Spring 1-1-2015

Uncertainty Quantification via Sparse Polynomial Chaos Expansion

Ji Peng
University of Colorado Boulder, jpeng.ustc@gmail.com

UNCERTAINTY QUANTIFICATION VIA SPARSE POLYNOMIAL CHAOS EXPANSION

by

JI PENG

B.S., University of Science and Technology of China, 2010

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Mechanical Engineering

2015

This thesis entitled:
Uncertainty Quantification via Sparse Polynomial Chaos Expansion
written by Ji Peng
has been approved for the Department of Mechanical Engineering

Prof. Alireza Doostan

Prof. Kurt Maute

Date

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

Peng, Ji (Ph.D., Mechanical Engineering)

Uncertainty Quantification via Sparse Polynomial Chaos Expansion

Thesis directed by Prof. Alireza Doostan

Uncertainty quantification (UQ) is an emerging research area that aims to develop methods for accurate predictions of quantities of interest (QoI's) from complex engineering systems, as well as quantitative validation of the associated mathematical models, with presence of random inputs. To perform a comprehensive UQ analysis, polynomial chaos expansion (PCE) is now a commonly used approach in which the QoI is represented in a series of multi-variate polynomials that are orthogonal with respect to the measure of the inputs. Traditional methods for PCE, such as Monte Carlo, stochastic collocation, least-squares regression, are known to suffer from either slow convergence rate or rapid growth of the computational cost (as the number of random inputs increases) in identifying the PCE coefficients. When the PCE coefficients are sparse, i.e., many of them are negligible, it has been shown that compressive sampling is an effective technique to identify the coefficients with smaller number of system simulations.

In the context of compressive sampling, this thesis presents new approaches which improve the accuracy of identifying PCE coefficients, and therefore the PCE itself. In detail, a weighted ℓ_1-minimization including *a priori* information about the PCE coefficients, a bi-fidelity ℓ_1-minimization, a bi-fidelity orthogonal matching pursuit (OMP), and an ℓ_1-minimization including the derivatives of QoI with respect to the random inputs are proposed. Both theoretical analyses and numerical experiments are presented to demonstrate that all the proposed approaches reduce the cost of computing a PCE.

For a QoI whose PCE with respect to the measure of the underlying random inputs is not sparse, a polynomial basis design is proposed where, in addition to the coefficients, the basis functions are also learned from the simulation data. The approach has been empirically shown to find the *optimal* basis which makes the PCE converge more rapidly, and enhances the accuracy of the PCE approximation.

DEDICATION

To my beloved father, who is fighting against pancreatic cancer

To my dear mother

ACKNOWLEDGEMENTS

The first person I am grateful to deeply in heart is my advisor, Prof. Alireza Doostan, under whose mentoring I conduct my research. His patience, enthusiasm, broad knowledge, and deep insight have impressed and helped me all the way along. He guided me to become a disciplined thinker, which I will be grateful for forever.

Meanwhile, I would like to thank my thesis committee members: Prof. Daven Henze, Prof. Brandon Jones, Prof. Kurt Maute, and Prof. Oleg Vasilyev, for their support and engagement. All the discussions I had with them are insightful and indispensable for the completion of this thesis.

Financial supports by the U.S. Department of Energy Office of Science, Office of Advanced Scientific Computing Research, under Award Number DE-SC0006402, and NSF grants DMS-1228359 and CMMI-1454601 are greatly acknowledged. This work utilized the Janus supercomputer, which is supported by the National Science Foundation (award number CNS-0821794) and the University of Colorado Boulder. The Janus supercomputer is a joint effort of the University of Colorado Boulder, the University of Colorado Denver and the National Center for Atmospheric Research.

Also, I thank the post-doctoral researchers in our group for their support and help in my research: Dr. Jerrad Hampton, and Dr. Dave Biagioni, as well as my fellow friends in the group: Mohammad Hadigol and Hillary Fairbanks.

Last but not least, I am grateful to my family for being supportive throughout. They have been and will be the motivation for my life.

CONTENTS

CHAPTER

TABLES

Table

FIGURES

Figure

CHAPTER 1

INTRODUCTION AND BACKGROUND

1.1 Uncertainty Quantification (UQ)

In the world of modeling and simulation of complex engineering systems, such as those exhibiting multiple physics and/or multiple scales, two stark realities exist: First, all models are approximations of their target phenomena and, second, uncertainties exist due to both lack of knowledge and inherent variability. Realistic analysis and design of such systems, therefore, require not only a deep understanding of the underlying physics and their interactions but also recognition of modeling errors, uncertainties, and their influences on quantities of interest (QoI's). The means to formally assess the predictive capability of a given simulation model has come to be known as model Verification and Validation (V&V). Uncertainty quantification (UQ) [2, 3, 4] enters this process through a number of avenues: (1) UQ tools are required to assimilate parameters of mathematical models (or often models themselves) based on their sparse and limited observations, (2) A challenging task is to efficiently propagate model uncertainties to estimate uncertainties in response quantities of interest (uncertainty propagation), and (3) UQ techniques are employed in the validation process to meaningfully compare model predictions with often sparse and limited experimental data. There are two types of uncertainty: epistemic and aleatory. Epistemic uncertainty is a potential deficiency that is due to a lack of knowledge, which can arise from assumptions introduced in the derivation of the mathematical model, for instance, turbulence model assumptions. Epistemic uncertainty can be reduced by increasing the knowledge about the system, such as a more accurate physical model or more experimental investigation. Aleatory uncertainty is the intrinsic variability present in the system. Unlike epistemic uncertainty, aleatory uncertainty cannot be reduced, and it can only be better characterized with additional information. Research on UQ has received recent attention in a variety research domains, such as computational fluid dynamics (CFD) [3, 5, 6, 7], computational structural mechanics [8, 9, 10], and fluid structure interaction (FSI) [11, 12, 13].

There are two means to propagate uncertainty: intrusive [14] and non-intrusive [15, 16, 17, 18]. In this thesis, non-intrusive methods are adopted as they employ the legacy codes as a black box to propagate uncertainty.

1.2 Polynomial Chaos Expansion (PCE)

Probability is a natural mathematical framework well suited for describing uncertainty, and so we assume that the uncertain system inputs are described by a vector of independent random variables, $\boldsymbol{\Xi}$, defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, which is formed by the product of d probability spaces, $(\mathbb{R}, \mathbb{B}(\mathbb{R}), \mathbb{P}_k)$ corresponding to each coordinate of $\boldsymbol{\Xi}$, denoted by Ξ_k; here $\mathbb{B}(\cdot)$ represents the Borel σ-algebra. We further assume that the random variable Ξ_k is continuous and distributed according to the density ρ_k implied by $\mathbb{P}_k$. Note that this entails $\Omega = \mathbb{R}^d$, $\mathcal{F} = \mathbb{B}(\mathbb{R}^d)$, that each Ξ_k is independently distributed, and that the joint distribution for $\boldsymbol{\Xi}$, denoted by ρ, equals the tensor product of the marginal distributions ρ_k.

Spectral methods [3, 4] are commonly utilized to represent the QoI, denoted by $u(\boldsymbol{\Xi})$, as a function of $\boldsymbol{\Xi}$, and in this work, we employ polynomial chaos expansions (PCEs) [19, 20]. Specifically, for each ρ_k we define $\{\psi_{i_k}\}_{i_k \geq 0}$ to be the complete set of orthonormal polynomials of degree i_k with respect to the weight function ρ_k [21, 20]. As a result, the multivariate orthonormal polynomials for $\boldsymbol{\Xi}$ are given by the products of the univariate orthonormal polynomials,

$$\psi_{\boldsymbol{i}}(\boldsymbol{\Xi}) = \prod_{k=1}^{d} \psi_{i_k}(\Xi_k), \tag{1.1}$$

where each i_k, representing the kth coordinate of the multi-index $\boldsymbol{i}$, is a non-negative integer. With these orthogonal polynomials, we can represent the QoI with finite variance as a function of Ξ, in the form of spectral expansion, consisting of orthogonal polynomials, $\{\psi_{\boldsymbol{i}}(\Xi)\}$:

$$u(\Xi) = \sum_{\boldsymbol{i}\in\mathcal{I}} c_{\boldsymbol{i}}\psi_{\boldsymbol{i}}(\Xi) \ , \tag{1.2}$$

where the set of d-dimensional multi-index $\mathcal{I} = \{(i_1, \ldots, i_d) : i_k \in \mathbb{N}\cup\{0\}\}$. For computation, we truncate the expansion in (1.2) to the set of P basis functions associated with the subspace of polynomials of total order not greater than p, that is $\sum_{k=1}^{d} i_k \leq p$. For convenience, we also order these P basis functions so that they are indexed by $\{1, \cdots, P\}$ as opposed to the vectorized indexing in (1.1). The basis set $\{\psi_j\}_{j=1}^{P}$ has the cardinality

$$P = \frac{(d+p)!}{d!p!}. \tag{1.3}$$

For the interest of presentation, we interchangeably use both notations for representing PCE basis. With the latter notation, the PCE in (1.2) and its truncation are then defined by

$$u(\Xi) = \sum_{j=1}^{\infty} c_j\psi_j(\Xi) \approx \sum_{j=1}^{P} c_j\psi_j(\Xi). \tag{1.4}$$

Though u is an arbitrary function in $L_2(\Omega, \mathbb{P})$, we are limited to an approximation in the span of our basis polynomials, and the error incurred from this approximation is referred as *truncation error*.

As the the PCE basis functions are orthogonormal,

$$\int_{\Gamma} \psi_m(\boldsymbol{\xi})\psi_n(\boldsymbol{\xi})\rho(\boldsymbol{\xi})d\boldsymbol{\xi} = \delta_{mn}, \tag{1.5}$$

where δ_{mn} is the Kronecker delta, the coefficients may be computed by non-intrusive spectral projection (NISP) [3], which projects $u(\Xi)$ onto the basis function $\psi_j(\Xi)$ such that

$$c_j = \int_{\Gamma} u(\Xi)\psi_j(\Xi)\rho(\Xi)d\Xi. \tag{1.6}$$

Assuming that for each k, ρ_k is known *a priori*, in Chapter 2-4, our study concentrates on the cases that probability densities for ρ_k are uniform and Gaussian; the corresponding polynomial bases are, respectively, Legendre and Hermite polynomials. In addition, in Chapter 5, Beta probability density and Jacobi polynomials are also employed. We note that there are more PCE basis functions and measure of Ξ could be utilized [20], depending on the given problems.

1.3 Problem Setup

Let the random vector Ξ characterize the input uncertainties, and we consider the solution of a partial differential equation defined on a bounded Lipschitz continuous domain $\mathcal{D} \subset \mathbb{R}^D$, $D \in \{1,\ 2,\ 3\}$, with boundary $\partial\mathcal{D}$. The uncertainty implied by Ξ may be represented in one or many relevant parameters, e.g., the diffusion coefficient, boundary conditions, and/or initial conditions. Letting $\mathcal{L}, \mathcal{I}$, and $\mathcal{B}$ depend on the physics of the problem being solved, the solution u satisfies the three constraints

$$\begin{aligned} \mathcal{L}(\boldsymbol{x}, t, \Xi; u(t, \boldsymbol{x}, \Xi)) &= 0, \qquad \boldsymbol{x} \in \mathcal{D}, \\ \mathcal{I}(\boldsymbol{x}, 0, \Xi; u(0, \boldsymbol{x}, \Xi)) &= 0, \qquad \boldsymbol{x} \in \mathcal{D}, \\ \mathcal{B}(\boldsymbol{x}, t, \Xi; u(t, \boldsymbol{x}, \Xi)) &= 0, \qquad \boldsymbol{x} \in \partial\mathcal{D}. \end{aligned} \tag{1.7}$$

In this work, we assume that conditioned on the ith random realization of Ξ, denoted by $\boldsymbol{\xi}^{(i)}$, the numerical solution to (1.7) may be calculated by a fixed deterministic solver; for some of our examples we use the finite element solver package FEniCS [22]. For any fixed $\boldsymbol{x}_0, t_0$, our objective is to reconstruct the solution $u(\boldsymbol{x}_0, t_0, \Xi)$ using N realizations $\{u(\boldsymbol{x}_0, t_0, \boldsymbol{\xi}^{(i)})\}$. For brevity we suppress the dependence of $u(\boldsymbol{x}_0, t_0, \Xi)$ and $\{u(\boldsymbol{x}_0, t_0, \boldsymbol{\xi}^{(i)})\}$ on $\boldsymbol{x}_0$ and t_0, and simply write them as $u(\Xi)$ and $\{u(\boldsymbol{\xi}^{(i)})\}$, respectively.

We use the realized samples $\boldsymbol{\xi}^{(i)}$, $i = 1, \ldots, N$, of Ξ to evaluate the PCE basis and identify a corresponding solution $u(\boldsymbol{\xi}^{(i)})$ to (1.7). This evaluated PCE basis forms a row of $\boldsymbol{\Psi} \in \mathbb{R}^{N\times P}$, that is $\boldsymbol{\Psi}(i, j) = \psi_j(\boldsymbol{\xi}^{(i)})$. The corresponding solution $u(\boldsymbol{\xi}^{(i)})$ is the associated element of the vector $\boldsymbol{u}$. Given these notations, PCE in (1.4) with these realized samples can be written as

$$\boldsymbol{u} = \boldsymbol{\Psi}\boldsymbol{c}, \tag{1.8}$$

where vector $\boldsymbol{c} \in \mathbb{R}^P$ consist of PCE coefficients c_j, $j = 1, \dots, P$. We are then faced with identifying the vector of PCE coefficients $\boldsymbol{c}$ in (1.8).

1.4 Numerical Methods for PCE

To calculate c_j in (1.6) numerically, sampling methods including Monte Carlo simulation [23] and pseudo-spectral stochastic collocation [24, 25, 26, 27] may be applied. In the following sections, these methods are briefly reviewed.

1.4.1 The Monte Carlo Method

The Monte Carlo method is a straightforward numerical technique. In the Monte Carlo method, a sequence realizations, $\boldsymbol{\xi}^{(1)}$, $\boldsymbol{\xi}^{(2)}$, $\dots$, $\boldsymbol{\xi}^{(N)}$, is sampled according to the probability distribution of $\boldsymbol{\Xi}$. Deterministic simulations are performed for each sampled $\boldsymbol{\xi}^{(i)}$ to obtain the realizations of the QoI, $u(\boldsymbol{\xi}^{(i)})$. The empirical integrals (1.6) of u may be calculated to approximate c_j,

$$\hat{c}_j = \frac{1}{N}\sum_{i=1}^{N} u(\boldsymbol{\xi}^{(i)})\psi_j(\boldsymbol{\xi}^{(i)}). \tag{1.9}$$

As the number of samples increases, these empirical integrals converges to the exact values asymptotically, i.e.,

$$\lim_{N\to\infty} \hat{c}_j = c_j. \tag{1.10}$$

The most significant advantage of the Monte Carlo method is that it is naturally insensitive to the dimensionality of the random input space, and does not suffer from the *curse of dimensionality*. Meanwhile, as different realizations do not depend on each other, the Monte Carlo method is inherently parallelable. In addition, the Monte Carlo method is robust, due to their simplicity. However, the major drawback of the Monte Carlo method is its slow convergence rate. The rate of convergence of (1.10) is governed by the central limit theorem, and is of the order of $1/\sqrt{N}$ without an exception. The implication of this slow convergence is high computational cost, or relatively large approximation error. Due to this limitation, the Monte Carlo method can only be applied when the desired accuracy is not high. If an accurate approximation of the QoI is desirable, using the Monte Carlo method can be extremely computationally costly.

1.4.2 Stochastic Collocation

The idea of stochastic collocation is to approximate the integral (1.6) by quadrature [28, 29]. According to the probability distribution of random variables Ξ, stochastic collocation selects a multi-dimensional grid $\boldsymbol{\xi}^{(1)},\ \boldsymbol{\xi}^{(2)},\ \ldots,\ \boldsymbol{\xi}^{(Q)}$, and approximate c_j by

$$\hat{c}_j = \sum_{i=1}^{Q} u(\boldsymbol{\xi}^{(i)}) w^{(i)}, \tag{1.11}$$

where the weights $w^{(i)} > 0$ depend on the probability density function $\rho(\Xi)$.

Commonly used multi-dimensional grid types include tensor grids and Smolyak sparse grids are constructed as nested one-dimensional grids following rules such as trapezoidal and Clenshaw-Curtis rules. The one-dimensional quadrature grid (such as Gauss-Legendre and Gauss-Hermite grid) points and weights can be computed by solving an eigenvaule problem.

The main advantage of stochastic collocation is a fast convergence. The error in $\hat{c}_j$ decreases approximately exponentially as the quadrature level in each dimension (if a tensor product grid is used) or the total level (if a sparse grid is used) increases. Similar to the Monte Carlo method, stochastic collocation solves a deterministic problem at each grid point. However stochastic collocation suffers from the *curse of dimensionality.* When the dimensionality d and the order p are both low, stochastic collocation performs well, but it may become impractical for high-dimensional random inputs as the cost asymptotically grows exponentially as d or p increases.

1.4.3 Least-squares Regression

Least-squares regression [30, 31, 32] can be employed to identify the PCE coefficients $\boldsymbol{c}$ in (1.8), by solving the following minimization problem,

$$\hat{\boldsymbol{c}} = \arg\min_{\boldsymbol{c}} \|\boldsymbol{\Psi}\boldsymbol{c} - \boldsymbol{u}\|_2^2, \tag{1.12}$$

By setting the derivative of $\|\boldsymbol{\Psi}\boldsymbol{c} - \boldsymbol{u}\|_2^2$ with respect to $\boldsymbol{c}$ to be $\mathbf{0}$, it can be shown that if $\boldsymbol{\Psi}^T\boldsymbol{\Psi}$ is full-rank, the solution to (1.12) is

$$\hat{\boldsymbol{c}} = (\boldsymbol{\Psi}^T\boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}^T\boldsymbol{u}. \tag{1.13}$$

As $\boldsymbol{\Psi}^T\boldsymbol{\Psi}$ needs to be full-rank, it generally requires $N > P$ solution realizations as a necessary condition to achieve a stable approximation of $\boldsymbol{c}$. Therefore, for large d or p, it may be computationally expensive to solve (1.7) for N times, particularly when the deterministic solver has high complexity.

1.4.4 Compressive Sampling

Compressive sampling (also known as compressive sensing or compressed sensing) is a signal processing technique for efficiently acquiring and reconstructing a signal, by finding solutions to underdetermined linear systems.

Conventional sampling methods follow the Nyquist/Shannon sampling theory [33], i.e., the sampling rate must be at least twice the highest frequency of the signal. Similarly, the fundamental theorem of linear algebra suggests that the number of collected samples (measurements) of a discrete finite-dimensional signal should be at least as large as its length (its dimension) in order to ensure reconstruction. However, a new paradigm in approximation theory has emerged, the so-called compressive sampling [34, 35], which states that sparse signals can be recovered from what was previously believed to be highly incomplete measurements. It has been successfully applied to a growing number of applications across a wide spectrum of disciplines including (medical) imaging [36, 37, 38, 39], seismic data-acquisition

[40, 41, 42], remote sensing [43, 44], and high-dimensional data analysis [45, 46], among others.

In the context of UQ, a sparse signal is the QoI that leads to a sparse PCE representation [47].

Sparse PCE

As $u(\Xi)$ has finite variance, the c_j in (1.4) necessarily converge to zero, and if this convergence is sufficiently rapid, then $u(\Xi)$ may be accurately approximated by

$$u(\Xi) = \sum_{j \in \mathcal{C}} c_j \psi_j(\Xi) + \varepsilon, \tag{1.14}$$

where the index set $\mathcal{C} \subset \{1, \ldots, P\}$ has few elements, and the truncation error ε is small. When (1.14) occurs we say that $u(\Xi)$ admits an approximately sparse PCE. The length of the set $\mathcal{C}$ is defined as the sparsity.

As the sparsity of the vector $\boldsymbol{c}$ defined in Section 1.3 implies the sparsity of the QoI, the practical advantage of representing the QoI with a small number of basis functions motivate a search for an approximate $\boldsymbol{c}$ supported on $\mathcal{C}$, which has few non-zero entries [48, 49, 47, 50, 51, 52, 53]. We seek to identify the small subset $\mathcal{C}$ and the corresponding c_j, therefore to achieve an accurate reconstruction of $u(\Xi)$ with a small number of samples, and so look to techniques from the field of compressive sampling [54, 55, 34, 56, 57, 35, 58, 59, 60].

Compressive sampling seeks a solution $\boldsymbol{c}$ with minimum number of non-zero entries by solving the optimization problem

$$\mathcal{P}_{0,\epsilon} \equiv \{\arg\min_{\boldsymbol{c}} \|\boldsymbol{c}\|_0 : \|\boldsymbol{\Psi c} - \boldsymbol{u}\|_2 \leqslant \epsilon\}. \tag{1.15}$$

Here $\|\boldsymbol{c}\|_0$ is defined as the number of non-zero entries of $\boldsymbol{c}$, and a solution to $\mathcal{P}_{0,\epsilon}$ directly provides an optimally sparse approximation in that a minimal number of non-zero entries are used to recover $\boldsymbol{u}$ to within ϵ in the ℓ_2 norm. In general, the cost of finding a solution to $\mathcal{P}_{0,\epsilon}$ grows exponentially in P [60]. To resolve this exponential dependence, approaches such as greedy methods and ℓ_1-minimization are employed.

Orthogonal Matching Pursuit

Orthogonal matching pursuit (OMP) is one of the commonly used greedy algorithms, and we employ it to approximate the solution to $(\mathcal{P}_{0,\epsilon})$ [61, 62]. Starting from $\boldsymbol{c}^{(0)} = \mathbf{0}$ and an empty active column set of $\boldsymbol{\Psi}$, at any iteration t, OMP identifies only one column to be added to the active column set. The column is chosen such that the ℓ_2-norm of the residual, $\|\boldsymbol{\Psi}\boldsymbol{c}^{(t)} - \boldsymbol{u}\|_2$, is maximally reduced. Having specified the active column set, a least-squares problem is solved to compute the solution $\boldsymbol{c}^{(t)}$. The iterations are continued until the error truncation tolerance ϵ is achieved. The following exhibit depicts a step-by-step implementation of the OMP algorithm.

Algorithm 1 Orthogonal matching pursuit (OMP)

```
Set $t = 0$, $\boldsymbol{c}^{(0)} = \mathbf{0}$, and $\boldsymbol{r}^{(0)} = \boldsymbol{u} - \boldsymbol{\Psi}\boldsymbol{c}^{(0)}$.
Set the initial solution support index set $\mathcal{J}^{(0)} = \emptyset$.
while $\|\boldsymbol{u} - \boldsymbol{\Psi}\boldsymbol{c}^{(t)}\|_2 > \epsilon$: do
    for all $j \notin \mathcal{J}^{(t)}$: do
        Evaluate $\epsilon(j) = \|\boldsymbol{\psi}_j\boldsymbol{\alpha}_j - \boldsymbol{r}^{(t)}\|_2$, with $\boldsymbol{\alpha}_j = \boldsymbol{\psi}_j^T\boldsymbol{r}^{(t)}/\|\boldsymbol{\psi}_j\|_2^2$.
    end for
    Set $t = t + 1$.
    Update the support index set $\mathcal{J}^{(t)} = \mathcal{J}^{(t-1)} \cup \{\arg\min_j \epsilon(j)\}$.
    Solve for $\boldsymbol{c}^{(t)} = \arg\min_{\boldsymbol{c}} \|\boldsymbol{u} - \boldsymbol{\Psi}\boldsymbol{c}\|_2$ subject to $\mathit{Support}\{\boldsymbol{c}\} = \mathcal{J}^t$.
    Update the residual $\boldsymbol{r}^{(t)} = \boldsymbol{u} - \boldsymbol{\Psi}\boldsymbol{c}^{(t)}$.
end while
Output the solution $\boldsymbol{c} = \boldsymbol{c}^{(t)}$.
```

Theoretical analyses have been provided to show that OMP is guaranteed to recover the PCE coefficients in problem $\mathcal{P}_{0,0}$, which is done via the concept of *mutual coherence.* The mutual coherence of a matrix $\boldsymbol{\Psi} \in \mathbb{R}^{N\times P}$ is defined [63] by

$$\mu(\boldsymbol{\Psi}) := \max_{1\leq i,j\leq P, i\neq j} \frac{|\boldsymbol{\psi}_i^T\boldsymbol{\psi}_j|}{\|\boldsymbol{\psi}_i\|\|\boldsymbol{\psi}_j\|}. \tag{1.16}$$

where $\boldsymbol{\psi}_i$ and $\boldsymbol{\psi}_j$ are two columns of $\boldsymbol{\Psi}$. The mutual coherence is a measure of how orthogonal a matrix is. For instance, when $\mu(\boldsymbol{\Psi}) = 0$, the matrix $\boldsymbol{\Psi}$ is unitary, while $\mu(\boldsymbol{\Psi}) = 1$, at least two columns of $\boldsymbol{\Psi}$ are identical. For any given $\boldsymbol{\Psi}$, $0 \leq \mu(\boldsymbol{\Psi}) \leq 1$. Specially, for under-determined case, $N < P$, the mutual coherence $\mu(\boldsymbol{\Psi})$ is strictly positive.

Following Theorem 6 in [60], we know that for problem $(\mathcal{P}_{0,0})$, where $N < P$, if a solution $\boldsymbol{c}_0$ exists satisfying

$$\|\boldsymbol{c}_0\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(\boldsymbol{\Psi})}\right), \tag{1.17}$$

OMP is guaranteed to recover $\boldsymbol{c}_0$ exactly. Furthermore, for problems with high-dimensional random inputs, Corollary 7.4 in [62] and Theorem 3.1 in [64] show that a upper bound of $\mu(\boldsymbol{\Psi})$ exists, and that $\mu(\boldsymbol{\Psi})$ is within this upper bound with a high probability, i.e.,

$$Prob\left[\mu(\boldsymbol{\Psi}) \geq \delta\right] \leq 2^{3/4} P^2 \exp\left(-\frac{N\delta^2}{2C3^{2p}}\right), \tag{1.18}$$

where rows of $\boldsymbol{\Psi}$ are the p-order Legendre polynomial chaos basis, independently realized in $d > p$ i.i.d. uniform random variables $\boldsymbol{\Xi}$, and the constant $C \approx 13.12$.

From (1.18), we can see that if the number of realizations N is larger, or the bound δ is higher, the mutual coherence is more tightly bounded, since the probability of $\mu(\boldsymbol{\Psi}) \geq \delta$ drops exponentially. In addition, for a fixed set of d and p, as N increases, the upper bound of $\mu(\boldsymbol{\Psi})$ may decrease, which enables OMP to recover a solution with more non-zero elements.

ℓ_1 minimization

The convex relaxation of $\mathcal{P}_{0,\epsilon}$ based on ℓ_1-minimization, also referred to as basis pursuit denoising (BPDN), has been proposed to approximate the solution to $\mathcal{P}_{0,\epsilon}$ [54, 55, 56, 34, 60]. Specifically, ℓ_1-minimization seeks to identify $\boldsymbol{c}$ by solving

$$\mathcal{P}_{1,\epsilon} \equiv \{\arg\min_{\boldsymbol{c}} \|\boldsymbol{c}\|_1 : \|\boldsymbol{\Psi c} - \boldsymbol{u}\|_2 \leqslant \epsilon\} \tag{1.19}$$

using convex optimization algorithms [54, 65, 66, 67, 68, 69, 70, 71]. In practice, $\mathcal{P}_{0,\epsilon}$ and $\mathcal{P}_{1,\epsilon}$ may have similar solutions, and the comparison of the two problems has received significant study, see, e.g., [60] and the references therein.

Note in (1.19) the constraint $\|\boldsymbol{\Psi c}-\boldsymbol{u}\|_2 \leqslant \epsilon$ depends on the observed $\boldsymbol{\xi}$ and $u(\boldsymbol{\xi})$. As a result, $\boldsymbol{c}$ may be chosen to fit the input data, and not accurately approximate $u(\boldsymbol{\xi})$ for unobserved realizations $\boldsymbol{\xi}$. To avoid this situation, we determine ϵ by cross-validation [47].

To solve problem $(P_{1,\epsilon})$, the majority of available solvers for ℓ_1-minimization are based on alternative formulations of $(P_{1,\epsilon})$, such as the ℓ_1-norm regularized least-squares problem

$$(QP_\lambda) \equiv \left\{ \arg\min_{\boldsymbol{c}} \left(\frac{1}{2} \|\boldsymbol{\Psi c}-\boldsymbol{u}\|_2^2 + \lambda \|\boldsymbol{c}\|_1 \right) \right\}, \tag{1.20}$$

or the LASSO problem, [72],

$$(LS_\tau) \equiv \{\arg\min_{\boldsymbol{c}} \frac{1}{2} \|\boldsymbol{\Psi c}-\boldsymbol{u}\|_2^2 : \|\boldsymbol{c}\|_1 \leq \tau\}. \tag{1.21}$$

It can be shown that for an appropriate choice of scalars ϵ, λ, and τ, the problems $(P_{1,\epsilon})$, (QP_λ), and (LS_τ) share the same solution [73, 60, 74]. Numerous solvers based on the *active set* [65, 75], *interior-point continuation* [76, 69] and *projected gradient* [66, 67, 77, 78, 79, 73, 80, 81] methods have been developed for solving the above formulations of the ℓ_1-minimization problem.

In this dissertation, we adopt the Spectral Projected Gradient algorithm (SPGL1) proposed in [73] and implemented in the MATLAB package `SPGL1` [70] to solve the ℓ_1-minimization problem $(P_{1,\epsilon})$ in (1.19). SPGL1 is based on exploring the so-called Pareto curve, describing the tradeoff between the ℓ_2-norm of the truncation error $\|\boldsymbol{\Psi c}-\boldsymbol{u}\|_2$ and the ℓ_1-norm of the solution $\|\boldsymbol{c}\|_1$, for successive solution iterations. At each iteration, the LASSO problem (1.21) is solved using the spectral projected gradient technique with a worst-case complexity of $\mathcal{O}(P \ln P)$ where P is the number of columns in $\boldsymbol{\Psi}$. Given the error tolerance ϵ, a scalar equation is solved to identify a value for τ such that the (LS_τ) solution of (1.21) is identical to that of $(P_{1,\epsilon})$ in (1.19). Besides being efficient for large-scale systems where $\boldsymbol{\Psi}$ may not be available explicitly, the SPGL1 algorithm is specifically effective for our application of interest as the truncation error $\|\boldsymbol{\Psi c}-\boldsymbol{u}\|_2$ is known only approximately.

To show that ℓ_1-minimization indeed finds a solution to $\mathcal{P}_{0,\epsilon}$, we investigate the case whose $\epsilon = 0$. We let $\boldsymbol{c}_0$ be the solution to $\mathcal{P}_{0,0}$, whose sparsity is $s := \|\boldsymbol{c}_0\|_0$, and $\boldsymbol{c}_1$ be the solution to $\mathcal{P}_{1,0}$. It has been shown in [64, 82] that when

$$N \geq C(1+\beta)3^p s \log(P), \tag{1.22}$$

where C is non-negative constant,

$$Prob\,[\boldsymbol{c}_1 = \boldsymbol{c}_0] \geq 1 - \frac{6}{P} - 6e^{-\beta}. \tag{1.23}$$

From (1.22) and (1.23), we can deduce that as the number of realizations N becomes larger, a higher probability ℓ_1-minimization achieves in recovering the coefficients $\boldsymbol{c}_0$.

1.5 Organization of the Thesis

Taking advantage of the sparse PCE and compressive sampling, the purpose of Chapters 2-4 is to develop the numerical methods for UQ, such that the QoI can be approximated significantly more accurately or cost-efficiently. For the QoI's that do not yield to sparse PCEs with traditional polynomial basis functions associated with the measure of random inputs, such as a quantity including sharp gradients or discontinuity, in Chapter 5 we seek to design an optimal basis that enhance the sparsity of the corresponding PCE, within the Jacobi polynomial family. The structure of this thesis is organized as follows:

In Chapter 2, we modify the standard ℓ_1-minimization algorithm, using *a priori* information about the decay of the PCE coefficients, when available, and refer to the resulting algorithm as *weighted ℓ_1-minimization*. We provide conditions under which we may guarantee recovery using this weighted scheme. Numerical tests are used to compare the weighted and non-weighted methods for the recovery of solutions to two differential equations with high-dimensional random inputs: a boundary value problem with a random elliptic operator and a 2-D thermally driven cavity flow with random boundary condition. When such *a priori* information is not available, we propose an alternative approach in Chapter 3.

In Chapter 3, we investigate bi-fidelity approaches for PCE, in which computationally economical low-fidelity solution is utilized to improve the surrogate approximation of a QoI.

PCE coefficients computed from low-fidelity solution is used as *a priori* information about the high-fidelity PCE coefficients, resulting in an improved accuracy in recovering the QoI. This *a priori* information is involved via weighted ℓ_1-minimization and a modified orthogonal matching pursuit (OMP), which is proposed as *bi-fidelity* OMP. Numerical experiments are provided to compare the bi-fidelity and standard methods, and they all show that bi-fidelity approaches admits solution recovery at improved accuracy.

In addition to *a priori*-relative approaches, in Chapter 4 we studied gradient-enhanced UQ, in which the derivatives of a QoI with respect to the uncertain parameters are utilized to improve the surrogate approximation. In detail, we investigate a gradient-enhanced ℓ_1-minimization, where derivative information is computed to accelerate the identification of the PCE coefficients. For this approach, stability and convergence analysis are lacking, and thus we address these here with a probabilistic result. In particular, with an appropriate normalization, we show the inclusion of derivative information will almost-surely lead to improved conditions, e.g. related to the null-space and coherence of the measurement matrix, for a successful solution recovery. Further, we demonstrate our analysis empirically via three numerical examples: a manufactured PCE, an elliptic partial differential equation with random inputs, and a plane Poiseuille flow with random boundaries. These examples all suggest that including derivative information admits solution recovery at reduced computational cost.

In Chapters 2-4, the type of polynomial bases are chosen based the probability measure of random inputs and from the so-called Askey family of orthogonal polynomials. However, for an arbitrary QoI such an *a priori* choice of basis may result in slow decaying expansion coefficients, which in turn may lead to large errors when small order PCEs are considered. Increasing the order of the truncated expansion may enhance the solution accuracy, however, at the expense of additional computation cost which may become prohibitive for complex systems. Alternatively, in Chapter 5, a design strategy is proposed to choose an *optimal* PCE basis, within the family of Jacobi polynomials, and the corresponding change of measure using (random) realizations of QoI, in an *a posteriori* manner. To this end, an alternating least squares (ALS) regression is proposed to estimate the parameters of the Jacobi basis and the expansion coefficients. It is demonstrated that the proposed PCE basis design leads to more rapidly decaying coefficients, hence reduces truncation error and enhances solution accuracy, relative to the PCE basis naturally orthogonal with respect to the probability measure of inputs. Several numerical tests, with QoI's exhibiting sharp gradients/discontinuities, are provided to illustrate the performance of this approach.

In the final chapter, Chapter 6, we briefly summarize the work involved in this thesis and present a outlook on future work.

CHAPTER 2

A WEIGHTED ℓ_1-MINIMIZATION APPROACH FOR SPARSE POLYNOMIAL CHAOS EXPANSIONS[1]

Abstract

This chapter proposes a method for sparse polynomial chaos (PC) approximation of high-dimensional stochastic functions based on non-adapted random sampling. We modify the standard ℓ_1-minimization algorithm, originally proposed in the context of compressive sampling, using *a priori* information about the decay of the PC coefficients, when available, and refer to the resulting algorithm as *weighted ℓ_1-minimization.* We provide conditions under which we may guarantee recovery using this weighted scheme. Numerical tests are used to compare the weighted and non-weighted methods for the recovery of solutions to two differential equations with high-dimensional random inputs: a boundary value problem with a random elliptic operator and a 2-D thermally driven cavity flow with random boundary condition.

2.1 Introduction

As we analyze engineering systems of increasing complexity, we must strategically confront the imperfect knowledge of the underlying physical models and their inputs, as well as the implied imperfect knowledge of a quantity of interest (QOI) predicted from these models. The understanding of outputs as a function of inputs in the presence of such uncertainty falls

[1] This chapter has been published by *J. Peng et al.* in *Journal of Computational Physics, 267(0):92 111, 2014.*

within the field of uncertainty quantification. The accurate quantification of the uncertainty of the QOI allows for the rigorous mitigation of both unfounded confidence and unnecessary diffidence in the anticipated QOI.

Probability is a natural mathematical framework for describing uncertainty, and so we assume that the system input is described by a vector of independent random variables, Ξ. If the random variable QOI, denoted by $u(\Xi)$, has finite variance, then the polynomial chaos (PC) expansion [19, 20] is given in terms of the orthonormal polynomials $\{\psi_j(\Xi)\}$ as

$$u(\Xi) = \sum_{j=1}^{\infty} c_j \psi_j(\Xi). \tag{2.1}$$

A more detailed exposition on the use of PC expansion in this work is given in Section 2.2.2.

To identify the PC coefficients, c_j in (2.1), sampling methods including Monte Carlo simulation [23], pseudo-spectral stochastic collocation [24, 25, 26, 27], or least-squares regression [30] may be applied. These methods for evaluating the PC coefficients are popular in that deterministic solvers for the QOI may be used without being adapted to the probability space. However, the standard Monte Carlo approach suffers from a slow convergence rate. Additionally, a major limitation to the use of the last two approaches above, in their standard form, is that the number of samples needed to approximate c_j generally increases rapidly with the dimension of the input uncertainty, i.e., the number of random variables needed to describe the input uncertainty. In particular, for asymptotically large dimensions, such growth is exponential, see, e.g., [3, 4, 83, 31, 18].

In this work, we use the Monte Carlo sampling method while considerably improving the accuracy of approximated PC coefficients (for the same number of samples) by exploiting the approximate sparsity of the coefficients c_j. As u has finite variance, the c_j in (2.1) necessarily converge to zero, and if this convergence is sufficiently rapid, then $u(\Xi)$ may be approximated by

$$\hat{u}(\Xi) = \sum_{j\in\mathcal{C}} c_j \psi_j(\Xi), \tag{2.2}$$

where the index set $\mathcal{C}$ has few elements. When this occurs we say that $\hat{u}$ is reconstructed from a sparse PC expansion, and that u admits an approximately sparse PC representation. By truncating the PC basis implied by (2.1) to P elements, we may perform calculations on the truncated PC basis. If we let $\boldsymbol{c}$ be a vector of c_j, for $j = 1, \dots, P$, then the approximate sparsity of the QOI (implied by the sparsity of $\boldsymbol{c}$) and the practical advantage of representing the QOI with a small number of basis functions motivate a search for an approximate $\boldsymbol{c}$ which has few non-zero entries [48, 49, 47, 50, 51, 52, 53]. We seek to achieve an accurate reconstruction with a small number of samples, and so look to techniques from the field of compressive sampling [54, 55, 34, 56, 57, 35, 58, 59, 60].

Let $\boldsymbol{\xi}$ represent a realization of $\boldsymbol{\Xi}$. We define $\boldsymbol{\Psi}$ as the matrix where each row corresponds to the row vector of P PC basis functions evaluated at sampled $\boldsymbol{\xi}$ with the corresponding $u(\boldsymbol{\xi})$ being an entry in the vector $\boldsymbol{u}$. We assume $N < P$ samples of $\boldsymbol{\xi}$, so that $\boldsymbol{\Psi}$ is $N \times P$, $\boldsymbol{c}$ is $P \times 1$, and $\boldsymbol{u}$ is $N \times 1$. Compressive sampling seeks a solution $\boldsymbol{c}$ with minimum number of non-zero entries by solving the optimization problem

$$\mathcal{P}_{0,\epsilon} \equiv \{\arg\min_{\boldsymbol{c}} \|\boldsymbol{c}\|_0 : \|\boldsymbol{\Psi}\boldsymbol{c} - \boldsymbol{u}\|_2 \leqslant \epsilon\}. \tag{2.3}$$

Here $\|\boldsymbol{c}\|_0$ is defined as the number of non-zero entries of $\boldsymbol{c}$, and a solution to $\mathcal{P}_{0,\epsilon}$ directly provides an optimally sparse approximation in that a minimal number of non-zero entries are used to recover $\boldsymbol{u}$ to within ϵ in the ℓ_2 norm. In general, the cost of finding a solution to $\mathcal{P}_{0,\epsilon}$ grows exponentially in P [60]. To resolve this exponential dependence, the convex relaxation of $\mathcal{P}_{0,\epsilon}$ based on ℓ_1-minimization, also referred to as basis pursuit denoising (BPDN), has been proposed [54, 55, 56, 34, 60]. Specifically, BPDN seeks to identify $\boldsymbol{c}$ by solving

$$\mathcal{P}_{1,\epsilon} \equiv \{\arg\min_{\boldsymbol{c}} \|\boldsymbol{c}\|_1 : \|\boldsymbol{\Psi}\boldsymbol{c} - \boldsymbol{u}\|_2 \leqslant \epsilon\} \tag{2.4}$$

using convex optimization algorithms [54, 65, 66, 67, 68, 69, 70, 71]. In practice, $\mathcal{P}_{0,\epsilon}$ and $\mathcal{P}_{1,\epsilon}$ may have similar solutions, and the comparison of the two problems has received significant study, see, e.g., [60] and the references therein.

Note in (2.4) the constraint $\|\boldsymbol{\Psi c}-\boldsymbol{u}\|_2 \leqslant \epsilon$ depends on the observed $\boldsymbol{\xi}$ and $u(\boldsymbol{\xi})$; not in general Ξ and $u(\Xi)$. As a result, $\boldsymbol{c}$ may be chosen to fit the input data, and not accurately approximate $u(\Xi)$ for previously unobserved realizations $\boldsymbol{\xi}$. To avoid this situation, we determine ϵ by cross-validation [47] as discussed in Section 2.2.5.

To assist in identifying a solution to (2.4), note that for certain classes of functions, theoretical analysis suggests estimates on the decay for the magnitude of the PC coefficients [84, 85, 29]. Alternatively, as we shall see in Section 2.4.2, such estimates may be derived by taking into account certain relations among physical variables in a problem. It is reasonable to use this *a priori* information to improve the accuracy of sparse approximations [86]. Moreover, even if this decay information is unavailable, each approximated set of PC coefficients may be considered as an initialization for the calculation of an improved approximation, suggesting an iterative scheme [1, 86, 87, 88, 51, 53].

In this work, we explore the use of *a priori* knowledge of the PC coefficients as a weighting of ℓ_1 norm in BPDN in what is referred to as *weighted ℓ_1-minimization* (or *weighted BPDN*),

$$\mathcal{P}_{1,\epsilon}^{(\boldsymbol{W})} \equiv \{\arg\min_{\boldsymbol{c}} \|\boldsymbol{Wc}\|_1 : \|\boldsymbol{\Psi c}-\boldsymbol{u}\|_2 \leqslant \epsilon\}, \tag{2.5}$$

where $\boldsymbol{W}$ is a diagonal weight matrix to be specified. Previously, ℓ_1-minimization has been applied to solutions of stochastic partial differential equations with approximately sparse $\boldsymbol{c}$ [48, 47, 51, 53], but these approximately sparse $\boldsymbol{c}$ include a number of small magnitude entries which inhibit the accurate recovery of larger magnitude entries. The primary goal of this work is to utilize *a priori* information about $\boldsymbol{c}$, in the form of estimates on the decay of its entries, to reduce this inhibition and enhance the recovery of a larger proportion of PC coefficients; in particular those of the largest magnitude. We provide theoretical results pertaining to the quality of the solution identified from the weighted ℓ_1-minimization problem $\mathcal{P}_{1,\epsilon}^{(\boldsymbol{W})}$.

This work specifically focuses on weights derived from *a priori* information for the anticipated solution vector as in Section 2.3.1. This *a priori* information, while not always available, may be produced from analytical convergence analysis as examined in Section 2.4.1 or scaling arguments as considered in Section 2.4.2. Additionally, as samples from a low-fidelity or low-order model are computationally cheaper to attain, an approximate PC solution may be cheaply computed. While not considered in this study, such a solution may be used to identify weights for the associated high-fidelity or high-order model of interest. The applicability of the present approach is restricted to cases where such *a priori* knowledge for the anticipated PC solution is available or can be cheaply generated. In the absence of such information the standard ℓ_1-minimization problem $\mathcal{P}_{1,\epsilon}$ may be applied.

We also present a theoretical contribution in Section 2.3.2 relating recovery in the weighted ℓ_1 setting to that in the standard non-weighted setting utilizing results concerning the Restricted Isometry Constant (RIC) [89, 90, 91]. These analyses yield conditions under which recovery of the weighted ℓ_1 setting is assured.

The utilization of weighted ℓ_1-minimization for the solution of PDEs with random inputs has been recently studied in [53]. However, as we shall describe in details in Section 2.3.1, our construction of the weights, a key step in this framework, is fundamentally different from that of [53]. Additionally, the numerical experiments of Section 2.4 suggest higher accuracies may be achieved by the approach presented in this work.

The rest of this paper is structured as follows. In Section 2.2, we introduce the problem of interest as well as our approach for the stochastic expansion of its solution. Following that, in Section 2.3, we present our results on weighted ℓ_1-minimization and its corresponding analysis for sparse PC expansions. In Section 2.4, we provide two test cases which we use to describe the specification of the weighted ℓ_1-minimization problem, and explore its performance and accuracy. In particular, in Section 2.4.2, we utilize a simple dimensional relation to derive estimates of PC expansion coefficients of the velocity field in a flow problem, using which we set the weights in $\mathcal{P}_{1,\epsilon}^{(\boldsymbol{W})}$.

2.2 Problem Statement and Solution Approach

2.2.1 PDE formulation

Let the random vector Ξ, defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, characterize the input uncertainties and consider the solution of a partial differential equation defined on a bounded Lipschitz continuous domain $\mathcal{D} \subset \mathbb{R}^D$, $D \in \{1,\ 2,\ 3\}$, with boundary $\partial\mathcal{D}$. The uncertainty implied by Ξ may be represented in one or many relevant parameters, e.g., the diffusion coefficient, boundary conditions, and/or initial conditions. Letting $\mathcal{L}, \mathcal{I}$, and $\mathcal{B}$ depend on the physics of the problem being solved, the solution u satisfies the three constraints

$$\begin{aligned} &\mathcal{L}(\boldsymbol{x}, t, \Xi; u(t, \boldsymbol{x}, \Xi)) = 0, && \boldsymbol{x} \in \mathcal{D}, \\ &\mathcal{I}(\boldsymbol{x}, \Xi; u(0, \boldsymbol{x}, \Xi)) = 0, && \boldsymbol{x} \in \mathcal{D}, \\ &\mathcal{B}(\boldsymbol{x}, t, \Xi; u(t, \boldsymbol{x}, \Xi)) = 0, && \boldsymbol{x} \in \partial\mathcal{D}. \end{aligned} \tag{2.6}$$

We assume that $(\Omega, \mathcal{F}, \mathbb{P})$ is formed by the product of d probability spaces, $(\mathbb{R}, \mathbb{B}(\mathbb{R}), \mathbb{P}_k)$ corresponding to each coordinate of Ξ, denoted by Ξ_k; here $\mathbb{B}(\cdot)$ represents the Borel σ-algebra. We further assume that the random variable Ξ_k is continuous and distributed according to the density ρ_k implied by $\mathbb{P}_k$. Note that this entails $\Omega = \mathbb{R}^d$, $\mathcal{F} = \mathbb{B}(\mathbb{R}^d)$, that each Ξ_k is independently distributed, and that the joint distribution for Ξ, denoted by ρ, equals the tensor product of the marginal distributions $\{\rho_k\}$.

In this work, we assume that conditioned on the ith random realization of Ξ, denoted by $\boldsymbol{\xi}^{(i)}$, the numerical solution to (2.6) may be calculated by a fixed solver; for our examples we use the finite element solver package FEniCS [22]. For any fixed $\boldsymbol{x}_0, t_0$, our objective is to reconstruct the solution $u(\boldsymbol{x}_0, t_0, \Xi)$ using N realizations $\{u(\boldsymbol{x}_0, t_0, \boldsymbol{\xi}^{(i)})\}$. For brevity we suppress the dependence of $u(\boldsymbol{x}_0, t_0, \Xi)$ and $\{u(\boldsymbol{x}_0, t_0, \boldsymbol{\xi}^{(i)})\}$ on $\boldsymbol{x}_0$ and t_0.

The two specific physical problems we consider are a boundary value problem with a random elliptic operator and a 2-D heat driven cavity flow with a random boundary condition.

2.2.2 Polynomial Chaos (PC) expansion

Our methods to approximate the solution u to (2.6) make use of the PC basis functions which are induced by the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ on which Ξ is defined. Specifically, for each ρ_k we define $\{\psi_{k,j}\}_{j\geq 0}$ to be the complete set of orthonormal polynomials of degree j with respect to the weight function ρ_k [21, 20]. As a result, the orthonormal polynomials for Ξ are given by the products of the univariate orthonormal polynomials,

$$\psi_{\boldsymbol{\alpha}}(\Xi) = \prod_{k=1}^{d} \psi_{k,\alpha_k}(\Xi_k), \tag{2.7}$$

where each α_k, representing the kth coordinate of the multi-index $\boldsymbol{\alpha}$, is a non-negative integer. For computation, we truncate the expansion in (2.1) to the set of P basis functions associated with the subspace of polynomials of total order not greater than q, that is $\sum_{k=1}^{d} \alpha_k \leq q$. For convenience, we also order these P basis functions so that they are indexed by $\{1, \cdots, P\}$ as opposed to the vectorized indexing in (1.1). The basis set $\{\psi_j\}_{j=1}^{P}$ has the cardinality

$$P = \frac{(d+q)!}{d!q!}. \tag{2.8}$$

For the interest of presentation, we interchangeably use both notations for representing PC basis. For any fixed $\boldsymbol{x}_0, t_0$, the PC expansion of u and its truncation are then defined by

$$u(\boldsymbol{x}_0, t_0, \Xi) = u(\Xi) = \sum_{j=1}^{\infty} c_j \psi_j(\Xi) \approx \sum_{j=1}^{P} c_j \psi_j(\Xi). \tag{2.9}$$

Though u is an arbitrary function in $L_2(\Omega, \mathbb{P})$, we are limited to an approximation in the span of our basis polynomials, and the error incurred from this approximation is referred as *truncation error.*

In this work we assume that, for each k, ρ_k is known *a priori.* Two commonly used probability densities for ρ_k are uniform and Gaussian; the corresponding polynomial bases are, respectively, Legendre and Hermite polynomials [20]. We furthermore set Ξ_k to be uniformly distributed on $[-1, 1]$ and our PC basis functions are constructed from the orthonormal Legendre polynomials. The presented methods, however, may be applied to any set of orthonormal polynomials and their associated random variables.

We use the samples $\boldsymbol{\xi}^{(i)}$, $i = 1, \ldots, N$, of Ξ to evaluate the PC basis and identify a corresponding solution $u(\boldsymbol{\xi}^{(i)})$ to (2.6). This evaluated PC basis forms a row of $\boldsymbol{\Psi} \in \mathbb{R}^{N\times P}$ in (2.4), that is $\boldsymbol{\Psi}(i,j) = \psi_j(\boldsymbol{\xi}^{(i)})$. The corresponding solution $u(\boldsymbol{\xi}^{(i)})$ is the associated element of the vector $\boldsymbol{u}$. We are then faced with identifying the vector of PC coefficients $\boldsymbol{c} \in \mathbb{R}^P$ in (2.9), which we address by considering techniques from compressive sampling.

