

# **Uncertainty Quantification**

---

# Computational Science & Engineering

The SIAM series on Computational Science and Engineering publishes research monographs, advanced undergraduate- or graduate-level textbooks, and other volumes of interest to an interdisciplinary CS&E community of computational mathematicians, computer scientists, scientists, and engineers. The series includes introductory volumes aimed at a broad audience of mathematically motivated readers interested in understanding methods and applications within computational science and engineering, monographs reporting on the most recent developments in the field, and volumes addressed to specific groups of professionals whose work relies extensively on computational science and engineering.

SIAM created the CS&E series to support access to the rapid and far-ranging advances in computer modeling and simulation of complex problems in science and engineering, to promote the interdisciplinary culture required to meet these large-scale challenges, and to provide the means to the next generation of computational scientists and engineers.

---

## Editor-in-Chief

Donald Estep  
Simon Fraser University

## Editorial Board

|                                 |  |                                 |
|---------------------------------|--|---------------------------------|
| Ben Adcock                      | Serkan Gugercin                          | David Keyes                     |
| Simon Fraser University         | Virginia Tech                            | Columbia University             |
| Daniela Calvetti                | Jan S. Hesthaven                         | Ralph C. Smith                  |
| Case Western Reserve University | Ecole Polytechnique Fédérale de Lausanne | North Carolina State University |
| Omar Ghattas                    | Johan Hoffman                            | Karen Willcox                   |
| University of Texas at Austin   | KTH Royal Institute of Technology        | University of Texas at Austin   |
| Chen Greif                      |  |                                 |
| University of British Columbia  |  |                                 |

Smith, Ralph C., *Uncertainty Quantification: Theory, Implementation, and Applications*, Second Edition

Basu, Samopriya, Butler, Troy, Estep, Don, and Panda, Nishant, *A Ramble through Probability: How I Learned to Stop Worrying and Love Measure Theory*

Demkowicz, Leszek F., *Mathematical Theory of Finite Elements*  
Rozza, Gianluigi, Stabile, Giovanni, and Ballarin, Francesco, *Advanced Reduced Order Methods and Applications in Computational Fluid Dynamics*

Gatto, Paolo, *Mathematical Foundations of Finite Elements and Iterative Solvers*

Adcock, Ben, Brugiapaglia, Simone, and Webster, Clayton G., *Sparse Polynomial Approximation of High-Dimensional Functions*

Hoffman, Johan, *Methods in Computational Science*

Da Veiga, Sébastien, Gamboa, Fabrice, Iooss, Bertrand, and Prieur, Clémantine, *Basics and Trends in Sensitivity Analysis: Theory and Practice in R*

Vidyasagar, M., *An Introduction to Compressed Sensing*

Antoulas, A. C., Beattie, C. A., and Gügencin, S., *Interpolatory Methods for Model Reduction*

Sipahi, Rifat, *Mastering Frequency Domain Techniques for the Stability Analysis of LTI Time Delay Systems*

Bardsley, Johnathan M., *Computational Uncertainty Quantification for Inverse Problems*

Hesthaven, Jan S., *Numerical Methods for Conservation Laws: From Analysis to Algorithms*

Sidi, Avram, *Vector Extrapolation Methods with Applications*

Borzi, A., Ciaramella, G., and Sprengel, M., *Formulation and Numerical Solution of Quantum Control Problems*

Benner, Peter, Cohen, Albert, Ohlberger, Mario, and Willcox, Karen, editors, *Model Reduction and Approximation: Theory and Algorithms*

Kuzmin, Dmitri and Hämäläinen, Jari, *Finite Element Methods for Computational Fluid Dynamics: A Practical Guide*

Rostamian, Rouben, *Programming Projects in C for Students of Engineering, Science, and Mathematics*

Smith, Ralph C., *Uncertainty Quantification: Theory, Implementation, and Applications*

Dankowicz, Harry and Schilder, Frank, *Recipes for Continuation*

Mueller, Jennifer L. and Siltanen, Samuli, *Linear and Nonlinear Inverse Problems with Practical Applications*

Shapiro, Yair, *Solving PDEs in C++: Numerical Methods in a Unified Object-Oriented Approach*, Second Edition

Borzi, Alfio and Schulz, Volker, *Computational Optimization of Systems Governed by Partial Differential Equations*

Ascher, Uri M. and Greif, Chen, *A First Course in Numerical Methods*

Layton, William, *Introduction to the Numerical Analysis of Incompressible Viscous Flows*

Ascher, Uri M., *Numerical Methods for Evolutionary Differential Equations*

Zohdi, T. I., *An Introduction to Modeling and Simulation of Particulate Flows*

Biegler, Lorenz T., Ghattas, Omar, Heinkenschloss, Matthias, Keyes, David, and van Bloemen Waanders, Bart, editors, *Real-Time PDE-Constrained Optimization*

Chen, Zhangxin, Huan, Guanren, and Ma, Yuanle, *Computational Methods for Multiphase Flows in Porous Media*

Shapiro, Yair, *Solving PDEs in C++: Numerical Methods in a Unified Object-Oriented Approach*

# Uncertainty Quantification

## Theory, Implementation, and Applications

### Second Edition

RALPH C. SMITH  
North Carolina State University  
Raleigh, North Carolina



Society for Industrial and Applied Mathematics  
Philadelphia

Copyright © 2024 by the Society for Industrial and Applied Mathematics

10 9 8 7 6 5 4 3 2 1

All rights reserved. Printed in the United States of America. No part of this book may be reproduced, stored, or transmitted in any manner without the written permission of the publisher. For information, write to the Society for Industrial and Applied Mathematics, 3600 Market Street, 6th Floor, Philadelphia, PA 19104-2688 USA.

No warranties, express or implied, are made by the publisher, authors, and their employers that the programs contained in this volume are free of error. They should not be relied on as the sole basis to solve a problem whose incorrect solution could result in injury to person or property. If the programs are employed in such a manner, it is at the user's own risk and the publisher, authors, and their employers disclaim all liability for such misuse.

Trademarked names may be used in this book without the inclusion of a trademark symbol. These names are used in an editorial context only; no infringement of trademark is intended.

MATLAB is a registered trademark of The MathWorks, Inc. For MATLAB product information, please contact The MathWorks, Inc., 3 Apple Hill Drive, Natick, MA 01760-2098 USA, 508-647-7000, Fax: 508-647-7001, [info@mathworks.com](mailto:info@mathworks.com), [www.mathworks.com](http://www.mathworks.com).

|                        |                     |
|------------------------|---------------------|
| Publications Director  | Kivmars H. Bowling  |
| Executive Editor       | Elizabeth Greenspan |
| Acquisitions Editor    | Elizabeth Greenspan |
| Developmental Editor   | Rose Kolassiba      |
| Managing Editor        | Kelly Thomas        |
| Production Editor      | David Riegelhaupt   |
| Copy Editor            | Susan Fleshman      |
| Production Manager     | Rachel Ginder       |
| Production Coordinator | Cally A. Shrader    |
| Compositor             | Cheryl Hufnagle     |
| Graphic Designer       | Doug Smock          |

#### Library of Congress Cataloging-in-Publication Data

Names: Smith, Ralph C., 1960- author.

Title: Uncertainty quantification : theory, implementation, and applications / Ralph C. Smith, North Carolina State University, Raleigh, North Carolina.

Description: Second edition. | Philadelphia : Society for Industrial and Applied Mathematics, [2024] | Series: Computational science and engineering ; 30 | Includes bibliographical references and index. | Summary: "Covers fundamental topics pertaining to sensitivity analysis and uncertainty quantification and illustrates them in the context of applications from science and engineering"-- Provided by publisher.

Identifiers: LCCN 2024011203 (print) | LCCN 2024011204 (ebook) | ISBN 9781611977837 | ISBN 9781611977844 (ebook)

Subjects: LCSH: Measurement uncertainty (Statistics) | Estimation theory.

Classification: LCC QA276.8 .S64 2024 (print) | LCC QA276.8 (ebook) | DDC 519.5/44--dc23/eng/20240326

LC record available at <https://lccn.loc.gov/2024011203>

LC ebook record available at <https://lccn.loc.gov/2024011204>

# Contents

|   |              |
|---|--------------|
| <b>Preface</b>  | <b>xi</b>    |
| <b>Notation</b>   | <b>xvii</b>  |
| <b>Acronyms and Initialisms</b>                                       | <b>xxi</b>   |
| <b>UQ Crimes</b>  | <b>xxiii</b> |
| <b>1 Introduction</b>   | <b>1</b>     |
| 1.1 Motivating Examples and Applications . . . . .                    | 1            |
| 1.2 Sources and Types of Uncertainties and Errors . . . . .           | 7            |
| 1.3 Sensitivity Analysis and Uncertainty Quantification Process . . . | 11           |
| 1.4 Omitted Topics . . . . .  | 15           |
| <b>I APPLICATIONS AND MODELS</b>                                      | <b>17</b>    |
| <b>2 Applications</b>   | <b>19</b>    |
| 2.1 Nuclear Power Plant Design . . . . .                              | 19           |
| 2.2 Quantitative Systems Pharmacology (QSP) Models . . . . .          | 25           |
| 2.3 Weather Models . . . . .  | 29           |
| 2.4 Climate Models . . . . .  | 34           |
| 2.5 Digital Twins . . . . .   | 41           |
| 2.6 Virtual Populations . . . . .                                     | 47           |
| <b>3 Models and Data</b>  | <b>51</b>    |
| 3.1 Algebraic Models . . . . .  | 51           |
| 3.2 ODE Models . . . . .  | 52           |
| 3.3 PDE Models . . . . .  | 59           |
| 3.4 Exercises . . . . .   | 65           |

|   |            |
|---|------------|
| <b>II PROBABILITY AND STATISTICS CONCEPTS</b>                                   | <b>67</b>  |
| <b>4 Topics from Probability and Statistics</b>                                 | <b>69</b>  |
| 4.1 Random Variables and Distributions . . . . .                                | 69         |
| 4.2 Estimators, Sampling Distributions, and Confidence Intervals . . . . .      | 82         |
| 4.3 Ordinary Least Squares and Maximum Likelihood Estimators . . . . .          | 86         |
| 4.4 Random Processes and Fields . . . . .                                       | 89         |
| 4.5 Markov Chains . . . . .   | 91         |
| 4.6 Random versus Stochastic Differential Equations . . . . .                   | 94         |
| 4.7 Statistical Inference . . . . .   | 96         |
| 4.8 Energy Statistics . . . . .   | 102        |
| 4.9 References . . . . .  | 104        |
| 4.10 Exercises . . . . .  | 105        |
| <b>5 Representation of Random Parameters and Fields</b>                         | <b>107</b> |
| 5.1 Random Model Inputs . . . . .   | 107        |
| 5.2 Representation of Random Fields . . . . .                                   | 108        |
| 5.3 Choices for the Covariance Function $c(x, y)$ . . . . .                     | 110        |
| 5.4 Truncation of the Karhunen–Loève Expansion . . . . .                        | 114        |
| 5.5 Sampled Covariance Function $c$ and Eigenpair $\lambda_n, \phi_n$ . . . . . | 115        |
| 5.6 Gaussian versus Non-Gaussian Random Fields . . . . .                        | 118        |
| 5.7 Exercises . . . . .   | 120        |
| <b>6 Observation Models</b>   | <b>123</b> |
| 6.1 Physical Observation Models . . . . .                                       | 124        |
| 6.2 Statistical Observation Models . . . . .                                    | 125        |
| 6.3 Mixed-Effects Models . . . . .  | 130        |
| 6.4 Properties and UQ Crimes Concerning Observation Models . . . . .            | 132        |
| 6.5 Exercises . . . . .   | 134        |
| <b>III PARAMETER SELECTION TECHNIQUES</b>                                       | <b>137</b> |
| <b>7 Parameter Identifiability and Influence</b>                                | <b>139</b> |
| 7.1 Examples . . . . .  | 140        |
| 7.2 Parameter Identifiability . . . . .   | 144        |
| 7.3 Parameter Correlation versus Identifiability . . . . .                      | 154        |
| 7.4 Parameter Influence . . . . .   | 156        |
| 7.5 Parameter Selection Techniques . . . . .                                    | 158        |
| 7.6 Notes and References . . . . .  | 159        |
| 7.7 Exercises . . . . .   | 160        |
| <b>8 Local Sensitivity Analysis</b>   | <b>161</b> |
| 8.1 Analytic Sensitivities . . . . .  | 164        |
| 8.2 Sensitivity Equations for ODEs . . . . .                                    | 165        |
| 8.3 Finite-Difference Approximations . . . . .                                  | 169        |
| 8.4 Complex-Step Approximations . . . . .                                       | 169        |

---

|           |  |            |
|-----------|--|------------|
| 8.5       | Automatic Differentiation . . . . .                              | 178        |
| 8.6       | Adjoint Methods . . . . .  | 179        |
| 8.7       | Scaling Techniques . . . . .                                     | 183        |
| 8.8       | Parameter Subset Selection (PSS) . . . . .                       | 184        |
| 8.9       | Exercises . . . . .  | 192        |
| <b>9</b>  | <b>Global Sensitivity Analysis</b>                               | <b>195</b> |
| 9.1       | Normalized Sample Spaces and Derivative Approximations . . . . . | 197        |
| 9.2       | Variance-Based Methods . . . . .                                 | 198        |
| 9.3       | Derivative-Based Global Sensitivity Indices . . . . .            | 210        |
| 9.4       | Morris Screening . . . . .                                       | 212        |
| 9.5       | Time- or Space-Dependent Responses . . . . .                     | 216        |
| 9.6       | Strategy for Large-Scale Problems . . . . .                      | 224        |
| 9.7       | Notes and References . . . . .                                   | 228        |
| 9.8       | Exercises . . . . .  | 230        |
| <b>10</b> | <b>Active Subspace Analysis</b>                                  | <b>233</b> |
| 10.1      | Classical Subspace Analysis for Linear Models . . . . .          | 234        |
| 10.2      | Active Subspace Analysis for Nonlinear Models . . . . .          | 242        |
| 10.3      | Activity Scores for Global Sensitivity Analysis . . . . .        | 246        |
| 10.4      | Response Surface Models . . . . .                                | 247        |
| 10.5      | Examples . . . . .   | 247        |
| 10.6      | Large-Scale Application: Neutronics . . . . .                    | 252        |
| 10.7      | Notes and References . . . . .                                   | 253        |
| 10.8      | Exercises . . . . .  | 254        |
| <b>IV</b> | <b>INVERSE AND FORWARD UNCERTAINTY QUANTIFICATION</b>            | <b>257</b> |
| <b>11</b> | <b>Frequentist Parameter Inference</b>                           | <b>259</b> |
| 11.1      | Linear Regression . . . . .                                      | 261        |
| 11.2      | Prediction Intervals for Linear Problems . . . . .               | 269        |
| 11.3      | Nonlinear Regression . . . . .                                   | 272        |
| 11.4      | Notes and References . . . . .                                   | 280        |
| 11.5      | Exercises . . . . .  | 280        |
| <b>12</b> | <b>Bayesian Parameter Inference</b>                              | <b>285</b> |
| 12.1      | Bayesian Inference . . . . .                                     | 287        |
| 12.2      | Markov Chain Monte Carlo (MCMC) Techniques . . . . .             | 290        |
| 12.3      | Metropolis and Metropolis–Hastings Algorithms . . . . .          | 291        |
| 12.4      | Stationary Distribution and Convergence Criteria . . . . .       | 301        |
| 12.5      | Delayed Rejection Adaptive Metropolis (DRAM) . . . . .           | 307        |
| 12.6      | Alternative Algorithms . . . . .                                 | 314        |
| 12.7      | Bayesian Inference on Active Subspaces . . . . .                 | 316        |
| 12.8      | Large-Scale Example: Wetland Methane Emission Model . . . . .    | 320        |

|           |   |            |
|-----------|---|------------|
| 12.9      | Notes and References . . . . .  | 324        |
| 12.10     | Exercises . . . . .   | 327        |
| <b>13</b> | <b>Uncertainty Propagation for Model Responses</b>                            | <b>331</b> |
| 13.1      | Direct Evaluation for Linear Models . . . . .                                 | 334        |
| 13.2      | Perturbation Methods Based on Linearization . . . . .                         | 337        |
| 13.3      | Frequentist Prediction Intervals for Nonlinear Parameterized Models . . . . . | 339        |
| 13.4      | Sampling Methods for Nonlinearly Parameterized Models . . . . .               | 340        |
| 13.5      | Examples . . . . .  | 342        |
| 13.6      | Large-Scale Examples . . . . .  | 347        |
| 13.7      | Notes and References . . . . .  | 349        |
| 13.8      | Exercises . . . . .   | 349        |
| <b>14</b> | <b>Model Discrepancy</b>  | <b>353</b> |
| 14.1      | Issues Pertaining to Model Discrepancy . . . . .                              | 356        |
| 14.2      | Additional Examples of Model Discrepancy . . . . .                            | 358        |
| 14.3      | Physics-Informed Techniques to Address Model Discrepancy in $f$ . . . . .     | 362        |
| 14.4      | Additive Model Discrepancy . . . . .  | 368        |
| 14.5      | Discrepancy in the Observation Model $g$ . . . . .                            | 373        |
| 14.6      | Notes and References . . . . .  | 373        |
| 14.7      | Exercises . . . . .   | 374        |
| <b>V</b>  | <b>SURROGATE AND REDUCED-ORDER MODELS</b>                                     | <b>377</b> |
| <b>15</b> | <b>Surrogate Models</b>   | <b>379</b> |
| 15.1      | High-Fidelity Mathematical Models and Surrogate Frameworks . . . . .          | 380        |
| 15.2      | Statistical Observation Framework . . . . .                                   | 383        |
| 15.3      | Choice of Coefficient Values $\mathbf{q}^m$ . . . . .                         | 384        |
| 15.4      | General Form of Surrogate Models . . . . .                                    | 388        |
| 15.5      | Exercises . . . . .   | 390        |
| <b>16</b> | <b>Numerical Surrogate Models</b>   | <b>391</b> |
| 16.1      | Polynomial Surrogates . . . . .   | 391        |
| 16.2      | Spectral Surrogates . . . . .   | 400        |
| 16.3      | Radial Basis Function Surrogates . . . . .                                    | 412        |
| 16.4      | Neural Network Representations . . . . .                                      | 413        |
| 16.5      | Spline-Based Surrogate Models . . . . .                                       | 415        |
| 16.6      | Large-Scale Example . . . . .   | 415        |
| 16.7      | Exercises . . . . .   | 417        |
| <b>17</b> | <b>Spectral Surrogates for Differential Equations</b>                         | <b>419</b> |
| 17.1      | Objectives for Spectral Representations . . . . .                             | 421        |
| 17.2      | Scalar Initial Value Problem . . . . .  | 422        |
| 17.3      | Boundary Value Problems and Stationary PDEs . . . . .                         | 428        |
| 17.4      | Time-Dependent PDEs . . . . .   | 432        |

---

|                     |  |            |
|---------------------|--|------------|
| 17.5                | Attributes of the Galerkin, Collocation, and Discrete Projection Methods . . . . . | 436        |
| 17.6                | Software Packages . . . . .  | 439        |
| 17.7                | Exercises . . . . .  | 439        |
| <b>18</b>           | <b>Statistical Surrogate Models</b>  | <b>441</b> |
| 18.1                | Gaussian Process Surrogates . . . . .  | 441        |
| 18.2                | Mean and Covariance Functions . . . . .  | 443        |
| 18.3                | Estimation of Hyperparameters . . . . .  | 446        |
| 18.4                | Gaussian Process Predictions . . . . .   | 448        |
| 18.5                | Large-Scale Applications . . . . .   | 456        |
| 18.6                | Software and Additional References . . . . .                                       | 459        |
| 18.7                | Exercises . . . . .  | 460        |
| <b>19</b>           | <b>Reduced-Order Models</b>  | <b>461</b> |
| 19.1                | Projection-Based Methods . . . . .   | 461        |
| 19.2                | Snapshot Sets and Greedy Sampling Algorithms . . . . .                             | 467        |
| 19.3                | Proper Orthogonal Decompositions (POD) . . . . .                                   | 469        |
| 19.4                | Large-Scale Examples . . . . .   | 479        |
| 19.5                | Omitted Topics and Additional References . . . . .                                 | 486        |
| 19.6                | Exercises . . . . .  | 486        |
| <b>20</b>           | <b>Numerical and Statistical Integration Techniques</b>                            | <b>489</b> |
| 20.1                | Statistical Integration Methods . . . . .  | 490        |
| 20.2                | Numerical 1-D Quadrature Techniques . . . . .                                      | 490        |
| 20.3                | Numerical Quadrature Techniques in $\mathbb{R}^p$ . . . . .                        | 500        |
| 20.4                | Sparse Grid Software . . . . .   | 507        |
| 20.5                | Exercises . . . . .  | 507        |
| <b>A</b>            | <b>Supporting Material</b>   | <b>509</b> |
| <b>Bibliography</b> |  | <b>511</b> |
| <b>Index</b>        |  | <b>539</b> |



# Preface

Uncertainty quantification serves a central role for simulation-based analysis of physical, engineering, and biological applications using mechanistic models. From a broad perspective, the field of uncertainty quantification can be described as the synthesis of mathematical, statistical, and computational theory and methods to quantify uncertainties associated with mechanistic models and their parameters, simulation codes, observed data, and predicted responses for applications whose complexity can preclude sole reliance on sampling-based methods. Hence the field is inherently interdisciplinary and can require the synthesis of theory inherent to considered applications.

This second edition of *Uncertainty Quantification: Theory, Implementation, and Applications* reflects both substantial advances in the field over the last decade and significant lessons learned while presenting topics pertaining to sensitivity analysis (SA) and uncertainty quantification (UQ) to colleagues and students. While this book draws upon the same foundations as the first edition, the scope differs significantly in the manner that it presents topics pertaining to parameter selection and sensitivity analysis, parameter inference and uncertainty propagation, and surrogate and reduced-order model development to illustrate the requisite theory and establish implementation techniques required for models employed in engineering and the physical and biological sciences. This includes the illustration of these techniques for several large-scale, partial differential equation (PDE) models for a variety of applications.

## Features of the Book

Advances in the field include both the development of new theory and algorithms and emerging applications for which sensitivity analysis and uncertainty quantification serve critical roles. The former includes the development of parameter subset selection algorithms, sensitivity analysis techniques, and active subspace theory to isolate subsets and subspaces of influential parameters for subsequent uncertainty analysis. We also illustrate highly effective algorithms for constructing surrogate and reduced-order models. Highlighted large-scale applications include digital twins – for applications including aerospace design, monitoring of nuclear power plants, and precision medicine – and virtual populations to support drug discovery and guide clinical trials. In both cases, we note the central role of sensitivity analysis and uncertainty quantification. To demonstrate the workflow regarding uncertainty

quantification for a large-scale application, we provide examples illustrating the end-to-end process in the context of PDE models employed for nuclear power plant analysis. We also illustrate surrogate model construction for MCNP simulations used to quantify radiation in an urban environment.

The suite of large-scale examples is augmented by the illustration of Bayesian inference and uncertainty quantification for a wetland methane emission model. We also include examples illustrating the construction of Gaussian process emulators for volcanic flows, reduced-order model implementation for a chemical vapor deposition reactor, and the construction of reduced-order models to optimize heat exchange capabilities for a thermal fin and facilitate uncertainty quantification for a subsonic rotor blade coupled to an unsteady flow. We note these separately since they represent research from a range of scientists in the field.

The topics and formulation into Parts I–V reflect the background, workflow, and supporting material generally required for sensitivity analysis and uncertainty quantification for engineering, physical, and biological applications.

- Part I focuses on large-scale applications and prototypical models, several of which are presented with experimental data. The prototypical models quantify the Helmholtz energy for smart materials, height dynamics (with data), exponential processes, spring dynamics, sprinter dynamics (with data), disease dynamics (with data), heat properties (with data), beam dynamics, and the Navier–Stokes equations for fluid flow. These models and data are employed extensively in the examples used throughout the book to motivate and illustrate concepts.
- Part II covers background concepts from probability and statistics, the representation of random inputs and fields, and the formulation of physical and statistical observation models.
- The next step in the workflow generally entails identifiability analysis, sensitivity analysis and active subspace techniques, as detailed in Part III. The objective is to determine subsets and subspaces of influential or identifiable parameters, which are employed for subsequent uncertainty analysis.
- Part IV focuses on theory and algorithms for parameter inference in frequentist and Bayesian frameworks. It also provides techniques to efficiently propagate uncertainties through models to provide distributions or prediction intervals for responses. The final chapter discusses issues associated with model discrepancy and summarizes frameworks, including machine learning, to address errors in both physical and observation models.
- The construction of numerical and statistical surrogate and reduced-order models is covered in Part V. Whereas this comprises the final component of the book, we illustrate via several large-scale examples that this analysis must often precede the sensitivity and uncertainty analysis in Parts III and IV to provide the efficiency required for many of the sampling-based algorithms.

We note that Parts III and IV cover the general end-to-end methodology for sensitivity analysis and uncertainty quantification, as supported by examples and models from Part I and concepts in Parts II and V.

A significant feature of the book is that essentially all concepts are motivated and illustrated by the large set of examples and models detailed in Part I. Most chapters start with one or more examples to motivate topics in that chapter. These and additional examples are subsequently revisited to illustrate the theory and numerical algorithms and motivate the manner in which readers can apply the results to problems arising in their disciplines. The inclusion of experimental data comprises a unique feature of many examples.

In addition to the extensive use of examples to motivate and illustrate concepts, a central theme throughout the book is the focus on numerical algorithms, which are highly robust and hence applicable for a wide range of applications. As we note in the **Online Material**, MATLAB code for examples is posted at the accompanying book website. We also provide links to Python packages, which can be employed for Bayesian inference and uncertainty propagation. Finally, we provide citations to selected R packages. In combination, this provides readers with substantial resources to apply the techniques to problems arising in their own disciplines.

The over 100 exercises strongly complement the examples while focusing on a blend of derivations and formulations in combination with numerous simulations. In many cases, readers and students can modify posted MATLAB codes for the examples to solve exercises. As with the examples, the inclusion of experimental data in exercises comprises a unique feature of the book.

Throughout the book, we list UQ crimes to help identify common misconceptions and guide those entering the field. These range in severity from UQ misdemeanors, which have probably been committed at some point by most scientists – often with minimal ramifications – to UQ felonies which should always be avoided.

## Changes from the First Edition

This book differs from the first edition in the following manner.

- The topics in Parts I–V have been significantly reordered from the first book to reflect a background and workflow now generally employed for sensitivity analysis and uncertainty quantification for a range of applications. This workflow is detailed in **Features of the Book**.
- We have added five new chapters on topics which are now considered fundamental for the field. These include chapters on random field representations, physical and statistical observation models, parameter identifiability and influence, active subspace techniques, and statistical surrogate models. We have also substantially increased the focus on numerical and statistical surrogate and reduced-order models, which now comprise five chapters in Section V of the book.
- The chapter on local sensitivity analysis has been completely rewritten and now focuses on the use of sensitivity equations, complex-step approximation, adjoint methods, and parameter subset selection techniques to ascertain parameter influence.

- The chapters on global sensitivity analysis, frequentist and Bayesian inference, uncertainty propagation for model responses, and model discrepancy have been substantially modified and extended to include new examples, improve their impact for courses, and illustrate large-scale applications.
- We have modified some of the notation to clarify concepts and more accurately reflect what is becoming more common in the uncertainty quantification literature.
- We have expanded the repertoire of examples both to illustrate large-scale applications, including digital twins and virtual populations, and to provide additional prototypical examples to more completely illustrate the workflow and topics.
  - As detailed in **Features of the Book**, we now illustrate the end-to-end parameter selection and uncertainty quantification process for a large-scale PDE model employed in nuclear power plant analysis.
  - The suite of new large-scale examples is further augmented by applications drawn from across the research community. These examples illustrate uncertainty quantification for a wetland methane emission model, the construction of Gaussian process emulators for volcanic flows, and reduced-order model construction for a chemical vapor deposition reactor, for a thermal fin, and to facilitate uncertainty analysis for a subsonic rotor blade in unsteady flow.
  - The set of prototypical models has been expanded through the inclusion of algebraic Helmholtz energy and height models, ODE models for sprinter and disease dynamics, and PDE models for incompressible and turbulent flow regimes.
  - We include measured data for the height, heat, sprinter, and disease models for use in examples and problems illustrating uncertainty analysis.
- The set of exercises has been expanded by over a factor of four to cover a broader range of topics.
- We have added UQ crimes throughout the text to identify common misconceptions and guide readers entering the field.

In addition to adding a number of topics now deemed central to uncertainty quantification and reordering the requisite workflow, we have also omitted certain topics which appeared in the first edition. This includes the abstract model frameworks, convergence analysis for sequences of random variables and Markov chains, and high-dimensional model representations (HDMR). We have also omitted the local forward and adjoint sensitivity analysis procedures and supporting functional analysis topics detailed in the Appendix. The omission of these topics is not due to lack of importance but rather to augment the end-to-end focus on sensitivity analysis and uncertainty quantification in this book. Where relevant, we cite this material as included in Supplemental Material for this book or provide external references to the topics.

## Course Options

This book can be used for either self-study or as a text for a one-semester upper undergraduate or graduate-level course. Both uses are facilitated by the large number of examples and exercises. The following schedule provides a prototype for a 15-week course. Topics: Motivating applications and prototypical models (0.5 week); Fundamental aspects of probability, random processes and statistics (1.5 weeks); Representation of random inputs (1 week); Parameter selection techniques; local/global sensitivity analysis, active subspace analysis (3 weeks); Frequentist and Bayesian model calibration (3 weeks); Uncertainty propagation in models (2 weeks); Model discrepancy (1 week); Surrogate and reduced-order model construction (2.5 weeks); Sparse grid techniques (0.5 week).

## Online Material

MATLAB code for most examples are posted online at

<https://bookstore.siam.org/cs30/bonus>

which accompanies the book. In addition to illustrating material in the body of the text, these codes can, in some cases, be modified to solve exercises. We will also post Supplemental Material, the Preface for the first edition, and an erratum at this website.

## Acknowledgments

It is stated in an African proverb that it takes a village to raise a child. A scientific version of this adage is that it takes a research and classroom community to help write a book in an evolving field. This book reflects significant input from students, colleagues, collaborators, and scientists in the field, with contributions from the latter illustrated in several examples highlighting large-scale applications.

Initial material for this book evolved from notes from a special topics course on “Validation, Verification, and Uncertainty Quantification,” which the author developed at North Carolina State University in 2008. This is now offered as the one-semester course “Uncertainty Quantification for Physical and Biological Models,” which includes the topics summarized under **Course Options**. Feedback from students motivated the inclusion of additional models and examples and significantly influenced the ordering of covered topics. In addition to noting numerous typos, feedback from students also guided a number of revisions that have improved the exposition.

The book has also benefitted significantly from feedback provided by colleagues, graduate students, and collaborators who read various chapters. The author sincerely thanks Alen Alexanderian, James Berger, Jenny Brynjarsdottir, Catherine Gorle, Pierre Gremaud, Mansoor Haider, Harley Hanes, Joey Hart, Adam Hetzler, Ilse Ipsen, Arsen Iskhakov, Joshua Kaizer, Min Kang, Kyle Mandli, Paul Miles, Sue Minkoff, Helen Moore, Jessica Notestine, Tony O’Hagan, Marco Panesi, Bruce Pitman, Walker Powell, Brian Reich, Arvind Saibaba, Jouni Susiluoto, Hien Tran, Brian Williams, and Alyson Wilson for candid and detailed feedback regarding

the manuscript and suggestions that have significantly improved both the scope and exposition. This book also reflects input and research contributions from graduate students, and postdocs, some of whose contributions are noted in the examples and citations. Sincere thanks are also extended to Brian Adams, Nate Burch, Amanda Coons, John Crews, John Harlim, Zhengzheng Hu, Dustin Kapraun, Zack Kenz, Christine Latten, Jerry McMahan Jr., Keri Rehm, Mami Wentworth, and Lucas Van Blaircum for their input and feedback for the first edition. Finally, the author sincerely thanks Karen Willcox and two anonymous reviewers whose feedback significantly improved the book. The danger in thanking individuals is the reality of inadvertently omitting key contributions. Sincere thanks and apologies are extended to those individuals who were inadvertently omitted.

The content and large-scale examples significantly reflect both research support and collaborations via consortia and individual support from funding agencies. The examples illustrate research performed via support from the Department of Energy Consortium for Advanced Simulation of Light Water Reactors (CASL) and the NNSA Consortium for Nonproliferation Enabling Capabilities (CNEC). The author also gratefully acknowledges the impact of support from the Air Force Office of Scientific Research (AFOSR), the National Aeronautics and Space Administration (NASA), and the National Science Foundation (NSF) on contents in this book.

Part of the first edition was written while the author was a Faculty Fellow in the 2011–12 Statistical and Applied Mathematical Sciences Institute (SAMSI) *Program on Uncertainty Quantification*. Collaboration and interactions during that year significantly influenced aspects of the first edition and the author very gratefully acknowledges the scientific and financial contributions from that program. Portions of this second edition were influenced by scientific input while the author was an invited participant during the research program on *Uncertainty Quantification for Complex Systems* at the Isaac Newton Institute, University of Cambridge, in 2018. It was also strongly influenced by collaborations during the SAMSI program on *Model Uncertainty: Mathematical and Statistical (MUMS)*, in 2018–19. The author very gratefully acknowledges scientific and financial contributions from those programs. Finally, the author extends his sincere thanks and gratitude to Elizabeth Greenspan of SIAM for her support, input, counsel and encouragement throughout the journey of writing this book.

Ralph C. Smith  
North Carolina State University  
Raleigh, NC  
February 15, 2024

# Notation

This compilation does not include all of the symbols used throughout the text, and we neglect those that appear one time in a specific context such as those in the models of Chapter 2. Instead, it is meant to clarify the role of symbols which are critical for the discussion in addition to symbols that have multiple definitions. Matrices and vectors are represented using bold fonts.

| Symbol                                   | Meaning   | Page     |
|--|---|----------|
| <i>Matrices and Vectors</i>              |   |          |
| $\beta$                                  | Auxiliary or design parameters  | 333, 380 |
| $\zeta$                                  | Active variables  | 243      |
| $\theta = [\theta_1, \dots, \theta_p]^T$ | Calibration parameters  | 97       |
| $\theta_{\sim i}$                        | Parameter vector $[\theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_p]^T$ | 199      |
| $\theta^0$                               | True but unknown parameter vector   | 96       |
| $\theta^*$                               | Nominal parameter vector  | 183      |
| $\theta^*$                               | Proposed Markov chain parameter   | 292      |
| $\theta^{k-1}$                           | Parameters at $k - 1$ step in Markov chain  | 292      |
| $\theta_{inf}, \theta_{noninf}$          | Influential and noninfluential parameters   | 225      |
| $\hat{\theta}, \theta$                   | Estimator and estimate for $\theta^0$   | 268, 275 |
| $\theta_{MAP}$                           | Maximum a posteriori parameter estimate   | 289      |
| $\theta_{MLE}$                           | Maximum likelihood estimate   | 88       |
| $\Lambda$                                | Eigenvalue matrix   | 243      |
| $\Sigma$                                 | Matrix of singular values of matrix $\mathbf{A}$                                    | 236      |
| $C$                                      | Covariance matrix with $[C]_{ij} = c(\mathbf{q}^i, \mathbf{q}^j)$                   | 444      |
| $\mathcal{C}$                            | Observation matrix or vector  | 55       |
| $F$                                      | Scaled Fisher information matrix  | 159      |
| $\mathcal{F}$                            | Fisher information matrix   | 152      |
| $I_n$                                    | Identity matrix ( $n \times n$ )  | 266      |
| $K$                                      | Stiffness matrix  | 435      |
| $M$                                      | Mass matrix   | 435      |

| Symbol  | Meaning  | Page     |
|---|--|----------|
| <i>Matrices and Vectors (Continued)</i>                           |  |          |
| $\mathbf{q} = [\boldsymbol{\theta}, \boldsymbol{\beta}]^T$        | Vector of $p$ calibration and auxiliary parameters                 | 333      |
| $\mathbf{q}^m$  | Parameter values to construct surrogate models                     | 380      |
| $\mathbf{q}^r$  | Quadrature points  | 402      |
| $\mathbf{Q}$  | Orthogonal matrix in QR factorization                              | 237      |
| $\mathbf{Q}^*$  | Matrix of test input values  | 448      |
| $\mathbf{R}$  | Upper triangular Cholesky factorization of $\mathbf{V}$            | 296      |
| $\mathbf{R}$  | Upper triangular matrix in QR factorization                        | 237      |
| $\mathbf{R}$  | Correlation matrix   | 446      |
| $\hat{\mathbf{R}}, \mathbf{R}$                                    | Residual estimator and estimate                                    | 263      |
| $\mathbf{S}$  | Sensitivity matrix   | 153      |
| $\mathbf{V}$  | Covariance matrix  | 81       |
| $\mathbf{V}_k$  | Chain covariance matrix  | 307      |
| $\mathbf{W}$  | Eigenvector matrix   | 243      |
| $\mathbf{X}$  | Deterministic $n \times p$ design matrix                           | 260      |
| $\mathbf{Y}, \mathbf{y}$  | Random vector, realizations for model response                     | 125      |
| <i>Probability, Statistics, Optimization, and Models</i>          |  |          |
| $\alpha(x, \omega)$   | Random field   | 108      |
| $\delta, \delta(\xi_i)$   | Model discrepancy or error   | 368      |
| $\varepsilon_i, \epsilon_i$                                       | Random and realized observation errors                             | 86, 261  |
| $\zeta$   | Random variables mapped to $\mathcal{N}(0, 1), \mathcal{U}(-1, 1)$ | 406      |
| $\mu$   | Mean   | 72       |
| $\mu_i, \mu_i^*$  | Derivative and Morris sensitivity indices                          | 210, 213 |
| $\xi_i$   | Independent variable, e.g., time $t_i$ or space $x_i$              | 123      |
| $\pi_0(\boldsymbol{\theta}), \pi(\boldsymbol{\theta} \mathbf{y})$ | Bayesian prior and posterior density                               | 98       |
| $\rho(\boldsymbol{\theta})$                                       | Probability density function for parameters $\boldsymbol{\theta}$  | 202      |
| $\sigma_0^2$  | True but unknown measurement error variance                        | 261      |
| $\hat{\sigma}^2, \sigma^2$  | Estimator and estimate for $\sigma_0^2$                            | 268, 275 |
| $\sigma_i$  | Morris sensitivity indices (standard deviation)                    | 213      |
| $a_i$   | Activity score   | 246      |
| $c(x, y), c(\mathbf{q}, \mathbf{q}')$                             | Covariance function  | 110, 442 |
| $d_i(q)$  | Morris elementary effects for $i^{th}$ input                       | 213      |
| $D, D_i, D_{ij}$  | Total and partial variances of response $Y$                        | 200      |
| $\mathbb{E}$  | Mean or expected value   | 72       |

| Symbol   | Meaning   | Page     |
|--|---|----------|
| <i>Probability, Statistics, Optimization, and Models (Continued)</i>             |   |          |
| $f_0, f_i, f_{ij}$   | Interaction terms   | 199      |
| $f_S(\mathbf{q}), f_S^K(\mathbf{q})$   | Surrogate models  | 380, 389 |
| $F_X(x)$   | Cumulative distribution function                          | 70       |
| $f_X(x)$   | Probability density function                              | 71       |
| $f(\xi_i, \boldsymbol{\theta})$  | Model outputs   | 123      |
| $J(\boldsymbol{\theta})$   | Least squares functional                                  | 185      |
| $J(\boldsymbol{\theta}^* \boldsymbol{\theta}^{k-1})$                             | Proposal or jumping distribution                          | 292      |
| $L(\boldsymbol{\theta} \mathbf{y}), \mathcal{L}(\boldsymbol{\theta} \mathbf{y})$ | Likelihood and log-likelihood function                    | 87       |
| $\mathcal{N}(\mu, \sigma^2)$   | Normal distribution with mean $\mu$ , variance $\sigma^2$ | 73       |
| $p(\mathbf{y} \boldsymbol{\theta})$  | Likelihood function                                       | 98       |
| $P$  | Probability measure                                       | 69       |
| $r$  | Acceptance ratio  | 293, 297 |
| $r(\mathbf{q}, \mathbf{q}')$   | Correlation function                                      | 444      |
| $s_j$  | Local sensitivity indices                                 | 164      |
| $s_j^*, \hat{s}_j, \tilde{s}_j$  | Scaled local sensitivities                                | 183      |
| $S_i, S_{ij}, S_{Ti}$  | Sobol' sensitivity indices                                | 201      |
| $SS_\theta$  | Sum of squares error                                      | 289      |
| $Y_i$  | Experimental observations                                 | 123      |
| <i>Operators, Functions, and Spaces</i>  |   |          |
| $\mathbf{1}_{[a,b]}(x)$  | Characteristic function on $[a, b]$                       | 74       |
| $\Gamma$   | Admissible and sample space for parameters                | 86, 199  |
| $\phi_i(x)$  | Spatial basis functions                                   | 429      |
| $\varphi$  | Adjoint variable  | 180      |
| $\Psi_k(\mathbf{q}), \psi_k(q)$  | Multivariate, univariate basis functions                  | 389, 391 |
| $\mathcal{A}(q, p)$  | Sparse grid quadrature operator                           | 506      |
| $H_i(\theta)$  | Hermite polynomials                                       | 401      |
| $\mathcal{H}(q, p)$  | Sparse quadrature grid                                    | 506      |
| $\mathbf{i}', \mathbf{j}', \mathbf{k}'$  | Multi-indices   | 403      |
| $\mathcal{I}^{(p)} f$  | Interpolation operator in $\mathbb{R}^p$                  | 396      |
| $I^{(p)} f$  | Integral operator in $\mathbb{R}^p$                       | 489, 500 |
| $L_m(q)$   | Lagrange interpolating polynomial                         | 392      |
| $L^2_{\rho_i}(\Gamma_i), L^2_\rho(\Gamma)$                                       | Square integrable functions on $\Gamma_i, \Gamma$         | 423      |

---

| Symbol  | Meaning   | Page     |
|---|---|----------|
| <i>Operators, Functions, and Spaces (Continued)</i> |   |          |
| $NI(\boldsymbol{\theta}), I(\boldsymbol{\theta})$   | Nonidentifiable, identifiable parameter subspaces | 234      |
| $\mathcal{NI}(\boldsymbol{\theta})$                 | Space of functionally noninfluential parameters   | 157      |
| $\mathcal{N}(\mathbf{A})$                           | Null space of the matrix $\mathbf{A}$             | 235      |
| $P_i(\theta)$                                       | Legendre polynomials                              | 402      |
| $\mathcal{Q}^{(p)} f$                               | Quadrature operator in $\mathbb{R}^p$             | 489      |
| $\mathcal{R}(\mathbf{A})$                           | Range of the matrix $\mathbf{A}$                  | 235      |
| $V, V^J$  | Spaces of spatial test functions                  | 429      |
| $w_r, w_i$  | Quadrature weights                                | 402, 490 |
| $Z, Z^K$  | Spaces of parameter test functions                | 429      |
| <i>Numerical Values</i>                             |   |          |
| $\sigma_j$  | Singular values of the matrix $\mathbf{A}$        | 236      |
| $M, M_\ell$   | Number of interpolation points                    | 396      |
| $n$   | Number of measurements or model evaluations       | 86       |
| $n, n_\ell$   | Number of tensor product quadrature points        | 501      |
| $\mathcal{N}$                                       | Number of sparse grid quadrature points           | 506      |
| $p$   | Number of parameters                              | 97       |
| $r$   | Rank of a matrix $\mathbf{A}$                     | 236      |

---

# Acronyms and Initialisms

| Term   | Meaning   | Page |
|--------|---|------|
| ABC    | Approximate Bayesian computation                                | 326  |
| ANOVA  | Analysis of variance  | 144  |
| cdf    | Cumulative distribution function                                | 70   |
| CESM   | Community Earth System Model                                    | 35   |
| CFD    | Computational fluid dynamics                                    | 23   |
| CFCs   | Chlorofluorocarbons   | 36   |
| CIPS   | Crud-induced power shift  | 21   |
| CMIP   | Coupled Model Intercomparison Project                           | 35   |
| DAKOTA | Design Analysis Kit for Optimization and Terascale Applications | 439  |
| DGSM   | Derivative-based global sensitivity measure                     | 210  |
| DRAM   | Delayed Rejection Adaptive Metropolis                           | 307  |
| DREAM  | DiffeRential Evolution Adaptive Metropolis                      | 316  |
| ECMWF  | European Centre for Medium-Range Weather Forecasts              | 33   |
| FNN    | Feedforward neural network                                      | 44   |
| GP     | Gaussian process  | 91   |
| gPC    | Generalized polynomial chaos                                    | 421  |
| GHS    | Greenhouse gas  | 40   |
| HMC    | Hamiltonian Monte Carlo   | 316  |
| HIV    | Human immunodeficiency virus                                    | 58   |
| iid    | Independent and identically distributed                         | 80   |
| IPCC   | Intergovernmental Panel on Climate Change                       | 35   |
| KL     | Karhunen–Loëve  | 108  |
| kde    | Kernel density estimation                                       | 78   |
| LHS    | Latin hypercube sampling  | 386  |
| LWR    | Light water reactor   | 20   |
| MAP    | Maximum a posteriori (estimate)                                 | 289  |

| Term  | Meaning   | Page |
|-------|---|------|
| MCMC  | Markov chain Monte Carlo                          | 290  |
| MC    | Monte Carlo                                       | 332  |
| MCNP  | Monte Carlo N-Particle                            | 19   |
| MLE   | Maximum likelihood estimate                       | 88   |
| MARS  | Multivariate adaptive regression splines          | 415  |
| NAMAC | Nearly Autonomous Management and Control          | 43   |
| NISP  | Nonintrusive spectral projection                  | 439  |
| OAT   | One-at-a-time                                     | 210  |
| ODE   | Ordinary differential equation                    | 52   |
| OLS   | Ordinary least squares                            | 86   |
| PBPK  | Physiologically-based pharmacokinetic             | 25   |
| PC    | Polynomial chaos                                  | 421  |
| PDE   | Partial differential equation                     | 59   |
| QSP   | Quantitative systems pharmacology                 | 25   |
| PSS   | Parameter subset selection                        | 184  |
| PCA   | Principal component analysis                      | 472  |
| pdf   | Probability density function                      | 71   |
| POD   | Proper orthogonal decomposition                   | 469  |
| PWR   | Pressurized-water reactor                         | 20   |
| PZT   | Lead-zirconate-titanate                           | 51   |
| Q-Q   | Quantile-quantile                                 | 77   |
| QoI   | Quantity of interest                              | 7    |
| RANS  | Reynolds averaged Navier–Stokes                   | 64   |
| RNN   | Recurrent neural network                          | 44   |
| RAVEN | Risk analysis and Virtual ENvironment             | 43   |
| RMSE  | Root mean squared error                           | 252  |
| SA    | Sensitivity analysis                              | 2    |
| SDE   | Stochastic differential equation                  | 95   |
| SEIR  | Susceptible, exposed, infected, recovered (model) | 57   |
| SIR   | Susceptible, infected, recovered (model)          | 56   |
| SVD   | Singular value decomposition                      | 236  |
| SSP   | Shared socioeconomic pathway                      | 40   |
| UQ    | Uncertainty quantification                        | 2    |
| UAV   | Unmanned air vehicle                              | 42   |

# UQ Crimes

We summarize here UQ crimes which are provided in the text to help readers recognize and avoid common mistakes and misconceptions. In many cases, we delineate those which often have minimal consequences from those which should always be avoided.

**UQ Crime 4.61:** Treating the concepts of frequentist confidence interval and Bayesian credible interval as synonymous. This should be avoided since they are based on different assumptions and quantify different behavior (Page 97).

**UQ Crime 4.62:** The support of the posterior distribution lies outside that of the prior distribution. This is theoretically not possible and indicates a mistake or inaccurate approximation during Bayesian inference (Page 99).

**UQ Crime 4.70:** Computing energy statistics using a kernel density estimate (kde) rather than the sampled data. The kde smooths the distribution and can diminish the accuracy of computed test statistics (Page 104).

**UQ Crime 6.12:** Employing longitudinal observation models for a single process or subject, evaluated at multiple values of an independent variable, for aggregate observations associated with multiple subjects. This can often be avoided by using mixed-effects or hierarchical approaches (Page 132).

**UQ Crime 6.13:** Sampling from a parameter distribution to construct synthetic data for observation models having additive observation errors. This violates the assumptions for the observation models and can be avoided by instead sampling the observation errors (Page 132).

**UQ Crime 6.15:** Sampling from Gaussian distributions for parameters having a sign constraint. Even for small variances, this can yield parameter values having the incorrect sign. This can be avoided by sampling from uniform or lognormal distributions (Page 133).

**UQ Crime 6.16:** Assuming independent errors when constructing observation models. For applications in which errors are dependent, this can degrade the accuracy of likelihoods and subsequent uncertainty analysis. Although often violated, the potential ramifications should be verified (Page 134).

**UQ Crime 7.22:** Assuming that correlated parameters are therefore nonidentifiable. These are different properties and should not be equated (Page 154).

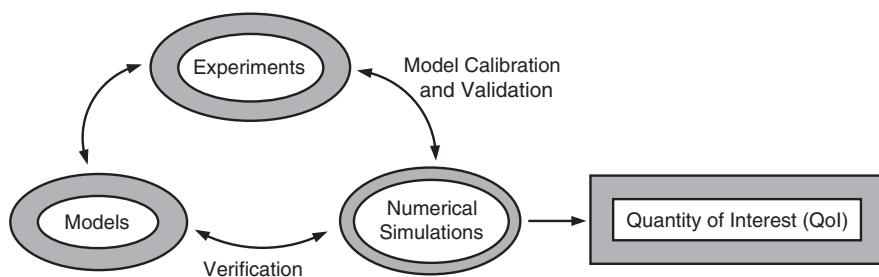
- UQ Crime 7.27:** The conditions *practically identifiable* and *functionally influential* are often used interchangeably when selecting parameters for subsequent inference or uncertainty propagation. Whereas both concepts can be employed for this role, their assumptions and interpretations differ so the terms should not be equated (Page 158).
- UQ Crime 8.27:** Employing a global scaled information matrix for parameter subset selection. Whereas this is valid when constructing active subspaces, it can yield incorrect results pertaining to parameter identifiability (Page 191).
- UQ Crime 9.11:** One occasionally encounters negative indices for partial variances computed using Sobol' analysis. This can often be addressed by employing a more robust approximation algorithm or increasing the number of Monte Carlo samples (Page 207).
- UQ Crime 9.13:** The assumption of independent parameters when computing Sobol' indices can yield erroneous rankings for intrinsically dependent parameters. In some cases, this can be addressed by incorporating the correlation structure (Page 207).
- UQ Crime 11.11:** Citing asymptotic results for small sample sizes can constitute a UQ crime if not substantiated (Page 266).
- UQ Crime 11.16:** Treating the concepts of frequentist confidence and prediction intervals for responses as synonymous. The latter provide estimates for new or future predictions (Page 270). See also **UQ Crime 4.61**.
- UQ Crime 12.14:** It is common to encounter MCMC results with no mention of burn-in or convergence. Such discussion is important to establish the validity of sampled posterior distributions and avoid a potential UQ crime (Page 304).
- UQ Crime 13.8:** It is sometimes stated that a 95% prediction interval contains 95% of *previously* sampled data. This interpretation should be avoided since a prediction interval instead quantifies the probability of including new or future predictions (Page 342).
- UQ Crime 16.15:** This UQ crime reiterates **UQ Crime 6.15**, which notes that employing normal distributions for parameters with a sign constraint can produce nonphysical behavior. This can be avoided by employing uniform or lognormal distributions. If a Gaussian distribution is required, one should monitor samples to avoid nonphysical coefficients (Page 409).
- UQ Crime 16.17:** We noted in **UQ Crime 9.13** that the assumption of independence can produce erroneous global sensitivity rankings for dependent parameter sets. This UQ crime applies the same warning when assuming independent parameters to facilitate surrogate model construction (Page 411).
- UQ Crime 17.6:** This reiterates **UQ Crimes 6.15** and **16.15** by noting nonphysical behavior which can result from the assumption of normally distributed parameters (Page 426).

# Chapter 1

# Introduction

The synthesis of modeling, large-scale simulations, and experiments has long been recognized as critical for understanding and advancing the state of science and technology. When considered in the broad sense of including requisite theory, these form the pillars of *predictive science*, as illustrated in Figure 1.1. In the context of predictive science, uncertainty quantification can be broadly defined as the science of identifying, quantifying, and reducing uncertainties associated with models, numerical algorithms, experiments, and predicted outcomes or quantities of interest. Aspects of this field, such as the quantification of measurement uncertainties and numerical errors, are well understood and are addressed by classical statistics and numerical analysis theory. However, the systematic quantification of uncertainties and errors in models, simulations, and experiments and the analysis of how they are propagated through complex models to affect predicted outcomes is more recent and constitutes both an active area of research and the subject of this text.

In Chapter 2, we detail five large-scale applications where model predictions with quantified uncertainties are critical for understanding and predicting scientific phenomena and making informed decisions and designs based upon these predictions. These applications are weather models, climate models, subsurface hydrology and geology models, nuclear reactor designs, and models for biological phenomena.



**Figure 1.1.** Modeling, numerical, and experimental components of predictive science with associated uncertainties and errors indicated in gray.

In the following example, we summarize aspects of the weather model detailed in Section 2.1 to motivate sources of uncertainty and indicate issues that must be addressed when making predictions.

### Example 1.1 (Weather Prediction).

The physical components of meteorological or weather models are constructed by quantifying the interactions between temperature and pressure gradients, wind, and precipitation using conservation of energy, mass, and momentum. When combined with conservation of water phases and aerosol concentrations, this yields the equations of atmospheric physics,

$$\begin{aligned} \frac{\partial \rho}{\partial t} + \nabla \cdot (\rho v) &= 0, \\ \frac{\partial v}{\partial t} &= -v \cdot \nabla v - \frac{1}{\rho} \nabla p - g \hat{k} - 2\Omega \times v, \\ \rho c_V \frac{\partial T}{\partial t} + p \nabla \cdot v &= -\nabla \cdot F + \nabla \cdot (k \nabla T) + \rho \dot{q}(T, p, \rho), \\ p &= \rho R T, \\ \frac{\partial m_j}{\partial t} &= -v \cdot \nabla m_j + S_{m_j}(T, m_j, \chi_j, \rho), \quad j = 1, 2, 3, \\ \frac{\partial \chi_j}{\partial t} &= -v \cdot \nabla \chi_j + S_{\chi_j}(T, \chi_j, \rho), \quad j = 1, \dots, J, \end{aligned} \tag{1.1}$$

where  $\rho$ ,  $v$ ,  $T$ ,  $p$ ,  $k$ , and  $c_V$  respectively denote the density, velocity, temperature, pressure, thermal conductivity, and specific heat of air. The concentration of water in solid, liquid, and gaseous phases is denoted by  $m_1$ ,  $m_2$ , and  $m_3$ , whereas the concentration of the  $j^{th}$  aerosol species is denoted by  $\chi_j$ .

For the reasons discussed in Section 2.1, one typically constructs phenomenological models for the source terms  $S_{m_j}(T, m_j, \chi_j, \rho)$  and  $S_{\chi_j}(T, \chi_j, \rho)$ . For example, it is established in (2.8) that  $S_{m_2}$  can be formulated as

$$S_{m_2} = S_1 + S_2 + S_3 - S_4, \tag{1.2}$$

where

$$S_1 = \bar{\rho} (m_2 - m_2^*)^2 \left[ 1.2 \times 10^{-4} + \left( 1.569 \times 10^{-12} \frac{n_r}{d_0 (m_2 - m_2^*)} \right) \right]^{-1} \tag{1.3}$$

requires the specification of the nonphysical parameters  $\bar{\rho}$ ,  $m_2^*$ ,  $n_r$ , and  $d_0$ . The remaining components have similar formulations.

**Model Errors or Discrepancies.** Both the conservation relations (1.1) and phenomenological closure equations (1.2) and (1.3) are approximations of the true underlying physics. Furthermore, phenomena such as the conversion of cloud droplets to rain drops, quantified by (1.3), occur on much smaller scales than the numerical grids employed when solving (1.1). The resulting model errors or discrepancies pro-

duce biased or systematic uncertainties that are typically difficult to quantify using a probabilistic framework.

**Input Uncertainties.** Parameters such as  $\bar{\rho}$ ,  $m_2^*$ ,  $n_r$ , and  $d_0$  in the phenomenological representation (1.3) for  $S_1$  are uncertain, as are initial conditions for the evolution equations (1.1). These comprise input uncertainties that are often amenable to probabilistic analysis.

**Numerical Errors and Uncertainties.** As detailed in Section 2.1.3, local meteorological models are numerically approximated on spatial grids having horizontal spacing on the order of 5 km and vertical spacing of approximately 200 m. This introduces numerical discretization or approximation errors. Furthermore, it introduces systematic uncertainties due to the fact that parameterized processes, such as aerosol-induced cloud formation and atmospheric turbulence, occur on much smaller, subgrid, scales.

**Measurement Errors and Uncertainties.** It is noted in Section 2.1.4 that meteorological data is comprised of earth-surface and atmospheric measurements. The latter is obtained from weather balloons, weather satellites, and aircraft. There are two primary sources of uncertainty: limited accuracy of the sensors and uncertainty associated with the time and location of measurements.

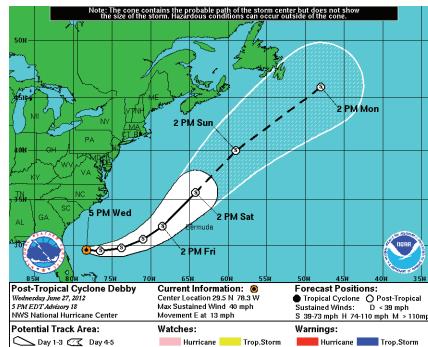
**Predictions for Weather Forecasts.** Uncertainty quantification for weather forecasting takes place in two steps. In the first, data assimilation—often performed in a Bayesian framework—is used to determine values and quantify uncertainties for inputs such as initial conditions and phenomenological parameters. This is the model calibration step. In the second step, the calibrated models are run forward in time to provide forecasts with quantified uncertainties.

To accommodate the effects of input, model, numerical, and measurement uncertainties, ensemble forecasts are computed by running multiple simulations from individual or multiple models with differing initial conditions or parameter values drawn from probability densities constructed during the calibration phase. Using the ensemble predictions, one computes statistical quantities of interest, such as the average temperature, relative humidity, or projected rain amounts.

Although uncertainties associated with quantities of interest are computed during the ensemble computations, they typically are not reported in forecasts. One exception is the prediction of large storms such as cyclones, tropical storms, or hurricanes. This is illustrated in Figure 1.2 by the predicted trajectory and uncertainty cones for the post-tropical cyclone Debby.

The following definitions quantify terms, introduced in Example 1.1, that play a fundamental role throughout the text.

**Definition 1.2 (Inputs).** The term inputs is used to designate parameters, initial conditions, boundary conditions, or exogenous forces that exhibit uncertainties which must be determined and propagated through models to construct predictions with quantified uncertainties. Known or fixed coefficients, independent and dependent variables, and control signals do not constitute inputs as defined here.



**Figure 1.2.** NOAA image of the trajectory and cone of uncertainty for the post-tropical cyclone Debby.

**Definition 1.3 (Quantity of Interest (QoI)).** The QoI designates the output of a simulation model or experiment that provides information necessary to make conclusions or decisions about the process. In many contexts, we employ the terms *model response* or *model output* in a synonymous manner. As illustrated in Example 1.1, we often consider statistical or probabilistic QoI to accommodate uncertainties intrinsic to the modeled process. Examples of probabilistic QoI include average temperatures, expected precipitation in a viewing region, average performance of a nuclear reactor, and expected impact of drilling in an environmentally fragile region. Chapter 2 details QoI for weather, climate, groundwater, nuclear reactor, and systems biology models.

**Definition 1.4 (Verification).** Verification refers to the process of quantifying the accuracy of simulation codes used to implement mathematical models.

**Definition 1.5 (Validation).** Validation describes the process of determining the accuracy with which mathematical models quantify the physical processes of interest. This necessarily involves the simulation code used to implement the model and experimental data from the process.

## 1.1 Nature of Uncertainties and Errors

In Example 1.1, we illustrated that uncertainties and errors arise in the modeling, simulation, and experimental components of applications. We detail the sources and nature of these uncertainties in this section.

### 1.1.1 Experimental Uncertainties and Limitations

*Experimental results are believed by everyone, except for the person who ran the experiment,* quoted by Max Gunzburger, Florida State University; original source unknown.

There are two fundamental sources of uncertainty and errors in experiments: limited or incomplete data and limited accuracy or resolution of sensors. The first

can be broadly interpreted as due to the fact that experiments are often surrogates or provide only partial measurements when we cannot fully observe the underlying application due to physical infeasibility or expense. Examples include the following.

- The meteorological data noted in Example 1.1 is obtained at discrete locations that can be uncertain.
- Wind tunnel tests are used as surrogates for flight tests. The limitations of using a scale model in lieu of a full aircraft must be incorporated in designs.
- Pharmaceutical and disease treatment strategies are often too dangerous or expensive for human tests or large segments of the population. For example, HIV trials are conducted with test subjects rather than a full population.
- Climate scenarios cannot be experimentally tested at the planet scale. Instead, forcing mechanisms such as those due to volcanic eruptions are tested using measurements such as the 1991 Mount Pinatubo data—see Section 2.2.1.
- In materials experiments, difficulties obtaining nano- and molecular-level time and spatial scale data limit multiscale testing of novel material designs.
- Subsurface hydrology data is very limited due to the expense and infeasibility of drilling large numbers of wells. As a result, there is significant uncertainty regarding specific subsurface structures—see Section 2.3.
- The harsh radioactive, thermal, and chemical environments in a nuclear reactor core limit the availability of measurements for performance improvement, nondestructive evaluation, and safety regulation—see Section 2.4.

Whereas several of these examples illustrate limited rather than statistically uncertain data, the associated deficiencies increase the reliance on models and augment model uncertainties due to lack of data.

The limited accuracy or resolution of sensors contributes statistical uncertainties that can produce parameter uncertainties during model calibration. These sensor uncertainties can occasionally be specified by sensor manufacturers and are often amenable to statistical analysis.

### 1.1.2 Model and Input Uncertainties

*Essentially, all models are wrong, but some are useful,* George E.P. Box, page 424 of [38].

Model uncertainties arise from two sources: model errors or discrepancies and input uncertainties due to uncertain parameters, forcing functions, and initial and boundary conditions.

#### Model Errors and Discrepancies

Modeling errors or discrepancies are due to approximate or imprecise representation of underlying physical, biological, economic, or social processes. The following examples illustrate sources of model error.

- Numerous components of weather and climate models—e.g., aerosol-induced cloud formation, greenhouse gas processes, turbulence—occur on scales that are much smaller than the numerical grids used to solve the atmospheric equations of physics. Moreover, many of these processes represent highly complex physics that is only partially understood. Subsequently, the processes are represented by phenomenological models with nonphysical parameters. Both the model form and the parameters exhibit uncertainty.
- Many biological applications are coupled, complex, highly nonlinear, and driven by poorly understood or stochastic processes. Moreover, they presently do not admit an encompassing set of governing relations analogous to those in physics. Hence, associated models are subject to significant uncertainty.
- The predicted production of greenhouse gases is highly dependent on projected economic and technological growth of nations. These processes are highly uncertain and difficult to model. This uncertainty is addressed in climate models by considering various economic and technology growth scenarios.

The quantification of model errors is typically problem-dependent since it necessitates obtaining additional knowledge about the problem. The development of general statistical techniques, such as construction of Gaussian processes for model discrepancies, constitutes a current research topic.

### **Input Uncertainties**

All models contain parameters that must be specified before the model can be used to represent or predict the behavior of the process. Moreover, differential equation models have initial or boundary conditions that must be designated in addition to potential exogenous forces. As noted in Definition 1.2, these components introduce input uncertainties that must be quantified and propagated through models.

- As indicated in Example 1.1 and shown in Sections 2.1–2.5, the phenomenological models used to represent processes such as turbulence in weather, climate, and nuclear reactor models have nonphysical parameters whose values and uncertainties must be determined using measured data.
- It is shown in Section 2.2 that forcing and feedback mechanisms in climate models serve as boundary inputs. These parameterized phenomenological relations introduce both model and parameter uncertainties.

The process of estimating model inputs based on measured data is typically termed *model calibration* or simply *parameter estimation* if inputs consist solely of parameters. The estimation of input uncertainties for a model using measured data is often referred to as *inverse uncertainty quantification*.

### **Coupled Systems**

The quantification of model discrepancies and input uncertainties is typically challenging for individual components of a system model. The difficulty grows

substantially as components are bidirectionally or tightly coupled to quantify multiscale or multiphysics phenomena. For such problems, the tight coupling generally prohibits a unidirectional propagation of discrepancies or uncertainties and instead necessitates a more global accommodation of these terms. Furthermore, the nature of inputs or parameters can change if they are actually states at another level in the process. The applications discussed in Sections 2.1–2.5 yield system models that are coupled to varying degrees, and the development of techniques to quantify model discrepancies and input uncertainties and construct prediction intervals for quantities of interest in such applications constitutes an active research area.

### 1.1.3 Numerical Errors and Uncertainties

*Computational results are believed by no one, except the person who wrote the code,* quoted by Max Gunzburger, Florida State University; original source unknown.

The characterization and regulation of numerical or algorithm errors is a topic in numerical analysis, and this is the least uncertain component of predictive sciences. Numerical errors and uncertainties include the following.

- Roundoff, discretization, or approximation errors.
- Bugs or coding errors.
- Bit-flipping, hardware failures, and uncertainty associated with future exascale and quantum computing.

Additionally, the 5–50 km grids required for numerical solution of field equations, in applications such as the weather model outlined in Example 1.1, are much larger than the scale of physics being modeled (e.g., turbulence or aerosol-induced cloud formation). This numerical requirement introduces uncertainty in phenomenological model relations.

### 1.1.4 Types of Uncertainty

The previous examples illustrate that modeling, experimental and numerical uncertainties, and errors can have various forms. The following definitions categorize uncertainties based on the degree to which they are inherent to the application or reflect lack of knowledge.

**Definition 1.6 (Aleatoric Uncertainty).** Also known as statistical, stochastic, or irreducible uncertainty, this is uncertainty inherent to a problem or experiment that in principle cannot be reduced by additional physical or experimental knowledge. Examples include uncertainties associated with nonphysical model parameters, subgrid atmospheric conditions such as wind gusts, subsurface microbe levels between wells, and initial conditions for weather models. Aleatoric uncertainties are typically unbiased and are often naturally defined in a probabilistic framework. Hence additional experiments or knowledge serve to better categorize the uncertainty.

**Definition 1.7 (Epistemic or Systematic Uncertainties).** Epistemic uncertainties are those that are due to simplifying model assumptions, missing physics, or basic lack of knowledge. Many of the previously mentioned modeling errors—e.g., phenomenological expressions for input or closure relations—and numerical errors are epistemic in nature. These uncertainties are often biased, and they are typically less naturally defined in a probabilistic framework.

If numerical errors are negligible, epistemic uncertainties are often termed *model errors*, *model discrepancies*, or *model inadequacies*. As detailed in Chapter 12, the quantification of these components for extrapolatory predictions outside the calibration domain constitutes an active research area.

The distinction between aleatoric and epistemic uncertainties is not always clear since lack of knowledge is relative and depends on current theory and experimental capabilities. One goal of uncertainty quantification is to reformulate epistemic uncertainties as aleatoric uncertainties where probabilistic analysis is applicable.

## 1.2 Predictive Estimation

A broad objective of predictive science is to use models, simulation codes, and experiments to predict system responses with quantified and reduced uncertainties. The probabilistic quantification of predicted experimental and computational outcomes with identified and quantified uncertainties is sometimes termed *predictive estimation*. As detailed in [52], predictive estimation is comprised of three components.

- Model Calibration: This involves the assimilation or integration of data to quantify and update input uncertainties associated with parameters, forcing functions, initial conditions, or boundary conditions.
- Model Prediction: Here one computes the response, or QoI, along with statistics, error bounds, or the probability density function (pdf) for the QoI.
- Estimation of the Validation Regime: This entails estimating contours of constant probability for the QoI to establish a domain for predictions with specified uncertainties.

To illustrate the predictive estimation process, we consider the mathematical model

$$y = f(\chi, q) \quad (1.4)$$

and statistical model

$$\Upsilon_i = f(\chi_i, q) + \delta(\chi_i) + \varepsilon_i, \quad (1.5)$$

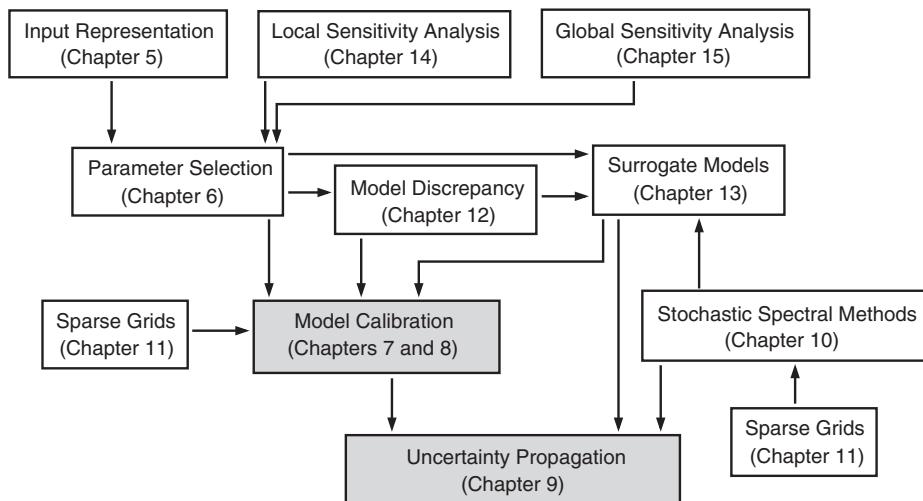
where  $q = [q_1, \dots, q_p]$  is a vector of inputs,  $y \in \mathbb{R}^\nu$  is the model output or QoI, and  $\chi$  denotes independent variables such as time  $t$  or space  $x$ . In the statistical model,  $\Upsilon_i$  and  $\varepsilon_i$  are random variables representing measurements and measurement errors and  $\delta(\chi_i)$  denotes biases due to epistemic or systematic uncertainties, as defined in

Definition 1.7. If numerical errors are negligible,  $\delta(\chi_i)$  quantifies the model error or model discrepancy.

In the following description, we indicate associated chapters in parentheses. The complete predictive estimation process is depicted in Figure 1.3.

### Predictive Estimation Process

- Input Representation (Chapter 5): One must first represent inputs in a probabilistic framework that facilitates model calibration and uncertainty propagation. This includes the construction of finite-dimensional representations for distributed or spatially varying parameters or initial conditions.
- Parameter Selection (Chapter 6): For many applications, the parameter dimension can be very large, e.g.,  $p = 100 - 10^6$ , which prohibits direct model calibration. Furthermore, parameters in many models are unidentifiable in the sense that they cannot be uniquely identified from the measured response. For such applications, one must employ parameter selection techniques to isolate a subset of most influential parameters  $\tilde{q} \in \mathbb{R}^{\tilde{p}}$ ,  $\tilde{p} < p$ , that are employed for subsequent analysis. This relies on the contents of Chapters 14 and 15.
  - Local Sensitivity Analysis (Chapter 14): Local sensitivity analysis focuses on the variability of the response as inputs are varied about a nominal value as quantified by the derivative  $\frac{\partial f}{\partial q_i}(q^*)$ .
  - Global Sensitivity Analysis (Chapter 15): Global sensitivity analysis quantifies how uncertainty in model responses can be apportioned to



**Figure 1.3.** Components of the predictive estimation process and relevant chapters. Model calibration and uncertainty propagation are the driving objectives. The remaining topics are required to achieve these objectives for large-scale applications.

uncertainties in inputs. We consider variance-based Sobol methods and screening algorithms based on approximation of the local index  $\frac{\partial f}{\partial q_i}(q^*)$  evaluated at random values  $q^*$  in the admissible parameter space.

- Surrogate Models (Chapter 13): For the large-scale applications detailed in Chapter 2, the complexity of simulation models generally precludes their direct use for model calibration and uncertainty quantification. This is addressed by constructing surrogate models  $y = \tilde{f}(\chi_i, \tilde{q})$  that encapsulate the primary behavior of the modeled process but are sufficiently efficient for model calibration, uncertainty propagation, and control implementation.
  - Stochastic Spectral Methods (Chapter 10): Stochastic Galerkin, collocation, or discrete projection methods provide one option for computing surrogate models.
  - Sparse Grid Methods (Chapter 11): Sparse grid methods are required to implement stochastic spectral methods and for direct implementation of Bayesian model calibration methods when parameter dimensions are moderate; e.g.,  $p = 8 - 50$ .
- Model Discrepancy (Chapter 12): For applications that exhibit epistemic or systematic uncertainties due to model discrepancy or numerical errors, one must quantify the bias term  $\delta(\chi_i)$  in (1.5) using physical, mathematical, or statistical analysis.
- Model Calibration (Chapters 7 and 8): Frequentist or Bayesian techniques are used to quantify the uncertainties associated with inputs  $q$  based on measured data  $y$ . In the Bayesian framework, inputs are formulated as random variables with associated pdf. For moderate input dimensions  $p$ , the sparse grid techniques of Chapter 11 can be used to directly evaluate Bayes' relation to avoid sampling-based Metropolis algorithms.
- Uncertainty Propagation (Chapter 9): The final objective of predictive estimation is to propagate input uncertainties through models to construct prediction intervals or pdf for QoI. For linearly parameterized problems, one can establish analytic relations for statistical moments. For mildly nonlinear problems, linearization using Taylor expansions can achieve the same objective. More generally, one can employ stochastic polynomial or sampling-based methods, such as those employed for model calibration, to compute moments or construct prediction intervals for QoI.

For large-scale applications, one must typically employ model selection techniques, address model discrepancies, and construct surrogate models before model calibration and uncertainty propagation can be achieved. For applications with identifiable parameter sets, negligible epistemic uncertainties, and highly efficient simulation codes, however, one can focus immediately on the model calibration and uncertainty propagation techniques detailed in Chapters 7, 8, and 9. We refer readers to [52] for details regarding the estimation of the validation regime.

