Adaptive Multi-Element Generalized Polynomial Chaos: Algorithms and Applications

by

Xiaoliang Wan

Sc.M., Peking University, China, 2001
B.S., Peking University, China, 1998

Thesis

Submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy in the Division of Applied Mathematics at Brown University

May 2007
INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

UMI Microform 3272079
Copyright 2007 by ProQuest Information and Learning Company. All rights reserved. This microform edition is protected against unauthorized copying under Title 17, United States Code.

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
This dissertation by Xiaoliang Wan is accepted in its present form by the Division of Applied Mathematics as satisfying the dissertation requirement for the degree of Doctor of Philosophy.

Date 04/06/07

George Em Karniadakis, Director

Recommended to the Graduate Council

Date 4/23/07

Jan Hesthaven, Reader

Date 03/15/07

Boris Rozovsky, Reader

Approved by the Graduate Council

Date 4/26/07

Sheila Bonde, Dean of the Graduate School

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
The Vita of Xiaoliang Wan

Education:

- Ph.D., Brown University, U.S.A., 2007
- Sc.M., Peking University, P. R. China 2001
- B.S., Peking University, P. R. China, 1998

Peer-Reviewed Journal Publications:


Peer-Reviewed Conference Publications:


Conference Presentations:

(' indicates the speaker)


Acknowledgments

I would like to thank my advisor Professor George Em Karniadakis for all his guidance and encouragement during the course of this work. George has kept introducing me into new research directions and inspiring me through numerous research discussions. I am also indebted to George for sharing his personal research experience with me, which provides a valuable reference for my future research career.

My gratitude goes to my committee members and readers, Professor Jan Hesthaven and Professor Boris Rozovsky, for reading and correcting my thesis. I have learned a lot about spectral methods from the excellent lectures given by Professor Hesthaven. I would like to thank Professor Chau-Hsing Su, whose talks at the group meeting helped me a lot in learning random processes. I would like to thank Professor Martin Maxey for his consideration to help me clarify my physical understanding of fluid mechanics when he served as an examiner for my preliminary and second-language exams.

My thanks go to many CRUNCHers, Sirod Sirisup, Dongbin Xiu, Jin Xu, Igor Pivkin, Guang Lin, Jasmin Foo, Eric Keaveny, Leopold Grinberg and James Noble, for all their help and many happy conservations. I would like thank all my other friends I have made at Brown, Ya Jin, Deyu Li, Yanhu Wei, Jiangfeng Fei, Weigu He, Fengtian Xue, Kongbin Kan, Qi Wen, Weifeng Shen, Shanshan Wu, Daniele Venturi and Pradya Prempraneerach for all happiness and help they bring to me. Many thanks are due to Ms. Madeline Browster, Ms. Jean Radican and Ms. Laura Leddy for being there whenever I need help.

I would like to thank my parents and my sisters for their constant love, support and faith in my abilities. I am grateful to my mother-in-law for her love and help in my life. Special thanks go to my dear wife, Yue Li, for her love and emotional support, which made my strenuous Ph.D. journey full of happiness.

This work was supported by the following grants:

- National Science Foundation, DMS-0510799
- National Science Foundation, CTS-0331031
- Department of Energy, DE-FG02-95ER25239
- Office of Naval Research, N00014-95-1-0256
- Office of Naval Research, N00014-04-1-0007
# Contents

Acknowledgments ........................................ vii

1 Introduction ........................................ 1

1.1 Stochastic modeling ............................. 1

1.2 Numerical methods in stochastic simulations ........................................ 3

1.2.1 Sampling techniques ........................ 3

1.2.2 Non-sampling techniques ................. 4

1.3 Overview of polynomial chaos methods ........................................ 5

2 Random inputs ................................ 7

2.1 Overview of simulations of random processes ........................................ 7

2.2 The Karhunen-Loève (K-L) expansion ........................................ 8

2.3 Numerical algorithms for the K-L expansion ........................................ 12

2.3.1 Discretization of the K-L eigenvalue problem ........................................ 12

2.3.2 Gaussian-type kernels: the fast Gauss transform (FGT) ...................... 14

2.3.3 Other kernels: the generalized Fast Multipole Method ................. 24

3 Multi-element generalized polynomial chaos (ME-gPC) ................. 27

3.1 Polynomial chaos ................................ 27

3.2 Generalized polynomial chaos ................ 29

3.3 Decomposition of random space ................ 32
3.4 Numerical orthogonal polynomials ................................................................. 35
   3.4.1 Algorithms ............................................................................................. 35
   3.4.2 Implementation ..................................................................................... 40
3.5 ME-gPC expansion of a general operator ..................................................... 48
   3.5.1 Galerkin projection .............................................................................. 48
   3.5.2 Collocation projection .......................................................................... 49
3.6 \( hp \)-convergence of ME-gPC ................................................................. 53

4 An adaptive ME-gPC method for ODEs with random coefficients ............ 61
   4.1 An adaptive criterion ............................................................................... 61
   4.2 Numerical implementation .................................................................... 64
      4.2.1 Uniform distributions ......................................................................... 64
      4.2.2 Arbitrary distributions ........................................................................ 66
   4.3 Low regularity in the parametric space: the Kraichnan-Orszag problem ..... 66
      4.3.1 Why gPC fails .................................................................................. 66
      4.3.2 Random inputs with uniform distributions ....................................... 71
      4.3.3 Random inputs with arbitrary distributions ..................................... 79

5 An adaptive ME-gPC method for elliptic problems with random coefficients 84
   5.1 Model problem ....................................................................................... 85
   5.2 An a-posteriori error estimator ............................................................... 86
      5.2.1 Local error \( e \) .................................................................................... 87
      5.2.2 Stochastic regularity .......................................................................... 89
      5.2.3 Estimation of \( \eta \) ............................................................................ 91
   5.3 Algorithm of \( h \)-type adaptive ME-gPC .................................................. 94
      5.3.1 Convergence rate of \( \eta \) ................................................................ 95
      5.3.2 Criteria of adaptivity .......................................................................... 96
      5.3.3 Importance of random dimension .................................................... 97
5.3.4 Revisiting the correction factor $\tilde{c}$ .................................................. 97
5.3.5 Discussion of cost ......................................................................................................... 100
5.4 Numerical results ......................................................................................................................... 101
5.4.1 The stochastic algebraic model .................................................................................. 101
5.4.2 The stochastic elliptic model ..................................................................................... 107

6 Long-term behavior of polynomial chaos ................................................................. 110
6.1 Error estimates for gPC ............................................................................................................ 111
6.2 Error estimates for ME-gPC ..................................................................................................... 114
6.3 Relation between $\epsilon_M$ and $\hat{\epsilon}_M$ ........................................................................ 115
6.4 Some discussions on general cases ............................................................................................ 118
6.4.1 Other distributions and high-dimensional random inputs .................................... 118
6.4.2 Initial conditions ............................................................................................................ 120
6.5 Numerical results ......................................................................................................................... 120

7 Adaptivity of deterministic spectral/$hp$ element method .............................................. 125
7.1 Introduction ................................................................................................................................... 126
7.2 Model problem and its discretization ..................................................................................... 127
7.2.1 Reference elements ......................................................................................................... 128
7.3 An element-residual error estimator ....................................................................................... 129
7.3.1 Revisiting the Ainsworth-Oden (A-O) procedure for flux splitting ................... 130
7.3.2 Explicit equilibrated fluxes ......................................................................................... 133
7.3.3 High-order equilibration conditions ............................................................................ 139
7.4 Non-homogeneous Dirichlet boundary conditions .............................................................. 139
7.5 Numerical examples ................................................................................................................... 140
7.5.1 Efficiency for smooth solutions .................................................................................. 141
7.5.2 Singular solution: the Motz problem ......................................................................... 143

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
8 Polynomial chaos methods and experimental data

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.1 Simulations of non-Gaussian random processes</td>
<td>149</td>
</tr>
<tr>
<td>8.2 The methodology for general random inputs</td>
<td>150</td>
</tr>
<tr>
<td>8.3 Error analysis of the proposed methodology</td>
<td>153</td>
</tr>
<tr>
<td>8.4 Density estimation</td>
<td>154</td>
</tr>
<tr>
<td>8.5 Numerical studies</td>
<td>156</td>
</tr>
<tr>
<td>8.5.1 Algebraic model</td>
<td>156</td>
</tr>
<tr>
<td>8.5.2 Elliptic model</td>
<td>159</td>
</tr>
</tbody>
</table>

9 Applications of ME-gPC in fluid mechanics

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.1 Uncertainty in three-dimensional diffusion</td>
<td>165</td>
</tr>
<tr>
<td>9.1.1 Discretization of governing equations</td>
<td>165</td>
</tr>
<tr>
<td>9.1.2 Numerical results</td>
<td>168</td>
</tr>
<tr>
<td>9.2 Noisy flow past a stationary circular cylinder</td>
<td>173</td>
</tr>
<tr>
<td>9.2.1 Random-frequency inflow noise</td>
<td>176</td>
</tr>
<tr>
<td>9.2.2 Random-amplitude inflow noise</td>
<td>183</td>
</tr>
<tr>
<td>9.3 Heat transfer in groove channel</td>
<td>184</td>
</tr>
<tr>
<td>9.3.1 Governing equations</td>
<td>185</td>
</tr>
<tr>
<td>9.3.2 Stochastic flow-rate condition</td>
<td>185</td>
</tr>
<tr>
<td>9.3.3 Discretization of governing equations</td>
<td>190</td>
</tr>
<tr>
<td>9.3.4 Numerical results</td>
<td>194</td>
</tr>
</tbody>
</table>

10 Summary and open issues
List of Tables

2.1 Comparison of $L_\infty$ error $\epsilon_{max}$ and error bounds in a cube with a single Gaussian at one vertex of the cube; $r = 1/\sqrt{2}$.

3.1 Correspondence of the type of Wiener-Askey polynomial chaos and their underlying random variables ($N \geq 0$ is a finite integer). The following polynomials are for the one-dimensional case, i.e., $n = 1$. For the high-dimensional case, $\phi_n$ can be obtained using tensor products of one-dimensional basis.

3.2 Error $\epsilon_{max}$ and time cost for a uniform mesh with $N$ elements. $p = 10$ and $g(x) = x^{10} + 1$. The relative error for the recurrence coefficients was set to be $10^{-13}$. $\alpha = 1$ and $\beta = 4$ for Beta distribution. The middle element $[-6, 6]$ was used for Gaussian distribution. The computations were performed on a 1.5GHz AMD CPU.

4.1 Maximum normalized errors of the variance of $y_1$, $y_2$ and $y_3$ at $t = 30$ with $c_0 = 1/2$. (The results given by ME-gPC with $\theta_1 = 10^{-7}$ and polynomial order $p = 5$ are used as exact solutions.).

4.2 Maximum normalized errors of the variance of $y_1$, $y_2$ and $y_3$ at $t = 20$ for case (i) of the K-O problem. $\alpha = 1/2$ and $Y_1$ is of beta distribution with $\alpha = 1$ and $\beta = 4$. (The results given by ME-gPC with $\theta_1 = 10^{-7}$ and polynomial order $p = 7$ are used as exact solutions.).
4.3 Maximum normalized errors of the variance of $y_1, y_2$ and $y_3$ at $t = 20$ for case (i) of the K-O problem. $\alpha = 1/2$ and $Y_1$ is of normal distribution. (The results given by ME-gPC with $\theta_1 = 10^{-7}$ and polynomial order $p = 7$ are used as exact solutions.).

5.1 Typical values of $\dim(Y_{p,q})$ with respect to the number $n$ of random dimensions. $p = 2$ and $q = 1$. The numbers in parenthesis denote the dimensions of the space $W^p$.

5.2 A posteriori error estimates given by $Y_{p,1}$ and $V^p$. $\dim(V^p)/\dim(Y_{p,1}) = 0.1$. Eigenvalues are from the exponential kernel.

5.3 Statistics of local effectivity indices for a mesh with 1060 random elements. $p = 2$ and $A = 1$.

7.1 Effectivity indices for cell pattern (a),(b) and (c). Linear elements and increment of polynomial order 1 are considered.

7.2 Effectivity indices for cell pattern (d),(e) and (f). Linear elements and increment of polynomial order 1 are considered.

7.3 Effectivity indices for cell pattern (a),(b) and (c). Linear elements and increment of polynomial order 2 are considered.

7.4 Effectivity indices for cell pattern (d),(e) and (f). Linear elements and increment of polynomial order 2 are considered.

7.5 Effectivity indices for cell pattern (a),(b) and (c). Spectral/hp elements and increment of polynomial order 2 are considered. Due to $p$-convergence, sixth-order polynomials are used for pattern (a) while eighth-order polynomials are used for pattern (b) and (c).

7.6 Effectivity indices for cell pattern (d),(e) and (f). Spectral/hp elements and increment of polynomial order 2 are considered. Due to $p$-convergence, sixth-order polynomials are used for pattern (e) and (f) while eighth-order polynomials are used for pattern (d).

7.7 Overall performance of the two a posteriori error estimators for the smooth problem.
7.8 Effectivity indices for different meshes when \( B_4C/B_4B_2 = 0.9, V_1B_4/V_1V_4 = 0.2 \) and 
\( V_1B_1/V_1V_2 = 0.001 \). The increments are 2 and 4 for \( p = 1 \) and \( p = 2 \), respectively, in 
the local finite element space \( V_h^K \).

9.1 Physical parameters used in this work. The lengths are non-dimensionalized with the 
channel half-width \( h \).

9.2 Skewness \( s_E \) and kurtosis \( k_E \) of \( E \) for different random inputs and Reynolds numbers.

9.3 The mean and standard deviation of \( E \) for different random inputs and the corre-
sponding deterministic \( E \) with the mean random inputs.

9.4 Effect of different frequencies on \( E \) at \( Re = 525 \). Case I: \( \Omega_{F,i} = 0.1, 0.15, 0.2 \).

9.5 Effect of different frequencies on \( E \) at \( Re = 525 \). Case II: \( \Omega_{F,i} = 0.11, 0.15, 0.18 \).

9.6 Effect of superharmonic frequencies on \( E \) at \( Re = 525 \) for case (III).
List of Figures

2.1 Ratio of the two error bounds at different polynomial orders for spatial dimension $d = 3$. ................................................................. 19

2.2 Left: $\varepsilon(\hat{\omega})$; Right: $\varepsilon(\hat{\lambda})$ of the first four approximated eigenvalues (largest). ................................. 22

2.3 Left: the third eigenvector; Right: the fourth eigenvector. .................................................. 22

2.4 First 50 eigenvalues for Gaussian kernel with $\delta = 4$. Diamonds represent the 3D spectrum while squares represent a corresponding 2D spectrum. ........................................ 25

2.5 Errors of the first 38 eigenvalues. Left: $\varepsilon(\hat{\lambda}_i)$ with $i = 1, 2, \ldots, 38$; Right: Normalized $L_2$ error $\varepsilon(\hat{\lambda})$. ................................................................. 25

2.6 Speed-up of the fast eigenvalue solver. The solutions given by the direct parallel solver serve as “exact ones”. ................................. 25

3.1 Orthogonal polynomials for random variable $X$, $\pi_1$ to $\pi_4$. ................................................. 42

3.2 Orthogonal polynomials for random variable $Y$, $\pi_1$ to $\pi_4$. ................................................ 43

3.3 Left: PDF of random variable $Y_t$ induced by Beta distribution with $\alpha = 1$ and $\beta = 4$; Right: Orthogonal polynomials for $Y_t$. ................................................ 44

3.4 Left: PDF of random variable $Y_r$ induced by Beta distribution with $\alpha = 1$ and $\beta = 4$; Right: Orthogonal polynomials for $Y_r$. ................................................ 44

3.5 Left: PDF of random variable $Y_m$ induced by Gaussian distribution; Right: Orthogonal polynomials for $Y_m$. ................................................ 45

3.6 Left: PDF of random variable $X_{\infty}$ induced by Gaussian distribution; Right: Orthogonal polynomials for $X_{\infty}$. ................................................ 46
4.1 Deterministic solutions of the Kraichnan-Orszag problem subject to different initial conditions. Left: 3D phase space; Right: 2D projection on $x_1-x_2$ plane. .......................... 69

4.2 Kraichnan-Orszag problem: Several deterministic solutions of $x_1$ versus time corresponding to different initial conditions. .......................................................... 70

4.3 Comparison of variance obtained from gPC and Monte Carlo simulations. (a): $c_0 = 0.94$; (b): $c_0 = 0.96$; (c): $c_0 = 0.98$; (d): $c_0 = 0.99$. ......................................................... 70

4.4 Evolution of the variance of $y_1$ for one-dimensional random input. .............................. 73

4.5 Adaptive meshes for the 1D random input with $c_0 = 1/2$. (a): $\theta_1 = 0.01, p = 3$; (b): $\theta_1 = 0.001, p = 3$; (c): $\theta_1 = 0.0001, p = 3$; (d): $\theta_1 = 0.0001, p = 5$. ....................... 74

4.6 Error versus cost of Monte Carlo simulations and ME-gPC with different polynomial orders (based on the $L_\infty$ error of the variance of $y_1$ in the time interval $[8, 30]$). Here we only count the average number of operations in one time step. ................................. 75

4.7 Error contribution of each random element given by two criteria with different $c_0$. $\theta_1 = 10^{-4}$ and $p = 5$. (a): $c_0 = 1/2, t = 50$; (b): $c_0 = 1/2, t = 100$; (c): $c_0 = 1/4, t = 50$; (d): $c_0 = 1/4, t = 100$. ......................................................... 76

4.8 The Kraichnan-Orszag problem with 2D random inputs. $c_0 = 1/2, \theta_1 = 0.1, 0.01, 0.001$ and $\theta_2 = 0.1$. (a): $\sigma_{y_1}^2$; (b): $\sigma_{y_2}^2$; (c): $\sigma_{y_3}^2$; (d): adaptive mesh for $\theta_1 = 0.001$ and $p = 5$. 77

4.9 Error versus cost of Monte Carlo simulations and ME-gPC with 2D Legendre-Chaos (based on the $L_\infty$ error of the variance of $y_1$ in the time interval $[4, 10]$). Here we only count the average number of operations in one time step. ................................. 78

4.10 Evolution of variance for the 3D Kraichnan-Orszag problem. $\theta_2 = 10^{-1}$. Left: $\sigma_{y_1}^2 = \sigma_{y_2}^2$; Right: $\sigma_{y_3}^2$. ......................................................... 79

4.11 Evolution of the element number for the 3D Kraichnan-Orszag problem. $\theta_1 = 10^{-3}$. 79

4.12 Adaptive mesh and corresponding random distribution for the K-O problem with 1D random inputs. $p = 5$ and $\theta_1 = 10^{-3}$. Left: Beta distribution with $\alpha = 1$ and $\beta = 4$; Right: Gaussian distribution. ......................................................... 82
4.13 Speedup for the K-O problem with 1D random inputs at $t = 20$. Left: Beta distribution with $\alpha = 1$ and $\beta = 4$; Right: Gaussian distribution. 

4.14 Adaptive meshes for case (ii) of the K-O problem. $p = 4$, $\theta_1 = 10^{-4}$ and $\theta_2 = 10^{-2}$.
Left: Beta distribution with $\alpha = 1$ and $\beta = 4$; Right: Gaussian distribution.

4.15 Speedup for case (ii) of the K-O problem at $t = 10$. Left: Beta distribution with $\alpha = 1$ and $\beta = 4$; Right: Gaussian distribution.

4.16 Evolution of the variance of $y_1$ for case (iii) of K-O problem. Left: Beta distribution with $\alpha = 1$ and $\beta = 4$. $c = 1$; Right: Gaussian distribution, $c = 0.3$.

5.1 Error contribution of each mode in $Y_{p,1}$ with $p = 1$ and $M = 10$. Eigenvalues are from kernel $\exp(|x - y|/A)$ in $[0, 1]$. The markers of the largest 10% terms for each $s_{1\alpha}^l$ are filled. Left: $A = 1$; Right: $A = 0.1$.

5.2 Error contribution of each mode in $Y_{p,1}$ with $p = 2$ and $M = 10$. Eigenvalues are from kernel $\exp(|x - y|/A)$ in $[0, 1]$. The markers of the largest 10% terms for each $s_{1\alpha}^l$ are filled. Left: $A = 1$; Right: $A = 0.1$.

5.3 The factor $u_2^2/p^{2\alpha}$ of each term in $Y_{p,1}$ with $p = 1$ and $M = 10$. Eigenvalues are from kernel $\exp(|x - y|/A)$ in $[0, 1]$. Left: $A = 1$; Right: $A = 0.1$.

5.4 The factor $u_2^2/p^{2\alpha}$ of each term in $Y_{p,1}$ with $p = 2$ and $M = 10$. Eigenvalues are from kernel $\exp(|x - y|/A)$ in $[0, 1]$. Left: $A = 1$; Right: $A = 0.1$.

5.5 Adaptive errors of the second-order moments for the stochastic algebraic model. $N$ is the number of random elements. $p = 2$, $M = 10$ and $A = 1$. Uniform distribution is considered. Left: exponential kernel; Right: Gaussian kernel.

5.6 Adaptive errors of the second-order moments for the stochastic algebraic model. $N$ is the number of random elements. $p = 2$ and $M = 10$. Uniform and Beta distributions are considered. Eigenvalues are from the exponential kernel. Left: $A = 1$; Right: $A = 0.1$. 

xvii

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
5.7 Adaptive errors of the second-order moments for the stochastic elliptic model. $N$ is the number of random elements, $p = 2$, $A = 1$ and $M = 10$. Uniform and Beta distributions are considered. Eigenvalues are from the exponential kernel. Left: $\dim(V^p)/\dim(Y_{p,1}) = 0.1$; Right: $\dim(V^p)/\dim(Y_{p,1}) = 0.2$.

6.1 Evolution of error estimates of gPC and ME-gPC. Here, $Y$ is uniform in $[-1,1]$.

6.2 Convergence of gPC and $J_{t+1/2}^2$ at $t = 2$. Here, $Y$ is uniform in $[-1,1]$.

6.3 Comparison of gPC and ME-gPC of the same polynomial order $p = 3$. Normalized numerical errors of the second moment are used. The time of gPC is multiplied by the number $N$ of random elements. Here, $Y$ is of uniform distribution $U[-1,1]$. Left: ME-gPC with $N = 2$; Right: ME-gPC with $N = 3$.

6.4 Comparison of gPC and ME-gPC of the same polynomial order $p = 3$. Normalized numerical errors of the second moment are used. The time of gPC is multiplied by the number $N$ of random elements. Here, $Y$ is of Beta distribution $Beta(0,1)$ on $[-1,1]$.

6.5 Comparison of gPC and ME-gPC of the same polynomial order $p = 3$. Normalized numerical errors of the second moment are used. The time of gPC is re-scaled by a constant. Here, $Y$ is of Gaussian distribution $N(0,0.2)$. Left: ME-gPC with $N = 4$; Right: ME-gPC with $N = 8, 16$.

7.1 Mapping between triangular elements and the reference element.

7.2 The patches $\mathcal{P}_n$ and $\mathcal{E}_n$ of elements and edges influenced by the basis function $\theta_n$ associated with an interior vertex located at $x_n$.

7.3 Left: a mesh with hanging nodes. Right: the patch $\mathcal{P}_n$ consisting of macro-elements associated with vertex $x_n$.

7.4 Different cell patterns consisting of triangle elements.

7.5 Geometry and boundary conditions for the Motz problem.

7.6 Effectivity index versus the deformation of linear elements for the Motz problem. Left: $p = 1$; Right: $p = 2$. 

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
8.1 Truncated auxiliary PDF (2D) of the random variables in a K-L expansion .......................... 154

8.2 Convergence of the mean and standard deviation. \( N \) is the number of samples for density estimation. Left: Chebyshev polynomials are used; Right: Numerical orthogonality and estimated PDFs are used. .......................... 157

8.3 Convergence of the mean and standard deviation. Estimated marginal PDFs are used for the numerical orthogonality. Left: 100 samples for density estimation. Right: 100,000 samples for density estimation. .......................... 158

8.4 PDF of Beta\(\left(\frac{1}{2},\frac{1}{2}\right)\) and estimated ones with different numbers of samples. .......................... 158

8.5 Estimated marginal PDFs using different numbers of sample points. \( \xi_i \) is Gaussian, \( \sigma = 0.3 \) and \( l = 5 \). Left: PDFs of \( \xi_1 \); Right: PDFs of \( \xi_i, i = 1, 2, 3 \). .......................... 161

8.6 Orthogonal polynomials based on the estimated marginal PDFs with polynomial order \( p = 2, \ldots, 5 \). \( \xi_i \) is Gaussian, \( \sigma = 0.3 \) and \( l = 5 \). Left: \( \phi_{1,i}(\xi_1) \); Right: \( \phi_{2,i}(\xi_2) \) .......................... 161

8.7 Mean and standard deviation of \( u(x) \) given by different methods. Five-term K-L expansion of \( a(x) \). \( \sigma = 0.2, l = 5 \). Left: mean; Right: standard deviation. .......................... 163

8.8 Difference of statistics between the proposed strategy and Monte Carlo simulations. Five-term K-L expansion of \( a(x) \) is employed. \( \sigma = 0.2 \) and \( l = 5 \). Left: mean; Right: standard deviation. .......................... 164

8.9 Difference of statistics between the proposed strategy and Monte Carlo simulations. Five-term K-L expansion of \( a(x) \) is employed. \( \sigma = 0.2 \) and \( l = 5 \). Left: mean; Right: standard deviation. .......................... 164

9.1 Half of the computational domain. .......................... 168

9.2 PDF of random diffusivity at different reference points based on the log-normal process for the steady case. .......................... 170

9.3 Contours of \( y = 0 \) for steady case. Left: mean field. Right: standard deviation. .......................... 171

9.4 Influence of approximated eigenvalues and eigenfunctions by FGT on the random diffusion. Statistics along the centerline are used. Left: mean; Right: standard deviation. .......................... 171
9.5 Comparison with Monte Carlo simulations statistics along the centerline, where a
direct parallel solver is used for the K-L decomposition. Left: mean; Right: standard
development. ................................................................. 171
9.6 Reference points on planes y=0 and y=1. ............................................ 172
9.7 PDF of random diffusivity at the reference points based on the random process for
the unsteady case. For reference, we include the PDF of the corresponding Gaussian
random variable with \( \sigma = 0.2 \) .......................................................... 173
9.8 Instantaneous contours of statistics on \( y = 0 \) for the unsteady case. Left: mean field.
Right: standard deviation. .......................................................... 174
9.9 Evolution of statistics at reference points on \( y = 0 \) for the unsteady random diffusion.
Left: mean solution. Right: COV. ...................................................... 174
9.10 Cross-correlation coefficients on reference points for the unsteady random diffusion.
Left: with respect to point A on \( y = 0 \). Right: between \( y = 0 \) and \( y = 1 \) along the
vertical centerline. ............................................................... 174
9.11 Temperature evolution at reference points on \( y = 0 \). ................................. 175
9.12 Normalized errors of statistics based on the approximated eigenvalues and eigen-
functions by FGT with \( p = 5 \). Left: standard deviation. Right: cross-correlation
coefficients between \( y = 0 \) and \( y = 1 \). .............................................. 175
9.13 Schematic of the domain for noisy flow past a circular cylinder. The size of the domain
is \([-15D, 25D] \times [-9D, 9D]\) and the cylinder is at the origin with diameter \( D = 1 \).
The mesh consists of 412 triangular elements. ..................................... 180
9.14 Evolution of mean (upper) and variance (lower) of lift coefficient. \( Y \) is uniform in
\([-\sqrt{3}, \sqrt{3}] \). \( \sigma = 0.1 \) .............................................................. 180
9.15 Evolution of mean (upper) and variance (lower) of drag coefficient. \( Y \) is uniform in
\([-\sqrt{3}, \sqrt{3}] \). \( \sigma = 0.1 \) .............................................................. 181
9.16 Evolution of relative errors of instantaneous variance of lift coefficients given by gPC and ME-gPC. The results obtained from ME-gPC with $N = 20$ and $M = 8$ are used as a reference. 181

9.17 Instantaneous spatial distribution of RMS of vorticity at $tU/D = 100$. $N = 20$ and $M = 8$. 182

9.18 Instantaneous RMS of lift coefficients. A uniform distribution in $[-\sqrt{3}, \sqrt{3}]$ and a Beta distribution $\text{Be}(2,2)$ in $[-\sqrt{7}, \sqrt{7}]$ are considered with $\sigma = 0.1$. 182

9.19 PDFs of $A_L$ given by different models: (left) uniform noise and (right) Beta noise. 182

9.20 Evolution of mean (left) and variance (right) of $C_L$ given by gPC and ME-gPC. Here, $\xi$ is a uniform random variable in $[-\sqrt{3}, \sqrt{3}]$ with $\sigma = 0.1$. 183

9.21 Schematic of the geometry with a periodicity length $L$. The domain is composed of the inflow boundary $\partial D_{in}$, the outflow boundary $\partial D_{out}$ and boundaries $\partial D_i$, $i = 1, 2, 3$. 185

9.22 A typical spectral/hp element mesh for the physical discretization. The dotted lines are based on quadrature points of Jacobi polynomials. The reference points $A-E$ defined along the straight line $x = 3.2$ are used later in the study of correlation. 194

9.23 Mean of $E$ with error bars indicating the standard deviation for uniform (upper) and Beta (lower) random inputs. $E(\bar{Y} = 0.5)$ is the deterministic solution for the mean input. $Re = 525$. 197

9.24 Dependence of the optimal frequency on the amplitude at $Re = 525$. 198

9.25 PDF of $E$ with uniform random inputs. $Re = 225$. 201

9.26 PDF of $E$ with uniform random inputs at $Re = 525, 825$. 201

9.27 PDF of $E$ with Beta random inputs at $Re = 525, 825$. 202

9.28 Curves of the enhancement ratio $E$ versus the excitation amplitude for different Reynolds numbers. 202

9.29 Spectra of the mean of $E$ at $Re = 525$ for case (I). $\Omega_{Ri} = 0.1, 0.15, 0.2$. 205

9.30 Spectra of the standard deviation of $E$ at $Re = 525$ for case (I). $\Omega_{Ri} = 0.1, 0.15, 0.2$. 205

9.31 Spectra of the mean of $E$ at $Re = 525$ for case (II). $\Omega_{Ri} = 0.11, 0.15, 0.18$. 206
9.32 Spectra of the standard deviation of $E$ at $Re = 525$ for case (II). $\Omega_{F, i} = 0.11, 0.15, 0.18, 0.206$

9.33 Correlation matrix $\rho_{ij}$ for velocity-temperature pairs $(v_i, T_j)$ on equidistant grid points along the straight line $x = 3.20, -1.89 \leq y \leq 0.73$. Uniform random inputs are used. $\Omega_F = 0.15$ and $Re = 525$.

9.34 Evolution of $\rho_{ij}$. $Y \sim U[0, 1]$ and $Re = 525$. Left: $\Omega_F = 0.15$; Right: $\Omega_F = 0.18$.

9.35 Mean (a1-a4) and standard deviation (b1-b4) of cross-flow velocity at $t = \hat{T}/4, \hat{T}/2, 3\hat{T}/4, \hat{T}$ with $\hat{T}$ being the period. $Y \sim U[0, 1]$, $\Omega_F = 0.15$ and $Re = 525$.

9.36 Mean (a1-a4) and standard deviation (b1-b4) of temperature at $t = \hat{T}/4, \hat{T}/2, 3\hat{T}/4, \hat{T}$ with $\hat{T}$ being the period. $Y \sim U[0, 1]$, $\Omega_F = 0.15$ and $Re = 525$. 
Chapter 1

Introduction

1.1 Stochastic modeling

Deterministic models are widely used in simulating physical and engineering applications, such as in solid and fluid mechanics. In deterministic models, fluctuations, such as small scales, microscopic effects and background noise, are usually ignored. Most of the success of deterministic models relies on the separation between systematic motions and fluctuations. In well-controlled experiments, the systematic behavior can also be isolated, up to a high degree of accuracy, from the uncontrolled fluctuations while the error bars of the measurements are usually necessary.

Fluctuations or random sources can be classified as internal, which is studied in thermodynamics, quantum mechanics and statistical mechanics, and external, which is usually noise in dynamical systems or experiments, for example, uncertainty in initial/boundary conditions, properties of material and driving force etc. Stochastic modeling of random sources has received a lot of attention since it is very important in physical, chemical and biological applications to predict the behavior of dynamical system more reliably.

There are two fundamentally methods to introduce noise into differential equations to model the stochastic dynamic behavior.

- **Type I.** Certain functions, coefficients and boundary or initial conditions in classical differen-
tial equations can be random. For example, the physical properties of heterogeneous materials, stochastic loading in structural mechanics, inflow boundary conditions in fluid mechanics can be modeled by spatial, temporal or spatial-temporal random processes. In reality, such random processes are usually stationary and characterized by the correlation function, which are often called “colored noise”. If the solutions have certain regularity properties, one can consider these problems as a family of classical deterministic problems for the individual realizations. This type of differential equations with uncertainty is of particular importance in engineering applications for more reliable prediction of dynamic behavior.

• **Type II.** This type is related to the so-called “white noise”, which is usually written formally as an ordinary (partial) differential equation

\[
\dot{X}_t = f(t, X_t) + G(t, X_t)\xi_t, \quad X_{t_0} = c. \quad (1.1.1)
\]

The white noise \(\xi_t\) is a stationary Gaussian random process with zero mean and a constant spectral density. Although such a process does not exist in the conventional sense, it is a very useful mathematical idealization for describing random influence that fluctuates rapidly. Mathematically the white noise can be identified with the Wiener process \(W_t\) and the above equation can be written as

\[
dX_t = f(t, X_t)dt + G(t, X_t)dW_t, \quad X_{t_0} = c. \quad (1.1.2)
\]

There are two types of integrals defined for equation (1.1.2): Itô’s and Stratonovich’s stochastic integrals. Itô’s integral is widely used by mathematicians since it preserves the martingale properties. However, if the white noise is approximated by smooth processes, i.e., a limit case of colored noise when the correlation length decreases to zero, the Stratonovich’s integral is preferred since the rules of classical Riemann-Stieltjes calculus can be applied.

In this dissertation, we focus on numerical methods for stochastic differential equations of type
1.2 Numerical methods in stochastic simulations

There are two types of numerical methods in stochastic simulations: sampling and non-sampling techniques.

1.2.1 Sampling techniques

The basic sampling technique, i.e., the Monte Carlo method, is the most general one in stochastic simulations due to its simplicity and robustness. The procedure of the Monte Carlo method is based on repeated sampling in the random space. According to the law of large numbers, the ensemble averages from the Monte Carlo method will converge at a rate $C N_{mc}^{-1/2}$ with $N_{mc}$ being the sample size, where $C$ is constant. One of the most appealing features of the Monte Carlo method is that the constant $C$ depends on the variance of the estimated random variables, but not on the integration dimension. The drawback of the Monte Carlo method is that the convergence rate is low, which means that Monte Carlo simulations can be very time-consuming. Many techniques have been developed to improve the convergence of the Monte Carlo method. The basic idea is to reduce the variance of the estimated random variables, in other words, to attack the constant factor $C$. Such methods are usually called variance reduction methods [33], and include importance sampling, control variates, stratified sampling and so on. Another technique to improve the convergence is the Quasi-Monte Carlo method [73], where the underlying source of random numbers from the Monte Carlo method is replaced by more uniformly distributed deterministic sequences, i.e., low-discrepancy sequences. In other words, the Quasi-Monte Carlo method is actually a deterministic method instead of a sampling method. Compared to Monte Carlo methods, Quasi-Monte Carlo methods provide a much better convergence rate $O(N_{mc}^{-1}(\log N_{mc})^{n-1})$, where $n$ is the number of dimensions.
1.2.2 Non-sampling techniques

Many non-sampling techniques have been developed. For problems of type I, typical methods include stochastic perturbation methods [60], Neumann series [120], polynomial chaos methods [113, 43, 116, 29, 14, 34, 70, 106], and probabilistic collocation methods [99, 59, 114, 13]; for problems of type II, the typical methods include moment equations and the probability density function (PDF) methods.

When both the random inputs and outputs are small, perturbation methods and Neumann expansion are popular approaches. Perturbation methods rely on the truncated Taylor expansions of the random field around its mean field. In practice, a second-order Taylor expansion is usually employed since the systems of equations may become extremely complicated beyond the second-order expansion. The Neumann expansion is based on the inverse of the stochastic operator in a Neumann series. Recently, the polynomial chaos methods have received a lot of attention in engineering applications. Polynomial chaos methods are based on spectral expansions of the random fields by proper orthogonal polynomials. For many cases, the fast (exponential) convergence can be maintained even for large perturbation when the solution has enough regularity. In the next section, we will present a detailed overview of methods along this line.

For the classical stochastic PDEs (type II), the moment equation methods and PDF methods can be employed. The moment equations methods rely on the derivation of equations of the statistical moments. However, the moment equations are not closed for nonlinear problems since the low-order moments depend on the higher-order ones. Furthermore, due to the nonlinearity, the hierarchy system of the moment equations may be very complicated. Thus, the moment equation methods work only for limited applications. The basic idea of the PDF method is to derive the evolution (Fokker-Plank) equation of the probability density of the solution. The PDF method usually results in a high-dimensional partial differential equation, which is difficult to handle in numerical simulations. Currently, the Monte Carlo methods are still the most popular numerical methods for the classical stochastic PDEs.
1.3 Overview of polynomial chaos methods

Polynomial chaos originated from the homogeneous chaos defined by Wiener [113], which was used to study the stochastic integral with respect to Wiener processes. Cameron and Martin [24] introduced a complete orthonormal set of functionals of Wiener process using Hermite polynomials such that every functional of Wiener process, which has a finite $L_2$ norm with respect to the Wiener measure, has a Fourier development in terms of this set which converges in the $L_2$ sense. Such a basis is often called Wiener-chaos or Hermite-chaos in the literature. Ghanem and Spanos first coupled Hermite-chaos and finite element methods [43] to study problems in solid mechanics [42, 40, 41]. Xiu and Karniadakis defined generalized polynomial chaos in [116], where orthogonal polynomials in the Askey family were employed to model non-Gaussian random fields based on the correspondence of PDFs of random variables and the weight functions of orthogonal polynomials. Due to the spectral expansion in the parametric space, polynomial chaos methods can achieve fast (exponential) convergence, i.e., $p$-convergence if the solution has enough regularity in the parametric space. Deb, Babuska and Oden applied finite element method in the parametric space in [29, 14] and obtained the same $h$-convergence rate as in the deterministic finite element methods. Another generalization, multi-element generalized polynomial chaos, was proposed in [102, 104, 106] based on numerical orthogonality, which yields $hp$-convergence in the parametric space. Sparse polynomial chaos bases were investigated in [34, 100]. Theoretical analysis of polynomial chaos methods can be found in [14, 34] for elliptic problems with random coefficients.

To this end, the polynomial chaos methods are proposed for problems of type I. The procedure is usually based on the Galerkin projection on polynomial chaos space, which results in a high-dimensional deterministic system of ODEs or PDEs. For linear problems, the unknown coefficients of polynomial chaos expansion can be decoupled and the obtained deterministic equations can be solved one by one. However, for nonlinear problems the unknown coefficients may be coupled, which results in a real challenge to design an efficient numerical solver.

To obtain more flexibility in the polynomial chaos methods, collocation projection has recently received more attention due to its efficiency of dealing with nonlinearity. In the collocation imple-
mentation of polynomial chaos methods, sparse grids are usually employed since they weakly depend on the dimensionality \[99, 59, 114\] and have a high degree of exactness for numerical integrations. The procedure of collocation projection is similar with that of the Monte Carlo methods. The difference is that all the sparse grids are deterministic, not sampled as in the Monte Carlo methods. Each grid corresponds to a deterministic ODE or PDE, which can be solved independently.

Wiener-chaos has also been employed to study problems of type II \[52\]. Wiener-chaos was applied to nonlinear filtering in \[66\]. Coupled with a recursive numerical scheme, Wiener-chaos demonstrates good efficiency. Wiener chaos was used to study linear parabolic stochastic PDEs in \[71\] and stochastic Navier-Stokes equations in \[72, 53\].
Chapter 2

Random inputs

In this Chapter, we discuss the representations of random inputs. We first present an overview of simulations of random processes. In the second section, we focus on the procedure of Karhunen-Loève (K-L) expansion due to its efficiency for the dimensionality reduction of a second-order random process. In the third section, we summarize the numerical algorithms for the eigen-value problem required by the K-L expansion. In particular, we develop a fast eigen-solver [107] for the Gaussian-type covariance kernels based on the fast Gauss transform (FGT) method.

2.1 Overview of simulations of random processes

The representation of stochastic processes is the essential part for numerical simulations as well as mathematical analysis when investigating a dynamic system subject to noisy excitation. Two typical methods are: the spectral representation method [85, 92, 118] and the Karhunen-Loève (K-L) expansion [65]. The spectral representation method is based on the well-known Wiener-Khintchine theorem. Based on the spectral density of covariance kernel, the spectral representation method can efficiently simulate uni-variate stationary Gaussian processes using a set of independent uniform random variables [94, 91, 92]. The computational efficiency of the spectral representation algorithm can be improved by the Fast Fourier Transform (FFT) [121, 122]. This algorithm was later extended to simulate multi-variate nonstationary non-Gaussian stochastic processes [119, 93, 82]. The K-L
expansion relies on the eigenvalues and eigenfunctions of the covariance kernel. A second-order stochastic process can be expressed by a set of uncorrelated random variables in the K-L expansion. For Gaussian stochastic processes, the uncorrelated random variables are actually independent since the Gaussian stochastic processes can be completely determined by up to second-order moments. The main difficulty in the K-L expansion is the eigenvalue problem, which corresponds to a large and dense matrix in the numerical discretization. The approximation to the eigenfunctions was investigated in [28, 96] using the Fourier series. The convergence of the K-L expansion is mainly affected by the regularity of covariance kernel and the correlation length [54, 34]. The K-L expansion is also employed to simulate nonstationary non-Gaussian stochastic processes [80]. We note that both methods can represent Gaussian stochastic processes efficiently. For non-Gaussian stochastic processes, higher-order moments are needed other than up to second-order moments, which are difficult to estimate from experimental data. Thus, the simulations of non-Gaussian stochastic are usually based on limited available probabilistic information: the covariance kernel and the marginal probability distribution functions.

2.2 The Karhunen-Loève (K-L) expansion

The Karhunen-Loève expansion provides the best linear approximation [30] in the mean square sense for a second-order random process and is easy to couple with polynomial chaos methods.

Let $\Omega, \mathcal{F}, P$ be a probability space, where $\Omega$ is the sample space, $\mathcal{F}$ is the $\sigma$-algebra of subsets of $\Omega$ and $P$ is a probability measure. We let $R(t)$ indicate a random process with $t$ being the index. For simplicity, we regard $t$ as time. We assume that $t \in [0, T]$, where $T \in \mathbb{R}^+$. We note that all the following discussions about the Karhunen-Loève expansion are also valid for space and space-time related second-order random processes.

Given a second-order process $R(t)$, i.e., $\mathbb{E}[R(t)^2] < \infty$, the covariance is defined as

$$K(t, s) = \mathbb{E}[R(t)R(s)] = \int_{\Omega} R(t)R(s)dP(\omega), \quad \forall (t, s) \in [0, T] \times [0, T], \quad (2.2.1)$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
where $P$ indicates the probability measure. We here assume that

$$\mathbb{E}[R(t)] = 0, \quad \mathbb{E}[R(t)^2] = 1, \quad \forall t \in [0, T], \quad (2.2.2)$$

which can be easily achieved by shifting and rescaling $R(t)$ using the mean and standard deviation. According to the Cauchy-Schwarz inequality, the covariance exists and is finite:

$$|K(t, s)| = |\mathbb{E}[R(t)R(s)]| \leq \mathbb{E}[R(t)^2]^{1/2}\mathbb{E}[R(s)^2]^{1/2} \leq 1.$$  

We claim that the covariance is a function of nonnegative-definite type, in other words, $K(t, s)$ is symmetric and for all $c_1, \ldots, c_n \in \mathbb{R}$ and $t_1, \ldots, t_n \in [0, T]$ we have

$$\sum_{i,j=1}^{n} c_i c_j K(t_i, t_j) = \sum_{i,j=1}^{n} c_i c_j \mathbb{E}[R(t_i)R(t_j)]$$

$$= \mathbb{E} \left[ \sum_{i,j=1}^{n} c_i c_j R(t_i)R(t_j) \right]$$

$$= \mathbb{E} \left[ \left( \sum_{i=1}^{n} c_i R(t_i) \right)^2 \right] \geq 0 \quad (2.2.3)$$

with $n \geq 1$ being an integer.

For such a covariance $K(t, s)$, there exists eigenpairs $\{ (\lambda_i, \varphi_i) \}_{i=1}^{\infty}$, where $\lambda_i \geq \lambda_j > 0$ for $i < j$, satisfying

$$\int_{[0,T]} K(t,s)\varphi_i(t)dt = \lambda_i \varphi_i(s), \quad (2.2.4a)$$

$$\int_{[0,T]} \varphi_i(t)\varphi_j(t)dt = \delta_{ij}. \quad (2.2.4b)$$

$\{ \varphi_i \}$ is a complete orthogonal set of functions on $[0, T]$. Any square integrable function can be expanded into a series in function $\varphi_i$. Moreover, by the Mercer's theorem the following expansion holds

$$K(t, s) = \sum_{i=1}^{\infty} \lambda_i \varphi_i(t)\varphi_i(s). \quad (2.2.5)$$
The series converges uniformly in both \( t \) and \( s \). Based on the eigenpairs \( \{(\lambda_i, \varphi_i)\}_{i=1}^{\infty} \), we have the following theorem

**Theorem 2.2.1** (Karhunen-Loève expansion [65]). A random function \( R(t) \) continuous in mean square on a closed interval \([0, T]\) has on \([0, T]\) an orthogonal expansion

\[
R(t) = \sum_{i=0}^{\infty} \sqrt{\lambda_i} \varphi_i(t) Y_i
\]

with

\[
E[Y_i Y_j] = \delta_{ij}, \quad \int_{[0,T]} \varphi_i(t) \varphi_j(t) dt = \delta_{ij},
\]

if, and only if, \( \lambda_i \) are the eigenvalues and \( \varphi_i(t) \) are the orthonormalized eigenfunctions of its covariance. Then the series converges in mean square uniformly on \([0, T]\).

**Remark 2.2.2.** From the above theorem we know that \( Y_i \) are mutually uncorrelated random variables with zero mean and unit variance. If \( R(t) \) is a Gaussian random processes, \( Y_i \) are mutually independent.

In numerical simulations, we need to truncate the series (2.2.6) up to a finite index number \( M \) as

\[
R_M(t) = \sum_{i=0}^{M} \sqrt{\lambda_i} \varphi_i(t) Y_i
\]

which yields an optimal mean square error [30]

\[
\int_{[0,T]} E[(R(t) - R_M(t))^2] dt = \sum_{i=M+1}^{\infty} \lambda_i.
\]

Such an error is completely determined by the decay rate of eigenvalues \( \lambda_i \) for a fixed \( M \). The decay rate of \( \lambda_i \) was examined theoretically for the covariance of stationary random processes, i.e., \( K(t, s) = K(t-s) \), in [34]. In particular, the exponential kernel \( K(t-s) = e^{-|t-s|/\Delta} \) and the Gaussian kernel \( K(t-s) = e^{-\frac{(t-s)^2}{\Delta^2}} \) were studied and compared, where \( \Delta \) indicates the correlation length.

The decay rate of \( \lambda_i \) depends on two factors: regularity of \( K(t-s) \) and the correlation length \( \Delta \).
For a fixed correlation length, the decay rate of eigenvalues is determined by the regularity of the covariance kernel $K(t-s)$. For example, the eigenvalues of the exponential kernel, which has a finite Sobolev regularity, have an algebraic decay rate; however, the eigenvalues of the Gaussian kernel, which is analytic, have an exponential decay rate. For a certain type of covariance kernel, the decay rate is mainly affected by the value of the correlation length. We consider two limit cases [56]: (1) When $A \rightarrow \infty$, $K(t-s) \rightarrow 1$. In other words, the stochastic process can be expressed as one random variable, i.e., fully correlated; (2) When $A \rightarrow 0$, $K(t-s) \rightarrow \delta(t-s)$. The spectral density of $K(t-s)$ will be a constant, i.e., white noise. Correspondingly, the eigenvalues of $K(t-s)$ will not decay. We note that the decay rate of eigenvalues influences significantly the number $M$ for a fixed truncation error and corresponds to the complexity of the problem.

Multiplying $\phi_j(t)$ to both sides of equation (2.2.6) and integrating on $[0,T]$, we can obtain

$$Y_j = \frac{1}{\sqrt{\lambda_j}} \int_{[0,T]} R(t) \phi(t) dt.$$  \hspace{1cm} (2.2.10)

It is clear that if $R(t)$ is Gaussian random process, $Y_j$ are Gaussian random variables with zero mean and unit variance. For non-Gaussian processes, the distribution of $Y_j$ depends on $R(t)$. We note that the integrand in (2.2.10) is a stationary random process. Such an integration can be defined as

$$\int_{[0,T]} R(t) \phi(t) dt = \lim_{\max_{|t_k-t_{k-1}| \rightarrow 0}} \sum_{k=1}^{n} r(t'_k) \phi(t'_k)(t_k - t_{k-1}),$$  \hspace{1cm} (2.2.11)

where $0 = t_0 < t_1 < \ldots < t_n = T$, $t_{k-1} \leq t'_k \leq t_k$. It is easy to obtain

$$E[\left(\int_{[0,T]} R(t) \phi(t) dt\right)^2] = \int_{[0,T]} \int_{[0,T]} K(t-s) \phi(t) \phi(s) dtds.$$  \hspace{1cm} (2.2.12)

Thus, a sufficient condition for the existence of the integral is

$$\int_{[0,T]} \int_{[0,T]} K(t-s) \phi(t) \phi(s) dtds < \infty,$$  \hspace{1cm} (2.2.13)

which is satisfied due to the properties of $K(t-s)$.  

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
2.3 Numerical algorithms for the K-L expansion

To obtain the K-L expansion, the eigenvalue problem (2.2.4) needs to be solved. However, analytical solutions are available only for few covariance kernels, such as one-dimensional exponential kernel. For most of the covariance kernels, we have to solve the eigenvalue problem numerically. In this section, we present some efficient algorithms for the eigenvalue problem. Some analytical solutions of equation (2.2.4) can be found in [43, 67].

2.3.1 Discretization of the K-L eigenvalue problem

The most commonly used numerical methods for the Fredholm integral equation (2.2.4) are projection methods and the Nyström method [11]. Let $D_e$ denote an finite element mesh of the solution domain $D$ and $e$ a certain finite element.

For the Galerkin projection method we consider a finite element space

$$S_h(D) = \text{span}\{b_i(x) : i = 1, \ldots, N\},$$

where the basis $b_i(x)$ are usually chosen as piece-wise polynomials. Projecting the eigenfunctions $\phi_i(x)$ onto the space $S_h(D)$ and taking Galerkin projection on both sides of the equation (2.2.4), we obtain the discrete form as

$$Sv = \lambda Qv,$$  \hspace{1cm} (2.3.1)

where

$$S_{ij} = \int_D \int_D b_i(x_1)K(x_1, x_2)b_j(x_2)dx_1dx_2$$

$$Q_{ij} = \int_D \int_D b_i(x_1)b_j(x_2)dx_1dx_2,$$

and $v_i$ are the expansion coefficients of eigenfunctions. Both $S$ and $Q$ are symmetric and positive definite.

On the other hand, the Fredholm integral equation (2.2.4) can be discretized by the Nyström
method resulting in the following form:

\[
\sum_{e \in D} \sum_{j=1}^{M_p^e} K(x, x_j^e) \varphi(x_j^e) w_j^e = \lambda \varphi(x),
\]

where \(M_p^e\) is the number of quadrature points in element \(e\), and \(w_j^e\) the weights on quadrature points in element \(e\) for \(p\)th-order polynomials. To find the solution at the quadrature points, we let \(x\) run through the entire set of quadrature points \(x_i\) to satisfy equation (2.3.2) using a collocation approach. This yields in global numbering:

\[
\sum_{j=1}^{M_p^e} K(x, x_j) \varphi(x_j) w_j = \lambda \varphi(x_i), \quad i = 1, \ldots, M_p^e,
\]

where \(M_p^e\) is the total number of quadrature points in the entire domain. By defining

\[
v_i = \varphi(x_i), \quad A_{ij} = K(x_i, x_j), \quad \text{and} \quad B_{ij} = w_i \delta_{ij} \quad i, j = 1, \ldots, M_p^e,
\]

we can rewrite the system (2.3.3) in matrix form:

\[
A B v = \lambda v.
\]

Let \(B^{-1}\) denote the inverse of \(B\); it is clear that \(B^{-1}\) exists because \(B\) is diagonal and \(w_i > 0\). Then, equation (2.3.4) can be transformed into a standard general eigenvalue problem as

\[
A \vartheta = \lambda B^{-1} \vartheta,
\]

where \(\vartheta = B v\). Since both \(A\) and \(B^{-1}\) are symmetric and positive, we know that all the eigenvalues are real and positive.