2.2.3 Sparse PC expansion

As the PC expansion in (2.9) is a sum of orthonormal random variables defined by $\psi_j(\Xi)$, the exact PC coefficients may be computed by projecting $u(\Xi)$ onto the basis functions $\psi_j(\Xi)$ such that

$$c_j = \mathbb{E}\left[u(\Xi)\psi_j(\Xi)\right] = \int_\Omega u(\boldsymbol{\xi})\psi_j(\boldsymbol{\xi})\rho(\boldsymbol{\xi})d\boldsymbol{\xi}.$$

To compute the PC coefficients non-intrusively, besides the standard Monte Carlo sampling, which is known to converge slowly, we may estimate this expectation via, for instance, sparse grid quadrature. While this latter approach performs well when d and q are small, it may become impractical for high-dimensional random inputs. Alternatively, $\boldsymbol{c}$ may be computed from a discrete projection, e.g., least-squares regression [30], which generally requires $N > P$ solution realizations to achieve a stable approximation.

We assume that $\boldsymbol{c}$ is approximately sparse, and seek to identify an appropriate $\mathcal{C}$, as in (2.2), having a small number of elements and giving a small truncation error. To this end we extend ideas from the field of compressive sampling. If the number of elements of $\mathcal{C}$, denoted by $|\mathcal{C}|$, is small, then using only the columns in $\boldsymbol{\Psi}$ corresponding to elements of $\mathcal{C}$ reduces the dimension of the PC basis from P to $|\mathcal{C}|$. This significantly reduces the number of PC coefficients requiring estimation and consequently the number of solution realizations N. We define $\boldsymbol{\Psi}_{\mathcal{C}}$ as the truncation of $\boldsymbol{\Psi}$ to those columns only relevant to the basis functions of $\mathcal{C}$, and similarly define $\boldsymbol{c}_\mathcal{C}$ as the truncation of $\boldsymbol{c}$. If $|\mathcal{C}| < N$, then the determination of $|\mathcal{C}|$ coefficients gives an optimization problem less prone to overfit the data [92], even when $N < P$. For example, the least-squares approximation of $\boldsymbol{c}_\mathcal{C}$, $\hat{\boldsymbol{c}}_\mathcal{C} = (\boldsymbol{\Psi}_\mathcal{C}^T\boldsymbol{\Psi}_\mathcal{C})^{-1}\boldsymbol{\Psi}_\mathcal{C}^T\boldsymbol{u}$, minimizing $\|\boldsymbol{\Psi}_\mathcal{C}\boldsymbol{c}_\mathcal{C} - \boldsymbol{u}\|_2$ is well-posed and will have a unique solution if $\boldsymbol{\Psi}_\mathcal{C}$ is of full rank.

Note that the identification of $\mathcal{C}$ is critical to the optimization problem $\mathcal{P}_{0,\epsilon}$ in (2.3). If we instead have a solution to $\mathcal{P}_{1,\epsilon}$, then we may infer a $\mathcal{C}$ by noting the entries of the approximated $\boldsymbol{c}$ which have magnitudes above a certain threshold. Motivated to obtain more accurate sparse solutions, we next introduce a compressive sampling technique which modifies $\mathcal{P}_{1,\epsilon}$ by weighting each c_j differently in $\|\boldsymbol{c}\|_1$. As we shall discuss later, these weights are generated based on some *a priori* information on the decay of c_j, when available.

2.2.4 Weighted ℓ_1-minimization

To develop a weighted ℓ_1-minimization $\mathcal{P}_{1,\epsilon}^{(\boldsymbol{W})}$ in (2.5), we do not consider any changes to the algorithm solving $\mathcal{P}_{1,\epsilon}$, but instead transform the problem with the use of weights, such that the same solver may be used. We define the diagonal weight matrix $\boldsymbol{W}$, with diagonal entries $w_j \geq 0$, and consider the new weighted problem $\mathcal{P}_{1,\epsilon}^{(\boldsymbol{W})}$ with

$$\|\boldsymbol{W}\boldsymbol{c}\|_1 = \sum_{j=1}^{P} w_j|c_j|. \tag{2.10}$$

If *a priori* information is available for c_j, it is natural to use it to define $\boldsymbol{W}$ [86]. Heuristically, columns with large anticipated $|c_j|$ should not be heavily penalized when used in the approximation, that is the corresponding w_j should be small. In contrast, $|c_j|$ which are not expected to be large should be paired with large w_j. This suggests allowing w_j to be inversely related to $|c_j|$, [1],

$$w_j = \begin{cases} |c_j|^{-p}, & c_j \neq 0, \\ \infty, & c_j = 0. \end{cases} \tag{2.11}$$

The parameter $p \in [0, 1]$ may be used to account for the confidence in the anticipated $|c_j|$. Large values of p lead to more widely dispersed weights and indicate greater confidence in these $|c_j|$ while small values lead to more clustered weights and indicate less confidence in these $|c_j|$. These weights deform the ℓ_1 ball, as Figure 2.1 shows, to discourage small coefficients from the solution and consequently enhance the accuracy. A detailed discussion of weighted ℓ_1-minimization and examples in signal processing are given in [1].

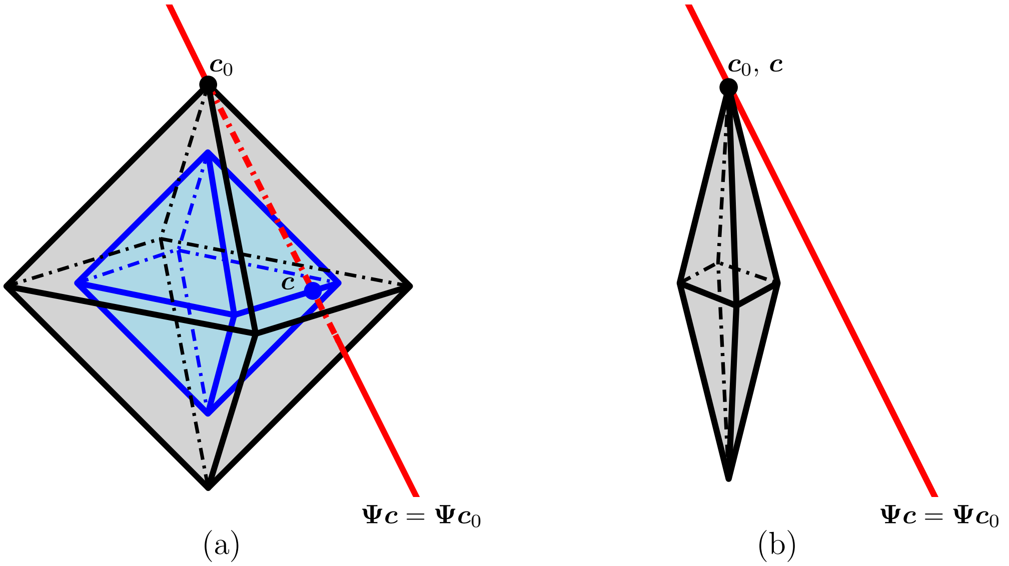

$\boldsymbol{c}_0$
$\boldsymbol{c}$
$\boldsymbol{\Psi c} = \boldsymbol{\Psi c}_0$
(a)
$\boldsymbol{c}_0$, $\boldsymbol{c}$
$\boldsymbol{\Psi c} = \boldsymbol{\Psi c}_0$
(b)

Figure 2.1: Schematic of approximation of a sparse $\boldsymbol{c}_0 \in \mathbb{R}^3$ via standard and weighted ℓ_1-minimization (based on [1]). (a) Standard ℓ_1-minimization where, depending on $\boldsymbol{\Psi}$, the problem $\mathcal{P}_{1,0}$ with $\boldsymbol{u} = \boldsymbol{\Psi c}_0$ may have a solution $\boldsymbol{c}$ such that $\|\boldsymbol{c}\|_1 \leq \|\boldsymbol{c}_0\|_1$. (b) Weighted ℓ_1-minimization for which there is no $\boldsymbol{c}$ with $\|\boldsymbol{Wc}\|_1 \leq \|\boldsymbol{Wc}_0\|_1$.

As in [87, 1], to insure stability, we consider a damped version of w_j in (2.11),

$$w_j = (|c_j| + \epsilon_w)^{-p}, \tag{2.12}$$

where ϵ_w is a relatively small positive parameter. In the numerical examples of this paper, we set $\epsilon_w = 5 \times 10^{-5} \cdot \hat{c}_1$ to generate w_j in $\mathcal{P}^{(\boldsymbol{W})}_{1,\epsilon}$, where $\hat{c}_1 = \frac{1}{N}\sum_{i=1}^{N} u(\boldsymbol{\xi}^{(i)})$ is the Monte Carlo estimate of the degree zero PC coefficient (or, equivalently, the sample average of u).

Remark 2.2.1 (Choice of p in (2.12))**.** *When defined based on the exact values $|c_j|$, the weights w_j in (2.12) together with (2.10) imply an ℓ_r-minimization problem of the form $\mathcal{P}_{r,\epsilon} \equiv \{\arg\min_{\boldsymbol{c}} \|\boldsymbol{c}\|_r : \|\boldsymbol{\Psi c} - \boldsymbol{u}\|_2 \leqslant \epsilon\}$ to solve for $\boldsymbol{c}$, where $r = 1 - p \in [0, 1]$. Depending on the value of r, such a minimization problem may outperform the standard ℓ_1-minimization,* *see, e.g., [87]. In practice, however, an optimal selection of r (or p) is not a trivial task and necessitates further analysis. In the present study, similar to [1], we choose $p = 1$.*

2.2.5 Choosing ϵ via cross validation

The choice of $\epsilon > 0$ for the optimization problems $\mathcal{P}_{1,\epsilon}$ or $\mathcal{P}_{1,\epsilon}^{(\boldsymbol{W})}$ is critical. If ϵ is too small, then $\boldsymbol{c}$ will overfit the data and give unfounded confidence in $u(\Xi)$; if ϵ is too large, then $\boldsymbol{c}$ will underfit the data and give unnecessary diffidence in $u(\Xi)$. In this work, following [47], the selection of ϵ is determined by cross-validation; here we divide the available data into two sets, a reconstruction set of N_r samples used to calculate $\boldsymbol{c}_r$, and a validation set of N_v samples to test this approximation. For the reconstruction set we let $\boldsymbol{c}_r(\epsilon_r)$ denote the calculated solution to $\mathcal{P}_{1,\epsilon}$ or $\mathcal{P}_{1,\epsilon}^{(\boldsymbol{W})}$ as a function of ϵ_r, and in this manner identify an optimal ϵ which is then corrected based on N_r and N_v. This algorithm is summarized below where the subscript indicates which data set is used in calculating the quantity: r for the reconstruction set; v for the validation set. We note that the optimal ϵ is dependent on the algorithm used to calculate $\boldsymbol{c}_r$ as well as the data input into that algorithm. In this paper we set $N_r = \lfloor \frac{4}{5} N \rfloor$ and $N_v = N - N_r$.

Algorithm 2 Algorithm for choosing ϵ using cross-validation.

```
Randomly divide the $N$ samples of $\Xi, u(\Xi)$ into two sets, a reconstruction set with $N_r$ samples and a validation set with $N_v$ samples.
Let $\epsilon^* = \arg\min_{\epsilon_r > 0} \|\boldsymbol{\Psi}_v \boldsymbol{c}_r(\epsilon_r) - \boldsymbol{u}_v\|_2$.
Return $\epsilon = \sqrt{\frac{N}{N_r}}\epsilon^*$.
```

2.3 Setting Weights w_j and Recovery Guarantees

We next introduce our approach for setting the weights w_j in (2.5) and present theoretical guarantees on computing the PC coefficients $\boldsymbol{c}$ via weighted ℓ_1-minimization problem $\mathcal{P}_{1,\epsilon}^{(\boldsymbol{W})}$.

2.3.1 Setting weights w_j

As the true $\boldsymbol{c}$ is unknown, an approximation of $\boldsymbol{c}$ must be employed to form the weights. In [87, 1, 93, 53] an iterative approach is proposed wherein these weights are computed from the previous approximation of $\boldsymbol{c}$. More precisely, at iteration $l+1$, the weights are set by

$$w_j = \left(|\hat{c}_j^{(l)}| + \epsilon_w \right)^{-1},$$

where $\hat{c}_j^{(l)}$ is the estimate of c_j obtained from $\mathcal{P}_{1,\epsilon}^{(\boldsymbol{W})}$ at iteration l and $w_j = 1$ at iteration $l = 1$. However, the solution to such iteratively re-weighted ℓ_1-minimization problems may be expensive due to the need for multiple $\mathcal{P}_{1,\epsilon}^{(\boldsymbol{W})}$ solves. Additionally, the convergence of the iterates is not always guaranteed [1]. Moreover, as we will observe from the results of Section 2.4, unlike the weighted ℓ_1-minimization, the accuracies obtained from the iteratively re-weighted ℓ_1-minimization approach are sensitive to the choice of ϵ_w. In particular, for relatively large or small values of ϵ_w, the iteratively re-weighted ℓ_1-minimization may even lead to less accurate results as compared to the standard ℓ_1-minimization.

Alternatively, to set w_j, we here focus our attention on situations when *a priori* knowledge on c_j in the form of bounds on $|c_j|$ or approximate $|c_j|$ is available. This includes primarily a class of linear elliptic PDEs with random inputs [84, 85, 29]. While not considered here, similar decay rates are also available for semi-linear elliptic, [94], and parabolic, [95, 96], PDEs with random inputs. We also provide preliminary results on a non-linear problem, specifically a 2-D Navier-Stokes equation, for which we exploit a physical dependency among solution variables along with a simple scaling argument to generate an approximate $|c_j|$. We notice that the success of our weighted ℓ_1-minimization depends on the availability of approximate $|c_j|$ and its ability to reveal *relative importance* of $|c_j|$ rather than their precise values. As we shall empirically illustrate in Section 2.4, when such information on c_j is used, the weighted ℓ_1-minimization approach outperforms the iteratively re-weighted ℓ_1-minimization.

To solve $\mathcal{P}_{1,\epsilon}^{(\boldsymbol{W})}$, the standard ℓ_1-minimization solvers may be used. In this work we use the MATLAB package SPGL1 [70] based on the spectral projected gradient algorithm [73]. Specifically, $\tilde{\boldsymbol{c}} = \boldsymbol{W}\boldsymbol{c}$ may be solved from $\mathcal{P}_{1,\epsilon}$ with the modified measurement matrix $\tilde{\boldsymbol{\Psi}} = \boldsymbol{\Psi}\boldsymbol{W}^{-1}$. We then set $\boldsymbol{c} = \boldsymbol{W}^{-1}\tilde{\boldsymbol{c}}$.

We defer presenting examples of setting w_j to Section 2.4 and instead provide theoretical analysis on the quality of the solution to the weighted ℓ_1-minimization problem $\mathcal{P}_{1,\epsilon}^{(\boldsymbol{W})}$. In particular, we limit our theoretical analysis to determining if $\mathcal{P}_{1,\epsilon}^{(\boldsymbol{W})}$ is equivalent to solving $\mathcal{P}_{0,\epsilon}$, finding an optimally sparse solution $\boldsymbol{c}$.

2.3.2 Theoretical recovery via weighted ℓ_1-minimization

Following the ideas of [90, 97, 98, 99, 34, 100], we consider analysis which depends on vectors in the kernel of $\boldsymbol{\Psi}$. We consider $\boldsymbol{c}_0$ to be a sparse approximation, such that $\boldsymbol{\Psi}\boldsymbol{c}_0 + \boldsymbol{e} = \boldsymbol{u}$ where $\|\boldsymbol{e}\|_2 \leq \epsilon$ indicates a small level of truncation error and/or noise is present, implying that exact reconstructions are themselves approximated by a sparse solution. Stated another way, $\boldsymbol{c}_0$ is a solution to $\mathcal{P}_{0,\epsilon}$. Let $\boldsymbol{c}_1$ be a solution to $\mathcal{P}_{1,\epsilon}^{(\boldsymbol{W})}$. Further, let $\mathcal{C} = \mathrm{Supp}(\boldsymbol{c}_0)$, and note that $s = |\mathcal{C}|$ is the sparsity of $\boldsymbol{c}_0$.

We are interested specifically in determining when $\boldsymbol{c}_1$ accurately approximates $\boldsymbol{c}_0$. Theorem 2.3.1 below provides a condition for recovery in terms of the Restricted Isometry Constant (RIC) – to be defined in (2.13) – when truncation error is present. Theorem 2.3.2 provides a result guaranteeing recovery in the absence of truncation error when a parameter is below a threshold. Related to this, Theorem 2.3.3 allows a bound on this parameter and leads to Corollary 2.3.1, which allows us to guarantee recovery with high probability when a sufficient number of samples are drawn.

The statement and proof of Theorem 2.3.1 are closely related to Theorem 1 of [90], providing a condition to compare a solution to $\mathcal{P}_{1,\epsilon}^{(\boldsymbol{W})}$ with a solution to $\mathcal{P}_{0,\epsilon}$. This is done in terms of the RIC δ_s, [89, 90], defined such that for any vector, $\boldsymbol{x} \in \mathbb{R}^P$, supported on at most s entries,

$$(1-\delta_s)\|\boldsymbol{x}\|_2^2 \leq \frac{1}{N}\|\boldsymbol{\Psi x}\|_2^2 \leq (1+\delta_s)\|\boldsymbol{x}\|_2^2. \tag{2.13}$$

While we follow Theorem 1 of [90] due to the simplicity of its proof, we note that improved conditions on the RIC have been presented in more recent studies [101, 102] , and we invite the reader to consult [90] for more motivation of the proof. Our contribution is the modest adaptation from non-weighted ℓ_1-minimization to weighted ℓ_1-minimization.

Theorem 2.3.1. *Let s be such that $\delta_{3s} + 3\delta_{4s} < 2$. Then for any approximate solution, $\boldsymbol{c}_0$, supported on $\mathcal{C}$ with $|\mathcal{C}| \leq s$, any solution $\boldsymbol{c}_1$ to $\mathcal{P}_{1,\epsilon}^{(\boldsymbol{W})}$ obeys*

$$\|\boldsymbol{c}_0 - \boldsymbol{c}_1\|_2 \leq C \cdot \epsilon,$$

where the constant C depends on s, $\max_{j\in\mathcal{C}} w_j$, and $\min_{j\in\mathcal{C}^c} w_j$. Here we utilize c as a superscript to denote the set complement.

Proof. Our proof is essentially an extension of the proof of Theorem 1 in [90] to account for the weighted ℓ_1 norm. Let $\boldsymbol{h} := \boldsymbol{c}_1 - \boldsymbol{c}_0$. Note that as $\boldsymbol{c}_1 = \boldsymbol{c}_0 + \boldsymbol{h}$ solves the weighted ℓ_1-minimization problem $\mathcal{P}_{1,\epsilon}^{(\boldsymbol{W})}$,

$$\|\boldsymbol{Wc}_0\|_1 - \|\boldsymbol{Wh}\|_{\mathcal{C},1} + \|\boldsymbol{Wh}\|_{\mathcal{C}^c,1} \leq \|\boldsymbol{W}\left(\boldsymbol{c}_0 + \boldsymbol{h}\right)\|_1 = \|\boldsymbol{Wc}_1\|_1 \leq \|\boldsymbol{Wc}_0\|_1,$$

where we use notation for an ℓ_r norm restricted to coordinates in a set $\mathcal{S}$ as $\|\boldsymbol{x}\|_{\mathcal{S},r}$. It follows that for some $0 \leq \beta \leq 1$,

$$\|\boldsymbol{Wh}\|_{\mathcal{C}^c,1} \leq \beta\|\boldsymbol{Wh}\|_{\mathcal{C},1}. \tag{2.14}$$

Sort the entries of $\boldsymbol{h}$ supported on $\mathcal{C}^c$ in descending order of their magnitudes, divide $\mathcal{C}^c$ into subsets of size M, and enumerate these sets as $\mathcal{C}_1, \cdots, \mathcal{C}_n$, where $\mathcal{C}_1$ corresponds to the indices of the M largest entries of sorted $\boldsymbol{h}$, $\mathcal{C}_2$ corresponds to the indices of the next M largest entries of sorted $\boldsymbol{h}$, and so on. Let $\mathcal{S} = \mathcal{C} \cup \mathcal{C}_1$, and note that the kth largest (in magnitude) entry of any $\boldsymbol{x}$ accounts for less than $1/k$ of the $\|\boldsymbol{x}\|_1$, so that

$$\|\boldsymbol{h}\|^2_{\mathcal{S}^c,2} = \sum_{k\in\mathcal{S}^c} h_k^2 \leq \|\boldsymbol{h}\|^2_{\mathcal{C}^c,1} \sum_{k=M+1}^{P} k^{-2} \leq \|\boldsymbol{h}\|^2_{\mathcal{C}^c,1} \cdot \frac{1}{M}.$$

We now bound the unweighted ℓ_1 norm from above by the weighted ℓ_1 norm, to achieve

$$\|\boldsymbol{h}\|^2_{\mathcal{C}^c,1} \cdot \frac{1}{M} \leq \|\boldsymbol{W}\boldsymbol{h}\|^2_{\mathcal{C}^c,1} \cdot \frac{1}{M \min_{i\in\mathcal{C}^c} w_i^2}.$$

From the condition (2.14),

$$\|\boldsymbol{W}\boldsymbol{h}\|^2_{\mathcal{C}^c,1} \cdot \frac{1}{M \min_{i\in\mathcal{C}^c} w_i^2} \leq \|\boldsymbol{W}\boldsymbol{h}\|^2_{\mathcal{C},1} \cdot \frac{\beta^2}{M \min_{i\in\mathcal{C}^c} w_i^2}.$$

Bounding the weighted ℓ_1 norm from above by the unweighted ℓ_1 norm gives,

$$\|\boldsymbol{W}\boldsymbol{h}\|^2_{\mathcal{C},1} \cdot \frac{\beta^2}{M \min_{i\in\mathcal{C}^c} w_i^2} \leq \|\boldsymbol{h}\|^2_{\mathcal{C},1} \cdot \frac{\beta^2 \max_{j\in\mathcal{C}} w_j^2}{M \min_{i\in\mathcal{C}^c} w_i^2}.$$

Bounding this by the ℓ_2 norm yields the desired inequality,

$$\|\boldsymbol{h}\|^2_{\mathcal{S}^c,2} \leq \|\boldsymbol{h}\|^2_{\mathcal{C},2} \cdot \frac{\beta^2 |\mathcal{S}| \max_{j\in\mathcal{C}} w_j^2}{M \min_{i\in\mathcal{C}^c} w_i^2}.$$

Let

$$\eta := \frac{\beta^2 |\mathcal{S}| \max_{j\in\mathcal{C}} w_j^2}{M \min_{i\in\mathcal{C}^c} w_i^2}.$$

It follows that

$$\|\boldsymbol{h}\|_2^2 = \|\boldsymbol{h}\|^2_{\mathcal{S}^c,2} + \|\boldsymbol{h}\|^2_{\mathcal{S},2} \leq (1+\eta)\|\boldsymbol{h}\|^2_{\mathcal{S},2}.$$

Following the proof from Theorem 1 of [90] we have that

$$\|\boldsymbol{\Psi}\boldsymbol{h}\|_2 \geq \left(\sqrt{1-\delta_{M+|\mathcal{C}|}} - \frac{|\mathcal{C}|}{M}\sqrt{1+\delta_M}\right) \|\boldsymbol{h}\|_{\mathcal{S},2},$$

and it follows that

$$\begin{aligned}\|\boldsymbol{h}\|_2 \leq \sqrt{1+\eta}\|\boldsymbol{h}\|_{\mathcal{S},2} &\leq \frac{\sqrt{1+\eta}}{\sqrt{1-\delta_{M+|\mathcal{C}|}} - \frac{|\mathcal{C}|}{M}\sqrt{1+\delta_M}} \|\boldsymbol{\Psi}\boldsymbol{h}\|_2, \\ &\leq \frac{2\sqrt{1+\eta}}{\sqrt{1-\delta_{M+|\mathcal{C}|}} - \frac{|\mathcal{C}|}{M}\sqrt{1+\delta_M}} \cdot \epsilon,\end{aligned}$$

which yields the proof with the remaining arguments from Theorem 1 of [90]. □

In the case of recovery with no truncation error, that is $\epsilon = 0$, we expand on the consideration of the parameter β in the above proof. We note that results for the case of $\epsilon = 0$ may not guarantee that a sparsest solution to $\mathcal{P}_{0,\epsilon}$ has been found, but may help to verify that as sparse as possible a solution to $\boldsymbol{u}_1 = \boldsymbol{\Psi}\boldsymbol{c}_1$ has been found. Stated another way, the computed solution that recovers $\boldsymbol{u}_1$ may have verifiable sparsity, where $\boldsymbol{u}_1$ is close to $\boldsymbol{u}$.

We show how $\boldsymbol{W}$ and $\mathcal{C}$ affect the recovery when $\epsilon = 0$ through the null-space of $\boldsymbol{\Psi}$. Specifically, recall that the difference between any two solutions to $\boldsymbol{\Psi}\boldsymbol{c} = \boldsymbol{u}$ is a vector in the null-space of $\boldsymbol{\Psi}$, denoted by $\mathcal{N}(\boldsymbol{\Psi})$. It follows that

$$\beta_{\boldsymbol{W}} = \max_{\boldsymbol{c}\in\mathcal{N}(\boldsymbol{\Psi})} \frac{\|\boldsymbol{W}\boldsymbol{c}\|_{\mathcal{C},1}}{\|\boldsymbol{W}\boldsymbol{c}\|_{\mathcal{C}^c,1}}, \tag{2.15}$$

is a bound on β in (2.14) for the case that $\epsilon = 0$.

When $\beta_{\boldsymbol{W}}$ is small we notice that adding to the sparse solution, $\boldsymbol{c}_0$, any vector $\boldsymbol{c} \in \mathcal{N}(\boldsymbol{\Psi})$ will induce a relatively small change in $\|\boldsymbol{W}(\boldsymbol{c}_0 + \boldsymbol{c})\|_{\mathcal{C},1}$ while inducing a larger change in $\|\boldsymbol{W}(\boldsymbol{c}_0+\boldsymbol{c})\|_{\mathcal{C}^c}$. We see that we may decrease $\beta_{\boldsymbol{W}}$ if we make w_j smaller for $j \in \mathcal{C}$, and larger for $j \in \mathcal{C}^c$, and this is consistent with our intuition regarding the identification of weights. As such, for small $\beta_{\boldsymbol{W}}$ we expect that $\|\boldsymbol{c}+\boldsymbol{c}_0\|_1 > \|\boldsymbol{c}_0\|_1$ for all $\boldsymbol{c} \in \mathcal{N}(\boldsymbol{\Psi})$, and the following theorem shows that a critical value for $\beta_{\boldsymbol{W}}$ is 1.

Theorem 2.3.2. *If $\beta_{\boldsymbol{W}} < 1$, then finding a solution to $\mathcal{P}_{1,0}^{(\boldsymbol{W})}$ is identical to finding a solution to $\mathcal{P}_{0,0}$. This result is sharp in that if $\beta_{\boldsymbol{W}} \geq 1$, a solution to $\mathcal{P}_{1,0}^{(\boldsymbol{W})}$, may not be identical to any solution of $\mathcal{P}_{0,0}$.*

Proof. Closely related to $\beta_{\boldsymbol{W}}$, we define the quantity $\gamma_{\boldsymbol{W}}$ given by

$$\gamma_{\boldsymbol{W}} = \max_{\boldsymbol{c}\in\mathcal{N}(\boldsymbol{\Psi})} \frac{\|\boldsymbol{W}\boldsymbol{c}\|_{\mathcal{C},1}}{\|\boldsymbol{W}\boldsymbol{c}\|_{1}}, \tag{2.16}$$

where the two constants are related by

$$\beta_{\boldsymbol{W}} = (\gamma_{\boldsymbol{W}}^{-1} - 1)^{-1}.$$

Recalling that $\boldsymbol{c}_0$ is supported on $\mathcal{C}$, we have that

$$\|\boldsymbol{W}(\boldsymbol{c}+\boldsymbol{c}_0)\|_1 = \|\boldsymbol{W}(\boldsymbol{c}+\boldsymbol{c}_0)\|_{\mathcal{C},1} + \|\boldsymbol{W}\boldsymbol{c}\|_{\mathcal{C}^c,1}.$$

Applying the reverse triangle inequality to $\|\boldsymbol{W}(\boldsymbol{c}+\boldsymbol{c}_0)\|_{\mathcal{C},1}$, we have that

$$\|\boldsymbol{W}(\boldsymbol{c}+\boldsymbol{c}_0)\|_1 \geq \|\boldsymbol{W}\boldsymbol{c}_{\mathbf{0}}\|_{\mathcal{C},1} - \|\boldsymbol{W}\boldsymbol{c}\|_{\mathcal{C},1} + \|\boldsymbol{W}\boldsymbol{c}\|_{\mathcal{C}^c,1}.$$

By the definition of $\gamma_{\boldsymbol{W}}$ in (2.16) we have that

$$\begin{aligned} \|\boldsymbol{W}\boldsymbol{c}\|_{\mathcal{C},1} &\leq \gamma_{\boldsymbol{W}}\|\boldsymbol{W}\boldsymbol{c}\|_1, \\ \|\boldsymbol{W}\boldsymbol{c}\|_{\mathcal{C}^c,1} &= \|\boldsymbol{W}\boldsymbol{c}\|_1 - \|\boldsymbol{W}\boldsymbol{c}\|_{\mathcal{C},1}, \\ &\geq (1-\gamma_{\boldsymbol{W}})\|\boldsymbol{W}\boldsymbol{c}\|_1. \end{aligned}$$

It follows that

$$\begin{aligned} \|\boldsymbol{W}(\boldsymbol{c}+\boldsymbol{c}_0)\|_1 &\geq \|\boldsymbol{W}\boldsymbol{c}_{\mathbf{0}}\|_{\mathcal{C},1} - \gamma_{\boldsymbol{W}}\|\boldsymbol{W}\boldsymbol{c}\|_{\mathcal{C},1} + (1-\gamma_{\boldsymbol{W}})\|\boldsymbol{W}\boldsymbol{c}\|_1, \\ &= \|\boldsymbol{W}\boldsymbol{c}_{\mathbf{0}}\|_{\mathcal{C},1} + (1-2\gamma_{\boldsymbol{W}})\|\boldsymbol{W}\boldsymbol{c}\|_1, \end{aligned}$$

which implies that when $\gamma_{\boldsymbol{W}} < 0.5$, or equivalently when $\beta_{\boldsymbol{W}} < 1$,

$$\|\boldsymbol{W}(\boldsymbol{c}+\boldsymbol{c}_0)\|_1 > \|\boldsymbol{W}\boldsymbol{c}_{\mathbf{0}}\|_{\mathcal{C},1} = \|\boldsymbol{W}\boldsymbol{c}_{\mathbf{0}}\|_1,$$

and as such $\boldsymbol{c}_0$ solves $\mathcal{P}_{1,0}^{(\boldsymbol{W})}$. To show sharpness, let $\boldsymbol{W}$ be the identity matrix. For $\alpha > 0$ define $\boldsymbol{\Psi}$ and $\boldsymbol{u}$ by

$$\boldsymbol{\Psi} = \begin{pmatrix} \alpha & 0 & 1 \\ 0 & \alpha & 1 \end{pmatrix}; \qquad \boldsymbol{u} = \begin{pmatrix} \alpha \\ \alpha \end{pmatrix}.$$

Note that the solution to $\mathcal{P}_{0,0}$ is always $(0\ 0\ \alpha)^T$, and as such $\beta_{\boldsymbol{W}} = \alpha/2$. If $\beta_{\boldsymbol{W}} = 1$, corresponding to $\alpha = 2$, then $(0\ 0\ 2)^T$ or $(1\ 1\ 0)^T$ are both solutions to $\mathcal{P}_{1,0}^{(\boldsymbol{W})}$. If $\beta_{\boldsymbol{W}} > 1$, corresponding to $\alpha > 2$, the solution to $\mathcal{P}_{1,0}^{(\boldsymbol{W})}$ is $(1\ 1\ 0)^T$.

As an aside, we note that if $\beta_{\boldsymbol{W}} < 1$, corresponding to $\alpha < 2$, the unique solution to $\mathcal{P}_{1,0}^{(\boldsymbol{W})}$ is $(0\ 0\ \alpha)^T$ as guaranteed by the theorem. □

This result suggests $\beta_{\boldsymbol{W}}$ as a measure of quality of $\boldsymbol{W}$ with smaller $\beta_{\boldsymbol{W}}$ being preferable. The following bound is useful in relating the recovery via weighted ℓ_1-minimization of a particular $\boldsymbol{c}_0$ to a uniform recovery in terms of the one implied by the RIC.

Theorem 2.3.3. *Let*

$$c := \min_{i\in\mathcal{C}} w_i / \max_{i\in\mathcal{C}^c} w_i;$$

$$C := \max_{i\in\mathcal{C}} w_i / \min_{i\in\mathcal{C}^c} w_i.$$

It follows that,

$$c\beta_{\boldsymbol{I}} \leq \beta_{\boldsymbol{W}} = \max_{\boldsymbol{c}\in\mathcal{N}(\boldsymbol{\Psi})} \frac{\|\boldsymbol{W}\boldsymbol{c}\|_{\mathcal{C},1}}{\|\boldsymbol{W}\boldsymbol{c}\|_{\mathcal{C}^c,1}} \leq C\beta_{\boldsymbol{I}}. \tag{2.17}$$

Further,

$$\beta_{\boldsymbol{I}} \leq \frac{\sqrt{2}\delta_{2|\mathcal{C}|}}{1-\delta_{2|\mathcal{C}|}}, \tag{2.18}$$

where δ is a RIC.

Proof. We first note that (2.17) follows from the definition of $\beta_{\boldsymbol{W}}$ in (2.15). To show (2.18), note that by Lemma 2.2 of [103], it follows that for any vector $\boldsymbol{x}$ in the null space of $\boldsymbol{\Psi}$,

$$\|\boldsymbol{x}\|_{\mathcal{C},1} \leq \frac{\sqrt{2}\delta_{2|\mathcal{C}|}}{1-\delta_{2|\mathcal{C}|}}\|\boldsymbol{x}\|_{\mathcal{C}^c,1},$$

which shows the bound. □

To complete our discussion on the theoretical analysis of weighted ℓ_1-minimization, we require a sufficiently small RIC δ to bound $\beta_{\boldsymbol{I}}$ and $\beta_{\boldsymbol{W}}$ in Theorem 2.3.3, and hence β in (2.14). For this, we report the result of [91, Theorem 4.3] – on general bounded orthonormal basis $\{\psi_j\}$ – specialized to the case of multi-variate Legendre PC expansions.

Corollary 2.3.1. *Let $\{\psi_j\}_{1\leq j\leq P}$ be a Legendre PC basis in d independent random variables $\boldsymbol{\Xi} = (\Xi_1, \ldots, \Xi_d)$ uniformly distributed over $[-1,1]^d$ and with a total degree less than or equal* *to q. Let the matrix $\boldsymbol{\Psi}$ with entries $\boldsymbol{\Psi}(i,j) = \psi_j(\boldsymbol{\xi}^{(i)})$ correspond to realizations of $\{\psi_j\}$ at $\boldsymbol{\xi}^{(i)}$ sampled independently from the measure of Ξ. If*

$$N \geq C 3^q \delta^{-2} s \log^3(s) \log(P), \tag{2.19}$$

then the RIC, δ_s, of $\frac{1}{\sqrt{N}}\boldsymbol{\Psi}$ satisfies $\delta_s \leq \delta$ with probability larger than $1 - P^{-\gamma \log^3(s)}$. Here, C and γ are constants independent of N, q, and d.

Proof. The proof is a direct consequence of Theorem 4.3 in [91] by observing that $\{\psi_j\}_{1\leq j\leq P}$ admits a uniform bound $\sup_j \|\psi_j\|_\infty = 3^{\frac{q}{2}}$, see, e.g. [47]. □

Remark 2.3.1 (Weighted ℓ_1-minimization vs. ℓ_1-minimization)**.** *While our theoretical analyses provide insight on the accuracy of the solution to the weighted ℓ_1-minimization problem $\mathcal{P}_{1,\epsilon}^{(\boldsymbol{W})}$ relative to the solution to $\mathcal{P}_{0,\epsilon}$ or $\mathcal{P}_{0,0}$, they do not provide conclusive comparison between the accuracy of the solution to $\mathcal{P}_{1,\epsilon}^{(\boldsymbol{W})}$ and the standard ℓ_1-minimization problem $\mathcal{P}_{1,\epsilon}$. However, for cases where the choice of $\boldsymbol{W}$ is such that the constant C in (2.17) is sufficiently smaller than 1, more accurate solutions may be expected from $\mathcal{P}_{1,\epsilon}^{(\boldsymbol{W})}$ than $\mathcal{P}_{1,\epsilon}$.*

2.4 Numerical examples

In this section, we empirically demonstrate the accuracy of the weighted ℓ_1-minimization approach in estimating statistics of solutions to two differential equations with random inputs.

2.4.1 Case I: Elliptic equation with stochastic coefficient

We first consider the solution of an elliptic realization of (2.6) in one spatial dimension, defined by

$$\begin{aligned} &-\nabla \cdot (a(x,\Xi)\nabla u(x,\Xi)) = 1 \quad x \in \mathcal{D} = (0,1), \\ &u(0,\Xi) = u(1,\Xi) = 0. \end{aligned} \tag{2.20}$$

We assume that the diffusion coefficient $a(x,\Xi)$ is modeled by the expansion

$$a(x,\Xi) = \bar{a}(x) + \sigma_a \sum_{k=1}^{d} \sqrt{\lambda_k}\varphi_k(x)\Xi_k,$$

in which the random variables $\{\Xi_k\}_{k=1}^d$ are independent and uniformly distributed on $[-1,1]$. Additionally, $\{\varphi_k\}_{k=1}^d$ are the eigenfunctions of the Gaussian covariance kernel

$$C_{aa}(x_1,x_2) = \exp\left[-\frac{(x_1-x_2)^2}{l_c^2}\right],$$

corresponding to d largest eigenvalues $\{\lambda_k\}_{k=1}^d$ of $C_{aa}(x_1,x_2)$ with correlation length $l_c = 1/16$. In our numerical tests, we set $\bar{a}(x) = 0.1$, $\sigma_a = 0.021$, and $d = 40$ resulting in strictly positive realizations of $a(x,\Xi)$. Noting that d represents the dimension of the problem in stochastic space, the Legendre PC basis functions for this problem are chosen as in (2.7), where we use an incomplete third order truncation, i.e., $q = 3$, with only $P = 2500$ basis functions. The PC basis functions $\{\psi_j\}$ are sorted such that, for any given order q, the random variables Ξ_k with smaller indices k appear first in the basis. The quantity of interest is $u(0.5,\Xi)$, the solution in the middle of the spatial domain.

Setting weights w_j

Recently, work has been done to derive estimates for the decay of the coefficients $c_{\boldsymbol{\alpha}}(x)$ in the Legendre PC expansion of the solution $u(x,\Xi) \approx \sum_{\boldsymbol{\alpha}} c_{\boldsymbol{\alpha}}(x)\psi_{\boldsymbol{\alpha}}(\Xi)$ to problem (2.20), [104, 29, 105]. Such estimates allow us to identify *a priori* knowledge of $\boldsymbol{c}$ and set the weights w_j in the weighted ℓ_1-minimization approach. In particular, following [29, Proposition 3.1], the coefficients $c_{\boldsymbol{\alpha}}$ admit the bound

$$\|c_{\boldsymbol{\alpha}}\|_{H_0^1(\mathcal{D})} \leq C_0 \frac{|\boldsymbol{\alpha}|!}{\boldsymbol{\alpha}!} e^{-\sum_{k=1}^d g_k \alpha_k}, \quad g_k = -\log\left(r_k/(\sqrt{3}\log 2)\right), \tag{2.21}$$

for some $C_0 > 0$ and $\boldsymbol{\alpha}! = \prod_{k=1}^d \alpha_k!$. The coefficients r_k in (2.21) are given by $r_k = \frac{\sigma_a\sqrt{\lambda_k}\|\varphi_k\|_{L^\infty(D)}}{a_{\min}}$, where $a_{\min} = \bar{a} - \sigma_a \sum_{k=1}^d \sqrt{\lambda_k}\|\varphi_k\|_{L^\infty(D)}$. As suggested in [29], a tighter bound on $\|c_{\boldsymbol{\alpha}}\|_{H_0^1(\mathcal{D})}$ is obtained when the g_k coefficients are computed numerically using one-dimensional analyses instead of the theoretical values given in (2.21). Specifically, for each k, the random variables Ξ_j, $j \neq k$, in (2.20) are set to their mean values and the PCE coefficients c_{α_k} of the corresponding solution – now one-dimensional at the stochastic level – are computed via, for instance, least-squares regression or sufficiently high level stochastic collocation. Notice that the total cost of such one-dimensional calculations depends linearly on d. Using these c_{α_k} values, the coefficient g_k is computed from the one-dimensional version of (2.21), i.e., $|c_{\alpha_k}| \sim e^{-g_k \alpha_k}$. In the present study, we adopt this numerical procedure to estimate each g_k.

As depicted in Figure 2.2, the bound in (2.21) allows us to identify an anticipated $\boldsymbol{c}$, which we use for setting the weights w_j in the weighted ℓ_1-minimization approach. The magnitude of reference coefficients was calculated by the regression approach of [30] using a sufficiently large number of solution realizations. We see that the reference values $|c_j|$ associated with some of the second and third degree basis functions decay slower than anticipated, but that the estimate is a reasonable guess without the use of realizations of $u(x, \Xi)$.

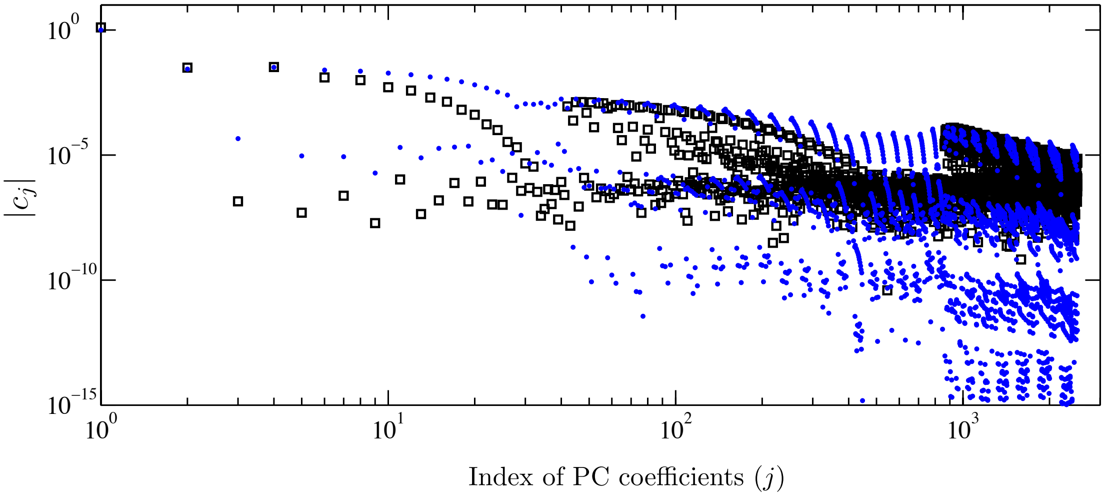


Figure 2.2: Polynomial chaos coefficients $\boldsymbol{c}$ of $u(0.5, \Xi)$ and the corresponding analytical bounds obtained from (2.21) (□ reference; ● analytical bound).

Results

To demonstrate the convergence of the standard and weighted ℓ_1-minimization, we consider an increasing number $N = \{81, 200, 1000\}$ of random solution samples. For each analysis, we estimate the truncation error tolerance ϵ in (2.4) based on the cross-validation algorithm described in Section 2.2.5. To account for the dependency of the compressive sampling solution on the choice of realizations, for each N, we perform 100 independent replications of standard and weighted ℓ_1-minimization, corresponding to independent solution realizations. We then generate uncertainty bars on solution accuracies based on these replications. Specifically, the symbols identify the average error values, and the uncertainty bars show the minimum and maximum error values among the 100 replications. We follow the same process for generating the error plots in Section 2.4.2.

Figure 2.3 displays a comparison between the accuracy of ℓ_1-minimization, weighted ℓ_1-minimization, iteratively re-weighted ℓ_1-minimization, and (isotropic) sparse grid stochastic collocation with Clenshaw-Curtis abscissas. The level one sparse grid contains $N = 81$ points. In particular, we observe that both ℓ_1-minimization and weighted ℓ_1-minimization result in smaller standard deviation and root mean square (rms) errors, compared to the stochastic collocation approach. Additionally, the weighted ℓ_1-minimization using the analytical decay of $|c_{\boldsymbol{\alpha}}|$ outperforms the iteratively re-weighted ℓ_1-minimization. Moreover, for small sample sizes N, the weighted ℓ_1-minimization outperforms the non-weighted approach. This is expected as the prior knowledge on the decay of $|c_{\boldsymbol{\alpha}}|$ has comparable effect on the accuracy as the solution realizations do. In fact, the trade-off between the prior knowledge (in the form of weights w_j) and the solution realizations (data) may be best seen in a Bayesian formulation of the compressive sampling problem $\mathcal{P}_{1,\epsilon}$. We refer the interested reader to [106, 107] for further information on this subject.

In Figure 2.4, we compare the accuracy of weighted ℓ_1-minimization and iteratively re-weighted ℓ_1-minimization in which the first iteration is performed via weighted ℓ_1-minimization instead of the standard ℓ_1-minimization. As can be seen from that plot, no significant difference is observed between the two approaches, thus suggesting additional iterations may not be necessary.

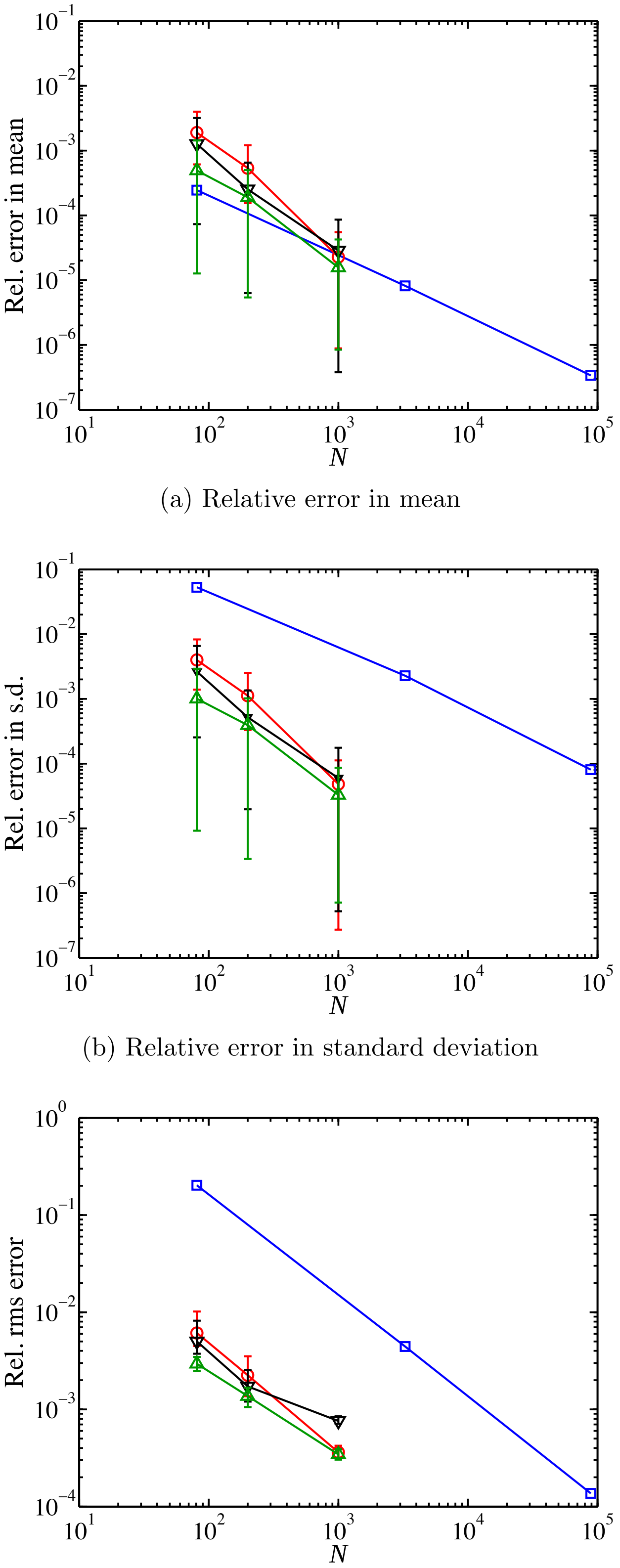
In the presence of the *a priori* estimates of the PC coefficients, one may consider solving a weighted least-squares regression problem $\mathcal{P}^{(\boldsymbol{W})}_{2,\epsilon} \equiv \{\arg\min_{\boldsymbol{c}_{\mathcal{C}}} \|\boldsymbol{W}\boldsymbol{c}_{\mathcal{C}}\|_2 : \|\boldsymbol{\Psi}_{\mathcal{C}}\boldsymbol{c}_{\mathcal{C}} - \boldsymbol{u}\|_2 \leqslant \epsilon\}$, in which $\boldsymbol{c}_{\mathcal{C}} \in \mathbb{R}^P$ denotes vectors supported on a set $\mathcal{C}$ with cardinality $|\mathcal{C}| \leq N$ identified based on the decay of PC coefficients. For example, to generate a well-posed weighted least-squares problem, $\mathcal{C}$ may contain the indices associated with $|\mathcal{C}| \leq \lfloor N/2 \rfloor$ largest (in magnitude) PC coefficients from (2.21). Stated differently, the estimates of PC coefficients may be utilized to form least-squares problems for *small* subsets of the PC basis function that are expected to be important. However, our numerical experiments indicate that, unlike in the case of weighted ℓ_1-minimization, the accuracy of such an approach is sensitive to the quality of the PC coefficient estimates, based on which $\mathcal{C}$ is set. Figure 2.5 presents an illustration of such observation.

2.4.2 Case II: Thermally driven flow with stochastic boundary temperature

Following [6, 3, 108], we next consider a 2-D heat driven square cavity flow problem, shown in Figure 2.6a, as another realization of (2.6). The left vertical wall has a deterministic, constant temperature $\tilde{T}_h$, referred to as the hot wall, while the right vertical wall has a stochastic temperature $\tilde{T}_c < \tilde{T}_h$ with constant mean $\bar{\tilde{T}}_c$, referred to as the cold wall. Both top and bottom walls are assumed to be adiabatic. The reference temperature and the reference temperature difference are defined as $\tilde{T}_{ref} = (\tilde{T}_h + \bar{\tilde{T}}_c)/2$ and $\Delta\tilde{T}_{ref} = \tilde{T}_h - \bar{\tilde{T}}_c$, respectively. In dimensionless variables, the governing equations (in the small temperature difference regime, i.e., Boussinesq approximation) are given by

$$\begin{aligned}
&\frac{\partial \boldsymbol{u}}{\partial t} + \boldsymbol{u}\cdot\nabla\boldsymbol{u} = -\nabla p + \frac{\mathrm{Pr}}{\sqrt{\mathrm{Ra}}}\nabla^2\boldsymbol{u} + \mathrm{Pr}T\hat{\boldsymbol{y}},\\
&\nabla\cdot\boldsymbol{u} = 0, \qquad\qquad (2.22)\\
&\frac{\partial T}{\partial t} + \nabla\cdot(\boldsymbol{u}T) = \frac{1}{\sqrt{\mathrm{Ra}}}\nabla^2 T,
\end{aligned}$$

where $\hat{\boldsymbol{y}}$ is the unit vector $(0,1)$, $\boldsymbol{u} = (u,v)$ is velocity vector field, $T = (\tilde{T} - \tilde{T}_{ref})/\Delta\tilde{T}_{ref}$ is normalized temperature ($\tilde{T}$ denotes non-dimensional temperature), p is pressure, and t is time. Non-dimensional Prandtl and Rayleigh numbers are defined, respectively, as $\mathrm{Pr} = \tilde{\mu}\tilde{c}_p/\tilde{\kappa}$ and $\mathrm{Ra} = \tilde{\rho}g\beta\Delta\tilde{T}_{ref}\tilde{L}^3/(\tilde{\mu}\tilde{\kappa})$, where the superscript tilde (˜) denotes the non-dimensional quantities. Specifically, $\tilde{\rho}$ is density, $\tilde{L}$ is reference length, g is gravitational acceleration, $\tilde{\mu}$ is molecular viscosity, $\tilde{\kappa}$ is thermal diffusivity, and the coefficient of thermal



(a) Relative error in mean

(b) Relative error in standard deviation

(c) Relative rms error

Figure 2.3: Comparison of relative error in statistics of $u(0.5, \boldsymbol{\Xi})$ for ℓ_1-minimization, weighted ℓ_1-minimization, and isotropic sparse grid stochastic collocation (with Clenshaw-Curtis abscissas) for the case of the elliptic equation. The uncertainty bars are generated using 100 independent replications for each samples size N (—○— ℓ_1-minimization; —△— weighted ℓ_1-minimization; —▽— iteratively re-weighted ℓ_1-minimization; —□— stochastic collocation).

