DESIGN AND ANALYSIS OF STOCHASTIC DYNAMICAL SYSTEMS WITH FOKKER-PLANCK EQUATION

A Dissertation

by

MRINAL KUMAR

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

December 2009

Major Subject: Aerospace Engineering

DESIGN AND ANALYSIS OF STOCHASTIC DYNAMICAL SYSTEMS WITH FOKKER-PLANCK EQUATION

A Dissertation

by

MRINAL KUMAR

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Co-Chairs of Committee,	Suman Chakravorty
	John L. Junkins
Committee Members,	Srinivas R. Vadali
	David C. Hyland
	Shankar P. Bhattacharyya
Head of Department,	Dimitris C. Lagoudas

December 2009

Major Subject: Aerospace Engineering

ABSTRACT

Design and Analysis of Stochastic Dynamical Systems

with Fokker-Planck Equation. (December 2009)

Mrinal Kumar, B.Tech., Indian Institute of Technology, Kanpur;

Co–Chairs of Advisory Committee: Dr. Suman Chakravorty
Dr. John L. Junkins

This dissertation addresses design and analysis aspects of stochastic dynamical systems using Fokker-Planck equation (FPE). A new numerical methodology based on the partition of unity meshless paradigm is developed to tackle the greatest hurdle in successful numerical solution of FPE, namely the curse of dimensionality. A local variational form of the Fokker-Planck operator is developed with provision for $h-$ and $p-$ refinement. The resulting high dimensional weak form integrals are evaluated using quasi Monte-Carlo techniques. Spectral analysis of the discretized Fokker-Planck operator, followed by spurious mode rejection is employed to construct a new semi-analytical algorithm to obtain near real-time approximations of transient FPE response of high dimensional nonlinear dynamical systems in terms of a reduced subset of admissible modes. Numerical evidence is provided showing that the curse of dimensionality associated with FPE is broken by the proposed technique, while providing problem size reduction of several orders of magnitude.

In addition, a simple modification of norm in the variational formulation is shown to improve quality of approximation significantly while keeping the problem size fixed. Norm modification is also employed as part of a recursive methodology for tracking the optimal finite domain to solve FPE numerically.

The basic tools developed to solve FPE are applied to solving problems in nonlin-

ear stochastic optimal control and nonlinear filtering. A policy iteration algorithm for stochastic dynamical systems is implemented in which successive approximations of a forced backward Kolmogorov equation (BKE) is shown to converge to the solution of the corresponding Hamilton Jacobi Bellman (HJB) equation. Several examples, including a four-state missile autopilot design for pitch control, are considered.

Application of the FPE solver to nonlinear filtering is considered with special emphasis on situations involving long durations of propagation in between measurement updates, which is implemented as a weak form of the Bayes rule. A nonlinear filter is formulated that provides complete probabilistic state information conditioned on measurements. Examples with long propagation times are considered to demonstrate benefits of using the FPE based approach to filtering.

To my parents, Sheela Sahai and Ashoka Kumar Sahai;

To my lovely wife and closest friend, Vasudha; and

To *the lost cause*, and all lives lost in its pursuit.

ACKNOWLEDGMENTS

I would like to give special thanks to my advisors Dr. Suman Chakravorty and Dr. John L. Junkins for their support and guidance throughout my stay in Texas A&M University. Dr. Junkins' unmatched experience and insight and Dr. Chakravorty's enviable enthusiasm were crucial ingredients of a memorable learning experience that will surely inspire me throughout my career. I also thank my committee members Dr. Srinivas Vadali, Dr. David Hyland and Dr. Shankar Bhattacharyya for their guidance. I am grateful to Dr. Daniele Mortari for all the teaching opportunities he gave me, all the engrossing discussions on celestial mechanics and indeed all the delightful pasta. Thanks to Dr. John Valasek and Dr. Suman Chakravorty for giving me opportunities to expand my research horizons through the project on morphing and reconfigurable systems. I am grateful to "life-saver" Lisa Willingham for managing almost anything and everything and the ever resourceful Karen Knabe for their assistance throughout my degree program.

I wish to thank my brilliant seniors Atul Ganpatye, Dr. Puneet Singla and Dr. Prasenjit Sengupta for their guidance and motivation, especially during the initial phase of my stay in College Station. Special thanks to Puneet for his guidance and all our discussions about "life, universe and everything" over uncountable cups of coffee.

I feel deeply fortunate to have had the company of wonderful friends who made my stay in College Station a special part of my life: Kevin Brink, Tarek Elgohary, Abhishek Halder, Monika Marwaha, Avinash Prabhakar, Roshmik Saha, Ashivni Shekhawat, Nipun Sinha, Siming Zhao and of-course, my road-trip friends: Gaurav Kumar, Yogesh, Dr. Pankaj Wahi, Sandipan, Dr. Anup Katake, Prabha, Ashwini Kumar and Kumar Avijit.

Finally, I offer my deepest gratitude to my parents and family, my loving in-laws; and last but not least, my lovely wife, Vasudha for all their love, support, encouragement and belief in me. Vasudha, also my closest "road-trip friend," had the misfortune of sharing with me also the bumpy ride that is graduate studies and making it the best time of my life. I hope the successful completion of this work is consolation for all the sacrifices she has made.

TABLE OF CONTENTS

CHAPTER Page

CHAPTER Page

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

FIGURE Page

FIGURE Page

FIGURE Page

FIGURE Page

CHAPTER I

INTRODUCTION

Analysis of stochastic dynamical systems is a mathematically challenging field that has inspired more than a century of researchers. Fokker-Planck equation (FPE) is a key equation encountered in the study of stochastic systems. It is a parabolic partial differential equation that captures the exact description of evolution of the state probability density function (pdf) through continuous dynamical systems perturbed by white noise excitation. An instance of pdf evolution is depicted in Fig.1(a), which shows propagation of an uncertainty cloud (depicted using points) through a stochastic system, clearly illustrating manifestation of underlying nonlinearity in the form of distortion of the initial Gaussian sample.

Unfortunately, analytical solutions of FPE are known to exist for only a handful of dynamical systems, which comprise a very small fraction of systems encountered in real life. Furthermore, due to several difficult issues, the greatest of which is the so called "curse of dimensionality", accurate numerical solutions of FPE for general high dimensional and/or nonlinear systems have remained elusive and drawn attention of numerous researchers for over half a century. Interest in FPE continues to thrive because it lies at the heart of the uncertainty propagation problem and nonlinear filtering. This dissertation considers the core problem of developing a robust numerical algorithm for solving FPE and explores applications in stochastic optimal control and nonlinear filtering. Two variants of the partition of unity based finite element meshless approach are developed to discretize the Fokker-Planck operator on a finite sized domain. Significant improvement over state-of-the-art is demonstrated

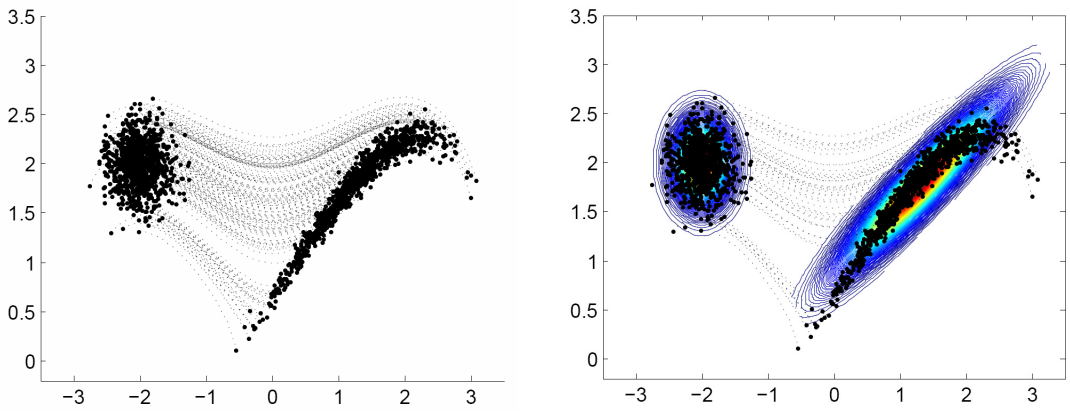
---

The journal model is *IEEE Transactions on Automatic Control*.

and numerical evidence for breaking of the curse of dimensionality is provided. Coupled with spectral analysis of the discretized FP operator, a robust numerical tool is developed that provides transient response of FPE in near real-time. The field of nonlinear filtering, wherein knowledge of the complete state-pdf (as opposed to first few moments) is worth its value in gold, especially stands to gain from such a robust solver. In addition, there are multiple problems in science, engineering and economics (detailed in section B) that can benefit from the methods developed in this work.

The remainder of this chapter is arranged as follows: Section A presents in detail the history of research in the field of stochastic dynamics, starting with its inception to the present state-of-the-art in solving FPE numerically. Section B presents immediate and potential application areas of this research in various fields of science, engineering and economics.

The remainder of this dissertation is organized as follows (see Fig.2 for an overview of structure): Chapter II presents a formal statement of problems considered in this dissertation, followed by a discussion of challenging research issues surrounding these problems. This chapter is linked with appendix A, which reviews important concepts in probability theory and stochastic dynamics with emphasis on physical underpinnings of abstract concepts. Chapter III discusses in length the details of various algorithms developed to solve the considered problems, including meshless partition of unity based discretization (PUFEM) and analytical integration of the resulting ODEs. The various modules of PUFEM and analytical integrator together constitute a semianalytical algorithm for solving FPE in near real-time. Chapter III also presents results for each algorithm discussed. Chapter IV presents recursive norm-modification techniques for solution refinement and domain tracking for nonlinear systems. Chapters V and VI discuss applications of various algorithms developed in Chapter III to problems in nonlinear stochastic optimal control and nonlinear fil-



(a) The state-pdf as a point cloud.

(b) Approximate characterization of the exact pdf

Fig. 1. The problem of uncertainty propagation. The x and y axes depict states of the dynamical system (e.g. x, $\dot{x}$).

tering respectively. Finally, chapter VII draws conclusions and discusses avenues for future extensions of work presented in this discourse.

A. History of Research in Stochastic Dynamics

The roots of stochastic dynamics go back to the investigations of Robert Maxwell and Ludwig Boltzmann into molecular properties of gases. The premise of their investigation was that heat in a medium is essentially random motion of its molecules [1]. Following heuristic arguments, Maxwell (1860) developed an expression for the steady state probability density of individual molecules as an exponential function of their kinetic energy (Maxwell distribution) [2, 3]. Boltzmann (1868) generalized Maxwell's result to include the case of gas molecules subjected to a conservative force-field [4]. The resulting steady-state probability density was then proposed to be an exponential function of the total energy (potential + kinetic), known as the Maxwell-

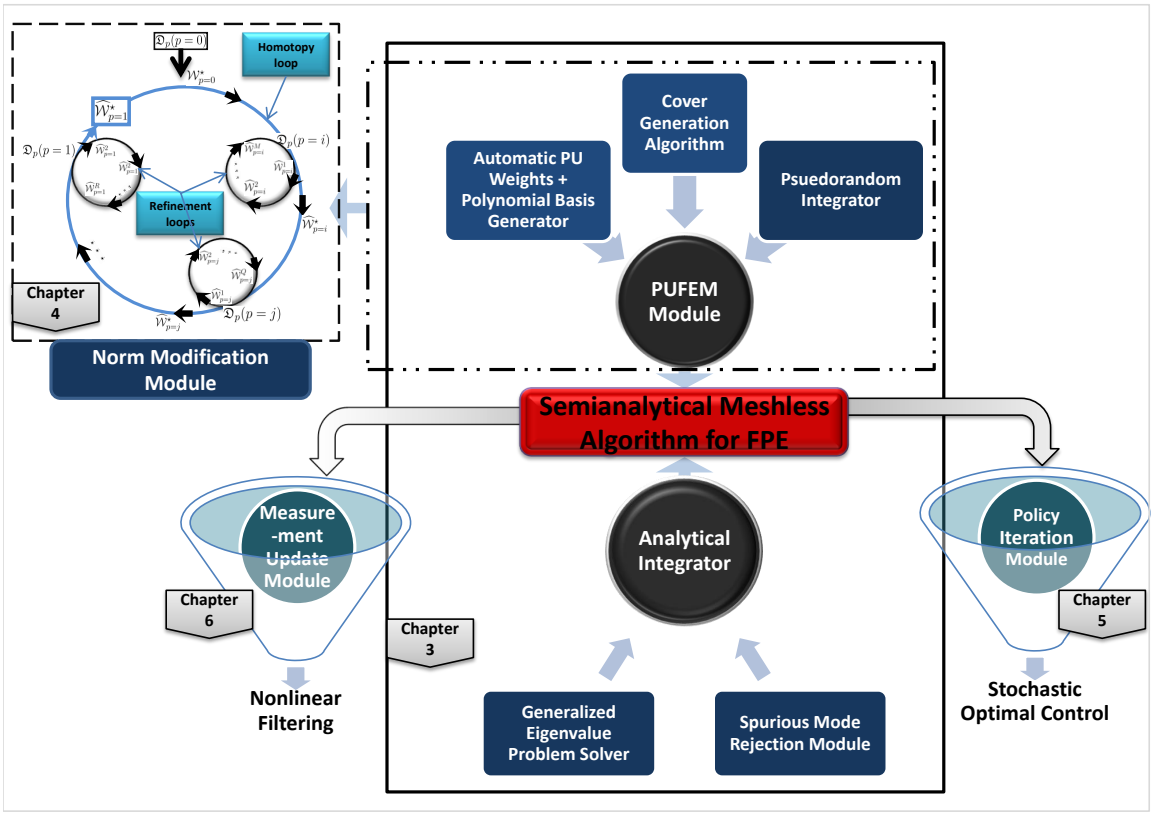


Fig. 2. An overview of different parts of the dissertation.

Boltzmann distribution and argued that a gas with any arbitrary initial distribution would eventually approach the Maxwell-Boltzmann distribution. Together, their work laid the foundation for stochastic dynamics, even though uncertainty entered only via random initial conditions in their studies. The systems they considered were autonomous and conservative with no stochastic forces. These assumptions led to contradictory results and paradox-like situations in their conclusions, e.g., reversibility paradox, recurrence paradox [1].

The inclusion of continuous random disturbances into dynamic analysis occurred around the end of the nineteenth century. Rayleigh (1880) was the first to study the problem of random walk [5, 6] and obtained a partial differential equation (PDE) governing evolution of the probability density of displacement, the "random variable" (1894, 1899) [7]. In 1891, he applied similar analysis to the theory of gases and obtained a PDE for the probability density of velocity of gas molecules [8]. This was the first instance of Fokker-Planck equation in physics (of course, with a different name). Bachelier (1900) constructed a model of the French stock exchange [9], obtaining in the process a simple form of FPE. Between 1910 and 1912, he related this work to the problem of gambler's ruin and obtained a more general version of FPE. The works of Rayleigh and Bachelier went largely unnoticed and the equations they derived were not identified as the "Fokker-Planck equation" [1].

In a second wave of interest in the subject, Albert Einstein brought together the Maxwell-Boltzmann theory with the approach of random walks in his famous 1905 paper on Brownian motion [10]. Einstein's Nobel prize winning work led to popularization of the theory of Brownian motion and formalization of the associated theory of stochastic dynamics. Langevin introduced the formal stochastic differential equation (SDE) in a 1908 paper for modeling dynamical systems perturbed by random disturbances (Langevin's equations) [11]. Fokker studied state-dependent white noise

for a first order system (motion of an electron in a field of random radiation) [12, 13]. Planck applied Fokker's equation to quantum physics and generalized it considerably, handling N-th order dynamical systems with state-dependent white noise [14], leading to the eventual christening, the *Fokker-Planck equation*.

Kolmogorov (1931) made significant contributions to the abstract theory of FPE, tying it together with the theory of Markov processes [15]. To honor his contributions, FPE has also been named the Fokker-Planck-Kolmogorov equation (FPKE), Kolmogorov forward equation and Kolmogorov's second equation. Kolmogorov's first equation, also known as the Backward-Kolmogorov equation is the formal adjoint of FPE. In 1933, Kolmogorov extended his work to vector processes and considered uniqueness properties of FPE [16].

Other notable progress was made by Uhlenbeck and Ornstein (1930) [17] and Chandrashekhar (1943) [18]. Barrett (1960) wrote an exposition on the application of FPE to control systems [19], following the works of Stratonovich [20] (electronic systems, 1963) and Chuang and Kazda [21] (nonlinear control systems, 1959). A more in-depth review of research literature related to application of FPE to stochastic control and nonlinear filtering is presented in chapters V and VI respectively. Ariaratnam (1960), Lyon (1960, 1961), Caughey and Dienes (1962), Caughey (1963) and Crandall (1963) were among the first to apply FPE to study random vibrations in nonlinear dynamical systems [22, 23, 24, 25, 26, 27, 28]. Since then, FPE has been employed to study a wide range of problems in several fields of science, engineering and economics; for example, particle physics, chemical mixture analysis, biomechanics, structural mechanics, astrodynamics, nonlinear filtering, stochastic optimal control, optimal stopping, stock-market analysis etc.

Currently, most of the fields mentioned above utilize approximate methods of stochastic analysis because of numerous problems associated with solving FPE. Some

popular approximate methods are Gaussian and higher-order closure techniques [29, 30, 31, 32, 33], equivalent and statistical linearization [34, 35, 36, 37, 38], and Monte Carlo analysis [39, 40, 41]. An illustration of Gaussian closure is shown in Fig.1(b), wherein Gaussian pdfs have been used to approximate (very well) the initial and (poorly) the final point clouds. Note that Gaussian closure is accurate up to the first two moments of the true pdf, while higher order closures also consider additional moments. All closure techniques are philosophically similar because they essentially characterize state uncertainty using a finite number of moments (two in case of Gaussian closure, equivalent to analysis of linearized system dynamics). Statistical linearization replaces the actual nonlinear system model with an equivalent linear system such that the mean square error with respect to parameters of the equivalent linear system is minimized [37]. All linearization/higher order based techniques work for small to moderate time durations of propagation, depending on the degree of nonlinearity.

The Monte Carlo method is simulation based and essentially involves sampling of the underlying probability space to generate a family of test points, which are individually numerically propagated forward through the exact nonlinear system. The pdf at any time step is then approximated by evaluating desired number of moments from the distribution of propagated sample points (see Fig.1). This method generally requires extensive computational resources and effort, and becomes increasingly infeasible for dynamical systems with high-dimensional state space and for long-term simulations. For a particular application under investigation, it is frequently possible to find a suitable approximate method that provides reasonable results if suitable assumptions are made. At the same time, it is well understood that the returns of developing a robust FPE solver for high dimensional nonlinear systems are enormous because it captures the true description of the uncertainty propagation problem.

Consequently, research efforts in this direction have persisted.

The first attempts to solve FPE numerically were made in the late 1950s. Researchers had realized that analytical solutions of the stationary and transient problems were possible for only a handful of dynamical systems which account for a very small percentage of the range of systems encountered in science and engineering. One of the earliest numerical attempts was made by Rosenbluth et al. in 1958 [42], who, while working on interaction of gas molecules under an inverse square gravitation field, expanded the related distribution function in terms of Legendre polynomials, resulting in an infinite set of integro-differential equations which they were claimed was solvable on a "computing machine." In a 1968 paper, Bhandari and Sherrer [43] used a Galerkin projection based method to determine the probability density function for one and two degree-of-freedom systems using global Hermite polynomial expansions. In 1973, Mayfield [44] presented a recursive sequence solution of FPE based on the parametrix approach for partial differential equations. Reif and Barakat (1973) [45] presented a Chebyshev polynomial expansion of FPE. Atkinson (1973) [46], Johnson and Scott (1979, 1980) [47, 48] and Risken et al. (1980) [49] looked at the eigenfunction expansion of FPE for first and second order nonlinear systems.

The earliest use of finite-difference methods date back to Killeen and Futch (1968) [50] and Whitney (1970) [51], involving the study of plasma. The finite element method (FEM) was used for the first time by Langley in 1985 [52, 53]. Since then, FEM in its various forms has become the most popular method for solving FPE. At the same time, several challenges have been identified, the most important of which are outlined in section C of chapter II. At the heart of finite difference and finite element methods lies discretization of a finite domain of solution, which harbors the most formidable numerical challenge, namely the curse of dimensionality. It is widely understood as an exponential increase in size of the discretized problem with increase

in dimensionality of the system, thus eventually placing it beyond the capability of machine computation. Due to this crippling obstacle, all results obtained using discretization techniques have been invariably restricted to dynamical systems with one {Palleschi et al. [54], Vanaja [55], Epperlein [56], Günel and Savacı [57] } and two { Volosov and Pekker [58], Pekker and Khudik [59], Wedig [60], Mirin [61], Langanten [62], Palleschi and de Rosa [63], Spencer and Bergman [64], Shiau and Wu [65], Muscolino et al. [66], Johnson et al.(1997) [67], Pardlwarter and Vasta [68], Zhang et al. [69], Zorzano et al. [70], McWilliam et al. [71], Wei [72], Fok et al. [73], Paola and Sofi [74], Masud and Bergman [75], Kumar et al. [76], Lambert et al. [77], Ujevic and Letelier [78], Attar and Vedula [79] } dimensional state-spaces. In many of these publications, theoretical extensions were presented for higher dimensional cases without actual examples due to infeasibility of numerical implementation. In other works, higher dimensional systems were first reduced to lower order models before using discretization techniques, thus restricting their applicability to a particular class of systems. For example, Wagner and Wedig [80] considered a 6-state example using global orthogonal functions (extended Laguerre polynomials). However, they reduced the 6-state problem to a 4-state problem via stochastic averaging and exploiting system properties before application of the global approximation. Such reduction is not always possible and global approximation is usually inadequate for general nonlinear systems. Soize (1988) [81] used global Hermite polynomials for a special class of systems represented by canonical state variables. A 12-state example was studied, although the actual polynomial expansion was applied to six decoupled 2-state systems, once again emphasizing the difficulty associated with problems in higher dimensions.

With growing computational resources, FPE for 3- and 4-state systems has been numerically solved using supercomputers recently. Notable work was done by

Bergman, Johnson, Wojtkiewicz and Spencer between 1992 and 2000, during which they published a series of papers analyzing the difficulties associated with the extension of discretization techniques (FD/FEM) to high dimensional spaces[82, 64, 67, 83, 84, 85]. In 1995, they used the finite element method on a supercomputer to obtain stationary distribution for three-state dynamical systems [82]. Wojtkiewicz and Bergman used the Cray Y-MP/464 supercomputer to solve a discretized problem of size 2.56 million to obtain the transient pdf for a four-state linear system [85]. Despite these results made possible by powerful computers, the curse of dimensionality remained a stumbling block because the problem size in FEM grows exponentially with dimensionality, assuming orders of 10^6 for the 4-state problem. Moreover, the bookkeeping involved in FEM for maintaining inter-element boundary information becomes increasingly cumbersome in high dimensions. In 2005, Masud and Bergman presented a multi-scale finite element method in which the final approximation was composed of a coarse level and a fine level approximation, in an attempt to curtail the growth of problem size. However, results were presented only for 2-state systems [75]. In 2006, the first attempt was made by Kumar et al. to use the meshless finite element approach for FPE [86]. The meshless paradigm is naturally suited to Fokker-Planck equation because of its numerous advantages over standard FEM, especially in the context of solving partial differential equations in higher dimensions. Two crucial advantages of meshless techniques over standard FEM are listed below:

1. ***Minimal bookkeeping:*** The meshless approach is essentially a "node-based" approximation, rather than an "element-based" approximation like FEM. Nodes used in the solution domain act as centers of local approximations, which are blended together to obtain the global approximation. On the other hand, an FEM approximation is constructed on the basis of shape of elements formed by

inter-connection of nodes. The information of inter-element boundaries is thus crucial, which becomes increasingly difficult to maintain in higher dimensions.

2. ***Ease of local refinement:*** In meshless techniques, the burden of enforcing conformity of the approximation space over the solution domain is assumed by special weight functions, which also define the local region of influence of various nodes. It is therefore possible to assign individual nodes independently selected local approximation spaces, making it easy to use higher order polynomial/non-polynomial basis functions in desired regions of the solution domain. This is known as local p-refinement.

This dissertation develops a robust numerical solver for FPE based on the partition of unity paradigm of meshless techniques. The key advantages mentioned above are exploited, thus furthering research in this field in the logical direction. The developed methods are shown to reduce problem size for relatively high dimensional systems, providing significant improvement over state-of-the-art. Coupled with spectral analysis, a semianalytical algorithm is developed, making possible near real-time solution of FPE transient response. Recursive norm-modification techniques are discussed to obtain further solution refinement while maintaining constant problem size, while also tracking the optimal sized domain for numerical solution. It is expected that the methods developed in this dissertation will permeate numerous applications in stochastic dynamics, several of which are described below.

B. Application Areas

Fokker-Planck equation lies at the core of several key problems in stochastic mechanics. Below, we enlist some important application areas of FPE.

- *Weak solution of stochastic differential equations (SDEs):* FPE governs the

time evolution of the pdf of the entire ensemble of weak solutions of SDEs. The mentioned pdf can be used for system analysis, e.g. stationary behavior, system reliability, first passage etc.

- *Uncertainty propagation:* FPE captures the exact description of the problem of nonlinear uncertainty propagation (Fig.1). Examples of particular applications include: prediction of probability of collision in space (spacecraft/spacecraft, asteroid/planet), study of nonlinear random vibrations in structural mechanics, weather prediction models and plume tracking following explosions/eruptions.

- *Nonlinear filtering:* A fast FPE algorithm would find immediate application in nonlinear filtering, especially in applications involving sparse measurements/and or highly nonlinear dynamics. The prediction step of the Bayes filter requires the solution of FPE for which closure schemes are currently employed, e.g. Gauss closure is the Kalman filter.

- *Stochastic optimal control:* A policy iteration algorithm can be set up using the adjoint of FPE, namely the backward Kolmogorov equation (BKE) to recursively solve the Hamilton-Jacobi Bellman (HJB) equation, thus converting the problem of solving a nonlinear PDE to a sequence of linear PDEs. It can also be used effectively for control law design in hybrid stochastic systems, e.g. morphing structures with multiple performance regimes.

- *Particle physics and quantum optics:* FPE finds widespread use in determining the stationary distribution of elementary particles under force fields. It is also a popular tool for handling noise in quantum optics, e.g. in the study of electric field of a laser.

Besides the above mentioned areas, several important problems in economics and finance have been formulated around Fokker-Planck equation, e.g. optimal stopping and modeling of the stock-market. It has also been used in biochemistry and neurosciences to study the behavior of the nervous system. An efficient algorithm for solving FPE is thus of great interest to the scientific and engineering community. This dissertation primarily deals with the problems of uncertainty propagation, nonlinear filtering and stochastic optimal control.

CHAPTER II

PROBLEM STATEMENT

A. Introduction

In this chapter, formal statement of problems considered in this dissertation is presented. As all these problems concern stochastic dynamic systems, a review is warranted, starting from basic concepts and leading up to the elements of stochastic differential equations. In order to maintain continuity, this review is presented in Appendix A and the reader is strongly encouraged to go through the discussed concepts. Appearing below are three important problems in stochastic dynamics considered in this work.

B. Problem Statement

This dissertation addresses the following three problems in stochastic dynamics:

Problem II.1 ***Meshless variational solution of Fokker-Planck equation***

Consider a continuous dynamical system modeled by the following stochastic differential equation (the state in all subsequent equations, which is a random variable, is written as $\mathbf{x}$):

$$d\mathbf{x} = \mathbf{f}(t,\mathbf{x})dt + \mathbf{g}(t,\mathbf{x})d\mathbf{B}(t), \quad \mathrm{E}[\mathbf{x}(0)] = \bar{\mathbf{x}}_0 \tag{2.1}$$

where, $\mathbf{B}$ represents a M-dimensional zero mean Brownian motion process with correlation function $\mathbf{Q}\delta(t_1 - t_2)$, and $\bar{\mathbf{x}}_0$ represents the mean initial state. Vector functions $\mathbf{f}(t,\mathbf{x}) : [0,\infty) \times \Re^N \mapsto \Re^N$ and $\mathbf{g}(t,\mathbf{x}) : [0,\infty) \times \Re^N \mapsto \Re^{N\times M}$ are measurable functions. The initial probability density of the state is assumed known and designated as $\mathcal{W}(0,\mathbf{x}) = \mathcal{W}_0(\mathbf{x})$, which captures the state uncertainty at time $t = 0$. For the

system in Eq.2.1, the following linear, parabolic partial differential equation describes the time evolution of the state-pdf:

$$\frac{\partial}{\partial t}\mathcal{W}(t,\mathbf{x}) = \mathcal{L}_{\mathcal{FP}}\mathcal{W}(t,\mathbf{x}) \tag{2.2}$$

on the domain $(t,\mathbf{x}) \in [0,\infty) \times \Re^N$, with the boundary condition

$$\mathcal{W}(t,\infty) = 0, \quad t \geq 0 \tag{2.3}$$

where,

$$\mathcal{L}_{\mathcal{FP}} = \left[-\sum_{i=1}^{\mathrm{N}} \frac{\partial}{\partial x_i}\mathrm{D}_i^{(1)}(\cdot,\cdot) + \sum_{i=1}^{\mathrm{N}}\sum_{j=1}^{\mathrm{N}} \frac{\partial^2}{\partial x_i \partial x_j}\mathrm{D}_{ij}^{(2)}(\cdot,\cdot)\right] \tag{2.4}$$

$$\mathrm{D}^{(1)}(t,\mathbf{x}) = \mathbf{f}(t,\mathbf{x}) + \frac{1}{2}\frac{\partial \mathbf{g}(t,\mathbf{x})}{\partial \mathbf{x}} Q \mathbf{g}(t,\mathbf{x}) \tag{2.5}$$

$$\mathrm{D}^{(2)}(t,\mathbf{x}) = \frac{1}{2}\mathbf{g}(t,\mathbf{x}) Q \mathbf{g}^{\mathrm{T}}(t,\mathbf{x}) \tag{2.6}$$

where, $\mathcal{L}_{\mathcal{FP}}$ is the Fokker-Planck operator, $\mathrm{D}^{(1)}$ is known as the drift coefficient vector and $\mathrm{D}^{(2)}$ is the diffusion coefficient matrix, both understood in the Stratonovich sense. In addition to satisfying Eq.2.2 and boundary conditions 2.3, the obtained solution $\mathcal{W}(t,\mathbf{x})$ must fulfill the following admissibility conditions for a valid pdf:

1. Positivity of the *pdf*: $\mathcal{W}(t,\mathbf{x}) \geq 0, \ \forall t \geq 0 \ \ \& \ \mathbf{x} \in \Re^N$.
2. Normalization constraint of the *pdf*: $\int\limits_{-\infty}^{\infty} \mathcal{W}(t,\mathbf{x}) dV = 1, \forall t \geq 0$.

Then, the first problem considered in this dissertation is to determine a finite dimensional meshless approximation $\widehat{\mathcal{W}}(t,\mathbf{x})$ of $\mathcal{W}(t,\mathbf{x})$ given by the following expansion on a finite sized domain $\Omega \triangleq \otimes_{i=1}^{N}[a_i,b_i]$:

$$\widehat{\mathcal{W}}(t,\mathbf{x}) = \sum_{i=1}^{\mathcal{D}} a_i(t)\Psi_i(\mathbf{x}) \tag{2.7}$$

where, $\Psi_i(\mathbf{x})$ are trial functions belonging to a conformal approximation space $\mathfrak{U}_{\mathcal{D}}$, which is a subspace of an infinite dimensional Hilbert space on Ω. The size of the finite dimensional approximation, $\mathcal{D}$, is known as the number of degrees of freedom of the approximation. The problem then reduces to determining coefficients $a_i(t)$ in Eq.2.7 such that the following variational form (a bilinear functional) is satisfied:

$$\int_{\Omega} \mathfrak{R}(\widehat{\mathcal{W}}, v) d\mathbf{x} = \alpha \int_{\Gamma} (\widehat{\mathcal{W}} - \mathcal{W}_{\Gamma}) v d\mathbf{x}, \quad \forall v \in \mathfrak{V}_{\mathcal{D}} \tag{2.8}$$

where, v is a test function belonging to another subspace $\mathfrak{V}_{\mathcal{D}}$ of a Hilbert space in Ω. The bilinear functional $\mathfrak{R}(\cdot, \cdot) : \mathfrak{U}_{\mathcal{D}} \times \mathfrak{V}_{\mathcal{D}} \mapsto \Re$ is essentially the projection of residual error resulting from substitution of approximation 2.7 into Eq.2.2 onto the test space $\mathfrak{V}_{\mathcal{D}}$. The boundary conditions of Eq.2.3 are enforced in soft form using a penalty parameter α.

Comments

The drift vector ($\mathrm{D}^{(1)}$) captures drifting apart of the mean of propagated pdf from the propagated mean of initial pdf. Generally, it increases with degree of nonlinearity of underlying dynamics, i.e. $\mathbf{f}(t, \mathbf{x})$. The diffusion matrix ($\mathrm{D}^{(2)}$) captures spreading out of the substantial portion of the pdf (e.g. the 3σ region) over state-space. In simple terms, it governs how flat (or diffuse) the pdf turns out to be. In case the underlying governing dynamics (Eq.2.1) is deterministic, i.e. $\mathbf{g}(t, \mathbf{x}) = \mathbf{0}$ and the source of uncertainty lies only in initial state, the diffusion matrix is identically zero and the reduced FPE is called the Liouville equation. An example of such a problem is uncertainty propagation through $2/N$-body equations of motion in celestial mechanics, wherein the dynamics are very well understood.

It is important to point out that Eq.2.5 represents the Stratonovich form of the drift vector. There exists another form known as the Itô form, which is generally different from the Stratonovich form and is considered by mathematicians to be the rigorously correct expression for $\mathrm{D}^{(1)}$. In engineering fields however, the Stratonovich form in preferred since it precludes the need for Itô calculus which is required to deal with the Itô form. The two forms are identical in case of state additive noise, i.e. when $\mathbf{g}(t, \mathbf{x}) = \mathbf{g}(t)$. This is typically the case with most real life stochastic systems.

Problem II.2 $\mathcal{H}_2$ ***Optimal Control Problem***

Consider the following specialized form (autonomous $\mathbf{f}$ and $\mathbf{g}$) of the governing dynamics of Eq.2.1 with a control influence term:

$$d\mathbf{x} = \mathbf{f}(\mathbf{x})dt + \mathbf{h}(\mathbf{x})\mathbf{u}dt + \mathbf{g}(\mathbf{x})d\mathbf{B}(t); \qquad \mathrm{E}[\mathbf{x}(0)] = \bar{\mathbf{x}}_0 \tag{2.9}$$

where, $\mathbf{u} \in \Re^m$ is the control input vector and $\mathbf{h}(\mathbf{x}) : \Re^N \rightarrow \Re^{N \times m}$ is the control influence matrix. Adjust the coordinate system such that $\mathbf{f}(\mathbf{0}) = \mathbf{0}$ and let $\varphi(t)$ be a trajectory that solves Eq.2.9. Determine an approximation for the optimal regulator $\mathbf{u}^*(\mathbf{x})$ that minimizes the following functional over infinite horizon:

$$J(\bar{\mathbf{x}}_0) = \mathrm{E}\left[\int_0^\infty [l(\varphi(t)) + \|\mathbf{u}(\varphi(t))\|^2_{\mathbf{R}}]e^{-\beta t}dt\right] \tag{2.10}$$

In the above relation, l is often referred to as the state-penalty function. Also, $\|\mathbf{u}\|^2_{\mathbf{R}}$ is defined as $\mathbf{u}^T\mathbf{R}\mathbf{u}$, where $\mathbf{R}$ is a positive definite matrix. The parameter β is a discount factor that ensures finite cost-to-go in the presence of random disturbances.

Problem II.3 ***Bayes Nonlinear Filtering Problem***

Consider again the dynamics of Eq.2.1 and augment to it the following discrete measurement model:

$$\mathbf{y}_k = \mathbf{h}(\mathbf{x}_k) + \mathbf{v}_k \tag{2.11}$$

where, $\mathbf{y} \in \Re^l$, $\mathbf{v}$ is a l-dimensional Brownian motion process (measurement noise), $\mathbf{h}(\mathbf{x}) : \Re^N \mapsto \Re^l$ is measurable and k denotes the time instant of measurement. The growth of information with successive measurements is denoted by the filtration $\mathcal{Y}^k = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_k\}$. Then, the problem is to approximate the state probability density function conditioned on the filtration $\mathcal{Y}^k$, i.e., $\mathcal{W}(\mathbf{x}_k|\mathcal{Y}^k)$ utilizing the following recursive equations (Bayes Filter):

$$\mathcal{W}(\mathbf{x}_k|\mathcal{Y}^{k-1}) = \int \mathcal{W}(\mathbf{x}_k|\mathbf{x}_{k-1})\mathcal{W}(\mathbf{x}_{k-1}|\mathcal{Y}_{k-1})d\mathbf{x}_{k-1} \tag{2.12}$$

$$\mathcal{W}(\mathbf{x}_k|\mathcal{Y}^k) = \frac{\mathcal{W}(\mathbf{y}_k|\mathbf{x}_k)\mathcal{W}(\mathbf{x}_k|\mathcal{Y}^{k-1})}{\int_\Omega \mathcal{W}(\mathbf{y}_k|\xi)\mathcal{W}(\xi|\mathcal{Y}^{k-1})d\xi} \tag{2.13}$$

Eq.2.12 represents the propagation part between two measurement updates. The left hand side of this equation is the prior state pdf, which can be obtained by integrating the associated Fokker-Planck equation in weak form (refer to problem II.1) between time labels t_{k-1} and t_k. The posterior state pdf, $\mathcal{W}(\mathbf{x}_k|\mathcal{Y}^k)$, can be obtained from the Bayes rule shown in Eq.2.13, in which $\mathcal{W}(\mathbf{y}_k|\mathbf{x}_k)$ represents the likelihood function, and is given by:

$$\mathcal{W}(\mathbf{y}_k|\mathbf{x}_k) = \frac{1}{\sqrt{(2\pi \det R)^m}} \exp\left(-\frac{1}{2}[\mathbf{y}_k - \mathbf{h}(\mathbf{x}_k)]^{\mathrm{T}}\mathbf{R}_k^{-1}[\mathbf{y}_k - \mathbf{h}(\mathbf{x}_k)]\right) \tag{2.14}$$

C. Research Issues

As observed in chapter I, any numerical approach for solving FPE faces numerous difficult issues, three of which are detailed below:

1. ***Solution constraints***: Problem statement II.1 stipulates conditions on the obtained solution in order to qualify as a valid pdf, namely, the positivity and normality constraints (Eqs.1, 2). In this dissertation, the normality constraint is enforced as a postprocessing step of re-normalization. The positivity con-

straint is not actively enforced, just checked to hold true within tolerance (e.g. negative probability mass $\leq 10^{-9}$). Positivity becomes relevant only in the tail regions of the pdf and its enforcement has thus far remained a tough proposition. Several researchers have used a log-transformation of FPE (the inverse exponential transform of the solution obtained ensures positive values) [66, 74]. However, this approach converts the linear PDE (Eq.2.2) into a nonlinear PDE, which is generally not desirable.

2. ***Domain of solution***: A numerical method based on discretization of state-space requires a finite sized domain of solution (Ω in problem II.1). Determining such a domain of appropriate size, shape and orientation is a challenging task because the theoretical domain of an N-state pdf is $\Re^N$. Moreover, the boundary conditions in problem II.1 are enforced at infinity, meaning that artificial boundary conditions need to be enforced on a "large enough solution domain," e.g. $\mathcal{W}(t, \Gamma) = 10^{-9}$, where Γ is the boundary of the chosen domain. In most published literature, heuristic methods are used to determine a finite domain for constructing the approximation. In this dissertation, a recursive norm-modification approach for error projection is presented to determine an appropriate solution domain for nonlinear systems. This approach also forms the basis for improving approximation accuracy while keeping the size of the discretized problem ($\mathcal{D}$ in Eq.2.7) small. This leads us to the final research issue discussed below.

3. ***Curse of dimensionality***: In the context of FPE, curse of dimensionality is widely understood as an exponential increase in size of the discretized problem ($\mathcal{D}$) with dimensionality of state-space (N). It remains the greatest challenge confronting the successful numerical solution of FPE and the state-of-the-art in

this field, namely the finite element method is known to suffer from this curse. This dissertation presents meshless approximation techniques based on the partition of unity finite element method (PUFEM) to tackle this research issue. Reduction in problem size by several orders of magnitude is demonstrated in comparison with standard FEM (e.g. 3 orders of magnitude reduction for systems in $\Re^4$) and numerical evidence for breaking of the curse of dimensionality is presented. All results presented in this dissertation have been obtained on a small workstation with modest computing resources.

CHAPTER III

SOLUTION METHODOLOGIES

A. Introduction

This chapter discusses methods of solution to address the problems described in chapter II. At the heart of each algorithm lies a robust numerical solver for Fokker-Planck equation and its formal adjoint, the backward Kolmogorov equation (BKE), which is developed using the meshless paradigm in finite elements. This paradigm was developed during the late 1980's and early 1990's through the efforts of numerous researchers striving to achieve the following two key attributes:

1. **Node based approximation:** The central idea behind any meshless technique is to construct an approximation based entirely on nodes distributed over the domain of solution, thus removing dependence on domain geometry. This is different from traditional finite element methods, where a mesh is constructed out of the nodes used for discretization. The nature of approximation (i.e. order of polynomial basis functions used) depends on the shape of elements in the mesh, i.e. how the nodes are connected to each other. For example, if the mesh comprises of triangular elements, i.e. three nodes per element, only first degree polynomials may be used to make sure that they satisfy inter-element continuity. It is therefore required to maintain a record of inter-connectivities among nodes to enforce regularity conditions across the solution domain. Because of this characteristic, traditional FEM is known to be an "element-based" technique and typically suffers from tedious book-keeping of elemental information. This is especially challenging for domains in higher than two-dimensional spaces. On the other hand, the meshless paradigm has built-in mechanisms for enforcing

solution regularity automatically, thus precluding maintenance of information about node interconnections. Each node acts as a center of approximation with a local region of influence, inside which basis functions associated with that node are defined.

2. **Local approximation enrichment:** Polynomial basis functions are typically used in FEM, the order of which depends on the shape of elements in the mesh. Local enrichment of approximation space is usually performed by addition of new nodes followed by re-meshing, a procedure widely known as h-refinement. A key benefit of the meshless node-based paradigm is that it greatly simplifies the use of non-polynomial basis functions. In addition, its underlying framework ensures that local approximations of individual nodes automatically satisfy continuity conditions (conformity), irrespective of the nature of basis functions used. This helps assign independently chosen basis functions (which may be polynomials or special "handbook" functions) to individual nodes. Additional basis functions may be added to selected nodes without significant overhead, thus facilitating "local p-refinement" in addition to h-refinement, making these techniques "multi-resultion" in addition to "meshless".

In brief, the meshless paradigm provides a flexible framework for solving PDEs in high dimensional spaces. Numerous versions have been developed, e.g. the element-free Galerkin method (EFGM) of Belytschko [87], which was one of the earliest meshless techniques, smooth particle hydrodynamics (SPH) of Monaghan [88, 89], hp-clouds of Duarte and Oden [90], reproducing kernel particle method (RKPM) of Liu et al. [91], meshless local Petrov-Galerkin methods (MLPG) of Atluri [92], extended FEM and generalized FEM (XFEM, GFEM) of Strouboulis and Babuška [93], partition of unity finite element method (PUFEM) of Babuška and Melenk [94], and

particle-PUFEM (pPUFEM) of Griebel and Schweitzer [95, 96, 97]. Related partition of unity (PU) methodology for piecewise continuous least squares approximation in N-dimensions were first developed by Junkins, Jancaitis and Miller in the 1970's [98, 99, 100]. In this dissertation, the PUFEM technique of Babuška is referred to as standard-PUFEM (sPUFEM) to distinguish it from the particle version of Griebel and Schweitzer. Variants of both these techniques are developed herein to solve FPE for high dimensional nonlinear stochastic systems.

The remainder of this chapter is arranged as follows: Section B describes various aspects of the PUFEM framework; including domain discretization, cover generation, construction of conformal approximation spaces, details of partition of unity weight functions, variational form equations and their integration using quasi Monte-Carlo techniques. Details for both the standard- and particle- versions of PUFEM are discussed. Results for stationary FPE are presented in section C and the issue of curse of dimensionality is discussed in elaborate detail. Numerical evidence for breaking of the curse of dimensionality is presented, although a rigorous theoretical statement still cannot be made. Section D couples the meshless discretization of FP operator with spectral analysis and spurious mode rejection leading to a semianalytic algorithm for near real-time solution of the transient FPE. It is shown that use of admissible eigenfunctions of the discretized FP operator as basis functions causes equation error to be bounded by an exponentially decaying envelope having greatest width at the initial time. Results for transient FPE are presented in section E.

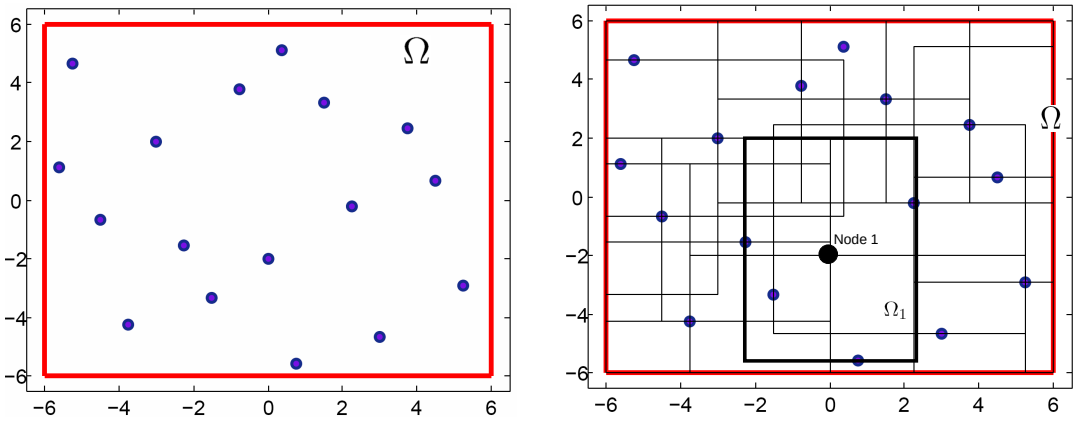
B. PUFEM Methodology

1. Domain Discretization

The partition of unity (PU) approach to meshless finite elements was first developed by Babuška. The technique was further developed and generalized by several researchers, most notably by Griebel and Schweitzer, who developed the particle version of PUFEM (pPUFEM). Since we are interested in solving FPE in this dissertation, we will consider only N-hypercuboids as domains of solution, which simplifies the process of discretization. This process differs greatly between s- and p- versions, but we will start by talking about domain discretization in PUFEM in general. To this end, consider a domain Ω and a set of overlapping subdomains Ω_i, $i = 1, 2, \ldots, P$, which form a cover for Ω. Each subdomain belongs to a node and acts as its region of influence inside which its local approximation will be defined (see Fig.3). The actual process of generating the cover will be discussed later. A "partition of unity" on Ω is a mathematical paradigm in which each of the overlapping subdomains Ω_i is associated with a compactly supported function $\varphi_i(\mathbf{x})$ called the PU pasting function (also, PU weight function), which is strictly zero outside Ω_i and has the property that:

$$\sum_{i=1}^{P} \varphi_i(\mathbf{x}) = 1, \quad \forall \mathbf{x} \in \Omega \tag{3.1}$$

The above paradigm represents the skeleton for a powerful meshless finite element method for solving PDEs on Ω. By assigning the subdomains Ω_i to individual nodes distributed over the global domain Ω, an implicit "discretization" is obtained (Fig. 3 shows an example in 2 dimensions), using which a local variational form of the PDE to be solved can be formulated. The discretization is not to be understood in the usual sense of traditional mesh-based FEM because of two primary reasons: (a) overlap



(a) Nodes distributed over the solution domain act as "centers of approximation".

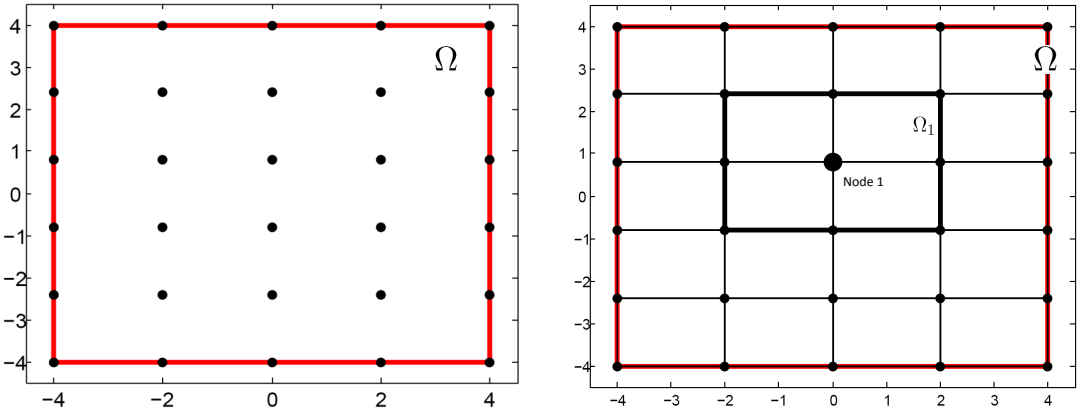
(b) Overlapping regions of influence of various nodes, Ω_i, form a cover for Ω. Subdomain for node indexed "1" is highlighted.

Fig. 3. A node-based meshless framework for solving PDEs.

among neighboring subdomains and (b) minimal role of inter-node or inter-element connectivity. By virtue of the latter property, this discretization is "meshless" and has immense advantage in application to high dimensional PDEs. The PU weights $\varphi_i(\mathbf{x})$ perform the important task of blending together local approximations smoothly in an unbiased manner, by virtue of the PU property (Eq.3.1). These functions are discussed in the next section. At this point, details of cover generation are considered for s- and p- versions of PUFEM.

a. Cover Generation in sPUFEM

The standard version of PUFEM as originally developed by Babuška works around a highly structured distribution of nodes - it requires them to be placed as if they lie on a rectangular grid overlaid on Ω, as shown in Fig.4. The meshless nature is intact because it is still not required to maintain information about node interconnections. However, the nature of PU weights used in sPUFEM require nodes



(a) Grid-like distribution of nodes in sPUFEM.

(b) Cover Generation in sPUFEM.

Fig. 4. Structured framework of standard-PUFEM.

to be placed on a grid whose outermost perimeters lie on the solution domain boundaries. As a result, if P_i nodes are used along each of the N dimensions of Ω, the total nodes appearing in the discretization are $\Pi_{i=1}^{N} P_i$, out of which $\Pi_{i=1}^{N}(P_i - 2)$ lie completely in the interior of Ω and $[\Pi_{i=1}^{N} P_i - \Pi_{i=1}^{N}(P_i - 2)]$ lie on its boundary, Γ. For such a structured assembly, cover generation is trivial because the subdomain of any given node can be a hypercuboid whose vertices house its 2^N neighboring nodes (see Fig.4(b)). This ensures that all subdomains put together form a cover for Ω. Of course, this approach suffers from the curse of dimensionality in h-refinement because with every added node, the size of discretization grows exponentially in N.

b. Cover Generation in pPUFEM

The particle version of PUFEM developed by Griebel and Schweitzer is a much more generalized form of PUFEM and allows nodes to be placed in a completely unstructured fashion on Ω. Cover generation is therefore considerably more complicated. Since no grid is involved, pPUFEM is free from the curse of dimensionality

in h-refinement. In Ref.[96], Griebel and Schweitzer developed a tree-based algorithm for assigning hypecuboid shaped subdomains to nodes that form a cover for Ω. However, if only the original set of nodes are used, this approach leaves voids in the domain which are filled by addition of further nodes. The number of additional nodes required grows exponentially in N ($\sim 2^N$), thus making this approach susceptible to the curse of dimensionality. This section provides an alternate method for generating hypercuboid subdomains whose union forms a cover for Ω, without introducing an exponential number of additional nodes in the domain:

Algorithm III.1 ***pPUFEM Cover Generation***

- Given: Solution domain, $\Omega = \otimes_k [a_k, b_k]$, $k = 1, \ldots, N$; and a set of nodes $\Xi = \{\mathbf{p}_i \in \Re^N \mid \mathbf{p}_i \in \Omega, i = 1, 2, \ldots, P\}$.

- For each node $\mathbf{p}_i \in \Xi$:

  1. Find the node $\mathbf{p}_m \in \Xi$ with least distance from $\mathbf{p}_i$. Let the displacement vector between $\mathbf{p}_i$ and $\mathbf{p}_m$ be $\mathbf{v}^\star_{im}$.

  2. Construct an initial estimate of subdomain of $\mathbf{p}_i$ as: $\Omega_i^0 = \otimes_k \, [{}^k\mathbf{p}_i - {}^k\mathbf{v}^\star_{im}, \; {}^k\mathbf{p}_i + {}^k\mathbf{v}^\star_{im}]$, where $k = 1, \ldots, N$, and the superscript k on the vectors denote their k^{th} component in $\Re^N$.

  3. Define *search index* for $\mathbf{p}_i$ as $\text{si} = \{1, 2, \ldots, P\} \setminus \{i, m\}$. Make two copies: $\text{si}^+ = \text{si}$, $\text{si}^- = \text{si}$.

  4. **Do** until $\text{si}^- = \phi$ or breaking condition is reached. Repeat this loop for each $k \in \{1, 2, \ldots, N\}$

     (a) Define ${}^k mx = \max_j \, [\{{}^k\mathbf{v}^-_{ij}, \; a_k\}]$, where ${}^k\mathbf{v}^-_{ij} \in \{{}^k\mathbf{v}_{ij} \mid {}^k\mathbf{v}_{ij} < 0\}$ and $j \in \text{si}^-$. Also, let $\arg\max\{j\} = l$.

