\; \frac{1}{2}\,x^{\mathsf{T}}\,\Sigma\,x \tag{4.141}$$

We seek a local coordinate transformation under which the nonlinear system realization (f, g, h) and corresponding energy functions L_c and L_o take an analogous form in a neighborhood of 0.

Assume that we already have computed the objects defined in Lemma 4.5.6. Define the change of coordinates (point-dependent scaling)

$$\eta : \widetilde{U} \to \bar{U}\,, \quad \widetilde{x} \to \bar{x} = \eta(\widetilde{x})\,, \quad \eta(0) = 0 \tag{4.142}$$

by

$$\bar{x} = \eta(\widetilde{x}) \stackrel{\triangle}{=} \Gamma(\widetilde{x})\,\widetilde{x} \qquad \widetilde{x} \in \widetilde{U} \tag{4.143}$$

where $\bar{U} = \eta(\widetilde{U})$ and

$$\Gamma(\widetilde{x}) \stackrel{\triangle}{=} \mathrm{diag}\left(\mu_1(\widetilde{x})^{\frac{1}{4}}, \ldots, \mu_n(\widetilde{x})^{\frac{1}{4}}\right) \tag{4.144}$$

Observe that η is linear and diagonal for each fixed $\widetilde{x}$. Composing η^{-1} with Φ_I yields the the balancing coordinate transformation

$$\Phi_B(\bar{x}) \stackrel{\triangle}{=} \Phi_I(\eta^{-1}(\bar{x})) = \Phi_M(\nu^{-1}(\eta^{-1}(\bar{x}))) \qquad \bar{x} \in \bar{U} \tag{4.145}$$

where $U = \Phi_B(\bar{U})$. This is summarized in the following.

Lemma 4.5.8 (Balanced Form) *There exist neighborhoods U and $\bar{U}$ of 0 and a local change of coordinates*

$$\Phi_B : \bar{U} \to U\,, \quad \bar{x} \to x = \Phi_B(\bar{x})\,, \quad \Phi_B(0) = 0 \tag{4.146}$$

such that

$$\bar{L}_c(\bar{x}) \stackrel{\triangle}{=} L_c(\Phi_B(\bar{x})) \;=\; \frac{1}{2}\,\bar{x}^T\,\bar{W}^{-1}(\bar{x})\,\bar{x} \tag{4.147}$$

$$\bar{L}_o(\bar{x}) \stackrel{\triangle}{=} L_o(\Phi_B(\bar{x})) \;=\; \frac{1}{2}\,\bar{x}^T\,\bar{W}(\bar{x})\,\bar{x} \tag{4.148}$$

where

$$\bar{W}(\bar{x}) \triangleq \operatorname{diag}\left(\sigma_1(\bar{x}), \ldots, \sigma_n(\bar{x})\right) \tag{4.149}$$

$$\sigma_i(\bar{x}) \triangleq \mu_i(\eta^{-1}(\bar{x}))^{\frac{1}{2}} \qquad i \in \underline{n} \tag{4.150}$$

Definition 4.5.9 *The functions* $\{\sigma_1(\cdot), \ldots, \sigma_n(\cdot)\}$ *are called the* singular value functions *of the affine nonlinear system.* □

Remark 4.5.10 *The terminology "singular value functions" was coined by Scherpen in [140, 141], where they were defined, somewhat differently, as*

$$\sigma_i(\bar{x}) = \mu_i\left(0, \ldots, 0, \eta^{-1}(\bar{x}_i), 0, \ldots, 0\right)^{1/2} \tag{4.151}$$

The difference is that the square root of μ_i *is evaluated at points on the* i*-th coordinate axis. This convention facilitates some of the subsequent calculations in [140, 141], but is inconsequential for our purposes here.* □

Remark 4.5.11 *In contrast to the LTI case, the singular value functions are not invariant under coordinate transformation. However, for a LTI system they specialize to the constant Hankel singular values.* □

Example 4.5.12 *We continue with the LTI system of Example 4.5.7. The balancing transformation is given by*

$$\bar{x} = \eta(\widetilde{x}) = \Sigma^{1/2}\,\widetilde{x} \tag{4.152}$$

$$x = \Phi_B(\bar{x}) = L\,T\,\Sigma^{-1/2}\,\bar{x} \tag{4.153}$$

resulting in $\bar{L}_c(\bar{x}) = \frac{1}{2}\,\bar{x}^{\mathsf{T}}\,\Sigma^{-1}\,\bar{x}$ *and* $\bar{L}_o(\bar{x}) = \frac{1}{2}\,\bar{x}^{\mathsf{T}}\,\Sigma\,\bar{x}$. □

The singular value functions $\sigma_1, \ldots, \sigma_n$ and the balancing transformation Φ_B are not unique for a given realization (f, g, h). This can be argued as a consequence

of the non-uniqueness of the Morse coordinate transformation. Thus, there exists a family of transformations Φ_B, each producing a balanced realization $\left(\bar{f}, \bar{g}, \bar{h}\right)$ from among a family of such balanced realizations.

The model reduction properties of nonlinear balancing, as with the LTI case, reside in a ranking of the singular value functions, i.e., the magnitude of $\sigma_i(\bar{x})$ relative to the others is an indication of the degree to which the i-th state component contributes to the input-to-output energy gain of the system. Since, in the nonlinear setting, the σ_i are functions of the state $\bar{x}$, we must be concerned with the neighborhood of 0 in which the functions do not intersect, i.e., switch places in the ranking. Furthermore, since they are not unique, there is the question of whether different collections of σ_i for (f, g, h) will result in different orderings by magnitude. We are not aware of any results addressing these issues.

4.5.4 Computation

In order to implement the nonlinear balancing procedure, we compute discretized approximations of the various functions and local coordinate transformations as described in the previous sections. Figure 4.3 illustrates the computational procedure. The inputs are the smooth functions f, g, and h in realization (f, g, h) and a suitable state-space grid $\mathcal{X}$. The outputs are discretized approximations of the functions $\bar{f}$, $\bar{g}$, and $\bar{h}$ in balanced realization $\left(\bar{f}, \bar{g}, \bar{h}\right)$ and neighborhoods of grid points U and $\bar{U}$ representing the neighborhoods on which the balancing transformation is defined. In this section we present the computational framework and the main algorithms for performing the required computations.

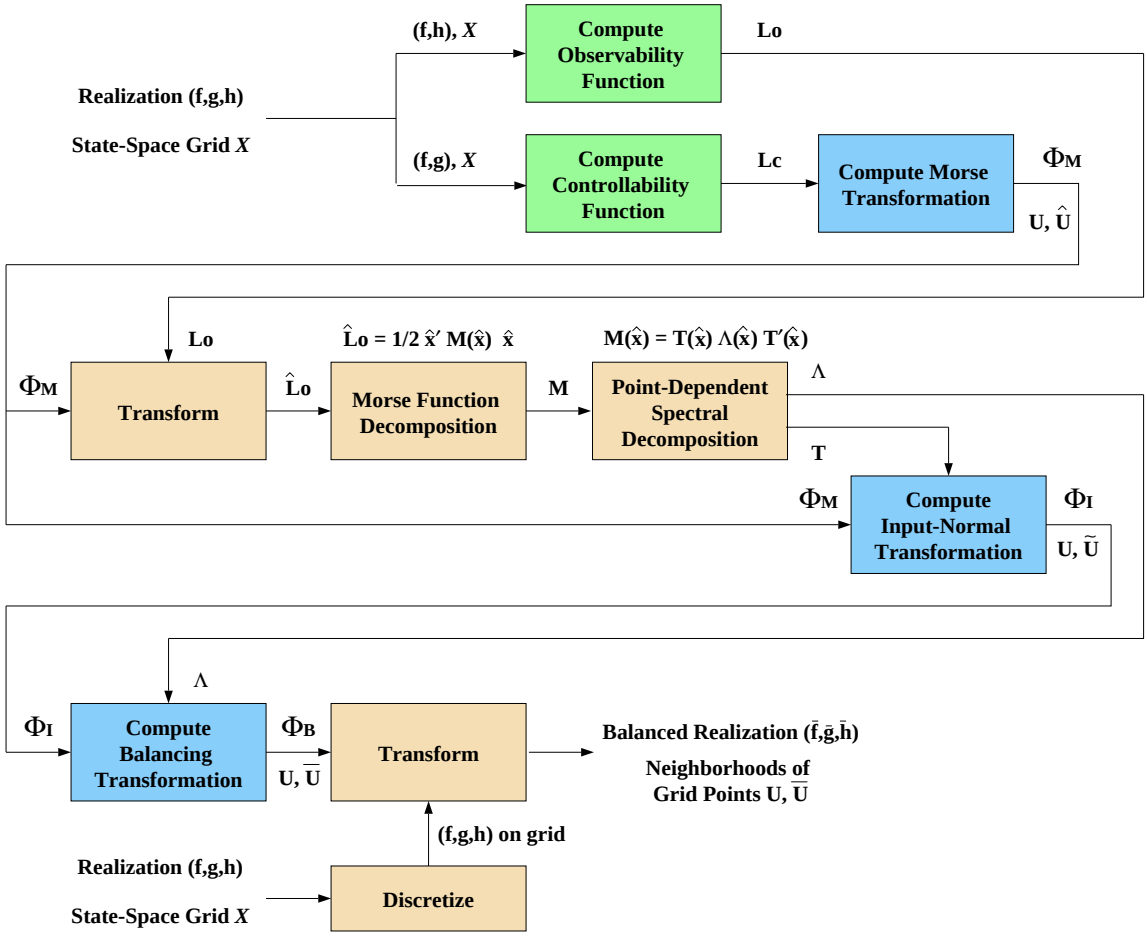


Figure 4.3: Overview of computational procedure for nonlinear balancing.

Computational Setting

In the computational setting, functions are evaluated at a pre-determined set of points on a state-space grid, i.e., they are discretized approximations. A neighborhood of 0 corresponds to a set of discrete grid points containing the point representing the origin. We do not address the problem of determining an appropriate discretization of the state-space. Rather, we assume that a grid, i.e., a collection of points denoted $\mathcal{X}$, of sufficient resolution has been constructed. Suppose that we have discretized the state-space in such a way that there are p evenly spaced grid points along each of the n dimensions. This means that there are p^n total discrete points in the state-space grid.

It is normally the case that the basic computational primitives of an algorithm are elementary operations such as floating point additions, multiplications, and so forth. Such a low-level viewpoint is unsuitable for our purposes here. Instead, we consider the primitives to be

- standard matrix-vector operations such as matrix multiplication, matrix inversion, matrix transposition, and spectral decomposition; and
- standard operations on a function of a real variable such as definite integration, partial differentiation, and multi-dimensional interpolation.

There exist standard algorithms for performing the above operations, each of which has an associated computational complexity (e.g., $o(m^3)$ elementary operations for multiplication of $m \times m$ matrices). However, for the balancing algorithms, the overall computational complexity is dominated by the system dimension as manifested in the grid resolution, i.e., the number of grid points at which computations are performed. For example, we use an algorithm called SPECTRAL-

DECOMP that takes a matrix-valued function, and returns the eigenvalues and eigenvectors (performs a spectral decomposition) at each point in the state-space grid. Thus, if the spectral decomposition algorithm has complexity $o(s(m))$, the algorithm SPECTRAL-DECOMP has complexity $o(s(m)\, p^n)$, i.e., is exponential in n. Point dependency dominates the computational complexity of the nonlinear balancing algorithms.

We adopt a point-wise data structure for storage of the objects of interest (e.g., grid points, controllability function). Although we do not program the balancing procedure using a database per se, it is a useful model for illustration. Consider a database, i.e., a collection of data records, each of which corresponds to a single point in the state-space grid. Each data record contains the value of each of the objects of interest at one particular grid point. Thus, a function is represented by one field of the entire database. The database structure is given in Table 4.1.

It is useful to define the inverses of the coordinate transformations Φ_M, Φ_I, and Φ_B by

$$\Psi_M = {\Phi_M}^{-1} \qquad \Psi_I = {\Phi_I}^{-1} \qquad \Psi_B = {\Phi_B}^{-1} \tag{4.154}$$

on $\Phi_M(U)$, $\Phi_B(U)$, and $\Phi_B(U)$, respectively. The inverse transformations are represented in the data structure, respectively, by the $\widehat{x}$, $\widetilde{x}$, and $\bar{x}$ data elements corresponding to grid point x.

Balanced Realization

The algorithm for computing a balanced realization is as follows.

Algorithm 4.5.13 (Balanced Realization)

BALANCE$(f, g, h, \mathcal{X})$
1 $Y \leftarrow$ SDE-SIM$(f, g, \mathcal{X})$
2 $L_c \leftarrow$ CTRB-FN-MONTE-CARLO$(Y, \mathcal{X})$

Data Record Structure

x	L_c	L_o	$\widehat{x}$	U	$\widehat{U}$	M	T	λ_1	$\ldots$	λ_n	$\widetilde{x}$	$\widetilde{U}$	Γ	$\bar{x}$	$\bar{U}$	σ_1	$\ldots$	σ_n

Data Fields

Data Element	Description	Type
x	Grid Point in Standard Coordinates	n-vector
L_c	Value of Controllability Function	scalar
L_o	Value of Observability Function	scalar
$\widehat{x}$	Grid Point in Morse Coordinates for L_c	n-vector
U	Neighborhood Membership Indicator	boolean
$\widehat{U}$	Neighborhood Membership Indicator	boolean
M	Intermediate Matrix for L_o	$n \times n$ matrix
T	Eigenvector Matrix for M	$n \times n$ matrix
λ_1	Largest Eigenvalue for M	scalar
$\ldots$	$\ldots$	$\ldots$
λ_n	Smallest Eigenvalue for M	scalar
$\widetilde{x}$	Grid Point in Input-Normal Coordinates	n-vector
$\widetilde{U}$	Neighborhood Membership Indicator	boolean
Γ	Scaling Matrix	$n \times n$ matrix
$\bar{x}$	Grid Point in Balanced Coordinates	n-vector
$\bar{U}$	Neighborhood Membership Indicator	boolean
σ_1	Largest Singular Value	scalar
$\ldots$	$\ldots$	$\ldots$
σ_n	Smallest Singular Value	scalar

Table 4.1: Database structure containing data elements for nonlinear balancing computational procedure. For a state-space grid with p^n points, there are p^n data records, each corresponding to a grid point.

3 $L_o \leftarrow$ OBSV-FN$(f, h, \mathcal{X})$
4 $\left(\phi, A, U, \widehat{U}\right) \leftarrow$ MORSE-PALAIS$(L_c, \mathcal{X})$
5 $\Psi_M \leftarrow$ SPHERICAL(ϕ, A, U)
6 $\widehat{L}_o \leftarrow$ TRANSFORM(L_o, Ψ_M, U)
7 $M \leftarrow$ MORSE-FN-DECOMP$\left(\widehat{L}_o, \widehat{U}\right)$
8 $(T, \Lambda) \leftarrow$ SPECTRAL-DECOMP$\left(M, \widehat{U}\right)$
9 $\left(\Psi_I, \widetilde{U}\right) \leftarrow$ INPUT-NORMAL-TRANS(T, Ψ_M, U)
10 $\left(\Psi_B, \bar{U}\right) \leftarrow$ BALANCING-TRANS(Λ, Ψ_I, U)
11 $\left(\bar{f}, \bar{g}, \bar{h}\right) \leftarrow$ TRANSFORM(f, g, h, Ψ_B, U)
12 RETURN $\bar{f}, \bar{g}, \bar{h}, \bar{U}$ □

We now present the computational methods and algorithms corresponding to the individual steps of Algorithm 4.5.13. Some of these have been addressed previously, although not within the framework of our computational setting.

Controllability Function

Methods for computing the controllability function have been described in Sections 4.2 and 4.3. In cases where we can derive an exact expression for L_c, e.g., via Theorem 4.3.14, the task is completed by discretizing the resulting function L_c. Otherwise, we use the Monte-Carlo approach, yielding the approximation (4.39), which improves as the number of experiments increases. We have developed versions of SDE-SIM and CTRB-FN-MONTE-CARLO, respectively, for numerical integration of SDEs and computation of the controllability function from the Monte-Carlo data. Numerical schemes for simulation of SDEs appear in Appendix C. We developed and used various multidimensional histogramming utilities to implement the Monte-Carlo approach.

Observability Function

We have not addressed computation of the observability function in detail. We implement OBSV-FN to compute the observability function via numerical integration of the natural response of the system and numerical integration of the output energy. The procedure is performed for each point on the state-space grid.

Algorithm 4.5.14 (Observability Function)

OBSV-FN$(f, h, \mathcal{X})$
1 **for** each point x in the state-space grid $\mathcal{X}$
2 $x_0 \leftarrow x$
3 $u \leftarrow 0$
4 $z \leftarrow$ ODE-SIM(f, u, x_0)
5 $y \leftarrow$ COMPOSE(h, z)
6 $T \leftarrow$ a number large enough so that the natural response of the stable system is nearly zero for $t > T$
7 $L_o[x] \leftarrow$ INTEGRAL$(y, 0, T)$
8 RETURN L_o □

Morse Coordinate Transformation

The algorithms for MORSE-PALAIS and MATRIX-SQUARE-ROOT are based on the computational procedures presented in Section 4.4.3.

Algorithm 4.5.15 (Morse Coordinate Transformation)

MORSE-PALAIS$(f, \mathcal{X})$
1 $H \leftarrow$ MORSE-FN-DECOMP$(f, \mathcal{X})$
2 $A \leftarrow$ ORIGIN-SELECT$(H, \mathcal{X})$
3 **for** each point x in the state-space grid $\mathcal{X}$
4 $B \leftarrow A \setminus H[x]$
5 **if** MATRIX-NORM$(\mathbb{1} - B) < 1$
6 $C \leftarrow$ MATRIX-SQUARE-ROOT(B)
7 $z \leftarrow C * x$
8 $\phi[x] \leftarrow z$
9 add x to the list of points in neighborhood U
10 add z to the list of points in neighborhood $\widehat{U}$
11 RETURN $\phi, A, U, \widehat{U}$ □

Remark 4.5.16 *The expression $B \leftarrow A \setminus H[x]$ is equivalent to solving $H(x) = A\,B$ for B.* □

The algorithm for MATRIX-SQUARE-ROOT is as follows.

Algorithm 4.5.17 (Matrix Square Root)

```
MATRIX-SQUARE-ROOT(B)
1   if MATRIX-NORM(𝟙 − B) < 1
2   C ← C_prev ← 𝟙
3   δ ← ε + 1
4   while δ ≥ ε
5       C ← C + 0.5 ∗ (B − C²)
6       δ ← MATRIX-NORM(C − C_prev)
7       C_prev ← C
8   else
9       error: algorithm will not converge
10  RETURN C
```
□

Remark 4.5.18 *The parameter ϵ represents an error tolerance. It would be set as a global constant or in some other appropriate manner.* □

Algorithm 4.5.15 returns discretized versions of the objects that appear in Theorem 4.4.7, i.e., the coordinate transformation ϕ, the invertible symmetric matrix A, and the neighborhoods U and $\widehat{U}$ of 0. Computation of Ψ_M from ϕ and A requires several additional straightforward steps including a standard spectral decomposition of A.

Algorithm 4.5.19 (Spherical Quadratic Form)

```
SPHERICAL(φ, A, U)
1   (V, Λ) ← EIG(A)
2   E ← ABS(Λ)
3   R ← E ∗ TRANSPOSE(V)
4   r ← the number of negative entries on the diagonal of Λ
5   for each point x in the collection of grid points U
6       ψ[x] ← R ∗ φ[x]
7   RETURN ψ, r
```
□

Remark 4.5.20 *When dealing with positive functions such as L_c and L_o, the index r is always zero so we ignore that parameter in the balancing procedure.* □

Function Decompositions

Algorithms 4.5.13 and 4.5.15 use the following algorithms for approximating the decompositions (4.89) and (4.93). Also, the decomposition $\widehat{L}_o = \widehat{x}^{\mathsf{T}} M(\widehat{x})\, \widehat{x}$ appears in the overall balancing procedure.

Algorithm 4.5.21 (Smooth Function Decomposition)

```
SMOOTH-FN-DECOMP(f, U)
1  for i = 1 to n
2     partialf[i] ← PARTIAL-DERIV(f, i, U)
3  for each point x in the collection of grid points U
4     for i = 1 to n
5        g[i][x] ← INTEGRAL(partialf[i], 0, x)
6  G ← VECTOR(g[1], ..., g[n])
7  RETURN G
```
□

Algorithm 4.5.22 (Morse Function Decomposition)

```
MORSE-FN-DECOMP(f, U)
1  G ← SMOOTH-FN-DECOMP(f, U)
2  for i = 1 to n
3     F[i] ← SMOOTH-FN-DECOMP(G[i], U)
4     for j = 1 to n
5        h[i, j] ← F[i][j]
6  H ← MATRIX(h[1, 1], ..., h[n, n])
7  RETURN H
```
□

Input-Normal and Balancing Coordinate Transformations

The algorithms for computing discretized approximations of the input-normal and balancing transformations are based on the computational procedures described in Sections 4.5.2 and 4.5.3.

Algorithm 4.5.23 (Input-Normal Coordinate Transformation)

INPUT-NORMAL-TRANS(T, ξ, U)
1 **for** each point x in the collection of grid points U
2 $\quad y \leftarrow \xi[x]$
3 $\quad z \leftarrow$ TRANSPOSE$(T[y]) * y$
4 $\quad \psi[x] \leftarrow z$
5 $\quad$ add z to the list of points in neighborhood $\widetilde{U}$
6 RETURN $\psi, \widetilde{U}$ □

Algorithm 4.5.24 (Balancing Coordinate Transformation)

BALANCING-TRANS(Λ, ξ, U)
1 **for** each point x in the collection of grid points U
2 $\quad y \leftarrow \xi[x]$
3 $\quad \Sigma \leftarrow$ SQUARE-ROOT(Λ, U)
4 $\quad \Gamma \leftarrow$ SQUARE-ROOT(Σ, U)
5 $\quad z \leftarrow \Gamma[y] * y$
6 $\quad \psi[x] \leftarrow z$
7 $\quad$ add z to the list of points in neighborhood $\bar{U}$
8 RETURN $\psi, \bar{U}$
□

MATLAB Toolbox

We have implemented the algorithms described in this section using the MATLAB [102] programming environment. The resulting collection of programs and utilities for performing various operations on multidimensional discretized functions, referred to as the *nonlinear balancing toolbox*, was used as the computational tool to apply nonlinear balancing to the examples in Section 4.6.

We performed simulations on a Sun Sparc Ultra-10 running the UNIX operating system. Running times for the various programs depend on grid resolution and system dimension. Roughly, the time required to compute a Morse transformation for systems of dimension 2, 3, and 4 is on the order of, respectively, seconds to minutes, minutes to hours, and hours to days. Computations for systems of

dimension 5 and higher are currently infeasible.

It is possible to increase the speed of computation by converting the MATLAB code to C, using a faster processor, and taking advantage of opportunities for parallelization and other economies. However, we did not pursue these options, since it is unlikely that the feasible dimension would increase significantly. Rather, we believe that new algorithms will be required for working with higher dimensional systems.

Utilities

The algorithms presented in this section use various utilities for performing computations with multidimensional discretized functions, standard vector-matrix operations, operations on a function of a real variable, and so forth. We briefly describe their purposes here so that the previous algorithms can be understood. The actual implementations in the nonlinear balancing toolbox do not necessarily reflect exactly the descriptions given below, nor is this list a complete compilation of toolbox utilities (e.g., TRANSFORM and INTEGRAL require the use of additional multidimensional interpolation utilities). Versions of some of these utilities are included as standard functionality in MATLAB. By a point-dependent matrix we mean a matrix-valued function on a grid.

SPECTRAL-DECOMP returns the point-dependent eigenvectors and eigenvalues of a point-dependent matrix

SQUARE-ROOT returns a point-dependent matrix whose entries are the square roots of the respective entries of a point-dependent matrix.

TRANSFORM takes one or more discretized mappings and returns its (their) values at the grid points after a coordinate transformation.

INTEGRAL returns the approximate definite integral of a discretized function.

ODE-SIM returns the sampled time evolution of a forced ODE given initial conditions and sampled input signal.

PARTIAL-DERIV returns an approximation to the i-th partial derivative of a discretized function.

COMPOSE returns a discretized function that represents the composition of two other discretized functions.

ORIGIN-SELECT returns the indices of the grid point that represents the origin in state-space.

EIG returns the eigenvalues and unit eigenvectors of an invertible matrix.

MATRIX-NORM returns the largest singular value of a matrix.

ABS returns a matrix whose entries are the absolute values of the entries of the original matrix.

TRANSPOSE returns the transpose of a matrix.

VECTOR assembles a collection of numbers or discretized functions into a vector.

MATRIX assembles a collection of numbers or discretized functions into a matrix.

4.6 Applications

In this section we illustrate the methods and algorithms presented in this Chapter by applying them to two examples of rigid link mechanical systems. We compute a balanced realization for a forced damped pendulum system, and take steps toward

balancing a forced damped double pendulum system. The material in this section relies heavily on the mathematical framework for mechanical systems as presented in Section 2.7.

4.6.1 A Balanced Realization for the Forced Damped Pendulum

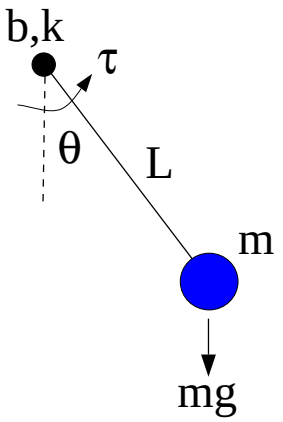
The first example that we consider is a simple pendulum system as illustrated in Figure 4.4. The system incorporates linear torsional damping, linear torsional stiffness, and a torque input at the rotary joint. We assume that the shaft is massless and that the pendulum moves only in the plane. We consider two cases for the system output: where the joint angle is measured (position read-out) and where the joint angular velocity is measured (velocity read-out). Figure 4.4 also provides values for each of the physical parameters that we use in numerical studies. Simulations were performed using routines from the nonlinear balancing toolbox described in Section 4.5.4.

It is beneficial to study the pendulum as an example because

- it is nearly linear, so we can use the LTI theory to obtain a good estimate of the correct results for comparison; and
- in previous sections we have studied second-order mechanical systems and obtained an exact formula for the controllability function.

State-Space Realization

We obtain a state-space realization (f, g, h) for the pendulum system via the Euler-Lagrangian mechanics outlined in Section 2.7. Let the generalized position q and



θ		joint angle (between shaft and vertical)
τ		torque applied at rotary joint
m	1/40	mass attached to end of shaft
b	2	torsional damping coefficient (friction)
k	1	torsional stiffness coefficient (spring constant)
L	20	length of shaft
G	10	gravitational acceleration

Figure 4.4: Planar pendulum system with massless shaft, linear torsional damping, linear torsional stiffness, and torque input applied at the rotary joint. Values of parameters are provided for the numerical studies that we conducted.

velocity $\dot{q}$ correspond to the joint angle θ and angular velocity $\dot{\theta}$, respectively. Let the generalized force F represent the applied joint torque τ. The kinetic, potential, and dissipation energies are given, respectively, by

$$K(q,\dot{q}) = \frac{1}{2} m L^2 \dot{q}^2 \tag{4.155}$$

$$U(q,\dot{q}) = \frac{1}{2} k q^2 - m G L \cos(q) \tag{4.156}$$

$$R(q,\dot{q}) = \frac{1}{2} b \dot{q}^2 \tag{4.157}$$

The Lagrangian L is given by

$$\begin{aligned} L(q,\dot{q}) &= K(q,\dot{q}) - U(q,\dot{q}) \\ &= \frac{1}{2} m L^2 \dot{q}^2 - \frac{1}{2} k q^2 + m G L \cos(q) \end{aligned} \tag{4.158}$$

We apply the Euler-Lagrange equation of motion (2.113), i.e.,

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{q}} - \frac{\partial L}{\partial q} = F - \frac{\partial R}{\partial \dot{q}} \tag{4.159}$$

to obtain the equation of motion for the pendulum system, given by

$$m\,L^2\,\ddot{q} + k\,q + m\,G\,L\sin(q) = F - b\,\dot{q} \tag{4.160}$$

The affine nonlinear control system is realized in coordinates $x = (x_1, x_2) = (q, \dot{q})$ by

$$f(x) = \begin{bmatrix} x_2 \\ \\ -\dfrac{G}{L}\sin(x_1) - \dfrac{k}{m\,L^2}\,x_1 - \dfrac{b}{m\,L^2}\,x_2 \end{bmatrix} \qquad g(x) = \begin{bmatrix} 0 \\ \\ \dfrac{1}{m\,L^2} \end{bmatrix} \tag{4.161}$$

and either $h(x) = x_1$ or $h(x) = x_2$ depending on whether we measure angular position or velocity.

Remark 4.6.1 *We need not explicitly realize the system in Hamiltonian coordinates (generalized positions and momenta). The results in Section 4.3.2 still apply, taking into account the different coordinates.* □

System Properties

The linearization (A, B, C) of (f, g, h) is given by

$$A = \begin{bmatrix} 0 & 1 \\ \\ -\dfrac{G}{L} - \dfrac{k}{m\,L^2} & -\dfrac{b}{m\,L^2} \end{bmatrix} \qquad B = \begin{bmatrix} 0 \\ \\ \dfrac{1}{m\,L^2} \end{bmatrix} \tag{4.162}$$

and either $C = [1 \quad 0]$ for position read-out or $C = [0 \quad 1]$ for velocity read-out.

Observe that the linearization is asymptotically stable since

$$\operatorname{spec}(A) = \left\{ \frac{1}{2\,m\,L^2}\left(-b \pm \sqrt{b^2 - 4\,m^2\,L^3\,G - 4\,m\,L^2\,k}\right) \right\} \in \mathbb{C}^- \tag{4.163}$$

Furthermore, the linearization is controllable and observable in both output cases since

$$\text{rank } [B \quad AB] = \text{rank} \begin{bmatrix} 0 & \dfrac{1}{m\,L^2} \\ \dfrac{1}{m\,L^2} & \dfrac{-b}{m^2\,L^4} \end{bmatrix} = 2 \tag{4.164}$$

and either

$$\text{rank} \begin{bmatrix} C \\ CA \end{bmatrix} = \text{rank} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = 2 \tag{4.165}$$

or

$$\text{rank} \begin{bmatrix} C \\ CA \end{bmatrix} = \text{rank} \begin{bmatrix} 0 & 1 \\ -\dfrac{G}{L} - \dfrac{k}{m\,L^2} & \dfrac{-b}{m\,L^2} \end{bmatrix} = 2 \tag{4.166}$$

Controllability and observability of the linearization, together with asymptotic stability, are sufficient to guarantee local asymptotic reachability and zero-state observability of the nonlinear system (see, e.g., [121]). Therefore, we can conclude that the controllability and observability functions exist in a neighborhood of 0 and are non-degenerate.

Substituting the values of the parameters given in Figure 4.4 gives the linearized realization

$$A = \begin{bmatrix} 0 & 1 \\ -0.6 & -0.2 \end{bmatrix} \qquad B = \begin{bmatrix} 0 \\ 0.1 \end{bmatrix} \tag{4.167}$$

and either $C = [1 \quad 0]$ or $C = [0 \quad 1]$. The Gramians and Hankel singular values are given, for the position and velocity output cases, respectively, by

$$W_c = \begin{bmatrix} 0.0417 & 0.0000 \\ 0.0000 & 0.0250 \end{bmatrix} \qquad W_o = \begin{bmatrix} 1.5000 & 0.0000 \\ 0.0000 & 2.5000 \end{bmatrix} \tag{4.168}$$

$\sigma_1 = 0.3671$ and $\sigma_2 = 0.2838$; and

$$W_c = \begin{bmatrix} 0.0417 & 0.0000 \\ 0.0000 & 0.0250 \end{bmatrix} \qquad W_o = \begin{bmatrix} 0.3671 & 0.0000 \\ 0.0000 & 0.2838 \end{bmatrix} \tag{4.169}$$

$\sigma_1 = 0.2500$ and $\sigma_2 = 0.2500$.

Controllability Function

The Hamiltonian H for the pendulum system is given by

$$\begin{aligned} H(x) &= K(x) + U(x) \\ &= \frac{1}{2} m L^2 x_2^2 + \frac{1}{2} k x_1^2 - m G L \cos(x_1) \end{aligned} \tag{4.170}$$

Furthermore, the equipartition of energy condition (4.65) is satisfied trivially for the 1-DOF system with ratio $\ell = b$. Applying Theorem 4.3.14, the controllability function L_c is given, exactly, by

$$L_c(x) = -2 b m G L \cos(x_1) + b k x_1^2 + b m L^2 x_2^2 + 2 b m G L \tag{4.171}$$

and with substituted values by

$$L_c(x) = -20 \cos(x_1) + 2 x_1^2 + 20 x_2^2 + 20 \tag{4.172}$$

Since we have an exact formula for L_c, we can study the performance of the Monte-Carlo approach by comparing L_c with an approximation computed via Equation (4.39) using an approximate stationary density. To this end, we simulated 50,000 sample paths for the pendulum system with approximate Gaussian white noise injected as the torque input. We assumed that steady-state was reached after 60 time units, 6 times the largest time constant of the system.

The results of the Monte-Carlo experiments are presented in Figure 4.5. We generated histograms for two grid resolutions: $\Delta x = 0.1$ (coarse) and $\Delta x = 0.05$

(fine). The accuracy of computations such as approximate differentiation and interpolation improves as grid resolution increases, i.e., becomes finer. However, it is crucial to obtain an approximation of L_c that is reasonably smooth and consequently has no local minima or maxima other than at 0. Smoothness of the approximation improves as the grid resolution becomes coarser. In this case, we use the coarse grid, which is roughly the highest resolution that provides a smooth approximation. By generating additional sample paths, we can increase the grid resolution while maintaining a smooth approximation.

We investigate the performance of the Monte-Carlo approach by comparing the approximate L_c with the exact L_c given by (4.171). Moreover, we check to see if L_c and the approximation satisfy the HJB equation (4.14). This is done by computing and plotting the *HJB residual*, i.e., the right hand side of (4.14), given by

$$\rho_c(x) = \frac{\partial L_c}{\partial x}(x)\ f(x) + \frac{1}{2}\frac{\partial L_c}{\partial x}(x)\ g(x)\ g^{\mathsf{T}}(x)\ \left[\frac{\partial L_c}{\partial x}(x)\right]^{\mathsf{T}} \tag{4.173}$$

The results are shown, for low and high resolution grids, respectively, in Figures 4.6 and 4.7. The large fluctuations in the residuals at the edges of the grid are due to numerical errors in computing derivatives at the edges and should be ignored. The residual is exactly zero everywhere on the grid for the exact controllability function, thus confirming that it exactly satisfies the HJB equation. The residual for the approximate controllability function fluctuates somewhat (more on the finer grid) but remains relatively close to zero at all grid points. The approximation is better at points close to the origin, since a region close to the origin contains most of the Monte-Carlo data. The performance of the Monte-Carlo approach and its numerical implementation in the nonlinear balancing toolbox appear to be good for this example.

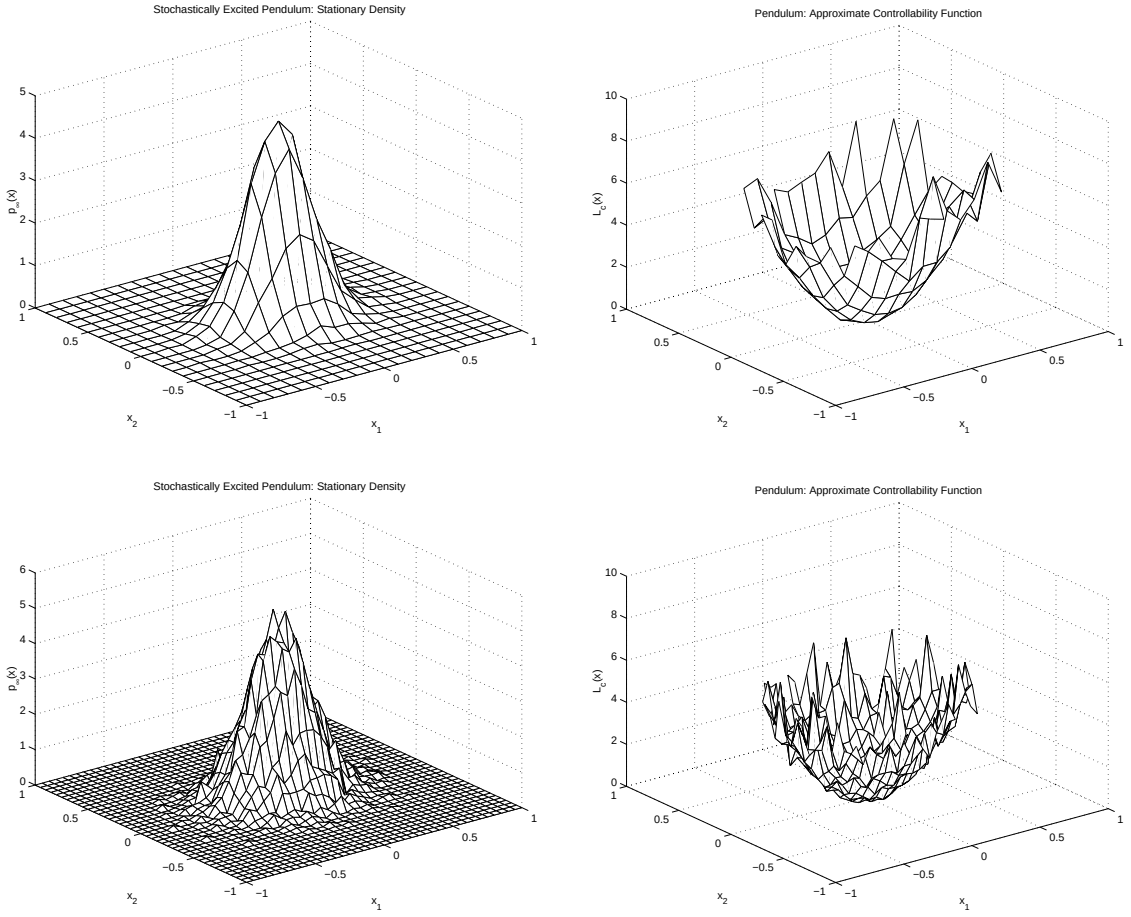


Figure 4.5: The stationary density and derived approximate controllability function for the pendulum system. Monte-Carlo approach used 50,000 sample paths for white noise driven system. Top left: approximate stationary density (coarse grid); Top right: approximate controllability function (coarse grid); Bottom left: approximate stationary density (fine grid); Bottom right: approximate controllability function (fine grid).

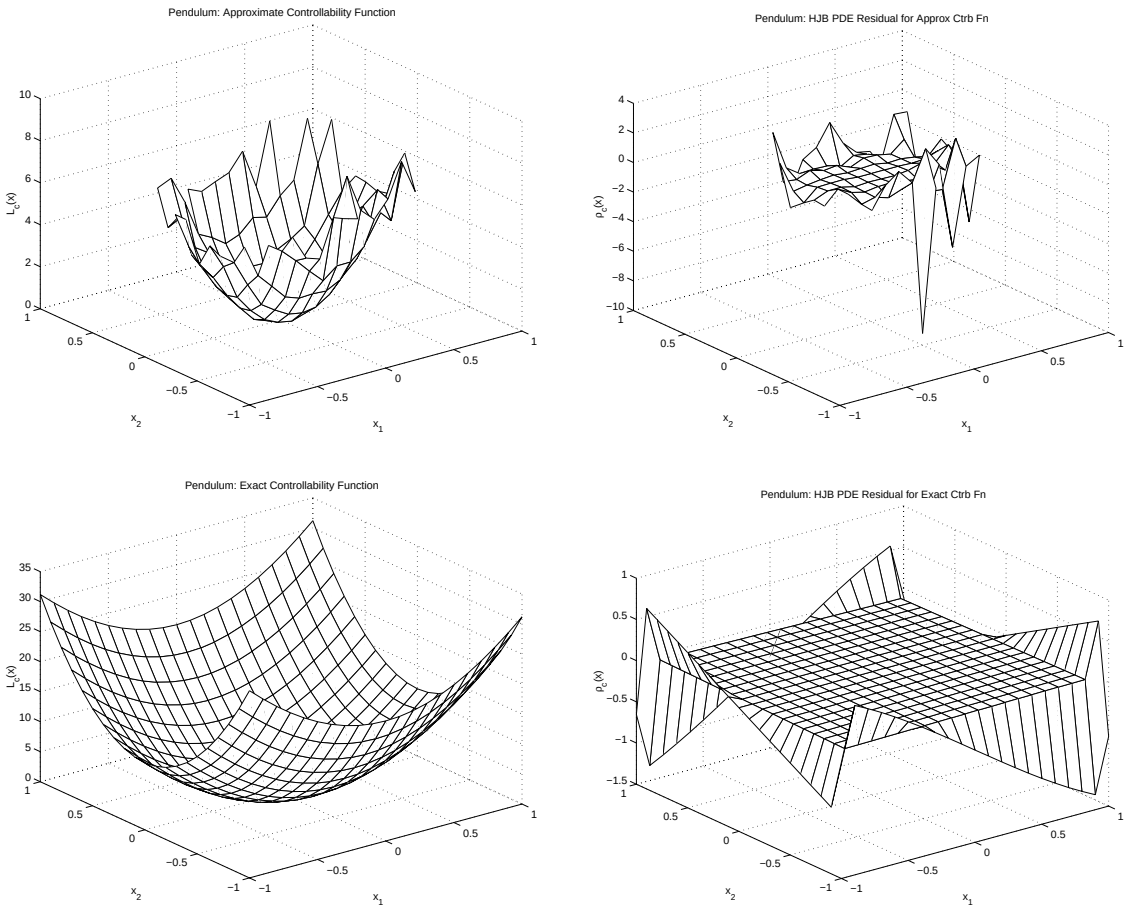


Figure 4.6: The controllability function and HJB residual for the pendulum system (low resolution grid). Top: Approximate controllability function (Monte-Carlo) and HJB residual; Bottom: Exact controllability function and HJB residual.

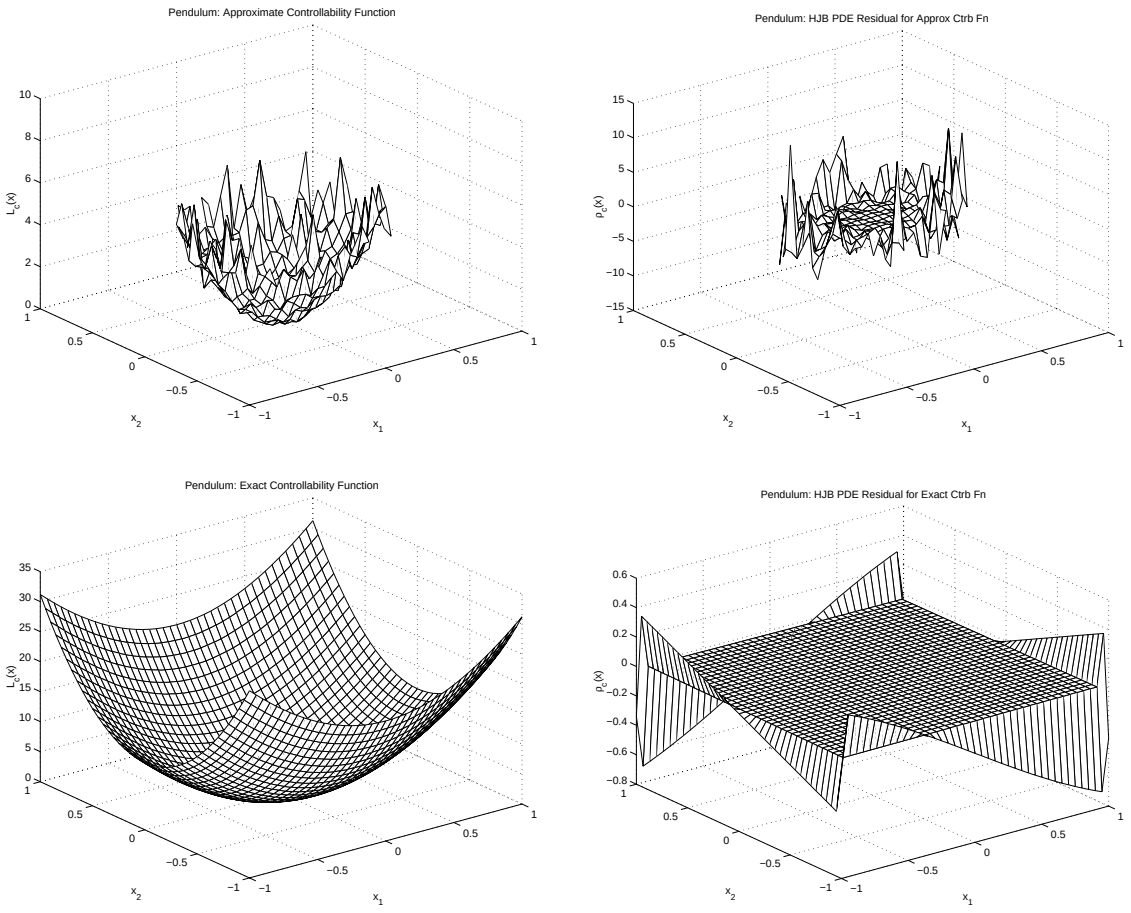


Figure 4.7: The controllability function and HJB residual for the pendulum system (high resolution grid). Top: Approximate controllability function (Monte-Carlo) and HJB residual; Bottom: Exact controllability function and HJB residual.

Observability Function

In the case where we measure angular velocity, i.e., $h(x) = x_2$, we can easily solve the Lyapunov-type PDE (4.15) satisfied by L_o to give an exact formula for the observability function. We obtain an expression for L_o, given by

$$L_o(x) = -\frac{m\,G\,L}{2\,b}\cos(x_1) + \frac{k}{4\,b}\,x_1^2 + \frac{m\,L^2}{4\,b}\,x_2^2 + \frac{m\,G\,L}{2\,b} \tag{4.174}$$

and with substituted values by

$$L_o(x) = -1.25\cos(x_1) + 0.125\,x_1^2 + 1.25\,x_2^2 + 1.25 \tag{4.175}$$

In the case where we measure the joint angle, i.e., $h(x) = x_1$, we cannot obtain a closed form solution of (4.15). Instead, we use Algorithm 4.5.14 to compute an approximation. In addition, to study the performance of the algorithm, we also use Algorithm 4.5.14 to compute an approximation in the case where $h(x) = x_2$. Again, for purposes of simulation, we assumed that steady-state was reached after 60 time units, 6 times the largest time constant of the system.

We investigate the performance of the algorithm by comparing the approximate L_o with the exact L_o (velocity output case) given by (4.174). Moreover, we compute and plot the *Lyapunov residual*, i.e., the right hand side of (4.15), given by

$$\rho_o(x) = \frac{\partial L_o}{\partial x}(x)\; f(x) + \frac{1}{2}\,h^{\mathsf{T}}(x)\; h(x) \tag{4.176}$$

The results are shown for the cases of velocity and position read-out, respectively, in Figures 4.8 and 4.9. As before, the large fluctuations in the residuals at the edges of the grid are due to numerical errors in computing derivatives at the edges and should be ignored. All residuals are zero or nearly zero at all grid points. Moreover, there is negligible difference between the exact observability

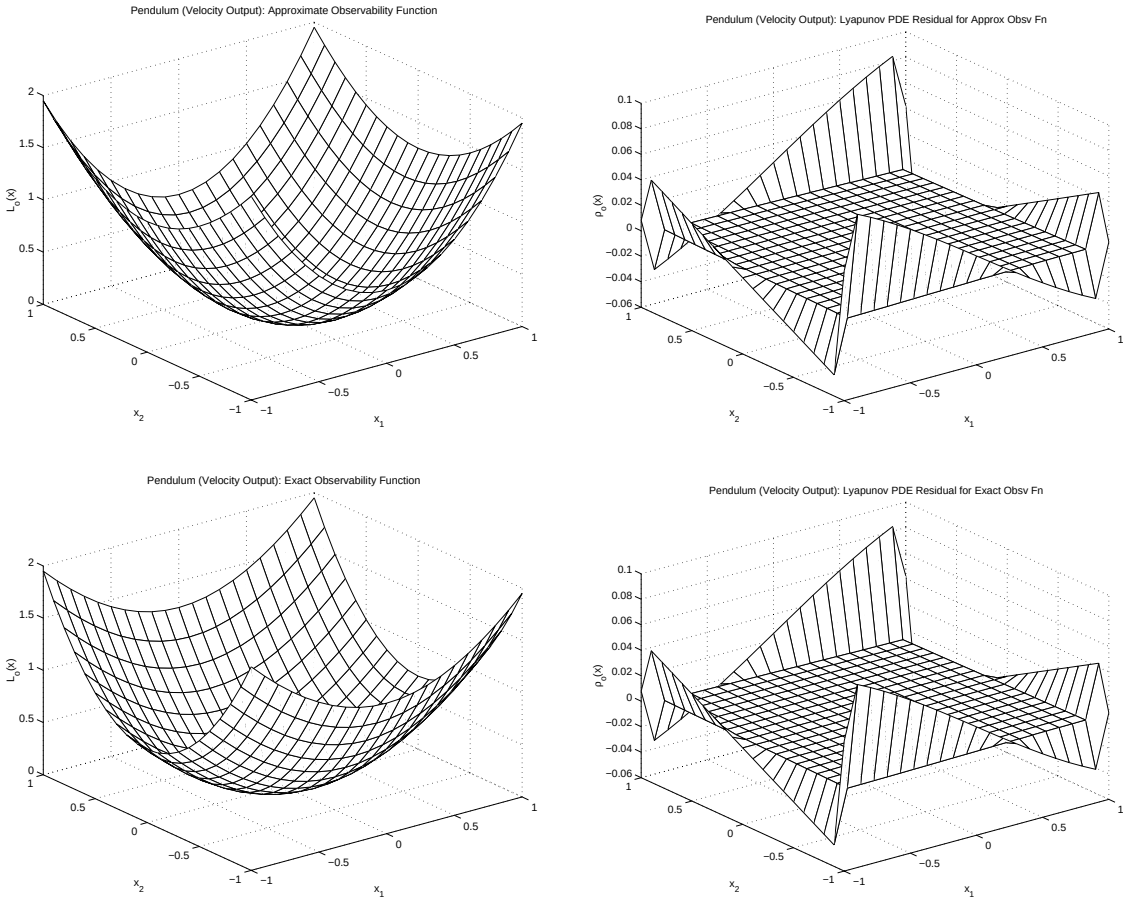


Figure 4.8: The observability function and Lyapunov residual for the pendulum system with velocity output. Top: Approximate observability function; Bottom: Exact observability function.

function computed via (4.174) and the approximate observability function computed using Algorithm 4.5.14. The performance of the algorithm and its numerical implementation in the nonlinear balancing toolbox appear to be good for this example.

Balanced Realization

We now use the previously computed L_c and L_o, and the algorithms presented in Section 4.5.4 and implemented in the nonlinear balancing toolbox, to compute

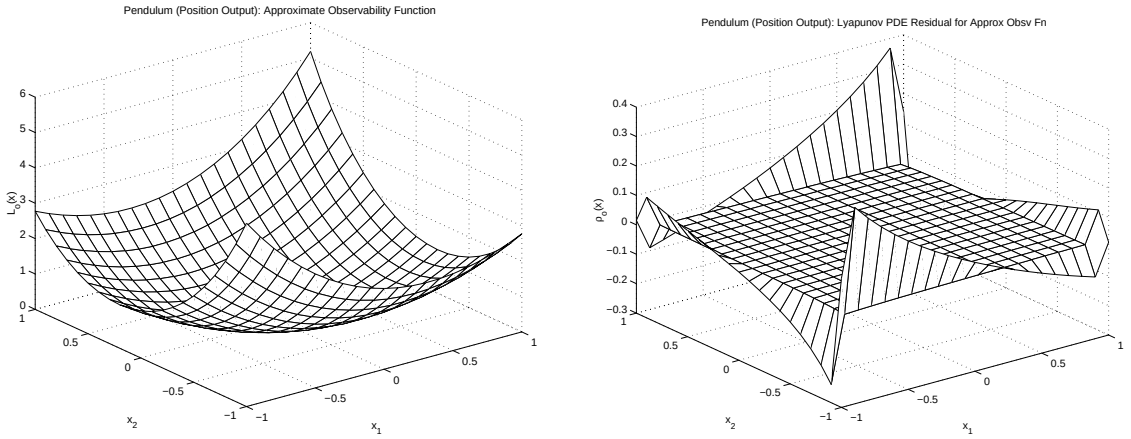


Figure 4.9: The observability function and Lyapunov residual for the pendulum system with position output.

balanced realizations for both pendulum systems, i.e., with position and velocity read-outs.