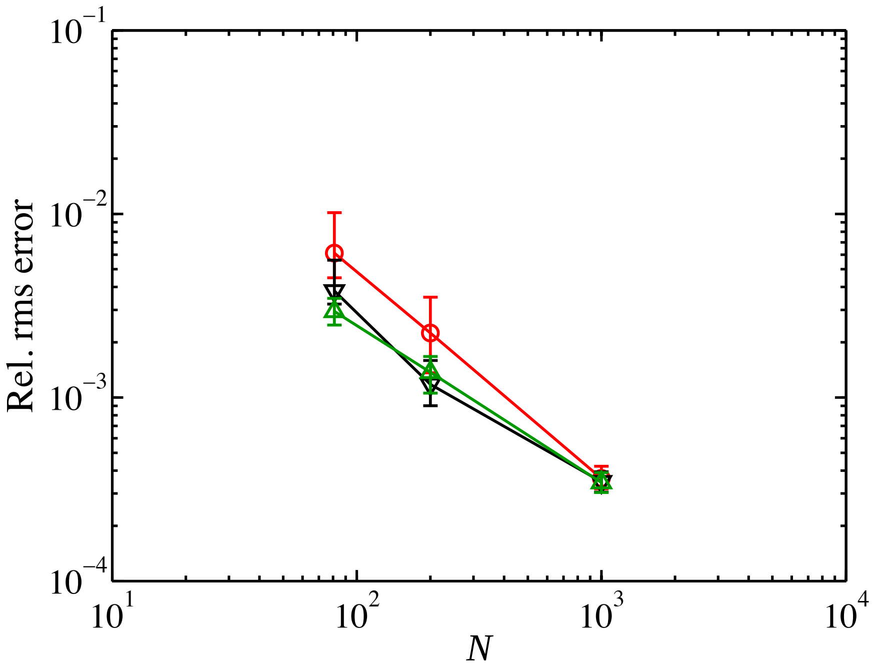


Figure 2.4: Comparison of relative rms error of $u(0.5, \boldsymbol{\Xi})$ for ℓ_1-minimization, weighted ℓ_1-minimization, and iteratively re-weighted ℓ_1-minimization in which the first iteration is performed via weighted ℓ_1-minimization. The uncertainty bars are generated using 100 independent replications for each samples size N (-○- ℓ_1-minimization; -△- weighted ℓ_1-minimization; -▽- iteratively re-weighted ℓ_1-minimization).

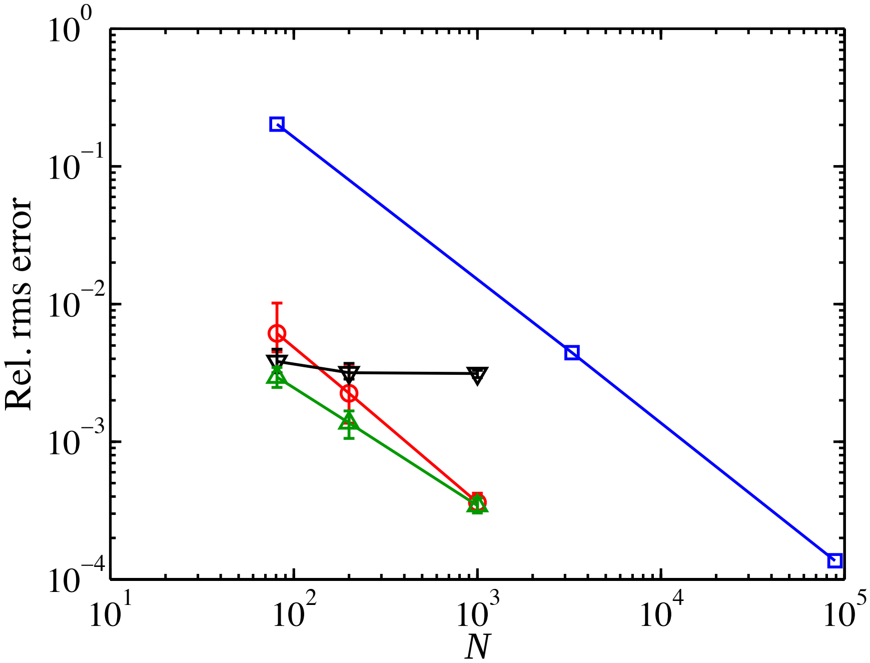


Figure 2.5: Comparison of relative rms error for ℓ_1-minimization, weighted ℓ_1-minimization, weighted least-squares regression, and sparse grid collocation for the case of the elliptic equation. In the weighted least-squares approach the set $\mathcal{C}$ with cardinality $|\mathcal{C}| = \lfloor N/2 \rfloor$ contains the indices of the largest (in magnitude) upper bounds on the PC coefficients (−○− ℓ_1-minimization; −△− weighted ℓ_1-minimization; −▽− weighted least-squares regression; −□− stochastic collocation).

expansion is given by β. In this example, the Prandtl and Rayleigh numbers are set to $\mathrm{Pr} = 0.71$ and $\mathrm{Ra} = 10^6$, respectively. For more details on the non-dimensional variables in (2.22), we refer the interested reader to [108, 6, 3].

On the cold wall, we apply a (normalized) temperature distribution with stochastic fluctuations of the form

$$\begin{aligned} T_c(x = 1, y, \Xi) &= \bar{T}_c + T_c', \\ T_c' &= \sigma_T \sum_{i=1}^{d} \sqrt{\lambda_i}\varphi_i(y)\Xi_i, \end{aligned} \tag{2.23}$$

where $\bar{T}_c$ is a constant mean temperature. In (2.23), Ξ_i, $i = 1, \dots, d$, are independent random variables uniformly distributed on $[-1, 1]$. $\{\lambda_i\}_{i=1}^{d}$ and $\{\varphi_i(y)\}_{i=1}^{d}$ are the d largest eigenvalues and the corresponding eigenfunctions of the exponential covariance kernel

$$C_{T_cT_c}(y_1, y_2) = \exp\left(-\frac{|y_1 - y_2|}{l_c}\right),$$

where l_c is the correlation length. Following [109], the eigenpairs $(\lambda_i, \varphi_i(y))$ in (2.23) are, respectively, given by

$$\lambda_i = \frac{2l_c}{l_c^2\omega_i^2 + 1},$$

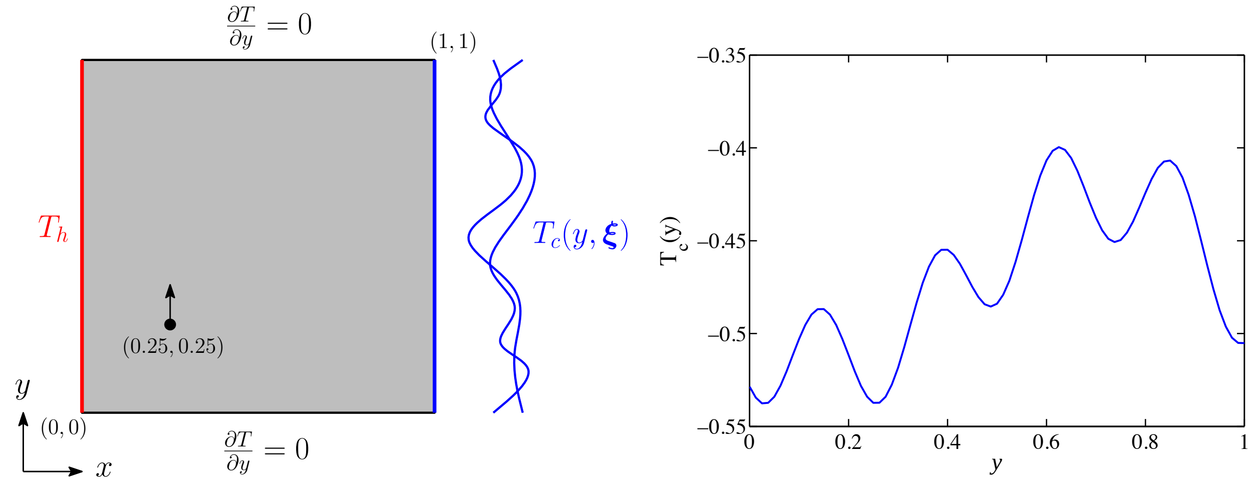
and

$$\varphi_i(y) = \begin{cases} \dfrac{\cos(\omega_i y)}{\sqrt{0.5 + \frac{\sin(\omega_i)}{2\omega_i}}}, & i \text{ is odd}, \\ \dfrac{\sin(\omega_i y)}{\sqrt{0.5 - \frac{\sin(\omega_i)}{2\omega_i}}}, & i \text{ is even}, \end{cases}$$

where each ω_i is a root of

$$\begin{aligned} (1/l_c) - \omega_i \tan(0.5\omega_i) = 0, &\quad i\text{is odd}, \\ \omega_i + (1/l_c)\tan(0.5\omega_i) = 0, &\quad i\text{is even}. \end{aligned}$$

In our numerical test we let $(T_h, \bar{T}_c) = (0.5, -0.5)$, $d = 20$, $l_c = 1/21$, and $\sigma_T = 11/100$. A realization of the cold wall temperature T_c is shown in Figure 2.6b. Our quantity of interest, the vertical velocity component at $(x, y) = (0.25, 0.25)$ denoted by $v(0.25, 0.25)$, is


$\frac{\partial T}{\partial y} = 0$
$(1,1)$
T_h
$T_c(y, \boldsymbol{\xi})$
$(0.25, 0.25)$
y
$(0,0)$
x
$\frac{\partial T}{\partial y} = 0$
–0.35
–0.4
–0.45
–0.5
–0.55
T_c(y)
0
0.2
0.4
0.6
0.8
1
y

(a) Schematic of the geometry and boundary conditions.

(b) A realization of $T_c(x = 1, y)$.

Figure 2.6: Illustration of the cavity flow problem.

expanded in the Legendre PC basis of total degree $q = 4$ with only the first $P = 2500$ basis functions retained, as described in the case of the elliptic problem. We seek to accurately reconstruct $v(0.25, 0.25)$ with $N < P$ random samples of $\boldsymbol{\Xi}$ and the corresponding realizations of $v(0.25, 0.25)$.

Approximate bound on PC coefficients

In order to generate the weights w_j for the weighted ℓ_1-minimization reconstruction of $v(0.25, 0.25)$, we derive an approximate bound on the PC coefficients of the velocity v in (2.22) at a fixed point in space.

For the interest of notation, we start by rewriting T_c' in (2.23) as

$$T_c'(y, \boldsymbol{\Xi}) = \sum_{i=1}^{d} \nu_i(y)\Xi_i, \tag{2.24}$$

where $\nu_i(y)$, $i = 1, \dots, d$, is given by

$$\nu_i(y) = \sigma_T \sqrt{\frac{\lambda_i}{0.5 + (-1)^{i-1}\sin(\omega_i)/2\omega_i}} \sin\left(\omega_i y + \frac{\pi}{2}\left((-1)^i + 1\right)\right).$$

We write the PC expansion of v as $v = \sum_j c_j \psi_j(\boldsymbol{\Xi})$ and seek approximate bounds on $|c_j|$ to set the weights w_j in the weighted ℓ_1-minimization results. By the orthonormality of the PC basis, c_j is

$$c_j = \int_{[-1,1]^d} v(\boldsymbol{\xi})\psi_j(\boldsymbol{\xi}) \left(\frac{1}{2}\right)^d d\boldsymbol{\xi}. \tag{2.25}$$

To approximately bound the coefficients c_j, we examine the functional Taylor series expansion of v around $v = v(\bar{T}_c)$. Note that by an appropriate definition of functional derivatives $\frac{\delta^k v}{\delta T_c^k}$ of v with respect to T_c, see, e.g., [110],

$$v(\boldsymbol{\Xi}) = \sum_{k=0}^{\infty} \frac{1}{k!} \int_{[0,1]^k} \frac{\delta^k v}{\delta \bar{T}_c^k}(\boldsymbol{y}, \boldsymbol{\Xi}) \prod_{j=1}^{k} T_c'(y_j, \boldsymbol{\Xi}) d\boldsymbol{y}, \tag{2.26}$$

where y_j is a copy of the spatial coordinate variable y. Plugging (2.26) in (2.25), we arrive at

$$c_j = \int_{[-1,1]^d} \psi_j(\boldsymbol{\xi}) \sum_{k=0}^{\infty} \frac{1}{k!} \int_{[0,1]^k} \frac{\delta^k v}{\delta \bar{T}_c^k}(\boldsymbol{y}, \boldsymbol{\xi}) \prod_{j=1}^{k} T_c'(y_j, \boldsymbol{\xi}) \left(\frac{1}{2}\right)^d d\boldsymbol{y} d\boldsymbol{\xi}. \tag{2.27}$$

To handle the functional derivatives, we consider the dimensional relation

$$\left|\frac{\delta^k v}{\delta \bar{T}_c^k}(\boldsymbol{y})\right| \approx C\left|\frac{v(\bar{T}_c)}{\left(\bar{T}_c\right)^k}\right|, \tag{2.28}$$

which we assume to hold uniformly in $\boldsymbol{y}$ and Ξ, for some constant $C \geq 0$. This, together with (2.24), allows us to derive the approximate bound

$$|c_j| \lessapprox C|v(\bar{T}_c)|\sum_{k=0}^{\infty}\frac{1}{k!|\bar{T}_c|^k}\left|\int_{[-1,1]^d}\psi_j(\boldsymbol{\xi})\left(\sum_{i=1}^{d}t_i\xi_i\right)^k\left(\frac{1}{2}\right)^d d\boldsymbol{\xi}\right|, \tag{2.29}$$

where $t_i = \int_0^1 \nu_i(y)dy$. In (2.29), the approximation comes from the assumption (2.28) on the functional derivatives. To evaluate the RHS of (2.29), we consider a finite truncation of the sum and a Monte Carlo (or quadrature) estimation of the integral.

In Figure 2.7, we display the approximate upper bound on $|c_j|$ of $v(0.25, 0.25)$ obtained from (2.27) by limiting k to 4. To generate a reference solution, we employ the least-squares regression approach of [30] with $N = 40,000$ random realizations of $v(0.25, 0.25)$. For the accuracies of interest in this study, the convergence of this reference solution was verified. For the sake of illustration, we normalize the estimated $|c_j|$ so that $|c_{\mathbf{0}}|$, the module of the approximate zero degree coefficient, matches its reference counterpart. Despite the rather strong assumption (2.28) on the functional derivatives, we note that the resulting estimates of $|c_j|$ describe the trend of the reference values qualitatively well. As we shall see in what follows, such qualitative agreement is sufficient for the weighted ℓ_1-minimization to improve the accuracy of the standard ℓ_1-minimization for small samples sizes N.

Remark 2.4.1. *We stress that the assumption (2.28), while here lead to appropriate estimates of $|c_j|$ for our particular example of interest, it may not give equally reasonable estimates for other problems or choices of flow parameters, e.g., larger* Ra *numbers. A weaker assumption on the functional derivatives in (2.28), however, requires further study and is the subject of our future work.*

Results

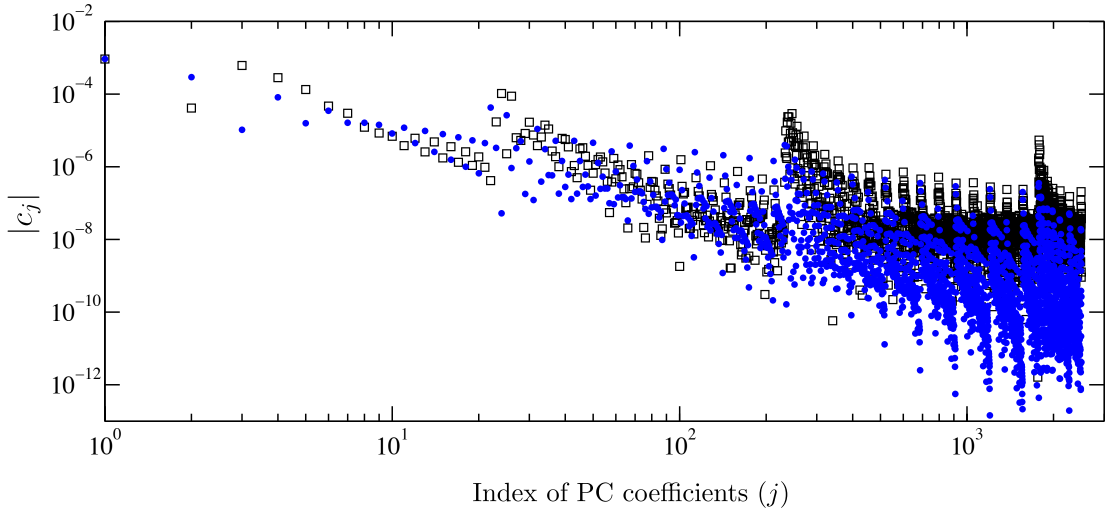


Figure 2.7: Approximate PC coefficients of $v(0.25, 0.25)$ vs. the reference coefficients obtained by least-squares regression using sufficiently large number of realizations of $v(0.25, 0.25)$ (□ reference; ● approximate bound).

We provide results demonstrating the convergence of the statistics of $v(0.25, 0.25)$ as a function of the number of realizations N. For this, we consider sample sizes $N = \{41, 200, 1000\}$ with $N = 41$ corresponding to the number of grid points in level one sparse gird collocation using Clenshaw-Curtis abscissas.

Fig. 2.8 displays comparisons between the accuracies obtained to approximate $v(0.25, 0.25)$. Similar to the previous example, the weighted ℓ_1-minimization approach achieves superior accuracy, particularly for the small sample size $N = 41$. The results obtained for the iteratively re-wighted ℓ_1-minimization correspond to $\epsilon_w = 5 \times 10^{-2} \cdot \hat{c}_1$, where $\hat{c}_1$ is the sample average of $v(0.25, 0.25)$. This leads to the smallest average rms errors among the trial values $\epsilon_w = \{5 \times 10^{-2}, 5 \cdot 10^{-3}, 5 \times 10^{-4}\} \cdot \hat{c}_1$. To show the sensitivity of this approach to the choice of ϵ_w, we present rms error plots in Figure 2.9 corresponding to multiple values of ϵ_w. In particular, for the cases of $\epsilon_w = \{5 \times 10^{-3}, 5 \times 10^{-4}\} \cdot \hat{c}_1$, when $N = 1000$ we observe loss of accuracy compared to the standard ℓ_1-minimization. On the other hand, the weighted ℓ_1-minimization results are relatively insensitive to the choice of ϵ_w, and best performance is obtained with $\epsilon_w = 5 \times 10^{-4} \cdot \hat{c}_1$, i.e., the smallest and most intuitive value among the trials.

Similar to the previous example, in Figure 2.10, we illustrate that weighted ℓ_1-minimization results in similar accuracies given by iteratively re-weighted ℓ_1-minimization in which the first iteration is performed via weighted ℓ_1-minimization. This implies that, given *a priori* knowledge on the decay of PC coefficients, the utilization of iteratively re-weighted ℓ_1-minimization may not be necessary.

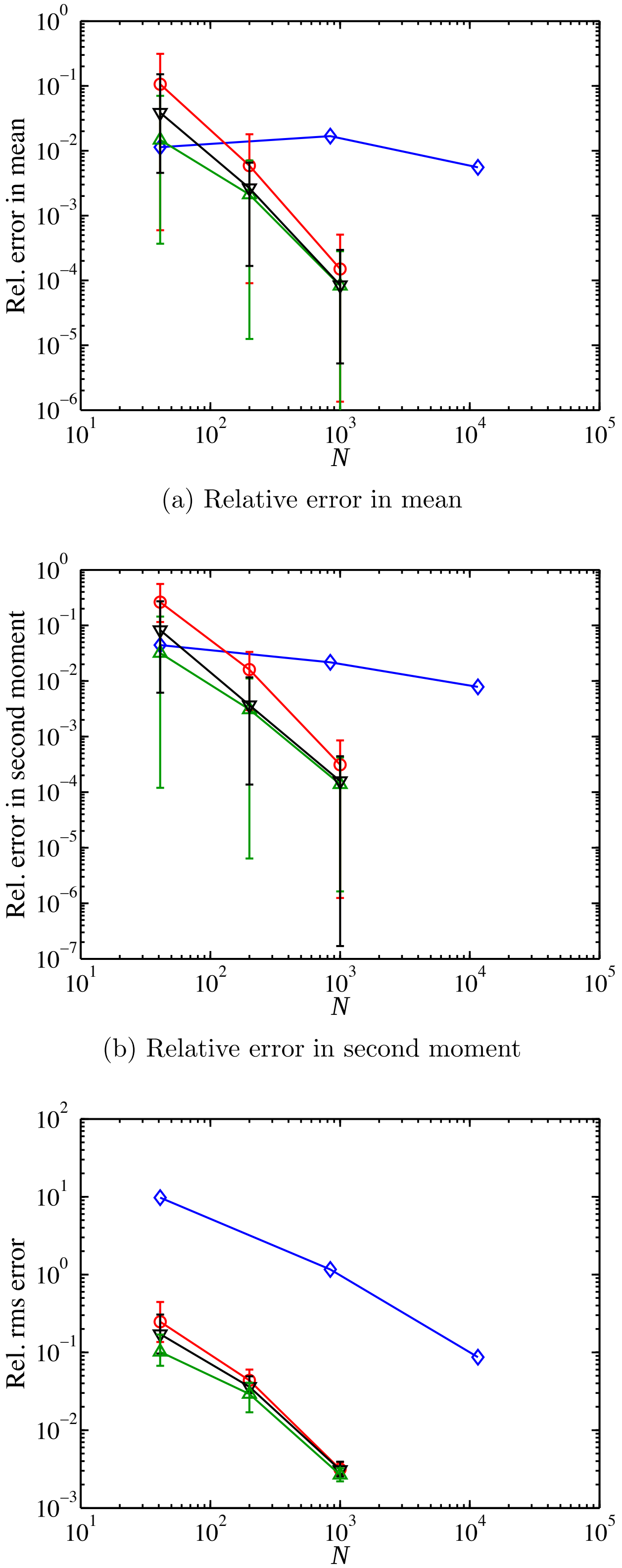
We note that the rather poor performance of the sparse grid collocation is due to the relatively large contributions of some of the higher order PC modes, as may be observed from Figure 2.7. Figure 2.11 shows the magnitude of PC coefficients of $v(0.25, 0.25)$ obtained using standard and weighted ℓ_1-minimization with $N = \{200, 1000\}$ samples. The better approximation quality of the weighted ℓ_1-minimization may be seen particularly from Figs. 2.11a and 2.11b. Finally, in Figure 2.12, we present a comparison between the rms errors obtained from ℓ_1-minimization, weighted ℓ_1-minimization, weighted least-squares regression, and sparse grid stochastic collocation. The weighted least-squares regression approach performs poorly for $N = \{200, 1000\}$ as some of the basis functions are selected incorrectly given the approximate bounds on the PC coefficients.

2.5 Conclusion

Within the context of compressive sampling of sparse polynomial chaos (PC) expansions, we introduced a *weighted ℓ_1-minimization* approach, wherein we utilized *a priori* knowledge on PC coefficients to enhance the accuracy of the standard ℓ_1-minimization. The *a priori* knowledge of PC coefficients may be available in the form of analytical decay of PC coefficients, e.g., for a class of linear elliptic PDEs with random data, or derived from simple dimensional analysis. These *a priori* estimates, when available, can be used to establish weighted ℓ_1 norms that will further penalize small PC coefficients, and consequently improve the sparse approximation. We provided analytical results guaranteeing the convergence of the weighted ℓ_1-minimization approach.

The performance of the proposed weighted ℓ_1-minimization approach was demonstrated through its application to two test cases. For the first example, dealing with a linear elliptic equation with random coefficient, existing analytical bounds on the magnitude of PC coefficients were adopted to establish the weights. In the second case, for a thermally driven flow problem with stochastic temperature boundary condition, we derived an approximate bound for the PC coefficients via a functional Taylor series expansion and a simple dimensional analysis. In both cases we demonstrated that the weighted ℓ_1-minimization approach outperforms the non-weighted counterpart. Furthermore, better accuracies were obtained using the weighted ℓ_1-minimization approach as compared to the iteratively re-weighted ℓ_1-minimization. Numerical experiments illustrate the sensitivity of the latter approach, unlike the former, with respect to the choice of a parameter defining the weights. Finally, we demonstrated that selection of subsets of PC basis and solving well-posed weighted least-squares regression may result in poor accuracies.

While our numerical and analytical results were for the case of Legendre PC expansions, our work may be extended to other choices of PC basis, such as those based on Hermite or Jacobi polynomials.



(a) Relative error in mean

(b) Relative error in second moment

(c) Relative rms error

Figure 2.8: Comparison of relative error in statistics of $v(0.25, 0.25)$ computed via ℓ_1-minimization, weighted ℓ_1-minimization, iteratively reweighted ℓ_1-minimization, and stochastic collocation. The error bars are generated using 100 independent replications with fixed samples size N (—○— ℓ_1-minimization; —△— weighted ℓ_1-minimization; —▽— iteratively reweighted ℓ_1-minimization; —□— stochastic collocation).

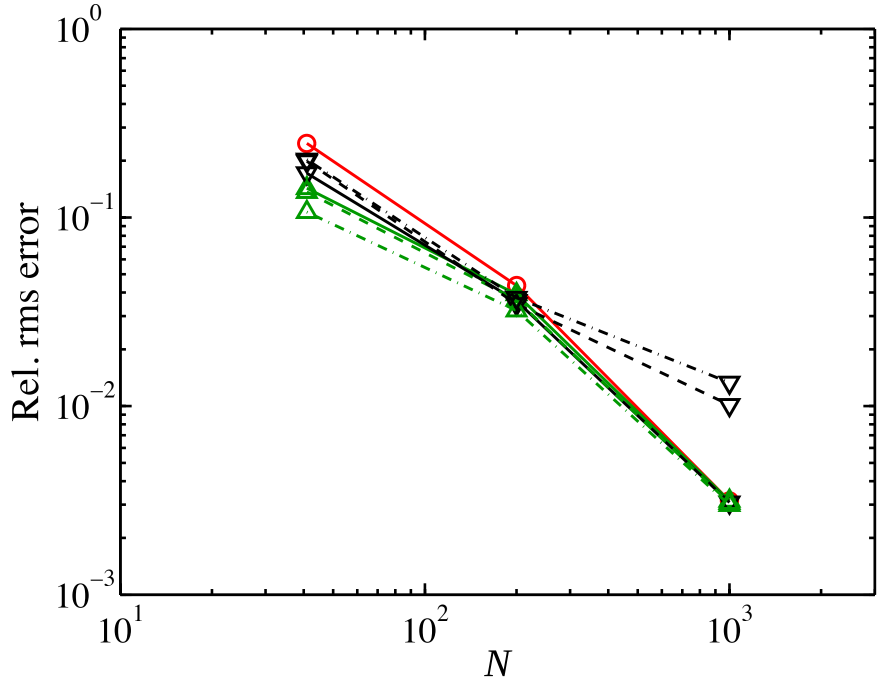


Figure 2.9: Relative average rms errors corresponding to multiple values of ϵ_w to set the weights w_j. The results demonstrate the sensitivity of the iteratively re-weighted approach to the choice of ϵ_p (—○— ℓ_1-minimization; —△— weighted ℓ_1-minimization; —▽— iteratively re-weighted ℓ_1-minimization; solid lines $\epsilon_w = 5 \times 10^{-2} \cdot \hat{c}_1$; dashed lines $\epsilon_w = 5 \times 10^{-3} \cdot \hat{c}_1$; dotted dashed lines $\epsilon_w = 5 \times 10^{-4} \cdot \hat{c}_1$). Here, $\hat{c}_1$ is the sample average of $v(0.25, 0.25)$.

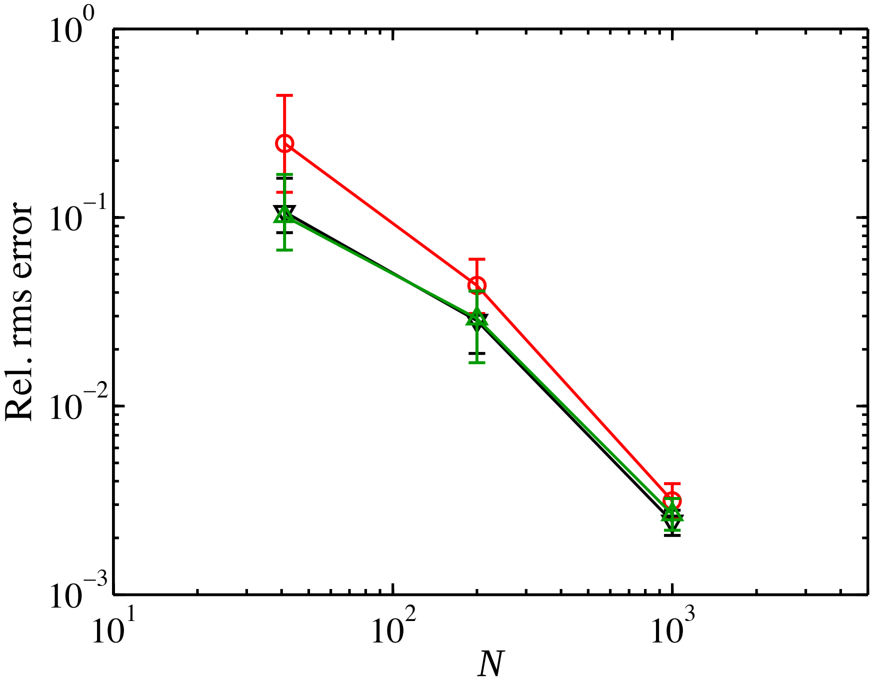
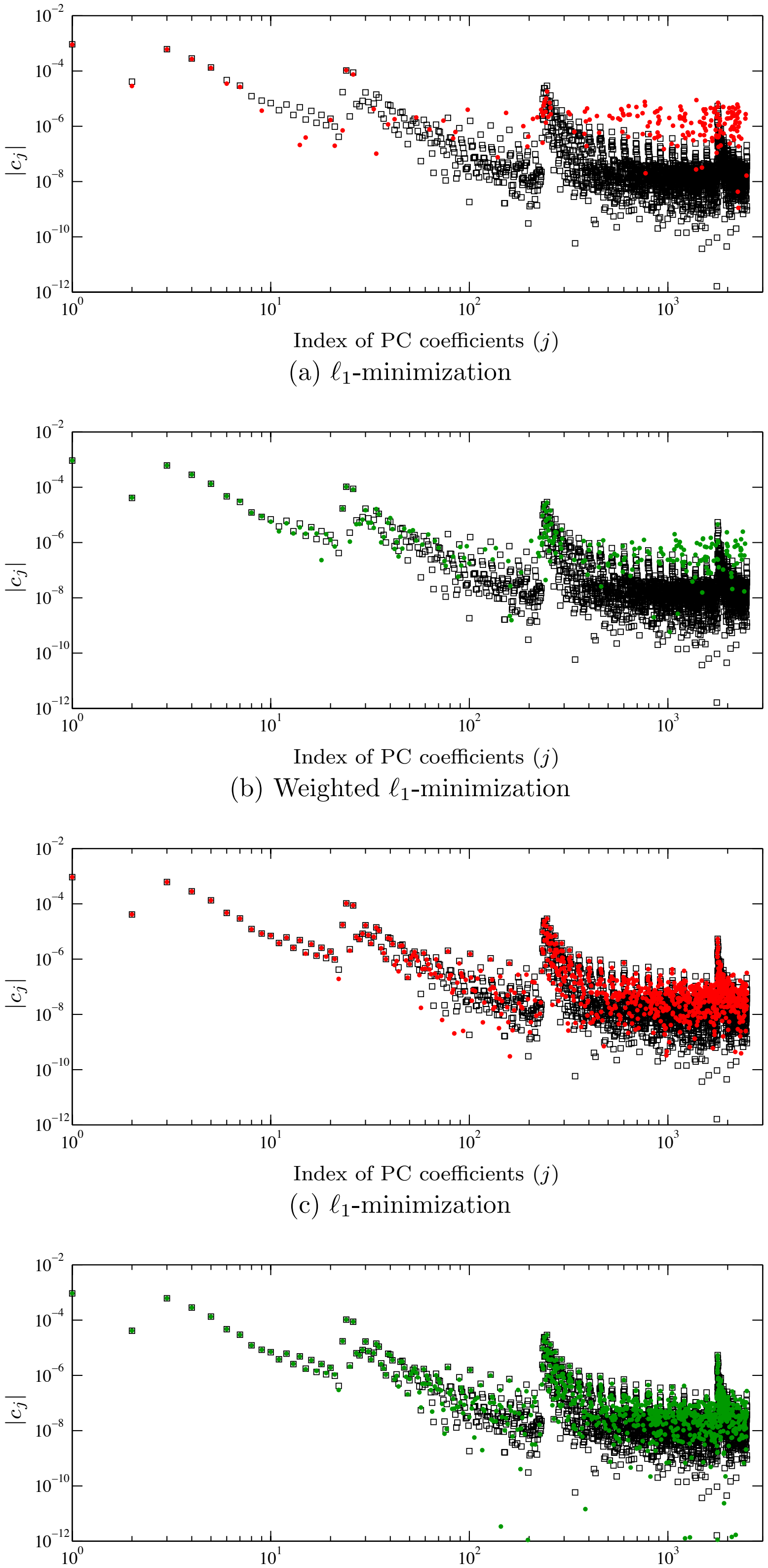


Figure 2.10: Comparison of relative rms error of $v(0.25, 0.25)$ for ℓ_1-minimization, weighted ℓ_1-minimization, and iteratively re-weighted ℓ_1-minimization using the analytical decay of $|c_{\boldsymbol{\alpha}}|$. The uncertainty bars are generated using 100 independent replications for each samples size N (○ ℓ_1-minimization; △ weighted ℓ_1-minimization; ▽ iteratively re-weighted ℓ_1-minimization).



(a) ℓ_1-minimization

(b) Weighted ℓ_1-minimization

(c) ℓ_1-minimization

(d) Weighted ℓ_1-minimization

Figure 2.11: Approximation of PC coefficients of $v(0.25, 0.25)$ using $N = 200$ samples (a), (b) and $N = 1000$ samples (c), (d) (□ reference; ● ℓ_1-minimization; ● weighted ℓ_1-minimization).

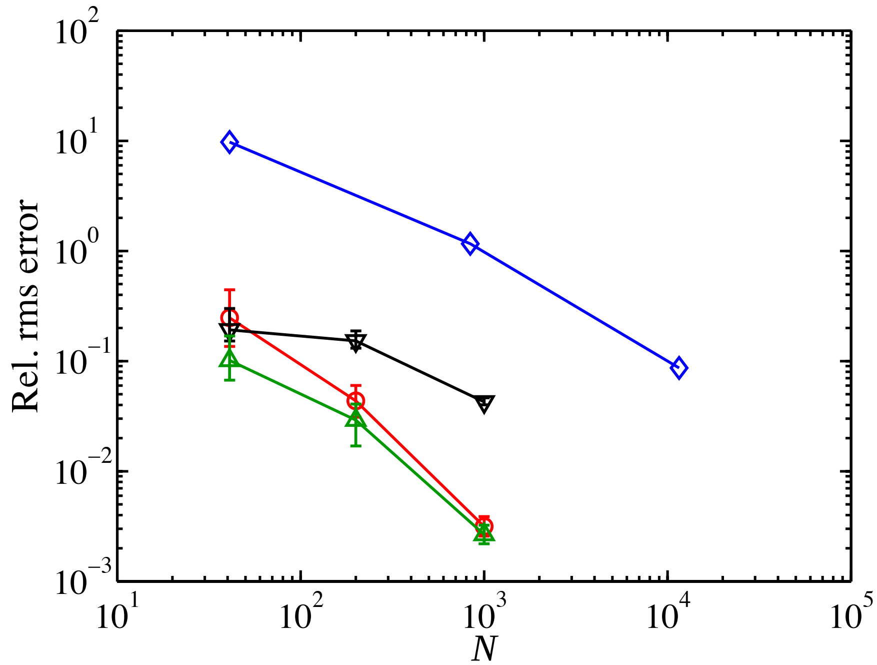


Figure 2.12: Comparison of relative rms error for ℓ_1-minimization, weighted ℓ_1-minimization, weighted least-squares regression, and sparse grid collocation for the cavity flow problem. In the weighted least-squares approach, the set $\mathcal{C}$ with cardinality $|\mathcal{C}| = \lfloor N/2 \rfloor$ contains the indices of the largest (in magnitude) approximate upper bounds on the PC coefficients (○ ℓ_1-minimization; △ weighted ℓ_1-minimization; ▽ weighted least-squares regression; □ stochastic collocation).

CHAPTER 3

A BI-FIDELITY TECHNIQUE VIA ℓ_1-MINIMIZATION AND ORTHOGONAL MATCHING PURSUIT FOR SPARSE POLYNOMIAL CHAOS EXPANSION[1]

Abstract

In this chapter, we investigate bi-fidelity approaches for sparse polynomial chaos expansion (PCE) in the context of compressive sampling, in which computationally economical low-fidelity simulations are utilized to improve the surrogate approximation of a quantity of interest (QoI). PCE coefficients computed from low-fidelity simulations are used as *a priori* information about the high-fidelity coefficients, resulting in an improved accuracy in recovering the solution, furthermore an improved quality in approximating the QoI. This *a priori* information is involved via weighted ℓ_1-minimization and a modified orthogonal matching pursuit (OMP), which is proposed as *bi-fidelity* OMP. Numerical experiments are provided to compare the bi-fidelity and standard methods, and they all show that bi-fidelity approaches admits solution recovery at an enhanced accuracy.

3.1 Introduction

Nowadays, engineering problems are described by highly complex models, in which stochastic variables are used to represent the uncertain parameter inputs, introduced by imperfect knowledge of the physics or inherent variability in the inputs. Uncertainty quantification

[1] This chapter is in preparation to be submitted by *J. Peng et al.* to *AIAA Journal.*

(UQ) [2, 3, 4] is a tool that aims at quantitatively understanding how the quantity of interest (QoI) performs under effects of these uncertain parameters, representing the QoI as a function of the uncertain inputs. These representations are commonly implemented via polynomial chaos expansions (PCEs) [19, 20].

To quantitatively analyze uncertainties via PCE, we consider a natural framework, probability, in which the uncertain inputs are modeled as a d-dimensional vector of independent random variables $\Xi := (\Xi_1, \ldots, \Xi_d)$. This random vector yields to probability density function $\rho(\Xi)$. The QoI that we seek to approximate is denoted by a scalar $u(\Xi)$, and assumed to have finite variance. Therefore, we represent $u(\Xi)$ by an expansion in multivariate orthogonal polynomials $\psi_j(\Xi)$,

$$u(\Xi) = \sum_{j=1}^{\infty} c_j \psi_j(\Xi) \approx \sum_{j=1}^{P} c_j \psi_j(\Xi). \tag{3.1}$$

We call (3.1) a PCE, in which $c_j, j = 1, 2, \ldots,$ are the corresponding PCE coefficients. As $u(\Xi)$ has finite variance, (3.1) usually converges rapid, thereby $u(\Xi)$ can be represented by finite P terms without a significant loss of accuracy. Usually P is determined by the dimension, d, and the highest total order of the PCE, p, which will be further explained in Section 3.2.2.

Typically, it is unnecessary to use all P terms to represent $u(\Xi)$, and we can restrict to the polynomials indexed by $j \in \mathcal{C}$, where the subset $\mathcal{C} \subset \{1, \ldots, P\}$ is unknown, such that $|\mathcal{C}|$, the number of elements in $\mathcal{C}$, is significantly smaller than P. Then, $u(\Xi)$ is approximated by

$$u(\Xi) \approx \sum_{j \in \mathcal{C}} c_j \psi_j(\Xi). \tag{3.2}$$

We call (3.2) a sparse PCE, and $|\mathcal{C}|$, the number of elements in $\mathcal{C}$, is referred to as the sparsity of the expansion. Identifying the small subset $\mathcal{C}$ and the values of the corresponding coefficients falls within the context of compressive sampling [34, 35, 47].

Compressive sampling seeks to identify the PCE coefficients by solving the optimization problem

$$\mathcal{P}_{0,\epsilon} \equiv \left\{ \arg\min_{\boldsymbol{c}} \|\boldsymbol{c}\|_0 : \|\boldsymbol{\Psi}\boldsymbol{c} - \boldsymbol{u}\|_2 \leq \epsilon \right\}, \tag{3.3}$$

where the vector $\boldsymbol{c} := (c_1, \ldots, c_P)$ contains PCE coefficients, and $\|\boldsymbol{c}\|_0$ denotes the number of non-zero entries of $\boldsymbol{c}$. In addition, the vector $\boldsymbol{u}$ and the matrix $\boldsymbol{\Psi}$ contains QoI and PCE basis function evaluations at realizations of the random inputs,

$$\boldsymbol{u} := \left(u(\boldsymbol{\xi}^{(1)}), \ldots, u(\boldsymbol{\xi}^{(N)})\right)^T; \tag{3.4}$$

$$\boldsymbol{\Psi}(i,j) := \psi_j\left(\boldsymbol{\xi}^{(i)}\right), \tag{3.5}$$

where $\boldsymbol{\xi}^{(i)}$ denotes the ith realization of Ξ, and N is the number of realizations.

A solution to $\mathcal{P}_{0,\epsilon}$ provides an optimally sparse approximation of $u(\Xi)$ within ϵ in the ℓ_2 norm. However, in general, solving $\mathcal{P}_{0,\epsilon}$ is an NP-hard problem, i.e., the cost of solving it grows exponentially in P [60]. To resolve this exponential dependence, approaches such as ℓ_1-minimization [54, 55, 56, 34, 60] and orthogonal matching pursuit (OMP) have been proposed.

3.1.1 ℓ_1-minimization

As a convex relaxation of $\mathcal{P}_{0,\epsilon}$, ℓ_1-minimization seeks to identify $\boldsymbol{c}$ by solving

$$\mathcal{P}_{1,\epsilon} \equiv \left\{ \arg\min_{\boldsymbol{c}} \|\boldsymbol{c}\|_1 : \|\boldsymbol{\Psi}\boldsymbol{c} - \boldsymbol{u}\|_2 \leq \epsilon \right\}, \tag{3.6}$$

via convex optimization algorithms [54, 65, 66, 67, 68, 69, 70, 71]. It is shown that in practice, the solution to $\mathcal{P}_{1,\epsilon}$ is usually similar to $\mathcal{P}_{0,\epsilon}$ [111]. For computation, $\mathcal{P}_{1,\epsilon}$ may be solved in form of the ℓ_1-norm regularized least-squares problem or the LASSO problem [72]. In this study we use the MATLAB package SPGL1 [70] based on the spectral projected gradient algorithm [73].

To show that ℓ_1-minimization indeed finds a solution to $\mathcal{P}_{0,\epsilon}$, we let $\boldsymbol{c}_0$ be the solution to $\mathcal{P}_{0,0}$, whose sparsity is $s := \|\boldsymbol{c}_0\|_0$, and $\boldsymbol{c}_1$ be the solution to $\mathcal{P}_{1,0}$. It has been shown in

[64, 82] that when

$$N \geq C(1+\beta)3^p s \log(P), \tag{3.7}$$

where C is non-negative constant,

$$Prob\left[\boldsymbol{c}_1 = \boldsymbol{c}_0\right] \geq 1 - \frac{6}{P} - 6e^{-\beta}. \tag{3.8}$$

From (3.7) and (3.8), we can deduce that as the number of realizations N becomes larger, a higher probability it achieves in recovering the coefficients $\boldsymbol{c}_0$. This is consistent with the observations in our numerical results in Section 3.4.

3.1.2 Orthogonal matching pursuit (OMP)

Orthogonal matching pursuit (OMP) is one of the commonly used greedy algorithms, which may be employed to approximate the solution to $\mathcal{P}_{0,\delta}$ [61, 62]. Denoting the PCE coefficients in the tth iteration by the superscript (t), we start from $\boldsymbol{c}^{(0)} = \boldsymbol{0}$ and an empty active column set of $\boldsymbol{\Psi}$. At any iteration t, OMP identifies only one column to be added to the active column set. The column is chosen such that the ℓ_2-norm of the residual, $\|\boldsymbol{\Psi}\boldsymbol{c}^{(t)} - \boldsymbol{u}\|_2$, is maximally reduced; or identically, the residual, $\boldsymbol{\Psi}\boldsymbol{c}^{(t)} - \boldsymbol{u}$, has highest inner product with the selected column. This active column selection process is also called the sensing part of OMP. Having specified the active column set, a least-squares problem is solved to compute the solution $\boldsymbol{c}^{(t)}$. The iterations are continued until the error truncation tolerance ϵ is achieved. In general, OMP is relatively fast compared to the ℓ_1-minimization algorithm (introduced in Section 3.1.1), both in theory and practice, but most of them deliver smaller recoverable sparsity compared to ℓ_1 minimization [47]. The following exhibit depicts an step-by-step implementation of the OMP algorithm.

Algorithm 3 Orthogonal matching pursuit (OMP)

```
Set $t = 0$, $\boldsymbol{c}^{(0)} = \boldsymbol{0}$, and $\boldsymbol{r}^{(0)} = \boldsymbol{u} - \boldsymbol{\Psi}\boldsymbol{c}^{(0)}$.
Set the initial solution support index set $\mathcal{I}^{(0)} = \emptyset$.
While $\|\boldsymbol{u} - \boldsymbol{\Psi}\boldsymbol{c}^{(t)}\|_2 > \epsilon$: do
    for all $j \notin \mathcal{I}^{(t)}$: do
        Evaluate $\varepsilon(j) = \left|\boldsymbol{\psi}_j^T \boldsymbol{r}^{(t)}\right|$.
    End for
    Set $t = t + 1$.
    Update the support index set $\mathcal{I}^{(t)} = \mathcal{I}^{(t-1)} \cup \{\arg\max_j \epsilon(j)\}$.
    Solve for $\boldsymbol{c}^{(t)} = \arg\min_{\boldsymbol{c}} \|\boldsymbol{u} - \boldsymbol{\Psi}\boldsymbol{c}\|_2$ subject to $Support\{\boldsymbol{c}\} = \mathcal{I}^{(t)}$.
    Update the residual $\boldsymbol{r}^{(t)} = \boldsymbol{u} - \boldsymbol{\Psi}\boldsymbol{c}^{(t)}$.
End while
Output the solution $\boldsymbol{c} = \boldsymbol{c}^{(t)}$.
```

Theoretical analyses have been provided to show that OMP is guaranteed to recover the PCE coefficients in problem $\mathcal{P}_{0,0}$, which is done via the concept *mutual coherence*. The mutual coherence of a matrix $\boldsymbol{\Psi} \in \mathbb{R}^{N \times P}$ is defined by

$$\mu(\boldsymbol{\Psi}) := \max_{1 \leq i,j \leq P, i \neq j} \frac{|\boldsymbol{\psi}_i^T \boldsymbol{\psi}_j|}{\|\boldsymbol{\psi}_i\| \|\boldsymbol{\psi}_j\|}. \tag{3.9}$$

where $\boldsymbol{\psi}_i$ and $\boldsymbol{\psi}_j$ are two columns of $\boldsymbol{\Psi}$. The mutual coherence is a measure of the orthogonality of a matrix. For instance, when $\mu(\boldsymbol{\Psi}) = 0$, the matrix $\boldsymbol{\Psi}$ is unitary, while $\mu(\boldsymbol{\Psi}) = 1$, at least two columns of $\boldsymbol{\Psi}$ are identical. For any given $\boldsymbol{\Psi}$, $0 \leq \mu(\boldsymbol{\Psi}) \leq 1$. Specially, for under-determined case, $N < P$, the mutual coherence $\mu(\boldsymbol{\Psi})$ is strictly positive.

Following Theorem 6 in [60], we know that for problem ($\mathcal{P}_{0,0}$), where $N < P$, if a solution $\boldsymbol{c}_0$ exists satisfying

$$\|\boldsymbol{c}_0\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(\boldsymbol{\Psi})}\right), \tag{3.10}$$

OMP is guaranteed to recover $\boldsymbol{c}_0$ exactly. Furthermore, for problems with high-dimensional random inputs, Corollary 7.4 in [62] and Theorem 3.1 in [64] show a upper bound of $\mu(\boldsymbol{\Psi})$

exists, and $\mu(\mathbf{\Psi})$ is within this upper bound with a high probability, i.e.,

$$Prob\,[\mu(\mathbf{\Psi}) \geq \delta] \leq 2^{3/4} P^2 \exp\left(-\frac{N\delta^2}{2C3^{2p}}\right), \tag{3.11}$$

where rows of $\mathbf{\Psi}$ are the p-order Legendre polynomial chaos basis, independently realized in $d > p$ i.i.d. uniform random variables $\mathbf{\Xi}$, and the constant $C \approx 13.12$.

From (3.11), we can see that if the number of realizations N is larger, or the bound δ is higher, the mutual coherence is more tightly bounded, since the probability of $\mu(\mathbf{\Psi}) \geq \delta$ drops exponentially. In addition, for a fixed set of d and p, as N increases, the upper bound of $\mu(\mathbf{\Psi})$ may decrease, which enables OMP to recover a solution with more non-zero elements. This is also in accordance with the numerical results in Section 3.4.

3.1.3 Motivations

The research on developing ℓ_1-minimization and OMP to improve the efficiency (increase the accuracy or decrease the computational cost) of compressive sampling has been thriving recently [112, 113, 114, 115]. For instance, in [114], a modified data dependent sensing dictionary has been unitized in OMP, the active column set is selected based on the projection of realized samples $\boldsymbol{u}$ onto the polynomial basis functions. In each iteration, the column with largest projection is added to the active set. Nevertheless superior performance has been shown in [114], due to the nonlinearity of the problem of our interest, no significant improvement has been observed. However, motivated by it, we modify the sensing part of OMP with *a priori* information on the PCE coefficient, which will be described in Section 3.3.2. In [112], a weighted ℓ_1-minimization has been proposed when *a priori* information is available, higher PCE accuracy is observed from the weighted ℓ_1-minimization, by solving the problem

$$\mathcal{P}_{1,\epsilon}^{(\boldsymbol{W})} \equiv \left\{\arg\min_{\boldsymbol{c}} \|\boldsymbol{W}\boldsymbol{c}\|_1 : \|\mathbf{\Psi}\boldsymbol{c} - \boldsymbol{u}\|_2 \leq \epsilon\right\}, \tag{3.12}$$

where $\boldsymbol{W}$ is a diagonal matrix to be specified using *a priori* information of the PCE coefficients, which may be obtained analyticall or by dimensional analysis. Nevertheless, in practice, this *analytical* information is not always available.