## Chapter 2

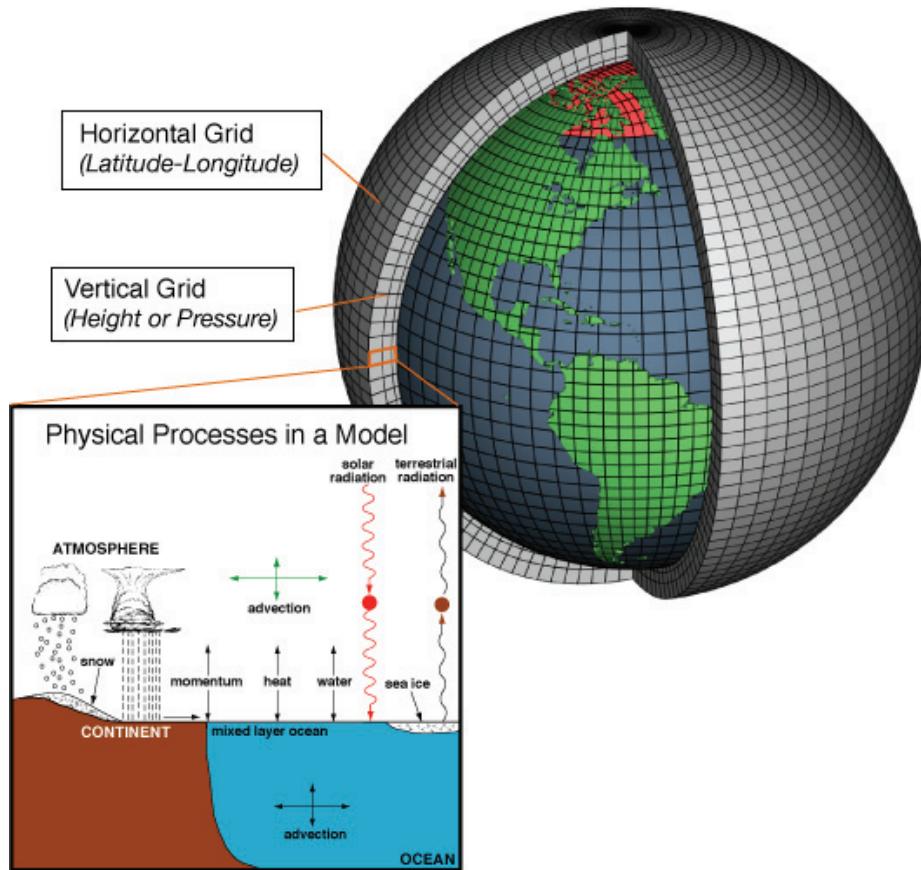
# Large-Scale Applications

In this chapter, we illustrate five applications where model predictions with quantified uncertainties are critical for understanding and predicting scientific phenomena and making informed decisions and designs based on these predictions. These applications are weather models, climate models, subsurface hydrology and geology models, nuclear reactor designs, and models for biological phenomena.

### 2.1 Weather Models

If asked to list areas of science in which uncertainty quantification is critical for predictive estimation, most people would likely include weather forecasting. Moreover, some would do so with a sense of derision while noting the occasional inaccuracy of forecasts. This negativity is due in part to the fact that the role of uncertainty in scientific predictions receives little attention in secondary and undergraduate curricula, which results in a potentially false sense of security regarding scientific predictions. This leads to a poor understanding of weather forecasts—e.g., surveys reveal that many people interpret a 50% chance of rain as meaning that half the viewing region will receive precipitation—and a general distrust of the science associated with weather forecasting. When one considers the complexity of the underlying phenomena and associated models, the inherently chaotic or unstable nature of the prediction process, and the uncertainties associated with the models, simulation codes, and data, the accuracy of forecasts with quantified uncertainties represents a major tour de force of physical modeling and scientific computation.

To motivate the complexity of modeled phenomena, one need only consider factors required to predict temperatures, precipitation, and winds. As illustrated in Figure 2.1, temperature in the atmosphere depends on the absorption and emission of radiation, latent heat release, advection due to winds, and convective heating or cooling at the earth’s surface. The temperature field thus depends on the horizontal and vertical distribution of small particulates and liquid droplets (excluding cloud droplets and precipitation) that are collectively termed *aerosols*. Additionally, it depends on wind patterns that produce warm and cold air advection, and the surface



**Figure 2.1.** Physical processes that must be incorporated in weather and climate models and the associated 3-D grid. Image courtesy of NOAA.

topography and heat profile. Furthermore, phase transitions between liquid, solid, and vapor phases for atmospheric moisture add or remove latent heat from the atmosphere, thus requiring that these effects be coupled with temperature models.

Aerosols are associated with atmospheric conditions, such as dust or smog levels, and are catalysts for phenomena such as cloud formation since they serve as condensation nuclei around which cloud droplets form. The smallest aerosols have radii on the order of  $0.1 \mu\text{m}$  and are typically attributed to chemical conversion of sulfate gases to liquids or solids. First principles models for these processes thus occur on relatively small space and time scales and require quantification of the associated chemical processes. Larger aerosols include wind-driven dust particles, particulates from volcanic reactions, and combustion byproducts.

Changes in temperature produce pressure gradients that in turn generate air movement and wind. Near the earth's surface, wind flow is significantly influenced

by the terrain and surface conditions such as temperature. This produces highly complex frictional and boundary layer effects. At higher altitudes, the situation is slightly less complicated but is still fully coupled with all of the previously mentioned phenomena. Hence high altitude flow patterns, such as the jet stream, tend to be somewhat more stable than surface wind patterns, but they still exhibit highly turbulent dynamics, instabilities, and bifurcations due to the highly nonlinear and complex coupling with atmospheric conditions.

### 2.1.1 Conservation Relations

The physical components of meteorological or weather models are based on conservation of mass, momentum, and energy in combination with conservation of water phases and constituent chemical spaces in aerosol models.

As detailed in [193, 213], conservation of energy using the first law of thermodynamics yields the partial differential equation

$$\rho c_V \frac{\partial T}{\partial t} + p \nabla \cdot v = -\nabla \cdot F + \nabla \cdot (k \nabla T) + \rho \dot{q}(T, p, \rho), \quad (2.1)$$

where  $T$  and  $v$  are the temperature and velocity at a point  $(x, y, z) \in \mathbb{R}^3$  and  $\rho, c_V, p$ , and  $k$  respectively denote the density, specific heat at constant volume, pressure, and thermal conductivity for an infinitesimal atmospheric volume  $V$ . Furthermore,  $F$  is the net radiative flux and  $\dot{q}(T, p, \rho)$  is the rate of internal heating or cooling associated with processes such as latent heat release due to phase changes in atmospheric moisture. The pressure, temperature, and density are typically related by the ideal gas law

$$p = \rho R T, \quad (2.2)$$

where  $R$  is the specific gas constant for air.

Atmospheric flow dynamics, such as the jet stream and local wind patterns, are quantified using conservation of mass and momentum, which yields the relations

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho v) = 0 \quad (2.3)$$

and

$$\frac{\partial v}{\partial t} = -v \cdot \nabla v - \frac{1}{\rho} \nabla p - g \hat{k} - 2\Omega \times v. \quad (2.4)$$

Here  $g$  is the force due to gravity and  $-2\Omega \times v$  is the Coriolis force due to the earth's rotation. The expansion of the nonlinear term  $v \cdot \nabla v$  and individual equations in spherical coordinates can be found in [196].

It is shown in [193] that the concentration of water, in the solid, liquid, and gaseous phases, and aerosols can also be quantified using conservation relations. If we let  $m_1, m_2$ , and  $m_3$  denote the mass of the solid, liquid, and gaseous water phases relative to the mass of air in the same volume, the concentration of each phase can be represented by the relation

$$\frac{\partial m_j}{\partial t} = -v \cdot \nabla m_j + S_{m_j}(T, m_j, \chi_j, \rho), \quad j = 1, 2, 3. \quad (2.5)$$

The source or sink terms  $S_{m_j}$  quantify the processes that govern phase transitions. These can be extremely complex due to their dependence on aerosol concentrations  $\chi_j$  and coupled atmospheric processes. The physics associated with the terms  $S_{m_j}$  is often difficult or impossible to establish on the grid scales illustrated in Figure 2.1, thus necessitating phenomenological parameterizations.

Aerosol concentrations are quantified in a similar manner. If we consider  $J$  species with concentrations  $\chi_j$ , conservation principles yield the relations

$$\frac{\partial \chi_j}{\partial t} = -v \cdot \nabla \chi_j + S_{\chi_j}(T, \chi_j, \rho), \quad j = 1, \dots, J, \quad (2.6)$$

where  $S_{\chi_j}$  incorporate changes in state, chemical reactions, and sedimentation. Like  $S_{m_j}$ , parameterizations are typically required to construct these terms.

The set of equations

$$\begin{aligned} \frac{\partial \rho}{\partial t} + \nabla \cdot (\rho v) &= 0, \\ \frac{\partial v}{\partial t} &= -v \cdot \nabla v - \frac{1}{\rho} \nabla p - g \hat{k} - 2\Omega \times v, \\ \rho c_V \frac{\partial T}{\partial t} + p \nabla \cdot v &= -\nabla \cdot F + \nabla \cdot (k \nabla T) + \rho \dot{q}(T, p, \rho), \\ p &= \rho R T, \\ \frac{\partial m_j}{\partial t} &= -v \cdot \nabla m_j + S_{m_j}(T, m_j, \chi_j, \rho), \quad j = 1, 2, 3, \\ \frac{\partial \chi_j}{\partial t} &= -v \cdot \nabla \chi_j + S_{\chi_j}(T, \chi_j, \rho), \quad j = 1, \dots, J, \end{aligned} \quad (2.7)$$

are often referred to as the *equations of atmospheric physics*. If one neglects the species relations for  $m_j$  and  $\chi_j$  and employs hydrostatic approximations to the momentum equation, one obtains what are often termed the *primitive equations* for Eulerian fluid motion. Various meteorological and climate models are constructed by employing simplified forms of these relations in combination with parameterizations for  $F, \dot{q}, S_{m_j}$ , and  $S_{\chi_j}$ .

We note that meteorological models for tropical dynamics are significantly more complex than those for midlatitude or extratropical regions, e.g., poleward from about  $30^\circ$  latitude. In the middle latitudes, the primary source of energy driving wind patterns is temperature-induced pressure gradients in balance with Coriolis forces, and latent heat release and radiative heating are secondary contributors to atmospheric dynamics. Here geostrophic or quasi-geostrophic theory, based on a balance of pressure gradients and Coriolis forces, provides simplified momentum relations for meteorological models.

In the tropics, however, temperature gradients are smaller and latent heat release associated with convective cloud systems is a more significant source of energy. Moreover, Coriolis forces are also smaller and there is a more significant coupling between atmospheric and ocean temperatures. Meteorological and regional climate models for these regions must thus incorporate the interaction between cumulus

convection and mesoscale and large-scale circulations as well as equatorial wave dynamics and air-ocean interactions. Resulting weather phenomena include monsoons, the trade winds, hurricanes, and El Niño. Readers are referred to Chapter 11 of [113] and the included references for more details regarding tropical weather phenomena and associated modeling issues.

### 2.1.2 Phenomenological Models

To numerically solve the coupled relations (2.7), expressions for the net radiative flux  $F$ , rate  $\dot{q}(T, p, \rho)$  due to latent heat release, and source terms  $S_{m_j}(T, m_j, \chi_j, \rho)$  and  $S_{\chi_j}(T, \chi_j, \rho)$  must be specified. For phenomena such as radiation transfer, modeling relations can be based on physical principles. However, the terms  $S_{m_j}$  and  $S_{\chi_j}$  are highly complex and occur on scales that are much smaller than computational grids. This necessitates the construction of phenomenological models having nonphysical parameters that are determined through model calibration via data assimilation or from independent experiments.

To illustrate, it is established in [193] that if  $m_2$  represents the concentration of water in liquid phase, then a simplified form of  $S_{m_2}$  is

$$S_{m_2} = S_1 + S_2 + S_3 - S_4, \quad (2.8)$$

where  $S_1$  represents the conversion of cloud droplets to form raindrops,  $S_2$  represents the accretion of cloud water by raindrops,  $S_3$  represents the melting of snow or ice to rain, and  $S_4$  quantifies the evaporation of rain. An accepted relation for  $S_1$  is

$$S_1 = \bar{\rho}(m_2 - m_2^*)^2 \left[ 1.2 \times 10^{-4} + \left( 1.569 \times 10^{-12} \frac{n_r}{d_0(m_2 - m_2^*)} \right) \right]^{-1}, \quad (2.9)$$

where  $\bar{\rho}$ ,  $m_2^*$ ,  $n_r$ , and  $d_0$  are constants that must be specified.

Parameterized phenomenological models are required for a range of atmospheric and terrestrial phenomena, including aerosol-induced cloud formulations, reactions that produce aerosols, turbulence at various levels in the atmosphere, and drag and surface effects due to mountains and attributes of the terrain. Quantification of uncertainties in parameters and phenomenological models is necessary for quantifying uncertainties in predictions of QoI.

### 2.1.3 Simulation Models

Essentially every weather person refers to predictions resulting from *computer models*. Care must be exercised when interpreting this phrase since it really implies simulations obtained using discretized physical models.

All approximation techniques yield atmospheric state values on a 3-D grid such as that depicted in Figure 2.1. The horizontal and vertical grid spacing is determined by a number of factors, including the spatial scales of modeled physics and available computing resources. Local meteorological models employ horizontal grids on the order of 5 km with vertical spacing of approximately 200 m. Global

meteorological and climate models necessarily employ larger horizontal grids that are on the order of 50–100 km.

The gridsizes required for weather and climate models constitute a significant source of uncertainty since parameterized processes such as aerosol-induced cloud formation, latent heat generation due to cloud formation, and atmospheric turbulence occur on much smaller, subgrid, scales.

Various discretization techniques can be employed to approximate the relations (2.7). Finite difference techniques are typically employed to discretize vertical spatial derivatives, whereas finite difference, or occasionally finite element, techniques are employed for the horizontal components of regional and some global models. Other global models exploit periodicity by employing spectral approximation methods.

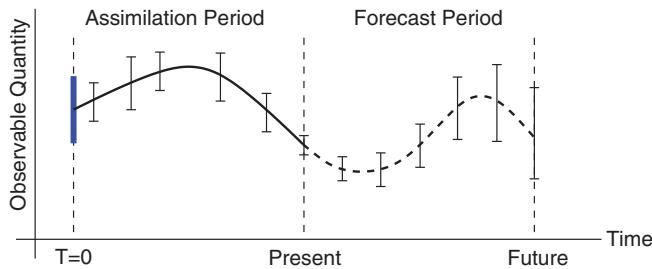
Semidiscretization in space yields very large, coupled, vector-valued systems of differential equations that must be numerically integrated forward in time to provide predictions. The use of explicit methods introduces stringent limits on temporal stepsizes due to stability or Courant–Friedrichs–Lowy (CFL) conditions that can limit the utility of algorithms. For example, to quantify gravity waves in jet streams that have maximum velocities of 200 m/s, horizontal grids of 100 km yield maximum time steps of approximately 8 minutes. To address this, numerical weather prediction (NWP) models commonly employ semi-Lagrangian integration techniques that approximate the path of air parcels (Lagrangian perspective) while predicting values on a fixed (Eulerian) grid. The advantage is that large temporal stepsizes can be employed without loss of stability. Moreover, semi-Lagrangian schemes can be constructed to ensure that species concentrations are conserved during advection.

The length of simulations varies for differing models. The UK Met Office Unified Model is run six days into the future, whereas the European Centre for Medium-Range Weather Forecasts’ (ECMWF) Integrated Forecast System provides 10 day predictions. The decline of accuracy for longer predictions is further discussed in the context of uncertainty quantification for weather forecasts.

### 2.1.4 Model Calibration, Data, and Data Assimilation

Meteorological models used for weather forecasting are essentially *initial value problems*. Based on measurements, one attempts to determine initial conditions and parameters so that models match present atmospheric conditions, as closely as possible, as illustrated in Figure 2.2. Data assimilation techniques are used to calibrate models by determining inputs (initial conditions and parameters) based on a wide range of measurements. Models are then numerically integrated forward in time to provide forecasts. Due to the nonlinear nature of the models, they are *chaotic* in the sense that uncertainties in initial conditions grow in an unstable manner, thus significantly diminishing the accuracy of forecasts for increasingly long future periods. This highlights the necessity of considering ensembles and statistical QoI, such as average temperatures, and quantifying the uncertainty in predictions.

A critical component of model calibration via data assimilation is globally distributed and frequently obtained earth-surface and atmospheric data. In the United



**Figure 2.2.** Assimilation period to estimate initial conditions and parameters so the model best fits data with measurement uncertainties. Forecast with quantified uncertainties.

States, there are approximately 1000 surface measurement locations, whereas the World Meteorological Organization (WMO) maintains approximately 10,000 land sites worldwide. Surface measurements on the ocean are obtained by moored and drifting buoys as well as ships on an array of routes. Atmospheric conditions are provided by radiosondes in weather balloons that rise into the stratosphere, weather satellites, commercial aircraft on prescribed routes, and reconnaissance aircraft that can be sent to regions yielding high impact data, e.g., regions where there is significant uncertainty or storms such as hurricanes.

In combination, this yields a fairly rich data environment. From the perspective of model construction, prediction, and uncertainty quantification, however, three properties of the data are important: it is measured on highly irregular grids and it is frequently distributed in time, the observations are not direct measurements of state variables, and there is uncertainty associated with all data. The limited accuracy of measuring devices constitutes one source of uncertainty. Second, several are moving, so there are varying degrees of uncertainty associated with the position and time of measurements.

To illustrate issues pertaining to data assimilation, we assume a grid of size  $432 \times 320 \times 50$  with a minimum of 8 variables comprising the wind speed, temperature, pressure, and moisture phases. This yields  $5.53 \times 10^7$  states  $x$ . Phenomenological components of the models can easily require in excess of 50 parameters in addition to initial conditions. Since observations are on the order of thousands, the problem is highly underdetermined.

To construct a functional to be minimized, we let  $y_i$  denote the vector of observations at times  $t_i$  where we assume  $n$  measurement times. States  $x$  are mapped to observations by the operator  $H$  so that  $y = H(x)$ . Prior information  $x^B$ , obtained from a previous model forecast, is often termed the *background*. The error and background error covariance matrices are denoted by  $H_i$  and  $B$ . Predictions  $x_{i+1}$  provided by the NWP model, based on current states  $x_i$ , can be represented as

$$x_{i+1} = M_{i+1,i}(x_i), \quad (2.10)$$

where  $M_{i+1,i}$  represents the discretization of the nonlinear model from time  $t_i$  to time  $t_{i+1}$ .

The data assimilation algorithm 4D-VAR is employed at several weather prediction centers, including the ECMWF and the UK Met Office as well as stations in Japan and Canada. In this algorithm, one minimizes the functional

$$J(x_0) = (x_0 - x_0^B)^T B^{-1} (x_0 - x_0^B) + \sum_{i=0}^n (y_i - H(x_i))^T R_i^{-1} (y_i - H(x_i)) \quad (2.11)$$

subject to the model dynamics

$$x_i = M_{i,0}(x_0).$$

A primary difference between 4D-VAR and the previous algorithm 3D-VAR is the use of adjoints to incorporate the times at which observations are made. Whereas ensemble Kalman filters are being investigated for NWP models, 4D-VAR is presently considered the state of the art.

We emphasize the fact that determination of initial conditions, often termed the *analysis*, and model parameters that provide a best fit to subsequent observations is central to NWP. This is in contrast to climate models which are essentially forced boundary value problems that are run until the transient effects of initial conditions are mitigated.

### 2.1.5 Sources and Nature of Uncertainties

There are four primary sources of uncertainty or errors in NWP models: model limitations, input uncertainties due to initial conditions, boundary conditions or parameters, numerical errors, and uncertainties in measurements. When combined with the highly nonlinear nature of models, these produce response uncertainties that grow as a function of time and must be quantified to provide a meaningful context in which to interpret predictions or forecasts.

Although they are based on physical conservation laws, the continuity, momentum, temperature, and species equations (2.7) are still approximations of the underlying physical phenomena. Moreover, the phenomenological relations used to model the flux and rate terms  $F$  and  $\dot{q}$  and source terms  $S_{m_j}$  and  $S_{\chi_j}$  are approximate representations for highly complex and often only partially understood physical phenomena such as turbulence, aerosol-induced cloud formation, and resulting latent heat release as precipitation forms. These uncertainties are primarily epistemic, as defined in Definition 1.7.

As illustrated in (2.9), the phenomenological components of the models contain parameters that are often nonphysical and hence cannot be correlated with measured data. Thus both their values and variability must be inferred through model calibration or data assimilation techniques. Even physical parameters such as the thermal conductivity  $k$  will exhibit some variability due to varying atmospheric conditions and the fact that they may partially accommodate unmodeled physics. In meteorological models, the initial atmospheric conditions constitute a second critical source of uncertainty that must be inferred from later observations. These uncertainties are aleatoric or stochastic in the sense defined in Definition 1.6.

The numerical discretization of the models introduces uncertainty and errors in two fundamental ways: the solution on a 3-D grid introduces significant uncertainty for parameterizations of subgrid scale physics, and the approximation techniques introduce discretization errors. The first is critical for a number of phenomena such as turbulence and aerosol-induced cloud formation which occur at the level of meters, whereas horizontal grids are on the order of tens to hundreds of kilometers. This uncertainty is obviously related to the previously mentioned model uncertainty. Discretization errors can often be asymptotically quantified using theory associated with finite difference, finite element, spectral, or semi-Lagrangian techniques. These are errors that are typically epistemic in nature.

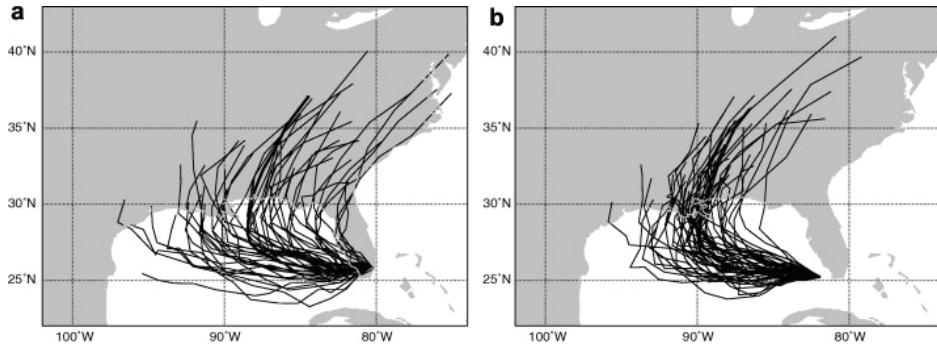
Uncertainty in the data can arise from two sources: limited accuracy of measurement devices and uncertainty associated with variability in the location and time of moving sensors. The first is often categorized as aleatoric, whereas the second is primarily epistemic.

### 2.1.6 Role of Uncertainty Quantification in Weather Forecasts

As illustrated in Figure 2.2, uncertainty quantification for weather forecasting takes place in two steps. In the model calibration step, values and uncertainties associated with inputs, such as initial conditions and parameters, are determined using data assimilation techniques such as 4D-VAR. As detailed in Chapter 8, this is often performed in a Bayesian framework, as evidenced by the priors  $x_0^B$  in (2.11). In the second step, the calibrated models are run forward in time to provide forecasts with quantified uncertainties.

In the 1970's and 1980's, it was recognized that forecasts obtained with a single model simulation or realization had limited utility due to the inherent uncertainty and chaos induced by the highly nonlinear initial value models. This led to the use of *ensemble* forecasts, which have been standard since the 1990's. In single model approaches, ensemble forecasts are obtained by running multiple simulations from an individual model with differing initial conditions or parameter values drawn from probability densities constructed during the calibration phase. Using the ensemble predictions, one constructs statistical QoI, such as the average temperature, relative humidity, or projected rain amounts. Using ensemble forecasts, a 50% chance of rain two days in the future means that given the present atmospheric conditions, half of the simulations predict measurable rain amounts at some random point in the specified area. Improved forecasts can be obtained using *multimodel ensemble forecasts* in which ensemble predictions from multiple models are used to construct QoI and uncertainty bounds. The reduction in variability of ensemble forecasts for the hurricane Katrina, obtained with an additional 12 hours of data, is illustrated in Figure 2.3. The hurricane position was very near the mode of the ensemble predictions when it made landfall near New Orleans.

Whereas uncertainty bounds or probability densities for QoI are constructed during the ensemble computations, they typically are not reported in forecasts. One exception is the prediction of large storms such as cyclones, tropical storms, or hurricanes. In these cases, forecasts usually include both the predicted trajectory and cones of uncertainty, as illustrated in Figure 2.4.



**Figure 2.3.** ECMWF ensemble forecasts made at (a) 00 UTC and (b) 12 UTC on August 26, 2005. Katrina made landfall near New Orleans at 12 UTC on August 29, 2005.

### 2.1.7 References

The topic of weather modeling and prediction is vast, and we have summarized only some aspects of the discipline to illustrate the role of uncertainty quantification for predictive estimation. Additional discussion regarding basic weather phenomena can be found in [28, 173], whereas detailed derivations of the atmospheric physics relations (2.7) and parameterized constitutive relations are provided in [123, 193, 213]. These references also contain an overview of numerical techniques for simulating weather models including semi-Lagrangian integration techniques. The implementation of data assimilation algorithms, such as 4D-VAR, is addressed in [129, 182], and [129, 150] provide further details about ensemble forecasting. Finally, the reader is referred to [263] for discussion regarding statistical issues and techniques associated with weather modeling and prediction.



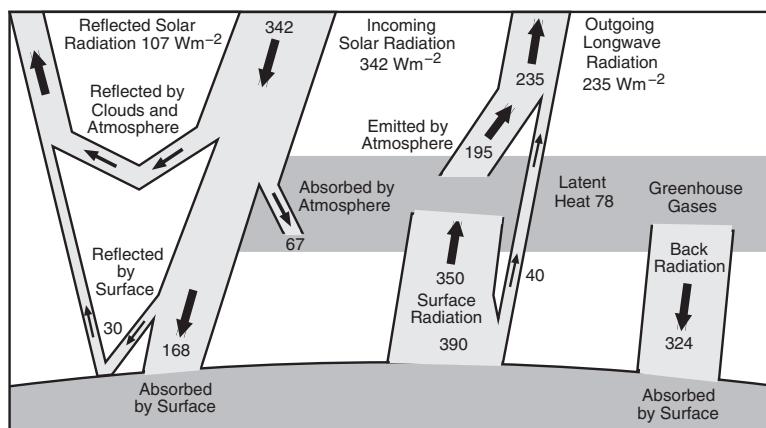
**Figure 2.4.** NOAA image of the trajectory and cone of uncertainty for Katrina.

## 2.2 Climate Models

The impression many people have of climate models is that they are simply weather models extended to much longer timescales of years, decades, centuries, or millennia. They subsequently conclude that since two-week weather predictions are typically poor, the accuracy of climate models must therefore be very suspect. Whereas climate and weather models both require the quantification of coupled atmospheric, land, ocean, and solar radiation processes, the vastly differing timescales dictate that different physical mechanisms be emphasized in the two modeling regimes.

As detailed in Section 2.1, weather models are highly nonlinear, initial value problems whose chaotic nature necessitates ensemble computations to predict statistical QoI whose accuracy is typically limited to 10 days to two weeks. Hence they are based on atmospheric conditions such as wind patterns, radiative, convective, and latent heat driven temperature changes, and aerosol and moisture levels on fairly short timescales. They can thus neglect seasonal effects, long-term anthropogenic and natural forcing terms such as increased CO<sub>2</sub> levels and volcanic ash, and influences such as deforestation. The emphasis is to use data assimilation techniques such as 4D-VAR to determine initial atmospheric conditions and parameter values so that models match recent and current conditions in a statistically accurate manner. Ensemble predictions are then used to compute future QoI such as temperature, precipitation levels, and storm tracks.

Climate models differ in the sense that they are required to accurately maintain a balance between absorbed solar energy and lower frequency infrared radiation emitted to space—typically termed the earth’s *global energy balance* or *energy budget*—for decades up to centuries. As illustrated in Figure 2.5, this energy budget is influenced by numerous natural and anthropogenic factors including greenhouse gas levels, seasonal effects, volcanic eruptions, deforestation, ocean dynamics, and polar ice coverage. The timescales dictate that the transient effects due to initial atmospheric and terrestrial conditions are essentially negligible. Instead, compu-



**Figure 2.5.** *Earth’s energy budget modified from [137].*

tation of the energy budget requires quantification of energy fluxes at the earth's surface and sources and sinks in the atmosphere. Hence climate models exhibit the dynamics of forced boundary value problems. *One ramification is that chaotic dynamics associated with weather models are largely mitigated in climate models.* As with weather models, one goal is to compute statistical QoI such as long-term change in CO<sub>2</sub> or temperature levels, with quantified uncertainties. The scope in climate models is much broader, however, since it additionally involves questions such as the following.

- Is the planet getting warmer and are manmade processes the cause?
- Are atmospheric and/or oceanic circulation patterns or currents changing and, if so, what will be the effect?
- Are the weather and climate becoming more extreme or variable and, if so, what are the ramifications?

Answers to these questions, with quantified uncertainties, are required to address societal questions.

- Will climate changes lead to improved agriculture and food supplies or reduced supplies due to widespread drought?
- Will sea-level changes threaten large civilization centers?
- Will changes in ozone levels significantly increase the incidence of cancer?

In subsequent discussion, we highlight ways in which uncertainty quantification is critical for obtaining climate predictions with uncertainties quantified in a manner that informs both scientists and policy makers.

### 2.2.1 Climate Forces and Feedback

The equations (2.7) quantify the basic processes associated with atmospheric physics. The difference in how these relations are used to construct weather and climate models lies in the simplifying assumptions and parameterized phenomenological models used to quantify the net radiative flux  $F$ , rate of internal heating  $\dot{q}$ , and source terms  $S_{m_j}$  and  $S_{\chi_j}$ . We summarize here physical phenomena and sources of uncertainty associated with the radiative fluxes since these are central to the global energy balance that drives climate models.

We illustrate in Figure 2.5 factors that influence the balance between solar energy absorbed by the atmosphere and earth and infrared radiation emitted to space. As detailed in [137], approximately 342 W/m<sup>2</sup> of solar radiation enters the earth's atmosphere, where it is absorbed or reflected by the atmosphere, ground, ocean, or surface ice. In the atmosphere, the primary absorbers are water vapor in the troposphere and ozone in the stratosphere. Clouds are the primary mechanism that scatter and reflect radiation. On the earth's surface, oceans and phenomena such as deforestation and changing polar icecaps directly affect absorption and reflection rates. Visible light constitutes the primary wavelength reaching the earth's surface with lesser amounts of UV and near infrared (heat) radiation.

Radiation due to heat at the earth's surface and reflected solar radiation both contribute to the longer wavelength infrared radiation that is emitted into space. The amount of infrared radiation is governed by cloud and water vapor levels along with greenhouse gas concentrations.

Latent heat associated with cloud formation and precipitation constitutes the primary nonradiative heat source in the atmosphere. A significant emphasis in climate modeling focuses on quantifying the mean behavior and uncertainties associated with these radiative and nonradiative processes.

Climate forcing mechanisms are defined as changes imposed on the earth's energy balance that produce changes in the climate. These can include external changes due to variability in the earth's orbit, fluctuations in solar radiation, and comet or meteor impacts, or internal factors such as volcanic eruptions, deforestation, or changes in aerosol and CO<sub>2</sub> levels. We note that the internal forces can be both natural and human-induced. Feedback processes are those in which changes in the climate state serve as forces that produce further climate changes. Examples include changes in cloud cover due to aerosols, changes in surface reflection due to melting polar icefields, and changing greenhouse gas levels due to increased temperature-induced evaporation. We point out that all of the climate forcing and feedback mechanisms are quantified using phenomenological models, often having a large number of nonphysical parameters. This introduces significant uncertainty that must be quantified in final climate model predictions.

Various natural and human-induced forcing mechanisms, along with uncertainties and a qualitative indication of the level of scientific understanding, are illustrated in Figure 2.6. We summarize next aspects of these mechanisms and indicate how associated uncertainties influence climate models and predictions.

### Natural External and Internal Forcing Mechanisms

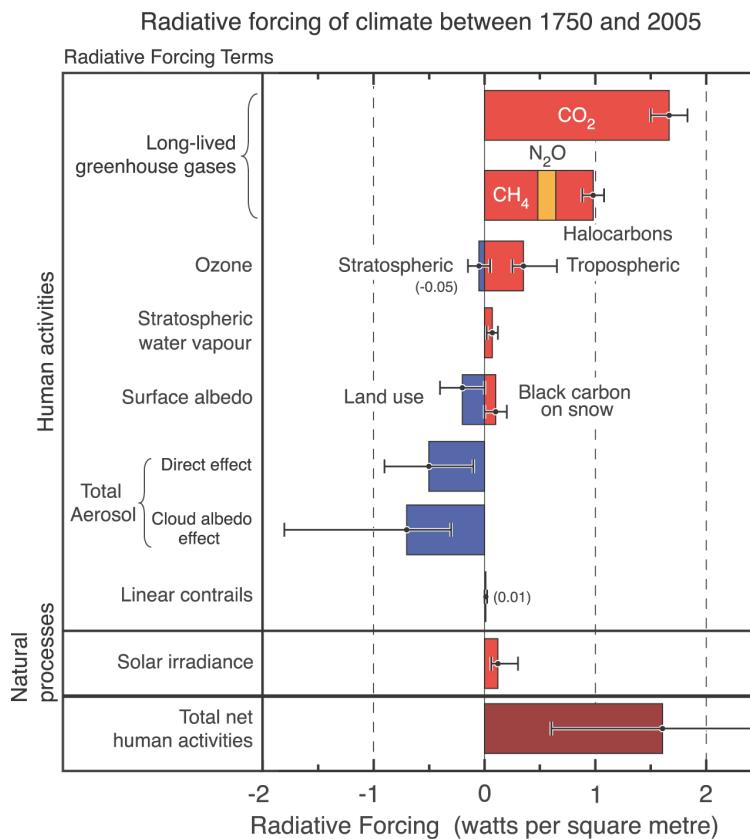
Whereas we cannot control natural forces, their quantification is necessary to determine their relative influence compared with anthropogenic forces that we can control.

#### *Solar Radiation*

There are two primary mechanisms that affect the level of solar radiation entering the earth's atmosphere: variability in the earth's orbit and fluctuations in solar activity. The periodicity of sunspot activity with approximately 11-year cycles has long been known, and the resulting variation in solar radiation is incorporated in models. However, the effect of this variability on weather and climate is still debated. Since the 1970's, radiation data has been measured by satellites and long-term variability is inferred from carbon data in tree rings. Sources of uncertainty include the parameterized models and measured data.

#### *Volcanic Eruptions*

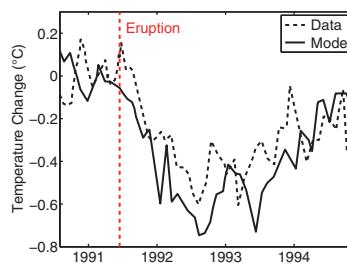
The eruption of Krakatoa in 1883 caused average global temperatures to drop by approximately 1 °C in the subsequent year and produced noticeable variability in weather for approximately 5 years. The eruption of Mount Pinatubo in 1991 occurred after the advent of a global monitoring network, so it significantly advanced



**Figure 2.6.** Contribution and uncertainties associated with climate forcing mechanisms from the 2007 IPCC Report [228, Figure SPM.2]. Level of scientific understanding – Long-lived greenhouse gases: High; Ozone: Med; Surface albedo: Med-Low; Total aerosol: Med-Low; Solar irradiance: Low.

our understanding of how large volcanic eruptions could force climate changes. Furthermore, it provided a unique opportunity to advance and test volcanic inputs to climate models.

Volcanic forcing is due to the high levels of particulates and gases that are introduced into the atmosphere. The manner in which these aerosols affect the earth's energy balance is largely dependent upon the height to which they are injected. Nonabsorbing aerosols reduce the amount of solar radiation that reaches the earth's surface, whereas greenhouse effects are increased by aerosols that absorb and emit in the infrared spectrum. The aerosols produced by Mount Pinatubo reduced the solar energy reaching the earth's surface by  $3\text{-}4 \text{ W/m}^2$  and cooled global temperatures by approximately  $0.5^\circ\text{C}$ , as illustrated in Figure 2.7. The climate changes due to volcanoes are relatively short-term unless they occur in conjunction with other climate forces or feedbacks.



**Figure 2.7.** Measured and predicted tropospheric temperatures surrounding the June 15, 1991 Mount Pinatubo eruption; data from [106].

### Human-Induced Forcing Mechanisms

The quantification of anthropogenic climate forcing mechanisms is of fundamental importance since we can control them.

#### Greenhouse Gas Emissions

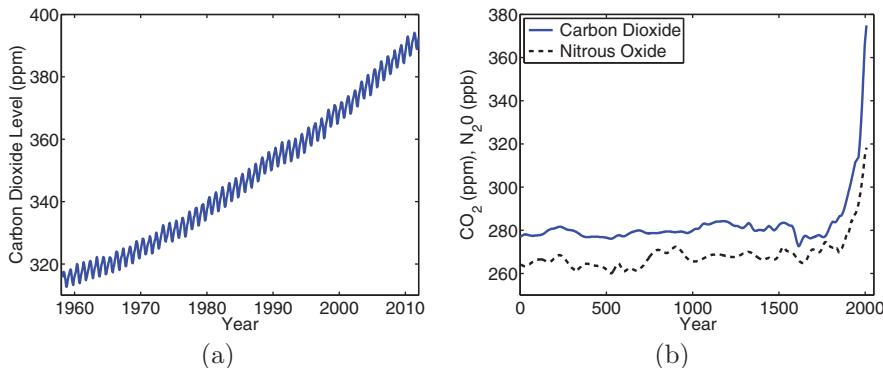
As illustrated in Figure 2.5, the *greenhouse effect* is the process in which thermal radiation from the earth's surface is absorbed and reflected by greenhouse gases such as water vapor, CO<sub>2</sub>, methane, ozone, and chlorofluorocarbons (CFCs) such as the refrigerant Freon. Because much of the thermal energy is radiated back to earth, increased levels of greenhouse gases produce an elevation in average surface temperatures. The name is somewhat of a misnomer since warming is due to changes in the absorption and reflection of radiated thermal energy rather than restriction of convective heat loss, as is the case in a glass greenhouse.

Warming due to increased greenhouse gas levels constitutes one of the most heavily studied areas of anthropogenic climate forcing, and it was stated in the 2007 IPCC Assessment Report that “It is very likely that greenhouse gas forcing has been the dominant cause of the observed warming of globally averaged temperatures in the last 50 years” [228].

Because CO<sub>2</sub> is the most abundant greenhouse gas after water vapor and because it is the byproduct of burning fossil fuels, its concentration levels have been extensively studied and documented. CO<sub>2</sub> concentrations monitored at Mauna Loa, Hawaii, since 1958 are plotted in Figure 2.8(a). The fluctuations have not adequately been explained but are likely due to complicated feedback with short-term atmospheric conditions.

Ice-core data has been used to establish CO<sub>2</sub> concentrations for the past 800,000 years at locations such as Antarctica and Greenland. Figure 2.8(b) illustrates the increase of CO<sub>2</sub> concentrations for the last 2000 years. Because measured carbon isotopes can be used to establish the sources, the figure also illustrates that the increase in CO<sub>2</sub> levels since the Industrial Revolution in the mid-nineteenth century is largely influenced by CO<sub>2</sub> emission from fossil fuels.

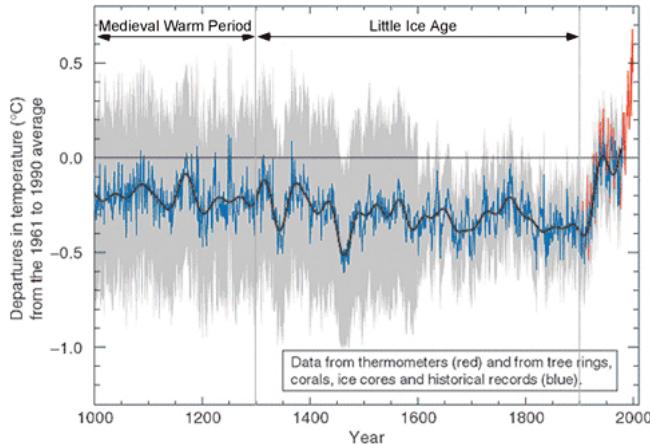
Based on data published by [164], the 2001 IPCC published Figure 2.9, which illustrated reconstructed temperatures from 1000 to 1980, measured temperatures from 1902 to 1999, 40-year smoothed averages, and two standard error limits [115].



**Figure 2.8.** (a) Concentration of  $CO_2$  measured at Mauna Loa; data from <http://www.esrl.noaa.gov/gmd/ccgg/trends/>. (b) Concentrations of greenhouse gases over the last 2000 years; data from 2007 IPCC report [228, Figure SPM.1].

The reconstructed temperatures are based on tree ring data, coral, ice core measurements, and historical records. CO<sub>2</sub> and temperature results of the type plotted in Figures 2.8 and 2.9 have been cited by numerous scientists and policy makers as evidence that increasing anthropogenic CO<sub>2</sub> levels are leading to unprecedented temperature increases which in turn produce global warming.

While accepted by a large percentage of the scientific community, this topic is still highly contentious and serious scientific arguments have been made questioning the data, conclusions, future predictions and resulting policies, and even the fundamental viability of mathematical and statistical models as a means of providing



**Figure 2.9.** Proxy temperatures from 1000–1980, measured temperatures from 1902–1999, 40-year smoothed averages, and two standard error limits; from 2001 IPCC Report [115, SPM, Figure 1].

meaningful climate predictions. While the sources of scientific disagreement are numerous, nearly all can be distilled to differing assumptions employed in statistical analysis or differences in how uncertainties are quantified (or neglected) in data, models, and predicted QoI. We summarize only a few representative examples of where uncertainty must be adequately quantified and incorporated to establish and predict the degree of anthropogenic global warming.

- The ice core and tree ring analysis, used to establish past CO<sub>2</sub> and temperatures, exhibit varying degrees of uncertainty or inaccuracy. For example, it is detailed in [49] that unaccommodated age-dependent variability in tree ring widths, for broad-leaved species such as oaks, can flatten longer-term climate fluctuations and potentially produce misleading data. Termed the *segment length curse*, this can cause serious underestimation of climate variability when the age of ancient timbers cannot be established through independent means. The initial debate regarding the accuracy of the temperature “hockey stick” plot centered on the statistical methods used to construct proxy past temperatures based on statistically blended tree ring measurements.
- In many cases, error bars or uncertainties are omitted, thus misrepresenting the validity of past values or future predictions. For example, the standard error limits shown in Figure 2.9 are not plotted in many of the reports that use this figure to argue that present temperatures are significantly higher than during the past millennium. When viewed with uncertainty measures on past proxy values, this conclusion is less dramatic. Moreover, when viewed with error measurements, it is difficult to correlate Figure 2.9 with reported past climate variations such as the *Medieval Warm Period* from roughly 950–1250, when Greenland was colonized by the Vikings, and the *Little Ice Age* from 1400–1700. However, the global nature of these events has not been established and is debated.
- Most greenhouse gas analysis focuses on the role of CO<sub>2</sub> and methane, since they are fossil fuel combustion byproducts, with some analysis of ozone and CFCs. However, water vapor is by far the most abundant greenhouse gas. Significantly less effort has focused on accurate measurement of water vapor over the last millennium. Moreover, it was noted in Section 3.1 that because moisture phase transitions exhibit complex aerosol and temperature dependencies and occur on subgrid scales, phenomenological models having nonphysical parameters will introduce substantial uncertainties in models and parameters.
- The specification of which data is utilized and which is neglected has introduced uncertainty into both present interpretations and future predictions. The conclusion drawn from Figures 2.8 and 2.9 is that present CO<sub>2</sub> levels and temperatures are higher now than in the past millennium. However, ice core measurements have demonstrated significantly higher CO<sub>2</sub> and temperature changes over the last 800,000 years—e.g., temperature variations of up to 15 °C [117, 127]. In this context, variations of 5 °C in the next century fit well within past levels. It must be noted, however, that whereas the earth

has exhibited significantly more extreme temperatures in the past, it was also not habitable by humans under those conditions.

- Climate model predictions must incorporate future CO<sub>2</sub> levels to predict greenhouse effects. This requires models that predict the growth of nations and their economies and technologies since these factors influence fossil fuel usage. This is very difficult and prone to significant uncertainty. For example, it is unlikely that models from 1990 would have accurately predicted current CO<sub>2</sub> emissions in China. As detailed in Section 2.2.3, this had led to predictions based on various population, economic, and technological scenarios.

Further details regarding the role of uncertainty quantification in climate models will be provided in Section 2.2.3.

#### *Aerosol Emission*

It was noted in Section 2.1 that aerosols critically affect cloud formation, which in turn affects the global energy balance. It was further noted that aerosol models are complicated by the fact that they contain complex forcing terms that must be quantified using parameterized phenomenological models; see, e.g., (2.9). In climate models, predicted aerosol levels are further complicated by the fact that, like greenhouse gas levels, they rely on socioeconomic growth models. Associated uncertainties are partially addressed by considering scenarios of the type discussed in Section 2.2.3.

#### *Deforestation and Desertification*

The absorption of CO<sub>2</sub> through photosynthesis, and absorption and emission of solar energy at the earth's surface, is strongly influenced by the nature of surface vegetation. At present, approximately 29% of the earth's land surface is forested and 11% is arable. However, this is changing as forests are cut—especially in the tropics such as the Amazon Basin in Brazil—croplands are urbanized, and sparse vegetation is grazed in semiarid areas. As with greenhouse gases and aerosols, models quantifying future deforestation and desertification are highly uncertain since they depend on demographic and socioeconomic factors.

### **Climate Feedback Mechanisms**

Forced climate changes are complemented by feedback mechanisms driven by changes in climate conditions such as temperature, precipitation levels, or aerosol concentrations.

#### *Ice Albedo Effects*

The percentage of solar energy reflected by a substance or object, termed the *albedo*, is an important factor in the global energy balance. On the earth's surface, ice and open ocean water have very different albedos, and the former reflects significantly more solar energy than the latter. The dependence of ice levels on temperatures is a *positive feedback* mechanism since increasing temperatures reduce the area covered by snow and ice, which subsequently produces further increase in temperatures. Whereas ice and snow albedo effects constitute an important

component in the global energy balance, they are also a source of uncertainty in climate models due to the complexity of associated physics and its highly nonlinear coupling in global models.

### *Greenhouse Water Vapor Levels*

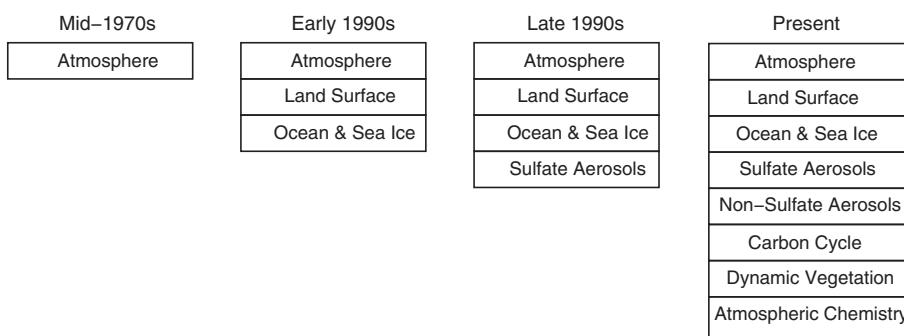
We noted previously that water vapor constitutes the most abundant greenhouse gas. Because increasing temperatures enhance evaporation that increases greenhouse gas levels, the interdependency between greenhouse gases and temperatures constitutes a positive climate feedback mechanism in addition to being a forced response.

## 2.2.2 Climate Models

The general framework used to construct climate models is similar to that described in Section 2.1.1 for weather models. Primitive equations derived from the equations of atmospheric physics (2.7) are numerically solved on a 3-D grid of the form illustrated in Figure 2.1. As for weather models, the numerical grid introduces significant uncertainty for processes such as aerosol-induced cloud formation and greenhouse gas levels that must be quantified using parameterized phenomenological models on subgrid scales.

It was noted in Section 2.2 that climate models differ fundamentally from weather models in the sense that they are forced boundary value problems rather than initial value problems. This introduces the problem of resolving complex climate forces and feedbacks over very long time scales (e.g., millennia) but significantly reduces the chaotic behavior associated with uncertain initial conditions.

Figure 2.10 illustrates the evolution of climate models over the last 30 years. Whereas ocean currents can be represented by primitive equations constructed using conservation of mass, momentum, and energy, the majority of processes such as cloud microphysics, radiation, surface energy fluxes, turbulence, aerosol levels, chemistry, sea ice formation, and dynamic vegetation levels are quantified using parameterized phenomenological models. As with weather models, values and un-



**Figure 2.10.** Development of climate models modified from the 2001 IPCC report [115].

certainties for these parameters must be determined using model calibration techniques.

Climate models are validated by testing their ability to simulate past climatic events (paleoclimates) such as the Cretaceous and Last Glacial Maximum (LGM) based on proxy, measured, or estimated climate forces. Their validity is also tested by simulating climate responses to current forces such as the eruption of Mount Pinatubo, as illustrated in Figure 2.7.

Examples of present climate models include the NSF, DOE, and NASA sponsored Community Earth System Model (CESM) and the Hadley Centre model HadCM3 developed in the United Kingdom. The package CESM1.0 is publicly available and includes atmosphere, land, sea ice, ocean, and land ice modules. The package HadCM3 was highly cited in the 2001 IPCC report [115] and includes atmospheric and ocean components, including sea ice. As noted in [178], the atmospheric package HadAM3 has on the order of 100 parameters of which approximately 29 are considered to control key atmospheric and surface processes.

### 2.2.3 Role of Uncertainty Quantification in Climate Modeling

It was detailed in Section 2.2.1 that the quantification of uncertainties in temperature and greenhouse gas data and models is critical for ascertaining the present levels of anthropogenic greenhouse effects and predicting future ramifications of potential warming. We summarize here other ways in which uncertainty quantification will be critical for making viable climate predictions.

For weather models, uncertainties associated with inputs were comprised primarily of those associated with initial conditions and nonphysical model parameters that were determined using data assimilation techniques such as 4D-VAR. For climate models, initial conditions are replaced by forced boundary conditions that introduce significant uncertainty for phenomena such as predicted greenhouse gas and aerosol emissions and rates of deforestation. Models for these phenomena are highly uncertain since they are based on socioeconomic and demographic factors that are also highly uncertain. Furthermore, since these forcing mechanisms and associated feedback loops occur in the future, data assimilation techniques are not viable for constructing associated error bounds or pdf. The uncertainties associated with parameters and models is augmented by the fact that numerous forcing and feedback mechanisms occur at subgrid scales and require parameterized phenomenological models to quantify complex or poorly understood physics.

Due to the highly uncertain nature of future aerosol and greenhouse gas emission levels, the IPCC reports predicted emission levels and climate changes for four *scenarios* representing a range of demographic, economic, and technological growth.

- A1: Rapid economic growth with increasingly efficient technology and a mid-century population peak—A1F1 represents fossil fuel intensive energy usage, whereas A1T and A1B respectively represent nonfossil and balanced energy resources.
- A2: Slow economic and technological growth and large population growth.

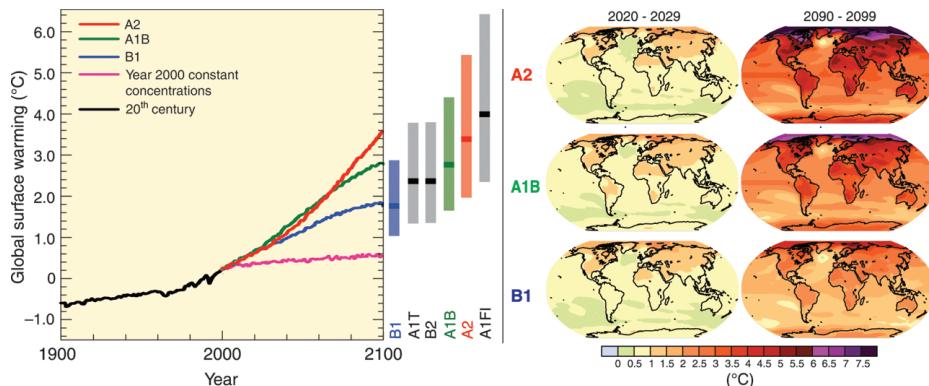
- B1: Same global population dynamics as A1 but more rapid change toward a service and information economy.
- B2: Intermediate population and economic growth with local solutions to environmental, social, and economic sustainability.

The report does not assign likelihoods to these scenarios.

The IPCC report [188] lists key uncertainties as well as robust findings. Representative uncertainties from that report include the following.

- “The magnitude of CO<sub>2</sub> emissions from land-use change and CH<sub>4</sub> emissions from individual sources remain as key uncertainties.”
- “Aerosol impacts on the magnitude of the temperature response, on clouds and on precipitation remain uncertain.”
- “Models differ considerably in their estimates of the strength of different feedbacks in the climate system, particularly cloud feedbacks, oceanic heat uptake and carbon cycle feedbacks, although progress has been made in these areas.”
- “Large-scale ocean circulation changes beyond the 21<sup>st</sup> century cannot be reliably assessed because of uncertainties in the meltwater supply from the Greenland ice sheet and model response to the warming.”
- “Projections of climate change and its impacts beyond 2050 are strongly scenario- and model-dependent, and improved projections would require improved understanding of sources of uncertainty and enhancements in systematic observation networks.”
- “Understanding of low-probability/high-impact events and the cumulative impacts of sequences of smaller events, which is required for risk-based approaches to decision-making, is generally limited.”

Since the objective of climate models is to predict trends for the future, it is natural to consider statistical QoI such as average temperatures, precipitation levels, or amounts of sea level rise. The manner in which QoI and associated uncertainties are reported depends on the highly varied nature of input uncertainties. When input uncertainties can be reasonable quantified, Monte Carlo simulations from input densities or confidence intervals are used to construct uncertainty bounds for QoI. For highly uncertain inputs such as future aerosol or greenhouse gas emissions, predictions based on the scenarios A1, A2, B1, and B2 are provided. For example, Figure 2.11, from the 2007 IPCC report [188], illustrates the predicted average change in global surface temperatures for these scenarios. The scenario A1F1, representing intensive reliance on fossil fuel energy sources, predicts a best estimate increase of 4 °C by 2100 with a likely assessed uncertainty range of 2.4–6.4 °C. To place this in perspective, it is predicted that a 4 °C temperature rise would cause approximately 35% reduction in African crop production, up to 50% less water available in the Mediterranean and Southern Africa, coastal flooding that could displace up to 300 million people annually, and loss of up to half the arctic tundra [49]. It is noted



**Figure 2.11.** *Levels of average global warming predicted using the scenarios A1, A2, B1, and B2; from the 2007 IPCC report [188, Figure 3.2].*

in the report that due to uncertainties in the ocean-atmosphere couplings, neither best estimate predictions nor uncertainty bounds could be reasonably established for sea level rise in 2100. Instead, ranges are reported.

Due to the impact that climate change can have on civilization, it is critical that physical, biological, and socioeconomic modeling, data collection and analysis, statistical and mathematical analysis, large-scale computing, and uncertainty quantification continue to be investigated in concert to advance the state of climate models so they can inform scientists in a manner that includes quantified uncertainties. This information must then be conveyed to policy makers and the general public in a manner that encourages economic and technological growth that minimizes the degree to which anthropogenic forces accelerate climate change.

## 2.2.4 Notes and References

As with weather modeling, we have focused only on aspects of climate modeling to highlight issues and motivate the central role of uncertainty quantification for understanding present climate trends and predicting future climate conditions. The basic atmospheric phenomena are the same as those for weather, and the underlying physical principles are detailed in [123, 147, 193, 213, 252]. The reader is referred to [172] for a perspective of research issues at the intersection between weather and climate and [163] for an overview of how the stochastic properties of turbulent dynamical systems pertain to climate models. The texts [28, 49, 169] provide very readable descriptions of the global energy balance, the natural and human-induced causes of climate change, issues associated with proxy measurements obtained using ice-core and tree ring data, required parameterized components of climate models, consequences of climate change, and debates pertaining to the subject. Details regarding various forcing mechanisms can be found in [136]. The 1995, 2001, and 2007 Intergovernmental Panel on Climate Change (IPCC) Assessment Reports on Climate Change [115, 188, 228] provide a comprehensive discussion of the scientific basis for and impact of climate change. These reports list key uncertainties and

robust findings which are defined as those that hold under a variety of assumptions, models, or approaches. Although listed as robust findings, conclusions such as “Anthropogenic warming over the last three decades has *likely* had a discernible influence at the global scale on observed changes in many physical and biological systems” still convey inherent uncertainty.

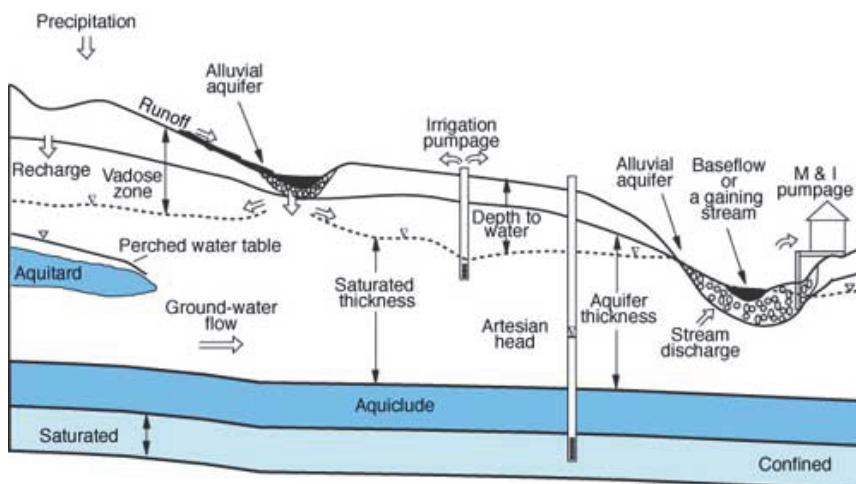
It was noted that whereas phenomena such as global warming are accepted by a majority of the scientific community, there is still serious scientific debate regarding the nature of models, data, and conclusions. Much of this has appeared in the popular literature and is difficult to judge since it does not include strict scientific rigor. The text [149] presents a fairly balanced presentation of issues that may detract from the validity of mainstream theory for climate change.

## 2.3 Subsurface Hydrology and Geology

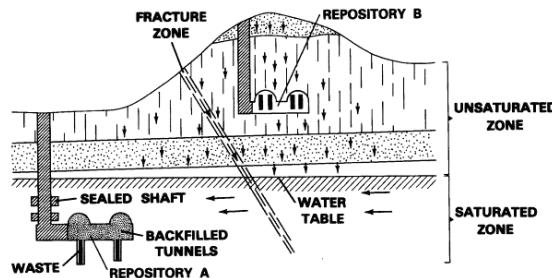
Uncertainty pervades subsurface hydrology and geology for the simple reason that the earth’s subsurface is largely inaccessible to observation or measurement except at a limited number of drilling sites or using mapping techniques such as seismic tomography. This places substantial demand on models and illustrates the criticality of obtaining accurate predictions with quantified and reduced uncertainties.

Subsurface hydrology addresses issues such as the availability and contamination of subsurface groundwater and has the goal of answering questions that include the following.

- Is it safe to sequester carbon dioxide in a depleted oil reservoir?
- How much groundwater is available in an aquifer?—Figure 2.12



**Figure 2.12.** Subsurface strata and the manner in which they affect aquifer properties; from [44].

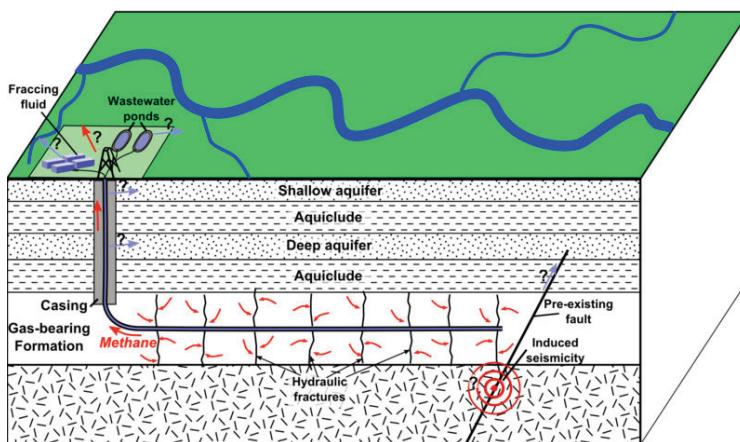


**Figure 2.13.** Representation of repositories in (A) saturated zones and (B) unsaturated zones from [209] and discussed for Yucca Mountain in [83].

- Are water tables sufficiently isolated in Yucca Mountain to permit safe storage of nuclear waste?—Figure 2.13
- Will hydraulic fracturing (fracking) contaminate water tables in a drilling region?—Figure 2.14
- Will microbial action naturally degrade contaminant levels to specified levels in a given time frame?
- What is the average time for transport of stored radioactive materials from a repository to a human environment?

The issues addressed in petroleum geology are similar.

- Does a given region offer substantial oil or gas reserves?
- Will drilling for oil have significant environmental risk?



**Figure 2.14.** Depiction of the potential impact of fracking on aquifers. Image courtesy of Mike Norton.

### 2.3.1 Models and Role of Uncertainty

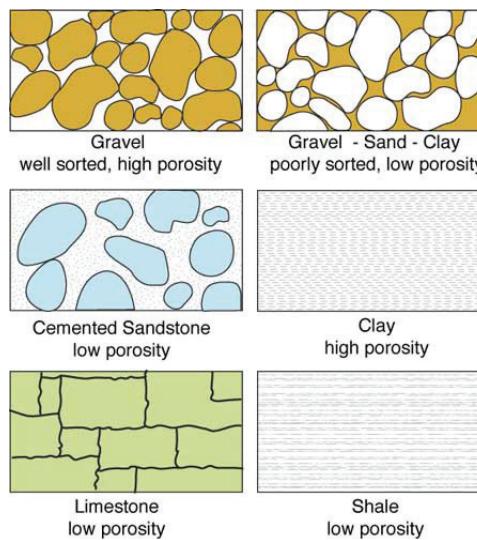
To illustrate issues associated with model and parameter uncertainty, we summarize a typical model used to characterize subsurface groundwater flow with  $N$  reactive species having concentrations  $c_n(t, x)$  and mass flux  $J_n(t, x)$ . As detailed in [245], the flow equations on a domain  $\Omega$  bounded by a surface  $\Gamma = \Gamma_D \cup \Gamma_N$ , where  $\Gamma_D$  and  $\Gamma_N$  denote Dirichlet and Neumann segments of the boundary, are

$$\begin{aligned} S_s \frac{\partial h}{\partial t} &= -\nabla \cdot q - f, \quad q = -K \nabla h, \\ \frac{\partial c_n}{\partial t} &= -\nabla \cdot J_n - R_n(c_1, \dots, c_N), \quad J_n = -D \nabla c_n + \frac{q}{\phi} c_n, \quad n = 1, \dots, N, \end{aligned} \quad (2.12)$$

with initial and boundary conditions

$$\begin{aligned} h(0, x) &= h_0; \quad h = H, x \in \Gamma_D; \quad n \cdot q = Q, \quad x \in \Gamma_N, \\ c_n(0, x) &= c_{n_0}; \quad c_n = C_n, \quad x \in \Gamma_D; \quad n \cdot J = Q_n, \quad x \in \Gamma_N. \end{aligned} \quad (2.13)$$

Here  $h(t, x)$  and  $q(t, x)$  denote the hydraulic head (equilibrium water elevation) and Darcy velocity due to uncertain sources or sinks  $f(t, x)$ . The specific storage  $S_s(x)$ , hydraulic conductivity tensor  $K(x)$ , porosity  $\phi(x)$ , and dispersion tensor  $D(x)$  are uncertain properties of the heterogeneous subsurface environment  $\Omega$ . For example, gravel, sand, and clay can yield different porosities, as illustrated in Figure 2.15. The chemical reactions  $R_n$  are assumed to have uncertain reaction rates  $\kappa = [\kappa_1, \dots, \kappa_N]$  and uncertain forcing terms. Finally,  $h_0(x)$  and  $c_{n_0}(x)$  denote the initial hydraulic head distribution and species concentration and  $H(x)$ ,  $C_n(x)$ ,  $Q(x)$ , and  $Q_n(x)$  are the hydraulic head, species concentration, and fluxes specified at the boundary.



**Figure 2.15.** Porosity of various subsurface strata; from [44].

To accommodate these uncertainties, parameters and forces are considered to be random fields or random processes if they are time varying. To relate parameter uncertainties to geologic uncertainties, one can decompose the domain  $\Omega$  into  $M$  nonoverlapping subdomains  $\Omega_i$ , so that  $\Omega = \cup_{i=1}^M \Omega_i$ , and represent the random fields on each subdomain; e.g.,

$$K(x) = \begin{cases} K_1(x) & , \quad x \in \Omega_1 \\ \vdots & \vdots \\ K_M(x) & , \quad x \in \Omega_M. \end{cases}$$

Uncertainty in the nature of the decomposition and number of subdomains can then be propagated through (2.12).

To determine uncertainty bounds or pdf for  $h$  or  $c_n$ , one must first specify uncertainties or pdf for the parameters and driving forces and then propagate these uncertainties through (2.12). In theory, one could parameterize the random fields in space and estimate pdf using model calibration techniques. In practice, however, one often invokes assumptions such as stationarity and assumes parametric pdf relations such as normal or lognormal [245]. In the final step of the analysis, best estimate values and uncertainties for the states  $h$  and  $c_n$  would be used to compute best estimate values and uncertainties for QoI necessary to answer the questions posed at the beginning of this section.

### 2.3.2 References

Details regarding the physical phenomena and models for subsurface hydrology and geology can be found in [194]. The reader is referred to [110] for information regarding model calibration, sensitivity analysis, and uncertainty quantification for groundwater and [144, 245, 260] and the included references for details regarding Markov chain Monte Carlo analysis and probabilistic risk assessment in the context of subsurface hydrology. Readers can find additional information regarding the PRA for nuclear waste disposal at Yucca Mountain in Chapter 3 of [250]. In addition to geological and geotechnical hazards, this includes aircraft crash hazards, industrial and military-related activity hazards, and potential hazards due to weather. Details regarding hydraulic fracturing, its associated models, and its potential environmental impact can be found in [187, 251, 269]. Finally, the reader is referred to [109] and the included references for an overview of uncertainty quantification issues concerning environmental and groundwater models.

## 2.4 Nuclear Reactor Design

Nuclear power constitutes a major source of sustainable energy, as evidenced by the fact that in the year 2000, there were 434 nuclear power stations producing 350,442 MWe worldwide [233]. This included 252 pressurized water reactors (PWR), including the Russian VVER reactors, 92 boiling water reactors (BWR), and 34 gas-cooled reactors (GCR). Additionally, there were 34 heavy water water-cooled reactors, 15 graphite moderated reactors (RBMK – Reaktor Bolshoy Moshchnosti Kanalniy), and 2 liquid-metal fast breeder reactors (LMFBR). Nuclear power plants

presently provide approximately 19% of the electricity in the United States and 14–15% of the world’s electricity.

Although the efficiency of nuclear power plants has improved over the last 30 years, there have not been significant changes in the fundamental designs. Furthermore, there has been a basic halt in construction of new plants in the United States for the last 30 years despite an increase in demand from 1980 when nuclear power plants provided approximately 11% of the nation’s electricity.

A number of resources are presently focused on the development of model and simulation-based design tools to improve reactor efficiency, reduce operating costs, reduce nuclear waste, and continue to enhance safety. A primary goal of the Department of Energy (DOE) funded Consortium for Advanced Simulation of Light Water Reactors (CASL) is to develop a “virtual reactor” environment that can predict the behavior of and interactions between the nuclear fuel, neutron transport, heat transfer, and thermal-hydraulic (coolant flow) components of a reactor using limited experimental measurements. This predictive environment would be used to improve present systems and enhance the design of next generation reactors. To motivate the central role of uncertainty quantification in this predictive design environment, we summarize the basic components of light water reactors and aspects of the models used to quantify the underlying physics.

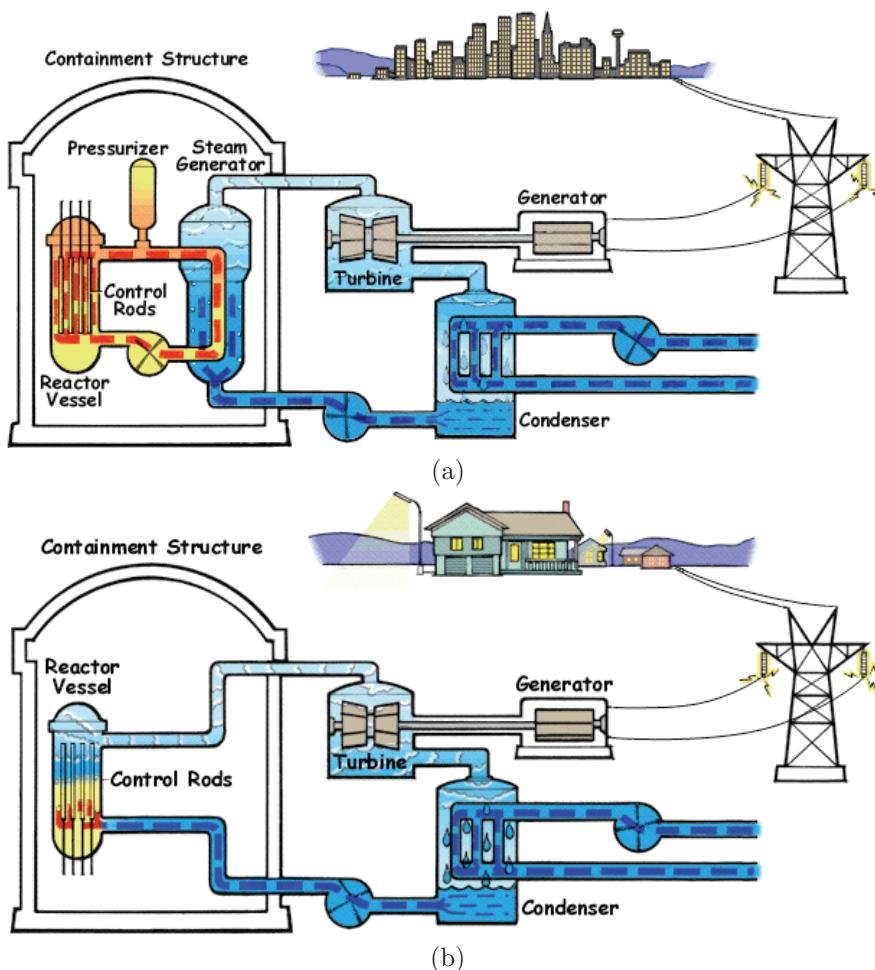
### 2.4.1 Light Water Reactors

Light water reactors employ ordinary water, termed light water, as a coolant and neutron moderator. This is in contrast to heavy water reactors employed in Canada (CANDU) and India (AHWR). The two primary designs, PWR and BWR, are illustrated in Figure 2.16. In the United States, there are presently 104 nuclear power plants of which 69 are PWR and 35 are BWR.

In all nuclear reactors, heat is produced by controlled nuclear fission in the reactor core. As depicted in Figure 2.17, the core contains nuclear fuel rods, control rods, and water-filled channels. The fuel rods, which are approximately 12 feet long, contain uranium or uranium oxide pellets. The control rods contain elements such as cadmium or hafnium that absorb neutrons, thus slowing chain reactions.

In a BWR, heat transferred through the fuel cladding turns the light water to steam, which directly drives turbines before being condensed using secondary lake, river, or ocean water. In a PWR, coolant is circulated through the core under high pressure so it remains liquid despite being raised to temperatures on the order of 315 °C. This hot primary coolant then flows through a heat exchanger where it generates steam in an isolated secondary coolant that in turn drives turbines. The separation of the two coolant sources prevents accidental radioactive contamination of the secondary coolant source, as was the case in the Fukushima plant, which is a BWR. We note that a 1100 MWe PWR core can contain over 50,000 fuel rods and 18 million fuel pellets in approximately 193 fuel assemblies.

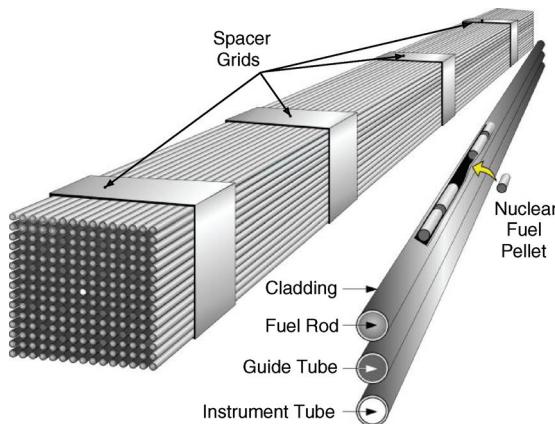
To sustain a chain reaction, fast fission neutrons must be slowed before they can interact with the uranium. This is termed moderation or thermalization and, in light water reactors, this function is directly provided by the coolant water. In the process of colliding with hydrogen atoms in the water, neutrons lose speed until



**Figure 2.16.** Schematics of (a) PWR and (b) BWR designs. Images courtesy of United States Nuclear Regulatory Commission.

their velocity is comparable with the thermal velocity of the nuclei, at which point they are called thermal neutrons.

An important stability feature of light water reactors is the fact that as temperatures increase, the water density decreases. This in turn decreases the slowing of neutrons, which slows the chain reaction. This negative feedback loop means that if there is a spike in the nuclear reaction, the coolant will moderate it. If coolant is lost due to an accident, loss of the moderator stops active fission. However, there will still be an approximately 5% *decay heat* for 1–3 years, which is hot enough to melt the core if sufficient cooling is not provided. While not as dangerous as active fission, the potential for decay heat is a risk factor associated with LWR designs.



**Figure 2.17.** Schematic of a nuclear fuel assembly (U.S. Department of Energy).

This is in contrast to the RBMK design used at Chernobyl, which employed graphite as a moderator. This design is less robust and is subject to rapid transients, which is considered one of the causes of the Chernobyl disaster.

There are two primary mechanisms for controlling the power generated by a PWR: insertion or retraction of the neutron absorbing control rods and varying the concentration of boric acid dissolved in the reactor coolant. The latter acts in a manner similar to the control rods since boron readily absorbs neutrons. Rather than using boric acid in the reactor coolant, BWR adjust the coolant flow rate to control reactor power.

Further advantages and disadvantages of light water reactor designs, including their use in nuclear ships and submarines, are detailed in [76, 233].

## 2.4.2 Reactor Models

Models for components of the reactor core reflect the complexity of the underlying physics. We summarize only aspects of the neutronics transport equations and thermal-hydraulic relations to illustrate issues that must be addressed for model verification, validation, and uncertainty quantification.

### Neutron Transport Equation

The quantification of neutron distributions in the reactor core is central to reactor models since neutron densities and energy levels govern the various nuclear reactions that occur in the reactor. The neutron interaction with the primary coolant is also critical since the coolant serves as a neutron moderator and the neutrons heat the coolant.

To specify the neutron distribution, one typically considers the angular neutron flux  $\varphi(r, E, \Omega, t)$ , where  $r = (x, y, z)$  is a position vector,  $v$  is the velocity,

$\Omega = v/|v|$  is a solid angle that specifies the direction of motion, and  $E$  is the energy level at time  $t$ . We note that this is a scalar flux as compared with vector-valued electromagnetic fluxes.

To construct the neutron transport equation, one balances neutron sources and loss mechanisms for an arbitrary volume  $V$ . Sources include fission, neutrons entering  $V$ , and neutrons with different  $E', \Omega'$  values that incur a scattering collision that changes their energy and direction to  $E, \Omega$ , which we denote by  $E' \rightarrow E, \Omega' \rightarrow \Omega$ . Losses are due to neutrons leaving  $V$  and neutrons suffering a collision.