(b) **if** ${}^k mx = \max_r[{}^r\mathbf{v}_{il}]$ or ${}^k mx = a_k$, set ${}^k\Omega_i^- = {}^k mx$ and **break**.

(c) **else** $\text{si}^- = \{1, 2, \ldots, P\} \setminus \{i, m, l\}$

5. **End Do**

6. (This is a similar step as 4-5, except for "right" nodes)

   Do until $\text{si}^+ = \phi$ or breaking condition is reached. Repeat this loop for each $k \in \{1, 2, \ldots, N\}$

   (a) Define ${}^k mn = \min_j\ [\{{}^k\mathbf{v}_{ij}^+,\ b_k\}]$, where ${}^k\mathbf{v}_{ij}^+ \in \{{}^k\mathbf{v}_{ij} \mid {}^k\mathbf{v}_{ij} > 0\}$ and $j \in \text{si}^+$. Also, let $\arg\min\{j\} = l$.

   (b) **if** ${}^k mn = \max_r[{}^r\mathbf{v}_{il}]$ or ${}^k mn = b_k$, set ${}^k\Omega_i^+ = {}^k mn$ and **break**.

   (c) **else** $\text{si}^+ = \{1, 2, \ldots, P\} \setminus \{i, m, l\}$

7. **End Do**

8. $\Omega_i = \otimes_k[{}^k\Omega_i^-,\ {}^k\Omega_i^+]$

A flowchart for the above algorithm is provided in Fig.5. This algorithm generates subdomains Ω_i for any given set of nodes placed in an N dimensional domain Ω. It starts by obtaining an initial estimate of Ω_i using its nearest neighbor (step 2). Note that the initial estimate, Ω_i^0, is not necessarily contained in Ω. The algorithm then expands (or contracts) the initial estimate both on the "left" and "right" along each dimension, until it meets another node, or hits the global boundary, Ω (steps 4a, 6a). For lack of appropriate terminology, "left" and "right" of a node are to be understood in the same sense as for a single dimension, e.g. if node $\mathbf{p}_i$ is on the "left" of node $\mathbf{p}_j$ in the k^{th} dimension, it implies that ${}^k\mathbf{v}_{ij} < 0$, where, $\mathbf{v}_{ij} \doteq \mathbf{p}_j - \mathbf{p}_i$. The above algorithm permits a neighboring node to block subdomain growth only if the component of the displacement vector has greatest absolute value in the direction of blockage (steps 4b, 6b). For example, the expansion of the subdomain of i^{th} node is blocked in

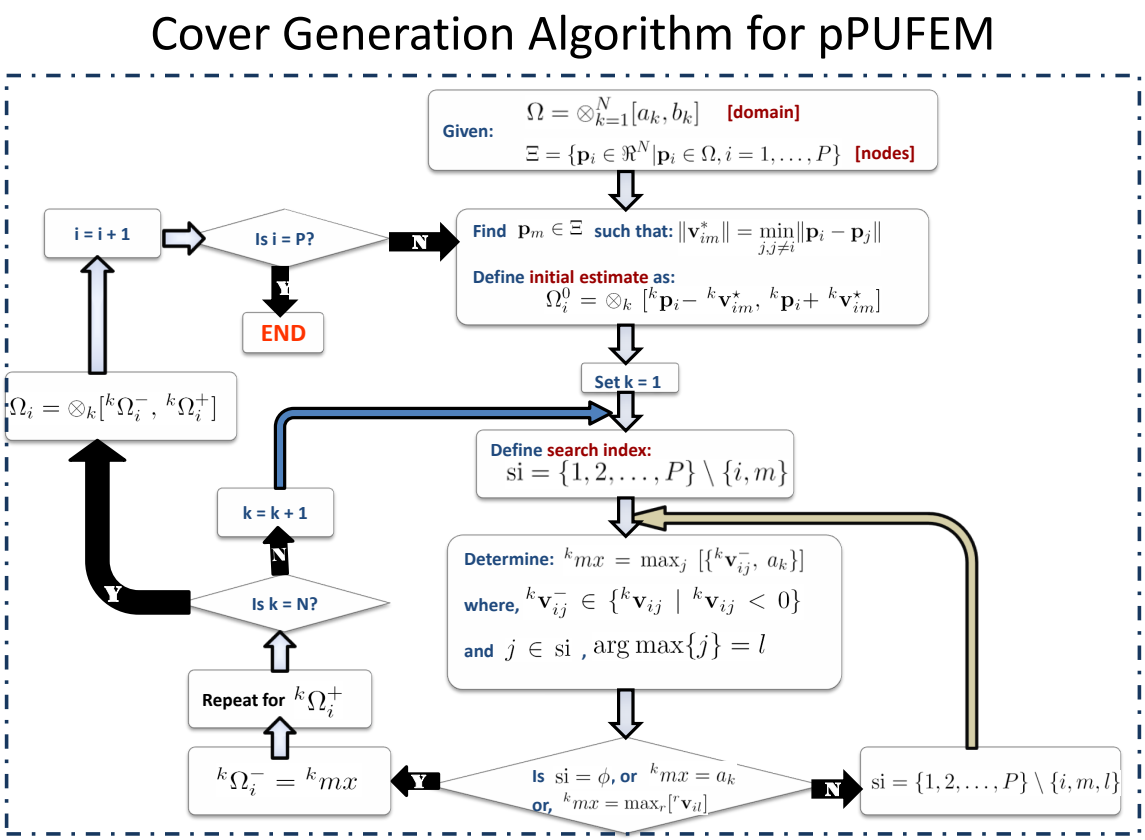


Fig. 5. Flowchart for the pPUFEM cover generation algorithm.

the k^{th} direction by the j^{th} node only if the the k^{th} component of the displacement vector $\mathbf{v}_{ij}$ has the largest absolute magnitude among all components of $\mathbf{v}_{ij}$. This is a very important step because it ensures that a node very far in a particular direction does not block expansion/contraction of subdomains in that direction. Therefore, if a particular node cannot extend all the way out to the domain, there is at least one other node present that would, hence increasing the chance of forming a cover. Thus expansion/contraction is continued until a valid blocking node is encountered or the global solution boundary is reached. An example of cover generation with this algorithm is shown in Fig.3(b) and the subdomain of the highlighted node is shown with solid lines. We note however that this algorithm cannot be proven to guarantee cover generation for a general N-dimensional solution domain. For certain node distributions in high dimensional domains (e.g. 4D, 5D), the algorithms was found to leave a very small fraction of the domain uncovered, never exceeding more than 0.1% of the total volume, near the boundary regions. These minor voids are easy to fill with a very small number of additional nodes.

2. Construction of Conformal Approximation Space

Once the node distribution and cover generation are established for Ω, it is possible to define local approximation spaces within individual subdomains. As mentioned previously, each node acts as a center of approximation, carrying local shape functions with compact support over its region of influence, Ω_i. In this section, we consider the problem of building a globally conformal approximation space out of local approximations. By conformity, we mean continuity (to desired order) of shape functions across boundaries of subdomains. To begin, consider the local approximation associated

with the i^{th} node, denoted by $\widehat{\mathcal{W}}_i(t, \mathbf{x})$:

$$\widehat{\mathcal{W}}_i(t, \mathbf{x}) = \sum_{j=1}^{Q_i} a_{ij}(t)\Psi_{ij}(\mathbf{x}), \quad \text{supp}(\Psi_{ij}) = \Omega_i; \ \ \forall j = 1, \ldots, Q_i \tag{3.2}$$

In the above equation, a total of Q_i shape functions have been used for the i^{th} node. It is important to mention again that the number (Q_i) and nature of shape functions can be different for different nodes by virtue of the flexibility of the PUFEM framework. The meshless nature of PUFEM guarantees that shape functions $\Psi_{ij}(\cdot)$ are conformal *by construction*, meaning that no additional steps are required to enforce continuity. Different meshless methods use different techniques for this purpose, most popular being the moving least squares (MLS) approach. PUFEM uses perhaps the simplest way of constructing conformal shape functions. The starting ingredient is a set of "basis functions" for a particular node, which when simply multiplied with the PU weight associated with that node, forms conformal shape functions. This is very easy to prove by considering derivatives of shape function on the inter-element boundaries. Note that this work differentiates between "basis functions" and "shape functions", the latter being the final form used in the approximation space and constructed "out of" basis functions in the following manner:

$$\Psi_{ij}(\mathbf{x}) = \varphi_i(\mathbf{x})\psi_{ij}(\mathbf{x}), \quad j = 1, \ldots, Q_i \tag{3.3}$$

By themselves, basis functions do not satisfy conformity and in essence, PUFEM delegates the burden of enforcing inter-element continuity to PU pasting functions so that the user is free to select basis functions purely on the criteria of local approximability (i.e. what functions best approximate local system behavior). The order of continuity of shape functions across local subdomains is inherited from the continuity properties of PU weights [94, 98, 99, 101]. PU weights also bring about an

unbiased average of local approximations (Eq.3.2) over regions of overlap by virtue of the PU property. Again, this is automatic because the weights are built into the shape functions. Typically, "tent-functions" are used as PU weights and they provide C^0 continuity. In this dissertation, higher order polynomial functions {Global/Local Orthogonal Mapping (GLO-MAP) weights} are used that provide any desired order of continuity across local boundaries and are considered in detail in the next section.

In comparison, basis functions are the same as shape functions in traditional FEM. As a result, they need to form a conformal space on their own which limits their range of selection. In most other meshless methods like SPH, MLPG etc, shape functions are constructed using data fitting algorithms like the moving least squares (MLS), using a pool of basis functions which may be non-polynomials. This is generally an elaborate process and makes independent selection of basis functions for individual nodes difficult.

Figure 6 illustrates the process of shape function construction in the PUFEM algorithm. In these figures, basis functions (ψ_{ij}) have been drawn using bold lines and the PU pasting functions (φ_i) using light lines. Also, all functions corresponding to odd-numbered nodes are drawn with solid lines and those corresponding to even-numbered nodes with dashed lines. The 1-D domain $[-1, 1]$ is discretized using 5 subdomains with tent-functions in Fig.6(a) and C^1 GLO-MAP weights in Fig.6(b). The use of quadratic polynomials as basis functions has been shown in all subdomains. Additionally, a sinusoidal function (which enriches the existing polynomial basis) has been introduced locally only in the third subdomain (corresponding to the highlighted node # 3). Clearly, these basis functions do not form a conformal space on their own. However, upon multiplication with PU pasting functions of the corresponding nodes, the resulting shape functions satisfy inter-element continuity as is clearly visible in Figs.6(c) and 6(d). Note the inheritance of continuity from

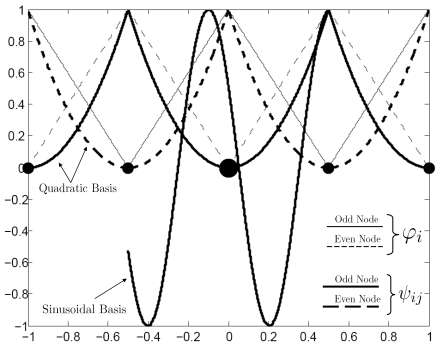
PU weights in Figs.6(e) and 6(f) as the latter figure illustrates continuous derivatives because the corresponding weight functions are C^1 continuous. The numerical solver developed in this dissertation has the capability to automatically generate a complete space of polynomial basis functions of p^{th} order on N dimensional domains. This feature was highlighted in Fig.2 in the introduction.

a. Partition of Unity Weights

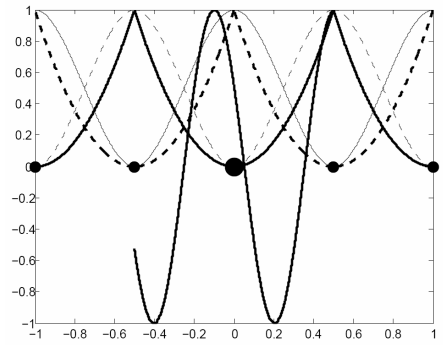
Clearly, PU weights are an integral part of the PUFEM approach. As with cover generation, their construction differs vastly between the s- and p- versions of PUFEM. However in both versions, their functionality is the same: they bring about an implicit domain discretization, merge together various local approximations by performing an unbiased average and determine their order of continuity across local boundaries [94]. Because of the requirement of PU constraint (Eq.3.1), it is generally a difficult task to construct pasting functions that enforce continuity of any desired order. Here, we consider weight construction for the s- and p- versions separately.

b. PU Weights for sPUFEM

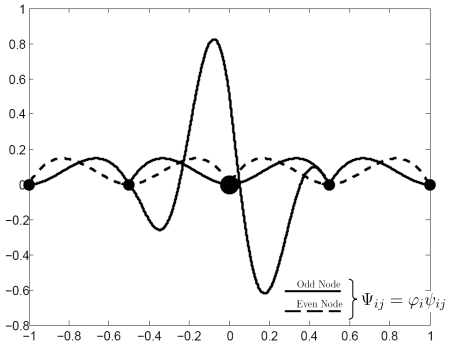
As previously mentioned, sPUFEM uses a structured grid-like distribution of nodes, which helps construct simple polynomial weights that satisfy Eq.3.1. Typically "tent-functions" are used, which provide C^0 continuity. Elsewhere, more sophisticated positive functions have been used after re-normalization to enforce the PU constraint in the following manner: $\varphi_i(.) = \frac{w_i(.)}{\sum\limits_j w_j(.)}$ (Shepard's functions). These functions however are generally difficult to integrate due to their rational-function form. The use of higher order polynomials as PU weights which could be automatically generated for a prescribed order of continuity has not been explored to a great extent in the PUFEM literature thus far. In this work, weight functions developed



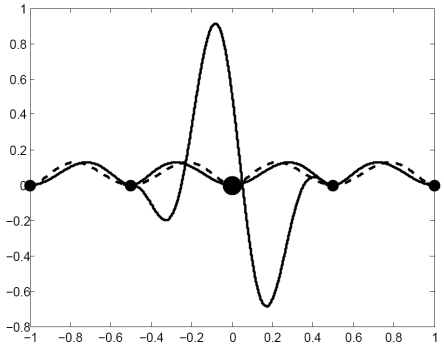
(a) PUFEM basis functions and the C^0 PU function.



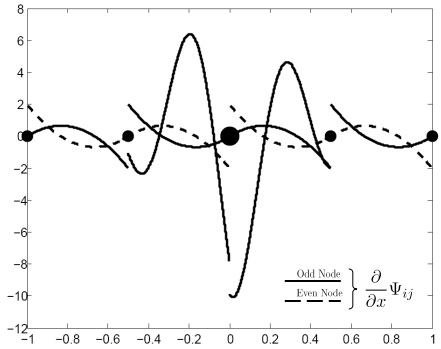
(b) PUFEM basis functions and the C^1 PU function.



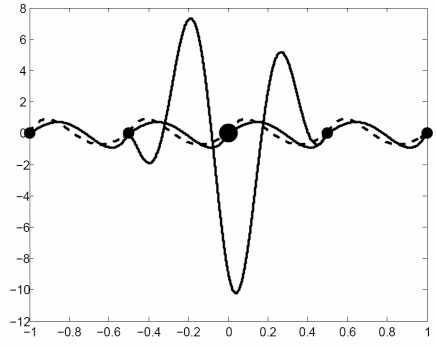
(c) PUFEM shape functions with C^0 pasting.



(d) PUFEM shape functions with C^1 pasting.

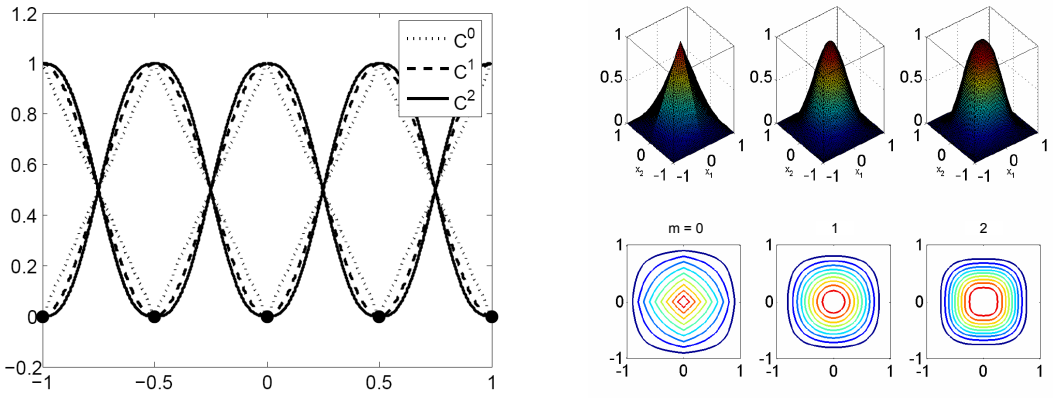


(e) Derivative of the shape functions - C^0 pasting.



(f) Derivative of the shape functions - C^1 pasting.

Fig. 6. 1D shape function construction in PUFEM. x axis: x, y axis: Function evaluated at x.



(a) GLO-MAP weight functions in 1-D, up to C^2 smoothness.

(b) GLO-MAP weight functions in 2-D, up to C^2 smoothness.

Fig. 7. GLO-MAP weights as PU pasting functions in standard-PUFEM.

by Jancaitis et al. [98, 99, 100], known as GLO-MAP weights are used as pasting functions for sPUFEM. These functions are of polynomial form, satisfy Eq.3.1, and have compact support - thus satisfying all requirements. Figs. 7(a) and 7(b) illustrate GLO-MAP weights upto C^2 continuity in 1 and 2 dimensions respectively.

The idea behind GLO-MAP weights is surprisingly simple - given a node belonging to a discretized domain, the polynomial function of lowest degree which assumes the value unity at its parent node and decays to zero at all its neighboring nodes with specified degree of smoothness satisfies the property of partition of unity on the global domain Ω. These conditions can be easily used to determine the coefficients of such a polynomial in one variable, assuming the following general form: (more details can be found in Ref. [101])

$$w_{(m)}(x) = 1 - y^{m+1}\left\{\frac{(-1)^m(2m+1)!}{(m!)^2}\sum_{k=0}^{m}\frac{(-1)^k}{2m-k+1}\binom{m}{k}y^{m-k}\right\}, \quad y = \frac{|x-x_i|}{2h} \tag{3.4}$$

where, m is the desired order of smoothness at the boundaries of Ω_i. In the above

equation, the function $w_m(x)$ is defined in terms of a normalized variable y, which has value unity at its parent node and zero at all neighboring nodes. Tent-functions are a special case with $m = 0$. Functions with higher degree of smoothness ($m = 1$ and $m = 2$) have also been shown, and their benefits illustrated in Fig.6. In the PUFEM framework, these weights come as an invaluable construct because of their several relevant properties [101]:

1. Polynomial form: By virtue of their polynomial form, GLO-MAP weights are easy to integrate. Additionally, if polynomial bases are used, the resulting weak form integrals can be evaluated analytically.

2. They satisfy the PU property. It is very easy to prove the fulfillment of this constraint when the GLO-MAP weights are written in local co-ordinates centered at the corresponding nodes and scaled with the inter-nodal distance along each dimension, $h^{(i)}$. This implies that in local co-ordinates, the central node is at the centroid of a N-hypercube and all its neighboring nodes are at the various 2^N vertices. The value of the GLO-MAP weights are 1 and 0 respectively at these locations.

3. They can provide any desired order of continuity across subdomain boundaries. This is very useful in applications which require the solution derivatives to satisfy error bounds.

4. Easy extension to higher dimensions: It is surprisingly easy to construct GLO-MAP weights in higher dimensions [99]. A simple continued product of 1-D weights written along the various dimensions gives the weight function in the higher dimensional space. E.g., $w_{(2)}(x_1, x_2) = w_{(2)}(x_1)w_{(2)}(x_2)$, i.e. a GLO-MAP weight in 2-D providing C^2 continuity is simply the continued product of

two 1-D weights providing the same level of smoothness.

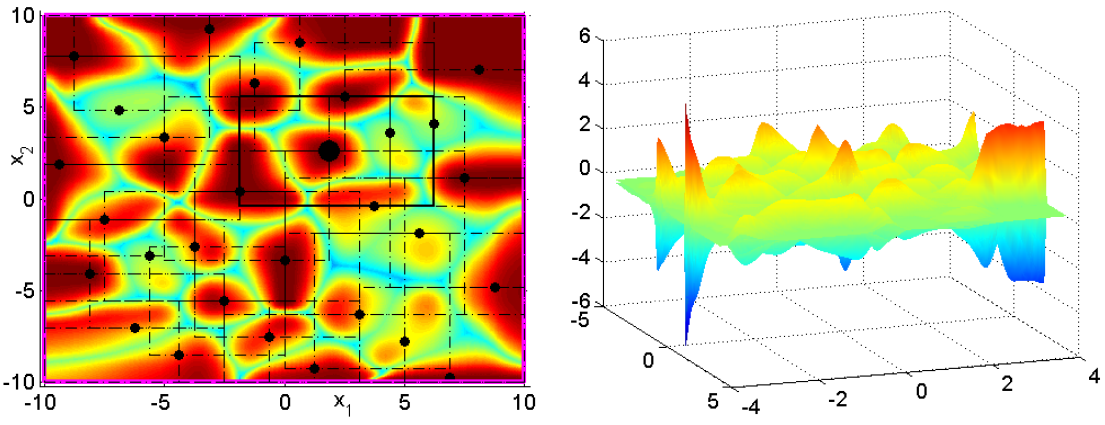
In summary, the generality provided by GLO-MAP weight functions and their easy extension to N-dimensions opens the path for implementation of the sPUFEM algorithm to solve high dimensional PDEs (including FPE). Furthermore, if basis functions orthogonal to these weight functions are used, we obtain an improvement in the condition number of the stiffness matrix, $\mathbf{K}$ (see Eq.3.9). The obvious limitation of these functions is that satisfaction of the PU constraint is contingent upon grid-like node distributions, implying that they can be used directly only in PDEs defined on N-hypercuboids. Domains of all other shapes would require a transformation into a hypercuboid. In the current application (FPE) however this is not a problem, because the domain of solution can be chosen to be an N-hypercube.

c. PU Weights for pPUFEM

Due to the unstructured nature of node distribution, building PU weights that satisfy Eq.3.1 is extremely difficult in pPUFEM. In fact, the only obvious way of building PU weights is by re-normalization as in Shepard's functions. Since subdomains in pPUFEM are hypercuboids, GLO-MAP weights can still be used after re-normalization:

$$\varphi_i(\mathbf{x}) = \frac{w_i(\mathbf{x})}{\sum_{j=1}^{P} w_j(\mathbf{x})} \tag{3.5}$$

where, $w_j(\mathbf{x})$ could be a GLO-MAP weight of desired order. Figure 8(a) shows Shepard PU weights constructed by re-normalization of C^2 GLO-MAP weights. Some comments are in order at this point. While GLO-MAP functions themselves are polynomials, Shepard functions (Eq.3.5) are not - they are rational polynomials. These functions are obviously much more difficult to integrate than piecewise polynomials. In fact, standard numerical integration techniques based on Gaussian quadrature



(a) Shepard's pasting functions constructed from C^2 GLO-MAP weights.

(b) x-Gradient of Shepard's pasting functions.

Fig. 8. Partition of unity weights and their x-derivatives for pPUFEM.

are usually inadequate since they are optimized for integrating polynomial like functions. Additionally, unlike standard GLO-MAP weights, Shepard functions may have steep gradients, also because of their rational polynomial nature, as is evident from Fig.8(b). At the same time, they provide great flexibility and allow construction of conformal approximation spaces even with staggered node distributions on Ω. This work utilizes quasi Monte-Carlo techniques to integrate shape functions constructed with Shepard's weights.

3. Variational Formulation

Once the approximation space is set up, the variational formulation of FPE is common in both versions of PUFEM. An approximation of the instantaneous pdf can now be written as:

$$\widehat{\mathcal{W}}(t,\mathbf{x}) = \sum_{i=1}^{P} \varphi_i(\mathbf{x}) \sum_{j=1}^{Q_i} a_{ij}(t)\psi_{ij}(\mathbf{x}) = \sum_{i=1}^{P}\sum_{j=1}^{Q_i} a_{ij}(t)\Psi_{ij}(\mathbf{x}) \tag{3.6}$$

where, $a_{ij}(t)$ are time-varying coefficients (to be determined) of local shape functions denoted by $\Psi_{ij}(\mathbf{x})$. Note that Eq.3.6 is equivalent to separation of time and space variables. Substitution of the above approximation into Eq.2.2 results in the following residual error (equation error):

$$\mathfrak{R}(t,\mathbf{x}) = \sum_{i=1}^{P}\sum_{j=1}^{Q_i}[\dot{a}_{ij}(t)\Psi_{ij}(\mathbf{x}) - a_{ij}(t)\mathcal{L}_{\mathcal{FP}}(\Psi_{ij}(\mathbf{x}))] \tag{3.7}$$

The objective of variational formulation, also known as weak formulation, is to minimize the residual error shown above *on average* over the solution domain, Ω. This can be done by the standard error projection technique using a finite dimensional space of test functions $\mathfrak{V} = \{v_{ij}(\cdot); i = 1,\dots P, j = 1,\dots Q_i\}$. In this work we follow the Galerkin approach, whereby test functions are chosen to be the same as the shape functions, i.e., $\mathfrak{V} = \{\varphi_i(\mathbf{x})\psi_{ij}(\mathbf{x})\}$. Normal equations of error projection then give:

$$\int_{\Omega_i} \mathfrak{R}(t,\mathbf{x})v d\mathbf{x} = \alpha \int_{\Gamma_i\cap\Gamma} (\widehat{\mathcal{W}}(t,\mathbf{x}) - \mathcal{W}_\Gamma(t,\mathbf{x}))v d\mathbf{x} \tag{3.8}$$

or,

$$\int_{\Omega_i} \frac{\partial}{\partial t}(\widehat{\mathcal{W}}(t,\mathbf{x}))v d\mathbf{x} - \int_{\Omega_i} \mathcal{L}_{\mathcal{FP}}(\widehat{\mathcal{W}}(t,\mathbf{x}))v d\mathbf{x} = \alpha \int_{\Gamma_i\cap\Gamma} (\widehat{\mathcal{W}}(t,\mathbf{x}) - \mathcal{W}_\Gamma(t,\mathbf{x}))v d\mathbf{x} \tag{3.9}$$

Note that integrals are computed over the local subdomain because of compact support of shape functions. The coefficients $a_{ij}(t)$ are the unknowns which can be determined using a sufficient number of test functions, v. It is clear from the Eq.3.6 that a total of $\sum_{i=1}^{P} Q_i$ linearly independent test functions are required. Boundary conditions are enforced using a penalty parameter "α" over the part of the local boundary that intersects with the global boundary, Γ. Although it is ideally desired to have $\mathcal{W}_\Gamma = 0$, we implement artificial boundary conditions (as $\mathcal{W}_\Gamma \approx 10^{-9}$) using the penalty parameter α. Putting together the projection equations from all local subdomains, we

are led to the following system of linear ODEs involving the mass matrix $\mathbf{M}$, stiffness matrix $\mathbf{K}$ and load vector $\mathbf{f}$:

$$\sum_{i=1}^{P}\sum_{j=1}^{Q_i}[\int_{\Omega_i}\{\dot{a}_{ij}(t)\varphi_i(\mathbf{x})\psi_{ij}(\mathbf{x})vd\mathbf{x} - a_{ij}(t)\mathcal{L}_{\mathcal{FP}}(\varphi_i(\mathbf{x})\psi_{ij}(\mathbf{x}))\}vd\mathbf{x} - \alpha\int_{\Gamma_i\cap\Gamma} a_{ij}(t)\varphi_i(\mathbf{x})\psi_{ij}(\mathbf{x})vd\mathbf{x}] = -\alpha\int_{\Gamma_i\cap\Gamma}\mathcal{W}_\Gamma vd\mathbf{x} \tag{3.10}$$

or,

$$\mathbf{M}\dot{\mathbf{a}}(t) = \mathbf{K}\mathbf{a}(t) + \mathbf{f} \tag{3.11}$$

where,

$$M_{ij} = \int_{\Omega_{\text{sub}}}\varphi_k(\mathbf{x})\psi_{kl}(\mathbf{x})\varphi_p(\mathbf{x})\psi_{pq}(\mathbf{x})d\mathbf{x} \tag{3.12}$$

$$K_{ij} = \int_{\Omega_{\text{sub}}}\mathcal{L}_{\mathcal{FP}}(\varphi_k(\mathbf{x})\psi_{kl}(\mathbf{x}))\varphi_p(\mathbf{x})\psi_{pq}(\mathbf{x})d\mathbf{x} - \alpha\int_{\Gamma_{\text{sub}}\cap\Gamma}\varphi_k(\mathbf{x})\psi_{kl}(\mathbf{x})\varphi_p(\mathbf{x})\psi_{pq}(\mathbf{x})d\mathbf{x} \tag{3.13}$$

$$f_i = -\alpha\int_{\Gamma_{\text{sub}}\cap\Gamma}\mathcal{W}_\Gamma(t,\mathbf{x})\varphi_p(\mathbf{x})\psi_{pq}(\mathbf{x})d\mathbf{x} \tag{3.14}$$

where, $i = \left(\sum_{s=1}^{k-1} Q_s + l\right)$ and $j = \left(\sum_{s=1}^{p-1} Q_s + q\right)$ and no implicit summation is implied by the repeated subscripts. We mention that the penalty parameter α used above requires minor tuning in order to prevent numerically induced ill-conditioning of the stiffness matrix. Also, note that irrespective of the underlying dynamical system, the norm of the load vector is exceedingly small ($\approx 10^{-6}$), because it involves integration of the pdf over the domain boundary. These facts will be important in section D which discusses the emergence of spurious modes and their elimination.

Note also that the diffusion term in FP operator contains second order derivatives, which may be extremely steep for Shepard function type PU weights used in

the particle version. It is thus good practice to convert the "volume" integrals involving second derivatives to "surface" integrals involving first derivatives utilizing Gauss divergence theorem:

$$\int_{\Omega} \frac{\partial^2 f}{\partial \mathbf{x}^2} d\mathbf{x} = \int_{\Gamma(\Omega)} \nabla . f d\mathbf{x} \tag{3.15}$$

Once the discretized form of FPE is obtained (Eq.3.11), it remains to solve the system of equations for coefficients $a_{ij}(t)$. In existing literature, the linear system of ODEs is integrated numerically, e.g. using a stabilized Crank-Nicholson scheme [67]. Primary reasons for this are typically large sizes of the matrices involved and possible ill-conditioning of the stiffness matrix. In this dissertation, spectral analysis of Eq.3.11 is performed leading to a semianalytical algorithm, which provides near real-time transient response for dynamical systems. This is treated in section D.

a. Considerations for Stationary FPE

On several occasions, it is only required to study the steady state solution of Eq.2.2, i.e. the following equation:

$$\mathcal{L}_{\mathcal{FP}}[\mathcal{W}(\mathbf{x})] = 0 \tag{3.16}$$

In other words, it is required to extract the null-space of the Fokker-Planck operator, which governs the long term behavior of stochastic dynamical systems. Conditions for existence of a nontrivial and unique null-space are well known and can be found in Fuller [1]. A necessary condition is time invariance of system dynamics, i.e. $\mathbf{f}(t, \mathbf{x}) = \mathbf{f}(\mathbf{x})$ and $\mathbf{g}(t, \mathbf{x}) = \mathbf{g}(\mathbf{x})$. Other necessary conditions include existence of finite intensity noise and the presence of at least one attractor in the system. More details can be found in Fuller [1]. The discretized form of stationary FPE reduces to the following algebraic equation: $\mathbf{Ka} = \mathbf{f}$. Note that if trivial boundary conditions

are chosen, i.e., $\mathcal{W}_\Gamma = 0$, the load vector becomes $\mathbf{f} = 0$ and the problem reduces to finding the null space of the stiffness matrix.

Theoretically, if the stationary solution exists, it is unique and globally asymptotically stable, meaning that the null-space of $\mathbf{K}$ should have unit dimensionality. However, this may not hold true for the numerical implementation shown above. If the parameter α is chosen to be too large, it may numerically induce a rank deficiency of greater than 1. In such event, one can study the equation error to determine the best solution. Alternatively, the penalty parameter α can be tuned to obtain a single dimensional (hence unique) null-space. Another approach is to implement the boundary conditions not as $\mathcal{W}_\Gamma = 0$, but a very small value, e.g. $\mathcal{W}_\Gamma = \epsilon$ ($= 10^{-9}$ was used), so that the RHS of Eq.3.11 is not zero. This approach gives highly acceptable results even with very coarse tuning of α. Note that rank deficiency in $\mathbf{K}$ may also be caused due to other factors besides α-tuning, like the failure to incorporate the constraints (1) and (2) mentioned in Sec. C of chapter II. Of course, the matrix $\mathbf{K}$ will always be ill-conditioned if α is not chosen judiciously.

4. Numerical Integration

The integrals contained in Eq.3.10 in general need to be computed numerically over N-dimensional local subdomains, due to which they are susceptible to the curse of dimensionality. Traditional Gaussian quadrature rules fail to compute these integrals effectively on two counts:

1. Gaussian quadrature by design is optimized to integrate polynomial/polynomial-like functions and its performance in integration of rational polynomials is largely inadequate.

2. Standard Gauss quadrature works in high dimensions by taking a continued

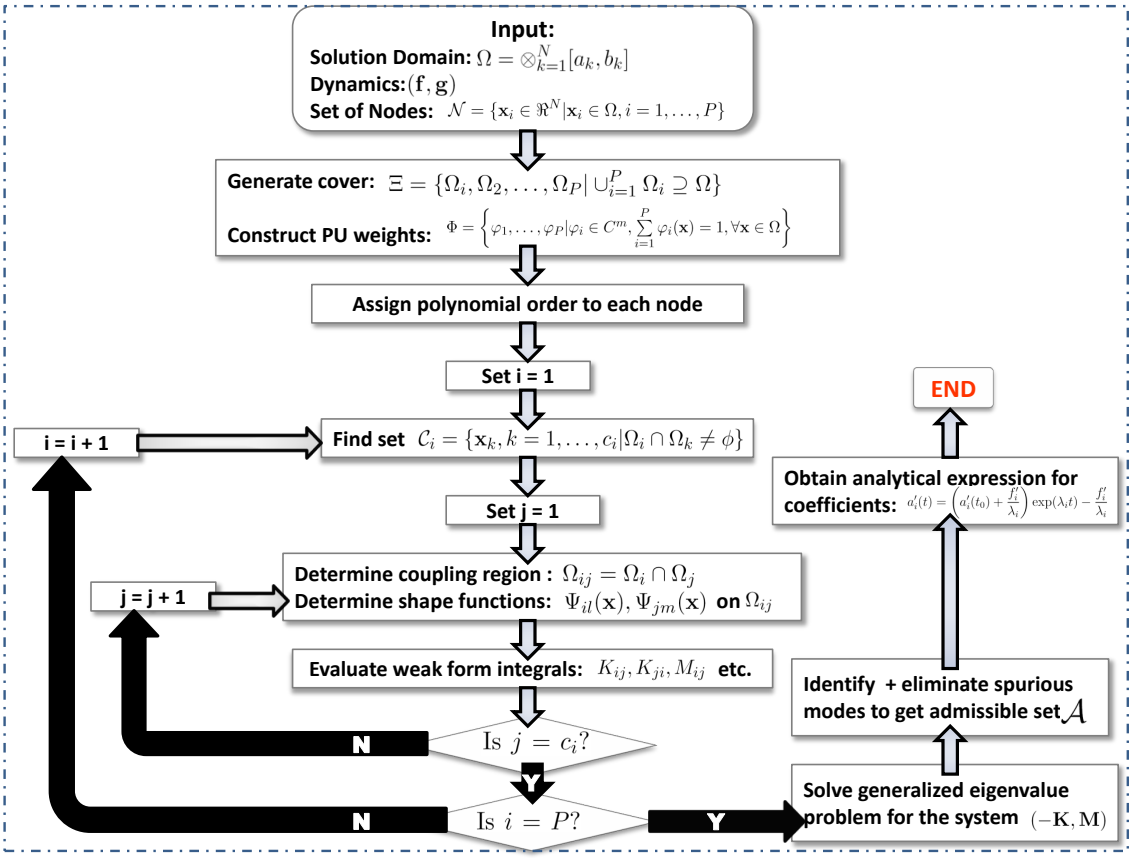


Fig. 9. A schematic of the PUFEM approach for FPE.

product of quadratures in individual dimensions, leading to grid formation - a procedure that suffers from the curse of dimensionality. For example, if 10 points are required to integrate in one dimension, 100 points will be required in 2 dimensions, 1000 in 3 dimensions and so on. This issue can be partially handled using Smolyak tensor product rules for quadrature construction in higher dimensions, but the actual savings remain small and subject to tuning.

The above issues are especially relevant for the particle version of PUFEM where PU weights are rational polynomials. Gaussian quadrature can potentially become a bottleneck in this scenario because even though it might be possible to discretize the solution domain and write FPE in variational form, the integrals might not be computable.

The randomization technique of Monte-Carlo integration is known to break the curse in numerical integration on the canonical domain $\otimes_{i=1}^{N}[0,1)$. In this approach, the integral is computed as a weighted sum of the integrand evaluated at a set of points, much like in standard Gauss quadrature techniques. In this case however, the "quadrature points" are drawn from a uniformly distributed sample and each point carries equal weight. This section briefly discusses a quasi-randomization technique, known as quasi Monte-Carlo (QMC) integration [102, 103], which exploits uniform-distribution like properties of pseudorandom numbers for numerical integration. A sequence of pseudorandom numbers is deterministically generated (i.e. by means of an algorithm), but appears to have a "near" uniform random distribution. In numerical integration, quasi randomization has several benefits over standard randomization: (1) It provides deterministic error bounds on numerical integration as opposed to probabilistic bounds, (2) it provides better convergence properties than standard Monte-Carlo integration and (3) it is much easier to implement because quadrature

points are generated using an algorithm as opposed to a random number generator.

The closeness of a pseudorandom sequence to a true uniform distribution is measured in terms of a discrepancy parameter, the most popular being "star discrepancy," defined as follows: For a psuedorandom sequence $\mathcal{S}_R = \{\mathbf{x}_1, \mathbf{x}_2, \dots \mathbf{x}_R\}$ in an N-dimensional unit cube $I = \otimes_{i=1}^{N}[0,1)$, star discrepancy is defined as the measure:

$$D_N^*(\mathcal{S}_R) \triangleq \sup_{B \in U^*} \left| \frac{A(B;\mathcal{S}_R)}{N} - \lambda(B) \right|, \tag{3.17}$$

where, U^* is the class of open sets of type $\otimes_{i=1}^{N}[0,a_i)$, $a_i \in [0,1)$; $A(B,\mathcal{S}_R)$ represents the number of points of $\mathcal{S}_R$ in B and $\lambda(B)$ is the Lebesgue measure of B in $\Re^N$. Clearly, a sequence with low star discrepancy is closer to a uniform distribution because it has low bias and best approximates the volume of the integration domain. It has been shown in existing literature that for a relatively large class of integrands (in terms of regularity), a low discrepancy sequence achieves the following convergence property:

$$\int_I f(\mathbf{x}) d\mathbf{x} = \lim_{R \to \infty} \frac{1}{R} \sum_{k=1}^{R} f(\mathbf{x}_k) \tag{3.18}$$

where, $\mathbf{x}_k \in I$ belong to a pseudorandom sequence $\mathcal{S}_R$. The above approximation achieves a deterministic error bound of $\mathcal{O}(\frac{(\log(R))^N}{R})$ for sequences with sufficiently low discrepancy, which is much better than the rate $\mathcal{O}(R^{-\frac{1}{2}})$ achieved by standard Monte-Carlo techniques based on uniformly distributed random numbers. The above integration scheme can easily be extended to hypercuboids of other sizes: $\Omega = \otimes_{i=1}^{N}[a_i, b_i]$, by a simple linear transformation of $\mathcal{S}_R$ as: $x_{k,i} \mapsto a_i + (b_i - a_i)x_{k,i}$, and altering the weights to $\frac{1}{R} \mapsto \frac{\text{area}(\Omega)}{R}$. There exist several algorithms for generating pseudorandom numbers with low discrepancy, e.g. the Halton sequence, Sobol sequence, Faure sequence, etc [102]. This dissertation utilizes the Halton sequence for integration of weak form integrals, the algorithm for which is provided in appendix B.

A flowchart of the PUFEM approach has been shown in Fig.9. In this figure, one stops after generating the variational matrices $(\mathbf{K}, \mathbf{M})$ if only the stationary behavior of the system is desired (see section below). Steps involving the generalized eigenvalue problem of Eq.3.11 are outlined in section D, and they lead to a near real-time solution of the transient FPE.

C. Results for Stationary FPE

This section presents results for stationary FPE for nonlinear dynamical systems using both s- and p- versions of PUFEM. The core issue of curse of dimensionality is addressed and numerical evidence is shown for breaking of the curse. A comparison with existing finite element methods shows reduction in problem size by several orders of magnitude for similar accuracy of approximation, e.g. three orders of magnitude for problems in 4-dimensional space. Results are presented separately for sPUFEM and pPUFEM. It is shown that pPUFEM provides unmatched flexibility and working accuracy of approximation with a very small number of degrees of freedom.

1. Curse of Dimensionality: Size of the Discretized Problem

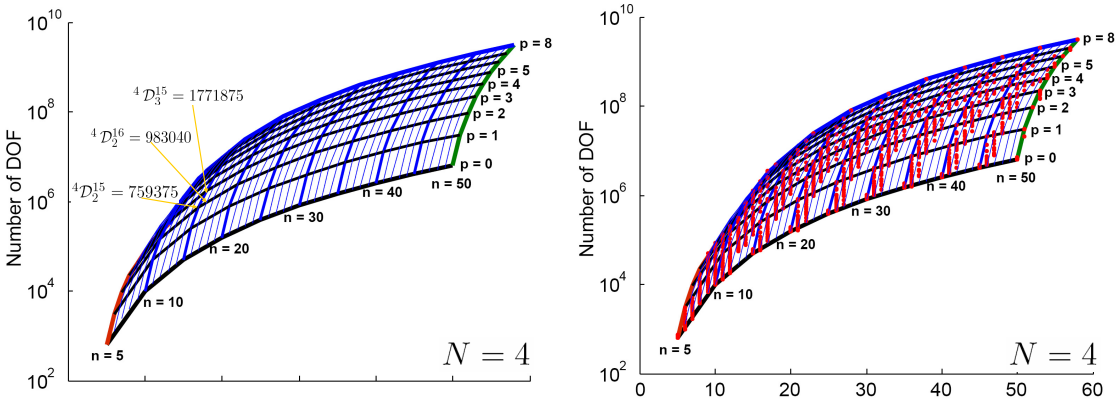
The size of the discretized problem, or, "degrees of freedom" of the approximation is the number of shape functions required to achieve a particular level of accuracy. In s-PUFEM, if n nodes are used to discretize each of the N dimensions of state-space and every node is endowed with a complete set of polynomial basis functions up to p^{th} order, the number of coefficients to determine, or the degrees of freedom (DOFs) of the approximation is given by the following expression: (depicted by ${}^{N}\mathcal{D}_{p}^{n}$)

$$ {}^{N}\mathcal{D}_{p}^{n} = n^{N} \times \sum_{k=0}^{p} \binom{N+k-1}{N-1} \tag{3.19} $$

Note that increasing the number of nodes (n) or the order of basis functions per node (p) both cause an explosive growth in problem size, ${}^{N}\mathcal{D}_{p}^{n}$, especially for large N. Therefore, in the absence of local p-refinement, sPUFEM would suffer from the curse of dimensionality. Local p-refinement (local p-enrichment) allows us to increase the order of basis functions at only selected nodes. In this scenario, we count the basis functions for each node individually leading to an expression for the total number of DOFs that is different from Eq. 3.19:

$$ {}^{N}\mathcal{D}_{loc_p}^{n} = \sum_{i=1}^{n^N} \sum_{k=0}^{p_i} \binom{N+k-1}{N-1} \tag{3.20} $$

The above equation shows that local p-refinement does not reduce the curse of dimensionality directly because the number of nodes still grows exponentially with N. However, it has been found to consistently lead to a significant reduction in number of nodes actually required per dimension because all nodes are not enriched equally. Although not broken rigorously, the curse of dimensionality is "indirectly" significantly curtailed. Its effect on problem size is illustrated in Fig.32. Fig.10(a) shows the growth in degrees of freedom in a standard PUFEM method without local p-refinement. The dimensionality of the underlying system is assumed to be 4. Horizontal curves are contours of constant polynomial order per node (varied from $p = 0$ to $p = 8$) while vertical curves represent contours of fixed number of nodes per dimension ($n = 5$ to $n = 50$). A grid is therefore formed and it is only possible to jump from one grid point to another. The y-axis shows the number of DOFs associated with each (n, p) pair. Three points are highlighted: ${}^{4}\mathcal{D}_{2}^{15} = 759375$, ${}^{4}\mathcal{D}_{2}^{16} = 983040$ and ${}^{4}\mathcal{D}_{3}^{15} = 1771875$. Clearly, the jump in DOFs per added node for each dimension or per polynomial order for each node are enormous, which greatly restricts flexibility. For example, it is not possible to build an approximation with 800000 degrees of freedom because it is does not lie on the grid. On the other hand, having the option of enrich-



(a) No local p-refinement available. (b) DOF growth with local p-refinement.

Fig. 10. Growth of DOF in sPUFEM with and without local p-refinement.

ing only selected nodes helps control this growth and balance between enrichment of the approximation space and growth in problem size. It makes the points in-between the grid-points in Fig.10(a) accessible (see Fig.10(b)), e.g. $^4\mathcal{D}^{15}_{\text{loc}_p} \approx 800000$. The curse would be truly broken if DOFs did not grow exponentially with added nodes. This feature is available in pPUFEM, wherein it is possible to use any desired number of nodes for any dimensional state-space (e.g. 5 nodes for 4D state-space). The resulting problem size is given by:

$$^N\mathcal{D}^P_{loc_p} = \sum_{i=1}^{P} Q_i \tag{3.21}$$

If special (non-polynomial) functions are used, Eq.3.21 is independent of N and the curse of dimensionality stands broken. Note that the number of DOFs in Eq.3.21 does not guarantee an approximation with a specified order of error. It merely presents a framework in which it is possible, given appropriate shape functions, to construct an approximation with a small number of unknowns in high dimensions. With this in mind let us look at specific results for several nonlinear dynamical systems. All

results presented below were obtained on a small computer (1.86 GHz Pentium M processor with 1 GB random access memory).

2. Stationary FPE Results with sPUFEM

a. Dynamical System 1: Example in Two Dimensions

We first consider the following 2 dimensional nonlinear dynamical system:

$$\ddot{x} + \eta\dot{x} + \alpha x + \beta x^3 = g(t)\mathcal{G}(t) \tag{3.22}$$

Eq.3.22 represents a stochastic Duffing oscillator with damping ($\eta > 0$) and is used widely for modeling of nonlinear vibrations. The expression for the true solution to the stationary FPE for this system is as follows:

$$\mathcal{W}_s(x,\dot{x}) \mid_{true} = \mathcal{C} \exp\left(-2\frac{\eta}{g^2 Q}\left(\frac{\alpha x^2}{2} + \frac{\beta x^4}{4} + \frac{\dot{x}^2}{2}\right)\right), \tag{3.23}$$

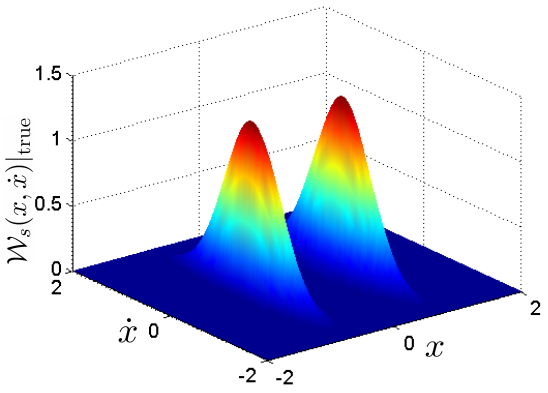
where $\mathcal{C}$ is a normalization constant. Note that the stationary pdf is an exponential function of the steady-state system energy (a Hamiltonian-like function), scaled by the parameter $-2\frac{\eta}{g^2 Q}$ [1]. For simulation purposes, we use $\alpha = -15$, $\beta = 30$, $\eta = 10$ along with $g = 1$. The stationary pdf corresponding to these parameter values is bimodal, shown in Fig. 11(a).

Following the discussion of rank deficiency of $\mathbf{K}$ in Sec.3, boundary condition was implemented as $W_\Gamma = \epsilon$ ($= 10^{-9}$), resulting in a non-zero load vector. Fig.11 shows the solution and error surfaces obtained using the PUFEM algorithm on a 18×18 rectangular grid with quadratic basis functions allocated to each node, i.e. $n = 18$ and $p = 2$ in Eq.3.19. This is equivalent to a stiffness matrix of size 1944×1944. In other words, the size of the discretized problem (or the number of coefficients to determine) is ${}^2\mathcal{D}_2^{18} = 1944$. The results of this discretization are shown in Figs.11(b)-11(c).

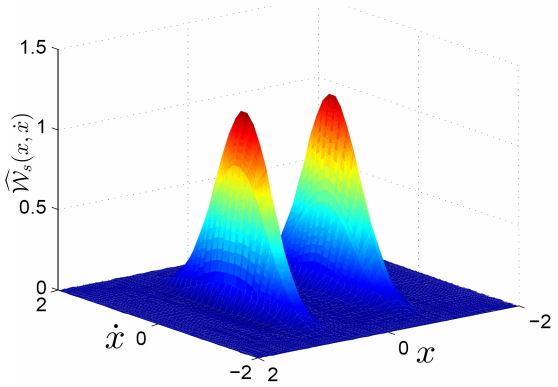
Table I. Numerical results using sPUFEM with local p-refinement: **Two-state Duffing oscillator**

Node Discretization	Polynomial Orders (p)	Problem Size (DOF)	RMS Error (e_2)	Max Error (e_∞)	Computation Time (t_{comp})
18×18	Boundary Nodes: $p = 2$ Interior Nodes: $p = 2$	${}^2\mathcal{D}_2^{18} = 1944$	6.856×10^{-3}	5.905×10^{-2}	30.1 s
9×9	Boundary Nodes: $p = 1$ Interior Nodes: $p = 4$	${}^2\mathcal{D}_{\text{loc}_p}^9 = 831$	5.998×10^{-3}	4.539×10^{-2}	10.7 s

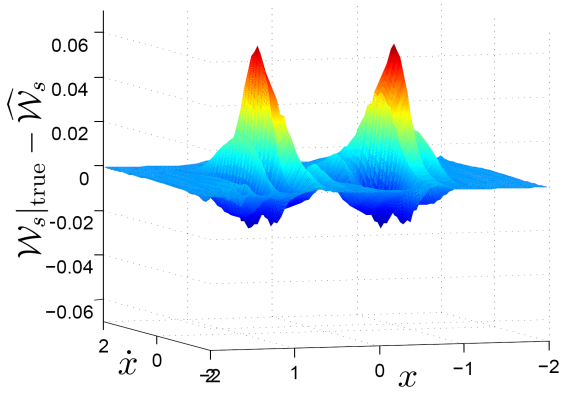
The power of local p-refinement can be illustrated by considering the size of the discretized problem versus approximation accuracy. It is reasonable to assume that using quartic polynomials (instead of quadratic) would lead to better accuracy of the approximation. If quartic polynomials were thus assigned to all nodes (global p-refinement), the resulting problem size would be ${}^2\mathcal{D}_4^{18} = 4860$. On the other hand, we know that the pdf is expected to be almost flat near the boundary of the global domain, and linear basis functions would likely be sufficient to capture its behavior in these regions. Following this reasoning, it is possible with the current approach to supply the nodes lying on the boundary with linear basis functions and the interior nodes with quartic basis functions (local p-refinement). The resulting discretized FPE contains 4044 DOFs, which is a sizeable reduction of about 17%. Depending on the extent and nature of a-priori information available about the particular problem at hand, it is possible to decide on the best polynomial assignment for every individual node such that an acceptable accuracy is obtained with a small number of DOFs. Table I shows one such example for the Duffing oscillator.



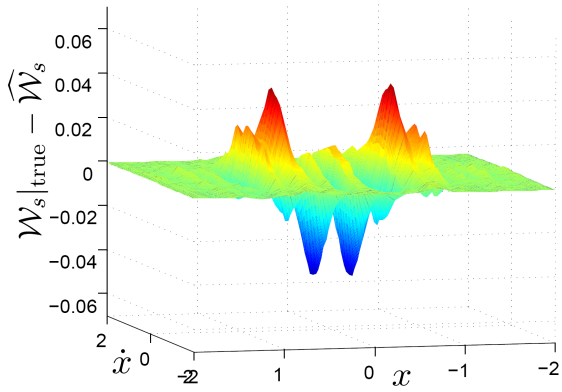
(a) True stationary pdf for the Duffing oscillator.