As a minor digression, we note that it is possible to calculate an expression to approximate the inverse Morse coordinate transformation of

$$L_c(x) = -20\cos(x_1) + 2\,x_1^2 + 20\,x_2^2 + 20 \tag{4.177}$$

around its non-degenerate critical point at 0. Expanding the cosine function yields

$$L_c(x) = 12\,x_1^2 - \frac{10}{6}\,x_1^4 + o(x_1^6) + 20\,x_2^2 = x^{\mathsf{T}}\,H(x)\,x \tag{4.178}$$

where

$$H(x) = \begin{bmatrix} 12 - \dfrac{10}{6}\,x_1^2 + o(x_1^4) & 0 \\ 0 & 20 \end{bmatrix} \tag{4.179}$$

Thus, the inverse Morse coordinate transformation is given, for $|(10/72)\,x_1^2 + o(x_1^4)| < 1$, by

$$\Psi_M(x) = \begin{bmatrix} \sqrt{12}\,\left(1 - \dfrac{10}{72}\,x_1^2 + o(x_1^4)\right)^{1/2}\,x_1 \\ \sqrt{20}\,x_2 \end{bmatrix} \tag{4.180}$$

Although we have calculated an expression to approximate the inverse Morse coordinate transformation, and an expression for its region of validity, we use Algorithm 4.5.15 within the overall balancing procedure to compute a balanced realization. Applying the remainder of the steps in Algorithm 4.5.13 produces discretized approximations of the singular value functions $\sigma_1(x)$ and $\sigma_2(x)$ and the functions $\bar{f}$, $\bar{g}$, and $\bar{h}$ in the balanced realization $\left(\bar{f}, \bar{g}, \bar{h}\right)$.

The computed singular value functions for the pendulum systems with position and velocity outputs, respectively, are shown in Figures 4.10 and 4.11. Because the pendulum system is nearly linear, we expect the singular value functions to be nearly constant at the value of the corresponding Hankel singular values of the linearized system. This is reflected in the computations.

In the case of measured joint angle, the singular value functions are nearly constant at grid points close to the origin, taking values close to 0.367 and 0.284. This closely matches the Hankel singular values of the linearized system. One state is roughly 1.3 times as important to the input-to-output behavior of the system.

In the case of measured angular velocity, the singular value functions are again nearly constant at grid points close to the origin. Here, the two functions are nearly equal, taking values close to 0.252 and 0.248. This closely matches the Hankel singular values of the linearized system. Both states are equally important to the input-to-output behavior of the system. This is expected from a physical standpoint, since there is a duality between the torque input and angular velocity output.

Remark 4.6.2 *Because this example system has only two states, we do not delete any states in order to produce a reduced model. Such a reduction would be dubious, since a system with only one state is qualitatively different from a two state model,*

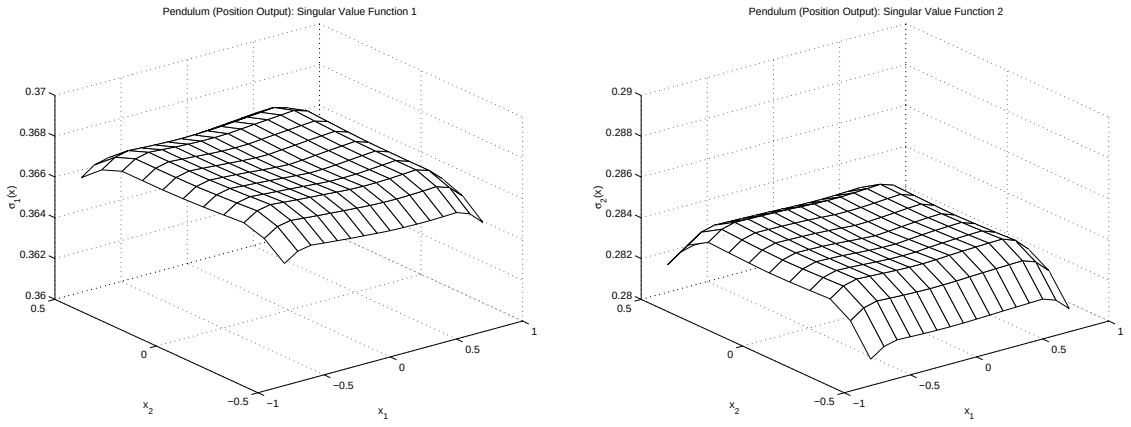


Figure 4.10: The singular value functions for the pendulum system with position output. Left: $\sigma_1(x)$ nearly constant 0.367; Right: $\sigma_2(x)$ nearly constant 0.284.

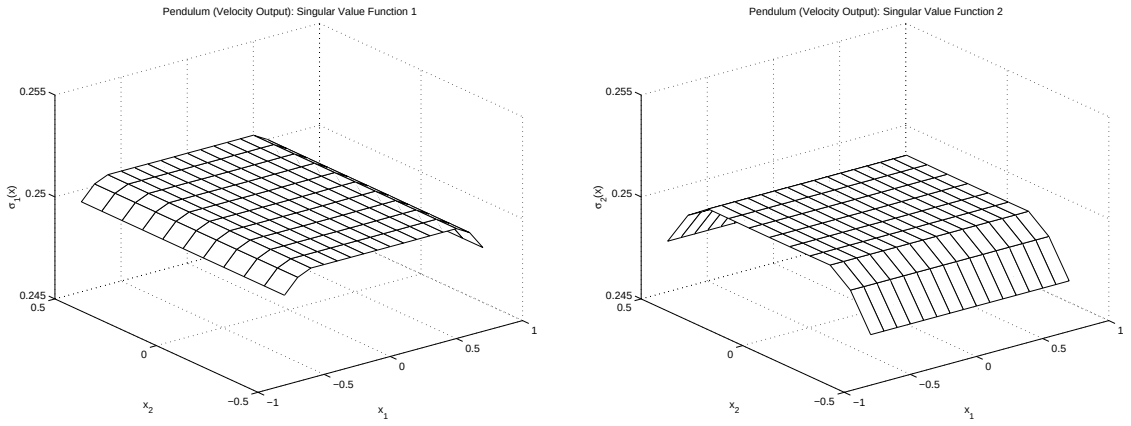


Figure 4.11: The singular value functions for the pendulum system with velocity output. Left: $\sigma_1(x)$ nearly constant 0.252; Right: $\sigma_2(x)$ nearly constant 0.248.

i.e., cannot exhibit the same behaviors, e.g., oscillation. □

Algorithm 4.5.13 produces discretized approximations to the functions $\bar{f}$, $\bar{g}$, and $\bar{h}$, i.e., gives their values at the grid points. In order to simulate the balanced system, we need explicit expressions for evaluating these functions anywhere in a region of 0. Therefore, we approximate the discretized functions with degree-4 polynomials using a linear least squares approximation scheme.

We computed balanced realizations for the pendulum systems with position

and velocity outputs in two ways: using the exact controllability function given by (4.171) and the approximate controllability function derived from Monte-Carlo data. We then simulated the eight systems (original and balanced coordinates; exact and approximate controllability function; position and velocity read-outs) using two input signals: $u \equiv 0$ (natural response) and $u(t) = 0.5 \sin(t/\pi)$. The output responses are shown, for the pendulum system with measured joint angle and angular velocity, respectively, in Figures 4.12 and 4.13.

Theoretically, the output responses of the original and balanced systems should be identical, since they are merely different representations of the same physical system. However, the computations introduce numerical error. We observe that by using the exact controllability function, the output responses of the original and balanced systems are virtually identical. Thus, the algorithms for computing the Morse, input-normal, and balancing transformations introduced negligible error. On the other hand, when using the approximate controllability function generated using Monte-Carlo data, the output responses of the original and balanced systems deviate somewhat. Thus, a better approximation may be desirable, which can be achieved by generating additional Monte-Carlo data.

4.6.2 Toward a Balanced Realization for the Double Pendulum

We now consider a double pendulum system as illustrated in Figure 4.14. As with the pendulum system of Section 4.6.1, the system incorporates linear torsional damping, linear torsional stiffness, and torque inputs at the rotary joints. We assume that the shafts are massless and that the pendulum moves only in the plane. We measure the horizontal position of the end-effector as the system output

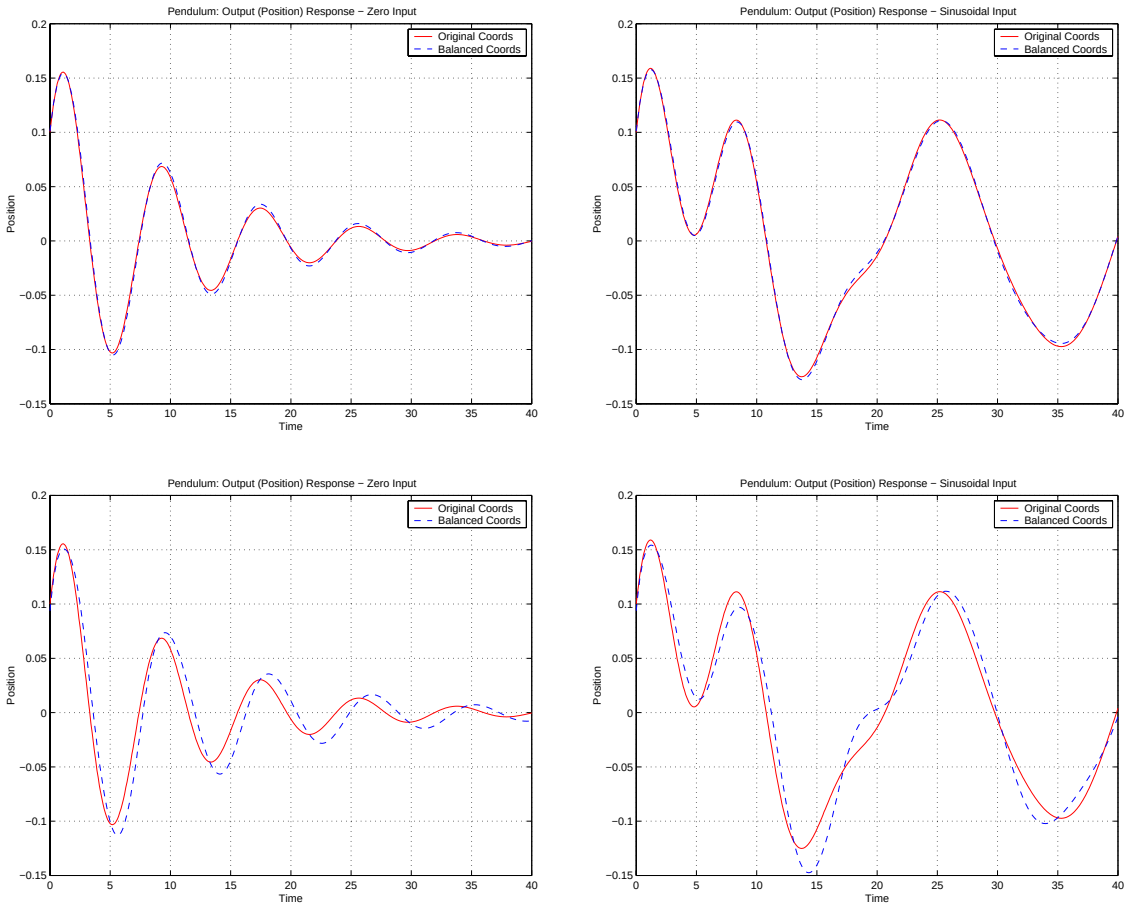


Figure 4.12: Output response for the pendulum system with position read-out: original coordinates (solid) vs. balanced coordinates (dashed). Top left: zero input, exact L_c; Top right: sinusoidal input, exact L_c; Bottom left: zero input, approximate L_c; Bottom right: sinusoidal input, approximate L_c.

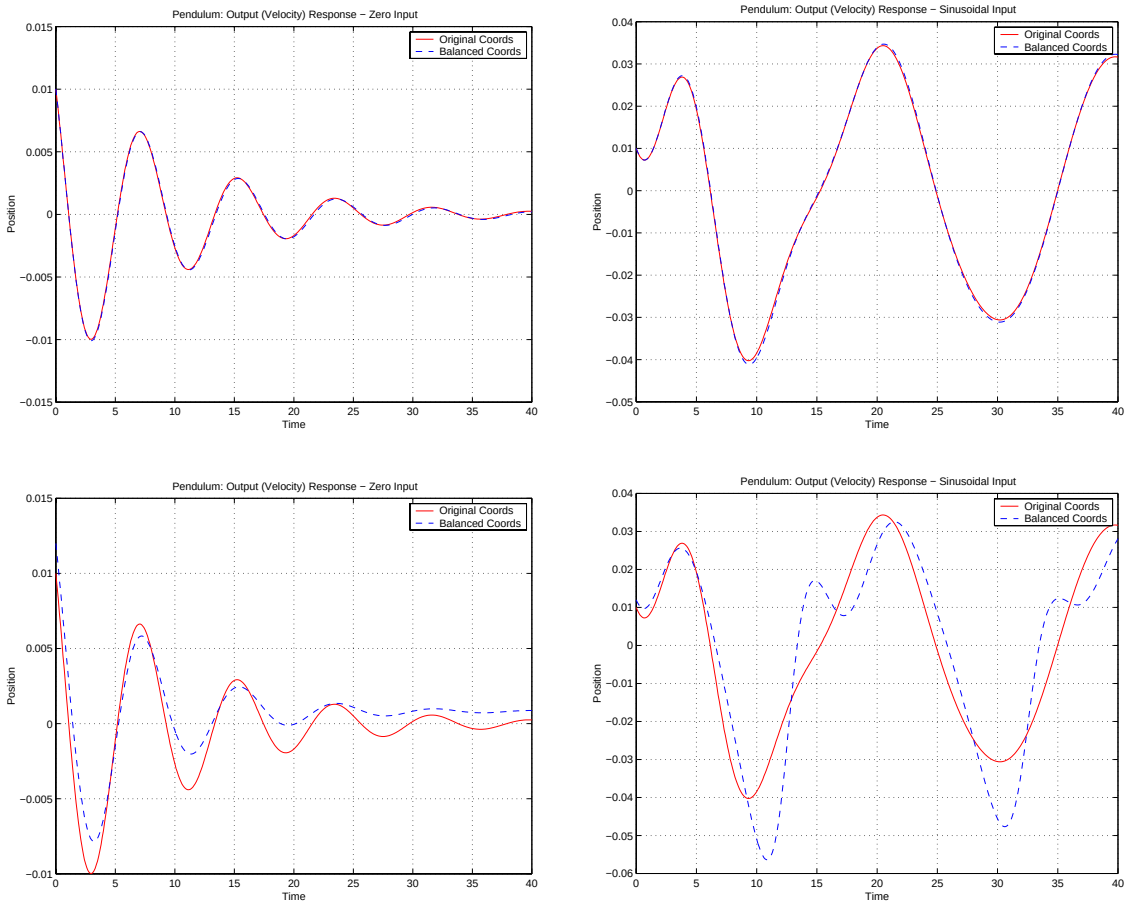
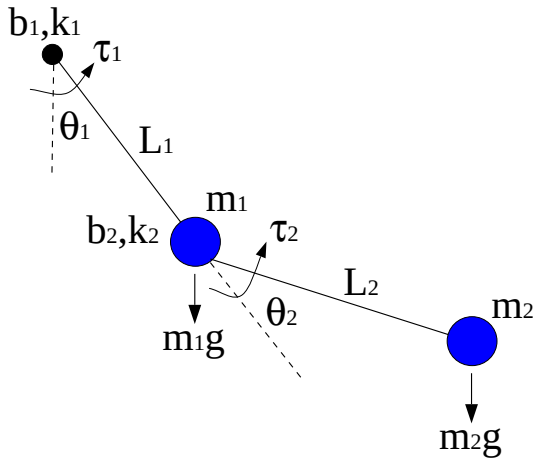


Figure 4.13: Output response for the pendulum system with velocity read-out: original coordinates (solid) vs. balanced coordinates (dashed). Top left: zero input, exact L_c; Top right: sinusoidal input, exact L_c; Bottom left: zero input, approximate L_c; Bottom right: sinusoidal input, approximate L_c.

(a nonlinear function of the state variables).



θ_1		joint 1 angle	θ_2		joint 2 angle
τ_1		joint 1 applied torque	τ_2		joint 2 applied torque
m_1	1	end mass 1	m_2	1	end mass 2
b_1	1	joint 1 tor. damping coeff.	b_2	1	joint 2 tor. damping coeff.
k_1	1	joint 1 tor. stiffness coeff.	k_2	1	joint 2 tor. stiffness coeff.
L_1	1	length of shaft 1	L_2	1	length of shaft 2
G	10	gravitational acceleration			

Figure 4.14: Planar double pendulum system with massless shafts, linear torsional damping, linear torsional stiffness, and torque input applied at the rotary joints. Values of parameters are provided for the numerical studies that we conducted.

State-Space Realization

As before, we obtain a state-space realization (f, g, h) for the pendulum system via the Euler-Lagrangian mechanics outlined in Section 2.7. Let $q = (\theta_1, \theta_2)$ and $\dot{q} = \left(\dot{\theta}_1, \dot{\theta}_2\right)$ denote the generalized positions and velocities corresponding, respectively, to joint angles and angular velocities. Let the generalized forces be given by the applied joint torques, i.e., $F = (\tau_1 - \tau_2, \tau_2)$. The kinetic, potential, and dissipation energies are given, respectively, by

$$
\begin{aligned}
K(q,\dot{q}) &= \frac{1}{2} m_1 L_1^2 \dot{q}_1^2 + \frac{1}{2} m_2 L_1^2 \dot{q}_1^2 \\
&+ \frac{1}{2} m_2 L_2^2 \left(\dot{q}_1 + \dot{q}_2\right)^2 + m_2 L_1 L_2 \cos(q_2)\ \dot{q}_1\ (\dot{q}_1 + \dot{q}_2) \qquad (4.181) \\
U(q,\dot{q}) &= \frac{1}{2} k_1 q_1^2 + \frac{1}{2} k_2 q_2^2 \\
&- (m_1 + m_2)\ G L_1 \cos(q_1) - m_2 G L_2 \cos(q_1 + q_2) \qquad (4.182) \\
R(q,\dot{q}) &= \frac{1}{2} b_1 \dot{q}_1^2 + \frac{1}{2} b_2 \dot{q}_2^2 \qquad (4.183)
\end{aligned}
$$

We apply the Euler-Lagrange equation of motion (2.113), i.e.,

$$
\frac{d}{dt}\frac{\partial L}{\partial \dot{q}} - \frac{\partial L}{\partial q} = F - \frac{\partial R}{\partial \dot{q}} \qquad (4.184)
$$

for $L = K - U$ to obtain the equation of motion for the double pendulum system, given by

$$
M(q)\ \ddot{q} + C(q,\dot{q}) + N(q,\dot{q}) = F \qquad (4.185)
$$

where

$$
M(q) \qquad (4.186)
$$

$$
= \begin{bmatrix} (m_1 + m_2)\ L_1^2 + m_2 L_2^2 + 2\, m_2 L_1 L_2 \cos(q_2) & m_2 L_2^2 + m_2 L_1 L_2 \cos(q_2) \\ m_2 L_2^2 + m_2 L_1 L_2 \cos(q_2) & m_2 L_2^2 \end{bmatrix}
$$

$$
C(q,\dot{q}) = \begin{bmatrix} -m_2 L_1 L_2 \sin(q_2)\ (2\,\dot{q}_1 \dot{q}_2\ + \dot{q}_2^2) \\ m_2 L_1 L_2 \sin(q_2)\ \dot{q}_1^2 \end{bmatrix} \qquad (4.187)
$$

$$
N(q,\dot{q}) = \begin{bmatrix} (m_1 + m_2)\ G L_1 \sin(q_1) + m_2 G L_2 \sin(q_1 + q_2) + k_1 q_1 + b_1 \dot{q}_1 \\ m_2 G L_2 \sin(q_1 + q_2) + k_2 q_2 + b_2 \dot{q}_2 \end{bmatrix} \qquad (4.188)
$$

The affine nonlinear control system is realized in coordinates

$x = (x_1, x_2, x_3, x_4) = (q_1, q_2, \dot{q}_1, \dot{q}_2)$ by

$$f(x) = \begin{bmatrix} \dot{q} \\ -M^{-1}(q)\ (C(q,\dot{q}) + N(q,\dot{q})) \end{bmatrix} \qquad g(x) = \begin{bmatrix} 0 \\ M^{-1}(q) \end{bmatrix} \tag{4.189}$$

and $h(x) = L_1 \sin(q_1) + L_2 \sin(q_1 + q_2)$.

System Properties

We have verified local accessibility and local observability of the system via calculations performed using the symbolic computation capabilities of MATLAB. The expressions for the various brackets, accessibility algebra, and observability codistribution are too lengthy and complicated to include here. The system is locally accessible at 0 since

$$\dim\left(\text{span}\left\{g_1, [f,g_1], [f,[f,g_1]], [f,[f,[f,g_1]]]\right\}|_{x=0}\right) = 4 \tag{4.190}$$

Furthermore, the system is locally observable at 0 since

$$\dim\left(\text{span}\left\{dh, dL_f h, dL_{[f,[g_1,[f,g_1]]]}h, dL_{[f,[g_1,[g_1,[f,[f,f]]]]]}h\right\}|_{x=0}\right) = 4 \tag{4.191}$$

Finally, we note that the system is asymptotically stable, since, for $A = \dfrac{\partial f}{\partial x}(0)$, $\text{spec}(A) \in \mathbb{C}^-$.

The linearization about 0 is given by

$$A = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -11 & 12 & -1 & 2 \\ 12 & -35 & 2 & -5 \end{bmatrix} \qquad B = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & -3 \\ -2 & 7 \end{bmatrix} \tag{4.192}$$

and $C = [2 \quad 1 \quad 0 \quad 0]$. The Hankel singular values of the linearization are 0.5029, 0.4702, 0.0249, and 0.0106.

Controllability Function

The double pendulum system is not integrable and does not, in general, satisfy the equipartition of energy condition. However, in the special case where $b_1 = b = b_2$ the equipartition of energy condition is satisfied with ratio $\ell = b$. Applying Theorem 4.3.14, the controllability function for the double pendulum system is given, exactly, by

$$L_c(x) = \tag{4.193}$$

$$b\,(m_1+m_2)\,L_1^2\,x_3^2 + b\,m_2\,L_2^2\,(x_3+x_4)^2 + 2\,b\,m_2\,L_1\,L_2\cos(x_2)\;x_3\;(x_3+x_4) +$$

$$b\,k_1\,x_1^2 + k_2\,x_2^2 - 2\,b\;(m_1+m_2)\;G\,L_1\cos(x_1) - 2\,b\,m_2\,G\,L_2\cos(x_1+x_2) +$$

$$2\,b\,G\;((m_1+m_2)\;L_1 + m_2\,L_2) \tag{4.194}$$

and after substitution of the parameter values given in Figure 4.14 by

$$\begin{aligned} L_c(x) \;\; = \;\; & 2\,x_3^2 + (x_3+x_4)^2 + 2\cos(x_2)\;x_3\;(x_3+x_4) + x_1^2 + x_2^2 \\ & -40\cos(x_1) - 20\cos(x_1+x_2) + 60 \end{aligned} \tag{4.195}$$

The controllability function L_c for the double pendulum system is shown in Figure 4.15.

Observability Function

We use Algorithm 4.5.14 to compute an approximation of the observability function L_o, for the double pendulum system, shown in Figure 4.16.

Balanced Realization

Figure 4.17 shows a 2-dimensional slice of each singular value function for the double pendulum system. At the origin, the singular value functions take the values,

respectively, 0.487, 0.444, 0.135, and 0.050. These values are reasonable close to what we expect from the Hankel singular values of the linearization. Two states of the balanced realization have considerably greater input-to-output importance than the other two states. We also observe that numerical errors are more prominent for the singular value functions of small magnitude, i.e., the oscillations that they display are likely caused by numerical error rather than being an accurate reflection of their actual behavior.

4.7 Remarks

We have presented methods and algorithms to compute the energy functions and coordinate transformations involved in the Scherpen theory and procedure for nonlinear balancing. We have shown that, under certain conditions, an exact formula for the controllability energy function can be derived. We applied our result to compute the controllability function for a 4-dimensional mechanical system. For other situations, we offer a Monte-Carlo approach for approximating the controllability function. For a 2-dimensional mechanical system, the Monte-Carlo approach yielded a good approximation.

We have presented an algorithm for a numerical implementation of the Morse-Palais lemma, which produces a local coordinate transformation under which a real-valued function with a non-degenerate critical point is quadratic on a neighborhood of the critical point. Application of the algorithm to the controllability function plays a key role in computing the balanced representation.

We have applied our methods and algorithms to derive balanced realizations for nonlinear state-space models of two example mechanical systems. Simulation results demonstrate that the algorithms produce accurate and useful approximations

to the energy functions and transformations involved in the nonlinear balancing procedure. For a 2-dimensional system, the approximate balanced realization produced input-to-output behavior that is nearly equivalent to that generated by the original, physically derived, realization. Thus, it serves as an equivalent representation, with the benefit that it provides a meaningful ranking of state components for purposes of model reduction.

The algorithms are currently too computationally intensive to be practical for high-order systems. It is likely that new algorithms will be required in order for the Scherpen procedure to become genuinely useful for model reduction of nonlinear systems.

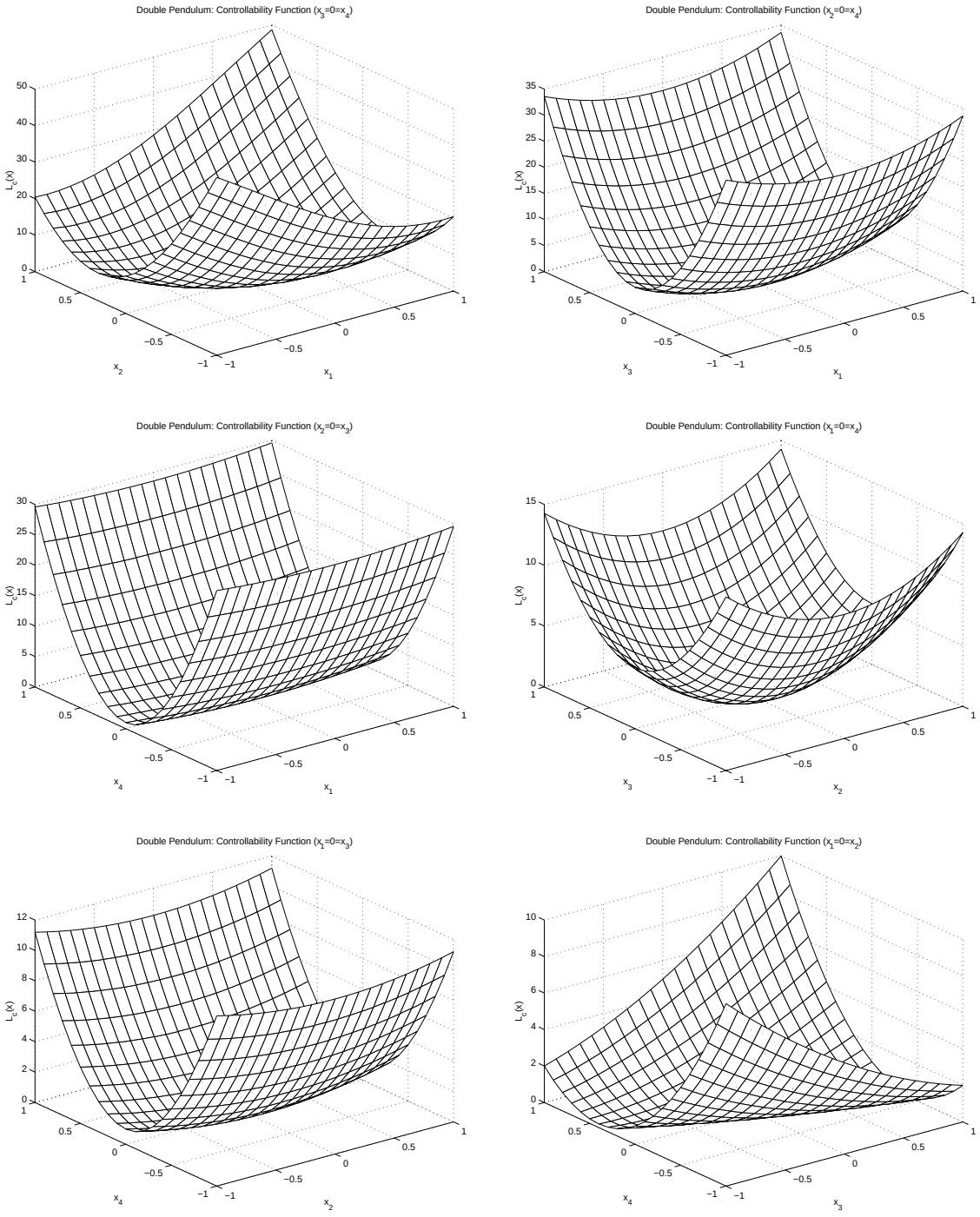


Figure 4.15: Controllability function for double pendulum (6 planes). Top left: $x_3 = 0 = x_4$; Top right: $x_2 = 0 = x_4$; Mid left: $x_2 = 0 = x_3$; Mid right: $x_1 = 0 = x_4$; Bottom left: $x_1 = 0 = x_3$; Bottom right: $x_1 = 0 = x_2$.

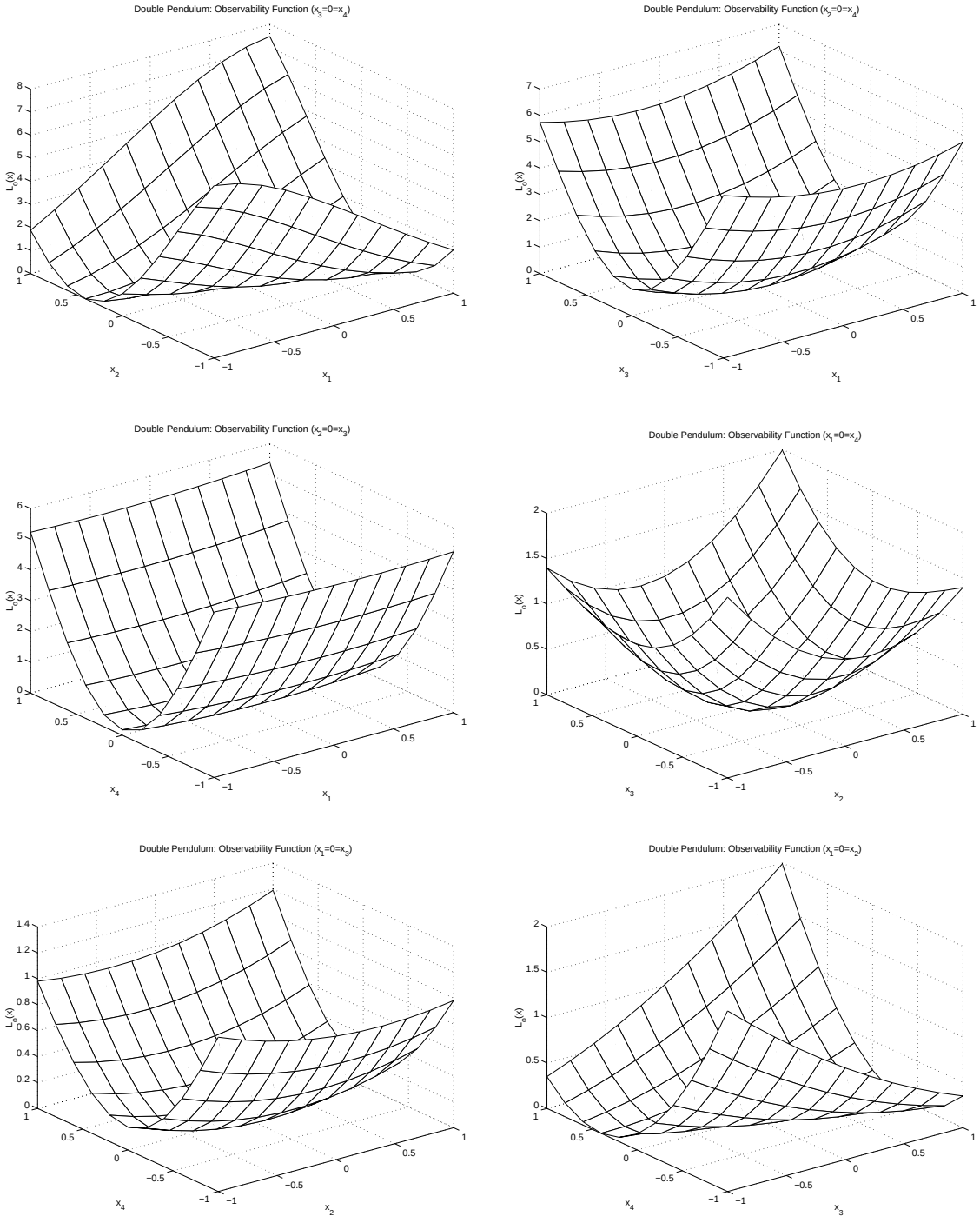


Figure 4.16: Observability function for double pendulum (6 planes). Top left: $x_3 = 0 = x_4$; Top right: $x_2 = 0 = x_4$; Mid left: $x_2 = 0 = x_3$; Mid right: $x_1 = 0 = x_4$; Bottom left: $x_1 = 0 = x_3$; Bottom right: $x_1 = 0 = x_2$.

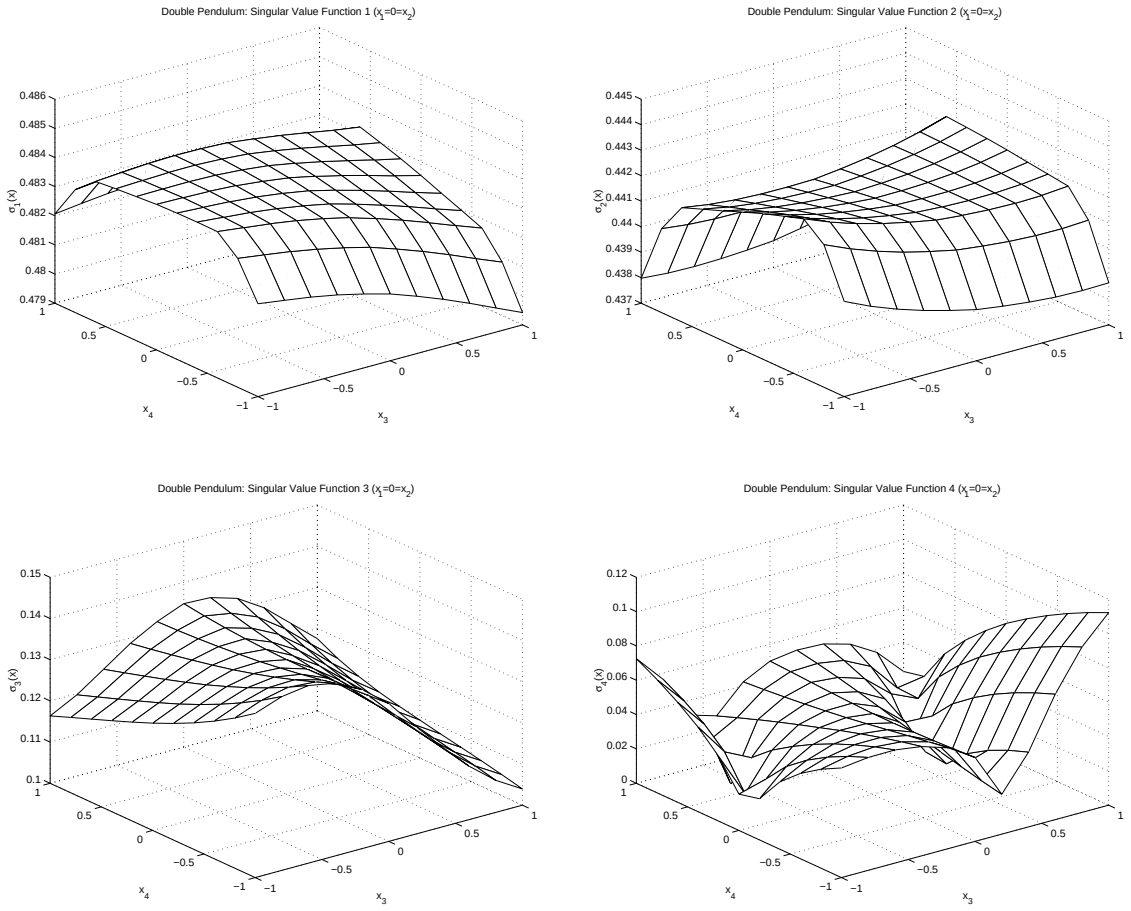


Figure 4.17: Singular value functions for double pendulum (x_3-x_4 plane). Top left: $\sigma_1(x)$; Top right: $\sigma_2(x)$; Bottom left: $\sigma_3(x)$; Bottom right: $\sigma_4(x)$.

Chapter 5

Modeling and Optimization for Silicon Growth via RTCVD

5.1 Introduction

This chapter addresses the problem of developing high-fidelity physical-chemical models for predicting the behavior and output of a commercial rapid thermal CVD reactor used for depositing thin films of Si and Si-Ge on silicon wafers. The problem is studied within the context of a joint project between the ISR of the University of Maryland, College Park, and Northrop Grumman ESSS (NG-ESSS), for modeling and optimization of epitaxial growth in the ASM Epsilon-1 rapid thermal CVD reactor [115, 116, 117]. The Epsilon-1 is a single-wafer lamp-heated CVD reactor manufactured by ASM America, Inc., Phoenix, AZ. It is used by NG-ESSS to deposit layers of epitaxial Si-Ge, epitaxial silicon (epi-Si), and polycrystalline silicon (poly-Si) on a silicon wafer.

The modeling of fundamental aspects of CVD involves both chemical kinetics and transport phenomena. Depending on the specific process and operating

conditions, it is often assumed that one or more factors has significantly more influence than all others over deposition product. In those cases, the factors that are considered less important are often completely or mostly neglected. This type of simplification has been adopted frequently in order to formulate relatively simple models describing specific components of the CVD process (e.g., heat transfer within and among solids) under a limited range of operating conditions (e.g., low-temperature regime) for a specific purpose (e.g., temperature control).

For example, NG-ESSS is interested in models that focus on low temperature silicon epitaxy. In the low temperature regime, growth rate is limited by surface reaction phenomena rather than by mass transport phenomena. Furthermore, surface reaction phenomena such as adsorption and desorption of reactant species are strongly dependent on temperature. For this reason, the low temperature regime is said to be thermally driven (kinetically limited). These simplifications motivate an approach in which physical-chemical models consist only of a simplified conjugate heat transfer model for thermal dynamics together with an Arrhenius law for growth rate in terms of temperature. In this approach, dynamics of transport phenomena are neglected, inlet conditions for gas phase species concentrations and temperature are assumed to hold throughout the process chamber, and chamber geometry plays no role.

However, as we show in this chapter, these simplifications seriously compromise the utility of such models for purposes of studying uniformity issues for thin film growth in the Epsilon-1. We demonstrate that chamber geometry and a variety of complex phenomena are essential elements, including depletion of reactants, non-uniform gas heating, gas phase chemistry, thermal diffusion, and gas flow patterns. This necessitates the incorporation of detailed models for three-dimensional

effects of gas flow, gas phase heat transfer, and transport of chemical species, in addition to the previously mentioned heat transfer phenomena. Furthermore, a more comprehensive model for surface reaction chemical kinetics is required to incorporate reactive intermediaries produced in the gas phase. Finally, it is crucial that the models reflect the coupling among these various phenomena in the process chamber.

We address these problems via development of a process-equipment model which accounts for the mechanisms and factors described above. The model is capable of predicting gas flow, heat transfer, species transport, and chemical mechanisms in the reactor given a process recipe for temperature, pressure, and flow rate set-points. It provides a platform for studying the effect of equipment settings and a relatively broad range of process conditions on deposition product characteristics, e.g., deposition rate and thickness uniformity.

The modeling effort includes development of physical and chemical models for fundamental CVD phenomena, experimental determination of growth parameters, and experimental validation of model predictions. Simulations are used as tools to predict deposition results, study the factors that affect deposition uniformity, and determine operating parameters for improved performance and product quality.

Specific applications include prediction and control of deposition rate and thickness uniformity; studying sensitivity of deposition rate to process settings such as temperature, pressure, and flow rates; and reducing the use of consumables via purge flow optimization. The implications of various simulation results are discussed in terms of how they can be used to reduce costs and improve product quality, e.g., thickness uniformity of thin films. We demonstrate that achieving deposition uniformity requires some degree of temperature non-uniformity to com-

pensate for the effects of other phenomena such as reactant depletion, gas heating and gas phase reactions, thermal diffusion of species, and flow patterns.

The semiconductor manufacturing environment in which this research was conducted is described in Section 5.2, including manufacturing objectives, equipment and materials involved, and a typical manufacturing situation of interest. The procedure and results of poly-Si growth experiments for studying equipment operation and deposition characteristics are given in Section 5.3. The overall process-equipment model for silicon growth in the Epsilon-1 is detailed in Section 5.4, including various components for prediction of the relevant transport phenomena and chemical mechanisms. In Section 5.5, we apply the model, via simulation, to study the factors that influence deposition rate and uniformity, and present the results and analysis. We summarize and make some additional remarks in Section 5.6.

5.2 Semiconductor Manufacturing Environment

The modeling and analysis presented in this chapter pertains specifically to the semiconductor manufacturing environment at NG-ESSS, e.g., silicon epitaxy using the Epsilon-1 reactor, and is motivated specifically by problems encountered within that environment. In this section we state the manufacturing objectives that motivated the modeling effort, provide relevant details about the equipment and processes, and offer some perspective and additional motivation through a case study describing a typical manufacturing situation of interest.

5.2.1 Manufacturing Objectives

The overall objective of this research is to improve manufacturing effectiveness for epitaxial growth of silicon and Si-Ge thin films on a silicon wafer in the Epsilon-1 reactor, a production tool currently in use at NG-ESSS. Improvement in product quality, increased flexibility of operation, and reduction of manufacturing costs are integral to achieving the overall objective. We provide specific details in the following explanations and descriptions, based on discussions with and demonstrations by NG-ESSS personnel [128].

Within the scope of this research, product quality is determined solely by deposition thickness uniformity. Other factors, such as film composition and resistivity uniformity, are important quality measures but are not considered here. Thickness variations of 5% are currently acceptable for most applications, although there is no guarantee that such a specification will remain stable. Currently, variations in the range of 2% are routinely achieved with the Epsilon-1. Improved results are always desirable.

The Epsilon-1 is capable of operation in several regimes for pressure, temperature, and flow rates, and deposition via injection of several types of precursor and carrier gases. Prediction of deposition rates and other film characteristics for a given combination of process conditions is key to taking advantage of the machine's flexibility. The manufacturer provides some predictive guidance and data. However, there is the desire for manufacturing "off-the-curve," i.e., operating in regimes and producing films with characteristics that do not appear in manufacturer provided information.

Performance of devices at high frequency is difficult to predict based on properties of the product and manufacturing parameters. To achieve a product with the

desired properties, NG-ESSS operates with a three to four month manufacturing cycle followed by a long testing cycle. It can take up to two years to converge on the desired product. Each manufacturing cycle requires an initial period of experimentation in which the necessary equipment settings and process conditions are determined. Once parameters are determined, and the customer is satisfied, the process is certified, and parameters are usually not changed for several years in order to provide the customer with a consistent product. It is possible, however, for drift of equipment characteristics over time to cause degradation, necessitating additional experimentation. In addition, the chamber tube is periodically cleaned and replaced, requiring a re-calibration of process settings. The result is that the various trial-and-error steps have a significant impact on time-to-manufacture and other production costs.

Additional cost concerns include operational integrity and "down-time" of equipment, and the use of consumables such as process gases. It is clear that reductions in experimental steps, equipment failure, and gas consumption will have a beneficial impact on manufacturing costs.

In light of manufacturing objectives, the modeling effort described in this chapter seeks to gain an understanding of the processes and equipment via physical and mathematical modeling, and use the resulting validated models for optimization of process conditions and equipment settings.

5.2.2 Equipment and Materials

The Epsilon-1 reactor is a radiantly heated, gas injected, single wafer processing system for CVD of doped or undoped epitaxial and polycrystalline layers on a 150 mm (6 in) diameter semiconductor wafer. In this section we provide some

descriptive background material necessary for model development, including characteristics of the process chamber and deposited and consumed materials; and an overview of reactor operation including typical processing recipes, operating conditions, equipment settings, and overall system structure.

Process Chamber

The process chamber is situated within the Epsilon-1 reactor system, accessible by the wafer handling system and between the parts of the lamp assembly, as shown in Figure 5.1. Also shown is a cross-sectional front view of the process chamber and wafer rotation apparatus. A cross-sectional side view of the process chamber and lamp assembly, and a top-down view of the wafer level apparatus, are shown, respectively, in Figures 5.2 and 5.3. The inlet and outlet sides of the reactor are referred to as the front (upstream) and rear (downstream), respectively.

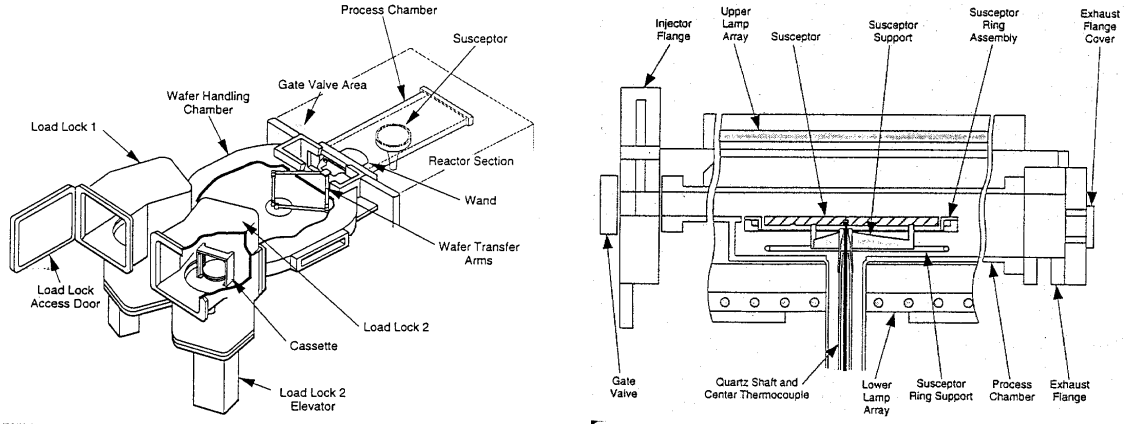


Figure 5.1: Epsilon-1 reactor system (left) and cross-section (front view) of the process chamber and wafer rotation apparatus (right). Source: ASM Epsilon-1 Reactor Manual.

Deposition takes place in the process chamber, which is a horizontally oriented quartz tube of lenticular shape, i.e., a cross-sectional view looking into the chamber front shows a flat bottom, short vertical sides, and curved top. The quartz shelves

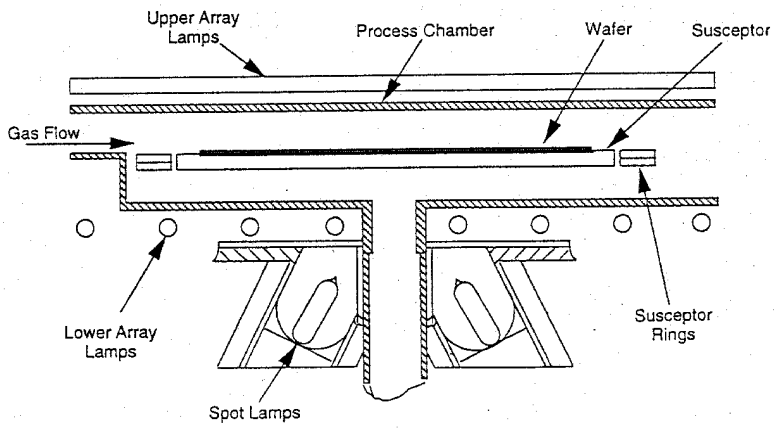


Figure 5.2: Cross-section (side view) of the Epsilon-1 process chamber and lamp assembly. Source: ASM Epsilon-1 Reactor Manual.

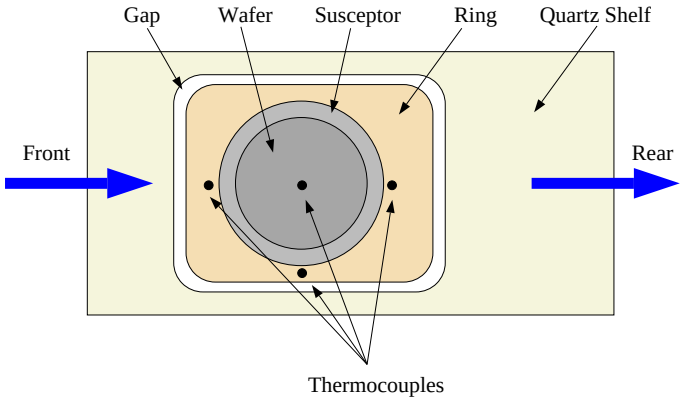


Figure 5.3: Overhead view of the Epsilon-1 at wafer level including thermocouple locations.

are connected to the quartz chamber walls to form a contiguous solid body.

The 150 mm (6 in) diameter wafer rests on small quartz pins attached to the pocket of a rotating susceptor that is surrounded by the susceptor ring. The susceptor and ring are constructed of graphite coated with silicon-carbide. The susceptor ring fits into a space within the quartz structure, leaving a thin gap between ring and shelf on all sides. The susceptor fits into the ring structure, and is supported and rotated by special apparatus located through and under the lower chamber section. There is also a gap, although much smaller, between the ring

and susceptor.

Process gases are pumped into the chamber through the inlet flange, flow horizontally through the chamber over the wafer surface, and are pumped out through the exhaust flange via pneumatic actuators. An optional flow guide can be used to force the inlet flow away from the chamber roof and toward the wafer. The inlet flange is designed to create a specialized sonic flow with possible swirling and mixing properties as the gas enters the process chamber. It has been observed by the manufacturer that the inlet flange design aids in achieving deposition uniformity [106].

The chamber is divided by the susceptor, ring, and a quartz shelf into upper and lower sections. Process gases enter and flow through the upper section; purge gases enter into the lower section. Thin gaps between the quartz shelf and the ring, and between the ring and susceptor, allow gas to flow between upper and lower chamber sections. In addition, diffusion of species from upper to lower and visa-versa can occur due to concentration and thermal gradients. For this reason, purge gases are pumped into the lower region through the susceptor rotation shaft and a purge inlet in the front wall. The purge flow prevents the process gases from escaping to the lower section, which can result in unwanted deposition on the back-side of the susceptor.

The wafer and chamber are heated by upper and lower arrays of linear tungsten-halogen lamps, and four spot lamps directed at the center of the susceptor (see Section 6.3 for details and analysis). The upper and lower lamp arrays illuminate, respectively, the top surface of the wafer and the bottom of the susceptor. Heat radiation is intensified by gold coated reflectors surrounding the process chamber on all sides. Four thermocouples measure the temperature at the center, front,

rear, and side of the susceptor. We note that while the center thermocouple is located in contact with the center of the susceptor, the other three thermocouples are located at the front, rear, and side of the ring that surrounds the susceptor. Thus, susceptor temperature is measured only approximately at points other than the center.

The quartz chamber and lamp-house are cooled by air flow. All components are contained in a stainless steel enclosure.

Product and Consumables

NG-ESSS uses the Epsilon-1 reactor to deposit thin films of epitaxial Si-Ge, epi-Si, and poly-Si. Epitaxial growth, or epitaxy, (see, e.g., [134, 162]) refers to the deposition of a thin layer of material onto the surface of a single-crystal substrate in such a manner that the layer is also single-crystal and has a fixed and predetermined crystallographic orientation with respect to the substrate. Epitaxial layers are deposited on silicon wafers that are either bare or covered with a patterned layer of silicon dioxide (SiO_2). Poly-Si is deposited on a layer of silicon dioxide.

The major source gases used to deposit epi-Si layers commercially are (see, e.g., [134, 156])

- silane (SiH_4) at low temperatures (< 1000 C); and
- silicon tetrachloride ($SiCl_4$), dichlorosilane (SiH_2Cl_2), and trichlorosilane ($SiHCl_3$) at higher temperatures.

This research deals with low temperature growth, for which NG-ESSS uses silane as the source gas. Germane (GeH_4) is added to the mixture for growth of Si-Ge. The carrier gas is either hydrogen (H_2) or nitrogen (N_2). Dopant precursors such as arsine (ArH_3) can also be added to the mixture.

There is a direct relationship between growth of poly-Si and epi-Si. The processing steps are identical, and growth rates are virtually identical [128]. Experiments are performed at NG-ESSS by depositing poly-Si rather than epi-Si, because they have tools for measuring poly-Si films (e.g., nanospec, ellipsometer) but not for epi-Si (e.g., SIMS). (It is the presence of the oxide boundary that allows for easier measurement of poly-Si.) Thus, poly-Si experiments allow for rapid process evaluation. For this reason, NG-ESSS sometimes performs epi-Si and poly-Si experiments in parallel, with measurements taken for the poly-Si films. In light of this, we performed experiments and simulations for growth of poly-Si. The process gases we used were a mixture of 2% SiH_4 diluted in H_2 as the silicon source, together with 20 slm H_2 as the carrier, except as noted otherwise.