In the present study, we extend the weighted ℓ_1-minimization approach. To acquire the *a priori* information, we first solve $\mathcal{P}_{1,\epsilon}$ in (3.12), with the vector $\boldsymbol{u}$ containing the evaluations of the QoI in low fidelity, and the solution is used to form $\boldsymbol{W}$ in $\mathcal{P}_{1,\epsilon}^{(\boldsymbol{W})}$. We refer to this extended weighted ℓ_1-minimization as *bi-fidelity* ℓ_1*-minimization*. Additionally, we modify the OMP algorithm to include the *a priori* obtained from low-fidelity solutions, and we call it *bi-fidelity OMP*. Two numerical examples are used to demonstrate the bi-fidelity approaches: an elliptical equation with stochastic diffusion coefficient; a steady state thermally driven cavity flow problem with random temperature boundary condition.

The rest of this manuscript is organized as follows. In Section 3.2, we state the general problem that is considered and introduce the formulation of PCE. In Section 3.3, we present the bi-fidelity approaches, and provide the theoretical conditions that guarantee the recovery of PCE coefficients via bi-fidelity ℓ_1-minimization. Two numerical experiments are used to demonstrate the bi-fidelity approaches in Section 3.4: a spatially two-dimensional elliptic PDE with stochastic diffusion coefficient and a thermally-driven cavity flow with random temperature boundary conditions.

3.2 Framework of the Problem

3.2.1 Problem statement

Considering an engineering system modeled by differential equations on a domain $\mathcal{D} \in \mathbb{R}^D$, $D \in \{1,2,3\}$, in which one or many uncertain parameters are characterized by the d-dimensional vector Ξ, e.g., boundary conditions and/or initial conditions. The solution u is governed by the equations

$$\begin{aligned} \mathcal{L}(\boldsymbol{x}, t, \Xi; u(\boldsymbol{x}, t, \Xi)) &= 0, \qquad \boldsymbol{x} \in \mathcal{D}, \\ \mathcal{I}(\boldsymbol{x}, \Xi; u(\boldsymbol{x}, 0, \Xi)) &= 0, \qquad \boldsymbol{x} \in \mathcal{D}, \\ \mathcal{B}(\boldsymbol{x}, t, \Xi; u(\boldsymbol{x}, t, \Xi)) &= 0, \qquad \boldsymbol{x} \in \partial\mathcal{D}, \end{aligned} \tag{3.13}$$

where $\mathcal{L}$, $\mathcal{I}$, and $\mathcal{B}$ are differential operators depending on the physics of the problem, the initial conditions, and the boundary conditions, respectively. We seek to approximate the solution for some fixed spatial location $\boldsymbol{x}_0$ and time t_0, $u(\boldsymbol{x}_0, t_0, \boldsymbol{\Xi})$, which is the QoI. For brevity, we drop $\boldsymbol{x}_0$ and t_0, and simply write the QoI and its evaluations as $u(\boldsymbol{\Xi})$ and $u(\boldsymbol{\xi})$, respectively. In the present work, we use the finite element methods project *FEniCS* [116] to solve (3.13) for the given problems in Section 3.4.

3.2.2 Polynomial chaos expansion (PCE)

To approximate the QoI, $u(\boldsymbol{\Xi})$, we rely on the PCE (3.1). The inputs, Ξ_k, are assumed to be independent random variables, and identically yield to probability density function ρ_k. The complete set of polynomial basis functions is defined by $\{\psi_{i_k}(\Xi_k)\}$, which contains polynomials of degree $i_k \in \mathbb{N}\cup\{0\}$ orthonormal with respect to the weight function ρ_k [21, 20]. Furthermore, the multivariate orthonormal polynomials in $\boldsymbol{\Xi}$ are given by the products of the univariate orthonormal polynomials,

$$\psi_{\boldsymbol{i}}(\boldsymbol{\Xi}) = \prod_{k=1}^{d} \psi_{i_k}(\Xi_k), \tag{3.14}$$

where the d-dimensional multi-index $\boldsymbol{i} \in \{(i_1, \ldots, i_d) : i_k \in \mathbb{N} \cup \{0\}\}$. For computation, the expansion in (3.1) is truncated to the set of basis functions, $\{\psi_j(\boldsymbol{\Xi})\}_{j=1}^{P}$, associated with the subspace of polynomials of total order not greater than p, i.e. $\sum_{k=1}^{d} i_k \le p$. The cardinality P can be calculated by

$$P = \frac{(d+p)!}{d!p!}. \tag{3.15}$$

For convenience, we also order these P basis functions so that they are indexed by $\{1, \ldots, P\}$, as in (3.1). We note that these two notations deliver the same PCE, and both of them may be used without confusion. Representing $u(\boldsymbol{\Xi})$ by PCE, the problem of approximating the QoI reduces to the problem of identifying the PCE coefficients.

3.3 Compressive Sampling via Bi-fidelity Technique

To increase the probability of accurately recover PCE coefficients, a larger number of samples, N is required [47, 64]. However, the deterministic solver of u is often computationally demanding, and hence it is expensive to compute the realized $u(\Xi)$. UQ including *a priori* on PCE coefficients has been shown to be accuracy-effective [112]. Analytical or approximate decaying information about the PCE coefficients has been used as the *a priori* information, however, it is not guaranteed that this kind of information is always available. In order to sustainably obtain *a priori* in numerical simulations, we consider the so called *bi-fidelity* approaches.

Usually, to accurately recover the PCE coefficients, the evaluations of the QoI are required to be calculated with high numerical accuracy, which may be computed on complex models, fine mesh, high-order accurate schemes, etc.. It is often computationally expensive to evaluate these *high-fidelity* samples, particularly for large scale complex simulations. Nevertheless, if we evaluate the samples on simple models, coarse mesh, or low-order accurate schemes, the computational complexity may be significantly reduced. Correspondingly, these *low-fidelity* computations usually lead to low-accurate realized QoI evaluations, since noise such as discretization errors, truncation errors are introduced. Therefore, we cannot trust the PCE coefficients recovered with the low-fidelity samples, but they may be used as the *a priori* information about the PCE coefficients.

3.3.1 Bi-fidelity ℓ_1-minimization

In bi-fidelity ℓ_1-minimization, we solve the weighted ℓ_1-minimization problem $\mathcal{P}_{1,\epsilon}^{(\boldsymbol{W})}$, in which the diagonal weight matrix $\boldsymbol{W}$ is defined with entries $w_j \geq 0$. We consider the weighted problem $\mathcal{P}_{1,\epsilon}^{(\boldsymbol{W})}$ in (3.12) with

$$\|\boldsymbol{W}\boldsymbol{c}\|_1 = \sum_{j=1}^{P} w_j |c_j|. \tag{3.16}$$

To identify w_j, we first solve $\mathcal{P}_{1,\epsilon}$ in (3.6) with low-fidelity realizations, giving the low-fidelity PCE coefficients, denoted by $\boldsymbol{c}^l$. We then use $\boldsymbol{c}^l$, as the *a priori* information, to define $\boldsymbol{W}$ [86],

$$w_j = \left(|c_j^l| + \delta_w\right)^{-q}, \tag{3.17}$$

where δ_w is a relatively small positive parameter. The parameter $q \in [0,1]$ may be used to account for the confidence in the anticipated $|c_j^l|$. These weights deform the ℓ_1 ball, as Figure 3.1 shows, to discourage small coefficients from the solution and consequently enhance the accuracy. We call $|c_j^l|$ the *a priori* information because the low-fidelity model is solved on a coarse mesh, and evaluating the QoI in low-fidelity is computationally cheap. Therefore, compared to solving $\mathcal{P}_{1,\epsilon}^{(\boldsymbol{W})}$ in high fidelity, the cost on solving for $|c_j^l|$ is usually negligible or small.

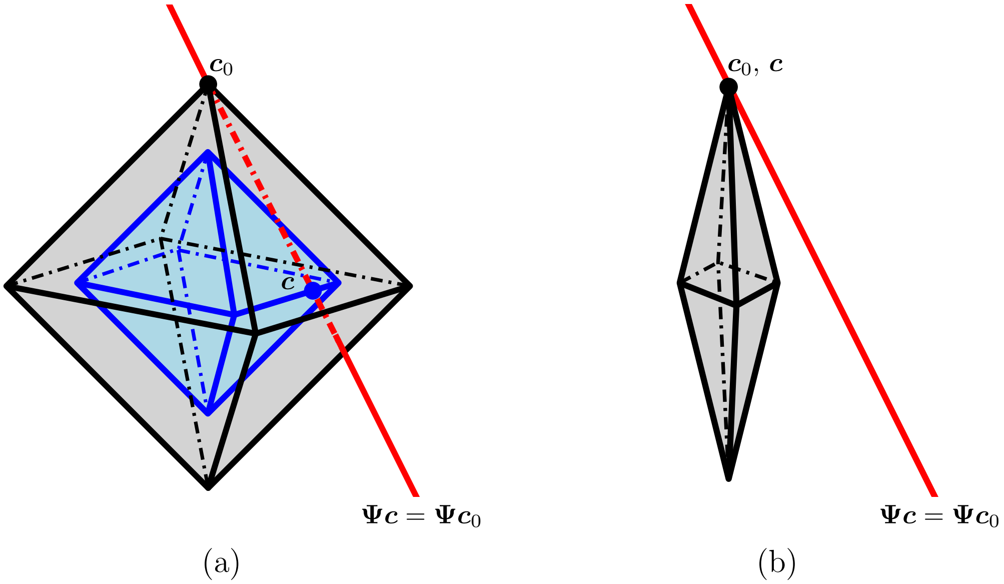

$\boldsymbol{c}_0$
$\boldsymbol{c}$
$\boldsymbol{\Psi c} = \boldsymbol{\Psi c}_0$
(a)
$\boldsymbol{c}_0$, $\boldsymbol{c}$
$\boldsymbol{\Psi c} = \boldsymbol{\Psi c}_0$
(b)

Figure 3.1: Schematic of approximation of a sparse $\boldsymbol{c}_0 \in \mathbb{R}^3$ via standard and weighted ℓ_1-minimization (based on [1]). (a) Standard ℓ_1-minimization where, depending on $\boldsymbol{\Psi}$, the problem $\mathcal{P}_{1,0}$ with $\boldsymbol{u} = \boldsymbol{\Psi c}_0$ may have a solution $\boldsymbol{c}$ such that $\|\boldsymbol{c}\|_1 \leq \|\boldsymbol{c}_0\|_1$. (b) Weighted ℓ_1-minimization for which there is no $\boldsymbol{c}$ with $\|\boldsymbol{Wc}\|_1 \leq \|\boldsymbol{Wc}_0\|_1$.

To solve $\mathcal{P}_{1,\epsilon}^{(\boldsymbol{W})}$, the standard ℓ_1-minimization solvers may be used. Specifically, $\tilde{\boldsymbol{c}} = \boldsymbol{Wc}$

may be solved from $\mathcal{P}_{1,\epsilon}$ with the modified measurement matrix $\tilde{\boldsymbol{\Psi}} = \boldsymbol{\Psi}\boldsymbol{W}^{-1}$. We then set $\boldsymbol{c} = \boldsymbol{W}^{-1}\tilde{\boldsymbol{c}}$.

Theoretical recovery via weighted ℓ_1-minimization

In this section, we provide a theoretical condition that guarantees the recovery of weighted ℓ_1-minimization, based on the Legendre polynomials chaos expansions. Assuming that the solution to $\mathcal{P}_{0,\epsilon}$ is $\boldsymbol{c}_0$, whose support is $\mathcal{C}$, and that $s = |\mathcal{C}|$, we are interested in the condition determining when $\boldsymbol{c}_1$ accurately approximates $\boldsymbol{c}_0$, where $\boldsymbol{c}_1$ is the solution to $\mathcal{P}_{1,\epsilon}$. This is done in the context of the Restricted Isometry Constant (RIC) [89, 90], which is defined such that for a given vector, $\boldsymbol{x} \in \mathbb{R}^P$, with at most s non-zero entries,

$$(1-\delta_s)\|\boldsymbol{x}\|^2 \leq \frac{1}{N}\|\boldsymbol{\Psi}\boldsymbol{x}\|^2 \leq (1+\delta_s)\|\boldsymbol{x}\|^2. \tag{3.18}$$

Here δ_s is the RIC.

Lemma 3.3.1. *Let $\{\psi_j\}_{1\leq j\leq P}$ be a Legendre PC basis in d independent random variables $\boldsymbol{\Xi} = (\Xi_1, \ldots, \Xi_d)$ uniformly distributed over $[-1,1]^d$ and with a total degree less than or equal to p. Let the matrix $\boldsymbol{\Psi}$ with entries $\boldsymbol{\Psi}(i,j) = \psi_j(\boldsymbol{\xi}^{(i)})$ correspond to realizations of $\{\psi_j\}$ at $\boldsymbol{\xi}^{(i)}$ sampled independently from the measure of $\boldsymbol{\Xi}$. If*

$$N \geq C3^p\delta^{-2}s\log^3(s)\log(P), \tag{3.19}$$

then the RIC, δ_s, of $\frac{1}{\sqrt{N}}\boldsymbol{\Psi}$ satisfies $\delta_s \leq \delta$ with probability larger than $1 - P^{-\gamma\log^3(s)}$. Here, C and γ are constants independent of N, p, and d.

Theorem 3.3.1. *Let s be a sparsity such that $\delta_{3s} + 3\delta_{4s} < 2$. Then for any sparse solution, $\boldsymbol{c}_0$, supported on $\mathcal{C}$ with $|\mathcal{C}| \leq s$, any solution $\boldsymbol{c}_1$ to $\mathcal{P}_{1,\epsilon}^{(\boldsymbol{W})}$ obeys*

$$\|\boldsymbol{c}_0 - \boldsymbol{c}_1\|_2 \leq C \cdot \epsilon,$$

where the constant C depends on s, $\max_{j\in\mathcal{C}} w_j$, and $\min_{j\in\mathcal{C}^c} w_j$.

Here $\mathcal{C}^c$ denotes the complement set of $\mathcal{C}$. We note that the proofs of Lemma 3.3.1 and Theorem 3.3.1 are presented in [112], and invite interested readers to consult for more detail.

3.3.2 Bi-fidelity OMP

Similar to bi-fidelity ℓ_1-minimization, we also consider applying low-fidelity PCE coefficients as the *a priori* information in OMP, where the low-fidelity PCE coefficients are calculated via OMP with low-fidelity realizations. In order to employ this information, we modify the OMP algorithm in the sensing part. In detail, we change the strategy how OMP selects the active column set: in the OMP, the active columns are selected totally depending on the residual reduction, and we modify the sensing process such that the *a priori* information about the PCE coefficients helps in selecting the active column set. Let vector $\boldsymbol{w} := (w_1, w_2, \ldots, w_P)^T$, where $w_j = |c^l_j|$, denoting the *a priori* information on the PCE coefficients, and we modify the OMP algorithm with $\boldsymbol{w}$ as the following exhibit shows:

Algorithm 4 OMP with *a priori* on PCE coefficients

```
Set $t = 0$, $\boldsymbol{c}^{(0)} = \boldsymbol{0}$, and $\boldsymbol{r}^{(0)} = \boldsymbol{u} - \boldsymbol{\Psi}\boldsymbol{c}^{(0)}$.
Set the initial solution support index set $\mathcal{I}^{(0)} = \emptyset$.
While $t \leq t_{max}$: do
    for all $j \notin \mathcal{I}^{(t)}$: do
        Evaluate $\varepsilon(j) = \left|\boldsymbol{\psi}_j^T \boldsymbol{r}^{(t)} \left(\|\boldsymbol{r}^{(t)}\|_2 - \lambda w_j\right)\right|$.
    End for
    Set $t = t + 1$.
    Update the support index set $\mathcal{I}^{(t)} = \mathcal{I}^{(t-1)} \cup \{\arg\max_j \epsilon(j)\}$.
    Solve for $\boldsymbol{c}^{(t)} = \arg\min_{\boldsymbol{c}} \|\boldsymbol{u} - \boldsymbol{\Psi}\boldsymbol{c}\|_2$ subject to $Support\{\boldsymbol{c}\} = \mathcal{I}^{(t)}$.
    Update the residual $\boldsymbol{r}^{(t)} = \boldsymbol{u} - \boldsymbol{\Psi}\boldsymbol{c}^{(t)}$.
End while
Output the solution $\boldsymbol{c} = \boldsymbol{c}^{(t)}$.
```

In the early iterations of Algorithm 4, the ℓ_2-norm of the residual, $\|\boldsymbol{r}^{(t)}\|_2$, is large, so for early selections of the active columns, $\boldsymbol{w}$ is less important. As the iteration develops, the residual is reduced, and columns are selected less by the reduction to the residual yet more by the *a priori* information. As the residual becomes very small, the choice will eventually depend less on the reduction in residual. The parameter λ indicates how trustable the *a priori* information, $\boldsymbol{w}$, is. A larger λ shows more reliability on $\boldsymbol{w}$, and the algorithm turns to $\boldsymbol{w}$ for the active column selection earlier. In this work, an appropriate λ is chosen via cross validation technique [47].

We note that in Algorithm 4, the stopping criterion is changed to $t \leq t_{max}$. This is because that the active column set in bi-fidelity OMP is not purely determined by the residual decrements, and thus a stopping criterion completely depending the residual may result in an considerable bias in the PCE basis functions selected. To avoid this, we set a maximum iteration number t_{max}. In practice, an optimal t_{max} may also be calculated via cross validation. In our numerical experiments, we set $t_{max} = N/2$.

3.4 Numerical Results

3.4.1 Case I: Two-dimensional elliptic PDE with random inputs

We first consider a two-dimensional elliptic PDE with stochastic coefficient:

$$\begin{aligned} &-\nabla \cdot (a(\boldsymbol{x}, \Xi)\nabla u(\boldsymbol{x}, \Xi)) = 1, && \boldsymbol{x} \in \mathcal{D} = (0,1) \times (0,1) \ , \\ &u(\boldsymbol{x}, \Xi) = 0, && \boldsymbol{x} \in \partial\mathcal{D} \ , \end{aligned} \tag{3.20}$$

where the uncertain diffusion coefficient $a(\boldsymbol{x}, \Xi)$ is modeled as

$$a(\boldsymbol{x}, \Xi) = \bar{a} + \sigma_a \sum_{k=1}^{d} \sqrt{\lambda_k}\phi_k(\boldsymbol{x})\Xi_k \ , \tag{3.21}$$

in which $d = 20$, and $\Xi_k, \ k = 1, \ldots, d$ independently yield to uniform distribution, $U(-1, 1)$. In addition, $\bar{a} = 0.1$, $\sigma_a = 0.05$, and $\{\phi_k\}_{k=1}^{d}$ are the eigenfunctions corresponding to the d largest eigenvalues $\{\lambda_k\}_{k=1}^{d}$ of the Gaussian covariance kernel

$$C_{aa}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp\left[-\frac{(x_1 - x_2)^2}{l_c^2} - \frac{(y_1 - y_2)^2}{l_c^2}\right] \ , \tag{3.22}$$

whose correlation length $l_c = 1/16$.

The QoI $u(\Xi)$ is chosen as the flux at boundary $x = 1$, i.e.,

$$u(\Xi) = \int_0^1 a\left((1, y), \Xi\right) \left.\frac{du\left((x, y), \Xi\right)}{dx}\right|_{x=1} dy. \tag{3.23}$$

Results

We first solve for $u(\boldsymbol{\xi}^{(i)})$ on a coarse 16×16 uniform FEM mesh, which are considered as the low-fidelity realizations. The high-fidelity realizations are performed on a 256×256 mesh, which yields accurate resolutions of the solution. The residual error from the numerical solver and computational time (in *seconds*) are shown in Figure 3.2.

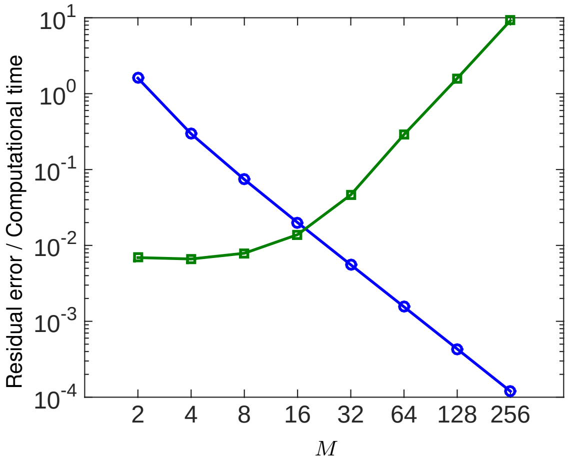
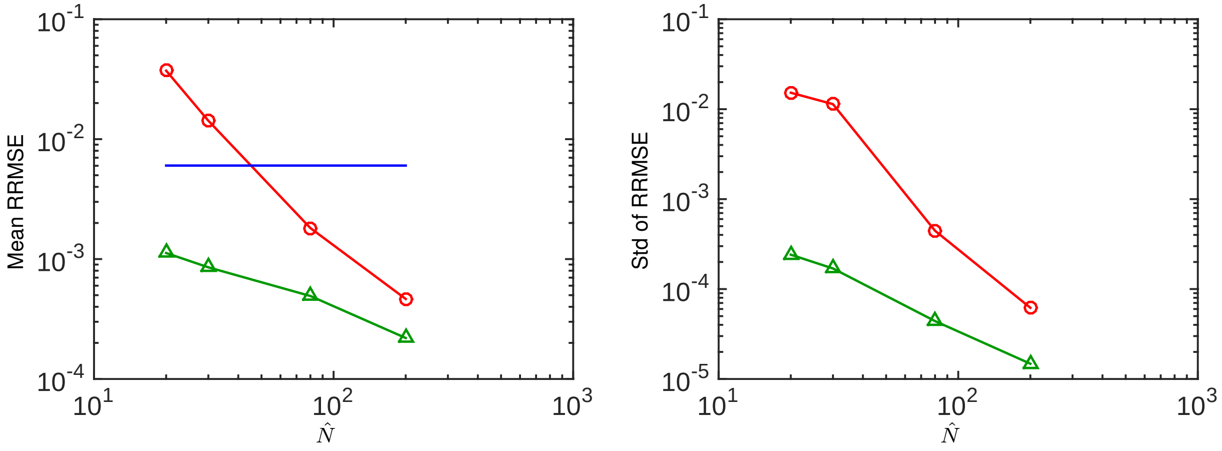


Figure 3.2: Relative residual error and computation time (*seconds*) on $M \times M$ meshes (—□— computation time, —○— relative residual error).

In this experiment, we seek to approximate $u(\Xi)$ by Legendre PCE with $P = 2500$ polynomial basis functions of total order $p \leq 4$. To compare the bi-fidelity approaches with standard methods, we define an equivalent sample size $\hat{N} = N_h + \nu N_l$, where N_h and N_l are the numbers of high-fidelity and low-fidelity realizations, respectively, and ν is non-negative number determined by the computational cost in a single low-fidelity simulation. In Fig. 3.2, it shows that the computational cost of a low-fidelity simulation is approximately 10^{-3} times that of a high-fidelity simulation, therefore, $\nu = 10^{-3}$. Hereby, $\hat{N} = N_h$ for high-fidelity PCE, while $\hat{N} = N_h + 10^{-3} N_l$ for bi-fidelity PCE. In addition, it is assumed that the computational cost on solving $\mathcal{P}_{0,\epsilon}$, $\mathcal{P}_{1,\epsilon}$, and $\mathcal{P}_{1,\epsilon}^{(\boldsymbol{W})}$ is negligible compared to the cost of deterministic solver.

In Figure 3.3, we compare the mean and standard deviation of the relative root mean square error (RRMSE) in reconstructing $u(\boldsymbol{\Xi})$ via ℓ_1-minimization, using 100 independent replications, with $\hat{N} \in \{20, 30, 80, 200\}$. The reference PCE coefficients are computed from high-fidelity solutions by least squares regression [31] with 10000 independent samples. It can be observed that although we use a large number of realizations $N_l = 400$ in low fidelity to reconstruct $u(\boldsymbol{\Xi})$, due to the significant simulation error, RRMSE in the reconstruction via ℓ_1-minimization is relatively high. However, when we use the PCE coefficients from low-fidelity simulations as the *a priori* information in bi-fidelity ℓ_1-minimization, with various $\hat{N}$, bi-fidelity ℓ_1-minimization outperforms the standard ℓ_1-minimization in high fidelity. We note that in bi-fidelity ℓ_1-minimization presented in Figure 3.3, the number of low-fidelity realizations is fixed as $N_l = 400$, we only change N_h in $\hat{N}$.

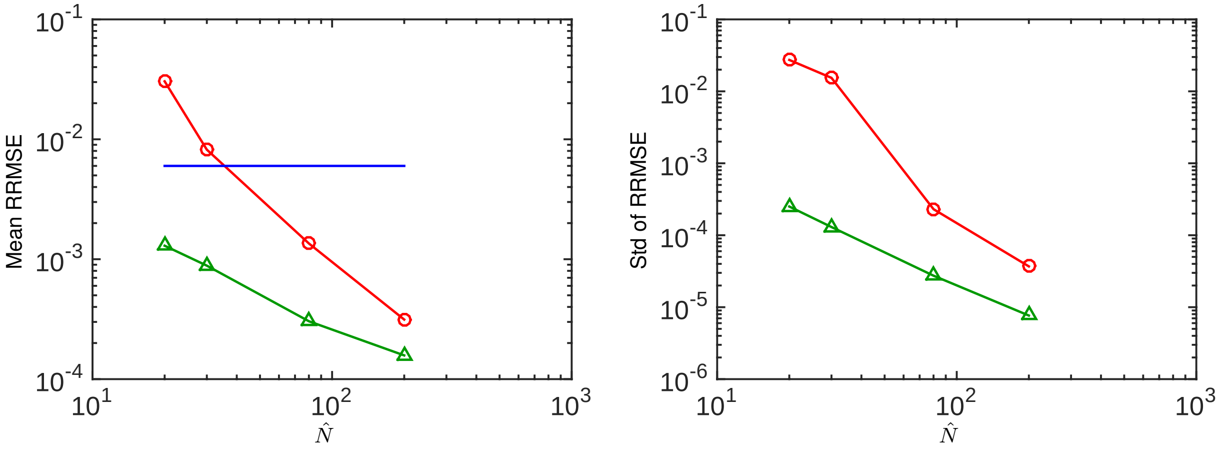


(a)

(b)

Figure 3.3: Comparison of the statistics of the RRMSE in reconstructing $u(\boldsymbol{\Xi})$ via ℓ_1-minimization. (a) Mean of RRMSE. (b) Standard deviation of RRMSE. (⊖ high-fidelity (256 × 256 mesh), △ bi-fidelity (16 × 16 and 256 × 256 mesh), — low-fidelity (16 × 16 mesh), with 400 realizations.)

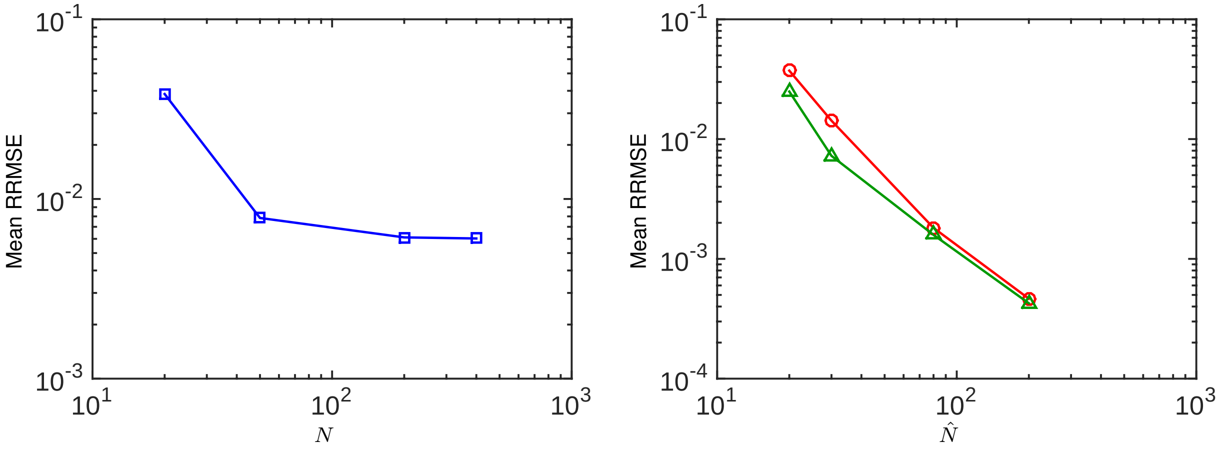
In Figure 3.4, we compare the mean and standard deviation of the RRMSE in reconstructing $u(\Xi)$ via OMP, using 100 independent replications, with $\hat{N} \in \{20, 30, 80, 200\}$. Similar to the bi-fidelity ℓ_1-minimization, the bi-fidelity OMP outperforms standard OMP in high fidelity in approximating $u(\Xi)$. Similar to bi-fidelity ℓ_1-minimization, in bi-fidelity OMP presented in Figure 3.4, the number of low-fidelity realizations is fixed as $N_l = 400$, we only change N_h in $\hat{N}$.



(a) (b)

Figure 3.4: Comparison of the statistics of the RRMSE in reconstructing $u(\Xi)$ via OMP. (a) Mean of RRMSE. (b) Standard deviation of RRMSE. (−○− high-fidelity (256×256 mesh), −△− bi-fidelity (16×16 and 256×256 mesh), — low-fidelity (16×16 mesh), with 400 realizations.)

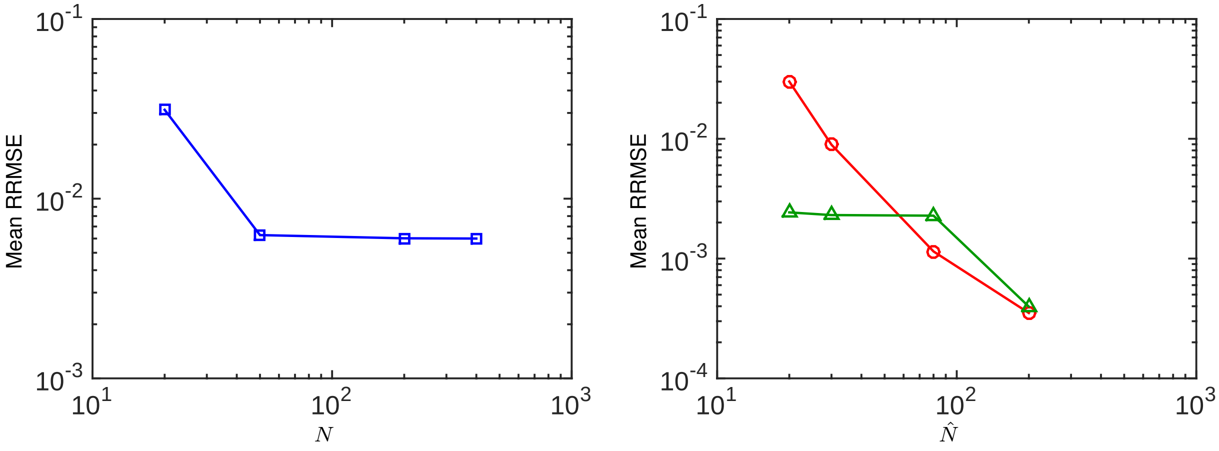
To study how the accuracy of the low-fidelity PCE coefficients affects the performance of bi-fidelity approaches, we repeat the same experiment with $N_l = 20$. According to Figure 3.5a and Figure 3.6a, the coefficients $\boldsymbol{c}^l$ calculated from these realizations are noticeably less accurate compared to the case with $N_l = 400$. We not that in both ℓ_1-minimization and OMP cases, the mean RRMSE stopped decreasing when $N \geq 50$, which is resulted from the numerical errors in low-fidelity simulation.



(a)

(b)

Figure 3.5: RRMSE in reconstructing $u(\mathbf{\Xi})$ via ℓ_1-minimization. (a) Mean RRMSE using 100 independent replications in low fidelity (16×16 mesh). (b) Comparison of mean RRMSE. (─○─ high-fidelity (256×256 mesh), ─△─ bi-fidelity (16×16 and 256×256 mesh) with $N_l = 20$).



(a)

(b)

Figure 3.6: RRMSE in reconstructing $u(\mathbf{\Xi})$ via OMP. (a) Mean RRMSE using 100 independent replications in low fidelity (16×16 mesh). (b) Comparison of mean RRMSE. (─○─ high-fidelity (256×256 mesh), ─△─ bi-fidelity (16×16 and 256×256 mesh) with $N_l = 20$).

From Figure 3.5b and Figure 3.6b, we can observe that with a less accurate $\boldsymbol{c}^l$, the accuracy in reconstructing the QoI via bi-fidelity methods decreases. This suggests that in practice N_l should be appropriately chosen depending on the desired accuracy of bi-fidelity approaches and the computational cost of low-fidelity simulations. We note that the accuracy of $\boldsymbol{c}^l$ may be enhanced by either increasing the fidelity of low-fidelity simulation or increasing the number of low-fidelity realizations. For any given problems, the strategy should be determined wisely, such that the computational cost on the deterministic simulations is minimized.

3.4.2 Case II: Steady-state thermally driven flow with stochastic boundary temperature

We next consider a 2-D heat driven cavity flow problem in steady-state, shown in Figure 3.7a. The left vertical wall has a deterministic, constant temperature $\tilde{T}_h$, referred to as the hot wall, while the right vertical wall has a stochastic temperature $\tilde{T}_c < \tilde{T}_h$ with constant mean $\bar{\tilde{T}}_c$, referred to as the cold wall. Both top and bottom walls are assumed to be adiabatic. The reference temperature and the reference temperature difference are defined as $\tilde{T}_{ref} = (\tilde{T}_h + \bar{\tilde{T}}_c)/2$ and $\Delta\tilde{T}_{ref} = \tilde{T}_h - \bar{\tilde{T}}_c$, respectively. In dimensionless variables, the governing equations (in the small temperature difference regime, i.e., Boussinesq approximation) are given by

$$\begin{aligned}
&\boldsymbol{u} \cdot \nabla \boldsymbol{u} = -\nabla p + \frac{\mathrm{Pr}}{\sqrt{\mathrm{Ra}}} \nabla^2 \boldsymbol{u} + \mathrm{Pr} T \hat{\boldsymbol{y}}, \\
&\nabla \cdot \boldsymbol{u} = 0, \\
&\nabla \cdot (\boldsymbol{u} T) = \frac{1}{\sqrt{\mathrm{Ra}}} \nabla^2 T,
\end{aligned} \tag{3.24}$$

where $\hat{\boldsymbol{y}}$ is the unit vector $(0,1)$, $\boldsymbol{u} = (u, v)$ is velocity vector field, $T = (\tilde{T} - \tilde{T}_{ref})/\Delta\tilde{T}_{ref}$ is normalized temperature ($\tilde{T}$ denotes non-dimensional temperature), p is pressure, and t is time. Non-dimensional Prandtl and Rayleigh numbers are defined, respectively, as $\mathrm{Pr} = \tilde{\mu}\tilde{c}_p/\tilde{\kappa}$ and $\mathrm{Ra} = \tilde{\rho} g \beta \Delta\tilde{T}_{ref} \tilde{L}^3/(\tilde{\mu}\tilde{\kappa})$, where the superscript tilde (˜) denotes the non-dimensional quantities. Specifically, $\tilde{\rho}$ is density, $\tilde{L}$ is reference length, g is gravitational acceleration, $\tilde{\mu}$ is molecular viscosity, $\tilde{\kappa}$ is thermal diffusivity, and the coefficient of thermal expansion is given by β. In this example, the Prandtl and Rayleigh numbers are set to $\mathrm{Pr} = 0.71$ and $\mathrm{Ra} = 10^6$, respectively. For more details on the non-dimensional variables in (3.24), we refer the interested reader to [108, 6, 3].

On the cold wall, we apply a (normalized) temperature distribution with stochastic fluctuations of the form

$$\begin{aligned} T_c(x=1, y, \Xi) &= \bar{T}_c + T_c', \\ T_c' &= \sigma_T \sum_{i=1}^{d} \sqrt{\lambda_i} \varphi_i(y) \Xi_i, \end{aligned} \tag{3.25}$$

where $\bar{T}_c$ is a constant mean temperature. In (3.25), Ξ_i, $i = 1, \ldots, d$, are independent random variables uniformly distributed on $[-1, 1]$. $\{\lambda_i\}_{i=1}^{d}$ and $\{\varphi_i(y)\}_{i=1}^{d}$ are the d largest eigenvalues and the corresponding eigenfunctions of the exponential covariance kernel

$$C_{T_c T_c}(y_1, y_2) = \exp\left(-\frac{|y_1 - y_2|}{l_c}\right),$$

where l_c is the correlation length. Following [109], the eigenpairs $(\lambda_i, \varphi_i(y))$ in (3.25) are, respectively, given by

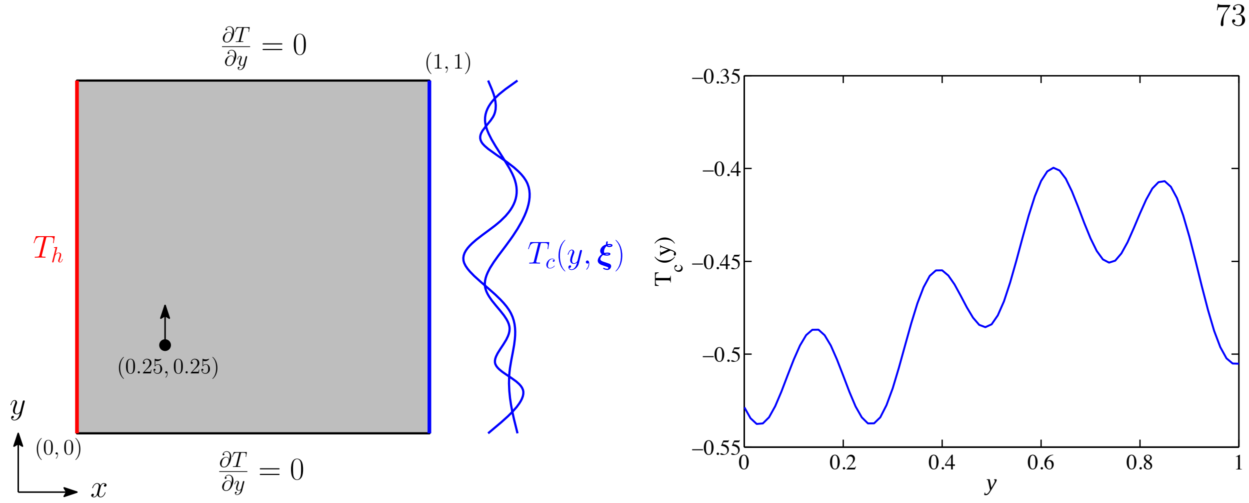
$$\lambda_i = \frac{2 l_c}{l_c^2 \omega_i^2 + 1},$$

and

$$\varphi_i(y) = \begin{cases} \dfrac{\cos(\omega_i y)}{\sqrt{0.5 + \frac{\sin(\omega_i)}{2\omega_i}}}, & i \text{ is odd,} \\ \dfrac{\sin(\omega_i y)}{\sqrt{0.5 - \frac{\sin(\omega_i)}{2\omega_i}}}, & i \text{ is even,} \end{cases}$$

where each ω_i is a root of

$$\omega_i + (1/l_c)\tan(0.5\omega_i) = 0.$$



(a) Schematic of the geometry and boundary conditions.

(b) A realization of $T_c(x = 1, y)$.

Figure 3.7: Illustration of the cavity flow problem.

In our numerical test we let $(T_h, \bar{T}_c) = (0.5, -0.5)$, $d = 20$, $l_c = 1/21$, and $\sigma_T = 7/25$. A realization of the cold wall temperature T_c is shown in Figure 3.7b. Our quantity of interest, the Nusselt number define by

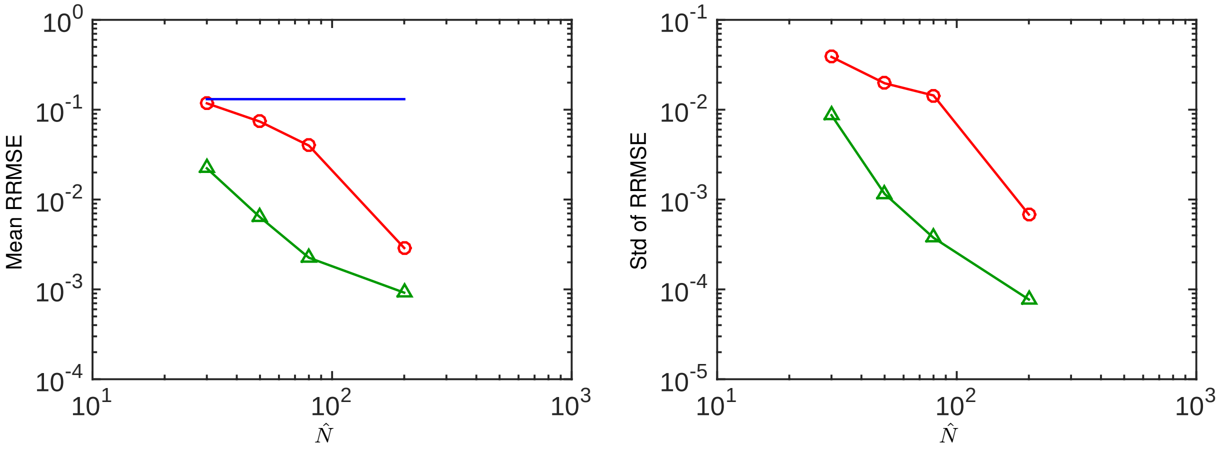
$$\mathrm{Nu} := -\int_0^1 \left.\frac{\partial T(x,y)}{\partial x}\right|_{x=1} dy, \tag{3.26}$$

is expanded in the Legendre PCE basis of total degree $p = 4$ with only the first $P = 2500$ basis functions retained, as described in the case of the elliptic problem. We seek to accurately reconstruct Nu with $N < P$ random samples of $\boldsymbol{\Xi}$ and the corresponding realizations of Nu.

Results

In this experiment, the low-fidelity simulations are performed on a uniform 16×16 FEM mesh, while 64×64 for high-fidelity mesh, with which we observe $\nu = 1/90$. Therefore, in this experiment, $\hat{N} = N_h + 1/15 N_l$ in bi-fidelity approaches, where N_l is fixed to be 400.

In Figure 3.8, we compare the mean and standard deviation of the relative root mean square error (RRMSE) in reconstructing Nu via ℓ_1-minimization, using 100 independent replications, with $\hat{N} \in \{30, 50, 80, 200\}$. The reference PCE coefficients are computed from high-fidelity solutions by least squares regression with 10000 independent samples. It can be observed that low-fidelity PCE leads to significant errors in reconstructing Nu, due to the numerical errors from the coarse mesh in the deterministic solver. However, when we use the PCE coefficients from low-fidelity simulations as the *a priori* information in bi-fidelity ℓ_1-minimization, with various $\hat{N}$, bi-fidelity ℓ_1-minimization outperforms the standard ℓ_1-minimization in high fidelity.

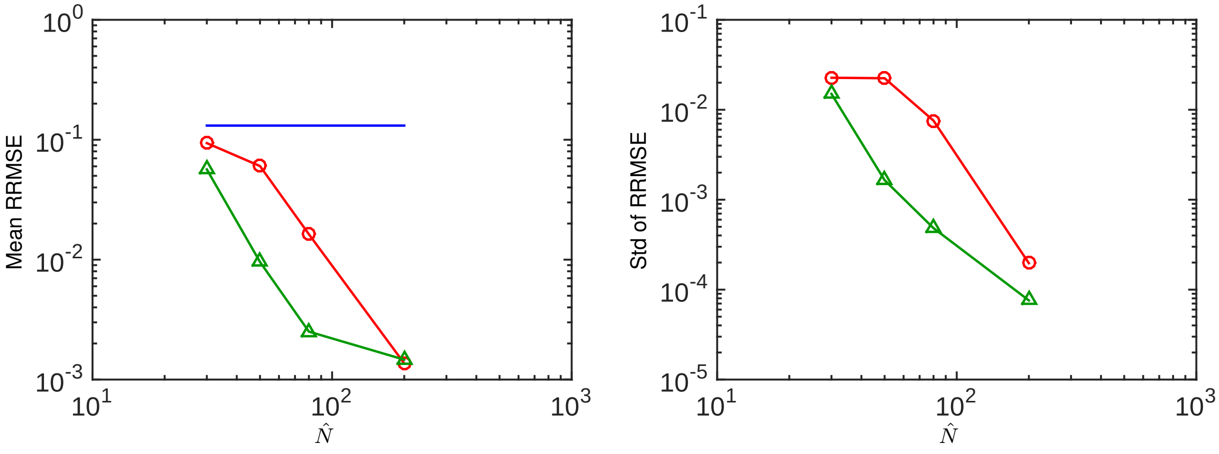


(a)

(b)

Figure 3.8: Comparison of the statistics of the RRMSE in reconstructing Nu via ℓ_1-minimization. (a) Mean of RRMSE. (b) Standard deviation of RRMSE. (–○– high-fidelity (64 × 64 mesh), –△– bi-fidelity (16 × 16 and 64 × 64 mesh), — low-fidelity (16 × 16 mesh), with 400 realizations.)

In Figure 3.9, we show the comparison in reconstructing Nu via OMP, using 100 independent replications, with $\hat{N} \in \{30, 50, 80, 200\}$. Similar to bi-fidelity ℓ_1-minimization, with various $\hat{N}$, bi-fidelity OMP outperforms the standard OMP in high fidelity.



(a)

(b)

Figure 3.9: Comparison of the statistics of the RRMSE in reconstructing Nu via OMP. (a) Mean of RRMSE. (b) Standard deviation of RRMSE. (○ high-fidelity (64 × 64 mesh), △ bi-fidelity (16 × 16 and 64 × 64 mesh), — low-fidelity (16 × 16 mesh), with 400 realizations.)

3.5 Conclusion

In this manuscript, we utilized bi-fidelity technique to provide *a priori* information about the polynomial chaos expansion (PCE) coefficients on the quantity of interest (QoI), within the context of compressive sampling. Furthermore, employed a weighted ℓ_1-minimization and modified the orthogonal matching pursuit (OMP) algorithm to include this *a priori*, therefore to improve the accuracy in approximating the QoI by PCE. In addition, we provide analysis on weighted ℓ_1-minimization, when the *a priori* is inaccurate, for Legendre polynomial chaos.

Numerical examples were shown to demonstrate the bi-fidelity methods: a two-dimensional elliptic equation with stochastic diffusion coefficient; a steady state thermally driven cavity flow with random temperature boundary condition. In all examples, bi-fidelity methods were consistently observed to improve the quality of solution recovery at the same computational cost. As the solutions recovered by bi-fidelity approaches are sensitive to the accuracy of the *a priori* information, which was observed in the first example, hence a wise strategy in determining the accuracy and number of realizations of low-fidelity simulation was suggested with concerns about the computational cost.

In addition, although all the examples are presented with a bi-fidelity enhancement, we note that the proposed approaches are not limited within bi-fidelity. One can easily employ the bi-fidelity approaches recursively with multiple levels of fidelity, depending on the availability of the low-cost deterministic solvers in each fidelity level.

CHAPTER 4

ON POLYNOMIAL CHAOS EXPANSION VIA GRADIENT-ENHANCED ℓ_1-MINIMIZATION[1]

Abstract

Gradient-enhanced Uncertainty Quantification (UQ) has received recent attention, in which the derivatives of a Quantity of Interest (QoI) with respect to the uncertain parameters are utilized to improve the surrogate approximation. Polynomial chaos expansions (PCEs) are often employed in UQ, and when the QoI can be represented by a sparse PCE, ℓ_1-minimization can identify the PCE coefficients with a relatively small number of samples. In this chapter, we investigate a gradient-enhanced ℓ_1-minimization, where derivative information is computed to accelerate the identification of the PCE coefficients. For this approach, stability and convergence analysis are lacking, and thus we address these here with a probabilistic result. In particular, with an appropriate normalization, we show the inclusion of derivative information will almost-surely lead to improved conditions, e.g. related to the null-space and coherence of the measurement matrix, for a successful solution recovery. Further, we demonstrate our analysis empirically via three numerical examples: a manufactured PCE, an elliptic partial differential equation with random inputs, and a plane Poiseuille flow with random boundaries. These examples all suggest that including derivative information admits solution recovery at reduced computational cost.

[1] This chapter has been submitted by *J. Peng et al.* to *Journal of Computational Physics.*

4.1 Introduction

In complex engineering system analysis, inherent variability in inputs and imperfect knowledge of the physics can lead to unfounded confidence or unnecessary diffidence in understanding Quantities of Interest (QoI). Uncertainty Quantification (UQ) [2, 3, 4] as a study aims to develop numerical tools that can accurately predict QoI and facilitate the quantitative validation of the simulation model.

To characterize uncertainty, probability is a natural framework. We model the uncertain inputs as a $d-$dimensional vector of independent random variables $\Xi := (\Xi_1, \dots, \Xi_d)$, with probability density function $\rho(\Xi)$. The QoI that we seek to approximate is denoted by a scalar $u(\Xi)$. Here we utilize polynomial chaos expansions (PCEs) [2, 20] to approximate $u(\Xi)$, assumed to have finite variance. In this case, $u(\Xi)$ can be represented as an expansion in multivariate orthogonal polynomials $\psi_j(\Xi)$, i.e.,

$$u(\Xi) = \sum_{j=1}^{\infty} c_j \psi_j(\Xi) \approx \sum_{j=1}^{P} c_j \psi_j(\Xi) + \epsilon_t(\Xi), \tag{4.1}$$

where $c_j, j = 1, 2, \dots,$ are the corresponding PCE coefficients, and ϵ_t is the truncation error associated with retaining P terms of a sorted basis. The PCE coefficients can be computed by the projection

$$c_j = \int u(\Xi)\psi_j(\Xi)\rho(\Xi)d\Xi = \mathbb{E}\left[u(\Xi)\psi_j(\Xi)\right], \tag{4.2}$$

where the operator $\mathbb{E}$ denotes the mathematical expectation. Here we assume that $\psi_j(\Xi)$ are normalized such that $\mathbb{E}\left[\psi_j^2(\Xi)\right] = 1$.

Typically, a full P term approximation is not necessary, and we can restrict to an unknown subset $\mathcal{C} \subset \{1, \dots, P\}$ such that $|\mathcal{C}|$, the number of elements in $\mathcal{C}$, is significantly smaller than P. In addition, $|\mathcal{C}|$ is referred to as the sparsity of the approximation. We approximate $u(\Xi)$ in (4.1) then by

$$u(\Xi) \approx \sum_{j\in\mathcal{C}} c_j \psi_j(\Xi). \tag{4.3}$$

This concept of using one basis, indexed by $\{1, \ldots, P\}$, to compute an approximation in a small but unknown subset of those basis functions, indexed by $\mathcal{C}$, falls within the context of compressive sampling [56, 34, 47, 112, 117].

To identify PCE coefficients, we consider non-intrusive sampling methods, in which deterministic solvers for the QoI are not modified. Such methods include Monte Carlo simulation [23, 3], pseudo-spectral stochastic collocation [24, 25, 3, 27], least squares regression [30, 118, 64], and ℓ_1-minimization [48, 47, 52, 53, 112, 117, 119, 120]. In this work, we adopt ℓ_1-minimization to estimate the coefficients, solving the problem

$$\arg\min_{\boldsymbol{c}} \|\boldsymbol{c}\|_1 \quad \text{subject to} \quad \|\boldsymbol{u} - \boldsymbol{\Psi}\boldsymbol{c}\|_2 \leq \delta, \tag{4.4}$$

where the vector $\boldsymbol{c} := (c_1, \ldots, c_P)$ contains PCE coefficients, while the vector $\boldsymbol{u}$ and the so-called *measurement matrix* $\boldsymbol{\Psi}$ contains function evaluations at realizations of the random input, $\boldsymbol{\Xi}$. Specifically, denoting the ith realization of $\boldsymbol{\Xi}$ as $\boldsymbol{\xi}^{(i)}$,

$$\boldsymbol{u} := \left(u(\boldsymbol{\xi}^{(1)}), \ldots, u(\boldsymbol{\xi}^{(N)})\right)^T; \tag{4.5}$$

$$\boldsymbol{\Psi}(i,j) := \psi_j\left(\boldsymbol{\xi}^{(i)}\right). \tag{4.6}$$

In (4.4), δ is a tolerance parameter necessitated by the truncation error, $\epsilon_t(\boldsymbol{\Xi})$, to reduce the effect of overfitting. For this work, δ is identified via cross-validation [121].