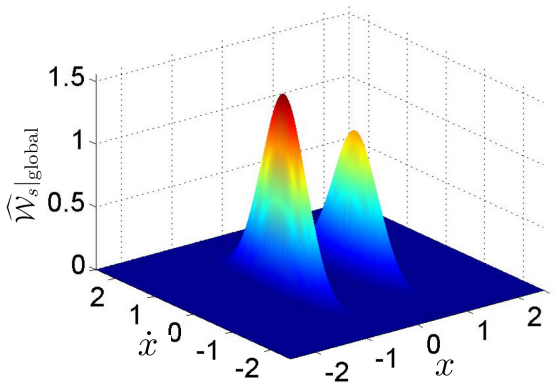
(b) Computed Solution: sPUFEM Algorithm, 18×18 Grid with quadratic basis. DOFs = 1944



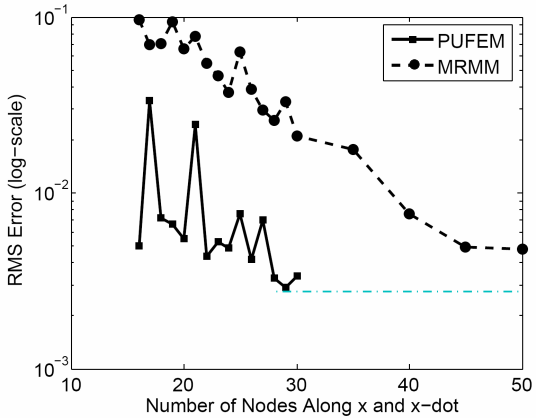
(c) Error surface: sPUFEM algorithm, 18×18 grid with quadratic basis.



(d) Error surface: sPUFEM algorithm, 9×9 grid with linear and quartic basis. DOFs = 831



(e) Computed solution (Global method, mean of the reference pdf = $(0, 0.0075)$).



(f) Comparative convergence characteristics of PUFEM and MRMM for the damped Duffing oscillator.

Fig. 11. Numerical results using the sPUFEM algorithm and global Galerkin approach.

Both discretizations shown in Table I result in approximations with comparable RMS error (defined as $e_2 \triangleq \sqrt{\frac{1}{r-1}\sum_{i=1}^{r}(\mathcal{W}(\mathbf{x}_i) - \widehat{\mathcal{W}}(\mathbf{x}_i))^2}$) and maximum error (e_∞) (see Figs. 11(c)&11(d)). However, the problem size for the second discretization (with local p-refinement) is less than half of the first, in addition to a reduction of about a third in the time of computation. This is an extremely important result, because it illustrates the fact that local p-refinement can provide same/better accuracy with a much smaller number of degrees of freedom, which augurs extremely well for higher dimensional problems.

For comparison, the same problem was solved using the global-Galerkin approach[66] with scaled Hermite polynomials as basis functions. It was found that although the global approximation is able to provide similar accuracy for this problem, it is not a suitable approach because it is extremely sensitive to certain tuning parameters. One such parameter is a reference pdf, which is used to determine the finite domain of solution, and attaches relative weights to different regions of the domain. A slight perturbation in the reference leads to a significant rise in the error, and the degree of tuning achieved in this study case may not be possible for general nonlinear systems. An example of such sensitivity is shown in Fig.11(e), in which the mean of the reference pdf was perturbed towards one of the modes, resulting in unbalanced weighting of the domain leading to significant errors. On the other hand, the PUFEM is not subject to such tuning issues. Moreover, there is no scope for local solution refinement in the global method.

In Kumar et al.[76], the above problem was solved using a multi-resolution meshless method (MRMM) based on the Meshless Petrov Galerkin approach (MLPG). The convergence characteristics of PUFEM were found to be significantly better than MRMM, as seen in Fig.11(f). Although convergence rate of the latter algorithm is faster, RMS error values are higher. The fast rate of convergence of MRMM is most

likely due to decrease in interpolation errors as the density of nodes is increased. Also, the PUFEM algorithm is considerably more computationally efficient, i.e. the time of execution of the PUFEM algorithm is much less than that for MRMM. This is primarily due to the fact that MRMM requires solution of several MLS problems (see Sec.B2) during evaluation of the weak form integrals. Thus for this particular problem, PUFEM provides improvement in accuracy and efficiency over other meshless methods based on MLS.

b. Dynamical System 2: Example in Two Dimensions

Consider now the following 2 dimensional quintic oscillator:

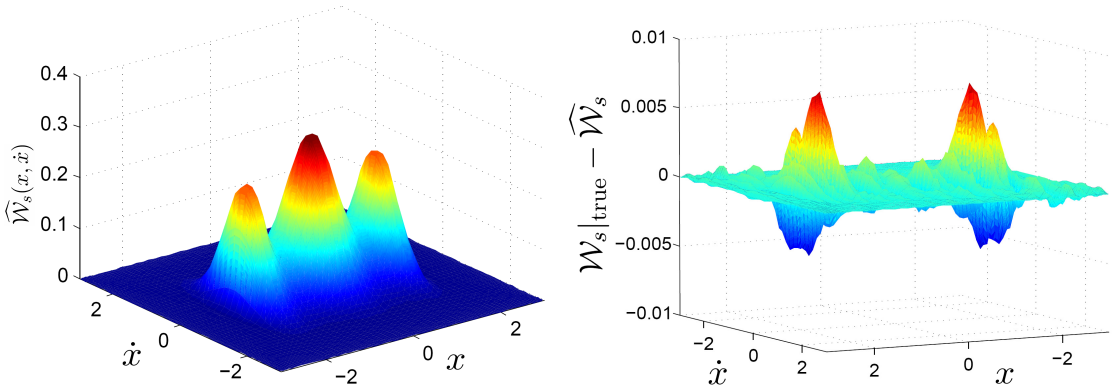
$$\ddot{x} + \eta\dot{x} + x(\varepsilon_1 + \varepsilon_2 x^2 + \varepsilon_3 x^4) = g(t)\mathcal{G}(t) \tag{3.24}$$

The stationary pdf for this system is given by the following expression:

$$\mathcal{W}_s(x,\dot{x}) \mid_{true} = \mathcal{C} \exp\left(-2\frac{\eta}{g^2 Q}\left(\frac{\varepsilon_1 x^2}{2} + \frac{\varepsilon_2 x^4}{4} + \frac{\varepsilon_3 x^6}{6} + \frac{\dot{x}^2}{2}\right)\right), \tag{3.25}$$

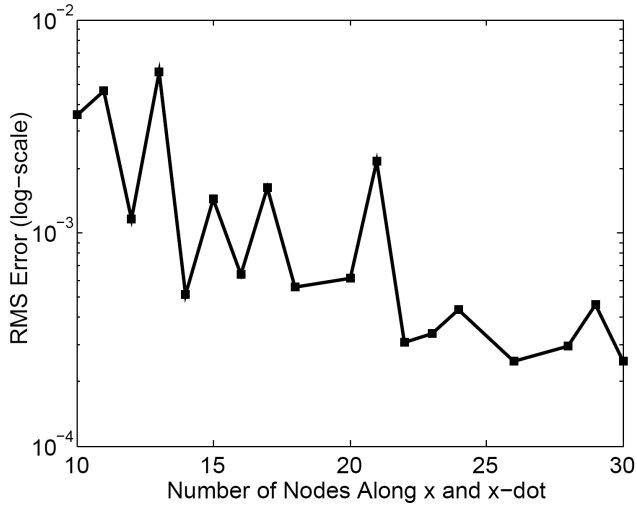
Values of various parameters used in this simulation were: $\varepsilon_1 = 1$, $\varepsilon_2 = -3.065$, $\varepsilon_2 = 1.825$, $\eta = 1.5$, $g = 1$. The stiffness matrix and load vector for this system are constructed exactly in the same manner as for system 1. From Fig.12 it is clear that the method is able to handle systems with high order nonlinearity with ease. Comparative convergence curves for this system, using PUFEM and MRMM show a similar trend as for system 1. Table II shows results for problem size reduction for this system. Clearly, these results are quite similar to the ones obtained for the Duffing oscillator. This example further reaffirms the ability of the current technique to handle highly nonlinear systems and multi-modal behavior accurately.

Study of the above two systems indicates that p-refinement (enrichment of basis) typically provides superior error-reduction than h-refinement (adding more nodes).



(a) Computed solution: sPUFEM algorithm, 18×18 grid with quadratic basis. DOFs = 1944

(b) Error surface: sPUFEM algorithm, 18×18 grid with quadratic basis.



(c) Convergence characteristics for the quintic oscillator using sPUFEM.

Fig. 12. Numerical results for the quintic oscillator using the sPUFEM algorithm.

Table II. Numerical results using sPUFEM with local p-refinement: **Two-state quintic oscillator**

Node Discretization	Polynomial Orders (p)	Problem Size (DOF)	RMS Error (e_2)	Max Error (e_∞)	Computation Time (t_{comp})
18×18	Boundary Nodes: $p = 2$ Interior Nodes: $p = 2$	${}^2\mathcal{D}_2^{18} = 1944$	6.121×10^{-4}	7.750×10^{-3}	33.2 s
9×9	Boundary Nodes: $p = 1$ Interior Nodes: $p = 4$	${}^2\mathcal{D}_{\text{loc}_p}^9 = 831$	7.469×10^{-4}	6.127×10^{-3}	12.9 s

It greatly curtails problem size for the same level of accuracy by reducing the number of nodes required along each dimension. This fact is illustrated effectively in Fig.13, which shows that p-refinement is clearly superior for error reduction. Note that it is possible to move towards the darker regions of low error on this graph by either h- or p-refinement. However, the latter approach clearly requires a very few number of nodes per dimension for achieving low approximation error. On the other hand, if one is constrained to work with fixed order polynomials (e.g. constant or linear polynomials), a very large number of nodes per dimension is required before the dark zone is reached, implying slower convergence. Such h-refinement is typically employed in standard FEM.

Overlaid on the error contours are contours of problem size, i.e. number of DOFs for a given n and p. As expected, problem size increases monotonically upon increasing n and/or p (see Eq.3.19). Therefore, as long as one remains underneath the DOF contour of a particular value, say the $\mathcal{D}^\star$-contour, the size of discretized problem remains less than $\mathcal{D}^\star$. This is very useful information, because now looking at the

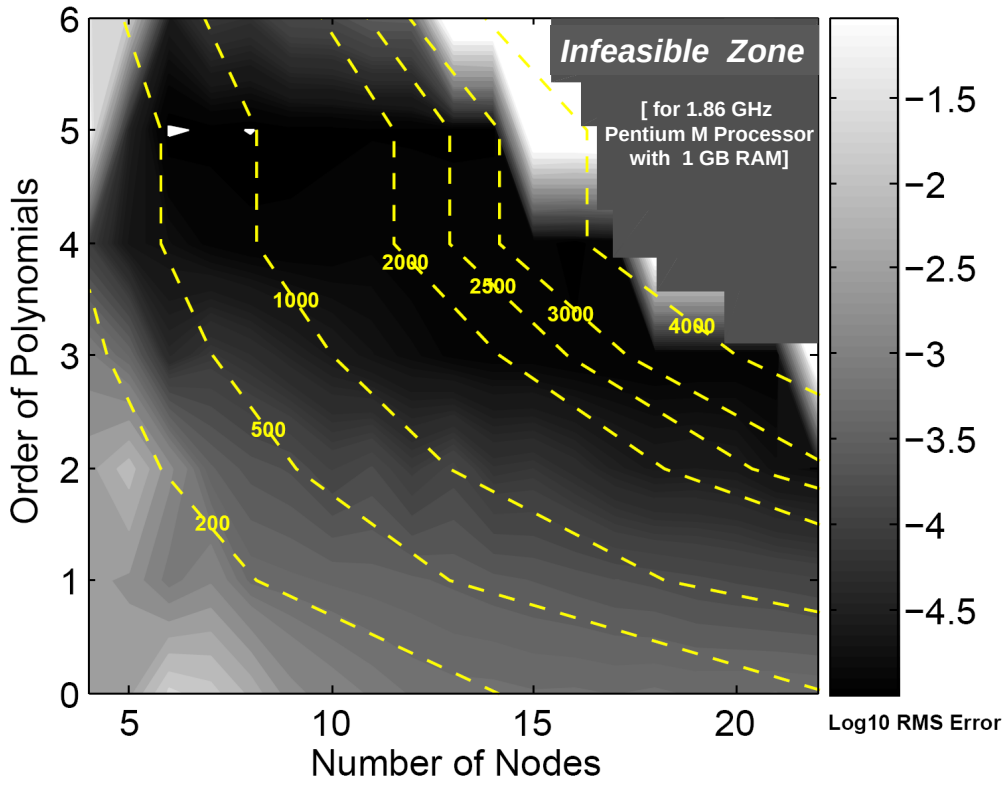


Fig. 13. An accuracy-feasibility contour map of standard-PUFEM for a two-state system.

composite contour-map in Fig.13, we see that p-refinement provides not only lower error, it also helps keep the problem size small. Combined together, it leads to high accuracy with less computational effort. This is the ultimate criteria for being able to attack problems in higher dimensions. Fig.13 has been referred to as an "accuracy-feasibility contour map" because it provides complete information about the number of nodes and order of polynomials required for desired accuracy, while keeping in mind available computational resources (i.e. size of the discretized problem to be solved). As previously mentioned, all results were obtained on a small workstation equipped with a 1.86 GHz Pentium M processor and 1 GB RAM. The infeasible domain for this machine (i.e. too many DOFs) are shown in the top-left section of the contour map.

At this point, it is important to state that sPUFEM is not claimed to be a remedy for the curse of dimensionality. It only helps ameliorate the curse, rather than cure it. Especially with the added flexibility of selective, local basis-enrichment, it is possible to keep the growth of problem size under tight check as system dimensionality increases. Besides the systems considered in this section, the PUFEM algorithm was applied to several other nonlinear oscillators in 2-D state-space, in all of which local p-refinement was seen to offer significant reduction in problem size. This aspect of the current method gives it advantage over both the global approach as well as traditional FEM.

Table III. Comparative results using sPUFEM with local p-refinement: **Three-state linear system**

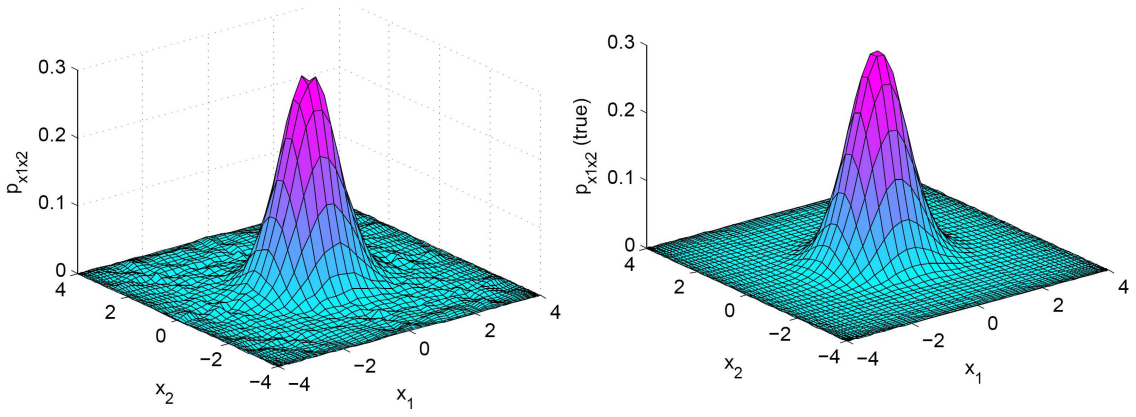
Discretization Method	Problem Size (DOF)	RMS Error (e_2)	Max Error (e_∞)	Computation Time (t_{comp})
$50 \times 50 \times 50$ Brick elements (FEM)	125000	1.133×10^{-4}	not available	not available
$7 \times 7 \times 7$ Nodes Boundary Nodes: $p = 1$ Interior Nodes: $p = 4$ (sPUFEM, local p-refinement)	${}^3\mathcal{D}^7_{\text{loc}_p} = 5247$	2.823×10^{-4}	4.037×10^{-3}	18 min, 42.3 s

c. Dynamic System 3: Example in Three Dimensions

Consider now the following 3-D linear system studied by Wojtkiewicz et.al.[82]:

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 & 0 \\ -\omega_0 & -2\zeta\omega_0 & 1 \\ 0 & 0 & -\alpha \end{bmatrix} \mathbf{x} + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} w(t) \tag{3.26}$$

The constants appearing in the above equation have the following values[82]: $\alpha = \omega_0 = 1, \zeta = 0.2$. The reason for studying a linear system is that its stationary distribution can be obtained easily by solving the corresponding algebraic Riccati equation. The stationary pdf for the above system was approximated by Wojtkiewicz et al. [82] using traditional FEM with "brick" elements in 3-D state-space. Comparative results have been shown in Table III. Approximation accuracy for both methods are approximately the same, but the current method holds a significant advantage in computational load. Fig.14 compares the computed $x_1 - x_2$ marginal pdf to the true marginal. Note that if we were constrained to use quartic polynomials on all



(a) Computed $x_1 - x_2$ marginal distribution.

(b) True $x_1 - x_2$ marginal distribution.

Fig. 14. Computed and true $x_1 - x_2$ marginal distributions for the linear three dimensional stochastic dynamics of Eq.4.45.

nodes, the resulting problem size would be ${}^3\mathcal{D}_4^7 = 12005$. This would most likely provide better accuracy, but at a much higher computational cost, likely beyond the capability of a small computer. Therefore, local p-refinement provides an attractive balance between approximation accuracy and computational cost. In this case, we obtain the same order of accuracy as traditional FEM with two orders of magnitude fewer DOFs, which is a significant improvement.

d. Dynamic System 4: Nonlinear Example in Three Dimensions

Consider now the following nonlinear system with three dimensional state space:

$$\begin{aligned} \dot{x} &= \sigma(y - x) + \zeta_1(t) \\ \dot{y} &= x(\rho - z) - y + \zeta_2(t) \\ \dot{z} &= xy - \beta z + \zeta_3(t) \end{aligned} \tag{3.27}$$

The above equations represent a noise-driven Lorenz attractor. The Lorenz attractor is a chaotic system that was originally used to model climatic changes and has been used since to study numerous physical systems, e.g. laser systems (Maxwell-Bloch model). Numerical values for various parameters appearing above are: $\sigma = 10, \rho = 1; \beta = 8/3$ and Q (noise intensity) $= 2$. Figures 15(a) and 15(b) show the $x - y$ marginal computed from the stationary pdf obtained by solving the static FPE corresponding to Eq. 3.50. The discretization utilized for this solution was a $6 \times 6 \times 6$ nodal grid with boundary nodes endowed with quadratic polynomials and interior nodes with quartic polynomials (corresponding problem size = 3760 DOFs). An analytical result is not available for this system, therefore error was quantified in terms of equation-error and is shown in Fig.15(c). In this figure, equation error has been integrated along the z-direction to obtain a $2D$ surface. The shown error surface has an RMS value of 6.668×10^{-6}.

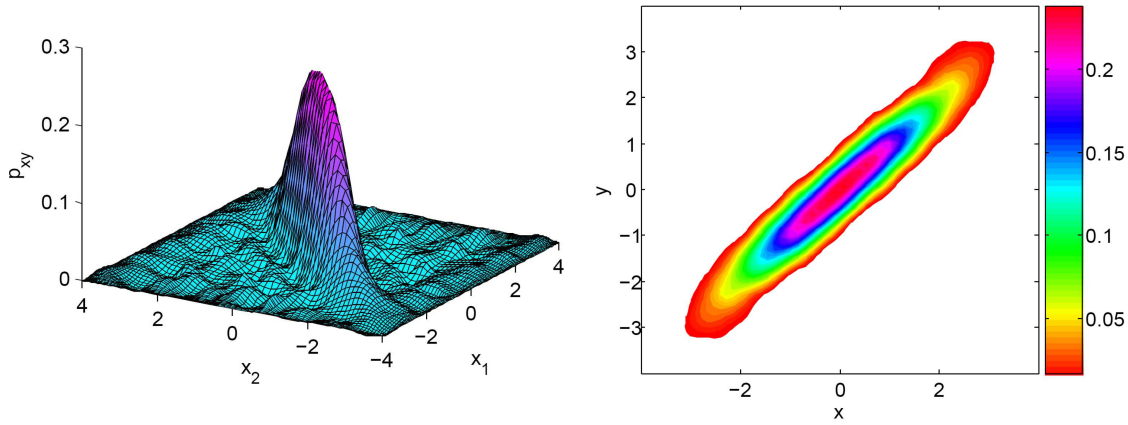
e. Dynamic System 5: Example in Four Dimensions

Here we look at the following linear dynamical system with a four dimensional state space studied by Wojtkiewicz et. al.[84]:

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -(k_1+k_2) & -c_2 & k_2 & 0 \\ 0 & 0 & 0 & 1 \\ k_2 & 0 & -(k_2+k_3) & -c_2 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix} \zeta(t) \tag{3.28}$$

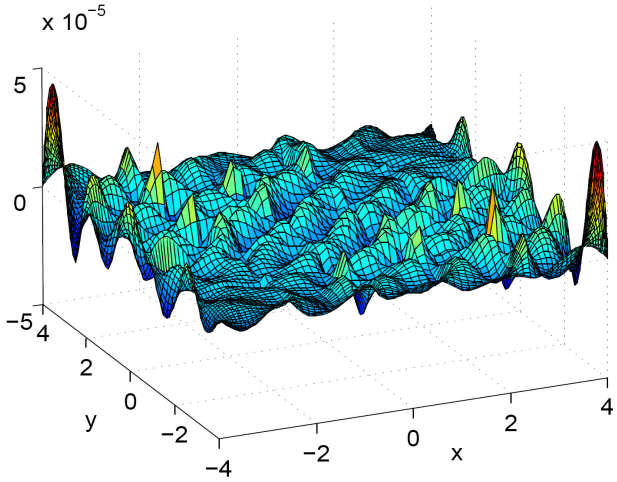
The Fokker Planck equation of concern is:

$$\begin{aligned} \frac{\partial p}{\partial t} = & -x_2 \frac{\partial p}{\partial x_1} - \frac{\partial}{\partial x_2}[(-(k_1+k_2)x_1 - c_1 x_2 + k_2 x_3)p] - x_4 \frac{\partial p}{\partial x_3} \\ & - \frac{\partial}{\partial x_3}[(k_2 x_1 - (k_2+k_3)x_3 - c_2 x_4)p] + D_1 \frac{\partial^2 p}{\partial x_2^2} + D_2 \frac{\partial^2 p}{\partial x_4^2} \end{aligned} \tag{3.29}$$



(a) Computed $x - y$ marginal for the Lorenz attractor.

(b) Contour plot of the computed $x - y$ marginal surface.



(c) Equation error surface integrated along the z-axis.

Fig. 15. Computed $x - y$ marginal distribution for the noise-driven Lorenz attractor of Eq.3.50.

Constants appearing above have the following values: $k_1 = k_2 = k_3 = 1$, $c_1 = c_2 = 0.4$, $D_1 = D_2 = 0.2$. Fig.16 shows the $(x_1 - x_2)$ marginal distribution computed with the meshless sPUFEM method alongside the true marginal surface for the above linear system. Available computing resources allowed the use of only 5 nodes along each of the 4 dimensions, with all interior nodes carrying cubic basis functions and boundary nodes linear polynomials. This leads to a total of 5555 DOFs, which is in sharp contrast to the approximately 2.56 million DOFs used by the standard FEM approach for the same problem. Table IV shows that the current method provides an accuracy of one order of magnitude less than FEM. We note that this is solely due to the limitation of computing resources currently utilized, with which it is not possible to deal with problems of size greater than about 5500. In the absence of local p-refinement, the problem size with cubic polynomials on a $5 \times 5 \times 5 \times 5$ grid would be ${}^4\mathcal{D}_3^5 = 21875$, which is well beyond the capability of a small computer.

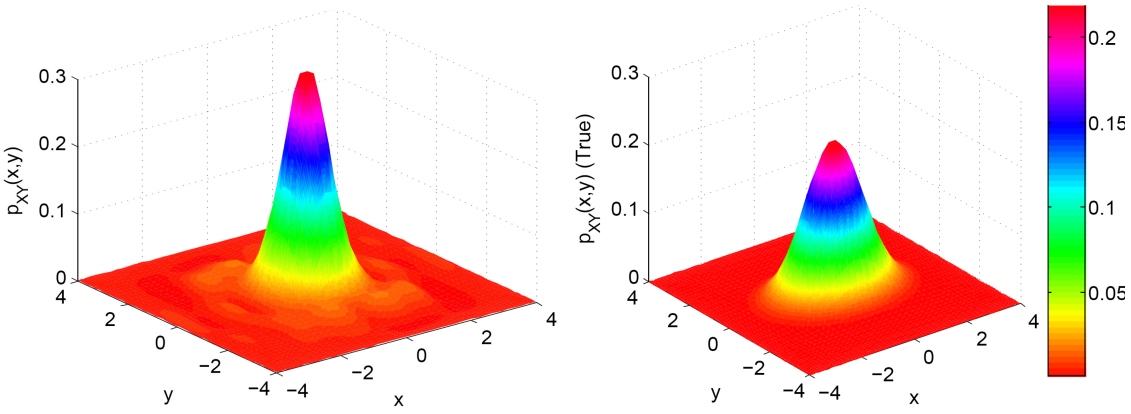
It is important to note therefore that the current approach provides highly acceptable accuracy (e.g. for use in the design of control laws via policy iteration) with better than three orders of magnitude reduction in problem size so that it can be solved on a small computer. Given the ability to deal with moderately larger matrices, it is reasonable to believe that excellent approximations can be obtained for high dimensional problems.

f. Remark on the Curse of Dimensionality

It has been demonstrated above through several examples that sPUFEM, coupled with local p-refinement has the ability to attenuate the effect of curse of dimensionality in FPE. Since the number of nodes required for discretization remains an exponential function of the system dimensionality, a claim to breaking of the curse cannot be made. At the same time, strong evidence towards curtailment of the curse has been

Table IV. Comparative results using sPUFEM with local p-refinement: **Four-state linear system**

Discretization Method	Problem Size (DOF)	RMS Error (e_2)	Max Error (e_∞)	Computation Time (t_{comp})
$40 \times 40 \times 40 \times 40$ 4D "Brick" elements (FEM)	2560000	5.237×10^{-5}	2.911×10^{-3}	not available
$5 \times 5 \times 5 \times 5$ Nodes: Boundary: $p = 1$ Interior: $p = 3$ (sPUFEM, local p-refinement)	${}^4\mathcal{D}^5_{\text{loc}_p} = 5555$	9.769×10^{-4}	8.870×10^{-2}	23 hr 36.4 min



(a) Computed $x_1 - x_2$ marginal surface. (b) True $x_1 - x_2$ marginal surface.

Fig. 16. Comparison of the computed $(x_1 - x_2)$ marginal for the four dimensional linear system with the truth.

presented. To further support this evidence, we look at Fig.17 and Table V. In Fig.17, several problems residing in dimensions $1-4$ have been solved with fixed accuracy ($\mathcal{O}(e_2) \approx 10^{-4}$) using the sPUFEM algorithm and the required number of DOFs is compared with the finite element method. For FEM, DOF requirement grows almost linearly in log-scale indicating the curse of dimensionality. On the other hand, the accuracy curve for sPUFEM tends to flatten out as dimensionality is increased. This trend is extremely encouraging for further progress to problems residing in 5, 6 and even higher dimensions. Approximate fits for the available data have also been shown in Fig. 17. No rigorous conclusions can be drawn from these fits because very little information (4 data points) is available. For FEM, it is well known (also seen in Fig.17) that an exponential fit best describes the problem size growth. On the other hand, a quartic-polynomial fit seems to capture DOF growth in sPUFEM. If indeed true, polynomial growth would mean breaking of the curse of dimensionality, but no formal conclusion is currently possible. Moreover, it is dangerous to extrapolate and no conclusion can yet be drawn for dimensions 5, 6 and higher. With greater computational resources, it would be possible to confirm such extension.

Table V shows the numbers appearing in Fig.17 along with feasibility on a small computer. Since all results for the current approach were obtained on a small computer, it is easy to anticipate that given greater computational ability, sPUFEM will likely handle problems residing in much higher dimensions. An important remark about problem size corresponding to $N = 4$ is due at this point. Table V shows that about 8200 DOFs are required to achieve mentioned accuracy for four dimensional systems. This is a projected number since the currently available resources do not allow solution to problems exceeding approximately 5500 in size. As shown in the results section above, the sPUFEM algorithm has been used to solve a linear problem in 4D state space, requiring 5555 DOFs. Based on this fact, and trends available

Table V. Growth in problem size with underlying system dimensionality: FEM and sPUFEM

N	Problem Size (DOF)		Feasible on a small computer?	
	FEM	sPUFEM	FEM	sPUFEM
1	100	50	Yes	Yes
2	2500	1200	Yes	Yes
3	1,25,000	5200	No	Yes
4	2,560,000	8200	No	No

from lower dimensional problems, it is projected that on average about 8200 DOFs will be required to solve more difficult nonlinear problems in 4D. Consequently, even though linear 4-state problems have been solved using sPUFEM, more general nonlinear problems may not be solvable on small computers, which is also reflected in Table V.

3. Stationary FPE Results with pPUFEM

This section presents results for stationary FPE using the particle version of PUFEM. In all these examples, the solution domain is discretized using randomly distributed nodes, implying that no a-priori knowledge about the system is utilized. It is shown that even with such domain discretization, extremely good approximations can be obtained with a very small number of degrees of freedom. The emphasis in this section is on obtaining approximations with very small number of DOFs, which may not always be possible with sPUFEM.

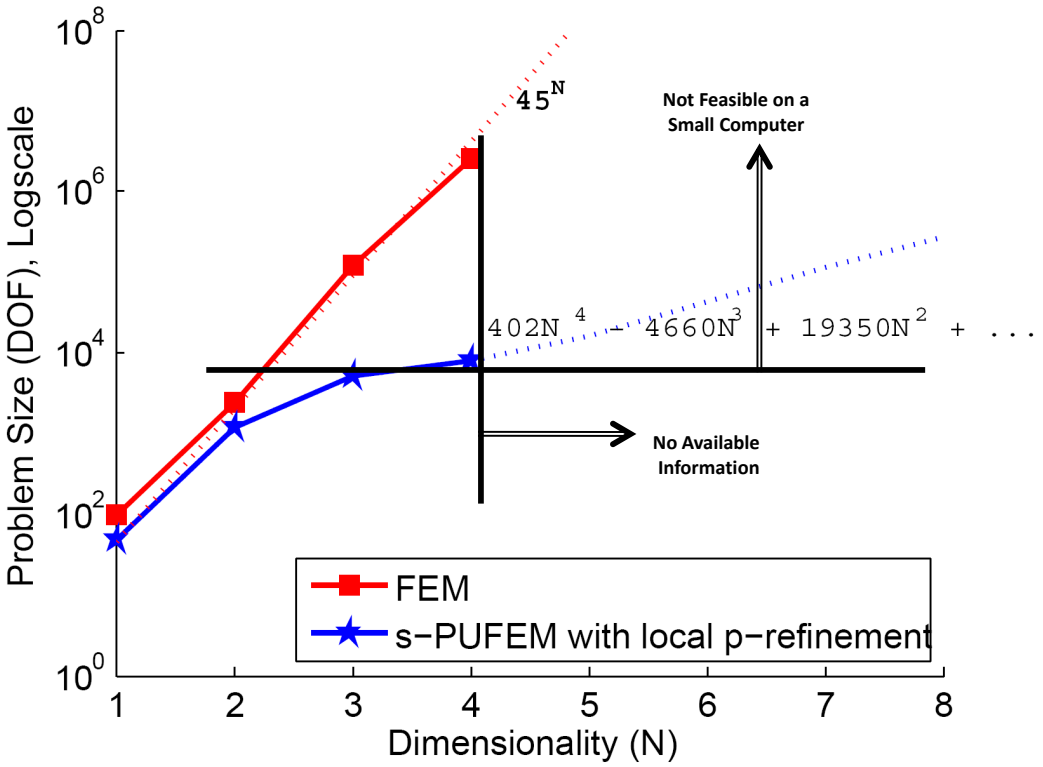


Fig. 17. Ameliorating the *curse of dimensionality* with sPUFEM.

a. Example 1: 2-State System

Consider first the following system in 2D state-space[66]:

$$\ddot{x} + \beta\dot{x} + x + \alpha(x^2 + \dot{x}^2)\dot{x} = g(t)\mathcal{G}(t) \tag{3.30}$$

The above system is known to admit a "volcano-shaped" stationary probability distribution shown in Fig.18. Two approximations using pPUFEM are shown in this section - one with high accuracy, and another with somewhat lower accuracy, but composed of an extremely small number of DOFs. Both approximations are built upon highly unstructured domain discretizations using nodes obtained from a Halton pseudorandom sequence.

b. Highly Accurate Approximation for Example 1

Consider the domain discretization and cover shown in Fig.19(a) using 100 nodes derived from a Halton sequence. Nodes near the boundary are assigned constant basis functions while those near the center are assigned quadratic polynomials, resulting in a problem size of 540 DOFs. The resulting approximation is shown in Fig.19(b) and corresponding error surface in Fig.19(c). The RMS value of the error surface is 1.13×10^{-4}, which is equivalent to a relative error of 0.5%. Note that even for this highly accurate solution, its problem size of 540 compares favorably over sPUFEM (see Table V).

c. Workable, Low Order Approximation for Example 1

The true power of pPUFEM lies in its ability to extract reasonable approximations with a small number of DOFs by means of unstructured node distributions. Fig.20(a) shows an implicit discretization using 32 nodes derived from a Halton se-

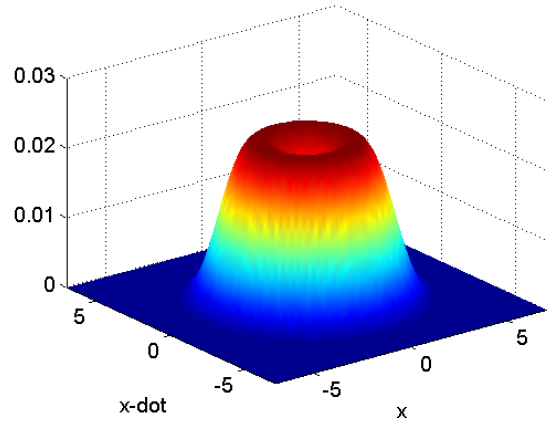
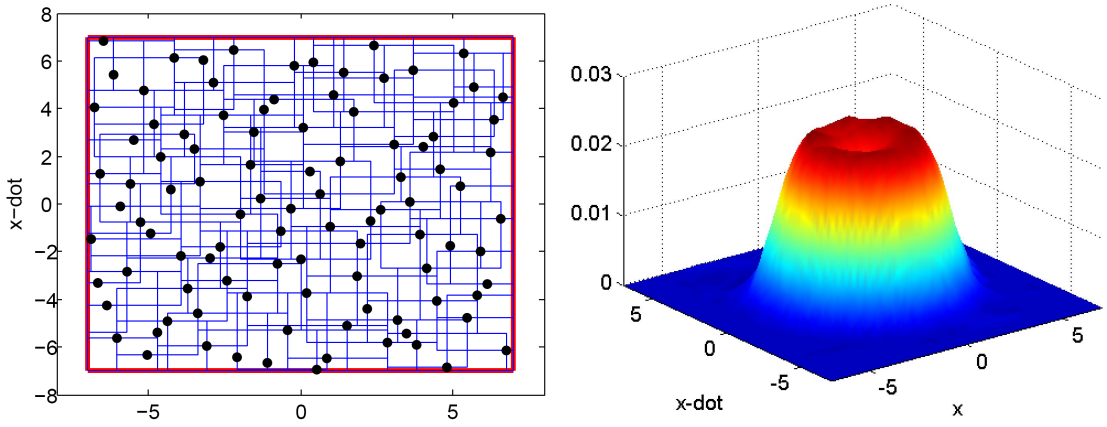


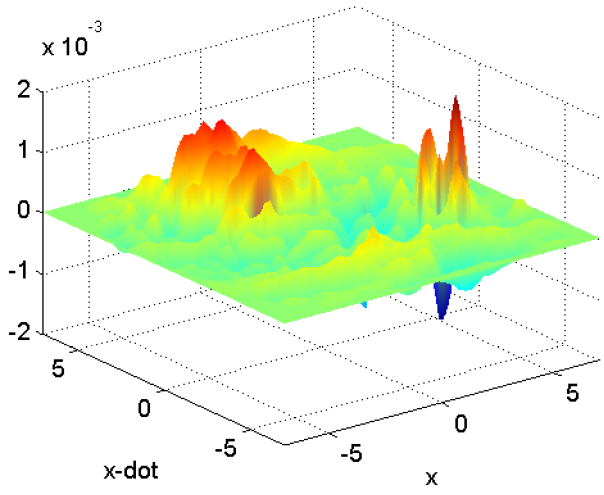
Fig. 18. True stationary pdf for system in Eq.3.30.

quence. Using local p-refinement as in the previous case, a problem size of 172 DOFs is obtained, which is almost an order of magnitude lower than sPUFEM (Table V), although with slightly lower accuracy. The resulting approximation and error surfaces are shown in Fig.20(b) and 20(c), which an RMS error value of 7.22×10^{-4}, equivalent to a relative error of about 3.1%. A pertinent question at this point is: how to these numbers compare with sPUFEM? To get an idea, the sPUFEM accuracy-feasibility plot for this system must be considered (like the one shown in Fig.13). It was found that the darkest point (i.e. lowest RMS error) around the 170-DOF contour on this graph reads an approximate value of $10^{-2.75}$, which corresponds to a relative error of about 7%. One can thus infer that pPUFEM performs better than sPUFEM even with a (pseudo) random domain discretization. The key is that there is minimal wastage of DOFs in pPUFEM because of its unstructured nature of discretization. The constraint of placing nodes on a grid can therefore prove expensive in sPUFEM, especially so in higher dimensions. As for sPUFEM, numerous other examples in $2-$ and $3-$D spaces have been carried out successfully using the pPUFEM algorithm.



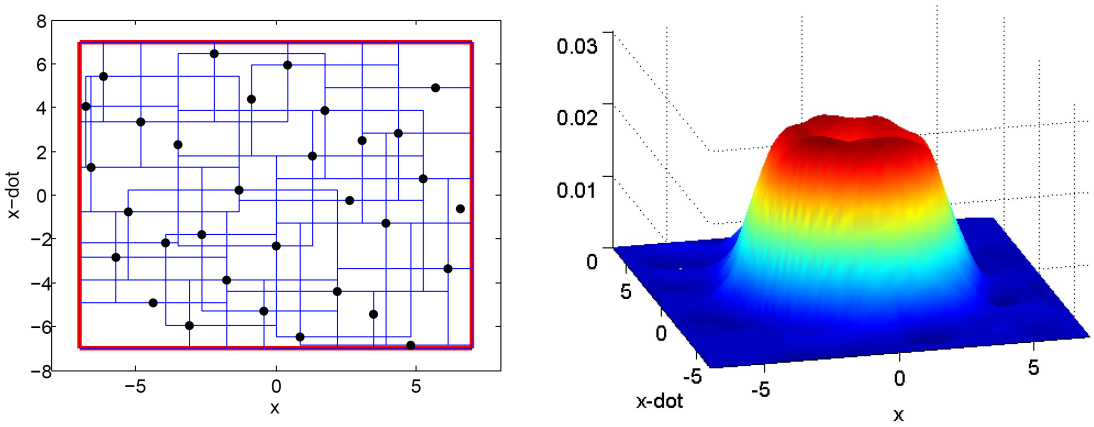
(a) Domain discretization using 100 nodes generated from the Halton sequence.

(b) Obtained approximation of stationary behavior.



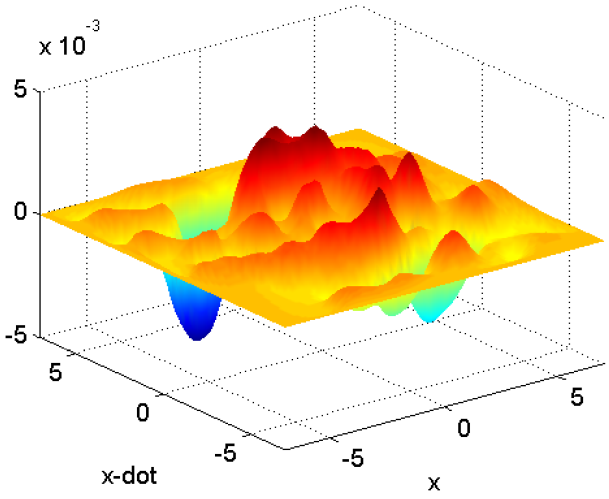
(c) Error surface.

Fig. 19. High accuracy approximation of stationary pdf using pPUFEM for the dynamical system in Eq.3.30.



(a) Domain discretization using 32 nodes generated from the Halton sequence.

(b) Obtained low order approximation of stationary behavior.



(c) Error surface.

Fig. 20. Low order approximation of stationary pdf using pPUFEM for the dynamical system in Eq.3.30.

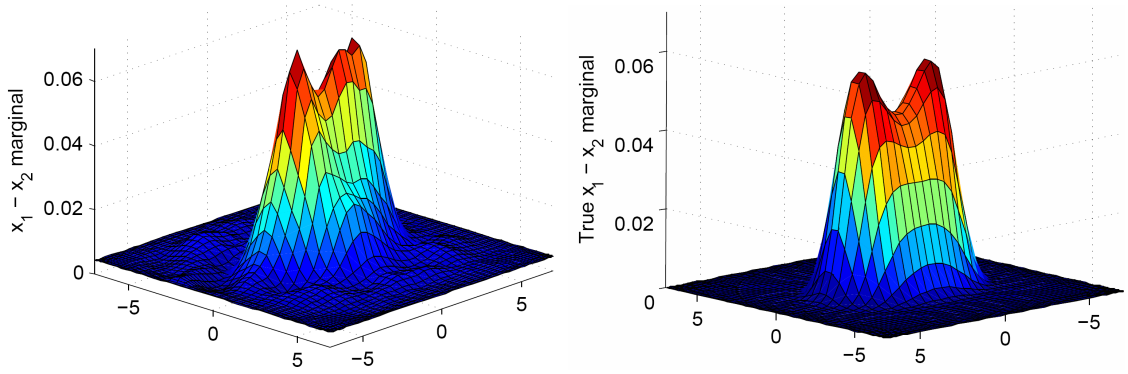
d. Example 2: 4-State System

We next consider a coupled two-degree-of-freedom (i.e. four-state) nonlinear vibration isolating suspension model considered by Ariaratnam [22]:

$$\begin{aligned} \dot{x_1} &= x_3 \\ \dot{x_2} &= x_4 \\ \dot{x_3} &= -\alpha x_3 - \frac{1}{M}\frac{\partial V}{\partial x_1} + \zeta_1 \\ \dot{x_4} &= -\beta x_4 - \frac{1}{I}\frac{\partial V}{\partial x_2} + \zeta_2 \end{aligned} \tag{3.31}$$

$$\tag{3.32}$$

The above system belongs to a small class of nonlinear systems possessing a Hamiltonian structure for which the true stationary distribution of FPE is known. In the above example, the coupled potential function V is given by [22]: $V(x_1, x_2) = k_1 x_1^2 + k_2 x_2^2 + \epsilon(\lambda_1 x_1^4 + \lambda_2 x_2^4 + \mu x_1^2 x_2^2)$. Values of various system parameters used are: $\alpha = 0.5$, $\beta = 1.0$, $k_1 = 0.5$, $k_2 = -0.5$, $\epsilon = 0.5$, $\lambda_1 = 0.25$, $\lambda_2 = 0.125$, $\mu = 0.375$. The noise process is two-dimensional with intensities $D_1 = 2$ and $D_2 = 4$. Note that parameters have been selected so that $\frac{D_1 M}{\alpha} = \frac{D_2 I}{\beta}$, because only under this condition can the analytical solution be written. Of course, the numerical approach does not require this condition to hold, and it has been made only for comparison purposes. Figure 21 shows the stationary $x_1 - x_2$ marginal obtained for this system using the pPUFEM approach. A total of 300 nodes were used to discretize the solution domain, chosen to be $[-6, 6] \otimes [-6, 6] \otimes [-6, 6] \otimes [-6, 6]$. Nodes close to the center of this domain were assigned quadratic polynomial basis and nodes closer to the boundary constant basis functions as in the above example, leading to a total of 2442 DOFs. The resulting stationary marginal is presented in Fig.21(a), which compares reasonably well with the true marginal shown in Fig.21(b). As a comparison, the stationary solution



(a) Computed $x_1 - x_2$ stationary marginal distribution.

(b) True $x_1 - x_2$ stationary marginal distribution.

Fig. 21. Computed and true $x_1 - x_2$ stationary marginal for the coupled 4-state nonlinear suspension model.

was obtained using sPUFEM for a linear 4-state system in the previous section using 5555 DOFs, and traditional FEM for the same linear problem requires 2.56 million DOFs. The advantage of pPUFEM is thus quite clear for this system and this example further underlines the fact that unstructured node distribution in the pPUFEM paradigm gives it an edge over sPUFEM.

e. Example 3: 5-State System

Finally, let us consider a linear example in five dimensional state space modeled by the following equations:

$$\begin{Bmatrix} \dot{x_1} \\ \dot{x_2} \\ \dot{x_3} \\ \dot{x_4} \\ \dot{x_5} \end{Bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ -\omega_1 & -2\xi_1\omega_1 & \alpha_1 & 0 & 0 \\ 0 & 0 & -\alpha & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & \alpha_2 & -\omega_2 & -2\xi_2\omega_2 \end{bmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{Bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{Bmatrix} \zeta_1 \\ \zeta_2 \\ \zeta_3 \end{Bmatrix} \tag{3.33}$$

Nominal values were used for various parameters appearing above and the pPUFEM discretization approach was used on a $[-5,5] \otimes [-5,5] \ldots [-5,5]$ domain with 301 nodes, and local p-refinement as above, leading to a problem size of 5496 DOFs. The resulting approximation was relatively coarse and is shown in Fig.22. Note that this result was obtained on a small computer (laptop with 1 GB computing memory) and thus can be easily improved with few additional DOFs. The result shown in this section illustrates the basic point that pPUFEM makes FPE solvable with extremely small number of DOFs, and problems that have so far only been attempted on supercomputers have been made solvable on the most basic computer available today. Given the capability to perform the described computations on larger computing platforms, it is expected that much more difficult problems can be solved, residing in much higher dimensional spaces.

The discussion on curse of dimensionality can be rounded off by revisiting Fig.17 and adding to it another curve: the one corresponding to pPUFEM (see Fig.23). Computing capability constraints remain the same for both figures, but the extra flexibility of pPUFEM in use of unstructured node distributions makes an additional dimension accessible to solution. Note from Fig.23 that problem size growth is much

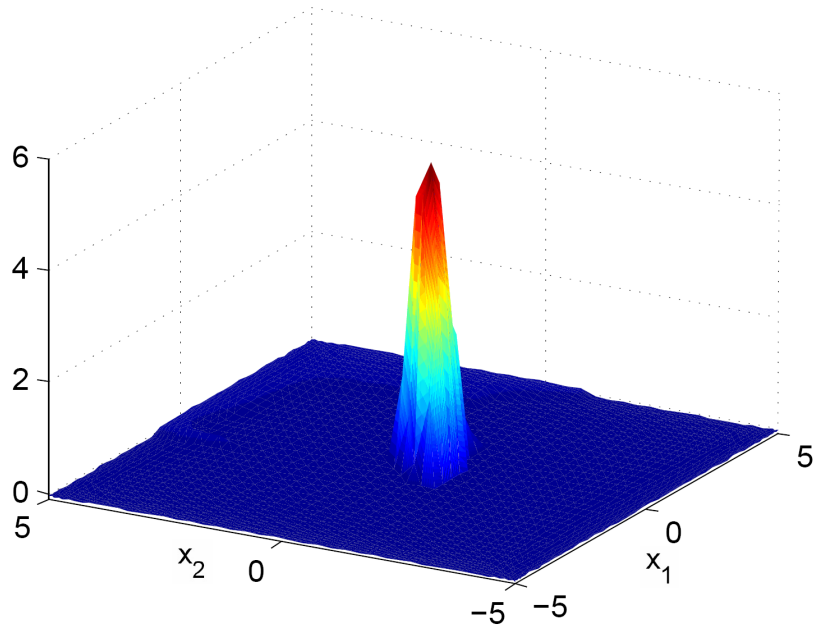


Fig. 22. Computed $x_1 - x_2$ marginal for the linear 5-state system.

slower for pPUFEM than sPUFEM and the amplitude of the leading term in a quartic-polynomial fit is one order of magnitude less than sPUFEM. This suggests definite advantages in higher dimensional applications. However, as mentioned in the discussion surrounding Fig.17, figure 23 offers additional encouraging numerical evidence, but still does not constitute a proof for breaking of the curse of dimensionality.

D. Semianalytical Approach for Transient FPE Response

This chapter thus far primarily considered spatial discretization of the FP operator and long term (i.e. stationary) behavior of the state probability density. The remainder of this chapter will focus on obtaining the transient FPE response, which must precede stationary behavior. This involves solving Eq.3.11 to obtain time varying coefficients $a_i(t)$, which parameterize the approximated transient response,

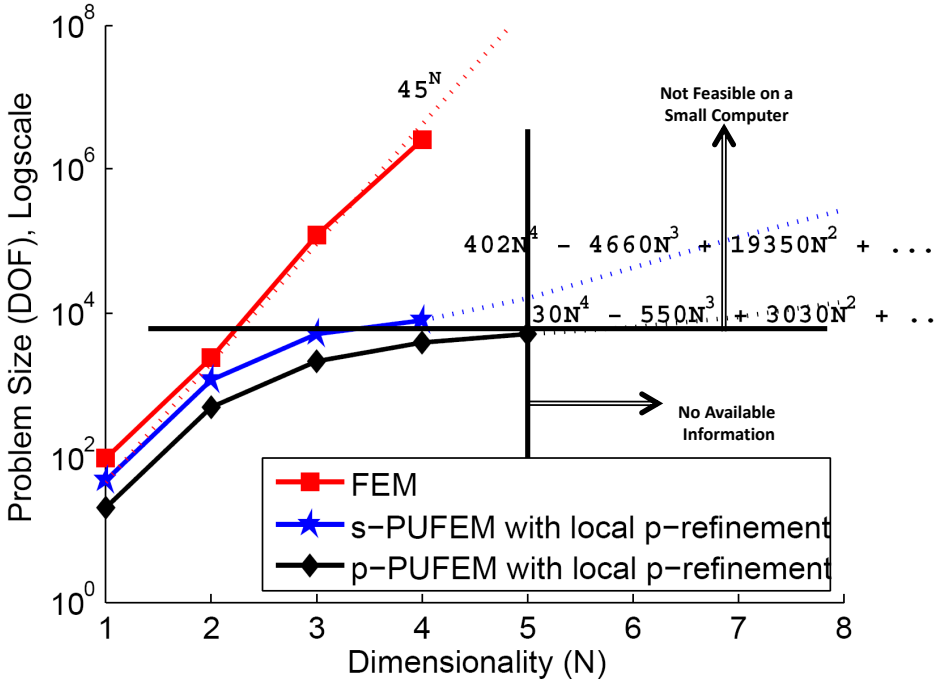


Fig. 23. Two variants of PUFEM (standard- and particle-) in the face of curse of dimensionality.

$\widehat{\mathcal{W}}(t, \mathbf{x})$ (Eq.3.6). It was mentioned in section B3 that in the existing literature, this is almost exclusively done by temporal discretization, for example, by implementing a stabilized Crank-Nicholson algorithm. In this section, an alternate algorithm is presented to develop an analytical solution for the transient FPE response. Coupled with the numerical discretization of section B, we obtain a semianalytical framework for solving transient FPE.

Note that in the absence of parametric uncertainty, Eq.3.11 represents a time-invariant system of linear equations (LTI). As is well known, it is often instructive to study the eigenstructure of an LTI system because this procedure provides valuable insight into the problem. However, the present problem involves several complicated numerical issues like ill-conditioning (due to large penalty parameter) and large size of matrices $\mathbf{M}$ and $\mathbf{K}$ in Eq.3.11. Additionally, the stiffness matrix is neither symmetric nor definite due to the non-normal nature of the FP operator. It is because of these issues that temporal discretization has been widely used despite the linear form of Eq.3.11. At the same time, extensive research has been conducted for such systems, e.g. Krylov space iterative methods[104, 105] and we can benefit from such studies. Thus considering Eq.3.11, we look at the following similarity transformation: $\mathbf{a}' = \mathbf{V}^{-1}\mathbf{a}$, where, $\mathbf{V}$ is a matrix of eigenvectors obtained by solving the generalized eigenvalue problem for the system $(\mathbf{K}, \mathbf{M})$, i.e.:

$$\mathbf{K}\mathbf{v} = \lambda \mathbf{M}\mathbf{v} \tag{3.34}$$

Assuming that Eq.3.34 can be solved, we obtain the familiar decoupled form of Eq.3.11:

$$\dot{\mathbf{a}}'(t) = \Lambda \mathbf{a}'(t) + \mathbf{f}', \tag{3.35}$$

where, Λ is a diagonal matrix containing the generalized eigenvalues of $(\mathbf{K}, \mathbf{M})$, and

$\mathbf{f}' = \mathbf{V}^{-1}\mathbf{M}^{-1}\mathbf{f}$ is the load vector in modal coordinates. Thus, if we can handle the numerical issues involved in computing $\mathbf{V}$ and Λ, transient FPE response of the system can be obtained analytically in modal space. Time history of the individual modal amplitudes, $a_i'(t)$, can be written as:

$$a_i'(t) = \left(a_i'(t_0) + \frac{f_i'}{\lambda_i}\right)\exp(\lambda_i t) - \frac{f_i'}{\lambda_i}. \tag{3.36}$$

Note that it is easy to deal with zero or near zero eigenvalues in the above equation by taking the limit $\lambda_i \to 0$. For example, if $\lambda_k \to 0$, we have:

$$\lim_{\lambda_k \to 0} a_k'(t) = a_k'(t_0) + f_k' t \tag{3.37}$$

It is known that the stationary distribution of FPE (if it exists) is the eigenfunction of the FP operator with zero eigenvalue. Additionally, when it exists, it is unique and globally asymptotically stable. Thus, there can be only one trivial eigenvalue. It is however not possible to recover an eigenvalue exactly equal to zero via numerical computation. Thus, in numerical computation, the stationary mode can be identified as the eigenvalue with minimum absolute magnitude. For the bulk of physically relevant models, this mode appears with a spectral gap, i.e. with separation in magnitude between the static mode eigenvalue and the remaining spectrum, thus making its identification fairly straightforward. In case the spectral gap is negligible (i.e. there appears a cluster of eigenvalues near the static eigenvalue, all close to zero), the mode with the least eigenvalue magnitude can be designated to be the static mode. For this eigenfunction, the modal amplitude should be time invariant, i.e. $a_k'(t) = a_k'(t_0)$, which is not consistent with Eq.3.37. However, note that time variation of the stationary mode is extremely slow (since $f_k \approx 10^{-6}$) and occurs due to enforcement of artificial boundary conditions. Therefore, one can force the identified static mode to be time invariant without causing noticeable error. For most

systems, this is actually not required because time rate of change of the static mode is typically several orders of magnitude slower than remaining modes. For more difficult systems, especially ones with negligible spectral gap, it may be required to force the static mode to be time invariant.