As detailed in [76], a balance of the source and loss terms yields the 3-D neutron transport equations

$$\begin{aligned} \frac{1}{|v|} \frac{\partial \varphi}{\partial t} + \Omega \cdot \nabla \varphi + \Sigma_t(r, E) \varphi(r, E, \Omega, t) \\ = \int_{4\pi} d\Omega' \int_0^\infty dE' \Sigma_s(E' \rightarrow E, \Omega' \rightarrow \Omega) \varphi(r, E', \Omega', t) \\ + \frac{\chi(E)}{4\pi} \int_{4\pi} d\Omega' \int_0^\infty dE' \nu(E') \Sigma_f(E') \varphi(r, E', \Omega', t), \end{aligned} \quad (2.14)$$

where  $\Sigma_t, \Sigma_f$  are macroscopic total and fission cross-sections—defined as the ratio of the reaction rate ( $s^{-1}$ ) to incident flux ( $cm^{-2}s^{-1}$ ) when a beam of particles impinges on a nucleus. The double differential scattering cross-section  $\Sigma_s$  characterizes scatter from  $(E', \Omega')$  to  $(E, \Omega)$  in the cone  $d\Omega$ . Finally,  $\chi(E)$  and  $\nu(E')$  respectively denote the fission spectrum and average number of fission neutrons produced by fission resulting from neutrons with energy  $E'$ .

We note that the integro-differential equation (2.14) is linear in the state  $\varphi$  but is a function of seven independent variables ( $r = x, y, z, E, \Omega = \theta, \phi, t$ ) in three dimensions. Posed in terms of a general source term  $s$ , the 1-D transport equation is

$$\begin{aligned} \frac{1}{v} \frac{\partial \varphi}{\partial t} + \mu \frac{\partial \varphi}{\partial x} + \Sigma_t \varphi(x, E, \mu, t) \\ = \int_{-1}^1 d\mu' \int_0^\infty dE' \Sigma_s(E' \rightarrow E, \mu' \rightarrow \mu) \varphi(x, E', \mu', t) + s(x, E, \mu, t), \end{aligned} \quad (2.15)$$

where  $\mu \equiv \cos \theta$ . Further details regarding the derivation and numerical implementation of the neutron transport equations can be found in [76, 233].

The transport equation provides a highly accurate description of neutron distributions in a reactor if one has correct macroscopic cross-section information. Hence from the perspective of uncertainty quantification, one must quantify uncertainties associated with the cross-sections and propagate them through the model to construct uncertainties for the QoI. Moreover, the dependence of cross-sections on  $r$  and  $E$  is complicated due to the complex reactor geometry and energy profiles, which includes resonant behavior and threshold effects.

Due to their central role in nuclear reactor design, however, numerous experiments have been performed to establish libraries of cross-section distributions for various stable and radioactive nuclei that can be employed for uncertainty quantification. It is noted in Chapter 17 of [51] that the establishment and use of these libraries for reactor analysis was an early example of model calibration based on

data assimilation, sensitivity analysis, and uncertainty quantification based on a systematic mathematical and physical framework.

Because the quantification of neutron distributions is so critical to nuclear reactor design, numerous commercial, government laboratory, and proprietary neutron transport codes have been developed. Deterministic codes include the massively parallel transport code DENOVO developed at Oak Ridge National Laboratory (ORNL) and the commercial code Atilla. Probabilistic software includes the ORNL code KENO and the Los Alamos National Laboratory (LANL) code MCNP.

### Thermal-Hydraulic Models

The characterization of primary coolant behavior in a nuclear reactor is highly complex since it includes the integration of high pressure, two-phase flow dynamics, heat conduction in the fuel, heat transfer, and neutron interactions in complex geometries. However, it is also critical for reactor design since accurate quantification of void fraction distributions and boiling transitions is essential for optimized performance and maintenance of safety margins. We illustrate only aspects of thermal-hydraulic models to illustrate some of the associated difficulties, and we refer the reader to [51, 121, 206] for detailed derivation and numerical analysis of these relations.

Due to the presence of both liquid and vapor phases in the reactor core channels, two-phase flow models are used to simulate transient and steady state coolant behavior. The two-phase mixture is modeled using conservation of mass, momentum, and energy in combination with conservation relations for compounds such as boron. Additionally, one employs closure relations commensurate with the operating conditions—e.g., one uses different phenomenological relations for laminar versus turbulent conditions or boiling versus high pressure nonboiling conditions.

To illustrate, we let  $\alpha_g, \alpha_f$  respectively denote the volume fraction of gas and fluid and let  $\rho_g, \rho_f$  and  $v_g, v_f$  denote the densities and velocities of the gas and fluid phases. The internal energies for the two phases are denoted by  $e_g, e_f$ . As detailed in (116)–(123) of Chapter 16 of [51], conservation of mass, momentum, and energy respectively yield the fluid phase relations

$$\begin{aligned} \frac{\partial}{\partial t}(\alpha_f \rho_f) + \nabla \cdot (\alpha_f \rho_f v_f) &= -\Gamma, \\ \alpha_f \rho_f \frac{\partial v_f}{\partial t} + \alpha_f \rho_f v_f \cdot \nabla v_f + \nabla \cdot \sigma_f^R + \alpha_f \nabla \cdot \sigma + \alpha_f \nabla p_f \\ &= -F^R - F + \Gamma(v_f - v_g)/2 + \alpha_f \rho_f g \end{aligned} \quad (2.16)$$

and

$$\begin{aligned} \frac{\partial}{\partial t}(\alpha_f \rho_f e_f) + \nabla \cdot (\alpha_f \rho_f e_f v_f + Th) &= (T_g - T_f)H + T_f \Delta_f \\ - T_g(H - \alpha_g \nabla \cdot h) + h \cdot \nabla T - \Gamma[e_f + T_f(s^* - s_f)] \\ - p_f \left( \frac{\partial \alpha_f}{\partial t} + \nabla \cdot (\alpha_f v_f) + \frac{\Gamma}{\rho_f} \right) \end{aligned} \quad (2.17)$$

with analogous coupled gas phase relations. Here energy, momentum, and mass exchange at the vapor/liquid interfaces are respectively modeled with the constitutive relations

$$\begin{aligned} H &= \kappa(T_f - T_g), \\ F &= \zeta(v_f - v_g), \\ \Gamma &= \gamma[(s_f - s_g)(T_{sat}(p_f) - T_g) - K_f], \end{aligned} \quad (2.18)$$

where  $T_f, T_g$  and  $s_f, s_g$  are the temperature and entropy density of the two phases,  $T_{sat}(p_f)$  is the saturation temperature at the continuous phase pressure  $p_f$ , and  $\kappa, \zeta$ , and  $\gamma$  are positive transport coefficients that must be estimated during model calibration. The viscous and heat transport coefficients are modeled using the constitutive relations

$$\sigma = -\eta \nabla v, \quad h = -\lambda \nabla T, \quad (2.19)$$

where  $\eta$  and  $\lambda$  must be estimated. Relations for  $K_f, T_f \Delta_f, s^*, \sigma^R$ , and  $F^R$  are provided in Chapter 16 of [51]. Further details regarding the derivation of the thermal-hydraulic equations and required constitutive relations for varying hypotheses regarding the operating conditions and carrier fluid and dispersed gas particles can also be found in [51, 122, 206].

Various packages have been developed to numerically solve the thermal-hydraulic equations including RELAP5 developed at Idaho National Laboratory (INL), CATHARE, FLICA4, and COBRA. As detailed in [206], RELAP5-3D provides multidimensional hydrodynamic and reactor kinetics modeling capabilities. The CATHARE package allows reactor coolant circuits to be modeled as interconnected submodules, whereas FLICA4 combines 3-D, two-phase fluid simulation capabilities with 1-D heat conduction for the fuel. COBRA is a subchannel code for which the subchannel spacing constitutes the finest lateral mesh. All four provide the capability for specifying various phenomenological closure conditions. Additionally, nuclear energy companies such as Westinghouse have developed and employed their own thermal-hydraulic codes, such as VIPRE-W.

### 2.4.3 Role of Uncertainty Quantification for Reactor Design

As with weather and climate models, there are four general sources of uncertainties and errors in nuclear reactor models: input uncertainties, model errors, numerical errors, and uncertainties in measurements. Like the atmospheric science relations (2.7), the thermal-hydraulic relations (2.16) and (2.17) are based on conservation of mass, momentum, energy, and species concentrations in combination with phenomenological closure relations and source terms. The resulting coupled system is highly nonlinear and numerically difficult to resolve and verify for small gridsizes. For example, the lateral mesh in COBRA cannot be resolved beyond the dimensions of the subchannels. It is noted in the RELAP5 code manual that the employed system of equations is ill-posed. This necessitates the use of artificial damping to obtain numerical solutions that are viable for nuclear reactor analysis.

In addition to numerical errors, the thermal-hydraulic relations exhibit varying degrees of uncertainty in the phenomenological models used to characterize com-

plex or poorly understood physics—e.g., (2.18) and (2.19), turbulence relations, and phenomenological models for quantifying subchannel void fractions—and the nonphysical parameters in these models. Unlike the cross-section uncertainties for the neutron transport model, which are fairly well characterized experimentally, the uncertainties for nonphysical thermal-hydraulic parameters must be estimated using model calibration techniques before they can be propagated through models. Finally, the harsh environment inside a nuclear reactor limits the number and type of measurements that can be obtained for model calibration and validation and experimental design.

A substantial challenge associated with uncertainty quantification for nuclear reactor designs is the necessity for propagating uncertainties through several linked simulation codes for all of the coupled subsystems. It has been illustrated that heat transfer, coolant flow, neutron distributions, and fission reaction rates are all tightly coupled to form a highly complex multicomponent, multiphysics system. The resulting models and simulation codes are computationally intensive and require the development of surrogate models to construct uncertainties for QoI. The synthesis of surrogate modeling and modular approaches for uncertainty quantification in large, multiphysics systems constitutes an area of active research.

As with weather, climate, and subsurface hydrology models, we are often interested in quantifying uncertainties associated with statistical QoI. Relevant QoI for nuclear reactor design include the following.

- Specify bounds on void fraction distributions and boiling transitions that guarantee specified performance levels and safety margins.
- Specify conditions that limit CRUD<sup>1</sup> on the outside of fuel cladding to within prescribed levels.
- Determine new cladding materials, fuel materials, and fuel pin geometries that provide an average specified improvement in performance and increased resistance to damage.
- Determine conditions that produce specified levels of radiation damage, mechanical thermal fatigue, and corrosion.

In all cases, measurement, input, model, and numerical uncertainties must be determined to provide predictive estimates for the QoI with quantified and reduced uncertainties.

**Remark 2.1.** An issue that complicates model calibration for many comprehensive simulation packages, including those considered for nuclear reactor simulations, is the fact that inputs and parameters are often hard-coded and inaccessible to users. Moreover, the values of hard-coded inputs are specified on a basis that is often not reported. For example, only one parameter is generally accessible in COBRA. This

---

<sup>1</sup>CRUD is a colloquial term that refers to corrosion and wear products that become radioactive when exposed to radiation. The acronym originated from references to *Chalk River Unidentified Deposit*.

is done to facilitate use by nonexperts and to minimize inadvertent changes to inputs. However, this necessitates significant code modification for model calibration and uncertainty quantification, which complicates the process.

#### 2.4.4 Probabilistic Risk Assessment (PRA)

The field of PRA grew out of the Three Mile Island nuclear power plant accident in 1979 and the Challenger space shuttle disaster in 1986. As detailed in [32, 245], structured PRA involves defining a system failure for complex multicomponent, multiphysics problems, identifying basic events that can cause a system failure, building a fault tree to relate component events to a system failure, and relating the joint probability of events to the probability of a system failure. The assignment of probabilities to events can be quantitative or qualitative in the sense that expert opinion is used to assign probabilities to various scenarios, as is the case with economic and technological growth in climate models. As detailed in [245], the related field of risk management is an example of optimization under uncertainty where the goal is to mitigate risk in the presence of uncertainty. This involves optimization techniques such as stochastic and fuzzy programming.

#### 2.4.5 References

The 1976 text [76] is fairly old, but it still provides a comprehensive and very readable exposition of issues pertaining to nuclear reactor design and associated models. This can be complemented by the text [233]. The text [151] is a good reference regarding computational techniques for the neutron transport equation, and the reader is referred to [122] for additional information pertaining to the thermal-hydraulic equations. The 2012, five-volume *Handbook of Nuclear Engineering* [51] is a great resource detailing the present state of the art for numerous aspects of the field. Specifically, readers are respectively referred to Chapters 1, 5, 7, 16, 17, and 28 for information regarding neutron cross-section measurements, the general principles of neutron transport, mathematics for nuclear engineering, multiphase flows, sensitivity and uncertainty analysis, and the scientific basis for nuclear waste management. We note that an eBook version is available if hard copies cannot be obtained.

### 2.5 Biological Models

The role of predictive estimation for biological applications ranging from molecular dynamics to ecosystems has burgeoned in the last 25 years and will become increasingly important as measurement techniques, models, numerical algorithms, computational architectures, and uncertainty quantification techniques continue to evolve. To focus the discussion of issues pertinent to predictive estimation, we consider the following biological scales.

- **Molecular.** Molecular biology focuses on the chemical compositions and interactions required for living organisms. This ranges from fundamental inves-

tigations of proteins and amino acids to understanding and cataloguing the hereditary and information-carrying capabilities of DNA and RNA. Significant recent advances have resulted from improved measurement capabilities, the development of statistical and mathematical modeling techniques, such as dynamic models for proteins and hidden Markov models for biological sequencing, and improved numerical algorithms and computational resources.

- **Cellular.** Cellular biology broadly focuses on self-replicating units such as viruses, bacteria, and animal and plant cells, along with their interactions and communication. The role of mathematical models has been established at this level for at least 30 years, due in part to the availability of relevant data. Models have successfully elucidated or helped to explain phenomena such as cellular kinetics, cell cycle dynamics, information processing and signal transduction in cells, tissue patterning, and cellular robustness and mutations.
- **Organisms.** This biological level focuses on the properties of multicellular tissues, organs, and organ systems that comprise organisms such as animal and plant species. This area has grown significantly due to improved non-invasive diagnostic techniques and the complementary development of comprehensive component and system models. Representative examples include coupled models for the heart and circulatory system and models supporting the development of synthetic organs such as artificial hearts and kidneys. This level also includes the interaction of organisms with the environment and their breakdown when robustness fails. This includes autoimmune afflictions such as rheumatoid arthritis and viral diseases such as the common cold, influenza strains, and the human immunodeficiency virus (HIV).
- **Populations.** Population biology focuses on the interactions, genetic variations, and disease spread among individuals of a single species. Deterministic and stochastic compartmental models have long played a significant role in quantifying processes such as disease dynamic and spread with more recent models focusing on the role of distributed attributes such as age or size.
- **Communities and Ecosystems.** Populations almost never live in isolation, and the final level addresses the interactions between various species and their environment. This ranges from classical models, such as the Lotka and Volterra predator-prey models developed in 1925–1926 [159, 258], to models quantifying the anthropogenic impact on the climate and environment, as discussed in Sections 2.2 and 2.3.

The delineation of biological processes into these scales facilitates our understanding of the processes but is somewhat misleading in the sense that it neglects the coupling that intrinsically occurs between scales. Issues that complicate, but are critical for, predictive estimation include the following.

- Biological phenomena often occur over vast space- and timescales with events at one scale strongly affecting those at other scales, e.g., interactions between viral cell dynamics and a host organism. The development of comprehensive multiscale models requires significant data across these scales.

- Biological systems can include a very large number of components—e.g., millions or more—with complex interactions between objects, e.g., intramolecular DNA and RNA interactions. These interactions are often far from equilibrium or steady state, and high-order interactions are often the rule.
- The feedback mechanisms associated with biological phenomena are often highly complex, involve multiple coupled components, have long timescales, exhibit significant variability between individuals, and are difficult to quantify using precise physical laws, e.g., enzyme responses to metabolism and immune responses to viral infection.
- The organization features of biological systems are typically highly complex and vary significantly among individuals. This introduces substantial uncertainty when bridging between individual and population properties, e.g., health trends such as obesity or disease dynamics.
- There are often more parameters than data which prohibits the use of statistical techniques developed for data-rich applications, e.g., modeling the disease dynamics of rare diseases.
- Data is often highly inaccurate, exhibits substantial noise, or contains significant uncertainties, e.g., determining HIV occurrence in countries where the stigma of the disease deters accurate reporting. Furthermore, data is often in a form that differs significantly from the state variables in models, e.g., photos of phenomena. This complicates model calibration and validation and can necessitate the development of new techniques for both.
- Phenomena are often inherently discrete, which prohibits the use of calculus and requires techniques such as information science or combinatorics, e.g., gene sequence analysis.
- Models for biological phenomena are often highly complex and have numerous inputs or parameters that cannot be measured directly but rather must be inferred through fits to data; e.g., see the HIV model discussed in Section 2.5.1. These parameters are often highly correlated and may not be identifiable.
- Models must balance accuracy with efficiency and utility to be useful for diagnostics and guiding treatment routines, policies, or general control design. Hence model discrepancy must be treated in a computationally tractable manner, e.g., models used to determine future influenza vaccines, HIV treatment schedules, or policies concerning anthropogenic environmental impact.
- The uncertainty in biological processes and variability among individuals often necessitates the use of stochastic forces or models to quantify uncertain processes, e.g., gene, bacteria, or viral evolution.

It is clear that incorporation and accommodation of uncertainties inherent to data, models, inputs, and numerical algorithms is critical to the success of predictive estimation for complex biological and biomedical applications. Furthermore, as illustrated in Figure 1.1, the design of calibration and validation experiments must

be tightly integrated with model development and numerical simulation to optimize the impact of predictive estimation for biological applications.

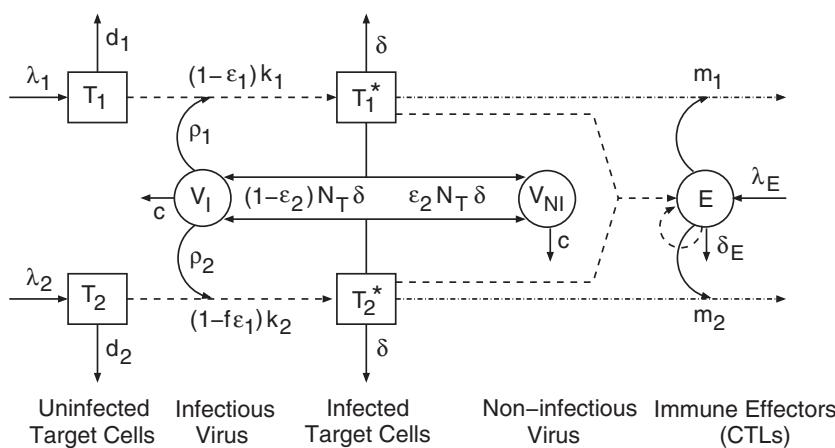
To illustrate the role of uncertainty quantification for predictive estimation in the context of disease dynamics, we summarize next an HIV model used to provide insights into the biological processes associated with this disease and to guide the establishment of treatment regimes.

### 2.5.1 HIV Model

The focus on HIV in the media has dwindled over the last decade due, in part, to the success of antiretroviral treatment regimes for slowing the progression of the disease and delaying the onslaught of acquired immunodeficiency syndrome (AIDS). However, it is estimated that approximately 0.5% of the world's population is presently living with the HIV infection [41] and the percentage of HIV positive individuals in the United States is increasing. Hence the development of data-driven models with reduced and quantified uncertainties is critical for understanding the disease, reducing its spread, and optimizing treatment strategies.

The scope of models ranges from the cellular to the population level with goals ranging from quantifying properties of the virus to predicting the disease spread in various cultures. We illustrate the role of uncertainty quantification in a model used to quantify aspects of HIV progression in an individual. An initial version of this model was proposed in [55], and it was extended significantly in [2, 3] to provide a framework for investigating facets of optimal treatment protocols for HIV.

As illustrated in Figure 2.18, the model is comprised of seven dynamic compartments where each compartment represents a specific cell concentration throughout the body. Here  $T_1$  and  $T_2$  denote uninfected type 1 and type 2 target cells which could respectively represent, for example, CD4 T-lymphocytes and macrophages. Infected target cells are denoted by  $T_1^*$  and  $T_2^*$ . Infectious and noninfectious free



**Figure 2.18.** Compartments in the HIV model modified from [3].

viruses are denoted by  $V_I$  and  $V_{NI}$ , and immune effector cells are denoted by  $E$ . We note that  $T_1, T_1^*, T_2, T_2^*$ , and  $E$  have magnitudes on the order of cells/ $\mu\text{l}$ , whereas  $V_I$  and  $V_{NI}$  have scales of cells/ml. The wide variation in scales exhibited by states is common in physical and biological ODE models and motivates the use of logarithmic scales in the manner detailed in Section 3.2.

Uninfected cells  $T_i$  become infected, with rates  $k_i$ , when they encounter infectious free virus  $V_I$ . The treatment factor  $\varepsilon_1(t)$  models a reverse transcriptase inhibitor (RTI) that blocks new infections. This is potentially more effective for type 1 target cells than for type 2, where the efficacy is represented by  $f\varepsilon_1(t)$  with  $f \in [0, 1]$ . Natural source and death rates for the two types of cells are represented by  $\lambda_i$  and  $d_i$ . Free viruses are produced by infected cells, and they leave the compartment by infecting cells or via natural death at rate  $c$ . Finally, the immune effectors  $E$  are produced in response to existing immune effectors and the presence of infected cells.

As detailed in [3], the resulting ODE model is

$$\begin{aligned} \dot{T}_1 &= \lambda_1 - d_1 T_1 - [1 - \varepsilon_1(t)] k_1 V_I T_1, \\ \dot{T}_2 &= \lambda_2 - d_2 T_2 - [1 - f\varepsilon_1(t)] k_2 V_I T_2, \\ \dot{T}_1^* &= [1 - \varepsilon_1(t)] k_1 V_I T_1 - \delta T_1^* - m_1 E T_1^*, \\ \dot{T}_2^* &= [1 - f\varepsilon_1(t)] k_2 V_I T_2 - \delta T_2^* - m_2 E T_2^*, \\ \dot{V}_I &= [1 - \varepsilon_2(t)] 10^3 N_T \delta (T_1^* + T_2^*) - c V_I \\ &\quad - [(1 - \varepsilon_1(t)) \rho_1 10^3 k_1 T_1 V_I + (1 - f\varepsilon_1(t)) \rho_2 10^3 k_2 T_2 V_I], \\ \dot{V}_{NI} &= \varepsilon_2(t) 10^3 N_T \delta (T_1^* + T_2^*) - c V_{NI}, \\ \dot{E} &= \lambda_E + \frac{b_E (T_1^* + T_2^*)}{(T_1^* + T_2^*) + K_b} E - \frac{d_E (T_1^* + T_2^*)}{(T_1^* + T_2^*) + K_d} E - \delta_E E, \end{aligned} \tag{2.20}$$

where the remaining parameters are defined in Table 2.1. The factors of  $10^3$  convert between microliter and milliliter scales.

For the experiments described in [3], data consisted of measurements of the total viral load  $V = V_I + V_{NI}$  and total CD4 $^{+}$  T-lymphocyte counts  $T_1 + T_1^*$ . In the control problem, inputs consist of the RTI treatment factor  $\varepsilon_1(t)$  and the action  $\varepsilon_2(t)$  of a protease inhibitor which causes infected cells to produce noninfectious viruses  $V_{NI}$ . The control objective is to reduce the viral load  $V$  using reasonable levels of  $\varepsilon_1(t)$  and  $\varepsilon_2(t)$ .

### 2.5.2 Source and Role of Uncertainties

#### Measurement Errors

As detailed in [3], viral loads  $V = V_I + V_{NI}$  were quantified using reverse transcriptase-polymerase chain reaction (RT-PCR) techniques which have a linear range of 400 to 750,000 copies/ml for the standard assay and 50 to 100,000 copies/ml for the ultrasensitive assay. For samples with viral loads that exceeded the upper

|               |                                 |             |                                      |
|---------------|---------------------------------|-------------|--------------------------------------|
| $\lambda_1$   | Target cell 1 production rate   | $\rho_1$    | Ave. virions infecting type 1 cell   |
| $\lambda_2$   | Target cell 2 production rate   | $\rho_2$    | Ave. virions infecting type 2 cell   |
| $d_1$         | Target cell 1 death rate        | $b_E$       | Max. birth rate immune effectors     |
| $d_2$         | Target cell 2 death rate        | $d_E$       | Max. death rate immune effectors     |
| $k_1$         | Population 1 infection rate     | $K_b$       | Birth constant, immune effectors     |
| $k_2$         | Population 2 infection rate     | $K_d$       | Death constant, immune effectors     |
| $c$           | Virus natural death rate        | $\lambda_E$ | Immune effector production rate      |
| $\delta$      | Infected cell death rate        | $\delta_E$  | Natural death rate, immune effectors |
| $\varepsilon$ | Population 1 treatment efficacy | $N_T$       | Virions produced per infected cell   |
| $m_1$         | Population 1 clearance rate     | $f$         | Treatment efficacy reduction         |
| $m_2$         | Population 2 clearance rate     |             |                                      |

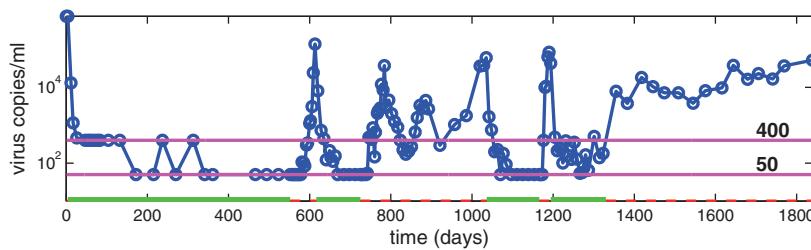
**Table 2.1.** Parameters used in the HIV model (2.20).

limit, the sample was diluted until the virus numbers were within the prescribed range and the measurements were scaled accordingly. Whereas this technique accommodated large viral loads, it is also a source of measurement errors. For viral loads below 50 copies/ml, this value was prescribed as a lower limit, as shown in Figure 2.19. Left-censoring of the data in this manner introduces unavoidable measurement error and must be accommodated when estimating parameters through a least squares fit to data.

### Model and Input Uncertainties

The model (2.20) is a vastly simplified representation of components in an extremely complex, and only partially understood, process. Hence its scope and accuracy are intrinsically limited, and it is unreasonable to expect it to provide direct answers regarding the pathogenesis of the HIV infection. When employed in conjunction with data, however, it can be used to elucidate properties of the viral transmission mechanisms and guide the development of control-based treatment regimes based on the reduction of the viral load.

It is noted in [2, 3] that whereas certain parameters, such as  $\delta_E$ ,  $c$ ,  $N_T$ ,  $\delta$ ,  $d_1$ , and  $d_2$ , can be estimated from human or macaque data, the remainder must be

**Figure 2.19.** Left-censored viral data from [3].

inferred through a fit to data. Furthermore, initial conditions for the states are generally unknown, so they too must be estimated. Due to the measurement errors and model discrepancy terms, these input terms will also have inherent uncertainty which must be quantified using the techniques of Chapters 7 and 8.

### 2.5.3 Robust Control Design

The objective of using the model to guide treatment regimes can be posed as a control problem in which one seeks to regulate the viral load  $V$  to zero using the RTI treatment factor  $\varepsilon_1(t)$  and protease inhibitor level  $\varepsilon_2(t)$  as control inputs. In [2], linear quadratic regulator (LQR) theory was used to construct continuous and structured treatment interaction strategies. Whereas LQR theory employs feedback to provide a degree of robustness with regard to state uncertainty, this control method is not explicitly designed to provide optimal robustness for uncertain systems.

Alternatively, one can consider the average viral load at time  $t$  to be the QoI with uncertainties quantified using the techniques detailed in Chapters 9 and 10. These quantified uncertainties can then be used to construct optimal gains for robust control techniques such as sliding mode or  $H_\infty$  control designs. This use of uncertainty quantification for improved robust control design constitutes an active research area.

### 2.5.4 References

The monograph [181] summarizes the status in 2005 and future directions recommended by the DOE Computational Biology Committee on Mathematical Sciences Research regarding research at the interface between mathematics and biology. Specifically, it delineates ways in which mathematics has impacted molecular, cellular, organismal, population, and community and ecosystem biology and recommends directions projected to have maximal impact. The reader is referred to [14] for details regarding models at all of these levels, disease models, and control strategies for various diseases in 1995. Mathematical modeling in biology, biomedical, and life sciences is treated in significant detail in the texts [114, 179], whereas [40] is devoted to mathematical models for communicable diseases. Details regarding models quantifying aspects of HIV pathogenesis can be found in [2, 3, 10, 55, 191], whereas models for the transmission dynamics of HIV are provided in [40] and the references therein.

## Chapter 3

# Prototypical Models

### 3.1 Models

The equations of atmospheric physics (2.7), groundwater flow (2.12), and thermal-hydraulic flow (2.16) and (2.17) are nonlinear and couple complex multiphysics systems. Similarly, structural and material applications require 2-D and 3-D shell and continuum models, whereas biological and biomedical systems are routinely quantified using coupled, nonlinear ODE and PDE models. The development of uncertainty quantification techniques for applications requiring this level of modeling is critical, and facets of current research are focused on techniques to address the complexities of high-dimensional, coupled, nonlinear ODE and PDE models with large numbers of parameters. This includes the construction of surrogate models, as detailed in Chapter 13.

To illustrate basic techniques, however, we consider a suite of simpler models, many of which have analytic solutions. Fundamental issues required to extend the methodologies to significantly more complex problems are noted where relevant.

**Example 3.1 (Exponential Processes).** We consider first the initial value problem

$$\begin{aligned}\frac{dz}{dt} &= az + b(t), \\ z(0) &= z_0,\end{aligned}\tag{3.1}$$

which models the exponential growth or decay of a process with forcing term  $b(t)$ . We collectively denote parameters by  $q = [a, z_0]$ . The solution

$$z(t, q) = e^{at} \left[ z_0 + \int_0^t e^{-as} b(s) ds \right]\tag{3.2}$$

depends on both the independent variable  $t$  and the parameters  $q$ . It is important to note that although the model (3.1) exhibits a linear dependence on the state or dependent variable  $z$ , the dependence of  $z(t, q)$  on the parameters is highly nonlinear.

Despite the simplicity of (3.1), exponential processes play a significant role in the applications of Chapter 2, including radioactive decay and cross-section relations for nuclear processes and certain chemical reactions for atmospheric species.

As detailed in Chapter 1, the uncertainty quantification problem has two components: determine uncertainties in the parameters  $q$  and forcing term  $b(t)$  and quantify the effect of these uncertainties on the state response  $z$ . The first component is addressed in Chapters 7 and 8, whereas Chapters 9 and 10 address the second.

To incorporate these random effects, it is illustrated in Chapters 4 and 5 that we can consider  $a(\omega)$  and  $z_0(\omega)$  to be random variables where  $\omega$  is an event in an underlying probability space. Similarly, we can consider  $b(t, \omega)$  and  $z(t, \omega)$  to be random processes in the manner detailed in Section 4.5 with the added assumption that  $b$  is continuous in  $t$  for each realization of  $\omega$ . The resulting model is the random differential equation

$$\begin{aligned} \frac{dz}{dt} &= a(\omega)z + b(t, \omega), \\ z(0) &= z_0(\omega), \end{aligned} \tag{3.3}$$

which has the solution

$$z(t, \omega) = e^{a(\omega)t} \left[ z_0(\omega) + \int_0^t e^{-a(\omega)s} b(s, \omega) ds \right]. \tag{3.4}$$

We note that sample paths specified by (3.4) are differentiable functions of  $t$ . As detailed in Section 4.7, this fundamentally delineates random differential equations from stochastic differential equations which are not differentiable in time and require Itô calculus for correct formulation and solution.

**Example 3.2 (Simple Harmonic Oscillator).** The simple harmonic oscillator model

$$\begin{aligned} m \frac{d^2z}{dt^2} + c \frac{dz}{dt} + kz &= f_0 \cos(\omega_F t), \\ z(0) = z_0, \quad \frac{dz}{dt}(0) = z_1 \end{aligned} \tag{3.5}$$

is a prototype for various damped, periodic processes. Here  $m > 0$ ,  $c \geq 0$ , and  $k > 0$  respectively denote the mass, damping, and stiffness coefficients and  $\omega_F$  is the driving frequency.

The solution of (3.5) is

$$z(t) = c_1 e^{r_1 t} + c_2 e^{r_2 t} + \frac{f_0}{\sqrt{m^2(\omega_0^2 - \omega_F^2)^2 + c^2 \omega_F^2}} \cos(\omega_F t - \delta), \tag{3.6}$$

where the natural frequency is  $\omega_0 = \sqrt{k/m}$ ,  $\delta$  is the solution to

$$\cos \delta = \frac{m(\omega_0^2 - \omega_F^2)}{\sqrt{m^2(\omega_0^2 - \omega_F^2)^2 + c^2 \omega_F^2}},$$

and

$$r_{1,2} = \frac{-c \pm \sqrt{c^2 - 4km}}{2m}$$

are the roots of the characteristic equation. The constants  $c_1$  and  $c_2$  are determined by the initial conditions.

Since  $m > 0$ ,  $e^{r_{1,2}t} \rightarrow 0$  as  $t \rightarrow \infty$ , so the solution limits to the periodic solution

$$z_p(t) = \frac{f_0}{\sqrt{m^2(\omega_0^2 - \omega_F^2)^2 + c^2\omega_F^2}} \cos(\omega_F t - \delta), \quad (3.7)$$

the maximum value of which is

$$Z_0 = \frac{f_0}{\sqrt{m^2(\omega_0^2 - \omega_F^2)^2 + c^2\omega_F^2}}. \quad (3.8)$$

In Chapter 9, we use (3.8) to illustrate the effects of nonlinear behavior near resonance. Further details regarding the use of this model to illustrate concepts from uncertainty quantification and statistical model validation can be found in [111].

A second case we will consider is the unforced problem  $f_0 = 0$ . For the case of underdamped motion,  $c^2 - 4km < 0$ , the solution can be expressed as

$$z(t) = e^{-(c/2m)t} \left[ c_1 \cos \left( \frac{\sqrt{4mk - c^2}}{2m} \cdot t \right) + c_2 \sin \left( \frac{\sqrt{4mk - c^2}}{2m} \cdot t \right) \right], \quad (3.9)$$

where

$$c_1 = z_0, \quad c_2 = \frac{c}{\sqrt{4mk - c^2}} z_0 + \frac{2m}{\sqrt{4mk - c^2}} z_1. \quad (3.10)$$

For the model (3.5),  $m, c, k, f_0, z_0$ , and  $z_1$  may be uncertain inputs and hence can be considered as random variables. This will yield a random solution  $z(t, \omega)$  whose distribution we wish to quantify based on the densities for the inputs.

When estimating and constructing densities for parameters, it is critical that parameters be identifiable in the sense that they can be uniquely determined from observations. One can easily see that this is not the case for (3.5) since the parameter sets  $q = [m, c, k, f_0]$  and  $q = [1, \frac{c}{m}, \frac{k}{m}, \frac{f_0}{m}]$  yield the same state values  $z(t)$ . Hence we also employ the formulation

$$\begin{aligned} \frac{d^2z}{dt^2} + C \frac{dz}{dt} + K z &= F_0 \cos(\omega_F t), \\ z(0) = z_0, \quad \frac{dz}{dt}(0) &= z_1, \end{aligned} \quad (3.11)$$

where  $C = \frac{c}{m}$ ,  $K = \frac{k}{m}$ , and  $F_0 = \frac{f_0}{m}$ . The forced and unforced solutions to (3.11) can be easily obtained from (3.6), (3.7), and (3.9) through the substitutions  $c \rightarrow C$ ,  $k \rightarrow K$ ,  $f_0 \rightarrow F_0$ , and  $m \rightarrow 1$ .

In general, we can measure only a subset of the states in a model. For example, a proximity sensor can be used to measure displacements  $z$ , whereas a laser vibrometer provides velocity measurements. However, it is unlikely that both would

be used to measure the full state  $u = [z, \dot{z}]^T$ . Displacement or velocity observations can be represented by

$$y = \mathcal{C}^T u, \quad (3.12)$$

where  $\mathcal{C}^T = [1, 0]$  or  $\mathcal{C}^T = [0, 1]$ .

However, other applications require more general responses or QoI, such as

$$y = \int_0^{t_f} \gamma(t) z(t) dt, \quad (3.13)$$

where  $\gamma(t)$  is a weight or filter over the time interval of interest. The full set of parameters in this case would be  $q = [C, K, z_0, z_1, \gamma(t)]$ .

To provide a framework that includes both (3.12) and (3.13), we denote observations or responses by

$$y = \mathcal{R}(u, q). \quad (3.14)$$

It will be illustrated in Section 3.3 that  $\mathcal{R}$  can be interpreted as a functional or operator defined on an appropriate Hilbert space.

The role of  $\mathcal{R}$  is analogous to that of  $f$  in (1.4). In (3.14), we highlight the fact that we are observing the state  $u$  by including it as an argument of  $\mathcal{R}$ . The role of  $u$  in (1.4) is not directly considered, and the output is instead formulated in terms of the inputs  $q$  and independent variables  $\chi$ .

**Example 3.3 (HIV Model).** To illustrate aspects of uncertainty quantification for a nonlinear system of ODEs arising in a biomedical application, we employ the model

$$\begin{aligned} \dot{T}_1 &= \lambda_1 - d_1 T_1 - (1 - \varepsilon) k_1 V T_1, \\ \dot{T}_2 &= \lambda_2 - d_2 T_2 - (1 - f\varepsilon) k_2 V T_2, \\ \dot{T}_1^* &= (1 - \varepsilon) k_1 V T_1 - \delta T_1^* - m_1 E T_1^*, \\ \dot{T}_2^* &= (1 - f\varepsilon) k_2 V T_2 - \delta T_2^* - m_2 E T_2^*, \\ \dot{V} &= N_T \delta(T_1^* + T_2^*) - c V - [(1 - \varepsilon) \rho_1 k_1 T_1 + (1 - f\varepsilon) \rho_2 k_2 T_2] V, \\ \dot{E} &= \lambda_E + \frac{b_E(T_1^* + T_2^*)}{T_1^* + T_2^* + K_b} E - \frac{d_E(T_1^* + T_2^*)}{T_1^* + T_2^* + K_d} E - \delta_E E. \end{aligned} \quad (3.15)$$

This is a slightly simplified version of (2.20) in which we neglect the contributions of noninfective virus  $V_{NI}$  and protease inhibitor  $\varepsilon_2(t)$ . As detailed in Section 2.5.1,  $T_1$  and  $T_1^*$  represent the populations of uninfected and infected T-lymphocytes,  $T_2$  and  $T_2^*$  are corresponding macrophage populations, and  $V, E$  denote the populations of free virus and immune effector cells.

The physical interpretation of the parameters is detailed in Table 2.1, and the values compiled in Table 3.1 are used in the examples of Chapters 8 and 9. The origin and units for these values are given in Table 1 of [2]. The initial conditions for subsequent examples are taken to be

$$T_1 = 0.9 \times 10^6, \quad T_2 = 4000, \quad T_1^* = 0.1, \quad T_2^* = 0.1, \quad V = 1, \quad E = 12. \quad (3.16)$$

|                             |                          |                          |                          |
|-----------------------------|--------------------------|--------------------------|--------------------------|
| $\lambda_1 = 1 \times 10^4$ | $d_1 = 0.01$             | $\varepsilon = 0$        | $k_1 = 8 \times 10^{-7}$ |
| $\lambda_2 = 31.98$         | $d_2 = 0.01$             | $f = 0.34$               | $k_2 = 1 \times 10^{-4}$ |
| $\delta = 0.7$              | $m_1 = 1 \times 10^{-5}$ | $m_2 = 1 \times 10^{-5}$ | $N_T = 100$              |
| $c = 13$                    | $\rho_1 = 1$             | $\rho_2 = 1$             | $\lambda_E = 1$          |
| $b_E = 0.3$                 | $K_b = 100$              | $d_E = 0.25$             | $K_d = 500$              |
| $\delta_E = 0.1$            |                          |                          |                          |

**Table 3.1.** Parameter values from Table 1 of [2].

Readers are referred to [2] for details regarding the use of this model to guide the development of treatment strategies for HIV.

**Example 3.4 (SIR Model).** Fundamental aspects of disease dynamics are quantified by the SIR model

$$\begin{aligned} \frac{dS}{dt} &= \delta N - \gamma k I S \quad , \quad S(0) = S_0, \\ \frac{dI}{dt} &= \gamma k I S - (r + \delta) I \quad , \quad I(0) = I_0, \\ \frac{dR}{dt} &= r I - \delta R \quad , \quad R(0) = R_0, \end{aligned} \tag{3.17}$$

where  $S(t)$ ,  $I(t)$ , and  $R(t)$  are the number of susceptible, infectious, and recovered individuals in a population of size  $N$ . Here  $\gamma$ ,  $k$ , and  $r$  respectively denote the infection coefficient, the interaction coefficient, which quantifies the probability that an individual comes in contact with others, and the recovery rate. The birth and death rates are assumed to equal with both denoted by  $\delta$ . It is observed that

$$\frac{dS}{dt} + \frac{dI}{dt} + \frac{dR}{dt} = 0$$

so that the total population  $S(t) + I(t) + R(t) = N$  is constant.

**Example 3.5 (Heat Equation).** The first law of thermodynamics and conservation of energy were used to quantify heat conduction and convection in the atmospheric equations (2.7) and thermal-hydraulic relation (2.17) used to model the flow and heating of coolant in a nuclear reactor. The following time-dependent and steady state heat equations encapsulate the prototypical behavior of both heat conduction and diffusion processes.

For the experimental configuration, which we will consider in Chapters 7 and 8 when quantifying parameter uncertainty, we consider copper and aluminum rectangular, insulated rods with cross-sectional dimensions  $a = b = 0.95$  cm and length  $L = 70$  cm. A heat source  $\Phi$  provides a fixed, but unknown, heat flux at  $x = 0$ . As detailed in [26], an energy balance yields the model

$$\rho c_p \frac{\partial T}{\partial t} = \frac{\partial}{\partial x} \left( k \frac{\partial T}{\partial x} \right) - \frac{2(a+b)h}{ab} [T(t, x) - T_{amb}] \quad , \quad 0 < x < L, \tag{3.18}$$

and boundary conditions

$$k \frac{dT}{dx}(t, 0) = \Phi \quad , \quad k \frac{dT}{dx}(t, L) = h[T_{amb} - T_s(t, L)]. \quad (3.19)$$

Here  $T$ ,  $\rho$ ,  $c_\rho$ ,  $k$ ,  $h$ , and  $T_{amb}$  respectively denote the temperature, density, specific heat, thermal conductivity, convective heat transfer coefficient, and ambient room temperature. Initial conditions are specified as

$$T(0, x) = T_0(x). \quad (3.20)$$

In experiments, temperatures are allowed to equilibrate to a steady state  $T_s(x)$  which is initially modeled by the boundary value problem

$$\begin{aligned} \frac{d^2T_s}{dx^2} &= \frac{2(a+b)}{ab} \frac{h}{k} [T_s(x) - T_{amb}], \\ \frac{dT_s}{dx}(0) &= \frac{\Phi}{k} \quad , \quad \frac{dT_s}{dx}(L) = \frac{h}{k} [T_{amb} - T_s(L)]. \end{aligned} \quad (3.21)$$

It is clear that the three parameters  $h$ ,  $k$ , and  $\Phi$  are not uniquely defined since (3.21) can be reformulated in terms of two parameters  $\bar{h} \equiv \frac{h}{k}$  and  $\bar{\Phi} \equiv \frac{\Phi}{k}$ . Because the thermal conductivity  $k$  is known for aluminum and copper, we respectively use the values  $k = 2.37 \frac{W}{cm \cdot C}$  and  $k = 4.01 \frac{W}{cm \cdot C}$  when modeling the aluminum and copper rods. These values lie within the ranges  $2.04\text{--}2.50 \frac{W}{cm \cdot C}$  and  $3.53\text{--}4.01 \frac{W}{cm \cdot C}$  reported for aluminum and copper. The source heat flux  $\Phi$  and convective heat transfer coefficient  $h$  are unknown, so the parameter set to be estimated and statistically analyzed is  $q = [\Phi, h]$ .

The solution of (3.21) is

$$T_s(x, q) = c_1(q)e^{-\gamma x} + c_2(q)e^{\gamma x} + T_{amb}, \quad (3.22)$$

where  $\gamma = \sqrt{\frac{2(a+b)h}{abk}}$  and

$$c_1(q) = -\frac{\Phi}{k\gamma} \left[ \frac{e^{\gamma L}(h + k\gamma)}{e^{-\gamma L}(h - k\gamma) + e^{\gamma L}(h + k\gamma)} \right] \quad , \quad c_2(q) = \frac{\Phi}{k\gamma} + c_1(q). \quad (3.23)$$

We suppress the parameter dependence of  $\gamma$  to clarify the notation. Observations for this experiment consist of temperature measurements at 15 equally spaced spatial locations  $x_j = x_0 + (j - 1)\Delta x$ ,  $j = 1, \dots, 15$ , where  $x_0 = 10$  cm and  $\Delta x = 4$  cm. Steady state temperature data for rectangular, uninsulated aluminum and copper rods is compiled in Tables 3.2 and 3.3.

We develop and analyze an extension of this steady state model in Example 12.4 of Section 12.2.

More generally, we will consider the model

$$\begin{aligned} \frac{\partial T}{\partial t} &= \frac{\partial}{\partial x} \left( \alpha(x) \frac{\partial T}{\partial x} \right) + f(t, x), \\ T(t, -1) &= T_\ell \quad , \quad T(t, 1) = T_r, \\ T(0, x) &= T_0(x), \end{aligned} \quad (3.24)$$

|           |       |       |       |       |       |       |       |       |
|-----------|-------|-------|-------|-------|-------|-------|-------|-------|
| $x$ (cm)  | 10    | 14    | 18    | 22    | 26    | 30    | 34    | 38    |
| Temp (°C) | 96.14 | 80.12 | 67.66 | 57.96 | 50.90 | 44.84 | 39.75 | 36.16 |
| $x$ (cm)  | 42    | 46    | 50    | 54    | 58    | 62    | 66    |       |
| Temp (°C) | 33.31 | 31.15 | 29.28 | 27.88 | 27.18 | 26.40 | 25.86 |       |

**Table 3.2.** Steady state temperatures measured at locations  $x$  for an aluminum rod.

where  $\alpha(x)$  is a spatially varying thermal diffusivity and  $f(t, x)$  is a distributed source term. Uncertainties in initial and boundary conditions can be interpreted as random variables, whereas it is illustrated in Section 4.5 that a random field  $\alpha(x, \omega)$  can be used to incorporate uncertainty in the spatially varying diffusivity.

|           |       |       |       |       |       |       |       |       |
|-----------|-------|-------|-------|-------|-------|-------|-------|-------|
| $x$ (cm)  | 10    | 14    | 18    | 22    | 26    | 30    | 34    | 38    |
| Temp (°C) | 66.04 | 60.04 | 54.81 | 50.42 | 46.74 | 43.66 | 40.76 | 38.49 |
| $x$ (cm)  | 42    | 46    | 50    | 54    | 58    | 62    | 66    |       |
| Temp (°C) | 36.42 | 34.77 | 33.18 | 32.36 | 31.56 | 30.91 | 30.56 |       |

**Table 3.3.** Steady state temperatures measured at locations  $x$  for a copper rod.

**Example 3.6 (Neutron Diffusion).** In Section 2.4.2, we summarized the neutron transport equations in the absence of diffusion. Steady state 1-D neutron diffusion in a material of width  $2a$  can be modeled as

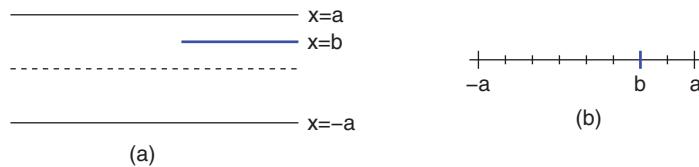
$$A_a \varphi - D \frac{d^2 \varphi}{dx^2} = S \quad , \quad x \in (-a, a), \quad (3.25)$$

where  $\varphi$ ,  $D$ ,  $A_a$ , and  $S$  respectively denote the neutron flux, a diffusion coefficient, a macroscopic absorption cross-section, and a constant distributed source [50]. Whereas the notation  $\Sigma_a$  is commonly employed in nuclear physics, we use  $A_a$  to avoid confusion with summations. The flux is assumed to vanish at  $x = \pm a$ , yielding the boundary conditions

$$\varphi(\pm a) = 0. \quad (3.26)$$

As illustrated in Figure 3.1(a), the response is measured using a detector with area  $A_d$  located at  $x = b$  so that

$$y = A_d \varphi(b). \quad (3.27)$$

**Figure 3.1.** (a) Material geometry and (b) finite difference grid with  $h = 2/N$ ,  $N = 8$ .

The parameters are  $q = [A_a, D, S, A_d]$ , and the solution to (3.25) with the boundary conditions (3.26) is

$$\varphi(x) = \frac{S}{A_a} \left( 1 - \frac{\cosh(xk)}{\cosh(ak)} \right), \quad k = \sqrt{A_a/D}. \quad (3.28)$$

It was noted in Section 2.4.2 that cross-section measurements and uncertainties for numerous materials have been catalogued for the last 70 years. Depending on the application, uncertainties in  $D$  and  $S$  can be determined using the techniques of Chapters 7 and 8.

To discretize (3.25), we consider a mesh  $x_i = -a + ih$ , where  $h = \frac{2a}{N}$  and  $N = 8$ , as shown in Figure 3.1(b). We assume that the observation location  $b$  is at  $\frac{a}{2}$  so that it coincides with  $x_6$ . Note that this choice of  $N$  and  $b$  was made solely to permit the specific description of the observation functional, and general choices for  $N$  and  $b$  are equally valid. Finally, we let  $\varphi_i \approx \varphi(x_i)$  denote approximate solutions so that the solution vector is  $\phi = [\varphi_1, \dots, \varphi_{N-1}]^T$ .

Discretization of the second derivative term using a central difference Taylor approximation and enforcement of boundary conditions yields

$$A_a \varphi_i - D \frac{\varphi_{i+1} - 2\varphi_i + \varphi_{i-1}}{h^2} = S, \quad i = 1, \dots, N-1,$$

which can be formulated as

$$\begin{bmatrix} h^2 A_a + 2D & -D & & \\ -D & h^2 A_a + 2D & -D & \\ & \ddots & \ddots & \ddots \\ & & -D & h^2 A_a + 2D \\ & & & -D & h^2 A_a + 2D \end{bmatrix} \begin{bmatrix} \varphi_1 \\ \varphi_2 \\ \vdots \\ \varphi_{N-2} \\ \varphi_{N-1} \end{bmatrix} = \begin{bmatrix} h^2 S \\ h^2 S \\ \vdots \\ h^2 S \\ h^2 S \end{bmatrix} \quad (3.29)$$

or

$$A(q)\phi = s(q). \quad (3.30)$$

The response (3.27) can be formulated as

$$y = \mathcal{C}^T(q)\phi = \mathcal{C}^T(q)A^{-1}(q)s(q). \quad (3.31)$$

For  $N = 8$  and  $b = \frac{a}{2}$ , the observation vector is

$$\mathcal{C}^T(q) = [0, 0, 0, 0, 0, A_d, 0, 0]. \quad (3.32)$$

**Example 3.7 (Beam Equation).** Consider a thin cantilever beam driven by a voltage spike  $V(t)$  applied to a surface mounted piezoelectric actuator, as shown in Figure 3.2. The beam is clamped at  $x = 0$ , and the location of the patch is designated by  $[x_1, x_2]$ . We let  $b$ ,  $h$ , and  $L$  respectively designate the width, thickness, and length of the beam and  $b_p, h_p$  denote the patch width and thickness. Transverse beam displacements are denoted by  $w(t, x)$ , and  $f(t, x)$  denotes a general transverse force applied to the beam.



**Figure 3.2.** Thin beam driven by a surface-mounted piezoelectric patch with displacements  $y_i$  measured at  $\bar{x} = 128$  mm.

It is shown in [225] that displacements can be modeled by the Euler–Bernoulli equation

$$\begin{aligned} \rho(x) \frac{\partial^2 w}{\partial t^2} + \gamma \frac{\partial w}{\partial t} + \frac{\partial^2 M}{\partial x^2} &= f(t, x) \quad , \quad 0 < x < L, \quad t > 0, \\ w(t, 0) = \frac{\partial w}{\partial x}(t, 0) &= 0 \quad , \quad M(t, L) = \frac{\partial M}{\partial x}(t, L) = 0 \quad , \quad t > 0, \\ w(0, x) = w_0(x) \quad , \quad \frac{\partial w}{\partial t}(0, x) &= w_1(x) \quad , \quad 0 < x < L, \end{aligned} \quad (3.33)$$

where  $\gamma$  is an air damping coefficient and the moment is

$$M(t, x) = YI(x) \frac{\partial^2 w}{\partial x^2} + cI(x) \frac{\partial^3 w}{\partial x^2 \partial t}.$$

To accommodate the differing geometry and material properties in the region covered by the patch, the density, stiffness, and damping terms are given by the piecewise constant relations

$$\begin{aligned} \rho(x) &= \rho h b + \rho_p h_p b_p \chi_p(x) \quad , \quad YI(x) = YI + Y_p I_p \chi_p(x), \\ cI(x) &= cI + c_p I_p \chi_p(x). \end{aligned} \quad (3.34)$$

Here  $\rho, \rho_p, Y, Y_p$ , and  $c, c_p$  are the density, Young's modulus, and Kelvin–Voigt damping coefficients for the beam and patch, and the characteristic function

$$\chi_p(x) = \begin{cases} 1 & , \quad x \in [x_1, x_2], \\ 0 & , \quad \text{else} \end{cases}$$

isolates the region covered by the patch. The moment of inertia  $I$  and constant  $I_p$  are given by

$$\begin{aligned} I &= b \int_{-h/2}^{h/2} z^2 dz = \frac{h^3 b}{12}, \\ I_p &= b_p \int_{h/2}^{h/2+h_p} z^2 dz = \frac{b_p}{3} \left[ (h/2 + h_p)^3 - (h/2)^3 \right]. \end{aligned} \quad (3.35)$$

The force generated by the application of a small voltage  $V(t)$  to the patch can be

| Component    | Geometry (mm)               | Material Properties                              |
|--------------|-----------------------------|--|
| Beam         | $393 \times 26 \times 1.25$ | $Y = 69 \text{ GPa}, \rho = 2700 \text{ kg/m}^3$ |
| PZT Actuator | $51 \times 26 \times 0.4$   |  |

**Table 3.4.** Geometry and material properties of the aluminum beam and QuickPack piezoelectric actuator.

approximated by

$$f(t, x) = \frac{\partial^2}{\partial x^2} [k_p \chi_p(x)] V(t).$$

To avoid issues associated with differentiating the piecewise constant characteristic function and to facilitate subsequent numerical implementation, one employs the weak model formulation

$$\begin{aligned} & \int_0^L \left[ \rho(x) \frac{\partial^2 w}{\partial t^2} + \gamma \frac{\partial w}{\partial t} \right] \phi dx + \int_0^L \left[ YI(x) \frac{\partial^2 w}{\partial x^2} + cI(x) \frac{\partial^3 w}{\partial x^2 \partial t} \right] \phi'' dx \\ &= k_p V(t) \int_{x_1}^{x_2} \phi'' dx, \end{aligned}$$

which must hold for all test functions  $\phi \in V = \{\phi \in H^2(0, L) \mid \phi(0) = \phi'(0) = 0\}$ .

The geometry and material properties of the beam and QuickPack piezoelectric actuator are compiled in Table 3.4, and the patch location is  $x_1 = 41$  mm and  $x_2 = 92$  mm. The values of  $\rho_p$  and  $Y_p$  are not well established since the QuickPack patch is a composite of a piezoelectric wafer embedded in an epoxy and Kapton matrix. Furthermore, the relations (3.34) and (3.35) neglect the glue layer. Hence we treat the linear density  $\tilde{\rho}_p = \rho_p h_p b_p$  and stiffness terms  $YI_p = Y_p I_p$  and  $YI_b = YI$  as parameters that must be estimated through a fit to data. Similarly, the parameters  $cI_b = cI$ ,  $cI_p = c_p I_p$ ,  $\gamma$ , and  $k_p$  must be estimated since there are no published values for these quantities. We fix  $\tilde{\rho}_b = \rho b = 0.08775$  to preserve the identifiability of the model. The parameter set is thus

$$q = [\tilde{\rho}_p, \gamma, YI_b, YI_p, cI_b, cI_p, k_p]. \quad (3.36)$$

Observations consist of displacement measurements  $y_i$  collected using a proximity sensor at the spatial location  $\bar{x} = 128$  mm. The corresponding model response is thus  $y(t_i, q) = w(t_i, \bar{x}, q)$ .

**Example 3.8 (Burgers' Equation).** The nonlinear equations (2.4) and (2.16), obtained by balancing momentum, quantify atmospheric and nuclear coolant flow processes. To provide a simple prototype, we consider the viscous Burgers' equation

$$\begin{aligned} & \frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} = \mu \frac{\partial^2 u}{\partial x^2}, \\ & u(t, 0) = u_\ell, \quad u(t, 1) = u_r, \\ & u(0, x) = u_0(x), \end{aligned} \quad (3.37)$$

which provides an example that has both nonlinear state and parameter dependence. It can exhibit uncertainty in the viscosity coefficient  $\mu$ , the boundary terms  $u_\ell$  and  $u_r$ , and the initial condition  $u_0(x)$ .

## 3.2 Evolution, Stationary, and Algebraic Models

The models in Section 3.1 can be generally categorized as evolution processes, stationary processes, or algebraic models. We summarize these general frameworks to provide the setting that we will employ for model calibration, uncertainty quantification, and model validation.

### Evolution Processes

The exponential model (3.1), simple harmonic oscillator model (3.5), HIV model (3.15), and SIR model (3.17) are finite-dimensional evolution processes, whereas the heat equations (3.18) and (3.24), beam equation (3.33), and Burgers' equation (3.37) are infinite-dimensional evolution models. All can be formulated as

$$\begin{aligned} \frac{du}{dt} &= g(t, u(t), q), \\ u(t_0) &= u_0, \end{aligned} \tag{3.38}$$

where  $q \in \mathbb{R}^p$  is a vector of parameter values and  $u$  is the state. For ODE models or spatial discretizations of PDEs (e.g., finite element, finite difference, or finite volume), the state  $u(t, q) \in \mathbb{R}^N$  is a finite-dimensional vector. More generally, one can consider  $u$  in a suitable Hilbert or Banach space in which case (3.38) additionally represents evolutionary PDEs [78, 190]. Since we are focusing on computational algorithms, we will consider only the finite-dimensional case in this chapter.

In general, we cannot observe all of the states  $u$  and instead consider continuous or discrete-time observations

$$y(t, q) = \mathcal{C}u(t, q)$$

or

$$y(t_j, q) = \mathcal{C}u(t_j, q), \quad j = 1, \dots, n. \tag{3.39}$$

We consider  $\mathcal{C}$  to be a  $\nu \times N$  matrix which yields  $\nu$  observations or responses  $y(t_j, q)$ . Note that  $\nu$  and  $N$  respectively designate the dimension of the response and the number of states. More general response relations are illustrated in (3.13) and Chapter 14.

We note that in most mathematical models for physical and biological applications, the solution  $u(t, q)$  will exhibit a nonlinear dependence on the parameters  $q$  even though the model may be linear with regard to the state  $u$ . Hence observations  $y(t, q)$  also typically exhibit a nonlinear parameter dependence.

As illustrated in Example 3.3, the individual states in coupled biological and physical models can vary over several orders of magnitude. This can stall the efficiency of numerical integration and optimization routines and produce unrealistic results if physically positive parameters become negative due to roundoff errors. In worst case scenarios, roundoff errors due to discrepant magnitude differences can

stop integration routines if they produce nonphysical growth or imaginary components.

Many of these issues can be avoided by replacing (3.38) by a log-transformed system. If we let  $\tilde{u}_i = \log_{10}(u_i)$ , this yields the system

$$\begin{aligned}\frac{d\tilde{u}_i}{dt} &= \frac{10^{-\tilde{u}_i}}{\ln(10)} g_i(t, 10^{\tilde{u}(t)}, q), \\ 10^{\tilde{u}_i}(t_0) &= \tilde{u}_0\end{aligned}\tag{3.40}$$

for  $i = 1, \dots, N$ . We employ transformed systems of this form in Chapters 7 and 8.

### Stationary Processes

The boundary value problems (3.21) and (3.25), used to model steady state heat conduction and neutron diffusion, are examples of stationary processes, as are elliptic PDEs. Stationary processes have the form

$$\begin{aligned}\mathcal{N}(u, q) &= F(q) \quad , \quad x \in \mathcal{D}, \\ B(u, q) &= G(q) \quad , \quad x \in \partial\mathcal{D},\end{aligned}\tag{3.41}$$

where  $\mathcal{N}$  is a linear or nonlinear differential operator,  $F$  denotes source terms,  $B$  and  $G$  denote boundary operators, and  $\mathcal{D}$  is a region in  $\mathbb{R}^1$ ,  $\mathbb{R}^2$ , or  $\mathbb{R}^3$ . The state  $u$  can be  $N$ -dimensional but, in our examples, we will consider  $u(x, q) \in \mathbb{R}^1$ . We also consider only the case when observations

$$y(x_j, q) = u(x_j, q)\tag{3.42}$$

are made at discrete points  $x_j$ ,  $j = 1, \dots, n$ .

For the steady state heat equation (3.21),  $u = T_s$ ,  $q = [\Phi, h]$ , and

$$\begin{aligned}\mathcal{N}(T_s, q) &= \frac{d^2T_s}{dx^2} - \frac{2(a+b)}{ab} \frac{h}{k} T_s, \\ F(q) &= -\frac{2(a+b)}{ab} \frac{h}{k} T_{amb}, \\ B(T_s, q) &= \begin{cases} \frac{dT_s}{dx} , & x = 0, \\ \frac{dT_s}{dx} + \frac{k}{h} T_s , & x = L, \end{cases} \quad , \quad G(q) = \begin{cases} \frac{\Phi}{k} , & x = 0, \\ \frac{k}{h} T_{amb} , & x = L. \end{cases}\end{aligned}$$

The observations are

$$y(x_j, q) = T_s(x_j, q) = c_1(q)e^{-\gamma x_j} + c_2(q)e^{\gamma x_j} + T_{amb},$$

where  $c_1(q)$  and  $c_2(q)$  are defined in (3.23). For the experimental data used for model calibration in later chapters, the points  $x_j$  are given in Table 3.2.

### Algebraic Models

Algebraic models arise when algebraic or polynomial relations are used to quantify physical or biological processes or result from the discretization of boundary value problems, as illustrated in Example 3.6. These models can be expressed as

$$\mathcal{N}(u, q) = 0,$$

where  $\mathcal{N}$  is a nonlinear or linear operator. Two special cases are

$$A(q)u = F(q), \quad A(u)q = F(q),$$

which represent linear dependence on the state and parameters.

### 3.3 Abstract Modeling Framework

The models in Examples 3.1–3.8 represent a range of algebraic relations, ODEs, and PDEs that exhibit both linear and nonlinear state dependencies. We provide here a general framework for the models that facilitate model calibration, propagation of uncertainties, and sensitivity analysis. Readers who are not familiar with functional analysis can skip this section without losing the physical understanding of the models.

#### 3.3.1 Linear Systems

We consider first systems that are linear in the state, or dependent variable, but typically exhibit a nonlinear dependence on inputs and parameters. Systems of  $N$  coupled equations can be formulated as

$$L(q)u = F(q(\chi)), \quad \chi \in \Omega, \quad (3.43)$$

where  $q(\chi) = [q_1(\chi), \dots, q_p(\chi)]^T$  are the parameters,  $u = [u_1(\chi), \dots, u_N(\chi)]^T$  is the state vector,  $L(q) = [L_1(q), \dots, L_N(q)]^T$  is a vector of operators that depend linearly on the state  $u$  and typically nonlinearly on the parameters  $q$ , and  $F(q) = [F_1(q), \dots, F_N(q)]^T$  are source terms. Potential spatial or temporal dependence is indicated by  $\chi = [x, t] \in \mathcal{D} \times \mathcal{T} \equiv \Omega$ , where  $\mathcal{D}$  is a subset of  $\mathbb{R}^1$ ,  $\mathbb{R}^2$ , or  $\mathbb{R}^3$  and  $\mathcal{T}$  is a subset of  $\mathbb{R}^1$ . If  $L$  represents differential operators, appropriate initial or boundary conditions are represented in operator form as

$$B(q)u = G(q), \quad \chi \in \partial\Omega, \quad (3.44)$$

where  $\partial\Omega$  is the boundary of  $\Omega$ .

The general observation or response is represented by

$$y = \mathcal{R}(u, q) \quad (3.45)$$

in the space  $H = H_u \times H_q$ , where  $H_u$  and  $H_q = \mathbb{Q}$  are Hilbert spaces for the state and parameters. Similarly, the sources  $F$  are assumed to be elements in the Hilbert space  $H_F$ . We note that in general,  $L$  will not be defined on all of  $H_u$  but rather on a dense subspace  $\text{dom}(L)$ . As indicated in Appendix A, for differential operators,  $\text{dom}(L)$  is often taken to be the subset of functions in a Sobolev space that satisfy boundary conditions.

**Example 3.9.** Consider the model (3.1) of Example 3.1. The state is  $u$  in the Hilbert space  $H_u = L^2(0, t_f)$  with the standard inner product. The parameters are

$q = [a, u_0, b(t)] \in \mathbb{R}^2 \times L^2(0, t_f)$  and  $H_F = L^2(0, t_f)$ . Since the independent variable is time, we have  $\chi = t \in [0, t_f]$ . With the operator definitions

$$L(q)u = \frac{du}{dt} - au, \quad B(q)u = u, \quad F(q(\chi)) = b(t), \quad G(q) = u_0,$$

the model can be formulated in the general framework (3.43)–(3.44).

**Example 3.10.** For the spring model of Example 3.2, appropriate operators are

$$L(q)u = \ddot{z} + C\dot{z} + Kz, \quad F(q) = F_0 \cos(\omega_F t), \quad B(q) = \begin{bmatrix} z \\ \dot{z} \end{bmatrix}, \quad G(q) = \begin{bmatrix} z_0 \\ z_1 \end{bmatrix}, \quad (3.46)$$

where  $u = z$  is considered in  $H_u = L^2(0, t_f)$  and  $\text{dom}(L) \subset H_u$ . The parameter set is taken to be  $q = [C, K, F_0, z_0, z_1]$  in the admissible parameter space  $\mathbb{Q} = [0, \infty) \times (0, \infty) \times \mathbb{R}^3$ .

Alternatively, one can consider the state  $u = [z, \dot{z}]$  in  $H^1(0, t_f) \times L^2(0, t_f)$  with the operators

$$L(q) = \begin{bmatrix} \frac{d}{dt} & -1 \\ K & \frac{d}{dt} + C \end{bmatrix}, \quad F(q) = \begin{bmatrix} 0 \\ F_0 \cos(\omega t) \end{bmatrix}, \quad B(q) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad G(q) = \begin{bmatrix} z_0 \\ z_1 \end{bmatrix}.$$

Whereas this definition isolates the action of the two states, we employ (3.46) in the sensitivity analysis of Chapter 14 since it facilitates construction of the adjoint operator.

**Example 3.11.** To pose the heat equation (3.24) in the framework (3.43)–(3.44), we take the parameters to be  $q = [\alpha, u_0, u_\ell, u_r] \in H^1(-1, 1) \times C[-1, 1] \times \mathbb{R}^2$  and note that  $\chi = [x, t] \in [-1, 1] \times [0, t_f]$ . We define the operators to be

$$\begin{aligned} L(q)u &= \frac{\partial u}{\partial t} - \frac{\partial}{\partial x} \left( \alpha \frac{\partial u}{\partial x} \right), & F(q) &= f, \\ B(q)u &= u, & G(q) &= \alpha_0 u_0 + \alpha_\ell u_\ell + \alpha_r u_r, \end{aligned}$$

where

$$\alpha_0 = \begin{cases} 1, & -1 < x < 1, t = 0, \\ 0, & \text{else,} \end{cases}, \quad \alpha_\ell = \begin{cases} 1, & x = -1, t \geq 0, \\ 0, & \text{else,} \end{cases}, \quad \alpha_r = \begin{cases} 1, & x = 1, t \geq 0, \\ 0, & \text{else.} \end{cases}$$

The regularity of the Hilbert space  $H_u$  depends on the regularity of  $H_F$ . For example, if  $u_\ell = u_r = 0$  and  $f \in L^2((0, t_f), H^{-1}(-1, 1))$ , where  $H^{-1}$  is the dual space of  $H_0^1$ , then  $u \in L^2((0, t_f), X)$ , where  $X = L^2(-1, 1)$ ; e.g., see [25, 207].

**Example 3.12.** For the boundary value problem (3.25) and response (3.27), the parameters are  $q = [A_a, D, S, A_d] \in \mathbb{R}^4$  and  $\chi = x \in [-a, a]$ . The operators are

$$L(q)u = A_a u - D \frac{d^2 u}{dx^2}, \quad F(q) = S, \quad B(q)u = u, \quad G(q) = 0,$$

and the response can be formulated as

$$y = \mathcal{R}(\varphi, q) = \int_{-a}^a A_d \varphi(x) \delta(x - b) dx.$$

Appropriate Hilbert spaces are  $H_u = H_F = L^2(-a, a)$  with the inner product

$$\langle f, g \rangle = \int_{-a}^a f(x)g(x) dx.$$

Here  $\text{dom}(L) \subset H_u$  can be taken to be  $H_0^1(-a, a) = \{\phi \in H^1(-a, a) | \phi(\pm a) = 0\}$ .

**Example 3.13.** The matrix system (3.30) can be posed in the abstract framework (3.43) by taking  $L = A$ ,  $F = f$ , and  $u = \phi$  with  $H_u$  and  $H_F$  taken to be the Euclidean space  $\mathbb{R}^{N-1}$  with the standard dot product. The adjoint operator is the matrix transpose  $L^* = A^T$ .

### 3.3.2 Nonlinear Systems

For models such as Burgers' equation, illustrated in Example 3.8, the general framework must be extended to accommodate the nonlinearity in the dependent variable. In this case the model can be formulated as

$$\mathcal{N}(u(\chi), q(\chi)) = F(q(\chi)) \quad , \quad \chi \in \Omega, \quad (3.47)$$

where  $\mathcal{N} = [\mathcal{N}_1(u, q), \dots, \mathcal{N}_N(u, q)]^T$  is an  $N$ -vector of nonlinear operators. The definitions of  $\chi, u, q$ , and  $F$  are the same as those employed in Section 3.3.1 for linear operators. Similarly, initial or boundary conditions are represented by

$$B(u, q) = G(q) \quad , \quad \chi \in \partial\Omega. \quad (3.48)$$

We refer the reader to [50, 53] for additional details and examples regarding the representation of nonlinear models in this framework.

**Example 3.14.** For the viscous Burgers' equation (3.37), the parameters are  $q = [u_0, \mu, u_\ell, u_r] \in C[0, 1] \times \mathbb{R}^3$  and the operators are

$$\begin{aligned} \mathcal{N}(u, q) &= \frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} - \mu \frac{\partial^2 u}{\partial x^2} \quad , \quad F(q) = 0, \\ B(u, q) &= u \quad , \quad G(q) = \alpha_0 u_0 + \alpha_\ell u_\ell + \alpha_r u_r, \end{aligned}$$

where  $\alpha_0, \alpha_\ell$ , and  $\alpha_r$  are defined in a manner analogous to Example 3.10.

## 3.4 Notation for Parameters and Inputs

The notation for parameters varies widely among disciplines. In statistics,  $\theta$  is commonly employed to denote *calibration parameters*, whereas general parameters are

commonly denoted by  $q$  and  $\alpha$  in the mathematics and nuclear engineering literature. Numerous other conventions are employed in other sciences and engineering.

We use the mathematics notation  $q$  for two reasons. The first is that it permits significant flexibility when discussing the associated random variables  $Q$  and admissible parameter spaces  $\mathcal{Q}$  and  $\mathbb{Q}$ . Second, it is not limited to solely representing calibration parameters as is the case for  $\theta$  in the statistics literature. This is important since we are often interested in propagating uncertainties associated with both calibration parameters, which are determined using the statistical techniques of Chapters 7 and 8, and physical parameters, whose uncertainties may be determined from prior experiments. We do not differentiate between calibration and physical parameters since the designation is typically indicated by the context. Whereas we use  $q$  to denote this combined parameter set, we warn readers that the notation in their respective disciplines will often differ.

As indicated in Definition 1.2, the term inputs is used to designate parameters, initial conditions, boundary conditions, or exogenous forces that exhibit uncertainties which must be determined and propagated through models to quantify response uncertainties. Once the techniques of Chapter 5 have been employed to represent random inputs, we often treat the terms inputs and parameters as synonymous.

## 3.5 Exercises

**Exercise 3.1.** Consider the heat equation (3.24) with constant diffusivity  $\alpha$  and  $f(t, x) = T_\ell = T_R = 0$ . For  $N + 1$  spatial gridpoints and temporal stepsize  $k$ , define the grid  $(x_i, t_j)$ , where  $x_i = -1 + ih$ ,  $h = \frac{2}{N}$ , for  $i = 0, \dots, N$ , and  $t_j = jk$ . Approximate solutions at gridpoints are denoted by  $T_{i,j} \approx T(x_i, t_j)$ . Use a central difference Taylor approximation in space and forward difference in time to obtain the discrete relation

$$\frac{T_{i,j+1} - T_{i,j}}{k} = \alpha \frac{T_{i+1,j} - 2T_{i,j} + T_{i-1,j}}{h^2}$$

or

$$T_{i,j+1} = (1 - 2\lambda)T_{i,j} + \lambda(T_{i+1,j} + T_{i-1,j}), \quad (3.49)$$

where  $\lambda = \frac{\alpha k}{h^2}$ . The initial conditions are  $T_{i,0} = T_0(x_i)$  and the boundary conditions yield  $T_{0,j} = T_{N,j} = 0$ . The relation (3.49) can be expressed as

$$\mathcal{T}^{j+1} = A\mathcal{T}^j = A^{j+1}\mathcal{T}^0, \quad (3.50)$$

where  $\mathcal{T}^j = [T_{1,j}, \dots, T_{N-1,j}]^T$  and

$$A = \begin{bmatrix} 1 - 2\lambda & \lambda & & & \\ \lambda & 1 - 2\lambda & \lambda & & \\ & \ddots & \ddots & \ddots & \\ & & \lambda & 1 - 2\lambda & \lambda \\ & & & \lambda & 1 - 2\lambda \end{bmatrix}. \quad (3.51)$$

For applications in which the initial heat distribution  $T_0(x)$  is unknown and treated as a random field, (3.50) provides a discretized form of the problem that depends linearly on the random vector  $\mathcal{T}^0$ .

## Chapter 4

# Fundamentals of Probability, Random Processes, and Statistics

We summarize in this chapter those aspects of probability, random processes, and statistics that are employed in subsequent chapters. The discussion is necessarily brief, and additional details can be found in the references cited in the text and noted in Section 4.9.

## 4.1 Random Variables, Distributions, and Densities

When constructing statistical models for physical and biological processes, we will consider parameters and measurement errors to be random variables whose statistical properties or distributions we wish to infer using measured data. The classical probability space provides the basis for defining and illustrating these concepts.