Specifications for the final product (thin film) include characteristics such as chemical composition, film thickness, dopant concentration, crystal structure, resistivity, and possibly other factors. Both aggregate and spatially distributed quantities are important. Usually, spatial uniformity is specified by a variation tolerance across the wafer surface, e.g., 5% allowable non-uniformity.

In general, aggregate characteristics such as average growth rate are determined by process conditions for temperature, pressure, and flow rates as set by the user in process recipes. The spatial distributions of the various properties, and hence uniformity, are mainly controlled by equipment settings such as thermocouple offsets and injector opening sizes. These two methods of equipment and process control are described below.

Process Conditions and Recipes

In order to achieve the desired aggregate characteristics, the process engineer designs step-by-step recipes. Each step performs a particular task such as etch, bake, purge, or deposit, for a specified amount of time. We consider only the deposition steps here. The process engineer specifies, for each deposition step, the choice of source, carrier, and purge gases, set-points for temperature, pressure, and flow rates, and time duration. We refer to these specifications as *recipe inputs*. They are programmed into the Epsilon-1 microprocessor and controlled automatically in-situ. For example, PID controllers and mass flow controllers (MFCs) regulate the thermocouple temperatures and inlet flow rates around their respective set-points.

Si-Ge films are deposited at a temperature of 675 C. This falls within the low temperature regime which is roughly 600–800 C. At low temperature, surface reactions are thermally activated and controlled by deposition kinetics. NG-ESSS also deposits some films in the high temperature regime which is roughly 900–1100 C. At high temperature, surface reactions are mass transport controlled. However, temperature regulation is still important, as it determines layer resistivity, and large temperature gradients can cause slip, i.e., mechanical damage to the wafer. All growth data in this study is restricted to the low temperature regime.

The Epsilon-1 reactor is capable of growth at atmospheric pressure (AP) and reduced pressure (RP) which is roughly 10–100 Torr. For this research, we performed deposition in the RP regime at 20 Torr and 40 Torr. The flow rate for each individual process and purge gas used is specified in standard liters per minute (slm) or standard cubic centimeters per minute (sccm). The process gases, e.g., hydrogen carrier and silane source, are mixed prior to injection into the chamber. The purge flow rate is set to prevent mixing between upper and lower chamber

sections, and is generally the same for each recipe. Since its impact is on equipment integrity rather than product characteristics, we treat it separately from the other recipe inputs.

Equipment Settings

Reactor operation can be adjusted ex-situ via several mechanisms that are included in certain components of the reactor. The process engineer can set the size of gas injector openings in the inlet flange, the relative power setting for each lamp group, PID feedback gains, susceptor rotation rate, and thermocouple offsets. We refer to these as *equipment settings*. In contrast to recipe inputs, the equipment settings are semi-permanent, i.e., they are not changed, in general, for each different process recipe. Rather, once an equipment setting is adjusted so that the reactor yields acceptable films, it remains fixed from run to run until process drift or tube replacement necessitates an adjustment. The equipment settings play a key role in achieving spatial uniformity of deposition thickness in the Epsilon-1. We elaborate on some of these settings here.

As stated earlier, wafer temperature is set as a recipe input. However, this one setting does not provide the capability to adjust the temperature distribution across the wafer surface. The necessary additional degrees of freedom are provided by the thermocouple offsets. There are three offsets, one each for thermocouples at the front, rear, and side of the susceptor. The center thermocouple has no offset, and its temperature is regulated about the recipe temperature set-point. The temperatures of the other thermocouples are regulated about the sum of the temperature set-point and the corresponding offset. For example, suppose the temperature set-point is given in the recipe as 700 C, and the front offset is given

as -20 C. Then the center temperature is regulated about 700 C and the front temperature is regulated about 680 C. Use of the offsets has the effect of creating four separate temperature set-points. However, even with the additional degrees of freedom, the authority to control the entire susceptor temperature profile is limited. The profile can be set only roughly at points other than at the thermocouple locations. Offsets are currently set via trial-and-error and heuristic methods.

In the inlet flange currently installed in the Epsilon-1 at NG-ESSS, there is a set of three gas injector slits with adjustable widths. These are used for adjusting the flow profile at the inlet to the process chamber. We note that the manual adjustment of slit widths is difficult, and the widths can be measured only approximately. In the future, this equipment will be replaced by a set of five injector port orifices with adjustable diameters. The new gas supply equipment will allow for tighter control and more degrees of freedom in determining the inlet flow profile. Either way, however, the authority to control the flow characteristics is once again limited. The manner in which the size of gas injector openings affect the flow profile is known only roughly. The size of gas injector openings are currently determined via trial-and-error and heuristic methods.

Wafer rotation is used to smooth non-uniform heating and other effects. It is typically set at 35 rpm for most, if not all, production runs.

Operating Structure

An overview of the general operating structure of the Epsilon-1 reactor, from the viewpoint of how recipe inputs and equipment settings affect reactor operation, is presented in Figure 5.4. Note that the internal details of individual blocks are not included here. They will be discussed whenever relevant later in this report.

Of particular interest is the process chamber block, which contains physical and chemical mechanisms for film growth.

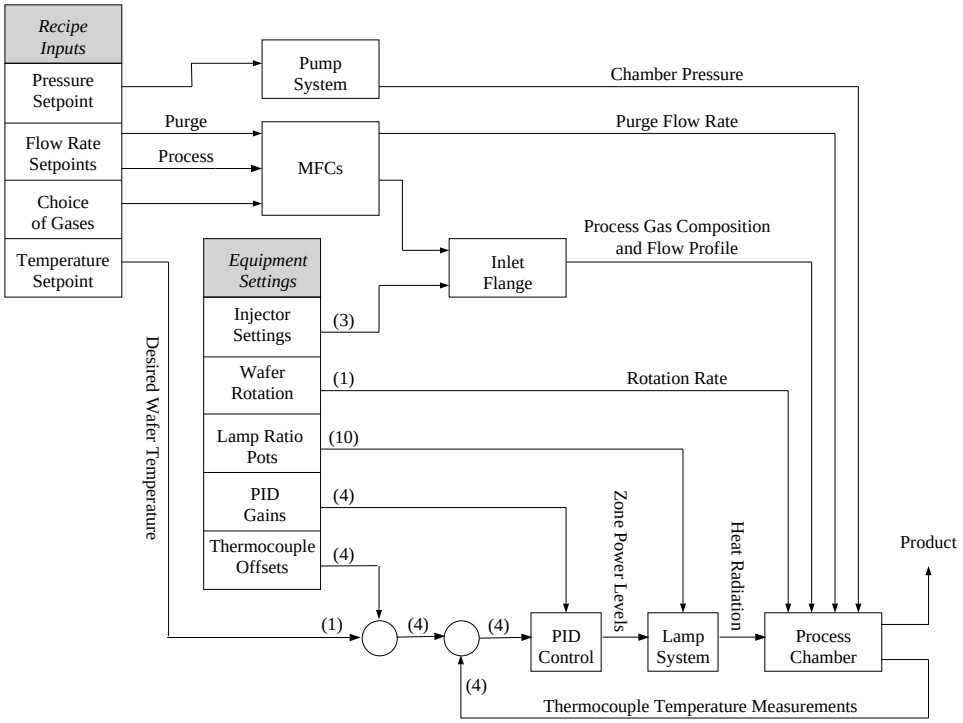


Figure 5.4: Overview of general operating structure of Epsilon-1 reactor, from the viewpoint of how recipe inputs and equipment settings affect reactor operation. Numbers in parentheses refer to the number of distinct signals in the associated path.

5.2.3 Uniformity Case Study

An essential purpose of the process-equipment model is to predict the steady state deposition rate with emphasis on the spatial distribution of film thickness. It is crucial, then, to identify those phenomena that are important to determining growth rate, and include the effects of those phenomena in the model. It is also necessary to identify and include relevant features of the reactor geometry and operation, and to incorporate sufficient spatial resolution and dimensionality.

For thermally activated thin film growth, it is usually assumed or implied in the literature that achieving deposition uniformity is tantamount to achieving temperature uniformity across the wafer surface. Optimization and control strategies are then designed to achieve the temperature uniformity objective via manipulation of lamp power settings, so that deposition uniformity is achieved via automatic lamp control. Examples of such studies are found in [28, 30, 63, 139, 151, 152]. Sometimes the assumption is justified by stating that wafer rotation will average out all other factors.

However, the assumption of equivalence between temperature uniformity and deposition uniformity does not necessarily hold, even for thermally driven processes and processes in which the wafer is rotating. For example, consider the experience of NG-ESSS with deposition of epi-Si or Si-Ge in the thermally driven regime (approximately 600–800 degrees C) in the Epsilon-1 reactor. The process engineer achieves thickness variations of less than 1.5% (considered acceptable uniformity) for a non-rotating wafer by setting thermocouple offsets at -25, -60, and -35 for front, rear, and side, respectively [128]. These values were determined via trial-and-error growth experiments. If growth rate were affected only by temperature and no other factors, then the 1.5% thickness variation that is achieved with those offsets would correspond to a 0.075% temperature variation across the wafer surface, or roughly 0.5 degrees C. However, for a recipe temperature set-point of 700 C, the corresponding thermocouple set-points are center at 700 C, front at 675 C, rear at 640 C, and side at 665 C, for a maximum deviation of 8.5%, as illustrated in Figure 5.5. Thus, the non-uniformity imposed by the offsets appears to be significantly greater than that indicated by actual growth rates.

The set-up of the reactor apparatus may partially explain the smaller variation

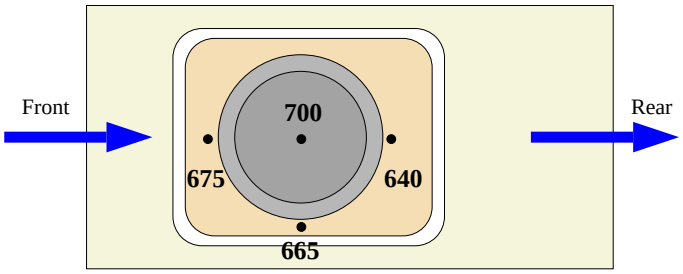


Figure 5.5: An example of how thermocouple offsets (front -25 C, rear -60 C, side -35 C) influence the temperature set-points around which the four thermocouples are regulated by PID controllers.

in actual growth rates. Recall that the front, rear, and side thermocouples are located in the ring surrounding the susceptor, rather than in the susceptor itself. Furthermore, the 150 mm (6 in) diameter wafer is resting on quartz pins at the center of the 225 mm (8.85 in) diameter susceptor. Therefore, it is reasonable to assume that the temperature variation across the wafer surface will be less than the variation across the entire susceptor. However, by similar reasoning, it is also intuitive that this could not entirely account for the thickness uniformity. For example, the front, rear, and side thermocouples are all the same distance from the center. But the offsets are not equal. The apparatus symmetry is not mirrored by the temperature set-points. Some other phenomena must be playing a role.

A quantifiable relationship between the offsets and the actual temperature field on the wafer surface is unknown. This is because there exists no reliable method for measuring temperature on the wafer surface. Poly–silicon growth rates are often used as a sensitive thermometer. While this method may be useful for measuring aggregate temperature, we argue here that it is flawed for measuring a temperature distribution across a surface, unless one can guarantee that other conditions across the surface (e.g., reactant concentrations) are perfectly uniform. The other alternative is to use an instrumented wafer, i.e., a wafer with attached thermocou-

ples. However, ASM America and NG-ESSS consider measurements taken using the instrumented wafer to be unreliable, especially while operating at process conditions for growth [106, 128]. Nevertheless, we report that experiments using an instrumented wafer indicate maximum temperature variations of 5 degrees C or 0.7% [128]. This non-uniformity is less than that predicted by considering just the offsets and more than that found by measuring growth rates. In that respect, it appears to fall in the correct range.

Finally, we note that wafer rotation reduces the growth rate variation from 1.5% to less than 1% and temperature variations as recorded by the instrumented wafer from 5 C to 1 C. Therefore, we may conclude that wafer rotation does have the intended flattening effect, but does not compensate entirely for temperature non-uniformity.

We have demonstrated anecdotally that thickness and growth rate uniformity in the thermally activated regime is achieved by setting three thermocouple offsets so that the temperature distribution across the susceptor is intentionally non-uniform. The actual relationships among offsets, temperature, and growth rate, and the other factors that affect them, are left to be determined.

5.3 Growth Experiments

In this section we describe silicon growth experiments performed to study the relationship between deposition rate and operating conditions such as temperature and flow rates. The experimental data is used later as a basis of comparison to validate the modeling results.

The experiments were conducted by the author and Mr. Paul Brabant of NG-ESSS using the Epsilon-1 reactor on site at NG-ESSS. In each experiment, we

deposited poly-Si on a silicon wafer coated with a layer of silicon-dioxide. These experiments are suitable for determining growth rates of both poly-Si and epi-Si, due to the fact that their growth rates are identical (see Section 5.2.2). The process gases used were a mixture of 2% SiH_4 diluted in H_2 as the silicon source, together with 20 slm H_2 as the carrier. Measurements of film thickness were taken using the available nanospec (Nanometrics 210 XP Scanning UV Nanospec/DUV Microspectrophotometer).

The main objective was to determine the relationship between growth rate and operating conditions such as temperature and flow rates, including finding the unknown parameters (e.g., activation energy) of an assumed Arrhenius relationship between wafer temperature and deposition rate. The relationships were studied under a range of typical operating conditions. The established relationship is used later for validation of process-equipment simulations (Section 5.5) and lamp heating models (Section 6.3).

Experimental Procedure

Thin films of polycrystalline silicon were deposited from the silane precursor over a five minute period at a pressure of 20 Torr. Deposition was performed under a combination of operating conditions consisting of four different wafer temperatures and three different silane flow rates (and hence three different silane mole fractions).

Temperatures were set in the surface-reaction controlled regime so that growth would be thermally activated. This regime is roughly from 600 C to 800 C for deposition of silicon from silane gas. We chose the following wafer temperatures at which to deposit silicon: 650 C, 700 C, 725 C, and 750 C.

Three different flow rates were used for the 2% silane in hydrogen precursor:

1.5 slm, 2.5 slm, and 3.5 slm. Considering the 2% dilution, these three flow rates correspond to 30 sccm, 50 sccm, and 70 sccm of silane, respectively. The silane-hydrogen precursor was again diluted in 20 slm of the carrier hydrogen (H_2) gas. Thus, the three flow rates correspond to three mole fractions 1.4 $\times 10^{-3}$, 2.2 $\times 10^{-3}$, and 3.0 $\times 10^{-3}$, respectively.

The reactor was operated in its usual, automatic mode (i.e., using PID control loops for temperature regulation and wafer rotation for uniformity), using pre-programmed recipes. Recipes were programmed to set chamber pressure at 20 Torr and to deposit silicon from silane precursor for five minutes onto the bare silicon wafers. Film thicknesses were measured later using the nanospec.

Experimental Results

We attempted twelve deposition experiments - one for each combination of the four wafer temperatures (650, 700, 725, 750 C) and three silane flow rates (30, 50, 70 sccm). Each of the twelve depositions was performed on a different wafer. At 650 C, there was no appreciable deposition for any of the flow rates. Thus, these three wafers provided no data for analysis. At 700 C and above, enough silicon was deposited so that measurements could be taken. Thickness was measured using the nanospec at five different points on the wafer surface (see [117] for the raw data). In the case where temperature was 700 C and silane flow rate was 30 sccm, deposited film thickness was less than 100 Angstroms over the five minute period, the minimum readable by the nanospec. Hence, growth rate was recorded as less than 20 A/min. The data is presented in Table 5.1.

A model that is useful for describing deposition kinetics in the thermally acti-

Poly-silicon Deposition Rate
As Function Of Temperature and Silane Flow Rate

Process Conditions	
Chamber Pressure	20 Torr
Carrier Gas	20 slm H_2
Source Gas	2% SiH_4 in H_2
Purge Gas	7 slm H_2

Growth Rate (A/min)

Wafer Temperature	Silane Flow Rate 30 sccm	50 sccm	70 sccm
700 C	< 20.00	65.68	80.08
725 C	73.12	106.60	138.72
750 C	118.28	171.32	216.68

Table 5.1: Measured deposition rate (Angstroms per minute): Five minute deposition; three wafer temperatures; three silane flow rates.

vated regime is the Arrhenius relationship

$$R_{\mathrm{Si}} = k_0 \exp\left(\frac{-E_a}{R_g\, T_w}\right) X_{\mathrm{SiH_4}} \tag{5.1}$$

where R_{Si} denotes deposition rate, k_0 denotes the pre-exponential constant, E_a denotes the activation energy, R_g denotes the gas constant, T_w denotes the wafer temperature, and $X_{\mathrm{SiH_4}}$ denotes the silane mole fraction. We call a plot of the logarithm of deposition rate versus inverse temperature an Arrhenius plot. The Arrhenius plots associated with the data we collected are shown in Figure 5.6.

According to equation (5.1), the slope of an Arrhenius plot gives the activation energy E_a while the intercept (along with knowledge of the silane mole fraction) gives the pre-exponential constant k_0. Computed parameters are given in Table 5.2. The activation energies calculated from the Arrhenius plots range from 1.57 eV to 1.69 eV depending on silane mole fraction. This range is very close to the activation energy of 1.82 eV determined experimentally by the manufacturer, ASM

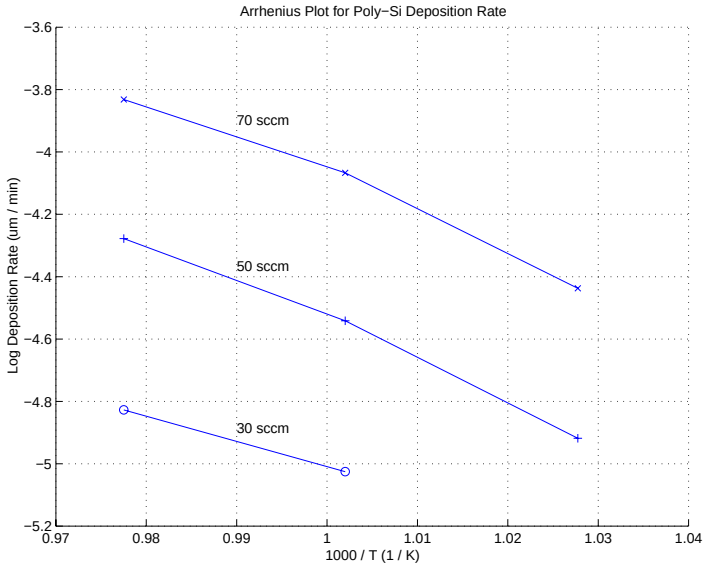


Figure 5.6: Arrhenius plots for silicon deposition from silane gas: each plot represents log of deposition rate (microns per minute) versus inverse absolute temperature for one of the three silane flow rates used.

America [105]. In addition, the pre-exponential constants range from 3.8×10^8 to 1.85×10^9, a range which includes the value of 7.9×10^8 predicted by the manufacturer.

5.4 Process-Equipment Model

This section motivates and describes the process-equipment model that we developed to predict process behavior (transient and steady-state) and product characteristics. We loosely describe the model as comprehensive because it accounts for a wide range of physical and chemical mechanisms, reactor geometry, material properties, and the effects of process conditions (pressure, temperature, flow rates, and gas composition) and equipment settings (injector sizes, thermocouple offsets). This does not imply that the model represents a complete description of process-equipment dynamics (if such a model is actually possible). Rather, we

Parameters For Arrhenius Relationship Describing Silicon Deposition Kinetics

Assumed Relationship

$$R_{Si} = k_0 \exp\left(\frac{-E_a}{R_g T_w}\right) X_{SiH_4}$$

Symbol	Description	Data		
V_{mix}	Silane/Hydrogen Mixture Flow Rate (slm)	1.5	2.5	3.5
V_{SiH_4}	Silane Flow Rate (sccm)	30	50	70
X_{SiH_4}	Silane Mole Fraction ($\times 10^{-3}$)	1.4	2.2	3.0
E_a	Activation Energy (eV)	1.69	1.67	1.57
	Activation Energy (J/mol) ($\times 10^5$)	1.63	1.61	1.51
E_a/R_g	Ratio (K) ($\times 10^4$)	1.96	1.94	1.82
k_0	Pre-exponential Constant (um/min) ($\times 10^9$)	1.85	1.30	0.38
k_0	Pre-exponential Constant (cm/sec) ($\times 10^3$)	3.08	2.16	6.54

Table 5.2: Parameters calculated by fitting experimental data to an assumed Arrhenius relationship for poly-Si growth rate as a function of temperature.

made choices so that the importance of a particular effect would be reflected in the model fidelity.

5.4.1 Modeling Approach

Models for silicon growth that cannot be coupled to gas phase transport phenomena and that use a simplified chemical kinetics model are inadequate for describing the essential physics and chemistry. For example, initial models for silicon growth in the Epsilon-1 presented in [117] considered the process-equipment state to be completely determined by a 1-dimensional (radial) wafer temperature profile. Growth rate was related to wafer temperature by a single nonlinear Arrhenius law. As stated earlier, this approach appears often in the literature (motivated by temperature control problems), but is inadequate for our purposes here.

Models incorporating more complete descriptions of transport phenomena, chemical mechanisms, couplings, 2-dimensional or 3-dimensional spatial effects, and non-symmetric geometries have been appearing recently in the literature. Authors have approached the modeling problem based on their specific objectives, process, and equipment, resulting in models with varying levels of detail and breadth of scope.

In [98] a dynamic simulator is presented which predicts the time–dependent behavior of equipment, process, sensors, and control systems for RTCVD of poly-Si from silane. This simulator is comprehensive in the respect that it provides the capability to predict aggregate values for deposition rate, film thickness, temperature, and gas flow, as well as cycle time, consumables volume, and reactant utilization. However, prediction of deposition uniformity requires high spatial resolution, rather than aggregate quantities, so their approach is not suitable here.

Axisymmetric cylindrical vertically oriented reactors are considered in [24, 49]. They incorporate coupled effects of 2-dimensional gas flow, mass transport, and heat transfer effects. In addition, [49] includes thermal diffusion and the effect of susceptor rotation on the gas flow. These models consider a relatively broad scope of effects for reactors with simple geometries that can be analyzed and simulated at high resolution in two spatial dimensions. They do not include models of chemical mechanisms for growth.

Models for reactors with non-symmetric geometries that require consideration of 3-dimensional effects are scarce. This is mainly due to the significant additional complexity of equations, boundary conditions, and solution techniques, along with burdensome computational demands. One strategy is to restrict the effort to one particular effect of interest. Two such models for commercial RTCVD chambers,

that are limited to heat transfer only, including complicated surface-to-surface radiation, are presented in [78, 85].

Another strategy is to use commercially available general purpose computational fluid dynamics (CFD) codes and software packages. These packages provide the necessary tools for modeling of transport phenomena coupled with some chemical mechanisms, including efficient numerical integration schemes and 3-dimensional grid generation for irregular geometries. The CFD approach provides a comprehensive and general process-equipment state description.

However, there are drawbacks to using general purpose CFD packages. There is an interface layer in the software that separates the user from the underlying computer code and variables. This is advantageous for setting up problems but makes it difficult to use CFD code in control loops or other specialized applications. The general purpose nature of the software results in some built-in limitations to the level of accuracy and detail that can be achieved in modeling specific aspects of a particular piece of equipment. It is unclear how to channel computational resources to areas in accordance with their importance, or to deal efficiently with phenomena that occur at vastly different spatial and temporal scales. CVD applications present special challenges, including modeling for transport of mass and momentum in a multicomponent gas mixture, heat radiation with spectral dependence, and surface chemistry.

Some of the above problems related to CVD applications were addressed by the ESPRIT ACCESS-CVD project funded by the European Commission to develop and implement a CFD code specifically designed for use in modeling CVD processes. The project resulted in a commercial code, PHOENICS-CVD, which makes it practical to include many of the important effects associated with CVD

processes. It consists of coupled dynamic sub-models for fluid flow, heat transfer, and multicomponent species transport in the gas phase, integrated with a model for conjugate heat transfer among lamps and other solid surfaces, databases and models for gas phase and surface chemistry for a large number of reactions, and databases and models for determining the time-varying, parameter dependent transport, thermodynamic, and optical properties of the involved materials.

The PHOENICS-CVD software was used to model a variety of CVD reactors in a semiconductor development line for 0.3 μm CMOS devices as presented in [163]. Most importantly for our purposes, the authors demonstrated the capability of PHOENICS-CVD as a tool for investigating uniformity issues in reactors with non-symmetric geometries.

Given manufacturing objectives, and in light of the complicated geometry and operation of the Epsilon-1 reactor, we implemented Epsilon-1 reactor simulations using PHOENICS-CVD. Figure 5.7 shows a general overview of the modeling framework. For a detailed exposition on the various aspects of this type of model see [82]. The idea is to produce a model that predicts the behavior of the process chamber block shown previously in Figure 5.4. Process recipe inputs and equipment settings enter the model via material parameters, boundary conditions on transport variables, and geometric construction of the solution grid. We note that even using the powerful PHOENICS-CVD tool, high-fidelity models that include most or all of the desired features previously described is an immensely time consuming undertaking. For this reason, various simplifications are still employed, which are described in the sequel as they are encountered.

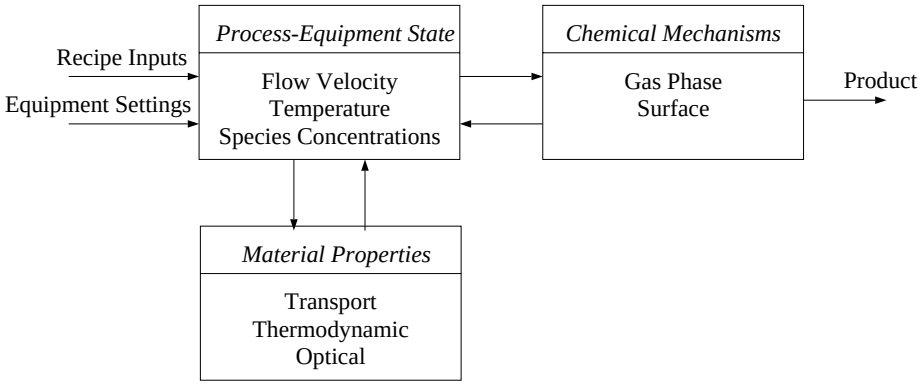


Figure 5.7: Overview of modeling framework. Process-equipment state components (gas flow, heat transfer, species transport) are coupled to each other, material properties, and chemical mechanisms.

5.4.2 Process-Equipment State

It has become apparent that spatial uniformity of deposition rate and film thickness is influenced by several variables, not limited to wafer temperature, even for thermally driven processes. These variables are included in what we refer to as the *process-equipment state*, which is the time-varying spatial distribution of flow velocity, temperature, and species concentrations throughout relevant portions of the reactor. Table 5.3 lists the essential variables. The time evolution and steady behavior of the process-equipment state is determined by the physical and chemical mechanisms of the CVD process, reactor geometry, material properties, recipe inputs, and equipment settings. The components of the process-equipment state interact with each other, the recipe inputs, and the equipment settings in a complex manner.

The process-equipment state is manifested in certain macroscopic phenomena that we believe have a significant influence on deposition uniformity. This is mainly due to the fact that they contribute to non-uniformity of reactant concentration profiles at or near the wafer surface. We wish to study these phenomena using the

Essential Variables Comprising Process-Equipment State

Symbol	Description
Gas Phase Transport Variables	
$\underline{v}$	Velocity Vector of Gas Mixture
P	Pressure
T_g	Temperature of Gas Mixture
ω_i	Mass Fraction of i-th Species in Gas Mixture
$\underline{j}_i$	Total Diffusive Mass Flux Vector of i-th Species in Gas Mixture
Gas Phase Thermal Properties	
ρ	Density of Gas Mixture
c_p	Specific Heat Capacity of Gas Mixture
k_c	Thermal Conductivity of Gas Mixture
Conjugate Heat Transfer Variables	
T_w	Temperature of Wafer
T_{wall}	Temperature of Chamber Wall
Thermal Properties of Solids	
ρ_w	Density of Wafer
c_{p_w}	Specific Heat Capacity of Wafer
k_w	Thermal Conductivity of Wafer

Table 5.3: Variables and material parameters comprising the process-equipment state. Dependencies on space and time have been suppressed.

process-equipment model. Here we describe some of the important effects that we focused on in the modeling effort.

Reactant Depletion

As precursor gases flow across the wafer surface, reactants are deposited, causing a gradual downstream reduction in their gas phase concentration. Thus, downstream portions of the wafer may be subject to lower concentrations of impinging reactants, and hence the growth rate may be lower there. The magnitude of the depletion effect varies depending upon process conditions. The degree to which wafer rotation compensates for reactant depletion is not accurately known.

Nonuniform Gas Heating and Gas Phase Reactions

Based on experimental data, gas phase reactions appear to be important in CVD processes, except for those under very low pressure (see [82] pp. 134). For example, at atmospheric pressure, growth rate of silicon from silane is strongly influenced by dissociative deposition of intermediate species formed in the gas phase. Furthermore, gas phase reaction rates can be strongly dependent on temperature. Typically, the gases heat up as they pass over the susceptor and wafer in the process chamber. This may cause a gradual downstream increase in gas phase reaction rates. Thus, downstream portions of the wafer may be subject to higher concentrations of impinging reactants. The overall effect depends on the gas composition and process conditions.

Thermal Diffusion of Species

The gas species in an initially homogeneous gas mixture will separate under the influence of a temperature gradient (see [82] pp. 110). Large, heavy molecules (e.g., silane) diffuse toward colder regions, whereas small, light molecules (e.g., hydrogen) diffuse toward hotter regions. Usually, the effect is small compared with ordinary concentration driven diffusion. However, due to the large thermal gradients in the cold-wall Epsilon-1 (e.g., 300 C difference between wafer and walls), thermal diffusion may have a significant effect. Thus, reactant concentration may be higher where the gas is cooler, e.g., upstream or in the lower section of the chamber. Reduction in growth rate by 20% to 30% caused by thermal diffusion has been observed in RTCVD chambers (see [82] pp. 164). Thermal diffusion is sometimes referred to as the Soret effect.

Flow Patterns

Calculations in [117] indicate a Reynolds number of approximately 27 for gas flow in the Epsilon-1. Thus, the flow is laminar, except possibly in and very close to the injector nozzles. Nevertheless, the flow may have some interesting characteristics that have an impact on deposition uniformity. Recirculation cells due to buoyancy effects are believed to occur in virtually all rapid thermal CVD (RTCVD) chambers due to the large thermal gradients present (see [23] pp. 339). Furthermore, three types of natural convection rolls are typically observed in horizontal CVD chambers: steady longitudinal, unsteady transversal, and steady transversal at the leading edge of the heated susceptor (see [82] pp. 162).

Remark 5.4.1 *For each of the above effects, the relationship between it, the process-equipment state, recipe inputs, equipment settings, and thickness uniformity is not*

well understood. Moreover, it is not well understood how to compensate for non-uniformity in species concentrations caused by these effects. The setting of gas injector opening sizes and thermocouple offsets to minimize these effects and to produce uniform thickness is done iteratively, usually requiring approximately five test recipes. This modeling effort is a first step toward understanding these relationships and developing a model-based systematic compensation method. □

5.4.3 Reactor Geometry and Finite Volume Mesh

The non-symmetric geometry of the Epsilon-1 necessitates genuine 3-dimensional modeling of transport phenomena in the process chamber. We adopt a Cartesian (x-y-z) coordinate system, since the lenticular chamber can be modeled roughly as long thin box with polygonal or curved sides.

We refer to the direction of flow from front to rear as the z direction, the bottom to top direction (perpendicular to the wafer) as the y direction, and the left side to right side direction (looking in through front) as the x direction. These coordinates are natural and convenient for chamber modeling but not for modeling the cylindrical wafer and susceptor, whose geometries must then be approximated.

PHOENICS uses a finite-volume mesh as the discretization of the spatial domain. Figure 5.8 shows a view of the overall mesh we developed for modeling the Epsilon-1 process chamber. The mesh is body-fitted and has dimensions of 25 by 27 by 52 volume elements in the x, y, and z directions, respectively.

The Epsilon-1 apparatus set-up and gas flows are not symmetric in the y- and z-directions. Although the exterior geometry appears to be y-symmetric, there are significant differences between upper and lower chamber sections. The chamber does have x-symmetry, with the center y-z plane serving as a symmetry plane

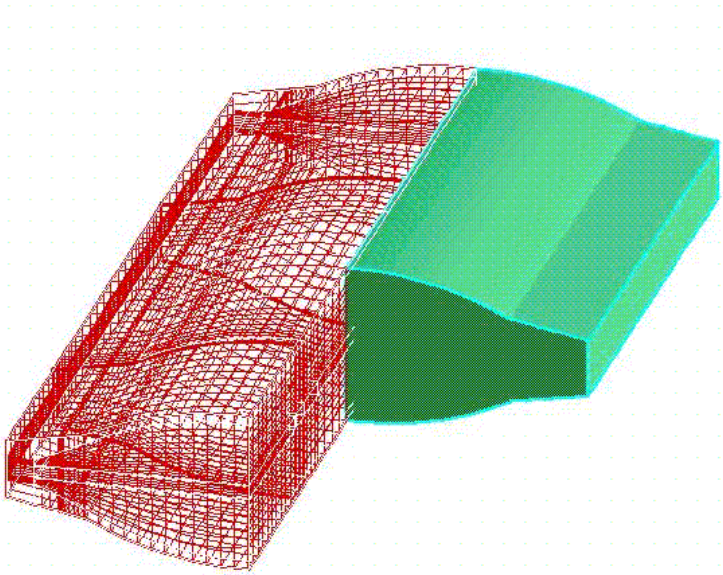


Figure 5.8: Overall body-fitted $25 \times 27 \times 52$ finite volume mesh for modeling the Epsilon-1 lenticular chamber. Solid cut-away figure at right is a viewing aide to show the full geometry of the chamber, but not part of the mesh. Inlet side faces viewer.

about which the geometry and values for all variables are mirrored exactly. The shaded portion on the right side of Figure 5.8 is not part of the actual mesh, but rather a viewing aide to show a portion of the overall chamber geometry. Only the left half of the chamber is modeled.

Figure 5.9 shows a top view of the x-z mid-plane level with the wafer surface. The surface geometry of the wafer, susceptor, and ring have been approximated by rectangular sections. It is possible to approximate the curved surfaces more accurately, either with additional rectangular volume elements arranged appropriately, or with irregularly shaped volume elements. However, irregularly shaped volume elements caused computational difficulties, and construction of the disk shape from regular elements required a large number of additional mesh elements in areas that

were not of particular interest. These drawbacks offset any advantages gained from improving the geometrical accuracy of the wafer.

Figure 5.10 shows a side view of the center y-z plane. The upper and lower chamber sections can be identified, respectively, above and below the wafer. Gases can flow and species can diffuse between the upper and lower chamber sections through thin gaps between the quartz shelf and the ring, and between the ring and susceptor. We model only the shelf-ring gap since it is significantly wider than the ring-susceptor gap, and assume that it accounts for all interaction between upper and lower sections.

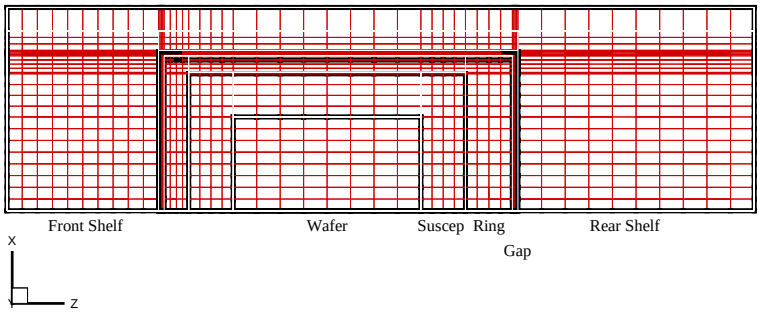


Figure 5.9: Top view of finite volume mesh at x-z mid-plane level with wafer surface.

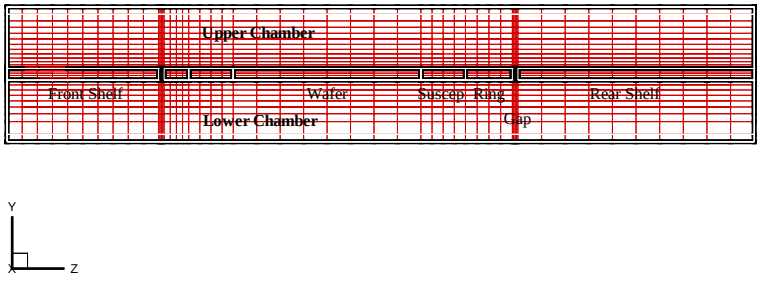


Figure 5.10: Side view of finite volume mesh at y-z mid-plane which serves as a symmetry plane.

In addition to the chamber model, we have also developed a simplified 3-

dimensional finite-volume mesh for the inlet flange. This mesh uses three thin gaps in a solid surface to model the three injector slits. The gas mixture flows vertically downward from an inlet opening through the slits until it reaches the bottom surface of the inlet flange, at which point it is forced to make a perpendicular change of direction toward the chamber entrance. The idea is to simulate the effects that the injector slits and the injector flange geometry has on the flow. For example, we are interested in seeing how the forced change of direction creates possible swirling effects, and how varying injector slit widths affects the flow profile as gases enter and flow through the chamber. This model is separate from the chamber model, which assumes a uniform flow profile at the inlet to the chamber.

5.4.4 Transport Phenomena

The process-equipment state is determined by the transport of mass, momentum, and heat energy in the process and purge gases, and heat energy in and among the solids that comprise the chamber walls, shelves, ring, susceptor, and wafer. The various effects are coupled through transport equations, state dependent material parameters, and boundary conditions. We provide here a brief overview of the main assumptions and equations used in the PHOENICS-CVD transport models and that we used in particular for modeling silicon growth in the Epsilon-1. Further details can be found in [82, 83].

Assumptions

The basic assumptions regarding the gas mixture are that it behaves as a continuum, is an ideal gas, and is transparent to infrared heat radiation. In addition, the flow is assumed to be laminar and the effects of viscous heating and pressure varia-

tions on the gas temperature is neglected. These assumptions are widely applicable to CVD systems and in particular are not limiting for modeling the Epsilon-1.

We also made assumptions regarding boundary conditions that are specific to modeling the Epsilon-1. The flow profile at the entrance to the chamber is assumed to be a uniform flow velocity in the direction normal to the entrance. All solids in the chamber are considered isothermal, i.e., constant temperature within each individual piece of apparatus and throughout the entire wafer. Chamber walls are assumed to be no-slip and stationary, even though in reality there are moving parts in the process chamber. We assume that the top surface of the wafer is the only surface on which chemical reactions occur. These assumptions can limit the scope of the predictive capability of the model. However, we believe that they do not seriously degrade model fidelity regarding prediction of steady-state phenomena, so long as they are accounted for in any investigation of the factors that influence uniformity. The assumptions are discussed further in Section 5.4.6.

Gas Phase Transport

We now give the basic transport equations for an N-component reacting gas mixture with K gas phase reactions. Gas flow in the reactor is governed by the familiar conservation equations for mass and momentum, i.e., the continuity equation

$$\frac{\partial \rho}{\partial t} = \nabla \cdot (\rho \underline{v}) \tag{5.2}$$

and the Navier-Stokes equation

$$\underbrace{\frac{\partial(\rho \underline{v})}{\partial t}}_{\text{transient}} = \underbrace{-\nabla \cdot (\rho \underline{v}\,\underline{v})}_{\text{inertial}} + \underbrace{\nabla \cdot \underline{\underline{\tau}}}_{\text{viscous}} - \underbrace{\nabla P}_{\text{pressure}} + \underbrace{\rho \underline{g}}_{\text{gravity}} \tag{5.3}$$

where ρ is the gas density, $\underline{v}$ is the gas velocity, P is the pressure, and $\underline{g}$ is gravity. The viscous stress tensor $\underline{\underline{\tau}}$ for a Newtonian fluid such as the gas mixture in a CVD

reactor takes the form

$$\underline{\underline{\tau}} = \mu \left(\nabla \underline{v} + (\nabla \underline{v})^{\mathsf{T}} \right) - \frac{2}{3} \mu \, (\nabla \cdot \underline{v}) \cdot \underline{\underline{\mathbb{I}}} \tag{5.4}$$

where μ is the dynamic viscosity of the gas. For CVD applications, the density ρ and viscosity μ are strongly dependent on the temperature, pressure, and mixture composition. For this reason, the gas flow equations are strongly coupled to the equations for transport of heat energy and species concentrations. In particular, temperature and concentration gradients cause variations in gas mixture density which are manifested in buoyancy effects.

Transport of heat energy in the reactor is governed by the familiar heat equation, with additional terms to account for effects that occur in chemically reacting multicomponent gases. In particular, heat is generated and consumed by the inter-diffusion of different species and by the various gas phase chemical reactions. Also, heat can flow due to the presence of a concentration gradient, which is referred to as the Dufour effect. The conservation equation for gas temperature is given by

$$\begin{aligned} \underbrace{c_p \frac{\partial(\rho T_g)}{\partial t}}_{\text{transient}} \;=\; & \underbrace{\nabla \cdot (k_c \nabla T_g)}_{\text{conduction}} + \underbrace{c_p \nabla \cdot (\rho \underline{v} T_g)}_{\text{convection}} + \underbrace{\nabla \cdot \left(R_g T_g \sum_{i=1}^{N} \frac{D_i^T}{m_i} \nabla \, (\ln f_i) \right)}_{\text{Dufour}} + \\ & \underbrace{\sum_{i=1}^{N} \frac{H_i}{m_i} \nabla \cdot \underline{j}_i}_{\text{inter-diffusion}} - \underbrace{\sum_{i=1}^{N} \sum_{k=1}^{K} H_i \, \nu_{ik} \left(R_k^g - R_{-k}^g \right)}_{\text{reactions}} \end{aligned} \tag{5.5}$$

where c_p is the specific heat capacity per unit mass of the gas, T_g is the gas temperature, k_c is the gas thermal conductivity, and R_g is the universal gas constant. Associated with the i-th gas species is the mole fraction f_i, molar mass m_i, thermal diffusion coefficient D_i^T, molar enthalpy H_i, and total diffusive mass flux $\underline{j}_i$. The stoichiometric coefficient of the i-th species in the k-th gas phase reaction is denoted ν_{ik} with forward reaction rate R_k^g and reverse reaction rate R_{-k}^g.

PHOENICS-CVD ignores the Dufour effect since it has been found to be very small in CVD systems. The density, viscosity, thermal conductivity, specific heat capacity, and thermal diffusion coefficient are dependent on temperature and gas mixture composition. For this reason, the heat transfer equation is strongly coupled to the gas flow and species concentration equations.

Gas species transport in the reactor is governed by a familiar diffusion-convection equation with an additional source term to account for the creation and destruction of species due to K reversible chemical reactions. The balance equation for the concentration of the i-th gas species is given by

$$\underbrace{\frac{\partial(\rho\,\omega_i)}{\partial t}}_{\text{transient}} = \underbrace{-\nabla\cdot(\rho\,\underline{v}\,\omega_i)}_{\text{convection}} - \underbrace{\nabla\cdot\underline{j}_i}_{\text{diffusion}} + \underbrace{m_i\sum_{k=1}^{K}\nu_{ik}\left(R_k^g - R_{-k}^g\right)}_{\text{reactions}}. \tag{5.6}$$

In the above, the concentration of the i-th gas species is a dimensionless mass fraction

$$\omega_i = \frac{\rho_i}{\rho} \tag{5.7}$$

There are $N-1$ independent species concentration equations of the form (5.6) since the mass fractions must sum to 1, i.e.,

$$\sum_{i=1}^{N}\omega_i = 1 \tag{5.8}$$

The diffusive mass fluxes are defined by

$$\underline{j}_i = \rho\,\omega_i\,(\underline{v}_i - \underline{v}) \tag{5.9}$$

with respect to the mass averaged velocity

$$\underline{v} = \sum_{i=1}^{N}\omega_i\underline{v}_i \tag{5.10}$$

and satisfy

$$\sum_{i=1}^{N}\underline{j}_i = 0 \tag{5.11}$$

again leaving $N-1$ independent variables.

Gas species diffusion is caused by concentration gradients, which we refer to as ordinary diffusion, and by temperature gradients, which we refer to as thermal diffusion, or the Soret effect. It is expressed as the sum of these two components

$$\underline{j}_i = \underline{j}_i^{\mathrm{C}} + \underline{j}_i^{\mathrm{T}} \tag{5.12}$$

where $\underline{j}_i^{\mathrm{C}}$ and $\underline{j}_i^{\mathrm{T}}$ denote the concentration driven and thermally driven diffusive fluxes, respectively. The ordinary diffusive mass fluxes can be computed via Fick's law, the Wilke approximation, or the full Stefan-Maxwell equations, depending upon the properties of the gas mixture, the desired degree of fidelity, and the available computational resources. The Stefan-Maxwell formulation is given by

$$\nabla \omega_i + \omega_i \nabla \left(\ln m\right) = \frac{m}{\rho} \sum_{j=1}^{N} \frac{1}{m_j D_{ij}} \left(\omega_i \underline{j}_j^{C} - \omega_j \underline{j}_i^{\mathrm{C}}\right) \tag{5.13}$$

with m the average mole mass of the mixture. The diffusive mass fluxes due to thermal diffusion are given by

$$\underline{j}_i^{\mathrm{T}} = -D_i^T \nabla \left(\ln T_g\right) \tag{5.14}$$

where the thermal diffusion coefficient D_i^T for each species is a function of temperature and gas mixture composition. In general $D_i^T > 0$ for large, heavy molecules and $D_i^T < 0$ for small, light molecules, resulting in the observed separation of species due to thermal gradients.

The last term in the species concentration equation (5.6) represents the creation and destruction of the i-th species due to homogeneous gas phase reactions. The forward and reverse reaction rates are given by

$$R_k^g = k_g(P,T) \prod_{\text{reactants}} \left(\frac{P\, f_i}{R_g\, T}\right)^{|\nu_{ik}|} \tag{5.15}$$

$$R_{-k}^g = k_{-g}(P,T) \prod_{\text{products}} \left(\frac{P\, f_i}{R_g\, T}\right)^{|\nu_{ik}|} \tag{5.16}$$

where k_g and k_{-g} are the forward and reverse reaction rate constants, and P the total pressure.

Boundary Conditions

For each of the gas phase transport equations there is an associated set of boundary conditions, which prescribe the state (or associated flux) at the inlet, outlet, chamber walls, and chamber apparatus including susceptor and wafer. The boundary conditions for temperature and species concentrations are responsible for coupling the gas phase transport phenomena to heat transfer in the solids and wafer surface chemical reactions, respectively. We elaborate further below.

For each inlet boundary (process and purge), we prescribe the inflow velocity of the gas mixture normal to the inflow opening, and the mass fraction for each of the gaseous species (e.g., silane and hydrogen). The values are set according to the process recipe we wish to simulate. The temperature of the gas mixture at the inlet is set to room temperature. There is no species diffusion through the inlet. These conditions are given by

$$\underline{n} \cdot \underline{v} = v_{\text{in}} \,, \quad \underline{n} \times \underline{v} = 0 \,, \quad T = T_{\text{room}} \,, \quad \omega_i = \omega_{i,\text{in}} \,, \quad \underline{n} \cdot \underline{j}_i = 0 \tag{5.17}$$

where n is the unit vector normal to the inlet opening.

For the outlet boundary, we impose zero gradient conditions for all variables. These conditions are given by

$$\underline{n} \cdot (\nabla \, (\rho \, \underline{v})) = 0 \,, \quad \underline{n} \times \underline{v} = 0 \,, \quad \underline{n} \cdot (k_c \, \nabla \, T_g) = 0 \,, \quad \underline{n} \cdot \underline{j}_i = 0 \tag{5.18}$$

where n is the unit vector normal to the outlet opening.

Boundary conditions at the solid-gas interfaces can be more complicated, mainly due to chemically reacting surfaces and heat transfer in the solids. First we consider

non-reacting surfaces, e.g., chamber walls, quartz shelves, ring, and susceptor. At these surfaces, the no-slip and impermeability conditions apply, i.e., flow velocities are set to zero. Also, the total mass flux normal to each non-reacting surface must be zero for each of the species. Note, however, that due to thermal diffusion, the concentration gradients normal to the surface will generally not be zero. These conditions are given by

$$\underline{v} = 0\,, \quad \underline{n} \cdot \underline{j}_i = 0 \tag{5.19}$$

where n is the unit vector normal to the non-reacting surface.

For a reacting surface, which in our case refers only to the top side of the wafer, there is a net mass production rate for each gaseous species. The velocity component normal to the surface is proportional to this rate, while the tangential component is zero. Furthermore, the total mass flux normal to the reacting surface is set equal to the production rate. These conditions for a process with L surface reactions are given by

$$\underline{n} \cdot \underline{v} = \frac{1}{\rho} \sum_{i=1}^{N} m_i \sum_{l=1}^{L} \sigma_{il}\, R_l^s\,, \quad \underline{n} \times \underline{v} = 0\,, \quad \underline{n} \cdot \left(\rho\, \omega_i\, \underline{v} + \underline{j}_i\right) = m_i \sum_{l=1}^{L} \sigma_{il}\, R_l^s \tag{5.20}$$

where n is the unit vector normal to the reacting surface, σ_{il} is the stoichiometric coefficient for the i-th gas species in the l-th surface reaction, and R_l^s is the reaction rate for the l-th surface reaction. The surface reaction rate is equal to the product of the collision rate of molecules with the wafer surface and the reaction probability, called the reactive sticking coefficient (RSC).

Thermal boundary conditions at the solid-gas interfaces can be complex due to heat transfer within and among the various solids in the reactor. This includes the effects of conduction within the solids, convective losses to the gas phase, and radiative transfer among the various surfaces. Heat radiation supplied by the lamps is especially important.

PHOENICS-CVD provides the capability for modeling heat transfer in the solids and coupling these effects to the gas phase transport phenomena via boundary conditions. Surface-to-surface radiation is modeled using viewfactor methods. However, due to the extremely complicated geometry of the Epsilon-1 lamp-house and reflector apparatus, and the very large number of solid surfaces with varying optical properties in the process chamber, we found the PHOENICS-CVD radiation modeling tool to be impractical for our purposes. This is discussed further in Section 5.4.6.

Instead of modeling heat transfer in the solids, we assumed that the wafer and susceptor were at a constant uniform temperature, and used anecdotal and experimental data from the manufacturer to estimate the temperature on other surfaces. The boundary conditions are given by

$$T_g = T_{\text{surf}} \tag{5.21}$$

for the case where the gas-solid interface is a isothermal surface and

$$\underline{n} \cdot \nabla T_g = 0 \tag{5.22}$$

when there is an adiabatic surface.

Specific values for the boundary conditions were set according to the process recipes that we were simulating. Values are provided as simulations are described in Section 5.5.

Material Properties

Models for describing the dependence of material properties on the process-equipment state are presented in [82, 83] and included in the PHOENICS-CVD software. Furthermore, PHOENICS-CVD provides databases containing any

necessary parameters for determining the transport, thermodynamic, and optical properties of most materials commonly used in CVD processes.

Transport properties of the gases include viscosity, thermal conductivity, and ordinary and thermal diffusion coefficients. Their dependence on temperature, pressure, and gas mixture composition is determined using the Lennard-Jones potential and kinetic theory. Lennard-Jones parameters for the individual gases are provided in a database. Properties of the gas mixture are calculated from the individual gas properties. For example, a semi-empirical relationship is employed for determining mixture viscosity.

Thermodynamic properties of the gases include specific heat capacity, standard heat of formation, and standard entropy. These properties are given as functions of temperature via polynomial approximations, with a different polynomial for each of three temperature ranges. Polynomial coefficients for individual gases are provided in a database. Again, properties of the gas mixture are calculated from the individual gas properties. For example, density is defined in terms of the mean molecular mass and specific heat is defined as the mass averaged value.

Optical properties of the solids include refractive indices and absorption coefficients. The temperature dependence of these properties in each of 60 spectral intervals is provided in a database. However, as stated earlier, we did not use the PHOENICS-CVD surface-to-surface radiation model, so the optical properties of the solids, e.g., quartz, do not play a role in our simulations.

5.4.5 Chemical Mechanisms for Growth

In Section 5.3 we presented the results of growth experiments and showed that, for a range of operating conditions, a simple Arrhenius law given by

$$R_{\mathrm{Si}} = k_0 \exp\left(\frac{-E_a}{R_g\, T_w}\right) X_{\mathrm{SiH_4}} \tag{5.23}$$

provides an accurate model for predicting growth rate as a function of wafer temperature. Parameters such as activation energy were calculated by fitting the experimental data to the model.