1. Spurious Modes

It was found that given the particular fineness of meshless discretization in use, the solution of the generalized eigenvalue problem of $(\mathbf{K}, \mathbf{M})$ contains spurious (false/ extraneous) modes which do not contribute to improvement of approximation accuracy. In fact, it is postulated that these modes contribute to the difficulties associated with solving Eq.3.11 in its original form. In the following section, evidence is provided to back this claim and it is shown that a small subset of admissible eigenfunctions can be identified which is sufficient to generate the transient FPE response in analytical form, irrespective of initial probability density of the state.

2. Identification and Elimination of Extraneous Modes

The spectrum of the discretized Fokker-Planck operator contains numerous spurious modes that do not correspond to physical reality and arise due to the details of numerical implementation. These extraneous modes can be classified into two groups. The first group (**G1**) emerges as an artifact of the penalty method used for boundary condition enforcement and comprises of eigenfunctions that display severe boundary condition violation. These modes are all either highly stable or unstable (i.e. have large negative or positive real parts), depending on the sign of the penalty parameter. In the former case, dynamics associated with these modes dies out almost instantly, while in the latter, cause the corresponding modal amplitudes to diverge. It is easy to prove that a large enough negative penalty parameter in Eq.3.9 guarantees the

existence of unstable modes. To this end, consider the following lemma:

Lemma III.1 *A matrix* $\mathbf{C} = \mathbf{AB}$*, where* $\mathbf{A}$ *is symmetric positive-definite and* $\mathbf{B}$ *asymmetric indefinite admits positive eigenvalues.*

Proof: Begin by assuming that $\mathbf{C}$ is negative definite, i.e., it admits only negative eigenvalues. Then, we have the following developments:

$$\begin{aligned} \mathbf{C} &= \mathbf{AB} \\ \mathbf{C}^{\mathrm{T}} &= \mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}} = \mathbf{B}^{\mathrm{T}}\mathbf{A} \\ \Rightarrow \mathbf{C} + \mathbf{C}^{\mathrm{T}} &= \mathbf{AB} + \mathbf{B}^{\mathrm{T}}\mathbf{A} \end{aligned} \tag{3.38}$$

Since $\mathbf{C}$ was assumed to be negative definite, we can write $\mathbf{C} + \mathbf{C}^{\mathrm{T}} = -\mathbf{Q}$, where $\mathbf{Q}$ is a symmetric positive-definite matrix. Hence, Eq.3.38 reduces to the following Lyapunov equation in $\mathbf{B}$ with positive-definite symmetric matrices $\mathbf{A}$ and $\mathbf{Q}$:

$$\mathbf{AB} + \mathbf{B}^{\mathrm{T}}\mathbf{A} + \mathbf{Q} = \mathbf{0} \tag{3.39}$$

Since $\mathbf{B}$ satisfies Eq.3.39, it must be Hurwitz, with all negative eigenvalues. This is clearly a contradiction because $\mathbf{B}$ is given to be indefinite. Hence, the assumption that $\mathbf{C}$ is negative definite is falsified. Matrix $\mathbf{C}$ therefore must admit positive eigenvalues. □

The above lemma leads to the desired result, stated as another lemma below:

Lemma III.2 *Given the structure of matrices* $\mathbf{M}$ *and* $\mathbf{K}$ *in Eqs.3.12 and 3.13 respectively, and a large enough negative penalty parameter* α*, the generalized eigenvalue problem* $\mathbf{Kv} = \lambda\mathbf{Mv}$ *admits positive eigenvalues.*

Proof: Proving the above lemma is equivalent to showing that the matrix $\mathbf{L} = \mathbf{M}^{-1}\mathbf{K}$ admits positive eigenvalues. This follows directly from Lemma III.1 because: (1)

Looking at Eq.3.12, we see that $\mathbf{M}$ is symmetric positive definite, implying that so is $\mathbf{M}^{-1}$, and (2) A large negative value of α coupled with the fact that the FP operator is non-normal ensures that $\mathbf{K}$ is asymmetric indefinite. $\square$

Eigenvalues belonging to group **G1** of spurious modes are easy to identify by virtue of their their large magnitude and unstable nature (for a negative penalty parameter). They appear as a distinct band in the spectrum with the largest absolute values and are easy to isolate (for example, see figure on page 85).

The second group (**G2**) of spurious modes comprises of unreliable eigenfunctions that have not converged for the particular spatial discretization in use. These poorly converged eigenfunctions are identified by evaluating the equation error in the eigenvalue problem of the FP operator in function space. In other words, while these functions satisfy the discretized eigenvalue problem exactly, they show large error in the original, un-discretized eigenvalue problem in function space. This is due to the fact that the spatial discretization in use is unable to capture these eigenfunctions sufficiently well. Thus we look at the norm of the following equation error:

$$\|\varepsilon_\phi(\mathbf{x})\| = \|\mathcal{L}_{\mathcal{FP}}(\phi(\mathbf{x})) - \lambda\phi(\mathbf{x})\|_{L_2(\Omega)} \tag{3.40}$$

All unconverged eigenfunctions ($\phi \in \mathbf{G2}$) appear as a distinct band showing high residual error and can be filtered out easily (for example, see on page 86).

Thus eliminating extraneous modes, we are left with a reduced set ($\mathcal{A}$) of "admissible" eigenfunctions that can be used to approximate the solution of transient FPE. Note that these selected modes, i.e. set $\mathcal{A}$, work for all possible initial distributions. As a result, once these modes are identified and isolated as a pre-processing step, FPE can be solved for any given initial distribution in real-time. In this sense, this algorithm is semianalytical - comprising of a numerical part (FPE discretization followed by spectral analysis) and a analytical part (writing the response in terms of

admissible modes). The identification of spurious modes is presented in the results section.

As stated above, the presented analysis holds true for any given initial probability distribution of the state. Before proceeding, we make the assumption that the set of eigenvalues corresponding to admissible modes does not contain any repeated eigenvalues. This is typically true for a numerically computed set of eigenvalues. Then, for a prescribed initial distribution, it may be possible to identify an even smaller set, $\mathcal{B} \subset \mathcal{A}$, that approximates the transient solution and provides further reduction in problem size. The set $\mathcal{B}$, identified by the process of elimination, is such that it approximates the initial state pdf sufficiently well in terms of the equation-error of FPE. In other words, the eigenfunctions of set $\mathcal{B}$ are chosen such that equation error in FPE is less than a prescribed tolerance at initial time. It can then be shown that equation-error at all subsequent times remains bounded by the initial equation-error. Therefore, the identified minimal admissible subset $\mathcal{B}$ is sufficient to generate the FPE response for the particular initial distribution for all time. This result can be stated as the following theorem:

Theorem III.1 *Let* $\mathcal{A} = \{\phi_i : Real(\lambda_i) < 0, \|\varepsilon_i\| = \|\mathcal{L}_{\mathcal{FP}}(\phi_i) - \lambda_i\phi_i\|_{L_2(\Omega)} < \delta\}$ *be the set of stable admissible eigenfunctions assumed to contain no repeated eigenvalues. Define equation error in FPE as* $e(t) = \left\| \frac{\partial}{\partial t}\widehat{\mathcal{W}}(t,\mathbf{x}) - \mathcal{L}_{\mathcal{FP}}(\widehat{\mathcal{W}}(t,\mathbf{x})) \right\|_{L_2(\Omega)}$. *If a subset* $\mathcal{B} = \{\phi_i^\star : \phi_i^\star \in \mathcal{A}, i = 1, \ldots, N_{\mathcal{B}}\}$ *of* $\mathcal{A}$ *can be identified such that the initial equation-error is within a specified tolerance, i.e.* $e(t_0) < \epsilon$, *then the equation-error at all subsequent times is bounded by the initial equation-error, i.e.* $e(t) < \epsilon$.

Proof: We assume that $\mathcal{W}_\Gamma = 0$, such that $\mathbf{f} = \mathbf{0}$. This assumption is not restrictive because the load vector is ideally zero, and in implementation turns out to be a small

value ($\|f\| \approx 10^{-6}$). Now, the equation-error in the FPE at time t is given by:

$$e(t) = \left\| \frac{\partial}{\partial t}\widehat{\mathcal{W}}(t,\mathbf{x}) - \mathcal{L}_{\mathcal{FP}}(\widehat{\mathcal{W}}(t,\mathbf{x})) \right\|_{L_2(\Omega)} \tag{3.41}$$

Using the proposed subset $\mathcal{B}$ of admissible eigenfunctions, we have the following expression for the instantaneous pdf (following the assumption that exist no repeated eigenvalues): $\widehat{\mathcal{W}}(t,\mathbf{x}) = \sum_{i=1}^{\text{card}(\mathcal{B})=N_{\mathcal{B}}} a_i'^{\star}(t)\phi_i^{\star}(\mathbf{x})$. Substituting this expression in Eq.3.41 we have the following development:

$$\begin{aligned}
e(t) &= \left\| \sum_{i=1}^{N_{\mathcal{B}}} \left[\frac{\partial}{\partial t}(a_i'^{\star}(t)\phi_i^{\star}(\mathbf{x})) - \mathcal{L}_{\mathcal{FP}}(a_i'^{\star}(t)\phi_i^{\star}(\mathbf{x})) \right] \right\|_{L_2(\Omega)} \\
&= \left\| \sum_{i=1}^{N_{\mathcal{B}}} [\dot{a}_i'^{\star}(t)\phi_i^{\star}(\mathbf{x}) - a_i'^{\star}(t)\mathcal{L}_{\mathcal{FP}}(\phi_i^{\star}(\mathbf{x}))] \right\|_{L_2(\Omega)} \\
&\underset{Eq.3.40}{=} \left\| \sum_{i=1}^{N_{\mathcal{B}}} [\dot{a}_i'^{\star}(t)\phi_i^{\star}(\mathbf{x}) - a_i'^{\star}(t)\{\lambda_i^{\star}\phi_i^{\star} + \varepsilon_i^{\star}(\mathbf{x})\}] \right\|_{L_2(\Omega)} \\
&= \left\| \sum_{i=1}^{N_{\mathcal{B}}} [\{\dot{a}_i'^{\star}(t) - \lambda_i^{\star}a_i'^{\star}(t)\}\phi_i^{\star}(\mathbf{x}) - a_i'^{\star}(t)\varepsilon_i^{\star}(\mathbf{x})] \right\|_{L_2(\Omega)} \qquad (3.42) \\
\Rightarrow e(t) &\underset{\substack{Eq.3.35,\\ \mathbf{f}=\mathbf{0}}}{=} \left\| \sum_{i=1}^{N_{\mathcal{B}}} a_i'^{\star}(t)\varepsilon_i^{\star}(\mathbf{x}) \right\|_{L_2(\Omega)} \qquad (3.43)
\end{aligned}$$

Noting that the eigenfunctions can be normalized to have unit $L_2(\Omega)$ norm, we get

$$e(t) \leq \sqrt{\sum_{i=1}^{N_{\mathcal{B}}} |a_i^{\star}(t)|^2} \tag{3.44}$$

The time history of modal amplitudes $a_i'^{\star}(t)$ (in general complex valued) is given by Eq.3.36, from which it is easy to show that $|a(t)| \leq |a(t_0)|$. We thus conclude from Eq.3.44 that $e(t) \leq e(t_0) \leq \epsilon \ \forall t \geq 0$. $\square$

It is important to note that the above theorem holds for $\mathcal{B} = \mathcal{A}$, and thus is not specific to particular initial distributions. In the special case where systems admit a stationary distribution, only one mode (corresponding to $\lambda = 0$, "stationary mode")

has non-zero amplitude as $t \to \infty$. A corollary to the above theorem is that equation-error is bounded by a monotonically decaying envelope which has greatest width at the initial time. This follows directly from Eq.3.43 and can be stated as follows:

Corollary III.1 *The equation error, $e(t)$ in FPE resulting from an admissible set $\mathcal{A}$ of eigenfunctions ($e(t)$ and $\mathcal{A}$ defined in Theorem III.1) is bounded by an exponentially decaying envelope which has its greatest width at the initial time:*

$$e(t) \leq e(t_0)\exp(\mu t) + e_s, \tag{3.45}$$

where, subscript 's' corresponds to the static mode, $e_s = ||a_s(t)\varepsilon_s(\mathbf{x})|| = ||a_s(t_0)\varepsilon_s(\mathbf{x})||$ is the equation-error in FPE resulting only from the static mode (stationary solution), $\mu = \max\{Real(\lambda_i),\ i \neq s\}$, and λ_i correspond to eigenfunctions $\phi_i \in \mathcal{A}$.

Proof: The proof is simple and follows directly from substituting the time variation of admissible modal coefficients in Eq.3.43:

$$\begin{aligned}
e(t) &\underset{\substack{Eq.3.35,\\ \mathbf{f}=\mathbf{0}}}{=} \left|\left| a_s'(t)\varepsilon_s(\mathbf{x}) + \sum_{\substack{i=1\\ i\neq s}}^{N_{\mathcal{A}}} a_i'(t_0)\exp(\lambda_i t)\varepsilon_i(\mathbf{x}) \right|\right|_{L_2(\Omega)} \\
&\leq \left|\left| \sum_{\substack{i=1\\ i\neq s}}^{N_{\mathcal{A}}} a_i'(t_0)\exp(\lambda_i t)\varepsilon_i(\mathbf{x}) \right|\right|_{L_2(\Omega)} + ||a_s'(t)\varepsilon_s(\mathbf{x})|| \\
&\underset{\substack{\mu=\max\{\mathrm{Real}(\lambda_i)\};\\ i\neq s;\ \text{and}\\ a_s'(t)=a_s'(t_0)}}{\leq} \left|\left| \sum_{i=1}^{N_{\mathcal{A}}} a_i'(t_0)\varepsilon_i(\mathbf{x}) \right|\right|_{L_2(\Omega)} \exp(\mu t) + e_s \\
&= e(t_0)\exp(\mu t) + e_s
\end{aligned} \tag{3.46}$$

Since λ_i correspond to $\phi_i \in \mathcal{A}$, $\mathrm{Real}(\lambda_i) < 0$. Thus, $\mu < 0$. $\square$

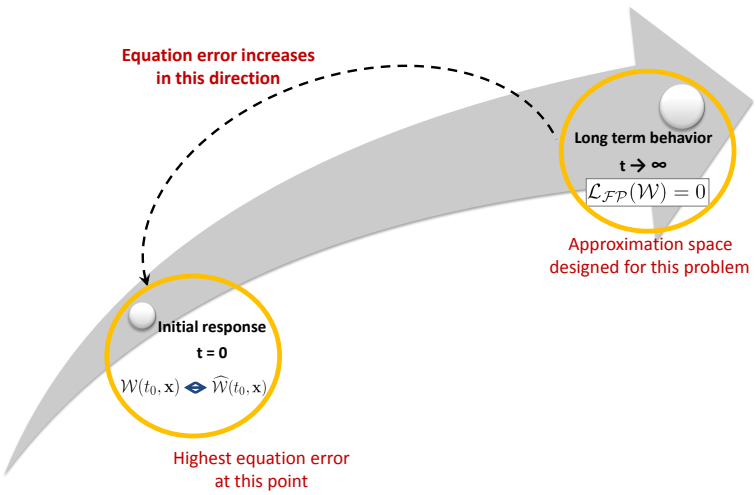


Fig. 24. Illustration of Theorem III.1.

3. Benefits of Spectral Analysis

The most important advantage of transformation to modal coordinates is that it allows for a robust solution of FPE in near-real time, given that the eigenvalue analysis is performed offline. The pre-processing step of identifying the set of admissible eigenfunctions makes the approach independent of initial state distribution. The use of eigenfunctions ensures that solution accuracy (equation error) improves with time and the approximation is at least as good as the approximation of the initial distribution in terms of equation-error in FPE. This is a significant step towards obtaining high-fidelity solutions of FPE. The idea of decreasing equation error is illustrated in Fig.24, which shows that modal basis functions are obtained essentially by studying long term characteristics of the FP operator. Therefore, these modes are intrinsically suited to approximate steady state response. On the other hand, initial conditions for modal coefficients are obtained by a standard least squares procedure

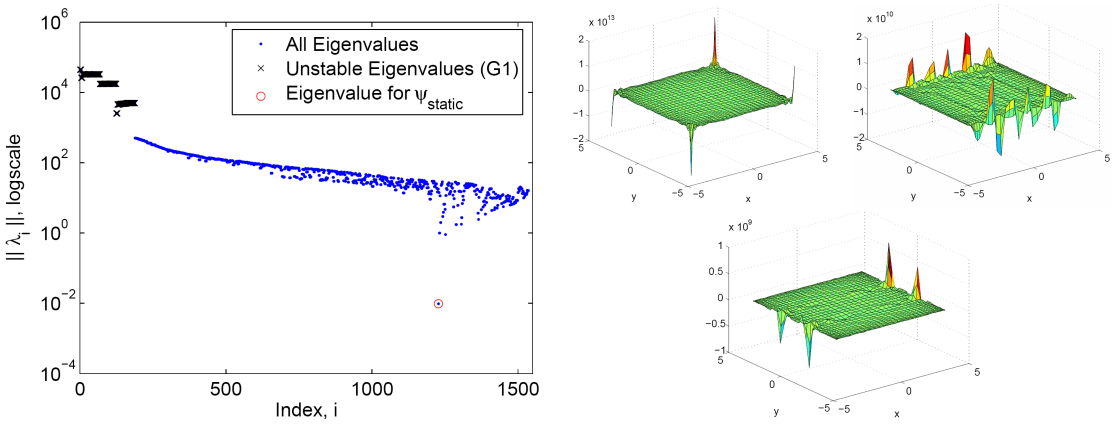
of function approximation of the initial density function, for which the reduced set of admissible eigenfunctions may not be the best choice of basis functions. Also note that the process of removing spurious modes leads to a significant reduction in size of the discretized problem, which is useful for high dimensional applications.

E. Results for Transient FPE

In this section, we present numerical results for various dynamical systems residing in 2 and 3 dimensions using the above outlined algorithm.

a. Dynamic System 1: Example in Two Dimensions

For the first example, we consider again the dynamical system of Eq.3.30. This system was studied by Muscolino et al. in Ref. [66], wherein global C-type Gram-Charlier expansions were used to obtain transient FPE response. In the current work, meshless sPUFEM discretization was implemented on a 14×14 grid. Local p-refinement was utilized to endow boundary nodes with constant basis functions and interior nodes with quadratic-polynomial basis, leading to a total of 916 undetermined coefficients in the approximation. The results of modal analysis are shown in Figs.25 and 26. Recall that the spurious modes of 'group 1' (**G1**) display severe boundary condition violation and those belonging to 'group 2' (**G2**) exhibit large equation-error in the eigenvalue problem of the FP-operator in function space. The modes belonging to group 1 are clearly distinguishable in Fig.25(a) by the large magnitude of their eigenvalues. Fig.25(b) shows some examples of these eigenfunctions and their exaggerated violation of boundary conditions is clearly visible. Eigenfunctions of group 1 constitute 25% of the total number of modes for this system. Note that the eigenvalue with the smallest magnitude is isolated from the rest of the spectrum (i.e. there exists



(a) Unstable eigenvalues (of Group 1) have the largest magnitudes and are easy to identify.

(b) Examples of group 1 spurious modes.

Fig. 25. Identification of spurious modes of the discretized FP operator: group **G1**.

a spectral gap) and the corresponding eigenfunction is the stationary distribution of the dynamical system under consideration. Note however that the stationary eigenfunction may not necessarily appear as an isolated mode for all dynamical systems, although most physical systems exhibit this behavior.

Recall that if a stationary solution exists, it is known to be unique, meaning it is globally asymptotically stable. As already mentioned, it is also often the case that the stationary mode appears in isolation with a spectral gap (as is evident from Fig.25(a)). Separation of the stationary solution from remaining modes endows exponential stability to the system, i.e. faster than asymptotic stability, the extent of which depends on the actual system under consideration. For systems where spectral gap is small, (i.e. there exists a cluster/continuum of modes in the neighborhood of the static mode) it is difficult to identify the stationary solution, and the mode with least eigenvalue can be assigned as the stationary mode. Note that despite this difficulty, the stationary solution, when it exists is unique and asymptotically stable.

The modes belonging to group 2 are identified in Fig.26, which shows two distinct bands of equation-error in the functional eigenvalue problem of FPE. The uniformly large magnitude of errors on this plot (see the scale on the y-axis) is due to the noise introduced in computing the second derivative of the approximation. The set $\mathcal{A}$ of admissible modes is shown in Fig.26 encircled with an ellipse. The smallest eigenvalue (corresponding to the unique stationary distribution) is highlighted and emerges as the eigenfunction with the best convergence. Fig.27 shows a possible initial distribution, namely a Gaussian pdf with mean at the origin. For this initial distribution, a much smaller set $\mathcal{B}$ was found, (shown with circles in Fig.26) which approximates the initial condition sufficiently well and contains only 20% of the total number of degrees of freedom originally used. This is a significant order reduction. Fig.27 provides a verification of the redundant nature of all eliminated modes ($\in \mathcal{B}^C$). In this figure, all initial modal amplitudes $a_i^\star(t_0)$ have been shown in the complex plane. It is clearly visible that initial amplitudes of all eliminated modes is negligible. In other words, they do not participate in the approximation of the initial distribution, which is a necessary condition for their redundant nature.

Furthermore, following Eq.3.43, we know that $|a_i^\star(t)| \leq |a_i^\star(0)|$, which in turn implies that these modes do not participate in the approximation at any later time and are therefore redundant.

Thus, we see that a significant reduction in the order of approximation is achievable for the transient problem. Fig.28(a) shows the evolution of modal amplitudes of all modes. Note that all transient modes decay to zero amplitude, while the amplitude of the static mode attains a non zero, (nearly) steady state value. We mention that the eigenvalue of the stationary mode is computed to have a very small real part due to numerical computations and hence has a finite decay rate. However, the decay rate of the static mode is several orders of magnitude lower than all other transient

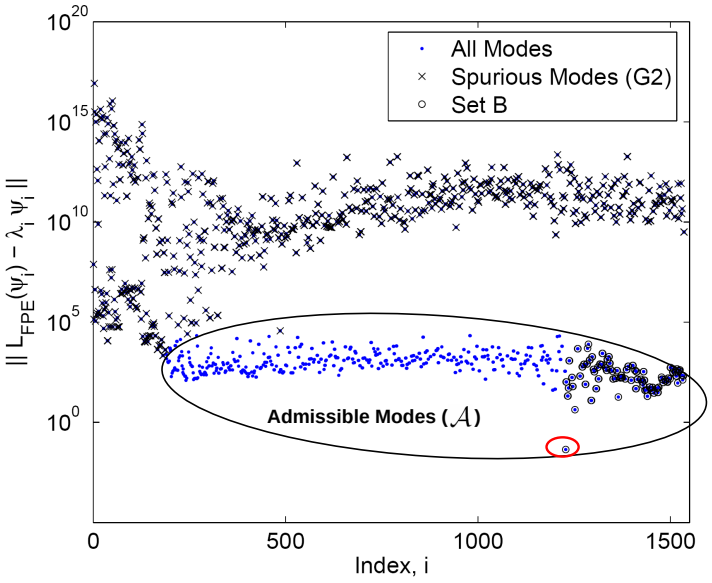


Fig. 26. Identification of spurious modes belonging to group **G2** (unconverged eigenfunctions).

modes and we obtain the stationary behavior for all practical purposes. This fact is visible in Fig.28(a). This result also verifies the discussion surrounding Eq.3.37 ($f_i \approx 10^{-6}$). Fig.28(b) confirms the exponential decrease in equation-error in accordance with Corollary III.1. The equation-error curve flattens out to a steady state value, including only the contribution from the static mode. Fig.29 shows time evolution of the initial distribution obtained from analytical integration of coefficients of the remaining modes. Several other chosen initial distributions led to the same stationary distribution, as expected.

b. Dynamic System 2: 2D Nonlinear Oscillator with Multiplicative Noise

Consider the following two-state nonlinear oscillator studied by Wojtkiewicz et al.[67]:

$$\ddot{x} + 2\eta\dot{x} + x - \epsilon x^3 = x\mathcal{G}_1(t) + \mathcal{G}_1(t) \tag{3.47}$$

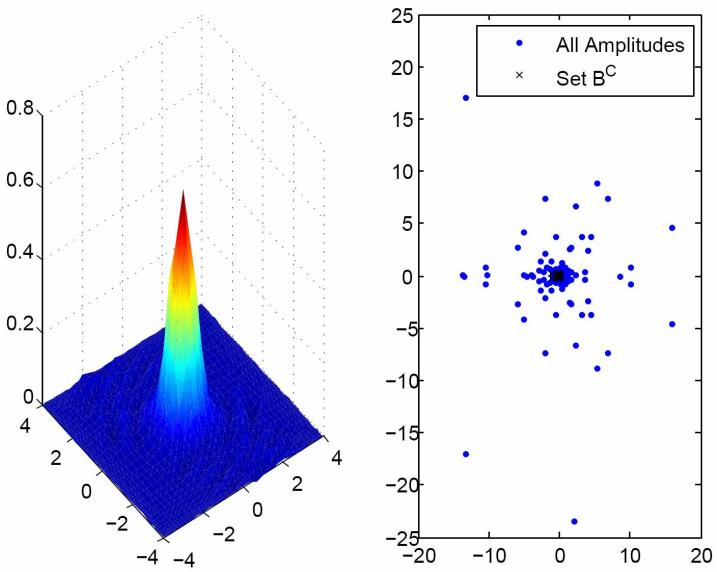
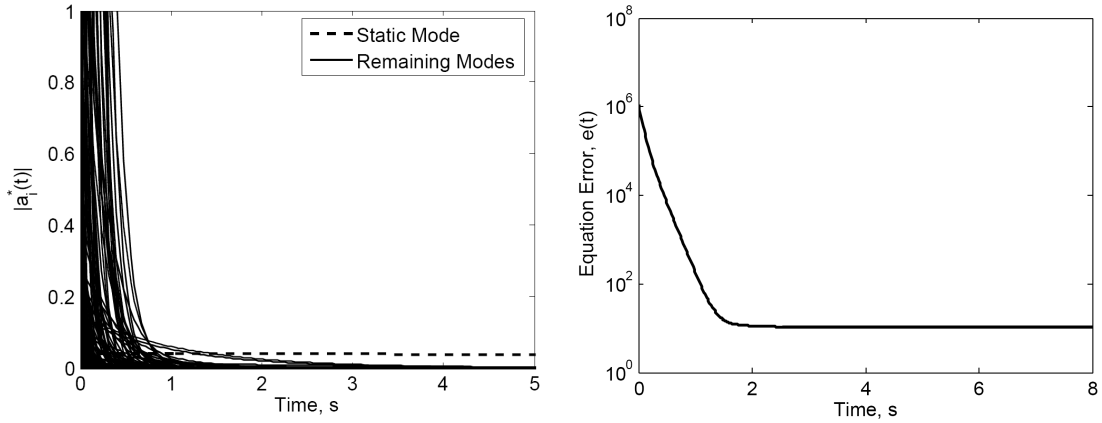


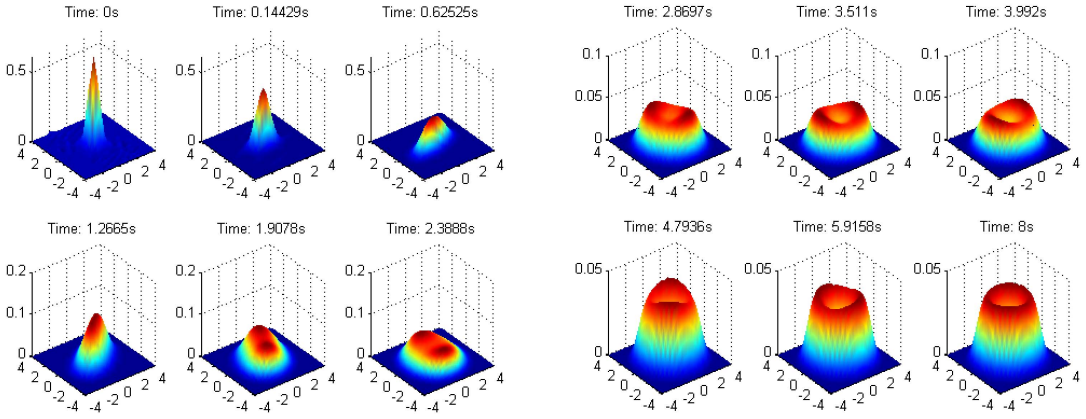
Fig. 27. Initial amplitudes of all modes $\in \mathcal{B}^{\mathrm{C}}$ are almost trivial for the shown initial distribution.



(a) Time history of individual modal coefficients.

(b) Time history of equation error, $e(t)$.

Fig. 28. Time history of modal coefficients and verification of Theorem III.1.



(a) Time evolution of *pdf*: Time: $0.0s$ – $2.4s$.

(b) Time evolution of *pdf*: Time: $2.8s$ – $8.0s$.

Fig. 29. Solution of transient FPE for a 2-state nonlinear oscillator starting with a Gaussian initial condition.

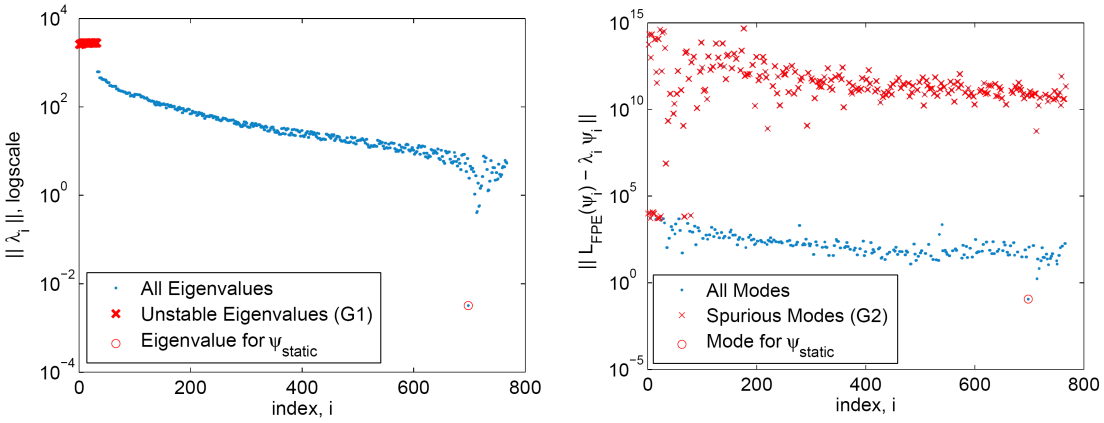
The above system represents a damped soft-spring Duffing oscillator with state multiplied noise. The noise driving the above system comprises of two independent components, and can be written as: (refer to Eq.2.1)

$$d\mathbf{B}(t) = \begin{bmatrix} d\mathcal{G}_1 \\ d\mathcal{G}_2 \end{bmatrix}, \mathbf{Q} = \begin{bmatrix} 2D_{11} & 0 \\ 0 & 2D_{22} \end{bmatrix} \tag{3.48}$$

where, D_{11} and D_{22} are intensities of the individual components, $d\mathcal{G}_1$ and $d\mathcal{G}_1$ respectively. We will consider the case of $D_{11} = 0.24$ ("high multiplicative excitation case" considered in Ref. [67]). Values of the other parameters used are: $\eta = 0.2$, $\epsilon = 0.1$ and $D_{22} = 0.4$ [67]. FPE for the above system can be written as:

$$\frac{\partial p}{\partial t} = \frac{\partial}{\partial x_1}[(2\eta x_2 - x_1 + \epsilon x_1^3)p] - x_2\frac{\partial p}{\partial x_1} + (D_{11}x_1^2 + D_{22})\frac{\partial^2 p}{\partial x_2^2} \tag{3.49}$$

The above problem was solved in Wojtkiewicz et al. [67] using 100 finite elements per dimension (problem size of $100^2 = 10000$). Computation time (including integration

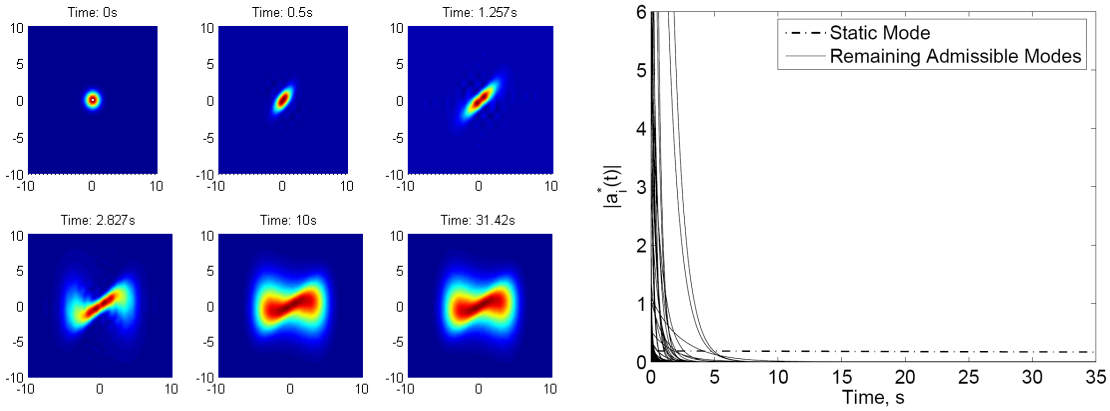


(a) Spurious modes of group **G1**.

(b) Spurious modes belonging to group **G2** identified via large equation error.

Fig. 30. Spectral analysis for the Duffing oscillator with state-multiplied noise.

of the discretized FPE) was reported to be 17 minutes on a CRAY Y-MP/464 supercomputer. In this example, sPUFEM meshless discretization with local p-refinement was implemented on a 9×9 grid, with constant basis functions endowed to the boundary nodes. All interior nodes were endowed with quartic shape functions, thus leading to a problem size of 767 degrees of freedom [106]. The computation time for the variational formulation and modal analysis detailed above was 2 minutes on a portable workstation with a 1.86 GHz Pentium M processor and 1 GB RAM. Results of modal analysis are shown in Fig.30. Note that only about 48% of the 767 DOFs are admissible, thus further reducing the problem size to a mere 366 DOFs. The time evolution of an initial probability distribution has been shown in Fig.31 alongside the evolution of modal coefficients of the admissible modes. Clearly, all transient modes die out, leaving the static mode as the only non-zero mode in long term.



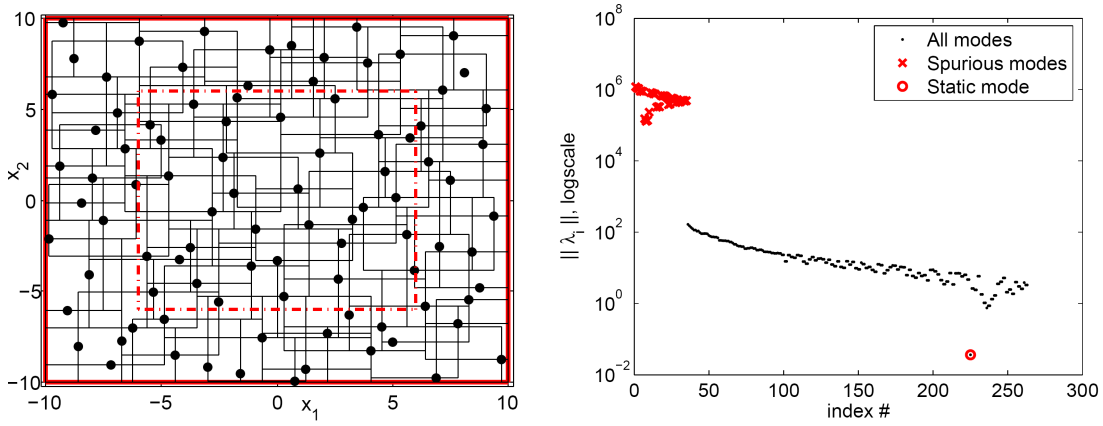
(a) Time evolution of *pdf*: Time: $0.0s - 31.42s$.

(b) Time histories of modal coefficients of the admissible eigenfunctions.

Fig. 31. Solution to FPE for the Duffing oscillator with state-multiplied noise starting with a Gaussian initial condition.

c. Approximation for Above System with pPUFEM

In this section, we consider the same 2-state duffing oscillator with high multiplicative noise and approach the problem with pPUFEM, aiming to achieve a smaller problem size. Figure 32(a) shows discretization of the solution domain $\Omega = [-10, 10] \otimes [-10, 10]$ using 93 nodes. These nodes were placed using the two dimensional Halton sequence; i.e. no pattern or a-priori knowledge was used for node placement. All nodes inside the bold broken box (34 of them) were assigned a complete set of quadratic polynomial basis functions and all nodes lying outside were assigned only the constant basis, i.e. $\Psi(\mathbf{x}) \equiv 1$. This was based on a-priori knowledge that the bulk of the pdf lies inside the box drawn with broken lines. Hence, the total number of degrees of freedom in this approximation is: $\mathcal{D}_{pPU} = 34 \times 6 + (93 - 34) \times 1 = 263$. The resulting approximation, following integration of variational equations and modal analysis is shown in Figs.32(b) and 33. Fig.32(b) shows the spectrum of the discretized



(a) 93 quasi uniformly random nodes used to discretize the solution domain for system in Eq.3.49.

(b) Spectrum of the discretized FP operator showing spurious modes and the isolated stationary mode for the system in Eq.3.49.

Fig. 32. pPUFEM discretization details for system 2.

FP operator. The spurious modes mentioned in Section D2 are marked out and their elimination leaves behind a mere 228 DOFs. The eigenfunction corresponding to the smallest eigenvalue (marked with a circle) is nothing but the stationary solution of FPE for this system. Note that there appears a sizeable spectral gap between the static mode and transient modes.

Fig.33 shows nine snapshots of the time evolution of the initial pdf for the system, which is assumed to be Gaussian, centered at $(1, 1)$. It is visible that the quality of approximation improves with time, as stated in Theorem III.1. This system was considered by Kumar et al.[107] using the standard-PUFEM and comparable results were obtained using about 700 DOFs. On the other hand, standard finite element method (FEM) requires 10,000 DOFs to deliver comparable results[67].

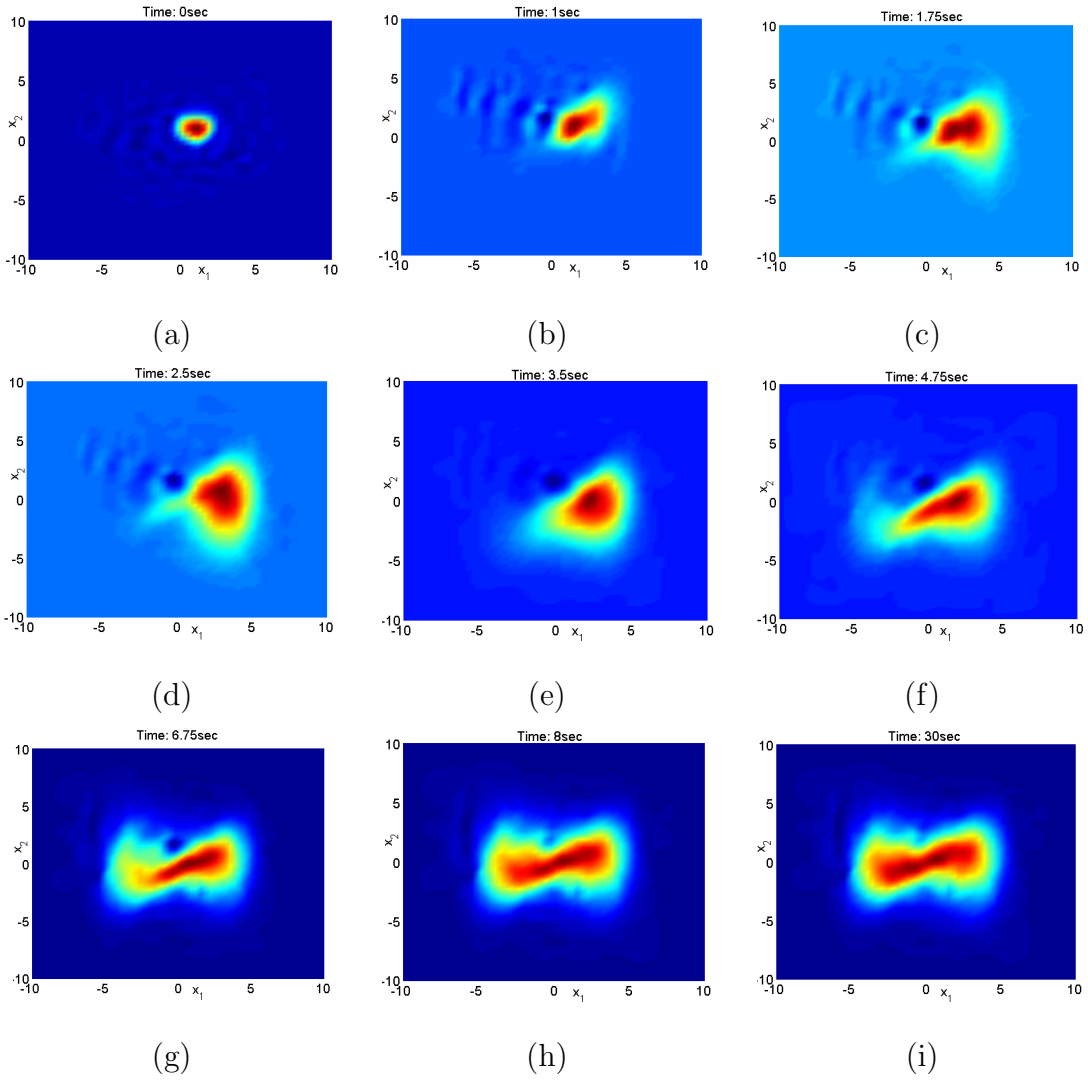


Fig. 33. Time evolution of an initial Gaussian pdf using pPUFEM approach for the dynamical system in Eq.3.49.

d. Dynamic System 3: 3D Nonlinear Oscillator (Lorenz Attractor)

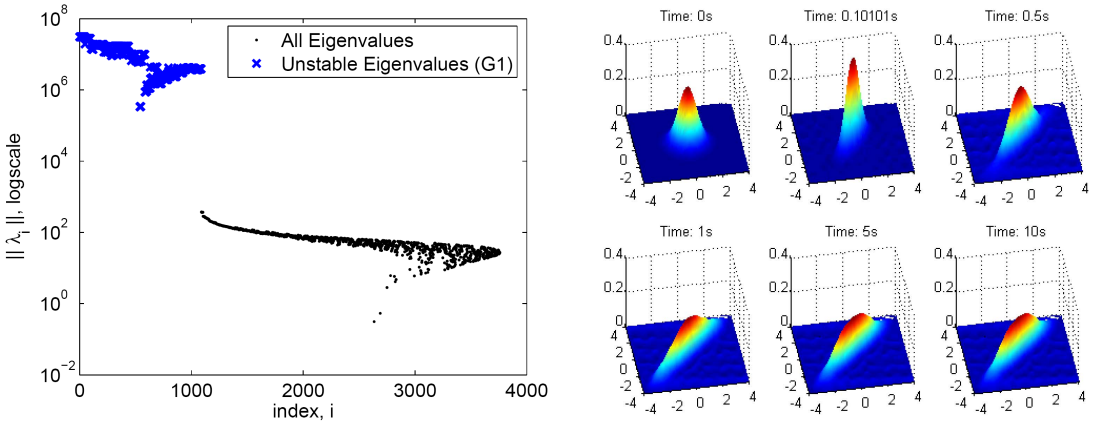
Consider now the following nonlinear system with three dimensional state space:

$$\begin{aligned} \dot{x} &= \sigma(y-x) + \zeta_1(t) \\ \dot{y} &= x(\rho - z) - y + \zeta_2(t) \\ \dot{z} &= xy - \beta z + \zeta_3(t) \end{aligned} \tag{3.50}$$

The above represents a noise-driven lorenz attractor. Numerical values for the various parameters appearing above are: $\sigma = 10, \rho = 1; \beta = 8/3$ and Q (noise intensity) $= 2$. This system was also studied in the previous section for its stationary distribution. We use the same discretization as in the first part, i.e. a $6 \times 6 \times 6$ nodal grid with the boundary nodes endowed with quadratic polynomials and interior nodes with quartic polynomials (corresponding problem size = 3760 DOFs). Fig. 34(a) shows the modal analysis for this system. About 28% of the total modes belong to group **G1**, which amounts to 1088 modes. Fig. 34(b) shows the time-evolution of the xy-marginal distribution. The (near) stationary distribution obtained from this analysis matches the stationary distribution obtained in the previous section to machine precision.

e. Dynamic System 4: 4-State Nonlinear Oscillator

To conclude this chapter, let us study the transient FPE response of the coupled nonlinear vibration isolation suspension model of Eq.3.31, which was considered in section C for its steady state response. A somewhat bigger domain of $\otimes_{i=1}^{4}[-8, 8]$ was considered and 350 locally enriched nodes were used under the pPUFEM framework, leading to a problem size of 3221 DOFs. The resulting spectrum and spurious modes are shown in Fig.35. The initial state density was assumed to be a Gaussian function centered at $[4, 4, 4, 4]^{\mathrm{T}}$ with unit covariance matrix. Time evolution of the $x_1 - x_2$



(a) Eigenvalues for the three state noise driven Lorenz attractor.

(b) Time evolution of the xy-marginal distribution of the Lorenz attractor.

Fig. 34. Numerical results for the transient Fokker-Planck Equation of the noise driven Lorenz attractor.

marginal probability density is shown in Fig.36. The system evolves at a fast rate in the initial stages of propagation shown in Fig.36(a), as bimodal behavior in the marginal density emerges. With the available computing resources, it is not possible to solve generalized eigenvalue problems of size much greater than that considered in this example. It is clear that even with a relatively few number of modes ($\sim$ 2000 admissible modes), the overall system behavior is captured remarkably well.

This is confirmed by comparing the shown evolution with Monte Carlo propagation of a few sample points (75 in all), shown overlaid on the surface plots. Note that there is considerable discrepancy between point propagation and pdf propagation in the time range 0.1 s to 2.0 s. This is due to the fact that process noise was severely attenuated for Monte Carlo propagation (intensity scaled down by 20 times) because it takes on the order of hours to integrate a single sample through nonlinear dynamics of Eq.3.31 if the same noise level used for FPE analysis is used in Monte Carlo simulations. As a result, in the absence of the actual dissipation level, the point cloud

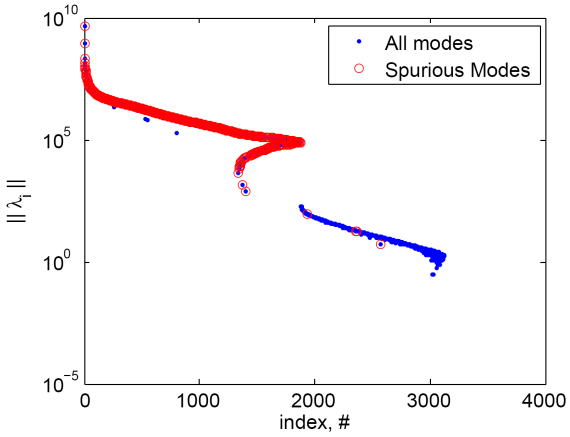
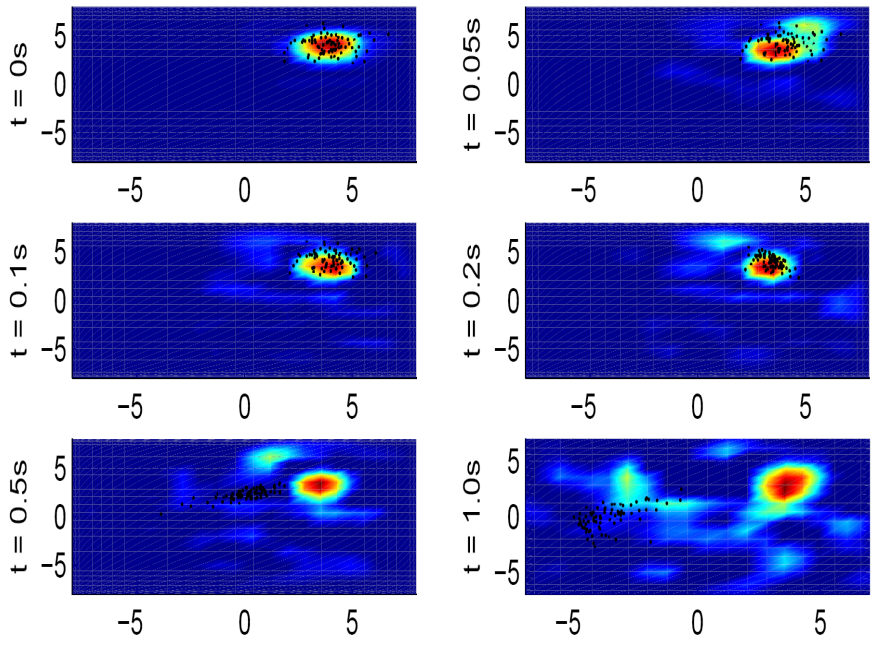


Fig. 35. Spectrum of the discretized FP operator for the four-state nonlinear vibration isolation suspension model.

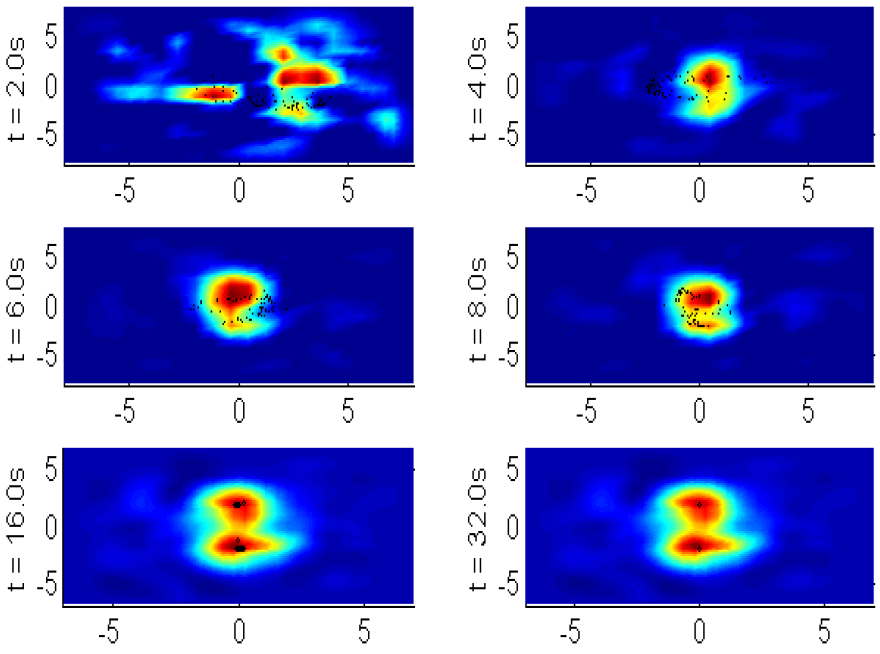
"drifts faster" than it should. Since long term behavior is globally asymptotically stable, the point cloud propagation and pdf propagation both eventually match up to the same probability distribution, albeit with mismatch in dissipation levels, clearly visible from Fig.36(b).

F. Summary

In this chapter, basic tools for a robust solution methodology for FPE have been developed. Two algorithms based on the standard and particle versions of PUFEM have been presented to discretize the Fokker-Planck operator while handling the associated curse of dimensionality. Spatial discretization is coupled with modal analysis and spurious mode rejection to obtain transient FPE response in real-time, independently and for all possible initial state probability distributions. In this section, we summarize the salient features of the tools developed in this chapter by comparing them to other existing techniques for solving FPE, like the global Galerkin method (GM), conventional FEM and the other meshless methods based on moving least



(a) Time evolution of *pdf*: Time: $0.0s - 1.0s$.



(b) Time evolution of *pdf*: Time: $2.0s - 32.0s$.

Fig. 36. Evolution of the $x_1 - x_2$ marginal of a 4-state vibration isolation suspension model with a Gaussian initial condition.

squares, like the one used in used in Kumar et al.[76] (MRMM):

- **Shape function selection:** It has been demonstrated through several examples that PUFEM (especially the particle version) offers great flexibility in the selection of independently chosen local approximation spaces (local p-refinement). This feature is also the primary cause behind weakening of the curse of dimensionality. Shape functions are constructed via the simple step of multiplying basis functions with PU pasting functions. In most other meshless methods, shape functions are constructed using data fitting algorithms like MLS. Consequently, while it is possible to use non-polynomial functions in the approximation space, it is a relatively difficult task to use different basis functions in different regions of the solution domain. Conventional FEM typically uses only polynomial shape functions in the approximation space, the order of which is determined by the shape of the finite element. Finally, there is no scope for local basis enrichment in global methods because of the nature of their formulation.