To this end, both the Galerkin projection and the Nyström method yield a standard eigenvalue problem

\[
V v = \lambda M v,
\]
where $V$ and $M$ are symmetric and positive matrixes. We note that the matrix $V$ is, in general, a dense matrix. In practice, the size of $V$ can be as large as $10^6$ and dense eigenvalue solvers are not applicable. Efficient numerical eigenvalue solvers are required.

### 2.3.2 Gaussian-type kernels: the fast Gauss transform (FGT)

The eigenvalue problem (2.3.6) for Gaussian-type kernels can be solved fast and accurately by employing the fast Gauss transform, which is an $O(N)$ algorithm in both time and memory. In this section we first present a new sharp error estimate of the fast Gauss transform, then construct an efficient numerical eigenvalue solver based on FGT.

#### A sharp error estimate of the fast Gauss transform

The first version of the FGT method was first proposed by Greengard & Strain in [47], and it was later improved in [95, 48] in terms of its computational speed-up. However, a wrong error estimate of truncated Hermite expansion was employed there. This was corrected later by Baxter & Roussos in [20] but this new correct estimate overestimates the error bound and is not as useful in high-dimensional (e.g. three-dimensional) problems because of its restriction on the side length of mesh.

Inspired by the work [20], we have constructed a new error estimate for the original FGT method.

Let $K(x_1, x_2)$ be the Gaussian kernel

\[
K(x_1, x_2) = e^{-\|x_1 - x_2\|_2^2/\delta},
\]

where $\| \cdot \|_2$ indicates a $L_2$ norm. Following the notations of [47, 20], we define for the one-dimensional case the Hermite function

\[
h_i(x) = e^{-x^2} H_i(x),
\]

where $H_i(x)$ are one-dimensional Hermite polynomials. A $n$-dimensional multi-index $\alpha = (\alpha_1, \ldots, \alpha_n)$ is a $n$-tuple of nonnegative integers, playing the role of a multi-dimensional index. For any multi-
index $\alpha$ and any $x \in \mathbb{R}^n$, we define

$$|\alpha| = \alpha_1 + \alpha_2 + \ldots + \alpha_n, \quad \alpha! = \alpha_1! \alpha_2! \ldots \alpha_n!, \quad x^\alpha = x_1^{\alpha_1} x_2^{\alpha_2} \ldots x_n^{\alpha_n}. \quad (2.3.9)$$

If $p$ is an integer, we say $\alpha \geq p$ if $\alpha_i \geq p$ for $1 \leq i \leq d$. We can now define the multi-dimensional Hermite polynomials and Hermite functions as

$$H_\alpha(x) = H_{\alpha_1}(x_1) \ldots H_{\alpha_n}(x_n),$$

$$h_\alpha(x) = e^{-|x|^2} H_\alpha x = h_{\alpha_1}(x_1) \ldots h_{\alpha_n}(x_n), \quad (2.3.10)$$

where $|x|^2 = x_1^2 + \ldots + x_n^2$.

**Theorem 2.3.1.** If a source point $y \in \mathbb{R}^d$ lies in a box $B$ with center $c_B$ and side length $r \sqrt{2d}$, then for any $x \in \mathbb{R}^d$

$$e^{-\|x_1 - x_2\|^2/\delta} = \sum_{\alpha < p} T_\alpha = \sum_{\alpha < p} \frac{1}{\alpha!} \left( \frac{x_2 - c_B}{\sqrt{\delta}} \right)^{\alpha} h_\alpha \left( \frac{x_2 - c_B}{\sqrt{\delta}} \right) \quad (2.3.11)$$

with error estimate

$$|E(p)| = \left| \sum_{\alpha \geq p} T_\alpha \right| \leq \sum_{i=1}^d \binom{d}{i} (K p^{-1/4} \frac{r_p}{1-r_p})^i, \quad (2.3.12)$$

where $K = 1.09 (2\pi)^{-1/4}$ and $r_p = r \sqrt{e/p} < 1$. Here $e$ denotes the base of natural logarithm and $p$ the order of Hermite polynomials in the truncated expansion.

**Proof.** Consider the $i$-th components of $x_1$, $x_2$ and $c_B$ with $1 \leq i \leq d$. By the properties of one-dimensional Hermite polynomials [1, equation (22.9.17)], it is easy to obtain

$$e^{-(x_{1,i}-x_{2,i})^2/\delta} = \sum_{n_{i}=0}^{\infty} \frac{1}{n_{i}!} \left( \frac{x_{2,i} - c_i}{\sqrt{\delta}} \right)^{n_{i}} h_{n_{i}} \left( \frac{x_{1,i} - c_i}{\sqrt{\delta}} \right).$$
We introduce the following definitions as in [20]:

\[ u_p^i(x_1,i, x_2,i, c_i) = \sum_{n_i=0}^{p-1} \frac{1}{n_i!} \left( \frac{x_2,i - c_i}{\sqrt{\delta}} \right)^{n_i} h_{n_i} \left( \frac{x_1,i - c_i}{\sqrt{\delta}} \right), \quad 1 \leq i \leq d, \]

\[ v_p^i(x_1,i, x_2,i, c_i) = \sum_{n_i=p}^{\infty} \frac{1}{n_i!} \left( \frac{x_2,i - c_i}{\sqrt{\delta}} \right)^{n_i} h_{n_i} \left( \frac{x_1,i - c_i}{\sqrt{\delta}} \right), \quad 1 \leq i \leq d, \]

which we can use to write the corresponding Gaussian kernel as

\[ e^{-\|x_1 - x_2\|^2/\delta} = \prod_{i=1}^{d} (u_p^i + v_p^i). \]

Using the fact that

\[ |u_p^i| - |v_p^i| \leq |u_p^i + v_p^i| = e^{-(x_1,i - x_2,i)^2/\delta} = e^{-(x_1,i - x_2,i)^2/\delta} \leq 1, \]

we have

\[ |E(p)| = \left| e^{-\|x_1 - x_2\|^2/\delta} - \prod_{i=1}^{d} u_p^i \right| \leq f(|v_p^1|, |v_p^2|, \ldots, |v_p^d|), \quad (2.3.13) \]

where

\[ f(|v_p^1|, |v_p^2|, \ldots, |v_p^d|) = \sum_{n=1}^{d} \sum_{1 \leq i_1 < \ldots < i_n} |v_p^{i_1}| |v_p^{i_2}| \ldots |v_p^{i_n}|. \]

For example, if \( d = 2, \)

\[ |E(p)| = |(u_p^1 + v_p^1)(u_p^2 + v_p^2) - u_p^2 u_p^2| \]

\[ = |(u_p^1 + v_p^1)v_p^1 + u_p^1 v_p^2| \]

\[ \leq |u_p^1 + v_p^1||v_p^1| + |u_p^1||v_p^2| \]

\[ \leq |v_p^1| + (1 + |v_p^1|)|v_p^2| \]

\[ = |v_p^1| + |v_p^2| + |v_p^1||v_p^2|. \]

Considering the inequality for Hermite functions, see [1, equation (22.14.17)], we have that

\[ \frac{1}{n!} |h_n(x)| \leq K_1 \frac{2^{n/2}}{\sqrt{n!}} e^{-x^2/2}, \quad n \geq 0 \text{ and } x \in \mathbb{R}, \quad (2.3.14) \]
where \( K_1 \) is a numerical constant less than 1.09 in value, and the Stirling's formula [1, equation (6.1.38)] for the factorial \( n! \), we can bound each \( |v^i_p| \) as

\[
|v^i_p| = \left| \sum_{n_i=p}^{\infty} \frac{1}{n_i!} \left( \frac{x_{2,i} - c_i}{\sqrt{\delta}} \right)^{n_i} h_{n_i} \left( \frac{x_{1,i} - c_i}{\sqrt{\delta}} \right) \right|
\]

\[
\leq K_1 \sum_{n_i=p}^{\infty} \frac{1}{\sqrt{n_i!}} \left( \frac{r}{\sqrt{2}} \right)^{n_i} 2^{n_i/2} e^{-\frac{x^2_i}{2}}
\]

\[
\leq K_1 \sum_{n_i=p}^{\infty} \frac{1}{\sqrt{n_i!}} r^{n_i}
\]

\[
= K_1 \sum_{n_i=p}^{\infty} r^{n_i} (2\pi)^{-1/4} n_i^{-1/2} e^{-\frac{n_i}{2} - \frac{r^2}{2n_i}}, \quad 0 < \theta < 1 \quad \text{(Stirling)}
\]

\[
\leq K_1 (2\pi)^{-1/4} \sum_{n_i=p}^{\infty} r^{n_i} \left( \frac{e}{n_i} \right)^{n_i/2} n_i^{-1/4}
\]

\[
\leq K_1 (2\pi)^{-1/4} \sum_{n_i=p}^{\infty} r^{n_i} \left( \frac{e}{p} \right)^{n_i/2} \leq K_1 (2\pi)^{-1/4} \frac{r^p}{1 - r_p}, \quad (2.3.15)
\]

where \( r_p = r \sqrt{e/p} \). Since \( f(|v^i_p|^p, |v^2_p|^p, \ldots, |v^{d_p}|^p) \) is an increasing function in terms of each \( |v^i_p|^p \), it can be bounded by replacing each \( |v^i_p|^p \) with the bound given in equation (2.3.15). Then the truncation error estimate (2.3.12) follows immediately.

Theorem 2.3.1 is based on a simple case with only one source point in a box; it can be easily generalized for the case with many source points in a box. Now we restate the lemma given by Greengard and Strain in [47] without proof, and apply a new error bound for the truncated Hermite expansion.

**Lemma 2.3.2** (Greengard and Strain, 1991). *Let \( N_B \) sources \( y_i \) with weight \( q_i \) lie in a box \( B \) with center \( c_B \) and side length \( r \sqrt{2\delta} \). Then the Gaussian field due to the sources in \( B \),

\[
G(x) = \sum_{i=1}^{N_B} e^{-\frac{\|x - x_i\|^2}{2\delta} q_i}, \quad (2.3.16)
\]

is equal to a single Hermite expansion about \( c_B \):

\[
G(x) = \sum_{\alpha \geq 0} B_{\alpha} h_{\alpha} \left( \frac{x - c_B}{\sqrt{\delta}} \right), \quad (2.3.17)
\]
The coefficients $B_\alpha$ are given by

$$B_\alpha = \frac{1}{\alpha!} \sum_{i=1}^{N_p} q_i \left( \frac{x_i - c_B}{\sqrt{\delta}} \right)^\alpha.$$  \hspace{1cm} (2.3.18)

If the Hermite expansion is truncated at order $p$, the error $E(p)$ satisfies the bound:

$$|E(p)| = \left| \sum_{\alpha \geq p} B_\alpha h_\alpha \left( \frac{x - c_B}{\sqrt{\delta}} \right) \right| \leq Q_B \sum_{i=1}^{d} \binom{d}{i} \left( K p^{-1/4} \frac{r_p^p}{1 - r_p} \right)^i,$$  \hspace{1cm} (2.3.19)

where $Q_B = \sum |q_i|$, $K = 1.09(2\pi)^{-1/4}$, and $r_p = r \sqrt{e/p}$.

The new error bound given in Theorem 2.3.1 is much sharper, especially in three-dimensions, than the one given by Baxter & Roussos in [20]; the latter has the form:

$$|E(p)| \leq (1 - r)^{-d} \sum_{i=0}^{d-1} \binom{d}{i} (1 - r^p)^i \left( K_1 r_p^p \frac{1}{\sqrt{p!}} \right)^{d-i}.$$  \hspace{1cm} (2.3.20)

If we ignore the high-order terms in both error bounds, we can get the corresponding simple versions

$$|E_{new}(p)| \leq dK p^{-1/4} \frac{r_p^p}{1 - r_p}$$  \hspace{1cm} (2.3.21)

and

$$|E_{BR}(p)| \leq dK_1 (1 - r)^{-d} (1 - r^p)^{d-1} \frac{r_p^p}{\sqrt{p!}}$$  \hspace{1cm} (2.3.22)

of our error estimate and Baxter & Roussos’s error estimate, respectively, which are the dominate parts of the two error estimates. To compare the two error bounds, we define a function $\rho$ as

$$\rho(p) = \frac{E_{BR}(p)}{E_{new}(p)}.$$  \hspace{1cm} (2.3.23)

It is easy to verify that $\rho$ is an increasing function in terms of $p$. In Fig. 2.1, we show the values of $\rho(p)$ at different $p$ when $d = 3$. We can see that $E_{new}(p)$ is much sharper than $E_{BR}(p)$, especially when $p$ is relatively large and $r$ is close to 1. Next we check the accuracy of the two error bounds...
by reconsidering the example in [20]. We assume that there is a single Gaussian at source point \( x = (0, 0, 0) \) with parameter \( \delta = 1 \) and weight \( q = 1 \). We evaluate \( G(x) \) on a \( 50 \times 50 \times 50 \) grid in the unit box \([0, 1]^3\). In table 2.1, we compare the two error bounds with the \( L_\infty \) error on the grid points. It can be seen that \( E_{new}(p) \) is about 4 times as big as the \( L_\infty \) error for different polynomial orders; however, the ratio of \( E_{BR}(p) \) over the \( L_\infty \) error keeps growing as \( p \) increases. For this case \( E_{BR}(p) \) overestimates the error by two orders of magnitude while \( E_{new}(p) \) is of the same order as the actual error. Thus, the new error bound is more reliable to be used in controlling the accuracy of FGT.

![Figure 2.1: Ratio of the two error bounds at different polynomial orders for spatial dimension \( d = 3 \).](image)

Table 2.1: Comparison of \( L_\infty \) error \( \epsilon_{max} \) and error bounds in a cube with a single Gaussian at one vertex of the cube; \( r = 1/\sqrt{2} \).

<table>
<thead>
<tr>
<th>( p )</th>
<th>6</th>
<th>8</th>
<th>10</th>
<th>12</th>
<th>14</th>
<th>16</th>
<th>18</th>
</tr>
</thead>
<tbody>
<tr>
<td>( E_{new}(p)/\epsilon_{max} )</td>
<td>4.03</td>
<td>3.89</td>
<td>3.86</td>
<td>3.88</td>
<td>3.92</td>
<td>3.97</td>
<td>4.02</td>
</tr>
<tr>
<td>( E_{BR}(p)/\epsilon_{max} )</td>
<td>68.02</td>
<td>79.06</td>
<td>87.06</td>
<td>93.46</td>
<td>98.62</td>
<td>103.19</td>
<td>107.16</td>
</tr>
</tbody>
</table>

In addition, the mesh restriction for the new error estimate is \( r_p = r \sqrt{e/p} < 1 \) instead of \( r < 1 \), which means that we can group the points in a relative bigger box if the polynomial order \( p \) is sufficiently large. For example, consider the three-dimensional case with \( r = 1 \) and \( \delta = 1 \). \( E_{BR}(p) \) will be infinity for this case; however, \( E_{new}(p) = 0.012 \) for \( p = 8 \). This is very important for three-dimensional problems because if we can increase the side length of mesh by a factor \( \eta \), where \( \eta > 1 \), the total number of boxes in a cube can be reduced significantly by a factor \( 1/\eta^3 \).
A fast eigenvalue solver for the Gaussian-type kernels

We develop a numerical scheme based on the Nyström method in conjunction with the spectral/hp element method [58]. We employ the implicitly restarted Arnoldi method to solve the standard general eigenvalue problem (2.3.5), where the package ARPACK [62, 63] is used. The essential part of the implicitly restarted Arnoldi method is the matrix-vector multiplication, which is also the most time-consuming part. Considering that the size of $A$ is $M_{p_s} \times M_{p_s}$, the work required for direct evaluation of $Av$ grows like $O(M_{p_s}^2)$ as $M_{p_s}$ increases; this makes it expensive to perform $Av$. To this end, we now employ FGT to compute $Av$ for each Arnoldi iteration.

According to Lemma 2.3.2, we can rewrite the matrix-vector multiplication in equation (2.3.5) as

$$Av \mapsto \sum_{j=1}^{M_{p_s}} e^{-\|x_i-x_j\|^2/\delta} q_j = \sum_{\alpha \geq 0} B_\alpha h_\alpha \left( \frac{x_i - cB}{\sqrt{\beta}} \right),$$

(2.3.24)

where $q_j = \phi(y_j)w_j$. If the Hermite series is truncated after $p^d$ terms, we can form the coefficients $B_\alpha$ first at a cost of $O(p^d M_{p_s})$ work and evaluate the Hermite series at all quadrature points, with a total cost of $O(2p^d M_{p_s})$.

Since the cost of matrix-vector multiplication is still dominant in the fast eigenvalue solver, the total cost is $O(p^d M_{p_s})$. As for the memory cost, the dominant part comes from the implicitly restarted Arnoldi method with $O(m M_{p_s})$, where $m$ is the number of Lanczos basis vectors. Thus, both the time cost and the memory cost are linear with respect to the total number $M_{p_s}$ of quadrature points. To this end, we construct a fast eigenvalue solver for the Gaussian kernel by employing FGT to accelerate the matrix-vector multiplication. We find in numerical experiments that the fast eigenvalue solver works better for larger eigenvalues, which means that for a certain accuracy a smaller order $p$ of Hermite polynomials is sufficient for a larger eigenvalue. Considering that only the first few largest eigenvalues are employed in practice for the K-L expansion, the fast eigenvalue solver can be effective in balancing speed-up and accuracy.
Example I: One-dimensional Gaussian kernel

To demonstrate the accuracy of the fast eigenvalue solver, we first consider the one-dimensional Fredholm integral equation

\[ \int_{-1}^{1} e^{-\frac{(x-y)^2}{\delta^2}} \phi(y) dy = \lambda \phi(x), \]

where \( \delta = 4.0 \). We use 80 Gauss-Legendre quadrature points to ensure that the numerical integral is sufficiently accurate and does not introduce extra errors into the fast eigenvalue solver. We first solve the standard eigenvalue problem (2.3.5) with direct summation of \( \mathbf{A} \mathbf{v} \) and let the results serve as "exact solutions". We then solve the standard eigenvalue problem (2.3.5) with the fast eigenvalue solver presented in the previous section.

Let \( \mathbf{b} \) be an approximated vector of \( b \) and \( \beta \) an approximated scalar of \( \beta \). We define the normalized error measures as

\[ \varepsilon(\mathbf{b}) = \frac{\| \mathbf{b} - \mathbf{\hat{b}} \|_2}{\| \mathbf{b} \|_2}, \quad \delta(\beta) = \frac{|\beta - \mathbf{\beta}|}{|\beta|}. \]

Let \( \mathbf{w} = \mathbf{A} \mathbf{v} \) denote the product by the direct summation and \( \mathbf{\hat{w}} \cong \mathbf{A} \mathbf{v} \) the product by FGT. Let \( \lambda \) be an exact eigenvalue and \( \hat{\lambda} \) an approximated eigenvalue from the fast eigenvalue solver. Fig. 2.2 shows the errors of \( \mathbf{\hat{w}} = \mathbf{A} \mathbf{v} \) with \( \mathbf{v} = (1, 1, \ldots, 1) \) on the left and the errors of the first four approximated eigenvalues on the right. The shape of \( \varepsilon(\mathbf{\hat{w}}) \) comes from the properties of Hermite polynomials. Hermite polynomials are odd functions with odd order \( p \) and even functions with even order \( p \). Thus, if we expand the Gaussian kernel with respect to the center of a box, some terms will cancel each other since the term \( \left( \frac{y - c_R}{\sqrt{\delta}} \right)^p \) in equation (2.3.18) has similar properties. For \( \delta(\hat{\lambda}) \), we only plot the results from Hermite polynomials with odd order \( p \). We see that both \( \varepsilon(\mathbf{\hat{w}}) \) and \( \delta(\hat{\lambda}) \) show exponential convergence as the order \( p \) of Hermite polynomials increases. As the eigenvalue decreases the accuracy also decreases for certain \( p \) because of the restriction of \( \varepsilon(\mathbf{\hat{w}}) \). In Fig. 2.3 the third and fourth normalized eigenfunctions are plotted. It can be seen that eigenfunctions are well approximated by FGT with a small order \( p \) of Hermite polynomials.
Figure 2.2: Left: $e(\bar{w})$; Right: $\hat{e}(\hat{\lambda})$ of the first four approximated eigenvalues (largest).

Figure 2.3: Left: the third eigenvector; Right: the fourth eigenvector.
Example II: three-dimensional Gaussian kernel

We assume that the Gaussian kernel $R_{hh}(x, y)$ in the Fredholm integral equation (2.2.4) takes the form

$$R_{hh}(x, y) = e^{-\frac{|x-y|^2}{\delta}} \quad (2.3.27)$$

with $\delta = 4.0$. The size of $A$ in the standard eigenvalue problem (2.3.5) is $W_P = 190,825$.

In Fig. 2.4 the first 50 largest eigenvalues are plotted. Due to the relative small correlation length, the eigenvalues decay slowly. For comparison, we include the eigenvalues of a corresponding two-dimensional Gaussian kernel defined on the middle plane of the computational domain. It can be seen that the eigenvalues of three-dimensional Gaussian kernel decay much more slowly than those of the two-dimensional case. We now define the eigenvalue vector $\hat{\lambda} = (\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_{38})$, where $\hat{\lambda}_i$ is the $i$-th approximated eigenvalue. In Fig. 2.5 the error $\varepsilon(\hat{\lambda})$ is plotted on the right and the error $\hat{\varepsilon}(\hat{\lambda}_i)$ on the left with $i = 1, 2, \ldots, 38$. It can be seen that the value of $\varepsilon(\hat{\lambda}_i)$ increases as the eigenvalue decreases. All the errors of the first 38 approximated eigenvalues are less than 0.02 when $p = 6$. When $p = 5$, the errors of the first 21 approximated eigenvalues are less than 0.01 and the errors of the rest 17 approximated eigenvalues fluctuate in the interval $(0, 0.16)$. When $p = 4$, the errors keep growing quickly after the fourth eigenvalue. Numerical results show that for odd $p$ big fluctuation occurs when the eigenvalue decreases to a certain value while for even $p$ no such fluctuation occurs. This should be due to the properties of Hermite polynomials, which give rise to the fluctuation in the convergence of $\varepsilon(\hat{\lambda})$. However, $\varepsilon(\hat{\lambda})$ still keeps exponential convergence on the whole. Also for odd $p$ or even $p$, the convergence rate can be given by the FGT error bound, which is plotted in the same figure as $\varepsilon(\hat{\lambda})$. Note that we use $r = 1.06$ here for which Baxter & Roussos's error bound does not work.

To demonstrate the efficiency of the fast Gauss eigenvalue solver, we want to compare it with a direct eigenvalue solver; the former introduces truncation errors whereas the latter does not. For high-dimensional stochastic inputs, it is not possible to run the direct eigenvalue solver on one processor due to the high cost in memory and time. Thus, we need to use a parallel eigenvalue
solver to estimate the cost of a direct eigenvalue solver on one processor. We first run the parallel eigenvalue solver on \( m \) identical processors to solve the eigenvalue problem in \( t \) seconds. We then solve the same problem again by the fast Gauss eigenvalue solver on \textit{only one} of the \( m \) processors in \( T \) seconds. We define the speed-up \( S \) as

\[
S = \frac{mt}{T},
\]

where we use the multiplication \( mt \) to estimate the cost of a direct solver on one processor. We note here that an ideal parallelization is assumed. Of importance is the speed-up for given accuracy as the fast Gauss eigenvalue solver is approximate, so we employ the approximation error as a parameter. In Fig. 2.6 we show the speed-up versus the approximation error based on the first 38 largest eigenvalues, where we use the solutions given by the direct parallel solver as the “exact” ones. We can see that the speed-up is more than 1000 for a \( O(10^{-2}) \) accuracy, which implies that the fast Gauss eigenvalue solver run on one CPU is almost \textit{four times} faster than the parallel solver run on 256 processors. Note here that for the parallel eigenvalue solver we do not store the whole matrix \( A \) in the memory, but compute each entry \( A_{ij} \) every time we need it. The reason is that in order to store the entire matrix \( A \) in double precision about 270GBs of memory are needed, which is beyond the capacity of the computer system we used. However, for the fast Gauss eigenvalue solver about 250MB of memory are enough for \( p = 15 \) and double-precision computation, i.e. we realize a reduction in memory by three orders of magnitude.

2.3.3 Other kernels: the generalized Fast Multipole Method

In [89] a generalized fast multipole method (gFMM) was proposed for the eigenvalue problem (2.3.6), which is a generalization of the Greengard-Rokhlin [21, 46, 86] method for the Coulomb potential. We summarize this method in this section.

The key idea of gFMM is to decouple the source points \( \mathbf{x}_1 \) of the kernel \( K(\mathbf{x}_1, \mathbf{x}_2) \) from the field points \( \mathbf{x}_2 \) as in the FGT method. The simplest approach of such a decoupling is the Taylor
Figure 2.4: First 50 eigenvalues for Gaussian kernel with $\delta = 4$. Diamonds represent the 3D spectrum while squares represent a corresponding 2D spectrum.

Figure 2.5: Errors of the first 38 eigenvalues. Left: $\varepsilon(\lambda_i)$ with $i = 1, 2, \ldots, 38$; Right: Normalized $L_2$ error $\varepsilon(\lambda)$.

Figure 2.6: Speed-up of the fast eigenvalue solver. The solutions given by the direct parallel solver serve as "exact ones".
expansion:

\[
K(x_1, x_2) = \sum_{(\mu, \nu) \in \mathbb{N}_0^d \times \mathbb{N}_0^d} (D^{\mu+\nu}K)(x_1 - x_2)(x_1 - x_2)^{\mu} (x_2 - x_2)^{\nu} \frac{\mu!}{\nu!},
\]

(2.3.29)

where we assume that the kernel is stationary. We note that the Taylor expansion coefficients have to be calculated by differentiation of the covariance kernel \( K(x_1, x_2) \). To avoid this we can consider the Chebyshev interpolation of the covariance kernel:

\[
K(x_1, x_2; x_1^0, x_2^0) = \sum_{\mu \in \mathbb{Z}^d, |\mu| < m} \hat{K}(x_1^0, x_2^0) T_\mu(\chi^{-1}(x_1 - x_2)),
\]

(2.3.30)

where \( m \) is the polynomial order and \( \chi(\xi) \) is an affine mapping defined as

\[
\chi(\xi) = \eta \| x_1^0 - x_2^0 \|_{L^\infty} \xi + x_1^0 - x_2^0, \quad 0 < \eta < 1
\]

(2.3.31)

to guarantee the convergence of Chebyshev expansion.

By consider the relation between the Chebyshev interpolation and the Taylor expansion we can finally obtain the approximation of the covariance kernel in the following form [89]

\[
K(x_1, x_2) = \sum_{(\mu, \nu) \in \mathbb{N}_0^d \times \mathbb{N}_0^d, |\mu+\nu| < m} \hat{K}(x_1^0, x_2^0) T_\mu(\chi^{-1}(x_1 - x_2)) \frac{(x_1 - x_2^0)^{\mu} (x_2 - x_2^0)^{\nu}}{\mu! \nu!} c_{\mu+\nu}(x_1^0, x_2^0),
\]

(2.3.32)

where \( m \) is the Chebyshev polynomial order and \( c_\mu \) are the coefficients of the interpolation polynomial defined by

\[
\sum_{\mu \in \mathbb{Z}^d, |\mu| < m} \hat{K}(x_1^0, x_2^0) T_\mu(\chi^{-1}(x_1 - x_2)) = \sum_{\mu \in \mathbb{N}_0^d, |\mu| < m} c_\mu(x_1^0, x_2^0)(x_1 - x_2)^\mu.
\]

(2.3.33)

Once we obtain the form (2.3.32), the idea of the fast summation algorithms, such as Fast Multipole Method, can be employed to obtain a fast algorithm for the matrix-vector multiplication needed by the eigenvalue solver. More details can be found in [89].
Chapter 3

Multi-element generalized polynomial chaos (ME-gPC)

In this chapter we propose a new polynomial chaos method [107, 104, 106]: multi-element generalized polynomial chaos method. In the first two sections, we overview the polynomial chaos (Hermite-chaos) and generalized polynomial chaos, respectively. In the third section we introduce the decomposition of random space, a similar technique as the deterministic finite element method. Due to the simultaneous decomposition of the probability density function, the local orthogonal polynomial chaos bases need to be reconstructed numerically. We discuss the algorithms and implementation of numerical orthogonality in the fourth section. In the fifth section, we apply the ME-gPC method to a general operator by Galerkin and collocation projections. We summarize the $hp$-convergence error estimates of ME-gPC in the sixth section.

3.1 Polynomial chaos

Polynomial chaos was first proposed by Wiener [113], which was employed to study stochastic integral with respect to the Brownian motion (Wiener process). Cameron and Martin [24] introduced a complete orthonormal set of functionals of Wiener process using Hermite polynomials such that
every functional of Wiener process, which has a finite $L_2$ norm with respect to the Wiener measure, has a Fourier development in terms of this set which converges in the $L_2$ sense. Such a basis is often called Wiener chaos in the literature.

Let $\{W_t, 0 \leq t \leq 1\}$ denote the Wiener process and $C$ the space of continuous functions introduced by the Wiener process. The basis of Wiener chaos is defined as follows:

**Definition 3.1.1 (Wiener chaos).** Let $\{b_i(t)\}$ be a complete orthonormal set of real functions, each belonging to $L_2(0,1)$. Let $\alpha \in \mathcal{I}$ denote a multi-index:

$$\mathcal{I} = \{\alpha = (\alpha_i, i \geq 1)|\alpha_i \in \mathbb{N}_0, |\alpha| = \sum_{i=1}^{\infty} \alpha_i\}$$ (3.1.1)

The basis functional can be defined as

$$\phi_\alpha = \prod_{i=1}^{\infty} H_{\alpha_i}(\int_0^1 b_i(t)dW_t),$$ (3.1.2)

where $H_{\alpha_i}(\cdot)$ are Hermite polynomials.

**Remark 3.1.2.** We note that in the above definition $Y_i = \int_0^1 b_i(t)dW_t$ is actually a Gaussian random variable. Using the orthogonality of $\{b_i(t)\}$ and properties of Wiener process, it is easy to verify that

$$E[Y_i] = 0, \quad E[Y_iY_j] = \int_0^1 b_i(t)b_j(t)dt = \delta_{ij},$$ (3.1.3)

where $\delta_{ij}$ is the Kronecker delta. In other words, $\{Y_i\}$ is a set of independent standard Gaussian random variables.

We now present the theorem of Fourier-Hermite expansion given by Cameron and Martin.

**Theorem 3.1.3 ([24]).** If $F$ is a functional of $L_2(C)$, satisfying

$$E[F^2] < \infty,$$ (3.1.4)
then the Fourier-Hermite expansion

\[ F = \sum_{|\alpha|=0}^{\infty} f_{\alpha}\phi_{\alpha}, \quad f_{\alpha} = E[F\phi_{\alpha}] \]  

converges in the \( L_2(C) \) sense.

### 3.2 Generalized polynomial chaos

The Cameron-Martin theory was generalized to other random processes, such as Poisson processes [76] and general-independent-increment processes [90], where orthogonal functionals were constructed correspondingly. The generalization is actually not straightforward. In particular, the result that \( Y_i = f_i^T b(t_i) dN_t \) (see Remark 3.1.2) are independent identically-distributed (i.i.d) random variables in general does not hold if \( N_t \) is not a Wiener process. The connections between stochastic processes and orthogonal polynomials are summarized in the monograph [88].

For a dynamical system with random inputs, orthogonal polynomials demonstrate good efficiency to model uncertainty propagation in many applications, especially when the random inputs can be modeled by a set of i.i.d. random variables. Ghanem and Spanos [43] first coupled the Hermite-chaos and the finite element method to study mechanic problems where the (Gaussian) random inputs was decomposed by the K-L expansion. Xiu and Karniadakis [116] generalized this idea and employed the Askey scheme of orthogonal polynomials to model non-Gaussian random inputs.

Let

\[ Y = (Y_1, \ldots, Y_n) : (\Omega, \mathcal{F}) \rightarrow (\mathbb{R}^n, B^n) \]  

be a \( \mathbb{R}^n \)-valued continuous random variable, where \( n \in \mathbb{N} \) and \( B^n \) is the \( \sigma \)-algebra of Borel subsets of \( \mathbb{R}^n \). A random function \( R(Y) \in L_2(\Omega, \mathcal{F}, P) \) can be expressed by generalized polynomial chaos (gPC) [116] as

\[ R(Y) = \sum_{|\alpha|=0}^{\infty} \hat{a}_\alpha \phi_{\alpha}(Y), \]  

where \( \alpha \in \mathbb{N}_0^n \), and \( \phi_{\alpha}(Y) \) denotes the gPC basis of degree \( |\alpha| = p \) in terms of the random variable
\( Y \). \( \phi_\alpha \) is, in general, a tensor product of one-dimensional basis \( \{ \phi_\alpha(Y_i) \} \). The family \( \{ \phi_\alpha \}_{\alpha=0}^\infty \) is an orthogonal basis in \( L_2(\Omega, \mathcal{F}, P) \) with orthogonality relation

\[
E[\phi_\alpha \phi_\beta] = E[\phi_\alpha^2] \delta_{\alpha\beta},
\]  

(3.2.3)

where \( \delta_{\alpha\beta} \) is the Kronecker delta and \( E \) denotes the expectation with respect to the probability density function (PDF) \( f(y) \) of \( Y \). From the properties of orthogonal polynomials and \( L_2 \) space [2, 97], we know that

**Theorem 3.2.1.** Let \( Y \in B \subset \mathbb{R}^n \). If \( B \) is compact, then \( \{ f(y)^{1/2} \phi_\alpha \} \) will be closed in the \( L_2 \) sense, where \( f(y) \) is the PDF of \( Y \).

Thus, the gPC expansion will converge to \( R(Y) \) in the \( L_2 \) sense:

\[
E[\{(R(Y) - \sum_{|\alpha|=0}^M \bar{a}_\alpha \phi_\alpha)^2\}] \to 0, \quad \text{as } M \to \infty.
\]  

(3.2.4)

We note that the theorem 3.2.1 in general does not hold if \( B \) is not compact. However, if the PDF decays fast enough when it goes to infinity, \( \{ f(y)^{1/2} \phi_\alpha \} \) can also be complete, e.g., the Laguerre and Hermite polynomials with respect to the exponential and Gaussian distributions, respectively [97].

Since gPC is actually a spectral expansion, the convergence of such an approximation is determined by the regularity of \( R(y) \) with respect to \( y \) in the range \( B \subset \mathbb{R}^n \). If \( R(y) \) is smooth enough, an fast (exponential) convergence rate can be achieved. Let \( I^n = (-1,1)^n \) be the \( n \)-dimensional cube. Let \( P_p(I^n) \) denote the collection of tensor products of one-dimensional Legendre polynomials of (separate) degree \( \leq p \) defined on \( I^n \). The norm \( \cdot \|_{H^m(I^n)} \) is defined as

\[
\| R(y) \|_{H^m(I^n)} = \left( \int_{I^n} \sum_{|\alpha| \leq m} (\partial^\alpha_y R(y))^2 \, dy \right)^{1/2}
\]  

(3.2.5)

We have the following theorem:
Theorem 3.2.2. If \( R(y) \in H^m(I^n) \), \( m \geq 0 \), then we have

\[
\| R - P_p R \|_{L_2(I^n)} \leq Cp^{-m} \| R \|_{H^m(I^n)},
\]  

(3.2.6)

where \( P_p \) indicates the Galerkin projection on the Legendre-chaos space.

Proof. The theorem can be proved by considering the Legendre expansion of \( R(y) \). Or we can conclude it directly from the approximation results of \( p \)-th order polynomials. See [12] for more details. \( \square \)

Remark 3.2.3. A general proof of the exponential convergence is not available for orthogonal polynomials with arbitrary measures. However, the approximation theory of orthogonal polynomials with respect to some particular measures were extensively investigated \([25, 98, 69, 12]\). Numerical verification of exponential convergence of the gPC expansion can be found in \([116, 106]\).

For some cases, it is convenient to map \( \alpha \) to a one-dimensional index. For example, the dimensionality of \( \{ \phi_\alpha \} \) satisfying \( |\alpha| \leq p \) is

\[
N_p = \frac{(p + n)!}{p!n!}.
\]  

(3.2.7)

Any \( \alpha \) corresponds to a unique index \( 0 \leq i \leq N_p - 1 \). Then the gPC expansion of \( R(\omega) \) takes the equivalent form

\[
R(\omega) = \sum_{i=0}^{N_p-1} \hat{a}_i \phi_i(Y(\omega)),
\]  

(3.2.8)

where the basis \( \phi_i \) satisfy the orthogonality

\[
\mathbb{E}[\phi_i \phi_j] = \mathbb{E}[\phi_i^2] \delta_{ij}.
\]  

(3.2.9)

Both expressions (3.2.2) and (3.2.8) will be used in this work for convenience.

For a certain \( \mathbb{R}^n \)-valued random variable \( Y \), the gPC basis \( \{ \phi_\alpha \} \) can be chosen in such a way that its weight function has the same form as the probability density function (PDF) of \( Y \). The corresponding type of classical orthogonal polynomials \( \{ \phi_\alpha \} \) and their associated random variable \( Y \)
are listed in table 3.1 [116]. For arbitrary probability measures, the orthogonality must be maintained numerically, as we explain in the next section.

Table 3.1: Correspondence of the type of Wiener-Askey polynomial chaos and their underlying random variables ($N \geq 0$ is a finite integer). The following polynomials are for the one-dimensional case, i.e., $n = 1$. For the high-dimensional case, $\phi_\alpha$ can be obtained using tensor products of one-dimensional basis.

<table>
<thead>
<tr>
<th>Random variables $Y$</th>
<th>Wiener-Askey chaos ${\phi_\alpha(Y)}$</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuous</td>
<td>Gaussian</td>
<td>$(-\infty, \infty)$</td>
</tr>
<tr>
<td></td>
<td>gamma</td>
<td>$[0, \infty)$</td>
</tr>
<tr>
<td></td>
<td>beta</td>
<td>$[a, b]$</td>
</tr>
<tr>
<td></td>
<td>uniform</td>
<td>$[a, b]$</td>
</tr>
<tr>
<td>Discrete</td>
<td>Poisson</td>
<td>${0, 1, 2, \ldots}$</td>
</tr>
<tr>
<td></td>
<td>binomial</td>
<td>${0, 1, \ldots, N}$</td>
</tr>
<tr>
<td></td>
<td>negative binomial</td>
<td>${0, 1, \ldots}$</td>
</tr>
<tr>
<td></td>
<td>hypergeometric</td>
<td>${0, 1, \ldots, N}$</td>
</tr>
</tbody>
</table>

3.3 Decomposition of random space

Let $Y \in B = \prod_{i=1}^{n} (a_i, b_i)$ be a $\mathbb{R}^n$-value random variable, where $a_i$ and $b_i$ are finite or infinite in $\mathbb{R}$ and the components of $Y$ are i.i.d. random variables. We define a non-overlapping decomposition $D$ of $B$ as

$$D = \begin{cases}
B_k = (a_{k,1}, b_{k,1}) \times (a_{k,2}, b_{k,2}) \times \cdots \times (a_{k,n}, b_{k,n}), \\
\bar{B} = \bigcup_{k=1}^{N} \bar{B}_k, \\
B_{k_1} \cap B_{k_2} = \emptyset, \text{ if } k_1 \neq k_2,
\end{cases} \quad (3.3.1)$$

where $k, k_1, k_2 = 1, 2, \ldots, N$. Let $I_{B_k}(Y)$ be an indicator random variable

$$I_{B_k}(Y) = \begin{cases}
1 & \text{if } Y \in B_k, \\
0 & \text{otherwise}.
\end{cases} \quad (3.3.2)$$

Thus, $\Omega = \bigcup_{k=1}^{N} I_{B_k}^{-1}(1)$ is a decomposition of the sample space $\Omega$, where

$$I_{B_i}^{-1}(1) \cap I_{B_j}^{-1}(1) = \emptyset, \text{ for } i \neq j. \quad (3.3.3)$$
Given any point \( q = (q_1, q_2, \cdots, q_d) \), we use \( Y \leq q \) to denote \( Y_i \leq q_i \) for \( i = 1, 2, \cdots, n \). According to the law of total probability, we can obtain

\[
\Pr(Y \leq q) = \sum_{k=1}^{N} \Pr(Y \leq q | I_{B_k}(Y) = 1) \Pr(I_{B_k}(Y) = 1).
\] (3.3.4)

Using Bayes’ rule, equation (3.3.4) implies that we can define a new random variable \( Y_k : I_{B_k}^{-1}(1) \rightarrow B_k \) on the probability space \((I_{B_k}^{-1}(1), \mathcal{F} \cap I_{B_k}^{-1}(1), P(|I_{B_k}(Y) = 1))\) subject to a conditional PDF

\[
f_k(y_k | I_{B_k}(Y) = 1) = \frac{f(y_k)}{\Pr(I_{B_k}(Y) = 1)}
\] (3.3.5)
in each random element \( B_k \), where

\[
\Pr(I_{B_k} = 1) = \int_{B_k} f(y) dy > 0.
\] (3.3.6)

Let \( u(x, t; Y) \in L_2(\Omega, \mathcal{F}, P) \) denote a second-order space-time related random field. For simplicity, we may drop \( x \) and \( t \).

**Proposition 3.3.1.** Let \( \mathcal{P}_p u(Y) \) denote the Galerkin projection of \( u(Y) \) onto the polynomial chaos basis \( \{ \phi_\alpha(Y) \} \) up to polynomial order \( p \), \( |\alpha| \leq p \). If \( \mathcal{P}_p u(Y) \) converges to \( u(Y) \) in the \( L_2 \) sense with respect to the PDF \( f(y) \), then \( \mathcal{P}_p u(Y_k) \) converges to \( u(Y_k) \) in the \( L_2 \) sense with respect to the conditional PDF \( f_k(y_k | I_{B_k}(Y) = 1) \), \( k = 1, 2, \cdots, N \).

**Proof.** According to the assumption, we know that

\[
\mathbb{E}[(u(Y) - \mathcal{P}_p u(Y))^2] = \int_B (u(y) - \mathcal{P}_p u(y))^2 f(y) dy \rightarrow 0, \quad \text{as } p \rightarrow \infty.
\]

By noting

\[
\mathbb{E}[(u(Y) - \mathcal{P}_p u(Y))^2] = \sum_{k=1}^{N} \Pr(I_{B_k}(Y) = 1) \int_{B_k} (u(y_k) - \mathcal{P}_p u(y_k))^2 f_k(y_k | I_{B_k}(Y) = 1) dy_k \quad (3.3.7)
\]
we obtain
\[ \int_{B_k} (u(y_k) - P_p u(y_k))^2 f_k(y_k | I_{B_k}(Y) = 1) dy_k \to 0, \quad \text{as } p \to \infty, \]

since both \( \Pr(I_{B_k}(Y) = 1) \) and the integrand on the right-hand side of equation (3.3.7) are positive for \( k = 1, 2, \ldots, N \).

To this end, we know that there exists a local polynomial approximation for each random element \( B_k \), which converges in the \( L_2 \) sense with respect to the local conditional PDF. For the orthogonal basis \( \{ \phi_\alpha(Y) \} \) on the entire random space, \( u(Y) - P_p u(Y) \) is orthogonal to the space \( V(p, Y) := \text{span}\{ \phi_\alpha(Y) | |\alpha| \leq p \} \). We note here that although \( P_p u(Y_k) \) converges to \( u(Y_k) \) in the \( L_2 \) sense within random element \( B_k \), \( u(Y_k) - P_p u(Y_k) \) is not orthogonal to the space \( V(p, Y_k) \) with respect to the conditional PDF \( f_k(\cdot | I_{B_k}(Y) = 1) \), since the orthogonality \( \mathbb{E}[\phi_\alpha \phi_\beta] = \delta_{\alpha\beta} \) is, in general, only valid on the entire random space of \( Y \) with respect to the PDF \( f(y) \). Due to the efficiency of orthogonal polynomials in representing stochastic processes, we will reconstruct the local polynomial chaos modes numerically to make them mutually orthogonal with respect to the local conditional PDF \( f(y_k | I_{B_k}(Y) = 1) \). According to [97, theorem 2.1.1], such an orthogonal polynomial system \( \{ \phi_k,\alpha(Y_k) \} \) always exists. Since in each random element we perform a spectral expansion as in gPC, we call this method "multi-element generalized polynomial chaos (ME-gPC)".

To approximate a random field \( u(Y) \) using ME-gPC, we expand the random field spectrally in each element \( B_k \), then re-construct the entire random field by the following proposition.

**Proposition 3.3.2.** Let \( \hat{u}_k(Y_k) \) be the local polynomial chaos expansion in element \( B_k \) using the local basis \( \phi_{k,\alpha}(Y_k) \) with \(|\alpha| \leq p \). The approximation on the entire random field can be defined as

\[
\tilde{u}^p(Y) = \sum_{k=1}^{N} \hat{u}_k(Y_k) I_{B_k}(Y) = \sum_{k=1}^{N} \sum_{|\alpha| \leq p} \hat{u}_{k,\alpha} \phi_{k,\alpha}(Y_k(Y)) I_{B_k}(Y), 
\tag{3.3.8}
\]

which converges to \( u(Y) \) in the \( L_2 \) sense, in other words,

\[
\int_B (u^p(y) - u(y))^2 f(y) dy \to 0, \quad \text{as } p \to \infty. 
\tag{3.3.9}
\]
Proof. It is easy to see that

\[
\int_B (u^p(y) - u(y))^2 f(y) dy = \sum_{k=1}^N \Pr(I_{B_k}(Y) = 1) \int_{B_k} (u^p(y_k) - u(y_k))^2 f_k(y_k | I_{B_k}(Y) = 1) dy_k.
\]

Since \( u^p(Y_k) = \hat{u}_k(Y_k) \) and the spectral expansion \( \hat{u}_k(Y_k) \) converges locally in the \( L_2 \) sense with respect to \( f(y_k | I_{B_k}(Y) = 1) \), it is clear that the right-hand side of above equation goes to zero as \( p \to \infty \). The conclusion follows immediately.

By Bayes’ rule and the law of total probability, any statistics can be obtained as

\[
\int_B g(u(y)) f(y) dy \approx \sum_{k=1}^N \Pr(I_{B_k}(Y) = 1) \int_{B_k} g(\hat{u}_k(y_k)) f_k(y_k | I_{B_k}(Y) = 1) dy_k,
\]

where \( g(\cdot) \in L_1(\Omega, \mathcal{F}, P) \) is a functional of random field \( u(Y) \).

3.4 Numerical orthogonal polynomials

3.4.1 Algorithms

We next discuss the numerical construction of orthogonal polynomials with respect to a conditional PDF \( f_k(\cdot | I_{B_k}(Y) = 1) \). For simplicity, we here only discuss the construction of one-dimensional orthogonal polynomials, since the high-dimensional basis can be obtained using tensor products of one-dimensional basis.

It is a distinctive feature of orthogonal polynomials, compared to other orthogonal systems, that they satisfy a three-term recurrence relation,

\[
\begin{align*}
\pi_{i+1}(\tau) &= (\tau - c_{i,0}) \pi_i(\tau) - c_{i,1} \pi_{i-1}(\tau), & i = 0, 1, \cdots, \\
\pi_0(\tau) &= 1, & \pi_{-1}(\tau) = 0,
\end{align*}
\] (3.4.1)
where \( \{ \pi_i(\tau) \} \) is a set of (monic) orthogonal polynomials,

\[
\pi_i(\tau) = \tau^i + \text{lower-degree terms}, \quad i = 0, 1, \cdots
\]  

(3.4.2)

and the coefficients \( c_{i,0} \) and \( c_{i,1} \) are uniquely determined by a positive measure, which corresponds to a probability measure in the construction we propose herein.

For a continuous measure \( m(\tau) \) there are two classical methods to compute the recurrence coefficients \( c_{i,0} \) and \( c_{i,1} \): the Stieltjes procedure and the modified Chebyshev algorithm [35]. The Stieltjes procedure uses the fact that the coefficients \( c_{i,0} \) and \( c_{i,1} \) can be expressed by the following simple formulas

\[
c_i,0 = \frac{\left( \pi_{i+1}, \pi_i \right)}{\left( \pi_i, \pi_i \right)}, \quad i = 0, 1, 2, \cdots
\]  

(3.4.3)

and

\[
c_0,1 = (\pi_0, \pi_0), \quad c_{i,1} = \frac{\left( \pi_{i+1}, \pi_i \right)}{\left( \pi_{i-1}, \pi_{i-1} \right)}, \quad i = 1, 2, \cdots,
\]  

(3.4.4)

where \((\cdot, \cdot)\) denotes the inner product in terms of the measure \( m(\tau) \). The above two formulas together with the recurrence relation (3.4.1) can be used to calculate recursively as many coefficients \( c_{i,0} \) and \( c_{i,1} \) as desired.

The modified Chebyshev algorithm is a generalization of the Chebyshev algorithm [35]. The Chebyshev algorithm relies on the fact that the first \( n \) pairs of recurrence coefficients \( c_{i,0} \) and \( c_{i,1} \), \( i = 0, 1, \cdots, n - 1 \), can be uniquely determined by the first \( 2n \) moments \( \mu_i \)

\[
\mu_i = \int_B \tau^i dm(\tau), \quad i = 0, 1, \cdots, 2n - 1.
\]  

(3.4.5)

Analytical formulas are known which express \( c_{i,0} \) and \( c_{i,1} \) in terms of Hankel determinants in these moments. However, this algorithm is not reliable for a big \( n \) due to the increasing sensitivity of these formulas to small errors. The modified Chebyshev algorithm replaces the power \( \tau^i \) with a properly.
chosen polynomial \( h_i(\tau) \) of degree \( i \)

\[
\nu_i = \int_B h_i(\tau) dm(\tau), \quad i = 0, 1, \cdots, 2n - 1. \quad (3.4.6)
\]

Generally, we can assume that \( h_i(\tau) \) are monic orthogonal polynomials satisfying a three-term relation

\[
h_{i+1}(\tau) = (\tau - \hat{c}_{i,0}) h_i(\tau) - \hat{c}_{i,1} h_{i-1}(\tau), \quad i = 0, 1, \cdots,
\]

\[
h_0(\tau) = 1, \quad h_{-1}(\tau) = 0. \quad (3.4.7)
\]

Using the \( 2n \) modified moments in equation (3.4.6) and the \( 2n - 1 \) pairs of recurrence coefficients \( \hat{c}_{i,0} \) and \( \hat{c}_{i,1}, i = 0, 1, \cdots, 2n - 2, \) in equation (3.4.7), the first \( n \) desired pairs of recurrence coefficients \( c_{i,0} \) and \( c_{i,1}, i = 0, 1, \cdots, n - 1, \) can be generated [35].

For a discrete measure

\[
dm_M(\tau) = \sum_{i=1}^{M} w_i \delta(\tau - \tau_i) d\tau, \quad i = 0, 1, \cdots, M \quad (3.4.8)
\]

with \( \delta \) being the Dirac delta function, we have another choice: the Lanczos algorithm [23, 35]. Given equation (3.4.8), there exists an orthogonal matrix \( Q^{(M+1)\times(M+1)} \) with the first column being \( [1, 0, \cdots, 0]^T \in \mathbb{R}^{(M+1)\times1} \) such that

\[
Q^T A_M Q = J_M, \quad (3.4.9)
\]

where

\[
A_M = \begin{bmatrix}
1 & \sqrt{w_1} & \sqrt{w_2} & \cdots & \sqrt{w_M} \\
\sqrt{w_1} & \tau_1 & 0 & \cdots & 0 \\
\sqrt{w_2} & 0 & \tau_2 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\sqrt{w_M} & 0 & 0 & \cdots & \tau_M
\end{bmatrix}, \quad (3.4.10)
\]
and $J_M$ is the Jacobian matrix

$$
J_M = \begin{bmatrix}
1 & \sqrt{c_{0,1}} & 0 & \cdots & 0 \\
\sqrt{c_{0,1}} & c_{0,0} & \sqrt{c_{1,1}} & \cdots & 0 \\
0 & \sqrt{c_{1,1}} & c_{1,0} & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & c_{M-1,0}
\end{bmatrix}
$$

(3.4.11)

The above algorithms were presented in [35] and a Fortran package based on these algorithms was proposed in [36]. To implement the Stieltjes procedure and the modified Chebyshev algorithm we need to evaluate the inner product with high precision. In practice, we usually employ a Gauss-type quadrature rule. Such a quadrature rule can be regarded as a discrete measure that yields the corresponding discrete versions of these two methods [35]. The stability of the Stieltjes procedure is not completely clear. One interesting aspect about this procedure is that it is reliable in many cases where a discretization for the inner product of a continuous measure is employed while it breaks down in some cases where the measure itself is discrete. In [32], the Stieltjes procedure for discrete measures is discussed and improved. The stability of the modified Chebyshev algorithm is determined by the condition of the map from the $2n$ modified moments to the $n$ pairs of recurrence coefficients. Such a condition was studied in [35, 31]. The Lanczos algorithm has good stability properties but it may be considerably slower than the Stieltjes procedure. In this work, we use the Stieltjes procedure and the Lanczos algorithm.