There are several assumptions and simplifications, both explicit and implicit, in the above Arrhenius model. It assumes that silicon growth is almost completely due to the heterogeneous decomposition of silane into silicon and hydrogen on the wafer surface. Thus, it models only a single surface reaction step. Furthermore, it is implicitly assumed that inlet conditions for silane mole fraction hold constant throughout the process chamber. This allows for a separation of the factor multiplying the exponential term into a mole fraction variable and a pre-exponential constant. The result is that growth rate is assumed to be dependent entirely on two process recipe inputs: wafer temperature set-point and silane mole fraction at the inlet; and two process dependent physical-chemical parameters: activation energy and pre-exponential constant.

The above approach takes the view that surface reactions are dominant and gas phase reactions are negligible. However, it was demonstrated by Coltrin and coworkers [31] that as chamber pressure increases, gas phase reactions play a greater role. They showed that at atmospheric pressure, silicon growth may be almost completely due to reactive intermediaries formed in the gas phase.

Kleijn develops a model for gas phase and surface chemistry in [81] for temperatures and pressures in an intermediate range, near the conditions at which

NG-ESSS deposits silicon in the Epsilon-1. It is a relatively closed subsystem of the full kinetic model that was used by Coltrin and co-workers. The key reaction is the homogeneous decomposition of silane which leads to the formation of silylene (SiH_2), and hydrogen. Further reactions produce disilane (Si_2H_6), trisilane (Si_3H_8), and silylsylene (Si_2H_4). The five step gas phase reaction mechanism is given by

$$SiH_4 \overset{\leftarrow}{\rightarrow} SiH_2 + H_2 \tag{5.24}$$

$$Si_2H_6 \overset{\leftarrow}{\rightarrow} SiH_4 + SiH_2 \tag{5.25}$$

$$Si_3H_8 \overset{\leftarrow}{\rightarrow} Si_2H_6 + SiH_2 \tag{5.26}$$

$$Si_2H_4 \overset{\leftarrow}{\rightarrow} SiH_2 + SiH_2 \tag{5.27}$$

$$Si_2H_6 \overset{\leftarrow}{\rightarrow} Si_2H_4 + H_2 \tag{5.28}$$

each of which has an associated forward and reverse reaction rate constant, respectively, k_g and k_{-g}. For silicon growth at the wafer surface from silane and the reactive intermediaries, Kleijn uses a set of five surface reactions given by

$$SiH_4(g) \rightarrow Si(s) + 2H_2(g) \tag{5.29}$$

$$SiH_2(g) \rightarrow Si(s) + H_2(g) \tag{5.30}$$

$$Si_2H_6(g) \rightarrow 2Si(s) + 3H_2(g) \tag{5.31}$$

$$Si_3H_8(g) \rightarrow 3Si(s) + 4H_2(g) \tag{5.32}$$

$$Si_2H_4(g) \rightarrow 2Si(s) + 2H_2(g) \tag{5.33}$$

each of which has an associated RSC. We refer to the reaction schemes (5.24)-(5.28) together with (5.29)-(5.33) as the *Kleijn model* for poly-Si deposition.

The reaction rate constants and RSCs in the Kleijn model are derived from studies by various investigators. The gas phase forward and reverse reaction rate

constants are, in general, temperature and pressure dependent, given by the expression

$$k_g = A\,P^{\kappa} \exp\left(\frac{-E_a}{R_g\,T_g}\right) \tag{5.34}$$

where parameters A and E_a were fitted to experimental data for temperatures from 300-1100 K and pressures from 10–100 Torr. Furthermore, each RSC is given by a different complicated function of wafer and gas temperature. Thus, the reaction scheme includes a complicated temperature dependence and is a function of a large number of physical-chemical parameters, e.g., multiple activation energies and multiple sticking coefficients.

PHOENICS-CVD provides the necessary tools to implement the Kleijn model, including a database of experimentally determined kinetics parameters. Thus, we used the Kleijn model to describe poly-Si chemical reaction kinetics in the Epsilon-1. In contrast to the initial simplified Arrhenius models, the chemistry model is coupled to the gas phase transport model, since gas phase reactions play an important role. Furthermore, no assumptions are made regarding the spatial distribution of temperature and reactant concentrations in the chamber.

5.4.6 Unmodeled Phenomena and Equipment

As stated earlier, even with powerful tools at our disposal, development of a comprehensive model that incorporates every relevant feature of the Epsilon-1 reactor is not practical. Here, we discuss some of the unmodeled features, phenomena, and processes that are relevant to growth in the Epsilon-1 but were not implemented in our models.

Si-Ge growth chemistry

Although reaction schemes including gas phase and surface reactions for growth of Si-Ge from silane and germane precursors have appeared recently in the literature, experimentally determined physical-chemical parameter values for such schemes are proprietary information and in general not widely available. We have contacted researchers at a government laboratory [107] regarding future experiments to determine rate constants and sticking coefficients for Si-Ge growth.

Development of models for Si-Ge growth is particularly complicated due to the large number of phenomena involved and the manner in which the deposited film depends on the process-equipment state. For example, epitaxial Si-Ge layers are deposited using either dichlorosilane, silane, or disilane, along with germane. Deposition rate and germanium content have been observed to be dependent on the choice of precursor gas [74] and germane concentration [73]. Furthermore, both of these effects have been observed to be temperature dependent [38]. Thus, uniformity may be affected in different ways by non-uniform concentrations of different reactants at various temperatures.

Radiative heat transfer, lamp-house, and reflectors

Radiative heat transfer modeling is implemented in PHOENICS-CVD via view-factor methods. This requires a discretization of the solid surfaces in the chamber into a large number of smaller surfaces that are considered isothermal and of constant optical properties. The predictive capability of the method depends on the number and size of the individual surfaces, which we refer to as the discretization resolution, as well as the accuracy of the chamber geometry implemented in the model.

Although the finite volume mesh used to model the chamber geometry is effective for capturing relevant gas phase transport phenomena, it neglects various features of the equipment which would have a significant effect on radiative heat transfer. Such features include apparatus containing wafer rotation machinery within the lower chamber section and the complicated lamp-house and reflector equipment. In the Epsilon-1, there are a variety of reflector designs, including both diffuse and specular, and some of a special parabolic shape. Also, certain parameters are known only roughly, such as the power supplied by the lamps, or equivalently the temperature of the filaments when they are turned on to 100% power. It is simply not practical to include the geometry and properties of these pieces of equipment in the finite volume reactor model.

Furthermore, in the Epsilon-1, there exist many transitions from one material to another, gaps between different pieces of the equipment, and a non-symmetric lenticular shape. This necessitates an extremely high discretization resolution for viewfactor modeling. For all of the above reasons, we believe that modeling of radiative heat transfer in the Epsilon-1 using the PHOENICS-CVD framework is too unwieldy and computationally expensive to be practical.

Currently, solid surfaces are modeled as isothermal, i.e., constant temperature within each individual piece of apparatus and throughout the entire wafer. Temperature values are set according to empirical data supplied by the manufacturer. In reality, temperature gradients exist within individual pieces of equipment, and the average steady-state temperature of any part of the reactor is known only roughly. However, it is not likely that implementation of a radiative heat transfer model using the PHOENICS-CVD framework would produce more accurate surface temperatures than manufacturer supplied empirical data.

A separate model restricted only to heat transfer within and among the solids in the Epsilon-1, and implemented using a more flexible programming environment, is presented in Chapter 6. It is desirable to incorporate a more comprehensive conjugate heat transfer model to reflect the relevant geometrical features at sufficiently high resolution to provide an accurate picture of the temperature distribution in the Epsilon-1 solids. Steady-state temperature values could then be used as boundary conditions in a refined process-equipment model.

Wafer rotation

The actual mechanical rotation of the wafer is not accounted for in the models. CFD software routines for modeling of rotating objects in a flow environment are available, but require axisymmetry of the entire domain. Therefore, a separate effort would be required to investigate the effect of mechanical wafer rotation on the chamber flow immediately surrounding the wafer. The effect of wafer rotation on growth can be studied, partially, by performing averaging calculations on simulation results, i.e., averaging deposition rates around the wafer surface.

Rotation shaft purge

There is an additional purge gas inlet through the wafer rotation shaft. Purge gases injected through the rotation shaft enter the chamber directly underneath the center of the wafer. Recall that the wafer rests on small pins attached to the top of the susceptor. The purpose of the shaft purge is to prevent the source gases from flowing between the wafer and susceptor which could result in back-side deposition. Typically, the shaft purge is set to 3 slm H_2. This may have an effect on the chamber flow in the vicinity of the wafer. In addition, the presence of the

rotation shaft would affect the chamber flow in the lower chamber section. The rotation shaft and related apparatus are not included in the process-equipment model.

Deposition on chamber walls

The process-equipment model treats all surfaces other than the top surface of the wafer as non-reacting. However, deposition on other surfaces does occur in the Epsilon-1. A build-up of such films is prevented by preceding the deposition step with a HCl etch clean. In Section 5.5.4, we slightly modify the process-equipment model so that the back-side of the susceptor becomes a reacting surface. This limited study could be expanded to study deposition on other chamber surfaces as well.

Chamber wall cooling

The quartz chamber and lamp-house are cooled by air flow. There is little data regarding characteristics of the air flow and convective losses from the outer wall of the chamber. These effects are not modeled. Instead, a constant temperature is set for each of the chamber walls.

5.5 Results and Applications

In this section we present results from poly-Si growth simulations using the Epsilon-1 process-equipment model. We use the simulation results to study various characteristics of the thin films and reactor operation that impact on manufacturing effectiveness. We show that the model can be used to predict growth rate and uniformity and to better understand the factors that influence these measures of

performance. We also present simulation results that provide guidance toward improved setting of purge gas flow rates.

Because deposition times are typically much longer than initial transients in radiative heating and gas phase transport in the Epsilon-1, it is reasonable to assume that all growth occurs during steady-state operation. For this reason, all simulations described in this section predict steady-state values of growth rate and other variables.

The PHOENICS input code, called a *Q1 file*, used for simulating poly-Si growth at 750 C temperature, 30 sccm silane flow rate, and 20 Torr chamber pressure, is presented in Appendix I. The PHOENICS code for the other simulations that we conducted is similar.

5.5.1 Deposition Rate Prediction

We have performed poly-Si growth simulations using the Epsilon-1 process-equipment model to study the relationship between growth rate and the various process recipe inputs. Prediction of growth rate given process conditions is important for taking advantage of the flexibility of the Epsilon-1. The manufacturer provides some predictive guidance and data, but the process-equipment model can allow for prediction of growth rates "off-the-curve" and also provide a tool for performing multiple trial-and-error steps and for understanding the factors that influence growth rate and uniformity.

Wafer Temperature Sensitivity

Here we investigate the relationship between silicon growth rate and wafer temperature in the Epsilon-1. We have already studied this relationship in Section 5.3,

where poly-Si growth rates were measured experimentally for a range of wafer temperatures (700, 725, 750 C) and silane flow rates (30, 50, 70 sccm) with pressure fixed at 20 Torr. It was shown that for the given range of operating conditions, a simple Arrhenius law provides an accurate model for predicting growth rate as a function of wafer temperature. Arrhenius model parameters such as activation energy were then calculated to fit the experimental data.

In contrast, the process-equipment model adopts the more complicated multi-step Kleijn model for silicon deposition. This growth model includes both gas phase and surface reactions, involves multiple reactive intermediaries, and is coupled to transport models. Because of the model's complicated nature, it is reasonable to expect difficulty in isolating the effect of wafer temperature on simulated growth rate. However, using reactor simulations, we show below that, like the experimentally determined growth rates, simulated growth rates can also be fitted to a simple Arrhenius law relating growth rate to wafer temperature. Furthermore, the single activation energy in the Arrhenius law fitted to simulated growth rates is nearly the same as that which was fitted to experimentally determined growth rates. Thus, there appears to be an underlying dominant chemical mechanism that obscures the effect of gas phase phenomena.

The predictive capability of the process-equipment model was tested by simulating poly-Si growth using operating conditions for pressure, temperature, and flow rate that are duplicates of conditions used for experiments presented in [117]. The boundary conditions used for the simulations are given in Table 5.4. The simulation results are presented in Table 5.5 together with corresponding experimental results from [117].

In order to test the fit of simulated growth rates to an Arrhenius law, we plot

Boundary Conditions Used In Epsilon-1 Reactor Simulations

Process Inlet Conditions
Carrier Gas H_2
Source Gas 2% SiH_4 in H_2

Condition	Values		
Flow Rate of Carrier (slm)	20	20	20
Flow Rate of Source (slm)	1.5	2.5	3.5
Flow Rate of Silane (sccm)	30	50	70
Mole Fraction of Silane ($\times 10^{-3}$)	1.40	2.22	2.98
Molar Mass of Silane (g/gmol)	32.12	32.12	32.12
Molar Mass of Gas Mixture (g/gmol)	2.06	2.08	2.11
Mass Fraction of Silane ($\times 10^{-2}$)	2.18	3.43	4.54
Density of Gas Mixture ($kg/m^3 \times 10^{-3}$)	2.25	2.28	2.30
Velocity of Gas Mixture (m/sec)	1.40	1.47	1.53
Temperature of Gas Mixture (C)	20	20	20

Purge Inlet Conditions
Purge Gas H_2

Condition	Value
Flow Rate of Purge Gas (slm)	7
Velocity of Purge Gas (m/sec)	0.45
Temperature of Purge Gas (C)	20

Solid-Gas Interface Conditions

Condition	Values		
Temperature of Wafer (C)	700	725	750
Temperature of Susceptor (C)	700	725	750
Temperature of Ring (C)	Conductive Solid		
Temperature of Front Quartz Shelf (C)	Conductive Solid		
Temperature of Rear Quartz Shelf (C)	Conductive Solid		
Temperature of Upper Chamber Wall (C)	400	425	450
Temperature of Lower Chamber Wall (C)	400	425	450

Table 5.4: Boundary conditions for process gas inlet, purge gas inlet, and solid surfaces used in simulations of poly-Si growth in the Epsilon-1 reactor.

Process-Equipment Model Predictive Capability
Growth Rate Temperature Dependence

Process Conditions	
Chamber Pressure	20 Torr
Carrier Gas	20 slm H_2
Source Gas	2% SiH_4 in H_2
Purge Gas	7 slm H_2

Growth Rate (A/min) Vs. Wafer Temperature and Silane Flow Rate

	Experiment			**Simulation**		
Wafer	Silane Flow Rate (sccm)			Silane Flow Rate (sccm)		
Temperature (C)	30	50	70	30	50	70
700		65.68	80.08	141.00	202.60	254.90
725	73.12	106.60	138.72	233.00	336.30	423.90
750	118.28	171.32	216.68	337.30	512.50	658.50

Ratio: Simulation / Experiment

Wafer	Silane Flow Rate (sccm)		
Temperature (C)	30	50	70
700		3.08	3.18
725	3.19	3.15	3.06
750	2.85	2.99	3.04

Mean	3.07
Standard Deviation	0.11

Note: Poly-silicon thickness for 750 C and 30 sccm was too small to be measured with available equipment.

Table 5.5: Results comparing poly-Si growth experiments with simulations. Growth rates are averaged over wafer surface.

the logarithm of growth rate as a function of inverse temperature, for each of the three flow rates used. The plots are shown, along with corresponding plots of experimental data, in Figure 5.11. We can fit the simulation data to a simple Arrhenius law, where the slope of each plot is proportional to the activation energy. More importantly, the slopes of all plots are consistent over the range of flow rates, for both simulation and experimental data. This means that the activation energy for simulated growth is nearly identical to the activation energy measured for actual growth in the Epsilon-1. Calculated Arrhenius parameters for experimental and simulation data are presented in Table 5.6.

We note that the calculated activation energies fall within the range of published activation energies for deposition of silicon from silane for the given range of process conditions (see, e.g., [92]). In particular, they lie between the activation energy for silane adsorption (125 kJ/mol), which is associated with with temperatures above 700 C, and hydrogen desorption (192 kJ/mol), which is associated with temperatures below 700 C. Thus, it is likely that these are the dominant activating mechanisms for both actual and simulated growth.

However, other phenomena also play a role, resulting in the consistent upward shift from experimentally determined to simulated growth rates observed in the Arrhenius plots. By a consistent upward shift, we mean that the ratio of simulated to experimentally determined growth rates is a constant over the given range of operating conditions. We calculated this constant offset factor relating simulation and experimental data to have mean value 3.07 with standard deviation 0.11, as indicated in Table 5.5. Thus, the process-equipment model, using the Kleijn model for poly-Si growth chemistry, predicts growth rates that are roughly three times greater than actual growth rates. More importantly, this factor is constant over

the selected range of temperatures and silane flow rates.

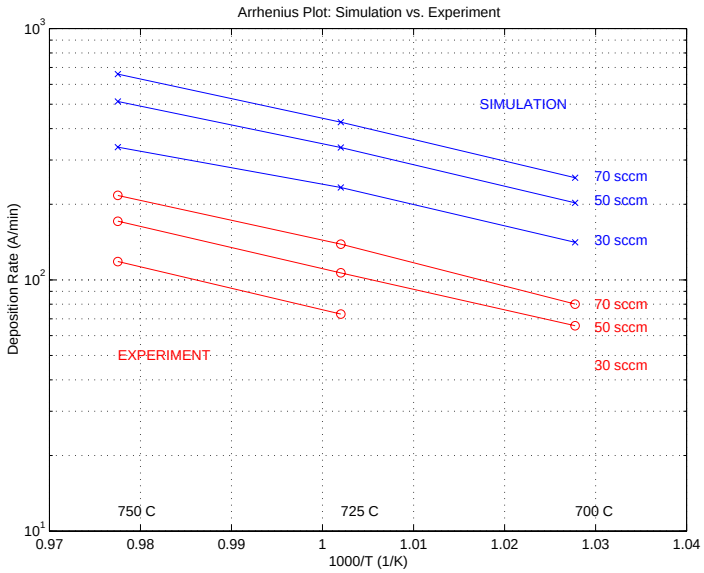


Figure 5.11: Plots illustrating Arrhenius relationship between poly-Si growth rate and wafer temperature in the Epsilon-1. Experimental and simulation data is taken for three silane flow rates (30, 50, and 70 sccm) and three temperatures (700, 725, 750 C) at 20 Torr. Simulated growth rates (top three plots) are a factor of 3.07 times greater than experimentally determined growth rates (bottom three plots) consistently over the given range of temperatures and flow rates.

We now offer some ideas toward a qualitative explanation of the presence of the offset factor. The Kleijn model is semi-empirical, i.e., it is based on phenomenological models and empirical data from growth experiments performed by various investigators. It is well known that rate constants and sticking coefficients for gas phase and surface reactions are difficult to measure, and reaction rates under nominally identical process conditions will vary among different reactors [128]. For example, Kleijn's study [81] used a cylindrical cold-wall chamber, which is very different from the lenticular hot-wall chamber in the Epsilon-1. Moreover, the process conditions used in the Kleijn study do not completely match those of our own. For example, Kleijn used pressures in the 1–10 Torr range (compared to our 20 Torr) and total flow rates on the order of 1 slm (compared to our $>$ 20 slm).

Parameters of Arrhenius Relationship Fitted to Data
Experiment vs. Simulation

Assumed Relationship
$R_{Si} = C \exp\left(\frac{-E_a}{R_g T_w}\right)$

Process Conditions	
Chamber Pressure	20 Torr
Carrier Gas	20 slm H_2
Source Gas	2% SiH_4 in H_2
Purge Gas	7 slm H_2

Symbol	Description	Experiment		
F_{SiH_4}	Silane Flow Rate (sccm)	30	50	70
E_a	Activation Energy (eV)	1.69	1.67	1.57
	Activation Energy (J/mol) ($\times 10^5$)	1.63	1.61	1.51
E_a/R_g	Ratio (K) ($\times 10^4$)	1.96	1.94	1.82
C	Pre-exponential Constant (A/min) ($\times 10^{10}$)	7.94	8.90	3.52
		Simulation		
F_{SiH_4}	Silane Flow Rate (sccm)	30	50	70
E_a	Activation Energy (eV)	1.30	1.48	1.55
	Activation Energy (J/mol) ($\times 10^5$)	1.26	1.43	1.49
E_a/R_g	Ratio (K) ($\times 10^4$)	1.51	1.72	1.80
C	Pre-exponential Constant (A/min) ($\times 10^{10}$)	0.08	1.01	2.81

Table 5.6: Parameters calculated by fitting experimental and simulation data for poly-Si growth rates to an assumed Arrhenius relationship.

The temperature ranges do coincide.

On the other hand, the consistency of the offset factor over a range of temperatures and silane flow rates indicates the likelihood that activation energies and sticking coefficients in the Kleijn model are close to those we would find by performing similar experiments in the Epsilon-1. It appears likely that the offset is due more to approximations and neglected effects in the process-equipment model. For example, discrepancies between actual and simulated gas phase flow, temperature, and species concentration distributions could alter the relative significance of each of the different reactive intermediaries. Thus, even if sticking coefficients for the five separate surface reactions are accurate, total growth rate would be shifted.

Beyond that, the coupling of gas phase reactions and transport phenomena with surface chemistry in the process-equipment model blurs any specific cause and effect relationships. It must also be emphasized that the model makes a number of approximations and assumptions whose cumulative effect is difficult to pinpoint. For example, the wafer geometry is approximated, so that the area of the wafer consuming reactants may not be modeled accurately.

By taking the offset factor of 3.07 into account, the model as it currently stands can be used to accurately predict silicon growth rate in the Epsilon-1 over a range of temperatures and silane flow rates typically used by NG-ESSS. However, it would be preferable to improve the transport component of the model, and to conduct a more extensive experimental study, similar to Kleijn's, in which chemical kinetics parameters for growth in the Epsilon-1 are measured over a wide range of operating conditions.

Chamber Pressure Sensitivity

Here we investigate the relationship between silicon growth rate and chamber pressure in the Epsilon-1. Actual growth rate is expected to increase as total pressure rises due to the increased number of molecular collisions on the wafer surface and the increased reaction rates of gas phase reactions. Simulation results presented below reflect this phenomenon.

We performed poly-Si growth simulations at a chamber pressure of 40 Torr using the same temperature and silane flow rate conditions as experiments and simulations performed at 20 Torr. However, due to time limitations, only two flow rates were used (50, 70 sccm). The simulation results for 20 Torr and 40 Torr are presented together in Table 5.7.

As before, we produced Arrhenius plots and calculated Arrhenius parameters in order to see how pressure affects growth rate. The results are presented in Figure 5.12 and Table 5.8. We see that growth rate increases by a factor of 1.26 as pressure increases from 20 Torr to 40 Torr. This offset factor is constant over the given range of operating conditions. This result is consistent with a study by Kleijn [81] in which growth rate increases as the logarithm (base 10) of pressure. Furthermore, activation energies for 40 Torr are slightly higher than those for 20 Torr but still within the expected range.

Flow Rate Sensitivity

We have studied the influence of silane flow rate on growth rate using the experimental and simulation data already presented. The recipe setting for silane flow rate directly affects two process variables concerning the gas mixture at the inlet, namely, silane mole fraction and overall gas velocity. As silane mole fraction in-

Growth Rate Pressure Dependence
Epsilon-1 Simulation

Process Conditions	
Carrier Gas	20 slm H_2
Source Gas	2% SiH_4 in H_2
Purge Gas	7 slm H_2

Simulated Growth Rate (A/min) Vs.
Wafer Temperature and Silane Flow Rate

Chamber Pressure	**20 Torr**		**40 Torr**	
Wafer	Silane Flow Rate (sccm)		Silane Flow Rate (sccm)	
Temperature (C)	50	70	50	70
700	202.60	254.90	270.00	315.40
725	336.30	423.90	423.40	510.00
750	512.50	658.50	651.90	820.90

Ratio: 40 Torr / 20 Torr

Wafer	Silane Flow Rate (sccm)	
Temperature (C)	50	70
700	1.33	1.24
725	1.26	1.20
750	1.27	1.25

Mean	1.26
Standard Deviation	0.04

Table 5.7: Results from poly-Si growth simulations comparing growth rates at 20 Torr pressure with growth rates at 40 Torr pressure. Growth rates are averaged over wafer surface.

Parameters of Arrhenius Relationship Fitted to Simulation Data
Pressure Dependence

Assumed Relationship
$R_{Si} = C \exp\left(\frac{-E_a}{R_g T_w}\right)$

Process Conditions	
Carrier Gas	20 slm H_2
Source Gas	2% SiH_4 in H_2
Purge Gas	7 slm H_2

Symbol	Description	20 Torr		40 Torr	
F_{SiH_4}	Silane Flow Rate (sccm)	50	70	50	70
E_a	Activation Energy (eV)	1.48	1.55	1.52	1.68
	Activation Energy (J/mol) ($\times 10^5$)	1.43	1.49	1.47	1.62
E_a/R_g	Ratio (K) ($\times 10^4$)	1.72	1.80	1.77	1.95
C	Pre-exponential Constant (A/min) ($\times 10^{10}$)	1.01	2.81	1.98	14.77

Table 5.8: Parameters calculated by fitting simulation data for poly-Si growth rates to an assumed Arrhenius relationship.

creases, the contribution of gas phase reactions to the overall deposition process will be enhanced [81]. We now briefly discuss the relationship between flow velocity and deposition rate.

It is typically assumed that the gas stream can be divided into two regions. In the region away from the wafer surface, the gas stream is assumed to flow with relatively constant velocity, while in the region next to the wafer surface, there exists a stagnant boundary layer where the flow velocity is zero. In this model, mass transfer of the reactant species through the stagnant layer is dominated by a diffusion process. The mass flux Ψ impinging upon the wafer surface is proportional to the diffusion coefficient D and the difference between the reactant concentration in the full flow C_g and at the surface C_s, and inversely proportional to the thickness

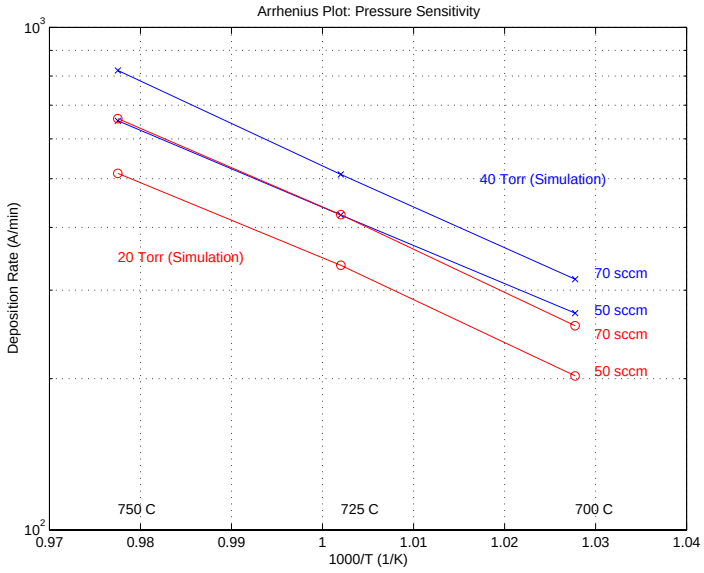


Figure 5.12: Plots illustrating Arrhenius relationship between poly-Si growth rate and wafer temperature in the Epsilon-1. Simulation data is taken for two silane flow rates (50, 70 sccm), three temperatures (700, 725, 750 C), and two chamber pressures (20, 40 Torr) . Growth rates for 40 Torr pressure are a factor of 1.26 times greater than growth rates for 20 Torr pressure consistently over the given range of temperatures and flow rates.

of the boundary layer δ, i.e.,

$$\Psi = \frac{D\,(C_g - C_s)}{\delta}. \tag{5.35}$$

Furthermore, the average boundary layer thickness δ is inversely proportional to the square root of the flow velocity V, i.e.,

$$\delta = C_1\, V^{-1/2} \tag{5.36}$$

The result is that the impinging flux of reactants Ψ is proportional to the square root of flow velocity, i.e.,

$$\Psi = C_2\, V^{1/2} \tag{5.37}$$

We express the relationship between deposition rate and flow velocity as a power law

$$R_{\mathrm{Si}} = C_3\, V^{1/2}. \tag{5.38}$$

If the power law (5.38) provides an accurate model of the relationship between growth rate and flow velocity in the Epsilon-1, then the slope of a plot of the logarithm of growth rate versus the logarithm of flow velocity should be approximately 0.5. This does not hold true (or come anywhere close) for our experimental and simulation data. However, we were able to determine an interesting relationship, by substituting silane flow rate $F_{\mathrm{SiH_4}}$ for flow velocity V in (5.38) and letting the power law exponent vary, i.e.,

$$R_{\mathrm{Si}} = C\, F_{\mathrm{SiH_4}}^{\alpha} \tag{5.39}$$

where α denotes the power law exponent, found by calculating the slope of $\log R_{\mathrm{Si}}$ versus $\log F_{\mathrm{SiH_4}}$. The log-log plots for experimental and simulation data over the range of temperatures we used are presented in Figure 5.13. The resulting values for α are given in Table 5.9. We see that the power law exponent α in (5.39) is roughly 0.7.

We also note that as the gas mixture flows through the process chamber, it heats up and consequently its density decreases and its velocity increases (see Section 5.5.3). This may partially account for the exponent being greater than 0.5.

Based on Equation (5.35), we also expect growth rate to be proportional to silane concentration and hence silane flow rate at the inlet. As shown in Figure 5.14, this relationship holds, with a temperature dependent proportionality constant, reflecting the fact that the process is thermally driven. This is in contrast to the temperature independent nature of the power law exponent. We emphasize that the power law is a purely mass transport controlled phenomenon.

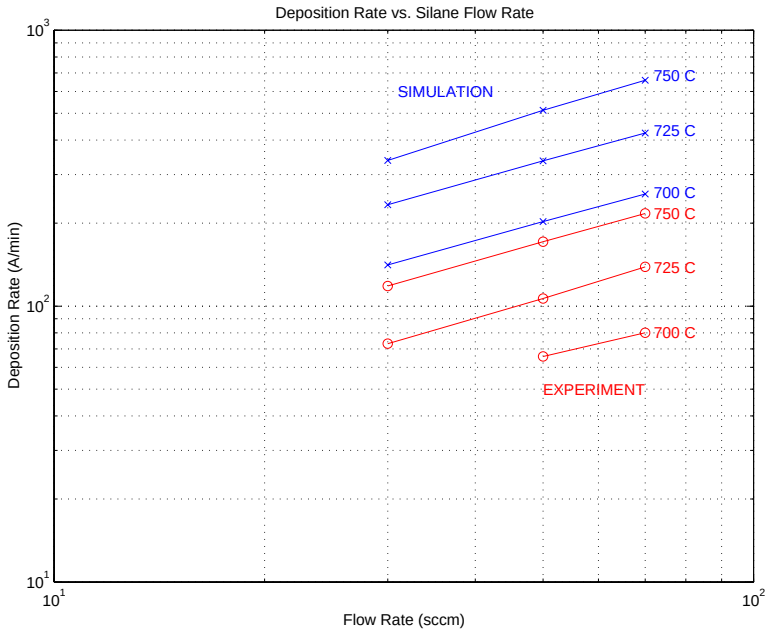


Figure 5.13: Plots illustrating power law relationship between poly-Si growth rate and silane flow rate in the Epsilon-1. Simulation and experimental data is taken for three silane flow rates (30, 50, 70 sccm) and three wafer temperatures (700, 725, 750 C). The power law exponent (slope of plots) is approximately 0.7, consistently over the given range of temperatures. The temperature independence of the power law exponent indicates a completely mass transport controlled phenomenon.

Carrier Gas Sensitivity

A preliminary investigation of the relationship between growth rate and carrier gas was conducted by simulating poly-Si growth using N_2 carrier gas instead of H_2 carrier gas. Wafer temperature was set at 750 C and silane flow rate was set at 70 sccm. The resulting growth rate was 1661 A/min which is a factor of 2.5 times greater than the corresponding simulated growth rate using H_2 carrier.

These simulation results are in accordance with a study by Kleijn [81]. There, use of nitrogen results in an increase in buoyancy effects which causes an increase in the average residency time of gases in the reactor. Thus, gases are heated for a longer period of time and the contribution of gas phase reactions becomes greater.

Relationship Between Growth Rate and Silane Flow Rate
Experiment vs. Simulation

Assumed Relationship	
$R_{Si} = C\, F_{\mathrm{SiH_4}}^{\alpha}$	
Process Conditions	
Chamber Pressure	20 Torr
Carrier Gas	20 slm H_2
Source Gas	2% SiH_4 in H_2
Purge Gas	7 slm H_2

Symbol	Description	Experiment			Simulation		
T_w	Wafer Temperature (C)	700	725	750	700	725	750
α	Power Law Exponent		0.76	0.71	0.70	0.71	0.79

Table 5.9: Power law exponent calculated by fitting experimental and simulation data for poly-Si growth rate to an assumed power law relationship between growth rate and silane flow rate.

Also, thermal diffusion effects are weaker in nitrogen than in hydrogen. Both of these phenomena cause a larger growth rate in nitrogen than in hydrogen.

5.5.2 Deposition Uniformity Prediction

In Section 5.2.3, we presented an argument, based on anecdotal evidence, that temperature uniformity does not produce deposition thickness uniformity in the Epsilon-1 reactor, even for thermally activated processes. On the contrary, thermocouple offsets are set so that the temperature distribution on the wafer surface is intentionally non-uniform. This occurs because of the various transport phenomena that couple with thermally activated chemical mechanisms to influence silicon deposition rate in the Epsilon-1.

In this section we use simulation results to illustrate the phenomenon. We simulated poly-Si growth using 20 Torr chamber pressure, 750 C wafer temperature, 70 sccm silane flow rate, 20 slm H_2 carrier, and 7 slm H_2 purge. The 750 C

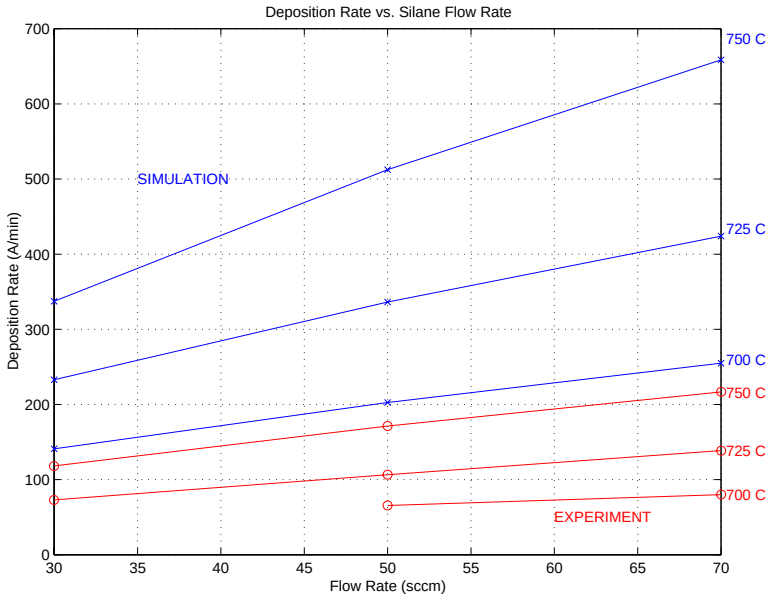


Figure 5.14: Plots illustrating linear relationship between poly-Si growth rate and silane flow rate in the Epsilon-1. Simulation and experimental data is taken for three silane flow rates (30, 50, 70 sccm) and three wafer temperatures (700, 725, 750 C). Slope of plots are temperature dependent, reflecting the fact that the process is thermally driven.

temperature is uniform across the entire surface of the wafer. Figure 5.15 shows a contour plot of the resulting steady-state growth rate on the wafer surface.

Simulated growth rate varies from a minimum of 628 A/min at the downstream side to a maximum of 681 A/min at the upstream outer edge. This represents an 8.4% maximum variation in growth rate across the wafer surface. If thermal activation were the sole contributor to growth rate, then the 8.4% growth rate variation would correspond to a 0.42% maximum variation in temperature. However, we know this is not the case, since we have imposed a perfectly uniform temperature profile on the wafer. The existence of other contributing factors is apparent. On the other hand, this does show that compensation for the other factors may be achievable with much smaller thermocouples offsets than those currently used, which create a maximum temperature variation of 8.5% between thermocouple

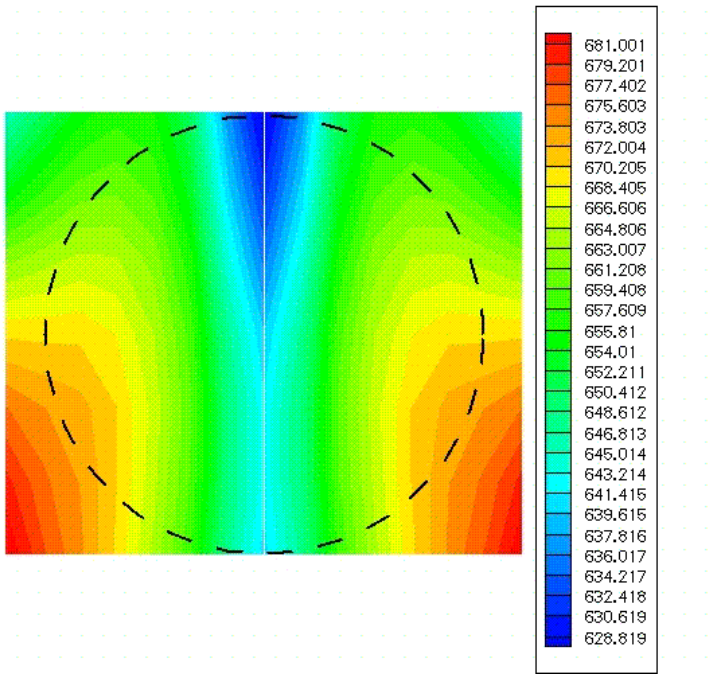


Figure 5.15: Spatial distribution of steady-state deposition rate (A/min) on wafer surface resulting from poly-Si growth with 750 C uniform temperature. The picture shows a non-uniform deposition rate despite the uniform temperature profile. Process conditions are 20 Torr pressure and 70 sccm silane flow rate. Gas flow is from bottom of picture (front/upstream) to top of picture (rear/downstream).

locations. The advantage to this would be reduced mechanical stress and a higher average temperature resulting in higher growth rates.

It is also worthwhile to examine the spatial distribution of the simulated growth rate non-uniformity. Growth rate appears to increase from wafer center to edges, and from front (upstream) side to rear (downstream) side. Thus, the expected depletion effect appears in poly-Si growth simulations. Non-uniformities in gas heating, resulting in non-uniform gas phase reactions and thermal diffusion may also be responsible for growth rate variations. Further simulations will be required to isolate those effects.

Thermocouple offset values described in Section 5.2.3 and used by NG-ESSS

to produce uniform growth create a temperature distribution that is hotter at the front (upstream) than the rear (downstream), and hotter at the center than at the side. The latter seems to match what we would expect given our simulation results, i.e., cool the side to reduce growth rate there. On the other hand, it is difficult to explain the former, since it would exacerbate any reactant depletion effect. Perhaps the simulation understates the effect of downstream gas phase reactions.

We emphasize that in actual operation, the wafer is rotating, so that growth rate variations are averaged, and the significance of front-to-rear variations becomes unclear. We cannot draw any further conclusions at this time. A further experimental study of the effect of thermocouple offsets on uniformity is necessary.

5.5.3 Process Chamber Transport Phenomena Prediction

In this section we study the gas phase transport phenomena in the process chamber of the Epsilon-1. As stated earlier, these effects play an important role in determining deposition rate and uniformity for silicon growth. In particular, we want to observe and analyze gas flow patterns and non-uniformities in the spatial distribution of reactant species.

Simulation results described in this section are for poly-Si growth at 20 Torr pressure, 750 C uniform wafer temperature, 450 C chamber wall temperature, 70 sccm inlet silane flow rate, 20 slm hydrogen carrier flow rate, and 7 slm hydrogen purge flow rate.

We first examine the flow field in the process chamber. Gases are pumped into the Epsilon-1 process chamber from two inlets: the process gas inlet in the upper chamber section and the purge gas inlet in the lower chamber section. They are

pumped out from the process chamber through one outlet located in the upper chamber section. Depending on process and purge inlet flow settings, it is possible for gases to flow from upper to lower chamber sections and visa-versa.

Figure 5.16 shows a view of the simulated flow field in the Epsilon-1 process chamber. Several features are of interest. We observe that there is no gas flow from the upper chamber through gaps to the lower chamber. Thus, the purge flow is effective in this regard. Also, gases from the lower chamber enter the upper chamber mainly through the gap toward the rear of the chamber and also somewhat through the gap near the side chamber wall. This causes the flow in the vicinity of the wafer to be directed from the side wall toward the center line of the chamber. In the y-direction, the flow takes a parabolic profile, i.e., slightly faster at the top of the chamber than near the wafer.

The contours in Figure 5.16 correspond to the flow speed in the z-direction, i.e., from front to rear. We observe that the gas velocity increases from front to rear. This is due to the fact that the gas heats up as it passes by the hot chamber walls and wafer level apparatus, causing the density of the gas mixture to decrease. However, differences in density do not cause any buoyancy driven recirculation cells in this simulation. This is because the flow velocity is relatively high and the temperature gradients in the hot wall chamber are not severe.

The heating of the gases is observed in Figure 5.17, which shows a contour plot of simulated temperature distribution in the Epsilon-1 process chamber. In the vicinity of the wafer the temperature increases from side to center and from wafer to chamber top wall, creating a highly non-uniform temperature field in the gas phase. We note that because the solid surfaces in the lower chamber section are also hot, the purge gas flowing up through the rear and side gaps does not cause

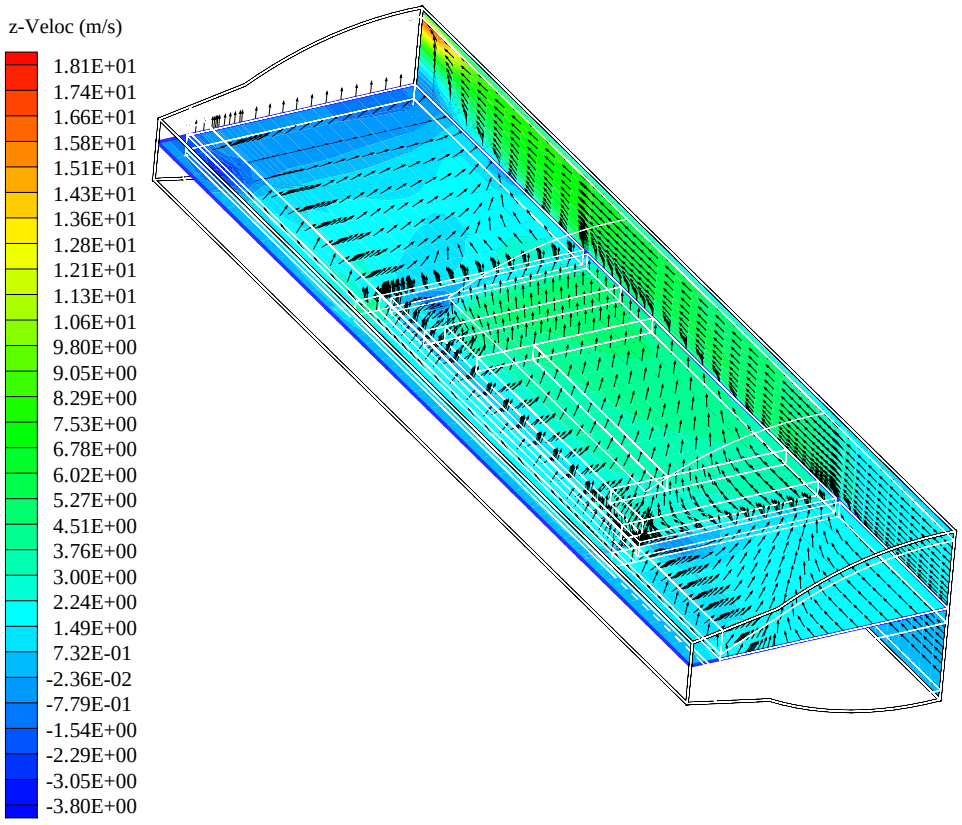


Figure 5.16: Steady state flow pattern of gases in the Epsilon-1 process chamber. Process conditions are 20 slm hydrogen carrier, 70 sccm silane source, 750 C wafer temperature, 450 C chamber wall temperature, and 20 Torr pressure.

a flow of cool gas to enter the upper chamber.

We now study the spatial distribution of the concentrations of the various reactant gases in the Epsilon-1. Silane enters the process chamber through the process inlet in the upper chamber section. It is also produced by one of the five gas phase reactions in the Kleijn model. The gas phase reactions also produce the reactive intermediaries: disilane, trisilane, silylsylene, and silylene. All of these gases are eventually diluted in the hydrogen carrier and hydrogen purge gas.

Figure 5.18 shows a contour plot of simulated silane mass fraction distribution. Silane mass fraction is a maximum at the inlet and becomes depleted by gas phase

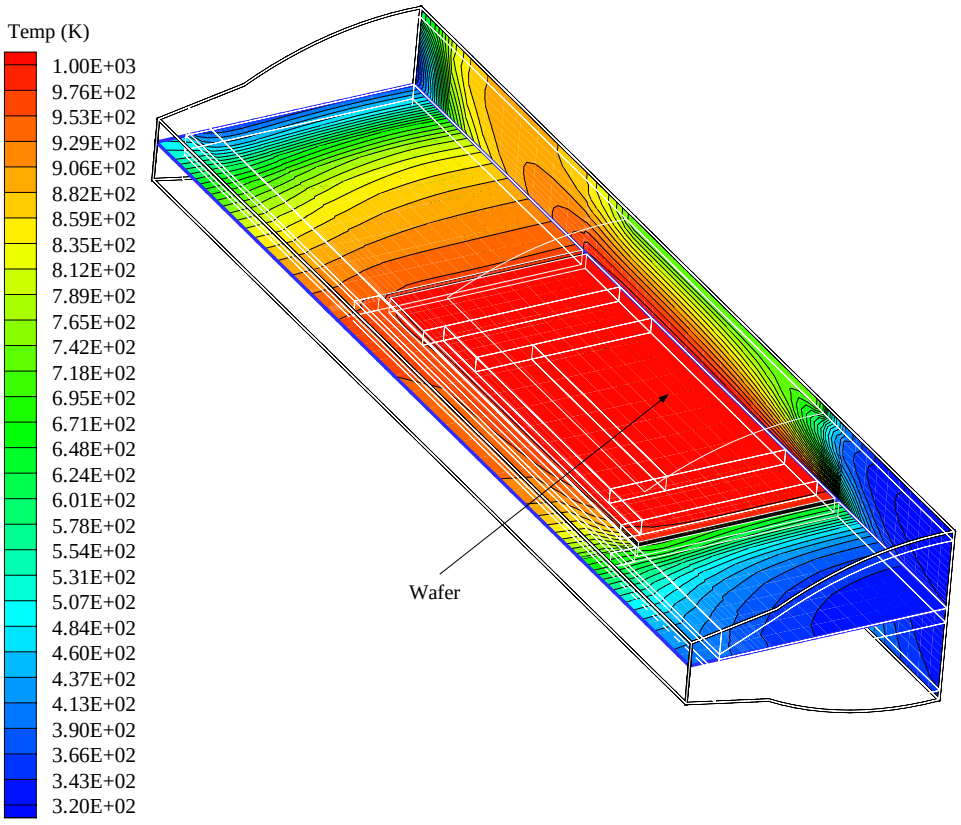


Figure 5.17: Cross-sectional view of the steady-state gas phase temperature distribution in the Epsilon-1 process chamber during growth of poly-Si. Process conditions are 750 C wafer temperature, 450 C chamber wall temperature, 20 slm hydrogen carrier, 70 sccm silane source, and 20 Torr pressure .

and surface reactions as the gas passes over the heated susceptor and wafer. It is at a minimum in locations where the hydrogen purge gas is flowing most heavily into the upper chamber section, in particular, at the side and rear of the ring.

The concentration distributions of the other reactive intermediaries are illustrated by contour plots in Figure 5.19. Because they do not enter at the inlet, these species appear in the flow only once the gas is hot enough for them to be produced, in this case at the front edge of the susceptor ring. Like silane, these species are depleted by surface reactions at the wafer surface. In fact, we observe

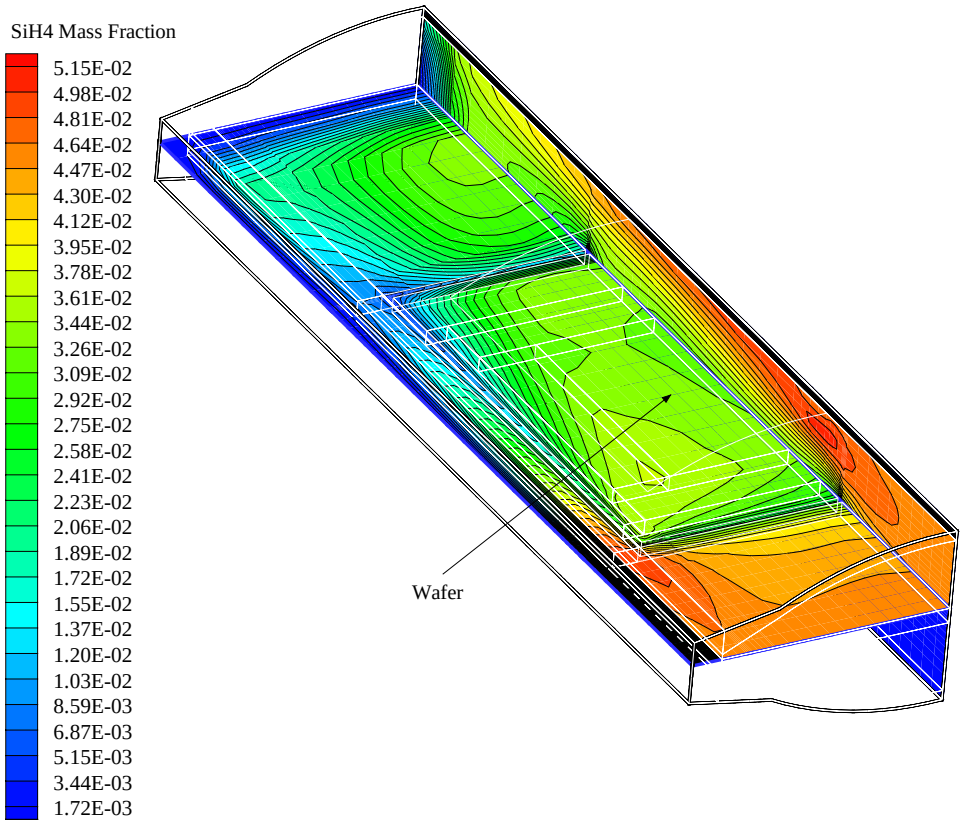


Figure 5.18: Steady-state silane mass fraction distribution in the Epsilon-1 process chamber during poly-Si growth. Process conditions are 750 C wafer temperature, 450 C chamber wall temperature, 20 slm hydrogen carrier, 70 sccm silane source, and 20 Torr pressure.

that silylene, silylsylene, and trisilane are almost completely consumed by surface reactions. On the other hand, some disilane remains just above the wafer surface, although it is at a maximum in areas surrounding the wafer perimeter.

It is clear that the spatial distribution of reactant species concentrations is strongly influenced by the flow field, the gas phase temperature distribution, and surface reactions. We suggested earlier in this report that thermal diffusion may also play a role. This effect is more difficult to isolate and identify.

Figure 5.20 shows two contour plots: the top plot is for silane mass fraction

and the bottom plot is for gas phase temperature. Both plots are snapshots of the $x - z$ plane approximately 2 mm above the wafer surface. In the area just above the wafer and susceptor, it is not possible to isolate any effect thermal diffusion may have, i.e., separate it from the depletion effect caused by gas phase and surface reactions. However, if we restrict attention to the area between the side ring-shelf gap and the side chamber wall, we observe a silane mass fraction gradient that may be due to the Soret effect. In particular, silane mass fraction increases steadily along the chamber side wall and side ring-shelf gap from front to rear. It is possible that the relatively heavy silane molecules have diffused toward the cooler area near the chamber side wall at the front ring-shelf gap and away from the hotter area toward the rear. Again, we emphasize that this speculation needs to be confirmed by conducting additional simulations and possibly actual experiments.

We also note that simulation results show no diffusion of silane or other reactive intermediaries into the lower chamber section. Evidently, the convective forces of the gas flow dominate through the gaps so that any heavy molecules diffusing toward the lower chamber are immediately swept back into the upper chamber.

5.5.4 Purge Flow Optimization

As we observed in Section 5.5.3, the 7 slm H_2 purge flow is effective in preventing any source gases from entering the lower chamber section. In particular, the mass fractions of silane and other reactive intermediaries in the lower chamber were zero for those simulations. This motivates an examination of the relationship between purge flow rate, reactant concentrations in the lower chamber, and possible deposition on the back-side of the susceptor. The objective is to optimize purge flow

rate, where the cost to be minimized is proportional to the amount of consumed H_2, and any back-side deposition is unacceptable.