**Definition 4.1 (Probability Space).** A probability space  $(\Omega, \mathcal{F}, P)$  is comprised of three components:

$\Omega$ : sample space is the set of all possible outcomes from an experiment;

$\mathcal{F}$ :  $\sigma$ -field of subsets of  $\Omega$  that contains all events of interest;

$P : \mathcal{F} \rightarrow [0, 1]$ : probability or measure that satisfies the postulates

(i)  $P(\emptyset) = 0$ ,

(ii)  $P(\Omega) = 1$ ,

(iii) if  $A_i \in \mathcal{F}$  and  $A_i \cap A_j = \emptyset$ , then  $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$ .

We note that the concept of probability depends on whether one is considering a frequentist (classical) or Bayesian perspective. In the frequentist view, probabilities are defined as the frequency with which an event occurs if the experiment is repeated a large number of times. The Bayesian perspective treats probabilities as a distribution of subjective values, rather than a single frequency, that are constructed or updated as data is observed.

**Example 4.2.** Consider an experiment in which we flip two individual coins (e.g., a quarter and nickel) multiple times and record the outcome which consists of an ordered pair. The sample space and  $\sigma$ -field of events are thus

$$\begin{aligned}\Omega &= \{(H, H), (T, H), (H, T), (T, T)\}, \\ \mathcal{F} &= \{\emptyset, (H, H), (T, H), (H, T), (T, T), \Omega, \{(H, T), (T, H), \dots\}\}.\end{aligned}\tag{4.1}$$

Note that  $\mathcal{F}$  contains all countable intersections and unions of elements in  $\Omega$ . If we flip the pair twice, two possible events are

$$A = \{(H, H), (T, H)\}, \quad B = \{(H, H), (H, T)\}.$$

For fair coins, the frequentist perspective yields the probabilities

$$P(A) = \frac{1}{2}, \quad P(B) = \frac{1}{2}, \quad P(A \cap B) = \frac{1}{4}, \quad P(A \cup B) = \frac{3}{4}.$$

We note that because the events are independent,  $P(A \cap B) = P(A)P(B)$ . We will revisit the probabilities associated with flipping a coin from the Bayesian perspective in Example 4.66 of Section 4.8.2.

We now define univariate random variables, distributions, and densities.

### 4.1.1 Univariate Concepts

**Definition 4.3 (Random Variable).** A random variable is a function  $X : \Omega \rightarrow \mathbb{R}$  with the property that  $\{\omega \in \Omega | X(\omega) \leq x\} \in \mathcal{F}$  for each  $x \in \mathbb{R}$ ; i.e., it is measurable. A random variable is said to be discrete if it takes values in a countable subset  $\{x_1, x_2, \dots\}$  of  $\mathbb{R}$ .

**Definition 4.4 (Realization).** The value

$$x = X(\omega)$$

of a random variable  $X$  for an event  $\omega \in \Omega$  is termed a realization of  $X$ .

We note that in the statistics literature, many authors employ the same notation for the random variable and realization and let the context dictate the meaning. For those who are new to the field, this can obscure the meaning and, to the degree possible, we will use different notation for random variables and their realizations.

**Definition 4.5 (Cumulative Distribution Function).** Associated with every random variable  $X$  is a cumulative distribution function (cdf)  $F_X : \mathbb{R} \rightarrow [0, 1]$  given by

$$F_X(x) = P\{\omega \in \Omega | X(\omega) \leq x\}.\tag{4.2}$$

This is often expressed as  $F_X(x) = P\{X \leq x\}$ , which should be interpreted in the sense of (4.2). The following example illustrates the construction of a cdf for a discrete random variable.

**Example 4.6.** Consider the experiment of Example 4.2 in which our event  $\omega$  consists of a single flip of a pair of coins. We define  $X(\omega)$  to be the number of heads associated with the event so that

$$\begin{aligned} X(H, H) &= 2, \\ X(H, T) &= X(T, H) = 1, \\ X(T, T) &= 0. \end{aligned}$$

For  $x < 0$ , the probability of finding an event  $\omega \in \Omega$  such that  $X(\omega) \leq x$  is 0, so  $F_X(x) = 0$  for  $x < 0$ . Similar analysis yields the cdf relation

$$F_X(x) = \begin{cases} 0 & , \quad x < 0, \\ 1/4 & , \quad 0 \leq x < 1, \\ 3/4 & , \quad 1 \leq x < 2, \\ 1 & , \quad x \geq 2, \end{cases}$$

which is plotted in Figure 4.1.

It is observed that, by construction, the cdf satisfies the properties

- (i)  $\lim_{x \rightarrow -\infty} F_X(x) = 0,$
- (ii)  $x_1 \leq x_2 \Rightarrow F_X(x_1) \leq F_X(x_2),$
- (iii)  $\lim_{x \rightarrow \infty} F_X(x) = 1.$

This is an example of a càdlàg (French “continue à droite, limite à gauche”) function that is right-continuous and has left limits everywhere. These functions also arise in stochastic processes that admit jumps.

For continuous and discrete random variables the probability density function and probability mass function are defined as follows.

**Definition 4.7 (Probability Density Function).** The random variable  $X$  is continuous if its cdf is absolutely continuous and hence can be expressed as

$$F_X(x) = \int_{-\infty}^x f_X(s)ds , \quad x \in \mathbb{R},$$

where the derivative  $f_X = \frac{dF_x}{dx}$  mapping  $\mathbb{R}$  to  $[0, \infty)$  is called the probability density function (pdf) of  $X$ .

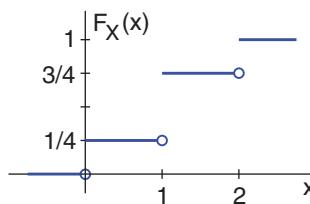


Figure 4.1. Cdf for Example 4.6.

**Definition 4.8 (Probability Mass Function).** The probability mass function of a discrete random variable  $X$  is given by  $f_X(x) = P(X = x)$ .

The pdf properties

- (i)  $f_X(x) \geq 0,$
- (ii)  $\int_{\mathbb{R}} f_X(x)dx = 1,$
- (iii)  $P(x_1 \leq X \leq x_2) = F_X(x_2) - F_X(x_1) = \int_{x_1}^{x_2} f_X(x)dx$

follow immediately from the definition and (4.3). The attributes of density functions can be further specified by designating their location or centrality, their spread or variability, their symmetry, and the contribution of tail behavior. In general, this information is provided by moments

$$\mathbb{E}(X^n) = \int_{\mathbb{R}} x^n f_X(x)dx$$

or central moments. For example, the mean

$$\mu = \mathbb{E}(X) = \int_{\mathbb{R}} xf_X(x)dx,$$

also termed the first moment or expected value, provides a measure of the density's central location, whereas the second central moment

$$\sigma^2 = \text{var}(X) = \mathbb{E}[(X - \mu)^2] = \int_{\mathbb{R}} (x - \mu)^2 f_X(x)dx \quad (4.4)$$

provides a measure of the density's variability or width. This typically is termed the variance of  $X$ , and  $\sigma$  is called the standard deviation. One often employs the relation

$$\sigma^2 = \mathbb{E}(X^2) - \mu^2,$$

which results directly from (4.4). We note that the third moment (skewness) quantifies the density's symmetry about  $\mu$ , whereas the fourth moment (kurtosis) quantifies the magnitude of tail contributions.

### Important Distributions for Inference and Model Calibration

We summarize next properties of the univariate normal, uniform, chi-squared, Student's  $t$ , beta, gamma, inverse-gamma, and inverse chi-squared distributions which are important for frequentist and Bayesian inference and model calibration.

**Definition 4.9 (Normal Distribution).** In uncertainty quantification, a commonly employed univariate density is the normal density

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, \quad -\infty < x < \infty.$$

The associated cdf is

$$F_X(x) = \int_{-\infty}^x f(s)ds = \frac{1}{2} \left[ 1 + \operatorname{erf}\left(\frac{x - \mu}{\sigma\sqrt{2}}\right) \right],$$

where the error function is defined to be

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-s^2} ds.$$

The notation  $X \sim N(\mu, \sigma^2)$  indicates that the random variable  $X$  is normally distributed with mean  $\mu$  and variance  $\sigma^2$ . For the normal density, 68.29% of the area is within  $1\sigma$  of the mean  $\mu$  and 95.45% is within  $2\sigma$ , as illustrated in Figure 4.2(a).

**Definition 4.10 (Continuous Uniform Distribution).** A random variable  $X$  is uniformly distributed on the interval  $[a, b]$ , denoted by  $X \sim U(a, b)$ , if any value in the interval is achieved with equal probability. The pdf and cdf are thus

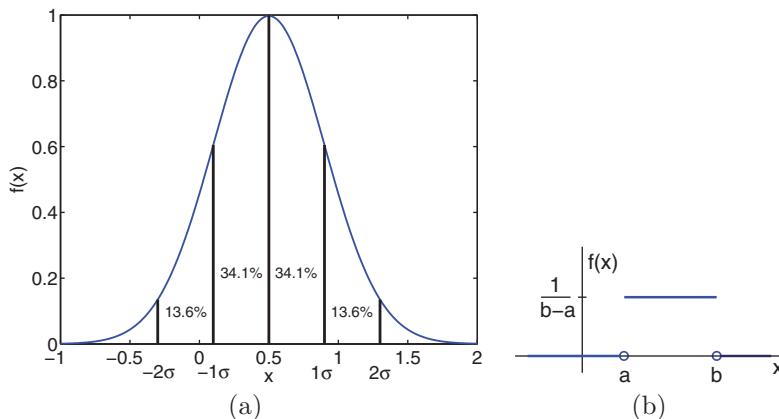
$$f_X(x) = \frac{1}{b-a} \chi_{[a,b]}(x) \quad (4.5)$$

and

$$F_X(x) = \begin{cases} 0 & , \quad x < a, \\ \frac{x-a}{b-a} & , \quad a \leq x < b, \\ 1 & , \quad x \geq b, \end{cases} \quad (4.6)$$

where the characteristic function  $\chi_{[a,b]}(x)$  is defined to be unity on the interval  $[a, b]$  and 0 elsewhere. The pdf is plotted in Figure 4.2(b). It is established in Exercise 4.1 that the mean and variance of  $X$  are

$$\mathbb{E}(X) = \frac{a+b}{2}, \quad \text{var}(x) = \frac{(b-a)^2}{12}, \quad (4.7)$$



**Figure 4.2.** (a) Normal density with  $\mu = 0.5$  and  $\sigma = 0.4$  and areas within  $1\sigma$  and  $2\sigma$  of  $\mu$ . (b) Uniform density on the interval  $[a, b]$ .

and the relationship between  $X \sim \mathcal{U}(a, b)$  and  $Z \sim \mathcal{U}(-1, 1)$  is established in Exercise 4.6. When prior information is lacking, it is often assumed that model parameters have a uniform density.

**Definition 4.11 (Chi-Squared Distribution).** Let  $X \sim N(0, 1)$  be normally distributed. The random variable  $Y = X^2$  then has a chi-squared distribution with 1 degree of freedom, denoted by  $Y \sim \chi^2(1)$ . Furthermore, if  $Y_i$ ,  $i = 1, \dots, k$ , are independent  $\chi^2(1)$  random variables, then their sum  $Z = \sum_{i=1}^k Y_i$  is a  $\chi^2$  random variable with  $k$  degrees of freedom, denoted by  $Z \sim \chi^2(k)$  or  $Z \sim \chi_k^2$ . The pdf

$$f_Z(z; k) = \begin{cases} \frac{z^{k/2-1} e^{-z/2}}{2^{k/2} \Gamma(k/2)} & , z \geq 0, \\ 0 & , z < 0, \end{cases} \quad (4.8)$$

can be compactly expressed in terms of the gamma function, where  $\Gamma(k/2) = \sqrt{\pi} \frac{(k-2)!!}{2^{(k-1)/2}}$  for odd  $k$ , and exhibits the behavior shown in Figure 4.3(a). The mean and variance of  $Z$  are

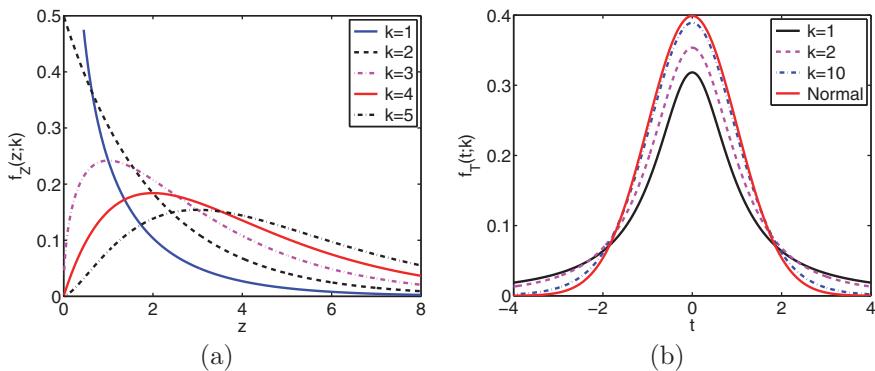
$$\mathbb{E}(Z) = k, \quad \text{var}(Z) = 2k.$$

Chi-squared distributions naturally arise when evaluating the sum of squares error between measured data and model values when estimating model parameters.

**Definition 4.12 (Student's  $t$ -Distribution).** Let  $X \sim N(0, 1)$  and  $Z \sim \chi^2(k)$  be independent random variables. The random variable

$$T = \frac{X}{\sqrt{Z/k}}$$

has a Student's  $t$ -distribution (or simply  $t$ -distribution) with  $k$  degrees of freedom.



**Figure 4.3.** (a) Chi-squared density for  $k = 1, \dots, 5$  and (b) Student's  $t$ -density with  $k = 1, 2, 10$  compared with the normal density with  $\mu = 0, \sigma = 1$ .

The pdf can be expressed as

$$f_T(t; k) = \frac{\Gamma((k+1)/2)}{\Gamma(k/2)\sqrt{k\pi}} \left(1 + \frac{t^2}{k}\right)^{-(k+1)/2},$$

where  $\Gamma$  again denotes the gamma function. Note that

$$f_T(t; 1) = \frac{1}{\pi(1+t^2)}$$

is a special case of the Cauchy distribution. As illustrated in Figure 4.3(b), the density is symmetric and bell-shaped, like the normal density, but exhibits heavier tails.

It will be shown in Section 7.2 that the  $t$ -distribution naturally arises when estimating the mean of a population when the sample size is relatively small and the population variance is unknown.

On a historic note, aspects of this theory were developed by William Sealy Gosset, an employee of the Guinness brewery in Dublin, in an effort to select optimally yielding varieties of barley based on relatively small sample sizes. To improve perception following the recent disclosure of confidential information by another employee, Gosset was allowed to publish only under the pseudonym “Student.” The importance of his work was advocated by both Karl Pearson and R.A. Fisher.

**Definition 4.13 (Gamma Distribution).** The gamma distribution is a two-parameter family with two common parameterizations: (i) shape parameter  $\alpha > 0$  and scale parameter  $\lambda > 0$  or (ii) shape parameter  $\alpha$  and inverse scale or rate parameter  $\beta = 1/\lambda$ . We employ the second since the inverse-gamma distribution formulated in terms of  $\alpha$  and  $\beta$  is a conjugate prior for likelihoods associated with normal distributions with known mean and unknown variance; see Example 4.69. For  $X \sim \text{Gamma}(\alpha, \beta)$ , the density is

$$f_X(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \quad x > 0,$$

and the expected value and variance are  $\mathbb{E}(X) = \alpha/\beta$  and  $\text{var}(X) = \alpha/\beta^2$ .

In MATLAB®, random values from a gamma distribution can be generated using the command `gamrnd.m`, which uses the first parameterization based on the shape and scale parameters  $\alpha$  and  $\lambda$ .

We point out that the one-parameter  $\chi_k^2$  distribution with  $k$  degrees of freedom is a special case of the gamma distribution with  $\alpha = \frac{k}{2}$  and  $\beta = \frac{1}{2}$ .

**Definition 4.14 (Inverse-Gamma Distribution).** If  $X$  has a gamma distribution, then  $Y = X^{-1}$  has an inverse-gamma distribution with parameters that satisfy

$$X \sim \text{Gamma}(\alpha, \beta) \Leftrightarrow Y \sim \text{Inv-gamma}(\alpha, \beta). \quad (4.9)$$

Hence the density is

$$f_Y(y; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{-(\alpha+1)} e^{-\beta/y}, \quad y > 0,$$

and the mean and variance are  $\mathbb{E}(Y) = \frac{\beta}{\alpha-1}$  for  $\alpha > 1$  and  $\text{var}(Y) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$  for  $\alpha > 2$ .

As noted in Definition 4.13 and illustrated in Example 4.69, the inverse-gamma distribution is the conjugate prior for normal likelihoods that are functions of the variance. The equivalence (4.9) can be used to generate random inverse-gamma values using the MATLAB Statistics Toolbox command `gamrnd.m`. Since  $x = \text{gamrnd}(\alpha, \lambda)$  is parameterized in terms of the scale parameter, one would employ the command  $y = \text{gamrnd}(\alpha, \beta)$ , with  $\beta = 1/\lambda$ , to generate realizations of  $Y \sim \text{Inv-gam}(\alpha, \beta)$ . A technique to construct random realizations from the inverse-gamma distribution, if `gamrnd.m` is not available, is discussed at the end of this section.

**Definition 4.15 (Inverse Chi-Squared Distribution).** The inverse chi-squared distribution is a special case of  $\text{Inv-gamma}(\alpha, \beta)$  with  $\alpha = \frac{k}{2}, \beta = \frac{1}{2}$ , so the density is

$$f_Y(y; k) = \frac{2^{-k/2}}{\Gamma(k/2)} y^{-(k/2+1)} e^{-1/2y}$$

for  $y > 0$ . This reparameterization can facilitate manipulation of conjugate families when constructing Bayesian posterior distributions.

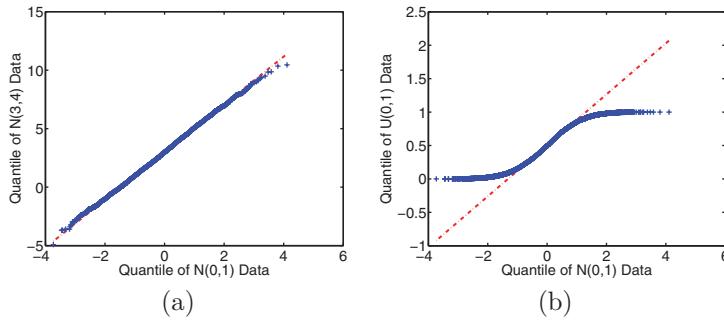
**Definition 4.16 (Beta Distribution).** The random variable  $X \sim \text{Beta}(\alpha, \beta)$  has a beta distribution if it has the density

$$f_X(x; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}$$

for  $x \in [0, 1]$ . As illustrated in Example 4.68, it is the conjugate prior for the binomial likelihood. It is observed that if  $\alpha = \beta = 1$ , the beta distribution is simply the uniform distribution which is often used to provide noninformative priors. Realizations from the beta distribution can be generated using the MATLAB command `betarnd.m`.

**Definition 4.17 (Quantile-Quantile (Q-Q) Plots).** A Q-Q plot is a graphical method for comparing data from two distributions by plotting their quantiles against each other. We will typically use this to determine the degree to which data is Gaussian, but the technique can be used to compare any distributions. If distributions are linearly related, Q-Q plots will be approximately linear. In MATLAB, Q-Q plots can be generated using the command `qqplot.m`.

To illustrate, we compare in Figure 4.4 realizations from  $N(3, 4)$  and  $\mathcal{U}(0, 1)$  distributions with data from a  $N(0, 1)$  distribution. The linearity in the first case illustrates that the two are from the same family, whereas the quantiles differ significantly in the comparison between the uniform and normal data.



**Figure 4.4.** *Q-Q plot for (a)  $N(3, 4)$  and (b)  $\mathcal{U}(0, 1)$  data as compared with  $N(0, 1)$  data.*

### Kernel Density Estimation

When estimating parameter densities in Chapter 8, we will determine the frequency with which values occur at the  $n$  points  $x_i$ . From this, we wish to compute density values  $f_X(x)$  at arbitrary points  $x$  in the sample space. We consider non-parametric estimation procedures that do not preassume a parametric form for the density.

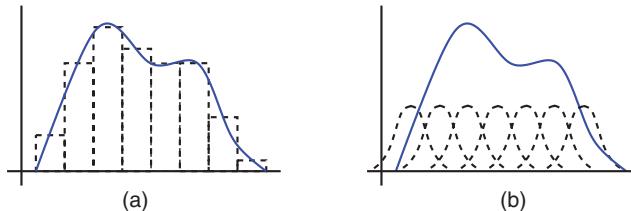
In principle, this can be achieved from a histogram of the computed values, as illustrated in Figure 4.5(a). After dividing the sample space into a set of  $N$  bins, the density is approximated using the relation

$$\tilde{f}(x) = \frac{1}{N} \frac{\text{Number of } x_i \text{ in same bin as } x}{\text{Width of bin}}.$$

Whereas this approach is simple to implement in one dimension, it has the following disadvantages: the choice of bin locations and numbers can determine the structure of the density, and it is difficult to implement in multiple dimensions.

Instead, one often employs kernel density estimation (kde) techniques in which densities are formulated in terms of known kernel functions, as shown in Figure 4.5(b). In one dimension, kernel density representations have the form

$$\tilde{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right), \quad (4.10)$$



**Figure 4.5.** (a) *Histogram and approximating density.* (b) *Kernel basis function and kernel density estimate.*

where  $K$  is a specified, symmetric pdf (e.g., normal) and  $h$  is a smoothing parameter termed the bandwidth [37, 221]. Representations in higher dimensions are analogous.

If one has access to the MATLAB Statistics Toolbox, the function `ksdensity.m` can be employed to construct kernel density estimates. Alternatively, the functions `kde.m` and `kde2d.m`, which implement automatic bandwidth selection, are available from the MATLAB Central File Exchange.

### Inverse Transform Sampling

In Definition 4.14, we discussed the use of the function `gamrnd.m`, from the MATLAB Statistics Toolbox, to construct random realizations from the inverse-gamma distribution. Here we summarize a technique to construct realizations of a general continuous random variable  $X$  with absolutely continuous distribution function  $F_X(x)$ .

For  $U \sim \mathcal{U}(0, 1)$ , we assume that we have a random number generator capable of generating realizations of  $U$ . We define the random variable  $Y = F_X^{-1}(U)$  which has the same distribution as  $X$  since

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(F_X^{-1}(U) \leq y) \\ &= P(U \leq F_X(y)) \\ &= F_X(y). \end{aligned} \tag{4.11}$$

To generate a realization  $x$  of  $X$ , we generate a realization  $u$  of  $U$  and define

$$x = F_X^{-1}(u).$$

One typically computes  $F_X^{-1}(u)$  using numerical algorithms. Even for an arbitrarily fine mesh, the cost of this procedure is typically low.

This technique can be used in lieu of calling `gamrnd.m` if the MATLAB Statistics Toolbox is unavailable.

### 4.1.2 Multiple Random Variables

For most applications, we have multiple parameters, responses, and measurements with each being represented by a random variable. We discuss here multiple random variables with associated distributions.

**Definition 4.18 (Random Vector).** Let  $X_1, \dots, X_n$  be random variables. The vector  $X : \Omega \rightarrow \mathbb{R}^n$  given by  $X = [X_1, X_2, \dots, X_n]$  is termed a random vector.

**Definition 4.19 (Joint CDF).** For a random vector  $X$ , the associated joint cdf  $F_X : \mathbb{R}^n \rightarrow [0, 1]$  is defined by

$$F_X(x_1, \dots, x_n) = P\{\omega \in \Omega | X_j(\omega) \leq x_j\}, \quad j = 1, \dots, n,$$

which is often written as  $F_X(x) = P\{X_1 \leq x_1, \dots, X_n \leq x_n\}$ .

Consider now the random variables  $X_1, \dots, X_n$  each having an expectation  $\mathbb{E}(X_i)$ . It follows immediately that

$$\mathbb{E}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i \mathbb{E}(X_i), \quad (4.12)$$

where  $a_1, \dots, a_n$  are real constants. Furthermore, if the  $n$  random variables are independent, then

$$\mathbb{E}(X_1 X_2 \cdots X_n) = \mathbb{E}(X_1) \mathbb{E}(X_2) \cdots \mathbb{E}(X_n). \quad (4.13)$$

**Definition 4.20 (Covariance and Correlation).** The covariance of random variables  $X$  and  $Y$  is the number

$$\text{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y), \quad (4.14)$$

and the correlation or Pearson correlation coefficient is

$$\rho_{XY} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}. \quad (4.15)$$

We note that if  $X$  and  $Y$  are independent, then  $\text{cov}(X, Y) = \rho_{XY} = 0$  and the random variables are uncorrelated. The converse is not true in general since the relation (4.15) quantifies only linear dependencies among random variables.

Returning to the case of  $n$  random variables, it is shown in [96] that

$$\text{var}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i^2 \text{var}(X_i) + 2 \sum_{i < j} a_i a_j \text{cov}(X_i, X_j), \quad (4.16)$$

which simplifies to

$$\text{var}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i^2 \text{var}(X_i) \quad (4.17)$$

if the random variables are pairwise uncorrelated.

**Theorem 4.21.** Let  $X_1, \dots, X_n$  be mutually independent, normally distributed, random variables with  $X_i \sim N(\mu_i, \sigma_i^2)$ , and let  $a_1, \dots, a_n$  and  $b_1, \dots, b_n$  be fixed constants. As proven in Corollary 4.6.2 of [62], it then follows that

$$Z = \sum_{i=1}^n (a_i X_i + b_i) \sim N\left(\sum_{i=1}^n (a_i \mu_i + b_i), \sum_{i=1}^n a_i^2 \sigma_i^2\right). \quad (4.18)$$

Like the univariate normal, the multivariate normal distribution plays a central role in many facets of uncertainty quantification.

**Definition 4.22 (Multivariate Normal Distribution).** The random  $n$ -vector  $X$  is said to be normally distributed with mean  $\mu = [\mu_1, \dots, \mu_n]$  and covariance matrix

$$V = \begin{bmatrix} \text{var}(X_1) & \text{cov}(X_1, X_2) & \cdots & \text{cov}(X_1, X_n) \\ \text{cov}(X_2, X_1) & \text{var}(X_2) & \cdots & \text{cov}(X_2, X_n) \\ \vdots & \vdots & & \vdots \\ \text{cov}(X_n, X_1) & \text{cov}(X_n, X_2) & \cdots & \text{var}(X_n) \end{bmatrix}, \quad (4.19)$$

designated  $X \sim N(\mu, V)$ , if the associated density is

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^n |V|}} \exp \left[ -\frac{1}{2}(x - \mu)V^{-1}(x - \mu)^T \right].$$

Here  $x = [x_1, x_2, \dots, x_n]$  and  $|V|$  is the determinant of  $V$ .

We use the next theorem when constructing proposal functions for the Metropolis algorithms detailed in Chapter 8.

**Theorem 4.23.** Let  $Y = [Y_1, \dots, Y_n]^T$  be a normally distributed random vector,  $Y \sim N(\mu, V)$ , where  $V$  is positive definite. Let  $Z \sim N(0, I_n)$ , where  $I_n$  is the  $n \times n$  identity. Then  $Y = (RZ + \mu)$ , where  $V = RR^T$  and  $R$  is a lower triangular matrix.

A proof of this theorem can be found in [96]. We note that the decomposition  $V = RR^T$  can be efficiently computed using a Cholesky decomposition.

Finally, the concepts of marginal and conditional distributions and densities will play an important role in statistical inference. We summarize the definitions for continuous random variables and refer the reader to [112, 171] for analogous definitions for discrete random variables.

**Definition 4.24 (Marginal PDF).** Let  $X_1$  and  $X_2$  be jointly continuous random variables with joint pdf  $f_X(x_1, x_2)$ . The marginal density functions of  $X_1$  and  $X_2$  are respectively given by

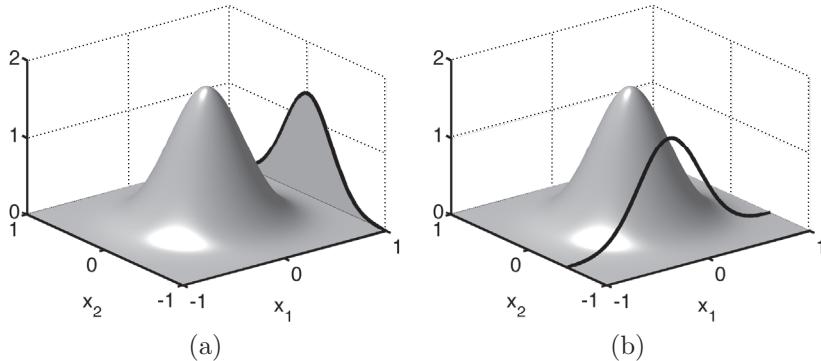
$$f_{X_1}(x_1) = \int_{\mathbb{R}} f_X(x_1, x_2) dx_2, \quad f_{X_2}(x_2) = \int_{\mathbb{R}} f_X(x_1, x_2) dx_1.$$

A representative marginal density is plotted in Figure 4.6(a). Similarly for jointly continuous random variables  $X_1, \dots, X_n$  with joint density function  $f_X(x_1, \dots, x_n)$ , the marginal pdf of  $X_1$  is

$$f_{X_1}(x_1) = \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} f_X(x_1, x_2, \dots, x_n) dx_2 \cdots dx_n.$$

**Definition 4.25 (Conditional PDF).** Let  $X_1$  and  $X_2$  be jointly continuous random variables with joint pdf  $f_X(x_1, x_2)$  and marginal pdf  $f_{X_1}(x_1)$  and  $f_{X_2}(x_2)$ . The conditional density of  $X_1$  given  $X_2 = x_2$  is

$$f_{X_1|X_2}(x_1|x_2) = \begin{cases} \frac{f_X(x_1, x_2)}{f_{X_2}(x_2)} & , \quad f_{X_2}(x_2) > 0, \\ 0 & , \quad \text{otherwise,} \end{cases}$$



**Figure 4.6.** (a) Marginal density  $f_{X_2}(x_2)$  and (b) conditional density  $f_{X_1|X_2}(x_1|x_2)$  at  $x_2 = -\frac{1}{2}$  for a normal joint density  $f_X(x_1, x_2)$  with covariance matrix  $V = 0.09I$ .

as plotted in Figure 4.6(b). We note that  $f_{X_1|X_2}(x_1|x_2)$  is a function of  $x_1$ . The definition for  $f_{X_2|X_1}(x_2|x_1)$  is analogous. Similarly, for  $n$  jointly continuous random variables  $X_1, \dots, X_n$  with joint density function  $f_X(x_1, \dots, x_n)$  and marginal density  $f_{X_1}(x_1)$ , the conditional pdf of  $X_2, \dots, X_n$  given  $X_1 = x_1$  is

$$f_{X_2, \dots, X_n | X_1}(x_2, \dots, x_n | x_1) = \frac{f_X(x_1, x_2, \dots, x_n)}{f_{X_1}(x_1)}.$$

**Definition 4.26 (iid Random Variables).** Random variables  $X_1, \dots, X_n$  are said to be independent and identically distributed (iid) with pdf  $g(x)$  if they are mutually independent and the marginal pdf  $f_{X_i}(x_i)$  for each  $X_i$  is the same function  $g(x) = f_{X_1}(x_1) = \dots = f_{X_n}(x_n)$ . The joint pdf for iid random variables is

$$f_X(x_1, \dots, x_n) = \prod_{i=1}^n f_{X_i}(x_i). \quad (4.20)$$

## 4.2 Estimators, Estimates, and Sampling Distributions

In this section, we summarize concepts pertaining to the estimation of unknown parameters through samples, observations, or measurements. In Section 4.3, we will detail specific techniques to estimate parameters in the context of model calibration. More general theory pertaining to frequentist and Bayesian inference is provided in Section 4.8 and Chapters 7 and 8.

**Definition 4.27 (Point and Interval Estimates).** Consider a fixed but unknown parameter  $q \in \mathbb{Q} \subset \mathbb{R}^p$ . A point estimate is a vector in  $\mathbb{R}^p$  that represents  $q$ . An interval estimate provides an interval that quantifies the plausible location of components of  $q$ . The mean, median, and mode of a sampling distribution are examples of point estimates, whereas confidence intervals are interval estimates.

**Definition 4.28 (Estimator and Sampling Distribution).** An estimator is a rule or procedure that specifies how to construct estimates for  $q$  based on random samples  $X_1, \dots, X_n$ . Hence the *estimator* is a random variable with an associated distribution, termed the *sampling distribution*, which quantifies attributes of the estimation process. The *estimate* is a realization of the estimator, so it is a function of the realized values  $x_1, \dots, x_n$ . An estimator is said to be *unbiased* if its mean is equal to the value of the parameter being estimated. Otherwise it is said to be biased. Two estimators that we will employ for model calibration are *ordinary least squares* and *maximum likelihood* estimators. We will also employ mean, variance, and interval estimators at various points in the discussion.

**Definition 4.29 (Statistic).** A statistic is a measurable function of one or more random variables that does not depend on unknown parameters.

**Example 4.30.** Let  $X_1, \dots, X_n$  be random variables associated with a sample of size  $n$ . Suppose we wish to estimate the population mean  $\mu$  and variance  $\sigma^2$ , which are assumed unknown. This can be accomplished using the estimators, or statistics,

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad , \quad S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2, \quad (4.21)$$

which are the sample mean and variance. We employ  $n-1$  rather than  $n$  in the expression for  $S^2$  to ensure that it is unbiased. If we additionally assume that  $X_i \sim N(\mu, \sigma^2)$ , it is illustrated in [171] that the sampling distributions for  $\bar{X}$  and  $S^2$  are

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad , \quad S^2 \sim \frac{\sigma^2}{n-1} \chi^2(n-1). \quad (4.22)$$

**Definition 4.31 (Interval Estimator and Confidence Interval).** The goal when constructing an interval estimate is to determine functions  $q_L(x)$  and  $q_R(x)$  that bound the location  $q_L(x) < q < q_R(x)$  of  $q$  based on realizations  $x = [x_1, \dots, x_n]$  of a random sample  $X = [X_1, \dots, X_n]$ . The random interval  $[q_L(X), q_R(X)]$  is termed an *interval estimator*. An interval estimator in combination with a confidence coefficient is commonly called a *confidence interval*. The confidence coefficient can be interpreted as the frequency of times, in repeated sampling, that the interval will contain the target parameter  $q$ . The  $(1-\alpha) \times 100\%$  confidence interval is the pair of statistics  $(q_L(X), q_R(X))$  such that for all  $q \in \mathbb{Q}$ ,

$$P[q_L(X) \leq q \leq q_R(X)] = 1 - \alpha. \quad (4.23)$$

As detailed on pages 418–419 of [62], it is important to note that the interval is the random quantity in (4.23), not the parameter.

**Example 4.32.** Consider a sequence of  $n$  random variables  $X_1, \dots, X_n$  from a normal distribution with known variance  $\sigma^2$  and unknown mean  $\mu$ ; that is,  $X_i \sim N(\mu, \sigma^2)$ . To determine information about the unknown mean, we consider the

sample mean  $\bar{X}$  given by (4.21) which has the sampling distribution given in (4.22). It follows that  $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$  so that

$$P\left(-2 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 2\right) \approx 0.9545$$

since 95.45% of the area of a normal distribution lies within two standard deviations of the mean. This implies that

$$P\left(\bar{X} - \frac{2\sigma}{\sqrt{n}} < \mu < \bar{X} + \frac{2\sigma}{\sqrt{n}}\right) \approx 0.9545.$$

Here  $[\bar{X} - 2\sigma/\sqrt{n}, \bar{X} + 2\sigma/\sqrt{n}]$  is an *interval estimator* for  $\mu$  where both endpoints are statistics since  $\sigma^2$  is considered known. The 95.45% confidence interval is  $[\bar{x} - 2\sigma/\sqrt{n}, \bar{x} + 2\sigma/\sqrt{n}]$ , where  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$  is the realized sample mean based on  $n$  measurements, or realizations,  $x_i$  of the random variables  $X_i$ .

**Example 4.33.** We now turn to the problem of determining the confidence interval for the mean  $\mu$  of a normal distribution when the variance  $\sigma^2$  is also unknown. To estimate  $\sigma^2$ , we employ the statistic  $S^2$  given by (4.21) which has the  $\chi^2$  distribution (4.22). We thus have

$$X = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim N(0, 1) \quad , \quad Z = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

so that the quotient

$$T = \frac{X}{\sqrt{Z/(n-1)}} = \frac{\sqrt{n}(\bar{X} - \mu)}{S}$$

has a  $t$ -distribution with  $n-1$  degrees of freedom; see Definition 4.12. To determine a  $(1-\alpha) \times 100\%$  confidence interval for a given value of  $n$ , we seek values  $a$  and  $b$  such that

$$P\left(a < \frac{\sqrt{n}(\bar{X} - \mu)}{S} < b\right) = 1 - \alpha.$$

In Figure 4.3(b), it is shown that the  $t$ -distribution is symmetric so that  $b = -a$ , which we denote by  $t_{n-1,1-\alpha/2}$  to reflect the  $n-1$  degrees of freedom and probability  $1-\alpha/2$ . It then follows that

$$P\left(\bar{X} - \frac{t_{n-1,1-\alpha/2}S}{\sqrt{n}} < \mu < \bar{X} + \frac{t_{n-1,1-\alpha/2}S}{\sqrt{n}}\right) = 1 - \alpha.$$

One can employ standard tables of  $t$ -distributions to determine  $t_{n-1,1-\alpha/2}$  given  $\alpha$  and  $n$  and thus specify the  $(1-\alpha) \times 100\%$  confidence interval  $[\bar{X} - t_{n-1,1-\alpha/2}S/\sqrt{n}, \bar{X} + t_{n-1,1-\alpha/2}S/\sqrt{n}]$ . We remind the reader that for  $\alpha = 0.05$ , this is a random interval that has a 95% chance of containing the unknown but fixed (deterministic) parameter  $\mu$ . The interval is constructed by obtaining measurements  $x_1, \dots, x_n$  and employing the realizations  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$  and  $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$  to obtain

$$\left[\bar{x} - \frac{t_{n-1,1-\alpha/2} s}{\sqrt{n}}, \bar{x} + \frac{t_{n-1,1-\alpha/2} s}{\sqrt{n}}\right].$$

We will use  $t$ -distributions in this manner in Chapter 7 to construct confidence intervals for model parameters determined using least squares estimators when  $\sigma$  is unknown and the degrees of freedom are relatively small.

## 4.3 Ordinary Least Squares and Maximum Likelihood Estimators

The process of model calibration entails estimating model parameters, and possibly initial and boundary conditions, based on measured data. More generally, the estimation of model parameters, based on observations, comprises a significant component of statistical inference which is further discussed in Section 4.8.

To motivate, consider the statistical model

$$\Upsilon_i = f(t_i, q_0) + \varepsilon_i \quad , \quad i = 1, \dots, n, \quad (4.24)$$

where  $\Upsilon_i$  are random variables whose realizations  $v_i$  are a set of  $n$  measurements from an experiment and  $f(t_i, q)$  is the parameter-dependent model response or QoI at corresponding times. The random variables  $\varepsilon_i$  account for errors between the model and measurements. Finally,  $q_0$  denotes the true, but unknown, parameter value<sup>2</sup> that we cannot measure directly but instead must infer from realizations of the random variables  $\Upsilon_i$ . We emphasize that in this context,  $q_0$  is *not* a random variable.

### 4.3.1 Ordinary Least Squares (OLS) Estimator

Consider (4.24) with the assumption that errors  $\varepsilon_i$  are iid, unbiased so that  $\mathbb{E}(\varepsilon_i) = 0$ , and have true but unknown variance  $\text{var}(\varepsilon_i) = \sigma_0^2$ . We assume that the true parameter  $q_0$  is in an admissible parameter space  $\mathcal{Q}$ , and we let  $\mathcal{Q}$  denote the corresponding sample space. As illustrated in the examples of Chapter 7, these spaces typically coincide.

The OLS estimator and estimate<sup>3</sup>

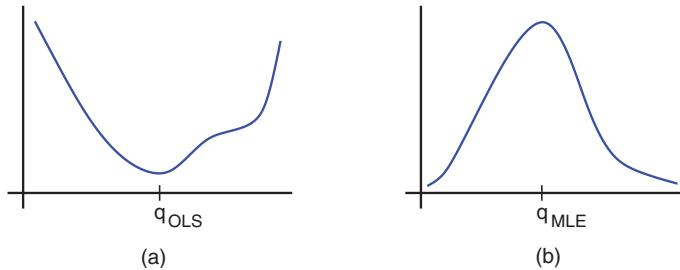
$$\begin{aligned} \hat{q}_{OLS} &= \underset{q \in \mathcal{Q}}{\operatorname{argmin}} \sum_{i=1}^n [\Upsilon_i - f(t_i, q)]^2, \\ q_{OLS} &= \underset{q \in \mathcal{Q}}{\operatorname{argmin}} \sum_{i=1}^n [v_i - f(t_i, q)]^2 \end{aligned} \quad (4.25)$$

are the random variable and realization in  $\mathbb{R}^p$  that minimize the respective sum of squares errors, as illustrated in Figure 4.7(a). Details regarding the distribution of

---

<sup>2</sup>As detailed in Section 3.4,  $\theta$  is typically employed to represent *calibration parameters* in the statistics literature, whereas other conventions are common in the mathematics, engineering, and science literature. We use the mathematics notation  $q$  due to the flexibility that it provides for representing both physical and calibration parameters as well as their admissible parameter spaces.

<sup>3</sup>The use of the notation  $\hat{q}_{OLS}$  to indicate the estimator is not universal, and many texts denote the least squares estimate by the hat-notation. Hence care must be taken to establish the convention employed in the specific text.



**Figure 4.7.** (a) Ordinary least squares solution  $q_{OLS}$  to (4.25) and (b) maximum likelihood estimate  $q_{MLE}$  given by (4.27).

$\hat{q}_{OLS}$  based on various assumptions regarding the distribution of the errors  $\varepsilon_i$  are provided in Chapter 7.

### 4.3.2 Maximum Likelihood Estimator

Maximum likelihood estimators can also be used to achieve the objective of estimating a parameter  $q$  based on random samples  $\Upsilon_1, \dots, \Upsilon_n$ .

**Definition 4.34 (Likelihood Function).** Let  $f_\Upsilon(v; q)$  be a parameter-dependent joint pdf associated with a random vector  $\Upsilon = [\Upsilon_1, \dots, \Upsilon_n]$ , where  $q \in \mathbb{Q}$  is an unknown parameter vector, and let  $v = [v_1, \dots, v_n]$  be a realization of  $\Upsilon$ . The likelihood function  $L : \mathbb{Q} \rightarrow [0, \infty)$  is defined by

$$L_v(q) = L(q|v) = f_\Upsilon(v; q), \quad (4.26)$$

where the observed sample  $v$  is fixed and  $q$  varies over all admissible parameter values. The notation  $L_v(q)$  is somewhat nonstandard, but it highlights the fact that the independent variable is  $q$ . Some authors use the notation

$$L(q) = L(q|d) = f_\Upsilon(d; q),$$

where  $d = [d_1, \dots, d_n]$  denotes the outcome from a random experiment, to reinforce this concept.

We note that because  $L$  is function of  $q$ , it is *not* a pdf, and the notation  $L(q|v)$ , while standard, should not be interpreted as a conditional pdf. If  $\Upsilon$  is discrete, then  $L_v(q)$  is the probability of obtaining the data  $v$  for a given parameter value  $q$ . For continuous  $\Upsilon$ , the fact that  $L$  is defined only to within a constant of proportionality can be combined with Riemann sum approximations of the integral to obtain a similar interpretation.

For  $n$  iid random variables, it follows from (4.20) that the likelihood function is

$$L(q|v) = \prod_{i=1}^n f_{\Upsilon_i}(v_i; q).$$

Finally, we denote the log-likelihood function by

$$\ell_v(q) = \ell(q|v) = \ln L(q|v).$$

**Example 4.35.** Consider the binomial distribution with probability of success  $q$ . The probability mass function

$$f_{\Upsilon}(v; q, n) = P(\Upsilon = v | n, q) = \binom{n}{v} q^v (1 - q)^{n-v}$$

quantifies the probability of obtaining exactly  $v = 0, 1, \dots, n$  successes in a sequence of  $n$  experiments. In this function,  $q$  and  $n$  are known and  $v$  is unknown. Although the likelihood

$$L(q|v, n) = \binom{n}{v} q^v (1 - q)^{n-v}$$

has the same functional form, the independent variable now is  $q$ , and  $v$  and  $n$  are known. Hence the likelihood function is continuous, whereas the probability mass function is discrete.

Estimates for  $q_0$  are commonly constructed by computing the value of  $q$  that maximizes the likelihood which is termed a *maximum likelihood estimate (MLE)*. For iid samples, the MLE is

$$q_{MLE} = \operatorname{argmax}_{q \in \mathbb{Q}} \prod_{i=1}^n f_{\Upsilon_i}(v_i | q).$$

To illustrate, we consider (4.24) with the assumption that errors are iid, unbiased, and normally distributed with true but unknown variance  $\sigma_0^2$  so that  $\varepsilon_i \sim N(0, \sigma_0^2)$  and hence  $\Upsilon_i \sim N(f(t_i, q_0), \sigma_0^2)$ . In this case,  $q$  and  $\sigma^2$  are both parameters and the likelihood function is

$$\begin{aligned} L(q, \sigma^2 | v) &= \prod_{i=1}^n \frac{1}{\sigma \sqrt{2\pi}} e^{-[v_i - f(t_i, q)]^2 / 2\sigma^2} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\sum_{i=1}^n [v_i - f(t_i, q)]^2 / 2\sigma^2}. \end{aligned} \tag{4.27}$$

The MLE for  $q_0$  and  $\sigma_0^2$  is

$$[q, \sigma^2]_{MLE} = \operatorname{argmax}_{\substack{q \in \mathbb{Q} \\ \sigma^2 \in (0, \infty)}} L(q, \sigma^2 | v), \tag{4.28}$$

where  $q_{MLE}$  is depicted in Figure 4.7(b).

Due to the monotonicity of the logarithm function, maximizing  $L(q, \sigma^2 | v)$  is equivalent to maximizing the log-likelihood

$$\ell(q, \sigma^2 | v) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n [v_i - f(t_i, q)]^2.$$

From a computational perspective, however, the log-likelihood is advantageous, so it is commonly employed in algorithms. For fixed  $\sigma^2$ , the condition  $\frac{\partial}{\partial q} \ell(q, \sigma^2 | v) = 0$  yields

$$\sum_{i=1}^n [v_i - f(t_i, q)] \nabla f(t_i, q) = 0, \tag{4.29}$$

where  $\nabla f$  denotes the gradient of  $f$  with respect to  $q$ . It is observed that with the assumption of iid, unbiased, normally distributed errors, the maximum likelihood solution  $q_{MLE}$  to (4.29) is the same as the least squares estimate  $q_{OLS}$  specified by (4.25). The equivalence between minimizing the sum of squares error and maximizing the likelihood will be utilized when we construct proposal functions for the MCMC techniques in Chapter 8.

In frequentist inference, the MLE  $q_{MLE}$  is the *parameter value that makes the observed output most likely*. It should *not* be interpreted as the *most likely* parameter value resulting from the data since this would require it to be a random variable which contradicts the tenets of frequentist analysis.

## 4.4 Modes of Convergence and Limit Theorems

There are several modes of convergence for sequences of random variables and distributions that are important for our discussion. We summarize the definitions and refer the reader to [62, 81, 82] for additional details, examples, and proofs of related theorems.

**Definition 4.36 (Convergence in Probability).** A sequence  $X_1, X_2, \dots$  of random variables converges in probability to a random variable  $X$ , written as  $X_n \xrightarrow{P} X$ , if for every  $\varepsilon > 0$ ,

$$\lim_{n \rightarrow \infty} P(|X_n - X| \geq \varepsilon) = 0 \quad \text{or, equivalently, } \lim_{n \rightarrow \infty} P(|X_n - X| < \varepsilon) = 1.$$

Note that  $X_1, X_2, \dots$  are typically not iid in this and the following definitions. This mode of convergence is weaker than *almost sure convergence*.

**Definition 4.37 (Almost Sure Convergence).** A sequence  $X_1, X_2, \dots$  of random variables converges almost surely to a random variable  $X$ , written as  $X_n \xrightarrow{a.s.} X$ , if for every  $\varepsilon > 0$ ,

$$P\left(\lim_{n \rightarrow \infty} |X_n - X| < \varepsilon\right) = 1.$$

Examples of sequences that converge in probability but not almost surely are provided in [62]. This is sometimes referred to as convergence with probability 1.

**Definition 4.38 (Convergence in Distribution).** Let  $X_1, X_2, \dots$  be a sequence of random variables with corresponding distributions  $F_{X_1}(x), F_{X_2}(x), \dots$ . If  $F_X(x)$  is a distribution function and

$$\lim_{n \rightarrow \infty} F_{X_n}(x) = F_X(x)$$

at all points  $x$  where  $F_X(x)$  is continuous, then  $X_n$  is said to have a limiting random variable  $X$  with distribution function  $F_X(x)$ . In this case,  $X_n$  is said to converge in distribution to  $X$ , which is often written as  $X_n \xrightarrow{D} X$ . Care must be taken when using this notation since the convergence of random variables is defined in terms

of the convergence of the distributions. Hence this mode of convergence is quite different from the previous two.

We note that almost sure convergence implies convergence in probability, which in turn implies convergence in distribution. Hence convergence in distribution is the weakest of the three concepts.

**Definition 4.39 (Consistent Estimator).** A sequence  $\hat{q}_n$  of estimators is said to be consistent, or weakly consistent, if it converges in probability to the value  $q_0$  of the parameter being estimated. In practice, we often construct estimators that are a function of the sample size  $n$ . In this case, the estimator is consistent if the sequence converges in probability to  $q_0$  as the number of samples tends to infinity.

### Law of Large Numbers and Central Limit Theorem

The *law of large numbers* and *central limit theorem* are two of the pillars of probability theory. To motivate them, we consider the problem of estimating the unknown mean  $\mu$  and variance  $\sigma^2$  of a population based on samples  $x_1, x_2, \dots$  and associated random variables  $X_1, X_2, \dots$ . An estimator for the mean is

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i, \quad (4.30)$$

so a natural question is the following: Does  $\lim_{n \rightarrow \infty} \bar{X}_n = \mu$ ? This is addressed by the strong and weak laws of large numbers.

**Theorem 4.40 (Strong Law of Large Numbers).** Let  $X_1, X_2, \dots$  be iid random variables with  $\mathbb{E}(X_i) = \mu$  and  $\text{var}(X_i) = \sigma^2 < \infty$ , and define  $\bar{X}_n$  by (4.30). Then for every  $\varepsilon > 0$ ,

$$P\left(\lim_{n \rightarrow \infty} |\bar{X}_n - \mu| < \varepsilon\right) = 1 \quad \text{or} \quad \bar{X}_n \xrightarrow{\text{a.s.}} \mu.$$

The formulation of the weak law of large numbers is similar except  $\bar{X}_n \xrightarrow{P} \mu$ . These laws are of fundamental importance since they establish that the random sample adequately represents the population in the sense that  $\bar{X}_n$  converges to the mean  $\mu$ .

Given the central role of the sample mean, it is natural to question the degree to which its sampling distribution can be established. In Example 4.30, we noted that if  $X_i \sim N(\mu, \sigma^2)$ , then  $\bar{X} \sim N(\mu, \sigma^2/n)$ . The requirement of normally distributed random variables is quite restrictive, however, so we relax this assumption and pose the same question in the context of iid random variables from an arbitrary distribution. The remarkable answer is provided by the central limit theorem.

**Theorem 4.41 (Central Limit Theorem).** Let  $X_1, \dots, X_n$  be iid random variables with  $\mathbb{E}(X_i) = \mu$  and  $\text{var}(X_i) = \sigma^2 < \infty$ . Furthermore, let  $\bar{X}_n$  be given by (4.30), and let  $G_n(x)$  denote the cdf of the random variable  $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ . Then

$$\lim_{n \rightarrow \infty} G_n = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy$$

so that the limiting distribution of  $\sqrt{n}(\bar{X}_n - \mu)/\sigma$  is a normal distribution  $N(0, 1)$ . The theorem is often expressed as

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{D} Z,$$

where  $Z \sim N(0, 1)$ .

Because

$$\bar{X}_n \xrightarrow{D} \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right), \quad (4.31)$$

$\bar{X}_n$  is approximately normal for sufficiently large  $n$ . This result is similar to that noted in Example 4.30 for  $X_i \sim N(\mu, \sigma^2)$  but with the major difference that (4.31) holds in an asymptotic sense for  $X_i$  from an *arbitrary distribution* as long as  $n$  is sufficiently large.

From a broad perspective, the combination of the law of large numbers and central limit theorem establishes that for sufficiently large  $n$ , samples are representative of the population (in the sense of the means) and the means of these samples behave asymptotically as normal distributions. The question as to how large  $n$  must be to ensure this asymptotic behavior is problem-dependent, and the assumption of approximate normality can be questionable when sample sizes are small.

We will invoke the asymptotic normality provided by the central limit theorem in Chapter 7 when constructing sampling distributions for model parameters.

## 4.5 Random Processes

In Section 4.1, we summarized the framework associated with random variables and random vectors. However, uncertainty quantification in the context of differential equation models can yield variables that exhibit time or space dependence in addition to randomness. This necessitates the discussion of stochastic or random processes and fields. We will also see that the Markov chain Monte Carlo (MCMC) techniques of Chapter 8 rely on the theory of stochastic processes.

To motivate our discussion of random processes, consider first the ODE

$$\begin{aligned} \frac{du}{dt} &= -\alpha(\omega)u, \quad t > 0, \\ u(0, \omega) &= \beta(\omega), \end{aligned} \quad (4.32)$$

where  $\alpha$  and  $\beta$  are random variables and  $\omega \in \Omega$  is an event in an underlying probability space. It was noted in Example 3.1 that for every time instance  $t$ , the random solution  $u(t, \omega)$  is an example of a stochastic or random process.

Now consider the PDE

$$\begin{aligned} \frac{\partial T}{\partial t} &= \frac{\partial}{\partial x} \left( \alpha(x, \omega) \frac{\partial T}{\partial x} \right) = f(t, x), \quad -1 < x < 1, \quad t > 0, \\ T(t, -1) &= T_\ell, \quad T(t, 1) = T_r, \quad t \geq 0, \\ T(0, x) &= T_0(x), \quad -1 \leq x \leq 1, \end{aligned} \quad (4.33)$$

which, as detailed in Example 3.5, models the flow of heat  $u$  in a structure having uncertain diffusivity  $\alpha$ . Here  $\alpha$  is an example of a random field and the solution  $T(t, x, \omega)$  is random for all pairs  $(t, x)$  of independent variables.

**Definition 4.42 (Stochastic Process).** A stochastic or random process is an indexed collection

$$X = \{X_t, t \in \mathbb{T}\} = \{X(t), t \in \mathbb{T}\}$$

of random variables, all of which are defined on the same probability space  $(\Omega, \mathcal{F}, P)$ . The index set is typically assumed to be totally ordered and often is taken to be time. Taking  $\mathbb{T}$  to be a subset of consecutive integers yields a discrete random process, whereas taking  $\mathbb{T}$  to be an interval of real numbers yields a continuous process.

The random solution  $u(t, \omega)$  to (4.32) is an example of a continuous random process. In the next section, we will devote significant discussion to Markov chains, which are discrete random processes, since they are central to the Metropolis methods used in the Bayesian analysis of Chapter 8 to quantify parameter densities.

Other ordered index sets can be considered, including spatial points or intervals. However, the ordering in dimensions greater than one is complicated, so we employ the terminology stochastic or random fields for spatially varying quantities.

A stochastic process can be interpreted three ways.

- (i)  $X$  is a function on  $\mathbb{T} \times \Omega$  with the realization  $X_t(\omega)$  for  $t \in \mathbb{T}$  and  $\omega \in \Omega$ .
- (ii) For fixed  $t \in \mathbb{T}$ ,  $X_t$  is a random variable.
- (iii) For an outcome  $\omega \in \Omega$ , the realization  $X_t(\omega)$  is a function of  $t$  that is often called the sample path or trajectory associated with  $\omega$ .

We note that continuous stochastic processes are infinite-dimensional and extreme care must be taken when extending finite-dimensional convergence results to these cases. The following class of random processes is important since the concepts of mean, covariance, and correlation functions are well defined for these processes.

**Definition 4.43 (Second-Order Stochastic Process).** A second-order stochastic process is one for which  $\mathbb{E}(X_t^2) < \infty$  for all  $t \in \mathbb{T}$ .

For second-order random processes, the random variable concepts of mean and covariance can be directly extended using the interpretation (ii). Specifically, the expectation and covariance functions of  $X$  are defined as

$$\begin{aligned} \mu(t) &= \mathbb{E}(X_t) , \quad t \in \mathbb{T}, \\ C(t, s) &= \text{cov}(X_t, X_s) = \mathbb{E}[(X_t - \mu(t))(X_s - \mu(s))] , \quad t, s \in \mathbb{T}. \end{aligned} \tag{4.34}$$

Hence  $\mu(t)$  quantifies the centrality of sample paths, whereas  $C(t, s)$  quantifies their variability about  $\mu(t)$ .

**Definition 4.44 (Gaussian Process).** A Gaussian process (GP) is a continuous-time stochastic process  $X$  such that all finite-dimensional vectors  $X_t = [X_{t_1}, \dots, X_{t_n}]$  have a multivariate normal distribution; that is,

$$X_t \sim N(\mu(t), C(t)),$$

where  $t = [t_1, \dots, t_n]$ ,  $\mu(t) = [\mathbb{E}(X_{t_1}), \dots, \mathbb{E}(X_{t_n})]$ , and  $[C(t)]_{ij} = \text{cov}(X_{t_i}, X_{t_j})$  for all  $1 \leq i, j \leq n$ . A GP is thus a probability distribution for a function.

The concept of stationarity is important in the theory of Markov chains since it provides criteria specifying when MCMC methods can be expected to converge to posterior distributions for parameters. We consider this in the context of a discrete index set  $\mathbb{T}$  but note that a similar definition holds for continuous index sets.

**Definition 4.45 (Stationary Random Process).** The random process  $X$  is said to be stationary if, for any  $t_1, t_2, \dots, t_n \in \mathbb{T}$  and  $s$  such that  $t_1+s, \dots, t_n+s \in \mathbb{T}$ , the random vectors  $[X_{t_1}, \dots, X_{t_n}]$  and  $[X_{t_1+s}, \dots, X_{t_n+s}]$  have the same distribution. For a stationary process,  $\mu(t)$  is constant for all  $t = [t_1, \dots, t_n]$  and  $C(t, s) = C(t-s)$  is a function only of the time difference  $|t - s|$ .

**Definition 4.46 (Autoregressive (AR) Models).** An AR(1) process, or time series,  $X$  satisfies

$$X_t = \rho_1 X_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma^2), \quad (4.35)$$

where  $\rho_1$  is a parameter. If  $|\rho_1| < 1$ , the process is said to be wide-sense stationary. In this case,  $\mathbb{E}(X_t) = \mathbb{E}(X_{t-1})$  so that  $\mathbb{E}(X_t) = 0$  and  $\text{var}(X_t) = \mathbb{E}(X_t^2) = \rho_1^2 \mathbb{E}(X_{t-1}^2) + \sigma^2$  so that  $\text{var}(X_t^2) = \frac{\sigma^2}{1-\rho_1^2}$ . We note that an AR(1) process smooths the output in the sense of a low-pass filter.

An AR( $p$ ) process satisfies

$$X_t = \sum_{k=1}^p \rho_k X_{t-k} + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma^2). \quad (4.36)$$

We note that AR( $p$ ) processes are a type of GP.

**Definition 4.47 (Random Field).** The concept of a random field generalizes that of a random process by allowing indices that are vector-valued or points on a manifold. Specifically, a random field is a collection

$$X = \{X_x, x \in \mathcal{X}\}$$

of random variables indexed by elements  $x$  in a topological space  $\mathcal{X}$ . For our applications, we will employ random fields to quantify uncertain spatially varying parameters such as  $\alpha(x, \omega)$  in (4.33).

For the definitions of random processes and random fields, we have considered indexed families of random variables which, for fixed values of the index, map  $\Omega$

to  $\mathbb{R}$ . When describing Markov processes, however, it is advantageous to generalize this concept to include random variables that map into a state space  $S$ . This is established in the following definitions.

**Definition 4.48 ( $S$ -Valued Random Variable).** Let  $S$  be a finite or countable set termed the *state space*. An  $S$ -valued random variable is a function  $X : \Omega \rightarrow S$  such that  $\{\omega \in \Omega | X(\omega) \leq x\} \in \mathcal{F}$  for each  $x \in S$  if there is an ordering on  $S$ . Note that this is exactly Definition 4.3 if  $S = \mathbb{R}$ .

**Definition 4.49.** A random process  $X$  is said to have a state space  $S$  if  $X_t$  is an  $S$ -valued random variable for each  $t \in \mathbb{T}$ .

## 4.6 Markov Chains

In Chapter 8, we will employ Markov chain Monte Carlo (MCMC) methods to construct posterior densities for model parameters. We summarize here the fundamental properties of Markov chains necessary for that development.

Broadly stated, a stochastic process is said to satisfy the Markov property if the probability of future states is dependent only on the present state rather than the sequence of past events that precede it. This is completely analogous to the state space concept of modeling in which a system is defined in terms of state variables that uniquely define the behavior at time  $t$ . When combined with dynamics encompassed in the model, the future state behavior can be completely defined. Both Markov processes and state space models are memoryless in the sense that the past history is not required to make future predictions. Whereas Markov processes can be defined for both continuous and discrete index sets  $\mathbb{T}$ , we focus solely on the latter since it provides the setting necessary for MCMC analysis. Discrete-time Markov processes are usually called *Markov chains*, although some authors also use this designation for continuous-time processes.

**Definition 4.50 (Markov Chain).** A Markov chain is a sequence of  $S$ -valued random variables

$$X = \{X_i, i \in \mathbb{Z}\}$$

that satisfy the Markov property that  $X_{n+1}$  depends only on  $X_n$ ; that is,

$$P(X_{n+1} = x_{n+1} | X_0 = x_0, \dots, X_n = x_n) = P(X_{n+1} = x_{n+1} | X_n = x_n), \quad (4.37)$$

where  $x_i$  is the state of the chain at time  $i$ .

A Markov chain is characterized by three components: a state space  $S$ , an initial distribution  $p^0$ , and a transition or Markov kernel. As indicated in Definition 4.48, the state space is the range of all random variables, so it is the set of all possible realizations. We assume a finite number  $k$  of discrete states, so  $S = \{x_1, \dots, x_k\}$ . The initial distribution quantifies the starting configuration for the chain, whereas the transition kernel quantifies the probability of transitioning from state  $x_i$  to  $x_j$ , so it establishes how the chain evolves. For our discussion, we

assume that the transition probabilities are the same for all time, which yields a *homogeneous* Markov chain.

We let  $p_{ij}$  denote the probability of moving from  $x_i$  to  $x_j$  in one step so that

$$p_{ij} = P(X_{n+1} = x_j | X_n = x_i).$$

The resulting transition matrix is

$$P = [p_{ij}] \quad , \quad 1 \leq i, j \leq k.$$

We will also be interested in the probability of transitioning between states in  $m$ -steps, which we denote by

$$p_{ij}^{(m)} = P(X_{n+m} = x_j | X_n = x_i)$$

with the corresponding  $m$ -step transition matrix

$$P_m = \left[ p_{ij}^{(m)} \right] = P^m.$$

The initial density, which is often termed mass when it is discrete, is given by

$$p^0 = [p_1^0, \dots, p_k^0],$$

where  $p_i^0 = P(X_0 = x_i)$ . Because  $p^0$  and  $P$  contain probabilities, their entries are nonnegative and the elements of  $p^0$  and rows of  $P$  must sum to unity. Matrices satisfying the property are termed *row-stochastic matrices*.

Given an initial distribution and transition kernel, the distribution after 1 step is  $p^1 = p^0 P$  and

$$p^n = p^{n-1} P = p^0 P^n$$

after  $n$  steps. We illustrate these concepts in the next example.

**Example 4.51.** Various studies have indicated that factors such as weather, injuries, and unquantifiable concepts such as hitting streaks lend a random nature to baseball [7]. We assume that a team that won its previous game has a 70% chance of winning their next game and 30% chance of losing, whereas a losing team wins 40% and loses 60% of their next games. Hence the probability of winning or losing the next game is conditioned on a team's last performance.

This yields the two-state Markov chain illustrated in Figure 4.8, where

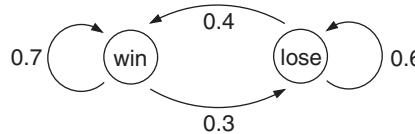
$$S = \{\text{win}, \text{lose}\}.$$

The resulting transition matrix is

$$P = \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix}.$$

There are a large number of teams in major league baseball, so

$$p^0 = [p_w^0, p_\ell^0], \quad p_w^0 + p_\ell^0 = 1$$



**Figure 4.8.** Markov chain quantifying the probability of winning or losing based on the last performance.

is the percentage of teams who won and lost their last games. To illustrate, we take  $p^0 = [0.8, 0.2]$ . We assume a schedule in which teams play at different times, so  $p_w^0$  and  $p_\ell^0$  do not both have to be 0.5.

The percentage of teams who win/lose their next game is given by

$$\begin{aligned} p^1 &= [0.8, 0.2] \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix} \\ &= [0.64, 0.36], \end{aligned}$$

so the distribution after  $n$  games is

$$p^n = [0.8, 0.2] \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix}^n.$$

The distributions for  $n = 0, \dots, 10$  are compiled in Table 4.1. These numerical results indicate that the distribution is limiting to a stationary value.

For this example, we can explicitly compute a limiting distribution  $\pi$  by solving the constrained relation

$$\begin{aligned} \pi &= \pi P, \quad \sum \pi_i = 1 \\ \Rightarrow [\pi_{win}, \pi_{lose}] \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix} &= [\pi_{win}, \pi_{lose}], \quad \pi_{win} + \pi_{lose} = 1 \end{aligned}$$

to obtain

$$\pi = [0.5714, 0.4286].$$

In general, however, we cannot solve explicitly for a stationary value and instead must establish the manner in which  $p^n$  limits to  $\pi$ . We next discuss the nature of this convergence and summarize criteria that guarantee the existence of a unique limiting value.

| $n$ | $p^n$            | $n$ | $p^n$            | $n$ | $p^n$            |
|-----|------------------|-----|------------------|-----|------------------|
| 0   | [0.8000, 0.2000] | 4   | [0.5733, 0.4267] | 8   | [0.5714, 0.4286] |
| 1   | [0.6400, 0.3600] | 5   | [0.5720, 0.4280] | 9   | [0.5714, 0.4286] |
| 2   | [0.5920, 0.4080] | 6   | [0.5716, 0.4284] | 10  | [0.5714, 0.4286] |
| 3   | [0.5776, 0.4224] | 7   | [0.5715, 0.4285] |     |                  |

**Table 4.1.** Iteration and distributions for Example 4.51.

As detailed in Section 4.4, it does not make sense to directly consider limits  $\lim_{n \rightarrow \infty} X_n$  of random variables. Instead, we consider the limit

$$\lim_{n \rightarrow \infty} p^n = \pi,$$

which is convergence in distribution. We note that if this limit exists, it must satisfy

$$\pi = \lim_{n \rightarrow \infty} p^0 P^n = \lim_{n \rightarrow \infty} p^0 P^{n+1} = \left( \lim_{n \rightarrow \infty} p^0 P^n \right) P = \pi P.$$

**Definition 4.52 (Stationary Distribution).** For a Markov chain with transition kernel  $P$ , distributions  $\pi$  that satisfy

$$\pi = \pi P \tag{4.38}$$

are termed equilibrium or stationary distributions of the chain. In a measure theoretic framework,  $\pi$  is an invariant measure.

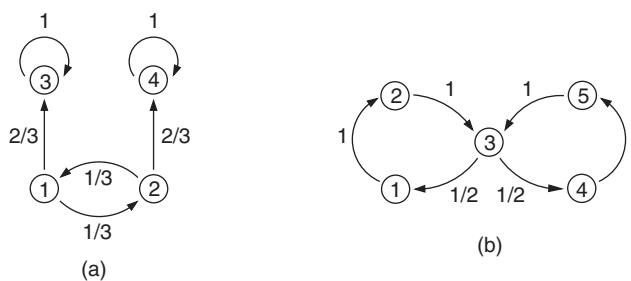
For every finite Markov chain, there exists at least one stationary distribution. However, it may not be unique and it may not be equal to  $\lim_{n \rightarrow \infty} p^n$ . Criteria necessary to establish a unique limiting distribution  $\pi = \lim_{n \rightarrow \infty} p^n$  are motivated by the following definitions and examples.

**Definition 4.53 (Irreducible Markov Chain).** A Markov chain is irreducible if any state  $x_j$  can be reached from any other state  $x_i$  in a finite number of steps; that is,  $p_{ij}^{(m)} > 0$  for all states in finite  $m$ . Otherwise it is reducible.

**Example 4.54.** Consider the Markov chain depicted in Figure 4.9(a) with the transition matrix

$$P = \begin{bmatrix} 0 & \frac{1}{3} & \frac{2}{3} & 0 \\ \frac{1}{3} & 0 & 0 & \frac{2}{3} \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

The chain is clearly reducible since  $p_{3j} = 0$  for  $j = 1, 2, 4$ . Furthermore, it is easy to verify that  $\pi = [0, 0, 1, 0]$  and  $\pi = [0, 0, 0, 1]$  are both stationary distributions. The property of irreducibility is required to guarantee that  $\pi$  is unique.



**Figure 4.9.** (a) Reducible chain for Example 4.54 and (b) periodic chain for Example 4.56.

**Definition 4.55 (Periodic Markov Chain).** A Markov chain is periodic if parts of the state space are visited at regular intervals. The period  $k$  is defined as

$$\begin{aligned} k &= \gcd \left\{ m \mid \pi_{ii}^{(m)} > 0 \right\} \\ &= \gcd \{ m \mid P(X_{n+m} = x_i | X_n = x_i) > 0 \}. \end{aligned}$$

The chain is aperiodic if  $k = 1$ .

**Example 4.56.** The Markov chain depicted in Figure 4.9(b) with the transition matrix

$$P = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

has the unique stationary distribution  $\pi = [1/6, 1/6, 1/3, 1/6, 1/6]$ . It is established in Exercise 4.8 that if  $p^0 = [1, 0, 0, 0, 0]$ , then  $p^3 = p^6 = p^9 = \dots = p^0$ , so the period is  $k = 3$ . Because mass cycles through the chain at a regular interval, it does not converge, so  $\lim_{n \rightarrow \infty} p^n$  does not exist. Furthermore, it is demonstrated in Exercise 4.9 that if the limit of a periodic chain exists for one initial distribution, other distributions can yield different limits. Hence aperiodicity is required to guarantee that the limit exists.

For infinite chains, one must additionally include conditions regarding the persistence or recurrence of states. However, we will focus on finite Markov chains for which it can be shown that if the chain is irreducible, all states are positive persistent [121].

Before providing a theorem that establishes the convergence  $\lim_{n \rightarrow \infty} p^n = \pi$ , we summarize relevant results from matrix theory.

**Definition 4.57.** A  $k \times k$  matrix  $A$  is

- (i) nonnegative, denoted by  $A \geq 0$ , if  $a_{ij} \geq 0$  for all  $i, j$ , and
- (ii) strictly positive, denoted by  $A > 0$ , if  $a_{ij} > 0$  for all  $i, j$ .

**Theorem 4.58 (Perron–Frobenius).** Let  $A$  be a  $k \times k$  nonnegative matrix such that  $A^m > 0$  for some  $m \geq M$ . Then

- (i)  $A$  has a positive eigenvalue  $\lambda_0$  with corresponding left eigenvector  $x_0$  where the entries of  $x_0$  are positive;
- (ii) if  $\lambda \neq \lambda_0$  is any other eigenvalue of  $A$ , then  $|\lambda| < \lambda_0$ ;
- (iii)  $\lambda_0$  has geometric and algebraic multiplicity 1.

There are several statements of the Perron–Frobenius theorem, and details and proofs can be found in [121, 130, 220].

**Theorem 4.59.** For all finite stochastic matrices  $P$ , the largest eigenvalue is  $\lambda_0 = 1$ .

See [121] for a proof of this theorem.

**Theorem 4.60.** Let  $P$  be a finite transition matrix for an irreducible aperiodic Markov chain. Then there exists  $M \geq 1$  such that  $P^m > 0$  for all  $m \geq M$ .

Further details are provided in [121], and the theorem is illustrated in Exercise 4.10. The following theorem establishes the convergence of the Markov chain.

**Theorem 4.61.** Every finite, homogeneous Markov chain that is irreducible and aperiodic, with transition matrix  $P$ , has a unique stationary distribution  $\pi$ . Moreover, chains converge in the sense of distributions,  $\lim_{n \rightarrow \infty} p^n = \pi$ , for every initial distribution  $p^0$ .

**Proof.** It follows from Theorems 4.58, 4.59, and 4.60 that the largest eigenvalue of  $P$  is  $\lambda_0 = 1$ , which has multiplicity 1. There is thus a unique left eigenvector  $\pi$  that satisfies  $\pi P = \pi$  and  $\sum \pi_i = 1$ . To establish the convergence, we first consider the eigendecomposition

$$UPV = \Lambda = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & \\ \vdots & & \ddots & \vdots \\ 0 & & \cdots & \lambda_k \end{bmatrix},$$

where  $1 > |\lambda_2| \geq \cdots \geq |\lambda_k|$  and  $V = U^{-1}$ . It follows that

$$\lim_{n \rightarrow \infty} P^n = \lim_{n \rightarrow \infty} V \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & \lambda_2^n & & \\ \vdots & & \ddots & \vdots \\ 0 & & \cdots & \lambda_k^n \end{bmatrix} U = V \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & & \\ \vdots & & \ddots & \vdots \\ 0 & & \cdots & 0 \end{bmatrix} U.$$

Furthermore, we observe that  $UP = \Lambda U$  implies that

$$\begin{bmatrix} \pi_1 & \cdots & \pi_k \\ \vdots & & \vdots \\ u_{k1} & \cdots & u_{kk} \end{bmatrix} \begin{bmatrix} & & \\ & P & \\ & & \end{bmatrix} = \begin{bmatrix} 1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix} \begin{bmatrix} \pi_1 & \cdots & \pi_k \\ \vdots & & \vdots \\ u_{k1} & \cdots & u_{kk} \end{bmatrix}$$

and  $V = U^{-1}$  implies that

$$UV = \begin{bmatrix} \pi_1 & \cdots & \pi_k \\ \vdots & & \vdots \\ u_{k1} & \cdots & u_{kk} \end{bmatrix} \begin{bmatrix} 1 & \cdots & v_{1k} \\ \vdots & & \vdots \\ 1 & \cdots & v_{kk} \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 1 \end{bmatrix}$$

since  $\sum \pi_i = 1$ . This establishes that the first column of  $V$  is all ones. Finally

$$\begin{aligned} \lim_{n \rightarrow \infty} p^n &= \lim_{n \rightarrow \infty} p^0 P^n \\ &= \lim_{n \rightarrow \infty} [p_1^0, \dots, p_k^0] \begin{bmatrix} 1 & \cdots & v_{k1} \\ \vdots & & \vdots \\ 1 & \cdots & v_{kk} \end{bmatrix} \begin{bmatrix} 1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_k \end{bmatrix} \begin{bmatrix} \pi_1 & \cdots & \pi_k \\ \vdots & & \vdots \\ u_{k1} & \cdots & u_{kk} \end{bmatrix} \\ &= [p_1^0 \ \cdots \ p_k^0] \begin{bmatrix} 1 & \cdots & v_{k1} \\ \vdots & & \vdots \\ 1 & \cdots & v_{kk} \end{bmatrix} \begin{bmatrix} 1 & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix} \begin{bmatrix} \pi_1 & \cdots & \pi_k \\ \vdots & & \vdots \\ u_{k1} & \cdots & u_{kk} \end{bmatrix} \\ &= [\pi_1, \dots, \pi_k] \\ &= \pi, \end{aligned}$$

thus establishing the required convergence. ■

Theorem 4.61 establishes that finite Markov chains which are irreducible and aperiodic will converge to a stationary distribution  $\pi$ . However, it is often difficult or impossible to solve for  $\pi$  using the relations  $\pi P = \pi$  subject to  $\sum \pi_i = 1$ . The detailed balance condition provides an alternative that is straightforward to implement in MCMC methods where the goal is to construct Markov chains whose stationary distribution  $\pi$  is the posterior distribution for parameters.