It can be seen that the basic operation in the above procedures is the inner product with respect to a given measure $m(\tau)$. Since an explicit formula is generally unavailable, we discretize the inner product using an appropriate quadrature rule. We know that a $n$-point Gauss-type quadrature rule can reach the highest algebraic accuracy of order $2n - 1$. However, such quadrature points and the corresponding weights are not in general explicitly known. In this work, we use the interpolatory quadrature rule [74] relative to the Legendre weight function on $[-1,1]$. A $n$-point interpolatory quadrature rule has an algebraic precision $n - 1$. Compared to the Gauss-Legendre quadrature rule...
rule, the advantage of the interpolatory quadrature rule is that the nodes and weights are explicitly known, so the expensive Newton-Raphson iterations are avoided.

Using the Stieltjes procedure or the Lanczos algorithm, the recurrence coefficients $c_{i,0}$ and $c_{i,1}$ are computed iteratively using the following stopping criterion \[35\]

$$|c_{i,1}^s - c_{i,1}^{s-1}| \leq \epsilon c_{i,1}^s, \quad i = 0, 1, \ldots, n - 1,$$

(3.4.12)

where $s$ is the iteration step and $\epsilon$ is the relative error.

If a measure $m(\tau)$ is given and orthogonal polynomials up to $p$-th order are needed, we first compute the recurrence coefficients $c_{i,0}$ and $c_{i,1}$, $i = 0, 1, \ldots, p - 1$. Once the recurrence coefficients are obtained, the orthogonal polynomials are uniquely determined. From the recurrence coefficients, the Gauss-type quadrature points and the corresponding integration weights can be efficiently derived [36].

It is well known that the gPC method with Galerkin projection usually relies on a three-dimensional table $E[\pi_l \pi_m \pi_n]$ for the Galerkin projection [43]. We here use the notation $E[\cdot]$ freely to denote the ensemble average with respect to a given measure $m(\tau)$ since $m(\tau)$ can be regarded as a probability measure in this work. If the recurrence coefficients are obtained — in other words, the orthogonal polynomials are uniquely determined — we can use any numerical integration formula to calculate the table $E[\pi_l \pi_m \pi_n]$. Here we consider the quadrature rule. We know that for the measure $m(\tau)$, there exists, for each $M \in \mathbb{N}$, a quadrature rule

$$\int_{\mathbb{M}} f(\tau) d\sigma(\tau) = \sum_{i=1}^{M} w_i f(\tau_i) + R_{M}(f),$$

(3.4.13)

where $R_{M}(f) = 0$ if $f$ is a polynomial of degree $\leq 2M - c$. The value of $c$ is determined by the type of quadrature used which can be either classical Gauss ($c = 1$), Gauss-Radau ($c = 2$) or Gauss-Lobatto ($c = 3$). If we do not consider the numerical accuracy of the recurrence coefficients, we need

$$M = \left\lfloor \frac{3p + c}{2} \right\rfloor$$

(3.4.14)
quadrature points to get the three-term integration $E[p_{i}p_{m}p_{n}]$ with zero error, where $[*]$ denotes the smallest integer no less than *. Thus, we need to compute all the recurrence coefficients $\alpha_i$ and $\beta_i$ with $i \leq M - 1$ although we employ polynomials only up to order $p$.

For the gPC method with collocation projection, a proper set of grids on the range of the random variable is needed. Usually the grids are chosen from the (Gauss) quadrature points corresponding to the PDF. More details about how to choose collocation grids are presented in section 3.5.

### 3.4.2 Implementation

Next we discuss issues related to the numerical implementation. Based on the decomposition of random space, it is straightforward to re-construct orthogonal polynomials with respect to the conditional PDFs of the local random variables $Y_k$. However, such a procedure is not robust in practice because $(\pi_i, \pi_i)$ may decay fast and cause a potential problem of overflow or underflow in computer.

**Proposition 3.4.1.** For any orthogonal polynomials defined on $[a, b]$, where $a$ and $b$ are positive, we have

$$
(p_{i+1}, p_{i+1}) \leq (b - a)^2 (p_i, p_i). \tag{3.4.15}
$$

**Proof.** From equation (3.4.3), we can obtain that $a \leq \alpha_i \leq b$. Using equation (3.4.1) and the Cauchy-Schwarz inequality, we know that

$$
(p_{i+1}, p_{i+1}) = ((\tau - \alpha_i)p_i, (\tau - \alpha_i)p_i) \leq ((\tau - \alpha_i)p_i, (\tau - \alpha_i)p_i)^{1/2}(p_{i+1}, p_{i+1})^{1/2}, \tag{3.4.16}
$$

which yields that

$$
(p_{i+1}, p_{i+1}) \leq ((\tau - \alpha_i)p_i, (\tau - \alpha_i)p_i). \tag{3.4.17}
$$

The proof concludes by noting $a \leq \alpha_i \leq b$. $\square$

Thus, if the element is small, the $L_2$ norm of $\pi_i$ will lead to underflow quickly. In practice, we
re-scale the random elements with *finite* boundaries by the following linear transformation

\[
Y_{k,i} = \frac{b_{k,i} - a_{k,i}}{2} \hat{Y}_{k,i} + \frac{b_{k,i} + a_{k,i}}{2},
\]

(3.4.18)

where we map the random variable \( Y_k \) defined in element \( k \) to a random variable \( \hat{Y}_k \) defined in \([-1,1]^d\). The PDF of \( \hat{Y}_k \) can be obtained as

\[
f(\hat{y}_k) = \det \left[ \frac{\partial y_k}{\partial \hat{y}_k} \right] f(y_k(\hat{y}_k)|I_{B_k}(Y) = 1) = \frac{f(y_k(\hat{y}_k))}{\Pr(I_{B_k}(Y) = 1)} \prod_{i=1}^{n} \frac{b_{k,i} - a_{k,i}}{2}.
\]

(3.4.19)

Compared to \( Y_k, \hat{Y}_k \) are much more tractable for the numerical construction of orthogonal polynomials as demonstrated in section 3.4.2. After such a transformation is employed, we can apply the ME-gPC scheme with respect to the new random variable \( \hat{Y}_k \) instead of the random variable \( Y_k \). Note that such a mapping is usually unnecessary for random elements with at least one *infinite* boundary because these elements can be related to a small probability \( \Pr(I_{B_k}(Y) = 1) \).

We first discuss the construction of orthogonal polynomials for random variables with some commonly used distributions. Due to the nice properties of uniform distribution, the orthogonality of Legendre-chaos can be naturally inherited in the decomposition of random space, which means that the polynomial construction is unnecessary for the uniform distribution. Here we focus on the Beta and Gaussian distributions, for which we need the on-the-fly polynomial construction.

**Problem of overflow and underflow**

We consider a random variable defined in \([-1,1]\) with the following Beta distribution

\[
f(x) = \frac{(1-x)^{\alpha}(1+x)^{\beta}}{2^{\alpha+\beta+1}B(\alpha+1,\beta+1)}, \quad -1 \leq x \leq 1.
\]

(3.4.20)

In a random element \([0,0.001]\), we define a random variable \( X \) with a conditional PDF

\[
f(x) = \frac{f(x)}{\int_0^{0.001} f(x)dx},
\]

(3.4.21)
and map $X$ to another random variable $Y$ defined in $[-1,1]$ with a PDF

$$f(y) = \frac{0.0005f(x(y))}{\int_0^1 f(x)dx},$$  \hspace{1cm} (3.4.22)

where

$$x(y) = 0.0005y + 0.0005. \hspace{1cm} (3.4.23)$$

Here we let $\alpha = 1$ and $\beta = 4$. In Fig. 3.1, we show the orthogonal polynomials for random variable $X$ from first-order up to fourth-order. We can see that the maximum values of these polynomials decrease very quickly approaching the machine accuracy. In Fig. 3.2, the corresponding orthogonal polynomials for random variable $Y$ are shown. It can be seen that their shapes are the same as those of the orthogonal polynomials for $X$, however, they are well scaled.

![Figure 3.1: Orthogonal polynomials for random variable $X$, $\pi_1$ to $\pi_4$.](image)

**Orthogonal polynomials for Beta distribution**

From the ME-gPC scheme, we know that the orthogonal basis in each random element depends on a particular part of the PDF of the random inputs. We now demonstrate such a dependence using
a Beta-type random variable $X$ with $\alpha = 1$ and $\beta = 4$. For simplicity, we only consider two random elements: $[-1, 0]$ and $[0, 1]$. We define a random variable $X_l$ in $[-1, 0]$ with a conditional PDF

$$f_l(x_l) = \frac{f(x_l)}{\int_{-1}^{0} f(x)dx}$$

and another random variable $X_r$ in $[0, 1]$ with a conditional PDF

$$f_r(x_r) = \frac{f(x_r)}{\int_{0}^{1} f(x)dx},$$

where $f(x)$ is the PDF of $X$. Due to the potential problem of overflow (underflow), we will not construct orthogonal basis according to the above two conditional PDFs, but first map $X_l$ and $X_r$ to $Y_l$ and $Y_r$, respectively, using the following transformation

$$X_l = \frac{1}{2} Y_l - \frac{1}{2}, \quad Y_r = \frac{1}{2} Y_r + \frac{1}{2},$$

then construct orthogonal polynomials in terms of $Y_l$ and $Y_r$. In Fig. 3.3 and 3.4, we show the PDFs of $Y_l$ and $Y_r$ on the left and the corresponding orthogonal polynomials on the right. It is clear that...
these two sets of orthogonal polynomials are quite different due to their different weight functions (i.e., local PDFs).

Figure 3.3: Left: PDF of random variable $Y_t$ induced by Beta distribution with $\alpha = 1$ and $\beta = 4$; Right: Orthogonal polynomials for $Y_t$.

Figure 3.4: Left: PDF of random variable $Y_r$ induced by Beta distribution with $\alpha = 1$ and $\beta = 4$; Right: Orthogonal polynomials for $Y_r$.

Orthogonal polynomials for Gaussian distribution

Some random distributions, e.g., Gaussian distribution and Gamma distribution, have long tails. Next, we demonstrate the decomposition of random space and the corresponding orthogonal polynomials for this type of random variables. Given a Gaussian random variable $X \sim N(0,1)$, we decompose the random space into three random elements $(-\infty, -a]$, $[-a, a]$ and $[a, +\infty)$, where $a$ is a positive constant. In the middle element $[-a, a]$, we define a random variable $X_m$ with a
conditional PDF

\[
f_m(x_m) = \frac{e^{-\frac{1}{2}x_m^2}}{\int_a^x e^{-\frac{1}{2}t^2} dt}.
\]

(3.4.27)

In one of the tail elements, say, \([a, +\infty)\), we define a random variable \(X_\infty\) with a conditional PDF

\[
f_\infty(x_\infty) = \frac{e^{-\frac{1}{2}x_\infty^2}}{\int_a^{\infty} e^{-\frac{1}{2}t^2} dt}.
\]

(3.4.28)

We choose \(a\) in such a way that \(\Pr(X \geq a) < \epsilon\), where \(\epsilon\) is a small positive constant. From equation (4.1.6), we know that the error contribution of tail elements can be effectively weakened by the small value of \(\Pr(X \geq a)\), thus we usually do not decompose the tail elements in applications, but the middle element \([-a, a]\). Furthermore, there is no overflow or underflow problem for the tail elements due to the infinite boundary. Here we take \(a = 5\), which yields \(\Pr(X \geq 5) = 2.87 \times 10^{-7}\). In Fig. 3.5, we show the PDF of the re-scaled random variable \(Y_m = X_m/a\) on the left and the corresponding orthogonal polynomial basis on the right. The orthogonal polynomials have similar shapes with Hermite polynomials, however, they are bounded now. In Fig. 3.6, the conditional PDF of \(X_\infty\) is shown on the left and the corresponding orthogonal basis on the right. Due to the infinite boundary the orthogonal polynomials look almost like Laguerre polynomials.

![Figure 3.5: Left: PDF of random variable \(Y_m\) induced by Gaussian distribution; Right: Orthogonal polynomials for \(Y_m\).](image)

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Efficiency of the solver for numerical orthogonal polynomials

It is clear that the ME-gPC method relies heavily on the numerical orthogonal polynomials. We now test the solver for the recurrence coefficients, quadrature points and integration weights. Due to the absence of explicit formulas, we use the following example. Given a weight function $w(x)$ defined in $[a, b]$, we decompose $[a, b]$ into small elements and construct numerical orthogonal polynomials in these elements. Then we map a simple function $g(x) = c_1 x^p + c_2$ onto such elements numerically and compute the errors on the quadrature points in each element. In the $k$-th element $[a_k, b_k]$, $g(x) = c_1 x^p + c_2$ should have the form

$$g_k(\tau) = c_1 \left( \frac{b_k - a_k}{2} \tau + \frac{b_k + a_k}{2} \right)^p + c_2,$$

(3.4.29)

since all the orthogonal polynomials in element $k$ are defined in $[-1, 1]$. Expressed by the numerical orthogonal polynomials, $g_k(\tau)$ has another form

$$\hat{g}_k(\tau) = \sum_{i=0}^{p} \nu_{k,i} \pi_{k,i}(\tau),$$

(3.4.30)

where the coefficients $\nu_{k,i}$ can be obtained by the inner product

$$\nu_{k,i} = \langle g_k(\tau), \pi_{k,i} \rangle / \langle \pi_{k,i}, \pi_{k,i} \rangle.$$

(3.4.31)
Next, we define the $L_\infty$ error on the quadrature points over all the elements

$$
\epsilon_{\text{max}} = \max_{k=1,2,\cdots,N} \left| \frac{g_k(\tau_j) - \hat{g}_k(\tau_j)}{g_k(\tau_j)} \right|, \quad j = 1, 2, \cdots, n,
$$

(3.4.32)

where $N$ is the number of elements and $n$ is the number of quadrature points in each element.

In Table 3.2, we show $\epsilon_{\text{max}}$ and the time cost for different uniform meshes. We record the time used for the computation of recurrence coefficients, quadrature points, integration weights and the table $E[\pi_{k,i} \pi_{k,j} \pi_{k,m}]$, which represents the total set-up requirement for the standard Galerkin gPC method. We can see that the biggest error $\epsilon_{\text{max}}$ occurs in a one-element mesh. As the number of elements increases, $\epsilon_{\text{max}}$ is almost of the same order as the relative error of recurrence coefficients, which implies that the numerical polynomials can achieve good orthogonality and high accuracy.

Furthermore, numerical experiments show that the time cost for the numerical polynomials is very small. For example, $10^5$ sets of orthogonal polynomials up to 10-th order can be computed within less than one and a half minutes on a 1.5GHz AMD CPU. In practice, we do not need so many elements and the time cost for the construction of orthogonal polynomials can be actually ignored compared to the time used by the gPC solver.

Table 3.2: Error $\epsilon_{\text{max}}$ and time cost for a uniform mesh with $N$ elements. $p = 10$ and $g(x) = x^{10} + 1$. The relative error for the recurrence coefficients was set to be $10^{-13}$. $\alpha = 1$ and $\beta = 4$ for Beta distribution. The middle element $[-6,6]$ was used for Gaussian distribution. The computations were performed on a 1.5GHz AMD CPU.

<table>
<thead>
<tr>
<th>$N$</th>
<th>Beta Distribution</th>
<th>Gaussian Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\epsilon_{\text{max}}$</td>
<td>time (sec)</td>
</tr>
<tr>
<td>1</td>
<td>7.98e-11</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>2</td>
<td>2.58e-12</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>4</td>
<td>2.38e-13</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>6</td>
<td>4.58e-13</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>8</td>
<td>3.53e-13</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>10</td>
<td>7.45e-13</td>
<td>0.01</td>
</tr>
<tr>
<td>$10^2$</td>
<td>9.45e-13</td>
<td>0.08</td>
</tr>
<tr>
<td>$10^3$</td>
<td>1.00e-12</td>
<td>0.79</td>
</tr>
<tr>
<td>$10^4$</td>
<td>3.69e-12</td>
<td>7.92</td>
</tr>
<tr>
<td>$10^5$</td>
<td>9.28e-13</td>
<td>76.43</td>
</tr>
</tbody>
</table>
3.5 ME-gPC expansion of a general operator

We consider a general (nonlinear) operator

\[ L(x, t, u; Y) = g(x, t; Y), \]

(3.5.1)

where the \( \mathbb{R}^n \)-value random variable \( Y \) with i.i.d. components is known. We assume that there exists a unique solution \( u(x, t; Y) \) for the above problem. When employing ME-gPC, we first decompose the parametric space of \( Y \) into \( N \) non-overlapping elements \( B_k \). In each element, we obtain the following local problem

\[ L_k(x, t, u_k; Y_k) = g_k(x, t; Y_k). \]

(3.5.2)

We note that the probability to get such a local problem is \( \Pr(I_{B_k}(Y) = 1) \). Since we usually map the range of \( Y_k \) from \( B_k \) to a cube \((-1, 1)^n\), the local operator \( L_k \) may have a different form from the global operator \( L \). However, the local gPC solver is the same for each element because the mapping of \( Y_k \) can only affect the random inputs, and all operations on \( x \) and \( t \) are unchanged. Thus, we only need to consider the local gPC expansion of the operator \( L_k \).

3.5.1 Galerkin projection

Given a local gPC basis \( \{ \phi_\alpha \}_|\alpha| \leq p \), we expand \( u_k \) spectrally as

\[ u_k = \sum_{|\alpha| \leq p} u_{k, \alpha} \phi_\alpha. \]

We substitute the gPC expansion of \( u_k \) into the local problem (3.5.2) and define the residual as

\[ R_k = L_k \left( x, t, \sum_{|\alpha| \leq p} u_{k, \alpha} \phi_\alpha; Y_k \right) - g_k(x, t; Y_k). \]

(3.5.3)
To obtain the equation for unknowns $u_{k,\alpha}$, we let the residual $R_k$ be orthogonal with the local gPC basis $\{\phi_\alpha\}_{|\alpha| \leq p}$, resulting in

$$E \left[ L_k \left( x, t, \sum_{|\alpha| \leq p} u_{k,\alpha} \phi_\alpha ; Y_k \right) \phi_\beta \right] = E [g_k(x, t; Y_k) \phi_\beta].$$  

(3.5.4)

The equation (3.5.4) is usually a deterministic ODE or PDE system of $u_{k,\alpha}$.

The advantage of Galerkin projection is that it can minimize the $L_2$ error by forcing the residual to be orthogonal with the gPC basis. The disadvantage of Galerkin projection is that if the operator $L_k$ is nonlinear, all the unknowns $u_{k,\alpha}$ will be coupled with each other, which introduces a real challenge to design an efficient deterministic solver.

### 3.5.2 Collocation projection

It is well known that collocation projection is an easy and efficient technique to deal with nonlinearity in deterministic numerical methods, e.g., spectral (element) methods. The collocation projection can also be used to construct a gPC solver [99, 114]. The procedure is very simple.

We choose a proper set of grid points $\{y^i_k\}$ on the range of $Y_k$ and let the equation (3.5.2) be satisfied on each grid point as

$$L_k(x, t, u_k; y^i_k) = g_k(x, t; y^i_k).$$  

(3.5.5)

We note that equation (3.5.5) is deterministic in terms of $u_k$. In other words, we can employ the solver for a deterministic case and do not have to construct a new gPC solver as we do for the Galerkin projection.

Although the procedure of collocation projection is similar with the Monte Carlo method, we should note that collocation projection is actually a deterministic algorithm since the grid points $y^i_k$ are prescribed, not sampled as in the Monte Carlo procedure.

If the solution is smooth enough in the parametric space with respect to $y_k$, the efficiency of collocation projection is determined by how to choose the grid points $y^i_k$. Quadrature points by full tensor products can be efficient for low-dimensional ($n \leq 3$) cases since they can provide the highest

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
degree of exactness for numerical integrations; however, they are not efficient for high-dimensional cases because the number of grid points increases exponentially, i.e., $M^n$. By choosing the quadrature points carefully, a high accuracy can be maintained with a much smaller number $M(\log M)^n$ of grid points, i.e., sparse grids.

**Sparse grids**

Let us consider the numerical integration of function $g(y)$ over the $n$-dimensional domain $\Lambda = \Lambda_1 \times \Lambda_2 \times \cdots \times \Lambda_n$, where $\Lambda_i$ can be possible unbounded. We consider the quadrature formula

$$Q_{m,\ell}(g) = \sum_{i=1}^{m} a_i g(y_i), \quad a_i \in \mathbb{R}, \ y_i \in \Omega$$

(3.5.6)

for the $n$-dimensional integral

$$I_n(g) = \int_{\Lambda} g(y)w(y)dy,$$

(3.5.7)

where $m$ is the number of grid points, $\ell$ is the degree of exactness of the quadrature formulas, and we assume that the weight function $w(y)$, i.e., PDF in ME-gPC, takes the tensor product form

$$w(y) = w_1(y_1) \cdots w_n(y_n).$$

(3.5.8)

Starting from a one-dimensional quadrature formula

$$U_j^i = \sum_{k=1}^{m_{ij}} g(y_k)a_k, \quad j = 1, 2, \cdots, d$$

(3.5.9)

we employ the Smolyak formulas

$$A(q, n) = \sum_{q-n+1 \leq ||i|| \leq q} (-1)^{q-||i||} \binom{n-1}{q-||i||} (U_1^{i_1} \otimes \cdots \otimes U_n^{i_n})$$

(3.5.10)

to compute the $n$-dimensional integrals, where $q \geq n$, $i \in \mathbb{N}^n$, $||i|| = i_1 + i_2 + \cdots + i_n$ and $m_{ij}$ is the number of quadrature points in dimension $j$. We use $m(q, n)$ to denote the total number of
quadrature points used by $A(q, n)$. Let

$$\tilde{m}(q, n) = \sum_{q-n+1 \leq ||l|| \leq q} m_{i_1} \cdots m_{i_n}.$$  

Let $S^j_i$ denote the set of quadrature points used by $U^j_i$. We note that $m(q, n) \leq \tilde{m}(q, n)$. We now consider a variation of the Smolyak formula $A(q, n)$. By defining a one-dimensional difference operator

$$\Delta^j_i = U^j_i - U^{i-1}_i, \quad U^0_j = 0,$$  

(3.5.11)

$A(q, n)$ can be rewritten as

$$A(q, n) = \sum_{||l|| \leq q} (\Delta^1_l \otimes \cdots \otimes \Delta^n_l).$$  

(3.5.12)

**Case 1: $S^j_i$ are nested**

We consider the Clenshaw-Curtis formulas, which uses non-equidistant explicit abscissas given as the zeros or the extreme points of the Chebyshev polynomials. When the extreme points are used, $S^j_i$ are nested, which means that

$$S^{j-1}_i \subset S^j_i.$$  

We choose

$$m_i = 2^{i-1} + 1, \quad i \geq 2.$$  

The abscissas are

$$y^i_k = -\cos \frac{\pi (k - 1)}{m_i - 1}, \quad k = 1, \ldots, m_i$$  

and the weights are

$$w^i_1 = w^i_{m_i} = \frac{1}{m_i (m_i - 2)}$$  

$$w^i_k = \frac{2}{m_i - 1} \left( 1 + 2 \sum_{k=1}^{(m_i-1)/2} \frac{1}{1 - 4k^2} \cos \frac{2\pi k(i - 1)}{m_i - 1} \right), \quad 2 \leq i \leq m_i - 1.$$
where $\sum'$ indicates that the last term of the sum is halved. We note that $U_j^i$ is exact for all polynomials of degree less than $n_i$ with $w_j(y_j) = 1$. For other cases, a linear system of equations may need to be solved for the correct weights.

**Proposition 3.5.1.** Let $q = \sigma n + \tau$ with $\sigma \in \mathbb{N}$ and $\tau \in \{0, 1, \ldots, n-1\}$. Then $\mathcal{A}(q, n)$ has a degree of exactness

$$\ell(q, n) = \begin{cases} 2(q - n) + 1 & \text{if } q < 4n; \\ 2^{\sigma-2}(n + \tau + 1) + 2n - 1 & \text{otherwise.} \end{cases}$$

**Proof.** See [75]. □

The advantage of nested formulas is that the $m(q, n)$ can be significantly reduced. If $n \gg 1$, we have

$$m(k + n, n) \approx \frac{2^k}{k!} n^k \text{ and } \tilde{m}(k + n, n) \preceq \frac{3^k}{k!} n^k. \quad (3.5.13)$$

Since we consider arbitrary PDF in ME-gPC, we have to modify $U_j^i$ to make it work for any PDF while maintaining the nested grid points $S_j^i$. The simplest way is to just modify the weights according to the given PDF. If so, we note that the degree of exactness of $U_j^i$ can not be $n_i$ any more since the integrand includes the PDF. If the number of random elements of ME-gPC is large enough, the PDF can be generally approximated very well by low-order polynomials, which implies that we can expect a degree of exactness close to $n_i$.

**Case II: $S_j^i$ are not nested**

We employ Gaussian formulas $U_j^i$ for the weights $w_j$. Gauss quadrature points for arbitrary weight functions can be computed efficiently, see, e.g., [35]. More discussions about the construction of orthogonal polynomials for Galerkin projection of ME-gPC can be found in [106]. We choose

$$m_i = 2^{i-1}$$
as the cardinality of $S_j^i$, which yields a degree
\[ e_i = 2^i - 1 \]
of exactness. For this case, $S_j^i$ is not nested any more and $m(q, n) \approx \tilde{m}(q, n)$. The following proposition holds:

**Proposition 3.5.2.** Let $q = \sigma n + \tau$ with $\sigma \in \mathbb{N}$ and $\tau \in \{0, 1, \cdots, n - 1\}$. Then $A(q, n)$ has a degree of exactness
\[ \ell(q, n) = \begin{cases} 
2(q - n) + 1 & \text{if } q < 3n; \\
2^{\sigma - 1}(n + \tau + 1) - 1 & \text{otherwise}. 
\end{cases} \]

**Proof.** See [75]. □

### 3.6 hp-convergence of ME-gPC

In this section we study the $hp$-convergence of ME-gPC, where $h$ denotes the side length of random element and $p$ the polynomial order. We consider the Legendre-chaos expansion, which corresponds to a uniform distribution.

We first quote the following approximation and scaling results.

**Lemma 3.6.1.** Let $A$ be an open subset of $\mathbb{R}^n$. There exist a constant $C(A)$ such that
\[ \forall v \in H^{m+1}, \inf_{\phi \in P_m(A)} \|v - \phi\|_{H^{m+1}} \leq C(A)\|v\|_{H^{m+1}}, \]
where the semi-norm $\|v\|_{H^{m+1}}$ is defined as
\[ |v|_{H^{m+1}} = \left( \sum_{|\alpha| = m+1} \int_A |\partial_y^\alpha v|^2 \, dy \right)^{1/2} \]

**Proof.** See theorem 3.1.1 in [27]. □

We also need the scaling argument.
Lemma 3.6.2. Let $A$ and $\hat{A}$ be two open subsets of $\mathbb{R}^n$ such that there exists an affine mapping $F(x) = B(x) + b$ of $\hat{A}$ onto $A$ and $F(\hat{A}) = A$. Let $\text{diam}(A) = 1$, $\rho_A = K$, $\text{diam}(\hat{A}) = h$, $\rho_{\hat{A}} = \tilde{K}h$. If a function $\hat{v}$ belongs to the Sobolev space $W^{m,q}(\hat{A})$ for some integer $m \geq 0$ and some number $q \in [1, \infty]$, the function $v = \hat{v} \circ F \in W^{m,q}(A)$ then

$$|\hat{v}|_{H^m(\hat{A})} \leq C h^{m-\frac{q}{2}} |v|_{H^m(A)} \quad (3.6.1a)$$

$$|\hat{v}|_{H^m(\hat{A})} \leq C h^{\frac{q}{2} - m} |v|_{H^m(A)} \quad (3.6.1b)$$

Proof. See theorem 3.1.2 in [27]. □

We are now ready to prove the $hp$-convergence of ME-gPC for Legendre-chaos.

Theorem 3.6.3. Let $u(y) \in H^m(I^n)$ and $D$ be a uniform decomposition of $I^n$ with $h$ being the side length of each element. Then the ME-gPC approximation $u^p(Y) \in P_p(Y)$ given by the Galerkin projection operator $P_p$ in each element converges to $u(Y)$ with an error

$$\|u^p - u\|_{L_2(I^n)} \leq C(m, n) p^{-m} h^m |u|_{H^m(I^n)}, \quad (3.6.2)$$

where we assume that $m \leq p + 1$.

Proof. We consider element $B_k$, where we define $y_k(y) : B_k \mapsto I^n$ by an affine mapping. Let $v(y_k) \in P_p(I^n)$. Using theorem 3.2.2 and lemma 3.6.2, we obtain

$$\|u(y_k) - P_p(y_k)\|_{L_2(I^n)} = \|u - v - P_p(u - v)\|_{L_2(I^n)}$$

$$\leq C p^{-m} \inf_{v \in P_p(I^n)} \|u - v\|_{H^m(I^n)}$$

$$\leq C p^{-m} h^{m - \frac{q}{2}} |u|_{H^m(B_k)}.$$
Using the formula (3.3.10), we obtain

\[
\|u(y) - u^p(y)\|^2_{L_2(I^n)} = 2^n \mathbb{E}[\|(u(y) - u^p(y))^2\|] \\
= 2^n \sum_{k=1}^N \Pr(I_{B_k}(Y) = 1) \mathbb{E}_k[(u(y_k) - P_P u(y_k))^2] \\
= \sum_{k=1}^N \text{vol}(B_k) \|u(y_k) - P_P u(y_k)\|^2_{L_2(I^n)} \\
\leq \sum_{k=1}^N C \|u(y)\|^2_{H^m(B_k)} p^{-2m} h^{2m} \\
\leq C p^{-2m} h^{2m} \sum_{k=1}^N \|u(y)\|^2_{H^m(B_k)} \\
= C p^{-2m} h^{2m} \|u(y)\|^2_{H^m(I^n)}.
\]

\[\square\]

**Remark 3.6.4.** If we take \(m = p + 1\) and consider the error of the second-order moment we recover the \(h\)-convergence rate

\[
|\mathbb{E}[u^2] - \mathbb{E}[P_p^2 u]| \leq Cr(p)N_1^{-2(p+1)}, \quad N_1 \sim 1/h,
\]

shown in [102, 29].

**Remark 3.6.5.** The convergence rate given in theorem 3.6.3 is similar with that of the deterministic \(hp\) polynomial approximation, which usually takes the form

\[
\|u - u^p\|_{L_2} \leq C p^{-m} h^\mu \|u\|_{H^m},
\]

where \(\mu = \min(p + 1, m)\) and \(h\) is the maximum size of finite elements in physical space. Since we will focus on the \(h\)-type adaptivity of ME-gPC, we do not pay much attention on the regularity parameter \(m\). We only use one parameter \(m\).

**Remark 3.6.6.** In the theorem 3.6.3, we use a full tensor-product basis. Such an approximation will suffer from the "curse of dimensionality". In practice, the importance of each random dimension is
usually different, sparse bases can be employed adaptively to reduce the cost while maintaining a good accuracy. Some analytical convergence rates of sparse rates are available in [100]. We will discuss the adaptivity of ME-gPC based on an a-posteriori error estimators later in Chapter 5.

Let $T^{(G)}_p$ indicate an interpolation operator on the space $P_p(I^n)$ based on full tensor-products of one-dimensional Gauss quadrature points. We know for the Legendre-chaos the following approximation results hold.

**Theorem 3.6.7.** If $u(y) \in H^m(I^n)$ and $m > \frac{1}{2}$, then we have

$$
\|u - T^{(G)}_p u\|_{L^2(I^n)} \leq C p^{-m} \|u\|_{H^m(I^n)}.
$$

(3.6.3)

Using this theorem and similar techniques in the proof of theorem 3.6.3, we can obtain the $hp$-convergence of ME-gPC based on collocation projection.

**Theorem 3.6.8.** Let $u(y) \in H^m(I^n)$ and $D$ be a uniform decomposition of $I^n$ with $h$ being the side length of each element. Then the ME-gPC approximation $u^p(Y) \in P_p(Y)$ given by the interpolation operator $T^{(G)}_p$ in each element converges to $u(Y)$ with an error

$$
\|u^p - u\|_{L^2(I^n)} \leq C p^{-m} h^m |u|_{H^m(I^n)},
$$

(3.6.4)

where we assume that $1/2 < m \leq p + 1$.

We see that the best $h$-convergence rate of ME-PCM is also $O(N_1^{-2(p+1)})$ for the second-order moment. We note that $T^{(G)}_p u$ is based on the full tensor-products of Gauss quadrature points, which is actually not practical in numerical implementation, especially for high-dimensional cases. Thus, it is necessary to investigate the relation between $h$-convergence rate of ME-PCM and different choices of quadrature rules, e.g., sparse grids.

We subsequently present some approximation results, which link the $h$-convergence rate of ME-PCM with the degree of exactness given by a general quadrature rule. We start from the one-dimensional case. Let $f(y) \in W^k,\infty(I)$. We consider the approximation behavior of a general quadra-
ature rule for the integral $\int_0^1 f(x) \, dx$. It is well known that the distinct points \( \{y_i; i = 1, 2, \ldots, m\} \) corresponds to a set of (Lagrange) polynomials \( \{l_i; i = 1, 2, \ldots, m\} \) up to polynomial order \( m - 1 \).

We have a polynomial approximation of \( f(x) \) as

$$
\mathcal{L}_m f = \sum_{i=1}^{m} f(y_i) l_i(y) \quad (3.6.5)
$$

corresponding to a quadrature rule

$$
\int_{-1}^{1} f(y) \, dy = \int_{0}^{1} \sum_{i=1}^{m} f(y_i) l_i(y) \, dy + R_m(f) \\
= \sum_{i=1}^{m} f(y_i) w_i + R_m(f), \quad (3.6.6)
$$

where \( w_i \) are the integration weights and \( R_m(f) \) is the residual. By definition, the quadrature rule (3.6.6) has the degree of exactness at least \( m - 1 \), i.e., \( R_m(f) = 0 \) for all \( f \in P_m(I) \). We know that the performance of the quadrature rule (3.6.6) is completely determined by the distribution of the collocation points \( \{y_i; i = 1, 2, \ldots, m\} \) and the highest degree of exactness is \( 2m - 1 \) given by the Gauss quadratures. For the quadrature formula \( Q_{m, \ell} \), it is obvious that \( m - 1 \leq \ell \leq 2m - 1 \). We examine the \( h \)-convergence rate of the quadrature rule \( Q_{m, \ell} \) for the integral \( \int_{-1}^{1} f(y) \, dy \).

**Theorem 3.6.9.** Let \( I = (-1,1) \), \( f(y) \in W^{k+1, \infty}(I) \), and \( I = \bigcup_{i=1}^{N} B_i \) be a non-overlapping mesh of \( I \), where \( B_i \cap B_j = \emptyset \), \( \forall i \neq j \). We assume that \( h = \max_i |b_i - a_i| \). If the quadrature rule \( Q_{m, \ell} \) is employed on each \( B_i \), we have

$$
| \int_{-1}^{1} f(y) \, dy - \sum_{i=1}^{N} Q_{m, \ell}^{B_i} f(y) | \leq C h^{k+1} | f |_{m+1, \infty, I}, \quad (3.6.7)
$$

where \( C \) is constant and \( Q_{m, \ell}^{B_i} \) is the corresponding quadrature rule of \( Q_{m, \ell} \) in element \( B_i \) deduced by a linear mapping.
Proof. For each interval \( B_i = (a_i, b_i) \), we begin with the Taylor formula

\[
f(x) = f(a_i) + \sum_{i=1}^{\ell} \frac{f^{(i)}(a_i)}{i!} (x - a_i)^i + \frac{1}{k!} \int_{a_i}^{x} (x - \theta)^k f^{(k+1)}(\theta) d\theta.
\]

(3.6.8)

According to the Peano theorem, we obtain that

\[
|R_m(f)|_{B_i} = \left| \int_{a_i}^{b_i} K_{\ell}(\theta)f^{(\ell+1)}(\theta)d\theta \right|,
\]

where \( K_{\ell}(\theta) \) is the \( \ell \)th Peano kernel for \( R_m \) in element \( B_i \). Then we obtain

\[
|R_m(f)|_{B_i} \leq \|K\|_{0,1,B_i} |f|_{\ell+1,\infty,B_i}
\]

Let us consider an affine mapping \( F(\hat{x}_i) = (b_i - a_i)\hat{x}_i + a_i \) from \((0,1)\) to \( B_i \). Using the scaling argument, we see that

\[
|R_m(f)|_{F^{-1}(B_i)} \leq \|K\|_{0,1,F^{-1}(B_i)} |f|_{\ell+1,\infty,(F^{-1}(B_i))}
\]

\[
\leq C h^\ell \|K\|_{0,1,B_i} |f|_{\ell+1,\infty,B_i}.
\]

Since

\[
\int_{-1}^{1} f(x) dx = \sum_{i=1}^{N} \frac{b_i - a_i}{2} \int_{0}^{1} f(\hat{x}_i) d\hat{x}_i,
\]

we obtain that

\[
|R_m(f)|_{I} \leq C h \sum_{i=1}^{N} |R_m(f)|_{F^{-1}(B_i)}
\]

\[
\leq C h^{\ell+1} \sum_{i=1}^{N} \|K\|_{0,1,B_i} |f|_{\ell+1,\infty,B_i}
\]

\[
\leq C h^{\ell+1} \|K\|_{0,1,I} |f|_{\ell+1,\infty,I}.
\]
We now generalize the above result to high-dimensional case.

**Theorem 3.6.10.** Let $I^n = (-1, 1)^n$, $f \in W^{m+1, \infty}(I^n)$, and $I^n = \bigcup_{i=1}^{N} B_i$ be a non-overlapping and uniform mesh of $I^n$, i.e., $B_i \cap B_j = \emptyset$, $\forall i \neq j$. Let $h$ indicate the side length of each element and $Q_{m, \ell}$ a quadrature rule with degree of exactness $\ell$ and $m$ collocation points in $I^n$. We use $Q^D_{m, \ell}$ denote the corresponding quadrature rule of $Q_{m, \ell}$ in the domain $D$ given by an affine linear mapping.

The following error estimate holds

\[
| \int_{I^n} f(y)dy - \sum_{i=1}^{N} Q_{m, \ell}^{B_i} f(y) | \leq Ch^{m+1}|f|_{m+1, \infty, I^n}, \tag{3.6.9}
\]

where $C$ is constant.

**Proof.** We define a linear functional

\[
E_D(f) = \int_D f(x)dx - Q_{m, \ell}^{D} f(x), \tag{3.6.10}
\]

whose norm is defined as

\[
\|E_D\|_{k, \infty, D} = \sup_{\|v\|_{k, \infty, D} \leq 1} |E_D(v)|. \tag{3.6.11}
\]

It is obvious that

\[
|E_D(f)| \leq C_D |f|_{0, \infty, D} \leq C_D \|f\|_{m+1, \infty, D}, \quad D \subseteq I^n
\]

which means that $\|E_D\|_{m+1, \infty, D}$ is bounded. Since

\[
E_D(v) = 0, \quad \forall v \in P_m(D),
\]

we know from the Bramble-Hilbert lemma that

\[
|E_D(f)| \leq C_D \|E_D\|_{m+1, \infty, D} |f|_{m+1, \infty, D}, \tag{3.6.12}
\]

where $C_D$ is constant. Let $I^n = F_i(B_i)$ be affine mappings. Using equation (3.6.12) and a scaling
argument, we obtain that

\[ \left| \int_{I^n} f(y) \, dy - \sum_{i=1}^{N} Q_{n,m}^i f(y) \right| \]

\[ = \left| \sum_{i=1}^{N} \int_{B_i} f(y) \, dy - Q_{n,m}^i f(y) \right| \]

\[ = \left| \sum_{i=1}^{N} \operatorname{vol}(B_i) \int_{F_i(B_i)} f(y) \, dy - Q_{n,m}^i f(y) \right| \]

\[ \leq \sum_{i=1}^{N} \operatorname{vol}(B_i) |E_{F_i(B_i)}(f)| \]

\[ \leq C\|E_{I^n}\|_{m+1,\infty,I^n} \sum_{i=1}^{N} \operatorname{vol}(B_i) |f|_{m+1,\infty,F_i(B_i)} \]

\[ \leq C\|E_{I^n}\|_{m+1,\infty,I^n} h^{m+1} \sum_{i=1}^{N} \operatorname{vol}(B_i) |f|_{m+1,\infty,B_i} \]

\[ \leq C\|E_{I^n}\|_{m+1,\infty,I^n} h^{m+1} |f|_{m+1,\infty,I^n} \]

\[ \square \]

**Remark 3.6.11.** In the proof of theorem 3.6.10, we note that \( \|E_{I^n}\|_{m+1,\infty,I^n} \) is actually the error bound of the quadrature rule \( Q_{m,e} \) in \( I^n \). For example, we can replace \( \|E_{I^n}\|_{m+1,\infty,I^n} \) with the error bounds [112, 75] of sparse grids to obtain the error estimates of ME-PCM methods.

**Remark 3.6.12.** In theorem 3.6.10, we do not include the number of points into the error bound, which may be improved by the detailed study on the value of \( \|E_{I^n}\|_{m+1,\infty,I^n} \) for different quadrature rules. If the analyticity condition is available instead of \( \| \cdot \|_{m+1,\infty} \), we may obtain a better error bound with exponential convergence.
Chapter 4

An adaptive ME-gPC method for ODEs with random coefficients

In this chapter we present an adaptive ME-gPC method [102] for ODEs with random coefficients. We first present a heuristic adaptive criterion, which is based on the decay rate of the coefficients of the local polynomial chaos expansions. In the second section, we discuss the numerical implementation for adaptivity. In the third section we provide a detailed study on the Kraichnan-Orszag three-mode model [77], for which gPC fails to converge. We demonstrate that the adaptive ME-gPC method can deal with the low regularity in the parametric space by the dual paths of convergence, i.e., $hp$-convergence.

4.1 An adaptive criterion

We consider an ODE system

$$\frac{du}{dt} = L(t,u,Y)$$
$$u(0) = u(0,Y),$$

(4.1.1)
where $\mathbf{Y} \in \mathbb{R}^n$ is a random variable and $L$ is nonlinear function.

For simplicity, we assume that $\mathbf{u} \in \mathbb{R}^1$. In the ME-gPC method, the spectral expansion of $u(t, \mathbf{Y})$ is

$$
\hat{u}_k(t, \mathbf{Y}_k) = \sum_{i=0}^{N_p-1} \hat{u}_{k,i} \phi_i(\mathbf{Y}_k).
$$

(4.1.2)

The approximate global mean can be expressed as

$$
\bar{u} = \sum_{k=1}^{N} \hat{u}_{k,0} \operatorname{Pr}(I_{B_k} = 1).
$$

(4.1.3)

From the orthogonality of gPC we can obtain the local variance approximated by polynomial chaos with order $p$

$$
\sigma_{k,p}^2 = \sum_{i=1}^{N_p-1} \hat{u}_{k,i}^2 \mathbb{E}[\phi_i^2],
$$

(4.1.4)

and the approximate global variance

$$
\bar{\sigma}^2 = \sum_{k=1}^{N} \left[ \sigma_{k,p}^2 + (\hat{u}_{k,0} - \bar{u})^2 \right] \operatorname{Pr}(I_{B_k} = 1).
$$

(4.1.5)

Let $\gamma_k$ be the error of the term $\sigma_{k,p}^2 + (\hat{u}_{k,0} - \bar{u})^2$. We obtain the exact global variance as

$$
\sigma^2 = \bar{\sigma}^2 + \sum_{k=1}^{N} \gamma_k \operatorname{Pr}(I_{B_k} = 1).
$$

(4.1.6)

We define the local decay rate of relative error of the gPC approximation in each element as follows

$$
\eta_k = \frac{\sum_{i=N_p-1}^{N_p-1} \hat{u}_{k,i}^2 \mathbb{E}[\phi_i^2]}{\sigma_{k,p}^2}.
$$

(4.1.7)

For $h$-type refinement, we consider two factors: the decay rate of the relative error $\eta_k$ in each element and the factor $\operatorname{Pr}(I_{B_k} = 1)$. We will split a random element into two equal parts when the following condition is satisfied

$$
\eta_k^\alpha \operatorname{Pr}(I_{B_k} = 1) \geq \theta_1, \quad 0 < \alpha < 1,
$$

(4.1.8)
where $\alpha$ is a prescribed constant.

When the random elements become smaller, (i.e., $\Pr(I_{B_k} = 1)$ becomes smaller), the value of $\eta_k$ satisfying the criterion will be bigger. Thus, the criterion relaxes the restriction on the accuracy of the local variance for smaller elements since the error contribution of small random elements will be dictated by their size. From equation (4.1.6) we can see that to achieve a certain level of accuracy, say $\beta$, we need $\sum_{k=1}^{N} \gamma_k \Pr(I_{B_k} = 1)/\sigma^2 \sim O(\beta)$. However, it is difficult to estimate such a global error since it is related to both $h$-type convergence and $p$-type convergence. By noting the hierarchical structure of orthogonal polynomial chaos basis, we replace $\gamma_k/\sigma^2$ with $\epsilon_k$ and use $\eta_k \Pr(I_{B_k} = 1)$ as an indicator of the error contribution of each element in this work.

There are two reasons to use the power of $\eta_k$ with respect to $\alpha$ in the criterion:

1. The decomposition of random space would terminate when $\Pr(I_{B_k} = 1) \sim \theta_1$. From the criterion, we can see that $\eta_k$ must satisfy $\eta_k \geq (\theta_1/\Pr(I_{B_k} = 1))^{1/\alpha}$ to trigger the decomposition of random space. If $\Pr(I_{B_k} = 1) < \theta_1$, $\eta_k$ must be greater than 1 and increase quickly as $\Pr(I_{B_k} = 1)$ becomes smaller further by noting that both $\theta_1/\Pr(I_{B_k} = 1)$ and $1/\alpha$ are greater than 1. It is, in general, hard to reach such a large $\eta_k$ in practice, even for problems involving low regularity in the parametric space. Thus, $\theta_1$ acts as a limit of the size of random elements.

In this work, we usually set $\alpha$ to be $1/2$.

2. In elements with low regularity the largest error contribution is $\eta_k \Pr(I_{B_k} = 1) \sim O(\Pr(I_{B_k} = 1)) \sim O(\epsilon_k)$ because the relative error $\eta_k$ could be almost $O(1)$. For such a case, we have to keep the error contribution of $O(\epsilon_k)$ because it is the best that gPC can do; however, we can weaken the error contribution of random elements in the smooth region. Note that $\eta_k \Pr(I_{B_k} = 1) \sim O(\eta_k^{1-\alpha} \theta_1)$, where $\theta_1$ is weighted by $\eta_k^{1-\alpha}$. Thus, in random elements with good regularity the error contribution will be much smaller than $\theta_1$ since $\eta_k < 1$ in these elements. Finally, the total error contribution $\sum_{k=1}^{N} \eta_k \Pr(I_{B_k} = 1)$ would be $O(m \Pr(I_{B_k} = 1)) \sim O(m \theta_1)$, where $m$ is the number of random elements with $O(\epsilon_k)$ error contribution. So, $\eta_k^{1-\alpha}$ works as a filter and $\theta_1$ also acts as an accuracy threshold besides the aforementioned limit of element size.

Furthermore, we use another threshold parameter $\theta_2$ to choose the most sensitive random di-
We define the sensitivity of each random dimension as

\[ r_i = \frac{(\hat{u}_{i,p})^2 E[\phi_{i,p}^2]}{\sum_{j=N_{p-1}}^{N_p-1} \hat{u}_j^2 E[\phi_j^2]}, \quad i = 1, 2, \ldots, d \]  

(4.1.9)

where we drop the subscript \( k \) for clarity and the subscript \( *_{i,p} \) denotes the mode consisting only of random dimension \( \xi_i \) with polynomial order \( p \). Let us consider a \( n \)-dimensional mode \( \phi_{(j_1,j_2,\ldots,j_n)}(Y) = \phi_{j_1}(Y_1)\phi_{j_2}(Y_2)\cdots\phi_{j_n}(Y_n) \) with \( \sum_{i=1}^{n} j_i \leq p \), which is a tensor product of one-dimensional modes \( \phi_{j_m}(Y_m) \), \( m = 1, 2, \ldots, n \). The subscript \( *_{i,p} \) means that \( \phi_{i,p}(Y) = \phi_p(Y_i) \), \( i = 1, 2, \ldots, n \). All random dimensions which satisfy

\[ r_i \geq \theta_2 \cdot \max_{j=1,\ldots,d} r_j, \quad 0 < \theta_2 < 1, \quad i = 1, 2, \ldots, d \]  

(4.1.10)

will be split into two equal random elements in the next time step while all other random dimensions will remain unchanged. Hence, we can reduce the total element number while gaining efficiency.

Considering that \( h \)-type refinement is efficient in practice, we only present results given by \( h \)-type refinement in this work. For some cases, say problems related to low regularity, \( h \)-type refinement may be the most effective choice since \( p \)-type convergence may not be maintained anymore. This is, of course, not surprising given what we know for deterministic problems [58].

### 4.2 Numerical implementation

#### 4.2.1 Uniform distributions

When \( h \)-type refinement is needed, we have to map the random field from one mesh of elements to a new mesh of elements. Suppose that the gPC expansion of the current random field is

\[ \hat{u}(\hat{Y}) = \sum_{i=0}^{N_p-1} \hat{u}_i \phi_i(\hat{Y}), \]  

(4.2.1)
then we assume that the gPC expansion in the next level takes the following form

\[ \tilde{u}(\tilde{Y}) = \tilde{u}\left( g\left( \tilde{Y} \right) \right) = \sum_{i=0}^{N_p-1} \tilde{u}_i \phi_i(\tilde{Y}), \]  

(4.2.2)

where \( \tilde{Y} \in [-1, 1]^d \). To determine the \( N_p \) coefficients \( \tilde{u}_i \), we choose \( N_p \) points \( \tilde{Y}_i, i = 0, 1, \cdots, N_p - 1 \), which are the uniform grid points in \([-1, 1]^d\) and solve the following linear system

\[
\begin{bmatrix}
\phi_{0,0} & \phi_{1,0} & \cdots & \phi_{N_p-1,0} \\
\phi_{0,1} & \phi_{1,1} & \cdots & \phi_{N_p-1,1} \\
\vdots & \vdots & \vdots & \vdots \\
\phi_{0,N_p-1} & \phi_{1,N_p-1} & \cdots & \phi_{N_p-1,N_p-1}
\end{bmatrix}
\begin{bmatrix}
\tilde{u}_0 \\
\tilde{u}_1 \\
\vdots \\
\tilde{u}_{N_p-1}
\end{bmatrix}
=
\begin{bmatrix}
\sum_{i=0}^{N_p-1} \tilde{u}_i \phi_i \left( g^{-1}\left( \tilde{Y}_0 \right) \right) \\
\sum_{i=0}^{N_p-1} \tilde{u}_i \phi_i \left( g^{-1}\left( \tilde{Y}_1 \right) \right) \\
\vdots \\
\sum_{i=0}^{N_p-1} \tilde{u}_i \phi_i \left( g^{-1}\left( \tilde{Y}_{N_p-1} \right) \right)
\end{bmatrix}
\]  

(4.2.3)

where \( \phi_{ij} = \phi_i \left( \tilde{Y}_j \right) \). We rewrite the above equation in matrix form as

\[ A \tilde{u} = \tilde{u}. \]  

(4.2.4)

Due to the hierarchical structure of the basis, \( A^{-1} \) exists for any \( N_p \) distinct points in \([-1, 1]^d\).