- **Convergence characteristics:** Although a very small body of theoretical results exist in this regard, convergence characteristics of (s,p)-PUFEM is expected to be superior to that of FEM, especially with the use of special (non-polynomial) shape functions. From our experience in the current application, we can conclude that convergence characteristics of PUFEM is better than that of MRMM, using the same set of basis functions (see Figs.11(f) and 12(c)). This could partially be due to the absence interpolation errors in PUFEM, whereas in MRMM additional errors are introduced due to interpolation required to find the solution at points other than the nodes used for discretization.

- **Computational load:** If we break up computational load into three stages - preprocessing (grid generation and approximation space construction), integration (evaluation of weak form integrals) and postprocessing (solving discretized FPE to obtain transient and stationary response), we get the following relative ordering: PUFEM and MRMM rank above FEM in the preprocessing stage because of minimal bookkeeping required in the former two methods. PUFEM and FEM are faster than most other meshless methods in the integration stage because the latter require solution to an MLS problem for every quadrature point used for numerical integration. In the currently developed algorithms, quasi-random integration schemes have been used to effectively integrate in high dimensions. PUFEM ranks above both FEM and other meshless methods in the postprocessing stage because it provides a functional form of the approximation; i.e. no interpolation is required to construct the solution at any given point in the domain. In addition, almost all other methods (global and local) apply temporal discretization to obtain transient FPE response. In the current work, modal analysis was utilized to glean admissible eigenfunctions of the discretized FP operator that provide an analytical solution of the transient problem. By virtue of local p-refinement, the effective memory usage is also smaller in PUFEM, especially pPUFEM. This has been amply demonstrated in the examples showing exaggerated benefits in problem size over other local techniques.

- **Application to high dimensional problems**: PUFEM and MRMM rank above conventional FEM in implementation to higher dimensional problems because mesh generation in 3 and higher dimensions is still not practical. Comparing MRMM and PUFEM in this respect, PUFEM has definite advantage

because of its simpler algorithm structure and much smaller time of execution. The easy implementation of local p-refinement makes the current approach extremely attractive for use in high dimensional nonlinear problems, as shown in the presented examples.

In summary, algorithms presented in this chapter compare favorably against existing techniques in terms of flexibility, ease of implementation, extension to higher dimensions and the ability to generate accurate approximations with small problem size. The method is robust in the sense that it does not involve several tuning parameters to achieve successful results. Modal analysis further improves robustness by retaining only useful eigenfunctions for generating the transient approximation, in the process also opening doors for nonlinear filtering with FPE, which is the subject matter of chapter VI.

CHAPTER IV

RECURSIVE SOLUTION REFINEMENT AND DOMAIN TRACKING

A. Introduction

In this chapter, an iterative approach for solution refinement of the stationary Fokker-Planck equation is presented. The recursive use of a modified norm induced on the solution domain by the most recent estimate of the stationary probability density function is shown to significantly improve the accuracy of the approximation over the standard L_2-norm based Galerkin error projection. The modified norm is argued to be naturally suited to the problem and hence preferable over the standard L_2-norm because the former requires substantially fewer degrees of freedom for the same order of approximation accuracy, making it immediately attractive for Fokker-Planck equation in higher dimensions. Additionally, it is shown that the modified norm can be utilized to progress through a homotopy of dynamical systems, $\mathfrak{D}_p$, in order to determine the domain of stationary distribution of a nonlinear system of interest, (corresponding to $p = 1$) by starting with a known dynamical system (corresponding to $p = 0$) and working upwards. The partition of unity finite element method is used for numerical implementation.

B. Solution Refinement

In chapter III, a PUFEM based variational solution methodology was developed for FPE. The standard Galerkin error projection approach was used, in which the test function space $\mathfrak{V}_{\mathcal{D}}$ is chosen to be the same as the trial space, $\mathfrak{U}_{\mathcal{D}}$. This results in an approximation of size $\mathcal{D}$ having a particular level of accuracy. If nothing else is changed, the only way to improve accuracy is to increase the number of DOFs,

$\mathcal{D}$. This section presents a methodology for improving approximation accuracy while keeping the problem sized fixed, by means of suitable modification of the test function space, $\mathfrak{V}_{\mathcal{D}}$. The idea as mentioned in the introduction, is to weight the test functions with the most recent approximation of the solution. This in essence modifies the norm under which error projection is performed and leads to a recursive scheme that progressively improves solution accuracy for a fixed value of $\mathcal{D}$.

1. Modification of the L_2 Inner Product

As mentioned above, Galerkin variational formulation of FPE employs the traditional L_2 inner product for error projection. It is claimed that there exists a natural measure which defines a modified inner product and which can be used to our advantage to obtain accurate approximations with a small number of degrees of freedom. This natural measure is characterized by the true solution of FPE. Besides redefining the inner product, it also implicitly defines the domain of solution by providing high weightage to only the significant regions of pdf (regions with most probability mass). The obvious problem with using the actual solution as a weight is that it is unknown. Therefore, an iterative scheme is developed in which the most recent approximation of the true solution is used to weight the L_2 inner product. As the first step, closeness of this approach to the optimal approximation of pdf obtained from normal equations of the Hilbert projection theorem is shown. To reiterate, the current problem of interest is restricted to the stationary FPE. We begin with the following assumptions on the approximation space $\mathfrak{U}_{\mathcal{D}}$ and the initial estimate $\widehat{\mathcal{W}_0}$ of the true solution $\mathcal{W}^\star$:

Assumption IV.1 *The true solution is exactly approximable by the trial space $\mathfrak{U}_{\mathcal{D}}$ using the Hilbert projection theorem, i.e. there exist $\{a_i\}$, $i = 1, 2, \ldots, \mathcal{D}$, such that:*

$$\mathcal{W}^\star(\mathbf{x}) = \sum_{i=1}^{\mathcal{D}} a_i \Psi_i(\mathbf{x}).$$

Assumption IV.2 *A "sufficiently" close approximation $\widehat{\mathcal{W}}_0$ of the true solution is available to start the iterative process:*

$$\|\widehat{\mathcal{W}}_0 - \mathcal{W}^\star\| < \epsilon.$$

Assumption IV.3 *Shape functions Ψ_i form an orthonormal set with respect to the standard Euclidean inner product, i.e. $L_2(\Omega)$.*

Assumption IV.1 has been made primarily for convenience and is equivalent to saying that the approximation space $\mathfrak{U}_\mathcal{D}$ equals $\mathfrak{U}_\infty$. It can be relaxed to read "sufficiently well approximable" (to within $\epsilon^\star$) instead of "exactly approximable," and the results would still hold but the mathematical development becomes tedious without adding significant insight. The stability proof that follows depends on the closeness of the starting approximation, i.e. assumption IV.2. Thus, if the L_2 error norm of the initial approximation is bounded above by ϵ, it is shown below that the error norm resulting from the next step of iteration is at most scaled by a constant factor. If the scaling factor (which depends on the particular system under consideration) is less than 1, a contraction mapping is obtained and convergence follows, but in general this might not be the case. The pdf for the first step of iteration can be obtained by using the Galerkin approach of chapter III, or other techniques such as stochastic averaging or statistical linearization. Finally, assumption IV.3 is made also purely for the sake of convenience of evaluating integrals, and the actual approximation space chosen need not satisfy this condition.

In the following, we set up equations for the Hilbert projection approach to find the coefficients a_i in assumption (IV.1). We redefine the inner product $\langle ., . \rangle$ as the following:

$$\langle \Psi_i, \Psi_j \rangle \quad \triangleq \quad \int_\Omega \Psi_i(\mathbf{x}) \Psi_j(\mathbf{x}) \mathcal{W}^\star(\mathbf{x}) d\mathbf{x}. \tag{4.1}$$

Then, the Hilbert coefficients, a_i for the true solution $\mathcal{W}^\star$ are given by the following equation:

$$\sum_{i=1}^{\mathcal{D}} a_i \langle \Psi_i, \Psi_j \rangle = \langle \mathcal{W}^\star, \Psi_j \rangle \qquad j = 1, 2, \ldots, \mathcal{D}. \tag{4.2}$$

Following assumption IV.3, $a_i = \langle \mathcal{W}^\star, \Psi_i \rangle$. Next, we define a new inner product, $\langle\langle ., . \rangle\rangle$, which is induced on the solution domain Ω by the current approximation $\widehat{\mathcal{W}}$, of the true solution $\mathcal{W}^\star$:

$$\langle\langle \Psi_i, \Psi_j \rangle\rangle \quad \triangleq \quad \int_{\Omega} \Psi_i(\mathbf{x}) \Psi_j(\mathbf{x}) \widehat{\mathcal{W}}(\mathbf{x}) d\mathbf{x}. \tag{4.3}$$

Using the new inner product defined above, projection equations 3.10 for variational formulation of stationary FPE can be rewritten as the following:

$$\begin{aligned} \langle\langle \mathcal{L}_{\mathcal{FP}}(\sum_{i=1}^{\mathcal{D}} a_i' \Psi_i), \Psi_j \rangle\rangle = \alpha \{ \langle\langle \sum_{i=1}^{\mathcal{D}} a_i' \Psi_i, \Psi_j \rangle\rangle_\Gamma - \\ \langle\langle \mathcal{W}^\star, \Psi_j \rangle\rangle_\Gamma \}, \quad j = 1, 2, \ldots \mathcal{D}, \end{aligned} \tag{4.4}$$

where, a_i' denote unknown coefficients of the approximation in the weighted Galerkin method. As before, α is a penalty parameter which has been introduced to enforce the boundary conditions, and $\langle\langle ., . \rangle\rangle_\Gamma$ denotes evaluation of the integral over the domain boundary.

2. Closeness of the Hilbert and Galerkin Approximations

In the above section, two sets of coefficients were discussed, a_i and a_i', corresponding to the Hilbert projection method and the Galerkin method weighted with the most recent approximation of the pdf respectively. Eq.4.4 represents a system of linear equations in a_i', which can be expressed as follows:

$$\mathbf{B}' \underline{\mathrm{A}}' + \mathbf{B}'_\Gamma \underline{\mathrm{A}}' = \underline{\mathrm{F}}'_\Gamma, \tag{4.5}$$

Following assumption IV.1, the Hilbert approximation of $\mathcal{W}^\star$ satisfies the weighted Galerkin variational form exactly, i.e.:

$$\langle \mathcal{L}_{\mathcal{FP}}(\sum_{i=1}^{\mathcal{D}} a_i \Psi_i), \Psi_j \rangle = \alpha \{ \langle \sum_{i=1}^{\mathcal{D}} a_i \Psi_i, \Psi_j \rangle_\Gamma - \langle \mathcal{W}^\star, \Psi_j \rangle_\Gamma \} \quad j = 1, 2, \ldots \mathcal{D}. \tag{4.6}$$

Note that the inner product in the Hilbert projection equation is weighted by the true solution, $\mathcal{W}^\star$, and hence the notation $\langle ., . \rangle$ is used. Eq.4.6 thus reduces to the following linear system:

$$\mathbf{B}\underline{\mathrm{A}} + \mathbf{B}_\Gamma \underline{\mathrm{A}} = \underline{\mathrm{F}}_\Gamma. \tag{4.7}$$

In Eq.(4.5), vector $\underline{\mathrm{A}}'$ represents the Galerkin coefficient vector while in Eq. (4.6), $\underline{\mathrm{A}}$ represents the Hilbert coefficient vector. Various other matrices and vectors are defined as follows:

$$\mathbf{B} = \left[\langle \mathcal{L}_{\mathcal{FP}}(\Psi_i), \Psi_j \rangle \right], \tag{4.8}$$

$$\mathbf{B}_\Gamma = -\alpha \left[\langle \Psi_i, \Psi_j \rangle_\Gamma \right], \tag{4.9}$$

$$\underline{\mathrm{F}}_\Gamma = -\alpha \left[\langle \mathcal{W}^\star, \Psi_j \rangle_\Gamma \right], \tag{4.10}$$

$$\mathbf{B}' = \left[\langle \langle \mathcal{L}_{\mathcal{FP}}(\Psi_i), \Psi_j \rangle \rangle \right], \tag{4.11}$$

$$\mathbf{B}'_\Gamma = -\alpha \left[\langle \langle \Psi_i, \Psi_j \rangle \rangle_\Gamma \right], \tag{4.12}$$

$$\underline{\mathrm{F}}'_\Gamma = -\alpha \left[\langle \langle \mathcal{W}^\star, \Psi_j \rangle \rangle_\Gamma \right]. \tag{4.13}$$

As the first step towards showing the closeness of $\underline{\mathrm{A}}'$ to $\underline{\mathrm{A}}$, we prove the proximity of Eq.4.5 to Eq.4.7 and write Eq.4.5 as:

$$\mathbf{B}\underline{\mathrm{A}}' + (\mathbf{B}_\Gamma + \Delta_3)\underline{\mathrm{A}}' = \underline{\mathrm{F}}_\Gamma + \Delta_1 + \Delta_2. \tag{4.14}$$

Comparing Eq.4.5 and Eq.4.14, we have:

$$\Delta_1 = \underline{\mathrm{F}}'_\Gamma - \underline{\mathrm{F}}_\Gamma, \tag{4.15}$$

$$\Delta_2 = \mathbf{B}\underline{\mathrm{A}}' - \mathbf{B}'\underline{\mathrm{A}}', \tag{4.16}$$

$$\Delta_3 = \mathbf{B}'_\Gamma - \mathbf{B}_\Gamma. \tag{4.17}$$

Then, we have the following lemma for upper bounds of various Δ_i :

Lemma IV.1 *Given the validity of assumptions IV.1 and IV.2, following inequalities hold:*

$$\|\Delta_1\| \leq K_1\epsilon,$$

$$\|\Delta_2\| \leq K_2\|\mathcal{L}_{\mathcal{FP}}\|\epsilon,$$

$$\|\Delta_3\| \leq K_3\epsilon,$$

where, K_1-K_3 are finite constants, $\|.\|$ represents the Euclidean norm for vectors Δ_1 and Δ_2, and matrix norm induced by the Euclidean norm for Δ_3, and $\|\mathcal{L}_{\mathcal{FP}}\|$ represents the operator norm of the Fokker-Planck operator.

Proof: Consider $\Delta_1 = [\delta_j^1]$:

$$\begin{aligned}
\delta_j^1 &= \alpha\{\langle \mathcal{W}^\star, \Psi_j \rangle_\Gamma - \langle\langle \mathcal{W}^\star, \Psi_j \rangle\rangle_\Gamma\} \\
&= \alpha \int_\Gamma \mathcal{W}^\star \Psi_j (\mathcal{W}^\star - \widehat{\mathcal{W}}) d\mathbf{x} && (4.18) \\
\Rightarrow |\delta_j^1|^2 &\leq |\alpha|^2 \int_\Gamma |\mathcal{W}^\star|^2 |\Psi_j|^2 |\mathcal{W}^\star - \widehat{\mathcal{W}}|^2 d\mathbf{x} && (4.19) \\
&\leq |\alpha|^2 \int_\Gamma |\Psi_j|^2 |\mathcal{W}^\star - \widehat{\mathcal{W}}|^2 d\mathbf{x} \int_\Gamma |\mathcal{W}^\star|^2 d\mathbf{x} && (4.20) \\
&\leq |\alpha|^2 \int_\Gamma |\Psi^\star|^2 |\mathcal{W}^\star - \widehat{\mathcal{W}}|^2 d\mathbf{x}.1 && (4.21) \\
&\leq |\alpha|^2 \int_\Gamma |\Psi^\star|^2 d\mathbf{x} \int_\Gamma |\mathcal{W}^\star - \widehat{\mathcal{W}}|^2 d\mathbf{x} && (4.22) \\
&\leq |\alpha|^2 .1. \epsilon^2 && (4.23) \\
\Rightarrow |\delta_j^1| &\leq |\alpha| \epsilon && (4.24)
\end{aligned}$$

In the above, Cauchy-Schwarz inequality has been applied in going from Eq.4.19 to Eq.4.20 and from Eq.4.21 to Eq.4.22. Additionally, weaker forms of assumptions IV.2 and IV.3 (since only boundary integrals are involved) have been used in Eq.4.21 and Eq.4.23. Thus, from Eq.4.24, we conclude that there exists $K_1 < \infty$ such that:

$$\|\Delta_1\| \leq K_1 \epsilon. \tag{4.25}$$

Next, looking at $\Delta_2 = [\delta_j^2]$, and following similar arguments as above, we obtain:

$$\begin{aligned}
\delta_j^2 &= \langle \sum_{i=1}^{\mathcal{D}} a_i' \mathcal{L}_{\mathcal{FP}}(\Psi_i), \Psi_j \rangle - \\
&\qquad \langle\langle \sum_{i=1}^{\mathcal{D}} a_i' \mathcal{L}_{\mathcal{FP}}(\Psi_i), \Psi_j \rangle\rangle \qquad (4.26) \\
&= \int_{\Omega} \sum_{i=1}^{\mathcal{D}} a_i' \mathcal{L}_{\mathcal{FP}}(\Psi_i)(\mathcal{W}^\star - \widehat{\mathcal{W}}) d\mathbf{x} \\
\Rightarrow |\delta_j^2|^2 &\leq \int_{\Omega} |\sum_{i=1}^{\mathcal{D}} a_i' \mathcal{L}_{\mathcal{FP}}(\Psi_i)|^2 |\widehat{\mathcal{W}} - \mathcal{W}^\star|^2 d\mathbf{x} \\
&\leq \|\mathcal{L}_{\mathcal{FP}}(\sum_{i=1}^{\mathcal{D}} a_i' \Psi_i)\|^2 \epsilon^2 \\
&\leq \|\mathcal{L}_{\mathcal{FP}}\|^2 \| \sum_{i=1}^{\mathcal{D}} a_i' \Psi_i \|^2 \epsilon^2 \\
&\leq \|\mathcal{L}_{\mathcal{FP}}\|^2 \|\underline{\mathrm{A}}'\|^2 \epsilon^2 \qquad (4.27)
\end{aligned}$$

In Eq.4.27, norm of the Galerkin coefficient vector, $\|\underline{\mathrm{A}}'\|$ is a finite quantity because it contains coefficients of various shape functions used to approximate pdfs that have well behaved functional forms (i.e. without δ-function like singularities). Hence, bounding it above by a finite quantity, we can show that there exists a $K_2 < \infty$ such that:

$$\|\Delta_2\| \leq K_2 \|\mathcal{L}_{\mathcal{FP}}\| \epsilon. \qquad (4.28)$$

Finally, considering $\Delta_3 = [\delta_{ij}^3]$:

$$\begin{aligned}
\delta_{ij}^3 &= \alpha \{ \langle \Psi_i, \Psi_j \rangle_\Gamma - \langle\langle \Psi_i, \Psi_j \rangle\rangle_\Gamma \} \\
&= \alpha \int_{\Gamma} \Psi_i \Psi_j (\mathcal{W}^\star - \widehat{\mathcal{W}}) d\mathbf{x} \\
\Rightarrow |\delta_{ij}^3|^2 &\leq |\alpha|^2 \int_{\Gamma} |\Psi_i|^2 |\Psi_j|^2 |\mathcal{W}^\star - \widehat{\mathcal{W}}|^2 d\mathbf{x} \\
&\leq |\alpha|^2 1.1. \epsilon^2 \qquad (4.29)
\end{aligned}$$

A weak form of assumption IV.3 (over the boundary) has been used in Eq.4.29. Thus, $\exists\ K_3 < \infty$, such that:

$$\|\Delta_3\| \leq K_3\epsilon \tag{4.30}$$

This completes the proof of the lemma. □

We now proceed to prove stability of the iterative approach by establishing an upper bound for the error of approximation resulting from the weighted Galerkin approach. We make the following additional assumptions:

Assumption IV.4 *The quantity ϵ is small enough such that $\|(\mathbf{B}+\mathbf{B_\Gamma})^{-1}\|\|\Delta_3\| \leq 1$.*

Assumption IV.5 *The operator norm of Fokker-Planck operator is bounded above as $\|\mathcal{L}_{\mathcal{FP}}\| = M < \infty$.*

This leads us to the following result:

Lemma IV.2 *Given the validity of assumptions IV.4 and IV.5, the following upper bound exists on the L_2 error norm between the weighted Galerkin and Hilbert approximations of FPE:*

$$\|\underline{\mathrm{A}}' - \underline{\mathrm{A}}\| \leq K\epsilon \tag{4.31}$$

Proof: Let us adopt the following notation: $\mathbf{B}' + \mathbf{B}'_\Gamma = \mathbf{B}_G$, and $\Delta_1 + \Delta_2 = \Delta_\Sigma$. Then, Eqs.4.7 and 4.14 become:

$$(\mathbf{B}_G - \Delta_3)\underline{\mathrm{A}} = \underline{\mathrm{F}}'_\Gamma - \Delta_\Sigma. \tag{4.32}$$

$$\mathbf{B}_G\underline{\mathrm{A}}' = \underline{\mathrm{F}}'_\Gamma. \tag{4.33}$$

Thus, we have:

$$\begin{aligned}
\underline{\mathrm{A}}' - \underline{\mathrm{A}} &= \mathbf{B}_G^{-1}\underline{\mathrm{F}}'_\Gamma - (\mathbf{B}_G - \Delta_3)^{-1}(\underline{\mathrm{F}}'_\Gamma - \Delta_\Sigma) \\
&= \{\mathbf{B}_G^{-1} - (\mathbf{B}_G - \Delta_3)^{-1}\}\underline{\mathrm{F}}'_\Gamma + (\mathbf{B}_G - \Delta_3)^{-1}\Delta_\Sigma
\end{aligned}$$

Taking standard L_2 norm on both sides and applying the triangle inequality,

$$\|\underline{\mathrm{A}}' - \underline{\mathrm{A}}\| \leq \|\mathbf{B}_G^{-1} - (\mathbf{B}_G - \Delta_3)^{-1}\|\|\underline{\mathrm{F}}'_\Gamma\| + \|(\mathbf{B}_G + \Delta_3)^{-1}\|\|\Delta_\Sigma\| \tag{4.34}$$

Furthermore, following assumption IV.4, we obtain the following expansion:

$$(\mathbf{B}_G - \Delta_3)^{-1} = \mathbf{B}_G^{-1} + \mathbf{B}_G^{-2}\Delta_3 - \dots$$

Thus, using the result for upper bound of Δ_3 from lemma IV.1 (Eq.4.30), we obtain:

$$\|(\mathbf{B}_G - \Delta_3)^{-1}\| \leq \|\mathbf{B}_G^{-1}\| + \|\mathbf{B}_G^{-1}\|^2 K_3\epsilon,$$
$$\|\mathbf{B}_G^{-1} - (\mathbf{B}_G - \Delta_3)^{-1}\| \leq \|\mathbf{B}_G^{-1}\|^2 K_3\epsilon.$$

Also, combining Eqs.4.25 and 4.28:

$$\|\Delta_\Sigma\| \leq K_\Sigma(1 + \|\mathcal{L}_{\mathcal{FP}}\|)\epsilon, \tag{4.35}$$

where $K_\Sigma = \max(K_1, K_2)$. Denoting $\|\mathbf{B}_G^{-1}\|$ as P, $\|\underline{\mathrm{F}}'_\Gamma\|$ as Q, and $\|\mathcal{L}_{\mathcal{FP}}\|$ as M, Eq. 4.34 becomes:

$$\|\underline{\mathrm{A}}' - \underline{\mathrm{A}}\| \leq QP^2K_3\epsilon + K_\Sigma(P + P^2K_3\epsilon)(1 + M)\epsilon.$$

Dropping out terms of order higher than $\mathcal{O}(\epsilon)$, we get:

$$\|\underline{\mathrm{A}}' - \underline{\mathrm{A}}\| \leq (QP^2K_3 + PK_\Sigma(1 + M))\epsilon,$$
$$\Rightarrow \quad \|\underline{\mathrm{A}}' - \underline{\mathrm{A}}\| \leq K\epsilon.$$

This completes the proof of the lemma. □

Therefore, we see that error in the next iteration of the refinement process is scaled by the constant K, which comprises of several norms associated with the underlying system. If this quantity is less than 1, we obtain a contraction mapping

and the error reduces to zero in the limit. However, this is not true in general. In either case, the method is stable for a finite number of iterations and will not lead to divergence (except in certain pathological cases discussed below). Superior convergence characteristics has been shown for dynamical systems in two and three dimensions through numerical simulations in the results section.

Looking closer at the constant K, we observe that the norm of the inverse of the Hilbert stiffness matrix, $\mathbf{B}_G$ appears in its expression. If this matrix is ill-conditioned or singular, the method loses its stability. This situation may arise in certain conditions (e.g. local methods which involve shape functions with compact support) and is discussed in detail below. On the other hand, the norm of the vector $\underline{\mathrm{F}}$ does not cause problems as it involves the integral of the true solution along the domain boundary, which is a very small quantity ($\sim 10^{-6}$ or lower). In the numerical examples shown below, we show that convergence is achievable using the PUFEM algorithm for variational formulation, in conjunction with suitable patching of solutions from successive iterations.

C. Domain Tracking

In the above section it was assumed that the solution domain on which iterations are performed is known a-priori. In general, this might not be the case, especially for nonlinear systems. In this section, the implementation of a space homotopy is demonstrated via a family of single parameter dynamical systems to track the domain of stationary distribution for the system of interest. The main underlying assumption is the existence of a family of dynamical systems, $\mathfrak{D}_p$ indexed by the homotopy parameter p:

$$\mathfrak{D}_p : d\mathbf{x} = \mathbf{f}(\mathbf{x}, p)dt + \mathbf{g}(\mathbf{x}, p)d\mathbf{B},\ p \in [0, 1], \tag{4.36}$$

where $\mathfrak{D}_1$ corresponds to the dynamical system of interest and $\mathfrak{D}_0$ corresponds to a stochastic dynamical system whose response is known, i.e., stationary FPE associated with it can be solved. Let $\mathcal{W}_p^\star(\mathbf{x})$ denote the true solution of FPE associated with dynamical system $\mathfrak{D}_p$. We make the following assumption about the family of dynamical systems $\mathfrak{D}_p$ and solutions of associated FPEs, $\mathcal{W}_p^\star$:

Assumption IV.6 *Given any $p \in [0,1]$, and any $\epsilon > 0$, there exists $\delta > 0$ such that for all $p' \in B_\delta(p)$ (open ball of radius δ centered at p), $||\mathcal{W}_p^\star(\mathbf{x}) - \mathcal{W}_{p'}^\star(\mathbf{x})|| \leq \epsilon$.*

In essence, the above assumption assumes the existence of a one parameter family of dynamical systems such that solutions to associated FPEs change smoothly over this parameter space - in other words, a homotopy exists. We next consider only those dynamical systems for which the constant K appearing in lemma IV.2 is less than unity (hence leading to a contraction mapping and ensuring convergence). The following obvious result can then be stated as a proposition:

Proposition IV.1 *Consider dynamical systems $\mathfrak{D}_p$ with $K < 1$ in lemma IV.2. Then, given $\widehat{\mathcal{W}}^1$, such that $||\widehat{\mathcal{W}}^1 - \mathcal{W}^\star|| \leq \epsilon$ and that ϵ is sufficiently small, a sequence of functions $\{\widehat{\mathcal{W}}^n\}_{n=1}^\infty$ can be constructed recursively, starting with $\widehat{\mathcal{W}}^1$ such that $||\widehat{\mathcal{W}}^n - \mathcal{W}^\star|| \to 0$ as $n \to \infty$, $\forall\ p \in [0,1]$.*

The proof is trivial because of lemma IV.2 and the contraction mapping argument for $K < 1$.

A note about notation: $\widehat{\mathcal{W}}^i$ refers to the i^{th} function of a sequence $\{\widehat{\mathcal{W}}^i\}_{i=1}^S$. On the other hand, $\mathcal{W}_i^\star$ refers to the true solution of FPE for the dynamical system $\mathfrak{D}_{p=i}$. In other words, the superscript i refers to member of a sequence while the subscript i refers to the homotopy parameter.

Then, with assumption IV.6 in mind, we have the following result pertaining to how the solution of FPE associated with the system of interest ($\mathfrak{D}_1$), i.e.,

$\mathcal{W}_1^\star(\mathbf{x}) = \mathcal{W}^\star(\mathbf{x})$ can be obtained recursively given the knowledge of solution of FPE for the system $\mathfrak{D}_0$. The result uses proposition IV.1 in conjunction with successive approximation.

Proposition IV.2 *Let ϵ_p be sufficiently small such that proposition IV.1 is satisfied for any $\widehat{\mathcal{W}}$ satisfying $||\widehat{\mathcal{W}} - \widehat{\mathcal{W}}_p^\star|| \leq \epsilon_p$. Let $\inf_{p\in[0,1]} \epsilon_p = \bar{\epsilon} > 0$. Then, under assumptions IV.1-IV.3, IV.4, IV.5 and IV.6, given $\mathcal{W}_0^\star$ (exact solution of FPE corresponding to $\mathfrak{D}_0$), there exists a finite sequence of functions $\{\widehat{\mathcal{W}}^n\}_{n=1}^M$ s.t. $\widehat{\mathcal{W}}^M = \mathcal{W}^*$. Moreover, this sequence can be obtained in a recursive fashion starting with $\mathcal{W}_0^\star$. (i.e. $\widehat{\mathcal{W}}^1 = \mathcal{W}_0^\star$)*

Proof: Let δ_p be such that if $p' \in B_{\delta_p}(p)$ then $||\mathcal{W}_p^\star - \mathcal{W}_{p'}^\star|| \leq \frac{\bar{\epsilon}}{2}$. Note that this is possible due to assumption IV.6. Consider the open covering $\bigcup_{p\in[0,1]} B_{\delta_p}(p)$ of the set $[0,1]$. Since $[0,1]$ is compact, there exists a finite subcover of $[0,1]$ given by $\bigcup_{i=1}^M B_{\delta_{p_i}}(p_i)$. Let $\delta_i \equiv \delta_{p_i}(p_i)$ and redefine $\mathcal{W}_{p_i}^\star \equiv \mathcal{W}_i^\star$.
Let us assume that $\mathcal{W}_i^\star$ is known and we need to obtain $\mathcal{W}_{i+1}^\star$. By definition, there exists a $\tilde{p}$ such that $|\tilde{p} - p_i| < \delta_i$ and $|\tilde{p} - p_{i+1}| < \delta_{i+1}$. Then, it follows from construction that

$$||\mathcal{W}_i^\star - \mathcal{W}_{i+1}^\star|| \leq ||\mathcal{W}_i^\star - \mathcal{W}_{\tilde{p}}^\star|| + ||\mathcal{W}_{\tilde{p}}^\star - \mathcal{W}_{i+1}^\star|| \leq \bar{\epsilon}. \tag{4.37}$$

Then, due to proposition IV.1, starting with $\mathcal{W}_i^\star$, it is possible to obtain $\mathcal{W}_{i+1}^\star$ in a recursive fashion. Note that the above holds for all $i = 0, 1, \ldots, M-1$. In this fashion we obtain the sequence $\{\mathcal{W}_0^\star, \mathcal{W}_1^\star, \ldots, \mathcal{W}_M^\star = \mathcal{W}^*\}$ recursively starting with $\mathcal{W}_0^\star$.
This completes the proof of the proposition. □

In summary, the development above (space homotopy in conjunction with solution refinement) can be presented as the following **algorithm**:

1. Find a homotopy of dynamical systems $\mathfrak{D}_p$, $p \in [0,1]$, such that $\mathfrak{D}_p(p=1)$ corresponds to the system of interest and $\mathfrak{D}_p(p=0)$ corresponds to a known system, in the sense that its stationary FPE can be solved.

2. Select a finite number of points $p_i \in [0,1], i = 1, \ldots, M$, that are "sufficiently" close. For the rest of the algorithm, refer to the dynamical system corresponding to p_i, namely $\mathfrak{D}_{p_i}$, by its index, i.e. as $\mathfrak{D}_i$, and the true solution of the associated stationary FPE, $\mathcal{W}^\star_{p_i}$ as $\mathcal{W}^\star_i$. Note that selection of points p_i can be done online, i.e. if p_{i+1} is found to be "not close enough" to p_i, it is possible to go back and redo the previous iteration.

3. Following the new "index based" notation, note that the exact solution for system $\mathfrak{D}_1$ is known ($\mathcal{W}^\star_1$). Also, the solution we are after (for $p=1$) is, in the new notation, $\mathcal{W}^\star_M = \mathcal{W}^\star$. Set $i = 2$.

4. Determine the solution $\mathcal{W}^\star_i$ in the following manner:

   (a) Set $j = 1$ and the current weight for norm modification, $\widehat{\mathcal{W}} = \mathcal{W}^\star_{i-1}$.

   (b) Using $\widehat{\mathcal{W}}$ as the weight in the modified norm approach, obtain $\widehat{\mathcal{W}}^j_i$, i.e. the j^{th} approximation for $\mathcal{W}^\star_i$.

   (c) **If** $\widehat{\mathcal{W}}^j_i = \mathcal{W}^\star_i$, goto step 5. **Else**, set $j = j+1$ and $\widehat{\mathcal{W}} = \widehat{\mathcal{W}}^{j-1}_i$ and goto step 4b.

5. **If** $i = M$, **stop**. **Else**, set $i = i+1$ and goto step 4a.

The above algorithm involves two loops. The outer loop runs over the homotopic sequence of dynamical systems, from the "known" to the "desired." The inner loop performs approximation refinements for each dynamical system via the modified-norm approach until a good enough approximation is achieved for its true solution. The

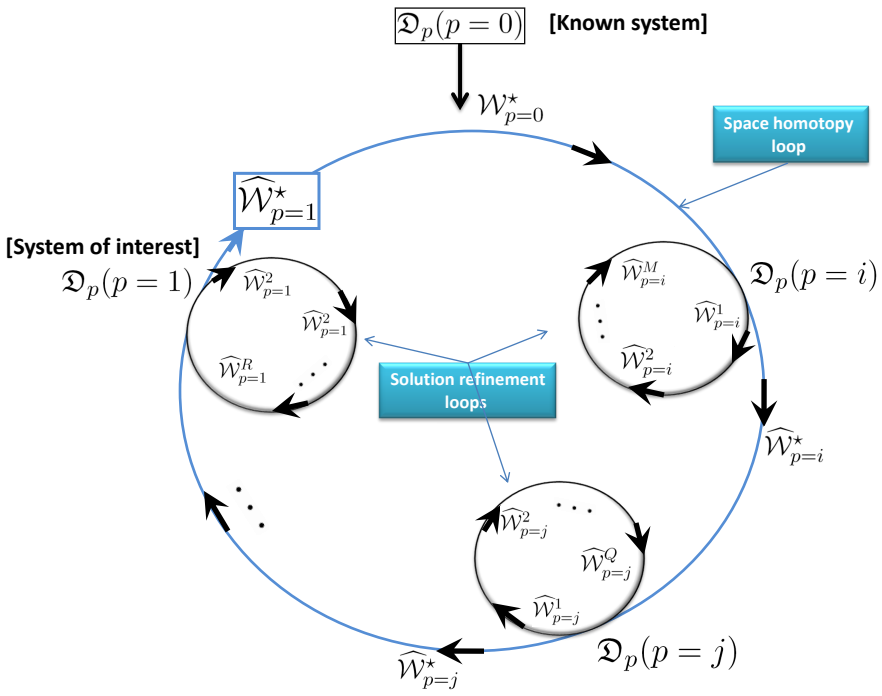


Fig. 37. A graphic illustration of the norm-modification algorithm.

algorithm starts with the known system ($p = 0$) whose solution of stationary FPE is available, serving as the first weight for norm-modification. A schematic of the above algorithm is presented in Fig.37, illustrating its nested loops. A point to note is that measuring error in the inner loop (closeness to the true solution for a particular dynamical system in the homotopy) is not a trivial exercise, because true solutions are not known except for $\mathfrak{D}_p(p = 0)$. In practice, equation error is used to measure closeness of approximation to the truth. A summary of the above algorithm along with results can also be found in the figure on page 132.

D. Numerical Implementation

The recursive norm modification techniques described in the above sections are not subject to a particular method of solving FPE in the variational form. Since PUFEM was used to develop Galerkin approximations in chapter III, the same will be used for solution refinement and domain tracking in the current chapter. The equation below represents the PUFEM approximation of the pdf in the n^{th} iteration, $\widehat{\mathcal{W}}^n$:

$$\widehat{\mathcal{W}}^n(\mathbf{x}) = \sum_{i=1}^{P} \sum_{j=1}^{Q_i} {}^n a'_{ij} \varphi_i(\mathbf{x}) \psi_{ij}(\mathbf{x}), \tag{4.38}$$

The above equation is similar to Eq.3.6 except with the superscript n to signify the current iteration in solution refinement. Note that to simplify notation in this chapter, the above equation has been written using a single index "i" instead of "i" and "j" and a single summation of shape functions, from 1 to $\mathcal{D}$ ($= \sum_{k=1}^{P} Q_k$). Then, the objective of solution refinement is to obtain a better approximation $\widehat{\mathcal{W}}^{n+1}$ by using $\widehat{\mathcal{W}}^n$ as a weight to modify the inner product, resulting in the following weighted Galerkin variational equation (see Eq.4.4):

$$\begin{aligned} \int_{\Omega_{\text{sub}}} \sum_{i=1}^{\mathcal{D}} \mathcal{L}_{\mathcal{FP}}({}^{n+1}a'_i \Psi_i) \Psi_j \widehat{\mathcal{W}}^n d\mathbf{x} = \alpha \int_{\Gamma_{\text{sub}} \cap \Gamma} \sum_{i=1}^{\mathcal{D}} {}^{n+1}a'_i \Psi_i \Psi_j \widehat{\mathcal{W}}^n d\mathbf{x} \\ -\alpha \int_{\Gamma_{\text{sub}} \cap \Gamma} \mathcal{W}^\star \Psi_j \widehat{\mathcal{W}}^n d\mathbf{x}, \quad j = 1, \ldots, \mathcal{D} \end{aligned} \tag{4.39}$$

The resulting elements of the matrices involved in the linear system of equations (4.5) are:

$$B'_{ij} = \int\limits_{\Omega_{\text{sub}}} \mathcal{L}_{\mathcal{FP}}(\Psi_i)\Psi_j \widehat{\mathcal{W}}^n d\mathbf{x}, \tag{4.40}$$

$$B'_{\Gamma ij} = -\alpha \int\limits_{\Gamma_{\text{sub}}\cap\Gamma} \Psi_i \Psi_j \widehat{\mathcal{W}}^n d\mathbf{x}, \tag{4.41}$$

$$F'_{\Gamma i} = -\alpha \int\limits_{\Gamma_{\text{sub}}\cap\Gamma} \mathcal{W}^\star \Psi_i \hat{\mathcal{W}}^n d\mathbf{x}, \tag{4.42}$$

1. Conditioning of the Stiffness Matrix

As seen in Sec.2, the condition number of stiffness matrix $\mathbf{B}_G$ is an important issue in considering stability of the above approach. Unfortunately, if a local approximation scheme such as PUFEM is used for either the weighted Galerkin or Hilbert approaches, the stiffness matrix invariably turns out to be ill-conditioned. This is because of the following reason - the current approximate pdf used as weight gives relative weightage to different regions of the domain, thus distinguishing regions of higher significance (close to the mean) from regions of low significance (e.g. regions beyond 3σ for a Gaussian distribution). Also, in a local scheme, shape functions and their coefficients (a_i or a'_i) have local influence. In other words, the integrals associated with shape functions close to the boundary are evaluated on local domains only near the boundary region. By virtue of the exponentially low weight given to these regions by the weighting pdf, these integrals get nearly "washed out" in comparison with the integrals evaluated on local domains close to the mean. Consequently, the entries in the stiffness matrix $\mathbf{B}$ corresponding to local shape functions defined near the boundary regions diminish severely in comparison with entries for shape functions in the interior. This effect makes the boundary coefficients unobservable, and

the resulting stiffness matrix numerically ill-conditioned.

2. A Numerical Fix

The section above argues that using a pdf as weight for modification of the inner product causes ill-conditioning of the stiffness matrix in local approximation techniques because it renders local coefficients near the boundary regions unobservable. A natural solution to this problem is to extract the portion of the stiffness matrix which has acceptable conditioning for inversion, and to retain the solution for the remaining coefficients from the previous iteration. In this manner, not all coefficients are modified in going from one iteration to another because coefficients close to the boundary do not change. This method also gives a simple way of trimming the domain of solution from one iteration to the next - by identifying and pruning regions which receive weightage below a specified tolerance from the weighting pdf. However, we mention that selective modification of coefficients usually leads to discontinuities and/or ripple formation in and around the concerned local domains. To counter this, the two sets of coefficients (from the current and previous iterations) are patched together to produce a smooth surface. This "patching" procedure can be done using the PUFEM algorithm with the help of smooth blending functions, such as those mentioned in section B2a in chapter III. This approach provided highly acceptable results as illustrated in the next section.

E. Results in Solution Refinement and Domain Tracking

In this section, numerical examples are presented for illustration of the theoretical ideas discussed above. It is shown that with the weighted norm approach, it is possible to obtain high accuracy while using a small number of degrees of freedom.

1. Solution Refinement of Stationary FPE: Results

We first consider two nonlinear systems residing in 2D state-space described below:

a. System 1: Example in 2D State-Space

Consider the following 2-D damped Duffing oscillator:

$$\ddot{x} + \eta\dot{x} + \alpha x + \beta x^3 = g\mathcal{G}(t) \tag{4.43}$$

We assign the parameters appearing above the following values: $\alpha = -15, \beta = 30, \eta = 10, g = 1$ (soft-spring case). The analytical solution of stationary FPE for this system is known and given in similar fashion as in Eq.3.22, shown in Fig.11(a), which is a bimodal pdf.

b. System 2: Example in 2D State-Space

Consider the following 2-D nonlinear oscillator[66]:

$$\ddot{x} + \beta\dot{x} + x + \alpha(x^2 + \dot{x}^2)\dot{x} = g\mathcal{G}(t) \tag{4.44}$$

We set the following values: $\alpha = 0.125, \beta = -0.5, g = 0.86$. This system was considered in section E in chapter III. We see that from the top-view (Fig.29(b)), the stationary distribution for this system looks like a ring. Note that stationary distributions for both systems considered above are exponentials of a polynomial function.

As the first exercise, we evaluate the various norms involved to ensure that the systems described above conform with the theory presented in section B1. In particular, we demonstrate that the numerical fix suggested in section D2 to tackle unobservability of boundary nodes indeed causes the constant K appearing in lemma IV.2 to be less than unity, thus leading to a contraction mapping, which in turn

Table VI. Approximate estimates of various norms and constants appearing in the theory, for systems 1 and 2.

Quantity/Norm	System 1	System 2
K_1	10^{-9}	10^{-9}
K_2	10^{-5}	10^{-5}
K_3	1	1
$\|\mathcal{L}_{\mathcal{FP}}\| = M$	41.13	7.87
$\|\mathbf{B}_G^{-1}\| = P$, before fix	2.25×10^8	7.92×10^9
$\|\mathbf{B}_G^{-1}\|_{fixed} = P_{fixed}$	1.96×10^3	1.04×10^4
$\|\underline{\mathbf{F}}'_\Gamma\| = Q$	2.44×10^{-9}	3.57×10^{-9}
K (Lemma IV.2)	0.835	0.961

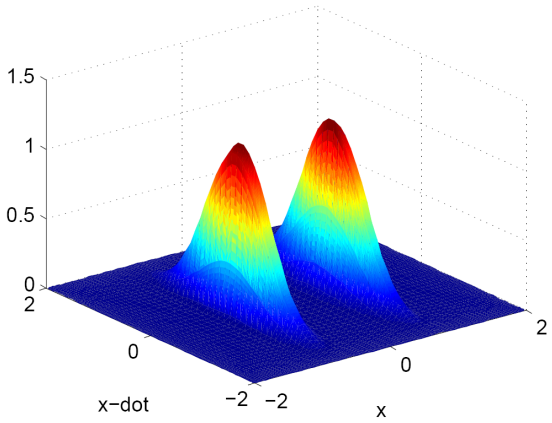
implies convergence. Table VI contains the various quantities that appear in lemmas IV.1 and IV.2. These are ballpark numbers and give order of magnitude estimates. The constants $K_1 - K_3$ appearing in lemma IV.1 have been computed by evaluating the various domain and boundary integrals. The operator norm, $\|\mathcal{L}_{\mathcal{FP}}\|$ has been computed via discretization. Note that using a pdf as weight to modify the L_2 norm causes the stiffness matrix to be nearly singular. However, the numerical fix suggested in section D2 attenuates the ill-conditioning significantly, enough to make the constant K appearing in lemma IV.2 less than unity. The actual numerical value of K suggests that convergence is expected to be faster for system 1 than system 2. This was indeed observed and is illustrated in convergence plots presented below.

Figs. 38(a)-38(f) show results for system 1 (soft-spring Duffing oscillator). Fig. 38(b) shows the error surface using the sPUFEM algorithm with standard L_2 inner product approach on a 16×16 nodal grid equipped with local quadratic basis functions. This error surface serves as a reference for the standard L_2 approach. We next perform the iterative refinement process, starting with the L_2 solution computed on a much coarser grid (12×12). This solution is also the weight for inner-product modification in the first iteration. Upon using the modified inner-product approach

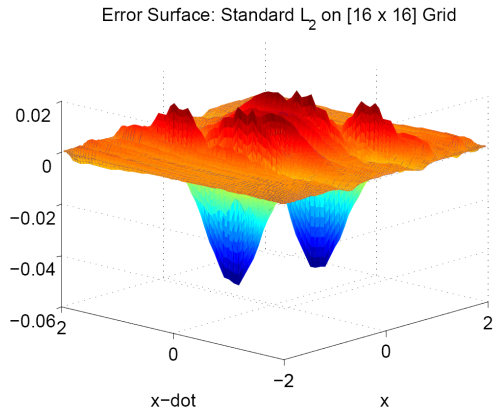
in conjunction with patching of neighboring iteration approximations, the accuracy improves significantly, which is evident in the error surface shown in Fig. 38(c).

The true power of this approach is illustrated in Fig.38(d), in which the convergence characteristics for three comparable methods have been shown. The graph corresponding to iterative sPUFEM confirms that the process is commenced on a coarse 12×12 grid, and the use of pdf obtained after every iteration to modify inner-product space for the subsequent iteration leads to significant drops in error. Once no further accuracy is possible with the 12×12 mesh, a switch is made to a finer grid (14×14), beginning with the last pdf obtained from the previous (12×12) grid as the weight for the first iteration on the new grid. The spacing between circles on the iterative sPUFEM graph illustrates the drop in error after individual iterations. Thus, huddling of circles signifies saturation on a particular grid, and a switch to a finer grid is made following such behavior. In the graph shown, iterations have been terminated after saturation of the (16×16) grid, and the final error surface is shown in Fig. 38(c). The most significant contribution of this result is that it shows that it is possible to achieve extremely accurate approximations with a small number of degrees of freedom. For example, compare in Fig. 38(d) the error after the final iteration on the 16×16 grid ($\equiv$ 1536 PUFEM DOFs using quadratic bases) with the error of the L_2 approach on a 30×30 grid ($\equiv$ 5400 PUFEM DOFs with quadratic bases).

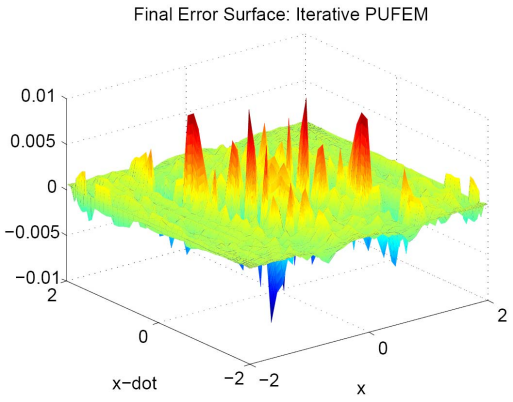
Fig. 38(e) illustrates (for iteration #3) the phenomenon of ripple formation when selective update is performed by pruning out unobservable coefficients that are weighted out by the *pdf*. As expected, ripples form on either side of the two weighty modes, where the pdf drops off suddenly to extremely small values on either side. However, patching of current and previous iterations described in section D2 smoothes out these ripples and a relatively better solution is obtained (Fig. 38(f)).



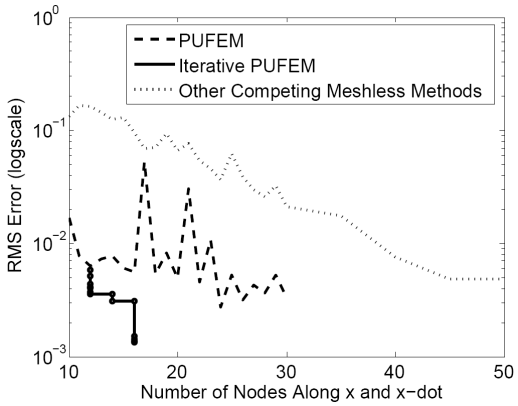
(a) True solution: Damped Duffing oscillator.



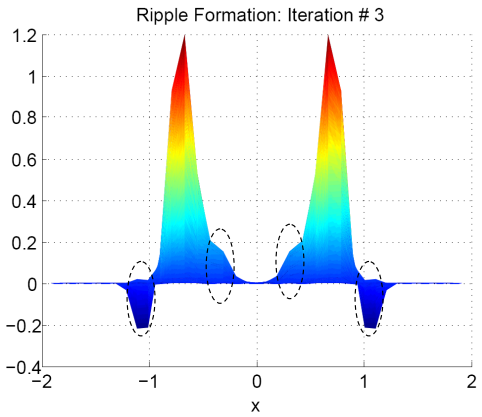
(b) Error surface: Standard L_2 approach.



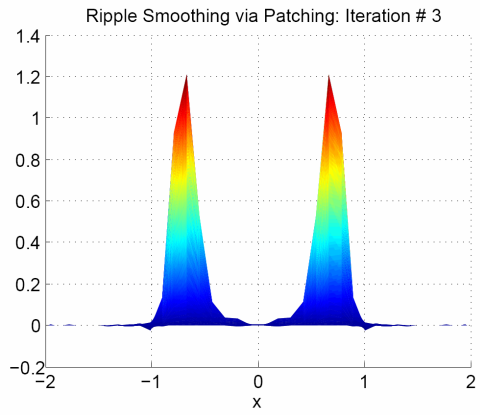
(c) Error surface at the end of the iterative process.



(d) Comparative convergence characteristics: Duffing oscillator.



(e) Ripple formation in the Duffing oscillator at the boundary of transition between low and high weightage regions of the weighting pdf.



(f) Smoothing of ripples by patching solutions from adjacent iterations.

Fig. 38. Simulation results for the damped Duffing oscillator.

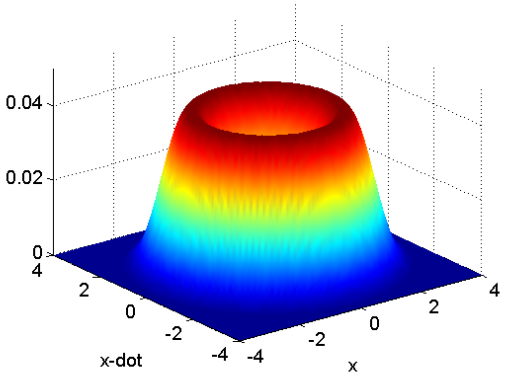
Similar results are obtained for system 2 (Fig. 39), and it is again possible to obtain high accuracy with a much smaller number of approximation nodes, as compared with the standard L_2 approach. However, the results are not as drastic here because for this system, it is possible to obtain fairly accurate results even with the standard L_2 approach. Also, the convergence rate is slower as visible in Fig. 39(c), which is also evident from the numerical value of constant K in Table VI. In addition to systems considered in this section, similar encouraging results for several other 2-D oscillators have been obtained.

c. System 3: Example in 3D State-Space

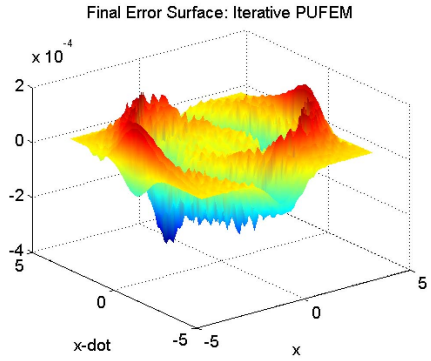
Consider the following dynamical systems studied in Wojtkiewicz et.al.[82]:

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 & 0 \\ -\omega_0 & -2\zeta\omega_0 & 1 \\ 0 & 0 & -\alpha \end{bmatrix} \mathbf{x} + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} w(t) \tag{4.45}$$

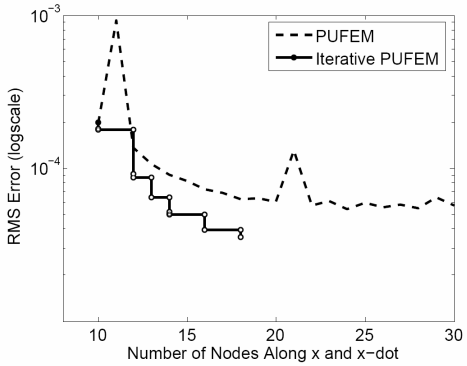
The above system is described in detail in section C in chapter III. Recall that stationary FPE for the system above was solved in [82] using the traditional FEM approach with $125,000$ "brick" elements, leading to RMS error (defined as $e_2 \triangleq \sqrt{\frac{1}{r-1}\sum_{i=1}^{r}(\mathcal{W}(\mathbf{x}_i) - \widehat{\mathcal{W}}(\mathbf{x}_i))^2}$, $r =$ number of test points) of $e_2(FEM) = 1.133\times10^{-4}$. The same problem was solved in this dissertation on a $6 \times 6 \times 6$ grid utilizing local p-refinement in the sPUFEM algorithm (see section C in chapter III). Cubic polynomials were allocated to nodes in the interior region of the domain and quadratic polynomials to boundary nodes. This polynomial assignment results in a problem size of 2800 DOFs. The standard L_2-norm approach results in an RMS error of $e_2(6 \times 6 \times 6,\ \text{cubic+quadratic}, L_2) = 1.336 \times 10^{-3}$. This approximation was used as the first weight to commence solution refinement with the modified-norm ap-



(a) True solution: System 2.