**Definition 4.62 (Detailed Balance).** A chain with transition matrix  $P = [p_{ij}]$  and distribution  $\pi = [\pi_1, \dots, \pi_k]$  is *reversible* if the detailed balance condition

$$\pi_i p_{ij} = \pi_j p_{ji} \tag{4.39}$$

is satisfied for all  $i, j$ . Since

$$\sum_i \pi_i p_{ij} = \sum_i \pi_j p_{ji} = \pi_j \sum_j p_{ji} = \pi_j,$$

it follows immediately that  $\pi P = \pi$  so that reversibility implies stationarity. Hence if the chains are irreducible and aperiodic, they will uniquely limit to this specified stationary distribution. In Chapter 8, we use the Metropolis algorithm to construct chains that satisfy (4.39) and converge to the posterior density.

## 4.7 Random versus Stochastic Differential Equations

We briefly illustrate here the difference between random differential equations, which we consider throughout this text, and stochastic differential equations. This is done in part to allay a growing trend in the uncertainty quantification community to treat these terms as synonymous when in fact they are distinctly different and they require completely different techniques for analysis and approximation.

**Definition 4.63 (Random Differential Equation).** Random differential equations are those in which random effects are manifested in parameters, initial or boundary conditions, or forcing conditions that are regular (e.g., continuous) with respect to time and space. An example is the ODE

$$\frac{dz}{dt} = a(\omega)z + b(t, \omega), \\ z(0) = z_0(\omega),$$

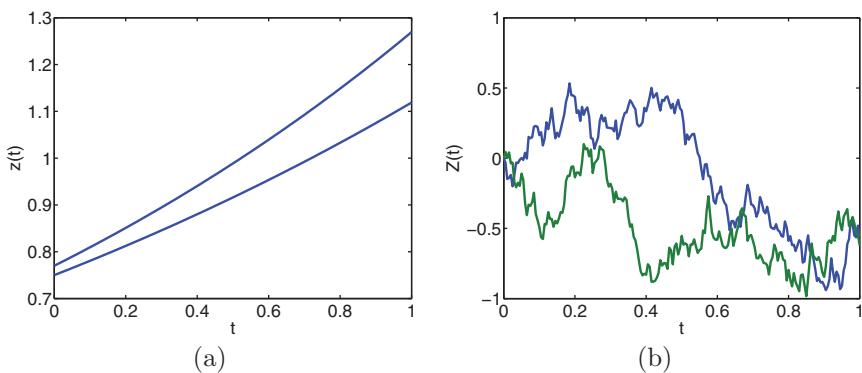
which has the solution

$$z(t, \omega) = e^{a(\omega)t} \left[ z_0(\omega) + \int_0^t e^{-a(\omega)s} b(s, \omega) ds \right].$$

We emphasize that  $b(t, \omega)$  is a random process, as defined in Definition 4.42, with the additional requirement that for an outcome  $\omega \in \Omega$ , the sample path  $b(t, \omega)$  is taken to be smooth, e.g., in  $C[0, t_f]$ . This guarantees that sample paths of the solution  $z(t, \omega)$  are at least differentiable functions, as illustrated in Figure 4.10.

In summary, for each realization of  $\omega$ , random differential equations are analyzed and solved sample path by sample path using the theory of standard differential equations [91, 138, 231]. The goal pursued in Chapters 9 and 10 is to determine distributions or uncertainty bounds for  $z(t, \omega)$  based on those of inputs such as parameters or initial and boundary conditions.

**Definition 4.64 (Stochastic Differential Equation).** The role of uncertainty is fundamentally different in stochastic differential equations (SDEs). In this case, the differential equations are forced by an irregular process such as a Wiener process or Brownian motion. SDEs are typically written symbolically in terms of stochastic differentials, but they are interpreted as Itô or Stratonovich stochastic integrals. For example, fluctuations in  $Z(t)$  due to a Wiener process  $W$  could be formulated



**Figure 4.10.** Realizations of (a) a random differential equation and (b) sample paths of an SDE.

as

$$dZ(t) = -aZ(t)dt + bdW(t),$$

which is interpreted as

$$Z(t) = Z_0 - \int_0^t aZ(s)ds + \int_0^t bdW(s),$$

where the second integral is an Itô stochastic integral.

As illustrated in Figure 4.10, the solutions of SDEs exhibit nondifferentiable sample paths due to the irregularity of the driving Wiener process. We do not further consider SDEs in this text but rather include this definition to delineate them from random differential equations. The reader is referred to [91, 138] for further details about SDEs.

## 4.8 Statistical Inference

The goal in statistical inference is to deduce the structure of, or make conclusions about, a phenomenon based on observed data. This often involves the determination of an unknown distribution based on observed data in which case the problem of statistical inference can be stated as follows. Given a set

$$S = \{x_1, \dots, x_n\}, \quad x_j \in \mathbb{R}^N,$$

of observed realizations of a random variable  $X$ , we want to infer the underlying probability distribution that produces the data  $S$ .

Statistical inference can be roughly categorized as being *parametric* or *nonparametric* in nature. In parametric approaches, one assumes that the underlying distributions can be adequately described in terms of a parametric relation having a relatively small number of parameters, e.g., mean and variance. The inference problem is to estimate those parameters or the distribution of those parameters. This approach has the advantage of a typically small number of parameters but the disadvantage of limited accuracy if the assumed functional relation is incorrect. In nonparametric approaches, one does not presuppose a functional form but instead describes or constructs the distribution based solely on properties of the observations. This avoids errors associated with incorrect parametric relations but requires that some structure be imposed on algorithms to ensure that reasonable distributions are determined.

### 4.8.1 Frequentist versus Bayesian Inference

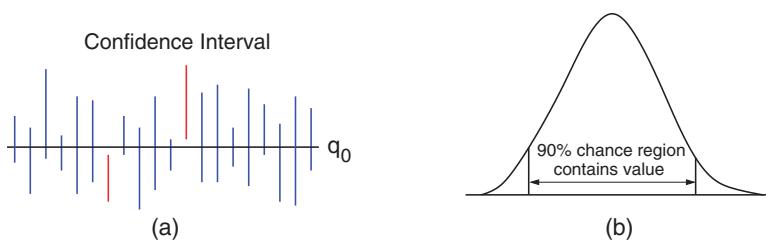
Frequentist and Bayesian inference differ in the underlying assumptions made regarding the nature of probabilities, models, parameters, and confidence intervals. As detailed in [30], each approach, or a hybrid combination of the two, is advantageous for certain problems or applications. Hence it is necessary that scientists understand both.

From a frequentist perspective, probabilities are defined as the frequencies with which an event occurs if the experiment is repeated a large number of times. Hence they are objective and are not updated as data is acquired. Parameters are considered to be unknown but fixed; hence they are deterministic. To statistically establish confidence in the estimation process, one constructs estimators, such as OLS estimators or maximum likelihood estimators, to estimate the parameters in the manner detailed in Section 4.3. Based on either the assumption of normality for the errors or asymptotic theory resulting from the central limit theorem, one can then construct sampling distributions and confidence intervals for the parameter estimators.

The interpretation of confidence intervals in the framework of frequentist inference is often a source of confusion. As detailed in Definition 4.31, a 90% confidence interval has the following interpretation: in repeated procedures, 90% of realized intervals would include the true parameter  $q_0$ . In model calibration, this means that if the estimation procedure is repeated a large number of times using data having the same error statistics, and a 90% interval estimate is computed each time, then 90% of the intervals would include  $q_0$ , as illustrated in Figure 4.11(a). The sampling distribution and confidence intervals thus quantify the accuracy and variability of the estimation procedure rather than providing a density for the parameter. Hence they do not provide a direct measure of parameter uncertainty.

Because parameters are fixed, but unknown, values in this framework, it cannot be directly applied to obtain parameter densities that can be propagated through models to quantify model uncertainty. In some problems, the sampling distributions may be similar to parameter distributions, but this needs to be verified either experimentally or using Bayesian analysis. This is discussed in more detail in Chapter 7.

Probabilities are treated as possibly subjective in the Bayesian framework, and they can be updated to reflect new information. Moreover, they are considered to be a distribution rather than a single frequency value. Similarly, parameters are considered to be random variables with associated densities and the solution of the parameter estimation problem is the posterior probability density. The Bayesian perspective is thus natural for model uncertainty quantification since it provides densities that can be propagated through models. The interpretation of interval estimates, termed credible intervals, is also natural in the Bayesian framework.



**Figure 4.11.** Interpretation of a (a) frequentist 90% confidence interval and (b) Bayesian 90% credible interval.

**Definition 4.65 (Credible Interval).** The  $(1 - \alpha) \times 100\%$  credible interval is that which has a  $(1 - \alpha) \times 100\%$  chance of containing the expected parameter. A 90% credible interval is illustrated in Figure 4.11(b).

We next provide details regarding Bayesian inference to provide the background necessary for Chapter 8.

### 4.8.2 Bayesian Inference

Bayesian inference is based on the supposition that probabilities, and more generally our state of knowledge regarding an observed phenomenon, can be updated as additional information is obtained. In the context of parametric models, parameters are treated as random variables having associated densities.

Bayes' formula

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

for probabilities provides a natural genesis for Bayesian inference. In the context of parameters  $Q = [Q_1, \dots, Q_p]$ <sup>4</sup> that are quantified based on observations  $v = [v_1, \dots, v_n]$ , one employs the relation

$$\pi(q|v) = \frac{\pi(v|q)\pi_0(q)}{\pi_Y(v)}, \quad (4.40)$$

where  $\pi_0(q)$  and  $\pi(q|v)$  respectively denote the prior and posterior densities,  $\pi(v|q)$  is a likelihood, and the marginal density  $\pi_Y(v)$  is a normalization factor. Here  $q = Q(\omega)$  denotes realizations of  $Q$ . The subscripts that indicate specific random variables are typically dropped from the prior and posterior in Bayesian analysis.

The prior density  $\pi_0(q)$  quantifies any prior knowledge that may be known about the parameter before data is taken into account. For example, one might have prior information based on similar previous models, data that is similar to previous data, or initial parameter densities that have been determined through other means, such as related experiments.

It is common in model calibration, however, that one does not have such prior information, so one uses instead what is termed a *noninformative prior*. A common choice of noninformative prior is the uniform density, or unnormalized uniform, posed on the parameter support. For example, one might employ

$$\pi_0(q) = \chi_{[0,\infty)}(q)$$

for a positive parameter. This choice is improper in the sense that the integral of  $\pi_0(q)$  is unbounded. It is recommended that a noninformative prior be used unless good previous information is known since it is shown in Example 4.66 that incorrect prior information can degrade (4.40) far more than a noninformative prior.

---

<sup>4</sup>Readers are referred to Section 3.4 for discussion regarding notation conventions, for parameters, in the mathematics, statistics, engineering, and science literature. For example,  $\theta$  is typically used to denote calibration parameters in statistics whereas  $q$  is commonly employed in the mathematics literature.

In “empirical Bayes” inference, one also encounters data-dependent priors in which priors estimated using frequentist techniques such as maximum likelihood are employed in the Bayesian model. It is argued in [35] that this double use of data is problematic with small sample sizes and is at odds with the tenets of Bayesian analysis.

The term  $\pi(v|q)$ , which is a function of  $q$  with  $v$  fixed, quantifies the likelihood  $L(q|v)$  of observing  $v$  given parameter realizations  $q$  as detailed in Section 4.3.2. We will illustrate various choices for the likelihood function in the examples at the end of this section and at the beginning of Chapter 8. The joint density is given by

$$\pi(q, v) = \pi(v|q)\pi_0(q)$$

and is normalized to unity by the marginal density function  $\pi_Y(v)$  of all possible observations.

Finally, the posterior density  $\pi(q|v)$  quantifies the probability of obtaining parameters  $q$  given observations  $v$ . It is the posterior density that we will be estimating using the Bayesian parameter estimation techniques of Chapter 8, and we point out that the data directly informs the posterior only through the likelihood. Finally, representation of  $\pi_Y(v)$  as the integral over all possible joint densities yields the Bayes relation

$$\pi(q|v) = \frac{\pi(v|q)\pi_0(q)}{\int_{\mathbb{R}^p} \pi(v|q)\pi_0(q)dq} \quad (4.41)$$

commonly employed for model calibration and data assimilation.

A significant issue, which will be discussed in detail in Chapter 8, concerns the evaluation of the normalizing integral. It can be analytically evaluated only in special cases, and classical tensored quadrature techniques are effective only in low dimensions; e.g.,  $p \leq 4$ . This has spawned significant research on high-dimensional quadrature techniques, including adaptive sparse grids for moderate dimensionality and Monte Carlo techniques for high dimensions; see Chapter 11.

**Example 4.66.** To illustrate (4.41) in a setting where the posterior density can be computed explicitly, we consider the results from tossing a possibly biased coin. The random variable

$$\Upsilon_i(\omega) = \begin{cases} 0 & , \quad \omega = T, \\ 1 & , \quad \omega = H, \end{cases}$$

represents the result from the  $i^{th}$  toss, and the parameter  $q$  is the probability of getting heads. We now consider the probability of obtaining  $N_1$  heads and  $N_0$  tails in a series of  $N = N_0 + N_1$  flips of the coin.

Because coin flips are independent events with only two possible outcomes, the likelihood of observing a sequence  $v = [v_1, \dots, v_N]$ , given the probability  $q$ , is

$$\begin{aligned} \pi(v|q) &= \prod_{i=1}^N q^{v_i} (1-q)^{1-v_i} \\ &= q^{\sum v_i} (1-q)^{N-\sum v_i} \\ &= q^{N_1} (1-q)^{N_0}, \end{aligned}$$

which is simply a scaled binomial density. We consider first a noninformative prior

$$\pi_0(q) = \begin{cases} 1 & , \quad 0 \leq q \leq 1, \\ 0 & , \quad \text{else}, \end{cases}$$

which yields the posterior density

$$\pi(q|v) = \frac{q^{N_1}(1-q)^{N_0}}{\int_0^1 q^{N_1}(1-q)^{N_0} dq} = \frac{(N+1)!}{N_0!N_1!} q^{N_1}(1-q)^{N_0}.$$

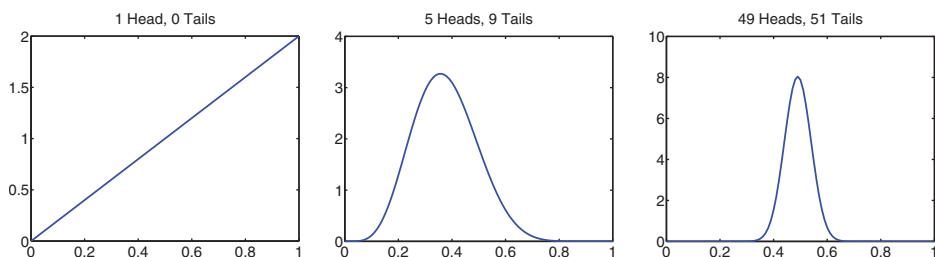
We note that in this special case, the denominator is the integral of a beta function which admits an analytic solution. In general, however, quadrature techniques must be employed to approximate the integral.

For a fair coin with  $q_0 = \frac{1}{2}$ , the posterior densities associated with various realizations  $N_1$  and  $N_0$  are plotted in Figure 4.12. It is first observed that Bayesian inference yields a posterior density with just one experiment, whereas frequentist analysis would specify a probability of either 0 or 1. It is also observed that the variability of  $\pi(q|v)$  decreases as  $N$  increases. Finally, the manner in which the data informs the density is illustrated by comparing the results with 5 Heads and 9 Tails, which has a mode of 0.36, to those of 49 Heads and 51 Tails, which has a mode of 0.495. This illustrates that the method is achieving the goal of having the data inform when there is no prior information.

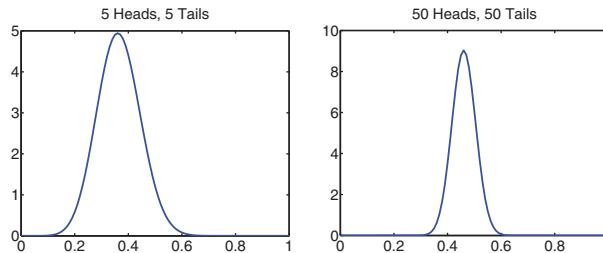
We next illustrate the effect of a poor choice for the prior density. For the same fair coin ( $q_0 = \frac{1}{2}$ ), we consider the choice

$$\pi_0(q) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(q-\mu)^2/2\sigma^2}$$

with  $\mu = 0.3$  and  $\sigma = 0.1$ . We cannot analytically evaluate the denominator in this case, so we instead employ Gaussian quadrature. As illustrated in Figure 4.13, even for a realization of 50 Heads and 50 Tails, the mode of the posterior is still smaller than  $q_0 = \frac{1}{2}$  but is significantly better than the result for 5 Heads and 5 Tails. This illustrates the manner in which a poor informative prior can have negative impact for a large number of observations. Hence if the validity of an informative prior is in doubt, it is recommended that a noninformative prior be used instead.



**Figure 4.12.** Posterior densities associated with a noninformative prior for three realizations of the coin toss experiment.



**Figure 4.13.** Posterior densities associated with a poor informative prior for two realizations of the coin toss experiment.

### Conjugate Priors

**Definition 4.67 (Conjugacy).** The property that the prior and posterior distributions have the same parametric form is termed conjugacy. When this occurs, the prior  $\pi_0(q)$  is termed a conjugate prior for the likelihood  $\pi(v|q)$ . Parameters in the prior relation are often termed *prior hyperparameters* to distinguish them from the model parameters  $q$ . The corresponding parameters in the posterior relation are called *posterior hyperparameters*.

The use of conjugate priors, when possible, is advantageous since closed-form expressions for the posterior are then available. This will be used when estimating densities for measurement errors in Chapter 8.

**Example 4.68.** Consider the binomial model

$$\pi(v|q) = q^{N_1} (1-q)^{N-N_1}, \quad N_1 = \sum_{i=1}^N v_i$$

used for the likelihood in the coin toss Example 4.66. We observe that if the prior is parameterized similarly, the product of the prior and likelihood will be in the same family. Specifically, we take  $\pi_0(q)$  to be a beta density with hyperparameters  $\alpha$  and  $\beta$  so that  $\pi_0(q) \propto q^{\alpha-1}(1-q)^{\beta-1}$ , as shown in Definition 4.16. It then follows that the posterior density satisfies

$$\begin{aligned} \pi(q|v) &\propto q^{N_1} (1-q)^{N-N_1} q^{\alpha-1} (1-q)^{\beta-1} \\ &= q^{N_1+\alpha-1} (1-q)^{N-N_1+\beta-1}, \end{aligned}$$

so it is a beta density with shape parameters  $N_1 + \alpha$  and  $N - N_1 + \beta$ . The beta prior distribution is thus a conjugate family for the binomial likelihood.

**Example 4.69.** Here we consider normally distributed random variables with known mean  $\mu$  and unknown variance  $\sigma^2$ . As detailed in Section 4.3.2, the likelihood of observing  $v = [v_1, \dots, v_n]$  iid measurements under these assumptions is

$$\pi(v|\sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-SS/2\sigma^2},$$

where the sum of squares error is

$$SS = \sum_{j=1}^n (v_j - \mu)^2.$$

This likelihood is in the inverse-gamma family, defined in Definition 4.14, so the conjugate prior is  $\pi_0(\sigma^2) \propto (\sigma^2)^{-(\alpha+1)} e^{\beta/\sigma^2}$ . The posterior density can then be expressed as

$$\begin{aligned}\pi(\sigma^2|v) &\propto \pi_0(\sigma^2)\pi(v|\sigma^2) \\ &\propto (\sigma^2)^{-(\alpha+1)}e^{-\beta/\sigma^2}(\sigma^2)^{-n/2}e^{-SS/2\sigma^2} \\ &= (\sigma^2)^{-(\alpha+1+n/2)}e^{-(\beta+SS/2)/\sigma^2}\end{aligned}$$

so that

$$\sigma^2|v \sim \text{Inv-gamma}(\alpha + n/2, \beta + SS/2).$$

As shown in Definitions 4.13 and 4.14, if  $X \sim \text{Gamma}(\alpha, \beta)$ , then  $Y = X^{-1} \sim \text{Inv-gamma}(\alpha, \beta)$ . This equivalence can be exploited so that the MATLAB command `gamrnd.m` can be used to generate random numbers from a gamma distribution which can then be used to construct random values from an inverse-gamma distribution.

## 4.9 Notes and References

This chapter provides an overview of statistical topics that play a role in uncertainty quantification, and we necessarily leave details to the following references. The text [62] provides a very accessible introduction to probability, point, and interval estimation, hypothesis testing, analysis of variance, and linear regression with clearly stated definitions. The texts [112, 171] are also excellent sources for obtaining an overview of probability and statistics at an upper undergraduate level. The book [96] delineates the difference between estimators and estimates by using different notation and is an excellent source for details regarding linear regression. Finally, [81, 82] are classics in the field of probability.

There are a number of excellent supplemental texts on random processes and Markov chains, including [99, 121, 126, 130, 158, 184, 254]. Additional theory, examples, and numerical algorithms for random equations and SDEs can be found in [91, 138, 186, 231]. We note that due to the mathematical nature of the underlying framework required for SDEs, these latter texts also provide a measure theoretic framework for random variables and other concepts discussed in this chapter. Additional details regarding a measure theoretic basis for aspects of this material can be found in [36].

The reader is referred to [56, 224] for introductory concepts and examples regarding Bayesian analysis and computing and [34, 92] for a more in-depth treatment of Bayesian inference and MCMC techniques. The text [128] provides an introduction to Bayesian inference in the context of inverse problems.

## 4.10 Exercises

**Exercise 4.1.** Use the definition of the mean and variance to prove the relations (4.12) and (4.16) when  $n = 2$ .

**Exercise 4.2.** Let  $X \sim \mathcal{U}(a, b)$  be a uniformly distributed random variable. Show that the mean and variance are

$$\mathbb{E}(X) = \frac{a+b}{2}, \quad \text{var}(X) = \frac{(b-a)^2}{12}.$$

**Exercise 4.3.** For  $z \in [-1, 1]$  and  $x \in [a, b]$ , show that the function  $f$  given by

$$x = f(z) = \frac{a+b}{2} + \frac{b-a}{2}z$$

is a one-to-one and onto mapping from  $[-1, 1]$  to  $[a, b]$ .

**Exercise 4.4.** Let  $Y$  be a random variable and  $c$  and  $d$  be real numbers. Show that

$$\mathbb{E}(cY + d) = c\mathbb{E}(Y) + d,$$

$$\text{var}(cY + d) = c^2\text{var}(Y).$$

**Exercise 4.5.** Let  $Y$  be a random  $p$ -vector,  $Z$  be a random  $n$ -vector, and  $A \in \mathbb{R}^{n \times p}$  be a deterministic and known matrix. Use (4.12) and (4.16) to establish that

$$\mathbb{E}(AY + Z) = A\mathbb{E}(Y) + \mathbb{E}(Z),$$

$$V(AY) = AV(Y)A^T,$$

where  $V(Y)$  denotes the covariance matrix for  $Y$ . This is Theorem 4.16 in [96].

**Exercise 4.6.** Let  $Z \sim \mathcal{U}(-1, 1)$  and  $X \sim \mathcal{U}(a, b)$  be uniformly distributed random variables with respective means  $\mu_z = 0$ ,  $\mu_x = (a+b)/2$  and variances  $\sigma_z^2 = 1/3$ ,  $\sigma_x^2 = (b-a)^2/12$ . Use the results of Exercises 4.2 and 4.3 to show that

$$X = \mu_x + \frac{\sigma_x}{\sigma_z}Z. \tag{4.42}$$

In this manner, a uniform random variable on the interval  $[a, b]$  can be expressed in terms of one defined on  $[-1, 1]$ . This is important since the Legendre polynomials employed in Chapter 10 and associated Gauss–Legendre quadrature formulae are defined on  $[-1, 1]$ .

**Exercise 4.7.** Let  $f_{X_1, X_2}(x_1, x_2)$  denote the pdf for a bivariate normal with covariance matrix  $V = \sigma^2 I$ , where  $I$  is the  $2 \times 2$  identity matrix. Compute the marginal density  $f_{X_2}(x_2)$  and conditional density  $f_{X_1|X_2}(x_1|\bar{x}_2)$  where  $\bar{x}_2$  is fixed. Compare your results with Figure 4.6.

**Exercise 4.8.** Consider the Markov chain from Example 4.56. Numerically verify that  $\pi = [1/6, 1/6, 1/3, 1/6, 1/6]$  is a stationary distribution. Write a program to show that if  $p^0 = [1/2, 0, 0, 1/2, 0]$ , then  $p^3 = p^6 = p^9 = \dots = p^0$  so that the period is  $k = 3$ . Because the Markov chain is periodic,  $\lim_{n \rightarrow \infty} p^n$  does not exist.

**Exercise 4.9.** Consider the Markov chain with the transition matrix

$$P = \begin{bmatrix} 0 & 1 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{2}{3} & 0 \\ 0 & 0 & 0 & 1 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{bmatrix}.$$

Write a program to investigate the limiting behavior of the chain, and consider the initial distribution  $p^0 = [1/4, 1/4, 1/4, 1/4]$ . Does the chain appear to converge? Now consider the initial distribution  $p^0 = [1, 0, 0, 0]$  and see if you get the same limit. Show that the chain is periodic, and determine its period. Note that existence of a limit with one initial distribution does not establish its existence for all initial conditions if the chain is periodic.

**Exercise 4.10.** Consider the Markov chain with the transition matrix

$$P = \begin{bmatrix} \frac{1}{3} & \frac{2}{3} \\ 1 & 0 \end{bmatrix},$$

which is nonnegative. Show numerically that the chain is irreducible and aperiodic and that  $P^m > 0$  for  $m \geq 2$ , as established in Theorem 4.60. Numerically determine  $\lim_{m \rightarrow \infty} P^m$ .

## Chapter 5

# Representation of Random Inputs

In Section 4.5, we illustrated that solutions of differential equations with uncertain parameters are random processes, whereas in Chapters 7 and 8, we provide statistical techniques to construct parameter densities using measured data. We detail here techniques for transforming random algebraic and differential equations into problems posed in terms of inputs—e.g., parameters, initial conditions, or boundary conditions—having densities that are constructed either experimentally or determined using the techniques of Chapters 7 and 8.

To motivate, consider the PDE

$$\begin{aligned} \frac{\partial T}{\partial t} &= \frac{\partial}{\partial x} \left( \alpha(x, \omega) \frac{\partial T}{\partial x} \right) + f(t, x) , \quad -1 < x < 1, \quad t > 0, \\ T(t, -1, \omega) &= T_\ell(\omega) , \quad T(t, 1, \omega) = T_r(\omega) , \quad t > 0, \\ T(0, x, \omega) &= T_0(\omega) , \quad -1 < x < 1, \end{aligned} \tag{5.1}$$

which, as detailed in Example 3.5, models the temperature  $T$  of a structure with uncertain diffusivity  $\alpha$ , boundary conditions  $T_\ell, T_r$ , and initial conditions  $T_0$ . In general, each parameter will have an associated probability space but, to simplify notation, we assume a single probability space  $(\Omega, \mathcal{F}, P)$ . Finally, the response  $y$  is taken to be the temperature  $T$  at each point  $(t, x)$  in the domain.

There are two problems associated with quantifying the randomness of inputs in (5.1):  $\alpha(\omega, x)$  is infinite-dimensional, and we cannot directly quantify the probability associated with events  $\omega \in \Omega$ . We address the first in Section 5.3 by assuming that random fields and processes can be adequately approximated by finite-dimensional expansions. To address the second issue, we illustrate in Section 5.1 techniques to pose the problem in terms of mutually independent random variables with associated densities.

## 5.1 Mutually Independent Random Parameters

We consider first problems with  $p$  mutually independent parameters  $Q = [Q_1, \dots, Q_p]$ , where each parameter  $Q_i(\omega) : \Omega \rightarrow \mathbb{R}$  has an associated density  $\rho_{Q_i}(q_i)$ . We denote

the range of  $Q_i$  by  $\Gamma_i = Q_i(\Omega) \subset \mathbb{R}$  and take  $\Gamma = \prod_{i=1}^p \Gamma_i$ . Since the parameters are assumed to be mutually independent, the joint density  $\rho_Q(q) : \Gamma \rightarrow \mathbb{R}$  is

$$\rho_Q(q) = \prod_{i=1}^p \rho_{Q_i}(q_i). \quad (5.2)$$

Because every realization  $\omega \in \Omega$  yields a value of the random vector  $Q$  in  $\Gamma$ , we can reformulate the problem in the image probability space  $(\Gamma, \mathcal{B}(\Gamma), \rho_Q(q)dq)$  rather than the abstract probability space  $(\Omega, \mathcal{F}, P)$ . Here  $\mathcal{B}(\Gamma)$  is the Borel  $\sigma$ -algebra on  $\Gamma$  and  $\rho_Q(q)dq$  is the measure of  $Q$ .

**Example 5.1.** Consider (5.1) with constant diffusivity  $\alpha(x, \omega) = \bar{\alpha}(\omega)$ . We seek  $T(t, x; Q) : [0, T_f] \times [-1, 1] \times \Gamma \rightarrow \mathbb{R}$ , which solves

$$\begin{aligned} \frac{\partial T}{\partial t} &= \bar{\alpha} \frac{\partial^2 T}{\partial x^2} + f(t, x) \quad , \quad -1 < x < 1, \quad t > 0, \\ T(t, -1, Q) &= T_\ell \quad , \quad T(t, 1, Q) = T_r \quad , \quad t > 0, \\ T(0, x, Q) &= T_0 \quad , \quad -1 < x < 1, \end{aligned} \quad (5.3)$$

where  $Q = [\bar{\alpha}, T_\ell, T_r, T_0]$  is the vector of mutually independent random inputs or parameters.

## 5.2 Correlated Random Parameters

The assumption of mutually independent parameters permits the representation (5.2) for the joint density  $\rho_Q(q)$ . This assumption is required for the stochastic Galerkin and discrete projection methods of Chapter 10 and is necessary for stochastic collocation unless the joint density can be directly constructed. This assumption also underlies many sampling methods and is required to establish the Gaussian behavior of responses constructed using perturbation methods.

Unfortunately, the parameters for even simple physical models are often correlated and hence dependent. For example, Figure 8.12 illustrates that  $Q$  and  $h$  in the steady state heat model (8.22) are highly correlated, whereas Figure 8.15 illustrates correlation between  $k_2, \delta$  and  $\lambda_1, d_1$  in the HIV model (8.24).

As detailed in Section 6.3, correlation between parameters is fundamentally different from parameter nonidentifiability and it typically cannot be addressed by reparameterization of the model. This is easily observed for the heat model.

One strategy for correlated parameters is to seek a transformation that yields a new set of independent parameters. As detailed in [266], this transformation is the Cholesky decomposition of the covariance matrix if parameters are normally distributed. For non-Gaussian parameter distributions, the transformations are typically nonlinear and specific techniques are determined by the available density information. For applications where marginal distributions and a correlation matrix are provided, but a joint density is not available, one can employ the Nataf transformation detailed in [74]. This is the method that is presently implemented

in the Sandia National Laboratories toolkit DAKOTA for use with stochastic spectral methods [4]. In the first step, the Nataf transformation is used to pose the problem in terms of correlated Gaussian random variables. Second, a Cholesky decomposition is applied to obtain a representation formulated in terms of mutually independent Gaussian random variables. We note that, in practice, an identity matrix is often employed in lieu of a correlation matrix if this information is not known. The degree to which this degrades the accuracy of the representation depends on the level of correlation. If the joint distribution is known, a Rosenblatt transformation is typically a better alternative [210]. However, this information is rarely available for complex problems.

For problems in which marginal densities and correlation matrices are unavailable, one can sample from the parameter chains constructed using the Markov chain techniques of Chapter 8 to construct densities or prediction intervals for responses or QoI. Random sampling based on the chain indices has the advantage that it eliminates the requirement of mutually independent parameters.

### 5.3 Finite-Dimensional Representation of Random Coefficients

We assume that infinite-dimensional random coefficients  $\alpha(t, x, \omega)$  can be adequately approximated by expansions of the form

$$\alpha(t, x, \omega) \approx \bar{\alpha}(t, x) + \sum_{n=1}^N Q_n(\omega) \Phi_n(t, x), \quad (5.4)$$

where  $\bar{\alpha}(t, x) = \mathbb{E}[\alpha(t, x, \omega)]$ ,  $Q = [Q_1(\omega), \dots, Q_N(\omega)] : \Omega \rightarrow \mathbb{R}$  is a vector of mutually independent random variables, and  $\Phi_n(t, x)$  are basis functions. There are three significant challenges associated with constructing these expansions: ensuring that the coefficients  $Q_n(\omega)$  are mutually independent, maintaining relatively small  $N$ , and determining densities  $\rho_{Q_n}(q_n)$  for each coefficient.

One approach is to specify appropriate basis functions  $\Phi_n(t, x)$ —e.g., splines or finite elements—and apply the Bayesian techniques of Chapter 8 to construct densities for the coefficients  $Q_n$ . The difficulty is that  $N$  will be very large if representing random processes in  $\mathbb{R}^3$ . For example, even a very coarse grid of  $N_0 = 10$  in each space and time dimension will yield  $N = 10^4$ .

Karhunen–Loève expansions provide an alternative for correlated random processes. This technique is also known as proper orthogonal decomposition (POD) and, in finite-dimensional settings, principal component analysis (PCA).

#### Karhunen–Loève Expansions

To illustrate, consider a correlated second-order random field  $\alpha(x, \omega)$  defined for  $x \in \mathcal{D}$  with mean  $\bar{\alpha}(x)$  and covariance function  $C(x, y)$ , as defined in (4.34). The Karhunen–Loève expansion of  $\alpha$  is

$$\alpha(x, \omega) = \bar{\alpha}(x) + \sum_{n=1}^{\infty} \sqrt{\lambda_n} \phi_n(x) Q_n(\omega), \quad (5.5)$$

where  $\lambda_n$  and  $\phi_n$  are the eigenvalues and orthonormal eigenfunctions of  $C$ ; that is, they solve the integral equation

$$\int_{\mathcal{D}} C(x, y) \phi_n(y) dy = \lambda_n \phi_n(x) \quad (5.6)$$

for  $x \in \mathcal{D}$ . From the orthogonality of  $\phi_n$ , it follows that the random variables  $Q_n(\omega)$  are given by

$$Q_n(\omega) = \frac{1}{\sqrt{\lambda_n}} \int_{\mathcal{D}} [\alpha(x, \omega) - \bar{\alpha}(x)] \phi_n(x) dx. \quad (5.7)$$

Hence they satisfy

$$\mathbb{E}(Q_n) = 0 \quad , \quad \mathbb{E}(Q_m Q_n) = \delta_{mn}, \quad (5.8)$$

so they are centered and uncorrelated.

For numerical implementation, one employs the truncated expansion

$$\alpha(x, \omega) = \bar{\alpha}(x) + \sum_{n=1}^N \sqrt{\lambda_n} \phi_n(x) Q_n(\omega), \quad (5.9)$$

where the choice of  $N$  is dictated by the decay rate of the eigenvalues  $\lambda_n$ . As detailed in [148, 266] and illustrated in Examples 5.2 and 5.3, the decay rate of  $\lambda_n$  is directly related to the smoothness of  $C$  and the correlation length  $L$  of the process. One typically chooses  $N$  so that the sum of the neglected terms is sufficiently small compared with the sum of the first  $N$  terms.

The second implementation issue concerns the specification of random parameters  $Q_n$ . Whereas (5.7) illustrates the statistical structure of the random variables, it is not useful for implementation in applications where  $\alpha(x, \omega)$  is unknown and must be constructed using statistical model calibration techniques.

For Gaussian processes, the situation is fairly simple since  $Q_n$  are Gaussian random variables. As in Section 5.1, we consider  $Q(\omega) = [Q_1(\omega), \dots, Q_N(\omega)] \in \Gamma \subset \mathbb{R}^N$  in the image probability space  $(\Gamma, \mathcal{B}(\Gamma), \rho_Q(q))$  where  $\rho_Q(q)$  is a multivariate normal density with mean zero. Furthermore, since uncorrelated and independent are equivalent properties for Gaussian random variables,  $Q_n$  will be mutually independent, which is important from the perspective of implementation.

For a non-Gaussian field  $\alpha(x, \omega)$ , uncorrelated does not necessarily imply independent, but we can still represent  $\alpha$  in terms of  $N$  random Gaussian parameters if a cumulative distribution  $F_\alpha$  for  $\alpha$  is known or can be approximated. If we define  $\gamma(x, \omega) = F_\alpha^{-1}(\beta(x, \omega))$ , where  $\beta(x, \omega)$  is a Gaussian random field, then arguments analogous to those in (4.11) establish that  $\gamma$  and  $\alpha$  have the same distribution. With a truncated Karhunen–Loëve expansion for  $\beta$ , the non-Gaussian field  $\alpha$  can be represented as

$$\alpha^N(x, \omega) = F_\alpha^{-1} \left( \bar{\beta}(x) + \sum_{n=1}^N \sqrt{\lambda_n} \phi_n(x) Q_n(\omega) \right),$$

where  $Q_n$  are normally distributed random variables. We note, however, that the random parameters constructed in this manner may not be independent, which can

be detrimental to implementation. Algorithms based on mappings of this nature are presented in [192].

**Example 5.2 (Uncorrelated and Fully Correlated Random Processes).** For uncorrelated random processes,  $C(x, y) = \delta(x - y)$  and (5.6) reduces to  $\phi_n(x) = \lambda_n \phi_n(x)$ , so  $\lambda_n = 1$  for all  $n$ . Hence the eigenvalues do not decay. Conversely,  $C(x, y) = 1$  for fully correlated processes, which yields

$$\int_{\mathcal{D}} \phi_n(y) dy = \lambda_n \phi_n(x).$$

In this case, one can specify  $\phi_1(x) = 1$ ,  $\lambda_1 = \text{length}(D)$ , and  $\lambda_n = 0$  for  $n > 1$ .

**Example 5.3.** We revisit the heat equation (5.1), where  $\alpha(x, \omega)$  is taken to be a Gaussian random field with mean  $\bar{\alpha}(x)$  and covariance function  $C_\alpha(x, y)$ . We also assume that the random variables  $T_\ell(\omega)$ ,  $T_r(\omega)$ , and  $T_0(\omega)$  are mutually independent.

We employ the Karhunen–Loève expansion

$$\alpha^N(x, Q) = \bar{\alpha}(x) + \sum_{n=1}^N \alpha_n(x) Q_n(\omega) \quad (5.10)$$

to approximate  $\alpha$ . The coefficients  $\alpha_n(x) = \sqrt{\lambda_n} \phi_n(x)$  are specified in terms of eigenvalues and eigenfunctions associated with the covariance function.

To illustrate the analytic computation of  $\lambda_n$  and  $\phi_n$  for a special choice of covariance function, we consider

$$C(x, y) = \frac{1}{2L} e^{-|x-y|/L}, \quad (5.11)$$

where the factor  $\frac{1}{2L}$  normalizes the integral of  $C$  to unity so that it is a density. The resulting integral equation is

$$\int_{-1}^1 e^{-|x-y|/L} \phi_n(y) dy = 2L \lambda_n \phi_n(x).$$

Results in [94] can be used to establish that the even and odd eigenvalues and eigenfunctions are

$$\lambda_{n_{even}} = \frac{1}{1 + L^2 \eta_{n_{even}}^2} \quad , \quad \lambda_{n_{odd}} = \frac{1}{1 + L^2 \eta_{n_{odd}}^2}$$

and

$$\phi_{n_{even}}(x) = \frac{\cos(\eta_{n_{even}} x)}{\sqrt{1 + \frac{\sin(2\eta_{n_{even}})}{2\eta_{n_{even}}}}} \quad , \quad \phi_{n_{odd}}(x) = \frac{\sin(\eta_{n_{odd}} x)}{\sqrt{1 - \frac{\sin(2\eta_{n_{odd}})}{2\eta_{n_{odd}}}}},$$

where  $\eta_{n_{even}}$  and  $\eta_{n_{odd}}$  are ordered solutions of the transcendental equation

$$[1 - L\eta_{n_{even}} \tan(\eta_{n_{even}})][L\eta_{n_{odd}} + \tan(\eta_{n_{odd}})] = 0. \quad (5.12)$$

Note that solutions  $\eta_{n_{even}}$  and  $\eta_{n_{odd}}$  to (5.12) can be easily computed by plotting the functions to obtain initial values for root-finding algorithms. In general, such closed-form eigenpair solutions cannot be computed and one relies instead on numerical solutions.

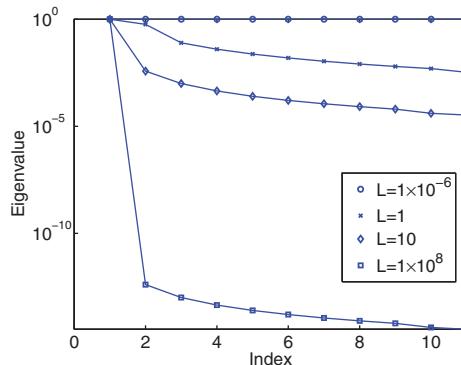
The eigenvalues obtained with various correlation lengths  $L$  are plotted in Figure 5.1. It is established in Exercise 5.1 that  $C(x, y)$  limits to the Dirac density  $\delta(x - y)$  in the limit  $L \rightarrow 0$ , and this is illustrated by the choice  $L = 1 \times 10^{-6}$ , which yields eigenvalues that are approximately unity. The choice  $L = 1 \times 10^8$  yields one eigenvalue of unity with the remaining eigenvalues limiting to zero, which is consistent with the observation that the process becomes fully correlated in the limit  $L \rightarrow \infty$ .

Based on the expansion (5.10), the finite-dimensional parameter set is

$$Q = [Q_1, \dots, Q_N, T_\ell, T_r, T_0],$$

so  $p = N + 3$ , and the approximate random system is

$$\begin{aligned} \frac{\partial T}{\partial t} &= \frac{\partial}{\partial x} \left( \alpha^N(x, Q) \frac{\partial T}{\partial x} \right) + f(t, x) \quad , \quad -1 < x < 1 , \quad t > 0, \\ T(t, -1, Q) &= T_\ell \quad , \quad T(t, 1, Q) = T_r \quad , \quad t > 0, \\ T(0, x, Q) &= T_0 \quad , \quad -1 < x < 1. \end{aligned} \tag{5.13}$$



**Figure 5.1.** First 11 eigenvalues for the covariance function (5.11) for various choices of the correlation length  $L$ .

## 5.4 Exercises

**Exercise 5.1.** Plot the covariance function  $C(0, y) = \frac{1}{2L} e^{-|y|/L}$  for various values of  $L$ , and illustrate that it behaves like the Dirac density  $C(0, y) \approx \delta(y)$  for small  $L$  and the constant function  $C(0, y) \approx 1$  for large  $L$ . For  $j = \frac{1}{L}$ , use the fact that  $C_j(0, y) = \frac{j}{2} e^{-j|y|}$  is a Dirac sequence [146] to prove that  $C_j(0, y) \rightarrow \delta(y)$  as  $j \rightarrow \infty$ .

## Chapter 6

# Parameter Selection Techniques

This chapter addresses techniques to isolate the set of identifiable or influential parameters in models. These techniques are often termed subset selection, active subspace, or essential subspace methods. Parameter selection is critical for the model calibration techniques detailed in Chapters 7 and 8 and to reduce the dimensionality of models for uncertainty propagation. For model calibration, unidentifiable parameters cannot be uniquely estimated by frequentist or Bayesian inference using noninformative priors. For example, we illustrated in Example 3.2 that the parameters  $q = [m, c, k]$  for the spring model (3.5) cannot be uniquely determined using displacement data, whereas the reformulated model parameters  $K = \frac{k}{m}$  and  $C = \frac{c}{m}$  are identifiable. For models such as those arising in systems biology or neutron transport, the number of parameters can be in the millions, so techniques to isolate influential inputs are critical to reduce the dimensionality of surrogate models used for model calibration or uncertainty propagation.

The terms *unidentifiable* and *noninfluential* are often used synonymously in the literature. As noted in the following definitions, however, the concepts differ. There are also misconceptions regarding the relation between the statistical concept of parameter correlation and the system input-output property of parameter identifiability; we address this in Section 6.3.

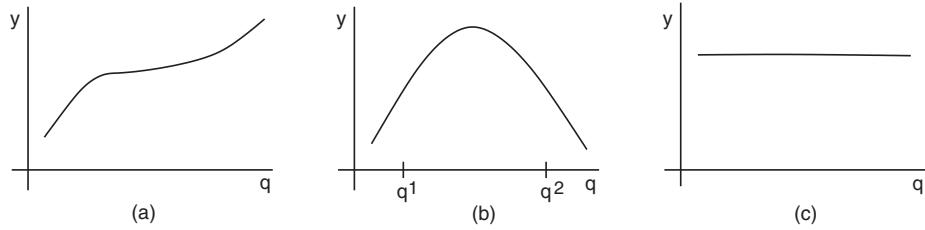
**Definition 6.1 (Identifiable Parameters).** Consider the input-output map

$$y = f(q), \quad q = [q_1, \dots, q_p].$$

The parameter set  $q$  is identifiable at  $q^*$  if  $f(q) = f(q^*)$  implies that  $q = q^*$  for any admissible  $q \in \mathbb{Q}$ . The parameter set  $q$  is identifiable with respect to a space  $I(q)$  if this holds for all  $q^* \in I(q)$ . We refer to  $I(q)$  as the identifiable subspace. The unidentifiable parameter space  $NI(q)$  is the orthogonal complement<sup>5</sup> of  $I(q)$  with

---

<sup>5</sup>Let  $S$  be a set of vectors in an inner product space  $X$ . The orthogonal complement  $S^\perp$  to  $S$  is the set of vectors in  $X$  that are orthogonal to all vectors in  $S$ . The space  $X$  can always be represented as the direct sum of a subspace  $S$  and its orthogonal complement  $S^\perp$ , which is expressed as  $X = S \oplus S^\perp$ .



**Figure 6.1.** Mapping  $y = f(q)$  for an (a) identifiable, (b) unidentifiable, and (c) noninfluential parameter.

regard to admissible parameter space  $\mathbb{Q}$  with the Euclidean inner product; that is,  $\mathbb{Q} = I(q) \oplus NI(q)$ . Intuitively, a parameter set is identifiable if it is uniquely determined by the observations, whereas it is unidentifiable if the same response  $y(q^1) = y(q^2)$  can be achieved with different parameter values  $q^1 \neq q^2$ , as illustrated in Figure 6.1(b).

**Definition 6.2 (Influential Parameters).** A parameter set  $q$  is termed noninfluential on the space  $\mathcal{NI}(q)$  if  $|y(q) - y(q^*)| < \varepsilon$  for all  $q$  and  $q^* \in \mathcal{NI}(q)$ . As illustrated in Figure 6.1(c), noninfluential parameters yield responses that are equal to within a specified tolerance when evaluated at all values in  $\mathcal{NI}(q)$ . These parameters can thus be fixed for subsequent model calibration and uncertainty propagation. The orthogonal complement is the space  $I(q)$  of influential parameters.

**Remark 6.3.** The space of noninfluential parameters is a subset of the space of unidentifiable parameters. This implies that if a parameter is identifiable, it is also influential. These hierarchies dictate the manner in which the terms can be interchanged. Finally, we note that the concepts of identifiable and influential parameters are the same for linearly parameterized problems.

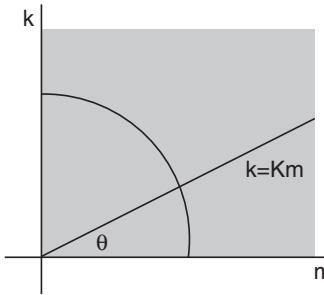
**Example 6.4.** Consider the spring model

$$\begin{aligned} m \frac{d^2z}{dt^2} + kz &= 0, \\ z(0) = z_0, \quad \frac{dz}{dt}(0) &= 0 \end{aligned} \tag{6.1}$$

discussed in Example 3.2. For  $q = [m, k]$ , the admissible parameter space is  $\mathbb{Q} = (0, \infty) \times (0, \infty)$ . From the solution  $z(t) = z_0 \cos(\sqrt{k/m} \cdot t)$ , we observe that  $q$  is not identifiable over all of  $\mathbb{Q}$  since the solution is constant along lines  $k = Km$ . Given displacements  $z(t)$ , determination of the slope  $K$  is equivalent to specifying the angle  $\theta$  illustrated in Figure 6.2. Hence the identifiable and unidentifiable subspaces of  $q$  are

$$I(q) = \{\theta = \arctan(k/m) \mid 0 < \theta < \pi/2\},$$

$$NI(q) = \left\{ r = \sqrt{k^2 + m^2} \mid r > 0 \right\},$$



**Figure 6.2.** Representation of the admissible parameter space  $\mathbb{Q}$  and identifiable and nonidentifiable subspaces  $I(q)$  and  $NI(q)$  for the spring model (6.1).

which are orthogonal complements. Hence the direct sum  $\mathbb{Q} = I(q) \oplus NI(q)$  is simply the representation of the first quadrant in terms of polar coordinates. The specification of the identifiable subspace  $I(q)$  is consistent with reformulation of the problem in terms of the parameter  $K = \frac{k}{m}$ .

For this problem, the noninfluential and influential subspaces are the same as the unidentifiable and identifiable subspaces. For a given value of  $K$ ,  $m$  and  $k$  can be fixed at any values which satisfy  $K = \frac{k}{m}$  and yield the correct displacements  $z(t)$  for all  $t$ .

The objective of this chapter is to provide techniques that can be used to construct these subspaces when the complexity of models precludes sole reliance on expert opinion to determine the identifiable and influential parameters. The relation between these concepts and the local and global sensitivity relations in Chapters 14 and 15 is illustrated in Figure 6.1. We first note that the local sensitivity must be zero at a point  $q^*$ ,  $\frac{\partial f}{\partial q}(q^*) = 0$ , in order for the parameter to be unidentifiable; however, this could be an inflection point, so this condition does not ensure that the parameter is unidentifiable. The difficulty with using local sensitivity measure  $\frac{\partial f}{\partial q}$  to establish identifiability is that it must be checked for all admissible parameters, which is typically infeasible. This motivates the linear algebra and statistical techniques discussed in this chapter and the global sensitivity methods of Chapter 15.

To simplify the discussion, we focus in this chapter solely on the relation between inputs  $q$  and output responses  $y$ . Modifications to accommodate independent variables  $t$  or  $x$  are discussed in Section 15.3.

## 6.1 Linearly Parameterized Problems

We consider first the linearly parameterized problem

$$y = Aq, \quad (6.2)$$

where  $q \in \mathbb{R}^p$ ,  $y \in \mathbb{R}^n$ , and  $A$  is an  $n \times p$  matrix. The following examples illustrate matrix subspaces that will be used to identify and isolate unidentifiable parameters.

**Example 6.5.** Consider the model

$$y_i = q_2 x_i, \quad i = 1, 2, 3,$$

with parameters  $q = [q_1, q_2]$ . The associated linear system is

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 0 & x_1 \\ 0 & x_2 \\ 0 & x_3 \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \end{bmatrix},$$

where  $\text{rank}(A) = 1$ . The unidentifiable and identifiable subspaces are specified by the relations

$$\begin{aligned} NI(q) &= \mathcal{N}(A) = c \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad c \in \mathbb{R}, \\ I(q) &= \mathcal{R}(A^T) = c \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad c \in \mathbb{R}, \end{aligned}$$

where  $\mathcal{N}(A)$  and  $\mathcal{R}(A^T)$  are the null space and range of  $A$  and its transpose. We note that the latter is the row space of  $A$  and that  $\mathcal{N}(A)$  and  $\mathcal{R}(A^T)$  are orthogonal complements. Furthermore, we observe that the unidentifiable and identifiable subspaces can also be specified by  $\mathcal{N}(A^T A)$  and  $\mathcal{R}(A^T A)$ . This is due to the property that

$$\mathcal{N}(A^T A) = \mathcal{N}(A), \quad \mathcal{R}(A^T A) = \mathcal{R}(A^T)$$

for all  $A \in \mathbb{R}^{n \times p}$ ; see Fact 4.21 of [119].

**Example 6.6.** Example 6.5 illustrates unidentifiable and identifiable subspaces that are aligned with the coordinate axes. To illustrate when this is not the case, consider the linear system

$$y = [2 \ 1] \begin{bmatrix} q_1 \\ q_2 \end{bmatrix}.$$

Here the unidentifiable and identifiable subspaces are

$$\begin{aligned} NI(q) &= \mathcal{N}(A) = c \begin{bmatrix} -\frac{1}{2} \\ 1 \end{bmatrix}, \quad c \in \mathbb{R}, \\ I(q) &= \mathcal{R}(A^T) = c \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \quad c \in \mathbb{R}. \end{aligned}$$

We again note that  $\mathcal{N}(A)$  and  $\mathcal{R}(A^T)$  are orthogonal complements. Example 6.4 illustrates subspaces that are not aligned with coordinate axes for a nonlinearly parameterized problem.

**Property 6.7.** For the linear problem (6.2), the unidentifiable and identifiable subspaces are specified by the null space and range relations

$$NI(q) = \mathcal{N}(A) = \mathcal{N}(A^T A),$$

$$I(q) = \mathcal{R}(A^T) = \mathcal{R}(A^T A).$$

The formulation in terms of  $A^T A$  is important since it relates identifiability to the Fisher information matrix  $\mathcal{F} = A^T A$  and covariance matrix  $V = \sigma^2(A^T A)^{-1}$  defined in Table 7.2. This motivates the observation that the covariance matrix will be singular for unidentifiable parameter sets.

### 6.1.1 Deterministic Algorithms

Algorithms to construct the identifiable and unidentifiable subspaces employ QR or singular value decompositions to compute the rank  $r$  of  $A$  along with  $\mathcal{N}(A)$  and  $\mathcal{R}(A^T)$ . We consider first deterministic representations that can be used when  $A$  is small to moderate in size; e.g.,  $n, p < 1000$ . We also focus on the case when there are more parameters than measurements,  $p \geq n$ , and  $A$  is rank deficient,  $r < \min\{p, n\}$ . Techniques to address the case  $n > p$  are analogous and are detailed in the references.

#### Singular Value Decomposition (SVD)

The SVD of  $A$  is

$$A = U \Sigma V^T, \quad (6.3)$$

where  $U \in \mathbb{R}^{n \times n}$  and  $V \in \mathbb{R}^{p \times p}$  are orthogonal and  $\Sigma \in \mathbb{R}^{n \times p}$  has the form  $\Sigma = [S \ 0]$ , where

$$S = \begin{bmatrix} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_r & \\ & & & 0 \end{bmatrix}, \quad \sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r \geq \varepsilon, \quad (6.4)$$

is a diagonal matrix comprised of the  $n$  ordered singular values  $\sigma_j$ . The numerical rank  $r$  is determined by the number of singular values greater than or equal to a specified tolerance  $\varepsilon$ . The columns of  $U$  and  $V$  are termed the left and right singular vectors of  $A$ .

The decomposition

$$\begin{aligned} U &= [U_r \ U_{n-r}], \quad U_r \in \mathbb{R}^{n \times r}, \quad U_{n-r} \in \mathbb{R}^{n \times (n-r)}, \\ V &= [V_r \ V_{p-r}], \quad V_r \in \mathbb{R}^{p \times r}, \quad V_{p-r} \in \mathbb{R}^{p \times (p-r)}, \end{aligned}$$

isolates the singular vectors corresponding to the nonzero singular values. Based on this decomposition,  $A$  can be expressed as

$$A = U_r S_r V_r^T, \quad (6.5)$$

where

$$S_r = \begin{bmatrix} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_r & \\ & & & 0 \end{bmatrix}.$$

As detailed in Section 4.3 of [119], the singular vectors in  $V_{p-r}$  provide a basis for  $\mathcal{N}(A)$  and those in  $V_r$  yield a basis for  $\mathcal{R}(A^T)$ .

Whereas this decomposition can in theory be used construct  $I(q) = \mathcal{R}(A^T)$ , it is illustrated in [120], for rank deficient Jacobian construction, that singular vectors can be inaccurate if  $A$  is close to a matrix of lower rank. This issue is avoided by rank-revealing QR algorithms.

### QR Algorithms

We focus on QR factorizations for  $A^T$  rather than  $A$  for two reasons:  $A^T \in \mathbb{R}^{p \times n}$ ,  $p \geq n$ , fits within the “tall and skinny” framework of the theory, and we are interested in  $\mathcal{R}(A^T)$  to construct the subspace  $I(q)$  of identifiable parameters. The QR factorization of  $A^T$  is

$$A^T = QR,$$

where  $Q \in \mathbb{R}^{p \times p}$  is orthogonal and  $R \in \mathbb{R}^{p \times n}$  is upper triangular. For full rank matrices, the first  $n$  columns of  $Q$  form an orthonormal basis for  $\mathcal{R}(A^T)$ . The next example illustrates that this is not generally true for rank deficient matrices where  $\text{rank}(A) = \text{rank}(A^T) = r < n$ .

**Example 6.8.** Consider the QR decomposition

$$A^T = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 0 & \frac{1}{\sqrt{2}} \\ 0 & \frac{1}{\sqrt{2}} \end{bmatrix} = QR, \quad (6.6)$$

where  $\text{rank}(A) = 1$ . We observe that neither column of  $Q$  forms a basis for  $\mathcal{R}(A^T) = [0, c]^T$ ,  $c \in \mathbb{R}$ .

This has been addressed by rank-revealing QR algorithms that introduce a permutation matrix  $P$  which pivots the columns of  $A^T$  so that the matrix  $R$  in the resulting QR factorization

$$A^T P = QR = Q \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix}$$

has an  $r \times r$  upper triangular block  $R_{11}$  whose diagonal elements are nonzero; see Section 5.4.1 of [97] or [63, 100]. This separates the linearly independent columns of  $R$  from those that are linearly dependent and ensures that the first  $r$  vectors of  $Q$  provide a basis for  $\mathcal{R}(A^T)$ . This can be accomplished using the MATLAB command `[Q,R,P] = qr(A')`.

**Example 6.9.** For  $A^T$  given in (6.6) the QR algorithm with column pivoting yields

$$P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad Q = \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}, \quad R = \begin{bmatrix} -1 & 0 \\ 0 & 0 \end{bmatrix},$$

so the first column of  $Q$  is a basis for  $\mathcal{R}(A^T)$ . The fact that  $-Q, -R$  yield the same result illustrates that the QR factorization is not unique.

### 6.1.2 Random Algorithms

The SVD or pivoted QR factorizations can be employed to construct the identifiable subspace  $I(q) = \mathcal{R}(A^T)$  for small to moderate size matrices  $A$ . The difficulty arises for models used for applications such as neutron transport or systems biology where  $n$  and  $p$  can be on the order of millions. For such problems, one cannot store the matrix  $A$  or compute the QR or SVD factorizations. This motivates the use of random algorithms to construct low-rank approximations to  $A$  or  $A^T$  that facilitate SVD or QR factorization of  $A$  or  $A^T$ . We summarize briefly the range-finding algorithm detailed in [105] as applied to  $A \in \mathbb{R}^{n \times p}$ .

This algorithm is comprised of two broad components.

**Stage 1.** Construct a low-dimensional subspace that adequately approximates the action of the matrix when only products  $y = Aq$  are available. To construct a basis for the range of  $A$ , we seek an orthonormal matrix  $Q$  with  $r = r(\varepsilon)$  columns whose  $r$ -dimensional range approximates  $\mathcal{R}(A)$  in the sense that

$$\|A - QQ^T A\| \leq \varepsilon, \quad (6.7)$$

where  $\|\cdot\|$  is the  $\ell_2$  operator norm. The objective is to make  $r$  as small as possible. This step is highly amenable to random sampling algorithms.

**Stage 2.** Use this low-rank  $Q$  to efficiently compute SVD or pivoted QR factorizations of  $A$  using the deterministic algorithms discussed in Section 6.1.1. To illustrate, one could form the matrix  $B = Q^T A$  whose SVD  $B = \tilde{U}\Sigma V^T$  can be efficiently computed. By forming  $U = Q\tilde{U}$ , it follows that  $A \approx U\Sigma V^T$ . Details regarding efficient factorization algorithms for this second stage are discussed in Section 5 of [105].

The heart of the algorithm focuses on the construction of a low-rank matrix  $Q$  whose range approximates that of  $A$ .

#### Algorithm 6.10 (Random Range Finder).

1. Choose  $\ell$  random inputs  $q^i$ , and compute outputs  $y^i = Aq^i$  which are compiled in the  $n \times \ell$  matrix  $Y$ .
2. Take a pivoted QR factorization  $Y = QR$  to construct a matrix  $Q$  whose columns form an orthonormal basis for the range of  $Y$ .

Details regarding the choice of  $\ell > r$ , the effect of the distribution for  $q$ , implementation when the numerical rank of  $A$  is unknown, and error estimates for the algorithm are provided in Section 4 of [105]. The use of the algorithm to construct hybrid methods for model reduction is detailed in [1].

**Example 6.11.** To illustrate the deterministic and random algorithms for isolating and constructing the identifiable subspace  $I(q) = \mathcal{R}(A^T)$ , we consider the function

$$y_i = \sum_{k=1}^p q_k \sin(2\pi k t_i)$$

evaluated at the  $n$  equally spaced points  $t_i = (i - 1)\Delta t$ ,  $\Delta t = \frac{1}{n-1}$ ,  $i = 1, \dots, n$ , on the interval  $[0, 1]$ . The resulting linear system is

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sin(2\pi t_1) & \cdots & \sin(2\pi p t_1) \\ \vdots & & \vdots \\ \sin(2\pi t_n) & \cdots & \sin(2\pi p t_n) \end{bmatrix} \begin{bmatrix} q_1 \\ \vdots \\ q_p \end{bmatrix}.$$

As we will illustrate, aliasing properties of the sine functions can be used to predict the rank of  $A$  and its factorizing matrices. This allows us to check the accuracy of linear algebra predictions for large parameter and response dimensions  $p$  and  $n$ .

### Case i: $n = p = 5$

To illustrate the effects of aliasing, we first consider the low-dimensional case  $p = n = 5$ . In Figure 6.3, we plot each column of  $A$  as a function of the points  $t_i$ . It is observed that  $\sin(2\pi k t_i) = 0$  for  $k = 2, 4$  and that  $\sin(2\pi t_i) = -\sin(2\pi \cdot 3t_i) = \sin(2\pi \cdot 5t_i)$  for this set of points  $t_i$ . Since there is only one linearly independent column, the rank of  $A$ , and hence the dimension of the identifiable subspace  $I(q)$ , is one. To illustrate this, we take the SVD and pivoted QR factorizations

$$A = U_r S_r V_r^T, \quad A^T P = Q R$$

and note that there is only one leading diagonal element in  $S_r$  and  $R$  greater than  $\text{tol} = 2 \times 10^{-15}$ . The first column of  $V_r$  or  $Q$ ,

$$V_r(:, 1) = Q(:, 1) = [-0.5774, 0, 0.5574, 0, -0.5774]^T,$$

provides a basis for the identifiable subspace  $I(q) = \mathcal{R}(A^T)$ .

### Case ii: $n = 101$ , $p = 1000$

We now turn to the moderate dimension problem where  $n = 101$  responses are constructed using  $p = 1000$  parameters. Because of aliasing, one can establish

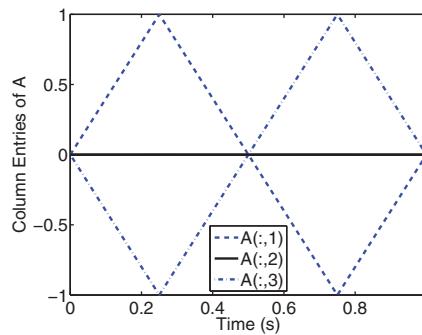
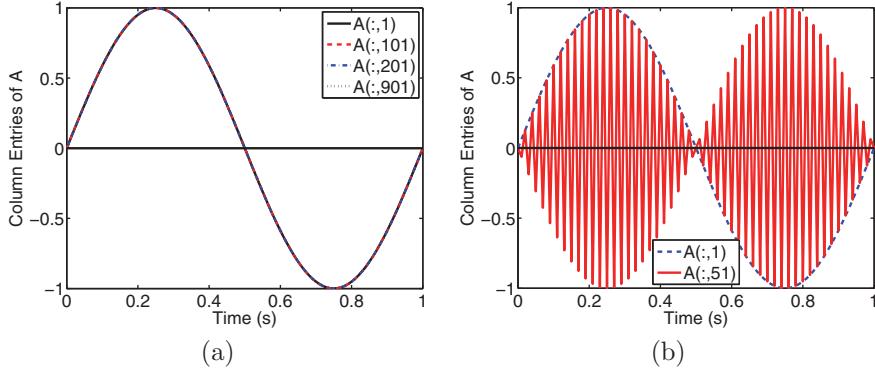


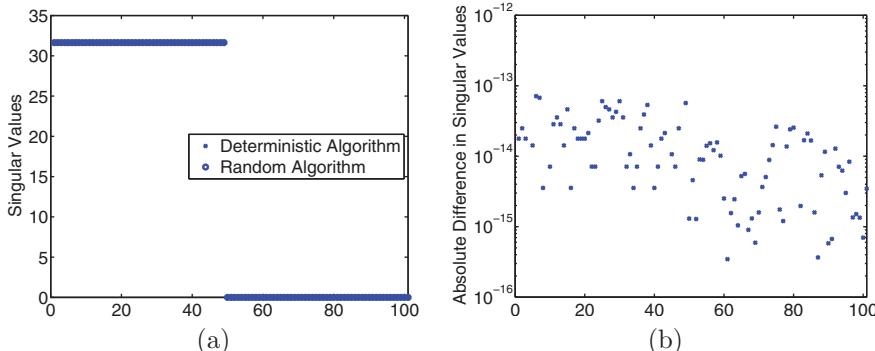
Figure 6.3. Columns of  $A$  versus the points  $t_i$ ,  $i = 1, \dots, 5$ .



**Figure 6.4.** Aliasing of  $\sin(2\pi kt_i)$  for (a)  $k = 1, 101, 201, 901$  and (b)  $k = 1, 51$ .

that  $\sin(2\pi t_i) = \sin(2\pi \cdot 101t_i) = \sin(2\pi \cdot 201t_i) = \dots = \sin(2\pi \cdot 901t_i)$ , as illustrated in Figure 6.4(a), with similar relations for  $p = 102, \dots, 199$ . This is reflected in the result  $\text{rank}(A) = 100$  returned by MATLAB. However, as illustrated by the singular values plotted in Figure 6.5(a), 49 of the 100 singular values are essentially zero, although their magnitude of  $10^{-12}$  is larger than the tolerance employed when computing the rank. This is again due to aliasing of the type illustrated in Figure 6.4(b), where it is illustrated that  $\sin(2\pi t_i) \approx \sin(2\pi \cdot 51t_i)$ . Hence the rank of  $A$  and dimension of the identifiable subspace  $I(q)$  are actually  $r = 49$ . The first 49 columns of  $Q$  or  $V_r$  provide a basis for this subspace.

We also employ the random range-finding Algorithm 6.10 with  $\ell = 75$  random parameter vectors drawn from a uniform distribution to construct an orthogonal matrix  $Q$  whose range approximates  $\mathcal{R}(A)$  in the sense of (6.7). The singular values of  $B = Q^T A$  are compared with those computed directly by the factorization  $A = U_r S_r V_r^T$  in Figure 6.5. The absolute differences illustrate the accuracy of the random algorithm.



**Figure 6.5.** (a) Singular values of  $A$  computed via the factorization  $A = U_r S_r V_r^T$  and of  $B = Q^T A$  computed using the random Algorithm 6.10 and (b) absolute difference between singular values.

### Case iii: $n = 101, p = 10^6$

Here we consider the performance of the random range-finding Algorithm 6.10 when  $A$  is too large to be stored or factored. For the reasons discussed in Case ii, the theoretical rank of  $A$  is 49. We again employ  $\ell = 75$  random vectors drawn from a uniform distribution to compute  $Y$  and hence  $Q$  and  $R$ . As in Case ii,  $R$  has 49 diagonal values that exceed a tolerance of  $10^{-3}$ , so the first 49 columns of  $Q$  can be used as a basis to approximate the range of  $A$ . Hence the random algorithm remains viable, whereas the deterministic algorithms will be infeasible.

## 6.2 Nonlinearly Parameterized Problems

The determination of identifiable and influential parameter subspaces for nonlinear problems

$$y = f(q), \quad q = [q_1, \dots, q_p]$$

is significantly more difficult than the linear case for two reasons: one cannot directly apply well-established techniques from linear algebra, and one must apply global rather than local analysis. We discuss two approaches for these problems. The first is to employ the global sensitivity analysis techniques of Chapter 15 to construct sensitivity indices which quantify the influence of parameter uncertainty on the response variance. This approach has the advantage that it requires no linearization or assumptions regarding monotonicity and it incorporates information about the parameter distributions. However, the computation of these sensitivity indices can be prohibitively expensive for large parameter dimensions since it requires quadrature over  $\mathbb{R}^p$ . In the second approach, one approximates the local sensitivities  $s_i = \frac{\partial f}{\partial q_i}$  to linearize the problem at values in the parameter space. The deterministic and random techniques of Section 6.1 are subsequently applied to the linearized problem. The primary disadvantage of this approach is that it is difficult to ensure global identifiability criteria.

### 6.2.1 Variance-Based Methods for Parameter Selection

To illustrate a nonlinear technique to select most influential parameters, we briefly summarize the variance-based global sensitivity analysis method that is detailed and illustrated with examples in Section 15.1.

We consider the scalar-valued model

$$Y = f(Q),$$

where  $Q = [Q_1, \dots, Q_p] \in \Gamma$  and  $Q_i$  are independent random variables that are assumed here to be uniformly distributed on  $[0, 1]$  so that  $\Gamma = [0, 1]^p$ . We also consider the second-order Sobol or HDMR expansion

$$f(q) = f_0 + \sum_{i=1}^p f_i(q_i) + \sum_{1 \leq i < j \leq p} f_{ij}(q_i, q_j),$$

where

$$\begin{aligned} f_0 &= \int_{\Gamma} f(q) dq, \\ f_i(q_i) &= \int_{\Gamma^{p-1}} f(q) dq_{\sim i} - f_0, \\ f_{ij}(q_i, q_j) &= \int_{\Gamma^{p-2}} f(q) dq_{\sim \{ij\}} - f_i(q_i) - f_j(q_j) - f_0. \end{aligned} \quad (6.8)$$

Here  $\Gamma^{p-1} = [0, 1]^{p-1}$ ,  $\Gamma^{p-2} = [0, 1]^{p-2}$ , and the notation  $q_{\sim i}$  denotes the vector having all the components of  $q$  except those in the set  $i$ .

The first- and second-order Sobol indices are

$$S_i = \frac{D_i}{D} \quad , \quad S_{ij} = \frac{D_{ij}}{D} \quad , \quad i, j = 1, \dots, p,$$

where the total variance  $D$  of the response  $Y$  is

$$D = \int_{\Gamma} f^2(q) dq - f_0^2$$

and the partial variances are

$$D_i = \int_0^1 f_i^2(q_i) dq_i \quad , \quad D_{ij} = \int_0^1 \int_0^1 f_{ij}^2(q_i, q_j) dq_i dq_j.$$

The total sensitivity indices

$$S_{T_i} = S_i + \sum_{j=1}^p S_{ij}$$

quantify the total effect of the parameter  $Q_i$  on  $Y$ .

It is noted in Remark 15.4 that the condition  $S_{T_i} \approx 0$  implies that  $Q_i$  is noninfluential and can be fixed for model calibration and uncertainty quantification. Hence the total sensitivity indices can be used to establish the set of influential and noninfluential parameters for nonlinear models. The extension of these relations to general densities and complete Sobol expansions is detailed in Section 15.1.2.

Whereas the Sobol indices  $S_i$ ,  $S_{ij}$ , and  $S_{T_i}$  provide comprehensive measures for quantifying the influence of parameter uncertainty on the variance of the response, their computation can be prohibitively expensive for large parameter dimensions due to the quadrature required to evaluate  $f_0$ ,  $f_i(q_i)$  and  $f_{ij}(q_i, q_j)$  in (6.8). Methods based on linearization of the problem provide an alternative technique to isolate and quantify influential parameters at significantly reduced computational cost.

### 6.2.2 Parameter Selection Based on Model Linearization

All linearization techniques employ either analytic or approximate values for the sensitivities  $s_i = \frac{\partial f}{\partial q_i}$  evaluated at values in the admissible parameter space. We first consider the  $n \times p$  sensitivity matrix defined componentwise by

$$\mathcal{X}_{ij}(q^*) = \frac{\partial f_i}{\partial q_j}(q^*),$$

where  $q^* = [q_1^*, \dots, q_p^*]$  is a nominal parameter value. The  $p \times p$  Fisher information matrix is

$$\mathcal{F} = \mathcal{X}^T \mathcal{X}.$$

For moderate dimensions  $n$  and  $p$ , the deterministic factorization techniques detailed in Section 6.1 provide measures to quantify *local* parameter identifiability and relative influence in neighborhoods of  $q^*$ . The performance of parameter selection algorithms based on  $\mathcal{X}$  and  $\mathcal{F}$  are illustrated for the HIV model (3.15) in [23].

To provide more *global* techniques to quantify parameter identifiability and influence, one can evaluate the sensitivities at random parameter values  $q^i$  specified by  $\rho_Q(q)$ . This provides the basis for the parameter selection methods employed in [1, 66]. To illustrate for the  $n = 1$  response, one employs  $r$  parameter realizations  $\{q^i\}_{i=1}^r$  to construct the  $p \times k$  sensitivity matrix

$$\mathcal{X} = [\nabla f(q^1), \dots, \nabla f(q^r)], \quad (6.9)$$

where  $\nabla f(q^i) = [\frac{\partial f}{\partial q_1}(q^i), \dots, \frac{\partial f}{\partial q_p}(q^i)]^T$ . The algorithm in [1] constructs a QR factorization of  $\mathcal{X}$  and employs the random range-finding techniques outlined in Section 6.1 to isolate and characterize the subspace of influential parameters. This algorithm is illustrated in the context of a neutron transport model employed for nuclear reactor design. The algorithm in [66] employs an SVD of  $\mathcal{X}$  to identify an active subspace for the parameter set which is then used to construct a kriging-based response surface surrogate model. The approach is illustrated in the context of a heat transfer model for a 3-D turbine blade with interior cooling holes.