When \( h \)-type refinement is implemented we divide the random space of a certain random dimension \( \tilde{Y}_i \) into two equal parts. For example, if \( \tilde{Y}_i \) corresponds to element \([a, b] \) in the original random space \([-1, 1]\), the elements \([a, \frac{a+b}{2}]\) and \([\frac{a+b}{2}, b]\) will be generated in the next level. However, due to the linearity of transformation, we do not have to perform such a map from the original random space, as we can just separate the random space of \( \tilde{Y}_i \), which is \([-1, 1]\), to \([-1, 0]\) and \([0, 1]\). Therefore, the matrix \( A \) will be the same for every \( h \)-type refinement, and we only need to compute \( A^{-1} \) once and store it for future use. When refinement is needed, we can obtain \( \tilde{u} \) easily by a matrix-vector multiplication

\[ \tilde{u} = A^{-1} \tilde{u}. \]  

(4.2.5)

For a relatively small polynomial order \( (p \leq 10) \), the mapping cost is small.
4.2.2 Arbitrary distributions

For arbitrary distributions, there are no invariants available. We need to do the following Galerkin projection whenever a refinement is needed. In other words, the coefficients in equation (4.2.2) can be written as

\[ \hat{u}_i = \int_{[-1,1]^n} \hat{u}(\bar{y}(\bar{y})) \tilde{\phi}_i(\bar{y}) d\bar{y}, \]

(4.2.6)

where we assume that the conditional random variable \( \bar{Y} \) is mapped to the cube \([-1,1]^n \) and \( \{\tilde{\phi}_i\} \) is the basis in the children elements which in general takes a different form from the basis \( \{\phi_i\} \) in the father element due to the decomposition of the PDF. Since the gPC basis is given by tensor products of one-dimensional ones, the above integration can be computed efficiently using the Gauss quadrature rule.

4.3 Low regularity in the parametric space: the Kraichnan-Orszag problem

The low regularity in the parametric space can severely weaken the effectiveness of gPC, resulting in a slow convergence or divergence in a short time. The typical problem is the Kraichnan-Orszag three-mode ODE model.

4.3.1 Why gPC fails

The Kraichnan-Orszag problem [77] is a nonlinear three-dimensional stochastic ODE system:

\[
\begin{align*}
\frac{dx_1}{dt} &= x_2x_3 \\
\frac{dx_2}{dt} &= x_1x_3 \\
\frac{dx_3}{dt} &= -2x_1x_2
\end{align*}
\]  

(4.3.1)
subject to stochastic initial conditions

\[ x_1(0) = x_1(0; \omega), \quad x_2(0) = x_2(0; \omega), \quad x_3(0) = x_3(0; \omega). \] (4.3.2)

We first check the deterministic solutions of equation (4.3.1). Given different initial conditions, deterministic solutions can be basically separated into four different groups \( g_i, \ i = 1, 2, 3, 4 \), which are shown in Fig 4.1. All these four groups of solutions are periodic. If the initial conditions are located on the planes \( x_1 = x_2 \) and \( x_1 = -x_2 \), the corresponding solutions would stay on these two planes forever due to two fixed points \((0, 0, \sqrt{2x_1^2(0) + x_2^2(0)})\) and \((0, 0, -\sqrt{2x_1^2(0) + x_2^2(0)})\). By considering the properties of elliptic functions \([61]\), we can obtain the analytic solutions of each group. Here we only give the analytic form of group \( g_1 \):

\[ x_1 = Pcn[q(t - t_0)], \quad x_2 = Qdn[q(t - t_0)], \quad x_3 = -Rsn[q(t - t_0)], \] (4.3.3)

where \( cn[\cdot], sn[\cdot] \) and \( dn[\cdot] \) are Jacobi's elliptic functions and \( P, Q, R, q \) and \( t_0 \) are constants to be determined. We now substitute equation (4.3.3) into equation (4.3.1) to obtain

\[ Pq = QR, \quad Qk^2q = PR, \quad Rq = 2PQ, \] (4.3.4)

where \( k \) is the modulus of elliptic functions. Since we have three initial conditions

\[
\begin{align*}
Pcn[q(t - t_0)] &= x_1(0; \omega), \\
Qdn[q(t - t_0)] &= x_2(0; \omega), \\
-Rsn[q(t - t_0)] &= x_3(0; \omega),
\end{align*}
\] (4.3.5)

we have six equations with six unknowns \( P, Q, R, k, q \) and \( t_0 \). Thus, we have obtained the exact general solution of the Kraichnan-Orszag problem.
We now consider the following initial conditions

\[ x_1(0) = c_0 + 0.01Y, \quad x_2(0) = 1.0, \quad x_3(0) = 1.0, \]  

(4.3.6)

where \( Y \) is a uniform random variable and \( c_0 \) is a constant. By solving equation (4.3.4) and (4.3.5), we can determine the unknowns as

\[ P^2 = f^2(Y) + \frac{1}{2}, \quad Q^2 = \frac{3}{2}, \quad R^2 = 2f^2(Y) + 1, \]

\[ p^2 = 3, \quad k^2 = \frac{2}{3}f^2(Y) + \frac{1}{3}, \quad t_0 = -dn^{-1}[\frac{1}{Q}/p], \]  

(4.3.7)

where \( f(Y) = c_0 + 0.01Y. \)

Next we examine the Fourier expansions of Jacobi’s functions:

\[
\begin{align*}
\text{sn}[u] &= \frac{2\pi}{kK} \left[ \frac{q^{1/2}/sin z}{1 - q} + \frac{q^{3/2}/sin 3z}{1 - q^3} + \frac{q^{5/2}/sin 5z}{1 - q^5} + \ldots \right], \\
\text{cn}[u] &= \frac{2\pi}{kK} \left[ \frac{q^{1/2}/cos z}{1 + q} + \frac{q^{3/2}/cos 3z}{1 + q^3} + \frac{q^{5/2}/cos 5z}{1 + q^5} + \ldots \right], \\
\text{dn}[u] &= \frac{\pi}{2K} + \frac{2\pi}{K} \left[ \frac{q_2 cos 2z}{1 + q^2} + \frac{q^2 cos 4z}{1 + q^4} + \frac{q^3 cos 6z}{1 + q^6} + \ldots \right],
\end{align*}
\]

(4.3.8)

where \( q = q(Y), \ K = K(Y), \) and \( z = z(Y,t). \) First, we can see that the frequency depends on the random variable \( Y. \) It is well known that this will reduce the effectiveness of gPC as the initial phase difference will be amplified very fast as time increases. In Fig. 4.2, we show how the period of \( x_1 \) change as \( x_1(0) \to 1. \) We can see that the period of \( x_1 \) will increase to infinity as \( x_1(0) \) goes to 1. Note here that if \( x_1(0) = 1, \) the initial point \((1,1,1)\) would be on the plane \( x_1 = x_2. \) Second, if \( q \) goes to 1, it is clear that we need more and more terms for the expansion of \( \text{sn}[u], \) which means that the order of polynomial chaos must increase correspondingly to resolve the solution.

From equations (4.3.6) and (4.3.7) we can see that if \( Y \) is uniform in \([-1,1], \) \( x_1 \) is uniform in \([c_0 - 0.01,c_0 + 0.01]\) and the range (non-uniform) of \( k(Y) \) is

\[
\left[ \sqrt{\frac{2}{3}(c_0 - 0.01)} + \frac{1}{3}, \sqrt{\frac{2}{3}(c_0 + 0.01)} + \frac{1}{3} \right].
\]
Let $k_r$ denote the upper bound of $k(Y)$. It is clear that if $c_0 \to 0.99$, $k_r \to 1$. By the properties of elliptic functions, we know that $q \to 1$ when $k \to 1$. Thus, for the same degree of perturbation gPC should work less efficiently when $c_0 \to 0.99$, because $k(Y)$ will be closer to 1. Now, we investigate four simple cases: $c_0 = 0.94, 0.96, 0.98, \text{and } 0.99$. For simplicity we only show the results for $x_1$, since the situation is similar for $x_2$ and $x_3$. In Fig. 4.3 we show how gPC fails when $c_0 \to 0.99$. It can be seen that in Fig. 4.3 (a)-(d) the valid range of polynomial chaos with order $p = 6$ becomes shorter as $c_0$ increases. If $c_0$ is strictly less than 0.99 corresponding to $q < 1$, increasing the polynomial order can efficiently improve the results of polynomial chaos. For the cases (a)-(c), the results of polynomial chaos with order $p = 20$ agree very well with the results of Monte Carlo with 100,000 realizations. However, if $c_0 = 0.99$ is included, the periods of stochastic solutions will change from a finite value to infinity and increasing the polynomial order hardly improves the results for this case. It is shown in (d) that the correct part of the variance given by polynomial chaos with order $p = 30$ is almost the same with that given by polynomial chaos with order $p = 6$. Therefore, it is at the bifurcation point where gPC fails to converge.

In general, if the initial random data does not intersect with the planes $x_1 = x_2$ and $x_1 = -x_2$, we can improve the results of polynomial chaos by increasing the polynomial order, otherwise, polynomial chaos will diverge even after a short time of integration.

![Figure 4.1](image-url): Deterministic solutions of the Kraichnan-Orszag problem subject to different initial conditions. Left: 3D phase space; Right: 2D projection on $x_1$-$x_2$ plane.
Figure 4.2: Kraichnan-Orszag problem: Several deterministic solutions of $x_1$ versus time corresponding to different initial conditions.

Figure 4.3: Comparison of variance obtained from gPC and Monte Carlo simulations. (a): $c_0 = 0.94$; (b): $c_0 = 0.96$; (c): $c_0 = 0.98$; (d): $c_0 = 0.99$. 

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
4.3.2 Random inputs with uniform distributions

One-dimensional random input: uniform distribution

Let us first study the low regularity of the Kraichnan-Orszag three-mode problem, which is introduced by one-dimensional random input. For computational convenience and clarity in the presentation we first perform the following transformation

\[
\begin{align*}
\frac{\sqrt{2}}{3} & \quad \frac{\sqrt{2}}{2} \\
\frac{\sqrt{2}}{2} & \quad \frac{\sqrt{2}}{2} \\
0 & \quad 0 \\
y_1 & = y_1 \\
y_2 & = y_2 + y_3 \\
y_3 & = y_3 \\
x_1 & = x_1 \\
x_2 & = x_2 \\
x_3 & = x_3
\end{align*}
\]

(4.3.9)

As a result, we will rotate the deterministic solutions by $\frac{\pi}{4}$ around to $x_3$ axis in the phase space.

Now the new system is

\[
\frac{dy_1}{dt} = y_1 y_3 \\
\frac{dy_2}{dt} = -y_2 y_3 \\
\frac{dy_3}{dt} = -y_1^2 + y_2^2
\]

(4.3.10)

subject to initial conditions

\[
y_1(0) = y_1(0; \omega), \quad y_2(0) = y_2(0; \omega), \quad y_3(0) = y_3(0; \omega).
\]

(4.3.11)

From now on, we will study this problem based on equation (4.3.10). Note that the regularity is low at the planes $y_1 = 0$ and $y_2 = 0$ after the transformation. Gaussian random variables are used as random inputs in [77]. Here, we use uniform random variables since the low regularity can be introduced similarly. Thus, we study the stochastic response subject to the following random input

\[
y_1(0; \omega) = 1, \quad y_2(0; \omega) = 0.1 Y(\omega), \quad y_3(0; \omega) = 0,
\]

(4.3.12)
where \( Y \sim U(-1,1) \). Since the random initial data \( y_2(0;\omega) \) can cross the plane \( y_2 = 0 \), we know from the aforementioned discussion that gPC will fail for this case.

In Fig. 4.4, we show the evolution of the variance of \( y_1 \) within the time interval \([0,30]\). For comparison we include the results given by gPC with polynomial order \( p = 30 \). It can be seen that comparing to the results given by Monte Carlo with 1,000,000 realizations, gPC with polynomial order \( p = 30 \) begins to lose accuracy at \( t \approx 8 \) and fails beyond this point while ME-gPC converges as \( \theta_1 \) decreases. In Table 4.1, we show the maximum normalized error of the variance of \( y_1, y_2 \) and \( y_3 \) at \( t = 30 \) given by ME-gPC and the corresponding number of random elements. It is seen that when the threshold parameter \( \theta_1 \) decreases, the accuracy becomes better and we can obtain almost \( O(\theta_1) \) error. As we mentioned before, the reason that errors are usually bigger than \( \theta_1 \) is due to the low regularity which can reduce the convergence of gPC. It can be seen that for the same polynomial order we need more random elements to get a better accuracy; on the other hand, for the same \( \theta_1 \) increasing the polynomial order can reduce the number of random elements.

In Fig. 4.5, we show four adaptive meshes. We can see that around the point \( Y = 0 \) in random space of \( Y \), where the regularity is low, the random elements are smallest, which means that the low regularity can be captured by small random elements. In Fig. 4.6, we show the errors of Monte Carlo and ME-gPC in terms of computational cost. The error is the \( L_\infty \) error of the variance of \( y_1 \) in the time interval \([8,30]\), where gPC fails. To implement gPC, we need to apply Galerkin projection onto the chaos basis, resulting in the ensemble average \( E[\phi_i \phi_j \phi_k] \) of three basis modes. Here, we count the operations of \( E[\phi_i \phi_j \phi_k] \) for ME-gPC in order to estimate its cost. For Monte Carlo, the number of realizations is employed in the cost evaluation. Let \( n \) denote the number of operations. If the data in Fig. 4.6 are approximated by a first-order polynomial in a least-squares sense, we can obtain accuracy proportional to \( n^{-0.49} \), \( n^{-2.26} \), \( n^{-2.99} \) and \( n^{-4.24} \), respectively, for Monte Carlo and ME-gPC with polynomial order \( p = 3, p = 4 \) and \( p = 5 \), respectively. The decay rate for Monte Carlo is about \( n^{-0.5} \) as expected. Comparing to Monte Carlo, the errors of ME-gPC show a much greater decay rate in terms of the cost. We can see that the speed-up increases for higher accuracy, which implies that ME-gPC is an efficient alternative to Monte Carlo for integration.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
where high-order accuracy is required. In Fig. 4.7, we show the error contribution of each random element. Here we compare two criteria with $c_0 = \frac{1}{2}$ and $c_0 = \frac{1}{4}$. It is seen that the shape of error distribution is like an isosceles triangle, i.e., a “Gibbs-like” behavior. On the apex of the triangle is the largest error contribution, where the regularity is low. The error contribution decreases quickly away from the region of low regularity, since $\eta_k \Pr(I_{B_k} = 1) \sim \eta_k^{1-c_0} \theta_1$ and $\eta_k$ is much smaller on the smooth part. Because gPC loses accuracy as time increases, the error contribution of each element will become larger with time and more random elements with relative errors of $O(1)$ would appear around the region of low regularity. For a smaller $c_0$, the error contribution near the region of low regularity decreases much faster.

![Figure 4.4: Evolution of the variance of $y_1$ for one-dimensional random input.](image)

Table 4.1: Maximum normalized errors of the variance of $y_1, y_2$ and $y_3$ at $t = 30$ with $c_0 = 1/2$. (The results given by ME-gPC with $\theta_1 = 10^{-7}$ and polynomial order $p = 5$ are used as exact solutions.)

<table>
<thead>
<tr>
<th>$\theta_1$</th>
<th>$N$</th>
<th>$\text{Error}$</th>
<th>$N$</th>
<th>$\text{Error}$</th>
<th>$N$</th>
<th>$\text{Error}$</th>
<th>$N$</th>
<th>$\text{Error}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$10^{-2}$</td>
<td>36</td>
<td>9.90e-2</td>
<td>74</td>
<td>3.24e-3</td>
<td>138</td>
<td>3.45e-4</td>
<td>286</td>
<td>2.31e-5</td>
</tr>
<tr>
<td>$10^{-3}$</td>
<td>28</td>
<td>7.24e-2</td>
<td>44</td>
<td>4.10e-3</td>
<td>78</td>
<td>2.90e-4</td>
<td>130</td>
<td>4.35e-6</td>
</tr>
<tr>
<td>$10^{-4}$</td>
<td>46</td>
<td>3.10e-2</td>
<td>106</td>
<td>2.32e-3</td>
<td>280</td>
<td>1.37e-4</td>
<td>820</td>
<td>2.87e-5</td>
</tr>
<tr>
<td>$10^{-5}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Figure 4.5: Adaptive meshes for the 1D random input with $c_0 = 1/2$. (a): $\theta_1 = 0.01$, $p = 3$; (b): $\theta_1 = 0.001$, $p = 3$; (c): $\theta_1 = 0.0001$, $p = 3$; (d): $\theta_1 = 0.0001$, $p = 5$. 

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Figure 4.6: Error versus cost of Monte Carlo simulations and ME-gPC with different polynomial orders (based on the $L_\infty$ error of the variance of $y_1$ in the time interval $[8, 30]$). Here we only count the average number of operations in one time step.

Two-dimensional random input: uniform distribution

In this section we use ME-gPC to study the Kraichnan-Orszag problem with two-dimensional random input

$$y_1(0; \omega) = 1, \quad y_2(0; \omega) = 0.1Y_1(\omega), \quad y_3(0; \omega) = Y_2(\omega),$$

(4.3.13)

where $Y_1$ and $Y_2$ are uniform random variables with unit standard deviation.

In Fig. 4.8, we show the evolution of the variance of $y_1$, $y_2$ and $y_3$ and an adaptive two-dimensional mesh. For comparison we include the result given by gPC with polynomial order $p = 10$. It can be seen that gPC with polynomial order $p = 10$ begins to diverge around $t \approx 4$ while ME-gPC with $p = 5$ Legendre-chaos shows good convergence to the results given by Monte Carlo with 1,000,000 realizations. From the final refined mesh, we can see that the results are more sensitive to $Y_1$, because $Y_1$ can cross the plane $y_2 = 0$ where the regularity is low. Note here that the region of low regularity is a line. In Fig. 4.9, we show the error of Monte Carlo and ME-gPC in terms of computational cost. Here we regard the results given by ME-gPC with $\theta_1 = 10^{-6}$ and $p = 5$ as exact solutions. From the empirical fit we obtain an accuracy proportional to $n^{-0.50}$, $n^{-1.72}$ and $n^{-2.56}$, respectively, for Monte Carlo and ME-gPC with $p = 3$ and $p = 5$. It is seen that ME-gPC is much
Figure 4.7: Error contribution of each random element given by two criteria with different $c_0$. $\theta_1 = 10^{-4}$ and $p = 5$. (a): $c_0 = 1/2, t = 50$; (b): $c_0 = 1/2, t = 100$; (c): $c_0 = 1/4, t = 50$; (d): $c_0 = 1/4, t = 100$. 

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
faster than Monte Carlo for higher accuracy. Comparing to the 1D case, however, the decay rate of relative error becomes smaller because both the random dimension and the region of low regularity become larger.

Figure 4.8: The Kraichnan-Orszag problem with 2D random inputs. $c_0 = 1/2, \theta_1 = 0.1, 0.01, 0.001$ and $\theta_2 = 0.1$. (a): $\sigma^2_{\gamma_1}$; (b): $\sigma^2_{\gamma_2}$; (c): $\sigma^2_{\gamma_3}$; (d): adaptive mesh for $\theta_1 = 0.001$ and $p = 5$.

Three-dimensional random input: uniform distribution

In this section we use ME-gPC to study the Kraichnan-Orszag problem with three-dimensional random input

$$ y_1(0) = Y_1(\omega), \quad y_2(0) = Y_2(\omega), \quad y_3(0) = Y_3(\omega), \quad (4.3.14) $$

where $Y_1, Y_2$ and $Y_3$ are uniform random variables with unit standard deviation.
Figure 4.9: Error versus cost of Monte Carlo simulations and ME-gPC with 2D Legendre-Chaos (based on the $L_{\infty}$ error of the variance of $y_1$ in the time interval [4, 10]). Here we only count the average number of operations in one time step.

In Fig. 4.10, we show the evolution of variance. Due to the symmetry of $y_1$ and $y_2$ in equation (4.3.10) and the symmetry of $y_1(0)$ and $y_2(0)$ in the random inputs, the variances of $y_1$ and $y_2$ are the same. Here we only show the results for $y_1$ and $y_3$. It can be seen that gPC diverges around $t \approx 1$ and fails subsequently while ME-gPC shows good convergence as before. For this case, the random space $[-1,1]^3$ of random inputs contains both $y_1 = 0$ and $y_2 = 0$ where the regularity is low. Comparing to the case with 2D random inputs, the region of low regularity is much larger. Thus, it is more difficult to resolve the 3D case. Based on the results given by ME-gPC with polynomial order 3 and $\theta_1 = 10^{-5}$, the $L_{\infty}$ errors of the variance of $y_1$ in the time interval [1.5, 6] are 0.16% and 0.21%, respectively, for Monte Carlo with 1,000,000 realizations and ME-gPC with polynomial order $p = 3$ and $\theta_1 = 10^{-3}$. Thus, these two errors are comparable. For this case, the speed-up of ME-gPC is much lower compared to the 2D problem. From the previous results, we know that this speed-up would increase for higher accuracy, but the increasing speed would be lower comparing to the 1D and 2D cases. In Fig. 4.11, we show the evolution of the random elements generated. It can be seen that to maintain the accuracy, the element number has to increase at a speed about 100 elements per time unit.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
In summary, ME-gPC shows good convergence when solving the Kraichnan-Orszag problem and it can achieve a desired accuracy at a cost much lower than Monte Carlo. However, ME-gPC loses efficiency for problems related to low regularity and high-dimensional random inputs, because the number of random elements has to increase fast to maintain a desired accuracy.

![Graph 1: Evolution of variance for the 3D Kraichnan-Orszag problem. \( \theta_2 = 10^{-1} \). Left: \( \sigma_y^2 = \sigma_{y_2}^2 \); Right: \( \sigma_y^2 \).](image1)

![Graph 2: Evolution of the element number for the 3D Kraichnan-Orszag problem. \( \theta_1 = 10^{-3} \).](image2)

### 4.3.3 Random inputs with arbitrary distributions

We now re-consider this problem with random inputs of Beta and Gaussian distributions. We study the following three different kinds of initial conditions:

(i) \( y_1(0) = 1, \quad y_2(0) = 0.1Y_1, \) and \( y_3(0) = 0; \)
(ii) \( y_1(0) = 1, \quad y_2(0) = 0.1Y_1, \) and \( y_3(0) = 0.1Y_2; \)

(iii) \( y_1(0) = cY_1, \quad y_2(0) = cY_2, \) and \( y_3(0) = cY_3, \)

where \( c \) is constant and \( Y_i \) are random variables of Beta or Gaussian distribution.

For all these three cases, the relative error of recurrence coefficients \( \alpha_i \) and \( \beta_i \) is set to be \( 10^{-12} \) when orthogonal local gPC basis is constructed numerically. For case (i) we show the convergence of adaptive ME-gPC in table 4.2 and 4.3, where \( Y_1 \) is of Beta distribution and Gaussian distribution, respectively. It can be seen that ME-gPC converges as \( \theta_1 \) decreases. For all the cases in table 4.2 and 4.3, we have recorded the time used for the construction of orthogonal polynomial chaos, which is less than 0.15% of the time used by the ME-gPC solver. Thus, the cost of the polynomial construction can be ignored compared with the total cost. In Fig. 4.12, we show the adaptive meshes for Beta distribution and Gaussian distribution. Due to the symmetry of distribution, the mesh for Gaussian distribution is symmetric in contrast to the asymmetric one for Beta distribution with \( \alpha = 1 \) and \( \beta = 4 \). From the adaptive criterion, we know that the element size is controlled by two factors: the relative error \( \eta_k \) and \( \Pr(I_{B_k} = 1) \). We can see that around the region \( (Y_1 = 0) \) of low regularity the mesh is refined because the error of gPC goes to \( O(1) \), and the mesh is coarser where \( \Pr(I_{B_k} = 1) \) is smaller. For the Gaussian distribution we only refine the middle element \([-6, 6]\) as before. In Fig. 4.13 we show the speedup of ME-gPC for case (i) comparing to the Monte Carlo (MC) method; \( n \) is defined as the number of operations. If the data in Fig. 4.13 are approximated by a first-order polynomial in a least-squares sense, the accuracy of ME-gPC can be obtained as \( O(n^{-3.23}), O(n^{-3.85}) \) and \( O(n^{-5.55}) \) corresponding to \( p = 3, p = 4 \) and \( p = 5 \), respectively, for Beta distribution and \( O(n^{-3.07}), O(n^{-5.16}) \) and \( O(n^{-6.33}) \) for Gaussian distribution. It can be seen that the adaptive ME-gPC converges much faster than the Monte Carlo method and the speedup increases with the polynomial order. However, since the error of gPC increases with time, the above speedups will decrease with time. The long-term behavior of gPC and ME-gPC was studied in [105].

In Fig. 4.14, we show the adaptive meshes of case (ii) for Beta and Gaussian distributions. Since the regularity is low at the line \( Y_1 = 0 \) in the random space for this case, it can be seen that the meshes are well refined along the line \( Y_1 = 0 \). It is not surprising that the random elements are

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
bigger where the PDF is smaller because $\Pr(I_{B_k} = 1)$ is relatively smaller in these elements. In Fig. 4.15, the speedup for case (ii) is shown. We can see that the Monte Carlo method is competitive for low accuracy and ME-gPC can achieve a good speedup for high accuracy.

In Fig. 4.16, the evolution of $y_1$ in case (iii) is shown. We take $c = 1$ for the Beta distribution and 0.3 for the Gaussian distribution. For the purpose of comparison, we include the results of gPC. For both cases, gPC with $p = 3$ begins to fail at $t \approx 1$. Since increasing polynomial order does not improve gPC [102], the results of gPC with a higher order are not shown. ME-gPC can maintain convergence by increasing random elements adaptively. However, for a comparable accuracy of $O(10^{-3})$, the Monte Carlo method is about twice faster than ME-gPC for the Beta distribution and about four times faster than ME-gPC for the Gaussian distribution. This is mainly due to two factors. First, the region of low regularity for case (iii) is large because it consists of two planes: $Y_1 = 0$ and $Y_2 = 0$. Second, the cost of gPC increases very quickly with the random dimension. Since ME-gPC is a dimension dependent method, its efficiency decreases as the random dimension increases. The Monte Carlo method is still a better choice to obtain a moderate accuracy for high-dimensional cases unless a high accuracy is really required.

Table 4.2: Maximum normalized errors of the variance of $y_1$, $y_2$ and $y_3$ at $t = 20$ for case (i) of the K-O problem. $\alpha = 1/2$ and $Y_1$ is of beta distribution with $\alpha = 1$ and $\beta = 4$. (The results given by ME-gPC with $\theta_1 = 10^{-7}$ and polynomial order $p = 7$ are used as exact solutions.)

<table>
<thead>
<tr>
<th>$\theta_1$</th>
<th>$N$ Error</th>
<th>$N$ Error</th>
<th>$N$ Error</th>
<th>$N$ Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p = 3$</td>
<td>24</td>
<td>2.98e-2</td>
<td>77</td>
<td>1.85e-3</td>
</tr>
<tr>
<td>$p = 4$</td>
<td>21</td>
<td>1.54e-2</td>
<td>45</td>
<td>1.05e-3</td>
</tr>
<tr>
<td>$p = 5$</td>
<td>16</td>
<td>6.25e-2</td>
<td>34</td>
<td>3.30e-3</td>
</tr>
</tbody>
</table>

Table 4.3: Maximum normalized errors of the variance of $y_1$, $y_2$ and $y_3$ at $t = 20$ for case (i) of the K-O problem. $\alpha = 1/2$ and $Y_1$ is of normal distribution. (The results given by ME-gPC with $\theta_1 = 10^{-7}$ and polynomial order $p = 7$ are used as exact solutions.)

<table>
<thead>
<tr>
<th>$\theta_1$</th>
<th>$N$ Error</th>
<th>$N$ Error</th>
<th>$N$ Error</th>
<th>$N$ Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p = 3$</td>
<td>38</td>
<td>2.79e-2</td>
<td>98</td>
<td>2.21e-3</td>
</tr>
<tr>
<td>$p = 4$</td>
<td>32</td>
<td>4.02e-2</td>
<td>62</td>
<td>9.07e-3</td>
</tr>
<tr>
<td>$p = 5$</td>
<td>28</td>
<td>4.63e-2</td>
<td>42</td>
<td>3.28e-3</td>
</tr>
</tbody>
</table>
Figure 4.12: Adaptive mesh and corresponding random distribution for the K-O problem with 1D random inputs. $p = 5$ and $\theta_1 = 10^{-3}$. Left: Beta distribution with $\alpha = 1$ and $\beta = 4$; Right: Gaussian distribution.

Figure 4.13: Speedup for the K-O problem with 1D random inputs at $t = 20$. Left: Beta distribution with $\alpha = 1$ and $\beta = 4$; Right: Gaussian distribution.

Figure 4.14: Adaptive meshes for case (ii) of the K-O problem. $p = 4$, $\theta_1 = 10^{-4}$ and $\theta_2 = 10^{-2}$. Left: Beta distribution with $\alpha = 1$ and $\beta = 4$; Right: Gaussian distribution.
Figure 4.15: Speedup for case (ii) of the K-O problem at $t = 10$. Left: Beta distribution with $\alpha = 1$ and $\beta = 4$; Right: Gaussian distribution.

Figure 4.16: Evolution of the variance of $y_1$ for case (iii) of K-O problem. Left: Beta distribution with $\alpha = 1$ and $\beta = 4$. $c = 1$; Right: Gaussian distribution. $c = 0.3$. 

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Chapter 5

An adaptive ME-gPC method for elliptic problems with random coefficients

In this chapter we propose an adaptive ME-gPC method [109] for elliptic equations with random coefficients of a moderate number ($O(10)$) of random dimensions. We employ low-order ($p \leq 3$) polynomial chaos and refine the solution using adaptivity in the random space. The model problem is set up in the first section. In the second section, we develop a posteriori error estimators using the solutions of the error equations in a larger approximation space. To reduce to the cost of solving the error equations, we construct a reduced space, where a much smaller number of terms in the enhanced polynomial chaos space are used to capture the errors of the local gPC approximation. The underestimated information will be recovered by modeling the variation in the coefficients of polynomial chaos expansions. In the third section, we propose a $h$-type adaptive ME-gPC algorithm based on local and global a posteriori error estimators. In the fourth section, we present some numerical results.
5.1 Model problem

Let \( D \) be a bounded, connected, open subset of \( \mathbb{R}^d (d = 1, 2, 3) \) with a Lipchitz continuous boundary \( \partial D \). We consider the following stochastic linear boundary value problem: find a stochastic function, \( u : \Omega \times \bar{D} \to \mathbb{R} \), such that almost surely (a.s.) the following equation holds:

\[
- \nabla \cdot (a(x; \omega) \nabla u(x; \omega)) = f(x) \quad \text{on } D,
\]
\[
u(x; \omega) = 0 \quad \text{on } \partial D,
\]
(5.1.1)

where \( a(x; \omega) \) is a second-order random processes satisfying the assumption

**Assumption 5.1.1.** Let \( a(x; \omega) \in L_\infty(D; \Omega) \) be strictly positive with lower and upper bounds \( a_{\text{min}} \) and \( a_{\text{max}} \), respectively,

\[
0 < a_{\text{min}} < a_{\text{max}} \text{ and } \Pr(a(x; \omega) \in [a_{\text{min}}, a_{\text{max}}], \forall x \in \bar{D}) = 1. \quad (5.1.2)
\]

For each random element \( B_k \), an independent problem of the same type is defined as

\[
- \nabla \cdot (a(x; y_k) \nabla u(x; y_k)) = f(x) \quad \text{on } D,
\]
\[
u(x; y_k) = 0 \quad \text{on } \partial D.
\]
(5.1.3)

All previous assumptions on \( a(x; \cdot) \) are satisfied with respect to the condition PDF \( f_k(y_k|I_{B_k} = 1) \).

We note that such a system is complete.

We define a bilinear form \( \mathcal{B}_k(\cdot, \cdot) \) as

\[
\mathcal{B}_k(u, v) = \mathbb{E}_k\left[ \int_D a \nabla u \cdot \nabla v \, dx \right]
\]
(5.1.4)

and a linear form \( \mathcal{L}_k(\cdot) \) as

\[
\mathcal{L}_k(u) = \mathbb{E}_k\left[ \int_D f u \, dx \right].
\]
(5.1.5)
We consider the following Hilbert space

\[ W_k = \left\{ (x; y_k(\omega)) \| v \|_{W_k}^2 = \mathbb{E}_k \left[ \int_D \nabla v \cdot \nabla v \, dx \right] < \infty, \quad \forall v \in H_0^1(D) \otimes L_2(I_{B_k}^{-1}(1)) \right\} \tag{5.1.6} \]

with an inner product \( (u, v)_{W_k} = \mathbb{E}[\int_D \nabla u \cdot \nabla v \, dx] \). For the entire random space, a similar Hilbert space \( W \) can be defined with an inner product \( (u, v)_W = \mathbb{E}[\int_D \nabla u \cdot \nabla v \, dx] \).

Due to the assumption 5.1.1, the following conditions are naturally satisfied

\[ \mathcal{B}_k(u, v) = a_{\max} \| u \|_{W_k} \| v \|_{W_k}, \quad \forall u, v \in W_k \text{ (Continuity);} \tag{5.1.7a} \]
\[ a_{\min} \| v \|_{W_k} \leq \mathcal{B}_k(v, v), \quad \forall v \in W_k \text{ (Coercivity).} \tag{5.1.7b} \]

Thus, we can claim by the Lax-Milgram theorem that an unique solution exists for the following weak form

\[ \mathcal{B}_k(u, v) = \mathcal{L}_k(v), \quad \forall v \in W_k. \tag{5.1.8} \]

Using the estimates of the approximation error and the Céa's lemma from the finite element theory, it is not difficult to prove the convergence of ME-gPC for the model problem.

### 5.2 An a-posteriori error estimator

In the deterministic finite element method (FEM), the local error estimator can be obtained either explicitly from the errors of a proper interpolation [101], or implicitly from a local problem [8]. In ME-gPC we use similar ideas to obtain a local error estimator. We focus on how to estimate the error of gPC efficiently with a relative low cost.

For simplicity we drop the subscript \( k \), since the following discussions are valid for every ME-gPC element. In practice, the space \( W \) needs to be truncated for an approximation of \( u(x; y) \). To focus on the error of polynomial chaos, we use the following assumption:

**Assumption 5.2.1.** Compared to the error introduced by the truncated polynomial chaos, the error
given by spatial discretization can be neglected.

Let \( \|u\|_E = \mathbb{E} \int_D a|\nabla u|^2 \) denote the energy norm.

### 5.2.1 Local error \( e \)

Since the dimension of \( W \) is infinite, we need to truncate it to a certain level for an approximate solution of equation (5.1.8). Let

\[
Q^p = \text{span}\{ \phi_\alpha(y), \ |\alpha| \leq p \}
\]
denote the truncated polynomial chaos space up to polynomial order \( p \). We define a truncated version of \( W \) as \( W^p = \{ v(x; y) | v \in H^1_0(D) \otimes Q^p \} \). Due to the assumption 5.2.1, we here keep \( H^1_0(D) \) unchanged although in numerical computation it will also be truncated. Let \( u^p \) satisfy the following equation

\[
\mathcal{B}(u^p, v) = \mathcal{L}(v), \quad \forall v \in W^p.
\]  

(5.2.1)

We define the approximation error as

\[
e = \| u - u^p \|_E.
\]  

(5.2.2)

Since \( Q^p \) is an hierarchic orthogonal space we introduce the following saturation assumption:

**Assumption 5.2.2.** There exists a constant \( \beta \in [0, 1) \) such that

\[
\| u - u^{p+q} \|_E \leq \beta \| u - u^p \|_E, \quad \forall q \in \mathbb{N}.
\]

Let \( e^* = u^{p+q} - u^p \), which satisfies that

\[
\mathcal{B}(e^*, v) = \mathcal{L}(v) - \mathcal{B}(u^p, v), \quad \forall v \in W^{p+q}.
\]  

(5.2.3)

Under the assumption 5.2.2, the true error \( e \) in the energy norm can be bounded from both ends by
However, it will be more expensive to solve equation (5.2.3) than the original problem. Thus, we decompose $e^*$ as

$$e^* = e_1 + e_2, \quad (5.2.5)$$

where $e_1 \in W^p$ and $e_2 \in Y_{p,q} := W^{p+q} - W^p$ satisfy

$$B(e_1, v) + B(e_2, v) = L(v) - B(u^p, v), \quad \forall v \in W^p \quad (5.2.6a)$$

$$B(e_1, w) + B(e_2, w) = L(w) - B(u^p, w), \quad \forall w \in Y_{p,q}. \quad (5.2.6b)$$

We note that $e_1$ and $e_2$ are coupled together. To obtain a reasonable approximation $\hat{e}$ to $e^*$, the coupling terms can be ignored such that

$$B(e_1, v) = 0, \quad \forall v \in W^p \quad (5.2.7a)$$

$$B(e_2, w) = L(w) - B(u^p, w), \quad \forall w \in Y_{p,q}, \quad (5.2.7b)$$

where $\hat{e} = \hat{e}_1 + \hat{e}_2 = \hat{e}_2$. By assuming that a strengthened Cauchy-Schwarz inequality holds, the error estimator $e^*$ in energy norm can be bounded from both ends as [8]

$$\|e\|_E \leq \|e^*\|_E \leq \frac{1}{\sqrt{1 - \gamma^3}} \|\hat{e}\|_E, \quad (5.2.8)$$

where $\gamma \in [0, 1)$ is a constant.

Based on the above discussion, we know that $\|\hat{e}\|_E$ can be used as a local error estimator. Due to the equivalence between the norm $\|\cdot\|_W$ and $\|\cdot\|_E$ we use $\eta = \|e\|_W \approx \|\hat{e}\|_W = \|\hat{e}\|_{L_2(B,H^1_0(D))}$ as a local error estimator.

The cost of obtaining the error estimator $\eta \approx \|\hat{e}\|_{L_2(B,H^1_0(D))}$ is determined by the number
Table 5.1: Typical values of \( \dim(Y_{p,q}) \) with respect to the number \( n \) of random dimensions. \( p = 2 \) and \( q = 1 \). The numbers in parenthesis denote the dimensions of the space \( W^p \).

\[
\begin{array}{|c|c|c|c|c|c|}
\hline
n & 2 & 4 & 6 & 8 & 10 \\
\hline
\dim(Y_{p,q}) & 4(6) & 20(15) & 56(28) & 120(45) & 220(66) \\
\hline
\end{array}
\]

\( \dim(Y_{p,q}) \). We set \( q = 1 \). In table 5.1 we show some typical values of \( \dim(Y_{p,q}) \) for \( p = 2 \), where we also give the dimensions of the solution space \( W^p \) in parenthesis for comparison. It is seen that \( \dim(Y_{p,q}) \) can be much larger than the dimension of \( W^p \), which implies that the cost to obtain \( \eta \) will be much more expensive than that to solve the original problem. Clearly, to make the numerical method practical we need to reduce the cost of obtaining the local error estimators.

### 5.2.2 Stochastic regularity

#### Taylor expansion

We first consider the Taylor expansion of \( u(\cdot, y) \) to examine the behavior of error contribution of each term \( y^\alpha \). Let

\[
\rho_i = \sqrt{\lambda_i} \| h_i \|_{L^\infty(D)}. \tag{5.2.9}
\]

In [100], the following proposition about the regularity of \( u(\cdot, y) \) is given.

**Proposition 5.2.3.** If \( u(\cdot, y) \) is a solution of the model problem, then

\[
\| \partial^\alpha_y u(\cdot, y) \|_{H^2(D)} \leq c_{\alpha,|\alpha|} |\rho|^{|\alpha|} \| u(\cdot, y) \|_{H^1(D)}, \quad \forall \alpha \in \mathbb{N}_0^n, \tag{5.2.10}
\]

where \( |\rho| = \prod_{i=1}^n \rho_i^{a_i} \) and \( c_{\alpha,|\alpha|} \) is a constant depending on the random process \( a(x; \omega) \) and \( |\alpha| \).

We now examine the Taylor expansion of \( u(\cdot, y) \) around \( y = 0 \), which takes the form

\[
u(\cdot, y) = \sum_{|\alpha| \leq p} \frac{\partial^\alpha_y u(\cdot, 0)}{\alpha!} y^\alpha + \sum_{|\alpha| > p} \frac{\partial^\alpha_y u(\cdot, 0)}{\alpha!} y^\alpha, \tag{5.2.11}\]

where \( y^\alpha = \prod_{i=1}^n y_i^{a_i} \). In particular, we consider the terms satisfying \( |\alpha| = p + 1 \), which contribute
most to the error estimate. Using Proposition 5.2.3, the terms with $|\alpha| = p + 1$ can be written as

$$
\| \sum_{\alpha=p+1} \frac{\partial^\alpha u(\cdot, 0)}{\alpha!} y^\alpha \|_{L^\infty(B, H^1_0(D))} \approx \sum_{|\alpha|=p+1} c_{\alpha, \alpha} T(\alpha) \rho^\alpha, \tag{5.2.12}
$$

where $T(\alpha) = |\alpha|!/\alpha!$.

**Spectral expansion**

We consider the spectral expansion of $u(\cdot, y)$ as

$$
u = \sum_{|\alpha|=0}^\infty u_\alpha \phi_\alpha(y),
$$

where $\{\phi_\alpha\}$ is the normalized orthogonal basis. By comparing the spectral expansion and the Taylor expansion, we obtain

$$
u_\alpha = \sum_{|\alpha|=0}^\infty \frac{\partial^\alpha u(\cdot, 0)}{\alpha!} E[y^\alpha \phi_\alpha(y)]. \tag{5.2.13}
$$

By noting the fact that

$$y^\alpha = \sum_{|\beta|=0}^{|\alpha|} a_\beta \phi_\beta(y),
$$

where the coefficients $a_\beta$ can be uniquely determined, and the orthogonality of $\{\phi_\alpha\}$, we obtain

$$
u_\alpha = \sum_{\beta \geq \alpha} \frac{\partial^\beta u(\cdot, 0)}{\beta!} E[y^\beta \phi_\alpha(y)]. \tag{5.2.14}
$$

For the error estimate we are interested in the terms with polynomial order $|\alpha| \geq p + 1$. Specifically we consider the dominant terms

$$
\| u - w^p \|_{L^2(B, H^1_0(D))} \approx \sum_{|\alpha|=p+1} \int_D (\nabla u_\alpha)^2 dx.
$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
where we assume that the polynomial chaos basis \( \{ \phi_\alpha \} \) has been normalized. If in the approximation of \( u_\alpha \) the term satisfying \( \beta = \alpha \) is dominant, we can rewrite the error approximation as

\[
\| u - u^p \|_{L^2(B, H^1(D))}^2 \approx \sum_{|\alpha| = p+1} c_{a,\alpha}^2 T^2(\alpha) \mathbb{E}[y^\alpha \phi_\alpha]^2 \rho^{2\alpha},
\]

where we separate the factor \( \rho^\alpha \) from \( \| \partial_y^\alpha u(-, 0) \|_{H^1(D)} \) motivated by the Proposition 5.2.3, and \( c_{a,\alpha} \) is constant satisfying

\[
c_{a,\alpha} = \frac{\| \partial_y^\alpha u(-, 0) \|_{H^1(D)}}{|\alpha|! \rho^\alpha}.
\]

Based on equation (5.2.15) we aim to reduce the cost to obtain the a posteriori error estimate \( \eta \).

### 5.2.3 Estimation of \( \eta \)

**Patterns of \( \alpha \)**

We first classify the patterns of \( \alpha \) for a given \(|\alpha| = p + 1\).

**Definition 5.2.4.** For \( \alpha \in \mathbb{N}_0^n \), the pattern of \( \alpha \) is defined as

\[
s_\alpha = [\alpha_{i_1}, \alpha_{i_2}, \ldots, \alpha_{i_n}], \quad \alpha_{i_j} \geq \alpha_{i_k}, \quad \forall j < k,
\]

where \( i_j = 1, 2, \ldots, n \). We let \( S_{|\alpha|} \) indicate the set \( \{ s_\alpha : |\alpha| \text{ is a constant} \} \), \( j = 1, 2, \ldots, n_S \), where \( n_S \) is the cardinality of \( S_{|\alpha|} \).

The cardinality \( n_S \) of \( S_{|\alpha|} \) is a function of \( n \) and \(|\alpha|\). For example, let \(|\alpha| = 3\) and \( n \geq 3 \). \( s_\alpha \) has at most the following three possibilities

\[
s_\alpha = [3, 0, 0, \ldots, 0], [2, 1, 0, \ldots, 0], [1, 1, 1, \ldots, 0].
\]
For $|\alpha| = 2, 3, 4, 5$, it can be easily verified that

\[
n_S = \begin{cases} 
2, 2, 3, & \text{if } n = 2 \\
2, 3, 4, 5, & \text{if } n \geq 3 
\end{cases} \tag{5.2.17}
\]

In other words, $n_S \leq |\alpha|$ for $|\alpha| \leq 5$.

Based on $s_{|\alpha|}$ we group the index $\alpha$ satisfying $|\alpha| = p + 1$. Let $Z_{|\alpha| = p+1} = \{\alpha| |\alpha| = p+1\}$. We define

\[
Z_{|\alpha| = p+1}^i = \{\alpha| s_\alpha = s_\alpha^i, s_\alpha^i \in S_{|\alpha| = p+1}\}, \quad i = 1, \ldots, n_S. \tag{5.2.18}
\]

Thus

\[
Z_{|\alpha| = p+1} = \bigcup_{i=1}^{n_S} Z_{|\alpha| = p+1}^i,
\]

where $Z_{|\alpha| = p+1}^i \cap Z_{|\alpha| = p+1}^j = \emptyset, \forall i \neq j$.

**A reduced space $V^p$**

Let $q = 1$. We now take a close look at the subspace

\[
Y_{p,1} = \{\phi_\alpha(y) | \phi_\alpha = \prod_{i=1}^{n} \phi_{a_i}(y_i), |\alpha| = p+1\}. \tag{5.2.19}
\]

We divide the space $Y_{p,1}$ according to the patterns of $\alpha$ as

\[
Y_{p,1} = \bigcup_{i=1}^{n_S} A_i, \quad A_i = \{\phi_\alpha| \alpha \in Z_{|\alpha| = p+1}^i\}. \tag{5.2.20}
\]

For each $A_i$, we define a subset $A_i' \subset A_i$

\[
A_i' = \{\phi_\alpha| \text{the corresponding } c_{a,\alpha}E[y^a \phi_\alpha] p^{2\alpha} \text{ are the largest } \theta_{n_A} \text{ ones for } \phi_\alpha \in A_i\},
\]
where $0 < \theta_s < 1$ is a prescribed constant and $n_{A_i}$ is the cardinality of $A_i$. We now construct the reduced space $V^p$ as

$$V^p = \bigcup_{i=1}^{ng} A'_i.$$  \hspace{1cm} (5.2.21)

It is obvious that we keep the most important modes $\phi_\alpha$ for each $s^i_\alpha \in S_{|\alpha|=p+1}$ based on the values $c_{a,\alpha}E[y^\alpha \phi_\alpha]^2 \rho^{2\alpha}$, and

$$\frac{\dim(V^p)}{\dim(Y_{p,1})} = \theta_s < 1.$$  

Instead of using the space $Y_{p,1}$, we want to estimate $\eta$ on the reduced space $V^p$. We note that the interaction between $V^p$ and the rest modes in $Y_{p,1}$ is ignored. Let $\tilde{\eta}$ indicate the error estimate given by $V^p$. Since the number of modes in $V^p$ can be significantly reduced compared to $Y_{p,1}$, the error $\eta$ will be underestimated if the eigenvalues $\rho_i$ decrease slowly. The simplest way to improve $\tilde{\eta}$ is to obtain a constant $\tilde{c}$ such that

$$\eta \sim \tilde{c}\tilde{\eta}.$$  

Estimation of $\tilde{c}$

Both values of $\eta$ and $\tilde{\eta}$ can be divided into $n_S$ parts according to the patterns $s^i_\alpha$ in $S_{|\alpha|=p+1}$, and $\eta$ can be rewritten as

$$\eta^2 = \sum_{i=1}^{n_S} \eta^2_{s^i_\alpha} \sim \sum_{i=1}^{n_S} \sum_{\phi_\alpha \in A_i} c_{a,\alpha}E[y^\alpha \phi_\alpha]^2 \rho^{2\alpha}. \hspace{1cm} (5.2.22)$$

Correspondingly, $\tilde{\eta}$ takes the form

$$\tilde{\eta}^2 = \sum_{i=1}^{n_S} \tilde{\eta}^2_{s^i_\alpha}. \hspace{1cm} (5.2.23)$$

Thus, using $\rho$ we can estimate the following ratios

$$\frac{\eta^2_{s^i_\alpha}}{\tilde{\eta}^2_{s^i_\alpha}} = \frac{\sum_{\phi_\alpha \in A_i} c_{a,\alpha}E[y^\alpha \phi_\alpha]^2 \rho^{2\alpha}}{\sum_{\phi_\alpha \in A'_i} c_{a,\alpha}E[y^\alpha \phi_\alpha]^2 \rho^{2\alpha}}. \hspace{1cm} (5.2.24)$$
We now take a look at the factor $T(\alpha)$ for each $s_\alpha$. We note that $T(\alpha) = |\alpha|! / \alpha!$, which depends on $s_\alpha$ only. If $T(\alpha)$ is the same for all $s_\alpha$, it is easy to verify that

$$\min \frac{\eta_{s_\alpha}^2}{\eta^{2}} \leq \frac{\eta^2}{\eta^{2}} \leq \max \frac{\eta_{s_\alpha}^2}{\eta^{2}},$$

which implies that we can choose $\tilde{c}$ as

$$\tilde{c}^2 = \max \frac{\eta_{s_\alpha}^2}{\eta^{2}}.$$  

We note that the terms $E[y^\alpha \phi_\alpha]$ are the same for a particular pattern $s_\alpha$ if the components of $y$ are i.i.d random variables. In ME-gPC for arbitrary probability measures the original PDFs of $y_i$ are decomposed simultaneously with the parametric space. $E[y^\alpha \phi_\alpha]$ will change correspondingly and can be regarded as a measure of such a change.

### 5.3 Algorithm of $h$-type adaptive ME-gPC

We define a global error as

$$\eta_g = \| u - u^p \|_{L_2(B, H^1_{\mu}(D))}$$

and a local error in element $k$ as

$$\eta_k = \| u - u^p \|_{L_2(B_k, H^1_{\mu}(D))}.$$  

We have the following lemma.

**Lemma 5.3.1.** The relation between $\eta_g$ and $\eta_k$ is

$$\eta_g^2 = \sum_{k=1}^N \eta_k^2 \Pr(I_{B_k=1}).$$  

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Proof.

\[ \eta_g^2 = \int_B \int_D (\nabla (u - u_p))^2 f(y) dy dx \]

\[ = \int_B \int_D (\nabla (u - \sum_{k=1}^{N} u_k(y_k) I_{B_k}))^2 f(y) dy dx \]

\[ = \sum_{k=1}^{N} \Pr(I_{B_k} = 1) \int_{B_k} \int_D (\nabla (u - u_k))^2 f_k(y_k | I_{B_k} = 1) dy dx_k \]

\[ = \sum_{k=1}^{N} \Pr(I_{B_k} = 1) \eta_k^2. \]

\[ \Box \]

5.3.1 Convergence rate of \( \eta_g \)

If the importance (degree of perturbation) of each random dimension is the same, we expect an \( h \)-type convergence rate

\[ \eta_g \sim O \left( N^{-(p+1)/n} \right) \]

where \( N \) is the total number of random elements [14, 100, 106]. We note that the algebraic index decreases by the factor \( \frac{1}{n} \). Under an analyticity assumption - the eigenvalues of the K-L expansion of \( a(x; y) \) decrease exponentially, the following convergence rate was proved in [100]

\[ \eta_g \sim O(N_{ace}^{-p+o(1)}), \quad n \to \infty \quad (5.3.4) \]

for sparse wavelet polynomial chaos bases, where \( N_{ace} \) is the number of deterministic PDEs to be solved. We note here that \( N_{ace}/N \) is constant since we use the same polynomial order in each random element. In this work, we relax the optimal assumption and implement adaptivity based on the a posteriori error estimator. In practice, the optimal convergence rate in equation (5.3.4) is difficult to maintain due to the low regularity of covariance kernel, e.g., the exponential kernel. The convergence rate can be weakened strongly by the dimensionality. Thus, we consider a moderate number (\( O(10) \)) of random dimensions in this work.
5.3.2 Criteria of adaptivity

The relation between $\eta_g$ and $\eta_h$ provides a natural strategy for the $h$-type adaptivity of ME-gPC. We define another local error indicator as

$$\hat{\eta}_k = \eta_k \Pr(I_{B_k} = 1)^{1/2} \approx \hat{\epsilon}_k \hat{\eta}_k \Pr(I_{B_k} = 1)^{1/2},$$

(5.3.5)

where $\hat{\epsilon}_k$ is the correction factor of the error estimator $\hat{\eta}_k$ from the reduced space $V^p_k$ in element $k$. We note that $\hat{\eta}_k$ includes an extra factor: the square root of $\Pr(I_{B_k} = 1)$. The value of $\hat{\epsilon}_k \hat{\eta}_k$ indicate the approximation error of the local gPC while $\Pr(I_{B_k} = 1)$ is the probability that the random variable $\mathbf{y}$ is located in the element $B_k$. These two factors should be balanced for the $h$-type refinement. For example, it may not be necessary to refine an element where $\eta_h$ is large while $\Pr(I_{B_k} = 1)$ is very small.