Figure 5.21 shows contours of silane mass fraction and flow streamlines resulting from two simulations, each using a different H_2 purge flow rate. The top and bottom figures correspond to 7 slm and 2 slm H_2 purge flow rates, respectively. We observe that purge flow rate has an effect on both the flow pattern in the process chamber and the distribution of reactant concentrations. For the 7 slm simulation, steady-state silane concentration in the lower chamber is zero, and streamlines indicate a regular smooth flow from inlets to outlet, with purge gases entering the upper chamber mainly through the rear ring-shelf gap. For the 2 slm simulation, silane concentration in the lower chamber is nonzero, and the flow field becomes irregular, including mixing between upper and lower chambers and recirculation cells in the lower chamber.

For the above simulations, we modeled both the front-side of the wafer and the back-side of the susceptor as reacting surfaces. We note that the wafer and susceptor have different material properties, but we modeled the back-side of the susceptor as if it were the back-side of a silicon wafer. Process conditions were set to 20 Torr pressure, 750 C wafer temperature, and 70 sccm silane flow rate at the upper chamber inlet. For the 7 slm purge flow simulation, no back-side deposition occurred, and average front-side deposition rate was 687 A/min. For the 2 slm purge flow simulation, back-side deposition rate varied from 205 to 359 A/min across the susceptor back-side surface, and average front-side deposition rate was 719 A/min. Apparently, the different flow pattern resulting from a reduction in purge flow rate also causes the front-side deposition rate to increase.

For purposes of optimization we performed one additional simulation with a

5 slm H_2 purge flow. Results qualitatively matched those for the 7 slm purge simulation, i.e., no back-side deposition and zero silane concentration in the lower chamber section. Based on simulation results, we can reduce the flow rate of H_2 purge from 7 slm to 5 slm, thus reducing the use of consumable gases while still maintaining purge effectiveness. However, a reduction to 2 slm is too much and results in unacceptable back-side deposition. The optimum purge flow rate is somewhere between 2 slm and 5 slm. We did not proceed further with this study.

5.6 Remarks

We have presented strong anecdotal evidence, based on thermocouple offsets used by NG-ESSS to produce uniform silicon growth in the Epsilon-1 reactor, that gas phase transport phenomena play an important role in determining deposition uniformity, even for thermally activated growth. This conjecture is in agreement with a study by Kleijn [81] using simulation and experimental data for silicon growth in a cold-wall cylindrical reactor. Models for silicon growth that cannot be coupled to gas phase transport phenomena and that use a simplified chemical kinetics model are inadequate for describing the essential physics and chemistry. This motivated the development of a 3-dimensional comprehensive process-equipment model for silicon growth in the Epsilon-1, incorporating as many relevant transport effects and chemical mechanisms as was feasible from a practical standpoint.

The process-equipment model provides a tool for prediction of deposition rate and other process variables, i.e., the process-equipment state, for a given set of recipe inputs (process conditions) and equipment settings. The predictive capability of the model was tested by comparing results of poly-Si growth simulations to experimental data. Simulations predict growth rates that are roughly three times

greater than actual growth rates, consistently over the given range of operating conditions.

Using the process-equipment model, we performed simulations in order to study the factors that influence deposition rate and uniformity for silicon growth in the Epsilon-1. The relationship between poly-Si growth rate and wafer temperature, chamber pressure, silane flow rate, and hydrogen carrier flow rate were investigated. Although we used a complicated model for poly-Si growth mechanisms developed by Kleijn [81], growth rate temperature sensitivity can be simplified to an Arrhenius relationship. Simulation results indicate that growth rate increases with the logarithm (base 10) of chamber pressure, in agreement with known relationships. We found a power law relationship connecting poly-Si growth rate with silane flow rate at the inlet, with power law exponent roughly 0.7. Finally, we demonstrated that substitution of nitrogen for hydrogen as the carrier gas results in a significantly increased deposition rate.

Simulation results showed that temperature uniformity does not guarantee deposition uniformity in the Epsilon-1. Simulations using a uniform wafer temperature in the thermally activated regime produced growth rates that were non-uniform across the wafer surface. Thus, it is apparent that achieving deposition uniformity requires some degree of temperature non-uniformity to compensate for the effects of other phenomena including reactant depletion, gas heating and gas phase reactions, thermal diffusion of species, and flow patterns.

We have taken steps toward achieving manufacturing objectives. Model predictions allow NG-ESSS to simulate growth experiments in advance, narrow parameter choices, and perform fewer actual experiments. Conditions and settings can be optimized off-line, taking into account simulation results and sensitivity analysis

for pressure, temperature, and flow rates. The effect of adjustments to wafer temperature set-point, chamber pressure, source gas flow rate, thermocouple offsets, and injector settings can be predicted and tuned off-line. Simulation results show that consumption of process gases can be reduced by decreasing the purge gas flow from 7 slm to 5 slm and possibly further without compromising the ability of the purge gas to prevent back-side deposition.

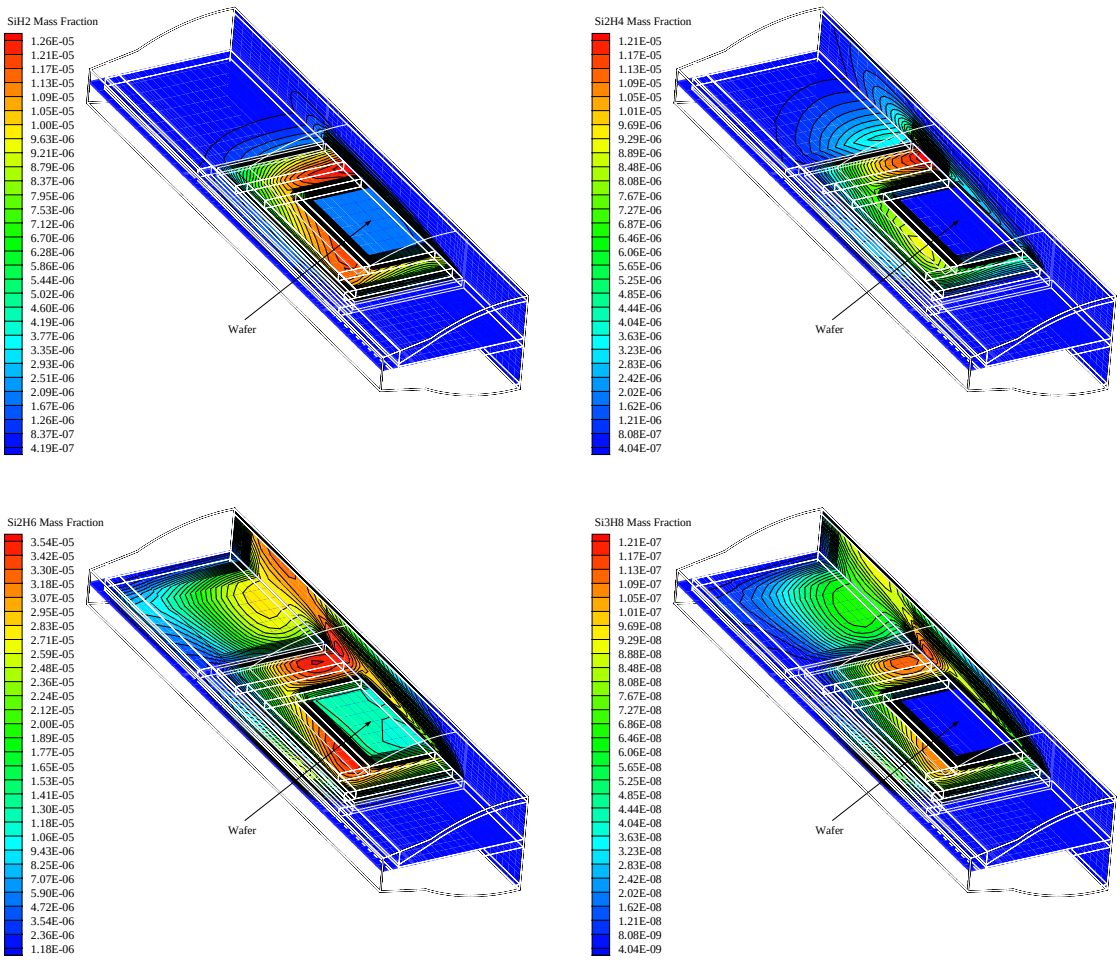


Figure 5.19: Steady-state mass fraction distribution for reactive intermediaries during poly-Si growth: silylene (top-left), silylsylene (top-right), disilane (bottom-left), trisilane (bottom-right). Process conditions are 750 C wafer temperature, 450 C chamber wall temperature, 20 slm hydrogen carrier, 70 sccm silane source, and 20 Torr pressure .

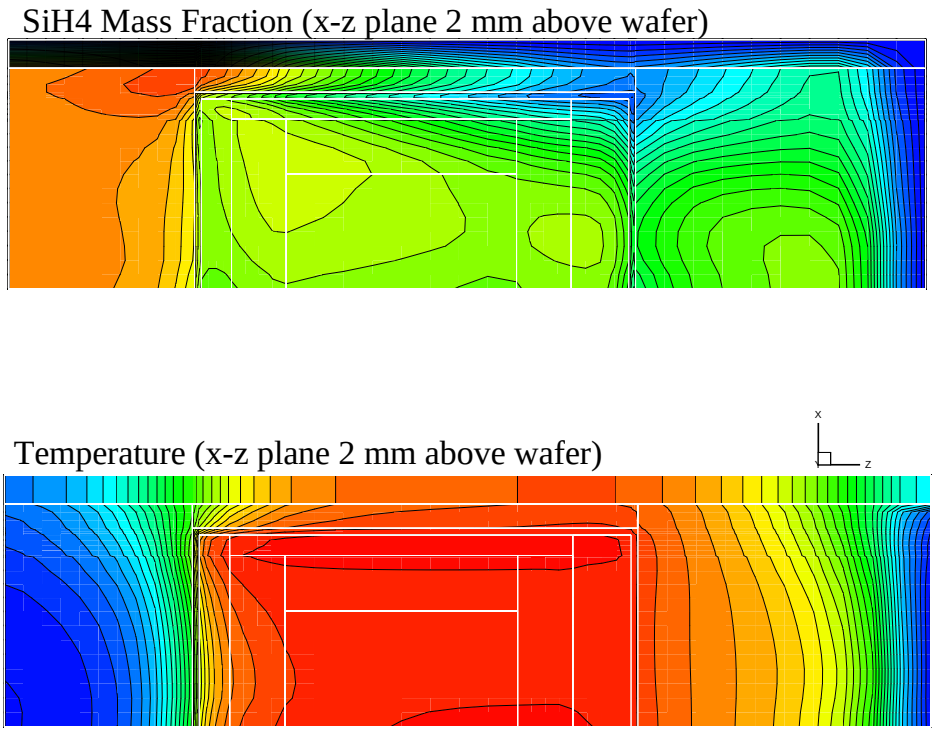


Figure 5.20: Illustration of thermal diffusion (Soret effect) in the Epsilon-1 process chamber. Contours of steady-state silane mass fraction (top) and temperature (bottom) for the $x - z$ plane approximately 2 mm above the wafer surface.

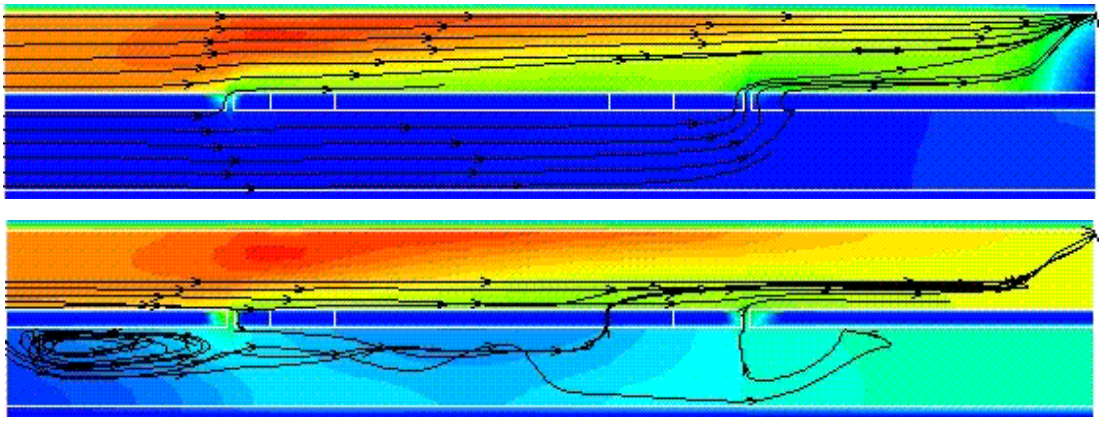


Figure 5.21: Comparison of flow streamlines and steady-state silane mass fraction contours for hydrogen purge flow rates of 7 slm (top) and 2 slm (bottom). The higher purge flow rate results in zero silane concentration in the lower chamber section, no back-side deposition, and regular flow from inlets to outlet. The lower purge flow rate is ineffective, producing non-zero silane concentration in the lower chamber section and some back-side deposition.

Chapter 6

Modeling and Reduction for RTP Heat Transfer

6.1 Introduction

This chapter addresses the problem of deriving low-order models for RTP control systems. We focus on one particular aspect of the overall process-equipment model for the Epsilon-1 reactor: heat transfer within the solid wafer and among the wafer, heat lamps, chamber walls, and flowing gases. A physical model of the wafer thermal dynamics is formulated in Section 6.2, accounting for conductive, radiative, and convective effects.

Control of the temperature distribution on the wafer surface is achieved through several independent lamp zone actuators, each of which causes a different set of tungsten-halogen lamps to irradiate the wafer. The lamp heating component of the heat transfer model is derived in Section 6.3, based on several simplifying assumptions and a detailed view factor analysis of the Epsilon-1 geometry and lamp system characteristics. We also present the results of growth experiments

that provide a degree of empirical validation for portions of the view factor analysis. Although contact measurements do not occur during an actual deposition run, we incorporate a set of thermocouples into the model to play the role of temperature sensors.

Given the RTP heat transfer model, we derive low-order approximations of the evolution equations via application of the reduction approaches detailed in Chapter 3, i.e., POD and balanced truncation. A comparative study is presented in Section 6.4, in which the effectiveness of the two approaches is examined through numerical simulations of the full and reduced RTP control system models. We summarize and make some additional remarks in Section 6.5.

6.2 Wafer Heat Transfer Model

The model for wafer heat transfer is a modified version of models presented in [2, 28, 97, 138, 139]. It is based on an energy balance for a heat conducting solid which emits and absorbs heat radiation at its boundary surfaces. The model takes into account simplified effects of conductive, radiative (including lamp heating), and convective heat transfer. Both a continuum model and a discretized version are presented here, based on identical principles of energy balance. Although the wafer is a continuous solid body, a discretized model is required for purposes of numerical solution.

Wafer Characteristics

We assume that the wafer shape is perfectly cylindrical. Its geometry is illustrated in Figure 6.1. The heat transfer model will be formulated in cylindrical coordinates with radial variable r, azimuthal variable θ, and axial variable z. The wafer radius

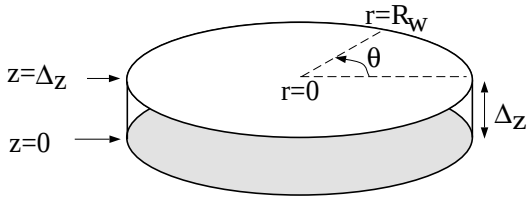


Figure 6.1: Wafer geometry used in the heat transfer model.

is denoted R_w and the wafer thickness is denoted Δ_z, so that the top surface of the wafer has z-coordinate Δ_z and the bottom surface has z-coordinate 0. Note that for purposes of this study, the wafer and susceptor have been combined into a single homogeneous solid body.

We assume that the wafer is pure silicon, and ignore, e.g., the layer of silicon dioxide on which poly-silicon is deposited. Thus, we use thermal and optical properties for pure silicon. The physical constants are given in Appendix G.

Continuum Model

The temperature field in the solid wafer is denoted $T_w = T_w(t, r, \theta, z)$ where t represents time. Time evolution of T_w is governed by a PDE (usually referred to as the *heat equation*) which models heat conduction within the wafer, together with boundary conditions (BCs) which model net heat flow to and from the wafer boundary surfaces (top, bottom, and edge). The PDE is given in cylindrical coordinates, for $t > 0$, $0 < r < R_w$, $0 \le \theta < 2\,\pi$, and $0 < z < \Delta_z$, by

$$\rho_w\, C_{p_w}\, \frac{\partial T_w}{\partial t} = \frac{1}{r}\, \frac{\partial}{\partial r} \left(k_w\, r\, \frac{\partial T_w}{\partial r} \right) + \frac{1}{r^2}\, \frac{\partial}{\partial \theta} \left(k_w\, \frac{\partial T_w}{\partial \theta} \right) + \frac{\partial}{\partial z} \left(k_w\, \frac{\partial T_w}{\partial z} \right) \qquad (6.1)$$

where ρ_w is the mass density of the wafer, C_{p_w} is the heat capacity of the wafer (the product $M_w = \rho_w\, C_{p_w}$ is often referred to as the wafer *thermal mass*), and k_w

is the thermal conductivity of the wafer. The associated BCs are given by

$$\frac{\partial T_w}{\partial r} = 0\,, \quad r = 0 \tag{6.2}$$

$$k_w \frac{\partial T_w}{\partial r} = q_{\text{edge}}(\theta, z)\,, \quad r = R_w \tag{6.3}$$

$$k_w \frac{\partial T_w}{\partial z} = -q_{\text{bottom}}(r, \theta)\,, \quad z = 0 \tag{6.4}$$

$$k_w \frac{\partial T_w}{\partial z} = q_{\text{top}}(r, \theta)\,, \quad z = \Delta_z \tag{6.5}$$

where the BC (6.2) results from symmetry about the wafer center, and q_{edge}, q_{bottom}, and q_{top} represent the net heat flow per unit surface area to and from the wafer edge, bottom, and top boundary surfaces, respectively, and will be described in more detail later.

For purposes of modeling film growth, we focus our attention on the top surface of the wafer where reactions take place. Invoking the assumption of azimuthal symmetry, so that no temperature gradients exist in the azimuthal direction (i.e., $\partial T_w/\partial\theta = 0$), time evolution of T_w at the wafer top surface is governed, for $t > 0$, $0 < r < R_w$, and $z = \Delta_z$, by

$$\rho_w\, C_{p_w}\, \frac{\partial T_w}{\partial t} = \frac{1}{r}\, \frac{\partial}{\partial r} \left(k_w\, r\, \frac{\partial T_w}{\partial r} \right) + \frac{\partial}{\partial z} \left(k_w\, \frac{\partial T_w}{\partial z} \right) \tag{6.6}$$

with BCs remaining the same as before except $q_{\text{edge}} = q_{\text{edge}}(z)$, $q_{\text{bottom}} = q_{\text{bottom}}(r)$, and $q_{\text{top}} = q_{\text{top}}(r)$. We also assume that the wafer thickness is sufficiently small so that no thermal gradients exist in the axial direction within the wafer interior. Therefore, we approximate the axial gradient term at the top surface by

$$\begin{aligned} \frac{\partial}{\partial z} \left(k_w\, \frac{\partial T_w}{\partial z} \right) &\simeq \frac{1}{\Delta_z} \left(k_w\, \frac{\partial T_w}{\partial z}\, |_{z=\Delta_z} - k_w\, \frac{\partial T_w}{\partial z}\, |_{z=0} \right) \\ &= \frac{1}{\Delta_z}\, (q_{\text{top}} + q_{\text{bottom}}) \end{aligned} \tag{6.7}$$

where we have made substitutions using BCs (6.4) and (6.5). The resulting PDE governs the evolution of the wafer top surface temperature field as a function of

time and radial position,

$$\rho_w \, C_{p_w} \, \frac{\partial T_w}{\partial t} = \frac{1}{r} \, \frac{\partial}{\partial r} \left(k_w \, r \, \frac{\partial T_w}{\partial r} \right) + \frac{1}{\Delta_z} \, (q_{\text{top}} + q_{\text{bottom}}) \tag{6.8}$$

with BCs

$$\frac{\partial T_w}{\partial r} = 0 \,, \quad r = 0 \tag{6.9}$$

$$k_w \, \frac{\partial T_w}{\partial r} = q_{\text{edge}}(\Delta_z) \,, \quad r = R_w \tag{6.10}$$

Now, we must find expressions for q_{top} and q_{bottom}, the net heat flow into the top and bottom surfaces of the wafer. For this study, we assume that the top and bottom surfaces are subject to identical heat transfer mechanisms, and let

$$q_{\text{top}} + q_{\text{bottom}} = q^{\text{em}} + q^{\text{ab}} + q^{\text{conv}} + q^{\text{chem}} \tag{6.11}$$

where the terms on the right hand side of (6.11) represent the flow of thermal energy to and from the wafer and are dependent on time, position, and wafer temperature. In particular, q^{em} is radiative energy emitted, q^{ab} is radiative energy absorbed, q^{conv} denotes energy losses due to convective heat transfer, and q^{chem} is energy transfer due to the heat generated by chemical reactions. For this study, we do not include the heat of chemical reactions and consequently ignore q^{chem}.

The term q^{em} represents radiative losses from the wafer. We assume q^{ab} depends on radiant heat flux from a uniform ambient, in this case the chamber walls, and radiant heat flux from the lamps, but without reflections or other effects. The individual terms are given by

$$q^{\text{em}} = -2 \, \epsilon_w \, \sigma_b \, T_w^4 \tag{6.12}$$

$$q^{\text{ab}} = 2 \, \alpha_w \, \sigma_b \, T_c^4 + \alpha_w \sum_{i=1}^{10} Q_i \, u_i \tag{6.13}$$

where σ_b denotes the Boltzmann constant, ϵ_w denotes the wafer emissivity, α_w denotes the wafer absorptivity, T_c denotes the uniform ambient temperature of

the chamber walls, $Q_i = Q_i(r)$ is a function of position describing the heat flux intensity incident on the wafer due to the i–th lamp group, and $u_i = u_i(t)$ is the time-varying actuated power level of the i-th lamp group.

The convective term is given by

$$q^{\text{conv}} = -h_v \, (T_w - T_g) \tag{6.14}$$

where h_v denotes the convective heat transfer coefficient and T_g denotes the temperature of the gas flowing past the wafer. Note that we have assumed a constant uniform gas temperature. In order to estimate h_v, we assume that flow in the process chamber is a laminar flow along a flat plate. The mean heat transfer coefficient is given in [124] (pp. 233-235) as

$$h_v = 2 \left[0.332 \, k_g \, Pr^{1/3} \left(\frac{Re^{1/2}}{L} \right) \right] \tag{6.15}$$

where k_g denotes the gas thermal conductivity, Pr denotes the gas Prandtl number, Re denotes the gas Reynolds number, and L denotes the length of the chamber. We have computed the Reynolds number Re to be approximately 27 for the flow in the ASM Epsilon-1 during a typical deposition run, thus confirming the laminar assumption. The calculated value of h_v was then validated using flow and temperature data from a corresponding CFD simulation.

Remark 6.2.1 *It is sometimes convenient to assume that the wafer is a graybody, so that $\epsilon_w = \alpha_w$ for all relevant wavelengths of radiation and wafer temperatures. However, we do not make this assumption here, and use different values for emissivity and absorptivity.* □

Remark 6.2.2 *The parameters ρ_w, C_{p_w}, and k_w in general have a nonlinear dependence on temperature, and can be modeled as polynomial functions of T_w. Like-*

wise, the parameters ϵ_w and α_w in general have a nonlinear dependence on temperature and deposition thickness. However, we invoke the assumption that mass density, heat capacity, thermal conductivity, emissivity, and absorptivity are constant, i.e., no variation with temperature, film thickness, position, or time. □

Given these simplifying assumptions, the PDE model specializes to

$$\begin{aligned}\frac{\partial T_w}{\partial t} = \frac{k_w}{\rho_w C_{p_w}} \frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial T_w}{\partial r} \right) &+ \frac{h_v}{\rho_w C_{p_w} \Delta_z} (T_g - T_w) + \frac{2 \sigma_b \epsilon_w}{\rho_w C_{p_w} \Delta_z} \left(T_c^4 - T_w^4 \right) \\ &+ \frac{\alpha_w}{\rho_w C_{p_w} \Delta_z} \sum_{i=1}^{10} Q_i \, u_i \end{aligned} \tag{6.16}$$

where we recall that $T_w = T_w(t, r)$, $Q_i = Q_i(r)$, and $u_i = u_i(t)$.

Since the guard ring insulates the wafer from radiation directed at its edge boundary surface, we assume zero heat transfer at the wafer edge so that

$$q_{\text{edge}} = 0 \tag{6.17}$$

giving the BCs

$$\frac{\partial T_w}{\partial r} = 0 \,, \quad r = 0 \tag{6.18}$$

$$\frac{\partial T_w}{\partial r} = 0 \,, \quad r = R_w \tag{6.19}$$

Discretized Model

The continuum model as given by the above PDE and BCs can be discretized using a suitable scheme, e.g., finite differences or finite elements. However, for our simplified model it is easier to formulate a discretization by applying the energy balance principles directly to individual annular elements of the wafer. The general idea, which divides the wafer into annular regions, is illustrated in Figure 6.2. Annular regions are numbered from 1 to n with element 1 being the innermost disk

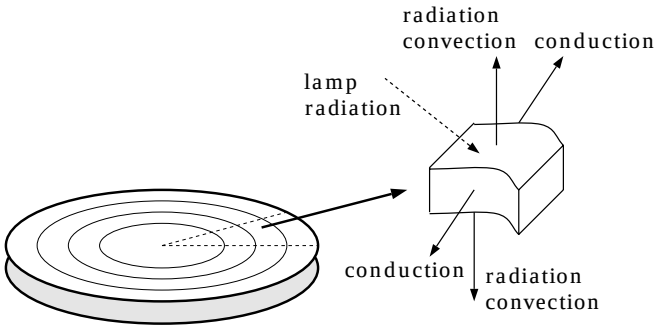


Figure 6.2: Heat transfer mechanisms affecting annular region of wafer.

and element n being the outermost annular region. The i-th element has mean radius $r(i)$, and is bounded by an outer cylinder of radius r_{out}, inner cylinder of radius r_{in}, top surface at $z = \Delta_z$, and bottom surface at $z = 0$. The discretization is uniform so that

$$\Delta_r = r(i) - r(j) \qquad i, j \in \underline{n} \tag{6.20}$$

is constant for all regions.

The usual symmetry assumptions are invoked so that temperature depends on radial position and time only. The discretized wafer temperature field is given by the n-vector $T_w(t)$, where the i-th entry of $T_w(t)$ represents the temperature at radial position $r(i)$ and time t.

The wafer heat transfer model is then given by the ODE

$$\dot{T}_w = A_c\, T_w + A_r\, T_w^4 + A_v\, T_w + \Gamma + B\, P \tag{6.21}$$

where A_c, A_r, and A_v are $n \times n$ matrices representing the effects of, respectively, conductive, radiative, and convective heat transfer mechanisms, Γ is a constant n-vector that accounts for the gas and chamber wall temperature, B is a $n \times m$ matrix derived from discretized lamp zone radiant intensity profiles, and $P = P(t)$ is a m-vector of control inputs corresponding to lamp zone power levels. We present the details of the ODE model below.

The top surface area, volume, and mass of annular region i are given, respectively, by

$$\begin{aligned} S(i) &= \pi\,(r_{\text{out}}(i)^2 - r_{\text{in}}(i)^2) \\ V(i) &= S(i)\,\Delta_z \\ m(i) &= \rho_w\,V(i) \end{aligned} \tag{6.22}$$

The matrix representing conductive heat transfer is then represented by the tridiagonal matrix given by the entries, for $i = 2, \ldots n-1$,

$$\begin{aligned} A_c(i,i) &= \frac{-2\,k_w}{\rho_w\,C_{p_w}\,\Delta_r}\,\frac{r_{\text{out}}(i) + r_{\text{in}}(i)}{r_{\text{out}}^2(i) - r_{\text{in}}^2(i)} \\ A_c(i,i+1) &= \frac{2\,k_w}{\rho_w\,C_{p_w}\,\Delta_r}\,\frac{r_{\text{out}}(i)}{r_{\text{out}}^2(i) - r_{\text{in}}^2(i)} \\ A_c(i,i-1) &= \frac{2\,k_w}{\rho_w\,C_{p_w}\,\Delta_r}\,\frac{r_{\text{in}}(i)}{r_{\text{out}}^2(i) - r_{\text{in}}^2(i)} \end{aligned} \tag{6.23}$$

and for the other other indices

$$\begin{aligned} A_c(1,1) &= \frac{-2\,k_w}{\rho_w\,C_{p_w}\,\Delta_r}\,\frac{1}{r_{\text{out}}(1)} \\ A_c(1,2) &= \frac{2\,k_w}{\rho_w\,C_{p_w}\,\Delta_r}\,\frac{1}{r_{\text{out}}(1)} \\ A_c(n,n) &= \frac{-2\,k_w}{\rho_w\,C_{p_w}\,\Delta_r}\,\frac{r_{\text{in}}(n)}{r_{\text{out}}^2(n) - r_{\text{in}}^2(n)} \\ A_c(n,n-1) &= \frac{2\,k_w}{\rho_w\,C_{p_w}\,\Delta_r}\,\frac{r_{\text{in}}(n)}{r_{\text{out}}^2(n) - r_{\text{in}}^2(n)} \end{aligned} \tag{6.24}$$

where we note that zero heat flux BCs have been incorporated into the model via boundary elements of matrix A_c.

The matrices representing radiative transfer from wafer surface to chamber walls and convective heat transfer from the process gases to wafer are given, respectively, by

$$A_r = \text{diag}\left(\frac{-\sigma_b\,\epsilon_w}{\rho_w\,C_{p_w}\,\Delta_z}, \ldots, \frac{-\sigma_b\,\epsilon_w}{\rho_w\,C_{p_w}\,\Delta_z}\right) \tag{6.25}$$

and

$$A_v = \text{diag}\left(\frac{-h_v}{\rho_w\, C_{p_w}\, \Delta_z}, \ldots, \frac{-h_v}{\rho_w\, C_{p_w}\, \Delta_z}\right) \tag{6.26}$$

where we note that A_r and A_v take a diagonal form as a result of the simplifications made in our model.

The effect of radiation from chamber wall to wafer and convective transfer from gas to wafer are incorporated into the constant vectors Γ_r and Γ_v whose entries are given by

$$\Gamma_r(i) = \frac{\epsilon_c\, \sigma_b\, \alpha_w}{\rho_w\, C_{p_w}\, \Delta_z}\, T_c^4 \qquad i \in \underline{n} \tag{6.27}$$

$$\Gamma_v(i) = \frac{h_v}{\rho_w\, C_{p_w}\, \Delta_z}\, T_g \qquad i \in \underline{n} \tag{6.28}$$

where ϵ_c is the emissivity of the quartz chamber walls. These effects are combined by summing into one constant vector

$$\Gamma = \Gamma_r + \Gamma_v \tag{6.29}$$

The lamp heating component is modeled by the term BP, where the matrix B is referred to as the influence matrix and P is the control input vector corresponding to lamp zone power levels. The influence matrix B is derived from the heat flux intensity profiles $Q_i(r), i = 1, \ldots, m$ associated with each of the independently actuated lamp zones. The details of these flux profiles, or influence functions, are given in Section 6.3.

The Epsilon-1 is equipped with four independently actuated lamp zones. However, the analysis in Section 6.3 (based on some simplifying assumptions) yields only three independent controls, i.e., two of the lamp zones produce identical flux profiles. Thus, we have $m = 3$ control inputs in our model. The flux profiles are suitably discretized and arranged in a matrix Q. They are then incorporated into

the influence matrix B given by

$$B = \frac{\alpha_w}{\rho_w \, C_{p_w} \, \Delta_z} \, Q \tag{6.30}$$

Computation

To avoid problems of scaling in computational work, we normalize variables and parameters so that all units cancel, i.e. write the model in dimensionless form. It is customary to adopt a notation for the dimensionless variables, e.g. T_w becomes $\widetilde{T}_w$. Instead, we denote the dimensionless variables by the same symbol as their dimensional counterparts and caution the reader to keep this in mind. The conversions are

$$T_w \to \frac{T_w}{T_c}\,, \quad Q_i \to \frac{Q_i}{Q_{ref}}\,, \quad t \to \frac{t}{\tau}\,, \quad r \to \frac{r}{R_w} \tag{6.31}$$

It has been observed in the ASM reactor that T_c, the chamber wall temperature, is approximately 300 K less than wafer temperature [105] during a typical processing run. As reference values we select a wafer temperature of 1000 K and chamber wall temperature of 700 K. The reference thickness h_{ref} of 1.0 micron was selected because it is on the order of the thickness of films we are interested in growing. The reference heat flux Q_{ref} of 29.24 W/cm^2 was computed using the lamp power specification of 6 kW radiating over one-half of a spherical surface area of radius 57.15 mm (2.25 in).

Using the dimensionless variables as given above, the parameters (matrices and vectors) in equation (6.21) become

$$\begin{gathered} A_c \to \frac{\tau}{R_w^2}\, A_c\,, \quad A_r \to \tau\, T_c^3\, A_r\,, \quad A_v \to \tau\, A_v\,, \\ \Gamma_r \to \frac{\tau}{T_c}\, \Gamma_r\,, \qquad \Gamma_v \to \frac{\tau}{T_c}\, \Gamma_v\,, \qquad B \to \frac{\tau}{T_c}\, B \end{gathered} \tag{6.32}$$

to yield a dimensionless ODE model equivalent to (6.21).

Given an initial temperature profile, $T_{w_0} = T_w(0)$, the temperature distribution on the wafer surface can be determined as a function of time and radial position by numerically integrating the ODE (6.21). Typically, the initial condition is a uniform temperature field set to ambient, i.e., $T_{w_0} = [700 \ldots 700]^{\mathsf{T}}$. We used a fourth order Runge-Kutta integration scheme to perform the numerical integrations. The discretization resolution was typically set at $n = 101$.

See Appendix G for values of all physical constants and parameters used in our simulations.

6.3 Lamp Heating Model

In this section we present the lamp heating component of the wafer heat transfer model presented in Section 6.2. This component enters as the control input term in the evolution equation (6.21) for wafer temperature. We describe additional details of the physical set-up of the Epsilon-1 lamp apparatus, derive the relationship between lamp power settings and heat flux incident on the wafer surface, and validate the analysis using experimental data.

Lamp Equipment

The Epsilon-1 reactor is equipped with 21 tungsten-halogen lamps for heating the wafer. There are 17 linear lamps (long, thin tubes) with a maximum power output of 6.0 kW and four spot lamps (spherical bulbs) with a maximum power output of 1.0 kW. The linear lamps are organized into two arrays, referred to as upper and lower. The layout is illustrated in Figure 6.3. The upper and lower lamp arrays are located outside the process chamber, respectively, above and below the top and bottom quartz walls. They illuminate, respectively, the top surface of the

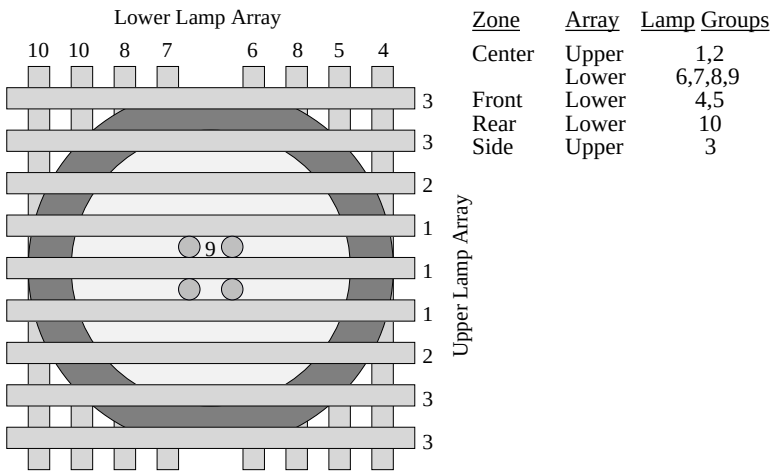


Figure 6.3: Organization (top view) of upper and lower lamp arrays and spot lamps: individual lamps are assigned to lamp groups and heat zones as shown.

wafer and the bottom surface of the susceptor. The upper lamps are arranged perpendicular to the lower lamps. The spot lamps are located directly below, and illuminate, the bottom of the center of the susceptor.

The power to individual lamps cannot be controlled in the Epsilon-1. Rather, the lamps are combined into groups, which are further combined into zones. The organization into groups and zones is also illustrated in Figure 6.3. The power to each of the four lamp heat zones is controlled independently via four PID feedback loops. In addition, the power to each of the ten lamp groups can be independently controlled but only via manual settings. However, it is the four lamp heat zones that normally serve as the actuators in the Epsilon-1 heating system. The time-varying percentages of full power P_i supplied to the respective zones are the four control inputs. The zone name roughly corresponds to the area of the wafer that receives the most intense illumination from the particular zone, e.g. center, front (upstream), rear (downstream), and side.

Flux Profiles

Each individual lamp creates a heat flux profile, i.e., a spatially distributed radiant intensity, that is incident across the wafer surface. We refer to these spatial profiles as influence functions, the value of which at a given point on the wafer surface is the heat flux intensity, measured in Watts per unit area, irradiated onto the given point.

For purposes of this study, we assume that the shape of the flux profile for a given lamp is determined solely by the geometry of the lamp and wafer apparatus. Other factors that play a role, but are unmodeled here, include the effect of reflectors, chamber walls, and any apparatus within the chamber enclosure that either reflects or absorbs heat radiation. The magnitude of a lamp's flux profile is determined completely by the maximum power output of the lamp, i.e., 6.0 kW for the linear lamps and 1.0 kW for the spot lamps.

We use view factor analysis to compute the flux profiles $Q_i(r, \theta)$, as a function of radial position r and azimuthal position θ, for the individual lamps in the Epsilon-1. For purposes of this analysis, we ignore the apparent symmetry breaking as described in Chapter 5, and assume that wafer rotation causes azimuthal symmetry of lamp radiation. Azimuthal averaging accounts for wafer rotation and results in profiles $Q_i(r)$ as functions of r only. The view factor calculations are lengthy, so we present the details and results of this procedure in Appendix H.

The individual flux profiles are combined using superposition, in accordance with the previously described organization of lamps into groups and zones, to produce four influence functions, each corresponding to one of the independently actuated lamp heat zones. They are shown in Figure 6.4. Each profile $Q_i(r)$ is modulated by a corresponding power setting $P_i(t)$ (control input), i.e., the pro-

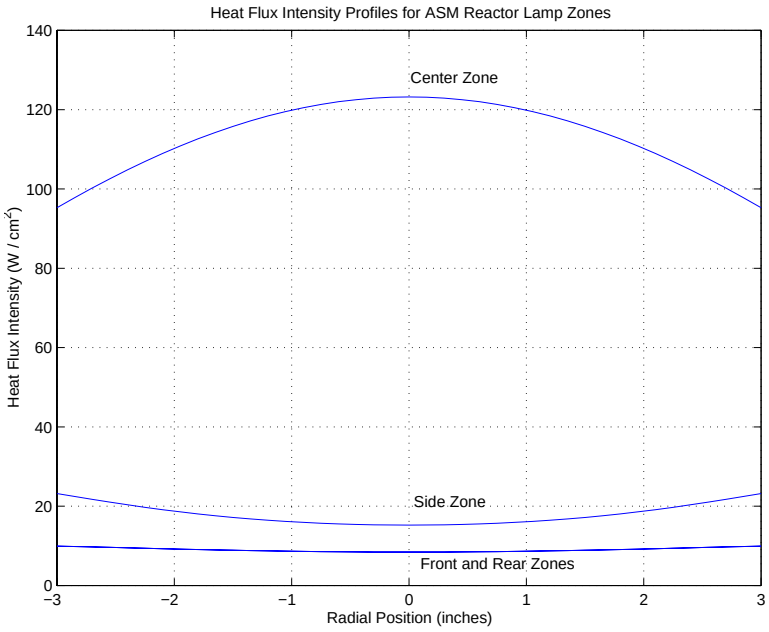


Figure 6.4: Heat flux intensity profiles for ASM Epsilon–1 heat zones: flux intensity (W/cm^2) versus radial position for the four heat zones.

portion of full power applied. The flux profiles Q_i and power settings P_i are then incorporated into the evolution equation (6.21) for wafer temperature as described in Section 6.2.

Remark 6.3.1 *As shown in Figure 6.4, the view factor analysis yields identical flux profiles for the front and rear zones. This is due to the geometry of the front and rear zones with respect to the wafer, wafer rotation, and the other simplifying assumptions invoked in the model development. Thus, we use only three independent lamp zone control inputs in our control system model for the Epsilon-1 RTP system.* □

Remark 6.3.2 *We note that, given accurate flux intensity or temperature measurements, one might be able to experimentally measure the flux profile for a given lamp zone. However, it has been the experience of both Northrop Grumman and*

ASM that an instrumented wafer used to take such measurements is unreliable. Moreover, the instrumented wafer that is available for these purposes is capable of measuring temperature at only nine points on the wafer surface, which is insufficient resolution for purposes of model development. Another alternative is to infer wafer temperature from growth rate of poly-silicon. However, this method is still impractical, since a prohibitively large number of measurements for wafer location and film thickness are required to achieve sufficiently high resolution. □

Experimental Validation

We do, however, use an experimental approach for purposes of a comparative validation of the flux profiles that were determined using view factor analysis. The procedure was as follows:

(i) We deposited poly-silicon on a non-rotating wafer for a fixed period of time τ with lamp group i manually set to power setting P. Thickness measurements were taken to give a thickness profile $h_{(i,P,\tau)}(r,\theta)$ and growth rate distribution $R_{(i,P)}(r,\theta) = h/\tau$. We then averaged azimuthally to give growth rate in terms of radial position for a rotating wafer, i.e., $R_{(i,P)}(r)$.

(ii) The Arrhenius law for growth rate (5.1) was inverted to determine temperature as a function of radial position, i.e.,

$$T_{(i,P)}(r) = \frac{E_a}{R_g}\left[\ln\left(k_0\, X_{\mathrm{SiH_4}}\right) - \ln\Big(R_{(i,P)}(r)\Big)\right]^{-1} \tag{6.33}$$

(iii) The temperature field $T_{(i,P)}(r)$ was substituted into the evolution equation for temperature (6.21) in the wafer heat transfer model. Applying the steady-state condition $\dot{T} = 0$, we can solve

$$0 = A_c\, T + A_r\, T^4 + A_v\, T + \Gamma + B_i\, P \tag{6.34}$$

for the discretized influence function $B_i(r)$.

Isolation of the individual lamp groups was achieved by operating the reactor in manual mode, i.e., with the automatic control loops for temperature regulation turned off. In manual mode, the lamp groups are no longer organized into four zones for the purpose of temperature control. Instead, each of the ten groups can be toggled on and off individually, and the power setting of each (between 0% and 100%) can be set manually. To isolate a particular lamp group, all others were turned off, while the power setting for the lamp group being tested is set manually to an appropriate level.

The wafer was heated with an individual lamp group, whose power setting was adjusted manually until at least one of the thermocouple readings reached the range where deposition would occur, which was approximately 700 C. The exact temperature readings were not important because in the next step temperature would be inferred from thickness data. Then, flow of silane in hydrogen carrier was started. Silicon was deposited for five minutes. Wafer rotation was turned off so that effects of asymmetry would appear in the resulting deposition.

This procedure was followed to test four of the lamp groups: 1, 8, 9, and 10. Lamp group 1 is in the upper lamp array and radiates directly toward the top center of the wafer. Using lamp group 1 alone, we were able to heat the wafer to a temperature sufficiently high for deposition to occur and to record sufficient data for analysis. Lamp groups 8, 9, and 10 are in the lower lamp array and radiate toward the bottom of the susceptor. Due to conduction and losses throughout the susceptor, it was more difficult to heat the wafer using each of these lamp groups alone. Of the lamp groups we isolated in the lower array, only lamp group 8 provided enough radiant energy to heat the wafer to a temperature sufficiently

high for deposition to occur. However, wafer temperature oscillated and was highly nonuniform in this case, causing the data to be unreliable. We focus now on the experiment that tested lamp group 1 from which reliable data was obtained.

Lamp group 1 was isolated and set to 45% of full power which brought the center thermocouple reading to 740 C, sufficiently high for deposition to occur. Silane flow rate was set at 30 sccm. After a five minute deposition period, the wafer was removed and thickness measurements were taken at 100 points on the wafer surface.

Figure 6.5 shows two views of the resulting polysilicon film thickness profile. Thermally activated growth using the isolated lamp group 1 produces a "hill" of polysilicon. The deposition pattern reaches a maximum in a line across the wafer center parallel to the lamps in lamp group 1, and decreases toward the wafer edges. Qualitatively, this result is what we would expect given the geometry of lamp group 1 with respect to the wafer.

The thickness data is then used to compute an empirically determined heat flux intensity profile for lamp group 1 as outlined previously. Figure 6.6 shows the result along with the analytically determined profile for purposes of comparison. The result indicates a reasonable agreement between the analytical model and experimental data.

6.4 Model Reduction: A Comparative Study

In this section, we apply the POD and balanced truncation approaches to derive low-order approximations from the RTP heat transfer control system model. We compare the effectiveness of the two approaches via numerical simulations using the full and reduced RTP control system models.

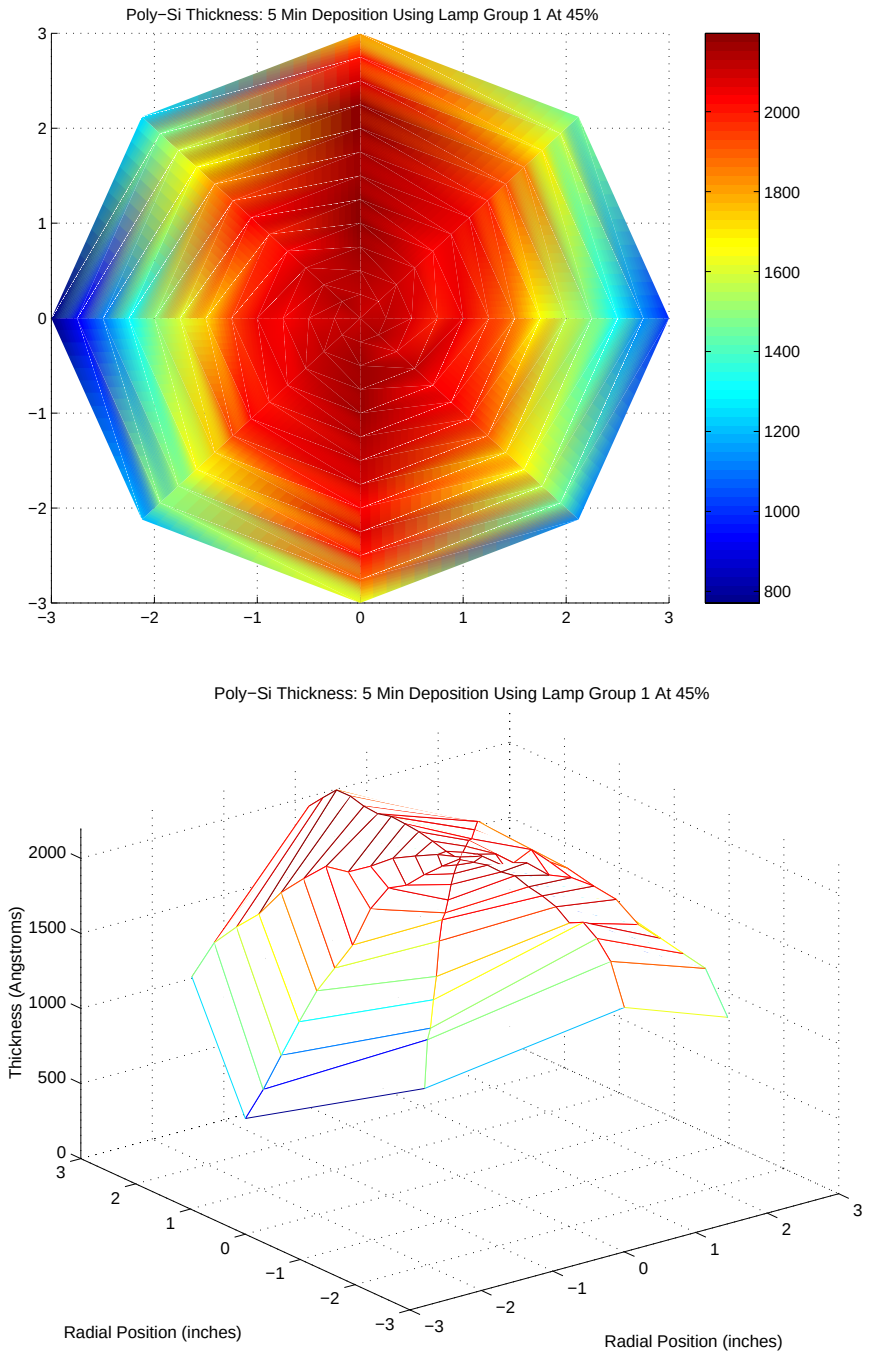


Figure 6.5: Two views of polysilicon film thickness profile resulting from 5 minute deposition using lamp group 1 at 45% power and silane flow rate of 30 sccm. Top figure shows contour map where colors/shades represent thicknesses. Bottom figure shows 3-dimensional view ("hill").

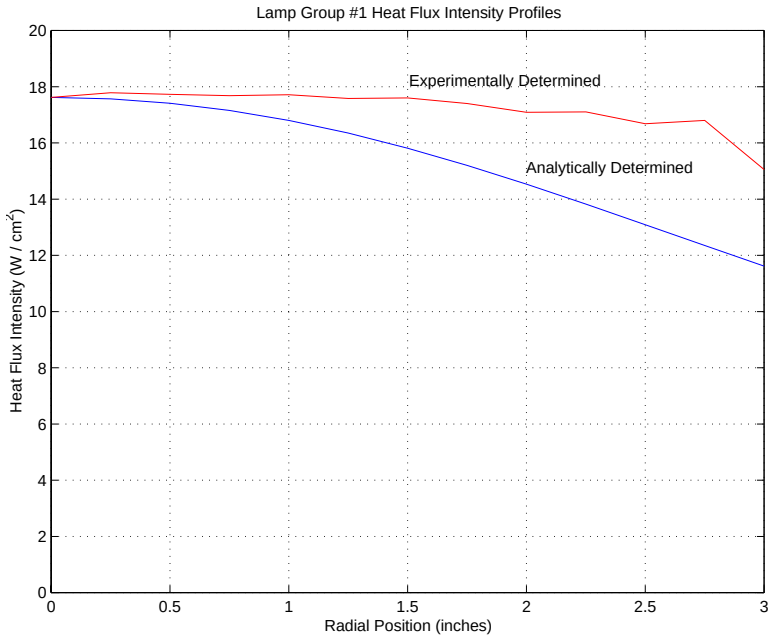


Figure 6.6: Experimentally determined heat flux intensity profile for lamp group 1 along with analytically determined profile for purposes of comparison.

6.4.1 RTP Control System

Recall that in Section 6.2 the evolution of the temperature field on the wafer surface was given by the ODE

$$\dot{T}_w = A_c\, T_w + A_r\, T_w^4 + A_v\, T_w + \Gamma + B\, P \tag{6.35}$$

To model the measurement of temperature at discrete points on the wafer surface via thermocouples, we augment the nonlinear state equation (6.35) with the linear output equation

$$T_{\rm tc} = C\, T_w \tag{6.36}$$

where $T_{\rm tc}$ is a p-vector of thermocouple measurements, and C is a $p \times n$ matrix with entries corresponding to thermocouple locations.

Under our modeling assumptions, there are $m = 3$ independent lamp zone control inputs. Although contact measurement does not occur during an actual

deposition run, we incorporate into the model a set of thermocouple sensors in contact with the wafer, by assuming that there are thermocouples in contact with $p = 3$ locations on the wafer surface: center, edge, and midpoint between center and edge. We ignore the actual placement of thermocouple sensors in the Epsilon-1 susceptor ring.

In Section 6.4.3, we will use a linearized version of (6.35). To linearize, first observe that

$$\dot{x} = A_c\, x + A_r\, (x + \Gamma)^4 - A_r\, \Gamma^4 + A_v\, x + B\, u \tag{6.37}$$

has an equilibrium point at $x = 0$ and is equivalent to (6.35) under the changes of variable $x = T_w - \Gamma$ and $u = P - P_{ss}$, where P_{ss} is the control input that results in a steady state temperature field of $T_w = \Gamma$. Linearizing (6.37) about the origin gives

$$\dot{x} = Ax + Bu \tag{6.38}$$

with

$$A = A_c + A_v + 4F \tag{6.39}$$

where

$$[F]_{ij} = [A_r]_{ij}\ \Gamma_j^3 \tag{6.40}$$

and x and u are translations of T_w and P, respectively. The output equation for the linearized control system is given by

$$y = C\, x \tag{6.41}$$

6.4.2 POD Approach

We follow the procedure for deriving reduced models via the POD method as detailed in Section 3.2.3. To generate empirical time series data, i.e., snapshots of the

wafer temperature field, the nonlinear control system (6.35)-(6.36) was simulated using two different types of control input recipes. (A recipe refers to a function giving the lamp power setting for each of the three lamp zones at each instance of time.) They are referred to as Ramp-Soak-Cool (RSC) and Perturbation-Of-Constant (POC).