The efficient evaluation of the gradient for multiple parameter values comprises a critical step when simulation codes are computationally expensive and  $p$  is large. For some applications, adjoint methods or automatic differentiation can be used for gradient computations. For highly nonlinear or complex codes, however, finite difference approximations are required to approximate the partial derivatives. For  $r$  parameter realizations, brute force computation of  $\mathcal{X}$  in (6.9) using finite differences would require  $2rp$  function evaluations, which is often prohibitive. This can be reduced to  $(p + 1)r$  using the Morris sampling strategy detailed in Section 15.2.

The objective of Morris screening algorithms is to use the difference relation

$$d_i^j = \frac{f(q^j + \Delta e_i) - f(q^j)}{\Delta} \quad (6.10)$$

to construct global sensitivity measures

$$\begin{aligned} \mu_i^* &= \frac{1}{r} \sum_{j=1}^r |d_i^j(q)|, \\ \sigma_i^2 &= \frac{1}{r-1} \sum_{j=1}^r \left( d_i^j(q) - \mu_i \right)^2, \quad \mu_i = \frac{1}{r} \sum_{j=1}^r d_i^j(q), \end{aligned}$$

which efficiently quantify the relative influence of inputs in high-dimensional prob-

lems. The stepsize  $\Delta$  is chosen from the set

$$\Delta \in \left\{ \frac{1}{\ell-1}, \dots, 1 - \frac{1}{\ell-1} \right\},$$

where  $\ell$  denotes the level, and  $e_i$  is a vector of zeros with one in the  $i^{th}$  component. Due to the magnitude of  $\Delta$ , the difference relation (6.10), which is termed the elementary effect, is a very coarse approximation to the local sensitivity. Hence the elementary effects can be used to rank the relative influence of parameters but not to resolve fine-scale gradient behavior.

For each index  $j = 1, \dots, r$ , one randomly samples a seed value  $q^*$  from  $\rho_Q(q)$  and then randomly specifies the  $p + 1$  parameter values, required to approximate the  $p$  elementary effects, using the random orientation matrix

$$B^* = \left[ J_{p+1,1} q^* + \frac{\Delta}{2} [(2B - J_{p+1,p}) D^* + J_{p+1,p}] \right] P^*.$$

Here  $D^*$  is a  $p \times p$  diagonal matrix whose elements are randomly chosen from the set  $\{-1, 1\}$  and the  $p \times p$  matrix  $P^*$  is constructed by randomly permuting the columns of a  $p \times p$  identity matrix.

The sensitivity matrix  $\mathcal{X}$  can be approximated in a similar manner if one employs a smaller stepsize  $\Delta$ .

Morris screening can thus be employed two ways for parameter selection in nonlinear problems. The first is based on the use of the random orientation matrix  $B^*$  to coarsely approximate the partial derivatives in the gradient relations (6.9). Alternatively, it is illustrated in Sections 15.2 and 15.3 that  $\mu^*$  and  $\sigma$  can often be used to rank the relative influence of parameters  $Q_i$  at a fraction of the cost required to construct the Sobol indices  $S_i$ ,  $S_{ij}$ , and  $S_{T_i}$ . However, the tradeoff is a characterization that can miss global properties of the nonlinear input-output map.

## 6.3 Parameter Correlation versus Identifiability

As noted in Definition 4.20, the Pearson correlation coefficient

$$\rho_{XY} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} \quad (6.11)$$

quantifies the degree to which the random variables  $X$  and  $Y$  are linearly dependent. Because  $\rho_{XY} \neq 0$  indicates a quantifiable statistical relationship between the variables, it is sometimes confused with the concept of parameter identifiability, which is a property of the input-output map  $Y = AQ$  or  $Y = f(Q)$ . This confusion is due in part to the fact that  $\rho_{XY} = \pm 1$  does indicate a linear algebraic relation between the variables, so they are not jointly identifiable. We illustrate the difference between parameter correlation and identifiability in the next example.

**Example 6.12.** Consider the linear model

$$Y_i = Q_1 + Q_2 x_i$$

with  $x_1 = 1$  and  $x_2 = 2$  so that the linear system is

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix}.$$

We first assume that  $Q = [Q_1, Q_2]$  is normally distributed with mean  $\mu = [0, 0]^T$  and covariance matrix

$$V = \begin{bmatrix} 0.3 & 0.5 \\ 0.5 & 1.0 \end{bmatrix}. \quad (6.12)$$

Here  $\text{cov}(Q_1, Q_2) = 0.5$ , so  $Q_1$  and  $Q_2$  are positively correlated, as illustrated by the scatterplot in Figure 6.6(a). Hence there is a statistical relation between the random variables.

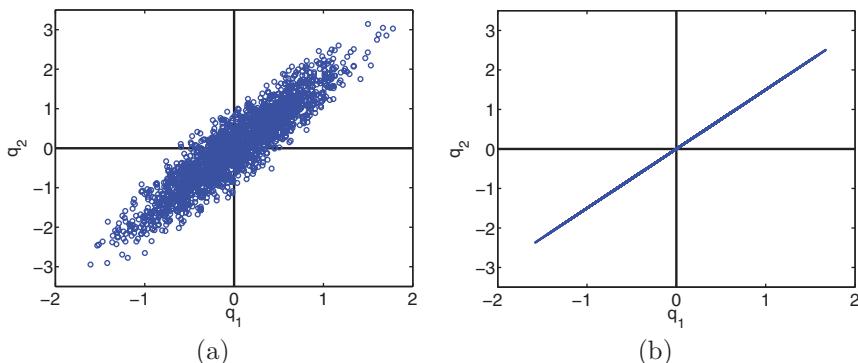
Because  $\text{rank}(A) = \dim(I(q)) = 2$ , both parameters are identifiable. Thus the parameters are correlated but can be uniquely determined from the response.

We now consider the same problem with  $Q_1 \sim N(0, 0.25)$  and  $Q_2 = \frac{3}{2}Q_1$ , which yields the scatterplot in Figure 6.6(b). The linear system in this case is

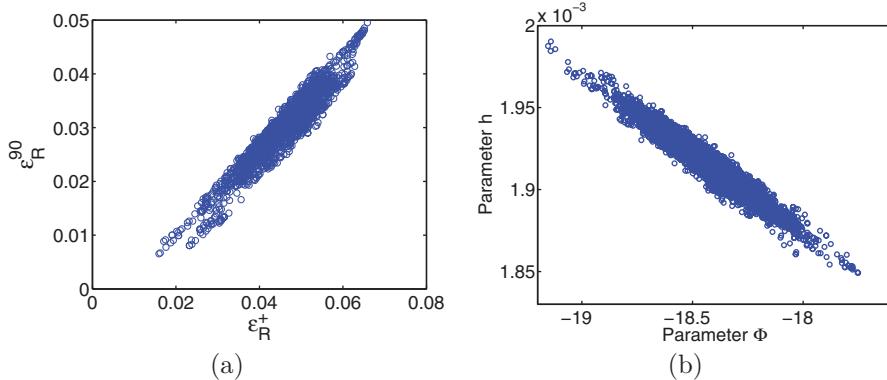
$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} \frac{5}{2} & 0 \\ 4 & 0 \end{bmatrix} \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix}.$$

Thus  $\text{rank}(A) = 1$ ,  $I(q) = c[0, 1]^T$  for  $c \in \mathbb{R}$ , and  $Q_2$  is not identifiable.

This example illustrates that correlation alone does not establish unidentifiability. However, scatterplots having essentially no width can indicate an algebraic dependence between parameters that could render them unidentifiable. Since scatterplots are constructed by sampling from the joint distribution, this is equivalent to stating that the joint distribution is nearly single-valued or exhibits a parametric functional form. For linearly dependent variables, this is manifested by values  $\rho_{XY} \approx \pm 1$  for the Pearson correlation coefficient (6.11). These functional relations need not be linear, however, since dependencies may be multiplicative; e.g.,



**Figure 6.6.** (a) Positive correlation of  $Q_1$  and  $Q_2$  for  $Q = [Q_1, Q_2] \sim N(\mu, V)$  with the covariance matrix (6.12). (b) Linear relationship  $Q_2 = \frac{3}{2}Q_1$  with  $Q_1 \sim N(0, 0.25)$ .



**Figure 6.7.** (a) Unidentifiable parameters  $\varepsilon_R^+$  and  $\varepsilon_R^{90}$  for the material model in [116] and (b) identifiable parameters  $\Phi$  and  $h$  from the heat equation in Examples 3.5 and 8.12.

$Q_2 = Q_1^2$ . This is consistent with the result  $S_{T_i} \approx 0$  for the global sensitivity indices, which indicates that variability in the random variable  $Q_i$  has minimal impact on the variability of the response  $Y$ .

Parameter sets can be directly reduced if the functional relation between parameters is evident, as for  $K = \frac{k}{m}$  in the spring model (6.1). However, this is rarely the case for complex models.

One can employ the Bayesian model calibration techniques of Chapter 8 to construct joint densities that are used to judge parameter identifiability. However, there are two difficulties with this approach. The first is that Bayesian analysis will exhibit the problems detailed in Section 8.5 if noninformative priors are used for inference with unidentifiable parameters. Moreover, one would like to select the identifiable parameter set *before* the computationally intense process of Bayesian model calibration. Second, it is often difficult to differentiate unidentifiable from identifiable parameters based on the width of scatterplots. This is illustrated in Figure 6.7 for parameters  $\varepsilon_R^+$  and  $\varepsilon_R^{90}$  arising in a smart material model [116] and  $\Phi$  and  $h$  in the heat model of Examples 3.5 and 8.12. The first model can be reformulated in terms of  $\Delta\varepsilon_R = \varepsilon_R^+ - \varepsilon_R^{90}$ , so that the pair is unidentifiable, whereas  $\Phi$  and  $h$  are identifiable. However, the widths of the scatterplots are similar. The techniques detailed in Sections 6.1 and 6.2 provide methods to ascertain identifiable or influential parameter sets before model calibration to avoid these difficulties.

## 6.4 Notes and References

The concepts of local and global identifiability are well established in the differential equations and systems literature due to the central role that they play in system identification and control theory. Bellman and Åström provided a framework for the concept they termed *structural identifiability* in their 1970 paper [33], and Åström and Eykhoff discuss the role of identification for control design in the

1971 survey paper [16]. The reader is referred to [25], and the references therein, for definitions and theory pertaining to stability and identification for distributed parameter systems. Details regarding the relation between the concepts of identifiability, controllability, and observability are provided in [253].

As deterministic and statistical system identification, parameter estimation, model calibration, global sensitivity analysis, and uncertainty quantification evolved, the terms *influential* and *uncorrelated* parameters have been employed by some researchers in a manner synonymous with *identifiable* parameters. The manner in which these concepts relate to but differ from the property of identifiability are detailed in this chapter.

Parameter selection techniques for small to moderate, linearly parameterized problems are based on SVD or pivoted QR factorizations. The theory and algorithms are detailed in [63, 97, 100, 119]. Random range-finding algorithms, such as those detailed in [105], can be employed for large problems where  $A$  cannot be stored and only the action of  $A$  or  $A^T$  is available.

Parameter selection for nonlinear problems is significantly more difficult due to the lack of a central unifying theory and the necessity of global rather than local analysis. As detailed in Chapter 15, Sobol indices quantify the global influence of parameters on the response. Whereas this approach accommodates fairly general nonlinear and nonmonotonic model behavior as well as general parameter densities, the computational cost can be prohibitive for very large input dimensions. The alternative is to approximate local sensitivities and use these linearized relations to establish influential parameters. To provide more global measures, local sensitivities are evaluated or approximated at points randomly chosen from the admissible parameter space. This is the basis for the Morris screening methods detailed in Chapter 15 and gradient-based methods of [1, 21, 66, 211]. These pseudoglobal methods will typically be ineffective for problems whose responses vary discontinuously with respect to parameters. Rather, they are predicated on the physically and theoretically motivated observation that responses often become increasingly smooth for very large parameter dimensions [47].

We have omitted a large literature on statistical methods for variable selection in high-dimensional problems. The survey paper [80] provides an overview of techniques in this category.

## 6.5 Exercises

**Exercise 6.1.** Verify that the concepts of identifiable and influential parameters are the same for linearly parameterized problems.

**Exercise 6.2.** Compute the SVD and pivoted QR expansions for the matrix  $A$  in the linear system

$$y = [2 \ 1] \begin{bmatrix} q_1 \\ q_2 \end{bmatrix}.$$

Show that the identifiable and unidentifiable subspaces are the same as those given in Example 6.5.

**Exercise 6.3.** Consider the linear function

$$Y = Q_1 + Q_2 x,$$

where  $Q = [Q_1 \ Q_2] \sim N(\bar{q}, V)$  with

$$\bar{q} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad V = \begin{bmatrix} 25 & 4 \\ 4 & 1 \end{bmatrix}.$$

Plot a scatterplot of the parameters  $Q_1$  and  $Q_2$  to illustrate the correlation. Now generate  $N = 100$  synthetic data values by taking  $h = \frac{1}{N}$  and  $x_i = ih$ ,  $i = 1, \dots, N$ , and sampling from  $N(\bar{q}, V)$  to construct corresponding values  $y_i$ . Using the theory of Section 7.2, compute a least squares fit using  $q = (X^T X)^{-1} X^T y$ , where

$$X = \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_N \end{bmatrix}^T.$$

Plot the synthetic data and least squares fit in the same figure. Finally, compute the condition number of  $X^T X$  to establish that both parameters are identifiable.

**Exercise 6.4.** Consider the steady state heat model

$$\begin{aligned} \frac{d^2 T_s}{dx^2} &= \frac{2(a+b)}{ab} \frac{h}{k} [T_s(x) - T_{amb}], \\ \frac{dT_s}{dx}(0) &= \frac{\Phi}{k}, \quad \frac{dT_s}{dx}(L) = \frac{h}{k} [T_{amb} - T_s(L)] \end{aligned}$$

detailed in Example 3.5. For fixed thermal conductivity  $k$ , use the techniques of Section 6.2 to establish that the parameters  $\Phi$  and  $h$  are identifiable.

## Chapter 7

# Frequentist Techniques for Parameter Estimation

The differential equation models of Section 3.2 can be classified as ODE systems

$$\begin{aligned} \frac{du}{dt} &= g(t, u(t), q) , \quad u(t_0) = u_0 , \quad u(t, q) \in \mathbb{R}^N, \\ y(t, q) &= \mathcal{C}u(t, q) , \quad \mathcal{C} \in \mathbb{R}^{\nu \times N}, \end{aligned} \tag{7.1}$$

stationary PDEs

$$\begin{aligned} \mathcal{N}(u, q) &= F(q) , \quad x \in \mathcal{D}, \\ B(u, q) &= G(q) , \quad x \in \partial\mathcal{D}, \\ y(x, q) &= \mathcal{C}u(x, q), \end{aligned} \tag{7.2}$$

or evolutionary PDEs

$$\begin{aligned} \frac{\partial u}{\partial t} &= \mathcal{N}(u, q) + F(q) , \quad x \in \mathcal{D}, \quad t \in [t_0, \infty), \\ B(u, q) &= G(q) , \quad x \in \partial\mathcal{D}, \quad t \in [t_0, \infty), \\ u(t_0, x, q) &= I(q) , \quad x \in \mathcal{D}. \end{aligned} \tag{7.3}$$

Here  $y$  and  $q$  denote observations and parameters and  $\mathcal{N}, F, B$ , and  $G$  denote differential operators, source terms, and boundary conditions.

Additionally, we considered algebraic models

$$A(q)u = F(q). \tag{7.4}$$

If  $A(q) \in \mathbb{R}^{n \times n}$  is invertible, we can represent the  $n$  observations by

$$y(q) = u(q) = A^{-1}(q)F(q). \tag{7.5}$$

Linear regression is a special case in which the parameter dependency is linear, so

$$y(q) = Xq.$$

For  $q \in \mathbb{R}^p$ ,  $X \in \mathbb{R}^{n \times p}$  is termed the design matrix.

In the statistics and inverse problems literature, the observed model response or QoI is often formulated as

$$y = f(\chi, q), \quad (7.6)$$

where  $\chi$  are independent variables—e.g.,  $t$  or  $x$ —or other known inputs. In the statistics literature,  $\chi$  are also referred to as explanatory or regressor variables. The function  $f$  generically denotes the map from the independent variables and parameters to the response. We assume that  $f$  is fixed and known in the sense that there exists a unique modeled response. For nonlinear ODE and PDE and algebraic models, however, one can rarely obtain analytic solutions and hence explicit formulations for  $f$ . Hence for most problems, we rely on numerical approximations for  $f$ . Finally, we note that parameters are often denoted by  $\theta$  in the statistics literature.

Throughout our discussion, we assume that we have observations  $(\chi_i, v_i)$ ,  $i = 1, \dots, n$ , where the measured quantity of interest  $v_i$  is corrupted by measurement errors  $\epsilon_i$  so that

$$v_i = f(\chi_i, q) + \epsilon_i \quad , \quad i = 1, \dots, n. \quad (7.7)$$

The mathematical inverse problem associated with parameter estimation can then be formulated as follows: given these noisy measurements, determine  $q$  in a stable manner. The associated statistical inverse problem—sometimes referred to as *inverse uncertainty quantification*—is to additionally quantify uncertainties associated with  $q$  due to the measurement errors. The assumptions required to approximate  $q$  and quantify its uncertainty define frequentist and Bayesian techniques for parameter estimation.

For sensitivity analysis and uncertainty propagation, the specific roles of the independent variables are typically of secondary importance and we are instead interested in how the model solution varies as a function of the parameters or inputs  $q$ . This is facilitated by the representation

$$v_i = f_i(q) + \epsilon_i \quad , \quad i = 1, \dots, n, \quad (7.8)$$

where  $f_i(q) \in \mathbb{R}^\nu$  denotes the observed model response and  $v_i \in \mathbb{R}^\nu$  again denotes measured data. For the models (7.1), (7.2), and (7.4), the model response can be expressed as  $n \times \nu$  vector

$$\begin{aligned} f(q) &= [f(t_1, q), \dots, f(t_n, q)]^T \quad , \quad \text{Evolution Processes,} \\ f(q) &= [f(x_1, q), \dots, f(x_n, q)]^T \quad , \quad \text{Stationary Processes,} \\ f(q) &= [f_1(q), \dots, f_n(q)]^T \quad , \quad \text{Algebraic Models.} \end{aligned} \quad (7.9)$$

Hence the dependence of the observed model response on the independent or regressor variables is suppressed in the notation  $f(q)$ . Readers are referred to Section 3.4 for discussion regarding alternative notation used for parameters in various disciplines.

For evolution models, we will have  $\nu \geq 1$  experimental measurements and model responses at each time  $t_j$ ,  $j = 1, \dots, n$ . For stationary processes and algebraic models, we consider scalar measurements and model evaluations, so  $\nu = 1$ .

## 7.1 Parameter Estimation from a Frequentist Perspective

We recall George E.P. Box's quote "Essentially, all models are wrong, but some are useful," page 424 of [38]. Thus the mathematical models will exhibit model errors, which we collectively denote by vector  $\delta = [\delta_1, \dots, \delta_n]^T$ , along with measurement errors. To accommodate these errors, we consider statistical models of the form

$$\Upsilon = f(q_0) + \delta + \varepsilon, \quad (7.10)$$

where  $\Upsilon = [\Upsilon_1, \dots, \Upsilon_n]^T$  is a random vector whose realization  $v = [v_1, \dots, v_n]^T$  is comprised of measurements from an experiment. Measurement errors are represented by the random vector  $\varepsilon = [\varepsilon_1, \dots, \varepsilon_n]^T$ , and errors resulting for a specific experiment are denoted by  $\epsilon = [\epsilon_1, \dots, \epsilon_n]^T$ .

As detailed in Section 4.8.1, a basic tenet of frequentist inference is the assumption that parameters are fixed but possibly unknown. Hence  $q_0$  represents the true but unknown value of the parameter set that generated the observations  $v = [v_1, \dots, v_n]^T$ . We emphasize that since  $q_0$  is not a random vector, the model response  $f(q_0)$  is a deterministic quantity.

If the quantification of modeling errors constitutes one of the goals, then it is necessary to consider the statistical model (7.10) and characterize the modeling errors in an efficient and statistically consistent manner, as detailed in Chapter 12. For many applications, however, the modeling and measurement errors can be collectively quantified by the random vector  $\varepsilon$ , in which case one would employ the statistical model

$$\Upsilon = f(q_0) + \varepsilon \quad (7.11)$$

in which errors are additive.

To construct likelihoods in the manner detailed in Section 4.3, we typically assume that the random variables  $\varepsilon_i$  are unbiased and iid, which is often not the case if they are comprised of both modeling and measurement errors. For example, we illustrate in Chapter 12 that residuals for a structural model are highly dependent on the magnitude of  $v$  even though the model is providing an accurate fit to measured data. Hence for some applications, the statistical model

$$\Upsilon_i = f_i(q_0)(1 + \varepsilon_i), \quad j = 1, \dots, n, \quad (7.12)$$

with multiplicative errors may be more appropriate since  $\text{var}(\Upsilon_i)$  will depend on the magnitude of  $f_i(q_0)$ .

The goal when calibrating models is to determine parameter estimates  $q$  so that the model response  $f(q)$  fits the data in some optimal sense. We showed in Section 4.3 that this can be achieved by constructing an estimator  $\hat{q}$  that estimates  $q_0$  in a statistically reasonable manner.<sup>6</sup> It was demonstrated that OLS estimators

$$\hat{q}_{OLS} = \underset{q \in \mathcal{Q}}{\operatorname{argmin}} \sum_{i=1}^n [\Upsilon_i - f_i(q)]^2 \quad (7.13)$$

---

<sup>6</sup>The notation  $\hat{q}$  for the estimator is not universal, and many texts denote the estimate by  $\hat{q}$ . Hence care must be taken to establish the convention employed in a specific text.

and maximum likelihood estimators both achieve this goal and are equivalent for certain assumptions regarding the distribution of errors  $\varepsilon_i$ .

**Remark 7.1.** Because the estimator  $\hat{q}$  is a random variable or random vector, it has a mean, covariance, and distribution termed the sampling distribution; see Definition 4.28. We will show that with appropriate assumptions regarding the distribution of  $\varepsilon_i$ ,  $\mathbb{E}(\hat{q}) = q_0$  and the covariance will quantify the variability of the errors. Furthermore, confidence limits for the sampling distribution can be used to *quantify the accuracy of the estimation process*.

What the sampling distribution *does not do* is provide a distribution for the model parameters since  $q_0$  is not a random variable in frequentist inference. We will illustrate that, for certain problems, the sampling distribution coincides with the parameter distribution constructed using Bayesian techniques. This makes it tempting to propagate the sampling distribution through the model, using the techniques of Chapters 9 and 10, to quantify the model or response uncertainty. However, this is problematic for two reasons. The first is that there is no convergence theory specifying an asymptotic relation between the sampling distribution and parameter distribution which relies on Bayesian assumptions. Second, the sampling distribution is Gaussian, which limits its accuracy for quantifying non-Gaussian parameter distributions. Hence this approach should be avoided unless additional analysis indicates an equivalence between the two distributions.

There are two alternatives. From a frequentist perspective, one can assume parametric forms (e.g., Gaussian or Johnson distributions) for the densities associated with model parameters and estimate the augmented parameter set using moment or distribution matching techniques [156, 157, 243]. For model responses of the form (7.3), this requires that errors  $\varepsilon_i$  be characterized from independent experiments. We do not provide further details about this approach but rather refer the reader to the cited references. Alternatively, the Bayesian techniques detailed in Chapter 8 can be used to construct parameter densities and moments that can be directly propagated through models.

The estimators  $\hat{q}$  can be determined explicitly only for linear parameter dependencies. Whereas applications such as convolution models for acoustics or image processing and X-ray tomography yield linearly parameterized models, general models typically exhibit a nonlinear dependence on  $q$ . To illustrate the derivation of relevant theory, we consider the linear regression (linear parameterization) problem first in Section 7.2. We return to the general problem posed here in Section 7.3.

## 7.2 Linear Regression

We illustrate here fundamental results regarding linear regression to motivate corresponding theory for the nonlinear least squares problem (7.13). Additional details can be found in [96].

We consider the statistical model

$$\Upsilon = X q_0 + \varepsilon, \quad (7.14)$$

where  $\Upsilon = [\Upsilon_1, \dots, \Upsilon_n]^T$  and  $\varepsilon = [\varepsilon_1, \dots, \varepsilon_n]^T$  are random vectors and the  $n \times p$  design matrix  $X$  is considered deterministic and known. We let  $q_0$  denote the vector of true but unknown parameters and let  $v = [v_1, \dots, v_n]^T$  denote realizations or observations from an experiment in which the realized errors are  $\epsilon = [\epsilon_1, \dots, \epsilon_n]$ . Throughout this discussion, we assume that there are more measurements than parameters so that  $n > p$ .

**Assumption 7.2.** We make the assumption that errors are unbiased and iid with variance  $\sigma_0^2$ ; hence for  $j = 1, \dots, n$ ,

- (i)  $\mathbb{E}(\varepsilon_i) = 0$ , and
  - (ii)  $\text{var}(\varepsilon_i) = \sigma_0^2$ ,  $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$  for  $i \neq j$ .
- (7.15)

In accordance with frequentist assumptions, the error variance  $\sigma_0^2$  is assumed fixed but unknown. At this point, we make no additional assumptions regarding the error distribution.

Our first objective is to construct unbiased estimators  $\hat{q}$  and  $\hat{\sigma}^2$  for the unknown parameters  $q_0$  and  $\sigma_0^2$ .

### 7.2.1 Parameter Estimator and Estimate

To construct an estimator  $\hat{q}$  for  $q_0$ , we seek  $q$ , which minimizes the OLS functional

$$\mathcal{J}(q) = (\Upsilon - Xq)^T(\Upsilon - Xq). \quad (7.16)$$

If (7.16) were scalar-valued, we would optimize it by setting the derivative with respect to  $q$  equal to 0 and solving for  $q$ . For vector-valued problems, this is achieved using the gradient  $\nabla_q \mathcal{J}$  of  $\mathcal{J}$  with respect to  $q$ . Specifically, one sets

$$\nabla_q \mathcal{J} = 2[\nabla_q(\Upsilon - Xq)^T][\Upsilon - Xq] = 0,$$

where

$$\nabla_q(\Upsilon - Xq)^T = -\nabla_q q^T X^T = -X^T,$$

to obtain the least squares estimator

$$\hat{q}_{OLS} = (X^T X)^{-1} X^T \Upsilon. \quad (7.17)$$

The realization

$$q_{OLS} = (X^T X)^{-1} X^T v \quad (7.18)$$

is the least squares estimate for the unknown true parameter  $q_0$ .

**Remark 7.3.** Throughout this chapter, we will discuss only OLS estimators and estimates. Hence to simplify notation, we will drop the subscript *OLS* and let  $\hat{q} = \hat{q}_{OLS}$  and  $q = q_{OLS}$  denote the least squares estimator and estimate.

Whereas the normal equations (7.17) provide an analytic minimum for (7.16), they are typically ill-conditioned for moderate to large numbers of parameters. Hence in practice, it is often numerically advantageous to solve the minimization problem (7.16) to avoid inaccurate results associated with numerically solving ill-conditioned linear systems.

### 7.2.2 Parameter Estimator Properties

**Result 7.4.** The parameter estimator  $\hat{q}$  has the mean and covariance matrix

$$\begin{aligned} \text{(i)} \quad & \mathbb{E}(\hat{q}) = q_0, \text{ and} \\ \text{(ii)} \quad & V(\hat{q}) = \sigma_0^2(X^T X)^{-1}. \end{aligned} \tag{7.19}$$

Relation (i) follows directly from (7.17) since

$$\mathbb{E}(\hat{q}) = \mathbb{E}[(X^T X)^{-1} X^T \Upsilon] = (X^T X)^{-1} X^T \mathbb{E}(\Upsilon) = q_0.$$

Hence  $q$  provides an unbiased estimate for the true parameter. To establish the covariance relation, we let  $A = (X^T X)^{-1} X^T$  and note that

$$\begin{aligned} V(\hat{q}) &= \mathbb{E}[(\hat{q} - q_0)(\hat{q} - q_0)^T] \\ &= \mathbb{E}[(q_0 + A\varepsilon - q_0)(q_0 + A\varepsilon - q_0)^T], \text{ since } \hat{q} = A\Upsilon = A(Xq_0 + \varepsilon) \\ &= A\mathbb{E}(\varepsilon\varepsilon^T)A^T \\ &= \sigma_0^2(X^T X)^{-1}. \end{aligned}$$

As noted previously, the error variance  $\sigma_0^2$  is assumed to be fixed but unknown. Hence to employ (7.19) to estimate the parameter covariance, we must construct an unbiased estimator  $\hat{\sigma}^2$  for  $\sigma_0^2$ .

### 7.2.3 Error Variance Estimator

**Result 7.5.** The unbiased error covariance estimator is

$$\hat{\sigma}^2 = \frac{1}{n-p} \hat{R}^T \hat{R}, \tag{7.20}$$

where

$$\hat{R} = \Upsilon - X\hat{q} \tag{7.21}$$

denotes the residual estimator.

To obtain this result, we first note that the residual can be expressed as

$$\hat{R} = (I_n - H)\Upsilon,$$

where  $I_n$  denotes the  $n \times n$  identity matrix and

$$H \equiv X(X^T X)^{-1} X^T.$$

It is straightforward to show that  $H$  satisfies the properties

$$\begin{aligned} H^T &= H \quad (\text{Symmetric}), \\ H^2 &= H \quad (\text{Idempotent}), \\ (I_n - H)^2 &= I_n - H, \\ (I_n - H)X &= 0. \end{aligned} \tag{7.22}$$

From (7.14) and (7.22), it follows that

$$\hat{R} = (I_n - H)\varepsilon$$

so that

$$\hat{R}^T \hat{R} = \varepsilon^T (I_n - H)\varepsilon. \tag{7.23}$$

If we generically denote the  $ij$  entry of  $I_n - H$  by  $h_{ij}$ , the quadratic form (7.23) can be expressed as

$$\hat{R}^T \hat{R} = \sum_{i=1}^n \sum_{j=1}^n h_{ij} \varepsilon_i \varepsilon_j.$$

It then follows that

$$\begin{aligned} \mathbb{E}(\hat{R}^T \hat{R}) &= \sum_{i=1}^n \sum_{j=1}^n h_{ij} \mathbb{E}(\varepsilon_i \varepsilon_j) \\ &= \sum_{i=1}^n \sum_{j=1}^n h_{ij} \text{cov}(\varepsilon_i, \varepsilon_j) \quad , \text{ follows from (4.14) with } \mathbb{E}(\varepsilon_j) = \mathbb{E}(\varepsilon_i) = 0 \\ &= \sum_{i=1}^n h_{ii} \text{var}(\varepsilon_i) \quad , \varepsilon_i \text{ independent} \\ &= \sigma_0^2 \text{tr}(I_n - H) \quad , \varepsilon \text{ identically distributed with variance } \sigma_0^2. \end{aligned}$$

Since the trace operator satisfies the properties  $\text{tr}(A + B) = \text{tr}(A) + \text{tr}(B)$  and  $\text{tr}(AB) = \text{tr}(BA)$ , it follows that

$$\begin{aligned} \text{tr}(I_n - H) &= n - \text{tr}[X(X^T X)^{-1} X^T] \\ &= n - \text{tr}[(X^T X)^{-1} X^T X] \\ &= n - p. \end{aligned} \tag{7.24}$$

Thus  $\hat{\sigma}^2 = \frac{1}{n-p} \hat{R}^T \hat{R}$  is an unbiased estimator for  $\sigma_0^2$ . Furthermore, we can conclude from (7.24) that the eigenvalues of  $H$  are 0 or 1.

**Example 7.6.** Consider the height-weight data from the 1975 World Almanac and Book Facts that is compiled in Table 7.1. To model this data, we employ the quadratic relation

$$\Upsilon_i = q_1 + q_2(x_i/12) + q_3(x_i/12)^2 + \varepsilon_i, \tag{7.25}$$

|                 |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|-----------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Height<br>(in)  | 58  | 59  | 60  | 61  | 62  | 63  | 64  | 65  | 66  | 67  | 68  | 69  | 70  | 71  | 72  |
| Weight<br>(lbs) | 115 | 117 | 120 | 123 | 126 | 129 | 132 | 135 | 139 | 142 | 146 | 150 | 154 | 159 | 164 |

**Table 7.1.** Height-weight data from the 1975 World Almanac and Book Facts [73].

where  $x_i$  is the height in inches and  $\Upsilon_i$  is the corresponding weight. Solution of the normal equations (7.18) yields the parameter values  $q = [261.88, -88.18, 11.96]^T$ . We note that the conditioning of the  $3 \times 3$  matrix  $X^T X$  is  $6.7 \times 10^7$ , thus illustrating the ill-conditioning of the normal equations. The variance estimate provided by (7.20) is  $\sigma^2 = 0.15$ , which yields the covariance matrix estimate

$$V = \begin{bmatrix} 634.88 & -235.04 & 21.66 \\ -235.04 & 87.09 & -8.03 \\ 21.66 & -8.03 & 0.74 \end{bmatrix}.$$

The estimated parameter values, plus and minus two standard deviations, are thus

$$\begin{aligned} q_1 &= 261.88 \pm 50.39 & q_1 &\in [211.48, 312.27] \\ q_2 &= -88.18 \pm 18.66 \Rightarrow q_2 &\in [-106.84, -69.51] \\ q_3 &= 11.96 \pm 1.72 & q_3 &\in [10.24, 13.68]. \end{aligned} \quad (7.26)$$

## 7.2.4 Sampling Distribution for $\hat{q}$

As detailed in Section 4.2, the estimator  $\hat{q}$  has a distribution, termed the sampling distribution, which we will use to construct confidence intervals for the estimation process. The assumptions required to specify a sampling distribution are more stringent than those in Assumption 7.2 and require either that errors be normally distributed or that samples be sufficiently large that the central limit theorem can be invoked for averaged error relations.

**Assumption 7.7.** The sampling distribution for  $\hat{q}$  can be directly specified for problems in which errors are iid and  $\varepsilon_i \sim N(0, \sigma_0^2)$ , where  $\sigma_0$  is fixed but likely unknown.

**Property 7.8 (Sampling Distribution for  $\hat{q}$ ).** With Assumption 7.7,  $\hat{q}$  has the sampling distribution  $\hat{q} \sim N(q_0, \sigma_0^2(X^T X)^{-1})$ . Furthermore, if we let  $\delta_k$  denote the  $k^{th}$  diagonal element of  $(X^T X)^{-1}$  and  $q_{0k}$  denote the  $k^{th}$  element of the true parameter vector  $q_0$ , then  $\hat{q}_k \sim N(q_{0k}, \sigma_0^2 \delta_k)$ .

To verify this property, we note from [96] that because each component  $\hat{q}_k$  is the linear combination of independent random variables  $\Upsilon_k$ , it follows that  $\hat{q}$  has a joint multivariate normal distribution. When combined with the fact that  $\mathbb{E}(\hat{q}) = q_0$  and  $\text{cov}(\hat{q}) = \sigma_0^2(X^T X)^{-1}$ , it follows that  $\hat{q} \sim N(q_0, \sigma_0^2(X^T X)^{-1})$ .

For numerous applications, errors may be iid with variance  $\sigma_0^2$  but not normally distributed. For sufficiently large sample sizes, asymptotic theory yields a result similar to Property 7.8.

**Property 7.9 (Asymptotic Sampling Distribution for  $\hat{q}$ ).** Consider the model (7.14) with errors which are iid with variance  $\sigma_0^2$ . For sufficiently large  $n$ , the sampling distribution for  $\hat{q}$  is asymptotically normal, which we denote by  $\hat{q} \xrightarrow{a} N(q_0, \sigma_0^2(X^T X)^{-1})$ .

Rather than provide a complete proof of Property 7.9, we instead summarize the approach and refer the reader to [219] for additional details. We first note that substitution of (7.14) into (7.17) yields  $\hat{q} - q_0 = (X^T X)^{-1} X^T \varepsilon$  so that

$$\sqrt{n}(\hat{q} - q_0) = \left( \frac{1}{n} X^T X \right)^{-1} \frac{1}{\sqrt{n}} X^T \varepsilon.$$

Because the first right-hand side term can be interpreted as an average, the law of large numbers is used to establish that

$$\frac{1}{n} X^T X \xrightarrow{P} \mathcal{Y},$$

where  $\mathcal{Y}$  is positive definite. Since  $\mathbb{E}(\frac{1}{\sqrt{n}} X^T \varepsilon) = 0$ , it follows that

$$\text{var}\left(\frac{1}{\sqrt{n}} X^T \varepsilon\right) = \mathbb{E}\left(\frac{1}{n} X^T \varepsilon \varepsilon^T X\right) \xrightarrow{P} \sigma_0^2 \mathcal{Y}.$$

The central limit theorem, discussed in Section 4.4, is then invoked to establish that

$$\frac{1}{\sqrt{n}} X^T \varepsilon \xrightarrow{D} Z,$$

where  $Z \sim N(0, \sigma_0^2 \mathcal{Y})$ , so that  $\sqrt{n}(\hat{q} - q_0) \xrightarrow{a} N(0, \sigma_0^2 \mathcal{Y}^{-1})$ . Finally, one shows that  $\frac{1}{n} X^T X$  is a strongly consistent estimator of  $\mathcal{Y}$  to obtain the asymptotic result.

An obvious practical question concerns the size  $n$  required to justify using these asymptotic results. This is problem-dependent, and alternative methods, such as Bayesian analysis, may be required to establish the normality of distributions when sample sizes are small.

### Confidence Intervals

It was shown in Section 3.3 that chi-squared and  $t$ -distributions are required to construct confidence intervals. This is established for our estimators in the next two properties.

**Property 7.10.** For  $\hat{\sigma}^2$  given by (7.20), the random variable  $\nu = \frac{(n-p)\hat{\sigma}^2}{\sigma_0^2}$  has a chi-squared distribution with  $n - p$  degrees of freedom.

To establish this, we note that

$$\begin{aligned}
\frac{(n-p)\hat{\sigma}^2}{\sigma_0^2} &= \frac{1}{\sigma_0^2} \hat{R}^T \hat{R} \\
&= \frac{1}{\sigma_0^2} \varepsilon^T (I_n - H) \varepsilon \\
&= \frac{1}{\sigma_0^2} \langle \varepsilon, U \Lambda U^T \varepsilon \rangle \quad , \quad I_n - H = U \Lambda U^T \text{ since symmetric} \\
&= \frac{1}{\sigma_0^2} \langle U^T \varepsilon, \Lambda U^T \varepsilon \rangle .
\end{aligned}$$

Since  $\text{tr}(I_n - H) = \text{rank}(I_n - H) = n - p$ , we can express  $\Lambda$  as

$$\Lambda = \begin{bmatrix} I_{n-p} & 0 \\ 0 & 0 \end{bmatrix},$$

where  $I_{n-p}$  is the  $n - p$  identity matrix. Moreover, it is proven in [96] that since  $U^T$  is an orthogonal matrix and  $\varepsilon \sim N(0, \sigma_0^2)$ , then  $u = U^T \varepsilon$  is a vector of  $N(0, \sigma_0^2)$  random variables. Because

$$\nu = \frac{(n-p)\hat{\sigma}^2}{\sigma_0^2} = \frac{\langle u, \Lambda u \rangle}{\sigma_0^2} = \sum_{i=1}^{n-p} \frac{u_i^2}{\sigma_0^2}$$

is the sum of squares of  $n - p$  independent  $N(0, 1)$  random variables, it thus has a chi-squared distribution with  $n - p$  degrees of freedom.

**Property 7.11.** The random variable

$$T_k = \frac{\hat{q}_k - q_{0_k}}{\hat{\sigma} \sqrt{\delta_k}}$$

has a  $t$ -distribution with  $n - p$  degrees of freedom.

To verify Property 7.11, we note from Property 7.8 that  $Z = \frac{\hat{q}_k - q_{0_k}}{\sigma_0 \sqrt{\delta_k}} \sim N(0, 1)$ . It then follows from Definition 4.12 that

$$\begin{aligned}
T_k &= \frac{\hat{q}_k - q_{0_k}}{\hat{\sigma} \sqrt{\delta_k}} \\
&= \frac{\hat{q}_k - q_{0_k}}{\sigma_0 \sqrt{\delta_k}} \cdot \frac{\sigma_0}{\hat{\sigma} \sqrt{n-p}} \cdot \sqrt{n-p} \\
&= \frac{Z}{\sqrt{\nu/(n-p)}} \quad , \quad Z \sim N(0, 1) \quad , \quad \nu \sim \chi^2(n-p),
\end{aligned}$$

has a  $t$ -distribution with  $n - p$  degrees of freedom.

To construct a  $(1 - \alpha) \times 100\%$  confidence interval, we employ the techniques of Example 4.33, with  $T_k = \frac{\hat{q}_k - q_{0_k}}{\hat{\sigma} \sqrt{\delta_k}}$ , to obtain

$$P\left(\hat{q}_k - t_{n-p, 1-\alpha/2} \cdot \hat{\sigma} \sqrt{\delta_k} < q_{0_k} < \hat{q}_k + t_{n-p, 1-\alpha/2} \cdot \hat{\sigma} \sqrt{\delta_k}\right) = 1 - \alpha.$$

We then employ the parameter estimate  $q = (X^T X)^{-1} X^T v$  and variance estimate  $\sigma^2 = \frac{1}{n-p} R^T R$ , where  $R = v - Xq$ , to obtain

$$\left[ q_k - t_{n-p,1-\alpha/2} \cdot \sigma \sqrt{\delta_k}, q_k + t_{n-p,1-\alpha/2} \cdot \sigma \sqrt{\delta_k} \right]. \quad (7.27)$$

We note that this is often expressed as

$$\left[ q_k - t_{n-p,1-\alpha/2} \cdot SE_k, q_k + t_{n-p,1-\alpha/2} \cdot SE_k \right], \quad (7.28)$$

where  $SE_k \equiv \sigma \sqrt{\delta_k}$  is termed the *standard error*. To construct (7.27) or (7.28), one uses a table of  $t$ -distributions or  $t$ -value calculator to look up or compute values of  $t_{n-p,1-\alpha/2}$  for specified values of  $n, p$ , and  $\alpha$  with  $n - p$  *degrees of freedom*. We caution the reader that whereas most tables are compiled in terms of one tail ( $1 - \alpha/2$ ), some provide values for both tails ( $1 - \alpha$ ). Hence care must be taken to employ  $\alpha$  consistent with the table.

**Example 7.12.** We revisit Example 7.6 and use the  $t$ -distribution to construct 90% confidence intervals for the parameters  $q_1, q_2$ , and  $q_3$  in the quadratic model (7.25). Here we have  $n = 15$  observations and  $p = 3$  parameters. For  $\alpha = 0.05$ , we obtain the value  $t_{n-p,1-\alpha/2} = 2.2$  from a table of  $t$ -values. This yields the 95% confidence intervals

$$\begin{aligned} q_1 &\in [206.45, 317.31], \\ q_2 &\in [-108.71, -67.65], \\ q_3 &\in [10.07, 13.86]. \end{aligned}$$

These intervals are slightly larger than those in (7.26) for two reasons: the intervals in (7.26) reflect  $2\sigma \approx 94.45\%$  confidence intervals, and the  $t$ -distribution has heavier tails than the normal distribution, as illustrated in Figure 4.3(b).

The statistical model, estimators, and statistical properties of the linear regression model are summarized in Table 7.2. This provides motivation and a basis for comparison for the nonlinear theory summarized in the next section.

### 7.3 Nonlinear Parameter Estimation Problem

We return to the evolutionary process model (7.1), stationary process model (7.2), and algebraic model (7.4), which exhibit nonlinear parameter dependencies, along with the associated statistical model

$$\Upsilon = f(q_0) + \varepsilon. \quad (7.29)$$

The model responses  $f(q_0)$  for the three regimes are summarized in (7.9). As before, we take  $q \in \mathbb{R}^p$  and let  $q_0$  designate the true but unknown parameter that generates the response  $v \in \mathbb{R}^n$ . As in Section 7.2, we assume that there are more measurements than parameters so that  $n > p$ . We let  $\mathbb{Q}$  denote the admissible parameter space and  $\mathcal{Q}$  denote the space associated with the estimator  $\hat{q}$ . Since both specify admissible parameter values,  $\mathbb{Q}$  and  $\mathcal{Q}$  will coincide for reasonable estimators.

Statistical Model:

$$\Upsilon = Xq_0 + \varepsilon, \quad q \in \mathbb{R}^p,$$

$$v = Xq_0 + \epsilon \quad (\text{realization})$$

Assumptions:  $\mathbb{E}(\varepsilon_i) = 0$ ,  $\varepsilon_i$  iid with  $\text{var}(\varepsilon_i) = \sigma_0^2$

Least Squares Estimator and Estimate:

$$\hat{q} = (X^T X)^{-1} X^T \Upsilon, \quad \mathbb{E}(\hat{q}) = q_0, \quad V(\hat{q}) = \sigma_0^2 (X^T X)^{-1},$$

$$q = (X^T X)^{-1} X^T v$$

Error Variance Estimator and Estimate:  $\hat{R} = \Upsilon - X\hat{q}$ ,  $R = v - Xq$

$$\hat{\sigma}^2 = \frac{1}{n-p} \hat{R}^T \hat{R}, \quad \sigma^2 = \frac{1}{n-p} R^T R$$

Covariance Matrix Estimator and Estimate:

$$V(\hat{q}) = \hat{\sigma}^2 (X^T X)^{-1}, \quad V = \sigma^2 (X^T X)^{-1}$$

Sampling Distribution: Requires  $\varepsilon_i \sim N(0, \sigma_0^2)$  or sufficiently large  $n$

- $\hat{q} \sim N(q_0, \sigma_0^2 (X^T X)^{-1})$
- $(1 - \alpha) \times 100\%$  Confidence Intervals:  $\delta_k = [(X^T X)^{-1}]_{kk}$   

$$\left[ q_k - t_{n-p, 1-\alpha/2} \sigma \sqrt{\delta_k}, \quad q_k + t_{n-p, 1-\alpha/2} \sigma \sqrt{\delta_k} \right]$$

**Table 7.2.** Statistical model, estimators, and statistical properties of the linear regression model. As noted in Remark 7.3,  $\hat{q} = \hat{q}_{OLS}$  and  $q = q_{OLS}$  are the OLS estimator and estimate.

As noted in Section 7.1, the OLS estimate for the scalar case is obtained by minimizing the functional

$$\mathcal{J}(q) = \sum_{i=1}^n [v_i - f_i(q)]^2 \tag{7.30}$$

subject to  $q \in \mathbb{Q}$ .

The difficulty is that analytic expressions for these minimizers generally cannot be obtained for nonlinearly parameterized problems. Instead, estimates must be obtained by minimizing the least squares functional. Rather than provide a detailed analysis of the nonlinear problem, we summarize results that are analogous to the linear theory and refer readers to [24, 26, 219] for details regarding the nonlinear problem.

### 7.3.1 Parameter and Error Variance Estimators—Scalar Observations

**Assumption 7.13.** To construct parameter and error variance estimators, we require  $\varepsilon_i$  to be iid with zero mean and fixed but unknown variance  $\sigma_0^2$ . With this assumption, it follows that  $\mathbb{E}(\Upsilon_i) = f_i(q_0)$  and  $\text{var}(\Upsilon_i) = \sigma_0^2$ .

#### Parameter Estimator and Estimate

Unlike the linear case, which can be solved explicitly using the normal equations, the determination of an OLS estimator and estimate,

$$\hat{q}_{OLS} = \underset{q \in \mathcal{Q}}{\operatorname{argmin}} \sum_{i=1}^n [\Upsilon_i - f_i(q)]^2, \quad q_{OLS} = \underset{q \in \mathbb{Q}}{\operatorname{argmin}} \sum_{i=1}^n [v_i - f_i(q)]^2, \quad (7.31)$$

requires numerical optimization techniques. The restriction  $q \in \mathbb{Q}$  can produce constraints that must be enforced during optimization.

It was noted in Example 3.3 that parameter values for physical or biological problems can easily vary over 10 orders of magnitude. The direct optimization of (7.30) using standard software will be highly inefficient or fail for such problems. To address this, we employ scaled parameters  $q_s = q./s$ , where  $./$  denotes componentwise division and  $s$  is a vector whose components are the scale or magnitude of each parameter. Point estimates for the scaled parameters are then given by

$$q_{OLS} = \underset{q_s \in \mathbb{Q}_s}{\operatorname{argmin}} \sum_{i=1}^n [v_i - f_i(q_s \times s)]^2, \quad (7.32)$$

where  $\times$  denotes componentwise multiplication and  $\mathbb{Q}_s$  is the scaled admissible parameter space. We employ (7.32) for physical problems where the magnitude of parameters vary significantly.

**Remark 7.14.** As noted in Remark 7.3, we will consider OLS estimators and estimates in this chapter. To simplify notation, we thus take  $\hat{q} = \hat{q}_{OLS}$  and  $q = q_{OLS}$  for the remainder of the discussion.

One approach for obtaining least squares estimates is to employ stochastic optimization techniques such as genetic algorithms, simulated annealing, and differential evolution [232]. These techniques reduce the reliance on accurate initial parameter estimates and, in theory, provide global convergence. However, their convergence rates are slower—they may require infinite time for convergence—and, because they are nondeterministic, multiple optimizations can yield varying final parameter values.

Alternatively, one can employ gradient-based methods such as the interior-reflective Newton, Levenberg–Marquardt, or sequential quadratic programming algorithms employed in the MATLAB routines `lsqnonlin` and `fmincon`. The efficiency and success of gradient-based optimization methods are predicated on determining good initial parameter estimates and being able to accurately determine

gradients. The advantage of gradient-based methods is that once they are near the minimum, they can exhibit quadratic convergence rates, which is vastly more efficient than stochastic optimization techniques. We note that one alternative is to employ the hybrid approaches in which the stochastic techniques are used to provide reasonable initial estimates for the gradient-based algorithms which then provide fast convergence to final parameter estimates.

### Parameter Estimator Mean and Variance

For the linear model with design matrix  $X$ , we showed in (7.19) that  $\mathbb{E}(\hat{q}) = q_0$  and  $V(\hat{q}) = \sigma_0^2(X^T X)^{-1}$ . In the nonlinear theory, linearization about  $q_0$  yields the approximate covariance relations

$$V(\hat{q}) \approx \sigma_0^2 [\mathcal{X}^T(q_0)\mathcal{X}(q_0)]^{-1} \approx \hat{\sigma}^2 [\mathcal{X}^T(q)\mathcal{X}(q)]^{-1}. \quad (7.33)$$

Here  $\mathcal{X}(q)$  denotes the  $n \times p$  sensitivity matrix whose elements are

$$\mathcal{X}_{ik}(q) = \frac{\partial f_i(q)}{\partial q_k}. \quad (7.34)$$

### Sensitivity Matrix Construction

The sensitivity matrix can be constructed using three techniques: (i) finite difference approximations, (ii) solution of sensitivity equations, or (iii) automatic differentiation. Ideally, one would compare matrices resulting from at least two of the methods to verify results.

The simplest conceptually is to approximate the derivatives using finite difference relations

$$\mathcal{X}_{ik}(q) = \frac{\partial f_i(q)}{\partial q_k} \approx \frac{f_i(q + h_k) - f_i(q)}{|h_k|}, \quad (7.35)$$

where  $h_k$  is a  $p$ -vector having a nonzero  $k^{th}$  element. The difficulty is that the accuracy of (7.35) is highly dependent on the choice of  $h_k$ , which also must be correctly scaled according to the magnitude of  $q$ . Hence the accuracy of results should be verified through comparison with the other techniques.

Sensitivity equations can be constructed using various techniques. In Chapter 14, we illustrate their formulation using Gâteaux differentials. More formally, they can be constructed by differentiating the evolution equation  $\frac{du}{dt} = g(t, u(t), q)$  with respect to the components  $q_k$  of  $q$ , and switching the order of integration, to obtain

$$\frac{\partial u_{q_k}}{\partial t} = \frac{\partial g}{\partial u} u_{q_k} + \frac{\partial g}{\partial q_k}, \quad (7.36)$$

where  $u_{q_k} \equiv \frac{\partial u}{\partial q_k}$ . The matrix component  $\mathcal{X}_{ik}(q) = \mathcal{C} \frac{\partial u(t_i, q)}{\partial q_k}$  is easily constructed once one has numerically integrated (7.36) to obtain  $u_{q_k}(t_i, q)$ . This approach has the advantage that it eliminates the uncertainty associated with choosing stepsizes  $h_k$  to provide accurate finite difference approximations. However, if the original system has  $N$  differential equations, the solution of (7.36) will involve  $N \cdot p$  additional

differential equations. Moreover, the analytic differentiation of the original system to construct the sensitivity equations is often difficult for complex systems.

For certain problems, automatic differentiation (AD) codes can be used to construct the sensitivity equations in a form that can be directly incorporated in ODE software. In such cases, the use of AD software to construct the sensitivity matrix  $\mathcal{X}(q)$  can avoid the inaccuracy associated with finite difference approximations and the potential for errors when formulating and solving the sensitivity equations.

### Error Variance Estimator

Since the error variance  $\sigma_0^2$  in (7.33) is unknown, we construct a variance estimator analogous to that in the linear case. Specifically, we consider the unbiased variance estimator and estimate

$$\hat{\sigma}^2 = \frac{1}{n-p} \hat{R}^T \hat{R} \quad , \quad \sigma^2 = \frac{1}{n-p} R^T R, \quad (7.37)$$

where  $\hat{R} = \Upsilon_i - f_i(\hat{q})$  and  $R = v_i - f_i(q)$  are the residual estimator and estimate. This yields the estimate

$$V = \sigma^2 [\mathcal{X}^T(q) \mathcal{X}(q)]^{-1} \quad (7.38)$$

for the covariance matrix.

### Sampling Distribution

To specify a sampling distribution for  $\hat{q}$ , we again require either Assumption 7.7, which stipulates that errors are iid and  $\varepsilon \sim N(0, \sigma_0^2)$ , or that  $n$  is sufficiently large that we can invoke the central limit theorem in the sense of Property 7.9. This directly or asymptotically establishes that

$$\hat{q} \sim N \left( q_0, \sigma_0^2 [\mathcal{X}^T(q_0) \mathcal{X}(q_0)]^{-1} \right), \quad (7.39)$$

where the covariance matrix is approximated by (7.38).

### Confidence Intervals

The construction of  $(1 - \alpha) \times 100\%$  confidence intervals is analogous to the formulation (7.27) or (7.28) for the linearly parameterized model. If we let  $\delta_k$  denote the  $k^{th}$  diagonal element of  $[\mathcal{X}^T(q) \mathcal{X}(q)]^{-1}$ , then the  $(1 - \alpha) \times 100\%$  confidence interval is

$$\left[ q_k - t_{n-p, 1-\alpha/2} \sigma \sqrt{\delta_k}, q_k + t_{n-p, 1-\alpha/2} \sigma \sqrt{\delta_k} \right], \quad (7.40)$$

where  $\sigma$  is given by (7.37). As noted in Section 7.2.4,  $t$ -calculators or tables can be used to calculate or look up  $t_{n-p, 1-\alpha/2}$  given values of  $n, p$ , and  $\alpha$ .

The properties of the least squares estimator  $\hat{q}$  for the nonlinear statistical model (7.13) are compiled in Table 7.3. These can be compared with analogous properties for the linear regression problem summarized in Table 7.2.

Statistical Model:

$$\Upsilon = f(q_0) + \varepsilon, \quad q \in \mathbb{R}^p,$$

$$v = f(q_0) + \epsilon \text{ (realization)}$$

Assumptions:  $\mathbb{E}(\varepsilon_i) = 0$ ,  $\varepsilon_i$  iid with  $\text{var}(\varepsilon_i) = \sigma_0^2$

Least Squares Estimator and Estimate:

$$\hat{q} = \underset{q \in \mathcal{Q}}{\operatorname{argmin}} \sum_{i=1}^n [\Upsilon_i - f_i(q)]^2, \quad q = \underset{q \in \mathbb{Q}}{\operatorname{argmin}} \sum_{i=1}^n [v_i - f_i(q)]^2$$

Error Variance Estimator and Estimate:  $\hat{R} = \Upsilon - f(\hat{q})$ ,  $R = v - f(q)$

$$\hat{\sigma}^2 = \frac{1}{n-p} \hat{R}^T \hat{R}, \quad \sigma^2 = \frac{1}{n-p} R^T R$$

Covariance Matrix Estimator and Estimate:  $\mathcal{X}_{ik}(q) = \frac{\partial f_i(q)}{\partial q_k}$

$$V(\hat{q}) = \hat{\sigma}^2 [\mathcal{X}^T(\hat{q}) \mathcal{X}(\hat{q})]^{-1}, \quad V = \sigma^2 [\mathcal{X}^T(q) \mathcal{X}(q)]^{-1}$$

Statistical Properties: Requires  $\varepsilon_i \sim N(0, \sigma_0^2)$  or sufficiently large  $n$

- $\hat{q} \sim N\left(q_0, \sigma_0^2 [\mathcal{X}^T(q_0) \mathcal{X}(q_0)]^{-1}\right)$
- $(1-\alpha) \times 100\%$  Confidence Intervals:  $\delta_k = [(\mathcal{X}^T(q) \mathcal{X}(q))^{-1}]_{kk}$   

$$\left[ q_k - t_{n-p, 1-\alpha/2} \sigma \sqrt{\delta_k}, \quad q_k + t_{n-p, 1-\alpha/2} \sigma \sqrt{\delta_k} \right]$$

**Table 7.3.** Statistical model, estimators, and statistical properties of the nonlinearly parameterized model (7.13) with scalar observations. As noted in Remark 7.14,  $\hat{q} = \hat{q}_{OLS}$  and  $q = q_{OLS}$  are the OLS estimator and estimate.

**Example 7.15.** Consider the spring model

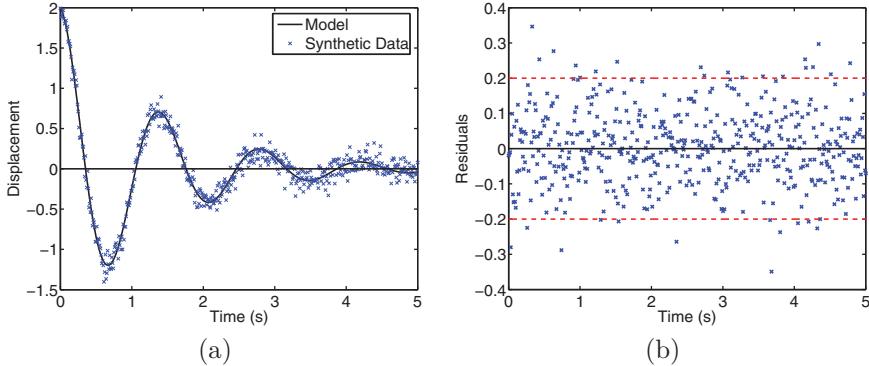
$$\begin{aligned} \ddot{z} + C\dot{z} + Kz &= 0, \\ z(0) &= 2, \quad \dot{z}(0) = -C \end{aligned} \tag{7.41}$$

with displacement observations so that

$$y = [1 \ 0] \begin{bmatrix} z \\ \dot{z} \end{bmatrix} = z.$$

We showed in Example 3.2 that (7.41) has the solution

$$z(t) = 2e^{-Ct/2} \cos\left(\sqrt{K - C^2/4} \cdot t\right) \tag{7.42}$$



**Figure 7.1.** (a) *Synthetic data and modeled displacement and (b) residuals at  $n = 501$  points.*

when  $C^2 - 4K < 0$ . We take  $K = 20.5$  to be known and let  $q = C$  be the parameter considered in the statistical analysis. We note that although the model exhibits a linear dependence on the states  $z$  and  $\dot{z}$ , the dependence of  $z(t, q)$  on  $q$  is nonlinear.

To numerically generate synthetic data, we employ  $C_0 = 1.5$  and add noise  $\varepsilon \sim N(0, \sigma_0^2)$ , where  $\sigma_0 = 0.1$ . The model and one realization of the data at  $n = 501$  points are plotted in Figure 7.1(a), and the residuals are plotted in Figure 7.1(b). By construction, the residuals are iid with 94.4% of the values lying with the  $2\sigma$  interval indicated by the horizontal lines.

The  $n \times 1$  sensitivity matrix (vector) is

$$\mathcal{X}(q) = \left[ \frac{\partial y}{\partial C}(t_1, q), \dots, \frac{\partial y}{\partial C}(t_n, q) \right]^T, \quad (7.43)$$

where

$$\frac{\partial y}{\partial C} = e^{-Ct/2} \left[ \frac{Ct}{\sqrt{4K - C^2}} \sin \left( \sqrt{K - C^2/4} \cdot t \right) - t \cos \left( \sqrt{K - C^2/4} \cdot t \right) \right] \quad (7.44)$$

results from differentiating (7.42). The construction of  $\mathcal{X}(q)$  by constructing and solving the corresponding sensitivity equations is addressed in Exercise 7.1.

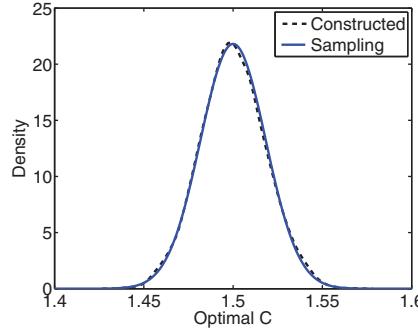
Because we know  $\sigma_0^2$ , we obtain the covariance value

$$V = \sigma_C^2 = \sigma_0^2 [\mathcal{X}^T(q) \mathcal{X}(q)]^{-1} = 3.35 \times 10^{-4}$$

so that  $\sigma_C = 0.0183$ . Since  $\varepsilon_i \sim N(0, \sigma_0^2)$ , the random variable  $\hat{C}$  has the sampling distribution

$$\hat{C} \sim N(C_0, \sigma_c^2), \quad (7.45)$$

which is plotted in Figure 7.2. The parameter estimated by minimizing (7.31) for the data plotted in Figure 7.1 is  $C = 1.4792$ , and the 95% confidence interval given by (7.40) is [1.4433, 1.5150].



**Figure 7.2.** Sampling density  $N(C_0, \sigma_C^2)$  for  $\hat{C}$  and density constructed from 10,000 simulations.

It was noted in Sections 4.8.1 and 7.1 that in frequentist inference, the 95% confidence interval has the following interpretation in the context of parameter estimation; if the procedure is repeated  $\ell$  times,  $0.95\ell$  of the computed intervals will contain the true parameter  $q_0$ . This is illustrated in Figure 4.11(a). To demonstrate for this example, we generated 10,000 sets of numerical data using the true parameter values  $C_0$  and  $\varepsilon_i \sim N(0, \sigma_0^2)$  with  $\sigma_0 = 0.1$ . For each data set, we optimized (7.31) to obtain a point estimate  $C$  and corresponding 95% confidence interval. In this set of numerical experiments, 9455 of the intervals contained  $C_0$ . Using the 10,000 estimated values of  $C$ , we used the kernel estimation techniques discussed in Section 4.1.1 to construct the density which is plotted in Figure 7.2. As expected, the kernel density estimate matches the representation (7.45) for the sampling distribution.

**Example 7.16.** We showed in Example 3.5 that the boundary value problem

$$\begin{aligned} \frac{d^2 T_s}{dx^2} &= \frac{2(a+b)}{ab} \frac{h}{k} [T_s(x) - T_{amb}], \\ \frac{dT_s}{dx}(0) &= \frac{\Phi}{k}, \quad \frac{dT_s}{dx}(L) = \frac{h}{k} [T_{amb} - T_s(L)] \end{aligned}$$

models the steady state temperature of an uninsulated rod with source heat flux  $\Phi$  at  $x = 0$  and ambient air temperature  $T_{amb}$ . The model parameters to be estimated and statistically analyzed are  $q = [\Phi, h]$ , where  $h$  is the convective heat transfer coefficient.

The rod used in these experiments was aluminum with cross-sectional dimensions  $a = b = 0.95$  cm and length  $L = 70$  cm. The temperature measurements  $v_i$ , compiled in Table 3.2, were made at 15 equally spaced spatial locations  $x_i = x_0 + (i-1)\Delta x$ , where  $x_0 = 10$  cm and  $\Delta x = 4$  cm. The observed solution is

$$y_i(q) = T_s(x_i, q) = c_1(q)e^{-\gamma x_i} + c_2(q)e^{\gamma x_i} + T_{amb},$$

where  $\gamma = \sqrt{\frac{2(a+b)h}{abk}}$  and

$$c_1(q) = -\frac{\Phi}{k\gamma} \left[ \frac{e^{\gamma L}(h + k\gamma)}{e^{-\gamma L}(h - k\gamma) + e^{\gamma L}(h + k\gamma)} \right], \quad c_2(q) = \frac{\Phi}{k\gamma} + c_1(q).$$

We suppress the parameter dependence of  $\gamma$  to clarify the notation. We employ the thermal conductivity value  $k = 2.37 \frac{W}{cm \cdot C}$  reported for aluminum and the measured ambient room temperature  $T_{amb} = 21.29^\circ C$ .

A least squares fit to the data yielded the parameter estimates  $\Phi = -18.41$  and  $h = 0.00191$ , and the model fit shown in Figure 7.3(a). We note that this value of  $h$  falls within the range  $2.8 \times 10^{-4} - 0.0023 \frac{W}{cm^2 \cdot C}$  reported for still air. The residuals plotted in Figure 7.3(b) exhibit no discernible pattern, thus motivating the assumption that the errors  $\varepsilon_i$  are iid. We assume that errors are normally distributed when constructing a sampling distribution.

The error variance estimate is  $\sigma^2 = 0.0627$ , and the covariance matrix, computed using analytic sensitivity relations, as derived in Exercise 7.4 and illustrated in Figure 7.4, is

$$V = \begin{bmatrix} 2.1034 \times 10^{-2} & -2.0286 \times 10^{-6} \\ -2.0286 \times 10^{-6} & 2.0972 \times 10^{-10} \end{bmatrix}. \quad (7.46)$$

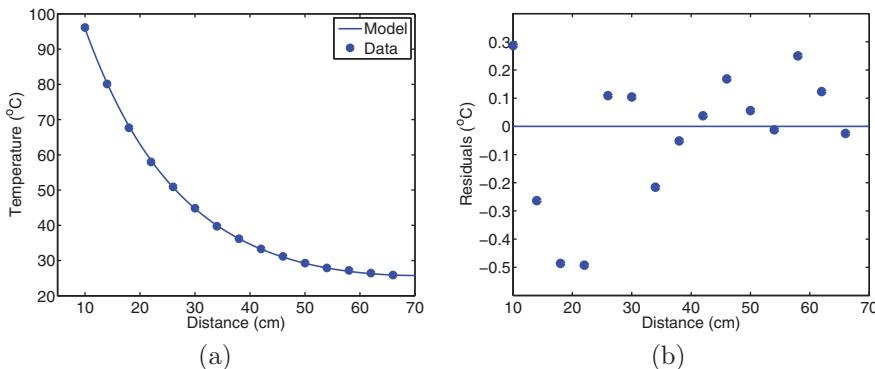
The standard deviations for the errors and sampling distribution are

$$\sigma = 0.2504, \sigma_\Phi = 0.1450, \sigma_h = 1.4482 \times 10^{-5}. \quad (7.47)$$

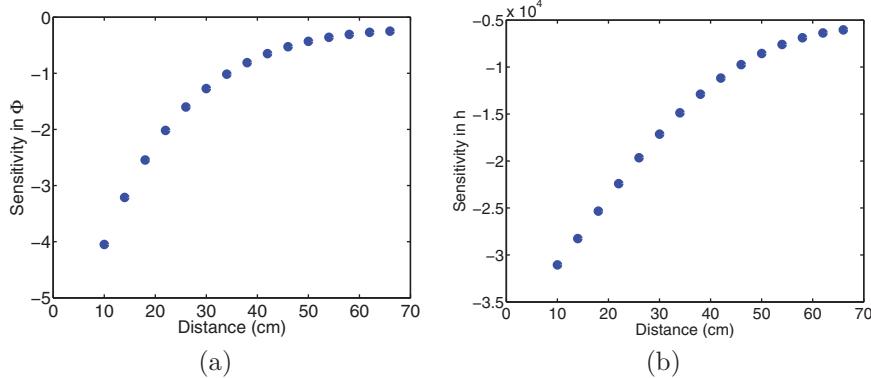
Since  $n = 15$  and  $p = 2$ , the 95% confidence intervals are

$$[-18.7233, -18.0967], [1.8787 \times 10^{-3}, 1.9413 \times 10^{-3}].$$

In Example 8.12, we revisit this example in the context of Bayesian analysis.



**Figure 7.3.** (a) Model fit to the steady state temperature data and (b) residuals at the 15 spatial locations.



**Figure 7.4.** Analytic sensitivity values: (a)  $\frac{\partial y}{\partial \Phi}(x_i, q)$  and (b)  $\frac{\partial y}{\partial h}(x_i, q)$ .

### 7.3.2 Parameter and Error Variance Estimators for Evolution Models—Multiple Responses

In this section, we consider the evolution equation (7.1) with  $\nu > 1$  data measurements and model responses specified by a  $\nu \times n$  matrix  $\mathcal{C}$ . The statistical model in this case is

$$\Upsilon_i = f(t_i, q_0) + \varepsilon_i, \quad j = 1, \dots, n,$$

where  $\Upsilon_i$  and  $\varepsilon_i$  are random  $\nu$ -vectors.

**Assumption 7.17.** To accommodate the possibility that error distributions associated with individual components of the observations could differ, we let  $\sigma_{0,j}^2$  denote the fixed but unknown variance of the error associated with the  $j^{th}$  observation. These values are compiled in the  $\nu \times \nu$  diagonal measurement error covariance matrix  $V_0 = \text{diag}[\sigma_{0,1}^2, \dots, \sigma_{0,\nu}^2]$ . As before, errors are assumed to be unbiased. We remind the reader that  $V_0$  is fixed but typically unknown.

The construction of parameter and covariance estimators is similar in theory to the scalar case  $\nu = 1$  but is complicated by the coupling induced by the potentially differing variances of the error components. We provide an overview of the estimators, estimates, and sampling distribution for  $\nu > 1$  and refer the reader to [24, 26] for details.

**Example 7.18.** It was noted in Example 3.2 that for vibrating systems modeled as a simple harmonic oscillator (3.11), displacements and velocities can be respectively measured using a proximity sensor and laser vibrometer. If both sets of measurements are available, the modeled observations will be

$$\begin{bmatrix} y_1(t_i, q) \\ y_2(t_i, q) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} z_1(t_i, q) \\ z_2(t_i, q) \end{bmatrix},$$

which is just the parameter-dependent states. Given the differing nature of the measurement devices, one would expect different error distributions to be associated

with the experimental measurements  $v_1$  and  $v_2$ . Hence we would employ

$$V_0 = \begin{bmatrix} \sigma_{01}^2 & 0 \\ 0 & \sigma_{02}^2 \end{bmatrix}. \quad (7.48)$$

**Example 7.19.** For the HIV model (3.15) of Example 3.3, one can typically measure only the total number  $T_1 + T_1^*$  of T-lymphocytes and the viral load  $V$ . Hence

$$\mathcal{C} = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

and  $y(t, q) \in \mathbb{R}^2$ . The error covariance matrix would again have the structure (7.48).

### Parameter and Error Covariance Estimators

The OLS estimator and estimate are taken to be

$$\begin{aligned} \hat{q}_{OLS} &= \underset{q \in \mathcal{Q}}{\operatorname{argmin}} \sum_{i=1}^n [\Upsilon_i - f(t_i, q)]^T V_0^{-1} [\Upsilon_i - f(t_i, q)], \\ q_{OLS} &= \underset{q \in \mathcal{Q}}{\operatorname{argmin}} \sum_{i=1}^n [v_i - f(t_i, q)]^T V_0^{-1} [v_i - f(t_i, q)], \end{aligned} \quad (7.49)$$

where  $V_0^{-1}$  weights the response components by the reciprocals of the corresponding error variance associated with each component. Since  $V_0$  is typically unknown, it too must be estimated. Motivated by (7.37), the estimate  $V \approx V_0$  is provided by the relation

$$V = \operatorname{diag} \left( \frac{1}{n-p} \sum_{i=1}^n [v_i - f(t_i, q_{OLS})] [v_i - f(t_i, q_{OLS})]^T \right). \quad (7.50)$$

Unlike the scalar response relations (7.31) and (7.37), the multiple response relations (7.49) and (7.50) are coupled due to the fact that  $V_0 \neq \sigma_0^2 I$ , and hence they must be solved as a coupled system.

### Sampling Distribution

To specify a sampling distribution, we need an assumption analogous to Assumption 7.7.

**Assumption 7.20.** Let  $\varepsilon_{ij}$  denote the error in the  $i^{th}$  component of  $\Upsilon_i$  at time  $t_i$ . We make the assumption that  $\varepsilon_{ij} \sim N(0, \sigma_{0j}^2)$  so that  $\varepsilon_i \sim N(0, V_0)$ . For  $n$  sufficiently large, the central limit theorem can be invoked in the manner detailed in Property 7.9 to obtain similar asymptotic results.

With this assumption, it is shown in [24, 26] that

$$\hat{q}_{OLS} \sim N(q_0, \mathcal{V}_0) \approx N(q_{OLS}, \mathcal{V}),$$

where

$$\mathcal{V}_0 \approx \left( \sum_{j=1}^n \mathcal{X}_j^T(q_0) V_0^{-1} \mathcal{X}_j(q_0) \right)^{-1}$$

is the  $p \times p$  covariance matrix and

$$\mathcal{X}_j(q) = \begin{bmatrix} \frac{\partial f_1(t_i, q)}{\partial q_1} & \dots & \frac{\partial f_1(t_i, q)}{\partial q_p} \\ \vdots & & \vdots \\ \frac{\partial f_\nu(t_i, q)}{\partial q_1} & \dots & \frac{\partial f_\nu(t_i, q)}{\partial q_p} \end{bmatrix} \quad (7.51)$$

is the  $\nu \times p$  sensitivity matrix at time  $t_i$ . For implementation,  $\mathcal{V}_0$  is approximated by

$$\mathcal{V} = \left( \sum_{j=1}^n \mathcal{X}_j^T(q_{OLS}) V^{-1} \mathcal{X}_j(q_{OLS}) \right)^{-1},$$

where (7.51) must be evaluated at each time step. The  $(1 - \alpha) \times 100\%$  confidence intervals are

$$[q_{OLS,k} - t_{n-p,1-\alpha/2} SE, q_{OLS,k} + t_{n-p,1-\alpha/2} SE],$$

where  $q_{OLS,k}$  is the  $k^{th}$  element of  $q_{OLS}$  and the standard error is

$$SE \approx \sqrt{\mathcal{V}_k}.$$

Here  $\mathcal{V}_k$  is the  $k^{th}$  diagonal element of  $\mathcal{V}$ .