When the global error estimator is larger than a threshold, then $h$-type refinement is needed; however, it is not wise to refine all the random elements. We examine the values of $\hat{\eta}_k$ for each element and refine the elements satisfying

Criterion I: $\hat{\eta}_k \geq \theta_H \max_{i=1,\ldots,N} \hat{\eta}_i, \quad 0 < \theta_H < 1,$

(5.3.6)

where $\theta_H$ is a prescribed constant.

Alternatively, we can sort the local error indicators $\hat{\eta}_k$

$$\hat{\eta}_{i_1}, \hat{\eta}_{i_2}, \ldots, \hat{\eta}_{i_N}, \quad \hat{\eta}_{i_m} \geq \hat{\eta}_{i_n}, \forall m < n,$$

(5.3.7)

where $(i_1, i_2, \ldots, i_N)$ is a permutation of $(1, 2, \ldots, N)$, and refine a certain percent of the total elements. This criterion can be expressed as

Criterion II: $\hat{\eta}_{i_j}$ with $j \leq \theta_{II} N, \quad 0 < \theta_{II} < 1.$

(5.3.8)
In other words, Criterion I refines the elements according to a threshold for the local error indicator without any restriction on the number of elements while Criterion II refines a certain percent of the total elements with the largest local error indicators.

5.3.3 Importance of random dimension

Another question about adaptivity is how to refine the local random element. It is too expensive to refine all the random dimensions due to the "curse of dimensionality", i.e., $2^n$. Thus we need to consider the importance of random dimension. In this work we use the value $\rho_i$ to measure the importance of each random dimension and choose the most important random dimensions to refine while leave the rest unchanged. If $\rho_i$ decreases fast, the random dimension corresponding to the largest $\rho_i$ should be refined; otherwise, we can relax the restriction and refine the first two or three most important random dimensions simultaneously. The support of each random dimension is divided into two equidistant elements.

5.3.4 Revisiting the correction factor $\tilde{c}$

In the construction of $V^p$ and the estimation of $\tilde{c}$, we include the constant $c_{a,\alpha}$. However, such a constant is, in general, unknown. Furthermore, in different random elements the random process $a(x, y)$ takes different forms $a_k(x, y_k)$:

$$a_k(x, y_k) = \mathbb{E}[a](x) + \sigma \sum_{i=1}^{M} \frac{b_{k,i} + \alpha_{k,i}}{2} \sqrt{\lambda_i} h_i(x) + \sigma \sum_{i=1}^{M} \frac{b_{k,i} - \alpha_{k,i}}{2} \sqrt{\lambda_i} h_i(x)y_{k,i},$$

where the element $x^M_{i=1} [a_{k,i}, b_{k,i}]$ is mapped to a unit hypercube $[-1, 1]^M$. Thus, $c_{a_k,\alpha}$ is also element dependent.

We first present our strategy to deal with $c_{a_k,\alpha}$ and explain it subsequently.

1. Compute $c_{a,\alpha}$ in the global random space using the gPC basis. In other words, we need to solve the error equation in the full space $Y_{p,1}$.

2. Construct the reduced space $V^p_k$ in each random element using the coefficients $c_{a,\alpha}$. 

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
3. Take into account the relative fluctuation of $c_{a,k}$ with respect to $c_{a,a}$.

We start from the following observation:

**Proposition 5.3.2.** If in element $\times_{i=1}^M [a_{k,i}, b_{k,i}]$, the linear mapping (3.4.18) is defined, then the following relation holds

$$
\| \partial^\alpha_y u(x, y) \|_{H^1_0(D)} \prod_{i=1}^M \left( \frac{b_{k,i} - a_{k,i}}{2} \right)^{\alpha_i} = \| \partial^\alpha_y u_k(x, 0) \|_{H^1_0(D)}
$$

(5.3.9)

where $u(x, y)$ and $u_k(x, y_k)$ are global and local solutions, respectively, and

$$
y_{c,i} = \frac{b_{k,i} + a_{k,i}}{2}.
$$

**Proof.** We note that $u(y) = u_k(y_k(y))$. The conclusion follows from the chain rule of differentiation.

□

Using the equations (5.2.16) and (5.3.9) we can obtain the relation between $c_{a,k,\alpha}$ and $c_{a,\alpha}$ as

$$
c_{a,k,\alpha} = \frac{\| \partial^\alpha_y u(x, y_c) \|_{H^1_0(D)}}{\| \partial^\alpha_y u(x, 0) \|_{H^1_0(D)}}
$$

(5.3.10)

We now examine the fluctuation of $\| \partial^\alpha_y u(x, y_c) \|_{H^1_0(D)}$ around $\| \partial^\alpha_y u(x, 0) \|_{H^1_0(D)}$. For simplicity and without loss of generality we consider the following one-dimensional model:

$$
- \nabla \cdot ((b(x) + \delta b(x)) \nabla u(x, \xi)) = f(x),
$$

(5.3.11)

where $\delta b(x)$ is a perturbation function around $b(x)$ and $\xi$ is a random variable with zero mean on $[-1, 1]$.

Applying the operator $\partial^m_\xi$ to both sides of equation (5.3.11), we obtain

$$
- \nabla \cdot ((b + \delta b) \nabla \partial^m_\xi u(x, \xi)) = m \nabla \cdot (\delta b \nabla \partial^{m-1}_\xi u(x, \xi)).
$$

(5.3.12)
When $\xi = 0$ equation (5.3.12) becomes

$$-\nabla \cdot (b \nabla \partial_{\xi}^{m} u(x, 0)) = m \nabla \cdot (\delta b \nabla \partial_{\xi}^{m-1} u(x, 0)).$$

For a particular value $\xi = \xi_{i}$, we consider the change of $\partial_{\xi}^{m} u(x, \xi_{i})$ with respect to $\partial_{\xi}^{m} u(x, 0)$. Let $u_{\xi}^{m}(x, \xi_{i}) = \partial_{\xi}^{m} u(x, \xi_{i})$ We write $\partial_{\xi}^{m} u(x, \xi_{i})$ as

$$u_{\xi}^{m}(x, \xi_{i}) = u_{\xi}^{m}(x, 0) + \delta u_{\xi}^{m}(x, 0),$$

which satisfies the following equation

$$-\nabla \cdot ((b + \delta b \xi_{i}) \nabla (u_{\xi}^{m}(x, 0) + \delta u_{\xi}^{m}(x, 0))) = m \nabla \cdot (\delta b \nabla (u_{\xi}^{m-1}(x, 0) + \delta u_{\xi}^{m-1}(x, 0))).$$

The above equation can be simplified as

$$-\nabla \cdot ((b + \delta b \xi_{i}) \nabla \delta u_{\xi}^{m}(x, 0)) = \nabla \cdot (\delta b \xi_{i} \nabla u_{\xi}^{m}(x, 0)) + m \nabla \cdot (\delta b \nabla \delta u_{\xi}^{m-1}(x, 0)).$$

For a low-order derivative and small perturbation, we neglect the second-order terms and obtain

$$-\nabla \cdot (b \nabla \delta u_{\xi}^{m}(x, 0)) = \nabla \cdot (\delta b \xi_{i} \nabla u_{\xi}^{m}(x, 0)), \quad (5.3.13)$$

which yields

$$(b \nabla \delta u_{\xi}^{m}(x, 0), \nabla \delta u_{\xi}^{m}(x, 0)) = (\delta b \xi_{i} \nabla u_{\xi}^{m}(x, 0), \nabla \delta u_{\xi}^{m}(x, 0)) \quad (5.3.14)$$

by Green's formula. The above equation implies that it is reasonable to use the degree of perturbation of the random inputs to model the fluctuation of $c_{a,a}$ with respect to $c_{a,a}$.

Finally, for the elliptic problem we take the value of $\bar{c}$ as

$$\bar{c}^{2} = 1 + \left( \frac{1}{n_{S}} \sum_{s_{a}^{2}} \frac{\eta_{s_{a}^{2}}}{\overline{c_{s_{a}^{2}}}} - 1 \right) \left( \frac{1 + \sigma_{a}}{1 - \sigma_{a}} \right)^{2}, \quad (5.3.15)$$
where $\sigma_a$ is degree of perturbation of $a(x,y)$.

**Remark 5.3.3.** We note that we take the mean of $\eta_{\sigma_a}^2 / \bar{\eta}_{\sigma_a}^2$ instead of the maximum since we have a new factor $(1 + \sigma_a)/(1 - \sigma_a)$. When the number of terms in the reduced space increases, $\tilde{c}$ goes to 1, in other words, $V_p$ becomes $Y_{p,1}$.

**Algorithm 1** [h-type adaptive ME-gPC]

1. Choose a global tolerance $\epsilon$ and a steering parameter $0 < \theta < 1$.
2. Compute $c_{a,\alpha}$ by solving the error equation on the space $Y_{p,1}$.
3. Construct local polynomial chaos basis, implement gPC element-by-element.
4. Sort the values $\rho_{k,t} \approx c_k, \sqrt{\lambda_i}$ in element $k$ to measure the importance of each random dimension.
5. Construct the reduced space $V^p_k$.
6. Compute the local error indicator $\tilde{\eta}_k$ for new random elements.
7. If the global error indicator $\eta_g$ satisfies

   $$\eta_g = \left( \sum_{k=1}^{N} \tilde{\eta}_k^2 \right)^{1/2} > \epsilon,$$

   then
8. If $\tilde{\eta}_k$ satisfies Criteria (I) or (II) then
9. Refine the first two or three leading random dimensions.
10. end if
11. Go to step 3.
12. else
13. Stop and exit.
14. end if

### 5.3.5 Discussion of cost

Since the PDE system for the polynomial chaos coefficients can be decoupled [14], we can use the number of PDEs that needs to be solved as a measure of the overall cost. We assume the adaptive ME-gPC simulation leads to $N$ random elements. If the space $Y_{p,1}$ is used, the total cost for the a posteriori error estimates is $O(\dim(Y_{p,1})NT_1)$, where $T_1$ is the time for resolving one deterministic PDE; if the reduced space $V^p$ is used the cost is $O(\dim(Y_{p,1})T_1 + \dim(V^p)(N - 1)T_1)$. When $N$ is large enough, the ratio of the two costs is about $\theta_*$ (see section 5.2.3). We note here that unlike the deterministic adaptive methods, we do not need to resolve the problem in the entire parametric space when splitting occurs; we only need to solve the local problems in the newly added random elements.
5.4 Numerical results

In this section we present a numerical study on the proposed a posteriori error estimators and the adaptive ME-gPC method using stochastic algebraic and elliptic model problems.

5.4.1 The stochastic algebraic model

To illustrate the effectiveness of the presented concepts, we first consider the following algebraic equation

\[
(\beta_0 + \sigma \sum_{i=1}^{M} \sqrt{\lambda_i} y_i) u = 1,
\]

where \(\beta_0\) and \(\sigma\) are constants, \(\lambda_i\) are eigenvalues from a second-order random process, and \(y_i \sim U[-1,1]\) are i.i.d. uniform random variables with zero mean and unit variance. To avoid the singularity at zero we assume that \(\beta_0 + \sigma \sum_{i=1}^{M} \sqrt{\lambda_i} y_i > 0\). We can appreciate that such an algebraic equation shares common properties with the elliptic model problem with random coefficients. For the algebraic model we consider the norm \(\| \cdot \|_{L^2(B)}\).

We consider the Taylor expansion of the solution. The term \(\partial_y^\alpha u(0)\) takes the form

\[
\partial_y^\alpha u(0) = \rho^\alpha \partial_z^{|\alpha|} u(z = 0),
\]

where \(z = \sigma \sum_{i=1}^{M} \sqrt{\lambda_i} y_i\) and \(\rho_i = \sigma \sqrt{\lambda_i}\). Thus, the Taylor expansion of the solution can be expressed as

\[
u(y) = \sum_{\alpha} \frac{\partial_y^\alpha u(0)}{\alpha!} y^\alpha = \sum_{\alpha} \frac{\partial_z^{|\alpha|} u(z = 0)}{\alpha!} \rho^\alpha y^\alpha.
\]

We note that given a particular point \(y_i\),

\[
\frac{\partial_y^\alpha u(y_i)}{\partial_y^\alpha u(0)} = \frac{\partial_z^{|\alpha|} u(z = z(y_i))}{\partial_z^{|\alpha|} u(z = 0)} = \text{constant},
\]

where the constant only depends on \(|\alpha|\) instead of \(\alpha\).

We consider two typical covariance kernels \(\exp(-(x - y)^2 / A)\) and \(\exp(-|x - y| / A)\), where \(x, y \in\)
[0, 1] and $A$ is the correlation length. We note that the decay rate of eigenvalues is exponential for the Gaussian kernel and asymptotically algebraic for the exponential kernel. In this work we focus on $n = 10$ random dimensions and consider two correlation lengths: $A = 1$ and $0.1$. When $A = 1$, $\lambda_{10}/\lambda_1 = 3.4 \times 10^{-3}, 2.8 \times 10^{-14}$ for exponential and Gaussian kernels, respectively; when $A = 0.1$, $\lambda_{10}/\lambda_1 = 1.1 \times 10^{-1}, 5.7 \times 10^{-6}$ correspondingly.

**Error estimates**

In figures 5.1 and 5.2 we present the error contribution $u_{\alpha}^2$ of each term in $Y_{p,1}$ for $p = 1, 2$, respectively, using the exponential kernel. It is observed that for each pattern $s_{\alpha}^i$, the error contribution shows an overall decreasing trend, which is determined by the factor $\rho^\alpha$ since the factors $T(\alpha)$ are comparable for each pattern $s_{\alpha}^i$. The decay rate is determined by the decay rate of eigenvalues. In figure 5.2, similar behaviors are observed. In figures 5.3 and 5.4, we plot the values $u_{\alpha}^2/\rho^{2\alpha}$, which are almost constant for each $s_{\alpha}^i$. This implies that it is reasonable to separate the factor $\rho^\alpha$ in the error estimate. Furthermore, it is observed that for a fixed polynomial order the profiles of $u_{\alpha}^2/\rho^{2\alpha}$ are almost the same for $A = 0.1, 1$. Such observations can be easily explained using equations (5.4.3) and (5.2.15). Also, this is the motivation in constructing the reduced space based on the pattern $s_{\alpha}^i$ since we do not need to consider the constants related to each pattern.

We now examine the effectiveness of the reduced space $V_p$. Let $\eta_W$ denote the error from the full space $Y_{p,1}$ and $\eta_V$ the error from the reduced space $V_p$. We use the largest 10% terms for each $s_{\alpha}^i$ in space $Y_{p,1}$, which are indicated in figures 5.1 and 5.2 by filled markers. In table 5.2 we show the values of $\eta$ and $\tilde{\eta}$ for different cases. It is observed the a posteriori error estimator $\tilde{\eta}$ is effective for the algebraic model with about 90% savings.

**Adaptive ME-gPC**

We now look at the behavior of the adaptive ME-gPC method for the algebraic problem. We let the polynomial order be $p = 2$. If the importance of each random dimension is the same, the $h$-convergence of ME-gPC for the second-order moment should be $O(N^{-2(2+1)/10}) = O(N^{-0.6})$ [29, 14, 106], which is close to the convergence rate of the standard Monte Carlo method.
Figure 5.1: Error contribution of each mode in $Y_{p,1}$ with $p = 1$ and $M = 10$. Eigenvalues are from kernel $\exp(|x - y|/A)$ in $[0,1]$. The markers of the largest 10% terms for each $s^p_0$ are filled. Left: $A = 1$; Right: $A = 0.1$.

Figure 5.2: Error contribution of each mode in $Y_{p,1}$ with $p = 2$ and $M = 10$. Eigenvalues are from kernel $\exp(|x - y|/A)$ in $[0,1]$. The markers of the largest 10% terms for each $s^p_0$ are filled. Left: $A = 1$; Right: $A = 0.1$.

Figure 5.3: The factor $u_0^2/\rho^2$ of each term in $Y_{p,1}$ with $p = 1$ and $M = 10$. Eigenvalues are from kernel $\exp(|x - y|/A)$ in $[0,1]$. Left: $A = 1$; Right: $A = 0.1$. 

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Figure 5.4: The factor $u_2^2 / \rho^{2\alpha}$ of each term in $Y_{p,1}$ with $p = 2$ and $M = 10$. Eigenvalues are from kernel $\exp(|x - y|/A)$ in $[0,1]$. Left: $A = 1$; Right: $A = 0.1$.

<table>
<thead>
<tr>
<th>$A$</th>
<th>$p$</th>
<th>$\eta_Y$</th>
<th>$\eta_W$</th>
<th>$\eta_Y / \eta_W$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A = 1$, $p = 1$</td>
<td></td>
<td>4.30e-3</td>
<td>4.20e-3</td>
<td>1.02</td>
</tr>
<tr>
<td>$A = 0.1$, $p = 1$</td>
<td></td>
<td>4.08e-3</td>
<td>4.06e-3</td>
<td>1.00</td>
</tr>
<tr>
<td>$A = 1$, $p = 2$</td>
<td></td>
<td>4.65e-4</td>
<td>4.47e-4</td>
<td>1.04</td>
</tr>
<tr>
<td>$A = 0.1$, $p = 2$</td>
<td></td>
<td>5.12e-4</td>
<td>5.01e-4</td>
<td>1.02</td>
</tr>
<tr>
<td>$A = 0.1$, $\lambda_1 = \lambda_1$, $p = 2$</td>
<td></td>
<td>1.93e-3</td>
<td>1.93e-3</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table 5.2: A posteriori error estimates given by $Y_{p,1}$ and $V^p$. $\dim(V^p) / \dim(Y_{p,1}) = 0.1$. Eigenvalues are from the exponential kernel.
It is known that if the eigenvalues decrease fast, the error contribution mainly comes from the first several random dimensions. For a limit case, if the correlation length goes to infinity, the random behavior can be described by one random variable. This implies that the most effective way to do $h$-type refinement is to refine the random dimension related to the largest $\rho_i$. Let $n_r$ denote the number of random dimensions to be refined. We present the evolution of errors for two different refinement strategies: $n_r = 1$ or 2, in figure 5.5. Both exponential and Gaussian kernels are examined. The local error estimators are obtained from spaces $V_p$ and $Y_{p,1}$, where the correction factor $\hat{c}$ is chosen using equation (5.2.26) since $\partial_{y_i}^\alpha u(y_i)/\partial_y^\alpha u(0) = \text{constant}$ for a given $|\alpha|$. We let $\theta_{II} = 0.1$ for Criterion II (see equation (5.3.8)), in other words, we refine 10% elements when necessary. It is observed that it is more effective to refine only one random dimension. All the curves can be roughly decomposed to two parts (indicated by the dotted lines). The $h$-convergence rate is much better in the first part where the difference between $\rho_i$ is large. As the $h$-refinement goes on, $\rho_i$ goes close to each other and the $h$-convergence rate becomes close to $O(N^{-2(p+1)/n})$. Since the eigenvalues of Gaussian kernel decrease very fast when $A = 1$, only the first two random dimensions contribute to the errors $\geq 10^{-10}$, and thus the $h$-convergence rate is $\approx O(N^{-2(2+1)/3})$. For the errors $\leq 10^{-10}$, the third random dimension must be included resulting in an $h$-convergence rate $\approx O(N^{-2(2+1)/3})$. Such observations are consistent with the optimal $h$-convergence rate (see equation (5.3.4)) given in [100].

In figure 5.6, we present the convergence of $h$-type adaptive ME-gPC for different random distributions. The eigenvalues are from the exponential kernel. Firstly, we note that the reduced space $V_p$ provides a comparable error estimator with the space $Y_{p,1}$. Secondly, the error estimator given by $V_p$ works well not only for the uniform distribution but also for other random distributions, although the PDFs for each random dimension in a certain random element may be different. This observation implies that the factor $E[y^\alpha \phi_\alpha] \rho^\alpha$ provides a good prediction.
Figure 5.5: Adaptive errors of the second-order moments for the stochastic algebraic model. $N$ is the number of random elements. $p = 2$, $M = 10$ and $A = 1$. Uniform distribution is considered. Left: exponential kernel; Right: Gaussian kernel.

Figure 5.6: Adaptive errors of the second-order moments for the stochastic algebraic model. $N$ is the number of random elements. $p = 2$ and $M = 10$. Uniform and Beta distributions are considered. Eigenvalues are from the exponential kernel. Left: $A = 1$; Right: $A = 0.1$. 

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
5.4.2 The stochastic elliptic model

We next consider a one-dimensional elliptic problem

$$\frac{-d}{dx} \left( a(x; \omega) \frac{du}{dx} \right) = 1, \quad x \in [0,1]$$

subject to homogeneous Dirichlet boundary conditions. Using an exponential kernel, we approximate $a(x; \omega)$ by the K-L decomposition as

$$a_M(x; \omega) = \mathbb{E}[a](x) + \sigma \sum_{i=1}^{M} \sqrt{\lambda_i} h_i(x) y_i,$$

where $\sigma$ is standard deviation and $y_i$ are i.i.d. random variables on $[-1,1]$. We let $\sigma = 0.35$ and $M = 10$. The random field is approximated by gPC while the physical space is discretized by the spectral element method [58].

Error estimates

We now test the effectiveness of the reduced space $V_p$, where the norm $\|e\|_{L^2(B,H^1_0(D))}$ is used for the error estimates. We define the effectivity index as

$$C_U = \frac{\tilde{\eta}}{\eta_{W^{p+1}}},$$

where $\eta_{W^{p+1}}$ is the error estimate from the space $W^{p+1}$ instead of $Y_{p,1}$ and $\tilde{\eta}$ is the error estimate from the reduced space $V_p$. We compute $\tilde{\eta}$ using the equation (5.3.15). Since the number of random dimension is 10, it is hard to generate a standard mesh in the parametric space. We solve the elliptic problem adaptively up to a 1060-element mesh of the parametric space, where we let $\theta_{NF} = 0.2$ and $n_r = 2$. Since the final meshes can be different for different PDFs of $y_i$, we collect the information in a statistical sense, in other words, we compute the mean and standard deviation of the local effectivity indices, which are shown in table 5.3. We can see that the local effectivity indices are distributed basically around the mean value since the standard deviation is very small. In other
Table 5.3: Statistics of local effectivity indices for a mesh with 1060 random elements. $p = 2$ and $A = 1$.

words, the reduced space $V_p$ can recover the information efficiently although only 10% or 20% terms in $Y_{p,1}$ are used. We note that the reduced space works well for nonlinear PDFs due to the factor $E[y^{n} \phi_{a}]$ (see equation (5.2.15)). The larger $V_p$ yields better local effectivity indices as expected.

Adaptive ME-gPC

We next examine the adaptive behavior of ME-gPC for the stochastic elliptic model problem. In figure 5.7 we plot the global adaptive errors of the second-order moments ($H_0^1(D)$ norm in the physical space) from the space $W_{p+1}$ and the reduced space $V_p$. We let $\theta_H = 0.1$ and $n_r = 1$. It appears that the a posteriori error estimator from $V_p$ provides a good prediction of the true errors with a much smaller overall cost. The curves show a similar two-part structure as we observed in the stochastic algebraic model, which implies that the splitting strategy is effective. Although the PDF is decomposed in ME-gPC and the local polynomial bases may be different between two random dimensions, the proposed a posteriori error estimator works well for all tested PDFs.
Figure 5.7: Adaptive errors of the second-order moments for the stochastic elliptic model. \( N \) is the number of random elements. \( p = 2, A = 1 \) and \( M = 10 \). Uniform and Beta distributions are considered. Eigenvalues are from the exponential kernel. Left: \( \dim(V^p)/\dim(Y_{\mu,1}) = 0.1 \); Right: \( \dim(V^p)/\dim(Y_{\mu,1}) = 0.2 \).
Chapter 6

Long-term behavior of polynomial chaos

In this chapter we study the long-term behavior of gPC and ME-gPC [105] using the following one-dimensional stochastic advection equation

\[ \frac{\partial u}{\partial t} + V(Y) \frac{\partial u}{\partial x} = 0 \]  

(6.0.1)

subject to the initial condition

\[ u(x) = u_0(x; Y), \]  

(6.0.2)

where \( Y \) is a one-dimensional uniform random variable defined on \([-1, 1]\) and the transport velocity \( V(Y) \) is finite. In particular, we assume that

\[ V(Y) = \bar{v} + \sigma Y, \quad u_0(x; Y) = \sin n\pi(1 + x), \quad x \in [-1, 1], \]  

(6.0.3)

where \( \sigma \) is a constant, \( \bar{v} \) is the mean of transport velocity and \( n \in \mathbb{N} \). It is easy to obtain the exact solution for this case as

\[ u(x, t; Y) = \sin n\pi(1 + x - (\bar{v} + \sigma Y)t), \]  

(6.0.4)
which shows that the frequency is random.

In the first two sections we provide the error estimates of approximations of gPC and ME-gPC, respectively, using the Legendre-chaos. In the third section we study the relation between the error evolution of gPC and ME-gPC in terms of the numbers elements in the ME-gPC mesh. In the fourth section, we present some discussions on general cases. We present some numerical verifications in the fifth section.

6.1 Error estimates for gPC

Let \{P_i(Y)\} denote the orthogonal basis of Legendre-chaos and \(P_m\) denote the projection operator as

\[
\mathcal{P}_M u(x, t; Y) = \sum_{i=0}^{M} u_i(x, t) P_i(Y),
\]

where

\[
u_i(x, t) = \frac{1}{\mathbb{E}[P_i^2(Y)]} \int_{-1}^{1} u(x, t; y) P_i(y) \frac{1}{2} dy.
\]

We study the convergence of \(\mathcal{P}_M u(x, t; Y)\) since it can demonstrate the main properties of numerical convergence.

**Theorem 6.1.1.** Let \(\epsilon_M\) denote the error of the second-order moment of \(\mathcal{P}_M u\). Given time \(t\) and polynomial order \(M\), \(\epsilon_M\) can be bounded as

\[
\epsilon_M \leq C(M) \frac{q_M^{2M+2}}{1 - q_M^2},
\]

where \(C(M)\) is a constant depending on \(M\) and

\[
q_M = \frac{\sigma \pi e t}{2M + 2} < 1.
\]

Here \(e\) is the base of natural logarithm.
Proof. According to the following formula in [45]

$$\sin c\pi(z + a) = \frac{1}{\sqrt{2c}} \sum_{i=0}^{\infty} (2i + 1)J_{i+1/2}(c\pi) \sin(c\pi a + \frac{1}{2} i\pi)P_i(z), \quad (6.1.5)$$

we obtain the polynomial chaos expansion of the exact solution (6.0.4) as

$$u = \frac{1}{\sqrt{2n\pi t}} \sum_{i=0}^{\infty} (2i + 1)J_{i+1/2}(\sigma n\pi t) \sin(n\pi(\bar{v}t - x - 1) + \frac{1}{2} i\pi)P_i(Y), \quad (6.1.6)$$

where $J_{i+1/2}$ are Bessel functions of the first kind. Using the orthogonality of Legendre polynomials, we obtain

$$E[u^2(x, t; Y)] = \frac{1}{2n\pi t} \sum_{i=M+1}^{\infty} (2i + 1)J_{i+1/2}(\sigma n\pi t) \sin^2(n\pi(\bar{v}t - x - 1) + \frac{1}{2} i\pi), \quad (6.1.7)$$

where $E[P_i(Y)P_j(Y)] = \delta_{ij}/(2i + 1)$ is employed. Then $\epsilon_M$ can be expressed as

$$\begin{align*}
\epsilon_M &= E[u^2] - E[(P_M u)^2] \\
&= \frac{1}{2n\pi t} \sum_{i=M+1}^{\infty} (2i + 1)J_{i+1/2}(\sigma n\pi t) \sin^2(n\pi(\bar{v}t - x - 1) + \frac{1}{2} i\pi). \quad (6.1.8)
\end{align*}$$

It is known (see [1]) that

$$\sqrt{\frac{\pi}{2\sigma n\pi t}}J_{i+1/2}(\sigma n\pi t) = \frac{(\sigma n\pi t)^i}{2^i i!} \int_0^\pi \cos((\sigma n\pi t)\cos(\theta)) \sin^{2i+1} \theta \, d\theta. \quad (6.1.9)$$

By substituting equation (6.1.9) into equation (6.1.8), $\epsilon_M$ can be approximated as

$$\epsilon_M = \sum_{i=M+1}^{\infty} \frac{(2i + 1)(\sigma n\pi t)^{2i}}{2^{2i+2}(i!)^2} A_i \sin^2(n\pi(\bar{v}t - x - 1) + \frac{1}{2} i\pi),$$

where

$$A_i = \left( \int_0^\pi \cos((\sigma n\pi t)\cos(\theta)) \sin^{2i+1} \theta \, d\theta \right)^2.$$
Using Stirling's formula [1] for the factorial \( i! \), we obtain that

\[
\epsilon_M \approx \sum_{i=M+1}^{\infty} \frac{(2i + 1)(\sigma \pi te)^{2i}}{8\pi i(2i)^{2i}} A_i \sin^2(\pi(\theta t - x - 1) + \frac{1}{2} \pi),
\]

where \( e \) is the base of natural logarithm. For a fixed time \( t \), the error \( \epsilon_M \) can be bounded as

\[
\epsilon_M \leq C_1 \sum_{i=M+1}^{\infty} \frac{(2M + 3)(\sigma \pi te)^{2i}}{8\pi(M + 1)(2M + 2)^{2i}} = C_1 \frac{(2M + 3)q_M^{2M+2}}{8\pi(M + 1)(1 - q_M^2)},
\]

where \( C_1 \) is a constant and \( q_M = \sigma \pi te/(2M + 2) \). Here the condition \( q_M < 1 \) is assumed for the convergence of summation. We subsequently check the constant \( C_1 \). Since \( \sin \theta \geq 0 \) in \( \theta \in [0, \pi] \), we obtain that

\[
A_i^{1/2} \leq \int_0^\pi \sin^{2i+1} \theta \, d\theta.
\]

Let \( B_i = \int_0^\pi \sin^{2i+1} \theta \, d\theta \). Using \( \sin^2 \theta + \cos^2 \theta = 1 \), the following relationship can be obtained

\[
B_i = \int_0^\pi \sin^{2i-1} \theta \, d\theta - \int_0^\pi \sin^{2i-1} \theta \cos^2 \theta \, d\theta = B_{i-1} - \int_0^\pi \sin^{2i-1} \theta \cos^2 \theta \, d\theta.
\]

Since the second term on the right-hand side is positive, we know that the sequence \( \{B_i\} \) is decreasing. Thus, we can bound \( A_i \) as

\[
A_i \leq B_i^2 \leq B_{i+1}^2, \quad i \geq (M + 1).
\]

Let \( C_1 = B_{M+1}^2 \) and \( C(M) = C_1(2M + 3)/8\pi(M + 1) \), then the conclusion follows immediately. \( \square \)

In Theorem 6.1.1, \( q_M < 1 \) is assumed for the convergence of summation in equation (6.1). For a general case we have the following corollary:

**Corollary 6.1.2.** Given time \( t \) and polynomial order \( M \), \( \epsilon_M \) can be bounded as

\[
\epsilon_M \leq \frac{1}{2\sigma t} \sum_{i=M+1}^{\tilde{M}} (2i + 1)J_{i+1/2}^2(\sigma \pi t) + C(\tilde{M})q_{\tilde{M}}^{2\tilde{M}+2} \quad \frac{1}{1 - q_{\tilde{M}}^2},
\]  

(6.1.10)
where $q_M$ is a function of $M$ defined as in equation (6.1.4) and $q_M < 1$.

6.2 Error estimates for ME-gPC

Let $\hat{P}_M$ denote the projection of $u(x, t; Y)$ onto the basis of ME-gPC.

**Theorem 6.2.1.** Given a decomposition of random space of $Y$ with element length $L_k = b_k - a_k$, $k = 1, 2, \cdots, N$, the error $\hat{e}_M$ of the second-order moment of $\hat{P}_M u$ can be bounded as

$$\hat{e}_M \leq C(M) \sum_{k=1}^{N} \frac{q_{k,M}^{2M+2}}{1 - q_{k,M}^2} \Pr(I_{B_k} = 1), \quad (6.2.1)$$

where $C(M)$ is a constant depending on $M$ and

$$q_{k,M} = \frac{\sigma \pi \epsilon L_k t}{2(2M + 2)} < 1. \quad (6.2.2)$$

**Proof.** We know that $\hat{P}_M$ can be expressed as

$$\hat{P}_M = \sum_{k=1}^{N} \hat{P}_{k,M} I_{B_k}, \quad (6.2.3)$$

where $\hat{P}_{k,M}$ is a local projection operator defined as

$$\hat{P}_{k,M} u(x, t; Y) = p_{k,M} u(x, t; \frac{b_k - a_k}{2} Y + \frac{b_k + a_k}{2}). \quad (6.2.4)$$

Then, the second-order moment can be expressed as

$$\mathbb{E}[(\hat{P}_M u(x, t; Y))^2] = \mathbb{E}\left[\left(\sum_{k=1}^{N} \hat{P}_{k,M} u(x, t; Y) I_{B_k}\right)^2\right] = \sum_{k=1}^{N} \mathbb{E}[\left(\hat{P}_{k,M} u(x, t; Y)\right)^2] \Pr(I_{B_k} = 1). \quad (6.2.5)$$
Thus, $\hat{\epsilon}_M$ takes the following form

$$\hat{\epsilon}_M = \sum_{k=1}^N \hat{\epsilon}_{k,M} \Pr(I_{B_k} = 1),$$

(6.2.6)

where $\hat{\epsilon}_{k,M}$ is the error of the second-order moment of $\hat{\mathcal{P}}_{k,M}u(x,t;Y)$. We now check the behavior of $\hat{\epsilon}_{k,M}$. Given a random element $B_k = [a_k, b_k]$, the local problem in ME-gPC is to find the solution of the following transformed equation

$$\frac{\partial u}{\partial t} + \left( \bar{v} + \sigma \left( \frac{b_k - a_k}{2} Y_k + \frac{b_k + a_k}{2} \right) \right) \frac{\partial u}{\partial x} = 0,$$

where $Y_k$ is the local uniform random variable defined on $[-1,1]$. The exact solution of equation (6.2) is

$$u(x,t;Y_k) = \sin n\pi \left[ 1 + x - \left( \bar{v} + \sigma \left( \frac{b_k - a_k}{2} Y_k + \frac{b_k + a_k}{2} \right) \right) t \right].$$

Using a similar procedure as in section (6.1), we can bound $\hat{\epsilon}_{k,M}$ as

$$\hat{\epsilon}_{k,M} \leq \frac{C(M)q_{k,M}^{2M+2}}{1 - q_{k,M}^2},$$

where

$$q_{k,M} = \frac{\sigma n \pi (b_k - a_k) t}{2(2M + 2)} < 1.$$

Using equation (6.2.6), the conclusion follows immediately.

$\square$

### 6.3 Relation between $\epsilon_M$ and $\hat{\epsilon}_M$

From theorem 6.1.1, we can see that $q_M$ increases linearly in terms of time $t$, which implies that gPC will lose $p$-convergence after a finite time. To keep a certain accuracy, the polynomial order of gPC must increase with time. Let

$$\delta = \frac{C(M)}{1 - q_M^2} q_M^{2M+2},$$

(6.3.1)
where $\delta$ denotes a desired accuracy. By solving such an equation we obtain that

$$t = \frac{1}{\sigma \eta \epsilon} \left( \frac{\delta(1 - q_M^2)}{C(M)} \right)^{1/(2M+2)} (2M + 2).$$

(6.3.2)

Since $[\delta(1 - q_M^2)/C(M)]^{1/(2M+2)} \to 1$ when $M \to \infty$, we obtain that

$$t \approx \frac{2M + 2}{\sigma \eta \epsilon},$$

(6.3.3)

which is a linear relation. It is instructive to define the increasing speed of polynomial order as

$$\frac{dM}{dt} \approx \frac{\sigma \eta \epsilon}{2},$$

(6.3.4)

which shows that to maintain an accuracy $\delta$ the polynomial order must increase at a speed $\sigma \eta \epsilon/2$; we note that $n/2$ is the wave number in the initial condition. We can see that the speed is proportional to the wave number and the degree of perturbation, which implies that gPC will quickly fail to converge for a problem with a large perturbation or wave number if a random frequency in time is involved.

**Theorem 6.3.1.** To maintain a certain accuracy of the second-order moment of $P_M u$, the polynomial order of gPC must increase with time and the following relation is satisfied

$$M \approx \frac{1}{2} \sigma \eta \epsilon t - 1.$$

(6.3.5)

We assume that a uniform mesh is employed and $p$-convergence is maintained, in other words, $q_M < 1$ and $q_{k,M} < 1$ with $k = 1, 2, \ldots, N$. Thus, we have

$$\dot{\epsilon}_M = C(M) \frac{q_M^{2M+2}}{1 - q_M^2}.$$

(6.3.6)
The ratio of $\hat{\epsilon}_M$ and $\epsilon_M$, for a fixed time $t$ and polynomial order $M$, is

$$\frac{\hat{\epsilon}_M}{\epsilon_M} = \left(\frac{1}{N}\right)^{2M+2} \frac{1 - q_{M}^{2}}{1 - (\frac{1}{N})^{2} q_{M}^{2}} \sim \left(\frac{1}{N}\right)^{2M+2}, \quad (6.3.7)$$

which is consistent with the $k$-convergence ($\epsilon \propto N^{-2(M+1)}$) of ME-gPC (see [29, 14, 102, 106]). Let us consider that gPC and ME-gPC of polynomial order $M$ reach accuracy of the same order. To satisfy this, we need to have

$$q_{M} = \hat{q}_{M}, \quad (6.3.8)$$

which yields

$$\hat{t}_{M} = N t_{M}, \quad (6.3.9)$$

where $\hat{t}_M$ and $t_M$ denote time for ME-gPC and gPC, respectively.

**Theorem 6.3.2.** Suppose that the error, $\epsilon$, of the second-order moment of gPC of order $M$ is maintained in the range $t \leq t_g$. Based on a uniform mesh with $N$ random elements, ME-gPC of order $M$ can maintain the accuracy $O(\epsilon)$ in the range $t \leq N t_g$. In other words, ME-gPC can extend the valid integration time of gPC linearly by a factor $N$.

If the mesh is non-uniform, the aforementioned linearity is still valid; however, the factor will be less than $N$. We assume that $\bigcup_{k=1}^{N} B_{k}$ is a decomposition for $Y$ of uniform distribution and the length of $B_{k}$ is an increasing series, $0 < l_{B_{1}} \leq l_{B_{2}} \leq \cdots \leq l_{B_{N}}$. From the proof of theorem 6.2.1, we know that

$$\hat{\epsilon}_M = \sum_{k=1}^{N} \hat{\epsilon}_{k,M} \Pr(I_{B_{k}} = 1) \leq C(M) \sum_{k=1}^{N} \frac{q_{k,M}^{2M+2}}{1 - q_{k,M}^{2}} \frac{l_{B_{k}}}{2}, \quad (6.3.10)$$

where

$$q_{k,M} = \frac{\sigma_{k} \pi e_{I_{B_{k}}} t}{2(2M + 2)} \quad \text{and} \quad \Pr(I_{B_{k}} = 1) = \frac{l_{B_{k}}}{2}. \quad (6.3.11)$$

We define a function

$$Q(z) = \frac{z^{2M+2}}{1 - q_{z}^{2}}.$$
where
\[ q_z = \frac{\sigma \pi \kappa t}{(2M + 2)^2} z. \]

It is easy to verify that \( Q(z) \) is an increasing function with respect to \( z \). Let \( z_k = \frac{t_{B_k}}{2} \). We can obtain
\[ Q(z_1) = \sum_{k=1}^{N} Q(z_k) z_k \leq \sum_{k=1}^{N} Q(z_k) z_k \leq \sum_{k=1}^{N} Q(z_N) z_k = Q(z_N), \]
where
\[ \sum_{k=1}^{N} z_k = 1, \quad 0 < z_1 \leq z_2 \leq \cdots \leq z_N. \]

To satisfy \( \epsilon_M = \epsilon_M \), we need
\[ Q(z_1) \leq Q(1) \leq Q(z_N), \]
which implies that
\[ \frac{2}{l_{B_1}} t_M \leq t_M \leq \frac{2}{l_{B_1}} t_M. \] (6.3.11)

### 6.4 Some discussions on general cases

#### 6.4.1 Other distributions and high-dimensional random inputs

In ME-gPC, the PDF of \( Y \) will be decomposed simultaneously with the random space; thus, the local orthogonality has to be maintained numerically. The only exception is the Legendre-chaos [102] due to the nice properties of uniform distribution. It is, in general, difficult to analyze theoretically the convergence for the numerical basis of ME-gPC. In this work, we compare the performance of gPC and ME-gPC numerically for other distributions.

The \( k \)-type convergence was shown theoretically in [14, 29] to be
\[ \|E[u] - E[u_M]\|_{L_2(D)} \leq Ck^{2(M+1)}, \quad \|E[u^2] - E[u_M^2]\|_{L_2(D)} \leq Ck^{2(M+1)} \] (6.4.1)

using a stochastic elliptic model problem, where \( C \) is a constant depending on \( M \), and \( k \) denotes the maximum size of random elements. We note here that \( D \) indicates the physical space. It was
shown in [102, 106] that the index of algebraic convergence of ME-gPC for the mean and variance goes asymptotically to \(2(M + 1)\) for a uniform mesh, which is consistent with equation (6.4.1). Note that the error bound (6.4.1) is independent of probability measures. Such observations imply that for any probability measure the following relation (see equation (6.3.7))

\[
\frac{\hat{\epsilon}_M}{\epsilon_M} \sim C(M)(\frac{\hat{\epsilon}_M}{N\epsilon_M})^{2M+2}
\]  

(6.4.2)

holds for the solution of equation (6.0.4), where the constant \(C\) depends on the polynomial order \(M\). Here we include the time \(t\) together with the number \(N\) of random elements because in the \(k\)-th random element of ME-gPC the solution takes the form

\[
u(x, t; Y_k) = \sin n\pi (1 + x - (\bar{b}_k + \frac{\sigma_t}{N}Y_k)),
\]  

(6.4.3)

where \(t\) and \(N\) can be treated together. From equation (6.4.2) we can see that if we re-scale the time of gPC by the number \(N\), \(\hat{\epsilon}_M/\epsilon_M\) will be a constant depending on the polynomial order. However, such a relation will be reached asymptotically because of the non-uniform random distribution [106].

It is easy to generalize the obtained results to high-dimensional random inputs. Since the high-dimensional basis of gPC is constructed by tensor products of one-dimensional basis, the error of chaos expansion should be dominated by the summation of errors of one-dimensional truncation. Thus, the results for the one-dimensional case should be still valid for a high-dimensional case. For example, if we have \(d\)-dimensional uniform random inputs, ME-gPC with \(N^d\) uniform elements should extend the valid integration time of gPC with the same polynomial order by a factor \(N\). However, in practice, the degree of perturbation in each random dimension is generally different, and only the random dimensions with large perturbations are needed to be refined. Such cases are much more difficult to analyze.
6.4.2 Initial conditions

For the initial condition \( u_0(x; Y) \), we intentionally employed functions such as cosine and sine waves, which introduce "random periodicity" in time for a given random transport velocity. Such solutions are often encountered in practice, e.g., random oscillators and simulations of unsteady turbulent or noisy flows. If the frequency is finite, we know that the random solution can be expressed by a Fourier transform in the time direction

\[
M/2 \quad u(x,t,Y) = \sum_{n=-M/2}^{M/2} u_n(x; Y)e^{\frac{2\pi in}{T}Yt}, \tag{6.4.4}
\]

where \( T(Y) \) is the random period. It is obvious that equation (6.0.4) represents the basic properties of each random mode. We note that gPC can effectively capture the random behavior for some other initial conditions, e.g., \( u_0(x; Y) = x^n \), if the polynomial order is large enough.

The long-term behavior for aforementioned initial conditions is similar to the spectral expansion of deterministic functions with high wave numbers (see [3]). Since gPC is indeed a spectral expansion in terms of certain random variables and the time plays a role similar to a wave number in chaos expansion, the order of gPC must increase with time to maintain a desired accuracy level.

6.5 Numerical results

Next we present some numerical results for \( \epsilon_M \) and \( \hat{\epsilon}_M \). Let \( n = 1, \vartheta = 0, \sigma = 1 \) in equation (6.0.3). Due to the periodic condition in physical space, we use a Fourier-collocation method to solve the deterministic PDEs introduced by the Galerkin projection in the gPC or ME-gPC method. It has been assumed that \( q_M < 1 \) for the convergence of summation in equation (6.1.4). However, \( q_M \) is an increasing function of \( t \), which means that the error \( \epsilon_M \) increases with time and it will reach \( O(1) \) values eventually. In figure 6.1, we present the evolution of the error bounds of gPC and ME-gPC, respectively. It can be seen that the error of ME-gPC increases at the same speed as that of gPC. However, since ME-gPC is much more accurate than gPC, it takes longer time for ME-gPC to reach \( O(1) \) error. The error \( \epsilon_M \) of eighth-order gPC is \( O(1) \) around \( t = 2 \). In figure 6.2, \( \epsilon_M \)
and the corresponding numerical errors are shown, where we also plot \( J_{i+1/2}^2(\sigma \pi t) \) for comparison. It is known that \( J_{i+1/2}^2(\sigma \pi t) \) decreases exponentially with \( i \) when \( i \) is much greater than \( \sigma \pi t \); otherwise, there is no \( p \)-convergence. It can be seen that \( p \)-convergence does not occur until \( M \geq 8 \) \((q_M < 1)\) and the rate of convergence is the same as the decreasing rate of \( J_{i+1/2}^2(\sigma \pi t) \) when \( i \to \infty \).

In figure 6.3 we demonstrate theorem 6.3.2 numerically. According to theorem 6.3.2, we know that the error of gPC at time \( t \) should be almost the same as the error of ME-gPC of the same polynomial order at time \( Nt \) for a uniform mesh. For gPC, we re-scale the time by a factor \( N \) while keeping the errors unchanged. It can be seen that the re-scaled error-time curve of gPC matches very well with the error-time curve of ME-gPC, and it appears that the errors of ME-gPC are always bounded by the shifted errors of gPC. We note that the random inputs are uniform.

To examine if the previous results extend to other random distributions, we consider Beta and Gaussian distributions. In figure 6.4, we plot the errors of gPC and ME-gPC versus time for a Beta distribution \( \text{Beta}(0,1) \) while the time for gPC is re-scaled as before. It is seen that the two curves for gPC and ME-gPC agree with each other very well. We also compare gPC and ME-gPC of polynomial order \( p = 3 \) in figure 6.5 when \( Y \) is of Gaussian distribution. For the ME-gPC method, we first decompose the support of the Gaussian distribution into three random elements: \((-\infty, -6], [-6, 6] \) and \([6, \infty)\). We subsequently decompose the middle element \([-6, 6]\) while keeping the tail elements unchanged since \( \Pr(Y \in (-\infty, -6) \text{ or } [6, \infty)) = 1.97 \times 10^{-9} \). The influence of tail elements can be observed only in the early stage, which is clearly shown in the left plot. Starting from \( t \approx 0.5 \), the tail elements will not affect the accuracy any more. Thus, we can drop the tail elements when we re-scale the error curves of gPC by the element number \( N \) due to their negligible error contribution. In the plot on the left, it can be seen that the error curve re-scaled by factor 2.5 matches the ME-gPC results better than the element number \( N = 4 \). The reason is that Gaussian distribution is non-uniform. If the middle element is decomposed to four equidistant ones, we know that \( \Pr(I_{-6,3}] = 1 \text{ or } I_{[3,6]} = 1) = 2.7 \times 10^{-3} \), which implies that the error of ME-gPC is mainly controlled by elements \([-3, 0] \text{ and } [0, 3]\) (see equation (6.2.6)). Thus the scale factor should be about 2. In the plot on the right, we can see that for larger element numbers, the re-scaled error curves of
Figure 6.1: Evolution of error estimates of gPC and ME-gPC. Here, $Y$ is uniform in $[-1, 1]$.

gPC agree well with the ME-gPC results, which implies the relation presented in theorem 6.3.2 will be valid asymptotically for the Gaussian distribution.
Figure 6.2: Convergence of gPC and $J_{t+1/2}$ at $t = 2$. Here, $Y$ is uniform in $[-1, 1]$.

Figure 6.3: Comparison of gPC and ME-gPC of the same polynomial order $p = 3$. Normalized numerical errors of the second moment are used. The time of gPC is multiplied by the number $N$ of random elements. Here, $Y$ is of uniform distribution $U[-1, 1]$. Left: ME-gPC with $N = 2$; Right: ME-gPC with $N = 3$. 

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Figure 6.4: Comparison of gPC and ME-gPC of the same polynomial order $p = 3$. Normalized numerical errors of the second moment are used. The time of gPC is multiplied by the number $N$ of random elements. Here, $Y$ is of Beta distribution $Beta(0, 1)$ on $[-1, 1]$.

Figure 6.5: Comparison of gPC and ME-gPC of the same polynomial order $p = 3$. Normalized numerical errors of the second moment are used. The time of gPC is re-scaled by a constant. Here, $Y$ is of Gaussian distribution $N(0, 0.2)$. Left: ME-gPC with $N = 4$; Right: ME-gPC with $N = 8, 16$. 

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Chapter 7

Adaptivity of deterministic spectral/hp element method

In this Chapter we present some results on a posteriori error estimators for deterministic elliptic problems. In the first section we present a short overview of a posteriori error estimators in the deterministic finite element methods. In the second section, we provide the polynomial approximation space based on two-dimensional spectral/hp element discretization. In the third section we present explicit equilibrated fluxes for the local pure-Neumann problems used in the a posteriori error estimator in [5]. We modify the target function for minimization to obtain new equilibrated fluxes in a weighted form. In the fourth section, we demonstrate how to deal with non-homogeneous boundary conditions in the a posteriori error estimate. In the fifth section, we investigate the performance of the new equilibrated fluxes in two-dimensional spectral/hp element approximations of an elliptic equation using conforming triangle elements. The new fluxes appear more robust and lead to substantial computational savings compared to the original numerical procedure in [5].
7.1 Introduction

A posteriori error estimation has become an important tool in the finite element method, as it can be used to control the numerical errors and guide goal-oriented mesh adaptivity strategies. Many a posteriori error estimators have been developed (see [101, 8, 16] and references therein), which are mainly based on the residual method [15, 6, 7, 19, 18] or on the recovery method [124, 125]. Recently, a new approach referred to as goal-oriented error estimation has been developed, which measures the error of a linear functional of the solution rather than the usual energy norm (see [8] and references therein).