It has been shown that as the number of samples, N, increases, the probability of accurately recovering PCE coefficients experiences a corresponding increase [117]. Often the deterministic solver of u is computationally demanding, and it is expensive to compute realized $u(\boldsymbol{\xi}^{(i)})$. In order to deal with this situation, coefficient estimation based on gradient-enhanced PCE has received recent attention [122, 123, 124, 11, 125, 120]. Specifically, the gradient information utilized is

$$\boldsymbol{u}_\partial := \left(\frac{\partial u}{\partial \Xi_1}\left(\boldsymbol{\xi}^{(1)}\right), \ldots, \frac{\partial u}{\partial \Xi_d}\left(\boldsymbol{\xi}^{(1)}\right), \ldots, \frac{\partial u}{\partial \Xi_1}\left(\boldsymbol{\xi}^{(N)}\right), \ldots, \frac{\partial u}{\partial \Xi_d}\left(\boldsymbol{\xi}^{(N)}\right)\right)^T \tag{4.7}$$

and

$$\boldsymbol{\Psi}_\partial((i-1)\cdot d + k, j) := \frac{\partial \psi_j}{\partial \Xi_k}(\boldsymbol{\xi}^{(i)}), \quad k = 1, \ldots, d. \tag{4.8}$$

With this gradient-enhancement, $\boldsymbol{c}$ is solved by minimizing $\|\boldsymbol{c}\|_2$, in least squares regression, or $\|\boldsymbol{c}\|_1$, in compressive sampling, subject to a constraint on

$$\left\| \begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{u}_\partial \end{pmatrix} - \begin{pmatrix} \boldsymbol{\Psi} \\ \boldsymbol{\Psi}_\partial \end{pmatrix} \boldsymbol{W}\boldsymbol{c} \right\|_2 , \tag{4.9}$$

where $\boldsymbol{W}$ is a positive-diagonal matrix depending on the basis functions and their derivatives, which we shall specify in Section 4.3.2. In this way, N realizations of the random inputs Ξ provide an $N(d+1) \times P$ matrix, where the gradients $\partial u / \partial \Xi_k,\ k = 1, \dots, d$, may be computed at each evaluation of Ξ from, e.g., direct or adjoint sensitivity equations [126], or automatic differentiation [127]. We assume that $u(\Xi)$ and its partial derivatives are square integrable with respect to the measure for Ξ, and are represented in the appropriate basis such that the QoI and its derivatives correspond to identical coefficients.

In a related context, a gradient-enhanced sparse approximation based on ℓ_1-minimization was proposed in [128], in which derivatives $\partial u / \partial \Xi_k,\ k = 1, \dots, d$, are projected onto Legendre PCE, with a corresponding coefficient vector $\boldsymbol{c}_\partial$.

4.1.1 Contribution of this work

In this work, we investigate a gradient-enhanced ℓ_1-minimization approach as seen in [120], in which the derivative information is empirically shown to improve the resolution of PCE coefficients.

In Section 4.3, a theoretical contribution concerning the Restricted Isometry Constant (RIC) is presented regarding recovery in ℓ_1-minimization with derivative information based on Hermite PCE. The RIC is a constant associated with the measurement matrix that provides fruitful (probabilistic) bounds for recovery of solutions via ℓ_1-minimization [103]. Additionally, under an appropriate normalization introduced by $\boldsymbol{W}$ in (4.9), it is guaranteed that null-space, measurement matrix column inner-products, and coherence related measures are almost-surely improved, implying that stability for solutions computed by (4.4) are not reduced by including derivative information. These analyses are original to the authors' best knowledge, and provide a framework for analysis of recovery in other PCE bases. Also, though not considered here, the approach may be extended to least squares polynomial chaos regression [64].

Three numerical experiments are used to demonstrate the gradient-enhanced ℓ_1-minimization in Section 4.4, among which a plane Poiseuille flow with random boundaries is simulated, and the derivative information is approximated via an adjoint sensitivity method. The empirical results agree with the theoretical analysis, and suggest that the inclusion of derivative information can improve solution recovery at a lower overall computational cost.

The structure of the manuscript is as follows. In Section 4.2, we state our problem and introduce the formulation of the gradient-enhanced ℓ_1-minimization approach. In Section 4.3, we present theoretical results concerning the stability and convergence of the gradient-enhanced ℓ_1-minimization approach, in the context of Hermite PCE. In Section 4.4, we demonstrate our analysis empirically via three numerical examples: a manufactured PCE, an elliptic partial differential equation with random inputs, and a plane Poiseuille flow with random boundaries. Section 4.5 presents the proofs to the results in Section 4.3.

4.2 Method Synopsis

4.2.1 Problem statement

We use differential equations to model engineering systems on a domain $\mathcal{D} \in \mathbb{R}^D$, $D \in \{1,2,3\}$, in which the uncertainty sources characterized by Ξ may be represented in one or many relevant parameters, e.g., boundary conditions and/or initial conditions. The solution u is governed by the equations

$$\begin{aligned}
\mathcal{L}(\boldsymbol{x}, t, \Xi; u(\boldsymbol{x}, t, \Xi)) = 0, &\qquad \boldsymbol{x} \in \mathcal{D}, \\
\mathcal{I}(\boldsymbol{x}, \Xi; u(\boldsymbol{x}, 0, \Xi)) = 0, &\qquad \boldsymbol{x} \in \mathcal{D}, \\
\mathcal{B}(\boldsymbol{x}, t, \Xi; u(\boldsymbol{x}, t, \Xi)) = 0, &\qquad \boldsymbol{x} \in \partial\mathcal{D},
\end{aligned} \tag{4.10}$$

where $\mathcal{L}$, $\mathcal{I}$, and $\mathcal{B}$ are differential operators depending on the physics of the problem, the initial conditions, and the boundary conditions, respectively. Our objective is to approximate the QoI, here $u(\boldsymbol{x}_0, t_0, \Xi)$, for some fixed spatial location $\boldsymbol{x}_0$ and time t_0. Since we denote the realizations of the random inputs by $\boldsymbol{\xi}^{(i)}$, the corresponding output is $u(\boldsymbol{x}_0, t_0, \boldsymbol{\xi}^{(i)})$. To reduce notation, we drop the reference to $\boldsymbol{x}_0$ and t_0, and simply write $u(\Xi)$ and $u(\boldsymbol{\xi}^{(i)})$.

4.2.2 Polynomial chaos expansion (PCE)

We rely on the PCE (4.1) to approximate the QoI, $u(\Xi)$. For convenience, we assume that the input random variables, Ξ_k, are independent and identically distributed according to the probability density function ρ_k, and define $\{\psi_{i_k}(\Xi_k)\}$ to be the complete set of polynomials of degree $i_k \in \mathbb{N} \cup \{0\}$ orthogonal with respect to the weight function ρ_k [21, 20]. Hence, the multivariate orthonormal polynomials in Ξ are given by the products of the univariate orthonormal polynomials,

$$\psi_{\boldsymbol{i}}(\Xi) = \prod_{k=1}^{d} \psi_{i_k}(\Xi_k), \tag{4.11}$$

where each $\boldsymbol{i} \in \{(i_1, \ldots, i_d) : i_k \in \mathbb{N} \cup \{0\}\}$ is a d-dimensional multi-index of non-negative integers. For computation, we truncate the expansion in (4.1) to the set of P basis functions associated with the subspace of polynomials of total order not greater than p, that is $\sum_{k=1}^{d} i_k \le p$. For convenience, we also order these P basis functions so that they are indexed by $\{1, \ldots, P\}$, as in (4.1), where there should be no confusion in using either notation. The basis set $\{\psi_j(\Xi)\}_{j=1}^{P}$ has cardinality

$$P = \frac{(d+p)!}{d!p!}. \tag{4.12}$$

4.2.3 ℓ_1-minimization with gradient information

We generate derivative information for the QoI, denoted by $\boldsymbol{u}_\partial$, and correspondingly evaluate the derivatives of $\psi_j(\Xi)$, $j = 1, \ldots, P$, at the realizations $\boldsymbol{\xi}^{(i)}$, $i = 1, \ldots, N$, stored in a matrix $\boldsymbol{\Psi}_\partial$, as in (4.7) and (4.8). For brevity, we define

$$\tilde{\boldsymbol{u}} = \begin{pmatrix} \boldsymbol{u} \\ \\ \boldsymbol{u}_\partial \end{pmatrix} \tag{4.13}$$

and

$$\tilde{\boldsymbol{\Psi}} = \begin{pmatrix} \boldsymbol{\Psi} \\ \\ \boldsymbol{\Psi}_\partial \end{pmatrix} , \tag{4.14}$$

where $\tilde{\boldsymbol{u}} \in \mathbb{R}^{N(d+1)\times 1}$, and $\tilde{\boldsymbol{\Psi}} \in \mathbb{R}^{N(d+1)\times P}$ is referred to as the *gradient-enhanced measurement matrix*. Gradient-enhanced ℓ_1-minimization solves the problem

$$\arg\min_{\boldsymbol{c}} \|\boldsymbol{c}\|_1 \quad \text{subject to} \quad \left\|\tilde{\boldsymbol{u}} - \tilde{\boldsymbol{\Psi}}\boldsymbol{W}\boldsymbol{c}\right\|_2 \leq \delta, \tag{4.15}$$

where δ generally differs from the choice in (4.4) and $\tilde{\boldsymbol{\Psi}}$ is assumed to be normalized such that $\mathbb{E}\left[N^{-1}\tilde{\boldsymbol{\Psi}}^T\tilde{\boldsymbol{\Psi}}\right] = \boldsymbol{I}$, the $P \times P$ identity matrix. Here, $\boldsymbol{W}$ is a positive-diagonal matrix, whose definition is deferred until Section 4.3.2. We next begin a theoretical development to justify this approach.

4.3 Theoretical Discussion

We present results supporting the premise that the inclusion of derivative information does not reduce the stability of solutions recovered via (4.15) when compared to solutions recovered via (4.4), i.e., in the absence of derivative information. We refer to these solutions and the methods to attain them using the adjectives gradient-enhanced and standard, respectively. We perform our analysis here with Hermite polynomials as they possess the convenient property that they and their derivatives are orthogonal with respect to the same measure, as by (5.5.10) of [129]. While we consider the Probabilists' polynomials here for exposition, the Physicists' polynomials would produce analogous results. Though we do not consider the details here, this analysis may be extended to the case of Laguerre or Jacobi polynomials where the derivative polynomials form an orthogonal system with respect to a measure that differs from the orthogonality measure, yet is still explicitly known, as by (5.1.15) and (4.21.7) of [129], respectively.

First, we motivate and summarize results for standard ℓ_1-minimization. Then we expand on these results in the case of gradient-enhanced ℓ_1-minimization. The path of this analysis flows through the Restricted Isometry Constant (RIC) [56], which is denoted by $\delta_s(\boldsymbol{\Phi})$ and is defined to be the smallest number satisfying

$$(1 - \delta_s(\boldsymbol{\Phi}))\|\boldsymbol{y}\|_2^2 \leq \|\boldsymbol{\Phi}\boldsymbol{y}\|_2^2 \leq (1 + \delta_s(\boldsymbol{\Phi}))\|\boldsymbol{y}\|_2^2. \tag{4.16}$$

Here, $\delta_s(\boldsymbol{\Phi})$ yields a uniform bound on the spectral radius of the submatrices of $\boldsymbol{\Phi}$ formed by selecting any s columns of $\boldsymbol{\Phi}$. Often, the matrix being considered is clear from context, and we then shorten $\delta_s(\boldsymbol{\Phi})$ to δ_s. Related to the RIC are restricted isometry properties that occur when the RIC reaches a small enough threshold, and guarantee that ℓ_1-minimization with the given matrix is a stable computation. An example of such a restricted isometry property is given in Theorem 4.3.1 from [130]. This theorem shows that if $\delta_{2s} < 3/(4+\sqrt{6})$, where s is dictated by the specific problem, then a stable recovery is assured.

Theorem 4.3.1. *[130] Let $\boldsymbol{c} \in \mathbb{R}^P$ represent a solution we seek to approximate, and let $\hat{\boldsymbol{c}}$ be the solution to (4.4). Let*

$$\boldsymbol{\eta} := \boldsymbol{\Psi}\boldsymbol{c} - \boldsymbol{u},$$

denote the contribution from sources of error, and let ϵ from (4.4), be chosen such that $\|\boldsymbol{\eta}\|_2^2 < \epsilon$. If

$$\delta_{2s}(\boldsymbol{\Psi}) < \delta_\star := 3/(4+\sqrt{6}) \approx 0.4652,$$

then the following error estimates hold,

$$\|\boldsymbol{c} - \hat{\boldsymbol{c}}\|_2 \leq \frac{c_1}{\sqrt{s}} \inf_{\|\boldsymbol{c}_s\|_0 \leq s} \|\boldsymbol{c}_s - \boldsymbol{c}\|_1 + c_2\epsilon;$$

$$\|\boldsymbol{c} - \hat{\boldsymbol{c}}\|_1 \leq c_3 \inf_{\|\boldsymbol{c}_s\|_0 \leq s} \|\boldsymbol{c}_s - \boldsymbol{c}\|_1 + c_4\epsilon\sqrt{s},$$

where c_1, c_2, c_3, and c_4 depend only on δ_{2s}, and $\|\cdot\|_0$ refers to the number of non-zero elements of the vector.

We note that, related to the discussion in Section 4.3.1, ϵ in Theorem 4.3.1 may be selected so that $\|\boldsymbol{\eta}\|_2^2 < \epsilon$ holds with high probability. Unfortunately, identifying the RIC for a given matrix requires a computation for every submatrix of s columns, which is intractable in most situations of interest. As a practical alternative we instead choose to bound the RIC in a probabilistic sense, allowing us to identify a probability that the RIC is below a chosen threshold. In this way we can guarantee that a restricted isometry property, such as the one of Theorem 4.3.1, holds with a certain probability. To do so, we introduce a definition of coherence, first considering the standard case, before expanding its definition to the gradient-enhanced case.

4.3.1 Standard ℓ_1-minimization analysis

We consider here an approach that uses arguments similar to those in [131, 117]. We note that those works did not proceed through the RIC as we do here. First, let $\mathcal{Q}$ be an arbitrary subset of the sample space for Ξ, that is the values which Ξ can take. Here, $\mathcal{Q}$ is used to truncate the domain to one on which the basis functions, here Hermite polynomials, can be uniformly bounded. The coherence parameter for the standard approach [131, 117] is defined to be

$$\mu_{\mathcal{Q}} := \sup_{k,\boldsymbol{\xi}\in\mathcal{Q}} |\psi_k(\boldsymbol{\xi})|_2^2, \tag{4.17}$$

which for a precompact $\mathcal{Q}$ is guaranteed to be finite. An example of a $\mathcal{Q}$ suitable for use with Hermite polynomials, and used in [117, 64], is

$$\mathcal{Q} := \{\boldsymbol{\xi} : \|\boldsymbol{\xi}\|_2^2 \leq (4 + \epsilon_{p,d})p + 2\}, \tag{4.18}$$

where $\epsilon_{p,d}$ is a positive constant, which may be arbitrarily small but close to zero in an asymptotic analysis of the behavior of Hermite polynomials. We note that this truncation has not been analyzed when the number of samples is exponentially greater than the number of basis functions; however, this is not an issue here where the number of samples is typically less than or not substantially greater than the number of basis functions. The definition of coherence parameter as in (4.18) leads to the following theorem, taken from Theorem 4.1 of [117].

Theorem 4.3.2. *For d-dimensional polynomials of order $p \geq 1$, the coherence in (4.17) is bounded by*

$$\mu_{\mathcal{Q}} \leq C_0 \cdot C_1^p, \tag{4.19}$$

where C_0 and C_1 are modest constants depending on $d, p, \epsilon_{p,d}$. As $p/d \to \infty$, C_0 decreases to 1, and C_1 decreases to a limit of $\exp(2 - \log(2)) \approx 3.7$.

This shows an exponential dependence of the coherence parameter on p. Let $\boldsymbol{X}_k$ denote a row vector consisting only of basis polynomial evaluations at $\boldsymbol{\xi}^{(k)}$. We bound the RIC for the matrix $\boldsymbol{\Psi}$ defined as in (4.6).

Theorem 4.3.3. *For any chosen $\mathcal{Q}$, we may bound the RIC in a probabilistic sense by*

$$\mathbb{P}(\delta_s < t) \geq \mathbb{P}(\mathcal{Q})^N - \exp\left(-C_{\mathcal{Q}}\frac{Nt}{s\mu_{\mathcal{Q}}} + s + \log(2s) + s\log(P/s)\right).$$

Remark 4.3.1. *We note that we do not generally require arbitrarily small δ_s, as is seen in Theorem 4.3.1. The constant $C_{\mathcal{Q}}$ scales with*

$$\epsilon_{\mathcal{Q}} := \left\|\mathbb{E}\left(\boldsymbol{X}^T\boldsymbol{X}|\boldsymbol{\xi} \in \mathcal{Q}\right) - \boldsymbol{I}\right\|_2, \tag{4.20}$$

which is a bias that is negligible in practical contexts for $\mathcal{Q}$ as in (4.18) [117]. Truncating in this way, neither $C_{\mathcal{Q}}$ nor $\mathbb{P}(\mathcal{Q})$ are problematic in practice.

Remark 4.3.2. *While our primary focus here is ℓ_1-recovery, the RIC corresponding to $s = P$ is useful for analyzing the stability of a least squares solution, and so this result is also applicable to ℓ_2-minimization. For ways in which this parameter may bound error from* *solutions computed via ℓ_2-minimization we point the interested reader to [132, 64]. In this case, a slight adjustment to the proof gives the bound*

$$\mathbb{P}(\delta_P < t) \geq \mathbb{P}(\mathcal{Q})^N - 2P \exp\left(-C_{\mathcal{Q}} \frac{Nt}{P\mu_{\mathcal{Q}}}\right).$$

The following corollary, which follows from a rearrangement of the result of Theorem 4.3.3, highlights the relationship between several quantities.

Corollary 4.3.1. *To insure that $\delta_s < \delta_\star$ with probability $p_\star$, it is sufficient to take $N_\star$ satisfying*

$$N_\star \delta_\star \geq \frac{s\mu_{\mathcal{Q}}}{C_{\mathcal{Q}}} \left[s + \log(2s) + s\log(P/s) - \log\left(\mathbb{P}(\mathcal{Q})^{N_\star} - p_\star\right)\right].$$

For example, using $\delta_\star$ as in Theorem 4.3.1, gives a guarantee for stability of solutions to (4.4) when the number of samples N satisfies

$$N \geq \frac{(4+\sqrt{6})s\mu_{\mathcal{Q}}}{3C_{\mathcal{Q}}} \left[s + \log(2s) + s\log(P/s) - \log\left(\mathbb{P}(\mathcal{Q})^{N} - p_\star\right)\right].$$

We note that N appears on both the left and right sides of the equation. This arises from the truncation, which has a technical issue when N is exponentially larger than P; specifically, the probability that at least one sample had fallen in $\mathcal{Q}^c$ becomes large, while the analysis relies on this event being rare. As N is very often smaller than P, and rarely chosen to be exponentially larger than P, this issue is not of practical concern. Next, we extend these results to the case where gradient information is included.

4.3.2 Gradient-enhanced ℓ_1-minimization analysis

In the previous case the rows of $\mathbf{\Psi}$ were independent, while $\tilde{\mathbf{\Psi}}$ has a more subtle independence structure, as only the sets of rows associated with the independent samples are independent. Related to this point, we let the $(d+1) \times P$ block of independent information related with the kth sample be given by $\boldsymbol{X}_k$, with a generic realization given by $\boldsymbol{X}$. Specifically,

$$\begin{aligned} \boldsymbol{X}(i,j) &= \frac{\partial \psi_j}{\partial \Xi_i}(\boldsymbol{\xi}), \quad i = 1, \dots, d; \\ \boldsymbol{X}(d+1,j) &= \psi_j(\boldsymbol{\xi}). \end{aligned}$$

That is, the last row corresponds to the realizations of the function, while the first d rows correspond to the derivative information. We note that there is no effect in the computations considered here by rearranging the rows of the matrix $\tilde{\boldsymbol{\Psi}}$. The RIC will necessarily be larger if the norms of the matrix columns are different. Adjusting for this can be done by adjusting the basis functions themselves. In the case of standard ℓ_1-minimization, we use orthonormal polynomials. Here, we will wish to include derivatives of those polynomials as well, requiring a different normalization. We use the Probabilists' Hermite polynomials, and the following lemma is used to identify this normalization.

Lemma 4.3.1. *For d-dimensional orthonormal Probabilists' Hermite polynomials $\psi_{\boldsymbol{i}}$ and $\psi_{\boldsymbol{j}}$ with order i_k, j_k in dimension k,*

$$\mathbb{E}\left(\psi_{\boldsymbol{i}}(\Xi)\psi_{\boldsymbol{j}}(\Xi) + \sum_{k=1}^{d} \frac{\partial \psi_{\boldsymbol{i}}}{\partial \Xi_k}(\Xi)\frac{\partial \psi_{\boldsymbol{j}}}{\partial \Xi_k}(\Xi)\right) = \delta_{\boldsymbol{i},\boldsymbol{j}}\left(1 + \sum_{k=1}^{d} i_k\right), \tag{4.21}$$

where $\delta_{\boldsymbol{i},\boldsymbol{j}}$ is the Kronecker Delta.

This suggests a different normalization of the basis functions to enforce that columns of the gradient-enhanced measurement matrix have the same expected ℓ_2-norm. We refer to this normalization as *gradient-normalization*. Specifically, we multiply the orthonormal basis function $\psi_{\boldsymbol{i}}$ by

$$\tilde{w}_{\boldsymbol{i}} := \left(1 + \sum_{k=1}^{d} i_k\right)^{-1/2}, \tag{4.22}$$

to gradient-normalize. We note that it is these weights that define the $\boldsymbol{W}$ in (4.9) and (4.15). In this work, we assume that when derivative information is included, that those basis functions are gradient-normalized, and that when derivative information is not included, that those basis functions are orthonormal, which we refer to as standard-normalization. This insures that the expected norms for the columns of the sampled matrices are consistent in both cases.

Recalling that N denotes the number of samples used, our analysis focuses on the Gramian matrix

$$\boldsymbol{M} := \frac{1}{N}\sum_{k=1}^{N} \boldsymbol{X}_k^T \boldsymbol{X}_k. \tag{4.23}$$

Here, $\boldsymbol{M} = N^{-1}\boldsymbol{\Psi}^T\boldsymbol{\Psi}$ for the standard approach, while $\boldsymbol{M} = N^{-1}\tilde{\boldsymbol{\Psi}}^T\tilde{\boldsymbol{\Psi}}$ for the gradient-enhanced approach. We now present some summarized results for the standard approach that will be compared to the results presented for the gradient-enhanced case. To analyze the spectrum of $\boldsymbol{M}$ in the case of gradient-enhanced ℓ_1-minimization, we use the following definition, which generalizes the ℓ_1-coherence as studied in [131, 117, 130] and defined in (4.17). Let $\mathcal{Q}$ be an arbitrary subset of Ω, which we use to truncate the sample space to insure a uniformly bounded polynomial system, e.g. as in [131, 117, 64], and let

$$\beta_{\mathcal{Q}} := \sup_{k,\boldsymbol{\xi}\in\mathcal{Q}} \|\boldsymbol{X}(:,k)\|_2^2. \tag{4.24}$$

This parameter is a generalization of $\mu_{\mathcal{Q}}$ in the case where rows are not independent, but sets of rows are. Note that in the case that $\boldsymbol{X}$ is a row vector, then this definition reduces to (4.17). Specifically, the results of Section 4.3.1 all hold when substituted for this parameter, which we highlight as a theorem.

Theorem 4.3.4. *The theorems of Section 4.3.1 hold for the gradient-enhanced case when $\mu_{\mathcal{Q}}$ is replaced by $\beta_{\mathcal{Q}}$.*

We conclude our analysis with three results which demonstrate that the inclusion of derivative information, coupled with gradient-normalization, does not reduce stability over the corresponding approach without derivative information.

The first result is an inequality concerning the coherence parameter defined in (4.17) and (4.24), showing that including derivative information does not require any weakening of the bounds of Section 4.3.1 in the case without derivative information. This inequality then directly applies to all of the theorems in Section 4.3.1.

The second result is a direct null-space comparison of the two different matrices, which is known to be fundamental for recovery of exactly-sparse solutions [56, 97, 98]. Specifically, as $\|\tilde{\boldsymbol{\Psi}}\boldsymbol{c} - \tilde{\boldsymbol{u}}\| < \delta$ (or $\|\boldsymbol{\Psi}\boldsymbol{c} - \boldsymbol{u}\| < \delta$) is enforced, the difference between potential solutions is close to an element of the null-space of $\tilde{\boldsymbol{\Psi}}$ (or $\boldsymbol{\Psi}$), so reducing the dimension of the null-space correspondingly reduces the space of potential solutions.

The third result concerns a bound on the inner-products of columns of the measurement-matrix. This is related to the RIC in that if the inner-product between several pairs of columns is of large absolute value then a linear combination of those columns will have small norm, resulting in a larger RIC. Similarly, if those inner-products are of small absolute value, then no linear combination will have a small norm. An analogous observation may be made regarding a linear combination of columns having a much larger norm. For this reason it is beneficial if the inner-product between columns is of small absolute value.

Theorem 4.3.5. *Let $\tilde{\boldsymbol{\Psi}}$ be a realized measurement matrix with derivative information that is gradient-normalized. Similarly, let $\boldsymbol{\Psi}$ be a realized measurement matrix with standard-normalization and no derivative information.*

Assume that $\boldsymbol{\Psi}$ and $\tilde{\boldsymbol{\Psi}}$ are formed from the same realized input samples, $\{\boldsymbol{\xi}^{(i)}\}_{i=1}^{N}$, so that up to row weighting, $\boldsymbol{\Psi}$ is a sub-matrix of $\tilde{\boldsymbol{\Psi}}$. Then the following statements related to the recovery of solutions via ℓ_1-minimization hold.

R1. *Using the definition in (4.24) for the two different approaches,*

$$\beta_{\mathcal{Q}}(\tilde{\boldsymbol{\Psi}}) \leq \mu_{\mathcal{Q}}(\boldsymbol{\Psi}),$$

and this inequality is almost-surely strict.

R2. *If $\mathcal{N}(\cdot)$ represents the null-space, then $\mathcal{N}(\tilde{\boldsymbol{\Psi}}) \subset \mathcal{N}(\boldsymbol{\Psi})$, and this is almost-surely a strict subset when $\boldsymbol{\Psi}$ is undersampled. Specifically, it almost-surely holds that $\dim(\mathcal{N}(\tilde{\boldsymbol{\Psi}})) = \max\{0, P - (d+1)N\}$ while $\dim(\mathcal{N}(\boldsymbol{\Psi})) = \max\{0, P - N\}$.*

R3. *If subscripts of matrices correspond to the columns, and i_k denotes the order of the basis polynomial in the $\boldsymbol{i}$th column in the kth dimension, then the associated inner product of columns is bounded by,*

$$\sup_{\boldsymbol{i}\neq\boldsymbol{j}} |(\tilde{\boldsymbol{\Psi}}_{\boldsymbol{i}}, \tilde{\boldsymbol{\Psi}}_{\boldsymbol{j}})| \leq \sup_{\boldsymbol{i}\neq\boldsymbol{j}} \frac{|(\boldsymbol{\Psi}_{\boldsymbol{i}}, \boldsymbol{\Psi}_{\boldsymbol{j}})|(1+\sum_{k=1}^{d}\sqrt{i_k j_k})}{\sqrt{(1+\sum_{k=1}^{d} i_k)(1+\sum_{k=1}^{d} j_k)}} \leq \sup_{\boldsymbol{i}\neq\boldsymbol{j}} |(\boldsymbol{\Psi}_{\boldsymbol{i}}, \boldsymbol{\Psi}_{\boldsymbol{j}})|. \tag{4.25}$$

Remark 4.3.3. *The theorem is presented for full gradient information to ease presentation, but the three points generalize to the case that derivative information is included for a fraction of samples. Specifically, (1) holds with an appropriate adjustment to the dimensionality, (2) holds with an adjustment to the basis dependent multiplicative constant, and (3) holds if $\tilde{\boldsymbol{\Psi}}$ has derivative information for only a few samples. In summary adding derivative information for even a percentage of the samples leads to bounds as in Theorem 4.3.5.*

4.3.3 A note on potentially contrasting solutions

It is of practical importance to note what functions are recovered in an asymptotic sense by the gradient-enhanced and standard approaches, as they may differ significantly. The gradient-enhanced method gives $\hat{u}$ approximating u in a Sobolev type loss function,

$$L(\hat{u}, u) := \|\hat{u} - u\|^2_{\ell_2(\boldsymbol{\Xi},N)} + \sum_{k=1}^{d} \left\| \frac{\partial(\hat{u}-u)}{\partial \Xi_k} \right\|^2_{\ell_2(\boldsymbol{\Xi},N)}, \tag{4.26}$$

where the $\ell_2(\boldsymbol{\Xi}, N)$ indicates the discrete ℓ_2 norm using N evaluations drawn from realizations of $\boldsymbol{\Xi}$, here $(\boldsymbol{\xi}^{(1)}, \dots, \boldsymbol{\xi}^{(N)})$. This norm is also normalized by N^{-1} so that as N goes to infinity this norm tends to $\mathcal{L}_2(\boldsymbol{\Xi})$, the standard $\mathcal{L}_2$ norm associated with the distribution of $\boldsymbol{\Xi}$. Specifically (4.15) guarantees that $NL(u, \hat{u}) < \delta$. In contrast, without derivative information using standard-normalization and producing a solution via (4.4), gives $\hat{u}$ approximating u such that $N\|\hat{u} - u\|^2_{\ell_2(\boldsymbol{\Xi},N)} < \delta$, that is the partial derivatives of the approximation need not be approximated by those of the target function. As these loss functions differ, so too does the limiting solution produced by each method for a finite expansion order p. However, the sequence of approximations given by the two loss functions and by using increasing p (and accordingly N) will converge to u in the $\mathcal{L}_2(\Xi)$ sense as $p, N \rightarrow \infty$.

4.4 Numerical Results

In this section, we empirically demonstrate the gradient-enhanced ℓ_1-minimization approach via three numerical examples: a manufactured PCE; an elliptic PDE with stochastic coefficient; and a plane Poiseuille flow with random boundaries. To compare the standard and gradient-enhanced ℓ_1-minimization solutions, we define an equivalent sample size $\tilde{N}$,

$$\tilde{N} := N_e + \nu N_g, \tag{4.27}$$

which accounts for the added cost of computing the derivative information. In (4.27), N_e is the number of samples without derivative information, N_g is the number of samples with derivatives (along all d directions), and ν is a positive parameter depending on the problem at hand and the approach employed to compute the derivatives. For the example of Section 4.4.2, the cost of generating d derivatives of the QoI, obtained by the adjoint sensitivity method, is roughly the same as that of evaluating the QoI, thus implying that $\nu = 2$. For transient problems for which the cost of solving the adjoint equations for derivative calculations may be considerably more than that of a single QoI evaluation, then $\nu > 2$. Nevertheless, we here present all cost comparisons in terms of the number of equivalent sample size $\tilde{N}$ in (4.27), for choices of ν that we shall specify. For simulations based on standard ℓ_1-minimization, we set $\tilde{N} = N_e$. Additionally, for the interest of convenience, we ignore the cost of solving the ℓ_1-minimization problems in (4.4) and (4.15). This is a valid assumption as often that cost is negligible relative to the cost of evaluating the QoI or its derivatives. For terminology, if $X\%$ of samples used in the computation of the solution contain derivative information, then we say that method is $X\%$ gradient-enhanced. In this way, the standard approach without derivatives would give a solution that is 0% gradient-enhanced, denoted as standard in the subsequent figures. Similarly, if all samples include derivative information, the associated solution would be 100% gradient-enhanced.

4.4.1 Case I: A manufactured PCE

First, we consider the reconstruction of a manufactured PCE, in which the sparsity and the entries of the coefficient vector $\boldsymbol{c}$ are *a priori* prescribed. Specifically, we set the dimension of the expansion to $d = 25$ and use a $p = 3$ order PCE (hence $P = 3276$ basis functions) to manufacture the QoI $u(\Xi)$. To generate $\boldsymbol{c}$, we first draw its P entries independently from the standard Gaussian distribution. We then retain $|\mathcal{C}| \in \{50, 150\}$ coefficients with largest magnitude and set the rest to zero. This gives a randomized sparsity support. Finally, the realizations of $u(\Xi)$ and its derivative with respect to Ξ_k, $k = 1, \dots, d$, are generated by

$$u(\boldsymbol{\xi}^{(i)}) = \sum_{j=1}^{P} c_j \psi_j(\boldsymbol{\xi}^{(i)}) \tag{4.28}$$

and

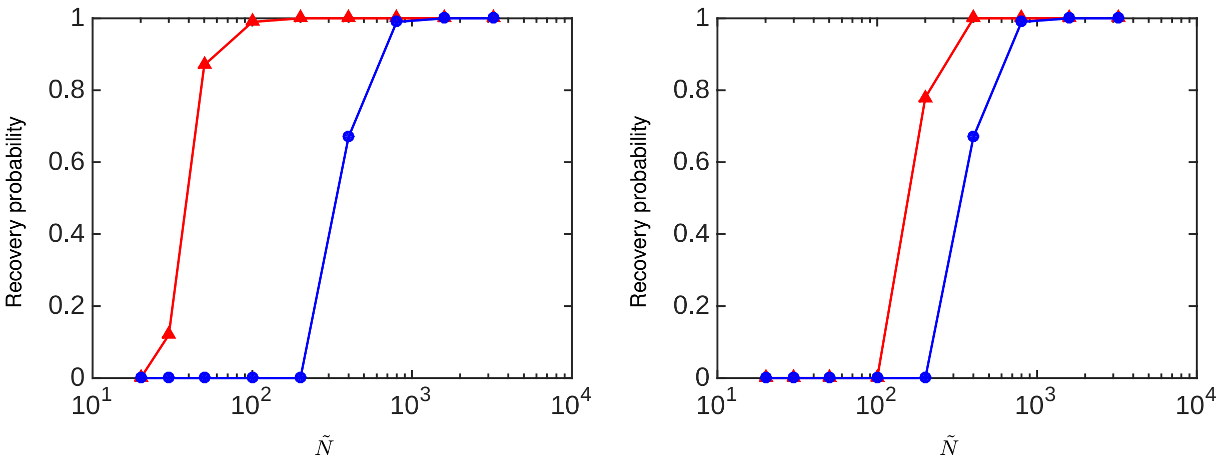
$$u_{\partial_k}(\boldsymbol{\xi}^{(i)}) := \sum_{j=1}^{P} c_j \frac{\partial \psi_j}{\partial \Xi_k}(\boldsymbol{\xi}^{(i)}), \tag{4.29}$$

respectively. We then approximate the PCE coefficients $\boldsymbol{c}$ via (4.4) and (4.15) from these generated data.

Results

Beginning with the $|\mathcal{C}| = 50$ case, we seek to recover the manufactured $\boldsymbol{c}$. If the computed solution, denoted by $\hat{\boldsymbol{c}}$, has a relative root-mean-square-error (RRMSE) below 0.01%, then we call it a successful recovery of $\boldsymbol{c}$.

We first consider the case where evaluations of $u(\Xi)$ and its derivatives are exact, i.e., noise free. In Figure 4.1, we compare the probability of successful recovery for gradient-enhanced and standard approaches, using 100 independent replications for each $\tilde{N}$. To set $\tilde{N}$, we pretend the cost of evaluating d derivatives of $u(\Xi)$ is the same as that of evaluating $u(\Xi)$, and therefore we set $\nu = 2$.



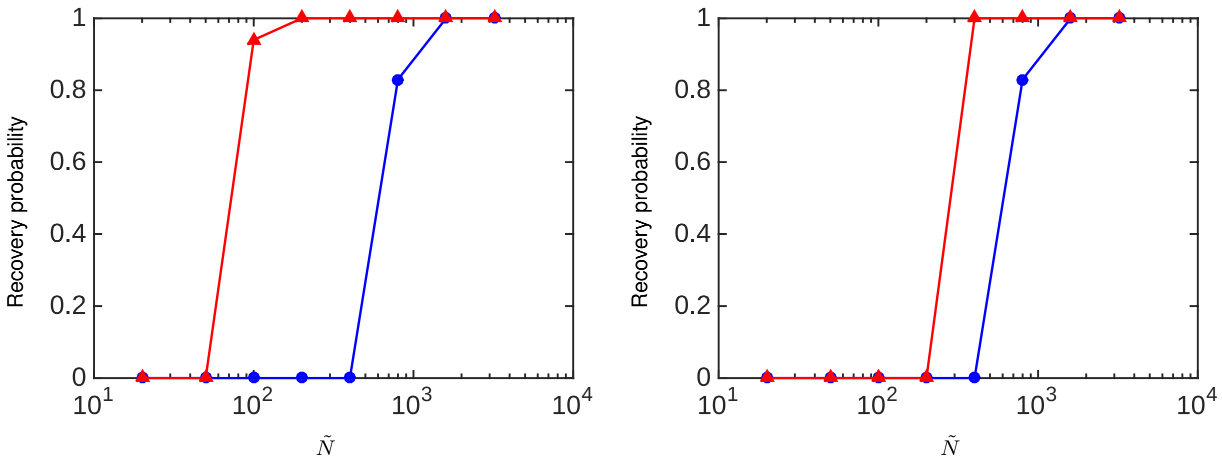
(a)

(b)

Figure 4.1: Probability of successful recovery of manufactures PCE with sparsity $|\mathcal{C}| = 50$ via gradient-enhanced and standard ℓ_1-minimization. (a) 100% vs. 0% gradient-enhanced. (b) 20% vs. 0% gradient-enhanced. (▲ gradient-enhanced, ● standard)

In Figure 4.1a, we see that 100% gradient-enhanced ℓ_1-minimization helps in reducing the computational effort to recover $\boldsymbol{c}$, while Figure 4.1b, demonstrates a notable, but less, improvement for 20% gradient-enhanced ℓ_1-minimization. The figure suggests that adding derivative information is a cost-effective means to increase the probability of successfully recovering $\boldsymbol{c}$ at low sample sizes $\tilde{N}$.

Figure 4.2 shows similar results for sparsity $|\mathcal{C}| = 150$. Since $|\mathcal{C}|$ is larger in this case, both standard and gradient-enhanced methods require more samples for recovery with a given success probability. This is consistent with the sampling rate presented in Corollary 4.3.1. We, however, note that the inclusion of derivative information still enhances the solution recovery.

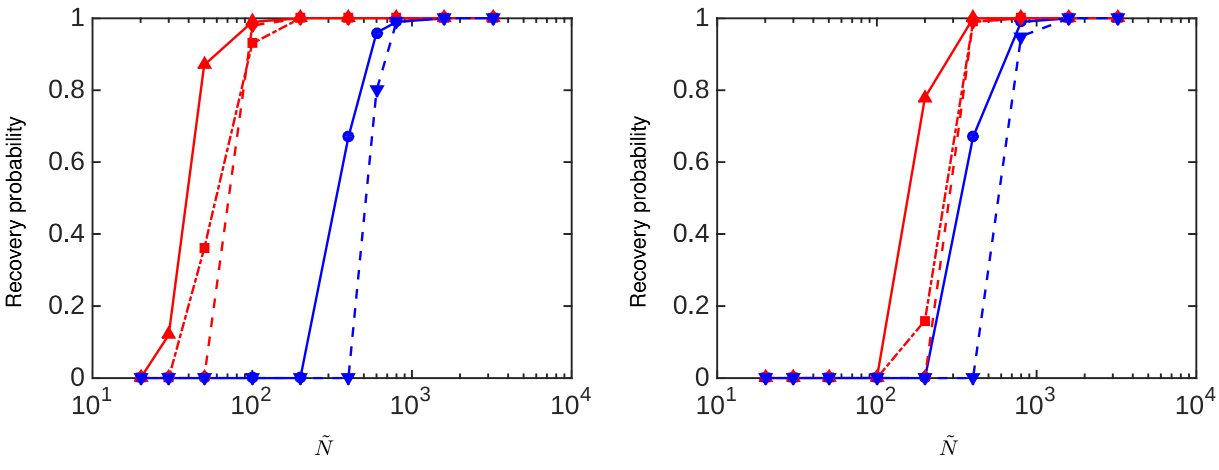


(a)

(b)

Figure 4.2: Probability of successful recovery of manufactures PCE with sparsity $|\mathcal{C}| = 150$ via gradient-enhanced and standard ℓ_1-minimization. (a) 100% vs. 0% gradient-enhanced. (b) 20% vs. 0% gradient-enhanced. (▲ gradient-enhanced, ● standard)

In practice, there is often error (or noise) in the evaluation of $u(\Xi)$ and its derivatives, with the latter being more prone to errors. To model such inaccuracies, here we multiplying the realizations of $u(\Xi)$ and its derivatives from (4.28) and (4.29), respectively, by independent realizations of $(1+\epsilon_N)$, where ϵ_N is a zero mean Gaussian random variable with variance 10^{-5}. In Figure 4.3, we consider the sparsity $|\mathcal{C}| = 50$ case and compute the probability of successful solution recovery as a function of $\tilde{N}$. Similar as in the previous test cases, we use an RRMSE error of 0.01% to identify a successful recovery. We observe that the inclusion of derivative information, while imprecise, still improves the performance of the standard ℓ_1-minimization.



(a) (b)

Figure 4.3: Probability of successful recovery of gradient-enhanced and standard ℓ_1-minimization for the manufactured PCE case with sparsity $|\mathcal{C}| = 50$. (a) 100% gradient-enhanced. (b) 20% gradient-enhanced (▲ gradient-enhanced, ■ gradient-enhanced with noisy $\boldsymbol{u}_\partial$ only, ♦ gradient-enhanced with both noisy $\boldsymbol{u}$ and $\boldsymbol{u}_\partial$, ● standard, ▼ standard with noisy $\boldsymbol{u}$)

4.4.2 Case II: Two-dimensional elliptic PDE with random coefficient

We next consider the two-dimensional (in space) elliptic PDE

$$\begin{aligned} &-\nabla\cdot(a(\boldsymbol{x},\boldsymbol{\Xi})\nabla u(\boldsymbol{x},\boldsymbol{\Xi})) = 1, \qquad && \boldsymbol{x}\in\mathcal{D}=[0,1]^2\ , \\ &u(\boldsymbol{x},\boldsymbol{\Xi}) = 0, && \boldsymbol{x}\in\partial\mathcal{D}\ , \end{aligned} \tag{4.30}$$

where the diffusion coefficient $a(\boldsymbol{x},\boldsymbol{\Xi})$ is modeled by the lognormal random field

$$a(\boldsymbol{x},\boldsymbol{\Xi}) = \exp\left[\bar{a} + \sigma_a\sum_{k=1}^{d}\sqrt{\lambda_k}\phi_k(\boldsymbol{x})\Xi_k\right]. \tag{4.31}$$

Here, $d = 30$, and $\Xi_k,\ k = 1,\dots,d$, are independent standard Gaussian random variables. In addition, $\bar{a} = 0.1$, $\sigma_a = 0.5$, and $\{\phi_k\}_{k=1}^d$ are the eigenfunctions corresponding to the d largest eigenvalues $\{\lambda_k\}_{k=1}^d$ of the Gaussian covariance kernel

$$C_{aa}(\boldsymbol{x}_1,\boldsymbol{x}_2) = \exp\left[-\frac{(x_1-x_2)^2}{l_c^2} - \frac{(y_1-y_2)^2}{l_c^2}\right]\ , \tag{4.32}$$

with correlation length $l_c = 1/16$.

The QoI $u(\boldsymbol{\Xi})$ is chosen as the solution to (4.30) at location $\boldsymbol{x} = (0.5, 0.5)$, i.e., $u(\boldsymbol{\Xi}) = u\left((0.5, 0.5), \boldsymbol{\Xi}\right)$. The derivatives $\partial u / \partial \Xi_k$ are computed using the adjoint sensitivity method explained in detail in Section 4.4.2. Both forward and adjoint solvers are implemented in the finite element method (FEM) project FEniCS [116].

Adjoint sensitivity derivatives

The adjoint sensitivity methods are commonly used in research areas such as sensitivity analysis [133, 134], optimization [135, 136], shape design [137, 138], etc., to compute the derivatives of the solutions of interest with respect to the underlying model parameters. In this work, we adopt the discrete adjoint sensitivity method to compute derivatives of the QoI at the $\boldsymbol{\Xi}$ samples.

In detail, for the interest of convenience, we consider the discrete formulation of the generic PDE in (4.10), at a fixed time, given by the residual equation

$$\boldsymbol{\mathcal{R}}(\boldsymbol{w}, \boldsymbol{\Xi}) = \boldsymbol{0}, \tag{4.33}$$

where $\boldsymbol{w} \in \mathbb{R}^{M \times 1}$ contains the discrete values of solution over the spatial domain $\mathcal{D}$, and M is the number of solution degrees-of-freedom. Recalling that $u(\boldsymbol{\Xi})$ (here a scalar functional of the solution) denotes the QoI, we seek to compute the sensitivity derivatives $du/d\Xi_k$ from

$$\frac{du}{d\Xi_k} = \frac{\partial u}{\partial \Xi_k} + \frac{\partial u}{\partial \boldsymbol{w}} \frac{\partial \boldsymbol{w}}{\partial \Xi_k}. \tag{4.34}$$

Taking the derivative of (4.33) with respect to Ξ_k, we have

$$\frac{\partial \boldsymbol{\mathcal{R}}}{\partial \Xi_k} + \frac{\partial \boldsymbol{\mathcal{R}}}{\partial \boldsymbol{w}} \frac{\partial \boldsymbol{w}}{\partial \Xi_k} = \boldsymbol{0}, \tag{4.35}$$

which results in $\partial \boldsymbol{w} / \partial \Xi_k = -\left(\partial \boldsymbol{\mathcal{R}} / \partial \boldsymbol{w}\right)^{-1} \partial \boldsymbol{\mathcal{R}} / \partial \Xi_k$. Plugging this in (4.34) gives

$$\frac{du}{d\Xi_k} = \frac{\partial u}{\partial \Xi_k} - \frac{\partial u}{\partial \boldsymbol{w}} \left(\frac{\partial \boldsymbol{\mathcal{R}}}{\partial \boldsymbol{w}}\right)^{-1} \frac{\partial \boldsymbol{\mathcal{R}}}{\partial \Xi_k},$$

which can be rewritten as

$$\frac{du}{d\Xi_k} = \frac{\partial u}{\partial \Xi_k} + \boldsymbol{\lambda}^T \frac{\partial \boldsymbol{\mathcal{R}}}{\partial \Xi_k}, \tag{4.36}$$

where $\boldsymbol{\lambda}$ is the solution to the discrete adjoint equation

$$\left(\frac{\partial \boldsymbol{\mathcal{R}}}{\partial \boldsymbol{w}}\right)^T \boldsymbol{\lambda} = -\left(\frac{\partial u}{\partial \boldsymbol{w}}\right)^T. \tag{4.37}$$

For the case of elliptic PDE (4.30), $\partial\boldsymbol{\mathcal{R}}/\partial\boldsymbol{w}$ is the symmetric stiffness matrix of the FEM discretization and $\partial\boldsymbol{\mathcal{R}}/\partial\Xi_k$ in (4.36) can be computed semi-analytically from (4.31) and the FEM formulation of (4.30). Here, we assume the inverse or a factorization of $\partial\boldsymbol{\mathcal{R}}/\partial\boldsymbol{w}$ is not stored when solving for $\boldsymbol{w}$, and that the total cost of obtaining $\partial\boldsymbol{\mathcal{R}}/\partial\Xi_k$ is smaller than that of computing $\boldsymbol{w}$. Therefore, the cost of solving for $\boldsymbol{\lambda}$ from (4.37) is roughly the same as solving for $\boldsymbol{w}$, which in turn suggests that $\nu = 2$ in (4.27).

Results

We approximate $u(\boldsymbol{\Xi})$ in a Hermite PCE with total degree $p = 3$, and seek to approximate the first 2500 coefficients. We sort the elements of $\{\psi_j\}$ such that, for any given total order basis, the random variables Ξ_k with smaller indices k contribute first to the basis.

To solve for the coefficients, we first generate the realizations $\boldsymbol{u}$ and $\boldsymbol{u}_\partial$ using a 256×256 uniform, linear FEM mesh, which resolves both quantities with *low* numerical errors. In Figure 4.4, we compare the mean and standard deviation of the RRMSE for solutions computed by the standard and gradient-enhanced ℓ_1-minimization, using 100 independent replications. The reference PCE coefficients are computed using least squares regression [64] with 10000 solution realizations and yields a relative error of 0.36% for 1000 additional validation samples. From Figure 4.4, we observe that higher accuracies are achieved by the gradient-enhanced ℓ_1-minimization with the same number of samples $\tilde{N}$. In Figure 4.5, we show the magnitude of the approximate PCE coefficients computed via standard and gradient-enhanced ℓ_1-minimization with $\tilde{N} = 80$ samples. More accurate coefficient estimates are obtained by the gradient-enhanced ℓ_1-minimization.