Control Input Recipe - Ramp-Soak-Cool

The RSC recipe mimics a typical processing recipe in which a lamp zone power setting is gradually ramped up from zero to full power, maintained at full power for a specified period of time, and then gradually ramped down from full to zero power, as shown in Figure 6.7. This recipe is applied to one of the lamp zones individually, while the other two zones are held at zero power. The RSC simulation is then repeated for each of the other two lamp zones. In this manner, the system response to excitation from an RSC recipe for each of the three lamp zones will appear in the time series data. The time series state-response data is shown in Figure 6.8.

The entire ensemble (three sets) of time series data is combined and arranged into a data matrix, each column of which represents one "snapshot" of the wafer temperature field. The POD basis elements and associated eigenvalues, or relative energy values, are then computed via SVD and ranked according to magnitude of relative energy. The basis elements with the four largest eigenvalues are shown in Figure 6.9. Corresponding relative energy values are compared with those from applying the balancing method in Table 6.2.

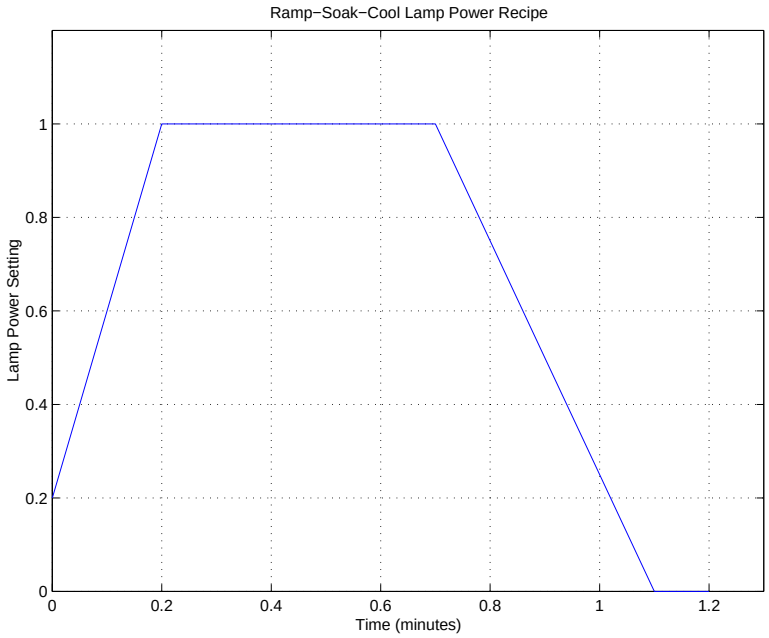


Figure 6.7: Lamp power settings for RSC recipe.

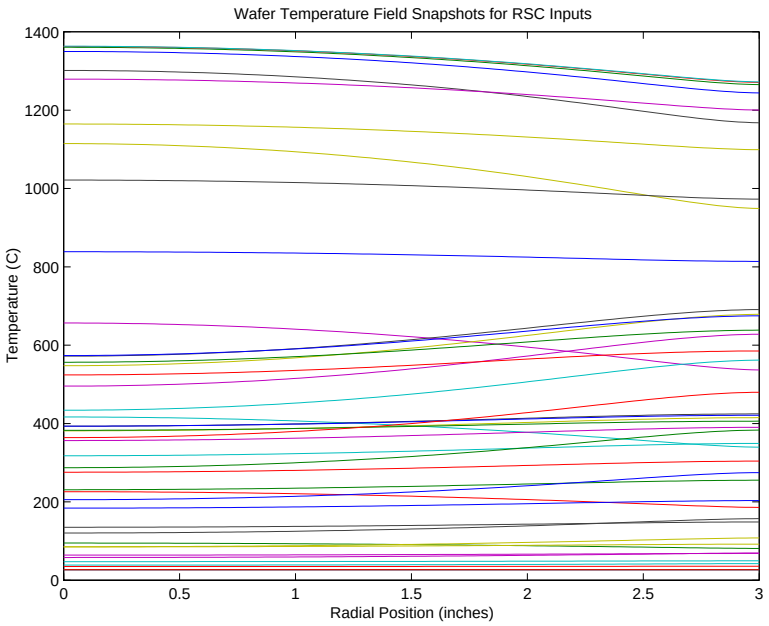


Figure 6.8: Snapshots of wafer temperature field with RSC input and uniform initial temperature.

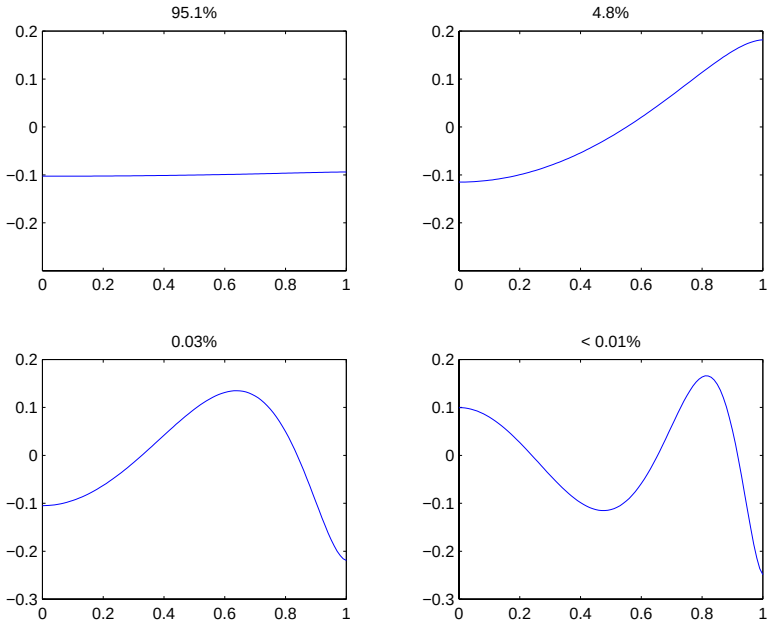


Figure 6.9: Basis elements computed using POD from RSC empirical data.

Control Input Recipe - Perturbation Of Constant

The POC recipe applies small perturbations of a pre-determined set of constant power settings which, if left unperturbed, would result in a uniform steady state temperature field of 1000K. The perturbations are achieved by adjusting the power setting of each lamp zone, one at a time, first to 110% and then to 90%, of the original setting. This results in 6 different control recipes, as shown in Table 6.1. Note that if the nominal constant power settings are used, then the wafer temperature field will evolve as a uniform field for all time. Thus, the perturbations are used to elicit a response that would be characterstic of the system behavior in response to certain types of disturbances.

The system response to excitation from each of the six POC recipes is sampled and combined as the time series data for computing POD basis elements. Time series snapshots are shown in Figure 6.10. Once again, POD basis elements are

Lamp Power Settings for POC Recipe

Note: Settings are constant for all time.

Recipe	Zone 1	Zone 2	Zone 3
P_{unif}	0.0798	0.4265	0.1965
$P^{(1)}$	0.0718	0.4265	0.1965
$P^{(2)}$	0.0878	0.4265	0.1965
$P^{(3)}$	0.0798	0.3838	0.1965
$P^{(4)}$	0.0798	0.4691	0.1965
$P^{(5)}$	0.0798	0.4265	0.1768
$P^{(6)}$	0.0798	0.4265	0.2161

Table 6.1: Lamp power settings for POC recipe.

computed and ranked by corresponding relative energy value. The basis elements with the four largest eigenvalues are shown in Figure 6.9. Corresponding relative energy values are given in Table 6.2.

6.4.3 Balancing Approach

We apply the balanced truncation procedure as detailed in Section 3.3.3 to the linearized control system model (6.38) and (6.41). We note that the realization (A, B, C) is nearly non-minimal, i.e., the condition numbers of the Gramians and their product are

$$\text{cond}(W_c) = 3.2 \times 10^{18} \tag{6.42}$$

$$\text{cond}(W_o) = 3.8 \times 10^{18} \tag{6.43}$$

$$\text{cond}(W_c W_o) = 9.4 \times 10^{18} \tag{6.44}$$

Remark 6.4.1 *The near non-minimality of the RTP control system is expected, since lamp influence functions and initial wafer temperature profiles are always smooth and relatively uniform. Hence, any non-smooth or spatially fluctuating*

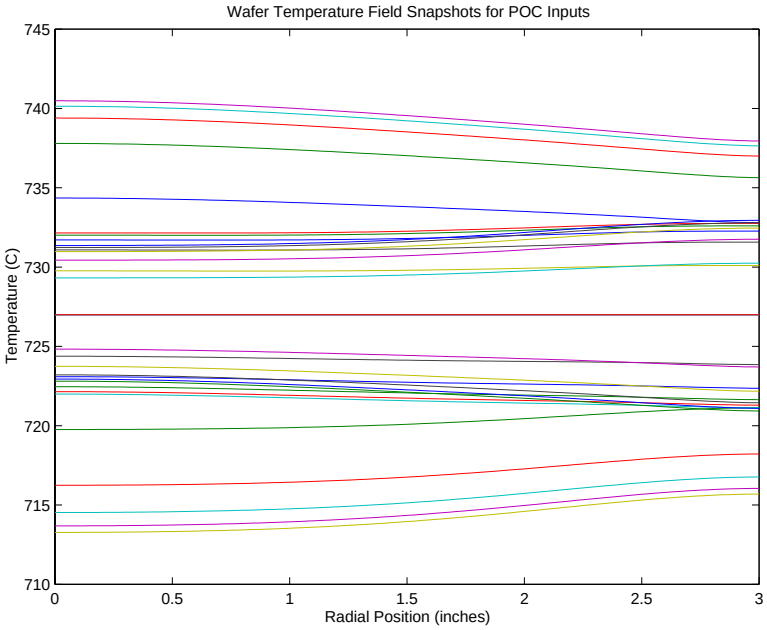


Figure 6.10: Snapshots of wafer temperature field with POC input and uniform initial temperature.

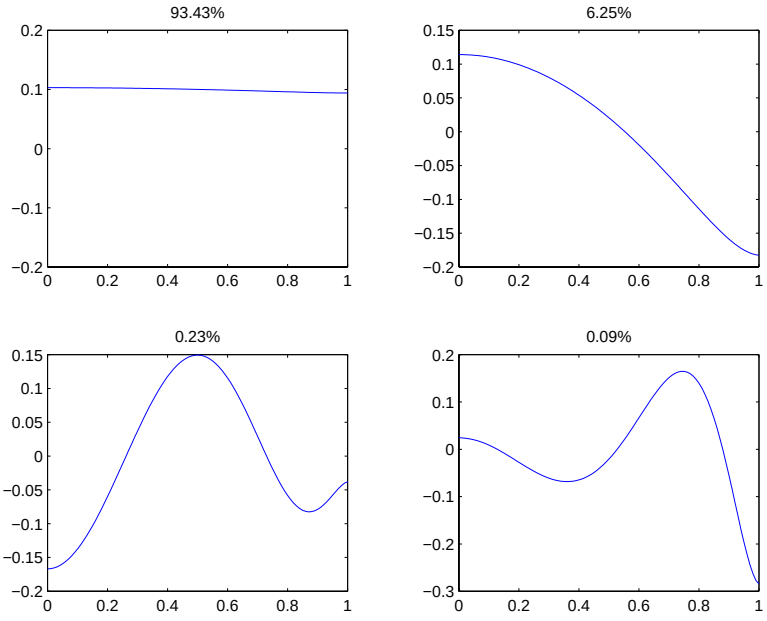


Figure 6.11: Basis elements computed using POD from POC empirical data.

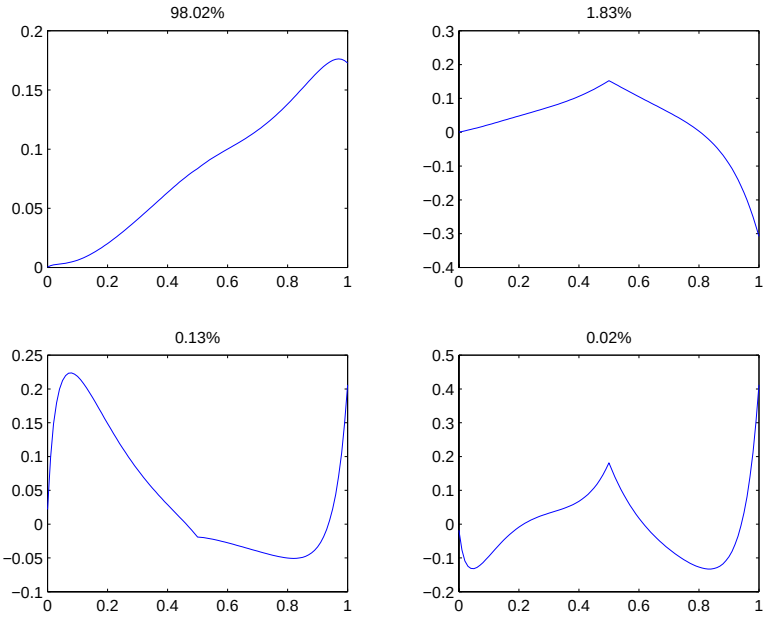


Figure 6.12: Left basis elements for balancing transformation.

temperature profiles are almost impossible to generate using the available control inputs, and likewise, to measure using three thermocouple sensors. □

Thus, it is necessary to use the Schur method of Safonov and Chiang (see Appendix E) in order to alleviate the numerical difficulties. Using this method, we derive a k-th order reduced model that is not necessarily balanced, but has transfer function $\widehat{G}(s)$ which is exactly the same as that for any k-th order balanced realization, thus enjoying the same attractive error bound.

Application of the Schur method to (A, B, C) yields left and right basis elements for a coordinate transformation, shown in Figures 6.12 and 6.13. Corresponding relative energy values are compared with those from the POD approach in Table 6.2. We note that the condition numbers for all of the matrices used in the Schur procedure are less than 1000.

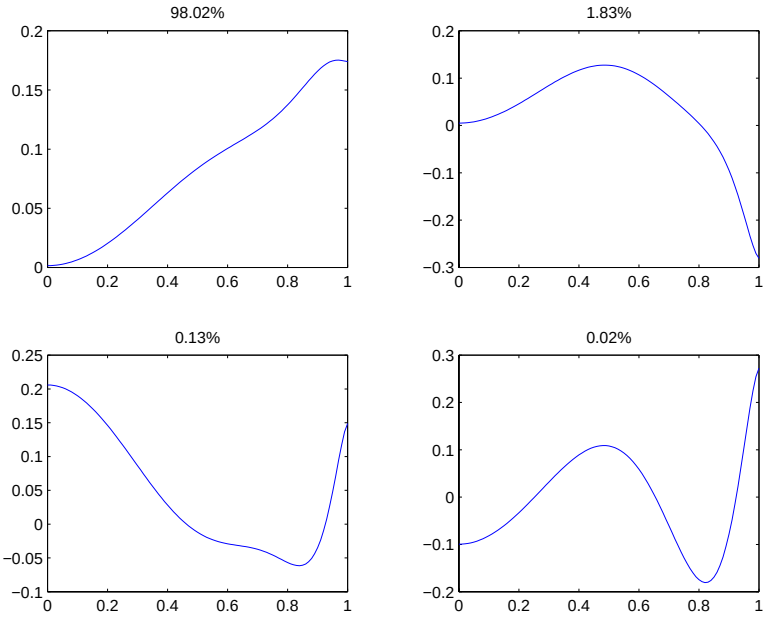


Figure 6.13: Right basis elements for balancing transformation.

6.4.4 Reduced Model Simulations

Validation of the predictive capability of the reduced models is accomplished by comparing simulation results using the original nth-order model with simulation results using reduced kth-order approximations for various values of the reduced model order k. In particular, the maximum deviation of the output signals, i.e., thermocouple readings, between the original and reduced models, are computed for each of the model reduction approaches we have previously described.

The test simulations use a uniform 700 C initial temperature field, and two different test recipes as the lamp control inputs. The test recipes are different from those used to generate the RSC and POC data ensembles.

Test Recipe 1 $\quad P = [0.5\ 0.5\ 0.5] \quad t \in [0, 1]$

Test Recipe 2 $\quad P = [1.0\ 0.0\ 0.0] \quad t \in [0, 0.4)$

$\qquad\qquad\quad P = [0.0\ 1.0\ 0.0] \quad t \in [0.4, 0.7)$

$\qquad\qquad\quad P = [0.0\ 0.0\ 1.0] \quad t \in [0.7, 1.0]$

The k-th order reduced models for $k = 1$, 2, 3, 4, and 5 and full $n = 101$ order model are numerically integrated using identical control recipes and initial states. Simulated thermocouple readings are recorded for each simulation. The reduced model fidelity, i.e., the error between the original and reduced models, is computed as in (1.1, via

$$e(k) = \|T_{\rm tc} - \widehat{T}_{\rm tc}\|_{max}\,, \quad k = 1, 2, 3, 4, 5 \tag{6.45}$$

where we define the norm $\|y\|_{max}$ for time–dependent p–vector $y(t)$ as

$$\|y\|_{max} = \max\{y_i(t)\ :\ 0 \le t < \infty\,,\ 1 \le i \le p\} \tag{6.46}$$

where $y_i(t)$ corresponds to the temperature reading of thermocouple i at time t. Thus, (6.45) gives the maximum deviation between actual and estimated thermocouple readings over the entire simulated time sequence and over all three thermocouples, i.e., a "worst case" error.

Remark 6.4.2 *Due to the shape of the lamp heat flux intensity profiles and the smoothing effect of the diffusion operator, the evolution of the wafer temperature field does not produce especially interesting behavior, e.g., spatial profiles whose fluctuations from the mean vary substantially in the mean square sense from the initial profile, assuming the initial profile is relatively smooth. Thus, we expect little difficulty in capturing the essence of the input-output behavior of the system in a low dimensional model. Our results show that this is indeed the case.* □

Percent Energy Associated With Transformation Basis Elements

Method	Mode 1	Mode 2	Mode 3	Mode 4	Mode 5
POD RSC	95.06	4.77	0.14	0.03	0.00
POD POC	93.43	6.25	0.23	0.09	0.01
Balancing	98.02	1.83	0.13	0.02	0.00

Table 6.2: Normalized eigenvalues, i.e., percent energy, corresponding to basis elements used in model reduction for POD method with RSC data, POD method with POC data, and balancing approach.

Tables 6.2 and 6.3 give the relative energy values for basis elements, and the maximum thermocouple temperature deviations for the original and reduced order models. Figures 6.14, 6.15, and 6.16 show simulated thermocouple readings resulting from simulations with test recipe 1, for the original $n = 101$ order model, and reduced models of order $k = 1$, 2, and 3. Figures 6.17, 6.18, and 6.19 show simulated thermocouple readings resulting from simulations with test recipe 2.

We observe that the output responses of the full and reduced systems are similar. In particular, using the test recipes as control inputs, the input-output behavior of the wafer heat transfer system can be reconstructed using reduced models of order 4 so that thermocouple readings are within 1 degree C of the readings using the original model. This holds whether the POD or balancing method is used, and for whichever set of empirical data was used for computing the POD transformation. Even reduced models of order 2 produce a reasonable approximation (but not suitable for control applications) with "worst case" errors less than 15 degrees C.

Maximum deviation (degrees C) between outputs of original and reduced models

Simulation	Reduction Method	Reduced Model Order 1	2	3	4	5
Test 1	POD RSC	27.23	2.68	0.58	0.11	0.01
	POD POC	26.85	1.26	1.13	0.10	0.05
	Balancing	50.68	7.03	0.44	0.08	0.02
Test 2	POD RSC	72.33	5.22	1.48	0.18	0.05
	POD POC	72.60	4.79	4.35	0.43	0.10
	Balancing	80.81	14.28	1.70	0.12	0.04

Table 6.3: Maximum deviation (degrees C) between outputs of original and reduced models for POD method with RSC data, POD method with POC data, and balancing approach.

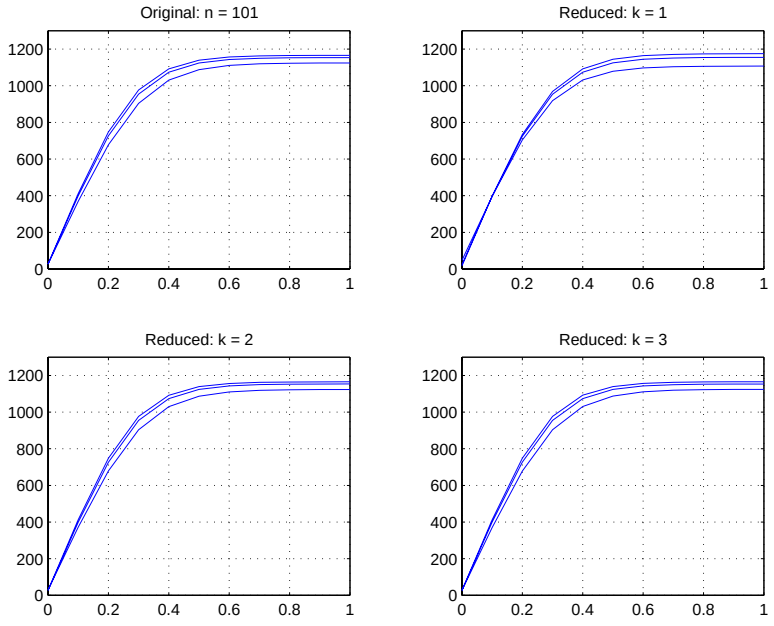


Figure 6.14: Thermocouple readings for original and reduced models with Test Recipe 1 using transformation from POD RSC.

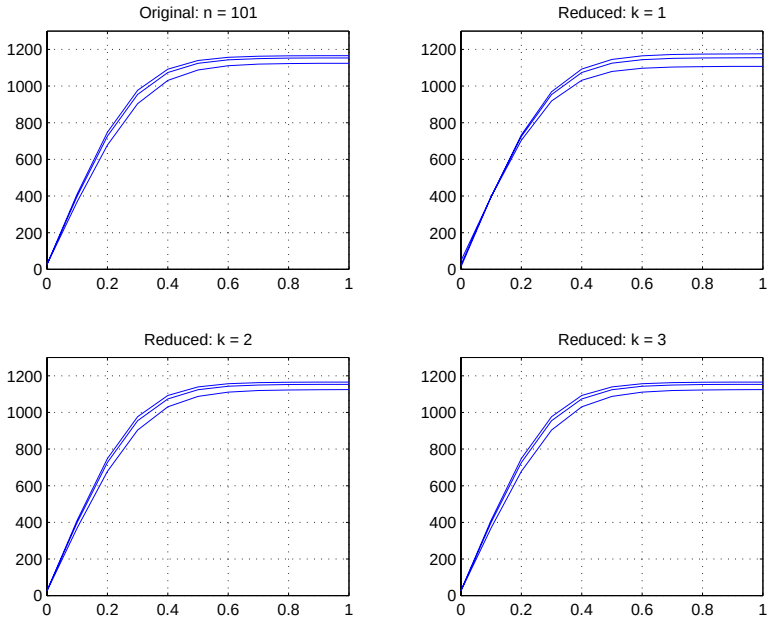


Figure 6.15: Thermocouple readings for original and reduced models with Test Recipe 1 using transformation from POD POC.

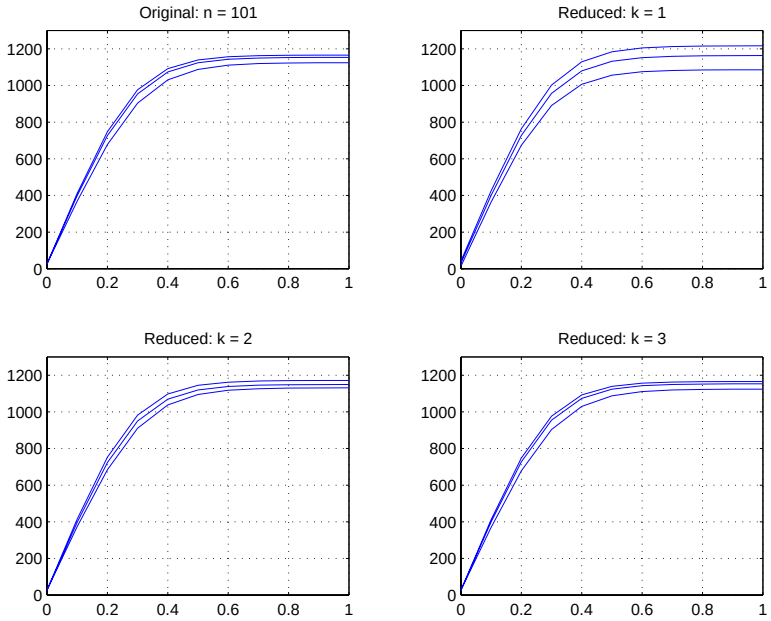


Figure 6.16: Thermocouple readings for original and reduced models with Test Recipe 1 using balancing transformation.

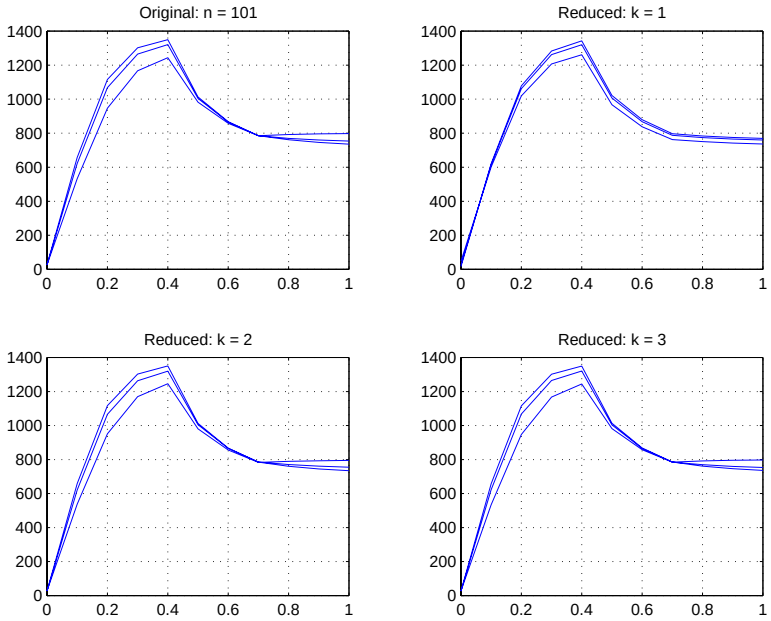


Figure 6.17: Thermocouple readings for original and reduced models with Test Recipe 2 using transformation from POD RSC.

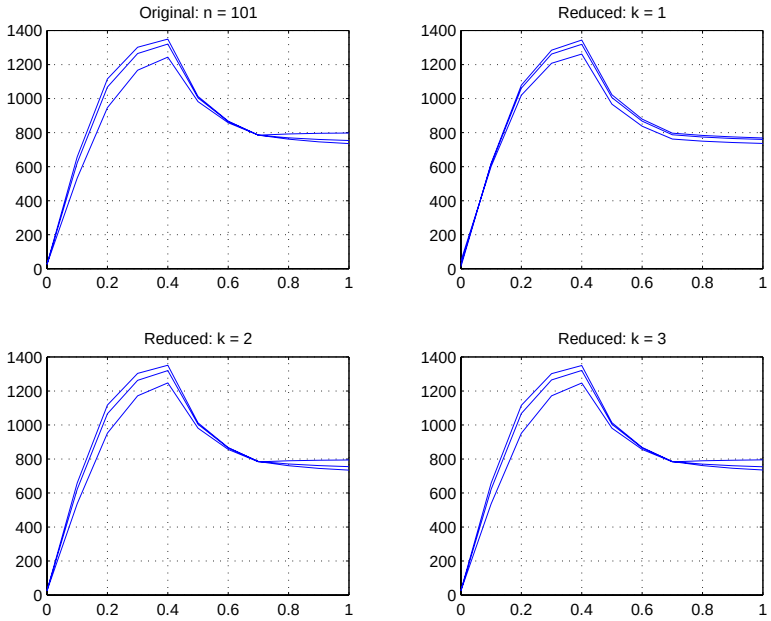


Figure 6.18: Thermocouple readings for original and reduced models with Test Recipe 2 using transformation from POD POC.

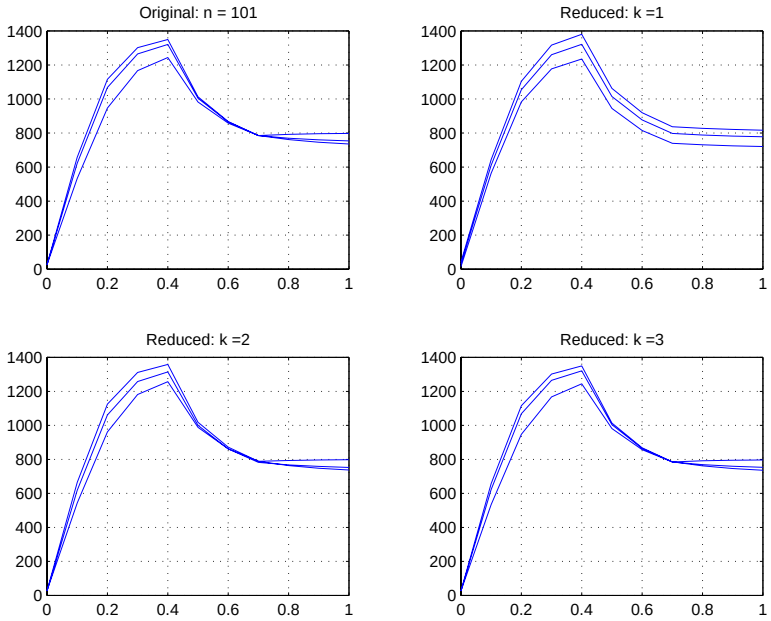


Figure 6.19: Thermocouple readings for original and reduced models with Test Recipe 2 using balancing transformation.

6.5 Remarks

We have presented physics-based and computational models for heat transfer in a silicon wafer, specifically pertaining to RTP in the Epsilon-1 reactor. The models account for the effects of conduction within the solid wafer, convective losses to the gas phase, and radiative losses to the ambient. We model radiative heat transfer from lamps to wafer by determining spatial profiles of radiant heat flux intensity for individual lamp groups and lamp zones. These radiant intensity profiles were computed analytically using view factor methods and then validated using data from poly-Si growth experiments. The model does not account for the effects of gas flow patterns, gas phase heat transfer, gas phase chemical reactions, and other phenomena affecting the rate and spatial distribution of transport of chemical species to the wafer surface, as detailed in Chapter 5.

The complexity and high computational demands of RTP models motivated

a study of model reduction techniques to derive low-order approximations. A comparative study of model reduction approaches examined the POD and balanced truncation methods, which were applied to the RTP control system model. Numerical simulations demonstrated that both the POD and balancing methods produced a change of coordinates that allows for a significant reduction in model dimensionality.

The POD method appears to have performed slightly better than the balancing method in this study, although both performed well. One reason for this result is that the balancing transformation was computed for the linearized system, while the validation tests were performed for the reduced order nonlinear system. Another reason is the relatively simple input-output behavior of this particular system, i.e., there is little difficulty in capturing the essential system behavior with time series state-response data. The empirical eigenfunctions of the flow, and their efficiency for purposes of representation, are relatively insensitive to the choice of inputs. However, the results are not decisive in terms of determining which of the two methods studied was more effective in reducing the order of the RTP control system.

Chapter 7

Conclusions and Future Research

We have motivated problems in state-space model reduction with a discussion and analysis of prominent state-of-the-art approaches, demonstrating the ad-hoc nature of their application to deriving low-order models for nonlinear systems. In the process we have emphasized computational issues and potential hazards in the hope that the exposition may serve as a useful guide.

In light of shortcomings associated with the aforementioned methodologies, we addressed the problem of computability pertaining to the Scherpen theory and procedure for balancing of nonlinear systems. We developed useful methods, tools, and algorithms to compute the associated energy functions and coordinate transformations. We applied our approach to derive, for the first time, balanced representations of nonlinear state-space models.

Because of the computational complexity of our algorithms, this research merely represents a first step toward making the balancing procedure a practical reduction tool. There is little use for order reduction of state-space models with four or fewer states. Faster algorithms will be required to balance and truncate high-order systems of interest to engineers and scientists. Moreover, we have explored

only a limited class of systems when seeking exact formulas for the controllability function. Clearly it would be beneficial to find results with broader applicability.

Our research in modeling of RTCVD for silicon growth reflects our focus on problems of practical interest to our industrial partner. Simulations using the process-equipment model provide for a certain degree of convergence on a suitable set of operating conditions for a particular process, thus avoiding some of the costly experimental trials. Furthermore, there is value in an enhanced understanding of the factors that influence deposition rate and uniformity. The economic advantages associated with accurate prediction of processing results and successful implementation of model-based control in the semiconductor industry will continue to drive the torrent of research in this area. Of particular interest for Si-Ge epitaxy on wafers with a pre-deposited oxide pattern, such as that performed by NG-ESSS, is recent work toward the integration of atomic level models for crystal growth with macroscale models for gas phase transport phenomena [131].

We have derived low-order models for an RTP heat transfer control system using ad-hoc versions of the POD and balanced truncation approaches. Although effective in this case, we believe that ultimately there is much to be gained from development of a more systematic methodology. Further research toward the practical application of balanced truncation for nonlinear systems appears to be a worthy goal.

Appendix A

Notation

Definition	Remarks
$\underline{n} \stackrel{\triangle}{=} \{1, \ldots, n\}$	set of natural numbers between 1 and positive integer n
$\sum_{i=1}^{\infty} a_i \stackrel{\triangle}{=} \lim_{N\to\infty} \sum_{i=1}^{N} a_i$	infinite series
$A = [A]_{ij} \stackrel{\triangle}{=} [a_{ij}]$	matrix with numbers or functions a_{ij} in the i-th row and j-th column
$\square$	end of definitions, theorems, remarks, etc.
$\blacksquare$	end of proofs

Definition	Remarks
$\dot{x} \triangleq \dfrac{d}{dt}\, x$	time derivative
$\dfrac{\partial f}{\partial x} \triangleq \left[\dfrac{\partial f}{\partial x_1}, \ldots, \dfrac{\partial f}{\partial x_n}\right]$	derivative vector
$\nabla f \triangleq \left[\dfrac{\partial f}{\partial x}\right]^{\mathsf{T}}$	gradient vector
$\nabla \cdot f \triangleq \sum_{i=1}^{n} \dfrac{\partial f}{\partial x_i}$	divergence
$\triangle f \triangleq \sum_{i=1}^{n} \dfrac{\partial^2 f}{\partial {x_i}^2} = \nabla \cdot \nabla f$	Laplacian
$Df \triangleq [Df]_{ij} = \left[\dfrac{\partial f_i}{\partial x_j}\right]$	derivative matrix (vector)
$D^2 f \triangleq \left[D^2 f\right]_{ij} = \left[\dfrac{\partial^2 f}{\partial x_i\, \partial x_j}\right]$	second derivative matrix (Hessian)
$x(t, t_0, x_0, \bar{u})$	solution of $\dot{x} = f(t, x, u)$ with $x(t_0) = x_0$ and $u = \bar{u}$
$x(\infty) \triangleq \lim_{t \to \infty} x(t)$	steady-state
$x(-\infty) \triangleq \lim_{t \to -\infty} x(t)$	steady-state (reverse-time system)

Hilbert Spaces

Notation	Remarks
$\mathcal{L}_2(a,b)$	square-integrable functions
ℓ_2	square-summable sequences
$\mathcal{H}_X$	linear operations on second-order random variable X

Norms

Norm	Argument	Definition
$\mathcal{L}_2(a,b)$	$f \in \mathcal{L}_2(a,b)$	$\| f \|_{\mathcal{L}_2(a,b)} = \left(\int_a^b \| f(t) \|^2 \, dt \right)^{1/2}$
Hankel	$G(s)$ stable	$\| G \|_H = \sup_{u \in \mathcal{L}_2(-\infty,0)} \frac{\| y \|_{\mathcal{L}_2(0,\infty)}}{\| u \|_{\mathcal{L}_2(-\infty,0)}}$
H_∞	$G(s)$ stable transfer function	$\| G \|_\infty = \sup_{\omega \in \mathbb{R}} \lambda_{max}^{1/2} \left(G(-j\omega)^\mathsf{T} G(j\omega) \right)$
Froebenius	$A \in \mathbb{R}^{n \times m}$	$\| A \|_F = \left(\sum_{i=1}^n \sum_{j=1}^m a_{ij}^2 \right)^{1/2} = \left(\mathrm{tr} \left(A \, A^\mathsf{T} \right) \right)^{1/2}$

Appendix B

Manifolds and Coordinates

We assume throughout this thesis that the state-space takes the form of a smooth manifold. A smooth manifold is a set which locally can be identified with $\mathbb{R}^n$ together with the intrinsic notion of differentiability defined on $\mathbb{R}^n$. In order to work with systems that evolve on a smooth manifold, we need the concepts of local coordinates, local representative of a map, and coordinate transformation. These and other related terminology are defined below.

The material contained in this section is standard. The following exposition is drawn from texts by Nijmeijer and van der Schaft [121], Isidori [69], and class notes in Geometric Control presented by Dayawansa [37] and Krishnaprasad [87] at the University of Maryland.

First we define terminology needed for characterizing functions of several real variables, i.e., functions defined on $\mathbb{R}^n$.

Definition B.0.1 (Homeomorphism) *A function $f : A \subset \mathbb{R}^n \to B \subset \mathbb{R}^n$ is said to be a* homeomorphism *if it is bijective (one-to-one and onto), and both f and f^{-1} are continuous.* □

Definition B.0.2 (Smooth Function) *Let A be an open subset of $\mathbb{R}^n$. A function $f : A \subset \mathbb{R}^n \to \mathbb{R}^m$ is said to be C^k or k* times continuously differentiable *if all mixed partial derivatives for $j \leq k$*

$$\frac{\partial^j f}{\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n}} \qquad \alpha_i \geq 0\,,\, i \in \underline{n} \qquad \sum_{i=1}^{n} \alpha_i = j\,,$$

exist and are continuous. The function f is said to be C^∞ or smooth *if f is C^k for all k.* □

Definition B.0.3 (Diffeomorphism) *A function $f : A \subset \mathbb{R}^n \to B \subset \mathbb{R}^n$ is said to be a* diffeomorphism *if f is a homeomorphism of A onto B, and both f and f^{-1} are smooth.* □

Definition B.0.4 (Coordinate Function) *The function $r_i : \mathbb{R}^n \to \mathbb{R}$ for $i \in \underline{n}$ defined by*

$$r_i(a_1, \ldots, a_n) = a_i \tag{B.1}$$

is called the i-th coordinate function *or* slot function *on $\mathbb{R}^n$.* □

The following terminology is related to functions defined on, and systems that evolve on, a smooth manifold. Some basic elements of point-set topology are required.

Definition B.0.5 (Topology) *Let M be a non-empty set. A collection T of subsets of M is said to be a* topology *on M if*

(i) M and the empty set belong to T;

(ii) The union of any number of subsets in T belongs to T;

(iii) The intersection of any two (and hence any finite number) subsets in T belongs to T.

□

The members of T are called *T-open sets*, or simply *open sets*, and the pair (M, T) is called a *topological space*. A *basis* for a topology T on M is a collection $S \subset T$ of open sets in T such that every open set can be wrtiten as a union of members of S. A *neighborhood* of a point p in M is any open set which contains p.

Definition B.0.6 (Hausdorff) *A topological space (M, T) is said to be* Hausdorff *if any two different points p_1 and p_2 have disjoint neighborhoods, i.e., there exist open sets $A_1, A_2 \in T$ such that $p_1 \in A_1$, $p_2 \in A_2$, and $A_1 \cap A_2$ is empty.* □

Definition B.0.7 (Continuous Mapping) *A mapping $F : M_1 \to M_2$ between topological spaces (M_1, T_1) and (M_2, T_2) is said to be* continuous *if $F^{-1}(A) \in T_1$ for all $A \in T_2$.* □

We redefine the notion of homeomorphism in the context of mappings defined on a topological space.

Definition B.0.8 (Homeomorphism) *A mapping $F : M_1 \to M_2$ between topological spaces (M_1, T_1) and (M_2, T_2) is said to be a* homeomorphism *if F is bijective and both F and F^{-1} are continuous.* □

Definition B.0.9 (Topological Manifold) *A Hausdorff topological space (M, T) with a countable basis is said to be a* topological manifold *of dimension n if for any point p in M there exists a homeomorphism ϕ from some neighborhood U of p onto an open subset of $\mathbb{R}^n$.* □

A smooth manifold will be defined as a topological manifold with some additional properties relating to differentiability. We use the following terminology.

Definition B.0.10 (Coordinate Chart) *Let (M, T) be a topological manifold of dimension n. A pair (U, ϕ) with $U \in T$ and ϕ a homeomorphism from U onto an open subset of $\mathbb{R}^n$ is called a* coordinate chart *or* coordinate neighborhood *for (M, T).* □

Definition B.0.11 (Local Coordinates) *Let (U, ϕ) be a coordinate chart for a topological manifold (M, T) of dimension n. The functions defined by*

$$x_i = r_i \circ \phi \qquad i \in \underline{n} \tag{B.2}$$

are called local coordinate functions *for (U, ϕ). For a point $p \in M$, the values $x_1(p), \ldots, x_n(p)$ are called the* local coordinates *of p.* □

Definition B.0.12 (Local Representative) *Let (U, ϕ) be a coordinate chart for a topological manifold (M, T) of dimension n. Let $f : M \to \mathbb{R}$ be a map. The function $\widehat{f} : \phi(U) \subset \mathbb{R}^n \to \mathbb{R}$ defined by*

$$\widehat{f} = f \circ \phi^{-1} \tag{B.3}$$

is called the local representative *of f.* □

Remark B.0.13 *By the definition of the local representative for f, we have*

$$f(p) = \widehat{f}\,(x_1(p), \ldots, x_n(p)) \tag{B.4}$$

for a point $p \in U$. We use the shorthand notation $f(x_1, \ldots, x_n)$ to denote both of the above functions, omitting the caret and suppressing the point-dependence. □

Definition B.0.14 (Coordinate Transformation) *Let (U, ϕ) and (V, ψ) be coordinate charts for a topological manifold (M, T) of dimension n such that $U \cap V$*

is not empty, i.e., the coordinate charts overlap. Let $x_i = r_i \circ \phi$ *and* $z_i = r_i \circ \psi$ *be the corresponding coordinate functions. The map* S *defined by*

$$S = \psi \circ \phi^{-1} : \phi(U \cap V) \to \psi(U \cap V) \tag{B.5}$$

is called the coordinate transformation *from local coordinates* $(x_1, \ldots, x_n)$ *to local coordinates* $(z_1, \ldots, z_n)$ *on* $U \cap V$*.* □

Definition B.0.15 (C^∞-Compatible) *Two coordinate charts* (U, ϕ) *and* (V, ψ) *are said to be* C^∞-compatible *if either* $U \cap V$ *is empty or the coordinate transformation* $S = \phi \circ \psi^{-1}$ *and the inverse coordinate transformation* $S^{-1} = \psi \circ \phi^{-1}$ *are both smooth.* □

Remark B.0.16 *The previous condition on* (U, ϕ) *and* (V, ψ) *is referred to as the* transition *property.* □

Definition B.0.17 (C^∞-Atlas) *Let* (M, T) *be a topological manifold of dimension* n*. An indexed collection of pairwise* C^∞*-compatible coordinate charts* $\mathcal{D} = \{U_\alpha, \phi_\alpha\}_{\alpha \in A}$ *is said to be a* C^∞-atlas *on* M *if* $\cup_{\alpha \in A} U_\alpha = M$*.* □

Remark B.0.18 *The previous condition on* $\mathcal{D}$ *is referred to as the* covering *property.* □

Definition B.0.19 (Maximal C^∞-Atlas) *Let* (M, T) *be a topological manifold of dimension* n*. A* C^∞*-atlas on* M *is said to be* maximal *if any coordinate chart* (V, ψ) *which is* C^∞*-compatible with every* $(U_\alpha, \phi_\alpha) \in \mathcal{D}$ *is also in* $\mathcal{D}$*.* □

Remark B.0.20 *The previous condition on* $\mathcal{D}$ *is referred to as the* maximality *property.* □

Definition B.0.21 (Smooth Manifold) *A topological manifold (M, T) is said to be a* smooth manifold *if there exists a maximal C^∞-atlas on M.* □

Remark B.0.22 *Thus, a smooth manifold possesses the properties of covering, transition, and maximality.* □

Finally, we redefine the notions of a smooth mapping and a diffeomorphism in the context of mappings defined on a smooth manifold.

Definition B.0.23 (Smooth Mapping) *Let M_1 and M_2 be smooth manifolds of dimension n_1 and n_2, respectively. A map $F : M_1 \to M_2$ is said to be* smooth *if for each $p \in M_1$ there exist coordinate charts (U, ϕ) of M_1 about p and (V, ψ) of M_2 about $F(p)$, such that the local representative $\widehat{F} = \psi \circ F \circ \phi^{-1}$ is smooth (as in Definition B.0.2) from $\phi(U) \subset \mathbb{R}^{n_1}$ into $\psi(V) \subset \mathbb{R}^{n_2}$.* □

Definition B.0.24 (Diffeomorphism) *Let M_1 and M_2 be smooth manifolds, both of dimension n. A map $F : M_1 \to M_2$ is said to be a* diffeomorphism *if F is a homeomorphism (as in Definition B.0.8) and both F and F^{-1} are smooth (as in Definition B.0.23).* □

Appendix C

Numerical Simulation of Stochastic Differential Equations

In this appendix we present a numerical scheme used for simulation of the system modeled by the white noise driven differential equation

$$\frac{d}{dt} X_t = f(t, X_t) + \sum_{i=1}^{m} g_i(t, X_t) \, (\zeta_t)_i \tag{C.1}$$

and discrete time approximation of white noise signals. This method appears in [94] and is based on results in [164, 165].

As stated earlier, in order to simulate (C.1) we numerically integrate the SDE

$$\begin{aligned} (dX_t)_i &= \left[f_i(t, X_t) + \frac{1}{2} \sum_{j=1}^{n} \sum_{k=1}^{m} \frac{\partial g_{ik}(t, X_t)}{\partial X_j} \, g_{jk}(t, X_t) \right] dt \\ &+ \sum_{i=1}^{m} g_i(t, X_t) \, (dW_t)_i \qquad i \in \underline{n} \end{aligned} \tag{C.2}$$

The justification for (C.2) leads to a computational method for numerical integration.

For simplicity, we consider the time-invariant, single-input ($m = 1$) case. The results extend without difficulty to the general case. Consider a sequence of Gaus-

sian processes $\left\{\zeta_t^{(n)},\, t \in \mathbb{R}^+\right\}$ which converges in some suitable sense to a white noise, and such that, for each n, the process has a well behaved sample path. Then, for each n, the initial value problem

$$\frac{d}{dt} X_t^{(n)} = f\left(X_t^{(n)}\right) + g\left(X_t^{(n)}\right) \zeta_t^{(n)} \qquad X_0^{(n)} = x_0 \tag{C.3}$$

can be solved, resulting in a sequence of processes $\left\{X_t^{(n)},\, t \in \mathbb{R}^+\right\}$. If the sequence $X_t^{(n)}$ converges to X_t then it is natural to say that X_t is the solution of (C.1).

The desired sequences are derived as follows. Consider a partition $0 = t_0 < t_1 < \cdots < t_r$ of the interval of integration, with maximum step size

$$\Delta = \max\left(t_{i+1} - t_i\right) \tag{C.4}$$

For each integration step, define a polygonal approximation of W_t

$$W_t^{\Delta} = W_{t_i}^{\Delta} + \frac{W_{t_{i+1}}^{\Delta} - W_{t_i}^{\Delta}}{t_{i+1} - t_i}\left(t - t_i\right) \qquad t_i \le t \le t_{i+1} \tag{C.5}$$

and a corresponding approximation to dW_t

$$dW_t^{\Delta} = W_{t_i}^{\Delta} - W_{t_{i+1}}^{\Delta} \qquad t_i \le t \le t_{i+1} \tag{C.6}$$

Then, since the polygonal approximation is piecewise differentiable with probability 1, the equation

$$dX_t^{\Delta} = f\left(X_t^{\Delta}\right) dt + g\left(X_t^{\Delta}\right) dW_t^{\Delta} \tag{C.7}$$

is an ODE for the sample function X_t^{Δ}. It is shown by Wong and Zakai [164, 165] that, under certain conditions,

$$\lim_{\Delta \to 0} \text{in q.m. } X_t^{\Delta} = X_t \tag{C.8}$$

where X_t is the unique solution of the SDE (C.2).

With the above justification in hand, we proceed to show how we can approximate a white noise process by a discrete-time signal, i.e., a sequence of random

numbers, for purposes of numerical integration. The key detail is in choosing the statistics of the random numbers in a manner consistent with the approximation scheme.

Consider the discrete-time approximation to a continuous-time white noise process

$$\zeta_t^\Delta = \frac{d}{dt} W_t^\Delta = \lim_{h\to 0} \frac{W_{t+h}^\Delta - W_t^\Delta}{h} \approx \frac{W_{t_{i+1}}^\Delta - W_{t_i}^\Delta}{\Delta} \qquad t_i \le t \le t_{i+1} \tag{C.9}$$

If Δ is sufficiently small then ζ_t^Δ is Gaussian with zero mean and variance $\frac{1}{\Delta}$. To see this, observe that, by definition of a Wiener process

$$E\left[\, W_{t_i} \,\right] = 0 \qquad E\left[\, W_{t_i}\, W_{t_{i+1}} \,\right] = t_i \qquad E\left[\, W_{t_{i+1}}\, W_{t_{i+1}} \,\right] = t_{i+1} \qquad i = 1, 2, \ldots \tag{C.10}$$

Thus $\mathrm{Var}\left(W_{t_{i+1}} - W_{t_i}\right) = \Delta$ and

$$\mathrm{Var}\left(\frac{W_{t_{i+1}}^\Delta - W_{t_i}^\Delta}{\Delta}\right) = \frac{1}{\Delta} \qquad i = 1, 2, \ldots \tag{C.11}$$

The desired discrete time signal is a sequence of Gaussian random variables with zero mean and variance $\frac{1}{\Delta}$.

It is typical that one has access to a random number generator that can generate a sequence of zero mean unit variance Gaussian random variables $\{Z_k,\ k = 1, 2, \ldots\}$. In this case, we set

$$dW_k = \left(\frac{1}{\Delta}\right)^{1/2} Z_k \qquad k = 1, 2, \ldots \tag{C.12}$$

as a discrete time approximation to dW_t within a suitable scheme for numerical integration of (C.2). A comparative study of numerical integration schemes for SDEs appears in [167]. We used a 4th-order Runge-Kutta scheme to integrate (C.2) with appropriately chosen time step Δ. For random number generation, we used the built-in linear congruential generator of MATLAB. According to the MATLAB

manual, it can generate all floating point numbers in the range $[2^{-53}, 2^{53}]$ and produce 2^{1492} values before repeating. This was easily sufficient for our purposes.