Among the residual type a posteriori error estimators the equilibrated method [6, 7] is now accepted as the most effective one, as it can be applied reliably to $hp$ finite element methods. In this paper we investigate the performance of the equilibrated residual method for the spectral/$hp$ element method [58]. The equilibrated residual method is based on the solutions of local Neumann type boundary-value problems. To obtain consistent flux conditions a singular linear system needs to be solved. In [7], a numerical procedure was proposed. To avoid solving such a singular linear system numerically, we derive the equilibrated fluxes in an explicit form for the two-dimensional finite elements. We subsequently introduce some weights into the target function for minimization and obtain new equilibrated fluxes. We compare the performance of the new fluxes with that given in [6] using elliptic problems with smooth and singular solutions. It appears that the new fluxes are less sensitive to the deformation of the finite elements with low-order polynomials; for $p$-version finite elements it is not so important how to determine the equilibration because the approximation solution can provide a good approximation of flux. Furthermore, the cost for the equilibrated fluxes is reduced to $O(N_{\text{patch}})$ for each interior vertex, where $N_{\text{patch}}$ is the number of elements forming the support of the corresponding linear shape function.
7.2 Model problem and its discretization

Let $\Omega \subset \mathbb{R}^2$ be a bounded domain with a Lipschitz boundary $\partial \Omega$. We consider the following elliptic boundary-value problem:

$$
-\Delta u + cu = f \quad \text{in } \Omega, \quad (7.2.1a)
$$

$$
\frac{\partial u}{\partial n} = g \quad \text{on } \Gamma_N, \quad (7.2.1b)
$$

$$
u = 0 \quad \text{on } \Gamma_D, \quad (7.2.1c)
$$

where the boundary segments $\Gamma_N$ and $\Gamma_D$ are disjoint with $\Gamma_N \cup \Gamma_D = \partial \Omega$ and $c$ is a nonnegative constant. We assume that $f \in L_2(\Omega)$ and $g \in L_2(\Gamma_N)$ guarantee a unique solution $u$. Let

$$
V = \{ v \in H^1(\Omega) : v|_{\Gamma_D} = 0 \}. \quad (7.2.2)
$$

The weak form for the model problem (7.2.1) is to find $u \in V$ such that

$$
\mathcal{B}(u, v) = \mathcal{L}(v) \quad \forall v \in V, \quad (7.2.3)
$$

where

$$
\mathcal{B}(u, v) = \int_{\Omega} (\nabla u \cdot \nabla v + cuv) \, dx \quad (7.2.4)
$$

and

$$
\mathcal{L}(v) = \int_{\Omega} f v \, dx + \int_{\Gamma_N} g v \, ds. \quad (7.2.5)
$$

Let $\mathcal{H}_h$ be a family of triangulations of $\Omega$ with straight edges. We assume that the family is regular, in other words, the minimal angle of all the triangles is bounded from below by a positive constant. However, the meshes are not assumed to be quasiuniform. We define the finite element space as

$$
V_h^K = \{ v : v \circ F_K^{-1} \in \mathcal{P}_p(\mathcal{R}) \}, \quad V_h = \{ v \in H^1(\Omega) : v|_K \in V_h^K, K \in \mathcal{H}_h \}, \quad (7.2.6)
$$
where $F_K$ is the mapping function for the element $K$ which maps the the reference element $R$ to element $K$ and $\mathcal{P}_p(R)$ denotes the set of polynomials of degree up to $p$ over $R$. We assume that $v_h|_{\Gamma_0} = 0, \forall v_h \in V_h$. The finite element solution $u_h$ satisfies:

Find $u_h \in V_h$ such that

$$\mathcal{B}(u_h, v_h) = \mathcal{L}(v_h), \quad v_h \in V_h. \quad (7.2.7)$$

### 7.2.1 Reference elements

In this work we employ the spectral/hp element spaces defined in [58] for $\mathcal{P}_p(R)$.

For a triangular reference domain, see figure 7.1, a hierarchical set of basis is given as

- **Vertex A:** \( \frac{1-\eta_1}{2} \frac{1-\eta_2}{2} \)
- **Vertex B:** \( \frac{1+\eta_1}{2} \frac{1-\eta_2}{2} \)
- **Vertex C:** \( \frac{1+\eta_2}{2} \)
- **Edge AB:** \( \frac{1-\eta_1}{2} \frac{1+\eta_1}{2} P_{p-1}^{1,1}(\eta_1) (\frac{1-\eta_2}{2})^{p+1} \quad (0 < p < P_1) \)
- **Edge AC:** \( \frac{1-\eta_1}{2} \frac{1+\eta_2}{2} P_{q-1}^{1,1}(\eta_2) \quad (0 < q < P_2) \)
- **Edge BC:** \( \frac{1+\eta_1}{2} \frac{1+\eta_2}{2} P_{q-1}^{1,1}(\eta_2) \quad (0 < q < P_2) \)
- **Interior:** \( \frac{1-\eta_1}{2} \frac{1+\eta_1}{2} P_{p-1}^{1,1}(\eta_1) (\frac{1-\eta_2}{2})^{p+1} \frac{1+\eta_2}{2} P_{q-1}^{p+1,1}(\eta_2) \)

\( (0 < p, q; p < P_1; p + q < P_2, P_1 \leq P_2) \)

subject to the mapping

\[ \eta_1 = \frac{1 + \xi_1}{1 - \xi_2} - 1, \quad \eta_2 = \xi_2, \]

where $P_p^{\alpha,\beta}$ denotes $p$-th order one-dimensional Jacobi polynomials on $(-1,1)$ with indexes $\alpha$ and $\beta$. 

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
7.3 An element-residual error estimator

Let $e = u - u_h$ denote the error for the finite element solution $u_h$. The error $e$ satisfies the following boundary value problem in element $K$

$$\mathcal{B}_K(e, v) = (f, v)_K - \mathcal{B}_K(u_h, v) + \int_{\partial K} v (n_K \cdot \nabla u) ds, \quad \forall v \in V(K), \quad (7.3.1)$$

where

$$V(K) = \{ v : v = w|_K, \quad \forall w \in V \}$$

$$\mathcal{B}_K(u, v) = \int_K (\nabla u \cdot \nabla v + cuv) dx$$

$$ (f, v)_K = \int_K f v dx.$$

We note here that the flux on $\partial K$ (last term in equation (7.3.1)) is, in general, unknown and needs to be constructed. Let $g_K \approx n_K \cdot \nabla u|_K$ be the approximation of the true flux on the boundary of element $K$. The equilibration condition

$$(f, v)_K - \mathcal{B}_K(u_h, v) + \int_{\partial K} g_K v ds = 0, \quad \forall v \in V_h^K \quad (7.3.2)$$

and the consistency condition

$$\sum_{K \in \mathcal{T}_h} \int_{\partial K} g_K v ds = \int_{\Gamma_N} g v ds, \quad \forall v \in V \quad (7.3.3)$$

Figure 7.1: Mapping between triangular elements and the reference element.
are used in [5] to obtain the equilibrated fluxes $g_K$. Once the equilibrated fluxes $g_K$ are obtained, we can define a well-posed local pure-Neumann problem in the local space $V(K)$

$$\mathcal{B}_K(\phi_K, v) = (f, v)_K - \mathcal{B}_K(u_h, v) + \int_{\partial K} v g_K \, ds, \quad \forall v \in V(K),$$

(7.3.4)

which has a unique solution $\phi_K$. Let $\| \cdot \|$ denote the global energy norm and $\| \cdot \|_K$ the local one. Based on $\{\phi_K\}$, the upper bound of $\|e\|$ can be guaranteed by the following theorem.

**Theorem 7.3.1** (\cite{8}). Suppose that $g_K$ are a set of equilibrated fluxes. The global error residual can be decomposed into local contributions

$$\mathcal{B}(e, v) = \sum_{K \in \mathcal{T}_h} \mathcal{B}_K(\phi_K, v), \quad v \in V,$$

where $\phi_K \in V(K)$ is the solution of the local problem (7.3.4). The global error in the finite element approximation may be bounded by

$$\|e\|^2 \leq \sum_{K \in \mathcal{T}_h} \|\phi_K\|^2_K.$$  

(7.3.5)

### 7.3.1 Revisiting the Ainsworth-Oden (A-O) procedure for flux splitting

For any $K \in \mathcal{T}_h$ we denote by $E(K)$ and $N(K)$ the sets of its edges and vertices, respectively. Let $E_h = \bigcup_{K \in \mathcal{T}_h} E(K)$ and $N_h = \bigcup_{K \in \mathcal{T}_h} N(K)$. We assume that all vertices in $N_h$ are regular nodes without hanging nodes. For each $x_n \in N_h$ we define an element patch $P_n$ as

$$P_n = \{K : x_n \in N(K), \quad \forall K \in \mathcal{T}_h\}$$

and an edge patch $E_n$ as

$$E_n = \{\gamma : x_n \in N(E), \quad \forall E \in \mathcal{E}_h\}. $$
Let $\mathcal{E}_n = \mathcal{E}_n^I \cup \mathcal{E}_n^N \cup \mathcal{E}_n^D$, where

$$
\mathcal{E}_n^N = \{ \gamma \in \mathcal{E}_n : \gamma \subset \partial K \cap \Gamma_N, K \in \mathcal{P}_n \}
$$

$$
\mathcal{E}_n^D = \{ \gamma \in \mathcal{E}_n : \gamma \subset \partial K \cap \Gamma_D, K \in \mathcal{P}_n \}
$$

and $\mathcal{E}_n' = \mathcal{E}_n \setminus \mathcal{E}_n^N \setminus \mathcal{E}_n^D$. We note here that there exists one-to-one correspondence between $x_n$ and $\mathcal{P}_n$.

We here focus on the fluxes for the linear shape functions $\{\theta_n\}$ corresponding to interior vertices. In figure 7.2, a patch associated with an interior vertex is shown.

![Figure 7.2: The patches $\mathcal{P}_n$ and $\mathcal{E}_n$ of elements and edges influenced by the basis function $\theta_n$ associated with an interior vertex located at $x_n$.](image)

Let $\mu_{K,n}^\gamma = \int_{\gamma} \theta_n \, ds$ denote the flux moment on edge $\gamma$. Then the equilibration (7.3.2) and consistency condition (7.3.3) are equivalent to the following two lemmas [7, 8], respectively.

**Lemma 7.3.2.** Let $\{\theta_n : n \in \mathcal{N}(K)\}$ be a basis for the local finite element space $V_h^K$ on element $K$.

The equilibration condition (7.3.2) holds on $K$ if and only if

$$
\sum_{\gamma \subset \partial K} \mu_{K,n}^\gamma = \Delta_K(\theta_n) \quad \forall n \in \mathcal{N}(K),
$$

where
with $\Delta_K(v)$ being the element residual

$$
\Delta_K(v) = B_K(u_h, v) - (f, v)_K.
$$

(7.3.7)

**Lemma 7.3.3.** The consistency condition (7.3.3) holds if and only if the fluxes $g_K$ satisfy

$$
g_K = g \text{ on } \partial K \cap \Gamma_N
$$

$$
g_K + g_{K'} = 0 \text{ on } \partial K \cap \partial K',
$$

which yields that

$$
\mu_{K,n}^\gamma = \int_{\gamma} g \theta_n ds, \quad \gamma = \partial K \cap \Gamma_N
$$

$$
\mu_{K,n}^\gamma + \mu_{K',n}^\gamma = 0, \quad \gamma = \partial K \cap \partial K'.
$$

(7.3.8)

Let $\tilde{\mu}_{K,n}^\gamma = \int_{\gamma} \theta_n (n_K \cdot \nabla u_h|_{\partial K}) ds$ denote the approximate flux moments. $\mu_{K,n}^\gamma$ related to the vertex modes are selected by the following problem [8]:

Minimize

$$
\frac{1}{2} \sum_{K \in P_n} \sum_{\gamma \in \partial K} (\mu_{K,n}^\gamma - \mu_{K,n}^\gamma)^2
$$

subject to

$$
\sum_{\gamma \in \partial K} \mu_{K,n}^\gamma = \Delta_K(\theta_n) \text{ for all } K \in P_n
$$

with

$$
\mu_{K,n}^\gamma = \int_{\gamma} g \theta_n ds \text{ on } \gamma \in \mathcal{E}_n^N \cap \partial K
$$

and

$$
\mu_{K,n}^\gamma + \mu_{K',n}^\gamma = 0 \text{ on } \gamma \in \mathcal{E}_n^s \cap \partial K \cap \partial K'.
$$

(7.3.9)

The problem (7.3.9) leads to the following conditions satisfied by the Lagrange multipliers $\{\sigma_{K,n} :$
\( K \in \mathcal{P}_n \) (see [8] for more details):

\[
\sum_{K', \partial K \cap \partial K' \in \mathcal{E}_h^t} (\sigma_{K,n} - \sigma_{K',n}) + 2 \sum_{\gamma \in \partial K \cap \mathcal{E}_h^P} \sigma_{K,n} = \bar{\Delta}_K(\theta_n), \forall K \in \mathcal{P}_n, \tag{7.3.10}
\]

where \( \bar{\Delta}_K(\theta_n) \) are modified element residuals

\[
\bar{\Delta}_K(\theta_n) = 2 \left[ \mathcal{B}_K(u_h, \theta_n) - (f, \theta_n)_K - \int_{\partial K} (\frac{\partial u_K}{\partial n_K}) \theta_n ds \right] \tag{7.3.11}
\]

and the average flux is defined as

\[
\frac{\partial u_K}{\partial n_K} = \begin{cases}
\frac{1}{2} \left( \frac{\partial u_n}{\partial n_K} |_K + \frac{\partial u_n}{\partial n_K} |_{K'} \right) & \gamma \in \mathcal{E}_n^t \cap \partial K \cap \partial K' \\
g & \gamma \in \mathcal{E}_n^N \cap \partial K \\
\frac{\partial u_n}{\partial n_K} & \gamma \in \mathcal{E}_n^D \cap \partial K.
\end{cases} \tag{7.3.12}
\]

Then the equilibrated fluxes can be expressed as

\[
\mu_{K,n} = \begin{cases}
\frac{1}{2}(\sigma_{K,n} - \sigma_{K',n}) + \frac{1}{2}(\bar{\mu}_{K,n} - \bar{\mu}_{K',n}) & \gamma \in \mathcal{E}_n^t \cap \partial K \cap \partial K' \\
\int_{\gamma} g \theta_n ds & \gamma \in \mathcal{E}_n^N \cap \partial K \\
\sigma_{K,n} + \bar{\mu}_{K,n} & \gamma \in \mathcal{E}_n^D \cap \partial K.
\end{cases} \tag{7.3.13}
\]

### 7.3.2 Explicit equilibrated fluxes

We denote the linear system (7.3.10) in matrix form as

\[
T_n \sigma_n = \bar{\Delta}_n, \tag{7.3.14}
\]

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
For an interior vertex shown in figure 7.2, $T_n$ takes the form

$$T_n = \begin{bmatrix} 2 & -1 & \cdots & -1 \\ -1 & 2 & -1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & -1 & 2 & -1 \\ -1 & \cdots & -1 & 2 \end{bmatrix}, \quad (7.3.15)$$

where $T_n$ is singular with a one-dimensional kernel $\dim(\ker(T_n)) = 1$. We note that $T$ is a circulant matrix corresponding to a finite difference scheme subject to periodic boundary conditions. The eigenvalues and eigenvectors of such a matrix can be found in [51]. We are interested in the explicit formulas of $(\sigma_{n,i} - \sigma_{n,i+1})$. It is easy to verify that the solvability condition $\sum_{K \in P_n} \hat{\Delta}_K = 0$ is satisfied due to the Galerkin projection. In [7] a procedure was proposed to pick up a solution in a least-squares sense. However, such a procedure would give rise to more cost compared to classical element residual estimators. Here we present the equilibrated fluxes in an explicit form for the two-dimensional case.

**Lemma 7.3.4.** For the patch shown in figure 7.2, the following relations satisfy equation (7.3.14)

$$\sigma_{n,i} - \sigma_{n,i+1} = \frac{\hat{\Delta}_{n,i} - \sum_{j=2}^{N-1} (N - j)\hat{\Delta}_{n,\text{mod}(j+i-1,N)}}{N}, \quad (7.3.16)$$

where $i = 1, 2, \cdots, N - 1$ and

$$\text{mod}(m, N) = \begin{cases} \text{mod}(m + kN, N) & \text{if } \text{mod}(m + kN, N) \neq 0; \\ N & \text{otherwise,} \end{cases}$$

with $k$ being any integer to make $m + kN > 0$. Furthermore, the relations (7.3.16) are uniquely determined by equation (7.3.14). From now on, we replace the symbol $K_i$ with $i$ in the subscripts of $\sigma_{K_i,n}$ and $\hat{\Delta}_{K_i,n}$ for convenience.
Proof. We first verify that the relations (7.3.16) satisfy the system (7.3.14). It is easy to obtain that

\[
-(\sigma_{n,i} - \sigma_{n,i+1}) + (\sigma_{n,i+1} - \sigma_{n,i+2}) = \frac{-\hat{\Delta}_{n,i} + \sum_{j=2}^{N-1} (N - j) \hat{\Delta}_{n,\text{mod}(j+i-1,N)}}{N} + \frac{\hat{\Delta}_{n,i+1} - \sum_{j=2}^{N-1} (N - j) \hat{\Delta}_{n,\text{mod}(j+i,N)}}{N} = \frac{N\hat{\Delta}_{n,i+1} - \sum_{j=1}^{N-1} \Delta_{n,i} \mod^j}{N} = \hat{\Delta}_{n,i+1}.
\]

Due to symmetry, we know that all equations in system (7.3.14) are satisfied. We next show that the relations (7.3.16) are uniquely determined by the system (7.3.14). Let

\[
\hat{\sigma}_n = A\sigma_n,
\]

where

\[
A = \begin{bmatrix}
1 & -1 & 0 & \cdots & 0 \\
0 & 1 & -1 & \cdots & 0 \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
0 & \cdots & 1 & -1 \\
0 & \cdots & 0 & 1
\end{bmatrix}, \quad \sigma_n = \begin{bmatrix}
\sigma_{n,1} - \sigma_{n,2} \\
\sigma_{n,2} - \sigma_{n,3} \\
\vdots \\
\sigma_{N-1} - \sigma_N \\
\sigma_N
\end{bmatrix}
\]

It is easy to obtain the inverse of \(A\) as

\[
A^{-1} = \begin{bmatrix}
1 & 1 & 1 & \cdots & 1 \\
0 & 1 & 1 & \cdots & 1 \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
0 & \cdots & 1 & 1 \\
0 & \cdots & 0 & 1
\end{bmatrix}
\]
Thus, the system (7.3.14) can be rewritten as

\[
\begin{bmatrix}
2 & 1 & 1 & \cdots & 1 & 0 \\
-1 & 1 & 0 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & \cdots & -1 & 1 & 0 \\
-1 & -1 & \cdots & -2 & 0 \\
\end{bmatrix}
\begin{bmatrix}
\sigma_1 \\
\sigma_2 \\
\vdots \\
\sigma_{N-1} \\
\sigma_N \\
\end{bmatrix}
= \begin{bmatrix}
\hat{\Delta}_{n,1} \\
\hat{\Delta}_{n,2} \\
\vdots \\
\hat{\Delta}_{n,N-1} \\
\hat{\Delta}_{n,N} \\
\end{bmatrix}.
\]

Since \(\dim(\ker(T_n)) = 1\), we only need to consider the first \(N - 1\) equations, which yield a linear system related to \(\sigma_{n,i} - \sigma_{n,i+1}\)

\[
B\tilde{\sigma}_n = \begin{bmatrix}
2 & 1 & 1 & \cdots & 1 \\
-1 & 1 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & \cdots & -1 & 1 \\
\end{bmatrix}
\begin{bmatrix}
\sigma_1 \\
\sigma_2 \\
\vdots \\
\sigma_{N-1} \\
\sigma_N \\
\end{bmatrix}
= \begin{bmatrix}
\hat{\Delta}_{n,1} \\
\hat{\Delta}_{n,2} \\
\vdots \\
\hat{\Delta}_{n,N-1} \\
\end{bmatrix}.
\]

Since \(\det(B) = N\), the conclusion follows immediately. \(\square\)

To this end, we have obtained the explicit forms of equilibrated fluxes for the interior vertex.

**Two-dimensional flux splitting based on weighted target functions**

In [6] the non-uniqueness of equilibrated fluxes is removed by minimizing the value of

\[
\frac{1}{2} \sum_{K \in P_n} \sum_{\gamma \in F_n \cap \partial K} (\mu_{K,n}^\gamma - \hat{\mu}_{K,n}^\gamma)^2. \tag{7.3.17}
\]

By noting that the function \(\theta_n\) introduces a factor \(l(\gamma)\) (length of \(\gamma\)) into \(\mu_{K,n}^\gamma\), we use a weighted version of equation (7.3.17) as

\[
\frac{1}{2} \sum_{K \in P_n} \sum_{\gamma \in F_n \cap \partial K} w_\gamma^2 (\mu_{K,n}^\gamma - \hat{\mu}_{K,n}^\gamma)^2, \tag{7.3.18}
\]

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
where we add weights \( w_i \) into equation (7.3.17). In this work we investigate a weight factor \( w_i = \frac{1}{l(\gamma)} \) with \( \gamma \in \mathcal{E}_n \cap \partial K_i \). By noting that \( \int_\gamma \theta_n ds = \frac{l(\gamma) - l_{(\gamma')}}{2} \), it is clear that we are minimizing the difference between the mean values of true fluxes \( \frac{\partial u}{\partial n_K} \) and numerical fluxes \( \frac{\partial u}{\partial n_K} \) with respect to \( \theta_n \) along the edges \( \gamma \in \mathcal{E}_n \).

Following a similar procedure as in [8], we can derive the equilibrated fluxes as

\[
\tilde{\mu}_{K,n}^\gamma = \begin{cases} 
\frac{1}{2} l_{(\gamma)}^2 (\sigma_{K,n} - \sigma_{K',n}) + \frac{1}{2} (\hat{\mu}_{K,n}^\gamma - \hat{\mu}_{K',n}^\gamma) & \gamma \in \mathcal{E}_n^l \cap \partial K \cap \partial K' \\
\int_\gamma g\theta_n ds & \gamma \in \mathcal{E}_n^d \cap \partial K \\
l_{(\gamma)}^2 \sigma_{K,n} + \hat{\mu}_{K,n}^\gamma & \gamma \in \mathcal{E}_n^b \cap \partial K.
\end{cases}
\]  

(7.3.19)

where \( \sigma_{K,n} \) must satisfy

\[
\frac{1}{2} \sum_{\gamma : \gamma \cap \partial K \cap \partial K' \in \mathcal{E}_n^l} l_{(\gamma)}^2 (\sigma_{K,n} - \sigma_{K',n}) + \sum_{\gamma : \gamma \cap \partial K \cap \partial K' \in \mathcal{E}_n^b} l_{(\gamma)}^2 \sigma_{n,i} = \hat{\Delta}_K (\theta_n). 
\]

(7.3.20)

For the patch \( P_n \) corresponding to an interior vertex, \( T_n \) in equation (7.3.14) takes the form

\[
T_n = \begin{bmatrix}
l_{(\gamma_1)}^2 + l_{(\gamma_2)}^2 & -l_{(\gamma_1)}^2 & \cdots & -l_{(\gamma_1)}^2 \\
-l_{(\gamma_2)}^2 & l_{(\gamma_2)}^2 + l_{(\gamma_3)}^2 & -l_{(\gamma_2)}^2 & \cdots & 0 \\
\vdots & \ddots & \ddots & \ddots & \vdots \\
0 & \cdots & -l_{(\gamma_{N-1})}^2 & l_{(\gamma_{N-1})}^2 + l_{(\gamma_N)}^2 & -l_{(\gamma_N)}^2 \\
-l_{(\gamma_1)}^2 & \cdots & -l_{(\gamma_N)}^2 & l_{(\gamma_N)}^2 + l_{(\gamma_1)}^2 & \end{bmatrix}
\]

Using the similar idea as in the proof of lemma 7.3.4, it can be proved that:

**Lemma 7.3.5.** For the patch shown in figure 7.2, the following relations are uniquely determined by equation (7.3.20)

\[
\sigma_{n,i} - \sigma_{n,i+1} = \frac{\hat{\Delta}_{n,i} - \sum_{j=2}^{N-1} \hat{\Delta}_{n,\text{mod}(j+i-1,N)} H_{i,j}}{	ext{diag}(l_{(\gamma_1)}^2, l_{(\gamma_2)}^2, \sum_{i=1}^{N-1} l_{(\gamma_i)}^2)},
\]

(7.3.21)
where
\[ H_{i,j} = \sum_{k=1}^{N-j} \frac{l_{h_n}^2}{l_{n+k|^2}}. \] (7.3.22)

**Computational Cost**

we consider here only the cost for the first-order equilibrated fluxes. We assume that the usual type of small angle condition is satisfied and the number of elements forming the support of any linear shape function \( \theta_n \) is bounded above by \( N_{\text{patch}} \) independently of the mesh size. To obtain \( (\sigma_{n,i} - \sigma_{n,i+1}) \) a cost of order \( N_{\text{patch}} \) is needed (see lemma 7.3.4). Using the relation \( \sigma_{n,i+1} - \sigma_{n,i+2} = \Delta_n,_{i+1} + (\sigma_{n,i} - \sigma_{n,i+1}) \) recursively, we can obtain the equilibrated fluxes moments \( \mu_{n,i}^\gamma \) for each patch by a total cost of \( O(2N_{\text{patch}}) \) compared to the cost of \( O(\frac{1}{3}N_{\text{patch}}^3 + \frac{1}{2}N_{\text{patch}}^2) \) given in [7], where \( \sigma_{n,i} \) is computed numerically. Thus, the cost can be reduced significantly by using the explicit solutions of equilibrated flux moments.

**Meshes with hanging nodes**

In the adaptive mesh refinement, it is convenient to use meshes with hanging nodes. We expect to obtain the explicit forms of equilibrated fluxes for such kinds of finite element meshes. Actually it can be done easily using the idea of macro-element defined in [8].

The macro-element is defined as follows (see figure 7.3 for an example):

**Definition 7.3.6 (Macro-element).** If \( x_n \) is a hanging node, the macro-element \( K_*^n = K \) with \( K \) being an element containing the node \( x_n \); otherwise, \( K_*^n \) is the domain formed from \( K \) and elements having a hanging node in common with \( K \), but not containing the node \( x_n \) itself.

We note that there are no hanging nodes any more on the edges of macro-elements. The point is that if we define the element patch \( P_n \) and \( E_n \) using the macro-elements we can get the explicit equilibrated fluxes on edges of the macro-elements for \( \theta_n \) using the formulas in lemmas 7.3.4 and 7.3.5. Subsequently, we can follow the procedure given in [8] to obtain explicitly the equilibrated fluxes for edges with hanging nodes.
7.3.3 High-order equilibration conditions

We now consider the equilibration condition for the edge modes. Due to the Galerkin projection, the high-order equilibration actually looks trivial, see [5, 8]. We here summarize the results as

**Lemma 7.3.7.** Given an element $K$ and edge modes $\theta_n$, the equilibrated flux moments can be expressed as

\[
\mu_{K,n}^\gamma = \Delta_K, \quad \gamma \in \partial K \cap (\mathcal{E}_n^l \cup \mathcal{E}_n^D) ,
\]

\[
\mu_{K,n}^\gamma = \int_\gamma g\theta_n ds, \quad \gamma \in \partial K \cap \mathcal{E}_n^N ,
\]

where

\[
\Delta_K(\theta_n) = \mathcal{B}_K(u_h, \theta_n) - (f, \theta_n)_K .
\]

7.4 Non-homogeneous Dirichlet boundary conditions

In last section we assume that the Dirichlet boundary conditions are homogeneous. However, the Dirichlet boundary conditions are, in general, non-homogeneous, which implies that the approximation error of Dirichlet boundary conditions should be taken into account for the a posteriori error estimate. In fact, such an error can affect the effectiveness of a posteriori error estimators significantly in some cases, e.g., linear finite elements. In this work, we use the method given in [4] to decompose the error $e$ into two parts as

\[
e = e_G + e_D
\]
where the Galerkin error, $e_G \in H^1_0(\Omega)$, satisfies

$$\mathcal{B}(e_G, v) = \mathcal{B}(e, v) \quad \forall v \in H^1_0(\Omega)$$ (7.4.1)

and the Dirichlet error, $e_D \in H^1(\Omega)$, satisfies

$$\mathcal{B}(e_D, v) = \mathcal{B}(e, v) - \mathcal{B}(e_G, v) = 0 \quad \forall v \in H^1_0(\Omega)$$

$$e_D = q - q_h \quad \text{on } \Gamma_D,$$ (7.4.2)

where $q$ and $q_h$ are the true and approximated Dirichlet boundary conditions. Following these definitions, it can be shown [4] that in the energy norm

$$\|e\|^2 = \|e_G\|^2 + \|e_D\|^2.$$ (7.4.3)

$\|e_G\|$ can be bounded using the equilibrated residual method described in last section; $\|e_D\|$ can be readily bounded by

$$\|e_D\|^2 \leq \sum_{K \in \mathcal{T}_h} \|\psi_K\|^2,$$ (7.4.4)

where $\psi_K$ is the solution of the local problems

$$\mathcal{B}_K(\psi_K, v) = 0 \quad \forall v \in H^1_0(K)$$

$$\psi_K = \begin{cases} 
  e_D & \text{on } \partial K \cap \Gamma_D \\
  0 & \text{on } \partial K \setminus \Gamma_D.
\end{cases}$$ (7.4.5)

Since such local problems are solved only in the Dirichlet boundary elements, the cost for $e_D$ is usually much smaller than that for $e_G$.

### 7.5 Numerical examples

In this section we demonstrate numerically the influence of weights we put into the A-O procedure.
7.5.1 Efficiency for smooth solutions

We first consider a symmetric elliptic operator with a smooth solution [5]:

\[-\Delta u + u = 0 \quad \text{in} \quad \Omega \quad (7.5.1)\]

subject to the boundary conditions

\[
\begin{align*}
    u(0, y) &= \left(e^{-\sqrt{1+4\pi^2}+1}\right)\sin 2\pi y, \quad 0 < y < 1/2, \\
    u(x, 0) &= u(0, 1/2) = 0, \quad 0 < x < 1/2, \\
    \frac{\partial u}{\partial n} &= 0, \quad x = 1/2, \quad 0 < y < 1/2.
\end{align*}
\]

The true solution for this problem is

\[
u(x, y) = \left(\exp \left((x - 1)\sqrt{1+4\pi^2}\right) + \exp \left(-x\sqrt{1+4\pi^2}\right)\right)\sin 2\pi y. \quad (7.5.3)\]

We study this problem numerically using the periodic cell patterns [17] in figure 7.4, which consist of triangle elements. The computation domain \(\Omega\) is decomposed by one of these cells periodically.

We control the deformation of triangle elements by the ratio \(a/b\). In this work we basically check two cases: \(a/b = 1\) and \(a/b = 4\). We denote the decomposition of the computation domain \(\Omega\) by a cell matrix \((M, N)\), where \(M\) is the number of cells in the \(y\)-direction and \(N\) is the number of cells in the \(x\)-direction. Thus, the total elements number is \(N_cM\), where \(N_c\) is the number of elements in a certain cell pattern. Let \(\|\|\|e_h\|\| / \|e\|\|\) be the effectivity index, which is the ratio of estimated error over the true error. We use \(E_{k,s}\) to denote the effectivity index from the weighted fluxes and \(E_{a,s}\) from the A-O procedure, where \(s\) indicates the cell patterns from (a) to (f).

For linear elements we enrich the local finite element space \(V^K_h(v)\) by polynomial order 1 or 2. In tables 7.1-7.4, we show the effectivity indices of the two a posteriori error estimators for different linear triangulations \(\mathcal{F}_h\) given by cell patterns (a)-(f). We observe the following:

- Both a posteriori error estimators can give asymptotically upper bounds of the error \(\|\|e\|\|\).
• When \( a/b = 1 \), both a posteriori error estimators show similar performance for all cell patterns (a)-(f); when \( a/b = 4 \), the a posteriori error estimator with weights appears more robust than that without weights, especially for the large deformation, e.g., cell patterns (b), (d) and (e).

The first observation is due to the theorem 7.3.1. The second one can be explained using the target function (7.3.18). If we use the factor \( \frac{1}{|\gamma|} \) as a weight function, \( \mu_{K,n} \) and \( \tilde{\mu}_{K,n} \) will have larger difference on a longer edge, which is reasonable since the linear function along a longer edge usually yields a larger approximation error. Thus, weights in equation (7.3.18) may have a large influence on the robustness of the a posteriori error estimator for the linear elements.

For spectral/hp elements we enrich the local finite element space \( V_h^K(u) \) by polynomial order 2. In tables 7.5 and 7.6, we show the effectivity indices for the spectral/hp elements. It can be seen that the two posteriori error estimators show similar performance for all different cell patterns and different deformation, which implies that for spectral/hp elements the flux \( \mu_{K,n}^{\gamma} \) will be mainly determined by the numerical approximation \( \frac{1}{2}(\tilde{\mu}_{K,n} - \tilde{\mu}_{K',n}) \) and the equilibration parts \( \frac{1}{2}(\sigma_{K,n} - \sigma_{K',n}) \) or \( \frac{1}{2}I_1^2(\sigma_{K,n} - \sigma_{K',n}) \) becomes less important compared to the linear elements. Thus, the a posteriori error estimators are not sensitive to the weights in equation (7.3.18) due to the fast p-convergence.
in spectral/hp elements.

To give an overall measurement of the performance of the two a posteriori error estimators, we show some statistics in table 7.7. We compute the mean and standard deviation of the effectivity index for the deformation $a/b = 1$ and $a/b = 4$. It can be seen that for the linear elements both mean and standard deviation of $E_k$ are smaller than those of $E_0$, which implies that the weights in equation (7.3.18) improve effectively the robustness of a posteriori error estimator. For the spectral/hp elements, $E_k$ and $E_0$ show similar mean and standard deviation.

Table 7.1: Effectivity indices for cell pattern (a),(b) and (c). Linear elements and increment of polynomial order 1 are considered.

<table>
<thead>
<tr>
<th>$(M,N)$</th>
<th>$E_{k,a}$</th>
<th>$E_{o,a}$</th>
<th>$E_{k,b}$</th>
<th>$E_{o,b}$</th>
<th>$E_{k,c}$</th>
<th>$E_{o,c}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1,1)</td>
<td>1.088</td>
<td>1.086</td>
<td>0.999</td>
<td>1.015</td>
<td>1.414</td>
<td>1.232</td>
</tr>
<tr>
<td>(2,2)</td>
<td>1.045</td>
<td>1.045</td>
<td>1.074</td>
<td>1.049</td>
<td>1.135</td>
<td>1.107</td>
</tr>
<tr>
<td>(3,3)</td>
<td>1.029</td>
<td>1.030</td>
<td>1.077</td>
<td>1.045</td>
<td>1.111</td>
<td>1.120</td>
</tr>
<tr>
<td>(4,4)</td>
<td>1.021</td>
<td>1.023</td>
<td>1.077</td>
<td>1.044</td>
<td>1.103</td>
<td>1.130</td>
</tr>
<tr>
<td>(4,1)</td>
<td>0.954</td>
<td>0.908</td>
<td>1.191</td>
<td>1.324</td>
<td>1.129</td>
<td>1.145</td>
</tr>
<tr>
<td>(8,2)</td>
<td>0.946</td>
<td>0.925</td>
<td>1.102</td>
<td>1.287</td>
<td>1.083</td>
<td>1.145</td>
</tr>
<tr>
<td>(12,3)</td>
<td>0.946</td>
<td>0.933</td>
<td>1.092</td>
<td>1.274</td>
<td>1.092</td>
<td>1.155</td>
</tr>
<tr>
<td>(16,4)</td>
<td>0.946</td>
<td>0.939</td>
<td>1.096</td>
<td>1.259</td>
<td>1.113</td>
<td>1.165</td>
</tr>
</tbody>
</table>

Table 7.2: Effectivity indices for cell pattern (d),(e) and (f). Linear elements and increment of polynomial order 1 are considered.

<table>
<thead>
<tr>
<th>$(M,N)$</th>
<th>$E_{k,d}$</th>
<th>$E_{o,d}$</th>
<th>$E_{k,e}$</th>
<th>$E_{o,e}$</th>
<th>$E_{k,f}$</th>
<th>$E_{o,f}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1,1)</td>
<td>1.931</td>
<td>1.859</td>
<td>1.009</td>
<td>1.010</td>
<td>1.048</td>
<td>1.045</td>
</tr>
<tr>
<td>(2,2)</td>
<td>1.309</td>
<td>1.603</td>
<td>1.031</td>
<td>1.031</td>
<td>1.030</td>
<td>1.029</td>
</tr>
<tr>
<td>(3,3)</td>
<td>1.233</td>
<td>1.514</td>
<td>1.024</td>
<td>1.024</td>
<td>1.021</td>
<td>1.021</td>
</tr>
<tr>
<td>(4,4)</td>
<td>1.196</td>
<td>1.452</td>
<td>1.020</td>
<td>1.020</td>
<td>1.017</td>
<td>1.018</td>
</tr>
<tr>
<td>(4,1)</td>
<td>0.965</td>
<td>2.081</td>
<td>1.214</td>
<td>1.379</td>
<td>1.148</td>
<td>1.032</td>
</tr>
<tr>
<td>(8,2)</td>
<td>1.016</td>
<td>1.812</td>
<td>1.275</td>
<td>1.495</td>
<td>1.089</td>
<td>1.012</td>
</tr>
<tr>
<td>(12,3)</td>
<td>1.028</td>
<td>1.679</td>
<td>1.327</td>
<td>1.537</td>
<td>1.068</td>
<td>1.008</td>
</tr>
<tr>
<td>(16,4)</td>
<td>1.030</td>
<td>1.586</td>
<td>1.362</td>
<td>1.555</td>
<td>1.057</td>
<td>1.007</td>
</tr>
</tbody>
</table>

7.5.2 Singular solution: the Motz problem

We now consider the Motz problem (see figure 7.5):

$$\Delta u = 0 \text{ in } \Omega$$

(7.5.4)
Table 7.3: Effectivity indices for cell pattern (a),(b) and (c). Linear elements and increment of polynomial order 2 are considered.

<table>
<thead>
<tr>
<th>(M,N)</th>
<th>$E_{k,a}$</th>
<th>$E_{o,a}$</th>
<th>$E_{k,b}$</th>
<th>$E_{o,b}$</th>
<th>$E_{k,c}$</th>
<th>$E_{o,c}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1,1)</td>
<td>1.192</td>
<td>1.191</td>
<td>1.057</td>
<td>1.056</td>
<td>1.553</td>
<td>1.294</td>
</tr>
<tr>
<td>(2,2)</td>
<td>1.163</td>
<td>1.163</td>
<td>1.176</td>
<td>1.160</td>
<td>1.339</td>
<td>1.252</td>
</tr>
<tr>
<td>(3,3)</td>
<td>1.148</td>
<td>1.149</td>
<td>1.201</td>
<td>1.175</td>
<td>1.326</td>
<td>1.277</td>
</tr>
<tr>
<td>(4,4)</td>
<td>1.141</td>
<td>1.144</td>
<td>1.219</td>
<td>1.185</td>
<td>1.324</td>
<td>1.294</td>
</tr>
<tr>
<td>(4,1)</td>
<td>1.126</td>
<td>1.106</td>
<td>1.398</td>
<td>1.590</td>
<td>1.608</td>
<td>2.448</td>
</tr>
<tr>
<td>(8,2)</td>
<td>1.183</td>
<td>1.180</td>
<td>1.325</td>
<td>1.573</td>
<td>1.493</td>
<td>2.509</td>
</tr>
<tr>
<td>(12,3)</td>
<td>1.205</td>
<td>1.205</td>
<td>1.333</td>
<td>1.582</td>
<td>1.468</td>
<td>2.502</td>
</tr>
<tr>
<td>(16,4)</td>
<td>1.215</td>
<td>1.218</td>
<td>1.350</td>
<td>1.584</td>
<td>1.473</td>
<td>2.488</td>
</tr>
</tbody>
</table>

Table 7.4: Effectivity indices for cell pattern (d),(e) and (f). Linear elements and increment of polynomial order 2 are considered.

<table>
<thead>
<tr>
<th>(M,N)</th>
<th>$E_{k,d}$</th>
<th>$E_{o,d}$</th>
<th>$E_{k,e}$</th>
<th>$E_{o,e}$</th>
<th>$E_{k,f}$</th>
<th>$E_{o,f}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1,1)</td>
<td>2.102</td>
<td>1.913</td>
<td>1.054</td>
<td>1.055</td>
<td>1.110</td>
<td>1.112</td>
</tr>
<tr>
<td>(2,2)</td>
<td>1.512</td>
<td>1.758</td>
<td>1.120</td>
<td>1.119</td>
<td>1.107</td>
<td>1.107</td>
</tr>
<tr>
<td>(3,3)</td>
<td>1.436</td>
<td>1.683</td>
<td>1.120</td>
<td>1.119</td>
<td>1.115</td>
<td>1.113</td>
</tr>
<tr>
<td>(4,4)</td>
<td>1.403</td>
<td>1.630</td>
<td>1.121</td>
<td>1.119</td>
<td>1.120</td>
<td>1.117</td>
</tr>
<tr>
<td>(4,1)</td>
<td>1.712</td>
<td>3.070</td>
<td>1.365</td>
<td>1.516</td>
<td>1.301</td>
<td>1.144</td>
</tr>
<tr>
<td>(8,2)</td>
<td>1.764</td>
<td>2.814</td>
<td>1.414</td>
<td>1.624</td>
<td>1.256</td>
<td>1.143</td>
</tr>
<tr>
<td>(12,3)</td>
<td>1.763</td>
<td>2.733</td>
<td>1.460</td>
<td>1.662</td>
<td>1.239</td>
<td>1.148</td>
</tr>
<tr>
<td>(16,4)</td>
<td>1.761</td>
<td>2.680</td>
<td>1.491</td>
<td>1.679</td>
<td>1.230</td>
<td>1.152</td>
</tr>
</tbody>
</table>

Table 7.5: Effectivity indices for cell pattern (a),(b) and (c). Spectral/hp elements and increment of polynomial order 2 are considered. Due to p-convergence, sixth-order polynomials are used for pattern (a) while eighth-order polynomials are used for pattern (b) and (c).

<table>
<thead>
<tr>
<th>(M,N)</th>
<th>$E_{k,a}$</th>
<th>$E_{o,a}$</th>
<th>$E_{k,b}$</th>
<th>$E_{o,b}$</th>
<th>$E_{k,c}$</th>
<th>$E_{o,c}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1,1)</td>
<td>1.099</td>
<td>0.991</td>
<td>1.097</td>
<td>1.013</td>
<td>1.037</td>
<td>0.988</td>
</tr>
<tr>
<td>(2,2)</td>
<td>1.039</td>
<td>1.030</td>
<td>1.067</td>
<td>1.043</td>
<td>0.996</td>
<td>0.999</td>
</tr>
<tr>
<td>(3,3)</td>
<td>1.047</td>
<td>1.040</td>
<td>1.064</td>
<td>1.035</td>
<td>0.999</td>
<td>1.003</td>
</tr>
<tr>
<td>(4,4)</td>
<td>1.050</td>
<td>1.043</td>
<td>1.062</td>
<td>1.033</td>
<td>1.002</td>
<td>1.006</td>
</tr>
<tr>
<td>(4,1)</td>
<td>0.925</td>
<td>0.920</td>
<td>1.038</td>
<td>1.042</td>
<td>1.172</td>
<td>1.148</td>
</tr>
<tr>
<td>(8,2)</td>
<td>0.951</td>
<td>0.927</td>
<td>1.080</td>
<td>1.039</td>
<td>1.215</td>
<td>1.237</td>
</tr>
<tr>
<td>(12,3)</td>
<td>0.964</td>
<td>0.932</td>
<td>1.099</td>
<td>1.035</td>
<td>1.227</td>
<td>1.259</td>
</tr>
<tr>
<td>(16,4)</td>
<td>0.972</td>
<td>0.936</td>
<td>1.112</td>
<td>1.032</td>
<td>1.232</td>
<td>1.271</td>
</tr>
</tbody>
</table>

Table 7.6: Effectivity indices for cell pattern (d),(e) and (f). Spectral/hp elements and increment of polynomial order 2 are considered. Due to p-convergence, sixth-order polynomials are used for pattern (e) and (f) while eighth-order polynomials are used for pattern (d).

<table>
<thead>
<tr>
<th>(M,N)</th>
<th>$E_{k,d}$</th>
<th>$E_{o,d}$</th>
<th>$E_{k,e}$</th>
<th>$E_{o,e}$</th>
<th>$E_{k,f}$</th>
<th>$E_{o,f}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1,1)</td>
<td>0.998</td>
<td>1.006</td>
<td>1.032</td>
<td>1.008</td>
<td>1.062</td>
<td>1.023</td>
</tr>
<tr>
<td>(2,2)</td>
<td>0.988</td>
<td>0.991</td>
<td>1.020</td>
<td>1.017</td>
<td>1.061</td>
<td>1.035</td>
</tr>
<tr>
<td>(3,3)</td>
<td>0.984</td>
<td>0.989</td>
<td>1.019</td>
<td>1.017</td>
<td>1.069</td>
<td>1.039</td>
</tr>
<tr>
<td>(4,4)</td>
<td>0.982</td>
<td>0.988</td>
<td>1.019</td>
<td>1.018</td>
<td>1.059</td>
<td>1.041</td>
</tr>
<tr>
<td>(4,1)</td>
<td>1.118</td>
<td>1.245</td>
<td>1.173</td>
<td>1.075</td>
<td>1.150</td>
<td>1.070</td>
</tr>
<tr>
<td>(8,2)</td>
<td>1.157</td>
<td>1.245</td>
<td>1.190</td>
<td>1.083</td>
<td>1.157</td>
<td>1.084</td>
</tr>
<tr>
<td>(12,3)</td>
<td>1.168</td>
<td>1.238</td>
<td>1.191</td>
<td>1.086</td>
<td>1.157</td>
<td>1.090</td>
</tr>
<tr>
<td>(16,4)</td>
<td>1.176</td>
<td>1.231</td>
<td>1.190</td>
<td>1.087</td>
<td>1.155</td>
<td>1.093</td>
</tr>
</tbody>
</table>
Table 7.7: Overall performance of the two a posteriori error estimators for the smooth problem.

| Increment 1 for $V_h^K(v)$ (Linear elements) | | Increment 2 for $V_h^K(v)$ (Linear elements) | | Increment 2 for $V_h^K(v)$ (Spectral/hp elements) |
|---|---|---|---|
| | $a/b = 1$ | $a/b = 4$ | $a/b = 1$ | $a/b = 4$ | $a/b = 1$ | $a/b = 4$ |
| $E_k$ | $E_0$ | $E_k$ | $E_0$ | $E_k$ | $E_0$ | $E_k$ | $E_0$ |
| Mean | 1.127 | 1.148 | 1.095 | 1.277 | 1.257 | 1.258 | 1.414 | 1.806 | 1.031 | 1.017 | 1.124 | 1.100 |
| Std. | 0.200 | 0.225 | 0.115 | 0.311 | 0.229 | 0.237 | 0.193 | 0.652 | 0.032 | 0.020 | 0.090 | 0.112 |

subject to the boundary conditions

\[
\begin{align*}
\frac{\partial u}{\partial n} &= 0, \quad 0 < r < 1, \\
\frac{\partial u}{\partial \theta} &= 0, \quad 0 < r < 1, \quad \theta = 0.
\end{align*}
\]  (7.5.5)

On the rest of the boundary $\partial \Omega$ we use the exact solution

\[
u(r, \theta) = r^{1/2} \cos\left(\frac{1}{2} \theta\right)
\]  (7.5.6)

as the Dirichlet boundary condition. For this problem there exist a singular point at the origin.

\[
u(r, \theta) = r^{1/2} \cos\left(\frac{1}{2} \theta\right)
\]

Figure 7.5: Geometry and boundary conditions for the Motz problem.
To study this problem we focus on the cell pattern (b) in figure 7.4. We set \( B_4C/B_4B_2 = 0.9 \) and \( V_1B_4/V_1V_4 = 0.2 \). Let \( V_1B_1/V_1V_2 = a \). We change the value of \( a \) to control the deformation of the elements. The domain \( \Omega \) is decomposed by a cell matrix \((2, 8)\), which yields 128 elements. It was shown in [5] that for such a singular problem a large increment of polynomial order in the local finite element space \( V^K_\Omega \) is necessary to get good error estimates. In this work, we consider an increment up to 4.

In figure 7.6 we plot the effectivity indices \( E_k \) and \( E_o \) versus the value of \( a \) for linear and quadratic elements with different increments in the local finite element space \( V^K_\Omega \). It is seen that for the linear elements \( E_o \) increases as \( a \) decreases while \( E_k \) remains almost the same (less than 2). For an increment of 4, \( E_o > 10 \) for \( a = 0.001 \), which means that \( E_o \) overestimates the error by an order. For this case, the weights improve the efficiency of the a posteriori error estimator significantly. Similar behavior is observed for the quadratic elements. However, for this case \( E_o \) increases much slower as \( a \) decreases, which implies that if the numerical boundary fluxes are more accurate the error estimator will be less sensitive to the equilibration of boundary fluxes. We note that for this case an increment of 4 is necessary for \( E_k \) to give an upper bound of the true error.

![Figure 7.6: Effectivity index versus the deformation of linear elements for the Motz problem. Left: \( p = 1 \); Right: \( p = 2 \)](image)

We study next the dependence of the a posteriori error estimators on the meshes. We set \( B_4C/B_4B_2 = 0.9, V_1B_4/V_1V_4 = 0.2 \) and \( a = 0.001 \). The effectivity indices for different meshes are shown in table 7.8. It is seen that the weights can, in general, improve the efficiency of the a posteriori error estimator for a large deformation, especially in linear elements.
Table 7.8: Effectivity indices for different meshes when \( B_4 C / B_4 B_2 = 0.9 \), \( V_1 B_4 / V_1 V_4 = 0.2 \) and \( V_1 B_1 / V_1 V_2 = 0.001 \). The increments are 2 and 4 for \( p = 1 \) and \( p = 2 \), respectively, in the local finite element space \( V_h^k \).

<table>
<thead>
<tr>
<th>Mesh</th>
<th>( p = 1 )</th>
<th>( p = 2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( E_k )</td>
<td>( E_o )</td>
</tr>
<tr>
<td>(a)</td>
<td>1.453</td>
<td>1.729</td>
</tr>
<tr>
<td>(b)</td>
<td>1.216</td>
<td>2.381</td>
</tr>
<tr>
<td>(c)</td>
<td>1.283</td>
<td>3.135</td>
</tr>
<tr>
<td>(d)</td>
<td>1.278</td>
<td>2.495</td>
</tr>
<tr>
<td>(e)</td>
<td>1.251</td>
<td>5.331</td>
</tr>
<tr>
<td>(f)</td>
<td>1.307</td>
<td>3.193</td>
</tr>
</tbody>
</table>

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Chapter 8

Polynomial chaos methods and experimental data

In this Chapter, we propose a general strategy to deal with general random inputs using the polynomial chaos methods. In the first section, we overview the simulations of non-Gaussian random processes. In the second section we present a general strategy to deal with general random inputs based on the K-L expansion and the polynomial chaos methods. We assume that a simulator for the random inputs is available, from which we can obtain realizations easily. We first obtain the marginal PDFs of the random variables in the K-L expansion using the density estimation technique. Based on the i.i.d. assumption we define an auxiliary PDF and implement the polynomial chaos method. Finally, we compute the statistics using the joint PDF instead of the auxiliary one. We can also use the obtained polynomial chaos solution as an effective prediction for variance reduction in the Monte Carlo methods. In the third section, we point out the error sources in the proposed methodology. In the fourth section, we review the density estimation technique based on Gaussian kernels, which can be more efficiently coupled with algorithms for numerical orthogonality. In the fifth section, we present some numerical results.
8.1 Simulations of non-Gaussian random processes

In engineering applications, random processes are often assumed to be Gaussian both for simplicity and by the virtue of the central limit theorem. However, the Gaussian assumption is not always valid since observation data exhibit distinct non-Gaussian characteristics in many cases. Thus, the simulation of non-Gaussian processes is of practical importance.

Although simulations of Gaussian processes are well established, simulations of non-Gaussian processes are still limited and in development. The main difficulty lies in the characterization of the process: unlike Gaussian processes which are determined by solely through the first- and second-order probabilistic characteristics, one must know the entire family of joint distributions, which is never available in practice.

At present, simulations of non-Gaussian processes are mainly based on memoryless nonlinear transforms of the standard Gaussian process due to the analytical tractability and availability of Gaussian simulation methods (spectral representation [92], Karhunen-Loève expansion [65], and Wavelet [123] etc.). Recent attempts utilize the Hermite polynomial chaos method [87, 84, 83] by expressing a non-Gaussian process \( R(t) \) as

\[
R(t) = \sum_{i=1}^{\infty} a_i H_i(G(t)),
\]

(8.1.1)

where \( H_i \) is the Hermite polynomial of degree \( i \), \( G(t) \) is a standard stationary Gaussian process. In [80, 81], the Karhunen-Loève expansion was employed to simulation non-Gaussian processes.

All these simulation methods can be coupled directly with the Monte Carlo method to simulate the random response of a dynamical system. If the random inputs can be expresses by a standard Gaussian process, the polynomial chaos method can be employed directly to capture the uncertainty propagation. However, the following situations may limit the effectiveness of polynomial chaos method:

- The strongly non-Gaussian processes \( R(t) \) may need a high-order polynomial chaos expansion with respect to \( G(t) \). If \( R(t) \) is employed as random inputs, a high-order polynomial chaos
expansion of the solution field may also be necessary for convergence, which may not be possible due to the limitation of dimensionality and complexity.

- When the Karhunen-Loève expansion is employed to simulate the non-Gaussian random processes, the (generalized) polynomial chaos methods can not be used directly since all random variables in the K-L expansion are uncorrelated while not independent.