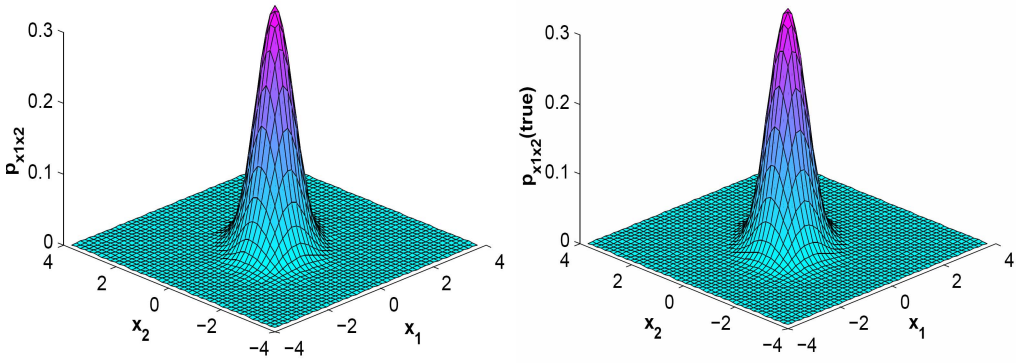


(b) Error surface at the end of the iterative process.



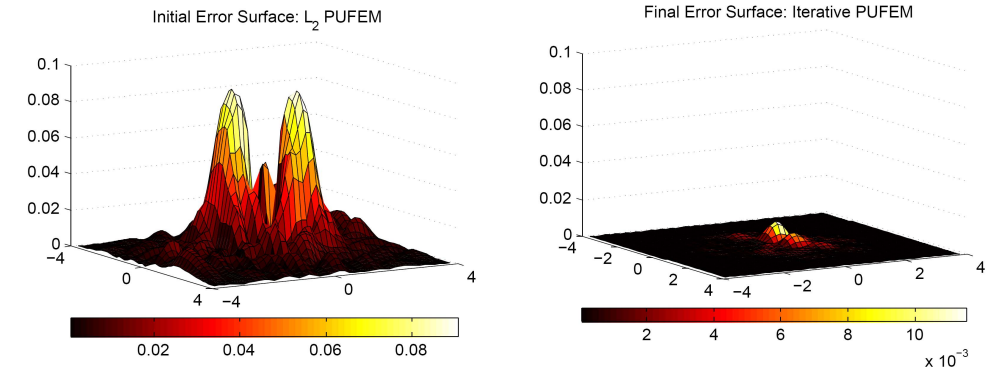
(c) Comparative convergence characteristics: System 2.

Fig. 39. Simulation results for system 2.



(a) Computed $x_1 - x_2$ marginal distribution for system 3 at the end of the iterative process.

(b) True $x_1 - x_2$ marginal probability density function for system 3.



(c) Error surface resulting from standard L_2 error projection using 2800 DOFs.

(d) Error surface at the end of the iterative process.

Fig. 40. Simulation results for system 3.

proach. The obtained results are shown in figure 40. Figures 40(a) and 40(b) show the converged $x_1 - x_2$ marginal surface alongside the true marginal distribution for the system. The iterative process was found to reduce the above stated RMS error of L_2 approach with 2800 DOFs by about one order of magnitude, down to $e_2(6 \times 6 \times 6, \text{cubic+quadratic}, \text{modified-norm}) = 1.984 \times 10^{-4}$. Figures 40(c) and 40(d) show error surfaces before and after the iterative refinement process. The error reduction is clearly visible in these plots.

Fig.41 compares convergence characteristics of the L_2-norm approach with the modified-norm method. The $x-$axis shows the size of discretized problem (i.e. $\mathcal{D}$) while the $y-$axis shows RMS error in the obtained approximation. The infeasible region in the right section of this figure demarcated by the dash-dot line depicts problem sizes that are beyond the capacity of computational resources available for this work. The dashed horizontal line clearly shows that the modified-norm approach provides slightly better accuracy than the best approximation obtained using the standard L_2 approach, with about half the number of DOFs (standard L_2 approach on a $7 \times 7 \times 7$ grid with quartic polynomials assigned to interior nodes and linear polynomials to boundary nodes results in an RMS error of $e_2(7 \times 7 \times 7, \text{quartic} + \text{linear}, L_2) = 2.823 \times 10^{-4}$ and the resulting problem size is 5247 DOFs). Furthermore, both approaches (standard L_2 PUFEM and modified-norm PUFEM) require three orders of magnitude less DOFs than the standard FEM approach for the same order of accuracy.

Similar results can be obtained for systems with 4 and higher dimensional state-space. The results shown in this work were obtained on a small workstation and each iteration of the 3D problem required about 45 minutes of computational time. If attempted on a more advanced computing platform, the modified norm approach can be utilized to solve problems in higher dimensions.

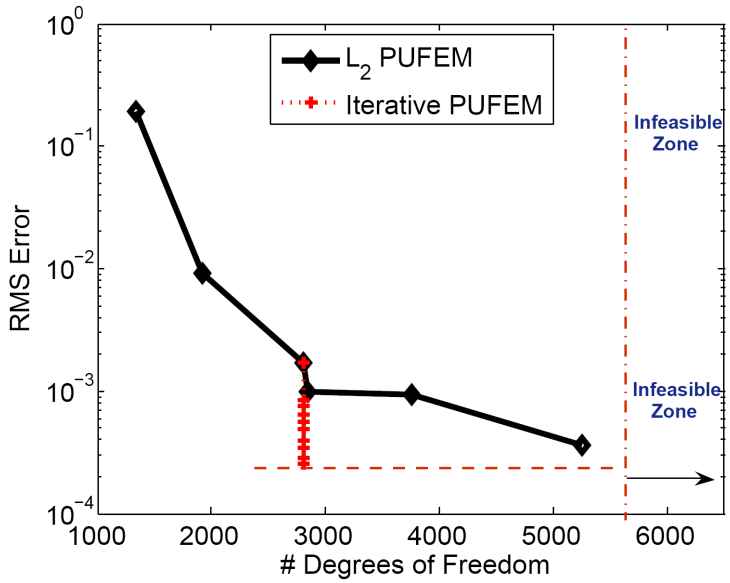


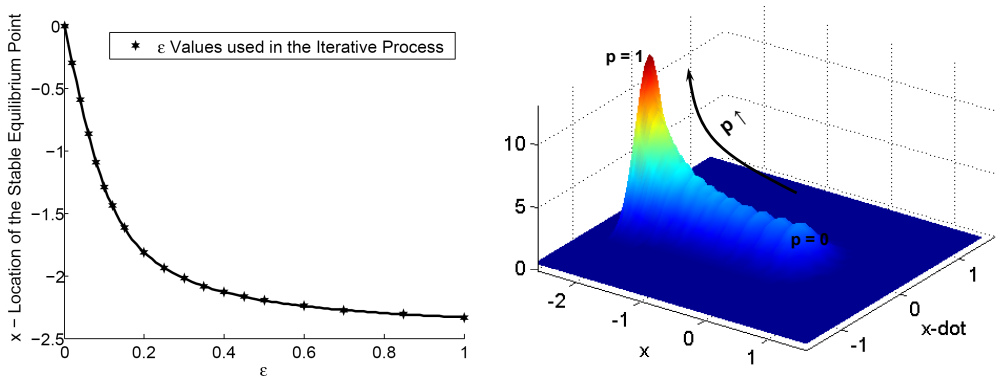
Fig. 41. Comparative convergence characteristics for the three dimensional system.

2. Space Homotopy: Results

In order to illustrate the use of homotopic approach for domain determination, we consider the following Duffing oscillator which is a modified version of the one used in the previous section (Eq.4.43):

$$\ddot{x} = -\alpha x - \beta \dot{x} + \epsilon(x^3 + \sigma) + w \tag{4.46}$$

The homotopy parameter p in the above system is ϵ, variation in which generates a family of dynamical systems of varying nonlinearity. The role of parameter σ is to shift the domain of significant portion of the pdf as p is varied. Its presence allows us to validate the fact that the proposed method can successfully track movement in the domain as ϵ changes from 0 to 1. Also, α is assumed to be positive (corresponding to a hard spring with a solitary stable equilibrium point). Fig.43 shows results for



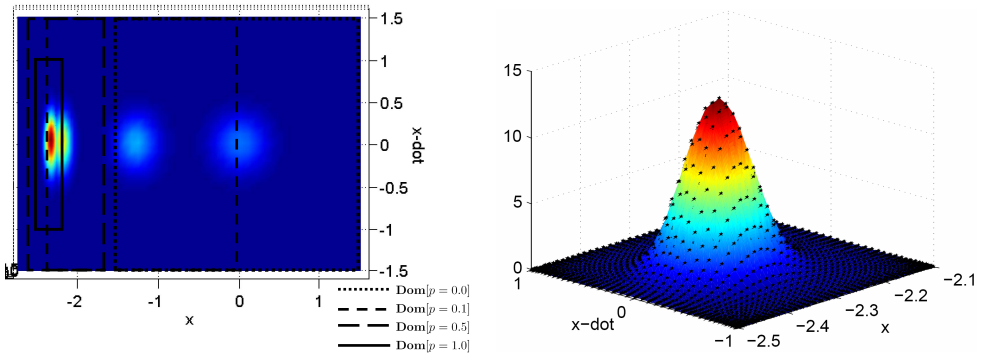
(a) Variation of the x-coordinate of the stable equilibrium with the homotopy parameter ($p = \epsilon$).

(b) Progression of iterations from a known dynamical system, $\mathcal{D}_0$, to the unknown, $\mathcal{D}_1$.

Fig. 42. Variation of the homotopy parameter, p.

the above system. In Fig.42(a), variation of x-coordinate of the stable equilibrium point is shown with ϵ. The marked values (stars) on this curve depict values of the parameter used enroute to the desired dynamical system, corresponding to $\epsilon = 1$. It was found necessary to take small steps initially (see figure) in order to satisfy the assumptions stated in lemma IV.2. In general, the nature of this variation will depend on how the homotopic parameter influences the particular system under consideration. As described in the double-loop algorithm, execution of space homotopy also involves carrying out the solution refinement process. Before one can proceed from a particular value of ϵ to the next, it must be ensured that the approximation obtained for the current ϵ has converged to within acceptable tolerance, which requires refinement iterations illustrated above.

Fig.42(b) shows the smooth variation of converged solutions obtained for each ϵ $(= p)$. It is visible in Fig.43(a) that the domain inside which dynamical system $\mathfrak{D}_p(p = 1)$ is solved is completely disconnected from the domain for the initially known system, $\mathfrak{D}_p(p = 0)$. However, via the iterative process of homotopic approximations



(a) Movement of the domain of solution along the iterative pProcedure.

(b) Comparison of the final iteration with the true solution.

Fig. 43. Illustration of space homotopy by variation of dynamical systems, $\mathcal{D}_0 - \mathcal{D}_1$.

(of which only 4 are shown in this figure), the desired result is achievable. Finally, Fig.43(b) shows the closeness of the final iteration (corresponding to $p = \epsilon = 1$, drawn surface) with the analytical result (shown with crosses), which is known in this case.

F. Summary

The ideas presented in this chapter can be summarized using a schematic of the homotopic recursive algorithm shown in Fig.44. The top-left plot in this schematic illustrates the problem of domain determination for FPE. For a general nonlinear dynamical system, it is difficult to determine the appropriate location and size of the finite sized domain on which to solve the FPE numerically. Using an extremely large sized conservative domain can lead to wastage of computational resources, possibly making higher dimensional problems infeasible. On the other hand, a domain too small can result in significant errors because it may not accommodate the entire probability density function. The method of modified norms discussed in this chapter has been shown to resolve this issue while also improving approximation accuracy.

In terms of Fig.44, the described approach obtains the desired solution shown in the top-left portion, starting with the solution for a known system. This involves setting up an iterative procedure using the best available solution for the current value of the homotopic parameter, p. Starting with $p = 0$, the available solution is successively refined using modified-norm error projection as shown in the top-right plot. In terms of the algorithm described in section C, this constitutes the inner loop, which reduces approximation error while keeping size of the discretized problem small. Once the desired accuracy is met, value of the homotopic parameter is increased, moving on to the next dynamical system, progressively towards the desired system, i.e. $p = 1$ (bottom-left plot). This constitutes the outer loop of the algorithm. The final result is a highly accurate approximation of the stationary distribution for the nonlinear dynamical system of interest. The bottom-right plot shows the error-reduction path taken by the current approach, as compared with the standard L_2 approach. Note that the only way to reduce the error with L_2 approach is to change the size of the discretized problem for given a domain size, whereas the modified-norm approach can reduce approximation error for a fixed problem size by changing the definition of the norm. The resulting advantage of the modified-norm approach is clearly apparent from the bottom-right plot in Fig.44.

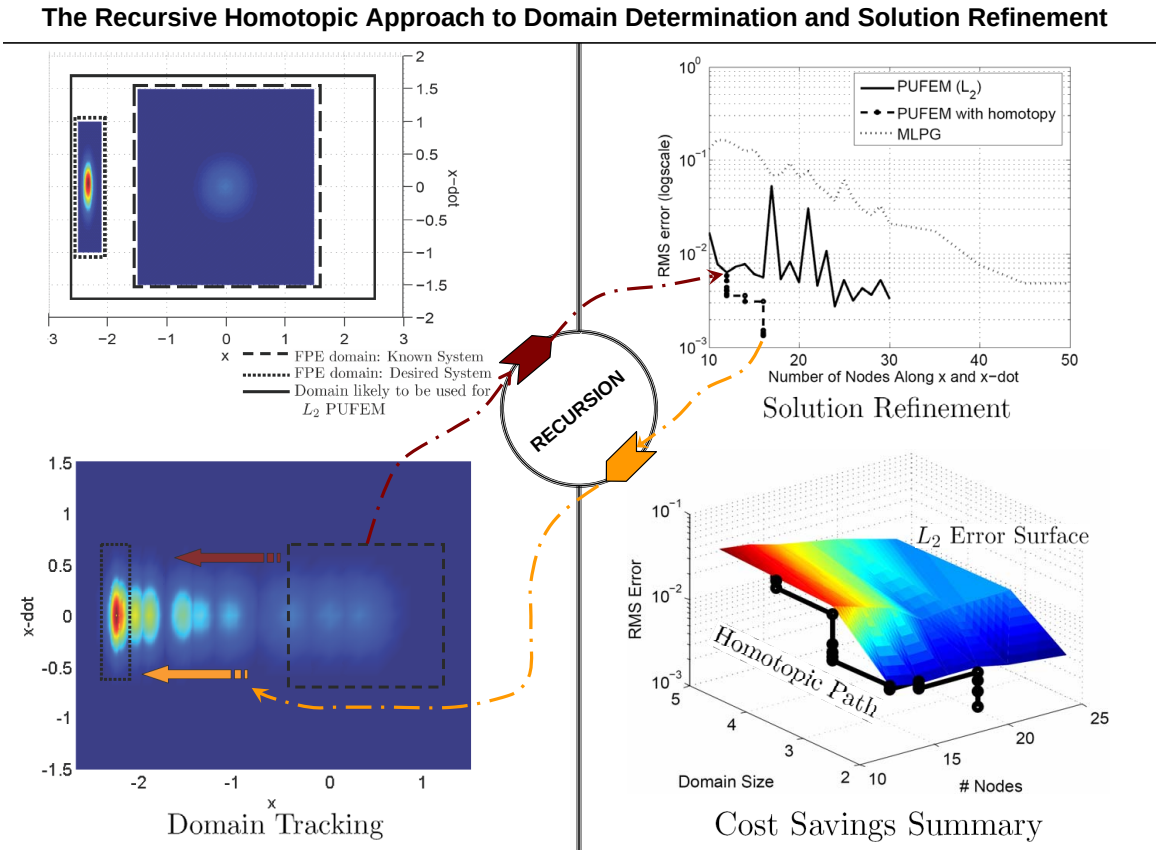


Fig. 44. A schematic of the combined process of homotopic domain tracking and iterative solution refinement.

CHAPTER V

COMPUTATIONAL STOCHASTIC OPTIMAL CONTROL

A. Introduction

One of the greatest benefits of having a robust solver for FPE lies in the field of stochastic analysis, design and control of nonlinear systems. In existing literature, Markov decision processes (MDPs) have long been one of the most widely used methods for discrete time stochastic control. However, dynamic programming equations underlying an MDP suffer from the curse of dimensionality [108, 109, 110]. Various approximate dynamic programming (ADP) methods have been proposed in the past several years for overcoming the curse [111, 110, 112, 113, 114], and can broadly be categorized under one of several "functional reinforcement learning" techniques, such as value function approximation methods, [110], policy gradient/approximation methods [112, 113] and actor-critic methods [111, 114]. These are essentially model free techniques for approximating the optimal control policy in stochastic optimal control problems. They attempt to reduce the dimensionality of the DP problem through a compact parametrization of the value function (with respect to a policy) and the policy function. The difference among these methods lies mainly in the actual parametrization employed to achieve the above goal, and range from nonlinear function approximators such as neural networks [111], to linear approximation architectures [110, 115]. The optimal control policy is learnt by repeated simulations on a dynamical system and thus, can take a long time to converge to a good policy, especially when the problem has continuous state and control spaces.

In contrast, the control methodology proposed in this chapter is model-based, and uses a finite dimensional representation of the underlying diffusion operator to

parameterize both the value function as well as the control policy in the stochastic optimal control problem. Considering the low order finite dimensional controlled diffusion operator provides a computationally efficient recursive method for obtaining progressively better control policies.

The literature on computational methods for solving continuous time stochastic control problems is relatively sparse as compared to discrete time problems. One approach uses locally consistent Markov decision processes [116], wherein the continuous controlled diffusion operator is approximated by a finite dimensional Markov chain which satisfies certain local consistency conditions, namely that its drift and diffusion coefficients match that of the original process locally. The resulting finite state MDP is solved by standard DP techniques such as value iteration and policy iteration. The method relies on a finite difference discretization and thus, can be computationally very intensive in higher dimensional spaces. In another approach [117, 118], the diffusion process is approximated by a finite dimensional Markov chain through the application of generalized cell to cell mapping [119]. However, even this method suffers from the curse of dimensionality because it involves discretization of the state space into a grid which becomes increasingly infeasible as system dimensionality grows. Also, finite difference and finite element methods have been applied directly to the nonlinear Hamilton-Jacobi-Bellman partial differential equation [120].

The method proposed here differs in that it uses policy iteration in the original infinite dimensional function space, along with a finite dimensional representation of the controlled diffusion operator in order to solve the problem. Considering a lower order approximation of the underlying operator results in a significant reduction in dimensionality of the computational problem. Utilizing the policy iteration algorithm results in typically having to solve a sequence of a few linear equations (typically < 15) before practical convergence is obtained, as opposed to solving a high dimensional

nonlinear equation if the original nonlinear HJB equation is solved.

The literature for solving deterministic optimal control problems in continuous time is relatively mature as compared to its stochastic counterpart. The method of successive approximations/policy iteration has been widely used to solve deterministic optimal control problems. A variety of methods have been proposed for solving the policy evaluation step in the policy iteration algorithm, including Galerkin-projection based methods [121, 122] and neural-network based methods [123, 124, 125, 126, 127] among others. The methodology outlined in this chapter can be viewed as an extension of this vast body of work to the stochastic optimal control problem. However, it should be noted that this extension is far from trivial because the stochastic optimal control problem, especially the control of degenerate diffusions, has pathologies that have to be treated carefully in order to devise effective computational methods. As described in chapter II, we are here interested in the feedback control of dynamical systems, i.e., in the solution of HJB equation as opposed to solving the open loop optimal control problem based on Pontryagin's principle[128] that results in a two-point boundary value problem involving the states and co-states of the dynamical system. The interested reader is directed to see references [128, 129, 130] for more on this approach of solving the optimal control problem. Also, to the best of knowledge of the author, there is no stochastic equivalent of the two-point boundary value problem that result from Pontryagin's principle applied to deterministic dynamical systems.

The rest of this chapter is arranged as follows: Section B discusses the formal adjoint of Fokker-Planck equation, namely, the Backward-Kolmogorov equation (BKE), and shows their equivalence under the condition of asymptotic stability. Section 3 outlines the computational method used to obtain a finite dimensional representation of the infinite dimensional Fokker-Planck or forward Kolmogorov (FPE), and the backward Kolmogorov Equation (BKE). In Section 4, the finite dimensional repre-

sentations of the FPE and BKE operators are used to outline a recursive procedure, based on policy iteration, to obtain the solution to the stochastic nonlinear control problem. In Section 5, the methodology is applied to various different nonlinear control problems.

B. Forward and Backward Kolmogorov Equations

In this section, the formal adjoint of FPE (also known as forward Kolmogorov equation), namely the backward Kolmogorov equation (BKE) is introduced. While FPE is central to uncertainty propagation and nonlinear filtering of dynamical systems, the latter is paramount in solving stochastic control problems. It can be shown that if FPE is asymptotically stable, then the forward and backward equations are equivalent in a sense to be made precise later in this section. Recall the SDE of Eq.2.1: for this system, backward Kolmogorov equation can be written as:

$$\frac{\partial}{\partial t}\mathcal{W}(t,\mathbf{x}) = \mathcal{L}_{\mathcal{BK}}\mathcal{W}(t,\mathbf{x}) \tag{5.1}$$

where, the backward Kolmogorov operator appearing above can be written as:

$$\mathcal{L}_{\mathcal{BK}}(\cdot) = -\left[\sum_{i=1}^{\mathrm{N}} f_i(t,\mathbf{x})\frac{\partial}{\partial x_i} + \frac{1}{2}\mathbf{g}(t,\mathbf{x})Q\mathbf{g}^{\mathrm{T}}(t,\mathbf{x})\sum_{i,j=1}^{\mathrm{N}}\frac{\partial^2}{\partial x_i \partial x_j}\right](\cdot) \tag{5.2}$$

For the sake of simplicity, we consider the case of an autonomous single dimensional system in this section to establish the equivalence between FP and BK operators. The results can easily be generalized to multidimensional systems:

$$dx = f(x)dt + g(x)dB, \tag{5.3}$$

For the above system, the associated forward Kolmogorov equation or Fokker-Planck

Equation (FPE) is:

$$\frac{\partial \mathcal{W}}{\partial t} = -\frac{\partial (f\mathcal{W})}{\partial x} + \frac{1}{2}\frac{\partial^2 (g^2\mathcal{W})}{\partial x^2}, \tag{5.4}$$

and the backward Kolmogorov equation (BKE):

$$\frac{\partial \mathcal{W}}{\partial t} = f\frac{\partial \mathcal{W}}{\partial x} + \frac{g^2}{2}\frac{\partial^2 \mathcal{W}}{\partial x^2}. \tag{5.5}$$

FPE governs the evolution of state probability density function of the stochastic system Eq.5.3 forward in time while the BKE governs its evolution backward in time. FPE is said to be asymptotically stable if there exists a unique pdf $\mathcal{W}_\infty$ such that any initial condition, deterministic or probabilistic, decays to the pdf $\mathcal{W}_\infty$ as time goes to infinity. In the following, the equivalence of BKE and FPE is shown, under the condition of asymptotic stability of the FPE.

Since FPE is assumed to be asymptotically stable, let the pdf at any time be given as follows:

$$\mathcal{W}(t,x) = \mathcal{W}_\infty(x)\bar{\mathcal{W}}(t,x), \tag{5.6}$$

i.e., as a product of the stationary pdf and a time varying part. Substituting into FPE, we obtain:

$$\frac{\partial \mathcal{W}_\infty \bar{\mathcal{W}}}{\partial t} = -\frac{\partial (f\mathcal{W}_\infty \bar{\mathcal{W}})}{\partial x} + \frac{1}{2}\frac{\partial^2 (g^2\mathcal{W}_\infty \bar{\mathcal{W}})}{\partial x^2}, \tag{5.7}$$

Expanding various terms in the above equation, we obtain the following identity:

$$\mathcal{W}_\infty \frac{\partial \bar{\mathcal{W}}}{\partial t} = -f\mathcal{W}_\infty \frac{\partial \bar{\mathcal{W}}}{\partial x} + \frac{1}{2}g^2\mathcal{W}_\infty \frac{\partial^2 \bar{\mathcal{W}}}{\partial x^2} + \frac{\partial \bar{\mathcal{W}}}{\partial x}\frac{\partial (g^2\mathcal{W}_\infty)}{\partial x}. \tag{5.8}$$

Now, make use of the fact that $\mathcal{W}_\infty$ is the stationary solution, i.e.,

$$\frac{\partial \mathcal{W}_\infty}{\partial t} = -\frac{\partial (f\mathcal{W}_\infty)}{\partial x} + \frac{1}{2}\frac{\partial^2 (g^2\mathcal{W}_\infty)}{\partial x^2} = 0, \tag{5.9}$$

Next, Consider the term $\left[-f\mathcal{W}_\infty \frac{\partial \bar{\mathcal{W}}}{\partial x} + \frac{\partial \bar{\mathcal{W}}}{\partial x}\frac{\partial (g^2\mathcal{W}_\infty)}{\partial x}\right]$ in Eq.5.8. Note that from Eq.5.9:

$$\frac{\partial}{\partial x}\left(-f\mathcal{W}_\infty + \frac{1}{2}\frac{\partial g^2\mathcal{W}_\infty}{\partial x}\right) = 0, \tag{5.10}$$

and hence,

$$-f\mathcal{W}_\infty + \frac{1}{2}\frac{\partial (g^2\mathcal{W}_\infty)}{\partial x} = const. \tag{5.11}$$

The above relation holds due to boundary conditions of the pdf at infinity. In fact, if we assume that the invariant distribution is well-behaved such that the terms $f\mathcal{W}_\infty \to 0$, and $\frac{\partial g^2\mathcal{W}_\infty}{\partial x} \to 0$ as $x \to \pm\infty$, i.e, if the stationary distribution is not heavy-tailed, then

$$-f\mathcal{W}_\infty + \frac{1}{2}\frac{\partial (g^2\mathcal{W}_\infty)}{\partial x} = 0, \tag{5.12}$$

and hence, substituting the above into Eq. 5.8, and dividing throughout by $\mathcal{W}_\infty(x)$, it follows that

$$\frac{\partial \bar{\mathcal{W}}}{\partial t} = f\frac{\partial (\bar{\mathcal{W}})}{\partial x} + \frac{g^2}{2}\frac{\partial^2 \bar{\mathcal{W}}}{\partial x^2}, \tag{5.13}$$

i.e, the time varying part of the pdf follows BKE.

Under the condition of asymptotic stability of FPE, and in view of the above development, BKE and FPE are equivalent in the following sense: if $\phi(x)$ is an eigenfunction of the BK operator $\mathcal{L}_{\mathcal{BK}}(.)$ with corresponding eigenvalue λ, then $\mathcal{W}_\infty(x)\phi(x)$ is an eigenfunction of the FP operator with the same eigenvalue λ. Hence, the knowledge of eigenfunctions of the FP operator along with the stationary distribution of FPE is sufficient to determine the eigenstructure of BKE, and vice-versa. This will be useful in the next section when we consider policy iteration for solving HJB.

C. Nonlinear Stochastic Control

In this section, an iterative approach for solving the Hamilton-Jacobi-Bellman (HJB) equation is presented for nonlinear stochastic dynamical systems. The problem of concern, as described in chapter II is known as the $\mathcal{H}_2$ optimal control problem. The infinite horizon scenario results in a stationary, i.e., time invariant control law. However, for the case of systems with additive noise, i.e., when $\mathbf{g}(\mathbf{x})$ is independent of $\mathbf{x}$, as would be the case for instance in the LQG problem, the cost-to-go is undefined since the trajectories of the system never decay to zero. In that case, discounting (via β in Eq.2.10) is a practical step to ensure a bounded cost-to-go function. The discount factor may also be interpreted as an artificial finite horizon over which the control law tries to optimize the system performance. For this problem, the optimal control law, $\mathbf{u}^*(\mathbf{x})$ is known to be given in terms of a *value function* $V^*(\mathbf{x})$ as follows:

$$\mathbf{u}^*(\mathbf{x}) = -\frac{1}{2}\mathbf{R}^{-1}\mathbf{h}^{\mathrm{T}}\frac{\partial V^*}{\partial \mathbf{x}}(\mathbf{x}) \tag{5.14}$$

where, the value function $V^*(x)$ solves the following well known (stationary) Hamilton-Jacobi Bellman equation:

$$\frac{\partial V^*}{\partial \mathbf{x}}^{\mathrm{T}}\mathbf{f} - \frac{1}{4}\frac{\partial V^*}{\partial \mathbf{x}}\mathbf{h}\mathbf{R}^{-1}\mathbf{h}^{\mathrm{T}}\frac{\partial V^*}{\partial \mathbf{x}} + \frac{1}{2}\mathbf{g}\mathbf{Q}\mathbf{g}^{\mathrm{T}}\frac{\partial^2 V^*}{\partial \mathbf{x}^2} + l - \beta V^* = 0, \quad V^*(\mathbf{0}) = 0 \tag{5.15}$$

As mentioned before, the above framework solving for the optimal control law, $u^*(x)$ is known as the $\mathcal{H}_2$ control paradigm. Note that obtaining the optimal control within the above framework requires solving a nonlinear PDE (Eq.5.15), which is in general a very difficult problem. It is however possible to restructure the above equations so that the central problem can be reduced to solving a sequence of linear PDEs, which is a much easier proposition. This is achieved via substitution of Eq.5.14 into the quadratic term involving the gradient of the value function in HJB (Eq.5.15).

Doing so gives us the following equivalent form the the HJB:

$$\frac{\partial V^*}{\partial \mathbf{x}}^{\mathrm{T}} (\mathbf{f} + \mathbf{h}\mathbf{u}) + \frac{1}{2}\mathbf{g}\mathbf{Q}\mathbf{g}^{\mathrm{T}} \frac{\partial^2 V^*}{\partial \mathbf{x}^2} + l + \|\mathbf{u}\|^2_{\mathbf{R}} - \beta V^* = 0 \tag{5.16}$$

The above form of HJB is known as the *generalized* HJB. Note that the substitution discussed above converted the nonlinear PDE to a linear PDE in the value function, $V^*(\mathbf{x})$. Eq.5.16 forms the core of a policy iteration algorithm in which the optimal control law can be obtained by iterating upon an initial stabilizing control policy as follows:

1. ***Let*** $u^{(0)}$ *be an initial stabilizing control law (policy) for the dynamical system (Eq.2.9), i.e., the FPE corresponding to the closed loop under* $u^{(0)}$ *is asymptotically stable.*

2. ***For*** $i = 0$ *to* ∞

   - **Solve for** $V^{(i)}$ **from:**

$$\frac{\partial V^{(i)}}{\partial \mathbf{x}}^{\mathrm{T}} (\mathbf{f} + \mathbf{h}\mathbf{u}^{(i)}) + \frac{1}{2}\mathbf{g}\mathbf{Q}\mathbf{g}^{\mathrm{T}} \frac{\partial^2 V^{(i)}}{\partial \mathbf{x}^2} + l + \|\mathbf{u}^{(i)}\|^2_{\mathbf{R}} - \beta V^{(i)} = 0 \tag{5.17}$$

   - **Update** policy **as:**

$$\mathbf{u}^{(i+1)}(\mathbf{x}) = -\frac{1}{2}\mathbf{R}^{-1}\mathbf{h}^{\mathrm{T}} \frac{\partial V^{(i)}}{\partial \mathbf{x}}(\mathbf{x}) \tag{5.18}$$

3. ***End*** **[Policy Iteration]**

The convergence of the above algorithm has been proven for the deterministic case ($\mathbf{Q} = 0$) in references [121, 122] and the result holds for the stochastic case as well[131]. However, conditions for existence of classical solutions of the linear elliptic PDE in the policy evaluation step (Eq.5.17), and asymptotic stability of the closed loop systems under the resulting control policies, in the sense that the associated FPE

is stable, have not been obtained to the best knowledge of the author. In this chapter, we assume that the policy evaluation step admits a classical solution and that the sequence of control policies generated asymptotically stabilize the closed loop system. In fact, these assumptions are borne out of the numerical examples considered in the next section. The greatest advantage of using the policy iteration algorithm over directly solving the nonlinear HJB equation is that the algorithm typically converges in a very few number of steps (≤ 15).

Let us take a closer look at Eq.5.17: writing it in operator form, we have:

$$(\mathcal{L} - \beta)V = q \tag{5.19}$$

where,

$$\mathcal{L} = (\mathbf{f} + \mathbf{hu})^{\mathrm{T}} \frac{\partial}{\partial \mathbf{x}} + \frac{1}{2}\mathbf{g}\mathbf{Q}\mathbf{g}^{\mathrm{T}} \frac{\partial^2}{\partial \mathbf{x}^2} \tag{5.20}$$

$$q = -(l + \|\mathbf{u}\|^2_{\mathbf{R}}) \tag{5.21}$$

Note that the operator $\mathcal{L}(\cdot)$ is identical to the BK operator of the closed loop under control policy $\mathbf{u}$. In other words, the policy iteration step reduces to recursively solving BKE for the controlled diffusion process. From section B, we saw the equivalence of FPE and BKE, under the conditions considered therein. Since a robust algorithm for FPE has already been developed, the policy iteration can easily be implemented to solve for the optimal control law. The unknown in this scenario is not a pdf, but the value function, V. As in Eq.3.6, an approximation for V can be written in terms of shape functions as follows:

$$\widehat{V}(\mathbf{x}) = \sum_{i=1}^{\mathcal{D}} a_i \Psi_i(\mathbf{x}) \tag{5.22}$$

where, $\Psi_i(\mathbf{x})$ could be local or global shape functions depending on the nature of

approximation being used. Then, following the Galerkin approach, residual error resulting from substitution of Eq.5.22 into the generalized HJB (Eq.5.17) can be minimized as follows:

$$\sum_{i=1}^{\mathcal{D}} a_i \int_{\Omega} (\mathcal{L} - \beta)[\Psi_i(\mathbf{x})]\Psi_j(\mathbf{x}) d\mathbf{x} = \int_{\Omega} q\Psi_j d\mathbf{x}, \quad j = 1, 2, \ldots, \mathcal{D} \tag{5.23}$$

Eq.5.23 is equivalent to the following linear system of algebraic equations in a_i:

$$\mathbf{K}\mathbf{a} = \mathbf{f} \tag{5.24}$$

where

$$\mathbf{K}_{ij} = \int_{\Omega} (\mathcal{L} - \beta)[\Psi_j]\Psi_i d\mathbf{x} \tag{5.25}$$

$$\mathbf{f}_i = \int_{\Omega} q\Psi_i d\mathbf{x} \tag{5.26}$$

Note that

$$\mathbf{K} = \mathbf{K}_{\mathcal{BK}} - \beta\mathbf{I}, \tag{5.27}$$

where $\mathbf{K}_{\mathcal{BK}}$ is a finite dimensional representation of the backward Kolmogorov operator. Some notes on admissibility of shape functions (Ψ_i) are due at this point. Note that the BK operator involves only derivatives of the unknown function, $V(x)$, and no terms containing the function itself. It is thus clear that when the discount factor (β) is zero, it is not possible to use a complete set of polynomial shape functions without making the stiffness matrix $\mathbf{K}$ singular ($\Psi \equiv 1$ causes a rank deficiency of 1). In theory, a nontrivial β allows us to include $\Psi \equiv 1$ in the basis set; but in practice it would require a large discount factor before $\mathbf{K}$ is reasonably well-conditioned for inversion; which in turn would make the solution far from being optimal because the control policy would become very "short sighted".

D. Numerical Examples

1. Two Dimensional System - Van der Pol Oscillator

Consider the following two dimensional stochastic Van der Pol oscillator:

$$\ddot{x} + v_1 x + v_2(1 - x^2)\dot{x} = u + gw(t) \tag{5.28}$$

with, $v_1 = 1, v_2 = -1, g = 1$. The objective is to minimize the following cost function:

$$J(x_0, \dot{x}_0) = \mathrm{E}\left[\int_0^\infty \frac{1}{2}\left[(x^2 + \dot{x}^2) + u^2\right] dt\right] \tag{5.29}$$

The approximation was constructed using a complete set of global polynomials up to fourth order (using polynomials up to the sixth order did not provide significant improvement in the approximation accuracy). It is notable that in the above formulation, no direct means of enforcing state or control input constraints have been used. In order to obtain reasonable bounds on the control input, the cost function (Eq.5.29) was scaled so as to modify the generalized HJB (Eq.5.17) as follows (in addition to including the discount factor (β)):

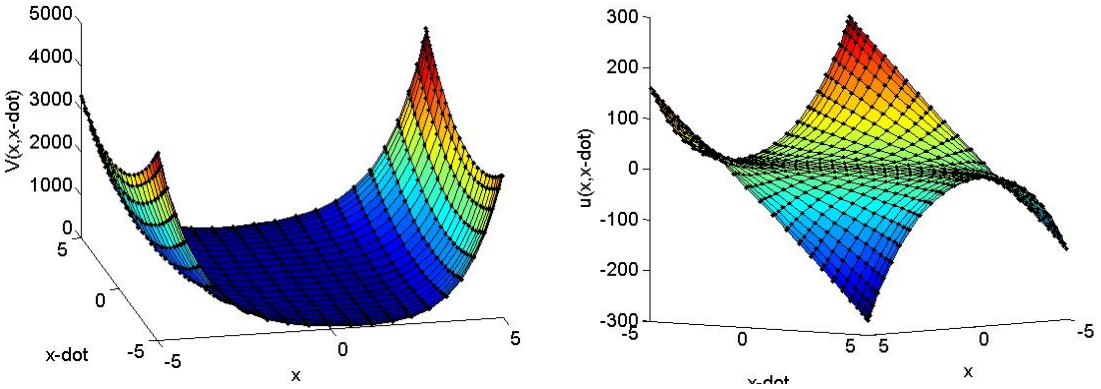
$$\frac{\partial V^{(i)}}{\partial \mathbf{x}}^{\mathrm{T}} (\mathbf{f} + \mathbf{h}\mathbf{u}^{(i)}) + \frac{1}{2}\mathbf{g}\mathbf{Q}\mathbf{g}^{\mathrm{T}} \frac{\partial^2 V^{(i)}}{\partial \mathbf{x}^2} + \beta V = -e^{\gamma}(l + \|\mathbf{u}^{(i)}\|_{\mathbf{R}}) \tag{5.30}$$

i.e.,

$$l(\cdot) \mapsto e^{\gamma} l(\cdot) \tag{5.31}$$

$$\mathbf{R} \mapsto e^{\gamma}\mathbf{R}, \quad \gamma < 0 \tag{5.32}$$

Fig.45 shows converged results obtained for the above system. A linear starting policy was used to begin the iteration process, i.e., $u^{(0)} = K\dot{x}$, with $K = -20$. Tuning parameters $\beta = 0.05$ and $\gamma = -0.15$ were found to give acceptable results after about 13 iterations of the control policy. Fig.46(a) shows the system response without a



(a) Converged value function surface.

(b) Converged control input surface. The maximum control required was modulated using the tuning parameter γ.

Fig. 45. Converged results for the stochastic Van der Pol oscillator.

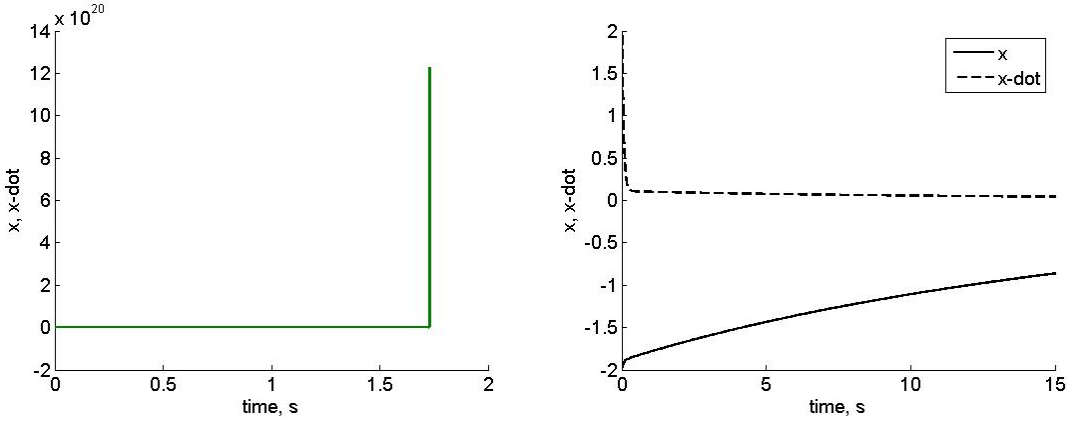
control input. Clearly, the states diverge. Fig.46(b) shows the system response with the initial stabilizing control. For both these results, nominal initial conditions were used inside the domain of operation, $\Omega = [-5, 5] \times [-5, 5]$. Fig.47 shows the optimal state trajectories obtained and the optimal control law.

2. Two Dimensional System - Duffing Oscillator

Next consider the following hard spring stochastic duffing oscillator:

$$\ddot{x} - x + \epsilon x^3 = u + gw(t) \tag{5.33}$$

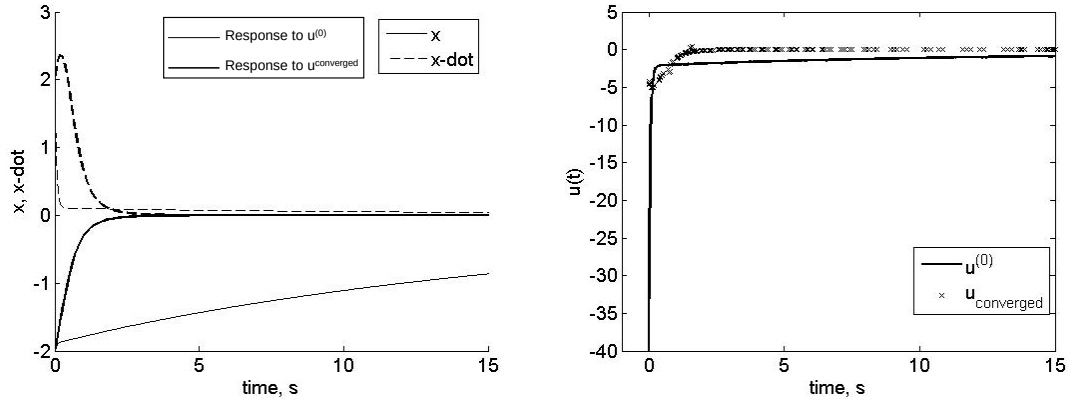
The above system represents the hard-spring case ($\epsilon = +1.5$) with an unstable equilibrium at $(x, \dot{x}) = (0, 0)$ and stable equilibria at $(\sqrt{\epsilon}, 0)$ and $(-\sqrt{\epsilon}, 0)$. The cost function to be minimized is the same as for the previous system, and we use a linear stabilizing control given by: $u^{(0)} = -30\dot{x}$ to begin policy iteration. Similar convergence characteristics as for the Van der Pol oscillator were obtained and are shown



(a) Uncontrolled system response for Eq.5.28.

(b) System response to initial stabilizing control, $u^{(0)} = -20\dot{x}$.

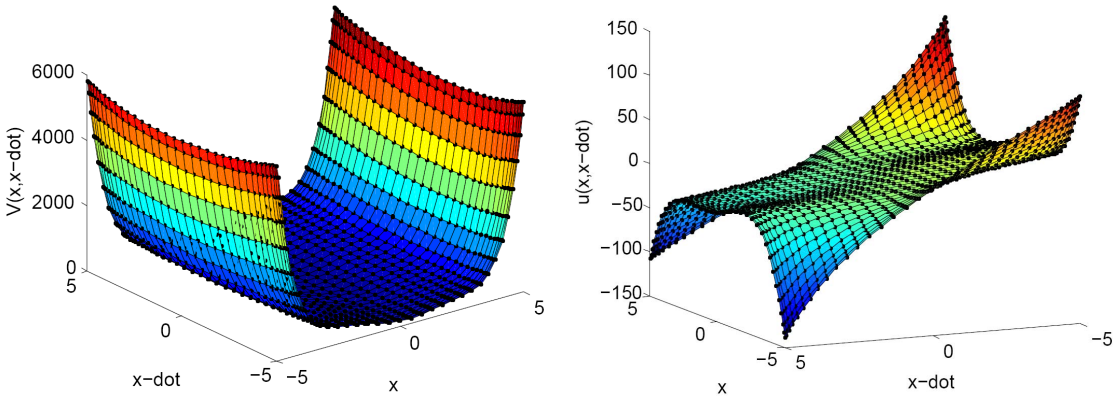
Fig. 46. System response of the stochastic Van der Pol oscillator I.



(a) Optimal path obtained, shown alongside response to $u^{(0)}$.

(b) Optimal control law converged upon and $u^{(0)}$.

Fig. 47. System response of the stochastic Van der Pol oscillator II.



(a) Converged value function surface.

(b) Converged control input surface. The maximum control required was modulated using the tuning parameter γ.

Fig. 48. Converged results for the stochastic hard spring Duffing oscillator.

in Figs. 48-50.

3. Four Dimensional System - Missile Pitch Control Autopilot

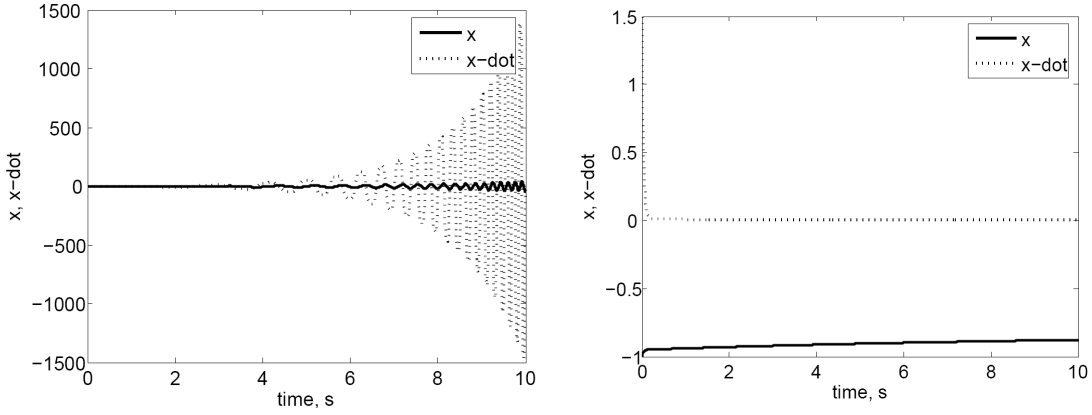
Finally, we consider a four-dimensional system that models the pitch motion control for a missile autopilot. The deterministic version of this model was considered in by Beard et al. [121, 122]. In the present model, noise terms have been added to all kinetic level equations:

$$\dot{q} = \frac{M_y}{I_y} + g_q w_1(t), \qquad M_y = C_m(\alpha, \delta) QSd \tag{5.34}$$

$$\dot{\alpha} = \frac{\cos^2 \alpha}{mU} F_z + q + g_\alpha w_2(t), \qquad F_z = C_n(\alpha, \delta) Sd \tag{5.35}$$

$$\ddot{\delta} = -2\zeta\omega_n \dot{\delta} + \omega_n^2(\delta_c - \delta) + g_\delta w_3(t) \tag{5.36}$$

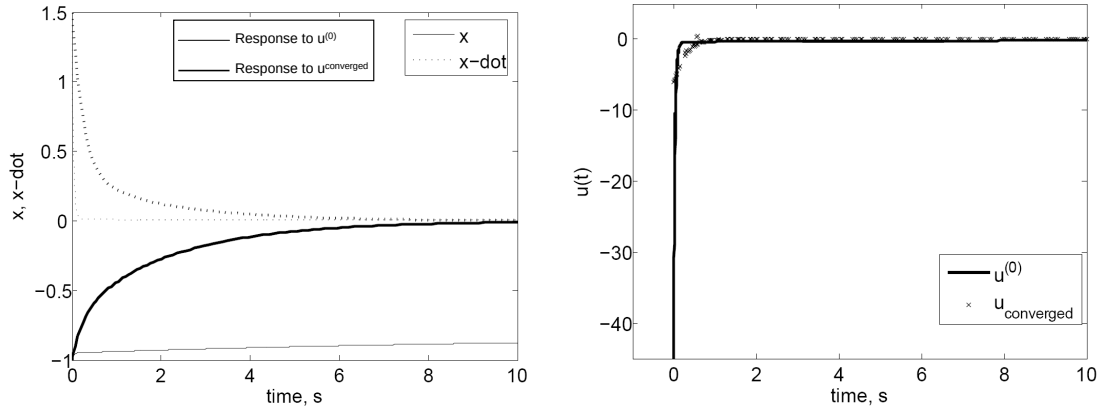
The control input in the above equations is $\delta_c(t)$, which is the commanded tail-fin deflection, while $\delta(t)$ denotes the actual tail-fin deflection. Pitch rate and angle of



(a) Uncontrolled system response of Eq.5.33.

(b) System response to initial stabilizing control, $u^{(0)} = -30\dot{x}$.

Fig. 49. System response of the stochastic Duffing oscillator I.



(a) Optimal path obtained, shown alongside response to $u^{(0)}$.

(b) Optimal control law converged upon and $u^{(0)}$.

Fig. 50. System response of the stochastic Duffing oscillator II.

attack are denoted by q and α respectively. $w_q(t), w_\alpha(t)$ and $w_\delta(t)$ are independent components of a 3 dimensional white noise process. The aerodynamic coefficients are given by the following nonlinear functions[122]:

$$C_m(\alpha,\delta) = b_1\alpha^3 + b_2\alpha|\alpha| + b_3\alpha + b_4\delta \tag{5.37}$$

$$C_n(\alpha,\delta) = a_1\alpha^3 + a_2\alpha|\alpha| + a_3\alpha + a_4\delta \tag{5.38}$$

Values of the various constants can be found in Beard et al. [122]. The cost function to be minimized is given by:

$$J(\mathbf{x}_0) = \mathrm{E}\left[\int_0^\infty \frac{1}{2}(q-q_{ss})^2 + 25(\alpha-\alpha_{ss})^2 + \left(\frac{F_z - F_{zss}}{m}\right)^2 + \frac{1}{2}\|u-\delta_{ss}\|^2_{\mathbf{R}}\right] \tag{5.39}$$

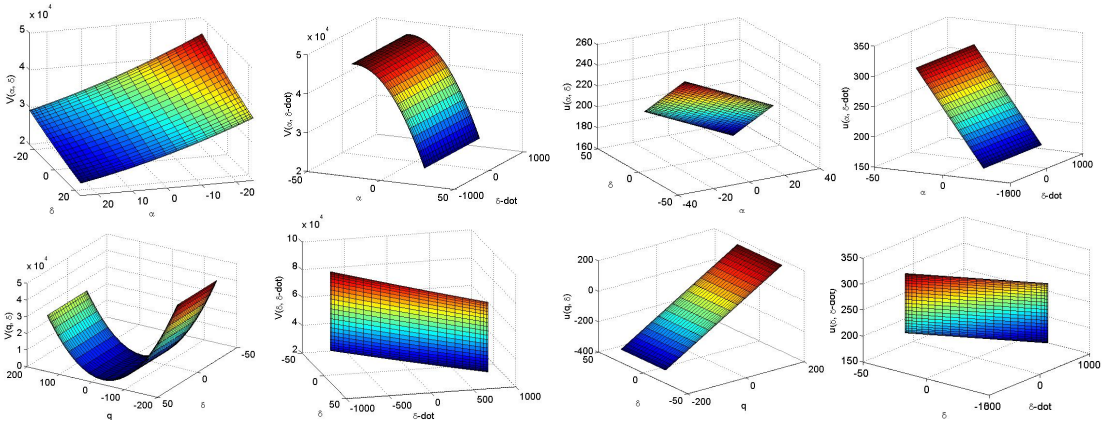
where the 'ss' subscripted quantities denote steady state values. We use the same starting stabilizing control as that used in Beard et al. [122]:

$$u^{(0)} = 0.08(q+q_{ss}) + 0.38(\alpha+\alpha_{ss}) + 0.37\frac{F_z}{m} \tag{5.40}$$

A complete basis of global quadratic shape functions was used to approximate the value function in the policy iteration algorithm. Various sections of the converged value function and optimal control surface are shown in Fig.51. The optimal trajectories are shown in Fig.51. Fig.52 illustrates system response to the optimal control law obtained. Note that pitch rate settles to zero and angle of attack acquires the desired steady state value.