## 7.4 Notes and References

The parameter estimation techniques discussed in this chapter are based on linear and nonlinear regression for which there are numerous excellent texts. The text [96] provides a very nice introduction to linear regression and has the advantage that the authors use different notation to delineate between random variables and their realizations. This is also a good resource for obtaining additional background regarding the confidence and prediction intervals discussed in Chapter 9. Asymptotic theory for nonlinear regression problems is detailed in the classic book [219]. We refer readers to [24, 26] for details regarding the construction of estimators and specification of sampling distributions for parameters in nonlinear evolution models.

For brevity, we do not discuss the following topics: infinite-dimensional inverse problems associated with parameter estimation, regularization, or optimization methods for inverse problems. The reader is referred to [25] for theory and estimation techniques for distributed parameter systems and [15, 128, 176, 244, 256] for details regarding regularization, computational algorithms, and case studies pertaining to parameter estimation and inverse problems. The texts [64, 131, 132, 232] cover a variety of optimization techniques that are appropriate for this class of problems.

## 7.5 Exercises

**Exercise 7.1.** Consider the unforced spring model (7.41) with displacement observations. Construct and solve the sensitivity equations for the damping parameter  $C$  and stiffness parameter  $K$ . Show that your observed solutions are the same as those obtained by differentiating the solution  $y(t, q)$ .

**Exercise 7.2.** For the spring model (7.41), approximate the time-dependent sensitivities using the finite difference relations (7.35) and compare your solutions to those obtained in Exercise 7.1 for various stepsizes  $h_K$  and  $h_C$ .

**Exercise 7.3.** Consider the unforced spring model (7.41) with displacement observations and initial conditions  $z(0) = z_0$ ,  $\dot{z}(0) = -C$ . Construct and solve the sensitivity equations for the initial condition  $z_0$ . Compare your answer to the solution obtained by differentiating  $y(t, q)$  with respect to  $z_0$ .

**Exercise 7.4.** Compute the analytic sensitivity relations  $\frac{\partial y}{\partial \Phi}$  and  $\frac{\partial y}{\partial h}$  for the steady state heat model in Example 7.16. Plot your solutions at the points  $x_i$  specified in the example.

**Exercise 7.5.** Use the finite difference expression (7.35) to approximate the sensitivity relations  $\frac{\partial y}{\partial \Phi}$  and  $\frac{\partial y}{\partial h}$  for the model in Example 7.16. Compare your solutions to the analytic sensitivity relations developed in Exercise 7.4 for various stepsizes. Discuss criteria that should be used to specify stepsizes.

**Exercise 7.6.** Repeat the analysis and numerical experiments detailed in Example 7.15 for the spring model

$$\ddot{z} + 0.15\dot{z} + Kz = 0,$$

$$z(0) = 2, \quad \dot{z}(0) = z_1$$

with displacement observations. Hence the parameters to be estimated are  $q = [K, z_1]$ .

**Exercise 7.7.** Repeat the analysis of Example 7.16 for the steady state heat model using the copper data in Table 3.3. Do your residuals appear to be iid? We will revisit this problem in Chapter 12.

**Exercise 7.8.** In this problem, we will model heat generated during the hardening of cement. Data from [104] is compiled in Table 7.4. Here  $v$  denotes heat with units of calories/gram cement and  $x_1-x_4$  respectively denote the percentage of tricalcium aluminate, tricalcium silicate, tetracalcium aluminoferrite, and dicalcium phosphate.

(a) Consider first the linear model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon.$$

Estimate the parameters, plot the residual, and determine confidence intervals of two standard deviations as well as 95% confidence intervals.

(b) Perform the same analysis using linear models that incorporate only  $x_1$  as well as  $x_1$  and  $x_2$ . How do your results compare with those obtained in (a)?

| Obs. No. | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $v$   |
|----------|-------|-------|-------|-------|-------|
| 1        | 7     | 26    | 6     | 60    | 78.5  |
| 2        | 1     | 29    | 15    | 52    | 74.3  |
| 3        | 11    | 56    | 8     | 20    | 104.3 |
| 4        | 11    | 31    | 8     | 47    | 87.6  |
| 5        | 7     | 52    | 6     | 33    | 95.9  |
| 6        | 11    | 55    | 9     | 22    | 109.2 |
| 7        | 3     | 71    | 17    | 6     | 102.7 |
| 8        | 1     | 31    | 22    | 44    | 72.5  |
| 9        | 2     | 54    | 18    | 22    | 93.1  |
| 10       | 21    | 47    | 4     | 26    | 115.9 |
| 11       | 1     | 40    | 23    | 34    | 83.8  |
| 12       | 11    | 66    | 9     | 12    | 113.3 |
| 13       | 10    | 68    | 8     | 12    | 109.4 |

**Table 7.4.** Cement data from [104].

## Chapter 8

# Bayesian Techniques for Parameter Estimation

For applications where modeling and measurement errors  $\varepsilon_i$  are unbiased and iid, we employ the statistical model

$$\Upsilon_i = f_i(Q) + \varepsilon_i, \quad i = 1, \dots, n, \tag{8.1}$$

where  $\Upsilon_i$ ,  $\varepsilon_i$ , and  $Q$  are random variables representing measurements, measurement errors, and parameters. As defined in (7.9),  $f_i(Q)$  denotes the parameter-dependent model response. We note that the measurement errors in this case are modeled as additive and mutually independent from  $Q$ . We also remind readers that calibration parameters and observed data are commonly denoted by  $\theta$  and  $y$  in the statistics literature.

## 8.1 Parameter Estimation from a Bayesian Perspective

As detailed in Section 4.8, the tenets of Bayesian inference differ significantly from the frequentist perspective described in Chapter 7. In the context of inverse problems involving parameter estimation, the Bayesian approach can be summarized as follows. Parameters are considered to be random variables  $Q$  with realizations  $q = Q(\omega)$  and associated densities that incorporate known information or information obtained as measurements are acquired. The solution of the inverse problem is the posterior density that best reflects the distribution of parameter values based on the sampled observations.

It was shown in Section 4.8.2 that the posterior is constructed in terms of a prior density and likelihood. The prior density  $\pi_0(q)$  incorporates any knowledge that we have about parameters prior to obtaining observations  $v$ . This could come from previous similar experiments or analysis regarding similar models. It was illustrated that if prior knowledge is of questionable accuracy, it is better to use a noninformative prior which is often taken as an improper uniform density posed on the parameter support; for example, one would employ  $\pi_0(q) = \chi_{(0,\infty)}(q)$  for positive parameters.

The likelihood function  $\pi(v|q) = L(q|v)$  incorporates information provided by the samples and constitutes the mechanism through which data informs the posterior density. As detailed in Section 4.3.2, the likelihood quantifies the probability of obtaining the observations  $v$  for a given value  $q$  of the parameter  $Q$ . Hence if we let  $\pi(q, v)$  denote the joint density of  $Q$  and  $\Upsilon$ , then the likelihood

$$\pi(v|q) = \frac{\pi(q, v)}{\pi_0(q)}$$

is the conditional probability of  $\Upsilon$  given a value of  $Q$ .

Once we have a measurement or observation  $v = v_{obs}$ , the conditional density

$$\pi(q|v_{obs}) = \frac{\pi(q, v_{obs})}{\pi(v_{obs})},$$

where we assume that

$$\pi(v_{obs}) = \int_{\mathbb{R}^p} \pi(q, v_{obs}) dq = \int_{\mathbb{R}^p} \pi(v_{obs}|q) \pi_0(q) dq \neq 0,$$

is the posterior density. The inverse problem in the Bayesian framework can thus be stated as follows: given measurements  $v_{obs}$ , find the posterior density  $\pi(q|v_{obs})$ . The complete formulation, which the authors of [128] refer to as *Bayes' theorem of inverse problems*, can be stated as follows.

**Result 8.1 (Bayes' Theorem of Inverse Problems).** We assume that the  $p$  random parameter variables  $Q$  have a known prior density  $\pi_0(q)$ , which can be noninformative, and we let  $v_{obs}$  be a realization of the random observation variable  $\Upsilon$ . The posterior density of  $Q$ , given the measurements  $v_{obs}$ , is

$$\pi(q|v_{obs}) = \frac{\pi(v_{obs}|q)\pi_0(q)}{\pi(v_{obs})} = \frac{\pi(v_{obs}|q)\pi_0(q)}{\int_{\mathbb{R}^p} \pi(v_{obs}|q)\pi_0(q) dq}. \quad (8.2)$$

When using (8.2), one implicitly assumes that observed data is used to construct the posterior density; hence we write  $v = v_{obs}$  in subsequent discussion so that (8.2) is the same as (4.41).

### 8.1.1 Likelihood Function

The specification of the likelihood function  $\pi(v|q)$  depends on the assumptions made regarding the distribution of errors. In Section 4.3.2, we showed that if we employ the statistical model (8.1) with the assumption that errors are iid and  $\varepsilon_i \sim N(0, \sigma^2)$ , where  $\sigma^2$  is fixed, then the likelihood function is

$$\pi(v|q) = L(q, \sigma^2|v) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-SS_q/2\sigma^2}, \quad (8.3)$$

where

$$SS_q = \sum_{i=1}^n [v_i - f_i(q)]^2 \quad (8.4)$$

is the sum of squares error. The construction of likelihoods for other error models, including multiplicative noise, is addressed in [128].

### 8.1.2 Maximum a Posteriori (MAP) Estimate

The posterior density  $\pi(q|v)$  provides the complete distribution of  $Q$  based on the observations  $v$ . From this, point estimates for the parameter values are provided by the mean, median, or mode. The latter is defined as the parameter value that maximizes  $\pi(q|v)$ . This value, termed the MAP estimate, is given by

$$q_{MAP} = \operatorname{argmax}_q \pi(q|v).$$

Since the normalization constant  $\pi(v)$  does not affect the maximizing argument, an equivalent formulation is

$$q_{MAP} = \operatorname{argmax}_q \pi(v|q)\pi_0(q). \quad (8.5)$$

For a uniform prior on  $\mathbb{R}$ ,  $q_{MAP}$  is thus equivalent to the maximum likelihood estimate  $q_{MLE}$  defined in (4.28). As detailed in Section 4.3.2, one would typically employ the log-likelihood function  $\ell(q, \sigma|v)$  in such cases since it facilitates optimization by eliminating the exponential.

### 8.1.3 Implementation Techniques

The formulation of the inverse problem in the Bayesian framework is concisely provided by Result 8.1. However, the implementation of (8.2) is extremely challenging if the dimensionality  $p$  of  $Q$  is large, as is often the case for physical or biological models. As illustrated in the next example, classical tensored quadrature rules can be applied for low dimensionality; e.g.,  $p \leq 6$ . Within the last ten years, significant research has focused on the development of adaptive sparse grid quadrature techniques for moderate dimensionality and Monte Carlo techniques for high dimensions. These techniques are discussed in Chapter 11.

Alternatively, one can construct Markov chains whose stationary distribution is the posterior density. We discuss Markov chain Monte Carlo (MCMC) techniques in Section 8.2.

**Example 8.2.** Consider the spring model

$$\begin{aligned} \ddot{z} + C\dot{z} + Kz &= 0, \\ z(0) = 2, \quad \dot{z}(0) &= -C, \end{aligned}$$

which, for  $C^2 - 4K < 0$ , has the solution

$$z(t) = 2e^{-Ct/2} \cos(\sqrt{K - C^2/4} \cdot t).$$

We assume displacement evaluation so that  $y(t_i, Q) = z(t_i, Q)$ . We consider  $K = 20.5$  to be known and treat  $Q = C$  as the unknown parameter to be estimated. To construct synthetic data, we take  $C_0 = 1.5$  and construct iid errors  $\varepsilon_i \sim N(0, \sigma_0^2)$ , where  $\sigma_0 = 0.1$ .

For this error distribution, the likelihood is given by (8.3). We employ the non-informative prior  $\pi_0(q) = \chi_{[0,\infty)}(q)$  to enforce  $C$  to be nonnegative. The posterior density can thus be expressed as

$$\pi(q|v) = \frac{e^{-SS_q/2\sigma_0^2}}{\int_0^\infty e^{-SS_\zeta/2\sigma_0^2} d\zeta} = \frac{1}{\int_0^\infty e^{-(SS_\zeta - SS_q)/2\sigma_0^2} d\zeta},$$

where  $SS_q$  is defined in (8.4) and  $SS_\zeta$  denotes the sum of squares defined in terms of the integration variable. The second formulation is necessary to avoid numerical  $\frac{0}{0}$  evaluation since  $e^{-SS_{q_{MAP}}} \approx 3 \times 10^{-113}$ . The use of a midpoint rule to approximate the integral yields

$$\pi(q|v) \approx \frac{1}{\sum_{i=1}^k e^{-(SS_{\zeta^i} - SS_q)/2\sigma_0^2} w^i}, \quad (8.6)$$

where  $\zeta^i, w^i$  respectively denote the quadrature points and weights.

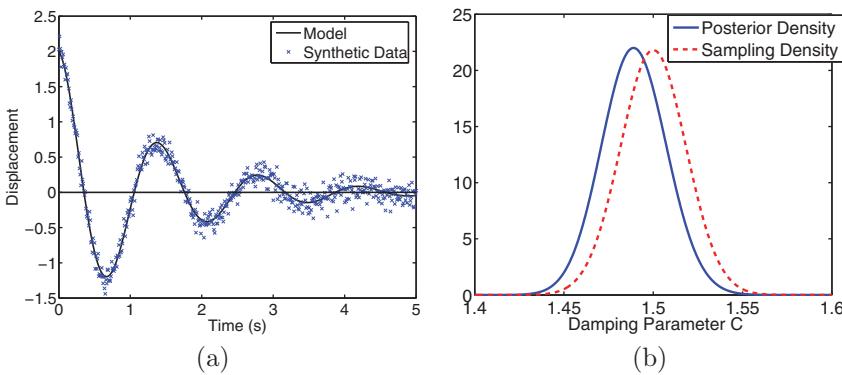
We generated one set of synthetic data  $v_i, i = 1, \dots, 501$ , which is plotted in Figure 8.1(a) along with the model response  $f_i(q_0)$ . The posterior density given by (8.6) is plotted in Figure 8.1(b). We note that the MAP estimate is  $q_{MAP} = 0.1489$ . Since we have employed a noninformative prior, this corresponds to the MLE. For the assumed error distribution, it also corresponds to the OLS estimate.

We showed in Chapter 7 that the OLS estimator has the sampling distribution

$$\hat{q}_{OLS} = \hat{C}_{OLS} \sim N(C_0, \sigma_0^2 [\mathcal{X}^T(C_0)\mathcal{X}(C_0)]^{-1}), \quad (8.7)$$

where  $\mathcal{X}(C_0)$  is given in (7.43) of Example 7.15. The sampling distribution is compared with the posterior density in Figure 8.1(b). We note that they have the same shape but the sampling distribution is centered at  $C_0$ .

We will revisit this example in Example 8.7, where we construct  $\pi(q|v)$  using MCMC methods.



**Figure 8.1.** (a) Synthetic data  $v_i$  and model response  $f_i(q_0)$ . (b) Posterior density and sampling distribution (8.7).

## 8.2 Markov Chain Monte Carlo (MCMC) Techniques

The evaluation of the posterior relation (8.2) using quadrature techniques requires the evaluation of densities over the region of  $\mathbb{R}^p$  where the posterior is defined. For moderate  $p$ , this necessitates the use of the sparse grid quadrature techniques discussed in Chapter 11, whereas Monte Carlo integration techniques are required for large dimensionality  $p$ . The difficulty of this approach is exacerbated by the fact that the support of the density is often part of the information that we are seeking.

An alternative is the following. Rather than using quadrature or Monte Carlo algorithms to specify parameter values at which we evaluate the density, we can use attributes of the density to specify parameter values that adequately explore the geometry of the distribution. This is achieved by constructing Markov chains whose stationary distribution, as defined in Definition 4.52, is the posterior density. By evaluating realizations of the chain, one thus samples the posterior and hence obtains a density for the parameter values based on observed measurements. This is the basis for the MCMC techniques employed here.

In Section 8.3, we summarize the Metropolis and Metropolis–Hastings algorithms and motivate their structure. The detailed balance condition defined in Definition 4.62 is used in Section 8.4 to establish that  $\pi(q|v)$  is the stationary distribution for the chain. We also discuss convergence criteria in that section. The role of parameter identifiability is discussed in Section 8.5, and the development of the delayed rejection adaptive Metropolis (DRAM) algorithm is detailed in Section 8.6. This is the algorithm that we employ in subsequent chapters. The DiffeRential Evolution Adaptive Metropolis (DREAM) algorithm is summarized in Section 8.7. The reader is referred to Section 4.6 for relevant definitions and theory pertaining to Markov chains.

## 8.3 Metropolis and Metropolis–Hastings Algorithms

Recall from Definition 4.50 that a Markov chain is a sequence of  $S$ -valued random variables that satisfy the Markov property that  $X_k$  depends only on  $X_{k-1}$ . The state space in this case is the set of possible parameter values, so we will be constructing chains based on parameters chosen according to the following strategy.

**Strategy 8.3.** Consider the parameter  $q^{k-1} \in \mathbb{R}^p$  to be specified.

- (i) Take the current chain realization to be  $X_{k-1} = q^{k-1}$ .
- (ii) Propose a new value  $q^* \sim J(q^*|q^{k-1})$ , where  $J$  is called the proposal or jumping distribution. The notation indicates that  $J$  specifies  $q^*$  based on the previous value  $q^{k-1}$ , and  $J(q^*|q^{k-1})$  should not be interpreted as a conditional density.
- (iii) With probability  $\alpha(q^*|q^{k-1})$ , determined by properties of the likelihood function and prior density, accept  $q^*$ ; i.e.,  $X_k = q^*$ . Otherwise, take  $X_k = q^{k-1}$ . We note that  $\alpha(q^*|q^{k-1})$  is not a conditional probability but rather specifies the probability of accepting  $q^*$  generated from the previous value  $q^{k-1}$ .
- (iv) Establish that the posterior density is the stationary distribution for the chain.

### 8.3.1 Metropolis Algorithm

We consider first the case when the proposal distribution is taken to be symmetric in the sense that  $J(q^*|q^{k-1}) = J(q^{k-1}|q^*)$ . We consider two possibilities for the proposal distribution:

$$\begin{aligned} J(q^*|q^{k-1}) &= N(q^{k-1}, V), \\ J(q^*|q^{k-1}) &= N(q^{k-1}, D). \end{aligned} \quad (8.8)$$

Here  $V$  is the covariance matrix for  $Q$ , whereas  $D$  is a diagonal matrix whose elements reflect the scale associated with each parameter value. The symmetry in the first case follows since

$$\begin{aligned} J(q^*|q^{k-1}) &= \frac{1}{\sqrt{(2\pi)^p |V|}} e^{-\frac{1}{2}[(q^* - q^{k-1})V^{-1}(q^* - q^{k-1})^T]} \\ &= \frac{1}{\sqrt{(2\pi)^p |V|}} e^{-\frac{1}{2}[(q^{k-1} - q^*)V^{-1}(q^{k-1} - q^*)^T]} \\ &= J(q^{k-1}|q^*). \end{aligned}$$

The analysis of the second choice is similar. We will provide further motivation for these choices after we summarize the algorithm.

#### Algorithm 8.4 (Metropolis Algorithm).

1. Initialization: Choose an initial parameter value  $q^0$  that satisfies  $\pi(q^0|v) > 0$ .
2. For  $k = 1, \dots, M$ 
  - (a) For  $z \sim N(0, 1)$ , construct the candidate

$$q^* = q^{k-1} + Rz,$$

where  $R$  is the Cholesky decomposition of  $V$  or  $D$ . As specified in Theorem 4.23, this ensures that

$$q^* \sim N(q^{k-1}, V) \text{ or } q^* \sim N(q^{k-1}, D).$$

Because the construction of  $q^*$  takes into account  $q^{k-1}$ , this is termed a *random walk* or *local Metropolis algorithm*.

- (b) Compute the ratio

$$r(q^*|q^{k-1}) = \frac{\pi(q^*|v)}{\pi(q^{k-1}|v)} = \frac{\pi(v|q^*)\pi_0(q^*)}{\pi(v|q^{k-1})\pi_0(q^{k-1})}. \quad (8.9)$$

- (c) Set

$$q^k = \begin{cases} q^* & , \text{ with probability } \alpha = \min(1, r), \\ q^{k-1} & , \text{ else.} \end{cases}$$

That is, we accept  $q^*$  with probability 1 if  $r \geq 1$  and we accept it with probability  $r$  if  $r < 1$ .

We first motivate the choice of acceptance criteria in steps 2(b) and (c). The first observation is that by forming the ratio of the posterior densities, we eliminate the normalization constant, which is difficult to compute when  $p$  is moderate or large. Now consider the case of a uniform prior and iid and normally distributed errors so that the likelihood is

$$\pi(v|q) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-SS_q/2\sigma^2}, \quad SS_q = \sum_{i=1}^n [v_i - f_i(q)]^2, \quad (8.10)$$

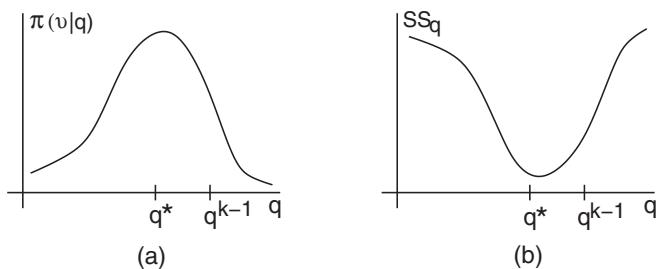
as established in (8.3) and (8.4). With these assumptions

$$r(q^*|q^{k-1}) = \frac{\pi(v|q^*)}{\pi(v|q^{k-1})} = \frac{e^{-SS_{q^*}/2\sigma^2}}{e^{-SS_{q^{k-1}}/2\sigma^2}} = e^{-[SS_{q^*} - SS_{q^{k-1}}]/2\sigma^2}, \quad (8.11)$$

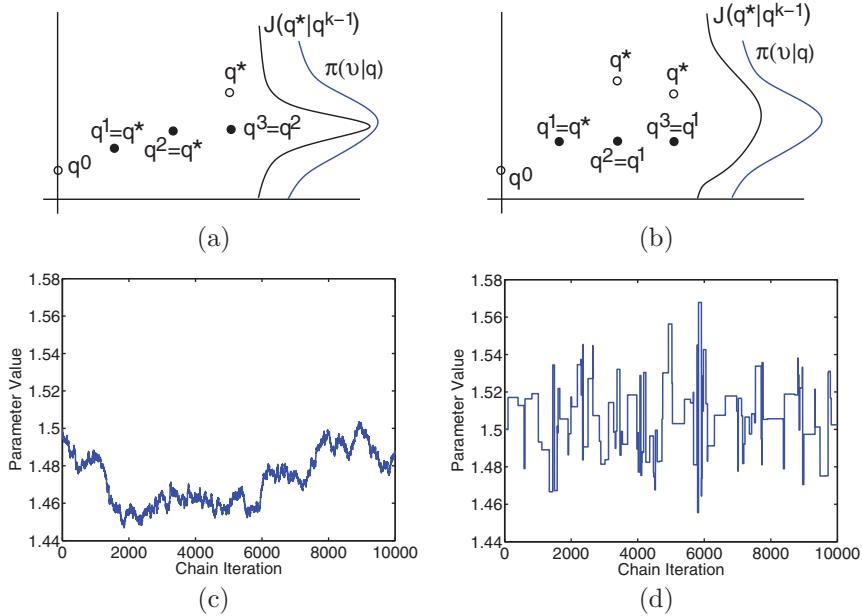
where the final step eliminates the potential for numerical  $\frac{0}{0}$  evaluation, as noted in Example 8.2. As illustrated in Figure 8.2, a candidate  $q^*$  that yields  $\pi(v|q^*) > \pi(v|q^{k-1})$  is equivalent to producing a smaller sum of squares error and this candidate is accepted with probability one. If  $q^*$  is such that  $\pi(v|q^*) < \pi(v|q^{k-1})$ , and hence the sum of squares error is increased, we accept the candidate with probability  $\alpha = r$ .

Properties of the proposal function and how they affect mixing are illustrated in Figures 8.3 and 8.4. If the variance is too large, a large percentage of the candidates will be rejected since they will have smaller likelihoods, and hence the chain will stagnate for long periods. The acceptance ratio will be high if the variance is small, but the algorithm will be slow to explore the parameter space.

As illustrated in Figure 8.4(a), if the posterior is highly anisotropic but the proposal distribution is isotropic, the efficiency with which the algorithm explores with respect to various components of the parameter vector will be highly nonuniform. The choices (8.8) for  $J(q^*|q^{k-1})$  address these issues by scaling the variability of each parameter component in the manner depicted in Figure 8.4(b). The goal is to achieve, to the degree possible, the efficiency of the univariate case.



**Figure 8.2.** (a) Likelihood and (b) sum of squares functions of  $Q$ .



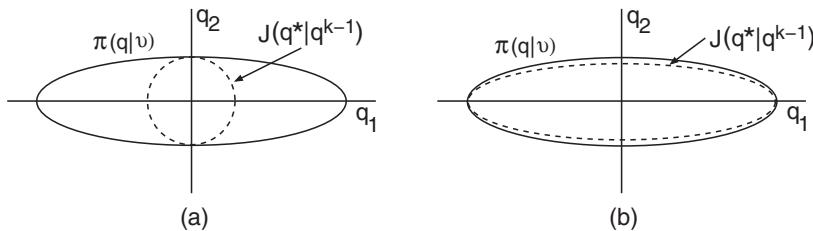
**Figure 8.3.** Generation of candidates  $q^*$  based on (a) narrow and (b) wide proposal functions  $J(q^*|q^{k-1})$ . Chains resulting from proposal functions that are (c) too narrow and (d) too wide.

The covariance matrix  $V$  is estimated in the manner detailed in Section 7.3. Specifically, we take

$$V = \sigma_{OLS}^2 [\mathcal{X}^T(q_{OLS})\mathcal{X}(q_{OLS})]^{-1},$$

$$\sigma_{OLS}^2 = \frac{1}{n-p} \sum_{i=1}^n [v_i - f_i(q_{OLS})]^2,$$
(8.12)

where  $\mathcal{X}_{ik}(q) = \frac{\partial f_i(q)}{\partial q_k}$ , as indicated in Table 7.3 on page 146.



**Figure 8.4.** Anisotropic posterior  $\pi(q|v)$  and (a) isotropic and (b) anisotropic proposal functions  $J(q^*|q^{k-1})$ .

### 8.3.2 Sample-Based Error Variance

The assumption that errors are iid and  $\varepsilon_i \sim N(0, \sigma^2)$  yields the likelihood function (8.10) and acceptance ratio (8.11) formulated in terms of  $\sigma^2$ . In most applications, however,  $\sigma^2$  is fixed but unknown. One solution is to employ the estimate (8.12) for  $\sigma^2$ . Alternatively one can treat it as an additional random parameter whose density is sampled through realizations of the Markov chain.

As illustrated in Example 4.69, the likelihood

$$\pi(v, q | \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-SS_q/2\sigma^2}$$

is in the inverse-gamma family, detailed in Definition 4.14, so the conjugate prior is

$$\pi_0(\sigma^2) \propto (\sigma^2)^{-(\alpha+1)} e^{\beta/\sigma^2}. \quad (8.13)$$

The hyperparameters  $\alpha$  and  $\beta$  can be treated as design parameters. The resulting posterior density representation is

$$\pi(\sigma^2 | q, v) \propto (\sigma^2)^{-(\alpha+1+n/2)} e^{-(\beta+SS_q/2)/\sigma^2}$$

so that

$$\sigma^2 | (v, q) \sim \text{Inv-gamma} \left( \alpha + \frac{n}{2}, \beta + \frac{SS_q}{2} \right). \quad (8.14)$$

An equivalent representation is

$$\sigma^2 | (v, q) \sim \text{Inv-gamma} \left( \frac{n_s + n}{2}, \frac{n_s \sigma_s^2 + SS_q}{2} \right), \quad (8.15)$$

where  $n_s = 2\alpha$  and  $\sigma_s^2 = \frac{\beta}{\alpha}$ . As noted in [92],  $n_s$  can be interpreted as representing the number of observations that provided the information encoded in the prior, whereas  $\sigma_s^2$  represents the mean squared error of the observations. In practice, one often takes  $n_s$  to be small (e.g.,  $n_s = 0.01$  to 1), which is consistent with a noninformative prior.

As noted in Example 4.69 and Definitions 4.13 and 4.14, random numbers from an inverse gamma distribution can be generated using the MATLAB Statistics Toolbox command `gamrnd.m` and then exploiting the equivalence between the gamma and inverse gamma distributions. If `gamrnd.m` is not available, one can use the inverse transform techniques of Section 4.1.1 to generate samples from the inverse gamma distribution.

We summarize in Algorithm 8.5 the random walk Metropolis algorithm with sampling-based error variance and noninformative prior. Some implementations employ the OLS estimate  $\sigma_{OLS}^2$  from (8.12) or the previous estimate  $s_{k-1}^2$  for  $\sigma_s^2$ . Whereas this is in the spirit of “empirical Bayes” inference as discussed in Section 4.8.2, it is noted in [35] that use of the present data to inform the prior can be problematic with small sample sizes and is at odds with the tenets of Bayesian analysis.

The issues associated with subjectively specifying  $n_s$  and  $\sigma_s^2$  based on prior knowledge can be avoided by using the Jeffreys prior

$$\pi_0(q, \sigma^2) = \frac{1}{\sigma^2}.$$

It is illustrated in Section 3.3.3 of [34] that this relation results from specification of a prior based on the Fisher information matrix for problems with mutually independent location and scale parameters  $q$  and  $\sigma^2$ . Alternatively, it can be obtained from (8.13) in the limit  $\alpha \rightarrow 0$ ,  $\beta \rightarrow 0$  for the hyperparameters.

**Algorithm 8.5 (Random Walk Metropolis with Noninformative Prior).**

1. Set number of chain elements  $M$  and design parameters  $n_s, \sigma_s$
2. Determine  $q^0 = \arg \min_q \sum_{i=1}^n [v_i - f_i(q)]^2$
3. Set  $SS_{q^0} = \sum_{i=1}^n [v_i - f_i(q^0)]^2$
4. Compute initial variance estimate:  $s_0^2 = \frac{SS_{q^0}}{n-p}$
5. Construct covariance estimate  $V = s_0^2 [\mathcal{X}^T(q^0) \mathcal{X}(q^0)]^{-1}$  and  $R = \text{chol}(V)$
6. For  $k = 1, \dots, M$ 
  - (a) Sample  $z_k \sim N(0, I_p)$
  - (b) Construct candidate  $q^* = q^{k-1} + Rz_k$
  - (c) Sample  $u_\alpha \sim \mathcal{U}(0, 1)$
  - (d) Compute  $SS_{q^*} = \sum_{i=1}^n [v_i - f_i(q^*)]^2$
  - (e) Compute
$$\alpha(q^* | q^{k-1}) = \min \left( 1, e^{-[SS_{q^*} - SS_{q^{k-1}}]/2s_{k-1}^2} \right)$$
  - (f) If  $u_\alpha < \alpha$ ,
    - Set  $q^k = q^*$  ,  $SS_{q^k} = SS_{q^*}$
    - else
    - Set  $q^k = q^{k-1}$  ,  $SS_{q^k} = SS_{q^{k-1}}$
    - endif
  - (g) Update  $s_k^2 \sim \text{Inv-gamma}(a_{val}, b_{val})$ , where
$$a_{val} = 0.5(n_s + n) , b_{val} = 0.5(n_s \sigma_s^2 + SS_{q^k})$$

**Remark 8.6.** We noted in (7.32) that for models in which the parameter scales vary by several orders of magnitude, one typically employs the scaled parameter  $q_s = q./s$  in optimization routines. Here  $s$  is a vector whose elements are the magnitude of each parameter and  $./$  denotes componentwise division. The same

scaling can improve the efficiency of optimization routines used to determine  $q^0$ , the conditioning of  $V$ , and the efficiency of Algorithm 8.5. Specifically, one would employ the alternative steps:

2. Determine  $q_s^0 = \arg \min_{q_s} \sum_{i=1}^n [v_i - f_i(q_s \cdot \times s)]^2$  and  $q^0 = q_s^0 \cdot \times s$ .
5. Construct covariance estimate  $V = s_0^2 [\mathcal{X}^T(q_s^0 \cdot \times s) \mathcal{X}^T(q_s^0 \cdot \times s)]^{-1}$  and  $R = \text{chol}(V)$ .
6. (b) Construct candidate  $q_s^* = q_s^{k-1} + Rz_k$ , and set  $q^* = q_s^* \cdot \times s$ .
6. (f) Additionally set  $q_s^k = q_s^*$  or  $q_s^k = q_s^{k-1}$ .

Note that the unscaled parameters  $q$  are employed in all model evaluations  $f_i(q)$ .

### 8.3.3 Metropolis–Hastings Algorithm

The Metropolis–Hastings algorithm generalizes the Metropolis algorithm to include nonsymmetric jumping or proposal functions  $J(q^*|q^{k-1})$ . For example, this includes Cauchy distributions

$$J(q^*|q^{k-1}) = \frac{1}{\pi[1 + (q^*)^2]}$$

and  $\chi^2(k)$  distributions

$$J(q^*|q^{k-1}) = \kappa(q^*)^{k/2-1} e^{q^*/2}.$$

In this case, candidates are accepted with probability  $\alpha = \min(1, r)$ , where the acceptance ratio is

$$\begin{aligned} r(q^*|q^{k-1}) &= \frac{\pi(q^*|v)/J(q^*|q^{k-1})}{\pi(q^{k-1}|v)/J(q^{k-1}|q^*)} \\ &= \frac{\pi(v|q^*)\pi_0(q^*)J(q^{k-1}|q^*)}{\pi(v|q^{k-1})\pi_0(q^{k-1})J(q^*|q^{k-1})}. \end{aligned} \tag{8.16}$$

For symmetric proposal functions  $J(q^*|q^{k-1}) = J(q^{k-1}|q^*)$ , (8.16) reduces to (8.9). We focus primarily on the Metropolis algorithm with symmetric proposal functions and refer the reader to [92] for details regarding the Metropolis–Hastings algorithm.

**Example 8.7.** In Examples 7.15 and 8.2, we considered the estimation of the parameter  $C$  for the spring model

$$\begin{aligned} \ddot{z} + C\dot{z} + Kz &= 0, \\ z(0) = 2, \quad \dot{z}(0) &= -C \end{aligned}$$

from a frequentist perspective and direct implementation of Bayes' relation (8.2). Here we illustrate the random walk Metropolis Algorithm 8.5. Recall that for  $C^2 - 4K < 0$ , the solution is

$$z(t) = 2e^{-Ct/2} \cos(\sqrt{K - C^2/4} \cdot t).$$

We consider displacement measurements at  $n = 501$  points in the time interval  $[0, 5]$  so that  $y_i(Q) = z(t_i, Q)$ , where  $t_i = 0.01i$ . Synthetic data  $v_i$  is simulated with errors  $\varepsilon_i \sim N(0, \sigma_0^2)$ , where  $\sigma_0 = 0.1$  is considered unknown when implementing the MCMC algorithm.

**Case i.** To compare with Example 8.2, we first take  $K = 20.5$  to be known and generate the displacement response with  $C_0 = 1.5$ . Hence the random parameters are  $Q = [C, \sigma^2]$ . We consider chains of length  $M = 10,000$ .

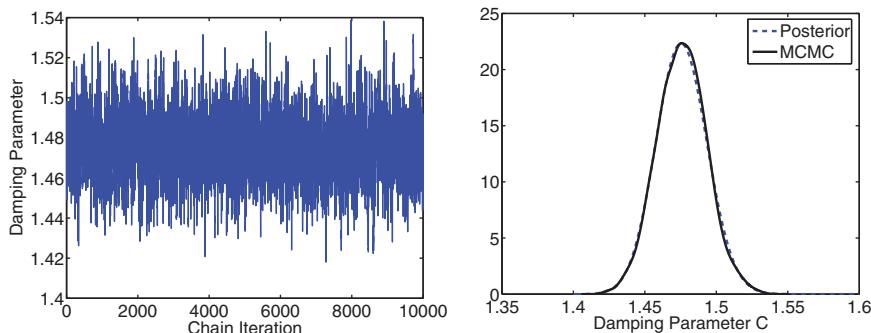
The chain, or marginal path, and kernel density estimate for  $C$ , computed using `kde.m`, are plotted in Figure 8.5. The comparison between the density computed using the random walk Metropolis algorithm and the direct posterior evaluation detailed in Example 8.2 shows that the two are nearly identical. The MCMC kernel will converge in the sense of distributions as  $M$  is increased and quadrature errors when computing the normalization constant are decreased. We note that the marginal path for  $C$  provides a baseline for comparison when applying the algorithm for multiple model parameters.

**Case ii.** Second, we consider the estimation of densities for  $Q = [C, K, \sigma^2]$  using synthetic data generated with  $K_0 = 20.5$ ,  $C_0 = 1.5$ , and  $\sigma_0 = 0.1$ . We consider first the choice  $J(q^*|q^{k-1}) = N(q^{k-1}, V)$  for the proposal distribution. Here

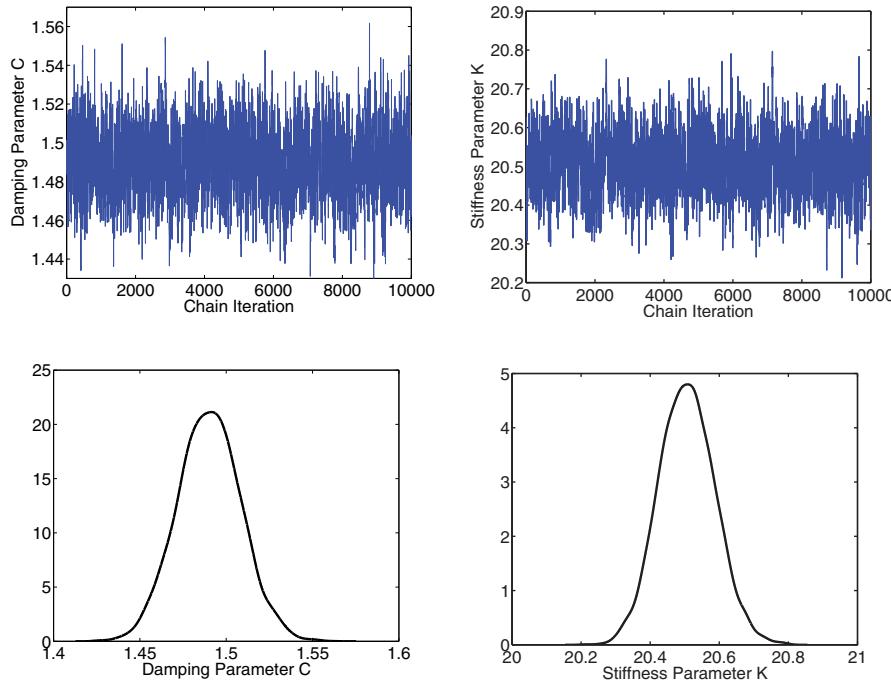
$$V = \begin{bmatrix} 0.000345 & 0.000268 \\ 0.000268 & 0.007071 \end{bmatrix}$$

is the covariance matrix given by (8.12) which is constructed using the analytic sensitivity relations

$$\begin{aligned} \frac{\partial y}{\partial C} &= e^{-Ct/2} \left[ \frac{Ct}{\sqrt{4K - C^2}} \sin(\sqrt{K - C^2/4} \cdot t) - t \cos(\sqrt{K - C^2/4} \cdot t) \right], \\ \frac{\partial y}{\partial K} &= \frac{-2t}{\sqrt{4K - C^2}} e^{-Ct/2} \sin(\sqrt{K - C^2/4} \cdot t). \end{aligned}$$



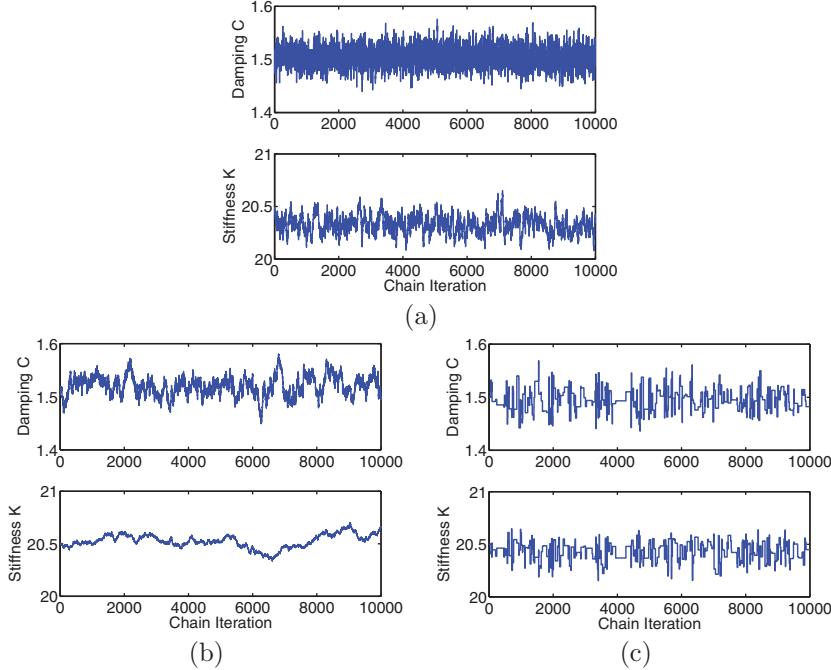
**Figure 8.5.** Marginal path and density for the damping parameter  $C$ .



**Figure 8.6.** Marginal paths and densities for the damping parameter  $C$  and stiffness parameter  $K$  obtained with  $J(q^*|q^{k-1}) = N(q^{k-1}, V)$ .

The marginal paths and densities are plotted in Figure 8.6. It is observed that  $2\sigma_C \approx 0.04$  so that  $\sigma_C^2 \approx 0.4 \times 10^{-3}$ , whereas  $2\sigma_K \approx 0.18$  so that  $\sigma_K^2 \approx 0.0081$ . These are close to the variance values in the covariance matrix  $V$ , which illustrates how it incorporates the general anisotropy exhibited by the posterior density. As a result, the marginal paths in Figure 8.6 exhibit essentially the same degree of mixing as the previous case  $Q = [C, \sigma^2]$  plotted in Figure 8.5(a).

To contrast, we illustrate the results obtained with the isotropic proposal function  $J(q^*|q^{k-1}) = N(q^{k-1}, sI)$  in Figure 8.7 for three choices of  $s$ . The mixing in Figure 8.7(a), which was obtained with  $s = 9 \times 10^{-4}$ , is reasonable but is not as rich as that obtained with the anisotropic proposal function  $J(q^*|q^{k-1}) = N(q^{k-1}, V)$ . Figure 8.7(b) illustrates the results obtained using a narrower proposal function constructed with  $s = 9 \times 10^{-6}$ . This yields substantial mixing but poor exploration of the parameter space for  $K$ . Conversely, the choice  $9 \times 10^{-2}$  yields poor mixing and chain stagnation since a large number of candidates are rejected. This illustrates the advantage of using the covariance matrix when it can be accurately constructed. Alternatively, in Section 8.6, we will discuss delayed rejection adaptive Metropolis methods that can be used to update the proposal distribution as candidates are accepted and the geometry of the posterior is determined.



**Figure 8.7.** Sample paths obtained with the proposal functions  $J(q^*|q^{k-1}) = N(q^{k-1}, sI)$ : (a)  $s = 9 \times 10^{-4}$ , (b)  $s = 9 \times 10^{-6}$ , and (c)  $s = 9 \times 10^{-2}$ .

## 8.4 Stationary Distribution and Convergence Criteria

The random walk Metropolis algorithm provides a Markov chain whose state space is the set of admissible parameter values. The initial distribution is provided by an OLS fit to calibration data. However, it is important to note that the Markov chain is based on samples from the “wrong” distribution in the sense that it is constructed using the proposal density rather than the sought after posterior density. In this section, we address two questions: (i) why should we expect the chain to have a stationary distribution that coincides with the posterior density, and (ii) what criteria indicate that the chain has converged to this distribution?

In Definition 4.62, we showed that the detailed balance condition  $\pi_{k-1} p_{k-1,k} = \pi_k p_{k,k-1}$  was a sufficient (but not necessary) requirement for stationarity. Since we want to show that the posterior density is the stationary distribution, we take  $\pi_k = \pi(q^k|v)$ . Similarly, we consider

$$p_{k-1,k} = P(X_k = q^k | X_{k-1} = q^{k-1}),$$

which is the probability of transitioning from parameter  $q^{k-1}$  to  $q^k$ . The detailed balance condition in this context can thus be expressed as

$$\begin{aligned} \pi_{k-1} p_{k-1,k} &= \pi_k p_{k,k-1} \\ \Rightarrow \pi(q^{k-1}|v) p_{k-1,k} &= \pi(q^k|v) p_{k,k-1}. \end{aligned}$$

Since  $p_{k-1,k} = P(\text{proposing } q^k)P(\text{accepting } q^k)$ , it follows from the definition of the proposal distribution  $J(q^k|q^{k-1})$  and acceptance probability  $\alpha$  that

$$\begin{aligned} p_{k-1,k} &= J(q^k|q^{k-1})\alpha(q^k|q^{k-1}) \\ &= J(q^k|q^{k-1}) \min \left( 1, \frac{\pi(q^k|v)J(q^{k-1}|q^k)}{\pi(q^{k-1}|v)J(q^k|q^{k-1})} \right). \end{aligned}$$

From the relation

$$v \min(1, x/v) = \min(x, v) = x \min(1, v/x),$$

which is established in Exercise 8.1, it follows that

$$\begin{aligned} \pi(q^{k-1}|v)p_{k-1,k} &= \pi(q^{k-1}|v)J(q^k|q^{k-1}) \min \left( 1, \frac{\pi(q^k|v)J(q^{k-1}|q^k)}{\pi(q^{k-1}|v)J(q^k|q^{k-1})} \right) \\ &= \pi(q^k|v)J(q^{k-1}|q^k) \min \left( 1, \frac{\pi(q^{k-1}|v)J(q^k|q^{k-1})}{\pi(q^k|v)J(q^{k-1}|q^k)} \right) \\ &= \pi(q^k|v)p_{k,k-1}. \end{aligned} \quad (8.17)$$

Hence the detailed balance condition is satisfied for the Metropolis–Hastings acceptance relation and the posterior density is the stationary distribution. We note that the transition kernel for the Markov chain can be defined as

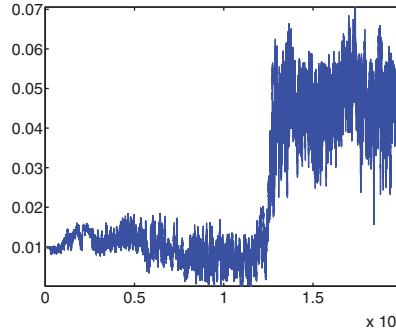
$$\begin{aligned} p_{ij} &= J(q^j|q^i) \min \left( 1, \frac{\pi(q^j|v)J(q^i|q^j)}{\pi(q^i|v)J(q^j|q^i)} \right), \quad i \neq j, \\ p_{ii} &= 1 - \sum_{j \neq i} p_{ij}. \end{aligned}$$

The detailed balance result (8.17) establishes that if chains are run sufficiently long, they will produce samples from the posterior density. However, the question of how long chains must be run to converge to and adequately sample from the posterior is difficult and analytic convergence and stopping criteria are lacking. It is noted in [42] that the convergence, or burn-in, of MCMC algorithms can be falsified but, in general, not completely verified.

Despite the lack of analytic convergence theory, there are various tests that can be used to establish confidence in simulations. We summarize only aspects of these tests and refer readers to [42, 92] for details regarding convergence or burn-in of MCMC simulations.

The most direct method for assessing burn-in or convergence is to visually or statistically monitor the marginal paths associated with each parameter, as illustrated in Figures 8.5, 8.6, and 8.7. The initial period during which means appear to transition is often termed the *burn-in period*, and these values are excluded when computing parameter or response densities since they are not sampled from the stationary or posterior distribution.

The difficulty is that chains can appear stationary for a very large number of simulations and then change in the manner shown in Figure 8.8 for a parameter from a transductive material model [116]. Because MCMC algorithms will determine



**Figure 8.8.** Shift in the marginal path after 130,000 iterations due to a local minimum in the sum of squares; see [116].

global minima, if run sufficiently long, this can be due to initial sampling in a local minimum before determining another with a lower residual. However, there is no guarantee that this is a global minimum and the chain could transition again if another, lower, minimum is found.

In some cases, the parameter density constructed using the burned-in MCMC chain can be compared with that directly computed using Bayes' relation. Whereas this is feasible only for a moderate number of parameters, which may require adaptive sparse grid quadrature, it can be used to verify (or falsify) the MCMC results.

From a statistical perspective, the percentage of accepted points, termed the *acceptance ratio*, is often used to quantify whether or not the chain is adequately sampling from the posterior. Because the optimal acceptance ratio depends on the geometry of the posterior, the range of reasonable acceptance ratios is quite large; e.g., values between 0.1 and 0.5 are often considered acceptable. The acceptance ratio is often used to tune the proposal density  $J(q^*|q^{k-1})$  to improve mixing. For example, a small acceptance ratio can produce stagnation, as shown in Figures 8.3(d) and 8.7(c). This can be addressed by decreasing the variance of affected parameters to narrow the proposal function.

A second commonly employed statistical test is to check the autocorrelation

$$R(k) = \frac{\sum_{i=1}^{M-k} (q_i - \bar{q})(q_{i+k} - \bar{q})}{\sum_{i=1}^M (q_i - \bar{q})^2} = \frac{\text{cov}(q_i, q_{i+k})}{\text{var}(q_i)} \quad (8.18)$$

between components in the chain that are  $k$  iterations apart. Because adjacent components are likely correlated due to the Markov property, this test can be used to establish that the chain is producing iid samples from the posterior. As detailed in [56], low autocorrelation is often indicative of fast convergence.

The reader is warned that care must be exercised when interpreting MCMC results reported in the literature. One commonly encounters parameter densities with no mention of the burn-in period or illustration of marginal sample paths. In such cases, it is difficult to verify that the reported density is truly indicative of the posterior. Second, it is not uncommon for authors employing computationally

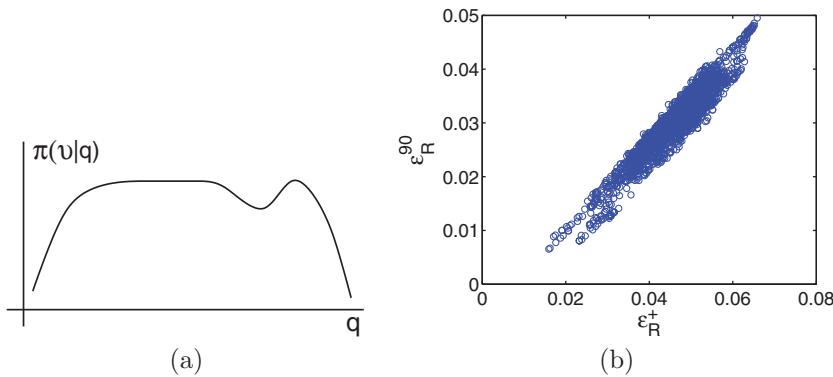
intensive codes to discuss very short burn-in periods. While this may be the case, it can also reflect the fact that the reported length reflects the largest number of simulations that could be run with the code.

## 8.5 Parameter Identifiability

We noted in Definition 6.1 that the concept of parameter identifiability quantifies the uniqueness of the input-output map between parameters and responses. *Hence parameter identifiability is a property of the model and observations rather than of the inference or estimation procedure.* For example, we illustrated in Example 3.2 that we could not uniquely determine  $q = [m, c, k]$  in the spring model (3.5) given displacement measurements  $z(t)$ . Instead we had to reformulate the model in terms of the parameters  $K = \frac{k}{m}$  and  $C = \frac{c}{m}$ . As detailed in Chapter 6, one must reformulate the model or fix certain parameter values to address lack of parameter identifiability.

From the perspective of the likelihood, unidentifiability produces flat regions in the likelihood function or multiple maxima having the same value, as illustrated in Figure 8.9(a). From a Bayesian perspective, we noted in Section 6.3 that unidentifiability can be manifested as posterior joint densities that are nearly single-valued for parameters having independent priors. Figure 8.9(b) illustrates the correlation exhibited by unidentifiable material parameters  $\varepsilon_R^{90}$  and  $\varepsilon_R^+$  in the model of [116]. The fact that multiple parameter values yield the same maximum likelihood value can also cause chains to jump in the manner shown in Figure 8.8. For noninformative priors, this can slow or stop the convergence of the chains to the posterior density. As detailed in Section 6.3, however, it is often difficult to differentiate between identifiable and unidentifiable parameters based solely on the width of joint densities, so this criterion should be interpreted as merely an indicator that parameter values may not be uniquely determined by the data.

In the random walk Metropolis algorithm, it follows from Property 6.7 that lack of identifiability is manifested by a singular covariance matrix  $V$  constructed



**Figure 8.9.** (a) Likelihood for an unidentifiable parameter set and (b) correlation of unidentifiable parameters; see [116].

from the sensitivity relations  $\mathcal{X}(q)$ . This is illustrated for the simple harmonic oscillator model in Exercise 8.5. This reinforces the relation between  $V$  and the Fisher information matrix  $\mathcal{F}$  which quantifies the information content of an experiment.

Whereas unidentifiable parameters cannot be uniquely determined using OLS estimators or maximum likelihood estimators, they can in some cases be determined using Bayesian estimators with informative priors for the unidentifiable parameters. Hence Bayesian inference can sometimes be successful for overparameterized or unidentifiable models if informative priors are available.

## 8.6 Delayed Rejection Adaptive Metropolis (DRAM)

Whereas the choices (8.8) for the proposal distribution incorporate aspects of parameter scaling and variability, they do not provide mechanisms to incorporate information learned about the posterior distribution as candidate parameters are accepted and the chain progresses. Such mechanisms are provided by various adaptive Metropolis algorithms [11, 103, 208, 255], including the DRAM algorithm [102], which we summarize here.

We note that because adaptive algorithms employ part of the chain history to update the proposal function, they are no longer Markovian processes, which requires that states depend only on the previous state. Hence the convergence criteria discussed in Section 8.4 do not apply and alternative ergodicity properties must be established to guarantee convergence to the posterior density. General criteria, such as the *diminishing adaptation* and *bounded convergence conditions*, that adaptive methods must satisfy to establish convergence to a stationary distribution are provided in [11, 103, 208].

### 8.6.1 Adaptive Metropolis

In principle, the adaptive Metropolis (AM) step employed in the DRAM algorithms is quite straightforward. During a nonadaptive period of length  $k_0$ , chain values  $q^0, q^1, \dots, q^{k-1}$  are computed using the initial covariance matrix  $V_0 = V$  or  $V_0 = D$  employed in the random walk Metropolis Algorithm 8.5. Once adaptation commences, the updated chain covariance matrix at the  $k^{th}$  step is taken to be

$$V_k = s_p \text{cov}(q^0, q^1, \dots, q^{k-1}) + \varepsilon I_p. \quad (8.19)$$

Here  $s_p$  is a design parameter that depends on the dimension  $p$  of the parameter space. As detailed in [102], a common choice is  $s_p = 2.38^2/p$ . The length  $k_0$  of the adaptation interval is chosen to balance mixing with providing sufficient diversity in points to ensure a nonsingular covariance matrix in the initial stages of the chain progression. Shorter adaptation intervals typically produce better mixing and higher acceptance ratios since they accelerate the rate at which information regarding the posterior is incorporated. In practice,  $k_0$  is often specified to be approximately 100. The term  $\varepsilon I_p$ , where  $\varepsilon \geq 0$  and  $I_p$  is the  $p$ -dimensional identity matrix, ensures that  $V_k$  is positive definite. One can often take  $\varepsilon = 0$ .

In theory,  $\text{cov}(q^0, \dots, q^{k-1})$  can be computed using the empirical covariance formula

$$\text{cov}(q^0, \dots, q^{k-1}) = \frac{1}{k-1} \left( \sum_{i=0}^{k-1} q^i (q^i)^T - k \bar{q}^k (\bar{q}^k)^T \right),$$

where  $\bar{q}^k = \frac{1}{k} \sum_{i=0}^{k-1} q^i$  and  $q^i$  are column vectors. However, this becomes increasingly inefficient as  $k$  becomes large. Instead, one employs the recursive relation

$$V_{k+1} = \frac{k-1}{k} V_k + \frac{s_p}{k} [k \bar{q}^{k-1} (\bar{q}^{k-1})^T - (k+1) \bar{q}^k (\bar{q}^k)^T + q^k (q^k)^T + \varepsilon I_p]. \quad (8.20)$$

In a similar manner, the sample mean can be computed recursively as

$$\begin{aligned} \bar{q}^{k+1} &= \frac{1}{k+1} \sum_{i=0}^k q^i \\ &= \frac{k}{k+1} \cdot \frac{1}{k} \sum_{i=0}^{k-1} q^i + \frac{1}{k+1} q^k \\ &= \frac{k}{k+1} \bar{q}^k + \frac{1}{k+1} q^k. \end{aligned}$$

It is noted in [102] that the efficiency of the algorithm is improved if adaptation occurs at prescribed intervals and, in Algorithm 8.8, we employ intervals of length  $k_0$ . The ergodicity of this adaptive algorithm is established in [103].

### 8.6.2 Delayed Rejection

In the standard Metropolis algorithm, chain candidates  $q^*$  are accepted with probability

$$\alpha(q^* | q^{k-1}) = \min \left( 1, \frac{\pi(q^* | v) J(q^{k-1} | q^*)}{\pi(q^{k-1} | v) J(q^* | q^{k-1})} \right) = \min \left( 1, \frac{\pi(q^* | v)}{\pi(q^{k-1} | v)} \right)$$

and, if rejected, the prior chain value  $q^{k-1}$  is retained. The delayed rejection (DR) algorithm provides a mechanism for constructing alternative candidates  $q^{*j}$  if  $q^*$  is rejected rather than initially retaining the previous value.

As detailed in [102], a second-stage candidate  $q^{*2}$  is chosen using the proposal function

$$J_2(q^{*2} | q^{k-1}, q^*) = N(q^{k-1}, \gamma_2^2 V_k),$$

where  $V_k = R_k R_k^T$  is the covariance matrix produced by the adaptive algorithm. The notation  $J_2(q^{*2} | q^{k-1}, q^*)$  indicates that we are proposing  $q^{*2}$  having started at  $q^{k-1}$  and rejected  $q^*$ . The software discussed in Remark 8.9 employs  $\gamma_2 = \frac{1}{5}$ , but other values are reasonable. Because  $\gamma_2 < 1$ , the second-stage proposal function is narrower than the original, which increases mixing. The probability of accepting

the second-stage candidate, having started at  $q^{k-1}$  and rejected  $q^*$ , is

$$\begin{aligned}\alpha_2(q^{*2}|q^{k-1}, q^*) &= \min\left(1, \frac{\pi(q^{*2}|v)J(q^*|q^{*2})J_2(q^{k-1}|q^{*2}, q^*)[1 - \alpha(q^*|q^{*2})]}{\pi(q^{k-1}|v)J(q^*|q^{k-1})J_2(q^{*2}|q^{k-1}, q^*)[1 - \alpha(q^*|q^{k-1})]}\right) \\ &= \min\left(1, \frac{\pi(q^{*2}|v)J(q^*|q^{*2})[1 - \alpha(q^*|q^{*2})]}{\pi(q^{k-1}|v)J(q^*|q^{k-1})[1 - \alpha(q^*|q^{k-1})]}\right)\end{aligned}\quad (8.21)$$

due to the symmetry of  $J_2$ .

The form of  $\alpha_2$  can be motivated as follows. It was noted in Section 8.4 that for the Metropolis–Hastings algorithm,

$$\begin{aligned}p_{k-1,k} &= P(X_k = q^k | X_{k-1} = q^{k-1}) \\ &= P(\text{proposing } q^k)P(\text{accepting } q^k) \\ &= J(q^k|q^{k-1})\alpha(q^k|q^{k-1}).\end{aligned}$$

We now consider the case when we accept  $q^k = q^{*2}$  having rejected  $q^*$  so that

$$\begin{aligned}p_{k-1,k} &= P(\text{proposing } q^*)P(\text{rejecting } q^*)P(\text{proposing } q^k)P(\text{accepting } q^k) \\ &= J(q^*|q^{k-1})[1 - \alpha(q^*|q^{k-1})]J_2(q^k|q^{k-1}, q^*)\alpha_2(q^k|q^{k-1}, q^*).\end{aligned}$$

To satisfy the detailed balance condition  $\pi(q^{k-1}|v)p_{k-1,k} = \pi(q^k|v)p_{k,k-1}$ , we thus require

$$\begin{aligned}\pi(q^{k-1}|v)J(q^*|q^{k-1})[1 - \alpha(q^*|q^{k-1})]J_2(q^k|q^{k-1}, q^*)\alpha_2(q^k|q^{k-1}, q^*) \\ = \pi(q^k|v)J(q^*|q^k)[1 - \alpha(q^*|q^k)]J_2(q^{k-1}|q^k, q^*)\alpha_2(q^{k-1}|q^k, q^*).\end{aligned}$$

The condition (8.21) guarantees that the detailed balance condition is satisfied and that  $\alpha \leq 1$ .

If  $q^{*2}$  is rejected, a third-stage candidate and acceptance condition can be constructed, and recursive relations to construct  $j^{th}$ -stage candidates  $q^{*j}$  and probabilities  $\alpha_i(q^{*j}, \dots, q^{*2}, q^*, q^{k-1})$  are provided in [102]. We employ only a second-stage candidate in Algorithm 8.8 since this is the default in the referenced software.

In combination, DR and AM provide two different but complementary mechanisms to modify the proposal function. The AM provides feedback in the sense that information learned about the posterior through accepted chain candidates is used to update the proposal via the chain covariance matrix. The DR is an open loop mechanism that alters the proposal function in a predetermined manner to improve mixing. The modifications from DR are temporary and have the goal of stimulating mixing, whereas the AM mechanism enacts permanent changes that reflect information learned about the posterior. In Algorithm 8.8, we summarize the DRAM algorithm implemented in the referenced software with a second-stage rejection mechanism. The reader is referred to [102] for details and other combinations of the DR and AM components.

**Algorithm 8.8 (Delayed Rejection Adaptive Metropolis Algorithm with Noninformative Prior [102]).**

1. Set design parameters  $n_s, \sigma_s^2, k_0$  and number of chain iterates  $M$
2. Determine  $q^0 = \arg \min_q \sum_{i=1}^n [v_i - f_i(q)]^2$
3. Set  $SS_{q^0} = \sum_{i=1}^n [v_i - f_i(q^0)]^2$
4. Compute initial variance estimate:  $s_0^2 = \frac{SS_{q^0}}{n-p}$
5. Construct covariance estimate  $V = s_0^2 [\mathcal{X}^T(q^0) \mathcal{X}(q^0)]^{-1}$  and  $R = \text{chol}(V)$
6. For  $k = 1, \dots, M$ 
  - (a) Sample  $z_k \sim N(0, I_p)$
  - (b) Construct candidate  $q^* = q^{k-1} + Rz_k$
  - (c) Sample  $u_\alpha \sim \mathcal{U}(0, 1)$
  - (d) Compute  $SS_{q^*} = \sum_{i=1}^n [v_i - f_i(q^*)]^2$
  - (e) Compute
$$\alpha(q^* | q^{k-1}) = \min \left( 1, e^{-[SS_{q^*} - SS_{q^{k-1}}]/2s_{k-1}^2} \right)$$
  - (f) If  $u_\alpha < \alpha$ ,
    - Set  $q^k = q^*$ ,  $SS_{q^k} = SS_{q^*}$
    - else
      - Enter DR Algorithm 8.10
    - endif
  - (g) Update  $s_k^2 \sim \text{Inv-gamma}(a_{val}, b_{val})$ , where
$$a_{val} = 0.5(n_s + n), \quad b_{val} = 0.5(n_s \sigma_s^2 + SS_{q^k})$$
  - (h) if  $\text{mod}(k, k_0) = 1$ 
    - Update  $V_k = s_p \text{cov}(q^0, q^1, \dots, q^k)$
    - else
      - $V_k = V_{k-1}$
  - (i) Update  $R_k = \text{chol}(V_k)$

**Remark 8.9.** MATLAB software for Algorithm 8.8 of [102] is available at the websites <https://wiki.helsinki.fi/display/inverse/Adaptive+MCMC> and <http://helios.fmi.fi/~lainema/mcmc/>.

**Algorithm 8.10 (Delayed Rejection Component of DRAM with Noninformative Prior).**

1. Set the design parameter  $\gamma_2 = \frac{1}{5}$
2. Sample  $z_k \sim N(0, I_p)$
3. Construct second-stage candidate  $q^{*2} = q^{k-1} + \gamma_2 R_k z_k$
4. Sample  $u_\alpha \sim \mathcal{U}(0, 1)$
5. Compute  $SS_{q^{*2}} = \sum_{i=1}^n [v_i - f_i(q^{*2})]^2$
6. Compute  $\alpha_2(q^{*2} | q^{k-1}, q^*)$  using (8.21)
7. If  $u_\alpha < \alpha$ ,
  - Set  $q^k = q^{*2}$ ,  $SS_{q^k} = SS_{q^{*2}}$
  - else
  - Set  $q^k = q^{k-1}$ ,  $SS_{q^k} = SS_{q^{k-1}}$
  - endif

**Remark 8.11.** It was noted in Remark 8.6 that the performance of the algorithm can be significantly enhanced by using scaled parameters  $q_s = q./s$  if physical parameter values vary significantly. This can be implemented with the modified steps:

2. Determine  $q_s^0 = \arg \min_{q_s} \sum_{i=1}^n [v_i - f_i(q_s \cdot \times s)]^2$  and  $q^0 = q_s^0 \cdot \times s$ .
5. Construct covariance estimate  $V = s_0^2 [\mathcal{X}^T (q_s^0 \cdot \times s) \mathcal{X}^T (q_s^0 \cdot \times s)]$  and  $R = \text{chol}(V)$ .
6. (b) Construct candidate  $q_s^* = q_s^{k-1} + R z_k$ , and set  $q^* = q_s^* \cdot \times s$ .
6. (f) Additionally set  $q_s^k = q_s^*$ .
- DR 3. Construct second-stage candidate  $q_s^{*2} = q_s^{k-1} + \gamma_2 R_k$ , and set  $q^{*2} = q_s^{*2} \cdot \times s$ .
- DR 7. Additionally set  $q_s^k = q_s^{*2}$  or  $q_s^k = q_s^{k-1}$ .

**Example 8.12.** In Example 7.16, we used frequentist analysis to construct sampling distributions for the parameters  $q = [\Phi, h]$  in the model

$$\begin{aligned} \frac{d^2 T_s}{dx^2} &= \frac{2(a+b)}{ab} \frac{h}{k} [T_s(x) - T_{amb}], \\ \frac{dT_s}{dx}(0) &= \frac{\Phi}{k}, \quad \frac{dT_s}{dx}(L) = \frac{h}{k} [T_{amb} - T_s(L)] \end{aligned} \tag{8.22}$$

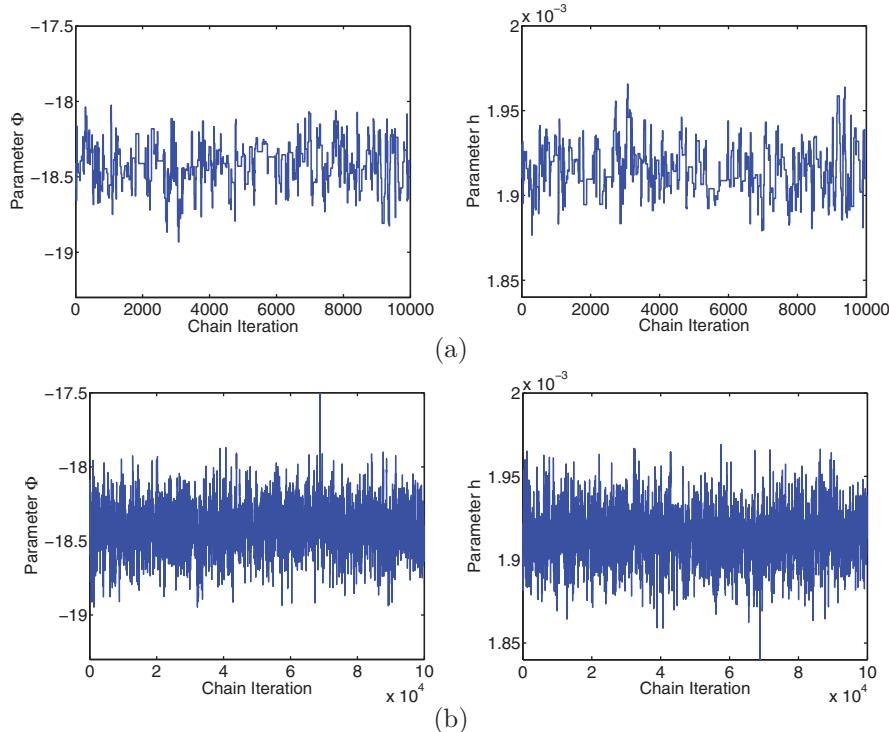
for steady state heat conduction in an uninsulated aluminum rod of length  $L$  with a source heat flux  $\Phi$  at  $x = 0$ . As detailed in Example 3.5,  $T_s$  is the steady state temperature,  $h$  is a convective heat transfer coefficient, and  $k = 2.37 \text{ W}\cdot\text{cm}^{-1}\cdot^\circ\text{C}^{-1}$  is the thermal conductivity coefficient for aluminum.

Here we construct densities for  $\Phi$  and  $h$  using both the random walk Metropolis Algorithm 8.5 and the delayed rejection adaptive Metropolis Algorithm 8.8. The residual plot in Figure 7.3(b) motivates the assumption that errors are iid and unbiased. We further assume that they are normally distributed with fixed but unknown variance  $\sigma_0^2$ . With these assumptions, we can employ the likelihood relation (8.10). We employ the covariance estimate

$$V_0 = \begin{bmatrix} 2.1034 \times 10^{-2} & -2.0286 \times 10^{-6} \\ -2.0286 \times 10^{-6} & 2.0972 \times 10^{-10} \end{bmatrix} \quad (8.23)$$

of (7.46) as the initial proposal function.

We first employ the nonadaptive algorithm to provide a baseline to illustrate advantages of the DRAM algorithm. The marginal paths, obtained using the random walk Metropolis algorithm with  $M = 10^4$  and  $M = 10^5$  Monte Carlo iterations, are plotted in Figure 8.10. Because  $V_0$  incorporates the anisotropy due to the dif-



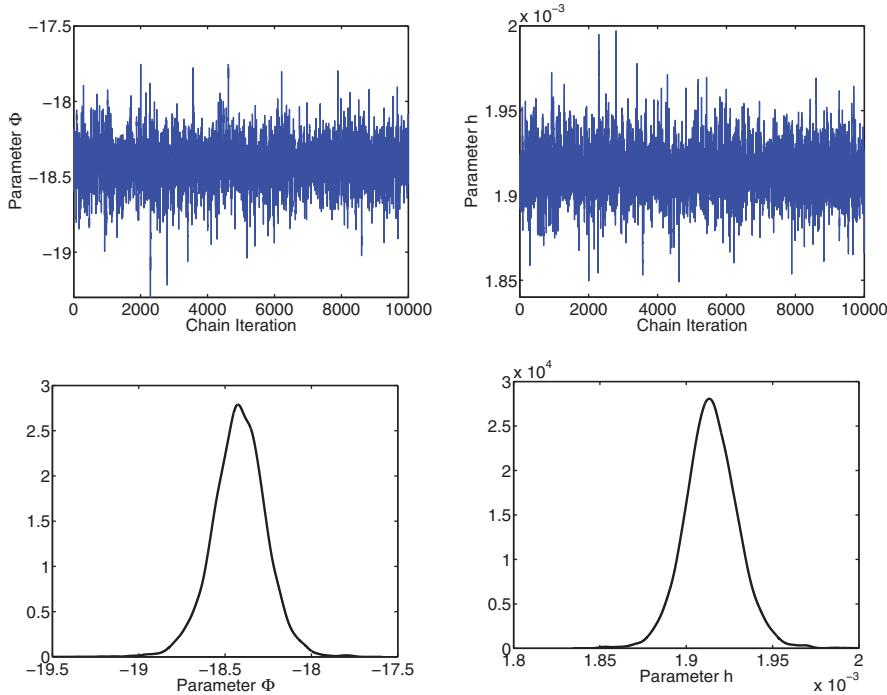
**Figure 8.10.** Sample paths obtained using the nonadaptive random walk Metropolis Algorithm 8.5 with (a)  $M = 10^4$  and (b) and  $M = 10^5$  iterations.

ferring variances of  $\Phi$  and  $h$ , the two chains exhibit similar mixing. However, the acceptance ratio is 0.056, which is smaller than the targeted range  $0.1 - 0.5$ , and the plots obtained with  $M = 10^4$  iterations exhibit regions where the chains briefly stagnate. As illustrated in Figure 8.3, this indicates that narrower proposal functions should improve mixing. We note that the stagnation regions are not visible in the plot of  $M = 10^5$  iterates, thus motivating the necessity of checking the acceptance ratio, which remains 0.056. The stationarity of the chains indicates that they have burned-in and are sampling from the posterior density.

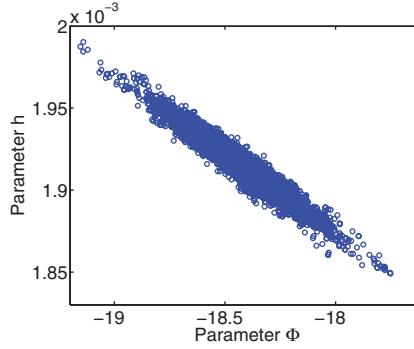
The marginal paths and kernel density estimates constructed using  $M = 10^4$  DRAM iterates are plotted in Figure 8.11. A comparison with the chains plotted in Figure 8.10(a) and (b) illustrates that the algorithm is achieving the goal of enhancing mixing and accelerating burn-in. The chain covariance matrix is

$$V = \begin{bmatrix} 2.4101 \times 10^{-2} & -2.3211 \times 10^{-6} \\ -2.3211 \times 10^{-6} & 2.3869 \times 10^{-10} \end{bmatrix},$$

which is very close to the original covariance matrix  $V_0$  in (8.23) which was provided by OLS theory. This demonstrates that for this problem, the narrowing of the proposal function in the DR step has a more substantial impact than the proposal modifications in the AM step. The algorithm provides the estimate  $\sigma^2 = 0.0678$  for



**Figure 8.11.** Marginal paths and densities for the parameters  $\Phi$  and  $h$  obtained with the DRAM algorithm.



**Figure 8.12.** Joint sample points for  $\Phi$  and  $h$ .

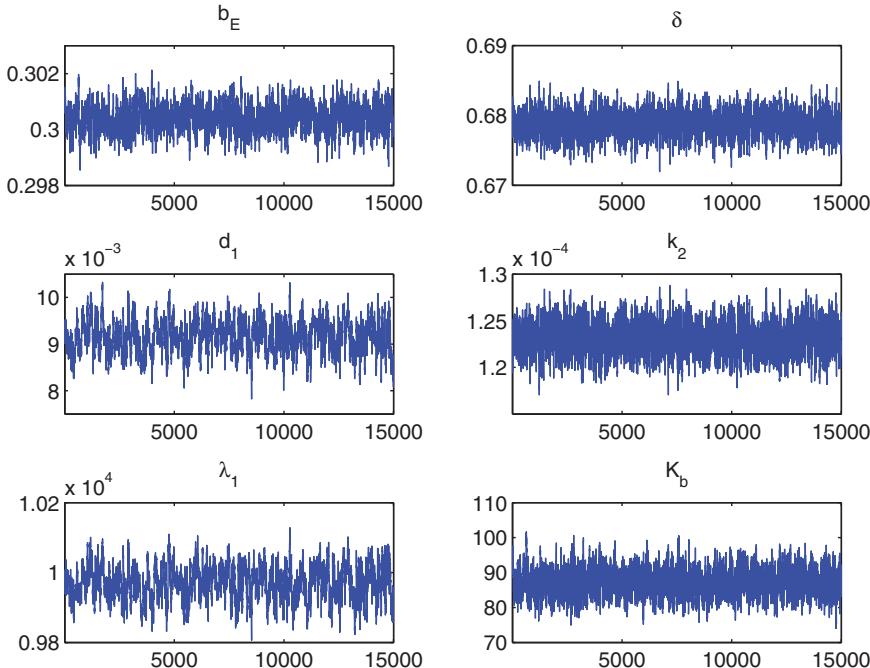
the error variance, so  $\sigma = 0.2604$ . The estimated standard deviations for  $\Phi$  and  $h$  are  $\sigma_\Phi = 0.1552$  and  $\sigma_h = 1.5450 \times 10^{-5}$ . It is observed in the marginal density plots in Figure 8.11 and joint density plotted in Figure 8.12 that two standard deviations represent approximately 95% of the density.