8.2 The methodology for general random inputs

We use the following stochastic elliptic problem as a model problem: find a stochastic function, $u : \Omega \times \overline{D} \rightarrow \mathbb{R}$, such that almost surely (a.s.) the following equation holds:

$$-\nabla \cdot (a(x;\omega)\nabla u(x;\omega)) = f(x) \quad \text{on } D,$$

$$u(x;\omega) = 0 \quad \text{on } \partial D,$$

(8.2.1)

where $a(x;\omega)$ is a second-order random processes satisfying the assumption

Assumption 8.2.1. Let $a(x;\omega) \in L_\infty(D;\Omega)$ be strictly positive with lower and upper bounds $a_{\min}$ and $a_{\max}$, respectively,

$$0 < a_{\min} < a_{\max} \quad \text{and} \quad \Pr(a(x;\omega) \in [a_{\min}, a_{\max}], \forall x \in \overline{D}) = 1. \quad (8.2.2)$$

This problem was studied in Chapter 5 using the adaptive ME-gPC method, where we imposed i.i.d. assumption on the random variables in the K-L expansion of $a(x)$ for convenience. We here relax this restriction. We only assume that a simulator of $a(x)$ is available, in other words, we can easily obtain a realization of $a(x)$.

We now present a methodology to deal with the general random inputs using the polynomial chaos methods and the Monte Carlo methods.

In practice, the most useful (non-Gaussian) processes are stationary processes. In all the simulation methods of non-Gaussian processes $a(x)$, the correlation function $K_a$ is required to be
maintained. It is known that we can always implement the Karhunen-Loève expansion if $K_a$ is available:

$$a(x) = E[a](x) + \sum_{i=1}^{\infty} \sqrt{\lambda_i} Y_i \psi_i(x),$$

(8.2.3)

where $\{Y_i\}$ is a set of uncorrelated random variables with zero mean and unit variance, and $\{(\lambda_i, \psi_i(x))\}$ is a set of eigenvalue-eigenfunction pairs (see chapter 2). Let $\tilde{a}(x)$ indicate the non-Gaussian random process given by the numerical simulations. We have

$$\tilde{a}(x) = E[\tilde{a}](x) + \sum_{i=1}^{M} \sqrt{\lambda_i} Y_i \psi_i(x),$$

(8.2.4)

from which we can obtained the random variable $Y_i$ as

$$Y_i = \frac{1}{\sqrt{\lambda_i}} \int_{\mathcal{D}} (\tilde{a}(x) - E[\tilde{a}](x)) \psi_i(x) dx.$$

(8.2.5)

Then the marginal PDF $\rho_i$ of $Y_i$ can be obtained by sampling and density estimation techniques (see section 8.4).

Based on the marginal PDFs $\rho_i$, we subsequently define an auxiliary PDF $\rho = \prod_{i=1}^{M} \rho_i$, where the K-L expansion is truncated up to the $M$th term and independence is assumed between $Y_i$ although they are only uncorrelated. The purpose to define $\rho$ is to construct orthogonal (generalized) polynomial chaos bases $\{\phi_{\alpha}\}$ for $Y = (Y_1, \ldots, Y_M)$ using the marginal PDFs $\rho_i$.

Using the polynomial chaos bases $\{\phi_{\alpha}\}$, we can model the solution field as

$$\tilde{u}(x; \tilde{a}(x)) = \sum_{|\alpha| \leq \rho} \tilde{u}_{\alpha}(x) \phi_{\alpha}(Y).$$

(8.2.6)

$\tilde{u}_{\alpha}$ can be obtained using Galerkin or collocation projection of polynomial chaos methods. We note here that both Galerkin and collocation projections will be based on the auxiliary PDF $\rho$ instead of the joint PDF of $Y_i$. However, if $\tilde{u}(x; Y)$ converges in the $L_\infty$ norm, we known that $\tilde{u}(x; Y)$ converges to $u(x; Y)$ with respect to the joint PDF of $Y_i$. In other words, we need enough regularity of $\tilde{u}(x; y)$ with respect to $y$ for convergence.
In the post-processing stage, we can not compute the desired statistics using the polynomial chaos solution since the auxiliary PDF is not equal to the joint PDF. However, since the simulator of \( a(x) \) is available we can sample the joint PDF of \( Y \) and compute the statistics from polynomial chaos solution \( \tilde{u}(x; Y) \). Given \( N_{mc} \) realizations of the random inputs \( \tilde{a}(x) \), we first compute the corresponding random variables \( Y_i \) using equation (8.2.5). Then the moments of \( u(x; \tilde{a}(x)) \) can be obtained as

\[
E[u^m(x; \tilde{a}(x))] = \frac{1}{N_{mc}} \sum_{i=1}^{N_{mc}} \tilde{u}^m(x; Y_i),
\]  

(8.2.7)

In other words, we sample the polynomial chaos solution \( \tilde{u}(x; \tilde{a}(x)) \) instead of the stochastic PDE. If the polynomial chaos solution provides a good point-wise approximation, we will obtain a good speed-up since sampling the stochastic PDE will be more time-consuming than the polynomial chaos methods.

Using the polynomial chaos solution as a prediction for variance reduction, we can further refine the results obtained from the last step by sampling directly the stochastic PDE. We implement Monte Carlo method to compute the moments as

\[
E[u^m(x; \tilde{a}(x))] = E[\tilde{u}^m(x; Y)] + \frac{1}{N_{mc}} \sum_{i=1}^{N_{mc}} [u^m(x; Y_i) - \tilde{u}^m(x; Y_i)],
\]  

(8.2.8)

where \( E[\tilde{u}^m(x; Y)] \) is computed from the joint PDF instead of the auxiliary PDF. Since we have the explicit form of \( \tilde{u}(x; Y) \), we can use a large number of samples to compute \( E[\tilde{u}^m(x; Y)] \) accurately. We expect that \( \tilde{u}^m(x; Y) \) is a good approximation of \( u(x; Y) \) and yields effective variance reduction. In other words, the value of \( N_{mc} \) can be reduced significantly for convergence of the Monte Carlo method.

**Remark 8.2.2.** The basic idea of the methodology is to compute the polynomial chaos expansion using the auxiliary PDF rather than the joint PDF. And the desired statistics are obtained by sampling the polynomial chaos solution rather than the stochastic PDE. Since the measure is changed, the issue on convergence needs to be addresses (see the next section).
8.3 Error analysis of the proposed methodology

In this section we present a general analysis of the errors in the methodology proposed in section 8.2. For simplicity, we use a two-dimensional case, in other words, we assume that a two-term K-L expansion is accurate enough. We first list the error sources as follows:

1. Truncated marginal PDFs. In general, the marginal PDF of $Y_i$ is defined on $y_i \in \mathbb{R}$. For efficient numerical construction of orthogonal polynomials, we truncate the range of $Y_i$. We assume that $\int_{|y_i|>1} |u(x, t, y)| \rho(y) dy \leq \epsilon_1$ for any $x$ and $t$, where $\rho(y)$ is the joint PDF for $Y$. Thus, the range of $Y \in [-1, 1]^2$ is compact, shown in figure 8.1.

2. Approximation of $\tilde{u}(x; t; Y)$ based on the auxiliary PDF. Here we consider the maximum interpolation error instead of the $L_2$ error. The reason that we use the $L_\infty$ norm is that it is in general impossible to obtain the joint PDF. On a compact support, $\tilde{u}(x; Y)$ will approximate $u(x; Y)$ in the $L_2$ sense for any given $x$ since the orthogonal polynomials are complete with respect to the auxiliary PDF. If $\tilde{u}(x; Y)$ can converge to $u(x; Y)$ in the $L_2$ sense with respect to the joint PDF $\hat{\rho}(y)$, it is necessary that $\| \hat{\rho}(y) \|_{L_\infty} < \infty$. However, it is in general impossible to obtain the information of $\hat{\rho}(y)$. Thus we consider the interpolation error of $\tilde{u}(x; y)$ in the $L_\infty$ norm using the analyticity of the function $u(\cdot; y)$ in the parametric space. We see that if $u(\cdot; y)$ is analytic, the interpolation error in the $L_\infty$ norm can be bounded. This is true for the model elliptic problem since the solution is regular in the parametric space (see proposition 5.2.3 in Chapter 5.) and we can always find the analyticity domain larger than the parametric.

We assume the interpolation error satisfies $\| u(x; y) - \tilde{u}(x; y) \|_{L_\infty} \leq \epsilon_2$ for any given $x$.

We now look at the error for given $x$

$$
\epsilon = \int_{y^2} |u(x; y) - \tilde{u}(x; y)| \rho(y) dy \\
= \int_{|y_i|>1} |u(x; y)| \rho(y) dy + \int_{y \in [-1, 1]^2} |u(x; y) - \tilde{u}(x; y)| \rho(y) dy \\
\leq \epsilon_1 + \epsilon_2 
$$

(8.3.1)
Then in the Monte Carlo simulations, we only need to estimate the statistics of a random variable of $O(\epsilon)$, which can result in a significant reduce of the size of realizations.

**Remark 8.3.1.** It is apparent that the choice of the auxiliary PDF is not unique. The reason we use the auxiliary PDF to construct the orthogonal polynomials is that we want to obtain better approximation at the region with a large probability density.

![Figure 8.1: Truncated auxiliary PDF (2D) of the random variables in a K-L expansion.](image)

8.4 Density estimation

Procedures of the K-L expansion and numerical orthogonality can be found in Chapters (2-3). In this section, we discuss the density estimation for the marginal PDFs of $Y_i$. To construct the orthogonal polynomials efficiently, we need that marginal PDFs are smooth because all the algorithms for numerical orthogonality rely on inner products with respect to the weight functions (PDFs). For smooth functions we can employ Gauss-type quadrature rules for an efficient integration.

We here employ the technique of kernel density estimator [78] to estimate the marginal PDFs $\rho_i$. For each realization of the approximated random field $\bar{a}(x)$ we can obtain a sample of $Y_i$ using equation (8.2.5). We assume that we have $N_{mc}$ samples $\{Y_{i,j}\}_{j=1}^{N_{mc}}$ for the random variable $Y_i$. Then the approximation $\hat{\rho}_i$ of $\rho_i$ takes the form

$$\hat{\rho}_i = \frac{1}{N_{mc}h} \sum_{j=1}^{N_{mc}} K_d\left(y_i - \frac{Y_{i,j}}{h}\right),$$  

(8.4.1)
where $h > 0$ is the bandwidth acting as a tuning parameter and $K_d(y)$ is a prescribed kernel function satisfying

$$K_d(y) \geq 0, \quad \int_{\mathbb{R}} K_d(y) dy = 1.$$  \hfill (8.4.2)

The most widely used kernel is the Gaussian kernel $K_d(y) = (2\pi)^{1/2} e^{-y^2/2}$. In this case the kernel density estimate can be written as

$$\hat{\rho}_i = \frac{1}{N_{mc} h \sqrt{2\pi}} \sum_{j=1}^{N_{mc}} e^{-(y_i - Y_{i,j})^2 / 2h^2},$$  \hfill (8.4.3)

The bandwidth $h$ is a scaling factor, which determines the quality of the approximate PDF $\rho_i$. To this end, we choose the optimal bandwidth $h_{\text{opt}}$, which minimize the asymptotic integrated mean square error (AMISE). AIMSE denotes the distance between two density functions, which can be written as

$$\text{AMISE}(\hat{\rho}_i, \rho_i) = \frac{1}{N_{mc}} \int_{\mathbb{R}} K_d^2 dy + \frac{1}{4} h^4 \int_{\mathbb{R}} y^2 K_d(y) dy \int_{\mathbb{R}} (\rho_i^\prime)^2 dy,$$  \hfill (8.4.4)

with $\rho_i^\prime$ being the second-order derivative of $\rho_i$. See [55, 111] and the references therein for numerical algorithms of nonparametric density estimation.

We note that the smoothness of $\hat{\rho}_i$ is inherited from the smoothness of the kernel $K_d$. If the Gaussian kernel is employed, we have $\hat{\rho}_i \in C^\infty(B)$, where $B$ denotes the range of $Y_i$. Thus $\hat{\rho}_i$ is proper for the algorithms of numerical orthogonality. From equation (8.4.1), we can see that the cost for the evaluation of $\hat{\rho}_i$ at a given point is determined by the size number $N_{mc}$ of samples. If $N_{mc}$ is large, the cost for numerical orthogonality is large. However, we note that $\hat{\rho}_i$ is nothing but a sum of a series of Gaussian functions, which implies that the fast Gauss transform (FGT) [47, 107] can be employed to accelerate the calculation. The best rate of convergence of the AMISE of the kernel density estimation is of order $O(N_{mc}^{-4/5})$ [55, 111].

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
8.5 Numerical studies

8.5.1 Algebraic model

We first investigate the convergence of the proposed strategy using a simple algebraic problem

\[(c + \sigma(\xi_1 + \xi_2))u = 1, \quad (8.5.1)\]

where \(c\) and \(\sigma\) are constant, and \(\xi_i, \ i = 1, 2\) are two uncorrelated random variables. For solvability, we assume that \(\xi_1 \in (-1,1)\) and \(c + \sigma(\xi_1 + \xi_2) > 0\).

For simplicity, we assume that \(\xi_2 = \xi_2(\xi_1)\) is a function of \(\xi_1\). Then the fact that \(\xi_1\) and \(\xi_2\) are uncorrelated implies orthogonality between \(\xi_1\) and \(\xi_2\) with respect to the PDF \(f(\xi_1)\) of \(\xi_1\). In other words, \(\xi_2\) can be expressed as an orthogonal polynomial of \(\xi_1\). We note that \(\xi_2\) is dependent on \(\xi_1\). In other words, in the phase space, \((\xi_1, \xi_2)\) is located on a curve \(\xi_2 = \xi_2(\xi_1)\); however, if we use the auxiliary PDF \(\rho(\xi_1, \xi_2)\), \((\xi_1, \xi_2)\) is on \((-1,1)^2\). This can be regarded as the worst case for our methodology since we need to approximate a one-dimensional problem using two-dimensional polynomials. It is obvious that the convergence on \((-1,1)^2\) in the \(L_2\) norm is not enough for the convergence on the curve \(\xi_2(\xi_1)\). A stronger metric to measure the convergence of polynomial chaos is needed, such as \(L_{\infty}\) norm. In this work, we investigate this problem numerically.

We now investigate different distributions of \(\xi_1\). We first sample \(\xi_1\) and \(\xi_2\) to get the density estimation. Based on the estimated marginal PDFs we construct numerical orthogonality and implement the gPC method. Finally we compute the statistics using the joint PDF, in other words, we sample \(\xi_1\) since we know that \(\xi_2\) is a function of \(\xi_1\).

\[\text{Beta}(1/2, 1/2)\]

We let \(\xi_2 = 2\xi_1^2 - 1\). It is easy to verify that \(\xi_2\) also has a Beta distribution on \((-1,1)\) with \(\alpha = \beta = \frac{1}{2}\). We assume that \(\xi_1\) and \(\xi_2\) are independent.

We have two choices to construct the polynomial chaos bases. First, we can construct the two-dimensional polynomial chaos bases using the one-dimensional Chebyshev polynomials since both
marginal PDFs are $\text{Beta}(\frac{1}{2}, \frac{1}{2})$. Second, we can estimate the marginal PDFs of $\xi_1$ and $\xi_2$, and then construct orthogonal polynomial chaos bases numerically. Both cases are investigated.

In figure 8.2 we show the convergence of both approaches. All statistics are computed using the Gauss-Chebyshev quadratures and the relation $\xi_2 = \xi_2(\xi_1)$. The left figure shows the convergence of mean and standard deviation for the joint PDF $\delta(\xi_2 - \xi_2(\xi_1))\rho_{(\frac{1}{2}, \frac{1}{2})}(\xi_1)$, where $\rho_{(\frac{1}{2}, \frac{1}{2})}(\cdot)$ is the PDF of $B(\frac{1}{2}, \frac{1}{2})$ distribution. It can be seen that the convergence rate is actually exponential. In the right figure we present the convergence with respect to the number of samples used for the density estimation. We use polynomial chaos with $p = 6$. It is shown that the convergence is not sensitive at all to the size of samples for density estimation. The reason is that the solution of the algebraic model is analytic with respect to $\xi_1$ and $\xi_2$, and the Taylor expansion converges in the $L_\infty$ norm. We present the different estimated marginal PDFs of $\xi_1$ in figure 8.4. If we use the solutions from numerical orthogonal polynomial chaos and the joint PDF $\delta(\xi_2 - \xi_2(\xi_1))\rho_{(\frac{1}{2}, \frac{1}{2})}(\xi_1)$ to compute the desired statistics, we also obtain fast (exponential) convergence, shown in figure 8.3. We note that if the marginal PDFs are well approximated, we obtain similar convergence behavior between numerical orthogonal polynomial chaos and Chebyshev-chaos (see the right one in figure 8.3).

![Figure 8.2](image-url)

**Figure 8.2:** Convergence of the mean and standard deviation. $N$ is the number of samples for density estimation. Left: Chebyshev polynomials are used; Right: Numerical orthogonality and estimated PDFs are used.
Figure 8.3: Convergence of the mean and standard deviation. Estimated marginal PDFs are used for the numerical orthogonality. Left: 100 samples for density estimation. Right: 100,000 samples for density estimation.

Figure 8.4: PDF of Beta(\(\frac{1}{2}, \frac{1}{2}\)) and estimated ones with different numbers of samples.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
8.5.2 Elliptic model

In this section we study the performance of the proposed methodology numerically using an elliptic problem with an random diffusivity

\[ -\frac{d}{dx} \left( a(x; \omega) \frac{du(x)}{dx} \right) = 1, \quad x \in (0, 1) \]
\[ u(x) = 0, \quad x = 0, 1, \quad (8.5.2) \]

where the random process \( a(x) \) takes the form

\[ a(x) = e^{G(x)} \quad (8.5.3) \]

and \( G(x) \) is a Gaussian random process satisfying an exponential kernel.

Karhunen-Loève expansion of \( a(x) \)

Let \( K_G(x_1, x_2) \) denote the covariance kernel of \( G(x) \) with the form

\[ K_G(x_1, x_2) = \sigma^2 e^{-|x_1 - x_2|/l}, \quad (8.5.4) \]

where \( \sigma \) is constant and \( l \) the correlation length. We know that the K-L expansion of \( G(x) \) is

\[ G(x) = \sigma \sum_{i=1}^{\infty} \sqrt{\lambda_{G,i}} h_{G,i}(x) \zeta_i, \quad (8.5.5) \]

where \( \{\lambda_{G,i}, h_{G,i}\}_{i=1}^{\infty} \) are eigenpairs and \( \{\zeta_i\} \) is a set of independent random variables with zero mean and unit variance. \( G(x) \) is a Gaussian random process and \( \zeta_i \) are We obtain the following statistics of \( a(x) \):

1. Mean:

\[ \bar{a}(x) = \mathbb{E}[a] = e^{\frac{1}{2} \sigma^2}. \quad (8.5.6) \]
2. Variance:

\[
\text{Var}(a) = E[(a - \bar{a})^2] = e^{2\sigma^2} - e^{\sigma^2}. \tag{8.5.7}
\]

3. Normalized covariance kernel:

\[
K_a(x_1, x_2) = \frac{E[(a(x_1) - \bar{a}(x_1))(a(x_2) - \bar{a}(x_2))]}{\text{Var}(a)} = \frac{e^{\sigma^2}e^{-|x_1 - x_2|/\lambda} - 1}{e^{\sigma^2} - 1}. \tag{8.5.8}
\]

The K-L expansion of \( a(x) \) takes the form

\[
a(x) = \bar{a} + \sigma_x \sum_{i=1}^{\infty} \sqrt{\lambda_{a,i}} h(a_i) \xi_i. \tag{8.5.9}
\]

We note that \( \xi_i \) are in general not independent. Although it is hard to obtain the joint PDF of \( \{\xi_i\}_{i=1}^{M} \), we can sample \( \{\xi_i\} \) to get indirectly the samples of \( \{\xi_i\}_{i=1}^{M} \).

**Numerical orthogonality**

Using equation (8.2.5), we can obtain the marginal PDFs of random variables \( \xi_i \) in the K-L expansion of the random process \( a(x) \). In figure 8.5, the estimated marginal PDFs of \( \xi_i \) are shown, where the underlying process \( G(x) \) is Gaussian process. For comparison, we include the normal distribution. It can be seen that the marginal PDFs of \( \xi_i \) are different from the normal distribution. There exist an apparent bias in the marginal PDF of \( \xi_1 \) towards the negative direction. We note that all the estimated PDFs are smooth.

In figure 8.6 we present some orthogonal polynomials with respect to the marginal PDFs of \( \xi_1 \) and \( \xi_2 \). It is seen that nonsymmetric structure in the marginal PDFs is clearly reflected in the orthogonal polynomials, in particular, the fifth-order polynomials.
Figure 8.5: Estimated marginal PDFs using different numbers of sample points. $\zeta_i$ is Gaussian, $\sigma = 0.3$ and $l = 5$. Left: PDFs of $\xi_1$; Right: PDFs of $\xi_i$, $i = 1, 2, 3$.

Figure 8.6: Orthogonal polynomials based on the estimated marginal PDFs with polynomial order $p = 2, \ldots, 5$. $\zeta_i$ is Gaussian, $\sigma = 0.3$ and $l = 5$. Left: $\phi_{1,i}(\xi_1)$; Right: $\phi_{2,i}(\xi_2)$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Convergence of the proposed strategy

Due to the difficulty to obtain accurate numerical solutions, we use the solutions given by Monte Carlo simulations as a reference.

Let $u^p$ be the polynomial chaos approximation of $u(x)$ based on the auxiliary PDF $\rho$. Let $N_\rho$ indicate the size of samples for marginal density estimation and $N_{mc}$ the number of realizations for Monte Carlo simulation. We compute the statistics using the following three methods:

(i) Implement Monte Carlo simulations with $10^5$ realizations.

(ii) Sample $u^p$ with respect to the auxiliary PDF $\rho$. We note that this method is only correct when the random variables in the K-L expansion of $a(x)$ are independent, otherwise, there exist a model error in the random inputs.

(iii) Sample $u^p$ with respect to the joint PDF $\hat{\rho}$.

(iv) Implement Monte Carlo simulations using $u^p$ for variance reduction.

We first consider correlation length $l = 5$. For this case, the eigenvalues decay fast and we keep the first five eigenvalues for the K-L expansion of the log-normal process $a(x)$. We use a 20-term K-L expansion to approximate the underlying Gaussian process $G(x)$ when Monte Carlo simulations are needed.

In figure 8.7 we show the mean and standard deviation of $u(x)$ given by the first three methods. It is seen that the results are close to each other, which implies that the correlation between $\zeta_i$ has a small effect on the statistics up to second order due to the Gaussian-like marginal PDFs.

We subsequently investigate the convergence of methods (iii) and (iv). Let $Q_{\text{ref}}(x)$ be a reference function and $Q(x)$ an approximation. We define the difference between $Q$ and $Q_{\text{ref}}$ as

$$\epsilon = \frac{|Q(x) - Q_{\text{ref}}(x)|}{\max_{0\leq x\leq 1} |Q_{\text{ref}}(x)|}.$$ (8.5.10)

We use the results given by Monte Carlo simulations with $10^5$ realizations as a reference. In figure 8.8 we show the convergence behavior with respect to $\epsilon$ for different polynomial orders $p$ and sample...
size $N_p$ used in the kernel density estimation. For the mean, it is clear that $\epsilon$ becomes smaller as the sample size $N_p$ for the density estimation increases. Similar behavior is observed for the standard deviation. However, $\epsilon$ of the standard deviation given by $N = 10^3$ is almost the same as that given by $N_p = 10^4$ when $p = 2$. Then we increase the polynomial order to $p = 3$ and obtain a smaller $\epsilon$, which implies that the error from polynomial chaos is dominant for $p = 2$ and $N_p = 10^3, 10^4$. Thus, the proposed strategy converges to the correct results when $N_p$ and $p$ increase. In contrary to the algebraic model, we see that the size of samples for density estimation has noticeable influence on the convergence. This is because that the long tails of PDFs can be approximated better if a larger number of samples is used.

However, when the number of random dimensions is large, it is not efficient to increase the polynomial order. Thus, it is necessary to consider the method (iv). We compare the results from methods (iii) and (iv) in figure 8.9. It is seen that a small number of realizations of the stochastic elliptic problems can significantly improve the convergence, which implies that the polynomial chaos solution based on the auxiliary PDF provides a good prediction for variance reduction.

![Figure 8.7: Mean and standard deviation of $u(x)$ given by different methods. Five-term K-L expansion of $a(x)$. $\sigma = 0.2$, $l = 5$. Left: mean; Right: standard deviation.](image-url)
Figure 8.8: Difference of statistics between the proposed strategy and Monte Carlo simulations. Five-term K-L expansion of $a(x)$ is employed. $\sigma = 0.2$ and $l = 5$. Left: mean; Right: standard deviation.

Figure 8.9: Difference of statistics between the proposed strategy and Monte Carlo simulations. Five-term K-L expansion of $a(x)$ is employed. $\sigma = 0.2$ and $l = 5$. Left: mean; Right: standard deviation.
Chapter 9

Applications of ME-gPC in fluid mechanics

In this chapter we present some applications of gPC and ME-gPC in fluid mechanics, including uncertainty in three-dimensional diffusion [110], noisy flow past a circular stationary cylinder [105, 64] and heat transfer in a grooved channel [103, 108].

9.1 Uncertainty in three-dimensional diffusion

9.1.1 Discretization of governing equations

Let $D$ be a bounded, connected, open subset of $\mathbb{R}^d (d = 1, 2, 3)$ with a Lipschitz continuous boundary $\partial D$ and denote by $n$ the unit outward normal direction on $\partial D$. Let $T = [0, \infty)$ be the time domain and $\kappa = \kappa(x; \omega)$ the stochastic diffusivity. Let $\{\partial D_d, \partial D_n\}$ be a partition decomposition of $\partial D$, where $\partial D_d$ is the Dirichlet boundary and $\partial D_n$ the Neumann boundary. The general form of the governing equation for the stochastic diffusion problem can be expressed as

\[
\alpha \frac{\partial u(x, t; \omega)}{\partial t} = \nabla \cdot [\kappa(x; \omega) \nabla u(x, t; \omega)],
\]

(9.1.1a)
where \((x, t; \omega) \in D \times T \times \Omega\), \(\alpha = 0, 1\) in steady and unsteady case, respectively, and the diffusivity \(\kappa : D \times \Omega \to \mathbb{R}\) are prescribed data. Without loss of generality and for simplicity, we neglect the source term in equation (9.1.1a). Note here that if \(\alpha = 0\), the initial condition is unnecessary. By applying the gPC expansion, we can expand the random process in equation (9.1.1a) in the following form:

\[
\begin{align*}
\phi_i(x, t) &= \sum_{i=0}^{M} \phi_i(Y), \\
\kappa(x; \omega) &= \sum_{i=0}^{M} \kappa_i(x) \phi_i(Y),
\end{align*}
\]  

where \(Y = (Y_1, \ldots, Y_n)\) with a dimensionality \(n\) determined by the random inputs. The number \((M + 1)\) is determined by the dimensionality of the chaos expansion \((n)\) and the highest order \((q)\) of the polynomials \(\{\phi_i\}\). Introducing the chaos expansion into the governing equation (9.1.1a), we obtain

\[
\alpha \sum_{i=0}^{M} \frac{\partial u_i(x, t)}{\partial t} \phi_i = \sum_{i=0}^{M} \sum_{j=0}^{M} \nabla \cdot [\kappa_i(x) \nabla u_j(x, t)] \phi_i \phi_j.
\]  

Note that \(M\) can be different for different processes, but for notational convenience we will keep the same \(M\) in all expansions. A Galerkin projection of the above equation onto \(\{\phi_k\}\) is then conducted in order to ensure that the error is orthogonal to the functional space spanned by the finite-dimensional basis \(\{\phi_1, \ldots, \phi_M\}\). Multiplying equation (9.1.3) by \(\phi_k\), evaluating its expectation and taking into account the orthogonality of the basis, we obtain

\[
\alpha \frac{\partial u_k(x, t)}{\partial t} E[\phi_k^2] = \sum_{i=0}^{M} \sum_{j=0}^{M} \nabla \cdot [\kappa_i(x) \nabla u_j(x, t)] e_{ijk},
\]  

where \(e_{ijk} = E[\phi_i \phi_j \phi_k]\), and \(k = 0, 1, \ldots, M\). Together with \((\phi_k^2)\), \(e_{ijk}\) can be evaluated analytically during the pre-processing stage. By defining \(b_{jk}(x) = \frac{1}{E[\phi_k^2]} \sum_{i=0}^{M} k_i(x) e_{ijk}\), we can rewrite the above equation as

\[
\alpha \frac{\partial u_k(x, t)}{\partial t} = \sum_{j=0}^{M} \nabla \cdot [b_{jk} \nabla u_j(x, t)] \quad k = 0, 1, \ldots, M.
\]
Upon expanding initial/boundary conditions (9.1.1b) and (9.1.1c) in a similar manner, we obtain a complete set of equations for each expansion coefficient.

**Steady state equation**

In this case, $\alpha = 0$. Since the size of system (9.1.5) is large for most applications, iterative schemes are preferred over direct schemes. In particular, we employ the block Gauss-Seidel iteration method proposed in [115]:

$$-\nabla \cdot (b_{kk} \nabla u_k^{n+1}(x)) = \sum_{j=0}^{k-1} \nabla \cdot [b_{jk} \nabla u_j^{n+1}(x)] + \sum_{j=k+1}^{M} \nabla \cdot [b_{jk} \nabla u_j^n(x)],$$

(9.1.6)

where the superscript $n$ denotes the iteration step. The convergence criterion is defined as

$$\frac{\|u_k^{n+1}(x) - u_k^n(x)\|}{\sup \|u_k^n(x) - u_k^0(x)\|} \leq \epsilon \quad k = 0, 1, \ldots, M,$$

where $\epsilon > 0$ is the error control. The reason we take $\sup \| \cdot \|$ here is that if the random dimensionality is high, the coefficients of high modes of the basis $\{\phi_k\}$ may be very small. In this paper, we use the $L_\infty$ norm and set $\epsilon = 10^{-9}$. The iterations normally converge within about 10 steps for the results presented in the next section.

**Unsteady equation**

In this case, $\alpha = 1$. A mixed explicit-implicit scheme is employed where we treat $u_k(x,t)$ on the right-hand implicitly and the rest explicitly. In addition, we employ a high-order stiffly stable integration scheme. This algorithm was first used in [58]:

$$\frac{\hat{u}_k(x) - \sum_{q=0}^{J-1} \alpha_q u_k^{n-q}(x)}{\Delta t} = \sum_{q=0}^{J-1} \beta_q \left[ \sum_{j=0,j\neq k}^{M} \nabla \cdot (b_{jk} \nabla u_j(x)) \right]^{n-q},$$

(9.1.7)

$$\frac{\gamma_0 u_k^{n+1}(x) - \hat{u}_k(x)}{\Delta t} = \nabla \cdot [b_{kk} \nabla u_k^{n+1}(x)],$$

(9.1.8)
where \( J \) is the order of accuracy in time and the superscript \( (n+1) \) and \( (n-q) \) denote the time level \( t^{n+1} \) and \( t^{n-q} \), respectively. The values of the coefficients \( \alpha_q, \beta_q \) and \( \gamma_0 \) in the scheme are for different temporal orders. Due to the decoupled system, the equations in (9.1.5) can be solved one-by-one now, so the restriction on each time step is significantly reduced. Spatial discretization can be obtained by any conventional method, e.g. finite difference, finite elements, etc. Here we employ the spectral/hp method based on a Jacobi polynomial tensor-product basis to achieve high accuracy in space and flexible element control [58]. Since the number of degrees-of-freedom is usually large in three-dimensional simulation with high-dimensional random inputs, we parallelize the spectral/hp solver to speed up the computation.

### 9.1.2 Numerical results

![Figure 9.1: Half of the computational domain.](image)

**Aspect of parallel computation**

Due to symmetry only half of the domain is shown in Fig. 9.1, but the computations were carried out in the entire domain. The boundary of the domain consists of four segments: the top \( \Gamma_T \), the bottom \( \Gamma_B \), the sides \( \Gamma_S \) and the boundaries of the cavity \( \Gamma_C \). For the parallel spectral/hp element solver in space, 832 hexahedral elements are used in the domain. Within each element, sixth-order (Jacobi)
polynomials are employed resulting in 190,825 degrees-of-freedom for one deterministic simulation. Second-order chaos and 38-term K-L decomposition were employed, with the smallest eigenvalue being 1.04% of the largest one. This results in a 780-term expansion for $n = 38, q = 2$; thus the total number of unknowns of the stochastic problem is 148,843,500.

To construct the K-L decomposition, the eigenproblem needs to be solved. If a normal eigensolver is employed for this problem, at least 270GBs of memory are needed for the storage and the cost of computation would be $O(190,825^2) \sim O(10^{10})$, which is prohibitively expensive. Thus, a fast eigensolver is necessary for the K-L decomposition, see chapter 2.

Since the stochastic system, which consists of 780 equations, is very large, we resort to parallel computation. By noting that all the stochastic modes are decoupled according to the numerical schemes (9.1.6) and (9.1.7), we can solve all the stochastic modes in parallel. Alternatively, we can parallelize the spectral/hp solver and solve all the stochastic modes one by one. In this paper, we implement the latter choice using MPI. All simulations were run on up to 256 400-MHz MIPS R12000 processors of SGI 3800.

**Steady 3D simulation: Hermite-chaos and log-normal inputs**

We now consider the 3D steady state diffusion with random diffusivity. Adiabatic boundary conditions are prescribed on $\Gamma_S$ and $\Gamma_B$. The Dirichlet boundary conditions are set to be 0 on $\Gamma_T$ and 1.0 on $\Gamma_C$. We assume that the random field $\kappa(x; \omega)$ is a log-normal process, which can be obtained as the exponential of a Gaussian process $g(x; \omega)$. The underlying Gaussian process $g(x; \omega)$ is from a 38-term K-L decomposition with $\sigma_g = 0.2$, and subsequently the log-normal process is approximated by a second-order Hermite-chaos. In Fig. 9.2, the PDF of three random variables from such an approximated log-normal process is shown. A second-order Hermite-chaos is employed to model the random solution. The contours of the stochastic solutions are plotted in Fig. 9.3. It is seen that the largest uncertainty, indicated by the standard deviation, occurs between the top and the cavity surface. We note here that the eigenvalues and eigenfunctions for the K-L expansion are from the direct parallel eigensolver.
We next perform the same simulations using the approximated eigenvalues and eigenfunctions by FGT for the K-L expansion and compare the stochastic solutions with the previous ones. In other words, we study the influence of the errors from the FGT on the steady random diffusion. We show the mean and standard deviation along the center line in Fig 9.4. It can be seen that the approximated eigenvalues and eigenfunctions for $p = 4, 5$ work very well for the mean. However, a big difference occurs for the standard deviation. For the standard deviation the normalized $L_2$ error along the center line is 0.59% for $p = 5$ and 4.00% for $p = 4$. This big difference comes from the error distributions of the approximated eigenvalues we showed in Fig. 2.5. In fact the 21-st eigenvalue is only 4.88% of the largest one and the 38-th eigenvalue is 1.04% of the largest one; the normalized $L_2$ error of the first 21 approximated eigenvalues for $p = 5$ is 0.32% and 1.93% for $p = 4$. This means that the approximated eigenvalues for $p = 5$ are much more accurate than those for $p = 4$ in the energy dominant part of all eigenvalues. This also implies that the difference should be small if we employed the first 21 eigenvalues for the K-L decomposition instead of 38 eigenvalues. In addition, the speed-up of the fast eigensolver for $p = 5$ is more than 600.

Monte Carlo simulations are also conducted to verify the results from Hermite-chaos. In Fig. 9.5, the mean and standard deviation along the vertical centerline are shown. We see that for the standard deviation obtained by the polynomial chaos expansion reaches a good agreement with Monte Carlo simulation based on 10,000 realizations.

![Figure 9.2: PDF of random diffusivity at different reference points based on the log-normal process for the steady case.](image-url)
Figure 9.3: Contours of \( y = 0 \) for steady case. Left: mean field. Right: standard deviation.

Figure 9.4: Influence of approximated eigenvalues and eigenfunctions by FGT on the random diffusion. Statistics along the centerline are used. Left: mean; Right: standard deviation.

Figure 9.5: Comparison with Monte Carlo simulations statistics along the centerline, where a direct parallel solver is used for the K-L decomposition. Left: mean; Right: standard deviation.
Unsteady 3D simulation: Legendre-chaos and non-Gaussian inputs

Here we set the Neumann boundary condition on $\Gamma_C$ to be $q|_{\Gamma_C} = 1$, and also employ adiabatic boundary conditions on $\Gamma_B$, $\Gamma_S$ and $\Gamma_T$. The initial condition is zero everywhere, and a zero source is prescribed. We assume that $\kappa(x; \omega)$ is a random field represented by the 38-term K-L decomposition with the underlying random variables having uniform distribution, and $\sigma_\kappa = 0.2$. The PDF of three random variables from such a random process is shown in Fig. 9.7. A second-order Legendre-chaos is employed for this case. Some reference points are shown in Fig. 9.6, where points without prime are located at $y = 0$ while their projection at $y = 1$ are denoted by prime. We are interested in the stochastic solutions at these points and their cross-correlation coefficients.

In Fig. 9.9, we show the evolution of the stochastic statistics at reference points at $y = 0$, with mean on the left and COV (coefficient of variance) defined as $\text{COV}(x, t) = \sigma_u(x, t)/E[u(x, t; \omega)]$ on the right. It is seen that the mean keeps growing over time while the COVs approach steady-state quickly. Relatively strong variation of COV is observed during the early transient period. In Fig. 9.10, the cross-correlation coefficients between reference points are plotted. It is also observed that the second-order statistics approach steady-state. Since the correlation length $\delta$ of stochastic diffusivity coefficient $\kappa$ is 4.0, which is relatively small, we can observe that as the distance away from $A$ increases, its influence to $B-F$ decreases up to about 15%. In Fig. 9.11, we show the evolution of mean solutions at some reference points at $y = 0$ with error bars indicating standard deviations.
It was shown that the approximated eigenvalues and eigenfunctions by FGT with $p = 4$ gave rise to a large error for the steady case. Subsequently, we employ only the approximated eigenvalues and eigenfunctions by FGT with $p = 5$ to perform the unsteady simulation again. Numerical experiments show that the normalized $L_2$ error of the standard deviation in the whole domain is 2.95% at $t = 1.5$. In Fig. 9.12, the error evolution of some statistics on the reference points is shown. The figure on the left is for the standard deviations and the figure on the right is for the cross-correlation coefficients. It can be seen that all errors reach steady-state quickly. Similar as the steady case, the mean solution is well approximated; thus the mean error is not plotted here. To this end, we show that the approximated eigenvalues and eigenfunctions for $p = 5$ is acceptable for both steady and unsteady problems with errors less than 5% for different statistics. Furthermore, we can obtain such eigenvalues and eigenfunctions by the fast eigensolver with small cost.

9.2 Noisy flow past a stationary circular cylinder

In chapter 6 we studied the long-term behavior of gPC and ME-gPC theoretically using a hyperbolic equation. In this section we further study this problem using a classical CFD model: flow past a stationary circular cylinder. We here relax the inflow boundary conditions to be uncertain, which is modeled by random variables or a random process. The inflow noise can affect the vortex street
Figure 9.8: Instantaneous contours of statistics on $y = 0$ for the unsteady case. Left: mean field. Right: standard deviation.

Figure 9.9: Evolution of statistics at reference points on $y = 0$ for the unsteady random diffusion. Left: mean solution. Right: COV.

Figure 9.10: Cross-correlation coefficients on reference points for the unsteady random diffusion. Left: with respect to point A on $y = 0$. Right: between $y = 0$ and $y = 1$ along the vertical centerline.
Figure 9.11: Temperature evolution at reference points on $y = 0$.

Figure 9.12: Normalized errors of statistics based on the approximated eigenvalues and eigenfunctions by FGT with $p = 5$. Left: standard deviation. Right: cross-correlation coefficients between $y = 0$ and $y = 1$. 

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
behind the cylinder significantly. Roughly, we classify the inflow noise as: random-frequency and random-amplitude inflow noise. Random-frequency inflow noise means that the inflow noise can affect the shedding frequency significantly while random-amplitude inflow noise not. Polynomial chaos shows very different long-term behavior for these two kinds of noise. We note that in general inflow noise is a mixture of these two simple kinds of noise.

9.2.1 Random-frequency inflow noise

We now simulate the two-dimensional noisy flow past a circular cylinder subject to the following random boundary conditions at the inflow

$$u = 1 + \sigma Y, \quad v = 0,$$

(9.2.1)

where $Y$ is a uniform random variable of zero mean and unit variance and $\sigma$ is a prescribed constant indicating the degree of perturbation. For each value of $Y$, there exists a corresponding Reynolds number, which determines a unique vortex shedding frequency. In other words, the shedding frequency in the stochastic simulation is random.

In figure 9.13 the mesh used for the discretization in physical space is shown. Neumann boundary conditions (zero flux) are employed at the outflow and periodic boundary conditions in the cross-flow direction. The numerical formulation of gPC for the incompressible Navier-Stokes equations was presented in [117], where spectral/$hp$ element methods were employed to solve the large deterministic PDE system produced by the Galerkin projection in gPC. The Reynolds number considered in this work is $Re = 100$.

**A stochastic harmonic model of lift coefficient $C_L$**

For a deterministic uniform flow past a stationary circular cylinder, the lift force $F_L$ on the cylinder can be modeled as harmonic in time at the shedding frequency $f_s$ as

$$F_L(t) = \frac{1}{2} \rho U^2 D C_{L}^d \cos(2\pi f_s t),$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
where $C_l^d$ is the maximum amplitude of the instantaneous deterministic lift coefficient. We want to extend such a harmonic model to the stochastic case. If the inflow velocity $u = U(t; \xi)$ is random, the corresponding lift force and vortex shedding frequency will also be random. We express the instantaneous stochastic harmonic model of $C_L$ as

$$
C_L(Y) = A_t(Y) \cos(2\pi f_s(Y)t),
$$

(9.2.2)

where $f_s(Y)$ is the random shedding frequency.

We assume that $f_s(Y)$ can be written as

$$
f_s(Y) = a_0 + a_1 X,
$$

(9.2.3)

where $a_i, i = 0, 1$ are constant and $X$ is a random variable with unit variance defined in $[a, b]$. Then the mean of $C_L$ is

$$
E[C_L] = \int_a^b g(x) \cos(2\pi(a_0 + a_1 x)t)dx,
$$

(9.2.4)

where $g(x) = f(x)A_t(\xi(x))$ with $f(x)$ being the PDF of $X$. By considering the following transform

$$
x = \frac{b - a}{2\pi} y + a, \quad t = \frac{\hat{t}}{a_1(b - a)}
$$

(9.2.5)

$E[C_L]$ can be written as

$$
E[C_L] = \int_0^{2\pi} (B_1(\hat{t}) \cos(\hat{t}y) - B_2(\hat{t}) \sin(\hat{t}y))g(y)dy,
$$

(9.2.6)

where

$$
|B_j| \leq \frac{2\pi}{b - a}, j = 1, 2.
$$

Thus, if $g(y)$ is piecewise,

$$
E[C_L] \to 0 \text{ as } t \to \infty,
$$

(9.2.7)
because
\[ \int_0^{2\pi} G(\hat{t}y)g(y)dy \to \frac{C}{t} \int_0^{2\pi} g'(y)dy \text{ as } t \to \infty, \]
\[ \int_0^{2\pi} G(\hat{t}y)g(y)dy \to -\frac{C}{t} \int_0^{2\pi} g'(y)dy \text{ as } t \to \infty, \]
where \( C \) is a constant and \( G(\hat{t}y) = \cos(\hat{t}y) \) or \( \sin(\hat{t}y) \). Similarly we can obtain that
\[ \mathbb{E}[C^2_L] \to \frac{1}{2} \int_a^b A(\xi(x))f(x)dx \text{ as } t \to \infty. \]

Numerical results

We first simulate a deterministic case with \( \sigma = 0 \) up to \( t = 1000 \) to obtain a fully developed flow and then introduce 10% noise at the inflow. We plot the instantaneous mean and variance of lift coefficient \( C_L \) in figure 9.14, and of drag coefficient \( C_D \) in figure 9.15. It can be seen that both the mean and variance of \( C_L \) given by ME-gPC oscillate periodically around a constant value with a decreasing amplitude after a short transient stage. This agrees with the developed stochastic model of lift coefficient, where \( C_L \) is modeled by a harmonic signal with a random frequency. Based on such a model, the mean of \( C_L \) goes to zero while the variance asymptotes a constant value. Good agreement between gPC and ME-gPC is observed only in the transient stage, after which gPC begins to diverge. Similar trends are observed for the drag coefficient \( C_D \). The study in chapter 6 shows that the polynomial order of gPC must increase at about a constant rate to maintain a certain accuracy if a random frequency is involved. In figure 9.16, we plot the normalized relative errors of the variance of the lift coefficient using the results given by ME-gPC with \( N = 20 \) and \( M = 8 \) as a reference. It can be seen that the errors of gPC increase quickly to \( O(1) \). ME-gPC with \( N = 20 \) and \( p = 6 \) reaches an error of \( O(10^{-2}) \) at \( t \approx 135 \). We note that errors less than \( 10^{-5} \) are not shown because the output data are truncated after the fifth digit.

In chapter 4 we have shown that the error of ME-gPC at a fixed time can be estimated from that of gPC of the same polynomial order but shifted by a factor \( N \). Here we cannot use this result directly because the decomposition of random shedding frequencies is not necessarily uniform although the noise at the inflow is uniform. However, we can estimate the scaling factor from the simulation results for the errors of gPC and ME-gPC of sixth order, which is about 12. Using this
value we know that the error of ME-gPC with $N = 20$ and $M = 8$ at $tU/D = 150$ should be roughly equal to the error of eighth-order gPC at $tU/D = 150/12$, which is $O(10^{-3})$. Thus, ME-gPC with $N = 20$ and $M = 8$ can provide accurate results in the range $tU/D \leq 150$, corresponding to about 20 shedding periods after the transient stage. In contrast, gPC provides accurate result up to less than two shedding periods.

In figure 9.17 the RMS of vorticity is plotted. The global structure is (approximately) symmetric and the values of RMS of vorticity are decreasing gradually from the front stagnation point, through the boundary layers, into the wake. This suggests that the vorticity behind the cylinder should contain a harmonic signal $A(x, Y) \cos(2\pi f_v(Y)t)$ with random frequencies $f_v$. The RMS of such a harmonic response will approach $\int_Y A^2(x, Y)f(Y)/2dY$ as $t \to \infty$, where $f(Y)$ is the PDF of $Y$. Since the flux of vorticity decreases in the $x$ direction due to viscous diffusion, the value of $A(x, Y)$ should also decrease in the $x$ direction. This explains qualitatively why we only observe decreasing RMS values of vorticity in the wake without the von Karman vortex street.

Subsequently, we study the random amplitude $A_L$ by incorporating the following models: (1) linear model: $A_L(\xi) = A_0 + c_1 \xi$ and (2) quadratic model: $A_L(\xi) = A_0 + c_2 \xi + c_3 \xi^2$, where $A_0 = 0.3372$ is the deterministic amplitude at $Re = 100$ given by direct numerical simulations and $c_i, i = 1, 2, 3$ are undetermined constants. Using the first- and second-order moments, we can resolve $c_i$ from the ME-gPC results. For the uniform noise, we obtain $c_i = 0.1402, 0.1100, 0.0109$; for the Beta noise, $c_i = 0.1414, 0.1101, 0.0112$. These two sets of numbers agree well because the above two models are independent of the PDF of $\xi$. We observe from the linear model that the response degree is $c_1/A_0 \approx 42\%$ in contrast to the 10% noise at the inflow, which implies that the lift coefficient is very sensitive to the upstream noise. Due to the quadratic term, the RMS of $C_L$ is bigger than the deterministic one at $Re = 100$ as observed in Fig. 9.18. In Fig. 9.19 the PDFs of $A_L$ are shown. We see that the PDFs of $A_L$ are well described by the quadratic model. This is consistent with the deterministic model of lift coefficient [22]. In contrast to the symmetric noise, the PDF of $A_L$ has a clear bias towards the side of lower Reynolds numbers.
Figure 9.13: Schematic of the domain for noisy flow past a circular cylinder. The size of the domain is $[-15D, 25D] \times [-9D, 9D]$ and the cylinder is at the origin with diameter $D = 1$. The mesh consists of 412 triangular elements.

Figure 9.14: Evolution of mean (upper) and variance (lower) of lift coefficient. $Y$ is uniform in $[-\sqrt{3}, \sqrt{3}]$. $\sigma = 0.1$. 

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Figure 9.15: Evolution of mean (upper) and variance (lower) of drag coefficient. $Y$ is uniform in $[-\sqrt{3}, \sqrt{3}]$. $\sigma = 0.1$.

Figure 9.16: Evolution of relative errors of instantaneous variance of lift coefficients given by gPC and ME-gPC. The results obtained from ME-gPC with $N = 20$ and $M = 8$ are used as a reference.
Figure 9.17: Instantaneous spatial distribution of RMS of vorticity at $tU/D = 100$. $N = 20$ and $M = 8$.

Figure 9.18: Instantaneous RMS of lift coefficients. A uniform distribution in $[-\sqrt{3}, \sqrt{3}]$ and a Beta distribution $Be(2, 2)$ in $[-\sqrt{7}, \sqrt{7}]$ are considered with $\sigma = 0.1$.

Figure 9.19: PDFs of $A_L$ given by different models: (left) uniform noise and (right) Beta noise.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
9.2.2 Random-amplitude inflow noise

Next we consider another noisy boundary condition at the inflow

\[ u = 1 + \sigma Y \cos 2\pi f_{in} t, \quad v = 0, \quad (9.2.10) \]

where we add a harmonic signal with a random amplitude into the inflow. We use \( f_{in} = 0.75f_s \), where \( f_s \) is the vortex shedding frequency at \( Re = 100 \). Let \( Y \) be a uniform random variable with zero mean and unit variance. We set \( \sigma \) to 0.1.

In figure 9.20 we compare the mean and variance of \( C_L \) given by gPC and ME-gPC of the same order \( M = 8 \), where \( N = 10 \) for the ME-gPC. We see that eighth-order gPC can capture all the statistics up to the second-order in the range of \( tU/D \leq 150 \) in contrast to the fast divergence of gPC for the random-frequency noise. Numerical experiments show that for the boundary condition (9.2.10) the frequency of \( C_L \) is not sensitive to the boundary noise, where the vortex shedding frequency at \( Re = 100 \) is dominant. Thus, the error of gPC increases much slower than the first case. A similar example is noisy flow past an oscillating circular cylinder [68], where the frequency is also not sensitive to the noise and thus gPC can do a good job. Such observations imply that the presence of random frequencies can cause a significant degradation of the performance of gPC, and thus employing ME-gPC is necessary for convergent results.
9.3 Heat transfer in groove channel

Convective heat transfer in grooved channels is a prototype problem, representative of a wide spectrum of engineering applications, such as heat exchangers, biomedical devices, cooling of microelectronic components, etc. It has been found in [38, 39] that inducing oscillations in the driving flow is an effective approach in improving substantially heat transfer rates. In the past two decades, this method, which is often called resonant heat transfer enhancement, has been widely studied numerically and experimentally [38, 39, 79, 10, 9, 49, 50, 26]. In these studies precise sinusoidal excitations were considered and precise knowledge of thermal boundary conditions was assumed. We now present a new modelling approach for unsteady convective heat transfer, where such constraints are relaxed and the excitations are described as stochastic processes. In particular, we study enhancement of heat transfer in a periodic grooved channel due to stochastic excitation. The stochastic simulations allow us to obtain lower and upper bounds of the Nusselt number as well as probability distributions, thus making the comparison of the simulation results with the experimental results more meaningful. In addition, the stochastic analysis readily provides measures of the sensitivity of the system while the richer and high-dimensional stochastic output can aid significantly in gaining deeper understanding of the momentum and heat transport mechanisms.