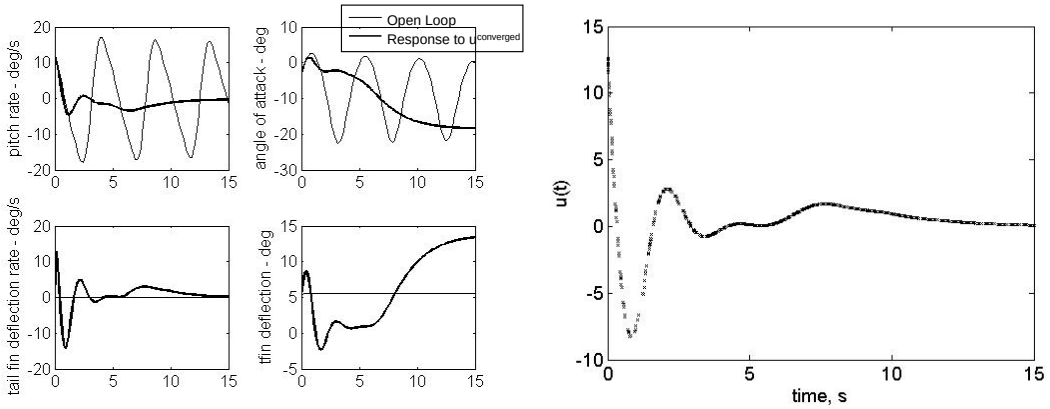
4. Notes on Modal Analysis

It is easy to see that any constant function, $\Psi \equiv c$, is a stationary solution of the BK equation, and an eigenfunction of the BK operator with trivial eigenvalue. It was mentioned in section C that the constant basis function, $\Psi \equiv 1$, is not admissible while solving for the coefficients, a_i. However, when considering the eigenvalue problem for



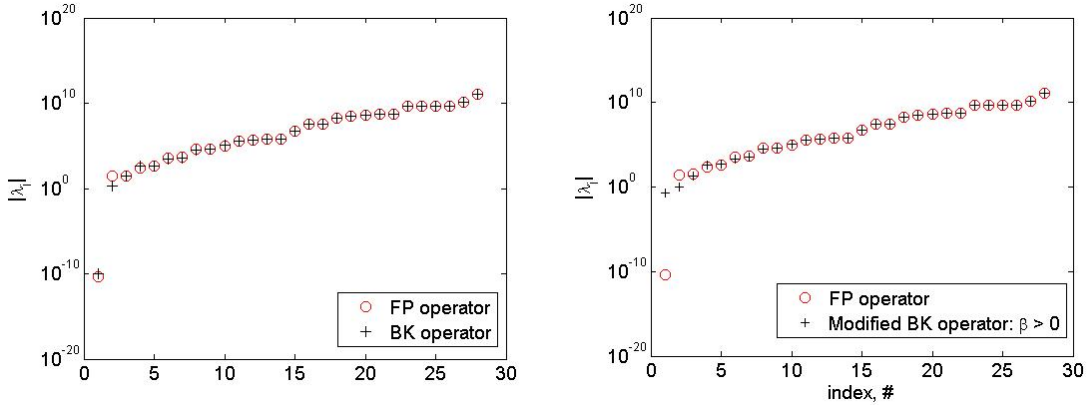
(a) Converged value function surface. (b) Converged control input surface.

Fig. 51. Converged results for the missile pitch controller, showing various sections of the four-D state space.



(a) Optimal path obtained, shown alongside open loop trajectory. (b) Optimal control law obtained.

Fig. 52. System response of the missile pitch controller.



(a) Comparison of spectra of the BK and FP operators for the Van der Pol oscillator.

(b) The modified BK operator does not contain a trivial eigenvalue in its spectrum ($\beta = 0.02$).

Fig. 53. Spectra of the BK, modified BK and FP operators for the Van der Pol oscillator.

the BK operator, $\mathcal{L}\psi = \lambda\psi$, it is imperative that the constant basis function be included in the set; because the stationary solution of the BK equation is nothing but the constant function. The modified BK operator (with $\beta > 0$), however, does not admit an eigenfunction with a trivial eigenvalue: for example, see Fig.53 for the Van der Pol oscillator.

E. Summary

In this chapter, backward Kolmogorov equation was utilized to formulate a policy iteration technique for solving closed-loop stochastic optimal control problems. The computational methodology was also tested on a number of test cases where it was shown to have satisfactory performance.

CHAPTER VI

NONLINEAR FILTERING

A. Introduction

This chapter considers the problem of nonlinear filtering described in chapter II for continuous dynamical systems and discrete measurement updates. The developed filter is based on the robust and fast solver for Fokker-Planck equation developed in chapter III, while the measurement update is performed as a weak form of Bayes rule. Nonlinear filtering has been a subject of intense research over past decades with the objective of developing accurate filters for nonlinear systems and/or sparse measurement scenarios. The Kalman filter [132] and extended Kalman filter (EKF) have been the standard tools for state estimation since the 1960's. While Kalman filter is an optimal state estimator for linear systems, the extended Kalman filter deals with nonlinear systems by considering first order Taylor series expansion of underlying system dynamics and measurement model. This is also known as Gaussian closure because the state probability density function is approximated by a Gaussian pdf at all times, fully characterized by its mean vector and covariance matrix. EKF can lead to poor performance if the degree of nonlinearity is high, the quality of measurements is poor, or the time duration between measurement updates is long.

The true description of nonlinear state estimation is given by Fokker-Planck equation for the propagation part, in conjunction with the Bayes measurement update rule. Unfortunately, as amply observed in this dissertation, FPE is a partial differential equation which is equivalent to an infinite dimensional system of nonlinear ordinary differential equations in the moments of the state pdf - a practically intractable problem [116, 133, 134, 20]. An extensive amount of research has been

conducted leading to a broad gamut of approximate filters for nonlinear systems - most notable being the unscented Kalman filter of Julier and Uhlmann, particle filters, Gaussian sum filters, higher order moment filters and sigma-point filters for discrete systems. In sigma-point filters, the various terms of interest for the optimal Kalman recursion are obtained using a semi-global technique known as stochastic or statistical linearization [37, 135]. The method essentially consists of discretizing the domain of the random variable into a set of weighted sigma points and transforming these through the nonlinear map in order to obtain the distribution characteristics of the transformed random variable. The various sigma point algorithms differ from each other in their choice of the sigma points, for instance the unscented Kalman filter (UKF) [136, 137], the central difference Kalman filter (CDKF) [138], the square root form UKF and CDKF [139] and so on. The Gaussian sum filter uses Gaussian pdfs as basis functions to solve moment propagation equations derived from FPE [140]. They have been especially useful in capturing multi-modal behavior in dynamical systems.

In this chapter, continuous dynamical systems are considered with measurements made after discrete time gaps. This allows us to use the elegant formalism of Fokker-Planck equations to evaluate the prediction terms in the nonlinear filtering recursions. In 1997, Beard presented a fast technique for solving Fokker-Planck equation using global shape functions, coupling it with discrete cosine transforms to obtain the so-called nonlinear projection filter [141]. Unfortunately, due to multiple problems related to solving FPE, the application of this filter was limited to a restricted class of systems in two dimensional state-space.

Here, the semianalytical method of solving FPE in near real-time developed in chapter III is utilized to develop a filter for systems with high nonlinearity and long durations of propagation in between measurement updates. Measurement updates are implemented as a variational (weak) form of the Bayes update rule. The methodology

developed in this work provides two major advantages:

1. The preprocessing numerical step of meshless discretization of the FP operator followed by spectral analysis and spurious mode rejection gives the ability to generate transient FPE response in near real-time, independent of initial probability distribution of the state. This provides the missing link in filtering theory in the form of a fast nonlinear propagator.

2. The equation error of transient response obtained using the modal basis reduces with time, implying that propagation is more accurate if time between measurements is longer. This counter intuitive result makes the developed solver ideally suited for filtering problems with sparse measurements.

B. Nonlinear Filter Based on FPE

Nonlinear state estimation as described in chapter II involves two key steps: (1) obtaining the prior pdf by integrating the associated FPE and (2) obtaining the posterior pdf by incorporating new measurements via the Bayes rule. As a convention, we will assume that the pdf at initial time is a "posterior," i.e. the filter is put into motion by state propagation.

The various steps of the current filter are presented in tabular form in Table VII. As previously mentioned the model comprises of nonlinear continuous dynamics and discrete measurements. Further details are can be obtained from description of Problem II.3 in chapter II. Before the filter can be initiated, it is required to execute the preprocessing steps of the semianalytical algorithm described in chapter III. This includes meshless discretization of the associated FP operator on a finite sized domain followed by spectral analysis and spurious mode rejection. In Table VII, it is assumed that these steps have been completed, and an admissible set of eigenfunctions, $\mathcal{A}$ is at

hand to describe the transient FPE response. The initial state density is assumed to be given ($\mathcal{W}(\tilde{\mathbf{x}}_0)$, where $\tilde{\mathbf{x}}$ denotes measured state), or can be found via initial state estimation techniques. Typically, this is a Gaussian pdf.

1. Filter Initialization

Following completion of preprocessing steps mentioned above, the conditional state pdf can now be parameterized by coefficients of the admissible eigenfunctions, $\phi^\star(\mathbf{x}) \in \mathcal{A}$ as follows:

$$\widehat{\mathcal{W}}(t, \mathbf{x}|\mathcal{Y}) = \sum_{i=1}^{\text{card}(\mathcal{A})} \acute{a}_i(t)\phi^\star(\mathbf{x}), \quad \phi^\star(\cdot) \in \mathcal{A} \tag{6.1}$$

Note change in terminology: the above equation uses $\acute{a}_i(t)$ instead of $a_i'(t)$ as in chapter III due to appearance of other superscripts starting from Eq.6.2 below. This holds for all similarly accented symbols throughout this chapter. Because of the above modal expansion, the entire filtering problem can now be formulated around propagation and update rules for the coefficients $\acute{a}_i(t)$. The first step is to determine the initial conditions $\acute{a}_i(t_0)$, which can be found by a standard least squares procedure of approximating the (given) function $\mathcal{W}(\tilde{\mathbf{x}}_0)$ with modal basis functions $\phi^\star \in \mathcal{A}$, leading to the following approximation:

$$\widehat{\mathcal{W}}(\mathbf{x}_0|\mathcal{Y}^0) = \sum_{i=1}^{\text{card}(\mathcal{A})} \acute{a}_i^+(t_0)\phi^\star(\mathbf{x}) \tag{6.2}$$

The superscript "+" signifies that the above pdf is a "posterior," in accordance with the assumption made earlier that the filter is set into motion by propagation rather than update. In other words, if there is a measurement update involved at $t = 0$, it is assumed to be built into the function $\mathcal{W}(\tilde{\mathbf{x}}_0)$.

Table VII. FPE based nonlinear filter

Model	$d\mathbf{x} = \mathbf{f}(t,\mathbf{x})dt + \mathbf{g}(t,\mathbf{x})d\mathbf{B}(t)$, $d\mathbf{B}(t) \sim N(\mathbf{0},\mathbf{Q})$ $\mathbf{y}_k = \mathbf{h}(\mathbf{x}_k) + \mathbf{v}_k$, $\mathbf{v}_k \sim N(\mathbf{0},\mathbf{R})$ Measurement updates at: $\{t_1, t_2, \ldots, t_m\}$
Initialize	$\mathcal{W}(\mathbf{x}_0\|\mathcal{Y}^0) = \mathcal{W}(\tilde{\mathbf{x}}_0)$ (assumed given) $\widehat{\mathcal{W}}(\mathbf{x}_0\|\mathcal{Y}^0) = \sum_i \acute{a}_i^+(t_0)\phi_i^\star(\mathbf{x})$, $\phi_i^\star(\mathbf{x}) \in \mathcal{A}$ Set $k = 1$
Propagate	$\acute{a}_i^-(t_k) = \left(\acute{a}_i^+(t_{k-1}) + \frac{\acute{l}_i}{\lambda_i^\star}\right)\exp(\lambda_i^\star t_k) - \frac{\acute{l}_i}{\lambda_i^\star}$ $\left\{\equiv \frac{\partial}{\partial t}\mathcal{W}(t,\mathbf{x}\|\mathcal{Y}^{k-1}) = \mathcal{L}_{\mathcal{FP}}\mathcal{W}(t,\mathbf{x}\|\mathcal{Y}^{k-1})\right\}$, $\widehat{\mathcal{W}}(\mathbf{x}_k\|\mathcal{Y}^{k-1}) = \sum_i \acute{a}_i^-(t_k)\phi_i^\star(\mathbf{x})$
Gain	$K_k = \mathbf{M}^{-1}\mathbf{Y}_k$ $[\mathbf{M}]_{ij} = \langle\phi_i^\star, \phi_j^\star\rangle_{L_2(\Omega)}$, $[\mathbf{Y}_k]_{ij} = \langle\phi_i^\star, \phi_j^\star\rangle_{L_2(d\mathcal{W}(\mathbf{y}_k\|\mathbf{x}))}$ $\mathcal{W}(\mathbf{y}_k\|\mathbf{x}) \equiv$ Likelihood function (Eq.2.14)
Update	$\acute{\mathbf{a}}^+(t_k) = \frac{1}{\mathbf{v}_k^{\mathrm{T}}\acute{\mathbf{a}}^-(t_k)}K_k\acute{\mathbf{a}}^-(t_k)$, $[\mathbf{v}_k]_i = \int_\Omega \phi_i^\star(\mathbf{x})\mathcal{W}(\mathbf{y}_k\|\mathbf{x})d\mathbf{x}$ $\left\{\equiv \mathcal{W}(\mathbf{x}_k\|\mathcal{Y}^k) = \frac{\mathcal{W}(\mathbf{y}_k\|\mathbf{x}_k)\mathcal{W}(\mathbf{x}_k\|\mathcal{Y}^{k-1})}{\int_\Omega \mathcal{W}(\mathbf{y}_k\|\xi)\mathcal{W}(\xi\|\mathcal{Y}^{k-1})d\xi}\right\}$ $\widehat{\mathcal{W}}(\mathbf{x}_k\|\mathcal{Y}^k) = \sum_i \acute{a}_i^+(t_k)\phi_i^\star(\mathbf{x})$; $\mathcal{Y}^k = \{\mathcal{Y}^{k-1}, \mathbf{y}_k\}$, $k \mapsto k+1$, until $k = m$

2. Filter Propagation

Once the initial parameters, $\acute{a}_i^+(t_0)$ are obtained, it is required to propagate them up to the time instant when the next measurement comes in, i.e., t_1. Since preprocessing steps are already complete, this is extremely easy because it follows directly from Eq.3.36 of chapter III. Written in current notation, we obtain:

$$\acute{a}_i^-(t_k) = \left(\acute{a}_i^+(t_{k-1}) + \frac{\acute{l}_i}{\lambda_i^\star}\right)\exp(\lambda_i^\star t_k) - \frac{\acute{l}_i}{\lambda_i^\star}, \tag{6.3}$$

where, $k = 1$ for the first propagation phase. More notational changes: i^{th} element of the load vector in modal coordinates, written as f_i' in Eq.3.36 appears above changed into $\acute{l}_i$ to prevent confusion between load vector and system dynamics, $\mathbf{f}(t, \mathbf{x})$. Also, $\lambda_i^\star$ are eigenvalues corresponding to admissible eigenfunctions $\phi_i^\star$. The superscript "$-$" in Eq.6.3 signifies that the measurement update at t_k has not yet taken place and the coefficients will thus characterize the prior pdf. Note that the above analytic expression is equivalent to solving FPE for the conditional state density, i.e. $\frac{\partial}{\partial t}\mathcal{W}(t, \mathbf{x}|\mathcal{Y}^{k-1}) = \mathcal{L}_{\mathcal{FP}}\mathcal{W}(t, \mathbf{x}|\mathcal{Y}^{k-1})$ and the resulting approximation of the prior pdf can be written as:

$$\widehat{\mathcal{W}}(\mathbf{x}_k|\mathcal{Y}^{k-1}) = \sum_{i=1}^{\text{card}(\mathcal{A})} \acute{a}_i^-(t_k)\phi_i^\star(\mathbf{x}) \tag{6.4}$$

The final step in state estimation is to obtain the posterior pdf by incorporating new incoming information accumulating in the filtration as $\mathcal{Y}^k = \{\mathcal{Y}^{k-1}, \mathbf{y}_k\}$.

a. Exponentially Decaying Modal Basis Functions

An important benefit of using eigenfunctions of FP operator as shape functions for approximating the state pdf is that their time constants of decay can be exploited to reduce problem size. The figures on page 160 show a typical spectrum associated with the FP operator. Note that the magnitude of eigenvalues (along y-axis) gives

a measure of how fast the corresponding eigenfunctions would decay and stop participating in the approximation. Therefore, the minimum known time gap between measurements ($\inf(t_k - t_{k-1})$) can be used as a guide to retain only those eigenfunctions that will participate in uncertainty propagation, hence removing several relatively fast-decaying modes from analysis. Note however, that such a step may reduce the accuracy of approximation of the posterior pdf (discussed in the next section) because the measurement update step may require additional shape functions in order to implement the weak form of the Bayes rule.

3. Measurement Update

End result of the propagation phase is the state probability density at the current time, conditioned on all previous measurements. The last step in recursive state estimation is to obtain the state probability density conditioned on all available information up to the present time. In the current filter, this is performed via the Bayes rule (Eq.2.13) enforced in weak form, which can be written as: (much like in Beard et al. [141])

$$\sum_{i=1}^{\mathrm{card}(\mathcal{A})} \acute{a}_i^+(t_k) \int_\Omega \phi_i^\star(\mathbf{x}) v d\mathbf{x} = \frac{\sum_{i=1}^{\mathrm{card}(\mathcal{A})} \acute{a}_i^-(t_k) \int_\Omega \phi_i^\star(\mathbf{x}) \mathcal{W}(\mathbf{y}_k|\mathbf{x}) v d\mathbf{x}}{\sum_{i=1}^{\mathrm{card}(\mathcal{A})} \acute{a}_i^-(t_k) \int_\Omega \phi_i^\star(\mathbf{x}) \mathcal{W}(\mathbf{y}_k|\mathbf{x}) d\mathbf{x}} \tag{6.5}$$

Following Galerkin's approach, we have $\mathfrak{V} = \{v_j\} = \{\phi_j^\star\}$. Using Beard's notation, the above equation can be represented in matrix form as:

$$\mathbf{a}^+(t_k) = \frac{\mathbf{M}^{-1}\mathbf{Y}_k\mathbf{a}^-(t_k)}{\mathbf{v}_k^{\mathrm{T}}\mathbf{a}^-(t_k)} \tag{6.6}$$

where,

$$[\mathbf{M}]_{ij} = \int_{\Omega} \phi_i^\star(\mathbf{x})\phi_j^\star(\mathbf{x})d\mathbf{x} = \langle \phi_i^\star, \phi_j^\star \rangle_{L_2(\Omega)} \tag{6.7}$$

$$[\mathbf{Y}_k]_{ij} = \int_{\Omega} \phi_i^\star(\mathbf{x})\phi_j^\star(\mathbf{x})\mathcal{W}(\mathbf{y}_k|\mathbf{x})d\mathbf{x} = \langle \phi_i^\star, \phi_j^\star \rangle_{L_2(d\mathcal{W}(\mathbf{y}_k|\mathbf{x}))} \tag{6.8}$$

$$[\mathbf{v}_k]_i = \int_{\Omega} \phi_i^\star(\mathbf{x})\mathcal{W}(\mathbf{y}_k|\mathbf{x})d\mathbf{x} \tag{6.9}$$

Note that in the above equations, only mass matrix, $\mathbf{M}$ can be pre-computed, while the measurement stiffness matrix, $\mathbf{Y}_k$ and measurement load vector, $\mathbf{v}_k$ need to be computed online as new measurements come in. This challenge makes the current approach not suitable for problems with very fast measurement updates, especially when using modest computing resources. Note however, that there is an alternate way of implementing a fast measurement update based on a function approximation approach. The RHS of Eq.2.13 can be treated as a function to be approximated using the shape functions $\phi_i^\star(\mathbf{x})(\cdot)$ by sampling it on a large number of points over the solution domain. This leads us to the following normal equations:

$$\mathbf{Ma} = \mathbf{b} \tag{6.10}$$

where $\mathbf{M}$ is the mass matrix, which can be pre-computed and $\mathbf{b}$ is a load vector given by:

$$b_j = \langle \phi_j^\star(\mathbf{x}), \mathcal{W}(\mathbf{y}_k|\mathbf{x}_k)\mathcal{W}(\mathbf{x}_k|\mathcal{Y}^{k-1}) \rangle_{\Omega} \tag{6.11}$$

$$\equiv \langle \phi_j^\star, \mathcal{W}(\mathbf{x}_k|\mathcal{Y}^{k-1}) \rangle_{L_2(d\mathcal{W}(\mathbf{y}_k|\mathbf{x}))} \tag{6.12}$$

The above norms are computed in the discrete sense using the sampled points, as in any function approximation problem where the objective function is given in the form of a table. Note that Eq.6.12 represents a "weighted norm" form of Eq.6.11, in the spirit of chapter IV. Indeed, Eq.6.12 provides a guideline for sampling the posterior

pdf because it gives relative weightage to the various regions of the domain Ω based on where the measurements appear. In order to enforce the normality condition, the posterior pdf resulting from the above approximation needs to be re-normalized to unit probability mass. While this process may still not make the measurement update an instantaneous computation, it does provide considerable speed-up over the weak form implementation of Bayes rule.

In either case, we note that the current filter is more suited for problems where time duration between measurement updates is relatively high, for two reasons: (1) the extended Kalman filter and other approximate filters work very well for applications where fast measurement updates are available and the current approach would be somewhat of an overkill, and (2) the current FPE propagator actually works better (is more accurate) when time gap between measurements is longer.

C. Results

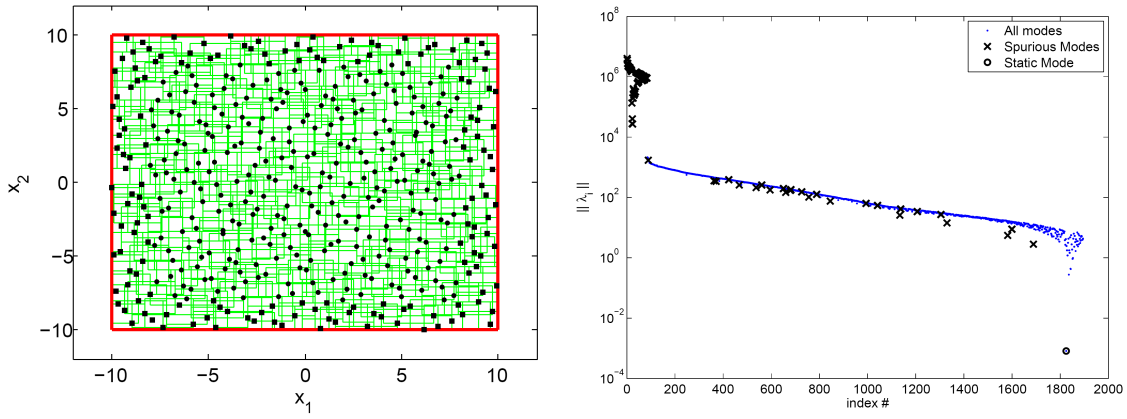
This section presents examples of nonlinear filtering using the above described methodology. Examples in 2, 3 and 4 dimensional space are considered.

1. Filtering in 2D: System 1

Consider the following 2-state nonlinear Duffing oscillator with state-multiplied noise:

$$\ddot{x} + 2\eta\dot{x} - x + \epsilon x^3 = x\mathcal{G}_1(t) + \mathcal{G}_2(t) \tag{6.13}$$

The two independent components of noise, $\mathcal{G}_1$ and $\mathcal{G}_1$ have intensities D_{11} and D_{22}. We consider the case of $D_{11} = 0.0$ for easy comparison with the extended Kalman filter. Note that the current method is fully equipped to deal with state multiplied noise. Values of other parameters used are: $\eta = 0.1$, $\epsilon = 0.5$ and $D_{22} = 0.4$. Figure



(a) Domain discretization with local p-refinement. (b) Spectrum of discretized FP operator.

Fig. 54. Domain discretization and spectral analysis for filtering of system in Eq.6.13.

54(a) shows discretization of the solution domain $\Omega = [-10, 10] \otimes [-10, 10]$ with interior nodes carrying quadratic polynomials and outer nodes carrying constant basis functions (shown with circles and squares respectively). Spectrum of ensuing generalized eigenvalue problem is shown in Fig.54(b) and spurious eigenfunctions are marked out. Two measurement models are considered:

1. Measurement model 1: The system state x is measured: $h(x, \dot{x}) = x$.

2. Measurement model 2: The system "energy" is measured: $h(x, \dot{x}) = x^2 + \eta\dot{x}^2 + \epsilon x^4$

Measurement noise in both cases is assumed to be $R = 2$ and measurements are assumed to arrive every 9 seconds. The initial state distribution is given by the following Gaussian pdf: $N(\{5, 5\}, 0.5\mathbf{I}_{2\times 2})$.

a. Results for Measurement Model 1

Tracking results using the FPE based nonlinear filter are shown in Fig.55. The "stars" appearing on the plot of true state trajectories denote times at which measurements were made. Figs.56(a) and 56(b) illustrate estimation errors and confidence bounds along with comparison to the extended Kalman filter. Since the actual state measurement is available in this case, EKF performs well, as expected. However, the confidence bound in estimates of $\dot{x}$ grow large because measurements include no information of this state. In comparison, the nonlinear filter performs with lower error and tighter confidence (see Fig.56(b)). It is important to note that the nonlinear filter developed in this chapter provides much more information than just mean and covariance estimates. Indeed, the complete state conditional pdf is available and can be used to derive any desired probabilistic information. In comparison, the mean and covariance plots comprise the complete information available from EKF. Fig.57 shows prior and posterior conditional pdfs at four time instants of measurement updates (excluding $t = 0$ where prior and posterior pdfs are the same) . Note that at $t = 9s$, the prior probability density is highly non Gaussian. However, measurement update converts it into a unimodal density function. At $t = 18s$, the prior density again assumes bimodal form at the end of the propagation phase, but the state measurement converts it back into a unimodal function, as expected. After this point, the filter enters a steady state and propagation-update steps repeat (see $t = 54s$).

b. Results for Measurement Model 2

Tracking results are much more interesting in the second case, wherein "energy" measurements are made. Fig.58 shows state estimates while filter-errors are plotted in Figs.59(a) and 59(b). Fig. 60 shows conditional densities obtained from the non-

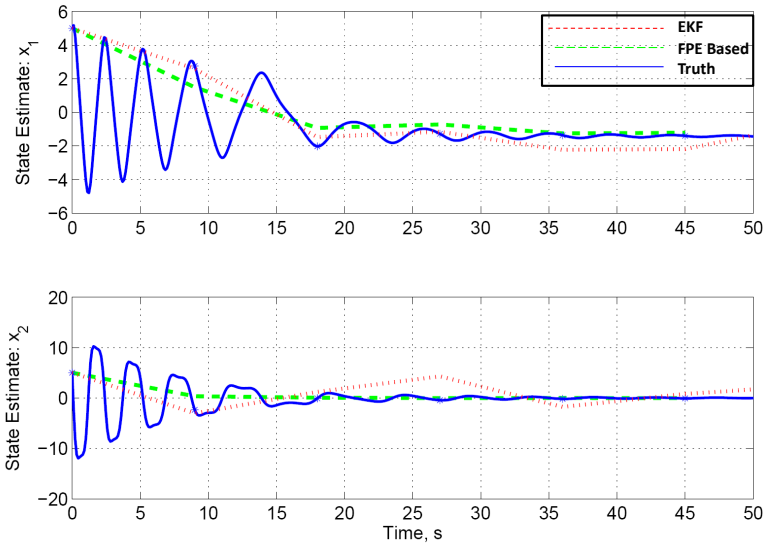
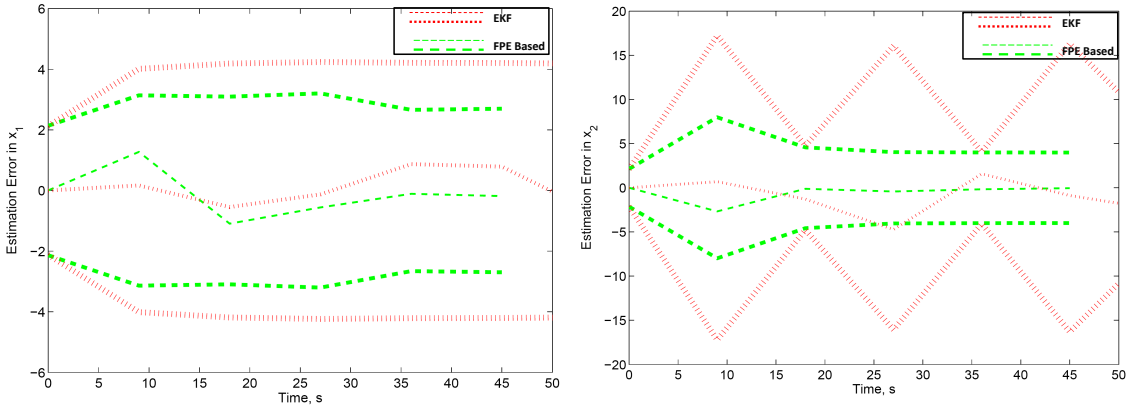


Fig. 55. State estimates for system 1 with measurement model 1.

linear filter at four time instants. It is instructive to consider Fig.60 first: since the measurement model does not provide unique information about the state (model is symmteric about the x and $\dot{x}$ axis), the measurement update is unable to convert the bimodal prior into a unimodal function. Due to the nature of the measurement model, the likelihood function is bimodal, and as a result, the update step only makes the two modes "sharper" (see Fig.60) rather than eliminating one of the modes. The resulting posterior means are therefore zero, as is visible from Figs.59(a) and 59(b). On the other hand, the EKF estimate of state x tends to drift away from the truth, and despite tuning efforts, leads to inconsistent behavior. This is clearly apparent in Fig.59(a), wherein the filter error in x breaks the $3-\sigma$ confidence boundary. This behavior is most likely due to the ambiguity introduced by the measurement model in conjunction with long propagation time.



(a) Error estimates for x (system 1, measurement model 1).

(b) Error estimates for $\dot{x}$ (system 1, measurement model 1.

Fig. 56. Filtering results (FPE based and EKF) for system 1 and measurement model 1.

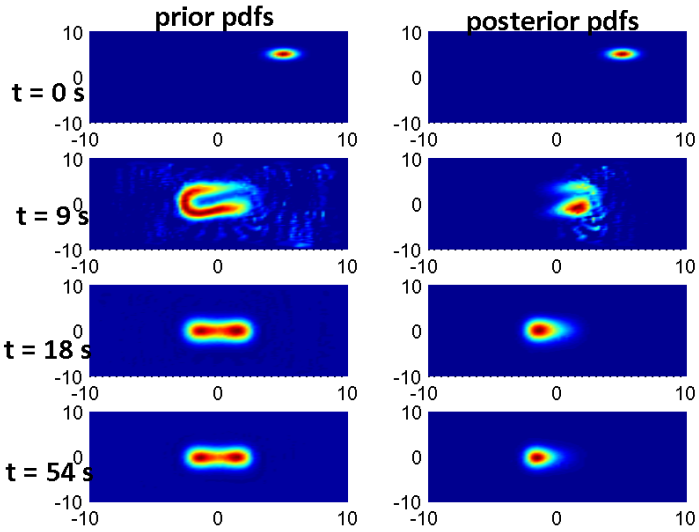


Fig. 57. Conditional pdf estimates with FPE based filter for system 1 and measurement model 1.

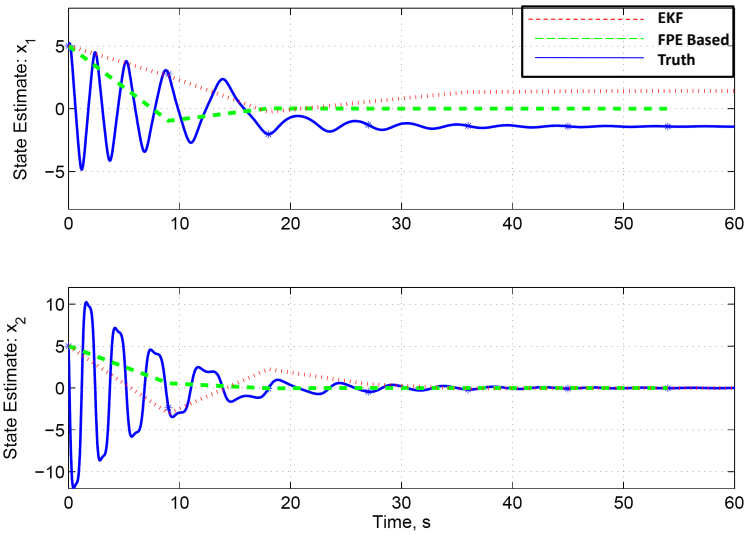
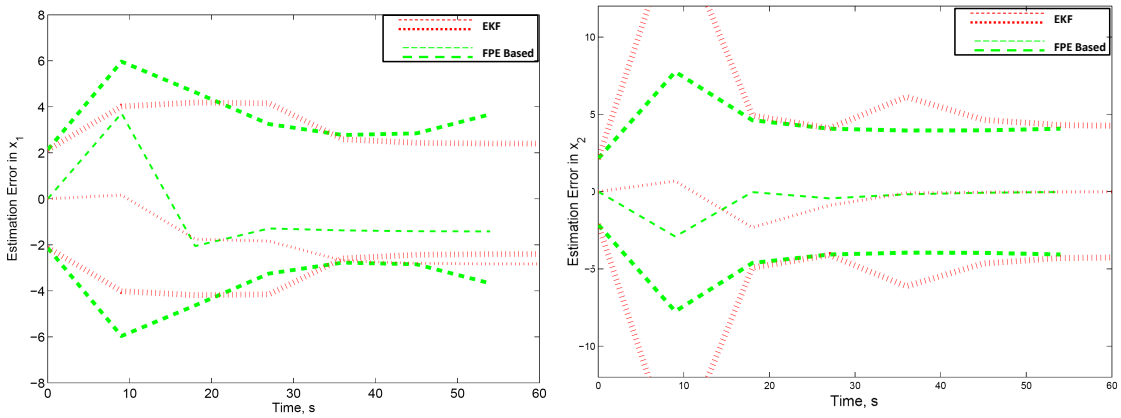


Fig. 58. State estimates for system 1 with measurement model 2.



(a) Error estimates for x (system 1, measurement model 2).

(b) Error estimates for $\dot{x}$ (system 1, measurement model 2).

Fig. 59. Filtering results (FPE based and EKF) for system 1 and measurement model 2.

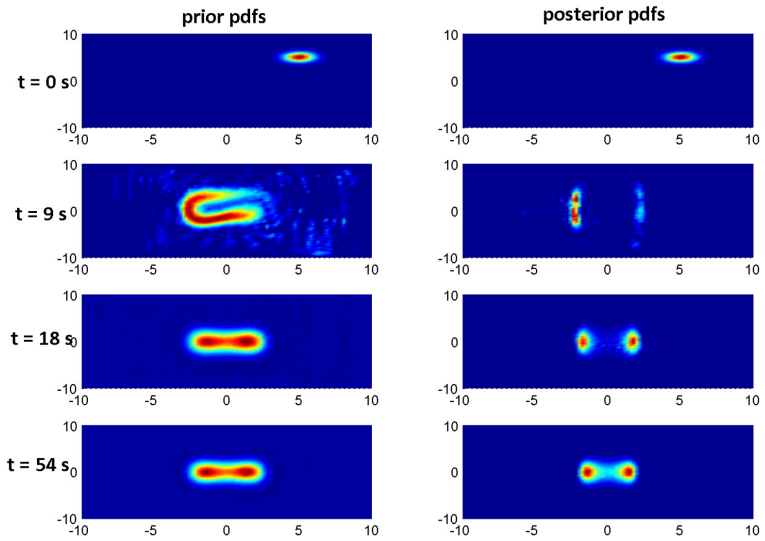


Fig. 60. Conditional pdf estimates with FPE based filter for system 1 and measurement model 2.

2. Filtering in 2D: System 2

We consider next the nonlinear Van der Pol oscillator given by the following equation:

$$\ddot{x} + x + \dot{x}(x^2 - 1) = g\zeta \tag{6.14}$$

Van der Pol equations are widely used in science and engineering for modeling several real life systems, e.g. in studying electrical and opto-electrical circuits, modeling geological faults between tectonic plates, and also in neurobiology for modeling behavior of neurons. The nonlinearity in a Van der Pol oscillator manifests in the form of a limit cycle. In this example, a displacement measurement model is used: $h(x) = x$. Measurements are assumed to arrive every 5 seconds with an error intensity of $R = 1$. The process noise (ζ) is assumed to have a high intensity of $Q = 10$. Resulting state estimates are shown in Fig.61 and error therein appears in Fig.62(a) and 62(b)

along with error estimates of the extended Kalman filter. Note that the FPE filter provides more accurate estimates with tighter confidence bounds, especially in $\dot{x}$. Because of the widely separated bimodal nature of prior conditional pdfs (see Figs.63 and 64), the estimated covariance of EKF is likely to be very high. This is clearly visible in error estimates of state $\dot{x}$ which grow with time, eventually making EKF estimates inconsistent (i.e. error $> 3\sigma$ bound). It is worth noting that no information is available about $\dot{x}$ from the measurement.

On the other hand, because of the measurement model, the likelihood function associated with this system is unimodal (see Eq.2.14). In the FPE based nonlinear filter, this is responsible for conversion of bimodal nature of prior pdfs to unimodal nature of the posteriors. Note that after 10 s, the prior pdfs always degenerate into the stationary solution of FPE, which is essentially a limit cycle with two regions of high probability density. However, the unimodal likelihood function discards one mode (for example, see posterior for $t = 15s$, $20s$, $80s$), or both modes ($t = 25s$) depending on the actual measurements, thus providing unambiguous information about the location of the state. State error estimates are never inconsistent for this filter.

3. System 3: Filtering in 3D (Lorenz Attractor)

We next consider the Lorenz attractor of Eq.3.50 considered in chapter III. The measurement model is considered to be similar to the previous examples, $h(\mathbf{x}) = x^2 + y^2 + z^2$, with $R = 2$. High measurement and process noise are chosen for this problem and results of state estimation are shown in Figs.65 and 66 for x and y. It is visible that the nonlinear filter estimates of the state settle to a steady state value with fixed error while the EKF estimates are uniformly high in error and tend to drift farther away from the truth as time progresses. This is likely due to the cumulative effect of high nonlinearity and high intensity process and measurement noise. Time

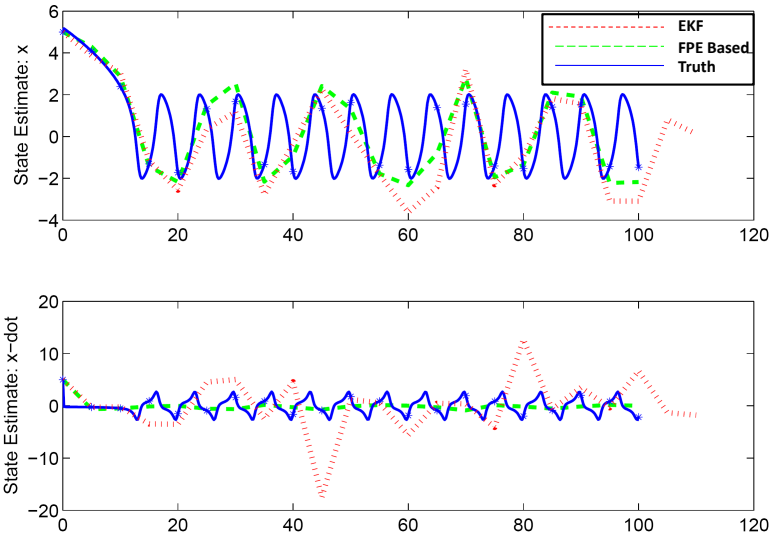
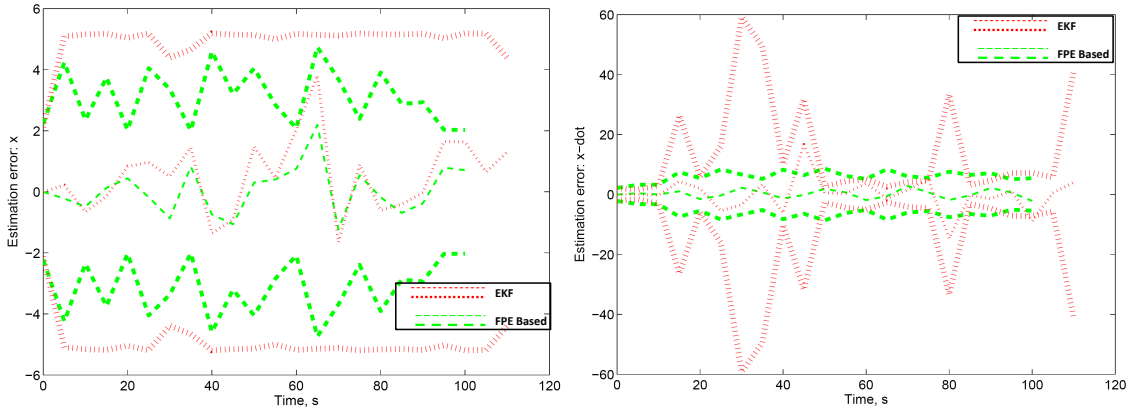


Fig. 61. State estimates for system 2.



(a) Error estimates for x (system 2). (b) Error estimates for $\dot{x}$ (system 2).

Fig. 62. Filtering results (FPE based and EKF) for system 2.

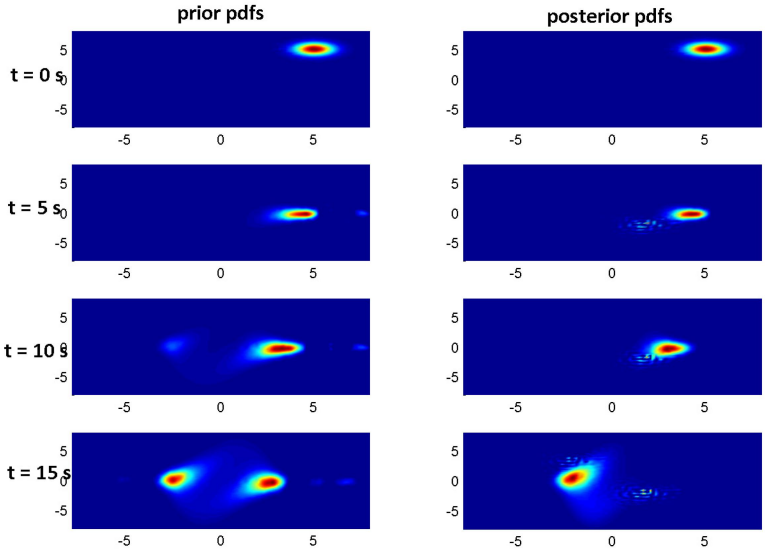


Fig. 63. Full state conditional pdf estimates with FPE based filter for system 2: $t = 0s$ to $t = 15s$.

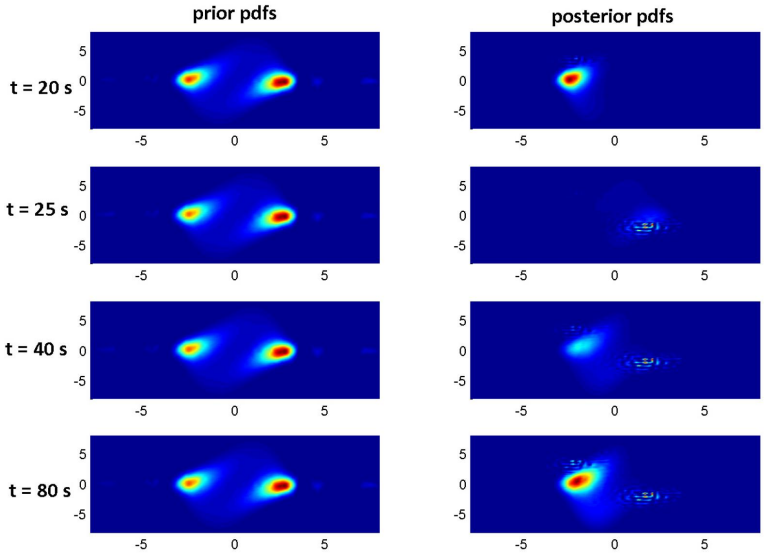


Fig. 64. Full state conditional pdf estimates with FPE based filter for system 2: $t = 20s$ to $t = 80s$.

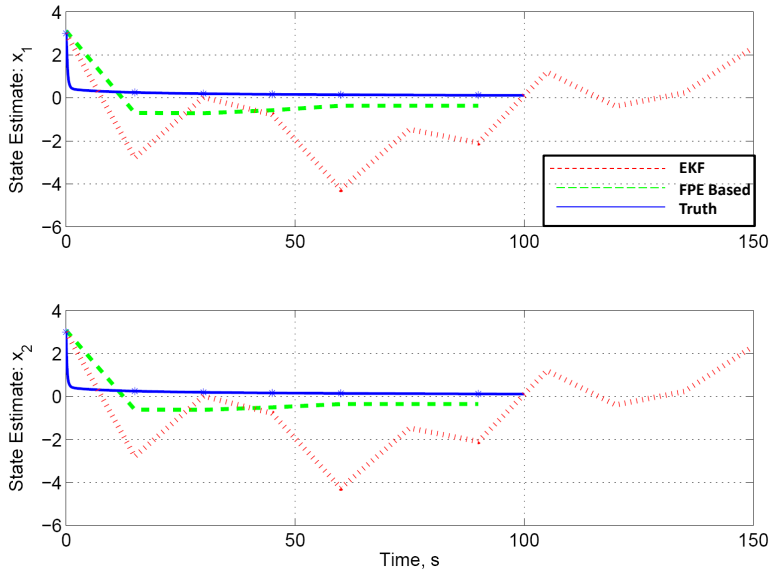
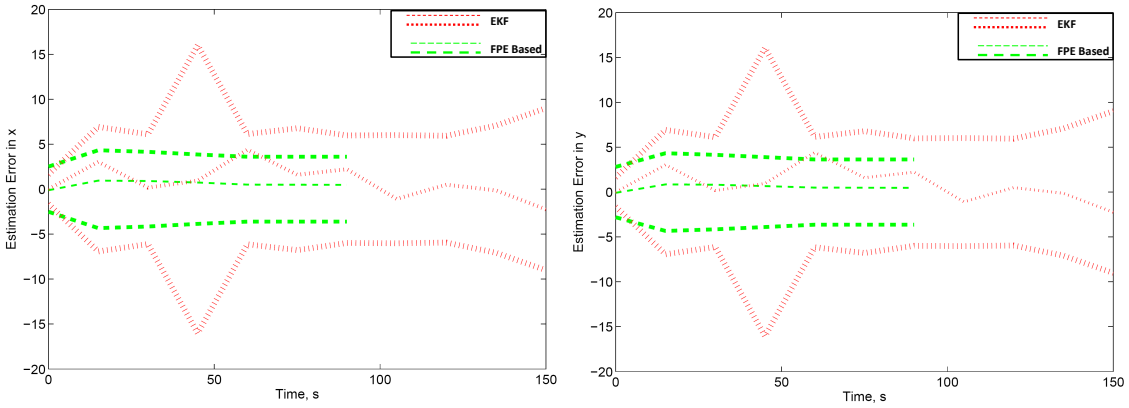


Fig. 65. State estimates for system 3 with "energy like" measurement model.

snapshots of the state conditional-pdf ($x - y$ marginals) have been shown in Fig.67, depicting unimodal behavior for this system. The steady state behavior of the system is also captured in these plots.

4. Filtering in 4D: Coupled Vibration Isolation Suspension

We finally consider nonlinear estimation for the two degree-of-freedom nonlinear vibration model of Eq.3.31 studied in chapter III. In this example, we consider a 2 dimensional measurement model in which the state-rates, $\dot{x_1}$ and $\dot{x_2}$, i.e. x_3 and x_4 are measured. The measurements are assumed to arrive every 4 seconds with an error covariance matrix of $5\mathbf{I}_{2\times 2}$. State estimates are shown in Fig.68 and error estimates in Figs.69. Note that the EKF estimate of state x_2 is erroneous because it settles over the incorrect mode. In fact, its behavior was observed to be unpredictable because different noise samples led to different steady state behaviors of x_2. In other



(a) Comparative error estimates for x, system 3.

(b) Comparative error estimates for y, system 3.

Fig. 66. Error estimates for system 3.

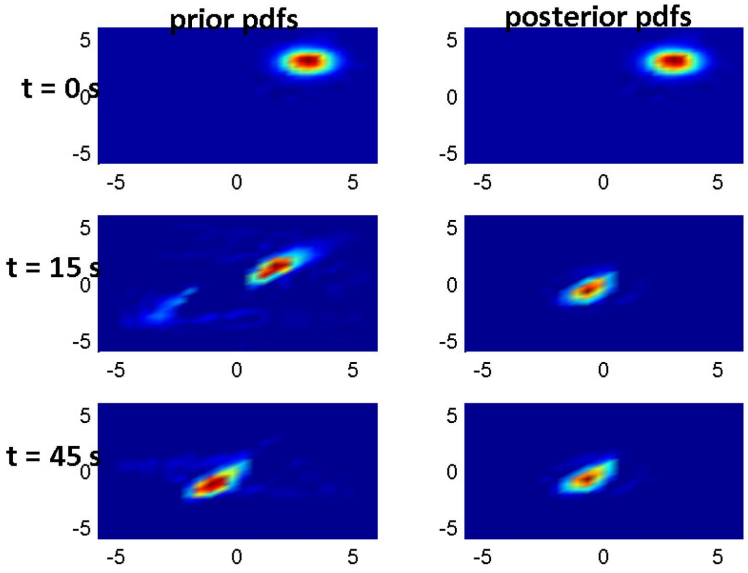
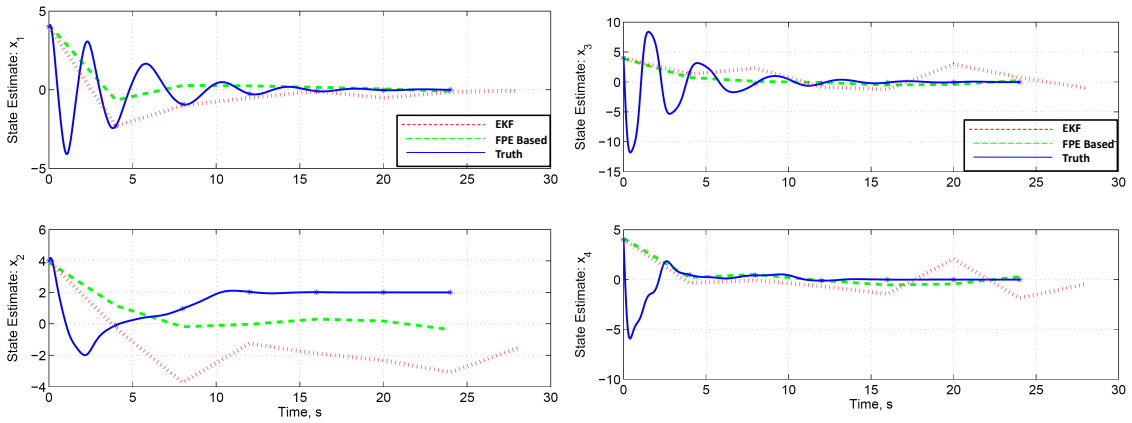


Fig. 67. Full state pdf tracking with FPE based filter for the Lorenz attractor.



(a) Comparative state estimates for x_1, x_2, system 3.

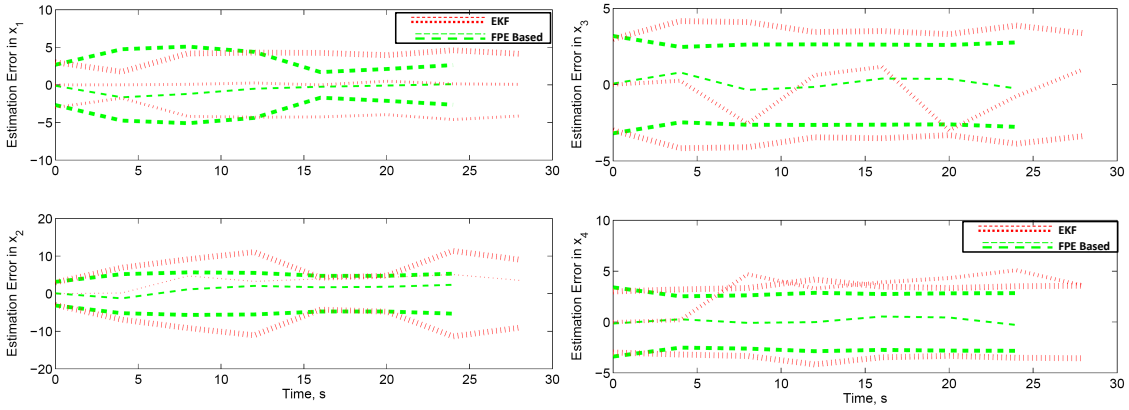
(b) Comparative state estimates for x_3, x_4, system 3.

Fig. 68. State estimation using FPE based nonlinear filter for system 3.

words, due to relative closeness of the two modes, the EKF chose one mode over the other depending on the particular noise sample used in the simulation. This is a result of high process noise coupled with relatively long propagation times, causing the EKF errors to border on inconsistency (see error estimates of states x_2 and x_4 in Figs.69(a),69(b)). On the other hand, with only the information available about the states x_3 and x_4, the FPE based filter is unable to decide between the two modes of the system. The x_1-x_2 marginal conditional densities illustrated in Fig.70 show that prior pdfs assume bimodal shape (as expected, from chapter III, section E) but the posterior conditional pdfs are unimodal centered at the origin as the rates x_3 and x_4 settle to zero.

D. Summary

In this chapter, a nonlinear filter based on FPE was developed and proposed to be suitable for problems involving long durations of propagation phase and/or



(a) Comparative error estimates for x_1 and x_2, system 3.

(b) Comparative error estimates for x_1 and x_2, system 3.

Fig. 69. Error estimates for system 3.