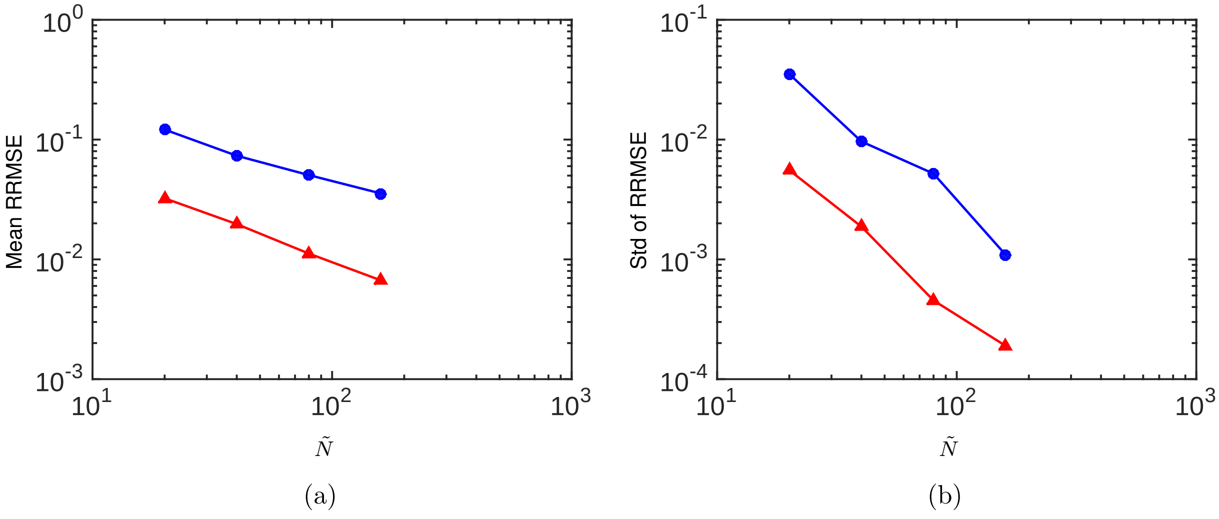


Figure 4.4: Comparison of the statistics of the RRMSE in reconstructing $u\left((0.5, 0.5), \Xi\right)$, where the realizations of u and its derivatives are computed on a uniform 256×256 FEM mesh. (a) Mean of RRMSE. (b) Standard deviation of RRMSE. (▲ 100% gradient-enhanced, ● standard)

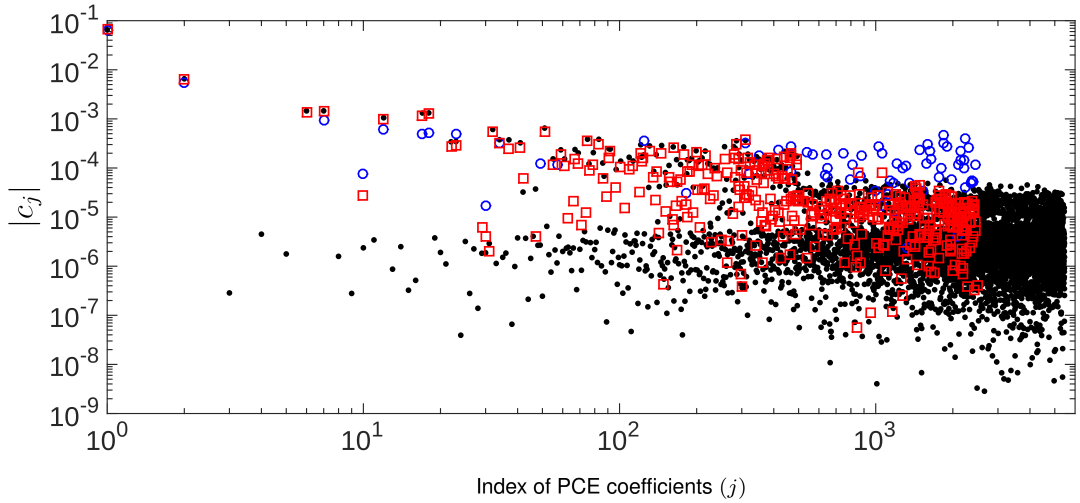
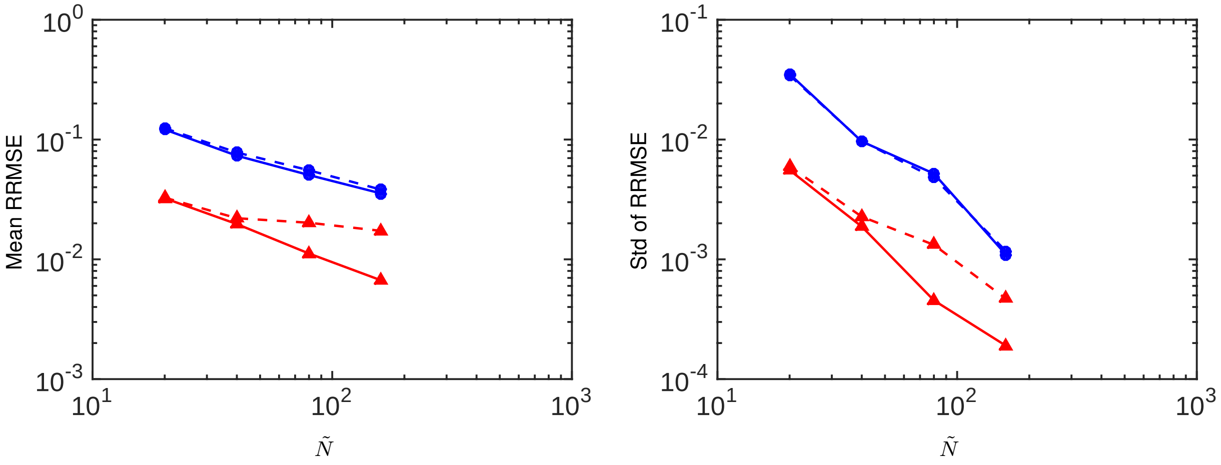


Figure 4.5: Approximate PCE coefficients of $u\left((0.5, 0.5), \Xi\right)$ with $\tilde{N} = 80$ vs. the reference coefficients obtained by least squares regression. (● reference, ○ standard ℓ_1-minimization, □ gradient-enhanced ℓ_1-minimization)

To study how the accuracy of derivative information affects the accuracy of solution obtained by the gradient-enhanced ℓ_1-minimization, we repeat this experiment on a coarser 16×16 mesh, where the derivative information is noticeably less accurate. We observe from Figure 4.6 that, the accuracy improvement achieved by the gradient-enhanced ℓ_1-minimization is not as considerable as in the case of 256×256 mesh. This suggests that high accuracy on the derivative samples $\boldsymbol{u}_\partial$ may be required for the gradient-enhanced ℓ_1-minimization to be most effective.



(a)

(b)

Figure 4.6: Comparison of the statistics of the RRMSE in reconstructing $u\left((0.5, 0.5), \boldsymbol{\Xi}\right)$ via gradient-enhanced and standard ℓ_1-minimization. (a) Mean of RRMSE. (b) Standard deviation of RRMSE. (−▲− 100% gradient-enhanced, −●− standard. Dashed and solid lines, respectively, correspond to the 16×16 and 256×256 mesh simulations.)

4.4.3 Case III: Plane Poiseuille flow with random boundaries

We consider a 2-D plane Poiseuille flow with random boundaries as depicted in Figure 4.7.

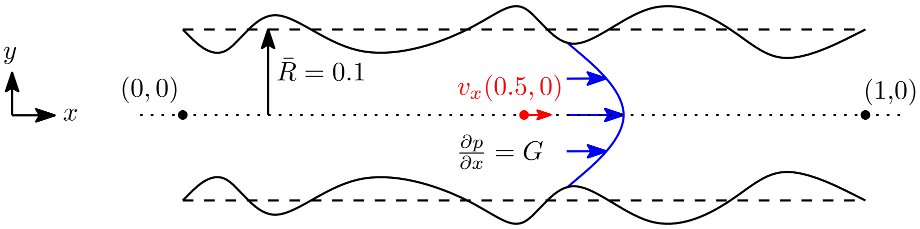


Figure 4.7: Schematic figure of plane Poiseuille flow with random boundaries.

The average width of the channel is $2\bar{R} = 0.2$, and the width of the channel is model as $2R(x) = 2\bar{R}+2r(x,\Xi)$, where $2r(x,\Xi)$ describes the random fluctuation of the channel width around $2\bar{R}$. We use $d = 20$ independent standard Gaussian random variables Ξ_k, $k = 1,\dots,d$, to represent the uncertainty in R, and let

$$r(x,\Xi) = \exp\left[\bar{r} + \sigma_r \sum_{k=1}^{d} \sqrt{\lambda_k}\phi_k(x)\Xi_k\right]. \tag{4.38}$$

Here, $\bar{r} = -4$, $\sigma_r = 0.5$, and $\{\lambda_k\}_{k=1}^d$ and $\{\phi_k(x)\}_{k=1}^d$ are the d largest eigenvalues and corresponding eigenfunctions of the exponential covariance kernel

$$C_{rr}(x_1,x_2) = \exp\left(-\frac{|x_1 - x_2|}{l_c}\right) \ , \tag{4.39}$$

where the correlation length $l_c = 1/21$.

We seek to investigate the steady state velocity field of the flow, which is governed by the incompressible steady state Navier-Stokes equations

$$\begin{aligned}
&(\boldsymbol{v}\cdot\nabla)\boldsymbol{v} - \frac{1}{Re}\nabla^2\boldsymbol{v} = -\nabla p, && \boldsymbol{x}\in\mathcal{D}(\Xi), \qquad (4.40)\\
&\nabla\cdot\boldsymbol{v} = 0, && \boldsymbol{x}\in\mathcal{D}(\Xi),\\
&\frac{\partial p}{\partial x} = G, && \boldsymbol{x}\in\mathcal{D}(\Xi),\\
&\boldsymbol{v} = 0, && y = \pm R,
\end{aligned}$$

where the Reynolds number $Re = 60$, and p denotes pressure. Notice that in (4.40), we assume that all physical quantities are non-dimensional. The flow is driven by a pressure gradient $G = -0.1$.

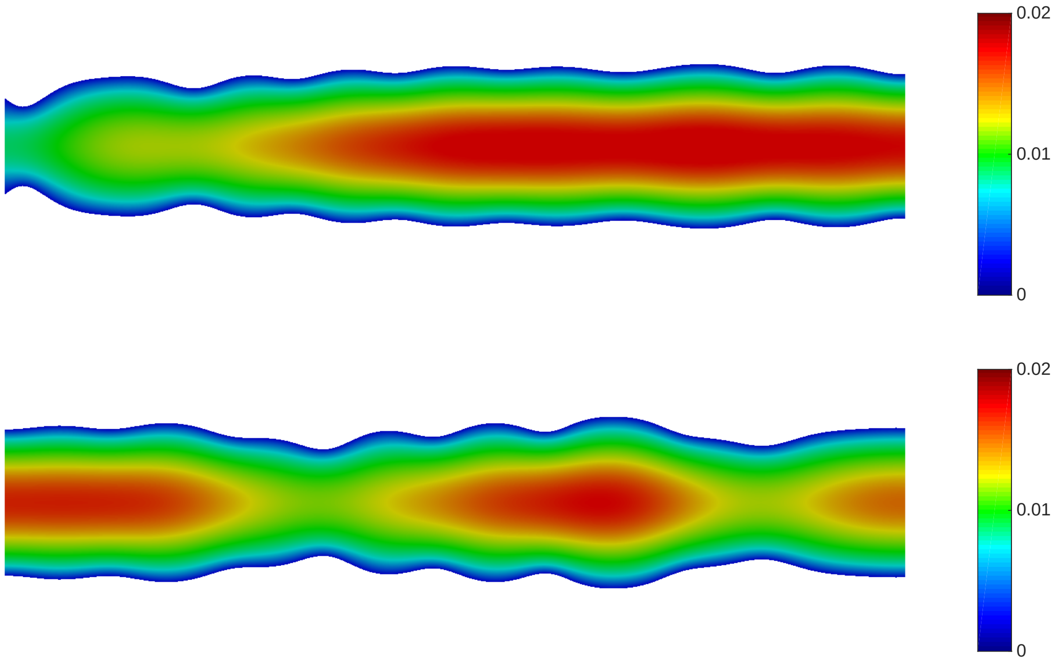


Figure 4.8: Velocity magnitude contours with two independent realizations of the random inputs.

The QoI is $v_x(0.5, 0)$, the horizontal velocity at $(0.5, 0)$, and we approximate it in a Hermite PCE of total degree $p = 3$.

Adjoint sensitivity derivatives

To compute the derivative information, we again adopt the adjoint sensitivity method, in which we approximate $\partial\boldsymbol{\mathcal{R}}/\partial\Xi_k,\ k = 1, \ldots, d$, in (4.36) via finite difference quotient

$$\frac{\partial\boldsymbol{\mathcal{R}}}{\partial\Xi_k} \approx \frac{\boldsymbol{\mathcal{R}}(\boldsymbol{w}, \boldsymbol{\Xi} + \Delta\boldsymbol{\Xi}^k) - \boldsymbol{\mathcal{R}}(\boldsymbol{w}, \boldsymbol{\Xi})}{\varepsilon}. \tag{4.41}$$

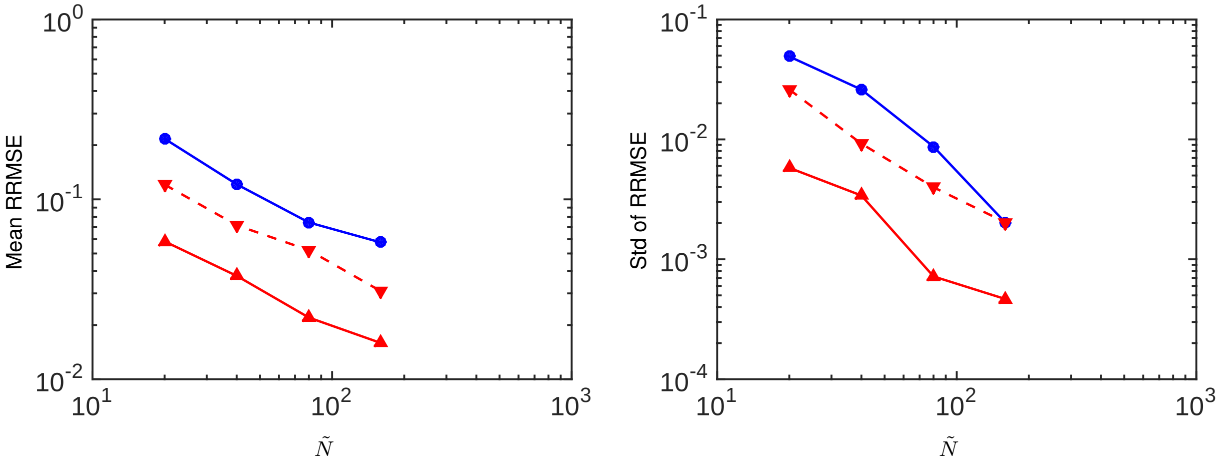
Here, the mth entry of the vector $\Delta\boldsymbol{\Xi}^k$ is defined as

$$\Delta\Xi_m^k = \begin{cases} \varepsilon, & m = k, \\ 0, & m \neq k. \end{cases} \tag{4.42}$$

To solve the non-linear problem (4.40), we employ a standard Newton solver, in which $\partial\boldsymbol{\mathcal{R}}/\partial\boldsymbol{w}$ from (4.37) is the Jacobian matrix. In (4.41), the discrete representation of the velocity field, $\boldsymbol{w}$, is kept unchanged; hence, (4.40) does not need to be solved again. The perturbed residual $\boldsymbol{\mathcal{R}}(\boldsymbol{w}, \Xi + \Delta\Xi^k)$ is computed by deforming the mesh to conform to the geometry corresponding to $\Xi + \Delta\Xi^k$, without recomputing $\boldsymbol{w}$. Compared to solving the adjoint equation (4.37), the cost of calculating (4.41) is negligible. Additionally, in our deterministic solver, computing $\boldsymbol{w}$ required on average 3 Newton steps. Therefore, the extra cost of computing derivative information is roughly equivalent to 1/3 of the cost of computing $\boldsymbol{w}$, which in turn suggests setting $\nu = 4/3$ in (4.27).

Results

We consider 0%, 20% and 100% gradient-enhanced ℓ_1-minimization, with $\tilde{N} \in \{20, 40, 80, 160\}$.



(a)

(b)

Figure 4.9: Comparison of the statistics of the RRMSE in reconstructing $v_x\left((0.5, 0), \Xi\right)$ via standard and gradient-enhanced ℓ_1-minimization, with 100 independent replications. (a) Mean of RRMSE. (b) Standard deviation of RRMSE. (▲ 100% gradient-enhanced, ▼ 20% gradient-enhanced, ● standard)

Figure 4.9 displays the comparisons of the mean and standard deviation of the RRMSE, with 100 independent replications of computed PCE coefficients, showing that gradient-enhanced ℓ_1-minimization again leads to cost-effective accuracy improvement.

4.5 Proofs

4.5.1 Theorem 4.3.4

We prove our results here in the case of $\beta_{\mathcal{Q}}$ defined as in (4.24), which due to the non-independent rows is a slight generalization of the results using (4.17). Adjusting the proofs present here to account for the latter case requires only the substitution of $\mu_{\mathcal{Q}}$ for $\beta_{\mathcal{Q}}$, showing this Theorem.

4.5.2 Lemma 4.3.1.

We begin by providing a brief proof for Lemma 4.3.1, which follows directly from the explicit form for the Hermite derivative as in (5.5.10) of [129].

Proof. Note that linearity of expectation allows us to take the expectation inside the sum, so that we may work on each term independently. For 1-dimensional orthonormal Hermite polynomials, where i represents the order of the polynomial (5.5.10) of [129] shows that

$$\frac{\partial \psi_i}{\partial \Xi}(\Xi) = \sqrt{i}\psi_{i-1}(\Xi). \tag{4.43}$$

As the tensor product of orthonormal polynomials is orthonormal, (4.21) for $\boldsymbol{i} = \boldsymbol{j}$ follows from the derivative being exactly in the direction of a Hermite polynomial. For $\boldsymbol{i} \neq \boldsymbol{j}$, (4.21) follows because each term in the sum is the expectation of a product of orthogonal polynomials that differ in at least one coordinate, and hence the integral is equal to zero. ■

4.5.3 Theorem 4.3.3

To prove Theorem 4.3.3, we appeal to a matrix variant of the Chernoff bound [139], which is similar to approaches taken in [132, 131, 140, 64].

Proof. Recall that $\mathcal{Q}$ represents a truncation for the domain of Ξ, and may be given by (4.18). We have that

$$\boldsymbol{I} = \mathbb{E}\left(\boldsymbol{X}^T\boldsymbol{X}|\boldsymbol{\xi} \in \mathcal{Q}\right)\mathbb{P}(\mathcal{Q}) + \mathbb{E}\left(\boldsymbol{X}^T\boldsymbol{X}|\boldsymbol{\xi} \in \mathcal{Q}^c\right)\mathbb{P}(\mathcal{Q}^c). \tag{4.44}$$

A brief calculation gives that

$$\begin{aligned} \epsilon_{\mathcal{Q}} &:= \left\|\mathbb{E}\left(\boldsymbol{X}^T\boldsymbol{X}|\boldsymbol{\xi} \in \mathcal{Q}\right) - \boldsymbol{I}\right\|_2 & (4.45)\\ &\leq \frac{\mathbb{P}(\mathcal{Q}^c)}{\mathbb{P}(\mathcal{Q})}\left(\left\|\mathbb{E}\left(\boldsymbol{X}^T\boldsymbol{X}\middle|\boldsymbol{\xi} \in \mathcal{Q}^c\right)\right\|_2 + 1\right), & (4.46)\end{aligned}$$

bounds the bias introduced from not accepting samples within $\mathcal{Q}^c$. We note that with the truncation in (4.18), and using this bound, $\epsilon_{\mathcal{Q}} \leq 0.1/\sqrt{P}$ [117], for all N considered here, and in most similar problems. Specifically, an analytic issue does not arise until consideration of exponential levels of oversampling. This analytic issue concerns with the truncated rare events being reliably observed due to the very large sample pool.

Restating (4.45), let $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ correspond to the smallest and largest eigenvalues of the argument matrix, respectively. Then for any arbitrary set of columns, denoted $\mathcal{S}$,

$$\begin{aligned} 1 - \epsilon_{\mathcal{Q}} &\leq \lambda_{\min}\left(\mathbb{E}\left(\boldsymbol{X}^T(:,\mathcal{S})\boldsymbol{X}(:,\mathcal{S})|\boldsymbol{\xi} \in \mathcal{Q}\right)\right) & (4.47)\\ &\leq \lambda_{\max}\left(\mathbb{E}\left(\boldsymbol{X}^T(:,\mathcal{S})\boldsymbol{X}(:,\mathcal{S})|\boldsymbol{\xi} \in \mathcal{Q}\right)\right) \leq 1 + \epsilon_{\mathcal{Q}}. & (4.48)\end{aligned}$$

From (4.24) we have that if each sample $\boldsymbol{\xi}^{(k)} \in \mathcal{Q}$, then for all k,

$$\|\boldsymbol{X}_k(:,\mathcal{S})\|_2^2 \leq s\beta_{\mathcal{Q}}, \tag{4.49}$$

holds uniformly for all choices of $\mathcal{S}$ such that $|\mathcal{S}| < s$. This provides an upper bound on the singular values of our independent self-adjoint matrices, $\boldsymbol{X}_k^T(:,\mathcal{S})\boldsymbol{X}_k(:,\mathcal{S})$, uniformly over all choices of $\mathcal{S}$ consisting of at most s elements. We define

$$\boldsymbol{M}_{\mathcal{S}} := \frac{1}{N}\sum_{k=1}^{N} \boldsymbol{X}_k^T(:,\mathcal{S})\boldsymbol{X}_k(:,\mathcal{S}). \tag{4.50}$$

An application of the Chernoff bound as in Theorem 1.1 of [139] and Theorem 1 of [132] gives that for $\delta \in [0, 1]$ and $|\mathcal{S}| \leq s$,

$$\mathbb{P}\left(\lambda_{\min}\left(\boldsymbol{M}_{\mathcal{S}}\right) \leq (1-\delta)(1-\epsilon_{\mathcal{Q}}) \middle| \boldsymbol{\xi}^{(k)} \in \mathcal{Q} \ \forall k\right) \leq s\left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\frac{N(1-\epsilon_{\mathcal{Q}})}{s\beta_{\mathcal{Q}}}}; \tag{4.51}$$

$$\mathbb{P}\left(\lambda_{\max}\left(\boldsymbol{M}_{\mathcal{S}}\right) \geq (1+\delta)(1+\epsilon_{\mathcal{Q}}) \middle| \boldsymbol{\xi}^{(k)} \in \mathcal{Q} \ \forall k\right) \leq s\left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\frac{N(1+\epsilon_{\mathcal{Q}})}{s\beta_{\mathcal{Q}}}}. \tag{4.52}$$

Note that

$$(1-\delta)(1-\epsilon_{\mathcal{Q}}) \geq 1-t \implies \delta \leq \frac{t-\epsilon_{\mathcal{Q}}}{1-\epsilon_{\mathcal{Q}}}; \tag{4.53}$$

$$(1+\delta)(1+\epsilon_{\mathcal{Q}}) \leq 1+t \implies \delta \leq \frac{t-\epsilon_{\mathcal{Q}}}{1+\epsilon_{\mathcal{Q}}}. \tag{4.54}$$

and so we have a critical δ, given by

$$\delta_t := (t-\epsilon_{\mathcal{Q}})/(1+\epsilon_{\mathcal{Q}}), \tag{4.55}$$

is such that for all $\delta < \delta_t$ the matrix $\boldsymbol{M}_{\mathcal{S}}$ is guaranteed to satisfy $\|\boldsymbol{M}_{\mathcal{S}} - \boldsymbol{I}\|_2 \leq t$. Note that for $0 \leq \delta < 1$,

$$\frac{e^{-\delta}}{(1-\delta)^{1-\delta}} \geq \frac{e^{\delta}}{(1+\delta)^{1+\delta}}, \tag{4.56}$$

and so we may bound the sum of the probabilities by

$$\mathbb{P}\left(\|\boldsymbol{M}_{\mathcal{S}} - \boldsymbol{I}\| \leq t \middle| \boldsymbol{\xi}^{(k)} \in \mathcal{Q} \ \forall k\right) \leq 2s\left(\frac{e^{-\delta_t}}{(1-\delta_t)^{1-\delta_t}}\right)^{\frac{N(1-\epsilon_{\mathcal{Q}})}{s\beta_{\mathcal{Q}}}}. \tag{4.57}$$

We now bound this probability. We note that if $\epsilon_{\mathcal{Q}} \leq t$ then

$$0 < \delta_t \leq \frac{t-\epsilon_{\mathcal{Q}}}{1+\epsilon_{\mathcal{Q}}} \leq t - \epsilon_{\mathcal{Q}}. \tag{4.58}$$

To create a bound without explicit dependence on $\epsilon_{\mathcal{Q}}$, we note that for

$$c_t := t - \epsilon_{\mathcal{Q}} + (t+\epsilon_{\mathcal{Q}})\log(t+\epsilon_{\mathcal{Q}});$$

$$\frac{e^{-\delta_t}}{(1-\delta_t)^{1-\delta_t}} \leq \exp(-c_t).$$

Thus for $t \in (0,1)$,

$$\mathbb{P}\left(\|\boldsymbol{M}_{\mathcal{S}} - \boldsymbol{I}\| \leq t \middle| \boldsymbol{\xi}^{(k)} \in \mathcal{Q} \ \forall k \right) \leq 2s \exp\left(-\frac{c_t(1-\epsilon_{\mathcal{Q}})N}{s\beta_{\mathcal{Q}}} \right); \tag{4.59}$$

$$\leq 2s \exp\left(-C_{\mathcal{Q}} \frac{tN}{s\beta_{\mathcal{Q}}} \right), \tag{4.60}$$

where $C_{\mathcal{Q}}$ is a reasonably large positive constant for most truncations. Via a union bound over the $\binom{P}{s}$ possibilities of subsets of P with cardinality s, it follows that

$$\mathbb{P}\left(\sup_{|\mathcal{S}| \leq s} \lambda_{\max}(\boldsymbol{M}_{\mathcal{S}} - \boldsymbol{I}) \geq t \right) \leq 2s \binom{P}{s} \exp\left(-C_{\mathcal{Q}} \frac{Nt}{s\beta_{\mathcal{Q}}} \right). \tag{4.61}$$

Recalling that

$$\sup_{|\mathcal{S}| \leq s} \lambda_{\max}(\boldsymbol{M}_{\mathcal{S}} - \boldsymbol{I}) = \delta_s,$$

gives

$$\mathbb{P}(\delta_s \geq t) \leq \exp\left(-C_{\mathcal{Q}} \frac{Nt}{s\beta_{\mathcal{Q}}} + \log\left(2s\binom{P}{s} \right) \right). \tag{4.62}$$

We assume that having any sample $\boldsymbol{\xi}^{(k)} \in \mathcal{Q}^c$ leads to an arbitrarily large $\lambda_{\max}$, hence yielding the bound,

$$\mathbb{P}(\delta_s \geq t) \leq 1 - \mathbb{P}(\mathcal{Q})^N + \exp\left(-C_{\mathcal{Q}} \frac{Nt}{s\beta_{\mathcal{Q}}} + \log\left(2s\binom{P}{s} \right) \right). \tag{4.63}$$

Using the relation, $\mathbb{P}(\delta_s < t) = 1 - \mathbb{P}(\delta_s \geq t)$, gives that

$$\mathbb{P}(\delta_s < t) \geq \mathbb{P}(\mathcal{Q})^N - \exp\left(-C_{\mathcal{Q}} \frac{Nt}{s\beta_{\mathcal{Q}}} + \log\left(2s\binom{P}{s} \right) \right). \tag{4.64}$$

Using the relation that

$$\binom{P}{s} \leq \left(\frac{eP}{s} \right)^s,$$

it follows that

$$2s\binom{P}{s} \leq 2se^s \left(\frac{P}{s} \right)^s; \tag{4.65}$$

$$\log\left(2s\binom{P}{s} \right) \leq \log(2s) + s + s\log(P/s), \tag{4.66}$$

which completes the proof. Remark 4.3.2 follows from taking $s = P$ and using that $\binom{P}{P} = 1$.
■

We note that Corollary 4.3.1 follows from Theorem 4.3.3, by substituting $p_\star$ for $\mathbb{P}(\delta_s < t)$; substituting $\delta_\star$ for t; and performing some algebraic manipulation.

4.5.4 Theorem 4.3.5

The proof of Theorem 4.3.5 relies on the properties of the measurement matrices $\tilde{\mathbf{\Psi}}$ and $\mathbf{\Psi}$ themselves. In an intuitive sense, the results follow from the two matrices having similar properties, but the gradient-enhanced matrix having more rows, yielding better conditioned Gramians.

Proof. We begin by showing **R1** for the one-dimensional case. Note that for arbitrary ξ, and i,

$$\frac{|\psi_i(\xi)|^2 + i|\psi_{i-1}(\xi)|^2}{1+i} \leq \max\left\{|\psi_i(\xi)|^2, |\psi_{i-1}(\xi)|^2\right\}$$

and that equality can only hold if $\psi_i(\xi) = \psi_{i-1}(\xi)$ for some i, which is an event that occurs with probability zero. As this inequality holds for all i and ξ, being strict for almost all ξ, **R1** follows for the one-dimensional case. The d-dimensional analogue follows as the d-dimensional polynomials are tensor-products of the one-dimensional polynomials.

To show **R2**, note that up to an invertible pre-multiplication, $\mathbf{\Psi}$ is a submatrix of $\tilde{\mathbf{\Psi}}$, and thus $\mathcal{N}(\mathbf{\Psi}) \subset \mathcal{N}(\tilde{\mathbf{\Psi}})$. Additionally, we notice that $\mathbf{\Psi}$, and $\tilde{\mathbf{\Psi}}$ are almost surely full rank matrices.

We next show **R3** in the case of 1-dimensional Hermite polynomials. Here subscripts of matrices refer to the column corresponding to that polynomial order. We have by the normalization (4.22) and (4.43) that

$$(\tilde{\mathbf{\Psi}}_i, \tilde{\mathbf{\Psi}}_j) = \frac{(\mathbf{\Psi}_i, \mathbf{\Psi}_j) + \sqrt{ij}(\mathbf{\Psi}_{i-1}, \mathbf{\Psi}_{j-1})}{\sqrt{(1+i)(1+j)}}. \tag{4.67}$$

It follows that,

$$|(\tilde{\mathbf{\Psi}}_i, \tilde{\mathbf{\Psi}}_j)| \leq \frac{|(\mathbf{\Psi}_i, \mathbf{\Psi}_j)| + \sqrt{ij}|(\mathbf{\Psi}_{i-1}, \mathbf{\Psi}_{j-1})|}{\sqrt{(1+i)(1+j)}}.$$

Applying a supremum,

$$\begin{aligned}\sup_{i\neq j} |(\tilde{\mathbf{\Psi}}_i, \tilde{\mathbf{\Psi}}_j)| &\leq \sup_{i\neq j} \frac{|(\mathbf{\Psi}_i, \mathbf{\Psi}_j)| + \sqrt{ij}|(\mathbf{\Psi}_{i-1}, \mathbf{\Psi}_{j-1})|}{\sqrt{(1+i)(1+j)}}, \\ &\leq \sup_{i\neq j} |(\mathbf{\Psi}_i, \mathbf{\Psi}_j)| \frac{1+\sqrt{ij}}{\sqrt{(1+i)(1+j)}},\end{aligned}$$

where we have used the inequality

$$\frac{1+\sqrt{(i-1)(j-1)}}{\sqrt{(1+(i-1))(1+(j-1))}} \leq \frac{1+\sqrt{ij}}{\sqrt{(1+i)(1+j)}}. \tag{4.68}$$

This shows the result for the one-dimensional case. The d-dimensional case leads to a decomposition in (4.67) with inner products of lower order along each dimension, and the inequality in (4.68), is replaced by d similar inequalities.■

4.6 Conclusion

Within the context of compressive sampling of sparse polynomial chaos expansions, we investigated ℓ_1-minimization when derivative information of a quantity of interest (QoI) is present. We provided analysis on gradient-enhanced ℓ_1-minimization for Hermite polynomial chaos, in which we showed that, for a given normalization, including derivative information will not reduce the stability of the ℓ_1-minimization problem. Further, we identified a coherence parameter that we used to bound the associated Restricted Isometry Constant, a useful and well-studied measure of the stability for solutions recovered by ℓ_1-minimization as in the context used here.

Furthermore, we observed improved solution accuracy from gradient-enhanced ℓ_1-minimization in three numerical examples: Manufactured polynomials; an elliptic equation; and a plane

Poiseuille flow with random boundaries. Consistently, gradient-enhanced ℓ_1-minimization was seen to improve the quality of solution recovery at the same computational cost, or equivalently achieve the same solution quality at a reduced computational cost. As the QoI derivatives are often more sensitive to discretization errors than the QoI itself, so too is the accuracy of the solution obtained by the gradient-enhanced ℓ_1-minimization. This was empirically observed in the second numerical example considered, thereby suggesting high accuracy requirements on derivative calculations for the gradient-enhanced ℓ_1-minimization to be most effective.

CHAPTER 5

DESIGN OF POLYNOMIAL CHAOS BASES FOR THE SOLUTION OF DIFFERENTIAL EQUATIONS WITH RANDOM INPUTS[1]

Abstract

Expansion of stochastic quantities of interest (QoIs) into a basis of orthogonal polynomials, referred to as polynomial chaos (PC), is now a standard technique for uncertainty quantification. The type of these polynomial bases has been conventionally chosen based the probability measure of random inputs and from the so-called Askey family of orthogonal polynomials. However, for an arbitrary QoI such an *a priori* choice of basis may result in slow decaying expansion coefficients, which in turn may lead to large errors when low order PC expansions are considered. Increasing the order of the truncated expansion may enhance the solution accuracy, however, at the expense of additional computation cost which may become prohibitive for complex systems. Alternatively, in this work, a design strategy is proposed to choose an *optimal* PC basis, within the family of Jacobi polynomials, and the corresponding change of measure using (random) realizations of QoI, in an *a posteriori* manner. To this end, a variable projection approach with alternating brute-force search is proposed to estimate the parameters of the Jacobi basis and the expansion coefficients. It is demonstrated that the proposed PC basis design leads to more rapidly decaying coefficients, hence reduces truncation error and enhances solution accuracy, relative to the PC basis naturally orthogonal with respect to the probability measure of inputs. Several numerical

[1] This chapter is in preparation to be submitted by *J. Peng et al.* to *Computer Methods in Applied Mechanics and Engineering.*

tests, with QoI's exhibiting sharp gradients/discontinuities, are provided to illustrate the performance of this approach.

5.1 Introduction

The credibility of computer simulations for design, analysis, and optimization of complex engineering systems is affected by the degree to which model uncertainties and their influences on quantities of interest (QoIs) are accounted for and measured. Model uncertainties, either parametric or structural, often arise due to, for instance, natural variability of the underlying physical quantities and/or our imperfect knowledge about them. The emerging field of uncertainty quantification (UQ) aims at developing numerical tools to characterize these uncertainties from the available information as well as efficiently propagating them for an accurate prediction of QoI and a quantitative validation of model predictions.

A common framework to represent uncertainty is based on the probability theory, where inputs are modeled by a vector of independent random variables $\boldsymbol{\xi} := (\xi_1, \ldots, \xi_d)$, with a probability density function (pdf) $f(\boldsymbol{\xi})$, defined on a suitable probability space with sample space Ω. The QoI, here a scalar function $u(\boldsymbol{\xi})$, therefore depends on $\boldsymbol{\xi}$, and the objective of uncertainty propagation is to approximate the map $\boldsymbol{\xi} \rightarrow u(\boldsymbol{\xi})$ directly and/or to estimate the statistics of $u(\boldsymbol{\xi})$. While several techniques are available for this purpose, see, e.g., [109, 3, 4], we here adopt an approach based on polynomial chaos (PC) expansions, [109, 20], where $u(\boldsymbol{\xi})$, assumed to have a finite variance, is expanded into a basis of multi-variate orthogonal polynomials $\Psi_i(\boldsymbol{\xi})$, i.e.,

$$u(\boldsymbol{\xi}) = \sum_{i=1}^{\infty} c_i \Psi_i(\boldsymbol{\xi}) \approx \sum_{i=1}^{P} c_i \Psi_i(\boldsymbol{\xi}). \tag{5.1}$$

Here, the size P of the truncated expansion is implied by the (total) degree p of $\Psi_i(\boldsymbol{\xi})$. The polynomials $\Psi_i(\boldsymbol{\xi})$ are conventionally selected to be orthogonal with respect to the measure $f(\boldsymbol{\xi})$ of the inputs $\boldsymbol{\xi}$, [20, 141]. For example, when $\boldsymbol{\xi}$ follows a jointly uniform or Gaussian distribution (with independent components), $\Psi_i(\boldsymbol{\xi})$ are multivariate Legendre or Hermite polynomials, respectively. For other instances of $f(\boldsymbol{\xi})$, $\Psi_i(\boldsymbol{\xi})$ may be chosen from the the so-called Askey family of orthogonal polynomials, [21, 20], or may be generated numerically, [142]. More details about the PC expansions considered in this work are provided in Section 5.2.2. The coefficients c_i specify the expansion in (5.1) and are given by the projection

$$c_i = \int u(\boldsymbol{\xi})\Psi_i(\boldsymbol{\xi})f(\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi} = \mathbb{E}[u(\cdot)\Psi_i(\cdot)], \tag{5.2}$$

where $\mathbb{E}$ denotes the mathematical expectation operator, and $\Psi_i(\boldsymbol{\xi})$ are assumed to be normalized such that $\mathbb{E}[\Psi_i^2(\cdot)] = 1$. A standard result from PC theory is that as $P \to \infty$, the truncated expansion in (5.1) converges to $u(\boldsymbol{\xi})$ in the mean-square sense, equivalently, $\sum_{P+1}^{\infty} c_i^2 \to 0$ as $P \to \infty$ and the rate of convergence depends how fast the coefficients c_i associated with higher order $\Psi_i(\boldsymbol{\xi})$ decay to zero.

In practice, the expectation in (5.2) is not available analytically; therefore, c_i has to be approximated via numerical integration, regression, or (Galerkin) projection when $u(\boldsymbol{\xi})$ is the solution to an operator equation, see, e.g., [3, 4]. In all these approaches, the accuracy of the estimates of c_i, denoted by $\hat{c}_i$, is limited by the level of noise/error in evaluating $u(\boldsymbol{\xi})$ as well as the truncation error $\epsilon_t(\boldsymbol{\xi}) := \sum_{i=P+1}^{\infty} c_i^2$, associated with the finite sum in (5.1). These errors may be reduced, respectively, by increasing the spatial/temporal resolution of deterministic solvers and the degree p of $\Psi_i(\boldsymbol{\xi})$ retained in (5.1). In this work we assume that the former error is negligible and focus on cases where the latter error is large, e.g., the c_i for the standard choice of $\Psi_i(\boldsymbol{\xi})$ does not decay to zero rapidly. Such scenarios arise, for instance, when the map $\boldsymbol{\xi} \to u(\boldsymbol{\xi})$ is *non-smooth*, e.g., exhibits discontinuities, sharp gradients, or bifurcations, [143, 144], or when $u(\boldsymbol{\xi})$ is obtained by a long-time integration of an operator equation, [145, 146]. We rewrite (5.1) as

$$\begin{aligned} u(\boldsymbol{\xi}) &= \sum_{i=1}^{P} c_i\Psi_i(\boldsymbol{\xi}) + \epsilon_t(\boldsymbol{\xi}) && (5.3)\\ &\approx \sum_{i=1}^{P} \hat{c}_i\Psi_i(\boldsymbol{\xi}) + \epsilon_t(\boldsymbol{\xi}), && (5.4)\end{aligned}$$

and highlight that, relative to the (optimal) c_i in (5.3), the accuracy of estimated $\hat{c}_i$ in (5.4) is in practice affected by $\epsilon_t(\boldsymbol{\xi})$. A result demonstrating such an influence is presented later in Theorem 5.2.1. The straightforward approach to enhance the accuracy is to increase the expansion order p; however, obtaining a higher order expansion requires additional computation cost which may become prohibitive for large p and d.

To improve the convergence of global PC expansions, a number of techniques have been proposed in recent years. In particular, the work in [147, 148, 149] proposes to enrich the global PC bases with basis functions that are specifically tailored to the non-smooth behavior, e.g., discontinuity, of the solution. Based on the observation that the pdf of the solution in an unsteady problem may considerably evolve from that of the initial solution, the work in [146] constructs PC basis orthogonal with respect to the solution pdf at selected time instances. Another approach – related to that of [146] – is the iterative generalized PC (i-gPC), [16, 150], which recursively uses the approximation of the QoI to generate a new probability density function and a corresponding set of orthogonal polynomials. In particular, i-gPC has been shown to significantly improve the convergence of the PC expansion for problems exhibiting discontinuities in the stochastic space.

Another class of techniques are based on (adaptive) partitioning of the support of $\boldsymbol{\xi}$ and *local* polynomial expansions – instead of the global expansion in (5.1) – over each partition, [151, 152, 153, 154, 155]. In this way, these methods seek to capture *features* of the solution present over a small subset of the support of $\boldsymbol{\xi}$ that may not be seen by global polynomial bases.

To enhance the convergence of PC expansions, we here present a different method that selects a PC basis directly using the solution realizations, so that the truncation error ϵ_t associated with a fixed expansion order p is reduced. The proposed PC *design* method selects a PC basis for which the coefficients $c_{\boldsymbol{i}}$ – and hence the truncation error – may converge towards zero more rapidly that for the standard PC expansion. The overall idea is to adaptively choose the proper orthogonal polynomials based on the realization of u.

In order for the problem to be tractable, we limit the choice of orthogonal polynomials to be from the Jacobi polynomial family. Therefore, the optimization problem becomes one of finding the optimal parameters that define the beta distribution with respect to which the Jacobi polynomials are orthogonal. We note that the Jacobi family includes familiar polynomial bases such as Legendre, Chebyshev, and Hermite. Therefore, when the random inputs follow a uniform distribution, for example, we anticipate that the proposed method performs in theory at least as well as Legendre PC approximation.

In the signal processing domain, several methods for adaptively designing collections of bases (often dubbed **dictionary matrices**) have gained recent attention. For instance, both the method of optimal directions (MOD) [156, 157] and K-singular value decomposition (K-SVD) [158, 159] have found successful application in image processing. The methods are of interest because the basis matrices used to represent the QoI are computed adaptively from the given data. However, these methods do not lead to *structured* basis matrices as in the PC expansions. More precisely, the resulting bases are not readily tensor (outer) products of one-dimensional orthogonal polynomials. Therefore, we do not consider such techniques for the problem at hand.

The method we propose in this manuscript is an alternating optimization approach to find the optimal parameters of the Jacobi basis dimension-wise, in which, according the variable projection technique, the PC coefficients are computed by the (regularized) least-squares regression. More specifically, we design optimal PC basis matrices within the Jacobi PC family with arbitrary measure of the random inputs. We compare our method with the traditional *a priori* way of choosing PC basis on several numerical experiments. Significant advantages to our approach is that it reduces the nominally high-dimensional optimization problem to a sequence of one-dimensional problems, and eliminates combination of linear and nonlinear optimization into nonlinear only. The results indicate that the proposed method improves the approximation accuracy over standard PC construction. We note here that different to i-gPC, the proposed approach builds optimal orthogonal PC basis directly with respect to the random inputs, while i-gPC iteratively builds orthogonal PC basis with respect to the approximated solution.

The rest of this paper is structured as follows. In Section 5.2, the problem setup and the details about PC expansion are given. In Section 5.3.1, we give the formulation of the beta distribution and Jacobi polynomials that we use to design the PC bases. The description of the algorithm and pseudo-code is given in Section 5.3.2. Three numerical experiments and their results are provided in Section 5.4.

5.2 Background

5.2.1 Problem statement

In this work, we consider systems modeled by differential equations defined on a domain $\mathcal{D} \in \mathbb{R}^D$, $D \in \{1, 2, 3\}$, in which the uncertainty characterized by the d-dimensional vector $\boldsymbol{\xi} = (\xi_1, \dots, \xi_d)$ may be represented in one or many relevant parameters, e.g., boundary conditions and/or initial conditions. Each coordinate of $\boldsymbol{\xi}$, denoted by ξ_k, $k = 1, \dots, d$, is defined on a probability space, such that $\boldsymbol{\xi}$ is defined on the probability space that is formed by their product. The solution u satisfies the following equations

$$\begin{aligned} &\mathcal{L}(\boldsymbol{x}, t, \boldsymbol{\xi}; u(\boldsymbol{x}, t, \boldsymbol{\xi})) = 0, && \boldsymbol{x} \in \mathcal{D}, \\ &\mathcal{I}(\boldsymbol{x}, 0, \boldsymbol{\xi}; u(\boldsymbol{x}, 0, \boldsymbol{\xi})) = 0, && \boldsymbol{x} \in \mathcal{D}, \\ &\mathcal{B}(\boldsymbol{x}, t, \boldsymbol{\xi}; u(\boldsymbol{x}, t, \boldsymbol{\xi})) = 0, && \boldsymbol{x} \in \partial\mathcal{D}, \end{aligned} \tag{5.5}$$

where $\mathcal{L}$, $\mathcal{I}$, and $\mathcal{B}$ are differential operators, depending on the physics of the problem. Our objective is to approximate the QoI, $u(\boldsymbol{x}, t, \boldsymbol{\xi})$, for some fixed spatial location $\boldsymbol{x}_0$ and time t_0. We denote the realizations of the random inputs by $\boldsymbol{\xi}^{(i)}$, thus the corresponding output is $u(\boldsymbol{x}_0, t_0, \boldsymbol{\xi}^{(i)})$. For brevity, we write the QoI and its realizations as $u(\boldsymbol{\xi})$ and $u(\boldsymbol{\xi}^{(i)})$, respectively.

5.2.2 Polynomial chaos (PC) expansion

We rely on PC expansions to approximate the QoI $u(\boldsymbol{\xi})$ to (5.5). In details, for the interest of presentation we assume that input random variables ξ_k are independent and identically distributed according to f_k, and define $\{\psi_{i_k}(\xi_k)\}$ to be the complete set of orthonormal polynomials of degree $i_k \in \mathbb{N} \cup \{0\}$ with respect to the weight function f_k [21, 20]. As a result, the orthonormal polynomials for $\boldsymbol{\xi}$ are given by the products of the univariate orthonormal polynomials,

$$\Psi_{\boldsymbol{i}}(\boldsymbol{\xi}) = \prod_{k=1}^{d} \psi_{i_k}(\xi_k), \tag{5.6}$$

where each $\boldsymbol{i} \in \{(i_1, \dots, i_d) : i_k \in \mathbb{N} \cup \{0\}\}$ is a d-dimensional multi-index of nonnegative integers. For computation, we truncate the expansion in (5.1) to the set of P basis functions associated with the subspace of polynomials of total order not greater than p, that is $\sum_{k=1}^{d} i_k \leq p$. For convenience, we also order these P basis functions so that they are indexed by $\{1, \dots, P\}$, as in (5.1), as opposed to the vectorized indexing in (1.1). The basis set $\{\Psi_i(\boldsymbol{\xi})\}_{i=1}^{P}$ has the cardinality

$$P = \frac{(d+p)!}{d!p!}. \tag{5.7}$$

Similarly, we define PC coefficients by the vector $\boldsymbol{c} = (c_1, \dots, c_P)^T$. We interchangeably use both notations for representing PC basis depending on the context.

5.2.3 PC expansion via least-squares regression

There are a number of sampling methods for estimating the PC coefficients including Monte Carlo simulation [23, 3], pseudo-spectral stochastic collocation [24, 25, 3, 27], least-squares regression [30, 160, 118, 161, 32], and ℓ_1-minimization for cases where $\boldsymbol{c}$ is approximately sparse [48, 47, 52, 53, 112, 162]. With these methods, the deterministic solvers for the QoI do not need to be adapted to the probability space and hence may be used in a black box fashion. In this work, we use the least-squares regression to compute the PC coefficients $\boldsymbol{c}$

by solving the optimization problem

$$\arg\min_{\boldsymbol{c}} \|\boldsymbol{u} - \boldsymbol{\Psi c}\|_2^2, \tag{5.8}$$

or its regularized formulation

$$\arg\min_{\boldsymbol{c}} \|\boldsymbol{u} - \boldsymbol{\Psi c}\|_2^2 + \lambda^2 \|\boldsymbol{c}\|_2^2, \tag{5.9}$$

where $\boldsymbol{u} := \left(u(\boldsymbol{\xi}^{(1)}), \dots, u(\boldsymbol{\xi}^{(N)})\right)^T$ is the vector of the realizations of the QoI u, and the entries of the $N \times P$ *measurement matrix* $\boldsymbol{\Psi}$ are the realizations of $\Psi_j(\boldsymbol{\xi})$ such that $\boldsymbol{\Psi}(i,j) = \Psi_j(\boldsymbol{\xi}^{(i)})$ for $i = 1, \dots, N$ and $j = 1, \dots, P$. Additionally, λ in (5.9) is the Tikhonov regularization parameter, which may be estimated using, for instance, the generalized cross-validation approach [163]. The solution to (5.8) and (5.9) may be computed from the normal equations $\boldsymbol{\Psi}^T\boldsymbol{\Psi c} = \boldsymbol{\Psi}^T\boldsymbol{u}$ and $(\boldsymbol{\Psi}^T\boldsymbol{\Psi} + \lambda^2\boldsymbol{I})\boldsymbol{c} = \boldsymbol{\Psi}^T\boldsymbol{u}$, respectively, where $\boldsymbol{I}$ is the $P \times P$ identity matrix.

The approximation of u is limited to the span of the basis polynomials of total degree at most p, and the error incurred from this approximation is referred to as the truncation error ϵ_t, specified in (5.3). As u has finite variance, the PC coefficients necessarily converge to zero when $p \to \infty$. How rapidly they converge to zero determines the accuracy of the truncated PC expansion as well as the required cost, here, the number of solution realizations. The following theorem, reported from [32], demonstrates a quality mean-squared convergence of the solution to (5.8), with high probability, when the PC basis is of Legendre type. More general presentations of this theorem, including results for Hermite PC expansions, may be found in [32].

Theorem 5.2.1 (Stability of Legendre PC expansions via least-squares regression, [32])**.** *Let $\boldsymbol{\xi}$ be a d-vector of independent random variables uniformly distributed over* $[-1,1]$ *and $u(\boldsymbol{\xi})$ a finite-variance function of $\boldsymbol{\xi}$. Let*

$$\hat{u}(\boldsymbol{\xi}) = \sum_{i=1}^{P} \hat{c}_i \Psi_i(\boldsymbol{\xi}) \tag{5.10}$$

be the PC expansion of $u(\boldsymbol{\xi})$ in Legendre polynomials $\Psi_i(\boldsymbol{\xi})$ of total degree not greater than p, where $\hat{\boldsymbol{c}} = (\hat{c}_1, \dots, \hat{c}_P)^T$ is computed from (5.8) using N realizations of $u(\boldsymbol{\xi})$ evaluated at independent samples of $\boldsymbol{\xi}$. It follows that

$$\mathbb{E}\left(\|u - \hat{u}\|^2_{L_2(\Omega,f)}\right) \leq \mathrm{Var}(\epsilon_t)\left(1 + \frac{4P\exp(2p)}{N}\right) \tag{5.11}$$

holds with probability $\mathbb{P} \geq 1 - 1/P - 2P\exp\left(-0.1N/\left(P\exp(2p)\right)\right)$, in which ϵ_t is the truncation error defined in (5.3), and $\mathrm{Var}(\epsilon_t)$ is the variance of ϵ_t.

We note that the expectation $\mathbb{E}$ in (5.11) corresponds to the variability of $\hat{u}(\boldsymbol{\xi})$ with respect to the random samples $\{u(\boldsymbol{\xi}^{(i)})\}_{i=1}^N$ used to solve (5.8). Following Theorem 5.2.1, the mean-squared error of the PC approximation grows proportional to $\mathrm{Var}[\epsilon_t]$. As mentioned earlier, the truncation error is determined by the decay rate of the PC coefficients, and so is its variance,

$$\mathrm{Var}(\epsilon_t) = \sum_{i=P+1}^{\infty} c_i^2. \tag{5.12}$$

When the total order p and the sample size N are fixed, if one can decrease the truncation error ϵ_t, then the error of the approximation may be reduced. A lower ϵ_t, and therefore a more accurate approximation of u, may be attained by reducing the magnitude of the truncated coefficients. However, such a reduction may not be achieved by adopting a different PC basis orthonormal with respect to $f(\boldsymbol{\xi})$, the pdf of $\boldsymbol{\xi}$. The following proposition demonstrates that the variance of ϵ_t is independent of the choice of PC basis that are orthonormal with respect to $f(\boldsymbol{\xi})$.

Proposition 5.2.1. *The truncation error associated with the approximation of $u(\boldsymbol{\xi})$ in any PC basis of maximum degree p and orthonormal with respect to $f(\boldsymbol{\xi})$ has the same variance.*

Proof. Let $u(\boldsymbol{\xi}) = \sum_{i=1}^P c_i\Psi_i(\boldsymbol{\xi}) + \epsilon_t(\boldsymbol{\xi})$ and $u(\boldsymbol{\xi}) = \sum_{i=1}^P \tilde{c}_i\tilde{\Psi}_i(\boldsymbol{\xi}) + \tilde{\epsilon}_t(\boldsymbol{\xi})$ denote, respectively, the PC expansions of $u(\boldsymbol{\xi})$ in two arbitrary PC bases $\{\Psi_i(\boldsymbol{\xi})\}$ and $\{\tilde{\Psi}_i(\boldsymbol{\xi})\}$ of total degree at most p. Both bases are assumed to be orthonormal with respect to $f(\boldsymbol{\xi})$, and $\epsilon_t(\boldsymbol{\xi})$

and $\tilde{\epsilon}_t(\boldsymbol{\xi})$ denote the corresponding truncation errors. It is straightforward to show that $\boldsymbol{c} = \boldsymbol{T}\tilde{\boldsymbol{c}}$, where $\boldsymbol{T}$ is a $P \times P$ orthogonal matrix with entries $\boldsymbol{T}(i,j) = \mathbb{E}(\psi_i \tilde{\psi_j})$, thus implying that $\|\boldsymbol{c}\|_2^2 = \|\tilde{\boldsymbol{c}}\|_2^2$. The statement of the Proposition follows by observing that $\mathrm{Var}(\epsilon_t) = \mathbb{E}(u^2) - \|\boldsymbol{c}\|_2^2 = \mathbb{E}(u^2) - \|\tilde{\boldsymbol{c}}\|_2^2 = \mathrm{Var}(\tilde{\epsilon}_t)$. □

Alternatively, the present study proposes a strategy to design the PC basis, along with a change of measure of inputs, that possibly make the corresponding coefficients decay more rapidly, thereby leading to a lower truncation error and enhanced PC approximation.