We employ ME-gPC to study heat transfer enhancement in a two-dimensional grooved channel. This geometry was introduced in [38, 39] and we use it here for comparisons with corresponding deterministic simulations reported therein. The excitation is an oscillation of random amplitude superimposed to the main deterministic flow. Based on the global energy balance presented in [37], we derive proper numerical schemes for the reduced coupled deterministic PDE system from polynomial chaos for fully developed flow. We consider two different stochastic inputs corresponding to uniform and Beta PDFs, and perform simulations at three different Reynolds numbers, i.e., $Re = 225, 525, 825$. We present results for the Nusselt number expressed as a heat transfer enhancement parameter $E$ for different excitation frequencies, where the optimal frequency for heat transfer enhancement is obtained from the mean stochastic response. We study, in particular, the PDFs of $E$ as well as its dependence on the amplitude and the frequency of the excitation. Using some statistics,
Figure 9.21: Schematic of the geometry with a periodicity length $L$. The domain is composed of the inflow boundary $\partial D_{in}$, the outflow boundary $\partial D_{out}$ and boundaries $\partial D_i, i = 1, 2, 3$.

e.g., variance and correlation, we demonstrate the uncertainty distribution and propagation in the flow field.

9.3.1 Governing equations

The schematic of computational domain $D \subset \mathbb{R}^d$ to be considered is shown in figure 9.21, which is periodical in the streamwise direction with a given periodicity length $L$ and Lipschitz boundary $\partial D$. The flow is assumed to be fully developed in the $x$-direction. Natural convection, variation of thermal properties and non-fully developed effects are all assumed to be negligible in this work.

9.3.2 Stochastic flow-rate condition

We consider a stochastic flow-rate condition

$$Q(t; \omega) = \int_{\partial D_{in}} u(x, y, t; \omega) dy = \frac{4}{3}(1 + \eta X(t; \omega)), \quad (9.3.1)$$

where $\eta$ is constant indicating the degree of perturbation and $X(t; \omega)$ is a second-order periodically correlated random process [44]. In other words, $X(t; \omega)$ satisfies

$$E[X(t)] = E[X(t + T)], \quad (9.3.2a)$$

$$K(t, s) = K(t + T, s + T), \quad (9.3.2b)$$
where $K(t, s)$ is the covariance kernel of $X(t; \omega)$ and $T$ is a period. The Karhunen-Loeve (K-L) decomposition of $X(t; \omega)$ takes a form [67]

$$X(t) = \sqrt{\frac{T}{\pi}} Y_{0,0} + \sqrt{\frac{2}{T}} \sum_{n=1}^{\infty} \sqrt{\lambda_n} \left[ Y_{n,1} \cos \frac{2n\pi}{T} t + Y_{n,2} \sin \frac{2n\pi}{T} t \right], \quad (9.3.3)$$

where $\{Y_{i,j}\}$ is a set of uncorrelated random variables. The eigenvalues $\{\lambda_i\}$ can be determined by the correlation length and Fourier coefficients of the kernel $K(t, s)$. Without loss of generality, we assume that the first eigenvalue $\lambda_0 = 0$ since the corresponding eigenfunction is constant. We note that the time average of $X(t; \omega)$ is zero due to the periodical properties of sine and cosine functions.

In engineering, such random inputs correspond to *amplitude modulation* of signals. Hence the time-averaged flow rate is $\bar{Q} = 4/3$, where

$$\bar{Q}(\omega) = \frac{1}{\tau} \int_{t_0}^{t_0+\tau} \left( \int_{\Omega} u(y, t; \omega) \cdot ndl \right) dt. \quad (9.3.4)$$

with $n$ being the outward normal. We note that the value of $Q(t; \omega)$ is independent of $x$ due to the mass conservation.

**Navier-Stokes equations**

For the velocity field we have the incompressible Navier-Stokes equations

$$\frac{\partial v}{\partial t} + (v \cdot \nabla)v = -\nabla p + Re \nabla^2 v \quad \text{in } D, \quad (9.3.5a)$$

$$\nabla \cdot v = 0 \quad \text{in } D, \quad (9.3.5b)$$

where $Re$ is the Reynolds number.

By taking into account the fully developed flow and periodical solution domain, we employ the following boundary conditions for the velocity

$$v(x, t; \omega) = 0 \quad \text{on } \partial D_i, i = 1, 2, 3 \quad (9.3.6a)$$
corresponding to no-slip and periodicity, respectively. For the pressure [37] we require that

\begin{equation}
\Pi(x, t; \omega) = -\Pi^x(t; \omega)x + \Pi(x, t; \omega),
\end{equation}

\begin{equation}
\Pi(x, t; \omega)|_{\partial D_{in}} = \Pi(x, t; \omega)|_{\partial D_{out}},
\end{equation}

where we extract a linear term in the x-direction and the coefficient \( \Pi^x \) will serve as a force term. Otherwise, a trivial solution will be obtained if only the periodical boundary conditions are imposed.

**Thermal convection equation**

For the temperature field \( T \)

\begin{equation}
\frac{\partial T}{\partial t} + \nabla \cdot (vT) = (RePr)^{-1} \nabla^2 T,
\end{equation}

where \( Pr \) is the Prandtl number. Due to the properties of convection we subtract a linear term from \( T \), which results in

\begin{equation}
T(x, t; \omega) = \gamma(\omega)x + \theta(x, t; \omega).
\end{equation}

For the deterministic case, a global energy conservation was used in [37] to determine the linear coefficient. Such a technique can be extended naturally to the stochastic case, since for each sample \( \omega \) the global energy conservation will be maintained. Along the same line, we obtain the linear coefficient \( \gamma(\omega) \) as

\begin{equation}
\gamma(\omega) = \frac{\alpha q'(\omega)}{\kappa \bar{Q}(\omega)L}
\end{equation}

where \( \kappa \) is the thermal conductivity of the fluid, \( \bar{Q}(\omega) \) is the time-averaged flow rate

\begin{equation}
\bar{Q}(\omega) = \frac{1}{\tau} \int_{t_0}^{t_0+\tau} \left( \int_{x=L} u(y, t; \omega) \cdot n dl \right) dt,
\end{equation}
and \( q' \) is the total heat transfer rate into the domain

\[
q'(\omega) = \int_{\partial D} q''(x; \omega) \, dl.
\]

(9.3.12)

Here \( q''(x; \omega) \) is the heat flux at the boundaries. In terms of \( \theta \), equation (9.3.8) can be written as

\[
\frac{\partial \theta}{\partial t} + \nabla \cdot \mathbf{H} = \alpha \nabla^2 \theta
\]

(9.3.13a)

\[
\mathbf{H} = \mathbf{v}(\theta + \gamma x)
\]

(9.3.13b)

subject to the following boundary conditions

\[
\kappa \nabla \theta \cdot \mathbf{n} = q''(x) - \kappa \gamma n_x \quad \text{on } \partial D_i, i = 1, 2, 3
\]

(9.3.14a)

\[
\theta(x, t; \omega)|_{\partial D_{in}} = \theta(x, t; \omega)|_{\partial D_{out}}
\]

(9.3.14b)

where \( n_x \) is the component of \( \mathbf{n} \) in the \( x \)-direction.

**ME-gPC expansions**

To employ ME-gPC we expand all the random fields spectrally in each random element \( B_k \). Given the local polynomial chaos basis \( \{ \phi_{k,i}(Y_k) \} \), the random fields, e.g., \( \mathbf{v} \), \( \Pi \) and \( \theta \), can be expressed in random element \( B_k \) as

\[
\mathbf{v} = \sum_{i=0}^{N_p-1} v_{k,i} \phi_{k,i}, \quad \Pi = \sum_{i=0}^{N_p-1} \Pi_{k,i} \phi_{k,i}, \quad \theta = \sum_{i=0}^{N_p-1} \theta_{k,i} \phi_{k,i}.
\]

(9.3.15)

Substituting (9.3.15) into equation (9.3.5) and (9.3.13) and performing a Galerkin projection in terms of each mode in \( \{ \phi_{k,i} \} \), we obtain the following deterministic PDE system

\[
\frac{\partial \mathbf{v}_{k,m}}{\partial t} + \mathbf{N}_{k,m} = -\hat{\Pi}_{k,m} + \nu \nabla^2 \mathbf{v}_{k,m} + \hat{\Pi}_{k,m} \hat{x}
\]

(9.3.16a)
\[ \nabla \cdot \mathbf{v}_{k,m} = 0 \]  
\[ (9.3.16b) \]

\[ \frac{\partial \theta_{k,m}}{\partial t} + \frac{1}{\langle \Phi_{k,m}^2 \rangle_f} \sum_{i=0}^{N_p-1} \sum_{j=0}^{N_p-1} e_{ijm}^k \nabla \cdot \mathbf{H}_{i,j} = \alpha \nabla^2 \theta_{k,m}, \]  
\[ (9.3.16c) \]

where \( e_{ijm}^k = \mathbb{E}_k[\phi_{k,i} \phi_{k,j} \phi_{k,m}] \), \( \hat{x} \) denotes the unit vector in the x-direction, \( \mathbf{H}_{i,j} = \mathbf{v}_{k,i}(\theta_{k,j} + \gamma_{k,j}x) \)

and

\[ N_{k,m} = \frac{1}{\langle \Phi_{k,m}^2 \rangle_f} \sum_{i=0}^{N_p} \sum_{j=0}^{N_p} c_{ijk}^k \left[(\mathbf{v}_{k,i} \cdot \nabla) \mathbf{v}_{k,j}\right], \]  
\[ (9.3.17) \]

\( m = 0, 1, \ldots, N_p - 1 \). The same procedure can be applied for the boundary conditions to complete the above PDE system.

In particular, we also need to project equation (9.3.10), the global energy conservation, onto ME-gPC basis. Using the following ME-gPC expansions

\[ \gamma(\omega) = \sum_{i=0}^{N_p-1} \gamma_{k,i} \phi_{k,i}, \quad q'(\omega) = \sum_{i=0}^{N_p-1} q'_{k,i} \phi_{k,i}, \quad \bar{Q}(\omega) = \sum_{i=0}^{N_p-1} \bar{Q}_{k,i} \phi_{k,i} \]  
\[ (9.3.18) \]

and a Galerkin projection, we obtain

\[ \sum_{i=0}^{N_p} \sum_{j=0}^{N_p} \gamma_{k,i} \bar{Q}_{k,i} e_{ijm}^k = \frac{\alpha}{\kappa L} q'_{k,m} \mathbb{E}_k[\Phi_{k,m}^2], \quad m = 0, 1, \ldots, N_p, \]  
\[ (9.3.19) \]

which is a linear system in terms of \( \gamma_{k,i} \).

**Postprocessing**

Once we obtain the local polynomial chaos expansion in each element \( B_k \), we can gather the local information to compute the global statistics. We subsequently use the temperature field \( T(\mathbf{x}, t; \omega) \) as an example to present the formulas for the commonly used statistics: mean \( M(T) \) and variance \( V(T) \).

Due to the orthogonality of local polynomial basis \( \{ \phi_{k,i} \} \), we know the local mean is \( T_{k,0} \). Using
the ME-gPC formulas, we obtain the global mean is

\[ M(T) \approx \sum_{k=1}^{N} \Pr(I_{B_k} = 1) T_{k,0}. \quad (9.3.20) \]

The global variance can be expressed as

\[ V(T) = \mathbb{E}[T^2] - M^2(T) \approx \sum_{k=1}^{N} \Pr(I_{B_k} = 1) \sum_{i=0}^{N_p} T_{k,i}^2 \mathbb{E}_{k}[\phi_{k,i}^2] - M^2(T). \quad (9.3.21) \]

Using the global mean and variance, we can obtain the global skewness and kurtosis, respectively,

\[ S(T) = V^{-3/2}(T) \mathbb{E}[(T - M(T))^3] \]
\[ = V^{-3/2}(T) \left[ \mathbb{E}[T^3] - 3M(T)V(T) - M^3(T) \right] \quad (9.3.22) \]
\[ = V^{-3/2}(T) \sum_{k=1}^{N} \Pr(\xi \in \mathcal{C}_k) \sum_{i=0}^{N_p} \sum_{j=0}^{N_p-1} \sum_{m=0}^{N_p-1} T_{k,i} T_{k,j} T_{k,m} \mathbb{E}_k[\phi_{k,i} \phi_{k,j} \phi_{k,m}] \]
\[ - V^{-3/2}(T)(3M(T)V(T) + M^3(T)), \]

\[ K(T) = V^{-2}(T) \mathbb{E}[(T - M(T))^4] \]
\[ = V^{-2}(T) \left[ \mathbb{E}[T^4] - 4M(T)\mathbb{E}[T^3] + 6M^2(T)V(T) + 3M^4(T) \right] \]
\[ = V^{-2}(T) \sum_{k=1}^{N} \Pr(\xi \in \mathcal{C}_k) \sum_{i=0}^{N_p-1} \sum_{j=0}^{N_p-1} \sum_{m=0}^{N_p-1} T_{k,i} T_{k,j} T_{k,m} \mathbb{E}_k[\phi_{k,i} \phi_{k,j} \phi_{k,m}] \]
\[ - 4V^{-2}(T)M(T) \sum_{i=0}^{N_p-1} \sum_{j=0}^{N_p-1} \sum_{m=0}^{N_p-1} T_{k,i} T_{k,j} T_{k,m} \mathbb{E}_k[\phi_{k,i} \phi_{k,j} \phi_{k,m}] \]
\[ + V^{-2}(T) \left[ 6M^2(T)V(T) + 3M^4(T) \right]. \quad (9.3.23) \]

### 9.3.3 Discretization of governing equations

#### Temporal discretization

In this section we present the semidiscrete formulation of equation (9.3.16). For the Navier-Stokes equations, we employ a splitting technique [58, 117], treating the convection term explicitly and the...
diffusion term implicitly, i.e.,

\[
\frac{\hat{v}_{k,m} - \sum_{q=0}^{J} \alpha_q v_{k,m}^{n-q}}{\Delta t} = - \sum_{q=0}^{J} \beta_q N_{k,m}^{n-q} + \Pi_{k,m}^{x,n+1/2} \hat{\gamma}
\]  

(9.3.24a)

\[
\frac{\dot{v}_{k,m} - \hat{v}_{k,m}}{\Delta t} = -\nabla \Pi_{k,m}
\]  

(9.3.24b)

\[
\frac{\gamma_0 v_{k,m}^{n+1} - \hat{v}_{k,m}}{\Delta t} = \nu \nabla^2 v_{k,m}^{n+1}
\]  

(9.3.24c)

where \( J \) is the order of accuracy in time, \( \alpha_q, \beta_q \) and \( \gamma_0 \) are integration weights. The discrete divergence-free condition for each velocity mode \( \nabla \cdot \mathbf{v}_{k,m}^{n+1} = 0 \) results in a set of consistent Poisson equations for each pressure mode

\[
\nabla^2 \Pi_{k,m} = \frac{1}{\Delta t} \nabla \cdot \hat{v}_{k,m}
\]  

(9.3.25)

with an appropriate boundary condition [57]

\[
\frac{\partial \Pi_{k,m}}{\partial n} = \mathbf{n} \cdot \left[ \sum_{q=0}^{J} \beta_q N_{k,m}^{n-q} + \nu \sum_{q=0}^{J} \beta_q \mathbf{L}(\mathbf{v}_{k,m}^{n-q}) \right] + n_x \Pi_{k,m}^{x,n+1/2},
\]  

(9.3.26)

where \( \mathbf{L}(\mathbf{v}_{k,m}) = -\nabla \times (\nabla \times \mathbf{v}_{k,m}) \). The scheme achieves third-order accuracy in time and the integration weights can be found in [58].

However, above numerical schemes are not implementable because \( \Pi_{k,m}^{x,n+1/2} \) are unknown. Thus, we need to figure out the condition that \( \Pi_{k,m}^{x,n+1/2} \) must satisfy. In equation (9.3.1), we prescribe a volume flow rate \( Q(t; \omega) \), which can be used to determine \( \Pi_{k,m}^{x,n+1/2} \) in a proper way. We here modify the technique described in [37], which is related to a Green’s function \( \mathbf{v}^g \)

\[
\frac{\tilde{\mathbf{v}}^g}{\Delta t} = \mathbf{\hat{x}}
\]  

(9.3.27a)

\[
\frac{\dot{\mathbf{v}}^g - \tilde{\mathbf{v}}^g}{\Delta t} = -\nabla \Pi^g
\]  

(9.3.27b)

\[
\gamma_0 v^g - \tilde{v}^g = \nu \nabla^2 v^g
\]  

(9.3.27c)
with the pressure boundary condition for $\Pi^g$ set to be

$$\frac{\partial \Pi^g}{\partial n} = n_x \quad (9.3.28)$$

and an associated flow rate

$$Q^g = \int_{\partial D_{\text{out}}} (v^g \cdot n) dA. \quad (9.3.29)$$

The Green's function $v^g$ can be solved in a pre-processing stage. Another important thing is that such $v^g$ does not depend on any local information in random element $k$ and the index $m$, which implies that it is the same for all the stochastic modes $v_{k,m}$ in all the random elements.

At each time step, we first employ the following scheme without including the effect of $\Pi^g_{k,m}$

$$\frac{\dot{v}^I_{k,m} - \sum_{q=0}^{J} \alpha_q v_{k,m}^{n-q}}{\Delta t} = -\sum_{q=0}^{J} \beta_q N_{k,m}^{n-q} \quad (9.3.30a)$$

$$\frac{\dot{v}^I_{k,m} - \ddot{v}^I_{k,m}}{\Delta t} = -\nabla \Pi^I_{k,m} \quad (9.3.30b)$$

$$\frac{\gamma_0 v^I_{k,m} - \ddot{v}^I_{k,m}}{\Delta t} = \nu \nabla^2 v^I_{k,m} \quad (9.3.30c)$$

with the pressure boundary condition set to be

$$\frac{\partial \Pi^I_{k,m}}{\partial n} = n \cdot \left[ \sum_{q=0}^{J} \beta_q N_{k,m}^{n-q} + \nu \sum_{q=0}^{J} \beta_q L(v_{k,m}^{n-q}) \right] \quad (9.3.31)$$

and an intermediate flow rate as

$$Q^I_{k,m} = \int_{\partial D_{\text{out}}} (v^I_{k,m} \cdot n) dA. \quad (9.3.32)$$

By the linearity of the numerical scheme, it can seen that the original scheme (9.3.24) can be easily recovered by the schemes (9.3.27) and (9.3.30), which implies that

$$v_{k,m}^{n+1} = v^I_{k,m} + \Pi^g_{k,m} v^g. \quad (9.3.33)$$
From equation (9.3.33), it is easy to obtain

\[ Q_{k,m}^{n+1} = Q_{k,m}^I + \Pi_{k,m}^{r,n+1/2} Q^g. \]  

(9.3.34)

Sine \( Q_{k,m}^{n+1} \), \( Q_{k,m}^I \), and \( Q^g \) are known, we can compute requisite pressure gradient \( \Pi_{k,m}^r \) using the relation (9.3.34), then obtain \( v_{k,m}^{n+1} \) from equation (9.3.33).

Similarly the scheme for the thermal convection equation is expressed as

\[ \frac{\hat{\theta}_{k,m} - \sum_{q=0}^{J} \alpha_q \theta_{k,m}^{n-q}}{\Delta t} = \sum_{i=0}^{N_x-1} \sum_{j=0}^{N_y-1} \nabla \cdot \mathbf{H}_{i,j} e_{ij}^{k} \theta_{k,m}^{n-q} \]  

(9.3.35a)

\[ \frac{\gamma_0 \theta_{k,m}^{n+1} - \hat{\theta}_{k,m}}{\Delta t} = \alpha \nabla^2 \theta_{k,m}^{n+1}. \]  

(9.3.35b)

We note here that we treat the convection term explicitly, which means that we only need the information of velocity at previous time steps \( n-q \), \( q = 0, 1, \cdots, J \).

**Spatial discretization**

For the spatial discretization, any classical methods, such as finite difference, finite element, etc., can be employed. In this work, we use the spectral\(/hp\) finite element method to achieve efficiency and flexible control on accuracy, where a modal basis [58], constructed by tensor products of one-dimensional Jacobi polynomials \( P_n^\alpha, \beta(\cdot) \), is employed for the spectral expansion in space. A typical mesh is shown in figure 9.22, where 42 elements and eighth-order Jacobi polynomials are used to obtain converged results.
9.3.4 Numerical results

Notation

To get an overall measure of the transport characteristics of the fully developed flow, we first define a time-averaged global Nusselt number $Nu$ as in [39]

$$
Nu(\omega) = L(1 + \frac{2a}{L})^2 \int_{\partial D_B} \langle \theta(\omega) - \theta_b(\omega) \rangle_t ds|_{\partial D_B},
$$

(9.3.36)

where $\theta_b$ is a reference temperature taken to be the (periodic part of the) mixed-mean temperature at $x = 0$,

$$
\theta_b(\omega) = \left\langle \int_{\partial D_B} \frac{u(x = 0, y, t; \omega)\theta(x = 0, y, t; \omega)dy}{Q(t; \omega)} \right\rangle_t,
$$

(9.3.37)

and $\langle \rangle_t$ refers to the time-average over a cycle of the flow, $t < t' < t + \Omega_F^{-1}$. The transport enhancement parameter is defined as

$$
E = \frac{Nu(Re, \eta, Pr)}{Nu(Re, \eta = 0, Pr)}.
$$

(9.3.38)

For the purpose of comparison, the parameters in table 9.1 are the same as those in [39].
Table 9.1: Physical parameters used in this work. The lengths are non-dimensionalized with the channel half-width $h$.

Mean values and bounds of $E$

We first assume that in the random flow rate condition (9.3.1) $X(t; \omega)$ takes a simple form

$$X(t; \omega) = \eta Y \sin 2\pi \Omega_F t,$$

(9.3.39)

where $Y$ is a random variable defined in $[0,1]$ and $\eta = 0.2$. Compared to $\tilde{Q}$, the mean amplitude $\eta \bar{Y} = 0.1$ indicates 10% perturbation. For the random amplitude the coefficient of variance is $\sigma_a/\mathbb{E}[\eta Y] = 58\%$ for uniform distribution and 38% for Beta$(2,2)$ distribution, where $\sigma_a$ is the standard deviation of $\eta Y$.

In figure 9.23, we show the mean of $E$ with “error bars” for uniform (upper) and Beta (lower) random inputs at different excitation frequencies, where both the upper and lower error ranges are equal to the standard deviation of $E$. ME-gPC up to $N = 3$ and $p = 5$ is employed to assure the numerical convergence. We first examine the results for the uniform random inputs. We observe that the mean of $E$ is sensitive to the frequency of excitation force, and the maximum value is reached at $\Omega_F \approx 0.15$. For comparison, we also show the corresponding deterministic solutions at $E(\bar{Y} = 0.5)$, where 0.5 is the mean of the uniform random variable $Y$. We can see that the mean of $E$ for different frequencies does not coincide with the corresponding deterministic solution, since the relation between $Y$ and $E$ is nonlinear. Furthermore, $E(\bar{Y} = 0.5) \geq \bar{E}$ when $\Omega_F \leq 0.16$; otherwise, $E(\bar{Y} = 0.5) \leq \bar{E}$. This means that the relation between $E$ and $Y$ changes for different frequencies. However, the deterministic and stochastic mean solutions give the same optimal frequency. As $\Omega_F$ approaches 0.15, the standard deviation increases gradually. This can be interpreted as a result of sensitivity of the heat transfer enhancement to the excitation frequency. Also, it is seen that the difference between $\bar{E}$ and $E(\bar{Y} = 0.5)$ reaches a maximum value at $\Omega_F = 0.15$. For Beta random

<table>
<thead>
<tr>
<th>Re</th>
<th>Pr</th>
<th>$L$</th>
<th>$l$</th>
<th>$a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(225-825)</td>
<td>1.0</td>
<td>6.6666</td>
<td>2.2222</td>
<td>1.1111</td>
</tr>
</tbody>
</table>
inputs, similar properties are observed.

In [38], it was found that the least-stable modes of grooved channel flow closely resemble Tollmien-Schlichting channel waves. The frequency $\Omega_1$ of the first least-stable mode can be predicted by the Orr-Sommerfeld dispersion relation; it is 0.142 for the geometry used in this work. In [39], it was verified that the optimal frequency for heat transfer enhancement is consistent with the Tollmien-Schlichting frequency for the linear system and shifts a little for the nonlinear system. The optimal frequency $\Omega_F = 0.15$ given by the polynomial chaos method agrees well with the aforementioned deterministic studies. We note that the standard deviation (indicated by the length of “error bars”) of $E$ does not reach the maximum value at $\Omega_F = 0.15$ as the mean does. The excitation frequency for the largest standard deviation is actually a little larger than $\Omega_F = 0.15$. This implies that the optimal frequency depends on the amplitude. Since the mean of $E$ has a maximum value at $\Omega_F = 0.15$, we know that

$$\frac{\partial E}{\partial \Omega_F}|_{\Omega_F=0.15} = \int_0^1 \frac{\partial E}{\partial \Omega_F} f(y) dy|_{\Omega_F=0.15} = 0, \quad (9.3.40)$$

where $f(y)$ is the PDF of $Y$. We now consider the first-order derivative of the variance of $E$ in terms of $\Omega_F$:

$$\frac{\partial (\sigma^2_E)}{\partial \Omega_F} = 2 \int_0^1 E \frac{\partial E}{\partial \Omega_F} f(y) dy - 2 \left( \int_0^1 E f(y) dy \right) \left( \int_0^1 \frac{\partial E}{\partial \Omega_F} f(y) dy \right). \quad (9.3.41)$$

Assuming that the optimal frequency does not depend on the amplitude, say, for any $Y$ the optimal frequency is $\Omega_F = \Omega_M$, we then have for any $Y$

$$\frac{\partial E}{\partial \Omega_F}|_{\Omega_F=\Omega_M} = 0, \quad (9.3.42)$$

which implies that both the mean and standard deviation of $E$ should reach the maximum values at $\Omega_F = \Omega_M$. However, this is a contradiction compared to the numerical results in figure 9.23. Thus, the optimal frequency should vary in terms of the amplitude. In figure 9.24, we demonstrate the dependence of optimal frequency on the amplitude using deterministic simulations, where it can

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
In [39] the excitation force with an amplitude 0.2 was studied and an optimal frequency for the nonlinear system was found to be 0.168, which agrees very well with our result 0.163. The above observations show that the choice of exact optimal frequency is a complicated problem, which is related to many factors. When randomness is present, a better way to select the optimal frequency is to check the mean response since it is more reasonable to assume that a factor, e.g., the amplitude, is random with a certain PDF.
Probability density functions of $E$

We now study the PDFs of the transport enhancement parameter $E$ under different types of random inputs. We study here only the case with $\Omega_f = 0.15$, which corresponds to the maximum response for $\eta = 0.2$. Based on the solutions of polynomial chaos, all PDFs shown in this section are obtained by a standard Monte Carlo method with 1,000,000 realizations.

First, we assume that $Y$ is a uniform random variable, i.e., $Y \sim U[0,1]$ and let $E = E(Y)$. Using the properties of PDF, we obtain the PDF of $E$

$$f_E(E) = E^{(1)}(Y(E))^{-1}, \quad (9.3.43)$$

where $(n)$ indicates the $n$-th order derivative. The first-order derivative of $f_E$ can be obtained from the chain rule of calculus as

$$f_E^{(1)}(E) = -\frac{E^{(2)}(Y(E))}{E^{(1)}(Y(E))^3}. \quad (9.3.44)$$

We now examine the PDF of $E$ at $Re = 225$, which is shown in figure 9.25. The PDF of $E$ is not uniform unlike the input and has a minimum at $E \approx 1.15$, which means that the second-order derivative of $E(Y)$ should be zero at $E \approx 1.15$ according to equation (9.3.44). Thus, the support of $E$ can be roughly divided into two parts:
Part I: $1 \leq E \leq 1.15$, where the probability density decreases and the first-order derivative of $f_E$ increases to zero.

Part II: $E \geq 1.15$, where the probability density increases and the first-order derivative of $f_E$ increases from zero.

Similar phenomena were shown in [39] for deterministic simulations at $Re = 225$. We subsequently check the local behavior of $E$. For $Y \ll 1$ ($E$ is around 1), we assume that

$$E - 1 \sim cY^\beta, \quad c \text{ is constant.} \quad (9.3.45)$$

It is easy to obtain the PDF of $E$ as

$$f_E(E) \sim c_E(E - 1)^{(1-\beta)/\beta}, \quad (9.3.46)$$

where $c_E$ is constant. Due to the singularity at $E = 1$, we obtain that $\beta > 1$ consistent with the deterministic estimation in [39], where $\beta$ is set to be 2 for small amplitudes. As the amplitude increases, such an estimate will be not valid. We know that $E^{(2)}(Y) = 0$ when $E \approx 1.15$. However, the second-order derivative of $E(Y) = cY^2$ is a nonzero constant, which is a contraction. Similarly, we analyze the part corresponding to large amplitudes with the assumption $E \sim c_1Y^\beta + c_2$, which yields that

$$f_E(E) \sim c_E(E - c_2)^{(1-\beta)/\beta}. \quad (9.3.47)$$

From figure 9.25, we can see that $(1 - \beta)/\beta \approx 1$ for large $E$, which yields that $\beta = \frac{1}{2}$. Such a result is the same as the deterministic estimation in [39].

In figure 9.26, we show the PDFs of $E$ at $Re = 525$ and 825. For comparison we include the results from high-order gPC. We use ME-gPC with a six-element mesh $[0,0.1,0.3,0.5,0.7,0.9,1]$ for $Re = 525$ and a eight-element mesh

$$[0,0.1,0.3,1.1/3,1.3/3,0.5,0.7,0.9,1.0]$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
for \( Re = 825 \). A similar two-part structure is observed in the PDF for \( Re = 525 \) but differences arise in the PDF for \( Re = 825 \). For large \( E \), the relation between \( E \) and probability density is not linear any more and becomes more complicated as the Reynolds number increases. For example, in the range \( E \geq 1.4 \), \( f_E^{(1)}(E) > 0 \) for \( Re = 525 \) while \( f_E^{(1)}(E) = 0 \) at \( E = 1.52 \) and 1.72 for \( Re = 825 \).

Using equation (9.3.44), we know that there exist two stationary points in the curve \( E = E(Y) \) for \( Re = 825 \) but no stationary points for \( Re = 525 \).

As a representative non-uniform random input, we consider the Beta distribution \( Beta(2, 2) \) in \([0, 1]\). In figure 9.27, the PDFs of \( E \) are shown for \( Re = 525 \) and 825. We use the same meshes as before for ME-gPC. Compared to the symmetric PDF of \( Beta(2, 2) \), the PDFs of \( E \) have a clear bias towards larger \( E \) and the bias shifts further as the Reynolds number increases. In table 9.2 we present the skewness \( s_E \) and kurtosis \( k_E \) of \( E \) for different random inputs and Reynolds numbers. In particular, we notice the negative \( s_E \) for \( Re = 525, 825 \) with the absolute values increasing with \( Re \). In other words, an asymmetric tail extending out to the left becomes stronger with \( Re \).

Although the PDF of \( E \) is not symmetric unlike the PDF of random inputs, it appears that we can estimate the mean of \( E \) using \( E(\bar{Y} = 0.5) \) for a moderate perturbation (\( \eta = 0.2 \)), where \( \bar{Y} = 0.5 \) is the mean of \( Y \). In table 9.3, we compare the mean of \( E \) and deterministic results \( E(\bar{Y} = 0.5) \), where the percentage of difference is also given. We can see that the difference is less than 10% for both uniform and Beta random inputs in the range \( Re \in [225, 825] \). For certain random inputs, the difference increases with the Reynolds number; for a fixed Reynolds number, the difference for the uniform inputs is larger than that for the Beta inputs because the PDF of \( Beta(2, 2) \) distribution has a maximum at \( Y = 0.5 \) and decreases to zero at the ends. However, such an approximate analysis is not expected to be valid for general random inputs with non-symmetric PDFs.

Since the stochastic amplitude is a random variable, any individual realization of the ME-gPC should correspond to a deterministic simulation; in figure 9.28 we plot the \( E \)-amplitude curves based on the ME-gPC solutions for different Reynolds number, where we also compare the ME-gPC predictions with deterministic simulations at some reference points. We see that indeed the deterministic solutions are included in the stochastic solutions. In deterministic simulations the common approach
to study the $E$-amplitude relation is to interpolate the results of $E$ at different amplitudes. However, such a procedure will usually involve many individual deterministic simulations to get a good approximation. Thus, ME-gPC provides an effective approach not only for stochastic studies but also for obtaining deterministic relations economically.
Figure 9.27: PDF of $E$ with Beta random inputs at $Re = 525, 825$.

Table 9.2: Skewness $s_E$ and kurtosis $k_E$ of $E$ for different random inputs and Reynolds numbers.

<table>
<thead>
<tr>
<th>$Re$</th>
<th>$s_E$(Uniform)</th>
<th>$k_E$(Uniform)</th>
<th>$s_E$(Beta)</th>
<th>$k_E$(Beta)</th>
</tr>
</thead>
<tbody>
<tr>
<td>225</td>
<td>0.10</td>
<td>1.65</td>
<td>0.05</td>
<td>2.13</td>
</tr>
<tr>
<td>525</td>
<td>-0.61</td>
<td>2.14</td>
<td>-0.88</td>
<td>3.51</td>
</tr>
<tr>
<td>825</td>
<td>-0.68</td>
<td>1.87</td>
<td>-1.20</td>
<td>3.30</td>
</tr>
</tbody>
</table>

Table 9.3: The mean and standard deviation of $E$ for different random inputs and the corresponding deterministic $E$ with the mean random inputs.

<table>
<thead>
<tr>
<th>$Re$</th>
<th>$E$(Uniform)</th>
<th>$\sigma_E$(Uniform)</th>
<th>$E$(Beta)</th>
<th>$\sigma_E$(Beta)</th>
<th>$E$(0.5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>225</td>
<td>1.1694(0.28%)</td>
<td>0.1196</td>
<td>1.1664(0.00%)</td>
<td>0.0829</td>
<td>1.1661</td>
</tr>
<tr>
<td>525</td>
<td>1.4045(3.65%)</td>
<td>0.2058</td>
<td>1.4344(1.60%)</td>
<td>0.1281</td>
<td>1.4577</td>
</tr>
<tr>
<td>825</td>
<td>1.5014(8.41%)</td>
<td>0.2766</td>
<td>1.5945(2.73%)</td>
<td>0.1945</td>
<td>1.6393</td>
</tr>
</tbody>
</table>

Figure 9.28: Curves of the enhancement ratio $E$ versus the excitation amplitude for different Reynolds numbers.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Effect of different frequencies on $E$

Given a periodic covariance kernel $K(t,s) = K(t + T, s + T)$, a second-order periodically correlated random process can be expressed by the Karhunen-Loeve (K-L) decomposition, which takes a form

$$X(t) = \sqrt{\frac{\lambda_0}{T}} Y_{0,0} + \sqrt{\frac{2}{T}} \sum_{n=1}^{\infty} \sqrt{\lambda_n} \left[ Y_{n,1} \cos \frac{2n\pi}{T} t + Y_{n,2} \sin \frac{2n\pi}{T} t \right],$$

(9.3.48)

where $T$ is the period and $\{Y_{i,j}\}$ is a set of uncorrelated random variables. The eigenvalues $\{\lambda_n\}$ can be determined by the correlation length and Fourier coefficients of the kernel $K(t,s)$ [67]. We assume that the first eigenvalue $\lambda_0 = 0$ since the corresponding eigenfunction is constant.

To focus on the interaction between $\Omega_F = 0.15$ and other frequencies, we first assume that the energy of each mode is of the same order and consider the following random condition

$$Q(t) = \frac{4}{3} \left(1 + \frac{0.2}{3} \sum_{i=1}^{3} a_i Y_i \sin(2\pi\Omega_{F,i} t)\right),$$

(9.3.49)

where $Y_i \sim U[0,1]$ are uniform i.i.d. random variables and $a_i = 0$ or $1$. In other words, we impose three frequencies with i.i.d. random amplitudes in the excitation force. The Reynolds number is set to $Re = 525$.

We investigate in detail two cases:

(I): $\Omega_{F,i} = 0.10, 0.15, 0.20$ and (II): $\Omega_{F,i} = 0.11, 0.15, 0.18$.

For both cases, we set $a_i = 1, i = 1, 2, 3$, which means that the standard deviation of each random amplitude is the same. We use ME-gPC with $N = 3 \times 4 \times 3 = 36$ and $p = 2$ to resolve the problem, where uniform meshes in each random dimension are employed. We select frequencies according to figure 9.23: frequencies that introduce smaller heat transfer enhancement are chosen as a perturbation of the frequency $\Omega_F = 0.15$.

We plot the spectra of mean and standard deviation of $E$ in figures 9.29-9.30 for case (I), and in figures 9.31-9.32 for case II, based on the ME-gPC solutions. We note that
<table>
<thead>
<tr>
<th>$(a_1, a_2, a_3)$</th>
<th>(1, 1, 1)</th>
<th>(1, 0, 0)</th>
<th>(0, 1, 0)</th>
<th>(0, 0, 1)</th>
<th>(3, 0, 0)</th>
<th>(0, 3, 0)</th>
<th>(0, 0, 3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E$</td>
<td>1.1793</td>
<td>1.0474</td>
<td>1.1488</td>
<td>1.0089</td>
<td>1.1998</td>
<td>1.4029</td>
<td>1.0683</td>
</tr>
<tr>
<td>$\sigma_E$</td>
<td>0.0834</td>
<td>0.0389</td>
<td>0.1142</td>
<td>0.0088</td>
<td>0.1290</td>
<td>0.2054</td>
<td>0.0556</td>
</tr>
</tbody>
</table>

Table 9.4: Effect of different frequencies on $E$ at $Re = 525$. Case I: $\Omega_F, \epsilon = 0.1, 0.15, 0.2$.

<table>
<thead>
<tr>
<th>$(a_1, a_2, a_3)$</th>
<th>(1, 1, 1)</th>
<th>(1, 0, 0)</th>
<th>(0, 1, 0)</th>
<th>(0, 0, 1)</th>
<th>(3, 0, 0)</th>
<th>(0, 3, 0)</th>
<th>(0, 0, 3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E$</td>
<td>1.1883</td>
<td>1.0523</td>
<td>1.1488</td>
<td>1.0142</td>
<td>1.2123</td>
<td>1.4029</td>
<td>1.1025</td>
</tr>
<tr>
<td>$\sigma_E$</td>
<td>0.0820</td>
<td>0.0423</td>
<td>0.1142</td>
<td>0.0133</td>
<td>0.1363</td>
<td>0.2054</td>
<td>0.0821</td>
</tr>
</tbody>
</table>

Table 9.5: Effect of different frequencies on $E$ at $Re = 525$. Case II: $\Omega_F, \epsilon = 0.11, 0.15, 0.18$.

- Frequency $\Omega_F = 0.15$ is dominant;
- A linear combination of other frequencies with $\Omega_F = 0.15$ is present.

It appears that there exists a nontrivial interaction between $\Omega_F = 0.15$ and other frequencies. For case (I), a subharmonic frequency 0.05 emerges in the spectra of mean and standard deviation; for case (II), more frequencies, e.g., 0.03 and 0.04, are involved. Since the energy from frequencies other than $\Omega_F = 0.15$ is small, the heat transfer enhancement will be determined mostly by the frequency $\Omega_F = 0.15$.

In tables 9.4 and 9.5, we compare the mean and standard deviation of $E$ for different random inputs. Since $\Omega_F = 0.15$ is the dominant frequency, we focus on the difference between one-dimensional random inputs with $\Omega_F = 0.15$ (corresponding to $(a_1, a_2, a_3) = (0,1,0)$) and the aforementioned two cases. The difference in mean is 2.7% and 3.4%, respectively, and the difference in standard deviation is 27% and 28%, respectively. In other words, the mean of $E$ is almost unchanged while the standard deviation is effectively reduced. Thus, it appears that the imposed frequencies, other than $\Omega_F = 0.15$, have a much stronger influence on the standard deviation of $E$ than on the mean value. Such a phenomenon should be the consequence of the previously mentioned interaction between $\Omega_F = 0.15$ and other frequencies.

In many cases, the periodicity exhibits its effect only as periodic variation at harmonics of a particular frequency. We subsequently consider the first four harmonics in the K-L expansion.
Figure 9.29: Spectra of the mean of $E$ at $Re = 525$ for case (I). $\Omega_{F_1} = 0.1, 0.15, 0.2$.

Figure 9.30: Spectra of the standard deviation of $E$ at $Re = 525$ for case (I). $\Omega_{F_1} = 0.1, 0.15, 0.2$. 

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Figure 9.31: Spectra of the mean of $E$ at $Re = 525$ for case (II). $\Omega_{F_{\Omega}} = 0.11, 0.15, 0.18$.

Figure 9.32: Spectra of the standard deviation of $E$ at $Re = 525$ for case (II). $\Omega_{F_{\Omega}} = 0.11, 0.15, 0.18$. 

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Table 9.6: Effect of superharmonic frequencies on $E$ at $Re = 525$ for case (III).

<table>
<thead>
<tr>
<th>$E(Q_1)$</th>
<th>$E(Q_2)$</th>
<th>$\sigma_E(Q_1)$</th>
<th>$\sigma_E(Q_2)$</th>
<th>$E(Q_1)/E(Q_2)$</th>
<th>$\sigma_E(Q_1)/\sigma_E(Q_2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.4261</td>
<td>1.4029</td>
<td>0.1675</td>
<td>0.2054</td>
<td>1.0165</td>
<td>0.8155</td>
</tr>
</tbody>
</table>

(9.3.48) using $T = 1/0.15$. Without loss of generality, we consider the following random input

(III): $Q(t) = \frac{4}{3}(1 + 0.1 \sin 2\pi \Omega_F t + \eta \sqrt{\frac{2}{T}} \sum_{i=1}^{4} \sqrt{\lambda_i} Y_i \sin(2\pi \Omega_F t))$, (9.3.50)

with $\Omega_F = 0.15$, where $Y_i \sim U[-1,1]$ are uniform i.i.d. random variables with zero mean. We consider the following eigenvalues obtained from the periodic covariance kernel

$\lambda_i = 1.2266, 0.6504, 0.3080$ and 0.1493.

We take $\eta = 0.1/\sqrt{2\lambda_1/T}$ and the flow rate (9.3.50) can be rewritten as

$Q(t) = \frac{4}{3}(1 + 0.2Y \sin 2\pi \Omega_F t + S(t, \omega))$,

where $Y$ is a uniform random variable in $[0,1]$ as before, and $S(t, \omega)$ can be regarded as the perturbation from the super harmonics. We use ME-gPC with $N = 8 \times 4 \times 2 \times 1 = 64$ and $p = 2$ to achieve numerical convergence. Let $Q_1$ indicate the above flow rate and $Q_2$ the flow rate without $S(t, \omega)$. In Table 9.6, we present the mean and standard deviation of $E$ for $Q_1$ and $Q_2$. We observe that the superharmonic frequencies increase $\bar{E}$ about 1.7% while decrease $\sigma_E$ about 18.5%, which is consistent with the observations from case (I) and (II).

In summary, for a given geometry and random (periodically correlated) flow rate condition, there exists a critical frequency which is most effective for heat transfer enhancement. Furthermore, to reach the maximum heat transfer enhancement, such a critical frequency should be related to the largest eigenvalue, i.e., the largest degree of perturbation. Other frequencies are, in general, not effective for increasing $\bar{E}$; however, they can significantly decrease $\sigma_E$. 

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Correlation between cross-flow velocity and temperature

In [39] the relation between cross-flow velocity and heat transfer enhancement was studied by deterministic simulations. The approach was to compare the optimal frequencies for cross-flow velocity and $E$. It was found that the two optimal frequencies agree with each other qualitatively although there exists a slight difference. We know that the cross-flow velocity is mainly due to the oscillation imposed in the inflow boundary condition [38]. Here we measure the relation between cross-flow velocity and temperature quantitatively using the correlation at reference points, which in turn reflects the influence of the imposed oscillating flow rate condition on the heat transfer enhancement.

Along the straight line $x = 3.20, -1.89 < y < 0.73$, we select 40 equidistant grid points, denoted as $y_i, i = 1, 2, \cdots, 40$. Specifically, we define the following five reference points (see figure 9.22): $A(3.20, -0.90), B(3.20, -1.89), C(3.20, -1.22), D(3.20, -0.80)$ and $E(3.20, 0.71)$. We use $v_i$ to denote the cross-flow velocity at grid $y_i$, and $T_j$ the temperature at grid $y_j$. Let $\rho$ denote the correlation between $v_i$ and $T_j$. For each pair $(v_i, T_j)$, we have the correlation

$$\rho_{ij} = \frac{E[(v_i - \bar{v}_i)(T_j - \bar{T}_j)]}{\sigma_{v_i}\sigma_{T,j}}, \quad (9.3.51)$$

where $\sigma_{v,i}$ denotes the standard deviation of the variable $v$. We know that $-1 \leq \rho_{ij} \leq 1$ and a large absolute values of $\rho_{ij}$ corresponds to a large correlation. In figure 9.33, we plot the correlation matrix $\rho_{ij}$ for uniform inputs with $\Omega_F = 0.15$ at $t = \hat{T}$, where $\hat{T}$ is the period of converged solutions. We can see that $|\rho_{ij}| \approx 1$ except a narrow region, which implies that the cross-flow velocity and temperature are perfectly correlated. For any fixed $v_i$, $\rho_{ij} \approx -1$ for $T_j$ at grid points $y_j \in [-1.89, -0.90]$, corresponding to the groove part; $\rho_{ij} \approx 1$ for $T_j$ at grid points $y_j \in [-0.62, 0.73]$, corresponding to the channel part; $\rho_{ij}$ increases quickly from $-1$ to $1$ for $T_j$ within the narrow interval $y_j \in [-0.90, -0.62]$. These observations are consistent with the flow physics. It is known that the cold fluid in the channel part would be forced into the downstream side of the cavity by the T-S traveling wave, then push the hot fluid from the upstream part of the groove into the channel part. Hence, if the cross-flow velocity $v_i$ at a certain point $y_i$ tends to increase, the temperature in the
Figure 9.33: Correlation matrix $\rho_{ij}$ for velocity-temperature pairs $(v_i, T_j)$ on equidistant grid points along the straight line $x = 3.20, -1.89 \leq y \leq 0.73$. Uniform random inputs are used. $\Omega_F = 0.15$ and $Re = 525$.

groove tends to decrease, corresponding to $\rho_{ij} \approx -1$, since the heat is from the groove to channel; the temperature in the channel tends to increase, corresponding to $\rho_{ij} \approx 1$. Thus, the correlation $\rho_{ij}$ indicates that the relation between the trends of $v_i$ and $T_j$ is almost linear, which implies that the evolution of temperature can be perfectly reflected by the evolution of cross-flow velocity.

In figure 9.34 we plot the time evolution of correlation between point $A$ and other defined points ($B, C, D$ and $E$) for $\Omega = 0.15$ on the left and for $\Omega_F = 0.18$ on the right. We see that the correlation value switches between $-1$ and $1$ periodically except for the points ($\rho_{AD}$) in the aforementioned narrow transient interval. The correlation for $\Omega_F = 0.18$ shows a sharper transition between $-1$ and $1$, which implies that the cross-flow velocity and temperature are more correlated for $\Omega_F = 0.18$. This does not conflict with the fact that the optimal frequency is $\Omega_F = 0.15$. Since the optimal frequency corresponds to a larger response in the cross-flow velocity, which implies that the induced uncertainty is also larger, it is reasonable that the transition region of $\rho_{ij}$ is larger for the optimal frequency. Since similar observations are obtained for other random inputs, the results are not presented here.
Figure 9.34: Evolution of $\rho_{ij}$, $Y \sim U[0,1]$ and $Re = 525$. Left: $\Omega_F = 0.15$; Right: $\Omega_F = 0.18$.

**Visualizations**

In this section we present the visualization of the statistics of velocity and temperature. In figure 9.35 and 9.36, we show the mean (a1-a4) and standard deviation (b1-b4) of velocity and temperature fields, respectively, for $Re = 525$ within one period. Since there exists qualitative similarity between the corresponding statistics for the different random inputs, we present here only results for the uniform random inputs.

In figure 9.35, significant mixing at the groove lip is observed in the mean of cross-flow velocity. The largest standard deviation is located near the groove lip and decreases into the channel and groove, which implies that the communication between the channel and groove is very sensitive to the excitation. In figure 9.36, a two-wave structure is observed in isocontours of the mean temperature field. Compared to corresponding deterministic results (see figure 3 in [39]), the two-wave structure is not as pronounced because the amplitude 0.2 was used in [39] while in our simulations the mean amplitude is 0.1 for a uniform random amplitude in $[0,0.2]$. The largest standard deviation is inside the groove and along the groove lip; however, the standard deviation is small in the channel. By noting the large cross-flow velocity along the groove lip and the property of correlation matrix $\rho_{ij}$ discussed in the previous section, we know that it is the convection process that introduces large uncertainty to the temperature in the groove.
Figure 9.35: Mean (a1-a4) and standard deviation (b1-b4) of cross-flow velocity at $t = T/4, T/2, 3T/4, T$ with $T$ being the period. $Y \sim U[0, 1]$, $\Omega_F = 0.15$ and $Re = 525$. 

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Figure 9.36: Mean (a1-a4) and standard deviation (b1-b4) of temperature at $t = \hat{T}/4, \hat{T}/2, 3\hat{T}/4, \hat{T}$ with $\hat{T}$ being the period. $Y \sim U[0, 1]$, $\Omega_F = 0.15$ and $Re = 525$. 

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Chapter 10

Summary and open issues

In this dissertation we developed a multi-element generalized polynomial chaos (ME-gPC) method to study differential equations with random inputs. In the ME-gPC method, the parametric space is decomposed into non-overlapping elements, and generalized polynomial chaos approximation is subsequently implemented in each element based on numerical orthogonality. The ME-gPC method provides dual paths of convergence, i.e., $hp$-convergence, which is valid for both Galerkin projection and collocation projection of ME-gPC. We subsequently proposed adaptive ME-gPC methods to improve the efficiency. Two adaptive criteria are developed. One is based on the decay rate of coefficients of local polynomial chaos expansion; the other one is based on a posteriori error estimators. We used the former (heuristic) one to study the Kraichnan-Orszag's problem, which demonstrates low regularity in the parametric space. The adaptive ME-gPC can capture the random behavior up to a desired accuracy by $h$-convergence in the region of low regularity and $p$-convergence in the smooth region. We studied elliptic problems with random coefficients using the latter adaptive criterion. To reduce the cost of solving the error equations, we developed a reduced space, which results in up to 90% saving in computation time. The a posteriori error estimator provides a good prediction of the approximation error of local polynomial chaos expansions. In the adaptive ME-gPC scheme, we first chose elements with the largest local a posteriori error estimators, then split random dimensions of the most importance. Numerical experiments show that the adaptive ME-gPC...
method demonstrates much better efficiency than Monte Carlo methods for a moderate number of random dimensions. Long-term integration is still an open problem for the efficiency of polynomial chaos methods, especially for hyperbolic problems. We compared the performance of generalized polynomial chaos (gPC) and ME-gPC for long-term integration using a hyperbolic model problem. The efficiency of gPC degrades due to the fast expansion of spectrum in the parametric space with time. Using the decomposition of the parametric space, the ME-gPC method can extend the valid integration time algebraically for a given accuracy. The ME-gPC method has also been applied to physical applications including heat diffusion in a 3D chip, noisy flow past 2D stationary circular cylinder and heat transfer enhancement in a grooved channel.

The adaptive ME-gPC methods can be much more efficient than Monte Carlo methods for a moderate number $O(10)$ of random dimensions, especially when the solution has enough regularity in the parametric space. When the number of random dimensions increases, the efficiency of the polynomial chaos methods will degrade and fail due to the curse of dimensionality. Monte Carlo methods are still a preferable choice for high-dimensional cases because its dependence on dimensionality is weak.

Many open issues remain in this field. We list some of them as follows:

- **Long-time integration.** Polynomial chaos methods work well for elliptic and parabolic problems. However, they often fail for hyperbolic problems in a short time due to the fast expansion of spectrum in the parametric space. So far, there are no effective non-sampling numerical techniques to deal with the random frequency problems rather than Monte Carlo methods.

- **Efficient adaptivity.** The fast convergence of polynomial chaos expansion or sparse grid interpolation relies on the smoothness of the solution in the parametric space with the assumption of equal importance in each direction. However, adaptivity is necessary for the region with low regularity or the direction with more importance. Effective adaptive strategies are necessary for both $h$- and $p$-convergence for certain problems.

- **General random inputs.** Random inputs in applications are often modeled by Gaussian
processes, which may be inaccurate in many applications. Simulations of non-Gaussian random processes have received more attention recently. It is an interesting problem how to couple the non-Gaussian simulations with polynomial chaos methods to accelerate the stochastic numerical simulations for general random inputs, e.g., experimental data.
Bibliography


Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.


Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.


[125] O. C. Zienkiewicz and J. Z. Zhu. The superconvergent path recovery and a posteriori error