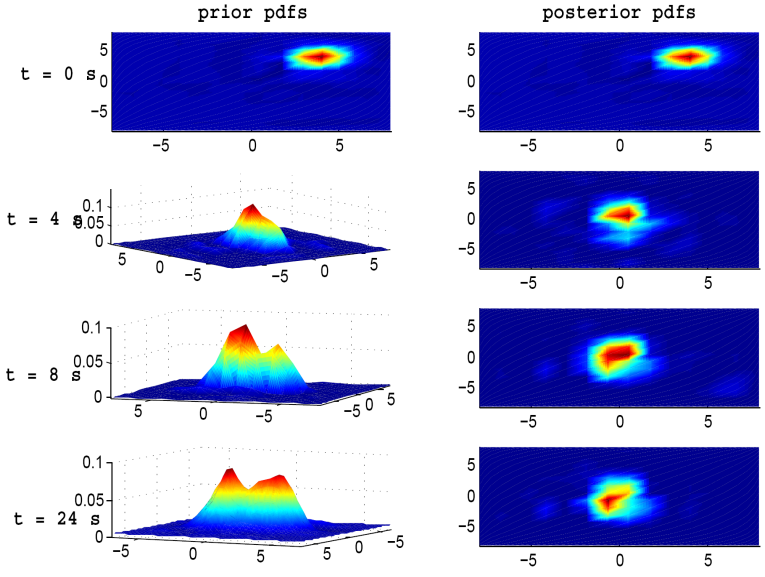


Fig. 70. Full state pdf tracking with FPE based filter for the four-state oscillator of Eq.3.31.

high process/measurement noise scenarios. The semianalytical algorithm developed in chapter III is used to trivially solve the propagation phase of the nonlinear filter. In fact, by virtue of modal basis functions used to characterize the conditional state density, equation error reduces with time (Theorem III.1 and Corollary III.1) implying that the current nonlinear filter works better when time duration of propagation is longer. Measurements are incorporated via a weak form of the Bayes update rule, which is a computationally heavy step and renders the current filter ineffective for applications involving high frequency measurement updates. The benefits of having the full-state conditional pdf at disposal for estimation are illustrated through several examples. It is observed that the FPE based nonlinear filter typically provides tighter confidence bounds than EKF most cases. It is never observed to be inconsistent in its predictions, primarily because it considers the actual conditional pdfs (as opposed to a few moments) to generate estimates. Because this information is not available to the EKF, it is often seen to be inconsistent or overly conservative, especially when propagation times are long. It is however noted that the discussed filter is still largely in its developmental phase and there remains tremendous scope for improvement. For example, the powerful technique of proper orthogonal decomposition (POD) can be utilized to significantly reduce the order of the filter by constructing highly accurate modal basis functions for approximation of the conditional pdf. The results presented in this chapter amply illustrate the power of the proposed FPE based filter under difficult state-estimation conditions and provide significant basis for optimism.

CHAPTER VII

CONCLUSIONS

This dissertation has presented a meshless numerical methodology for solving FPE and coupled it with spectral analysis to provide near real-time transient FPE response. A recursive norm-modification technique has been developed to improve approximation accuracy without changing problem size and track optimal solution domain for nonlinear systems. Applications of the developed algorithms has been discussed in the fields of computational stochastic optimal control and nonlinear filtering. This chapter reviews contributions made by this work and viable paths for extending this research to a wider span of applications.

A. Contributions of Research

Based on presented solution methodologies and results, a claim to the following contributions to the field of stochastic systems analysis and design can be made:

1. A robust algorithm for solving Fokker-Planck equation has been developed for nonlinear stochastic systems. The meshless approach has been introduced for the first time in FPE literature and shown to tackle the curse of dimensionality successfully, for which numerical evidence has been provided. The particle version of PUFEM is used to obtain working approximations with extremely small problem size. Significant improvement has been achieved over state-of-the-art techniques for solving FPE.

2. A semianalytic method has been developed by coupling meshless discretization with modal analysis and spurious mode rejection to solve the transient FPE in near-real time, independent of initial probability distribution. It is shown

that equation error in FPE is bounded by an exponentially decaying envelope with greatest width at the initial time, implying that approximation accuracy *improves over time* by use of modal basis functions. This approach will open new avenues in the area of nonlinear filtering.

3. A recursive norm-modification approach has been developed for solution refinement while maintaining small problem size. A homotopic approach has been developed to track solution domain for nonlinear dynamics.

4. An FPE perspective has been developed for stochastic optimal control problems and solution of the stationary nonlinear HJB equation has been obtained for several systems through a policy iteration recursion over the backward Kolmogorov equation.

5. The semianalytic algorithm has been used to solve the nonlinear filtering problem by propagating state-pdf in between measurement updates. The technique is especially suited for application with sparse measurements. The measurement update is performed by implementation of the Bayes rule in variational form.

B. Future Extensions of Conducted Research

Because of the core value of FPE in the field of randomly excited systems, there remains significant room for further progress in each of the problems considered in this dissertation. This section presents a few avenues for direct extensions in numerical and theoretical aspects of current research:

1. Extensions in Numerical Research

- *Large scale and parallel computing:* The true potential of meshless techniques can only be fulfilled by its implementation on large scale and parallel computing platforms. This is especially true for PU based methods because they are very highly amenable to parallelization of computation. The research presented in this dissertation has shown that problems previously attempted only on super-computers (e.g. FPE for four-state nonlinear systems) can be solved on a small workstation using PU based meshless techniques. With the emerging field of graphic processing unit (GPU) based computation, the PU based algorithm can be transformed into a truly potent tool for solving extremely high dimensional FPE, and an industrial standard PDE solver in general.

- *Randomization techniques and curse of dimensionality:* The primary reason for turning to meshless methods is to deal with problems in high dimensions and handle the curse of dimensionality associated with them. Randomization is a process in which a problem suffering from this curse is solved effectively by the use of a node-based approximation, typically drawn from a uniform random distribution. Popular examples include Monte-Carlo numerical integration in $\Re^N$ using uniform distributions and Discrete Decision Processes (a subclass of Markov Decision Processes). Randomly discretized solution domains used in this dissertation have shown (numerically) the weakening of the curse of dimensionality in solving FPE. The bigger question remains: Can randomly discretized domains, under the partition of unity paradigm of meshless methods, break the curse of dimensionality associated with solving PDEs? This will require extensive numerical and theoretical investigation.

- *Further development of approximation space refinement:* There is tremendous

scope for research in automatic generation of non-polynomial/special shape functions suited to a particular PDE. Also open is the problem of determining optimal placement of nodes for highest approximation accuracy.

2. Extensions in Theoretical Research

- *Convergence studies:* There still exist theoretical holes in determining convergence properties of the discussed meshless methods. This is currently the case with almost all existing meshless methods.

- *Rigorous proof for breaking the curse of dimensionality:* A rigorous proof for breaking the curse of dimensionality with optimal node placement and use of special functions is due. When developed, such a proof would be a ground breaking result in this field.

REFERENCES

[1] A. T. Fuller, "Analysis of nonlinear stochastic systems by means of the Fokker-Planck equation," *International Journal of Control*, vol. 9, no. 6, pp. 603–655, 1969.

[2] J. C. Maxwell, "Illustrations of the dynamical theory of gases," *Philosophical Magazine*, vol. 19, pp. 19–32, 1860.

[3] J. C. Maxwell, "On the dynamical theory of gases," *Philosophical Transactions of the Royal Society of London*, vol. 157, pp. 49–88, 1867.

[4] L. Boltzmann, "[german] studien Über das gleichgewicht der lebendigen kraft zwischen bewegten materiellen punkten," *Sitz. d.k. Akad. Wiss. Wien.*, vol. 58, pp. 517–560, 1868.

[5] L. Rayleigh, "On the resultant of a large number of vibrations of the same pitch and arbitrary phase," *Philosophical Magazine*, vol. 10, pp. 73–78, 1880.

[6] L. Rayleigh, "On James Bernoulli's theorem in probabilities," *Philosophical Magazine*, vol. 10, pp. 73–78, 1880.

[7] L. Rayleigh, *Theory of Sound, 2^{nd} edition*, Macmillan, London.

[8] L. Rayleigh, "Dynamical problems in the illustration of the theory of gases," *Philosophical Magazine*, vol. 32, pp. 424–445, 1891.

[9] L. Bachelier, "Théorie de la spéculation," *Annales Scientifiques de l'École Normale Supérieure*, vol. 3, pp. 21–86, 1900. English translation: P. H. Cootner, *The Random Character of Stock Market Prices*, MIT Press, Cambridge, pp. 17-78, 1964.

[10] A. Einstein, "[german] Über die von der molekularkinetischen theorie der wärme geforderte bewegung von in ruhenden flssigkeiten suspendierten teilchen, english title: On the movement of small particles suspended in stationary liquids required by the molecular-kinetic theory of heat," *Annalen der Physik*, vol. 17, pp. 549–560, 1905.

[11] P. Langevin, "[french] sur la théorie lu mouvement Brownien, english title: On the theory of Brownian motion," *C. R. Acad. Sci., Paris*, vol. 146, pp. 530–533, 1908.

[12] A. D. Fokker, *[German] Over Brownsche Bewegingen it Het Stralingsveld*, Ph.D. Dissertation, Leiden University, 1913.

[13] A. D. Fokker, "[german] die mittlere energie rotierender elektrischer dipole im strahlungsfeld," *Annalen der Physik*, vol. 348, pp. 810–820, 1913.

[14] M. Planck, "[german] zur theorie des rotationsspektrums," *Annalen der Physik*, vol. 357, pp. 491–505, 1917.

[15] A. N. Kolmogoroff, "[german] Über die analytischen methoden in der wahrscheinlichkeitsrechnung, english title: On analytical methods in probability theory)," *Mathematische Annalen*, vol. 104, pp. 415–458, 1931.

[16] A. N. Kolmogoroff, *[German] G̈rundbegriffe der Wahrscheinlichkeitsrechnung, (English title: Foundations of the theory of probability)*, Springer Berlin, 1933.

[17] G. E. Uhlenbeck and L. S. Ornstein, "On the theory of Brownian motion," *Physical Review*, vol. 36, pp. 823–841, 1930.

[18] S. Chandrashekhar, "Stochastic problems in physics and astronomy," *Reviews of Modern Physics*, vol. 15, no. 1, pp. 1–89, 1943.

[19] J. F. Barret, "Application of Kolmogorov's equations to randomly disturbed automatic control systems," in *1st International Congress of the IFAC on Automatic Control*, Moscow, USSR, 1960, Automatic and Remote Control, vol. 2, pp. 724–733.

[20] R. L. Stratonovich, *Topics in the Theory of Random Noise*, Gordon and Breach, New York, 1963.

[21] K. Chuang and L. F. Kazda, "A study of nonlinear systems with random inputs," *Transactions of the AIEE(Applications and Industry)*, vol. 78, Part II, pp. 100–105, 1959.

[22] S. T. Ariaratnam, "Random vibration of nonlinear suspensions," *Journal of Mechanical Engineering Science*, vol. 2, no. 3, pp. 195–201, 1960.

[23] R. H. Lyon, "On the vibration statistics of a randomly excited hard-spring oscillator," *Journal of the Acoustical Society of America*, vol. 32, no. 6, pp. 716–719, 1960.

[24] R. H. Lyon, "On the vibration statistics of a randomly excited hard-spring oscillator ii," *Journal of the Acoustical Society of America*, vol. 33, no. 10, pp. 1395–1403, 1961.

[25] T. K. Caughey and J. K. Dienes, "The behavior of linear systems with white noise input," *Journal of Mathematical Physics*, vol. 32, pp. 2476–2479, 1962.

[26] T. K. Caughey, "Derivation and application of the Fokker-Planck equation to discrete nonlinear dynamical systems subjected to white noise excitation," *Journal of the Acoustical Society of America*, vol. 35, no. 11, pp. 1683–1692, 1963.

[27] S. H. Crandall, *Random Vibration in Mechanical Systems*, Academic Press Inc., New York,, 1963.

[28] S. H. Crandall, "Perturbation techniques for random vibration of nonlinear systems," *Journal of the Acoustical Society of America*, vol. 35, no. 11, pp. 1700–1705, 1963.

[29] S. H. Crandall, "Non-Gaussian closure techniques for stationary random vibration," *International Journal of Nonlinear Mechanics*, vol. 20, no. 1, pp. 1–8, 1985.

[30] S. H. Crandall, "Non-Gaussian closure for random vibration of nonlinear oscillators," *International Journal of Nonlinear Mechanics*, vol. 15, pp. 303–313, 1980.

[31] W. F. Wu and Y. K. Lin, "Cumulant-neglect closure for nonlinear oscillators under parametric and external excitation," *International Journal of Nonlinear Mechanics*, vol. 19, no. 4, pp. 349–362, 1984.

[32] G.-K. Er, "Multi-Gaussian closure method for randomly excited nonlinear systems," *International Journal of Nonlinear Mechanics*, vol. 33, pp. 201–214, 1998.

[33] R. S. Park and D. Scheeres, "Nonlinear mapping of Gaussian state uncertainties: Theory and applications to spacecraft trajectory design," *Journal of Guidance Control and Dynamics*, vol. 29, no. 6, pp. 1367–1375, 2006.

[34] T. K. Caughey, "Equivalent linearization techniques," *Journal of the Acoustic Society of America*, vol. 35, pp. 1706–1711, 1963.

[35] D. C. Polidori and J. L. Beck, "Approximate solutions for nonlinear random vibration problems," *Probabilistic Engineering Mechanics*, vol. 11, pp. 179–185, 1996.

[36] H. J. Pradlwarter, "Nonlinear stochastic response distributions by local statistical linearization," *International Journal of Nonlinear Mechanics*, vol. 36, no. 7, pp. 1135–1151, 2001.

[37] J. B. Roberts and P. D. Spanos, *Random Vibration and Statistical Linearization*, New York, Dover Publication, 2003.

[38] H. J. Pradlwarter C. Proppe and G. I. Shuell, "Equivalent linearization and Monte Carlo simulation in stochastic dynamics," *Probabilistic Engineering Mechanics*, vol. 18, pp. 1–15, 2003.

[39] M. Shinozuka and Wen Y.-K., "Monte Carlo solution of nonlinear vibrations," *AIAA Journal*, vol. 10, no. 1, pp. 37–40, 1972.

[40] H. J. Pradlwarter and G. I. Shuell, "On advanced Monte Carlo simulation procedures in stochastic structural dynamics," *International Journal of Nonlinear Mechanics*, vol. 32, no. 2, pp. 735–744, 1997.

[41] L. A. Bergman E. A. Johnson, S. F. Wojtkiewicz and B. F. Spencer, "Observations with regard to massively parallel computation for Monte Carlo simulation of stochastic dynamical systems," *International Journal of Nonlinear Mechanics*, vol. 32, no. 4, pp. 721–734, 1997.

[42] W. M. MacDonald M. N. Rosenbluth and D.L. Judd, "Fokker-planck equation for an inverse square force," *Physical Review*, vol. 107, no. 1, pp. 350–355, 1957.

[43] R. G. Bhandari and R. E. Sherrer, "Random vibrations in discrete nonlinear dynamic systems," *Journal of Mechanical Engineering Science*, vol. 10, pp. 168–174, 1968.

[44] W. W. Mayfield, "A sequence solution to the Fokker-Planck equation," *IEEE Transactions on Information Theory*, vol. IT-19, no. 2, pp. 165–176, 1973.

[45] J. Reif and R. Barakatd, "Numerical solution of the Fokker-Planck equation via Chebyshev polynomial approximations with reference to the first passage time probability density functions," *Journal of Computational Physics*, vol. 23, pp. 425–445, 1977.

[46] J. D. Atkinson, "Eigenfunction expansions for randomly excited nonlinear systems," *Journal of Sound and Vibration*, vol. 30, no. 2, pp. 153–172, 1973.

[47] J. P. Johnson and R. A. Scott, "Extension ofeigenfunction-expansion solution of a Fokker-Planck equation-I. first order system," *International Journal of Nonlinear Mechanics*, vol. 14, pp. 315–324, 1979.

[48] J. P. Johnson and R. A. Scott, "Extension of eigenfunction-enpansion solution of a Fokker-Planck equation-II. second order system," *International Journal of Nonlinear Mechanics*, vol. 15, pp. 41–56, 1980.

[49] H. D. Vollmer H. Risken and H. Denk, "Calculation of eigenvalues for the kramers equation," *Physics Letters*, vol. 76A, no. 1, pp. 22–24, 1980.

[50] J. Killeen and A. H. Futch, "Numerical solution of the Fokker-Planck equations for a hydrogen plasma formed by a neutral injection," *IJournal of Computational Physics*, vol. 2, pp. 236–254, 1968.

[51] J. C. Whitney, "Finite difference methods for the Fokker-Planck equation," *Journal of Computational Physics*, vol. 6, pp. 486–509, 1970.

[52] R. S. Langley, "A finite element method for the statistics of random nonlinear vibration," *Journal of Sound and Vibration*, vol. 101, no. 1, pp. 41–54, 1985.

[53] R. S. Langley, "A variational formulation of the FPK equations with application to the first passage problem in random vibration," *Journal of Sound and Vibration*, vol. 123, no. 2, pp. 213–227, 1988.

[54] G. Marcozzi V. Palleschi, F. Sarri and M.R. Torquati, "Numerical solution of the Fokker-Planck equation: A fast and accurate algorithm," *Physics Letters A*, vol. 146, no. 7,8, pp. 378–386, 1990.

[55] V. Vanaja, "Numerical solution of a simple Fokker-Planck equation," *Applied Numerical Mathematics*, vol. 9, pp. 533–540, 1992.

[56] E. M. Epperlein, "Implicit and conservative difference scheme for the Fokker-Planck equation," *Journal of Computational Physics*, vol. 112, pp. 291–297, 1994.

[57] S Günel and F. A. Savacı, "Approximate stationary density of the nonlinear dynamical systems excited with white noise," May 23-26, 2005, vol. 232, pp. 4899–4902.

[58] V. N. Volosov and M. S. Pekker, "Numerical methods of solving the two dimensional problem for the Fokker-Planck equation," *U.S.S.R Computational Mathematics and Mathematical Physics*, vol. 20, no. 5, pp. 251–257, 1980.

[59] M. S. Pekker and V. N. Khudik, "Conservative finite difference scheme for the

Fokker-Planck equation," *U.S.S.R Computational Mathematics and Mathematical Physics*, vol. 24, no. 3, pp. 206–210, 1984.

[60] W. V. Wedig, "Generalized Hermite analysis of nonlinear stochastic systems," *Structural Safety*, vol. 6, no. 2-4, pp. 153–160, 1989.

[61] A. A. Mirin, "Massively parallel Fokker-Planck calculations," Apr 8-12,1990, pp. 426–432.

[62] H. P. Langanten, "A general numerical solution method for Fokker-Planck equations with applications to structural reliability," *Probabilistic Engineering Mechanica*, vol. 6, no. 1, pp. 33–48, 1991.

[63] V. Palleschi and M. de Rosa, "Numerical solution of the Fokker-Planck equation II. multidimensional case," *Physics Letters A*, vol. 163, pp. 381–391, 1992.

[64] Jr. B. F. Spencer and L. A. Bergman, "On the numerical solution of the Fokker-Planck equation for nonlinear stochastic systems," *Nonlinear Dynamics*, vol. 4, pp. 357–372, 1993.

[65] L.-C. Shiau and T.-Y. Wu, "A finite element method for analysis of a nonlinear system under stochastic,parametric and external excitation," *International Journal of Nonlinear Mechanics*, vol. 32, no. 2, pp. 193–201, 1996.

[66] G. Ricciardi G. Muscolino and M. Vasta, "Stationary and non-stationary probability density function for nonlinear oscillators," *International Journal of Nonlinear Mechanics*, vol. 32, no. 6, pp. 1051–1064, 1997.

[67] E. A. Johnson, S. F. Wojtkiewicz, L. A. Bergman, and B. F. SpencerJr., "Finite element and finite difference solutions to the transient Fokker-Planck equation,"

in *Proc. of a Workshop: Nonlinear and Stochastic Beam Dynamics in Accelerators - A Challenge to Theoretical and Numerical Physics*, A. Bazzani, J. Ellison, H. Mais, and G. Turchetti, Eds., Lüneburg, Germany, 1997, pp. 290–306.

[68] H. J. Pradlwarter and M. Vasta, "Numerical solution of the Fokker-Planck equation via Gaussian superposition representation," in *Structural Safety and Reliability*, N. Shairaishi, M. Shinozuka, and Y. K. Wen, Eds., Kyoto, 1997, ICOSSAR'97 - 7th International Conference on Structural Safety and Reliability, vol. 2, pp. 917–923.

[69] D. J. Kouri D. S. Zhang, G. W. Wei and D. K.Hoffman, "Distributed approximating functional approach to the Fokker-Planck equation: Eigenfunction expansion," *Journal of Chemical Physics*, vol. 106, no. 12, pp. 5216–5224, 1997.

[70] H. Mais M. P. Zorzano and L. Vazquez, "Numerical solution of two dimensional Fokker-Planck equations," *Applied Mathematics and Computation*, vol. 98, pp. 109–117, 1999.

[71] D. J. Kanppett S. McWilliam and C. H. J.Fox, "Numerical solution of the stationary FPK equation using Shannon wavelets," *Journal of Sound and Vibration*, vol. 232, no. 2, pp. 405–430, 2000.

[72] G. W. Wei, "A unified approach for the solution of the Fokker-Planck equation," *Journal of Physics A: Mathematical and General*, vol. 33, pp. 4935–4953, 2000.

[73] B. Guo J. C. M. Fok and T. Tang, "Combined Hermite spectral-finite difference method for the Fokker-Planck equation," *Mathematics of Computation*, vol. 71, no. 240, pp. 1497–1528, 2001.

[74] M. Di. Paola and A. Sofi, "Approximate solution of the Fokker-Planck-Kolmogorov equation," *Probabilistic Engineering Mechanics*, vol. 17, pp. 369–384, 2002.

[75] A. Masud and L. A. Bergman, "Application of multi-scale finite element methods to the solution of the Fokker-Planck equation," *Computational Methods in Applied Mechanics and Engineering*, vol. 194, pp. 1513–1526, 2005.

[76] M. Kumar, P. Singla, J.L. Junkins, and S. Chakravorty, "A multi-resolution meshless approach to steady state uncertainty determination in nonlinear dynamical systems," in 38^{th} *IEEE Southeastern Symposium on Systems Theory*, Cookeville, TN, Mar 5-7, 2006.

[77] F. E. Daum H. C. Lambert and J. L. Weatherwax, "A split-step solution of the Fokker-Planck equation for the conditional density," 2006, vol. 68, pp. 2014–2018.

[78] M. Ujevic and P. S. Letelier, "Solving procedure for a 25-diagonal coefficient matrix: Direct numerical simulations for the three dimensional linear Fokker-Planck equation," *Journal of Computational Physics*, vol. 215, pp. 485–505, 2006.

[79] P. J. Attar and P. Vedula, "Direct quadrature method for moments solution of the Fokker-Planck equation," *Journal of Sound and Vibration*, vol. 317, pp. 265–272, 2008.

[80] U. V. Wagner and W. V. Wedig, "On the calculation of stationary solutions of multi-dimensional Fokker-Planck equations with orthogonal functions," *Nonlinear Dynamics*, vol. 21, pp. 289–306, 2000.

[81] C. Soize, "Steady-state solution of the Fokker-Planck equation in higher dimension," *Probabilistic Engineering Mechanics*, vol. 3, no. 4, pp. 196–206, 1988.

[82] L. A. Bergman and B. F. Spencer Jr., "Robust numerical solution of the transient Fokker Planck equation for nonlinear dynamical systems," in *Proc. IUTAM Symposium on Nonlinear Stochastic Mechanics*, N. Bellomom and F. Casciati, Eds., Turin, Italy, 1992, pp. 49–60, Springer-Verlag.

[83] L. A. Bergman E. A. Johnson, S. F. Wojtkiewicz and B. F. Spencer, "Finite element and finite difference solutions to the transient Fokker-Planck equation," in *Nonlinear and Stochastic Beam Dynamics in Accelerators - A Challenge to Theoretical and Computational Physics*, H. Mais A. Bazzani, J. Ellison and G. Turchetti, Eds., 1997, pp. 290–306.

[84] L. A. Bergman S. F. Wojtkiewicz and B. F. Spencer Jr., "Computational issues arising in the numerical solution of the Fokker-Planck equation in higher dimensions: Use of iterative solution methods," Nov 24-28, 1997, pp. 851–858.

[85] S. F. Wojtkiewicz and L. A. Bergman, "Numerical solution of high dimensional Fokker-Planck equations," in *Proc. 8^{th} ASCE Specialty Conference on Probablistic Mechanics and Structural Reliability*, Notre Dame, IN, USA, 2000, pp. 167–172.

[86] S. Chakravorty M. Kumar, P. Singla and J. L. Junkins, "The partition of unity finite element approach to the stationary Fokker-Planck equation," in *AIAA/AAS Astrodynamics Specialist Conference*, Keystone, CO, USA, Aug. 21-24, 2006, pp. AIAA Paper Number 2006–6285.

[87] T. Belytschko, Y. Y. Lu, and J. Gu, "Element free Galerkin method," *International Journal of Numerical Methods in Engineering*, vol. 37, pp. 229–256, 1994.

[88] J. J. Monaghan, "An introduction to SPH," *Communications in Computational Physics*, vol. 48, pp. 89–96, 1988.

[89] J. J. Monaghan, "Smooth particle hydrodynamics," *Annual Review of Astronomy and Astrophysics*, vol. 30, pp. 543–574, 1992.

[90] A. Duarte and J. T. Oden, "Hp clouds - an h-p meshless method," *Numerical Methods for Partial Differential Equations*, vol. 12, pp. 673–705, 1996.

[91] S. Jun W. Liu and Y. Zhang, "Reproducing kernel particle methods," *International Journal for Numerical Methods in Fluids*, vol. 20, no. 8-9, pp. 1081–1106, 1995.

[92] S. N. Atluri and T. Zhua, "A new meshless local petrov-galerkin (MLPG) approach in computational mechanics," *Computational Mechanics*, vol. 22, pp. 117–127, 1998.

[93] K. Copps T. Strouboulis and I. Babuska, "The generalized finite element method," *Computational Methods in Applied and Mechanical Engineering*, vol. 190, pp. 4081–4193, 2001.

[94] I. Babuška and J. Melenk, "The partition of unity finite element method," *International Journal of Numerical Methods*, vol. 40, pp. 727–758, 1997.

[95] M. Griebel and M. A. Schweitzer, "A particle-partition of unity method for the solution of elliptic, parabolic and hyperbolic PDEs," *SIAM Journal on Scientific Computing*, vol. 22, pp. 853–890, 2000.

[96] M. Griebel and M. A. Schweitzer, "A particle-partition of unity method part ii: Efficient cover construction and reliable integration," *Sonderforschungsbereich 256, Institut für Angewandte Mathematik, Universität Bonn.*

[97] T. Gertsner and M. Griebel, "Numerical integration using sparse grids," *Numerical Algorithms*, vol. 18, pp. 209–232, 1998.

[98] G. W. Miller J. L. Junkins and J. R. Jancaitis, "A weighting function approach to modeling of geodetic surfaces," *Journal of Geophysical Research*, vol. 79, pp. 23, 1974.

[99] J. R. Jancaitis and J. L. Junkins, "Modeling N-dimensional surfaces using a weighting function approach," *Journal of Geophysical Research*, vol. 78, no. 11, pp. 1794–1803, 1973.

[100] G. W. Miller J. L. Junkins and J. R. Jancaitis, "A weighting function approach to modeling irregular surfaces," *American Geophysical Unions Transactions*, vol. 53, no. 4, pp. 346, 1972.

[101] P. Singla, *Multi-Resolution Methods for High Fidelity Modeling and Control Allocation in Large-Scale Dynamical Systems*, Ph.D. Dissertation, Department of Aerospace Engineering, Texas A&M University, College Station, 2006.

[102] H. Niederreiter, *Random number generation and quasi-Monte Carlo methods*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1992.

[103] V. E. Rosca and V. M. A. Leit ao, "Quasi Monte Carlo meshfree integration for meshless weak formulations," *Engineering Analysis with Boundary Elements*, vol. 32, pp. 471–479, 2008.

[104] Y. Saad, *Numerical Methods for Large Eigenvalue Problems*, Manchester University Press, UK, 1992.

[105] A. A. van der Vorst, *Iterative Krylov Methods for Large Linear Systems*, Cambridge Monographs on Applied and Computational Mathematics. Cambridge, UK, Cambridge University Press, 2003.

[106] P. Singla M. Kumar, S. Chakravorty and J. L. Junkins, "The partition of unity finite element approach with *hp*-refinement for the stationary Fokker-Planck equation," *Journal of Sound and Vibration, Elsevier*, vol. 327, no. 1-2, pp. 144–162, 2009.

[107] M. Kumar, S. Chakravorty, and J. L. Junkins, "A semianalytic meshless approach to the transient Fokker-Planck equation," *Probabilistic Engineering Mechanics, Elsevier*, under review.

[108] R. E. Bellman, *Dynamic Programming*, Princeton University Press, Princeton, New Jersey, 1957.

[109] D. P. Bertsekas, *Dynamic Programming and Optimal Control, Vol. I*, Athena Scientific, Cambridge, MA, 2000.

[110] D. P. Bertsekas, *Dynamic Programming and Optimal Control, Vol. II*, Athena Scientific, Cambridge, MA, 2000.

[111] P. Werbos, "Approximate dynmaic programming for real-time control and neural modeling," in *Handbook of Intelligent Control*. pp. 493–525, New York, Van Nostrand Reinhold.

[112] P. Marbach, *Simulation Based Optimization of Markov Reward Processes, PhD Dissertation*, Massachusetts Institute of Technology, Boston, MA, 1999.

[113] J. Baxter and P. Bartlett, "Infinite horizon policy-gradient approximation," *Journal of Artificial Intelligence Research*, vol. 15, pp. 319–350, 2001.

[114] Satinider Singh R. S. Sutton, David McAllester and Yishay Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Proc. 1999 Neural Information Processing Systems*, Denver, CO, 1999, vol. 12 of *Advances in Neural Information Processing Systems*, pp. 1057 – 1063.

[115] M. Lagoudakis and R. Parr, "Least squares policy iteration," *Journal of Machine Learning Research*, vol. 4, pp. 1107–1149, 2003.

[116] H. J. Kushner, "Numerical methods for stochastic control problems in continuous time," *SIAM Journal on Control and Automation*, vol. 28, pp. 999–1048, 1990.

[117] L. G. Crespo and J. Q. Sun, "Nonlinear stochastic control via stationary probability density functions," 2002, pp. 2029–2034.

[118] L. G. Crespo and J. Q. Sun, "Stochastic optimal control via Bellman's principle," *Automatica*, vol. 39, pp. 2109–2114, 2003.

[119] C. S. Hsu, *Cell to Cell Mapping: A method for the Global Analysis of Nonlinear Systems*, Springer-Verlag, New York, NY, 1987.

[120] F. B. Hanson, "Computational stochastic dynamic programming on a vector multiprocessor," *IEEE Transactions on Automatic Control*, vol. 36, no. 4, pp. 507–511, 1991.

[121] R. W. Beard and T. W. McLain, "Successive Galerkin approximation algorithms for nonlinear optimal and robust control," *International Journal of*

Control: Special Issue on Breakthroughs in the Control of Nonlinear Systems, vol. 71, no. 5, pp. 717–743, 1998.

[122] T. W. McLain and R. W. Beard, "Nonlinear optimal control design of a missile autopilot," Boston, MA, Aug. 10-12, 1998, pp. AIAA Paper 1998–4321.

[123] S. Jagannathan F. L. Lewis and A. Yesildirak, *Neural Network Control of Robotic Manipulators and Nonlinear Systems*, Series in Systems and Control. London, Taylor and Francis, 1999.

[124] M. Abu Kahlaf and F. L. Lewis, "Nearly optimal state feedback control of constrained nonlinear system using a neural network HJB," *Annual Reviews in Control*, vol. 28, pp. 239–251, 2004.

[125] J. Huang M. Abu-Khalaf and F. L. Lewis, *Nonlinear H2/ H-∞ Constrained Feedback Control: A Practical Approach Using Neural Networks*, New York, Springer Verlag, June 2006.

[126] Radhakant Padhi and S. N. Balakrishnan, "Proper orthogonal decomposition based optimal neurocontrol synthesis of a chemical reactor process using approximate dynamic programming," *Neural Netw.*, vol. 16, no. 5-6, pp. 719–728, 2003.

[127] S. Ferrari, *Algebraic and Adaptive Learning in Neural Control Systems, PhD Dissertation*, Princeton University, 2002.

[128] A. E. Bryson and Y.C. Ho, *Applied Optimal Control*, Hemsiphere Publsihing Co., Washington, D.C., 1975.

[129] I. M. Ross and F. Fahroo, "Pseudospectral knotting methods for solving nonsmooth optimal control problems," *AIAA Journal of Guidance Control and*

Dynamics, vol. 27, pp. 397–405, 2004.

[130] I. M. Ross and F. Fahroo, "Issues in real-time computation of optimal control," *Mathematical and Computer Modeling*, vol. 43, pp. 1172–1188, 2006.

[131] M. L. Puterman, *Markov Decision Processes*, Wiley InterScience, Hoboken, NJ, 2005.

[132] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Trans. ASME Journal of Basic Engineering*, pp. 35–45, 1960.

[133] H. J. Kushner, "The Cauchy problem for a class of degenerate parabolic equations and asymptotic propoerties of related diffusion equations," *Journal of Differential Equations*, vol. 6, pp. 209–231, 1969.

[134] A. H. Jazwinski, *Stochastic Processes and Filtering Theory*, Academic Press, New york, NY, 1970.

[135] H. Bruyninckx T. Lefebvre and J. De Schutter, "Comment on 'a new method for the nonlinear transformations of means and covariances in filters and estimators'," *IEEE Transactions on Automatic Control*, vol. 47, no. 8, Aug. 200.

[136] J. Uhlmann S. Julier and H. Durant-Whyte, "A new approach for the nonlinear transformation of means and covariances in filters and estimators," *IEEE Transactions on Automatic Control*, vol. AC-45(3), pp. 477–482, 2000.

[137] S. J. Julier and J. K. Uhlmann, "Unscented filtering and nonlinear estiamtion," 2004, vol. 92, pp. 401–422.

[138] K. Ito and K. Xiong, "Gaussian filters for nonlinear filtering problems," *IEEE Transactions on Automatic Control*, vol. 45, no. 5, pp. 910–927, May 2000.

[139] R. Van Der Merwe and E. A. Wan, "The square root unscented Kalman filter for state and parameter estimation," in *Proc. of the International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, Salt Lake City, UT, May 7-11, 2001, vol. 6, pp. 3461–3464.

[140] T. Singh G. Terejanu, P. Singla and P. D. Scott, "Gaussian sum filter with weight adaption between observations," *Signal Processing*, under review.

[141] J. Gunther J. Lawton R. Beard, J. Kenney and W. Stirling, "Nonlinear projection filter based on Galerkin approximation," *Journal of Guidance, Control and Dynamics*, vol. 22, no. 2, pp. 258–266, Mar-Apr 1999.

APPENDIX A

REVIEW OF STOCHASTIC DYNAMICAL SYSTEMS

The origins of probability theory can be traced back to the age of Renaissance, when mathematicians and gamblers alike were interested in determining the chance of winning hands. Notable contributions from researchers like Pierre de Fermat, Blaise Pascal and Christiaan Huygens led to development of the counting theory of probability, which dealt only with discrete variables and probabilities were expressed in terms of frequency of positive outcomes. The modernization of set theory during the nineteenth century and development of measure theory during the twentieth century led to the establishment of an axiomatic theory of probability, motivated by analytical concerns of dealing with both discrete and continuous random variables. The works of Andrey Kolmogorov, published in 1933 [16] are central to the formalization of the axiomatic theory of probability and below we review the basic concepts of this theory.

Definition VII.1 ***Sample space:*** *Denoted by* Ω*, sample space is the set of all possible outcomes of an experiment. An* event *is a subset of* Ω*.*

Definition VII.2 $\sigma-$***algebra:*** *Denoted by* $\mathcal{F}$*, a* σ*-algebra (also known as* σ*-field) is a collection of subsets of* Ω *such that:*

1. $\Omega \in \mathcal{F}$
2. $B \in \mathcal{F} \Rightarrow B^C \in \mathcal{F}$ *["closed under complementation"]*
3. $B_i \in \mathcal{F}, i \in \mathbb{N} \Rightarrow \bigcup_{i=1}^{\infty} \in \mathcal{F}$ *["closed under countable union"]*

The sample space, Ω, is the set of all possible outcomes of an experiment, keeping in mind that it could be an uncountable set. A σ-field is a measure of the amount of information available about the experiment. Of course, if everything is known, then $\mathcal{F} = \mathfrak{P}(\Omega)$, where $\mathfrak{P}(\cdot)$ denotes the power set. An important $\sigma-$algebra in stochastic dynamics is the Borel σ-algebra for $\Re^N$, which is defined as the σ-algebra generated by the open sets (N-hypercuboids) in $\Re^N$, and is denoted by $\mathcal{B}(\Re^N)$, or simply $\mathcal{B}$. Also, if $\mathcal{F}_1$ and $\mathcal{F}_2$ are two σ-fields, and $\mathcal{F}_1 \subseteq \mathcal{F}_2$, then $\mathcal{F}_1$ is called a sub-σ-algebra of $\mathcal{F}_2$. In terms of information, $\mathcal{F}_2$ carries at least as much information as $\mathcal{F}_1$, possibly depicting a nested sequence of information with subscripts 1 and 2 denoting time instances. Finally, consider a collection $\mathcal{U}$ of subsets of Ω, i.e. $\mathcal{U} \subseteq \mathfrak{P}(\Omega)$; then the σ-field *generated by* $\mathcal{U}$ is the smallest σ-field containing $\mathcal{U}$ and is denoted by $\sigma(\mathcal{U})$. With this background, we are in a position to define the probability space:

Definition VII.3 ***Probability space:*** *A triple* $(\Omega, \mathcal{F}, P)$ *is called a probability space, where:*

- $\Omega \neq \phi$, *and* $\mathcal{F}$ *is a* σ*-algebra of subsets of* Ω
- P *is a* ***probability measure****, i.e. a function* $P : \mathcal{F} \to [0, 1]$ *such that:*
  1. $P(\Omega) = 1$
  2. *If* $A_i \in \mathcal{F}$, $i \in \mathbb{N}$ *are disjoint; then*

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i) \ \text{[countable additivity, or } \sigma\text{-additivity]} \qquad \text{(A.1)}$$

The above two axioms associated with the probability measure P helped make probability theory rigorous in the 1930's. The word "measure" is key because the positive function P endows a sense of quantification to the available information, i.e. the σ-algebra $\mathcal{F}$. In fact, if the sample space is non-empty, it is always possible to define a probability measure on its σ-algebras. As a result, the pair $(\Omega, \mathcal{F})$, $\Omega \neq \phi$

is called a measureable space. The probability space is the foundation on which all subsequent concepts of stochastic dynamical systems are grounded. A very important idea related to growth of information within a probability space, especially relevant to filtering theory is that of the *filtration*:

Definition VII.4 ***Filtration:*** *Given a probability space,* $(\Omega, \mathcal{F}, P)$*, the collection* $\{\mathcal{F}_i : i \geq 0\}$ *of sub-*σ*-algebras of* $\mathcal{F}$ *is called a filtration if* $\mathcal{F}_n \subseteq \mathcal{F}_{n+1}$ *for all* n*.*

A filtration captures the idea that available information ($\mathcal{F}_i$) accumulates with time, but complete set of relevant information ($\mathcal{F}$) may never be known. This is central in filtering theory because the objective is to obtain the probability distribution of the state conditioned on all available information at the present time ($\sim$ posterior state density). The "state" under question at any given time is a random variable defined on the probability space that is governed by nonlinear dynamics perturbed by "noise". We are now in a position to define another important function on $(\Omega, \mathcal{F})$, namely the random variable:

Definition VII.5 ***Random variable:*** *Given a measureable space* $(\Omega, \mathcal{F})$*, a function* $X : \Omega \rightarrow \Re^N$ *is a random variable if*

$$X^{-1}(B) \in \mathcal{B}, \ \ \forall B \in \mathcal{B}(\Re^N) \tag{A.2}$$

In other words, a random variable is a measureable mapping, $X : (\Omega, \mathcal{B}) \mapsto (\Re^N, \mathcal{B}(\Re^N))$. Note that the inverse superscript on X is not to be understood as a function inverse, rather, as the inverse image. This general definition of random variables makes it applicable to both discrete and continuous measureable spaces. It can be roughly understood as a variable that can assume a range of values with different probabilities. Such probabilities can be quantified in terms of the so called probability distribution function of X, denoted by $F_X(V \in \Re^N)$. For a single dimensional random variable,

it can be defined as follows:

Definition VII.6 ***Probability distribution function:*** *The function* $F_X(x) \triangleq P(X \in (-\infty, x]), x \in \Re$ *is known as the probability distribution function of the one-dimensional random variable* X *and has the following properties:* F *is right continuous and monotone nondecreasing, if* $F(\infty) \triangleq \lim_{x\to\infty} F(x)$*, then* $F(\infty) = 1$*; and if* $F(-\infty) \triangleq \lim_{x\to-\infty} F(x)$*, then* $F(-\infty) = 0$*.*

Clearly, $F_X(x)$ gives the probability of finding X on the real line to the left of x, including x. This definition covers both discrete and continuous random variables and can be extended to multi-dimensional case in analogous fashion with ease. For continuous random variables, it is possible to define the derivative of the distribution function, known as the probability density function (pdf), denoted by $\mathcal{W}_X(x)$. The pdf is especially important because it depicts probability density over the real space (similar to mass density), the integration of which over the region of interest gives the probability of finding the random variable in that region. It is often desired to know the average value of a function of the random variable, which can be done by integrating (summing for discrete variables) the concerned function weighted by the probability density:

Definition VII.7 ***Expectation:*** *The expected value of a function* $f(X)$ *of a random variable* X *on* $(\Omega, \mathcal{F}, P)$ *is understood in the Lebesbue-Stieltjes sense of integration as:*

$$E[f(X)] \triangleq \int_{\Omega} f(\mathbf{x})\mathcal{W}_X(\mathbf{x})d\mathbf{x} \tag{A.3}$$

When $f(X) = X$, the integral is known as the mean value of the random variable while $f(X) = (X - E[X])^2$ is well known as the covariance. Expected value of the function over a smaller region $\Delta \subset \Omega$ can be obtained by changing the domain of

integration. We now proceed to the idea of the random process, which generalizes the concept of random variables by incorporating time in the framework.

Definition VII.8 ***Random process/stochastic process:*** *Let T ($\sim$ time) be an index set. A random process is a parameterized collection of random variables, $\{X_t : t \in T\}$ defined on a probability space $(\Omega, \mathcal{F}, P)$.*

Alternate ways of writing a stochastic process are X_t, $X_t(\omega)$ and $X(t, \omega)$. The state of a dynamical system perturbed by noise is modeled as a stochastic process. Note that for a fixed $t \in T$, the function $\omega \in \Omega \mapsto X_t(\omega)$ is a random variable which has its own pdf, $\mathcal{W}_{X_t}(x)$. For a fixed $\omega \in \Omega$, the function $t \in T \mapsto X_t(\omega)$ is called a path. Indeed, implicit is the fact that a stochastic process is really a spatiotemporal function: $X_t(\omega) : T \times \Omega \mapsto \Re^N$ with the special condition that for each $t \in T$, $X(t, \cdot) : \Omega \mapsto \Re^N$ is a random variable. Before we proceed to formulating stochastic differential equations for random processes, two more concepts must be considered:

Definition VII.9 ***Finite dimensional distributions:*** *Given a stochastic process $\{X_t\}_{t \in T}$, finite dimensional distributions are the measures $\mu_{t_1,\ldots,t_k}$ on $\Re^{Nk}$ defined as follows: For $t_1, \ldots, t_k \in T$ and $F_1, \ldots, F_k \in \mathcal{B}(\Re^N)$;*

$$\mu_{t_1,\ldots,t_k}(F_1 \times \ldots \times F_k) = P(X_{t_1} \in F_1, \ldots, X_{t_k} \in F_k) \tag{A.4}$$

From Eq.A.4, it is easy to see that for $k = 1$, the finite dimensional distribution is nothing but the probability distribution function for X_{t_1}. It is important to note that two stochastic processes with the same finite dimensional distributions can have completely different paths. We next consider a special random process known as Brownian motion:

Definition VII.10 ***Brownian motion:*** *The single-dimensional Brownian motion is a real valued stochastic process $\{B_t\}_{t \geq 0}$ such that*

1. *For* $t_0 < t_1 < \ldots < t_n$, $B(t_0), B(t_1) - B(t_0), \ldots, B(t_n) - B(t_{n-1})$ *are independent for any* n.
2. *For* s *and* $t \geq 0$, $B(s+t) - B(s) \sim N(0,t)$, *i.e. the increment process is normally distributed with zero mean and covariance* t.
3. *The paths* $t \mapsto B_t$ *are continuous with probability 1, i.e.* $P(t \mapsto B_t \text{ is continuous}) = 1$.

The above properties of Brownian motion are consistent with its use as a model for motion of pollen grains suspended in a liquid. The N-dimensional Brownian motion is a $\Re^N$-valued process whose components are individual one-dimensional Brownian motion processes. More importantly, its *formal* derivative is nothing but the white noise process, which can be written in terms of differentials as: $dB_t = W_t dt$. In fact, Brownian motion is the simplest *continuous* stochastic process that can be used to model noise. This leads us to dynamical systems perturbed by noise, written very roughly as:

$$\frac{d\mathbf{X}}{dt} = \mathbf{f}(t, \mathbf{X}) + \mathbf{g}(t, \mathbf{X}).\text{“white noise”} \tag{A.5}$$

Because of obvious difficulties associated with writing white noise as the rigorous derivative of Brownian motion (which is differential nowhere), the above equation can be written in terms of differentials, a form known popularly as the Itô stochastic differential equation (SDE):

$$d\mathbf{X}_t = \mathbf{f}(t, \mathbf{X}_t)dt + \mathbf{g}(t, \mathbf{X}_t)d\mathbf{B}_t, \quad 0 \leq t \leq T \tag{A.6}$$

Eq.A.6 represents a continuous dynamical system comprising a deterministic part $\mathbf{f}dt$, which in general may be nonlinear and a stochastic part $\mathbf{g}d\mathbf{B}_t$. The state of the system, $\mathbf{X}_t$ is a stochastic process (because of the perturbation term) which usually has a known initial probability density function, $\mathcal{W}_{\mathbf{X}_0}(\mathbf{x})$, or simply, $\mathcal{W}_0(\mathbf{x})$. It is

in general very difficult to obtain explicit solutions to Eq.A.6 and precise conditions of existence and uniqueness can be stated as follows: Let $T > 0$ and suppose $\mathbf{f} : [0,T] \times \Re^N \mapsto \Re^N$ and $\mathbf{g} : [0,T] \times \Re^N \mapsto \Re^{N \times M}$ are measureable functions such that for some constants C and D,

$$|\mathbf{f}(t,\mathbf{x})| + |\mathbf{g}(t,\mathbf{x})| \leq C[1 + |x|], \tag{A.7}$$

$$|\mathbf{f}(t,\mathbf{x}) - \mathbf{f}(t,\mathbf{y})| + |\mathbf{g}(t,\mathbf{x}) - \mathbf{g}(t,\mathbf{y})| \leq D|x - y| \tag{A.8}$$

where, $t \in [0,T]$ and $x, y \in \Re^N$. If $\mathbf{B}_t$ is a M dimensional Brownian motion and $\mathbf{X}_0$ is a random variable independent of $\mathcal{F}_T$ with finite second moments, i.e. $E[|\mathbf{X}_0|^2] < \infty$, then the SDE A.6 has an almost sure unique solution such that $E[\int_0^T |X_s|^2 ds] < \infty$. The "growth condition" of Eq.A.7 ensures that the solution $\mathbf{X}_t$ does not explode in finite time while Eq.A.8 represents a Lipschitz condition ensuring uniqueness of solution. Almost sure uniqueness (also known as strong uniqueness and pathwise uniqueness) means that if $\mathbf{X}_1$ and $\mathbf{X}_2$ satisfy Eq.A.6, then $\mathbf{X}_1(t,\omega) = \mathbf{X}_2(t,\omega)$ for all $t \leq T$ almost surely. It is important to note that these conditions of existence and uniqueness are very strong and explicit solutions to Eq.A.6 usually do not exist, or aren't very useful. Typically, it is most beneficial to solve for the time varying probability density function of $\mathbf{X}$, i.e. $\mathcal{W}_{\mathbf{X}}(t,\mathbf{x})$. This leads to the notion of weak uniqueness, under which if $\mathbf{X}_1$ and $\mathbf{X}_2$ are solutions of Eq.A.6, then they have the same probability density function. In this sense, it is sufficient to solve for the probability density function of the state $\mathbf{X}_t$. Once the time varying pdf is available, the average behavior of the system state can be analyzed in terms of its various expectations. We conclude this section by stating that for the SDE in Eq.A.6, there exists a parabolic partial differential equation known as Fokker-Planck equation, the solution to which gives the time varying pdf of state $\mathbf{X}_t$, and is considered in the next section.

APPENDIX B

HALTON SEQUENCE OF QUASI RANDOM NUMBERS

The following algorithm describes Halton's algorithm for generating quasi random numbers in $\Re^N$:

- Given domain of interest: $\Omega = \otimes_{i=1}^{N}[a_i, b_i]$.
- Start with a sequence of natural numbers: $\mathcal{S}_P = \{q_1, q_2, \ldots, q_P\}$. Define $\mathcal{G} = \{p_1, \ldots, p_N\}$ as the set of first N prime numbers.
- For each $\mathbf{q}_i \in \mathcal{S}_P$, do the following:

1. Set $n = 1$
2. Obtain base p_n representation of q_i as ${}^nB_i = \{b_1, b_2, \ldots, b_m\}$.
3. Reverse the above string to: ${}^n\tilde{B}_i = \{b_m, \ldots, b_2, b_1\} \triangleq \{c_1, c_2, \ldots, c_m\}$.
4. Obtain the n^{th} component of the psuedorandom number as: ${}^nx_i = \sum_{k=1}^{m} c_k p_n^{-k-1}$
5. Change $n \mapsto n + 1$, go back to step 2 until $n = N$.
6. Repeat above steps for each member of the sequence $\mathcal{S}_P$.

The above steps generates psuedorandom numbers in the canonical domain $I = \otimes_{i=1}^{N}[0, 1]$. To transform the sequence to the desired domain Ω, use the following linear transformation:

$$ {}^nx_i \mapsto a_n +^n x_i(b_n - a_n); \quad i = 1, 2, \ldots, P \tag{B.1} $$

The weights associated with each of the quadrature points is $w = \frac{\prod_{n=1}^{N}(b_n - a_n)}{P}$

APPENDIX C

DERIVATION OF THE 1-D RICCATI EQUATION FROM FPE

This appendix presents a derivation of the single dimensional Riccati equation from FPE. This simple exercise demonstrates that the Kalman filter propagation is a special case of the full nonlinear uncertainty evolution problem.

To start, consider the following unit dimensional linear stochastic system:

$$\dot{x} = ax + b\zeta \tag{C.1}$$

Then, FPE corresponding to the above dynamical system is:

$$\frac{\partial \mathcal{W}}{\partial t} = -ax\frac{\partial \mathcal{W}}{\partial x} - a\mathcal{W} + \frac{1}{2}b^2 q\frac{\partial^2 \mathcal{W}}{\partial x^2} \tag{C.2}$$

Since Eq.C.1 represents a linear system, the time varying pdf of state x is Gaussian, i.e. $\mathcal{W}(t,x) = N(\mu(t), \sigma(t))$, or,

$$\mathcal{W}(t,x) = \frac{\kappa}{\sigma}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \equiv \frac{\kappa}{\sigma}G(t,x) \tag{C.3}$$

where, $G(t,x) = \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$. Substituting Eq.C.3 in Eq.C.2, we the following partial derivatives:

$$\frac{\partial \mathcal{W}}{\partial t} = \frac{\mathcal{W}}{\sigma}\left[(z^2-1)\dot{\sigma} + z\dot{\mu}\right] \tag{C.4}$$

$$\frac{\partial^2 \mathcal{W}}{\partial x} = -\frac{\mathcal{W}}{\sigma}z \tag{C.5}$$

$$\frac{\partial^2 \mathcal{W}}{\partial x^2} = \frac{\mathcal{W}}{\sigma^2}(z^2-1) \tag{C.6}$$

where, $z = \frac{x-\mu}{\sigma}$ is an auxiliary variable. Substituting the above partials back into Eq.C.2, we get:

$$z^2\dot{\sigma} + z\dot{\mu} - \dot{\sigma} = \left(a\sigma + \frac{b^2q}{2\sigma}\right)z^2 + a\mu z - \left(a\sigma + \frac{b^2q}{2\sigma}\right) \tag{C.7}$$

Comparing coefficients of z^2, z and the constant term from both sides, we get:

$$\dot{\sigma} = a\sigma + \frac{b^2q}{2\sigma} \tag{C.8}$$

$$\dot{\mu} = a\mu \tag{C.9}$$

$$-\dot{\sigma} = -\left(a\sigma + \frac{b^2q}{2\sigma}\right) \tag{C.10}$$

Note that Eqs.C.8 and C.10 are exactly the same. Also, Eq.C.8 is not the desired form of covariance propagation. We are actually interested in knowing $\dot{\sigma^2}$, which is nothing but $\dot{\nu} = 2\sigma\dot{\sigma}$. So we get:

$$\dot{\mu} = a\mu \tag{C.11}$$

$$\dot{\nu} = 2a\nu + b^2q \tag{C.12}$$

The above equations are nothing but the Riccati equations of uncertainty propagation for a scalar linear system. the matrix version can be derived similarly.

VITA

Mrinal Kumar was born in Ranchi, India. After graduating from Delhi Public School in Ranchi, Mrinal attended the Indian Institute of Technology in Kanpur, India, earning a Bachelor of Technology degree in aerospace engineering in 2004. He joined the aerospace engineering department of Texas A&M University in fall 2004 to work towards a doctoral degree under the supervision of Dr. Suman Chakravorty and Dr. John L. Junkins. He has published 8 journal papers and won research awards from the American Institute of Aeronautics and Astronautics, American Astronautical Society and the George Bush Presidential Library Foundation.

He can be reached via email at mrinalkumar@gmail.com or by contacting Dr. Suman Chakravorty at 3141 TAMU, College Station, TX 77843-3141.