5.3 Design of PC basis: A Jacobi polynomial approach

In order to design a PC basis that leads to a lower truncation error, we generalize the least-square problem (5.8) to learn the basis matrix $\boldsymbol{\Psi}$, in addition to the coefficient vector $\boldsymbol{c}$, by solving the problem

$$\min_{\boldsymbol{c},\boldsymbol{\Psi}} \|\boldsymbol{u} - \boldsymbol{\Psi}\boldsymbol{c}\|_2^2 + \lambda\|\boldsymbol{c}\|_2^2. \tag{5.13}$$

Finding an optimal solution to (5.13) requires two special considerations. Firstly, the basis matrix $\boldsymbol{\Psi}$ must be *structured*; that is, it should consist of realizations of some multi-variate polynomials orthogonal with respect to some probability measure. Therefore, a direct minimization of (5.13) for an optimal $\boldsymbol{\Psi}$, as in [156, 157, 158, 159], is not possible. Alternatively, in this work, we limit the search for an optimal $\boldsymbol{\Psi}$ within the basis matrices corresponding to the family of Jacobi polynomials that are orthonormal with respect to an underlying beta pdf specified by d-vectors of parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. The problem (5.13) will therefore simplify to

$$\min_{\boldsymbol{c},\boldsymbol{\alpha},\boldsymbol{\beta}} \|\boldsymbol{u} - \boldsymbol{\Psi}^{(\boldsymbol{\alpha},\boldsymbol{\beta})}\boldsymbol{c}\|_2^2 + \lambda\|\boldsymbol{c}\|_2^2, \tag{5.14}$$

where the superscript $(\boldsymbol{\alpha},\boldsymbol{\beta})$ indicates the association of $\boldsymbol{\Psi}$ to the parameters of the underlying beta pdf's.

Secondly, both (5.13) and (5.14) are non-linear and non-convex programs which may not have unique solutions. While multiple approaches may be utilized to find an approximate (stationary) solution to (5.14), we here employ a variable projection approach [164, 165] in an alternating manner that allows us to find approximate solutions $\boldsymbol{c}, \boldsymbol{\alpha}, \boldsymbol{\beta}$ through a sequence of optimization problems with fewer number of unknowns. In detail, we make (5.14) be an optimization problem only dependent on the parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}$, by treating $\boldsymbol{c}$ as a function of them, then we iteratively search for optimal (α_k, β_k), $k = 1, \dots, d$ dimension-wise in a greedy manner, till the convergence criterion is satisfied. Such an alternating minimization approach requires a computation cost that grows linearly in d for the latter updates.

The following subsections present a detailed description of the proposed Jacobi PC design.

5.3.1 Beta random variables and Jacobi polynomials

Let $\tilde{\xi}$ denote a beta random variable with pdf $g^{(\alpha,\beta)}(\tilde{\xi})$ given by

$$g^{(\alpha,\beta)}(\tilde{\xi}) = \frac{(1-\tilde{\xi})^{\alpha}(1+\tilde{\xi})^{\beta}}{2^{\alpha+\beta+1}B(\alpha+1,\beta+1)}, \quad \tilde{\xi} \in [-1,1], \quad \alpha, \beta \in (-1, \infty), \tag{5.15}$$

where $B(\alpha,\beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$ is the beta function, and Γ is the gamma function. The parameters α and β make $g^{(\alpha,\beta)}$ and $\tilde{\xi}$ a family of densities and beta random variables, respectively.

The design method we propose restricts the PC basis to be within Jacobi PC family that are orthogonal with respect to $g^{(\alpha,\beta)}$. The corresponding univariate PC basis functions $\psi_i^{(\alpha,\beta)}(\tilde{\xi})$, parameterized by α and β, satisfy the three-term recurrence relation,

$$\begin{aligned} 2(i+1)(i+\alpha+\beta+1)(2i+\alpha+\beta)\psi_{i+1}^{(\alpha,\beta)}(\tilde{\xi}) = &\left[(2i+\alpha+\beta+1)(\alpha^2-\beta^2) + \frac{(2i+\alpha+\beta+2)!}{(2i+\alpha+\beta-1)!\tilde{\xi}}\right]\psi_i^{(\alpha,\beta)}(\tilde{\xi}) \\ &- 2(i+\alpha)(i+\beta)(2i+\alpha+\beta+2)\psi_{i-1}^{(\alpha,\beta)}(\tilde{\xi}), \end{aligned}$$

where i denotes the order of polynomials, and the first three polynomials are

$$\begin{aligned} &\psi_0^{(\alpha,\beta)}(\tilde{\xi}) = 1; \\ &\psi_1^{(\alpha,\beta)}(\tilde{\xi}) = \frac{1}{2}\left[2(\alpha+1) + (\alpha+\beta+2)(\tilde{\xi}-1)\right]; \\ &\psi_2^{(\alpha,\beta)}(\tilde{\xi}) = \frac{1}{8}\left[4(\alpha+1)(\alpha+2) + 4(\alpha+\beta+3)(\alpha+2)(\tilde{\xi}-1) + (\alpha+\beta+3)(\alpha+\beta+4)(\tilde{\xi}-1)^2\right]. \end{aligned} \tag{5.16}$$

Furthermore, these polynomials are normalized such that they are orthonormal with respect to $g^{(\alpha,\beta)}(\tilde{\xi})$,

$$\int \psi_i^{(\alpha,\beta)}(\tilde{\xi})\psi_j^{(\alpha,\beta)}(\tilde{\xi})g^{(\alpha,\beta)}(\tilde{\xi})\mathrm{d}\tilde{\xi} = \delta_{ij}, \tag{5.17}$$

where δ_{ij} is the Kronecker delta. We note that the Legendre and Chebyshev polynomials are special instances of Jacobi polynomials for various values of α and β, as shown in Table 5.1. In particular, when both α and β are large and approximately equal, beta distribution is approximately normal.

As in the standard PC construction in (5.6), the multivariate Jacobi PC functions are generated by the tensor product of univariate functions,

$$\Psi_{\boldsymbol{i}}^{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\tilde{\boldsymbol{\xi}}) = \prod_{k=1}^{d} \psi_{i_k}^{(\alpha_k,\beta_k)}(\tilde{\xi}_k). \tag{5.18}$$

To account for a possible anisotropic dependence of solution on $\tilde{\xi}_k$, we allow different parameters (α_k, β_k) for each $\tilde{\xi}_k$ in (5.18). Therefore, each Jacobi basis function $\Psi_{\boldsymbol{i}}^{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\tilde{\boldsymbol{\xi}})$ is uniquely identified by $2d$ parameters $(\boldsymbol{\alpha},\boldsymbol{\beta})$, $\boldsymbol{\alpha},\boldsymbol{\beta} \in (-1,\infty)^d$, which we learn from the observations of QoI, as described in Section 5.3.2.

Polynomial	α	β
Legendre	0	0
Chebyshev	-0.5	-0.5
Hermite	∞	∞

Table 5.1: Correspondence of the type of three known polynomials to the values of α and β in Jacobi polynomials.

5.3.2 Variable projection with alternating brute-force search

We find an approximate solution to the regression problem in (5.14) by a separable nonlinear least squares approach, *variable projection* [164, 165]. In variable projection, with given $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, $\boldsymbol{c}$ is obtained by solving the linear least-squares problem

$$\boldsymbol{c}(\boldsymbol{\alpha},\boldsymbol{\beta}) = \boldsymbol{\Psi}^{(\boldsymbol{\alpha},\boldsymbol{\beta})^\dagger}\boldsymbol{u}, \quad \boldsymbol{\Psi}^{(\boldsymbol{\alpha},\boldsymbol{\beta})^\dagger} = (\boldsymbol{\Psi}^{(\boldsymbol{\alpha},\boldsymbol{\beta})^T}\boldsymbol{\Psi}^{(\boldsymbol{\alpha},\boldsymbol{\beta})} + \lambda^2\boldsymbol{I})^{-1}\boldsymbol{\Psi}^{(\boldsymbol{\alpha},\boldsymbol{\beta})^T}, \tag{5.19}$$

which stands for the minimum-norm solution of the linear least-squares problem for fixed $(\boldsymbol{\alpha},\boldsymbol{\beta})$. Substituting (5.19) back to (5.14), the minimization problem takes the form of finding the optimal $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ by

$$\boldsymbol{\alpha},\boldsymbol{\beta} = \underset{\boldsymbol{\alpha},\boldsymbol{\beta}}{\arg\min} \left\| \left(\boldsymbol{I} - \boldsymbol{\Psi}^{(\boldsymbol{\alpha},\boldsymbol{\beta})}\boldsymbol{\Psi}^{(\boldsymbol{\alpha},\boldsymbol{\beta})\dagger}\right)\boldsymbol{u}\right\|_2^2. \tag{5.20}$$

The optimization problem in (5.20) is non-linear and non-convex. Motivated by the tensor-product structure of the basis $\Psi_{\boldsymbol{i}}^{(\boldsymbol{\alpha},\boldsymbol{\beta})}$, an alternating optimization approach may be devised to approximate (5.20) via a sequence of two-dimensional (non-linear) optimizations. In particular, for a given direction k, the pair α_k,β_k may be updated from

$$\alpha_k,\beta_k = \underset{\alpha_k,\beta_k}{\arg\min} \left\| \left(\boldsymbol{I} - \boldsymbol{\Psi}^{(\boldsymbol{\alpha},\boldsymbol{\beta})}\boldsymbol{\Psi}^{(\boldsymbol{\alpha},\boldsymbol{\beta})\dagger}\right)\boldsymbol{u}\right\|_2^2, \tag{5.21}$$

while fixing $(\alpha_{\hat{k}\neq k},\beta_{\hat{k}\neq k})$ at their current values. The optimization problem (5.21) is repeated for each direction k in each iteration ℓ, until the convergence criterion is reached, i.e., $\left\| \left(\boldsymbol{I} - \boldsymbol{\Psi}^{(\boldsymbol{\alpha},\boldsymbol{\beta})}\boldsymbol{\Psi}^{(\boldsymbol{\alpha},\boldsymbol{\beta})\dagger}\right)\boldsymbol{u}\right\|_2^2$ does not decrease throughout the $\boldsymbol{\alpha},\boldsymbol{\beta}$ updates.

Brute-force selection of $(\alpha_k^{\ell+1},\beta_k^{\ell+1})$ The optimization problem (5.21) is non-linear, non-convex, and possibly with multiple local minima. These limit the applicability of standard gradient-based algorithms, although gradients of $\boldsymbol{\Psi}^{(\boldsymbol{\alpha},\boldsymbol{\beta})}$ with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are available analytically, e.g., following [166]. We therefore adopt a brut-force approach where we search for an approximate global minimum $(\alpha_k^{\ell+1},\beta_k^{\ell+1})$ among a set of discrete values $\{(\alpha_k^{(i)},\beta_k^{(j)})\}$, $i,j=1,\dots,M$. That is, we set

$$(\alpha_k^{\ell+1},\beta_k^{\ell+1}) = \underset{\{(\alpha_k,\beta_k)\in(\alpha_k^{(i)},\beta_k^{(j)})\}}{\arg\min} \left\| \left(\boldsymbol{I} - \boldsymbol{\Psi}^{(\boldsymbol{\alpha},\boldsymbol{\beta})}\boldsymbol{\Psi}^{(\boldsymbol{\alpha},\boldsymbol{\beta})\dagger}\right)\boldsymbol{u}\right\|_2^2, \qquad i,j=1,\dots,M. \tag{5.22}$$

In our numerical experiments, we choose $\alpha_k^{(i)}$ and $\beta_k^{(j)}$ from a two-dimensional uniform grid of size $M\times M$ over the square domain $(-1,1)^2$. The upper limits of this search space may be extended to values beyond unity at the expense of increasing the cost of finding an optimum $(\alpha_k^{\ell+1},\beta_k^{\ell+1})$. We next delineate the approach we follow to solve (5.21) and an adaptation of it that empirically leads to more accurate basis design.

We note that there is a difference between the approach of (5.20)-(5.22) and the methods presented in [164, 165]. In particular, as opposed to the alternating minimization approach presented here, in [164, 165] all the parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}$ are computed simultaneously, which may result in expensive computation when d is large. Additionally, in the present work, the selection of optimal parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}$ relies on a brute-force search over a set of candidates, (5.22), while in [164, 165] this is achieved by a gradient-based optimization scheme. In Section 5.4.3, we show that for a case when $d = 8$, the present approach outperforms both simultaneous and alternating gradient-based methods. Furthermore, our PC basis design strategy relies on a change of measure associated with $\boldsymbol{\alpha}, \boldsymbol{\beta}$ updates that is not present in the standard variable projection technique. A detailed comparison between the two approaches is a subject of an ongoing study.

5.3.3 Change of measure

The standard PC expansion utilizes basis functions that are orthogonal with respect to the pdf of the inputs $\boldsymbol{\xi}$. While not necessary, the orthogonality of PC basis $\Psi_i(\boldsymbol{\xi})$ results in measurement matrices $\boldsymbol{\Psi}$ whose columns are orthogonal in expectation. This, along with other properties of $\Psi_i(\boldsymbol{\xi})$, ensures the stability of the least squares problem (5.8), as described in [32]. However, in the PC design approach of Section 5.3.2, the Jacobi basis functions are not known **a priori** and change throughout the $\boldsymbol{\alpha}, \boldsymbol{\beta}$ updates. Thus, to maintain orthogonality of the columns in $\boldsymbol{\Psi}$ (in expectation), our approach involves mapping the original random inputs ξ_k to beta random variables $\tilde{\xi_k}$ with parameters α_k and β_k, in which the Jacobi basis is constructed. For this purpose, we use the cumulative distribution functions (cdf's) of ξ_k and $\tilde{\xi_k}$, receptively detonated by $F(\xi_k)$ and $G^{(\alpha_k,\beta_k)}(\tilde{\xi_k})$. Specifically, for any trial values of $(\alpha_k^{(i)}, \beta_k^{(j)})$ in (5.22), we set

$$\tilde{\xi_k} = \left(G^{(\alpha_k^{(i)},\beta_k^{(j)})}\right)^{-1}(F(\xi_k)) = \left(\left(G^{(\alpha_k^{(i)},\beta_k^{(j)})}\right)^{-1} \circ F\right)(\xi_k). \tag{5.23}$$

We note that, although we change the measure of each random input ξ_k adaptively based on the values of $(\alpha_k^{(i)}, \beta_k^{(j)})$, the realizations of QoI u are untouched.

Remark 5.3.1. *While the Jacobi basis functions $\Psi_{\boldsymbol{i}}^{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\tilde{\boldsymbol{\xi}})$ in (5.18) are polynomials in $\tilde{\boldsymbol{\xi}}$, they are generally not polynomial functions of $\boldsymbol{\xi}$, given the nonlinear transformation (5.23). Therefore, by Jacobi 'polynomial' chaos expansion of u, we refer to the representation of u in a series of Jacobi polynomials in $\tilde{\boldsymbol{\xi}}$.*

Remark 5.3.2. *We note that the integral statistics of u, e.g., mean and variance, may be computed under the measure of either $\boldsymbol{\xi}$ or $\tilde{\boldsymbol{\xi}}$. In particular, let $\hat{u}(\boldsymbol{\xi}) = \sum_{i=1}^{P} \hat{c}_i \Psi_i^{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\tilde{\boldsymbol{\xi}})$ denote the optimal Jacobi PC expansion of u. The approximate mean and variance of u are given by $\mathbb{E}(\hat{u}) = \hat{c}_1$ and $Var(\hat{u}) = \sum_{i=2}^{P} \hat{c}_i^2$, respectively.*

5.3.4 Algorithm

We summarize the steps required for the implementation of the Jacobi PC design in Algorithm 5. As in the case of standard least-squares regression, the Jacobi basis design requires the specification of the order of the expansion p, as well as the samples $\boldsymbol{\xi}^{(i)}$ and the corresponding realizations of QoI $u(\boldsymbol{\xi}^{(i)})$, $i = 1, \ldots, N$. In the numerical examples of Section 5.4, the samples of $\boldsymbol{\xi}$ are generated according to its joint pdf $f(\boldsymbol{\xi})$. In Algorithm 5, the parameter $\boldsymbol{\alpha}, \boldsymbol{\beta}$ are updated from the optimization problem (5.22), while the PC coefficients $\boldsymbol{c}$ are computed via (5.19).

Algorithm 5 Jacobi PC Design via Variable Projection

```
Inputs: Order of expansion $p$; realizations of inputs, $\boldsymbol{\xi}^{(i)}$, and QoI, $u(\boldsymbol{\xi}^{(i)})$, $i = 1, \dots, N$; and joint cdf of $\boldsymbol{\xi}$, $F(\boldsymbol{\xi})$; a grid $\{(\alpha_k^{(i)}, \beta_k^{(j)})\}$, $i, j = 1, \dots, M$, for parameters $(\alpha_k, \beta_k)$, $k = 1, \dots, d$, of Jacobi basis.
• Set $\ell = 0$ and the initial guess for optimal parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta})$, e.g., $(\boldsymbol{\alpha}^\ell, \boldsymbol{\beta}^\ell) = (\mathbf{0}, \mathbf{0})$.
Repeat
    • $\tilde{\xi}_k \leftarrow \left(\left(G^{(\alpha_k^l, \beta_k^l)}\right)^{-1} \circ F\right)(\xi_k)$, $k = 1, \dots, d$,    (Change of measure from Eq. (5.23))
    • Generate $\boldsymbol{\Psi}^{(\boldsymbol{\alpha}^\ell, \boldsymbol{\beta}^\ell)}$ using realizations of $\tilde{\boldsymbol{\xi}}$
    for $k = 1 : d$ do
        for $(\alpha_k,\, \beta_k) \in \{(\alpha_k^{(i)}, \beta_k^{(j)})\}$ do
            • $\tilde{\xi}_k \leftarrow \left(\left(G^{(\alpha_k^{(i)}, \beta_k^{(j)})}\right)^{-1} \circ F\right)(\xi_k)$,    (Change of measure from Eq. (5.23))
            • Fix $(\alpha_{\hat{k}}, \beta_{\hat{k}})$, $\hat{k} \neq k$, at their latest values and update $\boldsymbol{\Psi}^{(\boldsymbol{\alpha}, \boldsymbol{\beta})}$ using realizations of $\tilde{\xi}_k$
            • $(\alpha_k^{\ell+1}, \beta_k^{\ell+1}) \leftarrow \arg\min_{(\alpha_k, \beta_k) \in \{(\alpha_k^{(i)}, \beta_k^{(j)})\}} \left\| \left(\boldsymbol{I} - \boldsymbol{\Psi}^{(\boldsymbol{\alpha}, \boldsymbol{\beta})} \boldsymbol{\Psi}^{(\boldsymbol{\alpha}, \boldsymbol{\beta})^\dagger}\right) \boldsymbol{u} \right\|_2^2$, with $\boldsymbol{\Psi}^{(\boldsymbol{\alpha}, \boldsymbol{\beta})^\dagger}$ defined in (5.19),
        End for
    End for
    • $\ell \leftarrow \ell + 1$
Until $\left\| \left(\boldsymbol{I} - \boldsymbol{\Psi}^{(\boldsymbol{\alpha}^\ell, \boldsymbol{\beta}^\ell)} \boldsymbol{\Psi}^{(\boldsymbol{\alpha}^\ell, \boldsymbol{\beta}^\ell)^\dagger}\right) \boldsymbol{u} \right\|_2^2$ remains unchanged or $\ell > \ell_{\max}$
Outputs: $\boldsymbol{\alpha}^\ell$, $\boldsymbol{\beta}^\ell$, and $\boldsymbol{c} = \boldsymbol{\Psi}^{(\boldsymbol{\alpha}^\ell, \boldsymbol{\beta}^\ell)^\dagger} \boldsymbol{u}$
```

We note that, in Algorithm 5, the optimal Tikhonov regularization parameter λ may change when the Jacobi basis is updated. However, here, we estimate λ using the GCV approach only at each time when k alternates throughout all d dimensions and use this estimate throughout the ℓth iteration.

Remark 5.3.3. *The search for optimal parameters* $(\alpha_k^{\ell+1}, \beta_k^{\ell+1})$ *in (5.22) requires the eval-*

uation of the cost function $\left\| \left(\boldsymbol{I} - \boldsymbol{\Psi}^{(\boldsymbol{\alpha},\boldsymbol{\beta})} \boldsymbol{\Psi}^{(\boldsymbol{\alpha},\boldsymbol{\beta})^{\dagger}} \right) \boldsymbol{u} \right\|_2^2$ *at all the grid points* $\{(\alpha_k^{(i)}, \beta_k^{(j)})\}$, *which can be done in parallel. The code used for the numerical experiments of Section 5.4 is implemented with multi-threading.*

5.4 Numerical experiments

In this section, we empirically demonstrate the accuracy of the Jacobi PC design in estimating statistics of solution to three differential equations with random inputs. In all three cases, the solution of interest features sharp gradients or discontinuities with respect to the random inputs, thus resulting in slow convergence of standard PC expansions. In the first test case, we also provide a comparison between the quality of the approximation obtained from the Jacobi PC design and the i-gPC approach.

5.4.1 Case I: Ordinary differential equation with stochastic coefficient

First, we consider the following ordinary differential equation, representing exponential population decay with a random reproduction rate [20, 146],

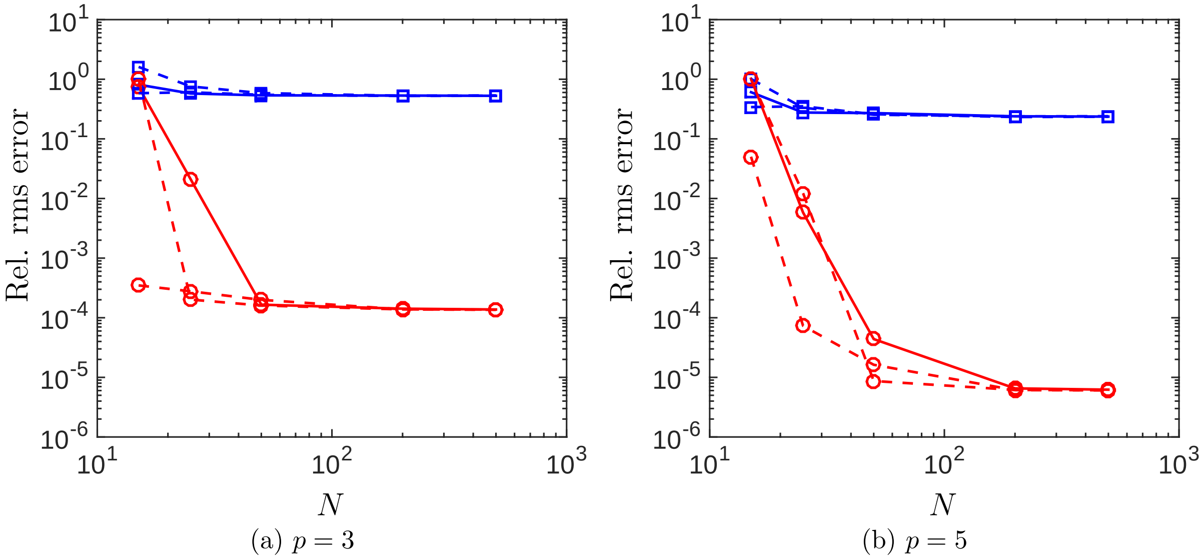
$$\frac{du(t,\xi)}{dt} + \kappa(\xi) u(t,\xi) = 0 \qquad (t,\xi) \in [0,T] \times \Omega, \tag{5.24}$$

with initial condition $u(t=0) = 1$. The reproduction rate $\kappa(\xi)$ is considered to be a random variable uniformly distributed over $[0,1]$ and given by $\kappa(\xi) = 1/2 + 1/2\xi$, where $\xi \sim U[-1,1]$. The QoI is the solution $u(t=25,\xi)$ at time $t = 25$. The analytic solution to (5.24) is given by

$$u(t,\xi) = e^{-\kappa(\xi)t}, \tag{5.25}$$

and thus the statistics of u can be computed exactly. It is known that the accuracy of a fixed degree Legendre expansion of $u(t,\xi)$ deteriorates as a function of time t, [146], i.e., the so-called *long-time integration* issue.

We next compare the quality of approximation obtained from Legendre PC and the Jacobi PC design approaches, using multiple numbers, $N = 50, 200, 500$, of random solution realizations. To examine the dependency on the choice of the solution realizations, for each N, we perform three independent replications of Legendre PC and the Jacobi PC design following Algorithm 5.



(a) $p = 3$

(b) $p = 5$

Figure 5.1: Comparison of relative root mean square (rms) error with 10,000 validation samples of $u(t = 25)$, with PC expansions of different orders p. (a) $p = 3$; (b) $p = 5$. (Legendre PC; Jacobi PC design; dashed lines denote two replications of expansion with independent solution realizations.)

To compare the proposed approach with Legendre PC, we examined the error in approximation of $u(t = 25)$ computed from $10,000$ validation samples, generated independently from the N training samples. From Fig. 5.1, we observe that Jacobi PC design improves the accuracy by multiple orders of magnitude. Additionally, the achieved accuracy becomes less dependent on the samples used when N is sufficiently large, depending on the order of expansion p. To further demonstrate this improvement, in Fig. 5.2, we compare the cdf of the solution at $p = 4$ and $N = 200$. In Fig. 5.3, we plot the PC expansion coefficients of $u(t = 25)$ in both Legendre and optimal Jacobi bases. We observe that the Jacobi expansion achieves a faster decay in the coefficients and hence smaller truncation error in accordance with the objective of the Jacobi PC design.

As mentioned earlier, change of measure contributes in improving PC approximation accuracy. In Figure 5.4, the optimal Jacobi basis functions are shown in terms of the original random input ξ. We note that, as opposed to the Legendre polynomials, these functions are mostly variable over $[-1, 0]$ where the solution of interest attains larger values.

Next we compare the results from i-gPC (with 10 iterations) with those obtained from the proposed method. For $p = 3, 5$ and for increasing values of N, we observe from Fig. 5.5 that, while i-gPC is more accurate than the Legendre PC, it is not as accurate as the optimal Jacobi PC approximation.

5.4.2 Case II: Woodward-Colella flow problem

We next consider the solution to a compressible channel flow problem with a forward facing step as shown in Fig. 5.6. This is a Riemann problem governed by the Euler equations and was previously studied in [167, 168]. The inflow Mach number, M, and heat capacity ratio, γ, are assumed to be independent, uniform random variables. Specifically, $M \sim U[2.4565, 3.0551]$ and $\gamma \sim U[1.35, 1.45]$, and to set up our experiment we consider $M = 2.7558 + 0.2993\xi_1$ and $\gamma = 1.40 + 0.05\xi_2$, where ξ_1, ξ_2 are independent and uniformly distributed over $[-1, 1]$. The QoI is chosen to be the pressure around the second Mach reflection point on the step, i.e., at location $(1.6, 0.2)$ in Fig. 5.6, where the left bottom corner point of the flow domain is assumed to be the origin $(0, 0)$. The density field corresponding to $(M, \gamma) = (2.57, 1.4)$ is also shown in Fig. 5.6.

For this experiment, we report the solution accuracy as a function of the expansion order $p = 2, \ldots, 6$ and using $N = 100, 300$ solution realizations. The exact QoI, as a function of (M, γ) is shown in Fig. 5.7a, from which we note the large gradients along γ leading to the Gibbs phenomenon in both PC representations, Figs. 5.7b and 5.7c. We used $p = 5$ and same samples of size $N = 200$ to generate the response surfaces in Figs. 5.7b and 5.7c. As may be observed from these results, the Gibbs phenomenon is significantly less severe in the case of the Jacobi PC design, as compared to the standard Legendre expansion. The more quantitative results in Fig. 5.8, obtained with $N = 100, 300$ samples, illustrate the higher accuracy of the proposed method relative to the Legendre approximation. In particular, with sample size $N = 300$, the Jacobi PC approximation improves the approximation error by about one order of magnitude. In this case, the solution shows less sensitivity to the actual samples utilized. We also compare the pdf of the approximate solutions in Fig. 5.9, from which we observe that the optimal Jacobi PC approach leads to a more accurate pdf.

Looking at the PC coefficients of both approximations in Fig. 5.10, we can see that those of the Jacobi PC decays slightly faster than for Legendre, leading to a lower truncation error. This is a direct consequence of the optimal Jacobi PC construction, where the solution residual is minimized further by allowing a basis possibly different from the one orthogonal to the measure of the inputs.

In Fig. 5.11, the univariate Jacobi basis functions obtained with $N = 300$ samples are shown in terms of the original uniform random variables ξ_1 and ξ_2. In particular, the basis functions along ξ_1, as displayed in Fig. 5.11a, while not actually polynomials are qualitatively similar to the Legendre basis. Along the same direction, the QoI varies smoothly as illustrated in Fig. 5.7a. However, the basis functions along ξ_2, shown in Fig. 5.11b, are drastically different and vary the most in the neighborhood of $\xi_2 = 0$, where the shockwave exists. The rather *local* variation of the Jacobi basis in this direction leads to an enhanced recovery of the QoI.

5.4.3 Case III: A kinetic problem with stochastic reaction rates

In this experiment, we apply our approach to a hydrogen oxidation problem previously studied in [169, 17]. Here the evolution of seven species, namely [OH], [H], [H_2O], [H_2], [O_2], [HO_2], and [H_2O_2], is governed by eight reversible reactions with stochastic reaction rates. Following [17], each of these reaction rates is modeled as an independent, uniformly distributed random variable, specified in Table 5.2. More precisely, $k_{f,j}$, $j = 1, \ldots, 8$, denotes the forward reaction rate, and the corresponding reverse rate $k_{r,j}$ is determined by the deterministic equilibrium constants $K_{c,j}$ via

$$k_{r,j} = K_{c,j}^{-1} \times k_{f,j}, \;\; j = 1, \ldots, 8.$$

The forward reaction rates $k_{f,j}$ are linear transformations of independent random variables ξ_j uniformly distributed in $[-1, 1]$, such that $k_{f,j}$ follow the distributions given in Table 5.2.

Index	Reaction	Distribution of $k_{f,j}$	Equilibrium constant $K_{c,j}^{-1}$
$j = 1$	$OH{+}H \rightleftharpoons H_2O$	$U[4.68 \times 10^{13}, 4.67 \times 10^{14}]$	0.3491×10^{-30}
$j = 2$	$H_2 + OH \rightleftharpoons H_2O + H$	$U[5.00 \times 10^{11}, 7.93 \times 10^{11}]$	0.4380×10^{-3}
$j = 3$	$H + O_2 \rightleftharpoons HO_2$	$U[5.26 \times 10^{13}, 1.31 \times 10^{14}]$	0.1045×10^{-08}
$j = 4$	$HO_2 + HO_2 \rightleftharpoons H_2O_2 + O_2$	$U[5.16 \times 10^{11}, 1.03 \times 10^{12}]$	0.9879×10^{-13}
$j = 5$	$H_2O_2 + OH \rightleftharpoons H_2O + HO_2$	$U[2.20 \times 10^{12}, 5.48 \times 10^{12}]$	0.3382×10^{-08}
$j = 6$	$H_2O_2 + H \rightleftharpoons HO_2 + H_2$	$U[8.48 \times 10^{10}, 3.39 \times 10^{11}]$	0.7723×10^{-05}
$j = 7$	$H_2O_2 \rightleftharpoons OH + OH$	$U[1.26 \times 10^{1}, 1.26 \times 10^{2}]$	$0.1589 \times 10^{+12}$
$j = 8$	$OH + HO_2 \rightleftharpoons H_2 + O_2$	$U[1.24 \times 10^{13}, 1.24 \times 10^{14}]$	0.3534×10^{-17}

Table 5.2: Random reaction model (Units are in the cm-mol-s-K system).

The evolution of species concentration is governed by the non-linear system of first-order ODEs,

$$\begin{aligned} \frac{d\boldsymbol{X}(t,\boldsymbol{\xi})}{dt} &= F(\boldsymbol{X}(t,\boldsymbol{\xi}),\boldsymbol{\xi}), \\ \boldsymbol{X}(0,\boldsymbol{\xi}) &= \boldsymbol{X}_0, \end{aligned} \tag{5.26}$$

where the operator F is governed by the reactions and the vector $\boldsymbol{X}$ consists of the species concentrations, i.e.,

$$\boldsymbol{X} = ([\mathrm{OH}],\; [\mathrm{H}],\; [\mathrm{H_2O}],\; [\mathrm{H_2}],\; [\mathrm{O_2}],\; [\mathrm{HO_2}],\; [\mathrm{H_2O_2}])\,.$$

As in [17], the following initial condition $\boldsymbol{X}_0$ is considered: $[H_2](t = 0) = 2.06{\times}10^{-6}\mathrm{mol/cm}^3$, $[O_2](t = 0) = 1.04{\times}10^{-6}\mathrm{mol/cm}^3$, and $[H_2O](t = 0) = 4.281{\times}10^{-3}\mathrm{mol/cm}^3$. All other initial concentrations are zero. All reactions are assumed to occur at fixed temperature $T = 823K$ and pressure $P = 246$ bar.

The QoI of the problem is chosen to be the concentration of $[H_2O_2]$ at the equilibrium condition. In Fig. 5.12, we are showing the coefficients in the Legendre expansion with total order $p = 6$, computed from 100,000 samples of $[H_2O_2]$ concentration. From Fig. 5.12, we note that the Legendre PC coefficients decay slowly. In other words, this representation has a high truncation error.

To explore the performance of the optimal Jacobi basis, we approximate the concentration of $[H_2O_2]$ with multiple values of p and accordingly with increasing sample sizes N. In Fig. 5.12, we contract the optimal Jacobi expansion coefficients against the Legendre counterpart, demonstrating a slightly faster decay and a smaller truncation error. To compare the two reconstructions quantitatively, we independently generate another 10,000 samples, and use the two expansions to predict these samples. In Fig. 5.13, we compare the relative rms error with the two approximations, from which we observe that with the optimal Jacobi expansion, the approximation accuracy is improved. As total order p increases, the accuracy is further improved. We note that the sample sizes N in that figure are sufficient, but not necessary, to get converged solutions. Besides the rms error, we also compare the probability density function of the QoI in Fig. 5.14. We observe that the optimal Jacobi PC leads to a more accurate probability density function than the Legendre PC. It is worthwhile highlighting that the Legendre PC expansion even gives highly negative, hence non-physical, realizations of the concentration that are due to large approximation errors. This is much less of an issue in the optimal Jacobi PC results.

Additionally, we also approximate the concentration of $[H_2O_2]$ with optimal Jacobi basis found by gradient-based optimization. In Fig. 5.15, we compare the relative rms error in predicting the 10,000 samples with Legendre and the optimal Jacobi basis, which are replications of the experiment shown in Fig. 5.13, except that the algorithms that are used to calculate the optimal parameters for Jacobi basis are replaced with gradient-based optimization. In Fig. 5.15, opposite to the situation shown in Fig. 5.13, no significant accuracy improvement is observed, which indicates that for this problem, gradient-based optimization does not perform as good as brute-force search in calculating the values of parameters associated with the optimal Jacobi basis. We note that in the gradient-bases optimization, the derivatives may be obtained both analytically and numerically, in these cases, we use finite difference to approximate the derivatives.

5.5 Conclusion

Within the context of polynomial chaos (PC) approximation of stochastic differential equations, we introduced an alternating least-squares (ALS) approach to design the PC basis within the Jacobi polynomial family. The PC basis is designed from an optimization over the Jacobi parameters without additional solution realizations. We argue that with the optimal Jacobi PC basis the approximation may have more rapidly decaying coefficients leading to a lower truncation error compared to conventional PC approximation. We noted that as conventional choices of basis such as Legendre and Chebyshev PC are a special cases of Jacobi family, the optimal Jacobi PC must have equal or better performance compared with these PC expansions. To test the performance of the proposed PC basis design approach, we applied it to three numerical test cases: an ordinary differential equation with stochastic coefficient, the Woodward-Colella flow problem, and a kinetic problem with stochastic reaction rates. In all three cases, the optimal Jacobi PC outperforms the Legendre PC, i.e. the polynomials orthogonal to the measure of the inputs.

Although, the optimal Jacobi PC improves the accuracy in these test cases, this approach may not apply to all stochastic differential equations. For QoIs that do not lend themselves to accurate representations in tensor product basis, the proposed PC design may not be effective.

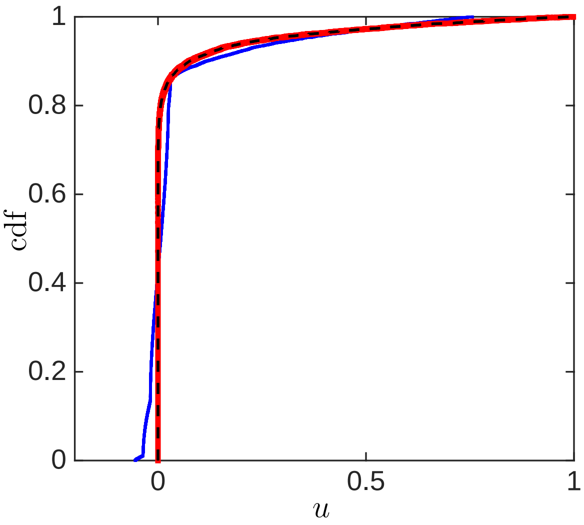


Figure 5.2: Comparison of cumulative distribution function (cdf) of Legendre and Jacobi approximations to $u(t = 25)$ with $p = 4$ and $N = 200$ (— Legendre PC; — Optimal Jacobi PC; ----- Exact).

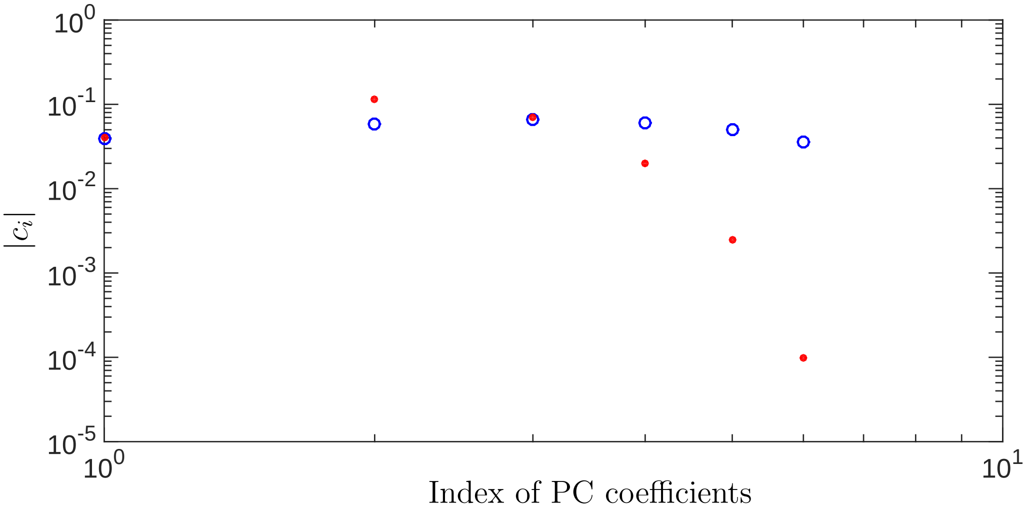


Figure 5.3: Comparison of relative rms error computed via Legendre PC and the optimal Jacobi PC with 10,000 validation samples of $u(t = 25)$, where $p = 5$ and $N = 200$ (○ Legendre PC; ● Jacobi PC design).

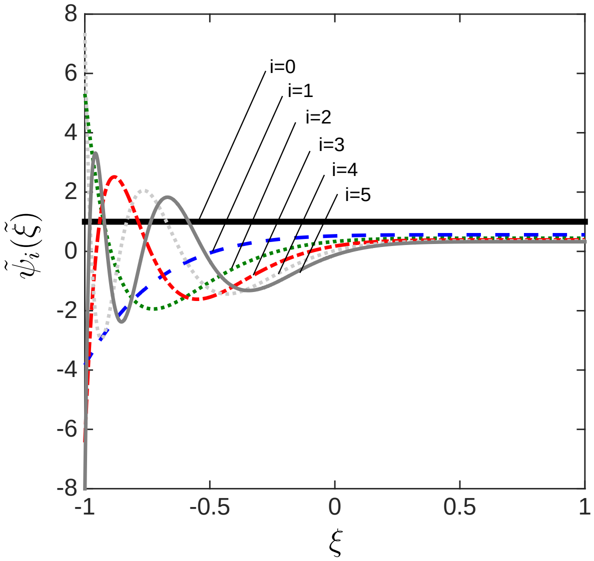
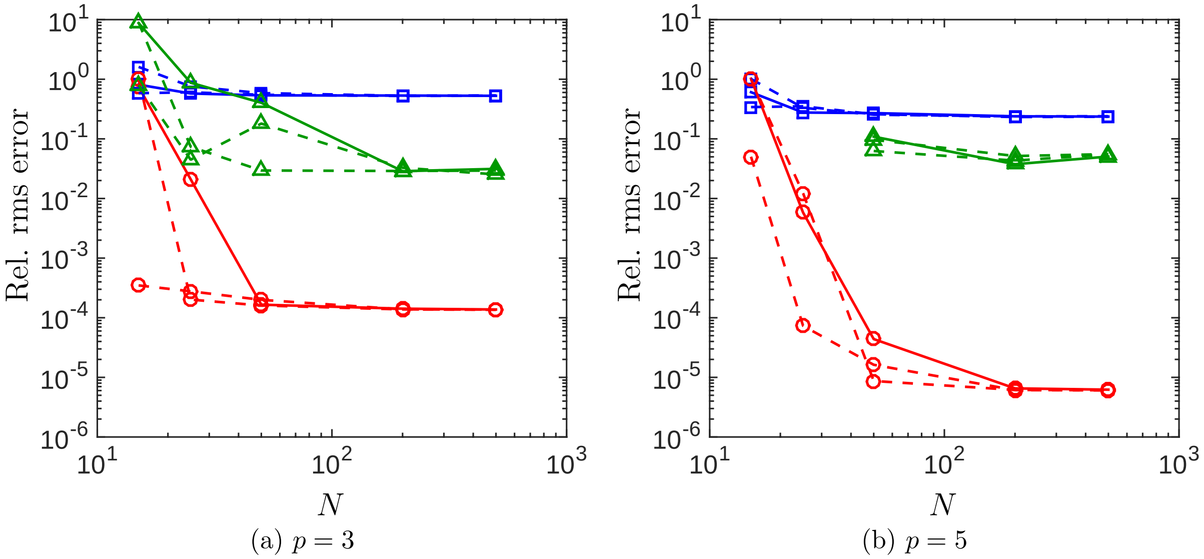


Figure 5.4: Optimal basis functions $\psi_i^{(\alpha,\beta)}(\tilde{\xi}) = \psi_i^{(\alpha,\beta)}\left(\left(G^{(\alpha,\beta)}\right)^{-1} \circ F(\xi)\right)$ obtained from Jacobi design with $p = 5$ and $N = 500$ for the solution to Case I.



(a) $p = 3$

(b) $p = 5$

Figure 5.5: Comparison of relative rms error computed via Legendre PC, the optimal Jacobi PC, and i-gPC expansions with 10,000 validation samples of $u(t = 25)$ (—□— Legendre PC; —○— Optimal Jacobi PC; —△— i-gPC; dashed lines denote tow independent replications with different choice of realizations).

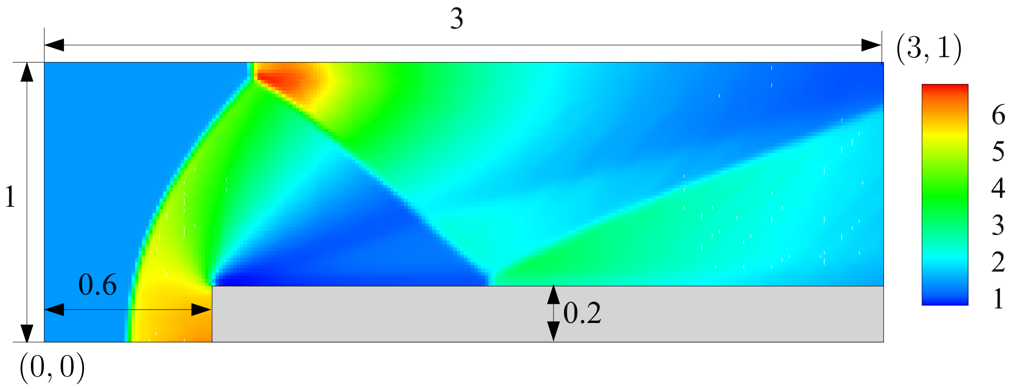
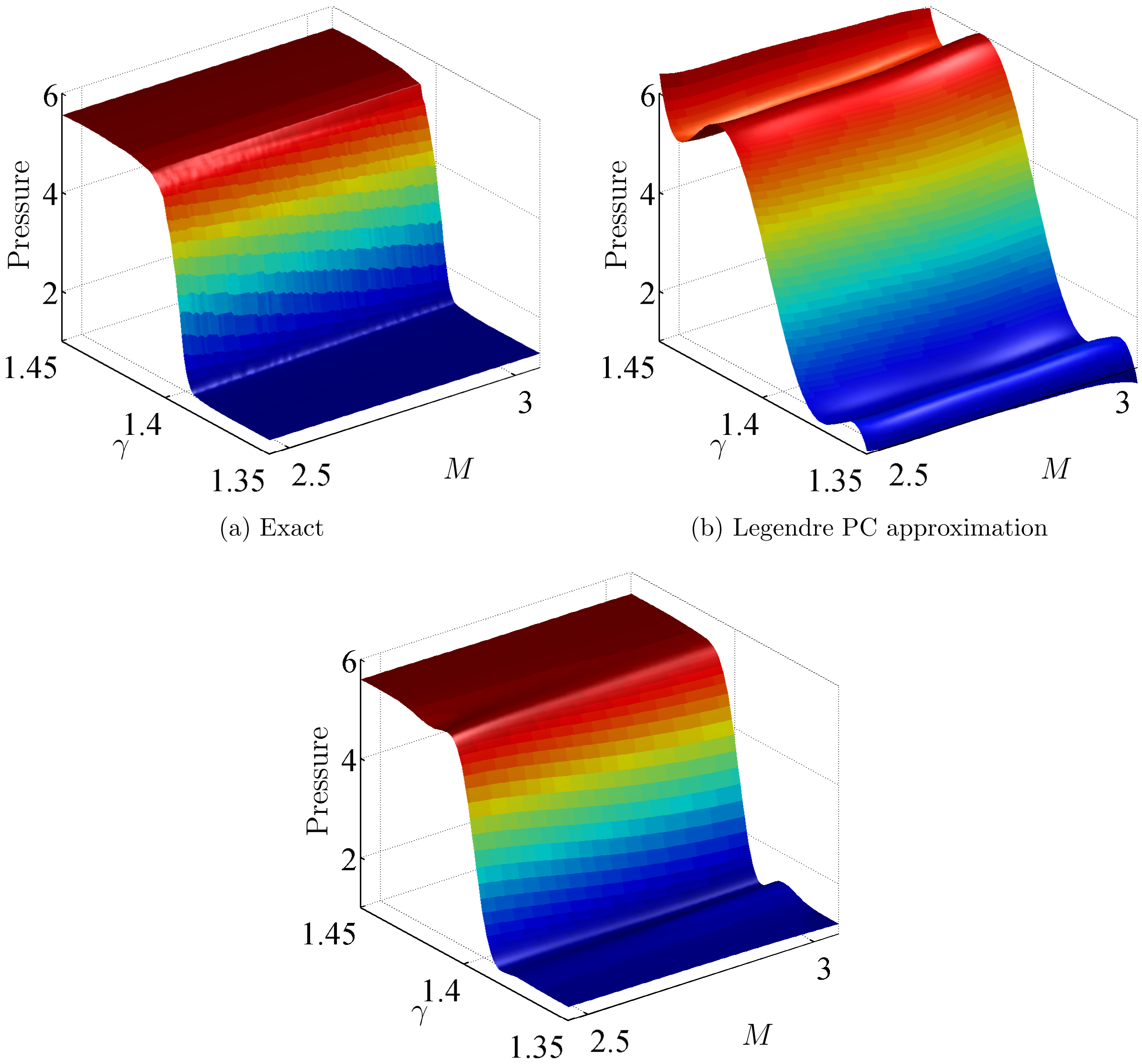


Figure 5.6: Schematics and density field of Woodward-Colella flow with $M = 2.57$ and $\gamma = 1.4$.

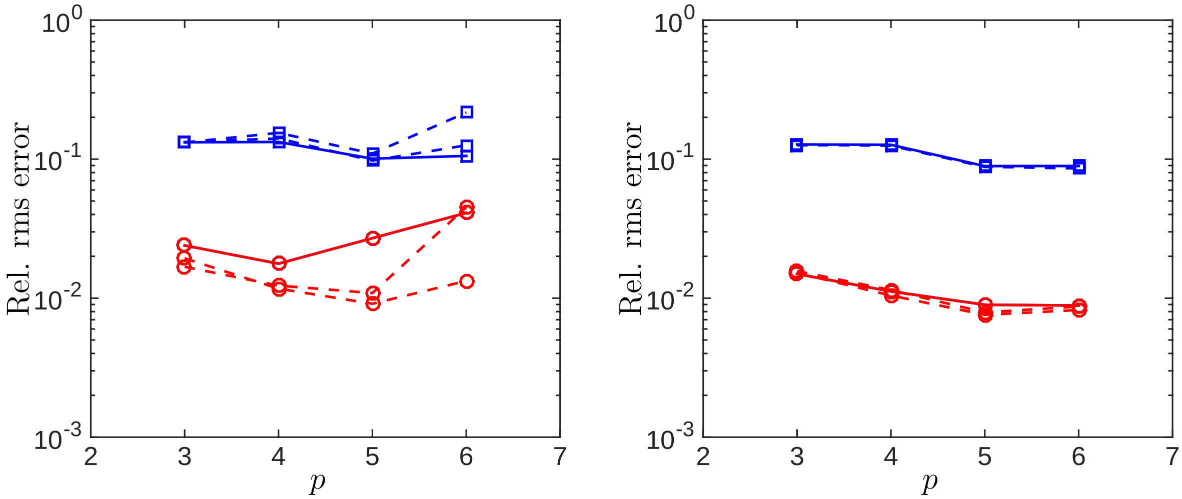


(a) Exact

(b) Legendre PC approximation

(c) Jacobi PC approximation

Figure 5.7: Response surface of the pressure at the second Mach refection point with respect to γ and M. The Legendre and Jacobi PC response surfaces are obtained with $p = 5$ and same solution realizations of size $N = 200$.



(a) $N = 100$

(b) $N = 300$

Figure 5.8: The error in predicting 10,000 independent samples for the pressure at the second Mach reflection point on the step. (—□— Legendre PC; —○— Optimal Jacobi PC; dashed lines denote two independent replications with different realizations of inputs).