Appendix D

Proof of the Proper Orthogonal Decomposition Theorem

Continuous Parameter POD (Theorem 3.2.5)

Proof For simplicity we assume without loss of generality that the process $\{X_t,\, t \in [a,b]\}$ is scalar-valued. Suppose that the functions $\{\phi_1, \phi_2, \ldots\}$ satisfy the integral equation (3.10) and that the random variables $\{a_1, a_2, \ldots\}$ are defined by (3.11) so that the orthonormality condition (3.8) holds. Define for each $N = 1, 2, \ldots$

$$S_N(t) \stackrel{\triangle}{=} \sum_{i=1}^{N} \sqrt{\lambda_i}\, a_i\, \phi_i(t) \tag{D.1}$$

The POD (3.7) is equivalent to the statement

$$\lim_{N\to\infty} \| X_t - S_N(t) \|^2_{\mathcal{H}_X} \stackrel{\triangle}{=} \lim_{N\to\infty} E\left[\, |X_t - S_N(t)|^2 \,\right] \equiv 0 \tag{D.2}$$

Observe that

$$E\left[\, |X_t - S_N(t)|^2 \,\right] = E\left[\, |X_t|^2 \,\right] + E\left[\, |S_N(t)|^2 \,\right] - 2\, E\left[\, X_t\, S_N(t) \,\right] \tag{D.3}$$

Computing the terms in (D.3) we get

$$
\begin{aligned}
E\left[\,|X_t|^2\,\right] &= R(t,t) && \text{(D.4)}\\
E\left[\,|S_N(t)|^2\,\right] &= E\left[\sum_{i=1}^{N}\sqrt{\lambda_i}\,a_i\,\phi_i(t)\sum_{j=1}^{N}\sqrt{\lambda_j}a_j\phi_j(t)\right]\\
&= \sum_{i=1}^{N}\sum_{j=1}^{N}\sqrt{\lambda_i}\,\sqrt{\lambda_j}\,E\left[\,a_i\,a_j\,\right]\,\phi_i(t)\,\phi_j(t)\\
&= \sum_{i=1}^{N}\sum_{j=1}^{N}\sqrt{\lambda_i}\,\sqrt{\lambda_j}\,\phi_i(t)\,\phi_j(t)\,\delta_{ij}\\
&= \sum_{i=1}^{N}\lambda_i\,|\phi_i(t)|^2 && \text{(D.5)}\\
E\left[\,X_t\,S_N(t)\,\right] &= E\left[X_t\sum_{i=1}^{N}\sqrt{\lambda_i}\,a_i\,\phi_i(t)\right]\\
&= \sum_{i=1}^{N}\sqrt{\lambda_i}\,\phi_i(t)\,E\left[\,X_t\,a_i\,\right]\\
&= \sum_{i=1}^{N}\sqrt{\lambda_i}\,\phi_i(t)\,E\left[X_t\left(\sqrt{\lambda_i}\right)^{-1}\int_a^b\phi_i(s)\,X_s\,ds\right]\\
&= \sum_{i=1}^{N}\phi_i(t)\int_a^b\phi_i(s)\,E\left[\,X_t\,X_s\,\right]\,ds\\
&= \sum_{i=1}^{N}\phi_i(t)\int_a^b R(t,s)\,\phi_i(s)\,ds\\
&= \sum_{i=1}^{N}\phi_i(t)\,\left(\lambda_i\,\phi_i(t)\right)\\
&= \sum_{i=1}^{N}\lambda_i\,|\phi_i(t)|^2 && \text{(D.6)}
\end{aligned}
$$

Substituting evaluated terms into equation (D.3) yields

$$
\begin{aligned}
E\left[\,|X_t - S_N(t)|^2\,\right] &= R(t,t) + \sum_{i=1}^{N}\lambda_i\,|\phi_i(t)|^2 - 2\sum_{i=1}^{N}\lambda_i\,|\phi_i(t)|^2\\
&= R(t,t) - \sum_{i=1}^{N}\lambda_i\,|\phi_i(t)|^2 && \text{(D.7)}
\end{aligned}
$$

The covariance function $R(\cdot,\cdot)$ is Hermitian symmetric and nonnegative definite by Propositions (2.5.19) and (2.5.20). It is also continuous on $[a,b]\times[a,b]$ by

q.m. continuity of $\{X_t,\, t \in [a,b]\}$. Therefore, the hypotheses of Mercer's theorem hold and the spectral decomposition exists, given by $R(t,s) = \sum_{i=1}^{\infty} \lambda_i \phi_i(t) \phi_i(s)$. Consequently

$$R(t,t) = \sum_{i=1}^{\infty} \lambda_i \, |\phi_i(t)|^2 \tag{D.8}$$

where convergence is uniform for $t \in [a,b]$. Equations (D.7) and (D.8) together imply the desired result, i.e., $\lim_{N\to\infty} E[|X_t - S_N(t)|^2] = 0$.

Conversely, suppose $\{X_t,\, t \in [a,b]\}$ has the stated expansion. Then,

$$\begin{aligned} R(t,s) &= E\,[\,X_t\,X_s\,] \\ &= E\left[\sum_{i=1}^{\infty} \sqrt{\lambda_i}\, a_i\, \phi_i(t) \sum_{j=1}^{\infty} \sqrt{\lambda_j}\, a_j\, \phi_j(t)\right] \\ &= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \sqrt{\lambda_i}\,\sqrt{\lambda_j}\, E\,[\,a_i\,a_j\,]\;\phi_i(t)\,\phi_j(t) \\ &= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \sqrt{\lambda_i}\,\sqrt{\lambda_j}\,\phi_i(t)\,\phi_j(t)\,\delta_{ij} \\ &= \sum_{i=1}^{\infty} \lambda_i\,\phi_i(t)\,\phi_i(s) \end{aligned} \tag{D.9}$$

Therefore,

$$\begin{aligned} \int_a^b R(t,s)\,\phi_i(s)\,ds &= \int_a^b \left(\sum_{j=1}^{\infty} \lambda_j\,\phi_j(t)\,\phi_j(s)\right)\,\phi_i(s)\,ds \\ &= \sum_{j=1}^{\infty} \lambda_j\,\phi_j(t)\int_a^b \phi_j(s)\,\phi_i(s)\,ds \\ &= \sum_{j=1}^{\infty} \lambda_j\,\phi_j(t)\,\delta_{ji} \\ &= \lambda_i\,\phi_i(t) \end{aligned} \tag{D.10}$$

■

Discrete Parameter POD (Theorem 3.2.16)

The POD theorem can be proved in the discrete parameter case in a similar fashion to the proof in the continuous parameter case, where the spectral theorem is used

instead of Mercer's theorem. However, we offer the following, simpler approach.

Proof Observe that the sampled data POD is written compactly in equation (3.29). Substitution of (3.30) together with orthogonality of Φ yields the desired identity. ■

Appendix E

Algorithms for Linear Balancing

Given a minimal realization (A, B, C) with A stable, a balanced realization $\left({S_{\text{bal}}}^{-1} A\, S_{\text{bal}}, {S_{\text{bal}}}^{-1} B, C\, S_{\text{bal}}\right)$ such that (3.89) holds can be obtained through the following standard algorithm:

Algorithm E.0.25 (Laub [90])

(L1) *Compute W_c and W_o (solve Lyapunov equations (3.83) and (3.84) via the Bartels-Stewart algorithm [14]).*

(L2) *Compute a matrix L_c such that $W_c = L_c\, {L_c}^{\mathsf{T}}$ (Cholesky decomposition).*

(L3) *Form the matrix ${L_c}^{\mathsf{T}}\, W_o\, L_c$ (matrix multiplications).*

(L4) *Compute an orthogonal matrix U and a diagonal matrix Σ such that ${L_c}^{\mathsf{T}}\, W_o\, L_c = U\, \Sigma^2\, U^{\mathsf{T}}$ (spectral decomposition).*

(L5) *Form the matrices $S_{\text{bal}} = L_c\, U\, \Sigma^{-1/2}$ and $S_{\text{bal}}^{-1} = \Sigma^{1/2}\, U^{\mathsf{T}}\, {L_c}^{-1}$ (matrix multiplications, matrix inversion).*

(L6) Form the balanced state-space matrices $\left({S_{\rm bal}}^{-1} A\, S_{\rm bal}, {S_{\rm bal}}^{-1} B, C\, S_{\rm bal}\right)$ *(matrix multiplications).*

□

Remark E.0.26 *An algorithm presented by Moore [109] is essentially the same except that a spectral decomposition replaces the Cholesky decomposition in step (L2).* □

The following improvement to the standard algorithm is more elegant numerically in that it computes the Cholesky factors without actually solving the Lyapunov equations for the Gramians, and computes the SVD of a product of Cholesky factors without explicitly forming their product. This is important in computing small singular values accurately.

Algorithm E.0.27 (Laub, et.al. [91])

(LH1) Compute matrices L_c *and* L_o *such that* $W_c = L_c\, {L_c}^{\mathsf{T}}$ *and* $W_o = L_o\, {L_o}^{\mathsf{T}}$*. The Cholesky decompositions are performed without ever forming the Gramians via the algorithm of Hammarling [64].*

(LH2) Compute orthogonal matrices U, V *and a diagonal matrix* Σ^2 *such that* ${L_o}^{\mathsf{T}}\, L_r = U\, \Sigma^2\, V^{\mathsf{T}}$*. The SVD is performed without ever forming the product* ${L_o}^{\mathsf{T}}\, L_r$ *via the algorithm of Heath, et.al. [65].*

(LH3) Form the matrices $S_{\rm bal} = L_c\, V\, \Sigma^{-1/2}$ *and* $S_{\rm bal}^{-1} = \Sigma^{-1/2}\, U^{\mathsf{T}}\, {L_o}^{\mathsf{T}}$ *(matrix multiplications).*

(LH4) Form the balanced state-space matrices (matrix multiplications).

□

Remark E.0.28 *The computational complexity of these algorithms are roughly the same, i.e., $O(n^3)$ with pre-multiplier in the 40-50 range.* □

Nearly Non-Minimal Systems

We now present an algorithm for dealing with nearly non-minimal systems.

Algorithm E.0.29 (Safonov and Chiang [136])

(SC1) Compute $n \times k$ matrices $V_{r,big}$ and $V_{l,big}$ whose columns form bases for the respective right and left eigenspaces of $W_c W_o$ associated with the "big" eigenvalues $\sigma_1^2, \ldots, \sigma_k^2$. This is done via the ordered Schur decomposition of $W_c W_o$ as follows.

(a) Compute W_c and W_o.

(b) Compute an orthogonal matrix V such that $V W_c W_o V^{\mathsf{T}}$ is upper triangular, i.e., put $W_c W_o$ into Schur form. The fact that W_c and W_o are real and symmetric ensures the existence of a real Schur transformation matrix V.

(c) Compute orthogonal transformations V_a and V_d which order the Schur forms in ascending and descending order, respectively,

$$V_a^T W_c W_o V_a = \mathrm{diag}(\lambda_{a,n}, \ldots, \lambda_{a,1}) + T_a \tag{E.1}$$

$$V_d^T W_c W_o V_d = \mathrm{diag}(\lambda_{d,1}, \ldots, \lambda_{d,n}) + T_d \tag{E.2}$$

where T_a and T_d are strictly upper triangular and such that

$$\{\lambda_{a,1}, \ldots, \lambda_{a,k}\} = \{\lambda_{d,1}, \ldots, \lambda_{d,k}\} = \{\sigma_1^2, \ldots, \sigma_k^2\} \tag{E.3}$$

$$\{\lambda_{a,k+1}, \ldots, \lambda_{a,n}\} = \{\lambda_{d,k+1}, \ldots, \lambda_{d,n}\} = \{\sigma_{k+1}^2, \ldots, \sigma_n^2\} \tag{E.4}$$

(d) Partition V_a and V_d as

$$V_a = \left[\overbrace{V_{r,small}}^{n-k} \,|\, \overbrace{V_{l,big}}^{k} \right] \tag{E.5}$$

$$V_d = \left[\overbrace{V_{r,big}}^{k} \,|\, \overbrace{V_{l,small}}^{n-k} \right] \tag{E.6}$$

(SC2) Form $E_{\text{big}} = V_{\text{l,big}}^T V_{\text{r,big}}$ and compute the SVD

$$E_{\text{big}} = U_{\text{E,big}} \Sigma_{\text{E,big}} V_{\text{E,big}}^T \tag{E.7}$$

(SC3) The not necessarily balancing transformations are

$$S_{\text{l,big}} = V_{\text{l,big}} U_{\text{E,big}} \Sigma_{\text{E,big}}^{-1/2} \tag{E.8}$$

$$S_{r,big} = V_{r,big} V_{E,big} \Sigma_{E,big}^{-1/2} \tag{E.9}$$

(SC4) The reduced model is given by

$$(\widehat{A}, \widehat{B}, \widehat{C}) = (S_{l,big}^T A S_{r,big}, S_{l,big}^T B, C S_{r,big}) \tag{E.10}$$

□

Gradient Flow Methods

Finally, we briefly describe the gradient flow method of Helmke and Moore [66]. Consider the usual linear transformation $x = S\,z$ and let $T = S^{-1}$ so that $z = T\,x$. To get a quantitative measure of how the Gramians change under the transformation, the authors use the cost function

$$\phi\,(T) = \text{tr}\left(T\,W_c T^{\mathsf{T}} + \left(T^{\mathsf{T}}\right)^{-1} W_o T^{-1} \right) \tag{E.11}$$

which corresponds to the sum of the eigenvalues of the transformed Gramians. Define the symmetric matrix $P = T^{\mathsf{T}}\,T$ so that the cost function

$$\psi\,(P) = \text{tr}\left(W_c\,P + W_o\,P^{-1} \right) \tag{E.12}$$

is equivalent to the cost function ϕ. The authors show that the cost function ψ has compact sublevel sets therefore ensuring the existence of a unique minimizing positive definite symmetric matrix P_∞, given by

$$P_\infty = W_c^{-1/2} \left(W_c^{1/2} W_o W_c^{1/2} \right)^{1/2} W_c^{-1/2} \tag{E.13}$$

Then $T_\infty = P_\infty^{1/2}$, and $S_\infty = {T_\infty}^{-1}$ is the unique symmetric positive definite balancing transformation for (A, B, C). The gradient flow $\dot{P}(t) = -\nabla \psi \left(P(t) \right)$ on the class of symmetric positive definite matrices is given by

$$\dot{P} = P^{-1} W_o P^{-1} - W_c \qquad P(0) = P_0 \tag{E.14}$$

For every initial condition $P_0 = {P_0}^\mathsf{T} > 0$, $P(t)$ exists for all $t \geq 0$ and converges exponentially fast to P_∞ as $t \to \infty$.

Appendix F

Proof of Theorem 4.2.15

Our proof for Theorem 4.2.15 is based on the framework of continuous time optimal control. Consider the dynamical system

$$\dot{x}(t) = f(x(t), u(t)), \quad t_0 \leq t \leq T \tag{F.1}$$

$$x(t_0) = x_0 \tag{F.2}$$

where $u(\cdot) \in U$ and

$$U = \{u : [0, T] \to \mathbb{R}^m \ : \ u(\cdot) \text{ is measurable } \} \tag{F.3}$$

is the set of admissible controls. Define the cost function

$$J_{x_0,t_0}(u(\cdot)) = \int_{t_0}^{T} \beta(x(t), u(t))\, dt + \gamma(x(T)) \tag{F.4}$$

and the value function

$$V(x_0, t_0) = \inf_{u(\cdot)\in U} J_{x_0,t_0}(u(\cdot)) \tag{F.5}$$

We use the following result presented by Evans [45] which states that the value function V satisfies a nonlinear Hamilton-Jacobi-Bellman PDE.

Theorem F.0.30 (Evans [45]) *The value function V is a weak solution of the Hamilton-Jacobi-Bellman PDE*

$$\frac{\partial V}{\partial t}(x,t) + H\left(\frac{\partial V}{\partial x}(x,t), x\right) = 0 \tag{F.6}$$

with boundary condition

$$V(x(T),T) = \gamma(x(T)) \tag{F.7}$$

where the Hamiltonian H is given by

$$H(p,x) = \min_{u\in U}\{f(x,u)\cdot p + \beta(x,u)\} \tag{F.8}$$

Theorem 4.2.15 (Scherpen)

Proof

To derive (4.14), fix $t_0 = 0$ and let $T \to -\infty$. Then in the framework of Evans, $L_c(x_0)$ corresponds to $V(x_o, 0)$,

$$J_{x_0,0}(u(\cdot)) = -\int_0^{-\infty} \frac{1}{2}\, u^T(t)\, u(t)\, dt \tag{F.9}$$

$$\beta(x,u) = -\frac{1}{2}\, u^T\, u \tag{F.10}$$

$$\gamma(x(T)) = 0 \tag{F.11}$$

and

$$H(p,x) = \min_{u\in U}\left\{p^T(f+gu) - \frac{1}{2}u^T u\right\} \tag{F.12}$$

where p^T corresponds to $\partial L_c/\partial x\,(x)$. The expression for H is minimized when $u = g^T p$ so that

$$H(p,x) = p^T f + \frac{1}{2}\, p^T\, g\, g^T\, p. \tag{F.13}$$

Also, for $u = g^T p$ we have

$$\dot{x} = f(x) + g(x)\, g^T(x)\, \frac{\partial^T L_c}{\partial x}(x). \tag{F.14}$$

Using the hypothesis that 0 is an asymptotically stable equilibrium of $-(f + g\,g^T\,\partial^T L_c/\partial x)$ we get $x \to 0$ as $t \to -\infty$. Finally, since t_0 is fixed we have

$$\frac{\partial L_c}{\partial t}(x,t) = 0 \tag{F.15}$$

and (F.6) becomes

$$\frac{\partial L_c}{\partial x}(x)\,f(x) + \frac{1}{2}\,\frac{\partial L_c}{\partial x}(x)\,g(x)\,g^T(x)\,\frac{\partial^T L_c}{\partial x}(x) = 0 \tag{F.16}$$

with boundary condition

$$L_c(0) = 0 \tag{F.17}$$

which is the desired equation.

To derive (4.15) we restrict the admissible controls to the singleton set $U_1 = \{u : u = 0\,, \quad 0 \le t \le \infty\}$. This would not be interesting for a real optimal control problem and is merely a device so that we can use the Evans framework. The observability function can then be written as

$$L_o(x_0) = \min\left\{\int_0^\infty h^T(x(t))\,h(x(t))\,dt\,,\,:\,u \in U_1\,,\,x(0) = x_0\right\} \tag{F.18}$$

Now fix $t_0 = 0$ and let $T \to -\infty$. In the Evans framework $L_o(x_0)$ corresponds to $V(x_o, 0)$,

$$J_{x_0,0}(u(\cdot)) = \int_0^{-\infty}\,\frac{1}{2}\,h^T(t)\,h(t)\,dt \tag{F.19}$$

$$\beta(x,u) = -\frac{1}{2}\,h^T\,h \tag{F.20}$$

$$\gamma(x(T)) = 0 \tag{F.21}$$

and

$$H(p,x) = \min_{u \in U_1}\left\{p^T(f + gu) + \frac{1}{2}h^T h\right\} \tag{F.22}$$

where p^T corresponds to $\partial\,L_o/\partial\,x\,(x)$. Since U_1 is trivial the minimization is trivial resulting in

$$H(p,x) = p^T\,f + \frac{1}{2}\,h^T\,h. \tag{F.23}$$

Using the hypothesis that 0 is an asymptotically stable equilibrium for $\dot{x} = f(x)$ we have $x \to 0$ as $t \to \infty$. Finally, since t_0 is fixed (F.6) becomes

$$\frac{\partial L_o}{\partial x}(x)\, f(x) + \frac{1}{2}\, h^T(x)\, h(x) = 0 \tag{F.24}$$

with boundary condition

$$L_o(0) = 0 \tag{F.25}$$

which is the desired equation.

■

Appendix G

Physical Constants

Listed here are the physical constants used in the models. The units have been selected for convenience and consistency. Properties of the wafer are those of pure silicon. Chamber wall properties are those of quartz. Properties of the process gases are those of hydrogen at 1000 K and 1 ATM. Chemical kinetics parameters are those experimentally determined from reactions involving thermally activated deposition of polysilicon from 30 sccm of 2% silane in hydrogen.

Constant	Description	Value	Units
k_0	Arrhenius Coefficient	3.0787×10^3	cm sec^{-1}
E_a	Activation Energy	1.6330×10^5	J mol^{-1}
R_g	Gas Constant	8.314	J mol^{-1} K^{-1}
h_{ref}	Reference Thickness	1.0×10^{-4}	cm
β_r	Rate Pre–Exponential Constant	1.8472×10^9	dimensionless
β_e	Rate Exponential Constant	2.8059×10^1	dimensionless
k_w	Thermal Conductivity of Wafer	0.22	W cm^{-1} K^{-1}
ρ_w	Mass Density of Wafer	2.3	g cm^{-3}
C_{p_w}	Heat Capacity of Wafer	2.3	J g^{-1} K^{-1}
σ_b	Boltzmann Constant	5.677×10^{-12}	W cm^{-2} K^{-4}
ϵ_w	Emissivity of Wafer	0.7	dimensionless
α_w	Absorptivity of Wafer	0.5	dimensionless
R_w	Radius of Wafer	7.62	cm
Δ_z	Thickness of Wafer	0.05	cm
h_v	Convective Heat Transfer Coeff	2.6474×10^{-4}	W cm^{-2} K^{-1}
Re	Reynolds Number of Gas Flow	27.2	dimensionless
k_g	Thermal Conductivity of Gas	4.40×10^{-3}	W cm^{-1} K^{-1}
Pr	Prandtl Number of Gas Flow	0.686	dimensionless
L	Chamber Length	50.8	cm
T_c	Chamber Wall (Ambient) Temp	700	K
ϵ_c	Emissivity of Chamber Wall	0.37	dimensionless
T_g	Gas Temperature	300	K
τ	Reference Time	60	seconds
Q_{ref}	Reference Heat Flux	29.24	W cm^{-2}

Appendix H

View Factor Analysis for Lamp Heating in the Epsilon-1

The approach we take to determine the heat flux spatial profiles is based on the concept of view factor [124, 145] which describes the radiation exchange between two or more surfaces separated by a non-participating medium that does not absorb, emit, or scatter radiation. The view factor between two surfaces represents the fraction of radiative energy leaving one surface that strikes the other surface directly.

In this method, the geometry of the chamber, including location and shape of lamps, susceptor, reflectors, and possibly other apparatus, is what determines the form of the resulting flux profiles. This geometric approach was adopted in [63], where the authors consider only a 2-dimensional slice of the chamber geometry, and includes the effect of reflectors behind the lamp banks. There, the 2-dimensional approach was reasonable, perhaps, since the lamp arrangement in the reactor under consideration was axisymmetric about the wafer center. This situation is, however, not the case in the Epsilon-1 reactor. Hence, our analysis is similar to that used

in [41], where the authors consider the chamber geometry from a 3-dimensional point of view. However, in that paper, as in this paper, the effect of reflectors is not included.

Assumptions

In the actual reactor, the internal surface of the chamber lid is gold plated to reflect infrared rays from the linear lamps, and the spot lamps are placed in gold plated parabolic reflectors. However, we do not consider the effect of reflections on the lamp heating of the wafer. In addition, the literature indicates that "virtual images", or radiation from the heated wafer to the reflectors and chamber walls which is reflected back to the wafer, will cause additional radiative effects. These effects are not included in the analysis here.

We consider all surfaces to be diffuse reflectors and diffuse emitters. Radiant intensity from the lamps is assumed to be independent of direction and constant across the length of the lamp. We assume that the quartz walls and the process gases transmit heat radiation from the lamps perfectly at the wavelengths of interest. Furthermore, we assume that the path from lamps to wafer (or lamps to susceptor) is completely free of any other obstacles.

Lamp Geometry

Figure H.1 shows a schematic of the upper lamp array superimposed over the susceptor and wafer, which is based on a description and diagram provided in [9]. For computational purposes, we consider each linear lamp to be a straight line segment of length 11.5 inches with the array consisting of parallel equally spaced lamps. The array begins directly above the susceptor edge, 5.0 inches (horizontally)

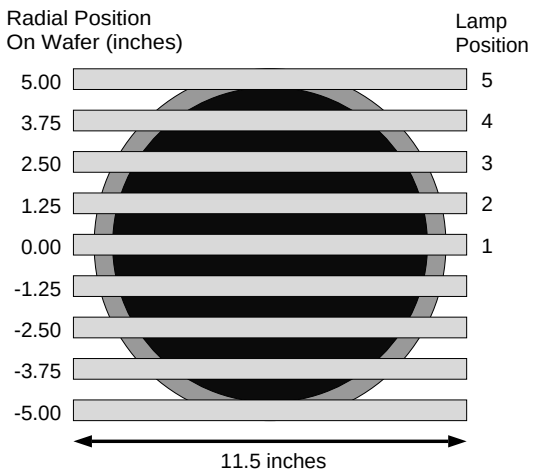


Figure H.1: Geometry (top view) of upper lamp array: radial position of each lamp is given in inches from center; lamps are identified by five uniquely distinguishable positions, numbered 1 through 5.

from the susceptor center. The distance between neighboring parallel lamps in the array is 1.25 inches. The vertical distance between wafer and upper lamp array is 2.25 inches, and the vertical distance between wafer and lower lamp array is 3.50 inches. We note that the distances given are estimates based on crude measurements taken on the reactor itself.

There are five lamp positions for the linear lamps that can be uniquely differentiated from the others. This is due to the wafer rotation. For example, two linear lamps equally distant from the center linear lamp have an identical irradiating effect on the wafer surface. The five lamp positions are numbered 1 through 5. The spot lamps have their own unique geometry and are analyzed separately later.

The source of radiation for each lamp is a tungsten filament, which we assume to be a straight line segment stretching the length of the lamp. Figure H.2 shows the geometry used to perform the analysis. We assume that for each filament the radiant intensity is independent of direction and constant across the length of the filament.

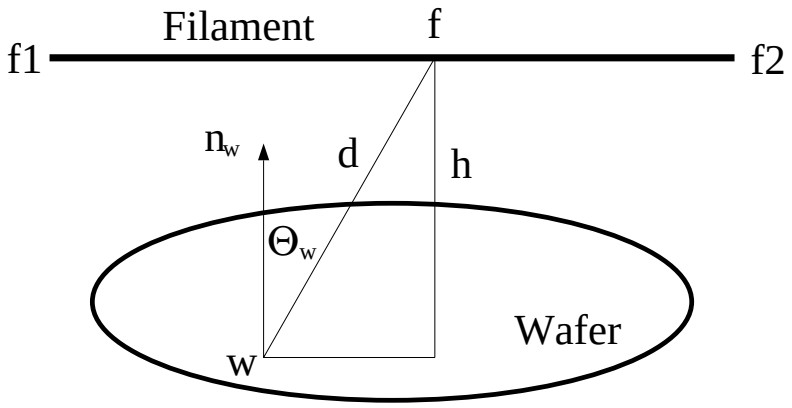


Figure H.2: Geometry for view factor analysis used to calculate heat flux intensity profiles for linear lamps.

For each point on the wafer, w, there is an irradiance contribution from each point on the filament, f, depending upon the distance between them, $d = \|w - f\|$, the angle θ_w formed by the vector $w - f$ and the vector n_w normal to the wafer surface, and the vertical distance, h, from wafer surface to filament. Note that on the filament diagram the endpoint values are $f_1 = -5.75$ inches and $f_2 = 5.75$ inches, and the vertical distance h is either 2.25 inches for the upper array or 3.50 inches for the lower array.

View Factor Analysis

For the derivation of the expression for heat flux radiant power per unit area on the wafer surface, we adopt the notation used in [124]. The rate of radiative energy dQ_f leaving a differential surface area dA_f (containing the point f) on the filament that strikes a differential surface area dA_w (containing the point w) on the wafer surface is given by

$$dQ_f = dA_f \, I_f \, cos(\theta_f) \, d\omega_{fw} \tag{H.1}$$

where I_f is the intensity of radiative energy leaving dA_f in all directions in hemispherical space (in dimensions of Watts per unit area per steradian), θ_f is the angle formed by the vector $w - f$ and a vector normal to dA_f, and $d\omega_{fw}$ is the solid angle subtended by dA_w from f given by

$$d\omega_{fw} = \frac{dA_w \, cos(\theta_w)}{d^2}. \tag{H.2}$$

Substituting (H.2) into (H.1) yields

$$dQ_f = dA_f \, I_f \, \frac{cos(\theta_f) \, cos(\theta_w) \, dA_w}{d^2}. \tag{H.3}$$

Now, the rate of radiation energy Q_f leaving the surface element dA_f on the filament in all directions over hemispherical space is [124]

$$Q_f = \pi \, I_f \, dA_f. \tag{H.4}$$

The elemental view factor $dF_{dA_f - dA_w}$ is defined as the ratio of the radiative energy leaving dA_f that strikes dA_w directly to the radiative energy leaving dA_f in all directions into the hemispherical space. Thus, we divide (H.3) by (H.4) to give the view factor

$$dF_{dA_f - dA_w} = \frac{dQ_f}{Q_f} = \frac{cos(\theta_f) \, cos(\theta_w) \, dA_w}{\pi d^2}. \tag{H.5}$$

Since we are assuming that filament radiant intensity is independent of direction, we take $\theta_f = 0$ independent of filament position f so that $cos(\theta_f) = 1$ and

$$dF_{dA_f - dA_w} = \frac{cos(\theta_w) \, dA_w}{\pi d^2}. \tag{H.6}$$

We are interested in the radiative energy illuminating a differential area on the wafer due to the entire filament. To compute the appropriate view factor, $F_{f - dA_w}$, we average (H.6) across the length of the filament

$$F_{f - dA_w} = \frac{dA_w}{|f_1 - f_2|} \int_{f_1}^{f_2} \frac{cos(\theta_w)}{\pi \, d^2} \, df. \tag{H.7}$$

Finally, we observe that

$$cos(\theta_w) = \frac{h}{d}$$

for the given geometry, so that

$$F_{f-dA_w} = \frac{dA_w}{|f_1 - f_2|} \int_{f_1}^{f_2} \frac{h}{\pi\, d^3}\, df \tag{H.8}$$

where we recall that $d = \|w - f\|$.

To determine the radiant heat flux profile for a given lamp, the view factor F_{f-dA_w} must be computed for each differential area dA_w on the wafer surface. For practical purposes, we discretize the wafer surface by choosing a cylindrical grid of wafer points $w = (r, \phi)$. We then assume that the differential area, dA_w, is constant for all wafer points w. Thus, (H.8) yields the view factor function $F_{f-dA_w}(r, \phi)$ which gives the fraction of radiative energy leaving the given lamp filament that strikes the given wafer point $w = (r, \phi)$ directly.

Now, we let P_f denote the radiant power supplied by the filament, so that P_f/dA_w gives the radiant heat flux intensity striking the differential area dA_w. The radiant heat flux intensity profile of the illumination due to the lamp filament is then given by

$$\begin{aligned} q_f(r,\phi) &= F_{f-dA_w}(r,\phi)\, \frac{P_f}{dA_w} && \text{(H.9)} \\ &= \frac{P_f\, h}{\pi\, |f_1 - f_2|} \int_{f_1}^{f_2} \frac{1}{d(r,\phi)^3}\, df && \text{(H.10)} \end{aligned}$$

where the value we use for P_f is provided by the manufacturer. In the case of the ASM Epsilon–1 reactor, the linear lamps supply 6000 Watts and the spot lamps supply 1000 Watts.

Since the wafer rotates at a uniform rate, this function is averaged over the circle (i.e., $0 \leq \phi < 2\,\pi$) at each radial position r on the wafer top surface

$$q_f(r) = \frac{P_f}{dA_w}\, \frac{1}{2\,\pi} \int_0^{2\pi} F_{fw}(r,\phi) d\phi \tag{H.11}$$

to give the heat flux profile

$$q_f(r) = \frac{P_f\, h}{2\,\pi^2\,|f_1 - f_2|} \int_0^{2\pi} \int_{f_1}^{f_2} \frac{1}{d(r,\phi)^3}\, df\, d\phi. \tag{H.12}$$

A similar analysis was performed for the spot lamps, except that each spot lamp was considered to be a point source of radiant energy, thus simplifying the analysis significantly.

The computational procedure was performed for each of the five different linear lamp positions for both upper and lower arrays, and for the spot lamps. Using the resulting heat flux intensity spatial profiles, we can then compute the desired profiles for each of the ten lamp groups, and the four heat zones of the Epsilon–1 reactor by appropriately combining the profiles determined from the individual lamps.

Results

Here we discuss some results of the analysis. Note that in what follows, the term "wafer surface" may represent the top surface of the wafer and exposed susceptor, or the bottom surface of the susceptor, depending upon the lamp group being considered.

Figures H.3 and H.4 show the heat flux irradiated on the wafer surface by lamps in positions 1, 2, 3, and 4 of the upper and lower array, respectively, and the spot lamp position. As expected, points on the wafer surface directly under (or over) the lamp filament receive the most intense illumination, i.e. the maximum flux value. Intensities are greater in magnitude for lamps in the upper array since it is physically closer to the wafer than the lower array and spot lamps. Spot lamps have lower flux intensities than linear lamps due to the smaller supplied power.

To account for wafer rotation, the flux intensity profiles are averaged around

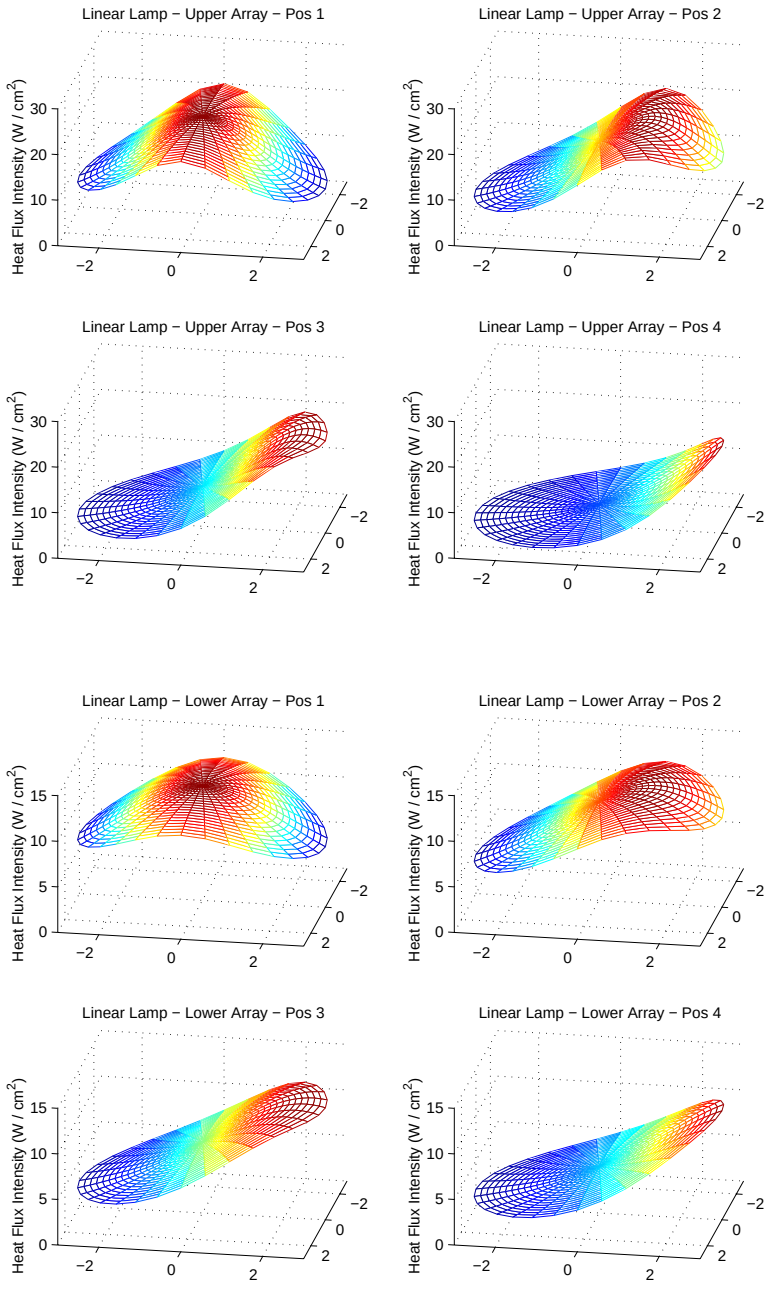


Figure H.3: Heat flux intensity profiles for linear lamps. Top: upper lamp array; Bottom: lower lamp array; (flux intensity (W/cm^2) versus position in two dimensions. Upper left: Position 1; Upper right: Position 2; Lower Left: Position 3; Lower right: Position 4).

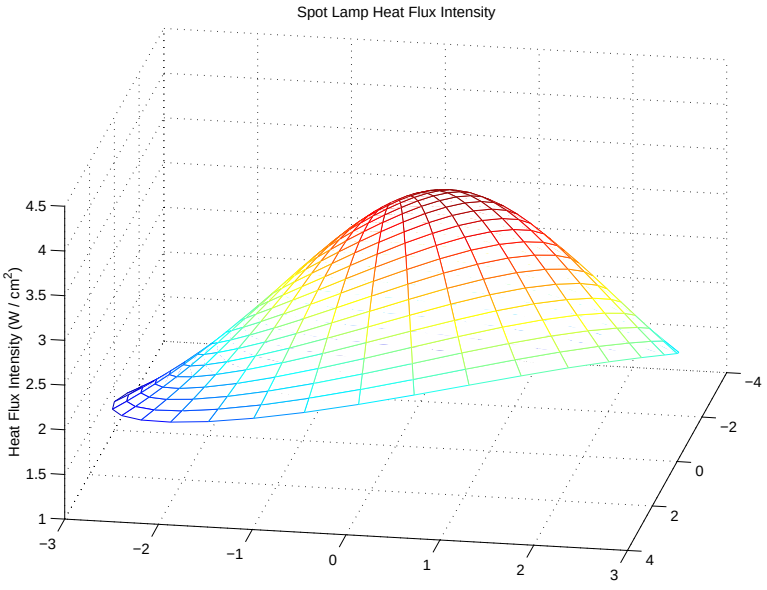


Figure H.4: Heat flux intensity profile for spot lamp: flux intensity (W/cm^2) versus position in two dimensions.

360 degrees resulting in profiles that are a function of radial position only. Figure H.5 shows the heat flux profiles, after averaging, for each of the individual lamp positions. Observe that as expected the lamp position directly over (or under) the wafer center irradiates the wafer center with greater intensity than the other positions. Lamp positions closer to the edge irradiate the edge with greater intensity than they irradiate the center.

Figure H.6 shows the heat flux profiles for each of the ten lamp groups. Figure H.7 shows the heat flux profiles for the four heating zones - center, front, rear, and side. The flux intensity for the center zone is significantly greater than for the others, indicating that it will have the greatest heating effect. Observe that profiles for front and rear zones are identical due to the symmetry assumptions and the way in which the individual lamps are organized to form the zones.

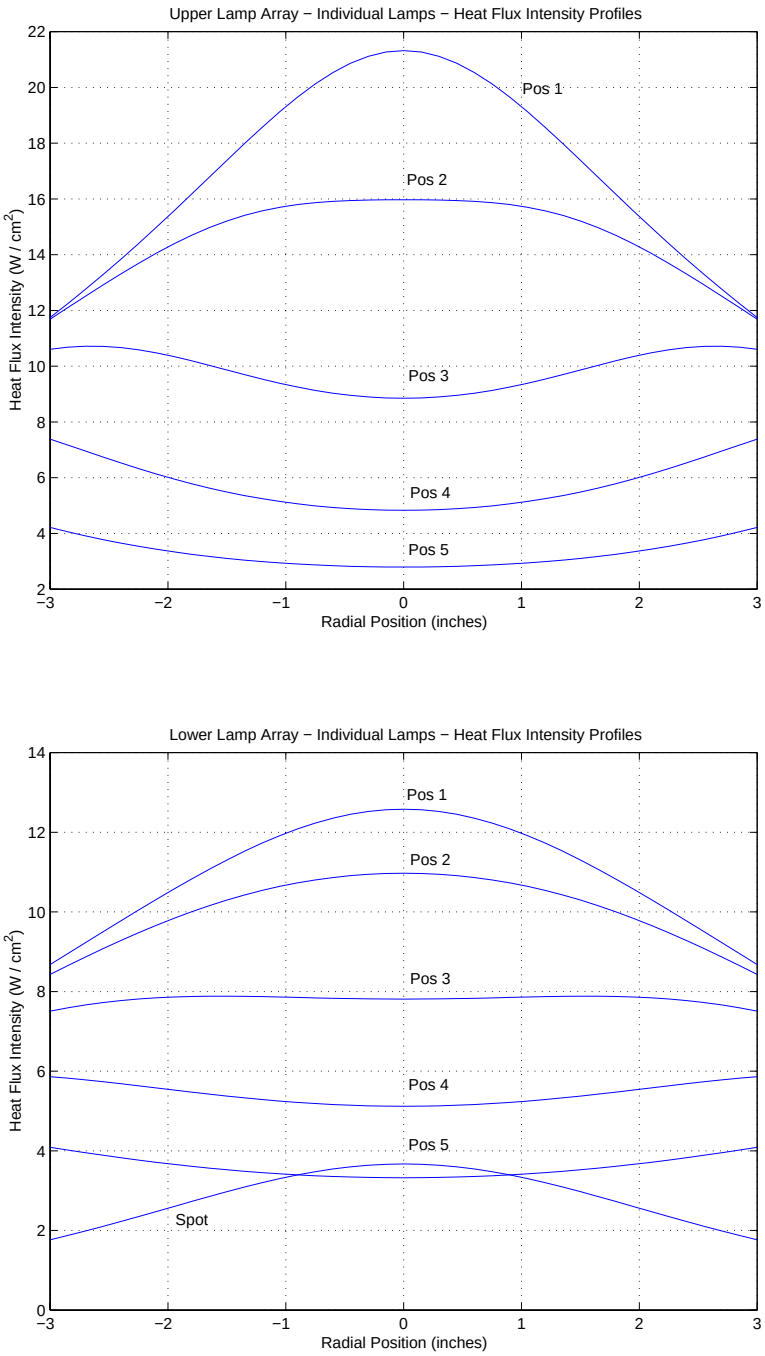


Figure H.5: Heat flux intensity profiles for individual lamps. Top: upper array; Bottom: lower array; (flux intensity (W/cm^2) versus radial position for the five uniquely distinguishable linear lamp positions and the spot lamp position).

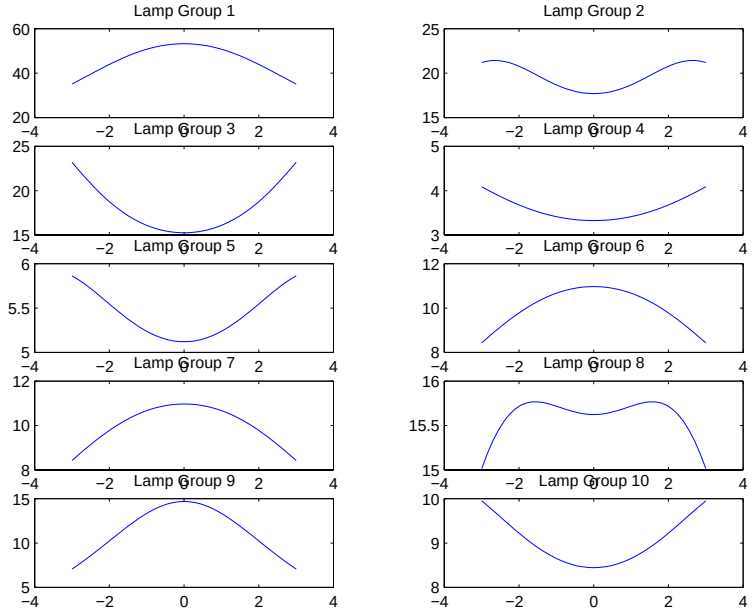


Figure H.6: Heat flux intensity profiles for ASM Epsilon–1 lamp groups: flux intensity (W/cm^2) versus radial position for the ten lamp groups.

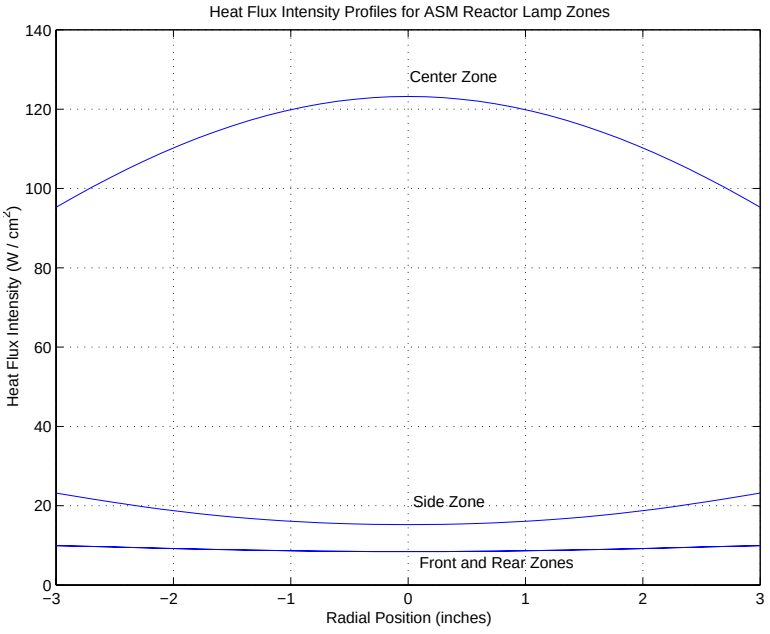


Figure H.7: Heat flux intensity profiles for ASM Epsilon–1 heat zones: flux intensity (W/cm^2) versus radial position for the four heat zones.