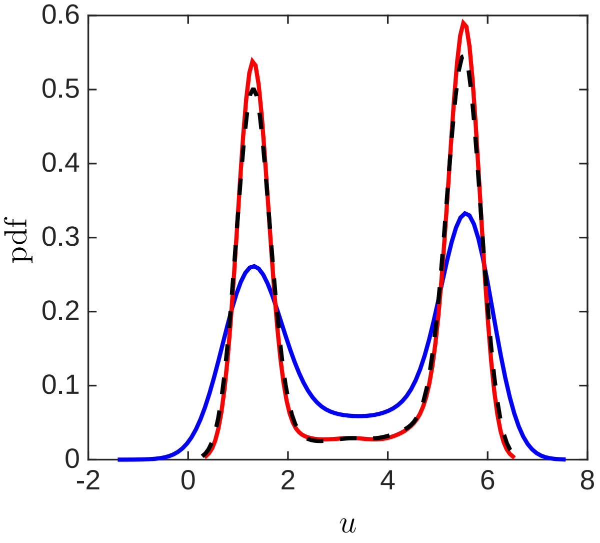


Figure 5.9: Comparison of probability density functions with Legendre and Jacobi expansions of the pressure at the second Mach reflection point, with $p = 6$ and $N = 300$ (— Legendre PC; — Optimal Jacobi PC; ----- Exact).

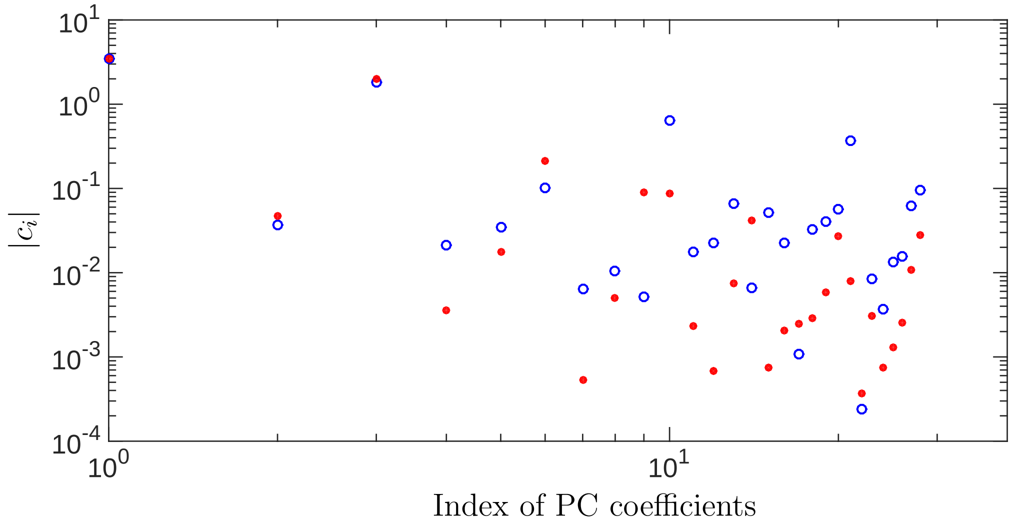


Figure 5.10: Comparison of PC coefficients obtained with $p = 6$ and $N = 300$ (○ Legendre PC; ● Optimal Jacobi PC).

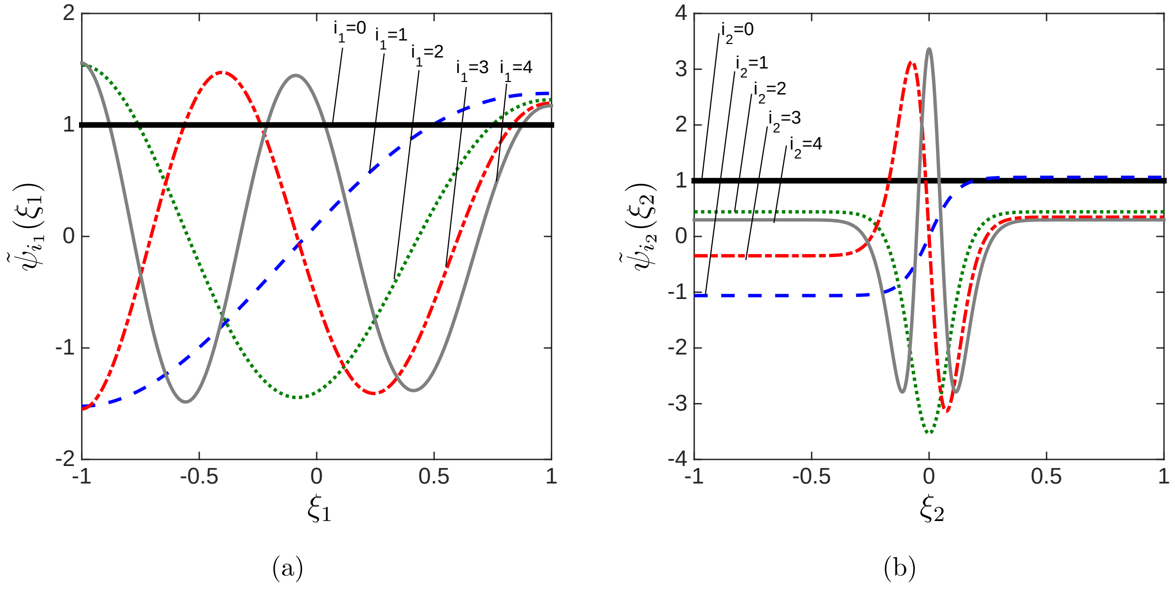


Figure 5.11: Optimal basis functions $\psi_{i_k}^{(\alpha_k,\beta_k)}(\tilde{\xi}_k) = \psi_{i_k}^{(\alpha_k,\beta_k)}\left(\left(G^{(\alpha_k,\beta_k)}\right)^{-1} \circ F(\xi_k)\right)$ for $k = 1$, (a), and $k = 2$, (b), obtained from Jacobi design with $p = 6$ and $N = 300$ for the solution to Case II. Only the first four basis functions are displayed.

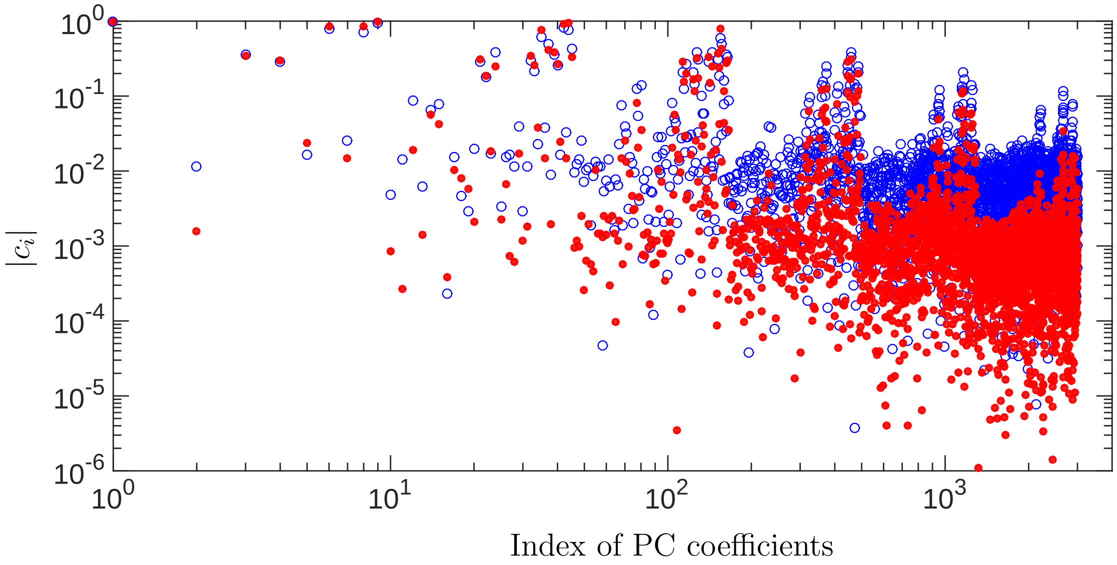


Figure 5.12: Coefficients of Legendre and optimal Jacobi expansions, where the coefficients are normalized such that $|c_1| = 1$ (○ Legendre PC; ● Optimal Jacobi).

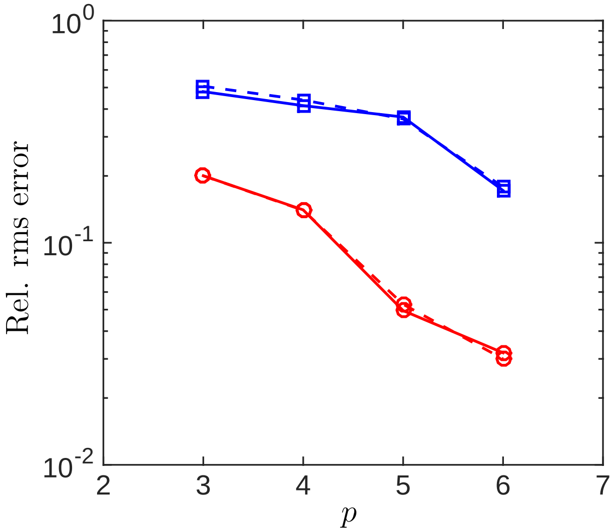
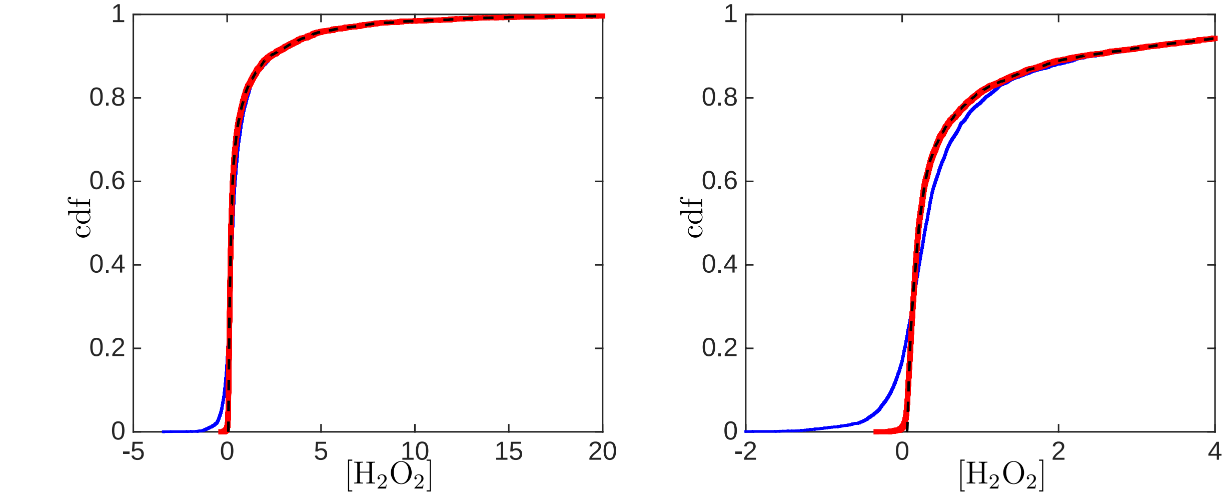


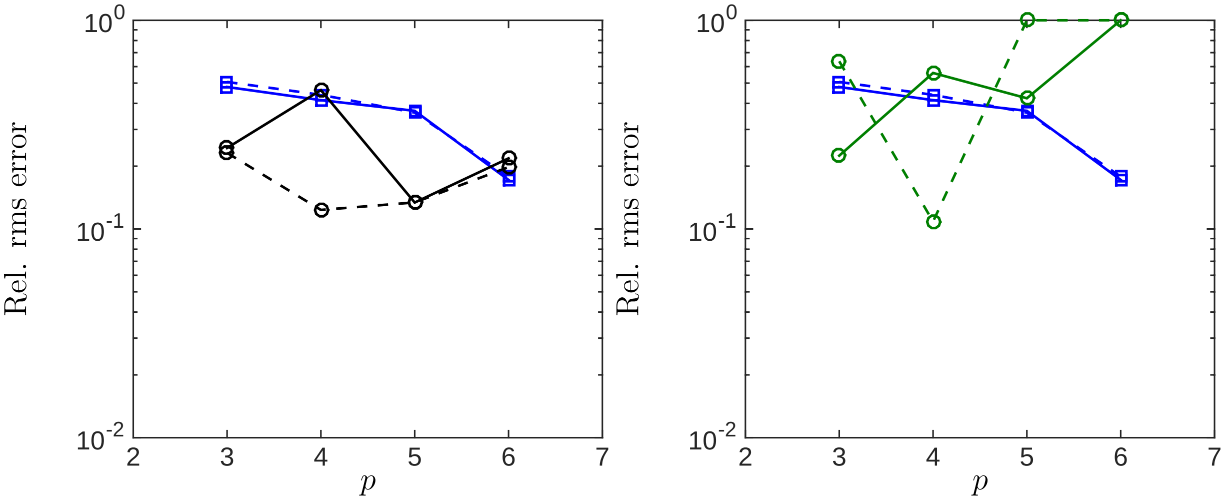
Figure 5.13: Comparison of relative rms error in predicting 10,000 independent samples. For Legendre and optimal Jacobi expansions of total order $p = 3, 4, 5$, and 6, we respectively used solution realizations of sizes $N = 500, 1000, 2000$, and 5000 (-□- Legendre PC; -○- Optimal Jacobi PC; dashed lines denote an independent replications of solution with independent input uncertainty).



(a) Entire cdf

(b) cdf zoomed around [H_2O_2]=0.

Figure 5.14: Comparison of cumulative distribution function (cdf) of [H_2O_2] approximated via Legendre PC and the optimal Jacobi PC with $p = 6$ and $N = 2000$ (— Legendre PC; — Optimal Jacobi PC; ----- Exact).



(a) Variable projection with gradient-based optimization simultaneously in all dimensions.

(b) Variable projection with dimension-alternating gradient-based optimization.

Figure 5.15: Comparison of relative rms error in predicting 10,000 independent samples. For Legendre and optimal Jacobi expansions of total order $p = 3, 4, 5,$ and 6, we respectively used solution realizations of sizes $N = 500, 1000, 2000,$ and 5000 (–□– Legendre PC; –○– Optimal Jacobi PC whose parameters are found simultaneously by gradient-based optimization; –○– Optimal Jacobi PC whose parameters are found dimension-wise by gradient-based optimization; dashed lines denote an independent replications of solution with independent input uncertainty).

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

6.1 Concluding remarks

Within the context of compressive sampling of sparse polynomial chaos expansions (PCEs), I introduced three approaches to enhance the quality in quantitatively approximating the quantity of interest (QoI). In addition, for QoI's that an *a priori* choice of basis may result in slow decaying expansion coefficients, I introduced a PCE basis design approach to enhance the sparsity of the corresponding coefficients, which in turn improve the quality of the approximation.

In Chapter 2, I introduced a *weighted ℓ_1-minimization* approach, wherein I utilized *a priori* knowledge on PC coefficients to enhance the accuracy of the standard ℓ_1-minimization. The *a priori* knowledge of PCE coefficients may be available in the form of analytical decay of PCE coefficients, e.g., for a class of linear elliptic PDEs with random data, or derived from simple dimensional analysis. These *a priori* estimates, when available, can be used to establish weighted ℓ_1 norms that will further penalize small PC coefficients, and consequently improve the sparse approximation. I provided analytical results guaranteeing the convergence of the weighted ℓ_1-minimization approach. The performance of the proposed weighted ℓ_1-minimization approach was demonstrated through its application to two test cases, and in both cases I demonstrated that the weighted ℓ_1-minimization approach outperforms the non-weighted counterpart.

In Chapter 3, it was assumed that the analytical or approximate *a priori* information about the PCE coefficients were not available, thus I utilized bi-fidelity technique to provide *a priori* information about the polynomial chaos expansion (PCE) coefficients on the quantity of interest (QoI), within the context of compressive sampling. Furthermore, I employed a weighted ℓ_1-minimization and modified the orthogonal matching pursuit (OMP) algorithm to include this *a priori* information, therefore to improve the accuracy in recovering the PCE coefficients. In addition, I provided analysis on weighted ℓ_1-minimization,for Legendre polynomial chaos. Numerical examples were shown to demonstrate the bi-fidelity methods, and in all cases, the bi-fidelity approaches were observed to outperform standard appraoches.

In Chapter 4, I investigated ℓ_1-minimization when derivative information of a QoI with respect to the random inputs is present. I provided analysis on gradient-enhanced ℓ_1-minimization for Hermite polynomial chaos, in which I showed that, for a given normalization, including derivative information will not reduce the stability of the ℓ_1-minimization problem. Further, I identified a coherence parameter that I used to bound the associated Restricted Isometry Constant, a useful and well-studied measure of the stability for solutions recovered by ℓ_1-minimization as in the context used here. Furthermore, I observed improved solution accuracy from gradient-enhanced ℓ_1-minimization in three numerical examples. Consistently, gradient-enhanced ℓ_1-minimization was seen to improve the quality of solution recovery at the same computational cost, or equivalently achieve the same solution quality at a reduced computational cost.

For all cases above, the type of polynomial bases is chosen *a priori* based the probability measure of random inputs and from the so-called Askey family of orthogonal polynomials. However, for an arbitrary QoI such an *a priori* choice of basis may result in slow decaying expansion coefficients, which in turn may lead to large errors when low order PCEs are considered. Hence, in Chapter 5, I introduced an alternating least squares (ALS) approach to design the PCE basis within the Jacobi polynomial family. The PCE basis is designed from an optimization over the Jacobi parameters without additional solution realizations. I showed that with the optimal Jacobi PCE basis the approximation may have more rapidly decaying coefficients leading to a lower truncation error compared to conventional PCE approximation. I noted that as conventional choices of basis such as Legendre and Chebyshev PCE are a special cases of Jacobi family, the optimal Jacobi PC must have equal or better performance compared with these PCE. To test the performance of the proposed PC basis design approach, I applied it to three numerical test cases, and in all cases, the optimal Jacobi PCE outperforms the Legendre PCE.

6.2 Future work

This section briefly describes future plans that may address the unresolved issues in the ongoing work or extend the present work.

Recovery Guarantees for Bi-fidelity ℓ_1-minimization

In Chapter 3, I demonstrated an improved quality in approximating the QoI via weighted ℓ_1-minimization, where the *a priori* is provided by the low-fidelity PCE coefficients. The theoretical recovery guarantee for weighted ℓ_1-minimization was presented under the assumption that the *a priori* information about the PCE coefficients are accurate. However, in our numerical experiments, it was observed that the low-fidelity PCE coefficients are often inaccurate, but they still improve the coefficients recovery in high fidelity via weighted ℓ_1-minimization.

To justify the observed phenomenon, theoretical supports are needed to show that even when the *a priori* information is inaccurate, the recovery via weighted ℓ_1-minimization is still feasible. Promisingly, this may be done via the concept *weighted restricted isometry constant* introduced in [170], and the *coherence* defined in [162].

PCE Basis Design via ℓ_1-minimization

In Chapter 5, the *optimal* PCE basis was designed via the alternating least squares (ALS) approach, which generally requires a large number of realizations $N \gg P$, where P is the number of basis functions. However, for a problem whose deterministic solver has high computational complexity, when the dimensionality of the random inputs is high, it is therefore unrealistic to evaluate the quantity of interest (QoI) N times. In addition, with large N, the design process becomes computationally costly too. Nevertheless, it has been observed that the *optimal* basis enhances the sparsity of the corresponding PCE, and that compressive sampling approaches recover the PCE coefficients efficiently with $N < P$. It is thus reasonable to consider replacing the ALS approach with a compressive sampling approach such as ℓ_1-minimization, to reduce the computational expense in both evaluating the QoI and basis design.

I seek to employ iteratively reweighted least squares (IRLS) approaches, in which the solution from previous iteration is utilized to weight the ℓ_2 norm of the current solution, in PCE basis design. It has been shown that in each iteration, IRLS has unique solution, and the solution converges to the solution to $\mathcal{P}_{1,\epsilon}$ in (1.19) [87, 66], i.e., IRLS can be used for ℓ_1-minimization.

Numerical experiments are needed to verify this idea of combining PCE basis design and ℓ_1-minimization. If positive results can be observed, theoretical justifications show the conditions that guarantee the success are needed.

Joint Approximation of QoI's via Sparse PCE

In all the methods and numerical experiments presented in this thesis, the QoI's were chosen at some fixed spatial location $\boldsymbol{x}_0$ and time t_0. In realistic, there are often quantities at multiple spatial and/or temporal points that are of interest, as they may have same or similar sparsity structure with respect to their PCEs. Taking advantages of these shared sparsity structure may be helpful in reducing the number of realizations required and/or lowering the computational cost in recovering the PCE coefficients.

For the case that the sparse PCE coefficients for all QoI's have the same support, it has been shown that algorithms concern with multiple measurement vectors (MMV) improve the performance of compressive sampling, both theoretically and empirically [171, 172]. However, the extent to the cases where the supports for the PCE coefficients are not exactly the same is missing. New approaches are in need to benefit from correlated sparse coefficients supports.

Furthermore, theorems and numerical experiments are needed to justify and verify these approaches.

BIBLIOGRAPHY

[1] E.J. Candès, M.B. Wakin, and S. Boyd. Enhancing sparsity by reweighted ℓ_1 minimization. Journal of Fourier Analysis and Applications, 14(5):877–905, 2008.

[2] R. Ghanem and P. Spanos. Stochastic Finite Elements: A Spectral Approach. Springer Verlag, 1991.

[3] O.P. Le Maitre and O. Knio. Spectral Methods for Uncertainty Quantification with Applications to Computational Fluid Dynamics. Springer, 2010.

[4] D. Xiu. Numerical Methods for Stochastic Computations: A Spectral Method Approach. Princeton University Press, 2010.

[5] R. Ghanem and S. Dham. Stochastic finite element analysis for multiphase flow in heterogeneous porous media. Transport in Porous Media, 32:239–262, 1998.

[6] O. LeMaitre, M. Reagan, H. Najm, R. Ghanem, and O. Knio. A stochastic projection method for fluid flow. ii: Random process. J. Comp. Phys., 181:9–44, 2002.

[7] H. Najm. Uncertainty quantification and polynomial chaos techniques in computational fluid dynamics. Annual Reviews, 41(1):35–52, 2009.

[8] J. L. Beck, E. Chan, A. Irfanoglu, and C. Papadimitriou. Multi-criteria optimal structural design under uncertainty. Earthquake Eng Struct Dynam, 28(7):741–761, 1999.

[9] W. Yao, X. Chen, W. Luo, M. van Toorenb, and J. Guo. Review of uncertainty-based multidisciplinary design optimization methods for aerospace vehicles . Progress in Aerospace Sciences, 47(6):450–479, 2011.

[10] I. Elishakoff, R. T. Haftka, and J. Fang. Structural design under bounded uncertainty–Optimization with anti-optimization. Computers & Structures, 53(6):1401–1405, 1994.

[11] JH de Baar, Thomas P Scholcz, Clemens V Verhoosel, Richard P Dwight, Alexander H van Zuijlen, and Hester Bijl. Efficient uncertainty quantification with gradient-enhanced kriging: Applications in fsi. In Proc. of the European Congress on Computational Methods in Applied Sciences and Engineering (ECCOMAS 2012), pages 10–14. Vienna, Austria, 2012.

[12] D. Ghiocel and R. Ghanem. Stochastic finite element analysis of seismic soil-structure interaction. ASCE, Journal of Engineering Mechanics, 128(1):66–77, 2002.

[13] Jeroen AS Witteveen, Sunetra Sarkar, and Hester Bijl. Modeling physical uncertainties in dynamic stall induced fluid–structure interaction of turbine blades using arbitrary polynomial chaos. Computers & structures, 85(11):866–878, 2007.

[14] M. Herzog, A. Gilg, M. Paffrath, P. Rentrop, and U. Wever. Intrusive versus non-intrusive methods for stochastic finite elements. In MichaelH. Breitner, Georg Denk, and Peter Rentrop, editors, From Nano to Space, pages 161–174. Springer Berlin Heidelberg, 2008.

[15] Leo Wai-Tsun Ng and MS Eldred. Multifidelity uncertainty quantification using nonintrusive polynomial chaos and stochastic collocation. In Proceedings of the 14th AIAA Non-Deterministic Approaches Conference, number AIAA-2012-1852, Honolulu, HI, volume 43, 2012.

[16] Gaël Poëtte and Didier Lucor. Non intrusive iterative stochastic spectral representation with application to compressible gas dynamics. Journal of Computational Physics, 231(9):3587–3609, 2012.

[17] Alen Alexanderian, OliverP. Matre, HabibN. Najm, Mohamed Iskandarani, and OmarM. Knio. Multiscale stochastic preconditioners in non-intrusive spectral projection. Journal of Scientific Computing, 50(2):306–340, 2012.

[18] A. Doostan, A. Validi, and G. Iaccarino. Non-intrusive low-rank separated approximation of high-dimensional stochastic models. Computer Methods in Applied Mechanics and Engineering, 263(1):42–55, 2013.

[19] R. Ghanem and A. Sarkar. Mid-frequency structural dynamics with parameter uncertainty. Comput. Methods Appl. Mech. Engrg., 191:5499–5513, 2002.

[20] D. Xiu and G.M. Karniadakis. The Wiener-Askey polynomial chaos for stochastic differential equations. SIAM Journal on Scientific Computing, 24(2):619–644, 2002.

[21] R. A. Askey and W. J. Arthur. Some basic hypergeometric orthogonal polynomials that generalize Jacobi polynomials, volume 319. AMS, Providence RI, 1985.

[22] Anders Logg, Kent-Andre Mardal, Garth N. Wells, et al. Automated Solution of Differential Equations by the Finite Element Method. Springer, 2012.

[23] M.T. Reagan, H.N. Najm, R.G. Ghanem, and O.M. Knio. Uncertainty quantification in reacting-flow simulations through non-intrusive spectral projection. Combustion and Flame, 132(3):545–555, 2003.

[24] L. Mathelin and M.Y. Hussaini. A stochastic collocation algorithm for uncertainty analysis. Technical Report NAS 1.26:212153; NASA/CR-2003-212153, NASA Langley Research Center, 2003.

[25] D. Xiu and J.S. Hesthaven. High-order collocation methods for differential equations with random inputs. SIAM J. Sci. Comput., 27(3):1118–1139, 2005.

[26] I. Babuška, F. Nobile, and R. Tempone. A stochastic collocation method for elliptic partial differential equations with random input data. SIAM J. Numer. Anal., 45(3):1005–1034, 2007.

[27] P. G. Constantine, M. Eldred, and E. Phipps. Sparse pseudospectral approximation method. Computer Methods in Applied Mechanics and Engineering, 229:1–12, 2012.

[28] X. Ma and N. Zabaras. An efficient bayesian inference approach to inverse problems based on an adaptive sparse grid collocation method. Inverse Problems, 25:035013+, 2009.

[29] J. Beck, F. Nobile, L. Tamellini, and R. Tempone. On the optimal polynomial approximation of stochastic PDEs by galerkin and collocation methods. Mathematical Models and Methods in Applied Sciences, 22(09):1250023, 2012.

[30] S. Hosder, R.W. Walters, and R. Perez. A non-intrusive polynomial chaos method for uncertainty propagation in CFD simulations. In 44th AIAA aerospace sciences meeting and exhibit, AIAA-2006-891, Reno (NV), 2006.

[31] A. Doostan and G. Iaccarino. A least-squares approximation of partial differential equations with high-dimensional random inputs. Journal of Computational Physics, 228(12):4332–4345, 2009.

[32] J. Hampton and A. Doostan. Coherence motivated sampling and convergence analysis of least-squares polynomial chaos regression. Computer Methods in Applied Mechanics and Engineering, 290:73–97, 2015.

[33] C. Shannon. The mathematical theory of communication. Bell System Journal, 1948.

[34] D.L. Donoho. Compressed sensing. IEEE Transactions on information theory, 52(4):1289–1306, 2006.

[35] E.J. Candès and T. Tao. Near optimal signal recovery from random projections: Universal encoding strategies? IEEE Transactions on information theory, 52(12):5406–5425, 2006.

[36] Michael Lustig, David Donoho, and John M. Pauly. Sparse mri: The application of compressed sensing for rapid mr imaging. Magnetic Resonance in Medicine, 58(6):1182–1195, 2007.

[37] M. Lustig, D.L. Donoho, J.M. Santos, and J.M. Pauly. Compressed sensing mri. Signal Processing Magazine, IEEE, 25(2):72–82, March 2008.

[38] Hong Jung, Kyunghyun Sung, Krishna S. Nayak, Eung Yeop Kim, and Jong Chul Ye. k-t focuss: A general compressed sensing framework for high resolution dynamic mri. Magnetic Resonance in Medicine, 61(1):103–116, 2009.

[39] Urs Gamper, Peter Boesiger, and Sebastian Kozerke. Compressed sensing in dynamic mri. Magnetic Resonance in Medicine, 59(2):365–373, 2008.

[40] Felix J. Herrmann and Gilles Hennenfent. Non-parametric seismic data recovery with curvelet frames. Geophysical Journal International, 173(1):233–248, 2008.

[41] Felix J. Herrmann, Haneet Wason, and Tim T.Y. Lin. Compressive sensing in seismic exploration: an outlook on a new paradigm. CSEG Recorder, 36(4):19–33, 04 2011.

[42] AC. Gurbuz, J.H. McClellan, and W.R. Scott. A compressive sensing data acquisition and imaging method for stepped frequency gprs. Signal Processing, IEEE Transactions on, 57(7):2640–2650, July 2009.

[43] Q. Huang, L. Qu, Bingheng Wu, and Guangyou Fang. Uwb through-wall imaging based on compressive sensing. Geoscience and Remote Sensing, IEEE Transactions on, 48(3):1408–1415, March 2010.

[44] Xiao Xiang Zhu and R. Bamler. Tomographic sar inversion by ℓ_1-norm regularization; the compressive sensing approach. Geoscience and Remote Sensing, IEEE Transactions on, 48(10):3839–3846, Oct 2010.

[45] Gregory A. Howland and John C. Howell. Efficient high-dimensional entanglement imaging with a compressive-sensing double-pixel camera. Phys. Rev. X, 3:011013, Feb 2013.

[46] David L. Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. Proceedings of the National Academy of Sciences, 106(45):18914–18919, 2009.

[47] A. Doostan and H. Owhadi. A non-adapted sparse approximation of PDEs with stochastic inputs. Journal of Computational Physics, 230:3015–3034, 2011.

[48] A. Doostan, H. Owhadi, A. Lashgari, and G. Iaccarino. Non-adapted sparse approximation of PDEs with stochastic inputs. Technical Report Annual Research Brief, Center for Turbulence Research, Stanford University, 2009.

[49] G. Blatman and B. Sudret. An adaptive algorithm to build up sparse polynomial chaos expansions for stochastic finite element analysis. Probabilistic Engineering Mechanics, 25(2):183–197, 2010.

[50] G. Blatman and B. Sudret. Adaptive sparse polynomial chaos expansion based on least angle regression. Journal of Computational Physics, 230:2345–2367, 2011.

[51] L. Mathelin and K.A. Gallivan. A compressed sensing approach for partial differential equations with random input data. Commun. Comput. Phys., 12:919–954, 2012.

[52] L. Yan, L. Guo, and D. Xiu. Stochastic collocation algorithms using ℓ_1-minimization. International Journal for Uncertainty Quantification, 2(3), 2012.

[53] X. Yang and G. E. Karniadakis. Reweighted ℓ_1 minimization method for stochastic elliptic differential equations. Journal of Computational Physics, 248:87–108, 2013.

[54] S.S. Chen, D.L. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. SIAM J. Sci. Comput., 20:33–61, 1998.

[55] S.S. Chen, D.L. Donoho, and M.A. Saunders. Atomic decomposition by basis pursuit. SIAM Rev., 43(1):129–159, 2001.

[56] E.J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. Information Theory, IEEE Transactions on, 52(2):489–509, 2006.

[57] E.J. Candès and J. Romberg. Quantitative robust uncertainty principles and optimally sparse decompositions. Found. Comput. Math., 6(2):227–254, 2006.

[58] E.J. Candès and J. Romberg. Sparsity and incoherence in compressive sampling. Inverse Problems, 23(3):969–985, 2007.

[59] E. Candès and M. Wakin. An introduction to compressive sampling. Signal Processing Magazine, IEEE, 25(2):21–30, 2008.

[60] A.M. Bruckstein, D.L. Donoho, and M. Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. SIAM Review, 51(1):34–81, 2009.

[61] J.A. Tropp. Greed is good: algorithmic results for sparse approximation. Information Theory, IEEE Transactions on, 50(10):2231 – 2242, oct. 2004.

[62] Holger Rauhut. Compressive sensing and structured random matrices, 2009.

[63] D.L. Donoho, M. Elad, and V.N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. IEEE Transactions on information theory, 52(1):6–18, 2006.

[64] Jerrad Hampton and Alireza Doostan. Coherence motivated sampling and convergence analysis of least squares polynomial chaos regression. Computer Methods in Applied Mechanics and Engineering, (0), 2015.

[65] M. R. Osborne, B. Presnell, and B. A. Turlach. A new approach to variable selection in least squares problems. IMA journal of numerical analysis, 20(3):389–403, 2000.

[66] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. Communications on Pure and Applied Mathematics, 57(11):1413–1457, 2004.

[67] P. L. Combettes and V. R. Wajs. Signal recovery by proximal forward-backward splitting. Multiscale Modeling & Simulation, 4(4):1168–1200, 2005.

[68] M. Figueiredo, R. D. Nowak, and S. J. Wright. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. Selected Topics in Signal Processing, IEEE Journal of, 1(4):586–597, 2007.

[69] S.-J Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky. An interior-point method for large-scale l1-regularized least squares. Selected Topics in Signal Processing, IEEE Journal of, 1(4):606–617, 2007.

[70] E. van den Berg and M. P. Friedlander. SPGL1: A solver for large-scale sparse reconstruction, June 2007. Available from http://www.cs.ubc.ca/labs/scl/spgl1.

[71] D.L. Donoho, A. Maleki, and A. Montanari. Message-passing algorithms for compressed sensing. Proceedings of the National Academy of Sciences, 106(45):18914–18919, 2009.

[72] R. Tibshirani. Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society, Series B, 58(1):267–288, 1996.

[73] E. van den Berg and M. P. Friedlander. Probing the Pareto frontier for basis pursuit solutions. SIAM Journal on Scientific Computing, 31(2):890–912, 2008.

[74] J.A. Tropp and S.J. Wright. Computational methods for sparse solution of linear inverse problems. Proceedings of the IEEE, 2010. in press.

[75] B. Efron, T. Hastie, L. Johnstone, and R. Tibshirani. Least angle regression. Annals of Statistics, 32:407–499, 2004.

[76] S.S. Chen, D.L. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. SIAM Rev., 43(1):129–159, 2001.

[77] E.T. Hale, W. Yin, and Y. Zhang. Fixed-point continuation for ℓ_1-minimization: Methodology and convergence. SIAM J. on Optimization, 19(3):1107–1130, 2008.

[78] K. Bredies and D.A. Lorenz. Linear convergence of iterative soft- thresholding. SIAM J. Sci. Comp., 30(2):657–683, 2008.

[79] J.M. Bioucas-Dias and M.A.T. Figueiredo. A new TwIST: Two-step iterative shrinking/thresholding algorithms for image restoration. IEEE Trans. Image Processing, 16(12):2992–3004, 2007.

[80] A. Beck and M. Teboulle. A fast iterative shrinkage-threshold algorithm for linear inverse problems. SIAM J. Imaging Sciences, 2:183–202, 2009.

[81] S. Becker, J. Bobin, and E. J. Candès. NESTA: A fast and accurate first-order method for sparse recovery. ArXiv e-prints, 2009. Available from http://arxiv.org/abs/0904.3367.

[82] E.J. Candes and Y. Plan. A probabilistic and ripless theory of compressed sensing. Information Theory, IEEE Transactions on, 57(11):7235–7254, Nov 2011.

[83] A. Doostan, G. Iaccarino, and N. Etemadi. A least-squares approximation of high-dimensional uncertain systems. Technical Report Annual Research Brief, Center for Turbulence Research, Stanford University, 2007.

[84] I. Babuška, R. Tempone, and G. Zouraris. Galerkin finite element approximations of stochastic elliptic partial differential equations. SIAM Journal on Numerical Analysis, 42(2):800–825, 2004.

[85] A. Cohen, R. DeVore, and C. Schwab. Convergence rates of best n-term galerkin approximations for a class of elliptic spdes. Foundations of Computational Mathematics, 10(6):615–646, 2010.

[86] O. Escoda, L. Granai, and P. Vandergheynst. On the use of a priori information for sparse signal approximations. IEEE Transactions in Signal Processing, 9:3468–3482, 2006.

[87] Rick Chartrand and Wotao Yin. Iteratively reweighted algorithms for compressive sensing. In in 33rd International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2008.

[88] M. A. Khajehnejad, W. Xu, A. S. Avestimehr, and B. Hassibi. Improved sparse recovery thresholds with two-step reweighted ℓ_1 minimization. In Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on, pages 1603–1607. IEEE, 2010.

[89] E.J. Candès and T. Tao. Decoding by linear programming. Information Theory, IEEE Transactions on, 51(12):4203–4215, 2005.

[90] E.J. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. Communications on Pure and Applied Mathematics, LIX:1207–1223, 2006.

[91] H. Rauhut and R. Ward. Sparse legendre expansions via ℓ_1-minimization. Journal of Approximation Theory, 164(5):517–533, 2012.

[92] P. C. Hansen. Rank-deficient and discrete ill-posed problems: numerical aspects of linear inversion. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1998.

[93] D. Needell. Noisy signal recovery via iterative reweighted $l1$-minimization. In Proc. Asilomar Conf. on Signals, Systems, and Computers, Pacific Grove, CA, Nov. 2009.

[94] M. Hansen and C. Schwab. Analytic regularity and nonlinear approximation of a class of parametric semilinear elliptic pdes. Mathematische Nachrichten, 2012.

[95] V. H. Hoang and C. Schwab. Sparse tensor galerkin discretization of parametric and random parabolic pdes—analytic regularity and generalized polynomial chaos approximation. SIAM Journal on Mathematical Analysis, 45(5):3050–3083, 2013.

[96] A. Kunoth and C. Schwab. Analytic regularity and gpc approximation for control problems constrained by linear parametric elliptic and parabolic pdes. SIAM Journal on Control and Optimization, 51(3):2442–2471, 2013.

[97] A. Juditsky and A. Nemirovski. Accuracy guarantees for ℓ_1-recovery. IEEE Trans. Inform. Theory, 57:7818–7839, 2011.

[98] A. Juditsky and A. Nemirovski. On verifiable sufficient conditions for sparse signal recovery via ℓ_1 minimization. Mathematical programming, 127(1):57–88, 2011.

[99] R. Gribonval and M. Nielsen. Sparse representations in unions of bases. Information Theory, IEEE Transactions on, 49(12):3320–3325, 2003.

[100] A. Cohen, W. Dahmen, and R. DeVore. Compressed sensing and best $k-$term approximation. J. Amer. Math. Soc., 22:211–231, 2009.

[101] Q. Mo and S. Li. New bounds on the restricted isometry constant δ_{2k}. Applied and Computational Harmonic Analysis, 31(3):460–468, 2011.

[102] J. Andersson and J. Strömberg. On the theorem of uniform recovery of random sampling matrices. arXiv preprint arXiv:1206.5986, 2012.

[103] E. J. Candès. The restricted isometry property and its implications for compressed sensing. Comptes Rendus Mathematique, 346(9):589–592, 2008.

[104] M. Bieri, R. Andreev, and C. Schwab. Sparse tensor discretization of elliptic sPDEs. Technical Report Research Report No. 2009-07, Seminar für Angewandte Mathematik, SAM, Zürich, Switzerland, 2009.

[105] G. Migliorati, F. Nobile, E. Schwerin, and R. Tempone. Analysis of the discrete L^2 projection on polynomial spaces with random evaluations. Technical report, Mathematics Institute of Computational Science and Engineering, Lausanne, Switzerland, 2011.

[106] M. E. Tipping. Sparse bayesian learning and the relevance vector machine. The Journal of Machine Learning Research, 1:211–244, 2001.

[107] S. Ji, Y. Xue, and L. Carin. Bayesian compressive sensing. Signal Processing, IEEE Transactions on, 56(6):2346–2356, 2008.

[108] P. Le Quéré. Accurate solutions to the square thermally driven cavity at high rayleigh number. Computers & Fluids, 20(1):29–41, 1991.

[109] R. Ghanem and P. Spanos. Stochastic Finite Elements: A Spectral Approach. Dover, 2002.

[110] V. Volterra. Theory of Functionals and of Integral and Integro-Differential Equations. Dover, 1959.

[111] David L. Donoho. For most large underdetermined systems of linear equations the minimal 1-norm solution is also the sparsest solution. Comm. Pure Appl. Math, 59:797–829, 2004.

[112] Ji Peng, Jerrad Hampton, and Alireza Doostan. A weighted -minimization approach for sparse polynomial chaos expansions. Journal of Computational Physics, 267(0):92 – 111, 2014.

[113] J. Peng, J. Hampton, and A. Doostan. On Polynomial Chaos Expansion via Gradient-enhanced ℓ_1-minimization. ArXiv e-prints, June 2015.

[114] A. Huang. A re-weighted algorithm for designing data dependent sensing dictionary. Int. J. Phys. Sci., 6(3):386–390, 2011.

[115] X. Yang, H. Lei, N. A. Baker, and G. Lin. Enhancing Sparsity of Hermite Polynomial Expansions by Iterative Rotations. ArXiv e-prints, June 2015.

[116] A. Logg, K. Mardal, and G Wells. Automated Solution of Differential Equations by the Finite Element Method. Springer Berlin Heidelberg, 2012.

[117] J. Hampton and A. Doostan. Compressive sampling of polynomial chaos expansions: Convergence analysis and sampling strategies. Journal of Computational Physics, 280:363–386, 2015.

[118] G Migliorati, Fabio Nobile, Erik von Schwerin, and Raúl Tempone. Approximation of quantities of interest in stochastic pdes by the random discrete lˆ2 projection on polynomial spaces. SIAM Journal on Scientific Computing, 35(3):A1440–A1460, 2013.

[119] B. Jones, N. Parrish, and A. Doostan. Postmaneuver collision probability estimation using sparse polynomial chaos expansions. Journal of Guidance, Control, and Dynamics, pages 1–13, 2015.

[120] J.D. Jakeman, M.S. Eldred, and K. Sargsyan. Enhancing ℓ_1-minimization estimates of polynomial chaos expansions using basis selection. Journal of Computational Physics, 289(0):18 – 34, 2015.

[121] Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. Statistics Surveys, 4:40–79, 2010.

[122] O. Roderick, M. Anitescu, and P. Fischer. Polynomial regression approaches using derivative information for uncertainty quantification. Nucl. Sci. Eng., 162(2):122–139, 2010.

[123] AK Alekseev, IM Navon, and ME Zelentsov. The estimation of functional uncertainty using polynomial chaos and adjoint equations. International Journal for Numerical Methods in Fluids, 67(3):328–341, 2011.

[124] Y. Li, M. Anitescu, O. Roderick, and F. Hickernell. Orthogonal bases for polynomial regression with derivative information in uncertainty quantification. International Journal for Uncertainty Quantification, 1(4), 2011.

[125] Brian Lockwood and Dimitri Mavriplis. Gradient-based methods for uncertainty quantification in hypersonic flows. Computers & Fluids, 85(0):27 – 38, 2013. International Workshop on Future of {CFD} and Aerospace Sciences.

[126] Vadim Komkov, Kyung K Choi, and Edward J Haug. Design sensitivity analysis of structural systems, volume 177. Academic press, 1986.

[127] L. B. Rall. Automatic differentiation: Techniques and applications, volume 120. Springer Berlin, 1981.

[128] Gary Tang. Methods for High Dimensional Uncertainty Quantification: Regularization, Sensitivity Analysis, and Derivative Enhancement. PhD thesis, Stanford University, 2013.

[129] G. Szegö. Orthongonal Polynomials. American Mathematical Society. American Mathematical Society, 1939.

[130] Holger Rauhut and Rachel Ward. Sparse legendre expansions via l1-minimization. J. Approx. Theory, 164(5):517–533, May 2012.

[131] E. J. Candés and Y. Plan. A probabilistic and ripless theory of compressed sensing. Information Theory, IEEE Transactions on, 57(11):7235–7254, 2010.

[132] Albert Cohen, Mark A. Davenport, and Dany Leviatan. On the stability and accuracy of least squares approximations. Foundations of Computational Mathematics, 13(5):819–834, 2013.

[133] JF Sykes, JL Wilson, and RW Andrews. Sensitivity analysis for steady state groundwater flow using adjoint operators. Water Resources Research, 21(3):359–371, 1985.

[134] Yang Cao, Shengtai Li, Linda Petzold, and Radu Serban. Adjoint sensitivity analysis for differential-algebraic equations: The adjoint dae system and its numerical solution. SIAM Journal on Scientific Computing, 24(3):1076–1089, 2003.

[135] Michael B Giles and Niles A Pierce. An introduction to the adjoint approach to design. Flow, turbulence and combustion, 65(3-4):393–415, 2000.

[136] Arthur Earl Bryson. Applied optimal control: optimization, estimation and control. CRC Press, 1975.

[137] Emmanuel Laporte and Patrick Le Tallec. Numerical methods in sensitivity analysis and shape optimization. Springer Science & Business Media, 2003.

[138] Antony Jameson. Aerodynamic shape optimization using the adjoint method. Lectures at the Von Karman Institute, Brussels, 2003.

[139] J. A. Tropp. User-friendly tail bounds for sums of random matrices. Foundations of Computational Mathematics, 12(4):389–434, 2012.

[140] Richard Baraniuk, Mark Davenport, Ronald DeVore, and Michael Wakin. A simple proof of the restricted isometry property for random matrices. Constructive Approximation, 28(3):253–263, 2008.

[141] C. Soize and R. Ghanem. Physical systems with random uncertainties: Chaos representations with arbitrary probability measure. SIAM Journal of Scientific Computing, 26(2):395–410, 2005.

[142] W. Gautschi. On generating orthogonal polynomials. SIAM Journal on Scientific and Statistical Computing, 3(3):289–317, 1982.

[143] R. Field and M. Grigoriu. On the accuracy of the polynomial chaos approximation. Probabilistic Engineering Mechanics, 19(1-2):65–80, 2004.

[144] W. Chen, R. Jin, and A. Sudjianto. Analytical Variance-Based Global Sensitivity Analysis in Simulation-Based Design Under Uncertainty. Journal of Mechanical Design, 127(5):875–886, 2005.

[145] X. Wan and G. Karniadakis. Long-term behavior of polynomial chaos in stochastic flow simulations. Computer methods in applied mechanics and engineering, 195(41):5582–5596, 2006.

[146] Marc Gerritsma, Jan-Bart van der Steen, Peter Vos, and George Karniadakis. Time-dependent generalized polynomial chaos. Journal of Computational Physics, 229(22):8333 – 8363, 2010.

[147] D. Ghosh and R. Ghanem. Stochastic convergence acceleration through basis enrichment of polynomial chaos expansions. International Journal for Numerical Methods in Engineering, 73:162–184, 2008.

[148] A. Nouy, A. Clement, F. Schoefs, and N. Moës. An extended stochastic finite element method for solving stochastic partial differential equations on random domains. Computer Methods in Applied Mechanics and Engineering, 197(51):4663–4682, 2008.

[149] C. Lang, A. Doostan, and K. Maute. Extended stochastic fem for diffusion problems with uncertain material interfaces. Computational Mechanics, 51(6):1031–1049, 2013.

[150] A. Birolleau, G. Poëtte, and D. Lucor. Adaptive bayesian inference for discontinuous inverse problems, application to hyperbolic conservation laws. Commun. Comput. Phys., 16:1–34, 2014.

[151] MK. Deb, I. Babuska, and JT. Oden. Solution of stochastic partial differential equations using Galerkin finite element techniques. Comput. Methods Appl. Mech. Engrg., 190:6359–6372, 2001.

[152] O.P. Le Maitre, H. Najm, R. Ghanem, and O. Knio. Multi-resolution analysis of Wiener-type uncertainty propagation schemes. J. Comp. Phys., 197(2):502–531, 2004.

[153] X. Wan and G. Karniadakis. An adaptive multi-element generalized polynomial chaos method for stochastic differential equations. J. Comp. Phys., 209:617–642, 2005.

[154] J. Foo and G. Karniadakis. Multi-element probabilistic collocation method in high dimensions. J. Comput. Phys., 229(5):1536–1557, 2010.

[155] D. Schiavazzi, A. Doostan, and G. Iaccarino. Sparse multiresolution regression for uncertainty propagation. International Journal for Uncertainty Quantification, 4(4):303–331, 2014.

[156] K. Engan, S.O. Aase, and J. Hakon Husoy. Method of optimal directions for frame design. In Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on, volume 5, pages 2443–2446 vol.5, 1999.

[157] K. Engan, S. O. Aase, and J.H. Husoy. Frame based signal compression using method of optimal directions (mod). In Circuits and Systems, 1999. ISCAS'99. Proceedings of the 1999 IEEE International Symposium on, volume 4, pages 1–4. IEEE, 1999.

[158] M. Aharon, M. Elad, and A Bruckstein. k -svd: An algorithm for designing overcomplete dictionaries for sparse representation. Signal Processing, IEEE Transactions on, 54(11):4311–4322, Nov 2006.

[159] M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. Image Processing, IEEE Transactions on, 15(12):3736–3745, 2006.

[160] M. Berveiller, B. Sudret, and M. Lemaire. Stochastic finite element: a non intrusive approach by regression. European Journal of Computational Mechanics/Revue Européenne de Mécanique Numérique, 15(1-3):81–92, 2006.

[161] Z. Gao and T. Zhou. Choice of nodal sets for least square polynomial chaos method with application to uncertainty quantification. Commun. Comput. Phys, pages 365–381, 2014.

[162] J. Hampton and A. Doostan. Compressive sampling of polynomial chaos expansions: convergence analysis and sampling strategies. Journal of Computational Physics, 280(3):263–386, 2015.

[163] G. Wahba. Practical approximate solutions to linear operator equations when the data are noisy. SIAM J. Numer. Anal., 14:651–667, 1977.

[164] G. Golub and V. Pereyra. The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate. SIAM Journal on numerical analysis, 10(2):413–432, 1973.

[165] G. Golub and V. Pereyra. Separable nonlinear least squares: the variable projection method and its applications. Inverse problems, 19(2):R1, 2003.

[166] Jochen Fröhlich. Parameter derivatives of the jacoby polynomials and the gaussian hypergeometric function. Integral Transforms and Special Functions, 2(4):253–266, 1994.

[167] Paul Woodward and Phillip Colella. The numerical simulation of two-dimensional fluid flow with strong shocks. Journal of Computational Physics, 54(1):115 – 173, 1984.

[168] D. Lucor, J. Witteveen, P. Constantine, D. Schiavazzi, and G. Iaccarino. Comparison of adaptive uncertainty quantification approaches for shock wave-dominated flows. In Proceedings of the Summer Program, page 219, 2012.

[169] B. D. Phenix, J. L. Dinaro, M. A. Tatang, J. W. Tester, J. B. Howard, and G. J. McRae. Incorporation of parametric uncertainty into complex kinetic mechanisms: Application to hydrogen oxidation in supercritical water. 112(1-2):132–146, 1998.

[170] Holger Rauhut and Rachel Ward. Interpolation via weighted minimization. Applied and Computational Harmonic Analysis, (0):–, 2015.

[171] Shane F Cotter, Bhaskar D Rao, Kjersti Engan, and Kenneth Kreutz-Delgado. Sparse solutions to linear inverse problems with multiple measurement vectors. Signal Processing, IEEE Transactions on, 53(7):2477–2488, 2005.

[172] Jie Chen and Xiaoming Huo. Theoretical results on sparse representations of multiple-measurement vectors. Signal Processing, IEEE Transactions on, 54(12):4634–4643, 2006.