Appendix I

PHOENICS Q1 Source File for Epsilon-1 Poly-Si Growth Simulation

This appendix contains the input code, called a Q1 file, for the PHOENICS CFD software package. The file was used to simulate deposition of poly-Si on a silicon wafer with wafer temperature of 750 C, silane flow rate of 30 sccm, and chamber pressure of 20 Torr. Input files for simulations with other process conditions are similar.

```
************************************************************
CPVNAM=CVD
************************************************************
IRUNN   =       1 ;LIBREF =      14
************************************************************
 Group 1. Run Title
TEXT(POLY-SI DEP  750 C  30 sccm SiH4  20 Torr  )
************************************************************
 Group 2. Transience
STEADY  =    T
************************************************************
```

```
 Groups 3, 4, 5  Grid Information
   * Overall number of cells, RSET(M,NX,NY,NZ,tolerance)
RSET(M,25,27,52)
   * Set overall domain extent:
   *        xulast  yvlast  zwlast
           name
XSI= 1.651000E-01; YSI= 1.094800E-01; ZSI= 6.068000E-01
RSET(D,EPS1    )
************************************************************
 Group 6. Body-Fitted coordinates
BFC=T
   * Set points
XPO= 7.6200E-02;YPO=-4.7000E-03;ZPO= 0.0000E+00;GSET(P,PW1 )
XPO= 7.6200E-02;YPO= 4.7000E-03;ZPO= 0.0000E+00;GSET(P,PW2 )
XPO= 0.0000E+00;YPO=-5.4740E-02;ZPO= 0.0000E+00;GSET(P,P00 )
XPO= 0.0000E+00;YPO=-4.8740E-02;ZPO= 0.0000E+00;GSET(P,P01 )
XPO= 0.0000E+00;YPO=-4.7000E-03;ZPO= 0.0000E+00;GSET(P,P02 )
XPO= 0.0000E+00;YPO= 4.7000E-03;ZPO= 0.0000E+00;GSET(P,P03 )
XPO= 0.0000E+00;YPO= 4.8740E-02;ZPO= 0.0000E+00;GSET(P,P04 )
XPO= 0.0000E+00;YPO= 5.4740E-02;ZPO= 0.0000E+00;GSET(P,P05 )
XPO= 1.1125E-01;YPO=-2.9718E-02;ZPO= 0.0000E+00;GSET(P,P10 )
XPO= 1.1125E-01;YPO=-2.3719E-02;ZPO= 0.0000E+00;GSET(P,P11 )
XPO= 1.1125E-01;YPO=-4.7000E-03;ZPO= 0.0000E+00;GSET(P,P12 )
XPO= 1.1125E-01;YPO= 4.7000E-03;ZPO= 0.0000E+00;GSET(P,P13 )
XPO= 1.1125E-01;YPO= 2.3719E-02;ZPO= 0.0000E+00;GSET(P,P14 )
XPO= 1.1125E-01;YPO= 2.9718E-02;ZPO= 0.0000E+00;GSET(P,P15 )
XPO= 1.1240E-01;YPO=-2.9506E-02;ZPO= 0.0000E+00;GSET(P,P20 )
XPO= 1.1240E-01;YPO=-2.3198E-02;ZPO= 0.0000E+00;GSET(P,P21 )
XPO= 1.1240E-01;YPO=-4.7000E-03;ZPO= 0.0000E+00;GSET(P,P22 )
XPO= 1.1240E-01;YPO= 4.7000E-03;ZPO= 0.0000E+00;GSET(P,P23 )
XPO= 1.1240E-01;YPO= 2.3198E-02;ZPO= 0.0000E+00;GSET(P,P24 )
XPO= 1.1240E-01;YPO= 2.9506E-02;ZPO= 0.0000E+00;GSET(P,P25 )
XPO= 1.2581E-01;YPO=-2.7021E-02;ZPO= 0.0000E+00;GSET(P,P30 )
XPO= 1.2581E-01;YPO=-1.7092E-02;ZPO= 0.0000E+00;GSET(P,P31 )
XPO= 1.2581E-01;YPO=-4.7000E-03;ZPO= 0.0000E+00;GSET(P,P32 )
XPO= 1.2581E-01;YPO= 4.7000E-03;ZPO= 0.0000E+00;GSET(P,P33 )
XPO= 1.2581E-01;YPO= 1.7092E-02;ZPO= 0.0000E+00;GSET(P,P34 )
XPO= 1.2581E-01;YPO= 2.7021E-02;ZPO= 0.0000E+00;GSET(P,P35 )
XPO= 1.3030E-01;YPO=-2.6188E-02;ZPO= 0.0000E+00;GSET(P,P40 )
XPO= 1.3030E-01;YPO=-1.5045E-02;ZPO= 0.0000E+00;GSET(P,P41 )
XPO= 1.3030E-01;YPO=-4.7000E-03;ZPO= 0.0000E+00;GSET(P,P42 )
XPO= 1.3030E-01;YPO= 4.7000E-03;ZPO= 0.0000E+00;GSET(P,P43 )
XPO= 1.3030E-01;YPO= 1.5045E-02;ZPO= 0.0000E+00;GSET(P,P44 )
XPO= 1.3030E-01;YPO= 2.6188E-02;ZPO= 0.0000E+00;GSET(P,P45 )
XPO= 1.4605E-01;YPO=-2.3270E-02;ZPO= 0.0000E+00;GSET(P,P50 )
```

```
XPO= 1.4605E-01;YPO=-7.8750E-03;ZPO= 0.0000E+00;GSET(P,P51 )
XPO= 1.4605E-01;YPO=-4.7000E-03;ZPO= 0.0000E+00;GSET(P,P52 )
XPO= 1.4605E-01;YPO= 4.7000E-03;ZPO= 0.0000E+00;GSET(P,P53 )
XPO= 1.4605E-01;YPO= 7.8750E-03;ZPO= 0.0000E+00;GSET(P,P54 )
XPO= 1.4605E-01;YPO= 2.3270E-02;ZPO= 0.0000E+00;GSET(P,P55 )
XPO= 1.6510E-01;YPO=-1.9740E-02;ZPO= 0.0000E+00;GSET(P,P60 )
XPO= 1.6510E-01;YPO=-7.8750E-03;ZPO= 0.0000E+00;GSET(P,P61 )
XPO= 1.6510E-01;YPO=-4.7000E-03;ZPO= 0.0000E+00;GSET(P,P62 )
XPO= 1.6510E-01;YPO= 4.7000E-03;ZPO= 0.0000E+00;GSET(P,P63 )
XPO= 1.6510E-01;YPO= 7.8750E-03;ZPO= 0.0000E+00;GSET(P,P64 )
XPO= 1.6510E-01;YPO= 1.9740E-02;ZPO= 0.0000E+00;GSET(P,P65 )
XPO= 5.0000E-02;YPO=-5.0929E-02;ZPO= 0.0000E+00;GSET(P,PA0 )
XPO= 5.0000E-02;YPO=-4.4929E-02;ZPO= 0.0000E+00;GSET(P,PA1 )
XPO= 5.0000E-02;YPO= 4.4929E-02;ZPO= 0.0000E+00;GSET(P,PA2 )
XPO= 5.0000E-02;YPO= 5.0929E-02;ZPO= 0.0000E+00;GSET(P,PA3 )
   * Set lines/arcs
GSET(L,LVW,PW1,PW2,5,1.0)
GSET(L,LV01,P00,P01,1,1.0)
GSET(L,LV02,P01,P02,8,.7)
GSET(L,LV03,P02,P03,5,1)
GSET(L,LV04,P03,P04,12,1.5)
GSET(L,LV05,P04,P05,1,1.0)
GSET(L,LV11,P10,P11,1,1.0)
GSET(L,LV12,P11,P12,8,.7)
GSET(L,LV13,P12,P13,5,1.0)
GSET(L,LV14,P13,P14,12,1.5)
GSET(L,LV15,P14,P15,1,1.0)
GSET(L,LV21,P20,P21,1,1.0)
GSET(L,LV22,P21,P22,8,.7)
GSET(L,LV23,P22,P23,5,1.0)
GSET(L,LV24,P23,P24,12,1.5)
GSET(L,LV25,P24,P25,1,1.0)
GSET(L,LV31,P30,P31,1,1.0)
GSET(L,LV32,P31,P32,8,.7)
GSET(L,LV33,P32,P33,5,1.0)
GSET(L,LV34,P33,P34,12,1.5)
GSET(L,LV35,P34,P35,1,1.0)
GSET(L,LV41,P40,P41,1,1.0)
GSET(L,LV42,P41,P42,8,.7)
GSET(L,LV43,P42,P43,5,1.0)
GSET(L,LV44,P43,P44,12,1.5)
GSET(L,LV45,P44,P45,1,1.0)
GSET(L,LV51,P50,P51,1,1.0)
GSET(L,LV52,P51,P52,8,.7)
GSET(L,LV53,P52,P53,5,1.0)
```

```
GSET(L,LV54,P53,P54,12,1.5)
GSET(L,LV55,P54,P55,1,1.0)
GSET(L,LV61,P60,P61,1,1.0)
GSET(L,LV62,P61,P62,8,.7)
GSET(L,LV63,P62,P63,5,1.0)
GSET(L,LV64,P63,P64,12,1.5)
GSET(L,LV65,P64,P65,1,1.0)
GSET(L,A00,P00,P10,12,1.0,ARC,PA0)
GSET(L,A01,P01,P11,12,1.0,ARC,PA1)
GSET(L,A04,P04,P14,12,1.0,ARC,PA2)
GSET(L,A05,P05,P15,12,1.0,ARC,PA3)
GSET(L,LH01,P02,PW1,8,1.0)
GSET(L,LH02,PW1,P12,4,1.0)
GSET(L,LH03,P03,PW2,8,1.0)
GSET(L,LH04,PW2,P13,4,1.0)
GSET(L,LH10,P10,P20,1,1.0)
GSET(L,LH11,P11,P21,1,1.0)
GSET(L,LH12,P12,P22,1,1.0)
GSET(L,LH13,P13,P23,1,1.0)
GSET(L,LH14,P14,P24,1,1.0)
GSET(L,LH15,P15,P25,1,1.0)
GSET(L,LH20,P20,P30,4,1.0)
GSET(L,LH21,P21,P31,4,1.0)
GSET(L,LH22,P22,P32,4,1.0)
GSET(L,LH23,P23,P33,4,1.0)
GSET(L,LH24,P24,P34,4,1.0)
GSET(L,LH25,P25,P35,4,1.0)
GSET(L,LH30,P30,P40,4,1.0)
GSET(L,LH31,P31,P41,4,1.0)
GSET(L,LH32,P32,P42,4,1.0)
GSET(L,LH33,P33,P43,4,1.0)
GSET(L,LH34,P34,P44,4,1.0)
GSET(L,LH35,P35,P45,4,1.0)
GSET(L,LH40,P40,P50,3,1.0)
GSET(L,LH41,P41,P51,3,1.0)
GSET(L,LH42,P42,P52,3,1.0)
GSET(L,LH43,P43,P53,3,1.0)
GSET(L,LH44,P44,P54,3,1.0)
GSET(L,LH45,P45,P55,3,1.0)
GSET(L,LH50,P50,P60,1,1.0)
GSET(L,LH51,P51,P61,1,1.0)
GSET(L,LH52,P52,P62,1,1.0)
GSET(L,LH53,P53,P63,1,1.0)
GSET(L,LH54,P54,P64,1,1.0)
GSET(L,LH55,P55,P65,1,1.0)
```

```
   * Set frames
GSET(F,F01,P00,-,P10,-,P11,-,P01,-)
GSET(F,F02,P01,-,P11,-,P12,PW1,P02,-)
GSET(F,F031,P02,-,PW1,-,PW2,-,P03,-)
GSET(F,F032,PW1,-,P12,-,P13,-,PW2,-)
GSET(F,F04,P03,PW2,P13,-,P14,-,P04,-)
GSET(F,F05,P04,-,P14,-,P15,-,P05,-)
GSET(F,F11,P10,-,P20,-,P21,-,P11,-)
GSET(F,F12,P11,-,P21,-,P22,-,P12,-)
GSET(F,F13,P12,-,P22,-,P23,-,P13,-)
GSET(F,F14,P13,-,P23,-,P24,-,P14,-)
GSET(F,F15,P14,-,P24,-,P25,-,P15,-)
GSET(F,F21,P20,-,P30,-,P31,-,P21,-)
GSET(F,F22,P21,-,P31,-,P32,-,P22,-)
GSET(F,F23,P22,-,P32,-,P33,-,P23,-)
GSET(F,F24,P23,-,P33,-,P34,-,P24,-)
GSET(F,F25,P24,-,P34,-,P35,-,P25,-)
GSET(F,F31,P30,-,P40,-,P41,-,P31,-)
GSET(F,F32,P31,-,P41,-,P42,-,P32,-)
GSET(F,F33,P32,-,P42,-,P43,-,P33,-)
GSET(F,F34,P33,-,P43,-,P44,-,P34,-)
GSET(F,F35,P34,-,P44,-,P45,-,P35,-)
GSET(F,F41,P40,-,P50,-,P51,-,P41,-)
GSET(F,F42,P41,-,P51,-,P52,-,P42,-)
GSET(F,F43,P42,-,P52,-,P53,-,P43,-)
GSET(F,F44,P43,-,P53,-,P54,-,P44,-)
GSET(F,F45,P44,-,P54,-,P55,-,P45,-)
GSET(F,F51,P50,-,P60,-,P61,-,P51,-)
GSET(F,F52,P51,-,P61,-,P62,-,P52,-)
GSET(F,F53,P52,-,P62,-,P63,-,P53,-)
GSET(F,F54,P53,-,P63,-,P64,-,P54,-)
GSET(F,F55,P54,-,P64,-,P65,-,P55,-)
   * Match a grid mesh
GSET(M,F01,+I+J,1,1,1,LAP5)
GSET(M,F02,+I+J,1,2,1,LAP5)
GSET(M,F031,+I+J,1,10,1,LAP5)
GSET(M,F032,+I+J,9,10,1,LAP5)
GSET(M,F04,+I+J,1,15,1,LAP5)
GSET(M,F05,+I+J,1,27,1,LAP5)
GSET(M,F11,+I+J,13,1,1,LAP5)
GSET(M,F12,+I+J,13,2,1,LAP5)
GSET(M,F13,+I+J,13,10,1,LAP5)
GSET(M,F14,+I+J,13,15,1,LAP5)
GSET(M,F15,+I+J,13,27,1,LAP5)
GSET(M,F21,+I+J,14,1,1,LAP5)
```

```
GSET(M,F22,+I+J,14,2,1,LAP5)
GSET(M,F23,+I+J,14,10,1,LAP5)
GSET(M,F24,+I+J,14,15,1,LAP5)
GSET(M,F25,+I+J,14,27,1,LAP5)
GSET(M,F31,+I+J,18,1,1,LAP5)
GSET(M,F32,+I+J,18,2,1,LAP5)
GSET(M,F33,+I+J,18,10,1,LAP5)
GSET(M,F34,+I+J,18,15,1,LAP5)
GSET(M,F35,+I+J,18,27,1,LAP5)
GSET(M,F41,+I+J,22,1,1,LAP5)
GSET(M,F42,+I+J,22,2,1,LAP5)
GSET(M,F43,+I+J,22,10,1,LAP5)
GSET(M,F44,+I+J,22,15,1,LAP5)
GSET(M,F45,+I+J,22,27,1,LAP5)
GSET(M,F51,+I+J,25,1,1,LAP5)
GSET(M,F52,+I+J,25,2,1,LAP5)
GSET(M,F53,+I+J,25,10,1,LAP5)
GSET(M,F54,+I+J,25,15,1,LAP5)
GSET(M,F55,+I+J,25,27,1,LAP5)
   * Copy/Transfer/Block grid planes
GSET(C,K11,F,K1,1,25,1,27,+,0,0,1.2306E-01,INC,1)
GSET(C,K15,F,K11,1,25,1,27,+,0,0,4.4960E-03,INC,1)
GSET(C,K19,F,K15,1,25,1,27,+,0,0,1.9960E-02,INC,1)
GSET(C,K23,F,K19,1,25,1,27,+,0,0,3.6195E-02,INC,1)
GSET(C,K31,F,K23,1,25,1,27,+,0,0,1.5240E-01,INC,1)
GSET(C,K35,F,K31,1,25,1,27,+,0,0,3.6195E-02,INC,1)
GSET(C,K39,F,K35,1,25,1,27,+,0,0,3.8100E-02,INC,1)
GSET(C,K43,F,K39,1,25,1,27,+,0,0,4.4960E-03,INC,1)
GSET(C,K53,F,K43,1,25,1,27,+,0,0,1.9189E-01,INC,1)
   **********
NONORT  =    T
   * X-cyclic boundaries switched
************************************************************
 Group 7. Variables: STOREd,SOLVEd,NAMEd
ONEPHS  =    T
   * Non-default variable names
NAME( 16) =S80  ; NAME( 17) =S140
NAME( 18) =S142 ; NAME( 19) =S145
NAME( 20) =S147 ; NAME( 21) =S158
NAME(141) =BLOK ; NAME(142) =WCRT
NAME(143) =VCRT ; NAME(144) =UCRT
NAME(145) =TEM1 ; NAME(146) =DEPO
NAME(147) =PRPS ; NAME(148) =ENUL
NAME(149) =RHO1 ; NAME(150) =EMIS
   * Solved variables list
```

```
SOLVE(P1  ,U1  ,V1  ,W1  ,S140,S142,S145,S147)
SOLVE(S158,TEM1)
   * Stored variables list
STORE(EMIS,RHO1,ENUL,PRPS,DEPO,UCRT,VCRT,WCRT)
STORE(BLOK,S80 )
   * Additional solver options
SOLUTN(P1  ,Y,Y,Y,N,N,Y)
SOLUTN(S140,Y,Y,Y,N,N,Y)
SOLUTN(S142,Y,Y,Y,N,N,Y)
SOLUTN(S145,Y,Y,Y,N,N,Y)
SOLUTN(S147,Y,Y,Y,N,N,Y)
SOLUTN(S158,Y,Y,Y,N,N,Y)
SOLUTN(TEM1,Y,Y,Y,N,N,Y)
IVARBK  =      -1 ;ISOLBK =       1

************************************************************
 Group 8. Terms \& Devices
DIFCUT  = 0.000E+00
NEWRH1  =    T
NEWENL  =    T
UDIFNE  =    T
USOURC  =    T
ISOLX   =       0 ;ISOLY  =       0 ;ISOLZ  =       0
************************************************************
 Group 9. Properties
RHO1    = GRND10
PRESS0  = 2.631E+03
TMP1A   = 2.930E+02 ;TMP1B  = 0.000E+00 ;TMP1C  = 0.000E+00
CP1     = GRND10
ENUL    = GRND10    ;ENUT   = 0.000E+00
PRNDTL(S140) = -GRND8     ;PRNDTL(S142) = -GRND8
PRNDTL(S145) = -GRND8     ;PRNDTL(S147) = -GRND8
PRNDTL(S158) = -GRND8     ;PRNDTL(TEM1) = -GRND10
TMP1A   = 2.930E+02
  *  List of user-defined materials to be read by EARTH
   MATFLG=T;IMAT=1
  *  Name
  *Ind. Dens.  Viscos.  Spec.heat  Conduct.  Expans.  Compr.
  *  <GAS\_MIXTURE>

   70 GRND8 GRND8 GRND8 GRND8 1.000 0.000
  *       constants for GRND option no 1
   0 0
  *       constants for GRND option no 2
   0 0
```

```
  *       constants for GRND option no 3
   0 0
  *       constants for GRND option no 4
   0 0
************************************************************
 Group 10.Inter-Phase Transfer Processes
************************************************************
 Group 11.Initialise Var/Porosity Fields
FIINIT(W1  ) =  1.000E+00 ;FIINIT(S140) =  2.185E-02
FIINIT(BLOK) =  1.000E+00 ;FIINIT(TEM1) =  2.980E+02
FIINIT(PRPS) =  7.000E+01

CONPOR(TOP     , -1.00,CELL  ,-\#1,-\#6,-\#5,-\#5,-\#1,-\#9)
INIT(TOP     ,BLOK, 0.000E+00, 2.000E+00)
INIT(TOP     ,PRPS, 0.000E+00, 1.060E+02)

CONPOR(BOT     , -1.00,CELL  ,-\#1,-\#6,-\#1,-\#1,-\#1,-\#9)
INIT(BOT     ,BLOK, 0.000E+00, 3.000E+00)
INIT(BOT     ,PRPS, 0.000E+00, 1.060E+02)

CONPOR(SIDE    , -1.00,CELL  ,-\#7,-\#7,-\#1,-\#5,-\#1,-\#9)
INIT(SIDE    ,BLOK, 0.000E+00, 4.000E+00)
INIT(SIDE    ,PRPS, 0.000E+00, 1.060E+02)

CONPOR(SHF     , -1.00,CELL  ,-\#1,-\#6,-\#3,-\#3,-\#1,-\#1)
INIT(SHF     ,BLOK, 0.000E+00, 5.000E+00)
INIT(SHF     ,PRPS, 0.000E+00, 1.060E+02)

CONPOR(SHR     , -1.00,CELL  ,-\#1,-\#6,-\#3,-\#3,-\#9,-\#9)
INIT(SHR     ,BLOK, 0.000E+00, 6.000E+00)
INIT(SHR     ,PRPS, 0.000E+00, 1.060E+02)

CONPOR(SHS     , -1.00,CELL  ,-\#6,-\#6,-\#3,-\#3,-\#2,-\#8)
INIT(SHS     ,BLOK, 0.000E+00, 7.000E+00)
INIT(SHS     ,PRPS, 0.000E+00, 1.060E+02)

CONPOR(RNGF    , -1.00,CELL  ,-\#1,-\#4,-\#3,-\#3,-\#3,-\#3)
INIT(RNGF    ,BLOK, 0.000E+00, 8.000E+00)
INIT(RNGF    ,PRPS, 0.000E+00, 1.110E+02)

CONPOR(RNGR    , -1.00,CELL  ,-\#1,-\#4,-\#3,-\#3,-\#7,-\#7)
INIT(RNGR    ,BLOK, 0.000E+00, 9.000E+00)
INIT(RNGR    ,PRPS, 0.000E+00, 1.110E+02)

CONPOR(RNGS    , -1.00,CELL  ,-\#4,-\#4,-\#3,-\#3,-\#4,-\#6)
```

```
INIT(RNGS    ,BLOK, 0.000E+00, 1.000E+01)
INIT(RNGS    ,PRPS, 0.000E+00, 1.110E+02)

CONPOR(SUSF    , -1.00,CELL  ,-\#1,-\#3,-\#3,-\#3,-\#4,-\#4)
INIT(SUSF    ,BLOK, 0.000E+00, 1.100E+01)
INIT(SUSF    ,PRPS, 0.000E+00, 1.110E+02)

CONPOR(SUSR    , -1.00,CELL  ,-\#1,-\#3,-\#3,-\#3,-\#6,-\#6)
INIT(SUSR    ,BLOK, 0.000E+00, 1.200E+01)
INIT(SUSR    ,PRPS, 0.000E+00, 1.110E+02)

CONPOR(SUSS    , -1.00,CELL  ,-\#2,-\#3,-\#3,-\#3,-\#5,-\#5)
INIT(SUSS    ,BLOK, 0.000E+00, 1.300E+01)
INIT(SUSS    ,PRPS, 0.000E+00, 1.110E+02)

CONPOR(WAF     , -1.00,CELL  ,-\#1,-\#1,-\#3,-\#3,-\#5,-\#5)
INIT(WAF     ,BLOK, 0.000E+00, 1.400E+01)
INIT(WAF     ,PRPS, 0.000E+00, 1.110E+02)

INIADD  =    F
************************************************************
 Group 12. Convection and diffusion adjustments
  No PATCHes used for this Group
************************************************************
 Group 13. Boundary \& Special Sources

INLET (BFCIN1  ,LOW   ,\#1,\#6,\#4,\#4,\#1,\#1,\#1,\#1)
VALUE (BFCIN1  ,P1  , GRND1    )
VALUE (BFCIN1  ,U1  , GRND1    )
VALUE (BFCIN1  ,V1  , GRND1    )
VALUE (BFCIN1  ,W1  , GRND1    )
VALUE (BFCIN1  ,S140, 2.185E-02)
VALUE (BFCIN1  ,WCRT, 1.400E+00)
VALUE (BFCIN1  ,TEM1, 2.980E+02)

INLET (BFCIN2  ,LOW   ,\#1,\#6,\#2,\#2,\#1,\#1,\#1,\#1)
VALUE (BFCIN2  ,P1  , GRND1    )
VALUE (BFCIN2  ,U1  , GRND1    )
VALUE (BFCIN2  ,V1  , GRND1    )
VALUE (BFCIN2  ,W1  , GRND1    )
VALUE (BFCIN2  ,WCRT, 4.500E-01)
VALUE (BFCIN2  ,TEM1, 2.980E+02)

PATCH (OU1     ,HIGH  ,\#1,\#6,\#4,\#4,\#9,\#9,\#1,\#1)
COVAL (OU1     ,P1  , FIXVAL   , 0.000E+00)
```

```
PATCH (REAR    ,HWALL ,\#1,\#6,\#2,\#2,\#9,\#9,\#1,\#1)
COVAL (REAR    ,U1  , GRND2    , 0.000E+00)
COVAL (REAR    ,V1  , GRND2    , 0.000E+00)

PATCH (SUSFT   ,VOLUME,\#1,\#3,\#3,\#3,\#4,\#4,1,1)
COVAL (SUSFT   ,TEM1, FIXVAL   , 1.023E+03)

PATCH (SUSRT   ,VOLUME,\#1,\#3,\#3,\#3,\#6,\#6,1,1)
COVAL (SUSRT   ,TEM1, FIXVAL   , 1.023E+03)

PATCH (SUSST   ,VOLUME,\#2,\#3,\#3,\#3,\#5,\#5,\#1,\#1)
COVAL (SUSST   ,TEM1, FIXVAL   , 1.023E+03)

PATCH (WAFT    ,VOLUME,\#1,\#1,\#3,\#3,\#5,\#5,\#1,\#1)
COVAL (WAFT    ,TEM1, FIXVAL   , 1.023E+03)

PATCH (TOPT    ,SOUTH ,\#1,\#3,\#5,\#5,\#4,\#6,\#1,\#1)
COVAL (TOPT    ,TEM1, FIXVAL   , 7.230E+02)

PATCH (BOTT    ,NORTH ,\#1,\#3,\#1,\#1,\#4,\#6,\#1,\#1)
COVAL (BOTT    ,TEM1, FIXVAL   , 7.230E+02)

PATCH (SURFWAF ,SOUTH ,1,8,15,15,23,30,\#1,\#1)
COVAL (SURFWAF ,P1  , 1.000E+00, GRND1    )
COVAL (SURFWAF ,S80 , FIXFLU   , GRND1    )
COVAL (SURFWAF ,S140, FIXFLU   , GRND1    )
COVAL (SURFWAF ,S142, FIXFLU   , GRND1    )
COVAL (SURFWAF ,S145, FIXFLU   , GRND1    )
COVAL (SURFWAF ,S147, FIXFLU   , GRND1    )
COVAL (SURFWAF ,S158, FIXFLU   , GRND1    )
COVAL (SURFWAF ,TEM1, FIXFLU   , GRND1    )

PATCH (RELT    ,PHASEM,1,25,1,27,1,52,1,1)
COVAL (RELT    ,S80 , GRND1    , SAME     )
COVAL (RELT    ,S140, GRND1    , SAME     )
COVAL (RELT    ,S142, GRND1    , SAME     )
COVAL (RELT    ,S145, GRND1    , SAME     )
COVAL (RELT    ,S147, GRND1    , SAME     )
COVAL (RELT    ,S158, GRND1    , SAME     )

PATCH (CHEM    ,VOLUME,1,25,1,27,1,52,1,1)
COVAL (CHEM    ,S80 , GRND1    , GRND1    )
COVAL (CHEM    ,S140, GRND1    , GRND1    )
COVAL (CHEM    ,S142, GRND1    , GRND1    )
```

```
COVAL (CHEM    ,S145, GRND1    , GRND1    )
COVAL (CHEM    ,S147, GRND1    , GRND1    )
COVAL (CHEM    ,S158, GRND1    , GRND1    )
COVAL (CHEM    ,TEM1, GRND1    , GRND1    )

PATCH (BUOYANCY,PHASEM,\#1,\#NREGX,\#1,\#NREGY,\#1,\#NREGZ,\#1,\#NREGT)
COVAL (BUOYANCY,U1  , FIXFLU   , GRND3    )
COVAL (BUOYANCY,V1  , FIXFLU   , GRND3    )
COVAL (BUOYANCY,W1  , FIXFLU   , GRND3    )

BUOYA   = 0.000E+00 ; BUOYB =-9.810E+00 ; BUOYC = 0.000E+00
BUOYD   = GRND10

BFCA    = 2.171E-03
************************************************************
 Group 14. Downstream Pressure For PARAB
************************************************************
 Group 15. Terminate Sweeps
LSWEEP  =     500
SELREF  =    T
RESFAC  = 1.000E-03
************************************************************
 Group 16. Terminate Iterations
************************************************************
 Group 17. Relaxation
RELAX(P1  ,LINRLX, 7.000000E-01)
RELAX(U1  ,FALSDT, 2.703210E-02)
RELAX(V1  ,FALSDT, 2.703210E-02)
RELAX(W1  ,FALSDT, 2.703210E-02)
RELAX(S140,FALSDT, 2.703210E+02)
RELAX(S142,FALSDT, 2.703210E+02)
RELAX(S145,FALSDT, 2.703210E+02)
RELAX(S147,FALSDT, 2.703210E+02)
RELAX(S158,FALSDT, 2.703210E+02)
RELAX(TEM1,LINRLX, 3.000000E-01)
************************************************************
 Group 18. Limits
VARMAX(U1  ) = 1.000000E+03 ;VARMIN(U1  ) =-1.000000E+03
VARMAX(V1  ) = 1.000000E+03 ;VARMIN(V1  ) =-1.000000E+03
VARMAX(W1  ) = 1.000000E+03 ;VARMIN(W1  ) =-1.000000E+03
VARMAX(S80 ) = 1.000000E+00 ;VARMIN(S80 ) = 1.000000E-20
VARMAX(S140) = 1.000000E+00 ;VARMIN(S140) = 1.000000E-20
VARMAX(S142) = 1.000000E+00 ;VARMIN(S142) = 1.000000E-20
VARMAX(S145) = 1.000000E+00 ;VARMIN(S145) = 1.000000E-20
VARMAX(S147) = 1.000000E+00 ;VARMIN(S147) = 1.000000E-20
```

```
 VARMAX(S158) = 1.000000E+00 ;VARMIN(S158) = 1.000000E-20
 VARMAX(TEM1) = 3.000000E+03 ;VARMIN(TEM1) = 2.600000E+02
 ************************************************************
  Group 19. EARTH Calls To GROUND Station
 NAMGRD  =CVD
 CSG10   ='Q1'
 SPEDAT(SET,CVD,THMDIF,L,T)
 SPEDAT(SET,CVD,THMOPT,I,1)
 SPEDAT(SET,CVD,THMFRQ,I,1)
 SPEDAT(SET,CVD,THMRLX,R,1.00000E+00)
 SPEDAT(SET,CVD,MCDOPT,I,2)
 SPEDAT(SET,CVD,BINOPT,I,4)
 SPEDAT(SET,CVD,MCPROP,I,3)
 SPEDAT(SET,CVD,CHMRLX,R,5.00000E-01)
 SPEDAT(SET,CVD,NGREAC,I,5)
 SPEDAT(SET,CVD,GREAC(1),I,6)
 SPEDAT(SET,CVD,GREAC(2),I,7)
 SPEDAT(SET,CVD,GREAC(3),I,9)
 SPEDAT(SET,CVD,GREAC(4),I,10)
 SPEDAT(SET,CVD,GREAC(5),I,16)
 SPEDAT(SET,CVD,NSREAC,I,5)
 SPEDAT(SET,CVD,SREAC(1),I,11)
 SPEDAT(SET,CVD,SREAC(2),I,12)
 SPEDAT(SET,CVD,SREAC(3),I,13)
 SPEDAT(SET,CVD,SREAC(4),I,14)
 SPEDAT(SET,CVD,SREAC(5),I,15)
 ************************************************************
  Group 20. Preliminary Printout
 ECHO    =    T
 ************************************************************
  Group 21. Print-out of Variables
 ************************************************************
  Group 22. Monitor Print-Out
 IXMON   =       8 ;IYMON  =      16 ;IZMON  =      27
 NPRMNT  =       1
 TSTSWP  =      -1
 ************************************************************
  Group 23.Field Print-Out \& Plot Control
   No PATCHes used for this Group
 ************************************************************
  Group 24. Dumps For Restarts
m
STOP
```

BIBLIOGRAPHY

[1] R. A. Adomaitis. Rapid thermal process model reduction via empirical eigenfunctions: A collocation approach. In *Proceedings of the AIChE Annual Meeting*, Miami, FL, 1995.

[2] R. A. Adomaitis. RTCVD model reduction: A collocation on empirical eigenfunctions approach. Technical Report T.R. 95-64, Institute for Systems Research, 1995.

[3] N. I. Akhiezer and I. M. Glazman. *Theory of Linear Operators in Hilbert Space*. Frederick Ungar Publishing Co., 1961.

[4] Ubaid M. Al-Saggaf and Gene F. Franklin. Model reduction via balanced realizations: An extension and frequency weighting techniques. *IEEE Transactions on Automatic Control*, 33(7):687–692, July 1988.

[5] H. Aling, Suman Banerjee, Anil K. Bangia, Vernon Cole, Jon Ebert, Abbas Emami-Naeini, Klavs F. Jensen, Ioannis G. Kevrekidis, and Stanislav Shvartsman. Nonlinear model reduction for simulation and control of rapid thermal processing. In *Proceedings of the 1997 American Control Conference*, pages 2233–2238, 1997.

[6] H. Aling, J.L. Ebert, A. Emami-Naeini, and R.L. Kosut. Application of a nonlinear model reduction method to rapid thermal processing reactors. In *Proceedings of the IFAC World Congress*, 1996.

[7] Ludwig Arnold. *Stochastic Differential Equations: Theory and Applications*. Wiley, 1974.

[8] Robert B. Ash and Melvin F. Gardner. *Topics in Stochastic Processes*. Academic Press, 1975.

[9] ASM America, Inc., Phoenix, AZ. *ASM Epitaxy Epsilon-1 Reactor Manual*, 1996.

[10] Karl J. Astrom. *Introduction to Stochastic Control Theory*. Academic Press, 1970.

[11] Nadine Aubry, Philip Holmes, John L. Lumley, and Emily Stone. The dynamics of coherent structures in the wall region of a turbulent boundary layer. *Journal of Fluid Mechanics*, 192:115–173, 1988.

[12] K.S. Ball, L. Sirovich, and L.R. Keefe. Dynamical eigenfunction decomposition of turbulent channel flow. *International Journal for Numerical Methods in Fluids*, 12:585–604, 1991.

[13] Suman Banerjee, J. Vernon Cole, and Klavs F. Jensen. Nonlinear model reduction strategies for rapid thermal processing systems. *IEEE Transactions on Semiconductor Manufacturing*, 11(2):266–275, May 1998.

[14] R. H. Bartels and G. W. Stewart. Solution of the matrix equation AX + XB = C. *Commun. Ass. Comput. Mach.*, 15:820–826, 1972.

[15] Gal Berkooz, Philip Holmes, and John L. Lumley. The proper orthogonal decomposition in the analysis of turbulent flows. *Annual Reviews of Fluid Mechanics*, 25:539–575, 1993.

[16] Garrett Birkhoff and Saunders MacLane. *A Survey of Modern Algebra*. Macmillan, 1977.

[17] P. A. Blelloch, D. L. Mingori, and J. D. Wei. Perturbation analysis of internal balancing for lightly damped mechanical systems with gyroscopic and circulatory forces. *Journal of Guidance*, 10(4):406–410, July 1987.

[18] Victor E. Borisenko and Peter J. Hesketh. *Rapid Thermal Processing of Semiconductors*. Plenum Press, 1997.

[19] Kenneth S. Breuer and Lawrence Sirovich. The use of the Karhunen-Loeve procedure for the calculation of linear eigenfunctions. *Journal of Computational Physics*, 96:277–296, 1991.

[20] Chris Brislawn and I. G. Rosen. Wavelet based approximation in the optimal control of distributed parameter systems. *Numerical Functional Analysis and Optimization*, 12(1 and 2):33–77, 1991.

[21] Roger W. Brockett. Nonlinear systems and differential geometry. *Proceedings of the IEEE*, 64(1):61–72, January 1976.

[22] Roger W. Brockett. Notes on stochastic processes on manifolds. In C. I. Byrnes, B. N. Datta, D. S. Gilliam, and C. F. Martin, editors, *Systems and Control in the Twenty-First Century*. Birkhauser, 1997.

[23] Stephen A. Campbell. Rapid thermal processing. In M. Meyyappan, editor, *Computational Modeling in Semiconductor Processing*, chapter 6, pages 325–350. Artech House, 1995.

[24] S. Chatterjee, I. Trachtenberg, and T. F. Edgar. Modeling of a single wafer rapid thermal reactor. *Journal of the Electrochemical Society*, 139:3682–3689, December 1992.

[25] Chien-Chong Chen and Hsueh-Chia Chang. Accelerated disturbance damping of an unknown distributed system by nonlinear feedback. *AIChE Journal*, 38(9):1461–1476, September 1992.

[26] Kuo-Ping Chiao. *Least Square Model Reduction for Vibration Control of Complex Mechanical Systems*. PhD thesis, University of Maryland, College Park, 1992.

[27] Byung-Jin Cho, Peter Vandenabeele, and Karen Maex. Development of a hexagonal-shaped rapid thermal processor using a vertical tube. *IEEE Transactions on Semiconductor Manufacturing*, 7(3):345–353, August 1994.

[28] Y. M. Cho and T. Kailath. Model identification in RTP systems. *IEEE Transactions on Semiconductor Manufacturing*, 6(3):233–245, August 1993.

[29] Y. M. Cho, Arogyaswami Paulraj, Thomas Kailath, and Guanghan Xu. A contribution to optimal lamp design in rapid thermal processing. *IEEE Transactions on Semiconductor Manufacturing*, 7(1):34–41, February 1994.

[30] Young Man Cho and Paul Gyugyi. Control of rapid thermal processing: A system theoretic approach. *IEEE Transactions on Control Systems Technology*, 5(6):644–653, November 1997.

[31] M. E. Coltrin, R. J. Kee, and J. A. Miller. A mathematical model of silicon chemical vapor deposition. *Journal of the Electrochemical Society*, 133:1206, 1986.

[32] M. G. Crandall and P. L. Lions. Two approximations of solutions of Hamilton-Jacobi equations. *Mathematics of Computation*, 43(167):1–19, July 1984.

[33] M. G. Crandall and P. L. Lions. Viscosity solutions of Hamilton-Jacobi equations. *Trans. Amer. Math. Soc.*, 282:487–502, 1984.

[34] Ruth F. Curtain and Keith Glover. Balanced realisations for infinite-dimensional systems. In H. Bart, I. Gohberg, and M. A. Kaashoek, editors, *Proceedings of the Workshop on Operator Theory and Systems*, pages 87–104. Birkhauser Verlag Basel, 1986.

[35] Ingrid Daubechies. *Ten Lectures on Wavelets*. SIAM, 1992.

[36] M. H. A. Davis. *Linear Estimation and Stochastic Control*. Halsted Press, 1977.

[37] W. P. Dayawansa. Class notes: Advanced nonlinear control. University of Maryland, 1996.

[38] W. B. de Boer and D. J. Meyer. Low temperature CVD of epitaxial Si and Si-Ge layers at atmospheric pressure. *Applied Physics Letters*, 58(12):1286–1288, March 1991.

[39] Dick de Roover, Abbas Emami-Naeini, Jon L. Ebert, Sarbajit Ghosal, and Gwen W. van der Linden. Model-based control of fast-ramp RTP systems. In *Proceedings of the 6th International Rapid Thermal Processing Conference, RTP98*, 1998.

[40] Anil E. Deane and Lawrence Sirovich. A computational study of Rayleigh-Benard convection. Part 1. Rayleigh-number scaling. *Journal of Fluid Mechanics*, 222:231–250, 1991.

[41] Jean-Marie Dilhac, Nicolas Nolhier, Christian Ganibal, and Christine Zanchi. Thermal modeling of a wafer in a rapid thermal processor. *IEEE Transactions on Semiconductor Manufacturing*, 8(4):432–439, November 1995.

[42] D. Elliott. *Controllable systems driven by white noise*. PhD thesis, University of California, Los Angeles, 1969.

[43] D. Elliott. Diffusions on manifolds arising from controllable systems. In D. Q. Mayne and R. W. Brockett, editors, *Geometric Methods in System Theory*. Reidel, Dordrecht, Holland, 1973.

[44] D. F. Enns. Model reduction with balanced realizations: An error bound and a frequency weighted generalization. In *Proceedings of the 23rd Conference on Decision and Control*, pages 127–132, 1984.

[45] Lawrence C. Evans. *Partial Differential Equations*. Berkley Mathematics Lecture Notes, 1994.

[46] K. V. Fernando and H. Nicholson. Singular perturbational model reduction of balanced systems. *IEEE Transactions on Automatic Control*, 27(2):466–468, April 1982.

[47] Wendell H. Fleming and H. Mete Soner. *Controlled Markov Processes and Viscosity Solutions*. Springer-Verlag, 1993.

[48] L. Fortuna, G. Nunnari, and A. Gallo. *Model Order Reduction Techniques with Applications in Electrical Engineering*. Springer-Verlag, 1992.

[49] Dimitrios I. Fotiadis, Shigekazu Kieda, and Klavs F. Jensen. Transport phenomena in vertical reactors for metalorganic vapor phase epitaxy: 1. Effects of heat transfer characteristics, reactor geometry, and operating conditions. *Journal of Crystal Growth*, 102:441–470, 1990.

[50] M. I. Friswell, J. E. T. Penny, and S. D. Garvey. The application of the IRS and balanced realization methods to obtain reduced models of structures with local nonlinearities. *Journal of Sound and Vibration*, 196(4):453–468, 1996.

[51] A. T. Fuller. Analysis of nonlinear stochastic systems by means of the Fokker-Planck equation. *International Journal of Control*, 9(6):603–655, 1969.

[52] David H. Gay and Harmon Ray. Identification and control of distributed parameter systems by means of the singular value decomposition. *Chemical Engineering Science*, 50(10):1519–1539, 1995.

[53] R. Genesio and M. Milanese. A note on the derivation and use of reduced-order models. *IEEE Transactions on Automatic Control*, 21(2):118–122, February 1976.

[54] Keith Glover. All optimal Hankel-norm approximations of linear multivariable systems and their L^∞-error bounds. *International Journal of Control*, 39(6):1115–1193, 1984.

[55] Keith Glover, Ruth F. Curtain, and Jonathan R. Partington. Realisation and approximation of linear infinite-dimensional systems with error bounds. *SIAM Journal of Control and Optimization*, 26(4):863–898, July 1988.

[56] Israel Gohberg and Seymour Goldberg. *Basic Operator Theory*. Birkhauser, 1981.

[57] Martin Golubitsky and Victor Guillemin. *Stable Mappings and Their Singularities*. Springer-Verlag, 1973.

[58] W. Steven Gray and Joseph Mesko. General input balancing and model reduction for linear and nonlinear systems. In *Proceedings of the 1997 European Controls Conference*, Brussells, Belgium, 1997.

[59] W. Steven Gray and Jacquelien M. A. Scherpen. Hankel operators and gramians for nonlinear systems. In *Proceedings of the 37th Conference on Decision and Control*, Tampa, FL, 1998.

[60] Victor Guillemin and Alan Pollack. *Differential Topology*. Prentice Hall, 1974.

[61] Morton E. Gurtin. *An Introduction to Continuum Mechanics*. Academic Press, 1981.

[62] Paul Gyugyi. *Model-Based Control Applied to Rapid Thermal Processing*. PhD thesis, Stanford University, June 1993.

[63] R. S. Gyurcsik, T. J. Riley, and F. Y. Sorrell. A model for rapid thermal processing: Achieving uniformity through lamp control. *IEEE Transactions on Semiconductor Manufacturing*, 4(1):9–13, February 1991.

[64] S. J. Hammarling. Numerical solution of the stable, non-negative definite Lyapunov equation. *IMA Journal of Numerical Analysis*, 2:303–323, 1982.

[65] M. T. Heath, A. J. Laub, C. C. Paige, and R. C. Ward. Computing the SVD of a product of two matrices. *SIAM J. Sci. Stat. Comp.*, 7:1147–1159, 1986.

[66] Uwe Helmke and John B. Moore. *Optimization and Dynamical Systems*. Springer, 1994.

[67] Philip Holmes, John L. Lumley, and Gal Berkooz. *Turbulence, Coherent Structures, Dynamical Systems, and Symmetry*. Cambridge University Press, 1996.

[68] A. Hu, E. Sachs, and A. Ingolfsson. Run by run process control: Performance benchmarks. In *IEEE/SEMI International Semiconductor Manufacturing Science Symposium*, pages 73–78, 1992.

[69] Alberto Isidori. *Nonlinear Control Systems*. Springer-Verlag, 1989.

[70] Kazufumi Ito and S. S. Ravindran. A reduced order method for simulation and control of fluid flows. *Journal of Computational Physics*, 143(2):403, July 1998.

[71] Edmond A. Jonckheere and Leonard M. Silverman. A new set of invariants for linear systems - application to reduced order compensator design. *IEEE Transactions on Automatic Control*, 28(10):953–964, October 1983.

[72] Thomas Kailath. *Linear Systems*. Prentice-Hall, 1980.

[73] T. I. Kamins and D. J. Meyer. Kinetics of Si-Ge deposition by APCVD. *Applied Physics Letters*, 59(2):90, July 1991.

[74] T. I. Kamins and D. J. Meyer. Effect of silicon source gas on silicon-germanium chemical vapor deposition kinetics at atmospheric pressure. *Applied Physics Letters*, 61(1):90, July 1992.

[75] Ted Kamins. *Polysilicon for IC Applications*. Kluwer Academic, 1988.

[76] K. Karhunen. Zur spektral theorie stochastischer prozesse. *Ann. Acad. Sci. Fennicae*, Ser. A 34, 1946.

[77] Tosio Kato. *A Short Introduction to Perturbation Theory for Linear Operators*. Springer-Verlag, 1982.

[78] A. Kersch. RTP reactor simulations. *The PHOENICS Journal of Computational Fluid Dynamics and its Applications*, 8(4):500–511, December 1995.

[79] Hassan Khalil. *Nonlinear Systems*. Macmillan, 1992.

[80] W. J. Kiether, M. J. Fordham, Seungil Yu, A. J. Silva Neto, K. A. Conrad, J. R. Hauser, F. Y. Sorrell, and J. J. Wortman. Three-zone rapid thermal processor system. In *Proceedings of the 2nd International Rapid Thermal Processing Conference, RTP94*, pages 96–101, Monterey, CA, 1994.

[81] C. R. Kleijn. A mathematical model of the hydrodynamics and gas-phase reactions in silicon LPCVD in a single-wafer reactor. *Journal of the Electrochemical Society*, 138(7):2190–2200, July 1991.

[82] C. R. Kleijn. Chemical vapor deposition processes. In M. Meyyappan, editor, *Computational Modeling in Semiconductor Processing*, chapter 4, pages 97–229. Artech House, 1995.

[83] C. R. Kleijn and K. J. Kuijlaars. The modelling of transport phenomena in CVD reactors. *The PHOENICS Journal of Computational Fluid Dynamics and its Applications*, 8(4):404–420, December 1995.

[84] C. R. Kleijn, K. J. Kuijlaars, and H. E. A. van den Akker. Modeling of a cold wall tungsten CVD reactor. *The PHOENICS Journal of Computational Fluid Dynamics and its Applications*, 8(4):465–490, December 1995.

[85] Karson L. Knutson, Stephen A. Campbell, and Floyd Dunn. Modeling of three-dimensional effects on temperature uniformity in rapid thermal processing of eight inch wafers. *IEEE Transactions on Semiconductor Manufacturing*, 7(1):68–72, February 1994.

[86] Robert L. Kosut, Dick de Roover, Abbas Emami-Naeini, and Jon L. Ebert. Run-to-run control of static systems. In *Proceedings of the 37th Conference on Decision and Control*, Tampa, FL, December 1998. IEEE.

[87] P. S. Krishnaprasad. Class notes: Geometric control theory. University of Maryland, 1998.

[88] Serge Lang. *Differential Manifolds*, chapter 7. Springer-Verlag, 1985.

[89] P. Langevin. *C. R. Hebd. Seanc. Academie des Sciences (Paris)*, 146:530, 1908.

[90] A. J. Laub. On computing "balancing" transformations. In *Proceedings of the 1980 Joint Automatic Control Conference*, San Francisco, CA, August 1980.

[91] Alan J. Laub, Michael T. Heath, Chris C. Paige, and Robert C. Ward. Computation of system balancing transformations and other applications of simultaneous diagonalization algorithms. *IEEE Transactions on Automatic Control*, 32(2):115–122, February 1987.

[92] M. Liehr, C. M. Greenlief, S. R. Kasi, and M. Offenberg. Kinetics of silicon epitaxy using SiH_4 in a rapid thermal chemical vapor deposition reactor. *Applied Physics Letters*, 56(7):629–631, February 1990.

[93] P. L. Lions. *Generalized Solutions of Hamilton-Jacobi Equations*. Pitman Advanced Publishing Program, 1982.

[94] Chang-Huan Liu. A comparison of optimal and suboptimal estimators and estimation lower bounds. Master's thesis, University of Texas, Austin, TX, 1978.

[95] M. Loeve. Functions de second ordre. *C. R. Academie des Sciences (Paris)*, 229, 1945.

[96] Michel Loeve. *Probability Theory*. D. Van Nostrand Company, Inc., 1963.

[97] H. A. Lord. Thermal and stress analysis of semiconductor wafers in a rapid thermal processing oven. *IEEE Transactions on Semiconductor Manufacturing*, 1(3):105–114, August 1988.

[98] Guangquan Lu, Monalisa Bora, Laura L. Tedder, and Gary W. Rubloff. Integrated dynamic simulation of rapid thermal chemical vapor deposition of polysilicon. *IEEE Transactions on Semiconductor Manufacturing*, 11(1):63–74, February 1998.

[99] John Lumley and Peter Blossey. Control of turbulence. *Annual Review of Fluid Mechanics*, 30:311–327, 1998.

[100] Vikram Manikonda. *Control and Stabilization of a Class of Nonlinear Systems with Symmetry*. PhD thesis, University of Maryland, College Park, 1998.

[101] Steven Marcus. Class notes: Stochastic control. University of Maryland, 1997.

[102] The Mathworks, Inc. Matlab ver. 5.3. Numeric Computation and Visualization Software, 1999.

[103] L. Meirovitch. *Methods of Analytical Dynamics*. McGraw-Hill, 1970.

[104] D. G. Meyer. Fractional balanced reduction - model reduction via fractional representation. *IEEE Transactions on Automatic Control*, 35(12):1341–1345, December 1990.

[105] Doug Meyer. Private communication. ASM America, Inc., Phoenix, AZ, March 1997.

[106] Douglas Meyer, Tony Komasa, and Rod Smith. Private communication. ASM America, Inc., Phoenix, AZ, April 1998.

[107] Meyya Meyyappan. Various private communications. NASA Ames, October-March 1998-1999.

[108] J. Milnor. *Morse Theory*. Princeton University Press, 1963.

[109] Bruce C. Moore. Principal component analysis in linear systems: Controllability, observability, and model reduction. *IEEE Transactions on Automatic Control*, 26(1):17–32, February 1981.

[110] M. Morse. *The Calculus of Variations in the Large*. AMS, 1934.

[111] C. T. Mullis and R. A. Roberts. Synthesis of minimum roundoff noise fixed point digital filters. *IEEE Transactions on Circuits and Systems*, 23:551–562, September 1976.

[112] Richard M. Murray, Zexiang Li, and S. Shankar Sastry. *A Mathematical Introduction to Robotic Manipulation*. CRC Press, 1994.

[113] Denis Mustafa and Keith Glover. Controller reduction by H_∞-balanced truncation. *IEEE Transactions on Automatic Control*, 36(6):668–682, June 1991.

[114] P. Narayan. Class notes: Random processes. University of Maryland, 1992.

[115] Andrew Newman and P. S. Krishnaprasad. Modeling and optimization for epitaxial growth: Transport and growth studies. Technical Report T.R. 99-19, Institute for Systems Research, March 1999.

[116] Andrew Newman, P. S. Krishnaprasad, Sam Ponczak, and Paul Brabant. Modeling and model reduction for epitaxial growth. In *Proceedings of the SEMATECH AEC/APC Workshop IX*, Lake Tahoe, NV, September 1997. SEMATECH.

[117] Andrew Newman, P. S. Krishnaprasad, Sam Ponczak, and Paul Brabant. Modeling and model reduction for control and optimization of epitaxial growth in a commercial RTCVD reactor. Technical Report T.R. 98-45, Institute for Systems Research, March 1998.

[118] Andrew J. Newman. Model reduction via the Karhunen-Loeve expansion Part I: An exposition. Technical Report T.R. 96-32, Institute for Systems Research, April 1996.

[119] Andrew J. Newman. Model reduction via the Karhunen-Loeve expansion Part II: Some elementary examples. Technical Report T.R. 96-33, Institute for Systems Research, April 1996.

[120] Andrew J. Newman and P. S. Krishnaprasad. Nonlinear model reduction for RTCVD. In *Proceedings of the 32nd Conference on Information Sciences and Systems*, pages 819–824, Princeton, NJ, March 1998.

[121] Henk Nijmeier and Arjan van der Schaft. *Nonlinear Dynamical Control Systems*. Springer-Verlag, 1990.

[122] Raimund Ober and Duncan McFarlane. Balanced canonical forms for minimal systems: A normalized coprime factor approach. *Linear Algebra and Its Applications*, pages 23–64, 1989.

[123] Raimund J. Ober. Balanced realizations: Canonical form, parameterization, model reduction. *International Journal of Control*, 46(2):643–670, 1987.

[124] M. Necati Ozisik. *Heat Transer: A Basic Approach*. McGraw-Hill, 1985.

[125] Richard S. Palais. Morse theory on Hilbert manifolds. *Topology*, 2:299–340, 1963.

[126] Athanasios Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, 2nd edition, 1984.

[127] Lars Pernebo and Leonard M. Silverman. Model reduction via balanced state space representations. *IEEE Transactions on Automatic Control*, 27(2):382–387, April 1982.

[128] Sam Ponczak, Thomas Knight, Michael O'Loughlin, and Paul Brabant. Various private communications. Northrop Grumman ESSS, Linthicum, MD, March-October 1998.

[129] Ywh Pyng. Private communication. Integrated Systems, Inc., September 1997.

[130] W. Fred Ramirez and Jan M. Maciejowski. Balanced realization for state space identification and optimal output regulation. *AIChE Journal*, 41(5):1217–1228, May 1995.

[131] Seth T. Rodgers and Klavs F. Jensen. Multiscale modeling of chemical vapor deposition. *Journal of Applied Physics*, 83(1):524–530, January 1998.

[132] Fred Roozeboom. Rapid thermal processing systems: A review with emphasis on temperature control. *Journal of Vacuum Science*, B8(6):1249–1258, November 1990.

[133] W. R. Runyan and K. E. Bean. *Semiconductor Integrated Circuit Processing Technology*, chapter 4. Addison-Wesley, 1990.

[134] W. R. Runyan and K. E. Bean. *Semiconductor Integrated Circuit Processing Technology*, chapter 7, pages 294–360. Addison-Wesley, 1990.

[135] Emanuel Sachs, George H. Prueger, and Roberto Guerrieri. An equipment model for polysilicon LPCVD. *IEEE Transactions on Semiconductor Manufacturing*, 5(1):3–13, February 1992.

[136] M. G. Safonov and R. Y. Chiang. A Schur method for balanced truncation model reduction. *IEEE Transactions on Automatic Control*, 34(7):729–733, July 1989.

[137] Jerrold E. Marsden Sanjay Lall and Sonja Glavaski. Empirical model reduction of controlled nonlinear systems. In *Proceedings of the 1999 IFAC Congress*, 1999.

[138] C. D. Schaper, Y. M. Cho, and T. Kailath. Low-order modeling and dynamic characterization of rapid thermal processing. *Applied Physics A*, 54:317–326, 1992.

[139] Charles Schaper and Thomas Kailath. Thermal model validation for RTCVD of polysilicon. *Journal of the Electrochemical Society*, 143(1):374–381, January 1996.

[140] Jacquelien M. A. Scherpen. Balancing for nonlinear systems. *Systems and Control Letters*, 21:143–153, 1993.

[141] Jacquelien M. A. Scherpen. *Balancing for Nonlinear Systems*. PhD thesis, University of Twente, The Netherlands, 1994.

[142] Jacquelien M. A. Scherpen and W. Steven Gray. Sufficient conditions for minimality of a nonlinear realization via controllability and observability functions. In *Proceedings of the 1998 American Control Conference*, Philadelphia, PA, June 1998.

[143] Adam L. Schwartz and E. Polak. *RIOTS: A Matlab Toolbox for Solving Optimal Control Problems*. University of California at Berkeley, Berkeley, CA, version 1.0 edition, 1996.

[144] Arthur Sherman. *Chemical Vapor Deposition For Microelectronics*. Noyes Publications, 1987.

[145] Robert Siegel and John R. Howell. *Thermal Radiation Heat Transfer*. Hemisphere, 1992.

[146] L. Sirovich and M. Kirby. Low-dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America A*, 4(3):519–524, March 1987.

[147] Lawrence Sirovich. Turbulence and the dynamics of coherent structures; Part I: Coherent structures. *Quarterly Appl. Math.*, 45(3):561–571, October 1987.

[148] Lawrence Sirovich. Turbulence and the dynamics of coherent structures; Part III: Dynamics and scaling. *Quarterly of Applied Mathematics*, 45(3):583–590, October 1987.

[149] Lawrence Sirovich. Empirical eigenfunctions and low dimensional systems. In Lawrence Sirovich, editor, *New Perspectives in Turbulence*, chapter 5, pages 139–163. Springer-Verlag, 1991.

[150] F. Smithies. *Integral Equations*. Cambridge University Press, 1962.

[151] F. Y. Sorrell, M. J. Fordham, M. C. Ozturk, and J. J. Wortman. Temperature uniformity in RTP furnaces. *IEEE Transactions on Electron Devices*, 39(1):75–79, January 1992.

[152] F. Yates Sorrell, Seungil Yu, and William J. Kiether. Applied RTP optical modeling: An argument for model-based control. *IEEE Transactions on Semiconductor Manufacturing*, 7(4):454–459, November 1994.

[153] Panagiotis E. Souganidis. Approximation schemes for viscosity solutions of Hamilton-Jacobi equations. *Journal of Differential Equations*, 59:1–43, 1985.

[154] Gilbert Strang. *Linear Algebra and Its Applications*. Harcourt Brace Jovanovich, Inc., 1988.

[155] John Douglas Stuber, Isaac Trachtenbert, and Thomas F. Edgar. Design and modeling of rapid thermal processing systems. *IEEE Transactions on Semiconductor Manufacturing*, 11(3):442–457, August 1998.

[156] S. M. Sze. *Physics of Semiconductor Devices*. Wiley Interscience, 1981.

[157] Artemis Theodoropoulou, Raymond A. Adomaitis, and Evanghelos Zafiriou. Model reduction for optimization of rapid thermal chemical vapor deposition systems. *IEEE Transactions on Semiconductor Manufacturing*, 11(1):85–98, February 1998.

[158] A. J. van der Schaft and J. E. Oeloff. Model reduction of linear conservative mechanical systems. *IEEE Transactions on Automatic Control*, 35(6):729–733, June 1990.

[159] M. van Hulzen. Pseudo-balancing the double pendulum. Master's thesis, University of Twente, Faculty of Applied Mathematics, 1994.

[160] Erik I. Verriest. Model reduction via balancing and connections with other methods. In Uday B. Desai, editor, *Modeling and Application of Stochastic Processes*, chapter 6, pages 123–154. Kluwer, 1986.

[161] M. Vidyasagar. *Nonlinear Systems Analysis*. Prentice-Hall, 1993.

[162] Pei-Jih Wang. Epitaxy. In C. Y. Chang and S. M. Sze, editors, *ULSI Technology*, chapter 3, pages 105–143. McGraw-Hill, 1996.

[163] Chr. Werner and M. Hierlemann. Application of PHOENICS-CVD to epitaxial Si-Ge, polysilicon, and silicon deposition in a range of CVD reactors. *The PHOENICS Journal of Computational Fluid Dynamics and its Applications*, 8(4):538–552, December 1995.

[164] E. Wong and M. Zakai. On the convergence of ordinary integrals to stochastic integrals. *Annals of Mathematical Statistics*, 36:1560–1564, 1965.

[165] E. Wong and M Zakai. On the relation between ordinary and stochastic differential equations. *International Journal of Engineering Science*, 3:213–229, 1965.

[166] Eugene Wong. *Stochastic Processes in Information and Dynamical Systems*. McGraw-Hill, 1971.

[167] D. J. Wright. The digital simulation of stochastic differential equations. *IEEE Transactions on Automatic Control*, 19:75–76, February 1974.

[168] Evanghelos Zafiriou, Raymond A. Adomaitis, and Gangadhar Gattu. An approach to Run-to-Run control for RTP. In *Proceedings of the 1995 ACC*, 1995.

[169] Moshe Zakai. A Lyapunov criterion for the existence of stationary probability distributions for systems perturbed by noise. *SIAM Journal of Control*, 7(3):390–397, August 1969.

[170] Kemin Zhou and John C. Doyle. *Essentials of Robust Control*, chapter 7, pages 105–128. Prentice-Hall, 1998.

[171] W. Q. Zhu and Y. Q. Yang. Exact stationary solutions of stochastically excited and dissipated integrable Hamiltonian systems. *Journal of Applied Mechanics*, 63:493–500, June 1996.