The frequentist analysis in Example 7.16 yielded the standard deviations  $\sigma = 0.2504$ ,  $\sigma_\Phi = 0.1450$ , and  $\sigma_h = 1.4482 \times 10^{-5}$ . We first note that the values of  $\sigma$  are within 4% despite the fact that  $\sigma = 0.2504$  is an estimate, whereas  $\sigma = 0.2604$  is the mean of a density sampled through realizations of the Markov chain. Furthermore, we observe that the values of  $\sigma_\Phi$  and  $\sigma_h$  obtained through Bayesian analysis are within 5% of the frequentist values for the sampling distribution.

From a mathematical perspective, the similarity of the sampling distribution and parameter distribution can be attributed to the normality of the estimated parameter densities. However, care must be exercised when interpreting this result since the sampling distribution is for the parameter estimator rather than the parameters. As detailed in Chapter 7, it thus quantifies uncertainty pertaining to the estimation procedure rather than uncertainty associated with the parameters.

**Example 8.13.** To illustrate the performance of the delayed rejection adaptive Metropolis algorithm for a system of coupled ODEs with multiple responses, we employ the model

$$\begin{aligned}
 \dot{T}_1 &= \lambda_1 - d_1 T_1 - (1 - \varepsilon) k_1 V T_1, \\
 \dot{T}_2 &= \lambda_2 - d_2 T_2 - (1 - f\varepsilon) k_2 V T_2, \\
 \dot{T}_1^* &= (1 - \varepsilon) k_1 V T_1 - \delta T_1^* - m_1 E T_1^*, \\
 \dot{T}_2^* &= (1 - f\varepsilon) k_2 V T_2 - \delta T_2^* - m_2 E T_2^*, \\
 \dot{V} &= N_T \delta(T_1^* + T_2^*) - c V - [(1 - \varepsilon) \rho_1 k_1 T_1 + (1 - f\varepsilon) \rho_2 k_2 T_2] V, \\
 \dot{E} &= \lambda_E + \frac{b_E(T_1^* + T_2^*)}{T_1^* + T_2^* + K_b} E - \frac{d_E(T_1^* + T_2^*)}{T_1^* + T_2^* + K_d} E - \delta_E E,
 \end{aligned} \tag{8.24}$$



**Figure 8.13.** Chains for  $Q = [b_E, \delta, d_1, k_2, \lambda_1, K_b]$ .

developed in [2, 3] to provide a framework to investigate control strategies for HIV. As detailed in Example 3.3,  $T_1$  and  $T_1^*$  represent the populations of uninfected and infected T-lymphocytes,  $T_2$  and  $T_2^*$  are corresponding macrophage populations, and  $V, E$  denote the populations of free virus and immune effector cells.

To construct synthetic data for all six states, we add noise to model solutions computed using the parameter values reported in [3]. We note that clinical data comprised of the total number  $T_1 + T_1^*$  of T-lymphocytes and viral load  $V$  can be found in [3].

For this example, we used the DRAM algorithm to construct chains and densities for the parameters  $Q = [b_E, \delta, d_1, k_2, \lambda_1, K_b]$  with the remaining parameters fixed at the values reported in [3]. After a burn-in period of 5000 iterates, the parameter chains, densities, and joint sample points obtained using 15,000 DRAM iterates are plotted in Figures 8.13–8.15. We note that this is a relatively short burn-in period despite the fact that parameter values vary over eight orders of magnitude. It is observed from the pairwise joint sample plots in Figure 8.15 that  $k_2$  and  $\delta$  are clearly correlated, as are  $\lambda_1$  and  $d_1$ . The fact that the parameters are not mutually independent proves important when we revisit this problem in Example 9.14, where we discuss propagation of uncertainty in models.

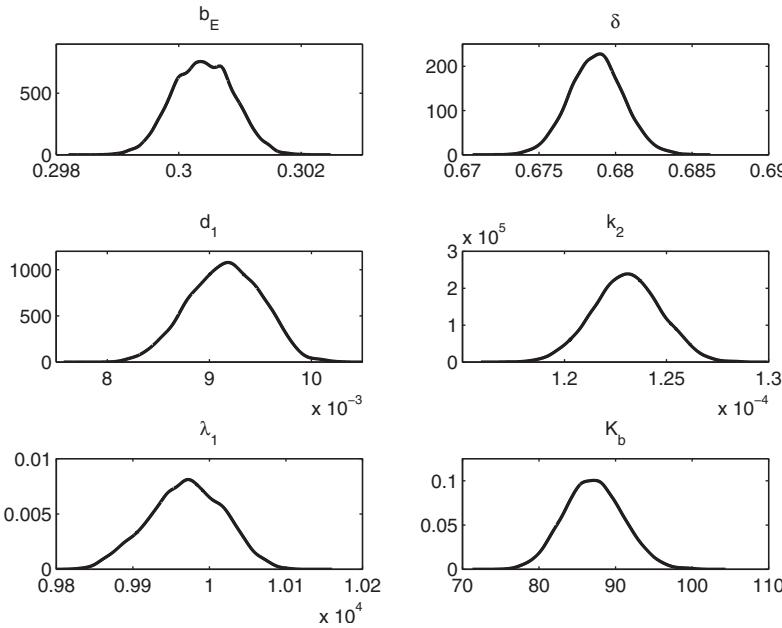


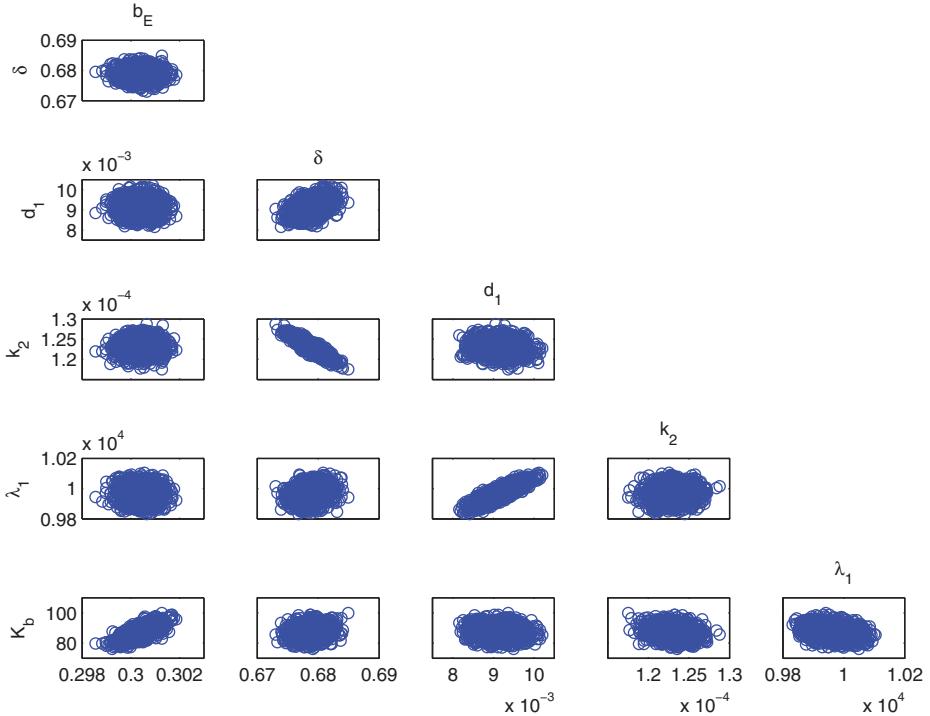
Figure 8.14. Marginal densities for  $Q = [b_E, \delta, d_1, k_2, \lambda_1, K_b]$ .

## 8.7 DiffeRential Evolution Adaptive Metropolis (DREAM)

It was illustrated in Section 8.3 that the scale and orientation of the proposal function critically affects the mixing and exploration of chains in standard random walk implementations of MCMC algorithms. The DRAM algorithm of Section 8.6 significantly improves performance through two mechanisms: adaptation updates the chain covariance matrix as information is obtained about the posterior density, and delayed rejection alters the proposal function in a predefined manner to improve mixing. For many models, that is sufficiently efficient for constructing parameter densities that can subsequently be employed for quantifying uncertainties in model responses or QoI.

However, there are various regimes for which DRAM algorithms are often not efficient. These include problems in which posterior densities are multimodal, are highly complex, or have heavy tails. For these cases, the single DRAM chain will be slow to traverse the posterior, which can significantly diminish its efficiency. Moreover, the computational overhead associated with complex models—such as the weather, climate, hydrology, and nuclear reactor models discussed in Chapter 1—often preclude the construction of burned-in single chains, whereas one can often compute shorter parallel chains using massively parallel architectures.

With the framework discussed in Section 8.6, these issues have motivated the development of parallel chain versions of the adaptive Metropolis algorithms. In the interchain adaptation approach detailed in [230], independent parallel chains,



**Figure 8.15.** Joint sample points for  $Q = [b_E, \delta, d_1, k_2, \lambda_1, K_b]$ .

with early rejection mechanisms, are used to adapt the proposal function which in turn influences the mixing and exploration of future chain elements. This approach is highly parallelizable, improves the convergence of individual chains, and has been applied to climate models.

Differential evolution Markov chain (DE-MC) methods provide an alternative that can be more efficient for problems with multimodal or heavy tailed densities [246]. This approach can be summarized as follows. For  $p$  parameters,  $N$  chains  $q_i^k$ ,  $i = 1, \dots, N$ , are simultaneously run in parallel and the present population is stored in an  $N \times p$  matrix  $X$ . Note that here the subscript designates the  $i^{th}$  chain rather than the  $i^{th}$  parameter component. In the original algorithm, candidates  $q_i^*$  were constructed by randomly choosing two chains from  $X$ , without replacement, and adding the weighted difference to  $q_i^k$ , that is,

$$q_i^* = q_i^k + \gamma (q_{i_1}^{k-1} - q_{i_2}^{k-1}) + e, \quad i_1 \neq i_2 \neq i,$$

for  $i = 1, \dots, N$ . Typically, one takes  $e$  as realizations of  $E \sim N(0, bI_p)$ , where  $b$  is chosen smaller than the variance of the posterior. An optimal choice for the weight is  $\gamma = 2.38/\sqrt{2p}$ . During implementation, one often specifies  $\gamma = 1$  at every 10th generation to permit direct jumping between modes. Candidates are then accepted

with probability  $\alpha = \min(1, r)$ , where  $r$  is the acceptance ratio given by (8.11) or (8.16).

The DE-MC algorithm differs from DRAM in the sense that it generates candidates based on current chain information stored in  $X$  rather than the covariance matrix  $V$  or chain covariance  $V_k$  defined in (8.19) which constitute the proposal function. The construction of  $q^*$  based on random members of the population facilitates the exploration of multimodal and heavy tailed posterior distributions and provides a mechanism for determining the appropriate scale, shape, and orientation of the proposal function. Further, the parallel chains in DE-MC algorithms learn from each other as compared with the DRAM implementation, where independent chains are used to adapt the proposal function. Theory establishing that Markov chains constructed in this manner have a unique stationary distribution and details regarding the implementation and performance of the DE-MC algorithm are provided in [246].

Further improvements in efficiency can be realized for many applications when similar evolution algorithms are combined with self-adaptive, randomized subspace sampling. This is the basis for the DiffeRential Evolution Adaptive Metropolis (DREAM) algorithm detailed in [261]. The candidates in this case are randomly generated using the algorithm

$$q_i^* = q_i^k + (I_p + f)\gamma(\delta, p') \left[ \sum_{j=1}^{\delta} q_{i_1(j)}^{k-1} - \sum_{n=1}^{\delta} q_{i_2(n)}^{k-1} \right] + e,$$

where  $i_1(j), i_2(n) \in \{1, \dots, N\}$  satisfy  $i_1(j) \neq i_2(n) \neq i$  for  $j, n = 1, \dots, N$ . Here  $\delta$  denotes the number of randomly sampled pairs and  $p'$  is the number of parameters that are jointly updated. Finally,  $f$  and  $e$  are realizations of uniform and normal random  $p$ -vectors; i.e.,  $F \sim \mathcal{U}_p(-b, b)$  and  $E \sim N(0, b^* I_p)$ , where  $b$  and  $b^*$  are small compared to the width of the posterior density.

For large parameter dimensions  $p$ , DREAM employs a random subspace sampling strategy that can decrease  $p'$  from its original values of  $p$ . This reduces the number of required parallel chains  $N$  so that, in theory, DREAM can run with  $N < p$  as compared with  $N = 2p$  required for DE-MC.

Convergence analysis and case studies for DREAM, DREAM<sub>(D)</sub>, and MT-DREAM<sub>(ZS)</sub> are provided in [144, 259, 261]. Specifically, [261] illustrates applications where DREAM exhibits superior performance to DRAM and DE-MC. For moderate- to high-dimensional problems with computationally intensive codes, DREAM shares the advantage of the parallel DRAM algorithms since both can be implemented on massively parallel architectures. For example, the use of MT-DREAM<sub>(ZS)</sub> to perform Bayesian inference for 241 parameters in a hydrologic model is illustrated in [144].

Because DREAM is newer than DRAM, there are fewer available MATLAB toolboxes that are configured for general usage. However, that will certainly change in the next few years and readers are advised to incorporate both in their libraries of Bayesian parameter estimation routines for large, nonlinear engineering and scientific problems.

## 8.8 Notes and References

MCMC methods have proven highly successful for numerous applications quantified by data-based or statistical models. This is due in part to improved computational resources and the success of techniques such as the use of conjugate priors and Gibbs samplers, which permit parameter by parameter sampling when conditional posterior distributions can be reasonably approximated. However, the application of these techniques to engineering, science, and mathematical models was initially hampered by the following issues:

- the complexity and highly nonlinear dependence on parameters in models prohibited efficient use of conjugate priors and Gibbs samplers;
- appropriate statistical models and likelihood functions were difficult to formulate;
- static proposal functions were ineffective for exploring high-dimensional and complex posterior densities;
- the computational time required for codes associated with phenomena such as nonlinear, coupled, or high-dimensional PDEs prohibited burn-in.

These issues have been addressed in part by the development of algorithms such as DRAM and DREAM that have the adaptive capabilities to explore complex and multimodal posterior distributions and are amenable to implementation on massively parallel architectures. As a result, these algorithms are presently being employed for PDE models such as those employed for climate simulations. It is anticipated that their use will grow substantially as toolboxes evolve and the success of these and emerging algorithms is established.

We have focused primarily on Metropolis algorithms as a prelude for discussing DRAM and DREAM, and hence we have neglected several techniques, such as Gibbs samplers, which have proven highly successful in other contexts. The reader is referred to [92] for details regarding Gibbs samplers and the associated software package BUGS (Bayesian inference Using Gibbs Sampling) which can be implemented from within the statistical package R but is not yet available in MATLAB. Details about importance sampling can also be found in this reference. Sequential Monte Carlo (SMC) methods [75], which are also known as particle filters, can provide a more accurate alternative to extended or unscented Kalman filters for data assimilation when the number of samples is sufficiently large [79]. We refer the reader to [56, 240] for general overviews of Bayesian analysis and [15, 128, 244] for discussion regarding the use of Bayesian inference for parameter estimation. The interpretation of Tikhonov regularization in a Bayesian context is detailed in [27, 128].

## 8.9 Exercises

**Exercise 8.1.** By considering the cases  $x < v$ ,  $x = v$  and  $x > v$ , establish the relation

$$v \min(1, x/v) = \min(x, v) = x \min(1, v/x).$$

**Exercise 8.2.** Consider the steady state heat model of Example 8.12 and the temperature data  $v$  compiled in Table 3.2. Use DRAM to compute chains and marginal densities for the parameters  $Q = [\Phi, h]$ . You should be able to reproduce the results in Example 8.12. Now compute the posterior density  $\pi(q|v)$  directly using Bayes' relation (8.2). You can approximate the integral using tensored 1-D quadrature relations, as detailed in Chapter 11. Compare your posterior density with the joint density obtained using DRAM. Now numerically integrate  $\pi(q|v)$  to construct marginal densities for  $\Phi$  and  $h$  and compare to those constructed using DRAM.

**Exercise 8.3.** Repeat the computations of Example 8.12 using DRAM with the default proposal function  $J(q^*|q^{k-1}) = N(q^{k-1}, D)$  rather than the choice  $J(q^*|q^{k-1}) = N(q^{k-1}, V)$  used to obtain the reported results. Plot the first 200 iterates to show the initial burn-in period. Compare your final chain covariance matrix to that reported in the example.

**Exercise 8.4.** Verify the recursive relation (8.20).

**Exercise 8.5.** Show that  $V$  is singular if we try to estimate  $q = [m, c, k]$  for the spring model

$$m \frac{d^2z}{dt^2} + c \frac{dz}{dt} + kz = 0,$$

$$z(0) = 2, \quad \frac{dz}{dt}(0) = -C.$$

**Exercise 8.6.** Here we are going to use DRAM to construct densities for parameters in the HIV model (8.24) detailed in Example 3.3 and illustrated in Example 8.13. Synthetic data is provided in the file `hiv-data` which can be downloaded from the website <http://www.siam.org/books/cs12>. The seven columns respectively contain the time and values for the six states  $T_1, T_2, T_1^*, T_2^*, V$ , and  $E$  measured every 5 days for 200 days. We are going to construct chains and densities for the parameters  $Q = [d_1, k_2, \delta, b_E]$ .

You should start by writing a MATLAB code that uses `fminsearch` to optimize  $Q$  based on this data. You can use the initial conditions (3.16) and remaining parameter values in Table 3.1.

Now use DRAM to compute densities for  $d_1, k_2, \delta$ , and  $b_E$ . You should monitor your chains to ensure that they have burned-in or converged. Plot the chains, marginal densities, and pairwise scatterplots. We will revisit this problem in Exercise 9.7.

## Chapter 9

# Uncertainty Propagation in Models

*“Prediction is very difficult, especially if it’s about the future,” Niels Bohr*

We noted in Chapter 1 that predictive estimation is comprised of three components: model calibration, model prediction, and estimation of the validation domain. In the first, measured data is used to quantify input uncertainties associated with parameters, initial or boundary conditions, and forcing functions. For certain applications, input uncertainties can be directly quantified from experiments. For phenomenological models with nonphysical parameters, however, experimental input distributions are typically unavailable and densities must be estimated using the Bayesian techniques presented in Chapter 8. This is often referred to as *inverse uncertainty quantification*.

For model prediction, one computes the mean and statistics, prediction intervals, or a pdf, for a model response or quantity of interest (QoI). As detailed in Chapter 1, QoI include average rainfall amounts for a specified area, expected temperature increase over a future time period, or bounds on void fraction distributions that guarantee specified performance levels and safety margins in a nuclear reactor. To construct uncertainty bounds for the QoI, one must propagate input uncertainties through the model while accounting for measurement errors. This is the topic of this chapter and Chapter 10. In Section 9.4, we illustrate the difference between confidence or credible intervals, which quantify the accuracy of model fits, and prediction intervals which incorporate both propagated uncertainties and measurement errors.

Techniques to propagate uncertainties through models include the following.

- Direct Evaluation for Linearly Parameterized Models: Whereas the models discussed in Chapters 2 and 3 typically exhibit nonlinear parameter dependencies, linear parameterized models arise in applications such as image processing and X-ray tomography. We illustrate linear models in Section 9.1 since response uncertainties can be computed explicitly in this case.
- Sampling Methods: These methods, including Monte Carlo techniques, are commonly employed to propagate uncertainties in nonlinearly problems. As

noted in Section 9.2, response and QoI uncertainties can, in certain cases, be constructed with no additional computational cost if the Bayesian techniques of Chapter 8 are used to determine parameter densities. This technique has the advantage of being independent from the number of parameters but the disadvantage that the method converges at a rate of  $\frac{1}{\sqrt{M}}$  where  $M$  is the number of simulations. This is due to the solely statistical nature of the method and the fact that it does not exploit regularity associated with the parameter space. *For problems with correlated parameters or sufficiently large parameter dimensions, however, this may be the best choice.*

- Perturbation Methods: We illustrate in Section 9.3 methods based on truncated Taylor expansions of the model response or QoI evaluated at the parameter mean. To facilitate implementation, first- or second-order expansions are typically employed, which limits the technique's accuracy for applications where the map from inputs to responses is highly nonlinear.
- Spectral Representations: The objective of stochastic Galerkin and collocation methods is to represent uncertain inputs in a manner that facilitates the evaluation of moments and distributions for QoI. This is achieved by employing spectral expansions that exploit the smoothness, often associated with high-dimensional parameter spaces, to improve the convergence of techniques used to specify realizations used to specify QoI. Due to its breadth, we devote Chapter 10 to this method.

As in Chapters 7 and 8, we consider statistical models of the form

$$\Upsilon_i = f_i(Q) + \varepsilon_i, \quad i = 1, \dots, n, \quad (9.1)$$

where  $\Upsilon_i$  and  $\varepsilon_i$  are random  $\nu$ -vectors representing observations and errors and  $f_i(Q) \in \mathbb{R}^\nu$  is the model response which is a random variable due to the random parameters  $Q$ . We assume that  $\varepsilon_i$  are iid and unbiased. We denote realizations of the model response and errors by  $f(q)$  and  $\epsilon$ . In general, the model evaluations  $f_i(q)$  will depend not only on the  $p$  parameters but also on the independent and dependent variables, as detailed in (7.9). We suppress the latter dependencies in the notation since sensitivity analysis and uncertainty quantification are governed by parameter dependence.

For uncertainty propagation, we assume that, at a minimum, mean parameter values  $\bar{q}_i$ , variances  $\text{var}(Q_i)$ , and covariances  $\text{cov}(Q_i, Q_j)$  have been obtained either experimentally or through statistical analysis. If the Bayesian techniques of Chapter 8 have been employed for inverse uncertainty quantification, one additionally has posterior densities  $\pi(q|v)$  for the parameters.

## 9.1 Direct Evaluation for Linear Models

Linearly parameterized models arise in applications including X-ray tomography, inverse heat models, as illustrated in Exercise 3.1, image processing, and acoustic phenomena quantified by convolution models [27, 176, 256]. Furthermore, their

analysis motivates theory required for nonlinearly parameterized models. We illustrate here uncertainty propagation for a linear model where one can construct explicit mean and variance relations.

Consider the linear multiple regression model

$$\Upsilon_i = Q_1 + \sum_{j=2}^p x_{ij} Q_j + \varepsilon_i \quad , \quad i = 1, \dots, n, \quad (9.2)$$

which has the matrix-vector representation

$$\Upsilon = XQ + \varepsilon, \quad (9.3)$$

where  $Q = [Q_1, \dots, Q_p]^T$  are random variables with means  $\bar{q} = [\bar{q}_1, \dots, \bar{q}_p]$ . The design matrix is

$$X = \begin{bmatrix} 1 & x_{12} & \cdots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n2} & \cdots & x_{np} \end{bmatrix}. \quad (9.4)$$

The modeled response is thus  $f(Q) = XQ$ . We note that for polynomial regression, the independent variables may represent powers—e.g.,  $x_{11} = (x/12)$  and  $x_{12} = (x/12)^2$  in Example 9.3—in which case (9.3) exhibits a nonlinear dependence on the independent variables but a linear dependence on parameters. Finally, we assume that the errors are iid and unbiased.

It follows from (4.12) and (4.16) that

$$\begin{aligned} \mathbb{E}[f_i(Q)] &= \bar{q}_1 + \sum_{j=2}^p x_{ij} \bar{q}_j, \\ \text{var}[f_i(Q)] &= \text{var}(Q_1) + \sum_{j=2}^p [x_{ij}^2 \text{var}(Q_j) + 2x_{ij} \text{cov}(Q_1, Q_j)] \\ &\quad + 2 \sum_{j < k}^p x_{ij} x_{ik} \text{cov}(Q_j, Q_k) \end{aligned} \quad (9.5)$$

for  $i = 1, \dots, n$ . For the linear model (9.3), the mean and variance of  $f(Q) = XQ$  can thus be computed directly using mean and covariance values of the parameters.

The development of analogous mean and variance relations for linear models  $\Upsilon = XQ + \varepsilon$  with general design matrices  $X$  is explored in Exercise 9.1.

**Remark 9.1.** The relations (9.5) can be used to construct confidence or credible intervals for the linear model  $f(Q) = XQ$ . As detailed in Section 9.4, this quantifies the accuracy of the model fit but does not indicate the uncertainty associated with subsequent predictions. Prediction intervals are constructed by noting that

$$\begin{aligned} \mathbb{E}(\Upsilon_i) &= \mathbb{E}[f_i(Q)], \\ \text{var}(\Upsilon_i) &= \text{var}[f_i(Q)] + \text{var}(\varepsilon_i) \end{aligned}$$

based on the assumption that the model response and measurement errors are mutually independent and that errors are unbiased.

**Remark 9.2.** Consider the linear model  $f(Q) = XQ$ , where  $Q_1, \dots, Q_p$  are mutually independent and normally distributed,  $Q_i \sim N(\bar{q}_i, \sigma_i^2)$ , so that  $V = \text{diag}(\sigma_1^2, \dots, \sigma_p^2)$ . For  $\bar{y} = X\bar{q}$ , it follows from Theorem 4.21 that

$$f(Q) \sim N(\bar{y}, X^T V X). \quad (9.6)$$

**Example 9.3.** Consider the height-weight data compiled in Table 7.1 of Example 7.6 with the regression model

$$f(Q) = Q_1 + Q_2(x/12) + Q_3(x/12)^2. \quad (9.7)$$

The Bayesian analysis of Chapter 8 yields the parameter and covariance estimates

$$q = [\bar{q}_1, \bar{q}_2, \bar{q}_3]^T = [260.65, -87.74, 11.92]^T,$$

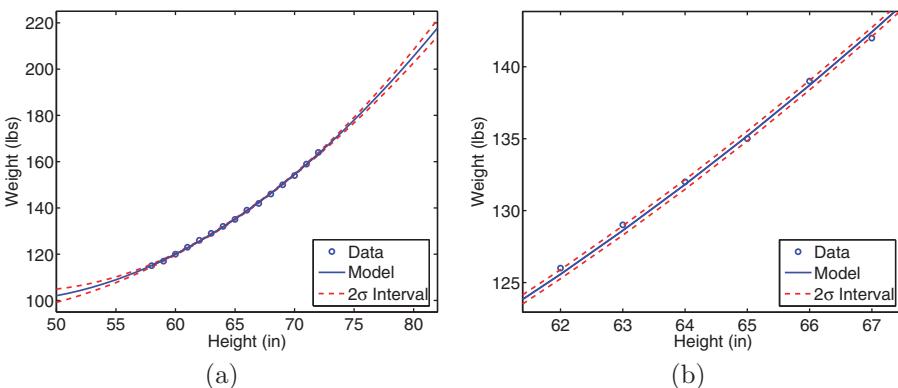
$$V = \begin{bmatrix} \text{var}(Q_1) & \text{cov}(Q_1, Q_2) & \text{cov}(Q_1, Q_3) \\ \text{cov}(Q_2, Q_1) & \text{var}(Q_2) & \text{cov}(Q_2, Q_3) \\ \text{cov}(Q_3, Q_1) & \text{cov}(Q_3, Q_2) & \text{var}(Q_3) \end{bmatrix} = \begin{bmatrix} 778.35 & -287.93 & 26.51 \\ -287.93 & 106.61 & -9.82 \\ 26.51 & -9.82 & 0.91 \end{bmatrix}.$$

The relations (9.5) then yield

$$\begin{aligned} \mathbb{E}[f(Q)] &= \bar{q}_1 + (x/12)\bar{q}_2 + (x/12)^2\bar{q}_3, \\ \text{var}[f(Q)] &= \text{var}(Q_1) + (x/12)^2\text{var}(Q_2) + (x/12)^4\text{var}(Q_3) \\ &\quad + 2(x/12)\text{cov}(Q_1, Q_2) + 2(x/12)^2\text{cov}(Q_1, Q_3) + 2(x/12)^3\text{cov}(Q_2, Q_3), \end{aligned} \quad (9.8)$$

which can be used to predict the expected weights and uncertainties for specified values of the independent height variable  $x$ .

To illustrate, the relations (9.8) were used to predict the expected weights and  $2\sigma$  credible intervals for input heights ranging from 50 to 80 inches. As illustrated in Figure 9.1, the  $2\sigma$  credible intervals are very tight in the region 58 inches to



**Figure 9.1.** (a) Model fit to data with  $2\sigma$  credible interval and (b) expanded perspective in the fitted region.

72 inches used to estimate model parameters. As expected, the credible intervals grow significantly when extrapolating for heights outside the calibration region. This will be further discussed in Section 9.4.2 in the context of prediction intervals.

## 9.2 Sampling Methods

For applications where distributions for measurement errors and input uncertainties—e.g., due to uncertain parameters, initial or boundary conditions, or forcing functions—have been determined either experimentally or using the Bayesian techniques of Chapter 8, sampling methods can often be used to construct distributions for responses or QoI. In principle, this approach is very intuitive and is implemented by randomly sampling from the measurement error and joint input distributions to construct an ensemble of responses from which response statistics and prediction intervals can be computed. The technique has the additional advantage that its efficiency is essentially independent of the number of parameters since one can simultaneously sample from each parameter distribution.

The disadvantage of sampling methods is that they typically exhibit relatively slow convergence rates, and thus a large number of response realizations are required to construct a reasonable statistical ensemble. For example, Monte Carlo techniques have an  $\mathcal{O}(\frac{1}{\sqrt{M}})$  convergence rate where  $M$  is the number of realizations. Hence the number of simulations must be increased by a factor of 100 to gain an additional place of accuracy. In this sense, Monte Carlo methods are decelerating since a factor of four increase in computational effort is required to obtain a factor of two improvement in accuracy. For models that require hours to days for a single evaluation, Monte Carlo sampling will be infeasible unless realizations can be constructed in parallel. Whereas Latin hypercube and quasi-Monte Carlo sampling methods exhibit higher convergence rates, they too are often infeasible for computationally complex problems such as applications involving coupled physics or multicomponent biological systems. Details regarding stochastic quadrature methods are provided in Section 11.1.1.

**Remark 9.4.** The success of sampling methods relies on reasonable representation of the joint density  $\rho_Q(q)$ . For problems in which the joint distribution is assumed to be normal or uniform, sampling from  $\rho_Q(q)$  is highly efficient. For more general distributions of *mutually independent* parameters, one can sample from the marginal distributions and employ (5.2) to construct  $\rho_Q(q)$ . *The difficulty arises for correlated, non-Gaussian, nonuniform parameters, which is typically the case for physical models.* If marginal distributions and a correlation matrix can be constructed, it was noted in Section 5.2 that one can employ the Nataf transformation in combination with a Cholesky decomposition to obtain a representation based on mutually independent Gaussian random variables.

If this information is not available, one can instead use the methods of Chapter 8 to construct prediction intervals. If the prediction domain lies within the calibration domain used for Bayesian parameter estimation, response ensembles can in some cases be computed using the model solutions employed to construct

the likelihoods central to the Metropolis algorithms. For these cases, input and response distributions have the same computational cost. If memory limitations prohibit storage of all ensembles, one can instead sample from the chain indices to construct response densities and prediction intervals. This is illustrated in Example 9.14 for an ODE system with correlated non-Gaussian parameters. *The fact that this technique applies to correlated parameter sets constitutes an advantage since the spectral representations discussed in Chapter 11 require independent parameters or representations of the joint density  $\rho_Q(q)$  which in turn rely on the accuracy and numerical feasibility of the transformation techniques discussed in Section 5.2.*

### 9.3 Perturbation Methods

For large or complex nonlinearly parameterized models, sampling-based uncertainty propagation is often computationally infeasible. In some cases, truncation of multi-dimensional Taylor expansions for  $f(Q)$  yields approximate uncertainty criteria whose accuracy is dictated by the order of the Taylor expansion [50].

For this development, we assume that the density for each component  $Q_i$  is symmetric about a nominal value  $\bar{q}_i$  and consider the representation

$$Q = \bar{q} + \delta Q = [\bar{q}_1 + \delta Q_1, \dots, \bar{q}_p + \delta Q_p]^T, \quad (9.9)$$

where  $\delta Q$  is the vector of perturbations or uncertainties about  $\bar{q}$ . One typically takes  $\bar{q}$  to be the expected parameter values and  $\delta Q$  to be one standard deviation for each of the parameters, but other choices are possible. It is assumed that realized values for  $\bar{q}$  and  $\delta Q$  have been determined either experimentally or using the model calibration techniques discussed in Chapters 7 and 8.

#### Single Response

Consider first the case of a single response so that  $n = \nu = 1$ . The  $N^{th}$ -order Taylor expansion of  $f(Q)$  about nominal values  $\bar{q}$  with perturbations  $\delta Q$  is

$$\begin{aligned} f(Q) &= f(\bar{q}_1 + \delta Q_1, \dots, \bar{q}_p + \delta Q_p) \\ &\approx f(\bar{q}) + \sum_{i_1=1}^p \frac{\partial f}{\partial Q_{i_1}} \Big|_{\bar{q}} \delta Q_{i_1} + \frac{1}{2} \sum_{i_1, i_2=1}^p \frac{\partial^2 f}{\partial Q_{i_1} \partial Q_{i_2}} \Big|_{\bar{q}} \delta Q_{i_1} \delta Q_{i_2} \\ &\quad + \dots + \frac{1}{N!} \sum_{i_1, \dots, i_N=1}^p \frac{\partial^N f}{\partial Q_{i_1}, \dots, \partial Q_{i_N}} \Big|_{\bar{q}} \delta Q_{i_1} \dots \delta Q_{i_N}. \end{aligned} \quad (9.10)$$

We subsequently consider the first-order (linear) expansion

$$f(Q) = \bar{y} + \sum_{i=1}^p s_i \delta Q_i, \quad (9.11)$$

which is reindexed since we are assuming a single response. Here  $\bar{y} = f(\bar{q})$  and  $s_i = \frac{\partial f}{\partial Q_i}(\bar{q})$  is the sensitivity of the response to the  $i^{th}$  parameter evaluated at  $\bar{q}$ .

For a random variable  $Q = [Q_1, \dots, Q_p]$  with joint pdf  $\rho_Q(q)$ , the parameter means, variances, and covariances can be expressed as

$$\begin{aligned}\mathbb{E}(Q_i) &= \bar{q}_i, \\ \text{var}(Q_i) &= \int_{\mathbb{R}^p} (q_i - \bar{q}_i)^2 \rho_Q(q) dq = \int_{\mathbb{R}^p} (\delta q_i)^2 \rho_Q(q) dq, \\ \text{cov}(Q_i, Q_j) &= \int_{\mathbb{R}^p} (q_i - \bar{q}_i)(q_j - \bar{q}_j) \rho_Q(q) dq = \int_{\mathbb{R}^p} \delta q_i \delta q_j \rho_Q(q) dq,\end{aligned}\tag{9.12}$$

where  $\delta q_i = q_i - \bar{q}_i$ . It follows from (9.11) that

$$\mathbb{E}[f(Q)] = \bar{y} \int_{\mathbb{R}^p} \rho_Q(q) dq + \sum_{i=1}^p s_i \int_{\mathbb{R}^p} (q_i - \bar{q}_i) \rho_Q(q) dq = \bar{y}.\tag{9.13}$$

The first integral is unity since  $\rho_Q(q)$  is a density, whereas the symmetry of the second integrand yields an integral of zero. The variance of  $f(Q)$  is

$$\begin{aligned}\text{var}[f(Q)] &= \mathbb{E}[(f(Q) - \bar{y})^2] \\ &= \int_{\mathbb{R}^p} \left( \sum_{i=1}^p s_i \delta q_i \right)^2 \rho_Q(q) dq \\ &= \sum_{i=1}^p s_i^2 \int_{\mathbb{R}^p} (\delta q_i)^2 \rho_Q(q) dq + \sum_{i=1}^p \sum_{j=1, j \neq i}^p s_i s_j \int_{\mathbb{R}^p} (\delta q_i)(\delta q_j) \rho_Q(q) dq \\ &= \sum_{i=1}^p s_i^2 \text{var}(Q_i) + \sum_{i=1}^p \sum_{j=1, j \neq i}^p s_i s_j \text{cov}(Q_i, Q_j).\end{aligned}\tag{9.14}$$

We note that (9.14) can be written as the matrix relation

$$\text{var}[f(Q)] = S^T V S,\tag{9.15}$$

where  $V$  is the covariance matrix for  $Q$  and  $S^T = [s_1, \dots, s_p]$  is the row vector of response sensitivities. The relation (9.15) is sometimes referred to as the *sandwich relation*.

**Remark 9.5.** If  $Q_1, \dots, Q_p$  are mutually independent and normally distributed,  $Q_i \sim N(\bar{q}_i, \sigma_i^2)$ , so that  $V = \text{diag}(\sigma_1^2, \dots, \sigma_p^2)$ , it follows from Theorem 4.21 that

$$f(Q) \sim N(\bar{y}, S^T V S)\tag{9.16}$$

since  $f(Q)$  given by (9.11) is linearly parameterized.

### Multiple Responses

The mean and variance relations can be directly extended to the case of  $n$  random variables  $\Upsilon = [\Upsilon_1, \dots, \Upsilon_n]^T$  and model responses  $f(Q)$ . The linear Taylor expansion of  $f(Q)$  about the nominal value  $\bar{q}$  with perturbations  $\delta Q$  is

$$f(Q) \approx f(\bar{q}) + S\delta Q,$$

where  $S$  is the  $n \times p$  sensitivity matrix having elements  $[S]_{ij} = \frac{\partial f_i}{\partial Q_j}$ . The mean

$$\mathbb{E}[f(Q)] = \bar{y} = f(\bar{q}) \quad (9.17)$$

and covariance matrix

$$\text{var}[f(Q)] = S^T V S \quad (9.18)$$

are obtained in a manner analogous to the single response relations (9.13) and (9.15); see Exercise 9.2.

This Taylor series-based technique is often referred to as *propagation of moments* or *propagation of errors*, and the relations (9.13)–(9.18) are often termed *moment propagation relations*. The first-order Taylor expansion (9.11) is linear and, as illustrated in Example 9.7, will have limited accuracy in highly nonlinear regimes. Expressions for propagating higher-order moments, computed by retaining additional terms in the Taylor expansion (9.10), are provided in [50]. However, their complexity can preclude their use unless the required accuracy justifies their inclusion.

Finally, we point out that unlike Monte Carlo techniques, which provide a density for the response, the Taylor-based method provides only an estimate for the mean and variance or covariance. Hence it provides a measure of the magnitude but not the shape of the uncertainty.

**Example 9.6.** We revisit the height-weight Example 9.3 with the linearly parameterized model (9.7). The sensitivity and covariance matrices are

$$S^T = \left[ \frac{\partial f}{\partial Q_1}, \frac{\partial f}{\partial Q_2}, \frac{\partial f}{\partial Q_3} \right] = [1, (x/12), (x/12)^2]$$

and

$$V = \begin{bmatrix} \text{var}(Q_1) & \text{cov}(Q_1, Q_2) & \text{cov}(Q_1, Q_3) \\ \text{cov}(Q_2, Q_1) & \text{var}(Q_2) & \text{cov}(Q_2, Q_3) \\ \text{cov}(Q_3, Q_1) & \text{cov}(Q_3, Q_2) & \text{var}(Q_3) \end{bmatrix}$$

so that

$$\begin{aligned} \mathbb{E}[f(Q)] &= \bar{q}_1 + (x/12)\bar{q}_2 + (x/12)^2\bar{q}_3, \\ \text{var}[f(Q)] &= \text{var}(Q_1) + (x/12)^2\text{var}(Q_2) + (x/12)^4\text{var}(Q_3) \\ &\quad + 2(x/12)\text{cov}(Q_1, Q_2) + 2(x/12)^2\text{cov}(Q_1, Q_3) + 2(x/12)^3\text{cov}(Q_2, Q_3). \end{aligned}$$

As expected, these results are identical to the directly computed mean and variance since the first-order Taylor expansion is exact for linear problems.

**Example 9.7.** In (3.8) of Example 3.2, we showed that the amplitude of the harmonic oscillator equation

$$m \frac{d^2 z}{dt^2} + c \frac{dz}{dt} + kz = f_0 \cos(\omega_F t),$$

$$z(0) = z_0, \quad \frac{dz}{dt}(0) = z_1$$

is  $Z_0(Q) = \frac{f_0}{\sqrt{m^2(\omega_0^2 - \omega_F^2)^2 + c^2\omega_F^2}}$ , where  $\omega_0 = \sqrt{k/m}$  is the natural frequency. In this example, we compare the sampling-based techniques of Section 9.2 with perturbation methods for computing variability in the response

$$y = f(\omega_F, Q) = \frac{Z_0(Q)}{f_0} = \frac{1}{\sqrt{(k - m\omega_F^2)^2 + (c\omega_F)^2}}$$

under the assumption that the parameters  $Q = [m, c, k]^T$  are normally distributed with mean  $\bar{q} = [2.7, 0.24, 8.5]$  and covariance matrix

$$V = \begin{bmatrix} 0.002^2 & 0 & 0 \\ 0 & 0.065^2 & 0 \\ 0 & 0 & 0.001^2 \end{bmatrix}. \quad (9.19)$$

We note that the largest relative uncertainty is associated with the damping parameter  $c$ , which is often the case in damped oscillating systems. Finally, the mean natural frequency is  $\bar{\omega}_0 = 1.7743$  Hz.

The sensitivities as a function of the drive frequency  $\omega_F$  are

$$\frac{\partial f}{\partial m} = \frac{(k - m\omega_F^2)\omega_F^2}{[(k - m\omega_F^2)^2 + (c\omega_F)^2]^{3/2}},$$

$$\frac{\partial f}{\partial c} = \frac{-c\omega_F^2}{[(k - m\omega_F^2)^2 + (c\omega_F)^2]^{3/2}},$$

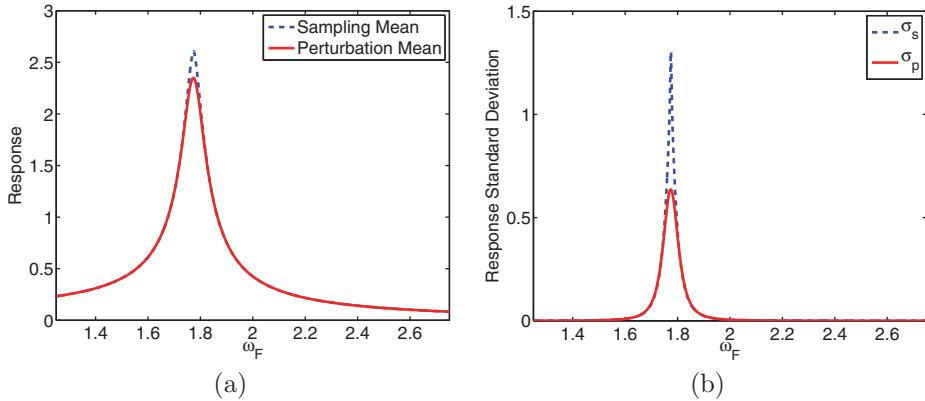
$$\frac{\partial f}{\partial k} = \frac{-(k - m\omega_F^2)}{[(k - m\omega_F^2)^2 + (c\omega_F)^2]^{3/2}},$$

and  $S^T = [\frac{\partial f}{\partial m}, \frac{\partial f}{\partial c}, \frac{\partial f}{\partial k}]$ . Since the parameters are mutually independent and normally distributed, it follows from Remark 9.5 that the response is also normally distributed for each input frequency  $\omega_F$  with mean  $\bar{y}_p = f(\bar{q})$  and variance  $\sigma_p^2 = S^T V S$ . We remind the reader that the normal response is due to the fact that we are using a first-order expansion (9.11) that is linear with regard to the parameters.

For the Monte Carlo sampling method, we computed  $M = 10,000$  realization of the response  $f(\omega_F, q)$  using random samples from the parameter distributions. For each value of  $\omega_F$ , we constructed mean and standard deviation values

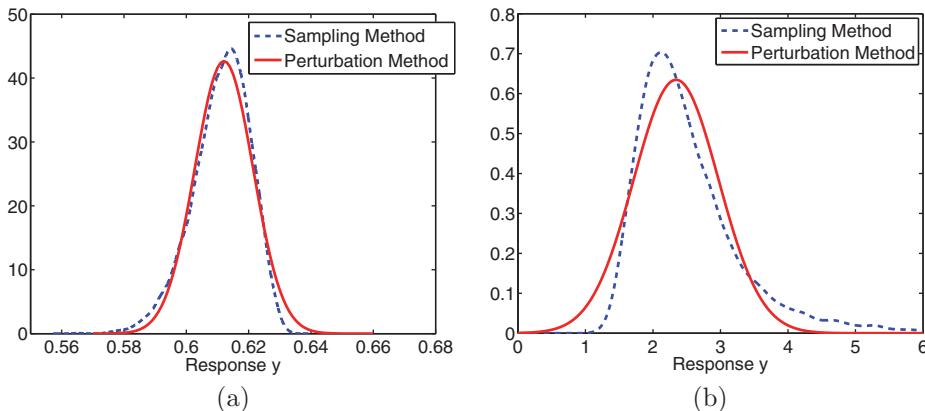
$$\bar{y}_s(\omega_F) = \frac{1}{M} \sum_{m=1}^M f(\omega_F, q^m), \quad (9.20)$$

$$\sigma_s(\omega_F) = \sqrt{\frac{1}{M-1} \sum_{m=1}^M [f(\omega_F, q^m) - \bar{y}_s(\omega_F)]^2}.$$

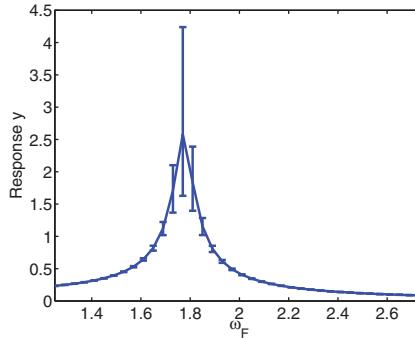


**Figure 9.2.** (a) Sampling mean  $\bar{y}_s$  from (9.20) and perturbation mean  $\bar{y}_p = f(\bar{q})$ , and (b) standard deviation  $\sigma_s$  from (9.20) and  $\sigma_p = S^T V S$ .

The mean and standard deviation, for a range of input frequencies, are compared in Figure 9.2 for the two methods. It is observed that the values agree for most frequencies but are significantly different near the natural frequency where the nonlinear parameter effects are most significant. In Figure 9.3, we compare the densities constructed using the two methods, near and away from resonance. The kde techniques of Section 4.1 were used to construct the sampling-based density, whereas the Gaussian perturbation method density is given by (9.16). It is observed that away from resonance, the sampling-based density is approximately normal, whereas it is skewed and relatively nonnormal at  $\omega_F = 1.77$  Hz, which is close to the natural frequency of  $\omega_0 = 1.7743$  Hz. This illustrates further ramifications of the nonlinear effects that are not accurately quantified using the linear perturbation relation.



**Figure 9.3.** Densities obtained using the sampling and linear perturbation methods at (a) 1.60 Hz and (b) 1.77 Hz.



**Figure 9.4.** 95% credible intervals computed using the sampling method.

The 95% credible intervals, obtained using the sampling method, are plotted in Figure 9.4. These are easily constructed by ordering the realizations in ascending order and determining the location of the 0.05 and 0.95 values. The asymmetry in credible values near resonance reflects the asymmetry of the densities due to nonlinear effects.

In summary, this illustrates the limited accuracy of the linear perturbation expansion when the input to response map is highly nonlinear. The accuracy of perturbation methods can be improved by including second- and higher-order contributions but at greater computational cost.

## 9.4 Prediction Intervals

The linear analysis, sampling techniques, and perturbation methods discussed in Sections 9.1–9.3 address the propagation of parameter uncertainties through models. If Bayesian techniques are used to quantify parameter uncertainties, these propagation methods will indirectly incorporate measurement errors  $\varepsilon_i$  since they influence the variability of parameters. However, the direct influence of measurement errors is neglected in these techniques and must be incorporated when constructing prediction intervals for QoI. To illustrate the difference between confidence, credible, and prediction intervals, we consider first linear regression.

### 9.4.1 Confidence versus Prediction Intervals

Consider the linear model

$$\Upsilon = Xq + \varepsilon, \quad (9.21)$$

where  $q_0 \in \mathbb{R}^p$  is assumed fixed but unknown. Errors  $\varepsilon = [\varepsilon_1, \dots, \varepsilon_n]$  are assumed to be independent and normally distributed with  $\varepsilon_i \sim N(0, \sigma_0^2)$ . An example is the regression equation

$$\Upsilon_i = q_1 + \sum_{j=2}^p x_{ij} q_j + \varepsilon_i, \quad i = 1, \dots, n, \quad (9.22)$$

where the design matrix  $X$  is specified in (9.4).

We assume that given measured data, (7.17) and (7.18) are used to compute the parameter estimator  $\hat{q}$  and estimate  $q$  for the true but unknown parameter  $q_0$ . We now consider two types of prediction at a point  $x_0$  in the domain of the independent variable  $x$  but not among the data used to estimate  $\hat{q}$  and  $q$ . For the regression equation (9.22),  $x_0 = [1, x_{02}, \dots, x_{0p}]^T$ .

The first is the prediction of a new observation  $\Upsilon_{x_0}$  at  $x_0$ , whereas the second is the prediction of the mean response  $\mu_{x_0} = \mathbb{E}(\Upsilon_{x_0})$ . As we will illustrate, the interval estimates differ for the two cases.

Consider first the estimation of the mean response  $\mathbb{E}(\Upsilon_{x_0})$ . We note that

$$\hat{\Upsilon}_{x_0} = x_0^T \hat{q}$$

is an unbiased point estimator for  $\mathbb{E}(\Upsilon_{x_0})$ . Furthermore, it follows from (4.17) and property (ii) of (7.19) that

$$\text{var}(\hat{\Upsilon}_{x_0}) = \sigma_0^2 [x_0^T (X^T X)^{-1} x_0]. \quad (9.23)$$

Since  $\sigma_0^2$  is typically unknown, we employ the variance estimator  $\hat{\sigma}^2$  specified in (7.20), which yields the estimator

$$\hat{\sigma}^2(\hat{\Upsilon}_{x_0}) = \hat{\sigma}^2 [x_0^T (X^T X)^{-1} x_0]. \quad (9.24)$$

With the assumption that  $\varepsilon_i \sim N(0, \sigma_0^2)$ , it follows from Property 7.8 that  $\hat{\Upsilon}_{x_0}$  is a linear combination of joint multivariate normal random variables, which implies that its sampling distribution is a normal distribution with mean  $\mu_{x_0}$  and variance (9.23) so that

$$\frac{\hat{\Upsilon}_{x_0} - \mu_{x_0}}{\sigma_0 \sqrt{x_0^T (X^T X)^{-1} x_0}} \sim N(0, 1).$$

From Definition 4.12, (7.20), and the independence of  $\hat{q}$  and  $\hat{\sigma}$ , it follows that

$$T = \frac{\hat{\Upsilon}_{x_0} - \mu_{x_0}}{\hat{\sigma} \sqrt{x_0^T (X^T X)^{-1} x_0}}$$

has a  $t$ -distribution with  $n - p$  degrees of freedom. The  $(1 - \alpha) \times 100\%$  interval estimator for  $\mu_{x_0}$  is thus

$$\left[ \hat{\Upsilon}_{x_0} \pm t_{n-p, 1-\alpha/2} \cdot \hat{\sigma} \sqrt{x_0^T (X^T X)^{-1} x_0} \right].$$

This is a *confidence interval* since  $\mu_{x_0}$  is a linear combination of parameters.

We now consider the construction of interval estimates for the new prediction  $\Upsilon_{x_0}$ . We again assume that the estimators  $\hat{q}$  and  $\hat{\sigma}^2$  have been computed using previous data in which case  $\Upsilon_{x_0}$  will be independent from  $\hat{q}$  and  $\hat{\sigma}$ . It thus follows that the random variable  $\Upsilon_{x_0} - \hat{\Upsilon}_{x_0}$  will be normally distributed with mean

$$\mathbb{E}(\Upsilon_{x_0} - \hat{\Upsilon}_{x_0}) = 0$$

and variance

$$\begin{aligned}\text{var}(\Upsilon_{x_0} - \hat{\Upsilon}_{x_0}) &= \text{var}(\Upsilon_{x_0}) + \text{var}(\hat{\Upsilon}_{x_0}) \\ &= \sigma_0^2 [1 + x_0^T (X^T X)^{-1} x_0].\end{aligned}\quad (9.25)$$

It follows immediately that

$$\frac{\hat{\Upsilon}_{x_0} - \Upsilon_{x_0}}{\sigma_0 \sqrt{1 + x_0^T (X^T X)^{-1} x_0}} \sim N(0, 1) \quad (9.26)$$

and

$$T = \frac{\hat{\Upsilon}_{x_0} - \Upsilon_{x_0}}{\hat{\sigma} \sqrt{1 + x_0^T (X^T X)^{-1} x_0}}$$

has a  $t$ -distribution with  $n - p - 1$  degrees of freedom. The interval estimator for  $\Upsilon_{x_0}$  is

$$\left[ \hat{\Upsilon}_{x_0} \pm t_{n-p, 1-\alpha/2} \cdot \hat{\sigma} \sqrt{1 + x_0^T (X^T X)^{-1} x_0} \right]. \quad (9.27)$$

This is termed the *prediction interval* for  $\Upsilon_{x_0}$ . It is constructed by using the point estimates  $v_{x_0}$  and  $\sigma$  for  $\hat{\Upsilon}_{x_0}$  and  $\hat{\sigma}$ .

**Definition 9.8 (Prediction Interval).** The  $(1 - \alpha) \times 100\%$  prediction interval for a random response  $\Upsilon_{x_0}$  to the linear model (9.21) is the pair of statistics  $[\Upsilon_L(X), \Upsilon_R(X)]$  constructed from a random sample  $X$  such that

$$P(\Upsilon_L(X) \leq \Upsilon_{x_0} \leq \Upsilon_R(X)) = 1 - \alpha,$$

where  $\Upsilon_{x_0}$  is a new observation at the point  $x_0$  that is independent of the data used to construct  $\Upsilon_L(X)$  and  $\Upsilon_R(X)$ .

**Remark 9.9.** In Section 7.3, we illustrated that linearization of the nonlinear regression model

$$\Upsilon = f(q_0) + \varepsilon$$

about  $q_0$  yielded estimators that were analogous to the linear case with  $X$  replaced by the sensitivity matrix  $\mathcal{X}(q)$  where  $\mathcal{X}_{ij}(q) = \frac{\partial f_i(q)}{\partial q_j}$ . As detailed in [219], the same is true for the prediction interval, which is

$$\left[ \hat{\Upsilon}_{x_0} \pm t_{n-p, 1-\alpha/2} \cdot \hat{\sigma} \sqrt{1 + x_0^T (\mathcal{X}^T \mathcal{X})^{-1} x_0} \right].$$

### 9.4.2 Extrapolation

To illustrate the increased uncertainty inherent to extrapolation, we consider the univariate regression problem

$$\Upsilon_i = q_1 + x_i q_2 + \varepsilon_i,$$

which is simply (9.22) with  $p = 2$ . The design matrix in this case is

$$X = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad (9.28)$$

so that

$$(X^T X)^{-1} = \frac{1}{n S_{xx}} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{bmatrix}, \quad (9.29)$$

where  $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$  and  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$  is the mean of the points used to estimate  $q_1$  and  $q_2$ ; see Exercise 9.3. The point estimator for  $\mathbb{E}(\Upsilon_{x_0})$  is now

$$\hat{\Upsilon}_{x_0} = \hat{q}_1 + \hat{q}_2 x_0$$

and, as developed in Exercise 9.4,

$$\text{var}(\hat{\Upsilon}_{x_0}) = \sigma_0^2 \left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right). \quad (9.30)$$

The confidence interval for  $\mu_{x_0} = \mathbb{E}(\hat{\Upsilon}_{x_0})$  is

$$\left[ \hat{\Upsilon}_{x_0} \pm t_{n-2, 1-\alpha/2} \cdot \hat{\sigma}^2 \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \right], \quad (9.31)$$

whereas the prediction interval for  $\Upsilon_{x_0}$  is

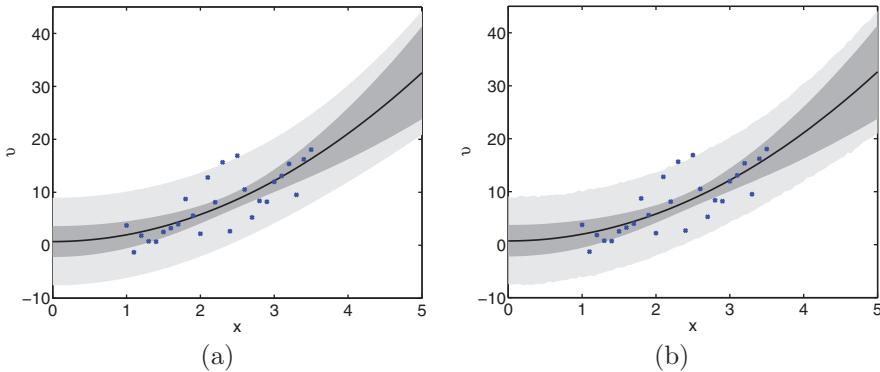
$$\left[ \hat{\Upsilon}_{x_0} \pm t_{n-2, 1-\alpha/2} \cdot \hat{\sigma}^2 \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \right]. \quad (9.32)$$

The relations (9.31) and (9.32) demonstrate that the quality of both the fit and prediction degrade as the distance  $|x_0 - \bar{x}|$  increases. In particular, care must be exercised if extrapolating with  $x_0$  outside the region  $[x_{\min}, x_{\max}]$ . Details regarding extrapolation for the multivariate problem can be found in [96].

**Remark 9.10.** Prediction intervals constructed in this manner are pointwise in the sense that they apply to individually specified future measurements. This is weaker than the concept of simultaneous  $(1 - \alpha) \times 100\%$  prediction intervals, which specify the probability that all future measurements or model evaluations lie within the interval. As detailed in [219], the Bonferroni method is one technique for constructing the wider simultaneous prediction intervals.

**Example 9.11.** To illustrate these concepts, consider the model

$$\Upsilon_i = q_1 + q_2 x_i^2 + \varepsilon_i,$$



**Figure 9.5.** Data, point estimates, and (a) 95% confidence and prediction intervals obtained using the linear theory of Section 9.4.1, and (b) 95% credible and prediction intervals produced by the Bayesian analysis discussed in Example 9.13.

where the true parameters are  $q_0 = [0.6, 1.2]$ . We specify  $\varepsilon_i \sim N(0, \sigma_0^2)$ , with  $\sigma_0 = 3$ , to construct synthetic data  $v_i$  at the 26 points  $x_i$  shown in Figure 9.5(a). In the same figure, we plot the 95% confidence and prediction intervals specified in Section 9.4.1 and (9.32) using the estimated parameter values  $q = [0.6920, 1.2752]$ . We note that both intervals are tightest at  $\bar{x} = 2.25$  and that the uncertainty associated with both the model fit and predictions grows more rapidly for extrapolation outside the region  $[x_{\min}, x_{\max}] = [1, 3.5]$ . The observation that 2 out of the 26 data points lie outside the 95% prediction interval is consistent with its definition.

### 9.4.3 Prediction Intervals for Uncertainty Quantification

The relations (9.27) and (9.32) for the prediction intervals are based on the sampling distribution (9.26) for  $\Upsilon_{x_0} - \hat{\Upsilon}_{x_0}$ , where  $\hat{\Upsilon}_{x_0}$  is the estimator for the fixed, but unknown, parameter  $\mu_{x_0}$ . As noted in Remark 7.1, sampling distributions do not always correspond with distributions for associated random variables, so care must be exercised if considering them for uncertainty quantification. For example, use of the relations (9.27) and (9.32) will clearly yield inaccurate prediction intervals when the actual distribution for the mean response is highly non-Gaussian since the relation (9.26) for the sampling distribution is Gaussian. Hence while the analysis of Section 9.4.1 motivates the general framework required to construct prediction intervals for uncertainty quantification, we must interpret the construction of response variances in a manner consistent with the assumption that inputs and QoI are random variables with associated distributions.

From (9.25), we note that the variance of  $\Upsilon_{x_0} - \hat{\Upsilon}_{x_0}$  is the sum of the measurement error variance  $\sigma_0^2$  and the variance  $\hat{\sigma}^2(\hat{\Upsilon}_{x_0})$  associated with the mean model response. The measurement error can be specified from experiments or estimated using the Bayesian techniques of Chapter 8. The distribution for the model response is constructed by propagating input uncertainties through the model using the techniques detailed in Sections 9.1–9.3 and Chapter 10. The model response

distribution yields the *credible interval*, which is a measure of the model fit. The sum of the propagated uncertainty and measurement errors provides the *prediction interval*, which is typically required for predictive estimation.

**Remark 9.12.** For the linear and linearized models discussed in Sections 9.1 and 9.3,  $\hat{\sigma}^2(\hat{Y}_{x_0})$  is specified by (9.24), whereas orthogonality properties of the Hermite or Legendre polynomials used to represent Gaussian or uniform distributions can be exploited to provide relations for  $\hat{\sigma}^2(\hat{Y}_{x_0})$  when using the spectral methods of Chapter 10. If no other moments are specified, this will yield Gaussian distributions for the propagated response and credible intervals that will likely agree with (9.24), which is based on the sampling distribution. The sampling methods of Section 9.3 can be used to directly construct credible intervals for non-Gaussian distribution. It is illustrated in Examples 9.13 and 9.14 that the prediction intervals constructed by sampling from the indices of the DRAM chains are consistent with the sampling distribution theory of Sections 9.4.1 and 9.4.2 for certain problems but permits the quantification of non-Gaussian response intervals for nonlinear problems with potentially non-Gaussian parameter distributions.

**Example 9.13.** We revisit Example 9.11 but, in this case, we employ the Bayesian theory of Chapter 8 to construct densities for the parameters  $Q_1$  and  $Q_2$  and measurement variance  $\sigma$ . We then propagate the input uncertainties through the model to construct credible intervals and prediction intervals comprised of the summed measurement and response uncertainties. The results obtained using the DRAM algorithm described in Section 8.6 are plotted in Figure 9.5(b). A comparison with Figure 9.5(a) illustrates that in this case, the propagated uncertainties agree with those computed using the sampling distribution theory of Sections 9.4.1 and 9.4.2. As illustrated in the next example, this will not be true in general.

**Example 9.14.** In Example 8.13, we illustrated the use of the DRAM algorithm to construct densities for the parameters  $Q = [b_E, \delta, d_1, k_2, \lambda_1, K_b]^T$  and error variance  $\sigma^2$  in the HIV model

$$\begin{aligned}\dot{T}_1 &= \lambda_1 - d_1 T_1 - (1 - \varepsilon) k_1 V T_1, \\ \dot{T}_2 &= \lambda_2 - d_2 T_2 - (1 - f\varepsilon) k_2 V T_2, \\ \dot{T}_1^* &= (1 - \varepsilon) k_1 V T_1 - \delta T_1^* - m_1 E T_1^*, \\ \dot{T}_2^* &= (1 - f\varepsilon) k_2 V T_2 - \delta T_2^* - m_2 E T_2^*, \\ \dot{V} &= N_T \delta (T_1^* + T_2^*) - c V - [(1 - \varepsilon) \rho_1 k_1 T_1 + (1 - f\varepsilon) \rho_2 k_2 T_2] V, \\ \dot{E} &= \lambda_E + \frac{b_E (T_1^* + T_2^*)}{T_1^* + T_2^* + K_b} E - \frac{d_E (T_1^* + T_2^*)}{T_1^* + T_2^* + K_d} E - \delta_E E\end{aligned}$$

of [2, 3]. Here  $T_1$  and  $T_1^*$  represent the populations of uninfected and infected T-lymphocytes,  $T_2$  and  $T_2^*$  are corresponding macrophage populations, and  $V, E$  denote the populations of free virus and immune effector cells. The pairwise joint

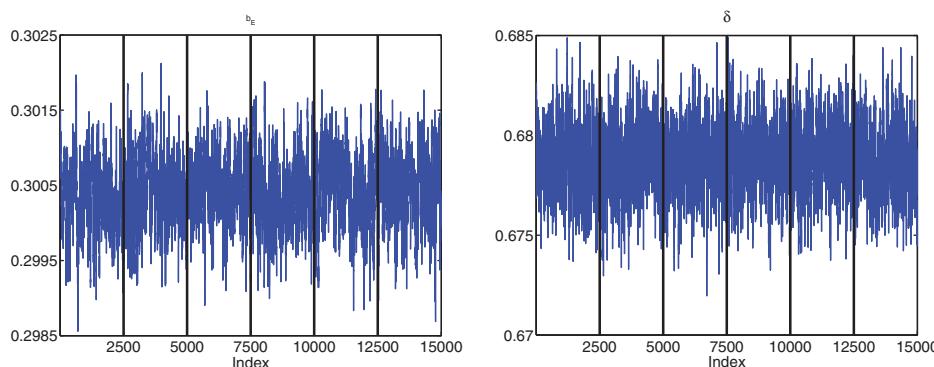
sample plots in Figure 8.15 illustrate that  $k_2, \delta$  and  $\lambda_1, d_1$  are correlated, which violates the assumption of mutually independent parameters, as generally required for the stochastic Galerkin, collocation, and discrete projection methods of Chapter 10. In this example, we illustrate the construction of credible and prediction intervals for the states by randomly sampling from the burned-in parameter chains illustrated in Figure 8.13, as illustrated for two chains in Figure 9.6.

To construct 95% credible intervals, we sampled by index from the parameter chains to construct 5000 realizations of the model. This uncertainty was added to the estimated error variance to construct prediction intervals. These credible and prediction intervals, along with the point estimates and synthetic data, are illustrated for the immune effector cell count  $E$  in Figure 9.7. We note that both intervals are asymmetric with respect to the point estimates between 30 and 50 s. This is due to the highly nonlinear nature of the problem in combination with the slightly non-Gaussian parameter densities shown in Figure 8.14. This asymmetric and non-Gaussian behavior of the response would not be quantified using the sampling distribution theory of Sections 9.4.1 and 9.4.2. This illustrates the necessity of propagating input uncertainties through the model to accurately quantify the distribution for the QoI.

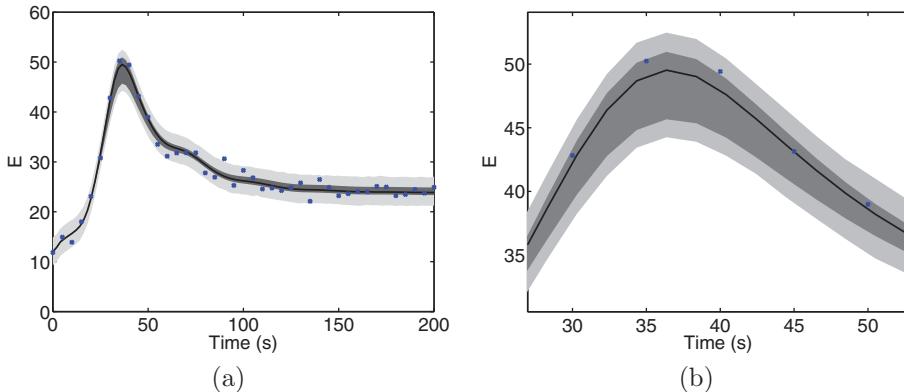
## 9.5 Notes and References

For certain applications, operator-based or moment methods can be used to quantify response uncertainties. Because these methods are quite problem-dependent, we do not focus on them but instead refer readers to [266] for details regarding this methods.

The reader is referred to [50, 53] for details regarding the perturbation techniques of Section 9.4 and further discussion of their use for sensitivity analysis and uncertainty quantification. These perturbation techniques are also employed in the “best-estimate” data assimilation framework detailed in [52, 54]. This framework is based on the principle of maximum entropy, and it provides best-estimate calibrated



**Figure 9.6.** Five index-based samples from the chains for  $b_E$  and  $\delta$ .



**Figure 9.7.** Mean, synthetic data, and 95% credible and prediction intervals for  $E$ : (a) full time interval and (b) reduced time interval to illustrate asymmetric and non-Gaussian behavior.

parameters and responses along with parameter, response, and parameter-response covariance matrices. The response covariance matrix can be directly used to construct credible intervals for the model response along with prediction intervals if the measurement errors are incorporated in the manner detailed in Section 9.4.3. The theory and applications in [50, 52, 53] are based on first-order, and hence linear, expansions which can limit their accuracy in highly nonlinear regimes. Theory and applications utilizing higher-order expansions are provided in [197].

## 9.6 Exercises

**Exercise 9.1.** Consider the linearly parameterized problem

$$\Upsilon = XQ + \varepsilon,$$

where  $Q = [Q_1, \dots, Q_p]$  and  $\varepsilon = [\varepsilon_1, \dots, \varepsilon_n]$  are random vectors and  $X \in \mathbb{R}^{n \times p}$  is a deterministic and known matrix. We assume that  $\varepsilon \sim N(0, V)$ , where  $V = \sigma^2 I_{n \times n}$ , and that the mean  $\bar{q}$  and covariance matrix  $V_q$  for  $Q$  are known. Use the results from Exercise 4.5 to establish mean and covariance relations for  $\Upsilon$  analogous to (9.5).

**Exercise 9.2.** Derive the relations (9.17) and (9.18) for the mean and variance of  $\Upsilon$  when there are  $n$  random variables and model responses.

**Exercise 9.3.** For the design matrix  $X$  defined in (9.28), use Cramer's rule to derive the inverse relation (9.29).

**Exercise 9.4.** Use (9.29) to derive (9.30) using the matrix relation (9.23).

**Exercise 9.5.** Consider the heat equation with constant diffusivity  $\alpha$  and  $f(t, x) = T_\ell = T_R = 0$ . It was established in Exercise 3.1 that a finite difference discretization

with an  $N + 1$  spatial gridpoint yields the vector relation

$$\mathcal{T}^{j+1} = A^{j+1} \mathcal{T}^0,$$

where  $\mathcal{T}^j = [T_{1,j}, \dots, T_{N-1,j}]^T$  and the  $(N - 1) \times (N - 1)$  matrix  $A$  is defined in (3.51). Here we take  $\alpha = 2$ ,  $x_i = -1 + ih$  with  $h = \frac{1}{4}$  and consider the initial heat distribution  $\mathcal{T}^0$  to be a random vector with mean  $\bar{\mathcal{T}}^0 = \sin[\frac{\pi}{2}(x_i + 1)]$ ,  $i = 1, \dots, 7$ , and covariance matrix  $V = 0.1^2 I_{7 \times 7}$ . The temporal stepsize is taken to be  $k = 0.01$  to satisfy the stability constraint

$$k \leq \frac{h^2}{2\alpha}.$$

Let  $\mathcal{T}^f$  denote the solution at  $t_f = \frac{1}{3}$ , and consider the statistical model

$$\mathcal{T}^f = \mathcal{A}\mathcal{T}^0 + \varepsilon,$$

where  $\mathcal{A} = A^{j+1}$  and the measurement error is iid and  $\varepsilon \sim N(0, 0.05^2 I_{7 \times 7})$ . Use the results of Exercise 9.1 to compute credible and prediction intervals for  $\mathcal{T}^f$ , and discuss their relation to the distribution for  $\mathcal{T}^0$  in the context of the diffusive nature of the heat equation.

**Exercise 9.6.** Consider the steady state heat model

$$\begin{aligned} \frac{d^2 T_s}{dx^2} &= \frac{2(a+b)}{ab} \frac{h}{k} [T_s(x) - T_{amb}], \\ \frac{dT_s}{dx}(0) &= \frac{\Phi}{k} \quad , \quad \frac{dT_s}{dx}(L) = \frac{h}{k} [T_{amb} - T_s(L)] \end{aligned}$$

detailed in Example 3.5 and illustrated in Example 8.12 in the context of Bayesian model calibration. Here we are going to employ a subset of the data from Table 3.2 to construct prediction intervals that extrapolate beyond the calibration domain. Employ the DRAM code discussed in Section 8.6 to construct densities for  $\Phi$ ,  $h$ , and  $\varepsilon$  using the data compiled in Table 9.1. You can use the thermal conductivity value  $k = 2.37 \frac{W}{cm \cdot C}$  for aluminum. By sampling from the densities, construct credible and prediction intervals for  $x \in [10, 66]$  and plot with the complete data set from Table 3.2. Does the correct percentage of data points outside the calibration domain lie within the prediction interval?

|           |       |       |       |       |       |       |       |       |       |
|-----------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| $x$ (cm)  | 22    | 26    | 30    | 34    | 38    | 42    | 46    | 50    | 54    |
| Temp (°C) | 57.96 | 50.90 | 44.84 | 39.75 | 36.16 | 33.31 | 31.15 | 29.28 | 27.88 |

**Table 9.1.** Steady state temperatures measured at locations  $x$  for an aluminum rod.

**Exercise 9.7.** We revisit the HIV model

$$\begin{aligned}\dot{T}_1 &= \lambda_1 - d_1 T_1 - (1 - \varepsilon) k_1 V T_1, \\ \dot{T}_2 &= \lambda_2 - d_2 T_2 - (1 - f\varepsilon) k_2 V T_2, \\ \dot{T}_1^* &= (1 - \varepsilon) k_1 V T_1 - \delta T_1^* - m_1 E T_1^*, \\ \dot{T}_2^* &= (1 - f\varepsilon) k_2 V T_2 - \delta T_2^* - m_2 E T_2^*, \\ \dot{V} &= N_T \delta (T_1^* + T_2^*) - cV - [(1 - \varepsilon) \rho_1 k_1 T_1 + (1 - f\varepsilon) \rho_2 k_2 T_2]V, \\ \dot{E} &= \lambda_E + \frac{b_E (T_1^* + T_2^*)}{T_1^* + T_2^* + K_b} E - \frac{d_E (T_1^* + T_2^*)}{T_1^* + T_2^* + K_d} E - \delta_E E\end{aligned}$$

illustrated in Examples 3.3, 8.12, and 9.14. Use your DRAM code from Exercise 8.6 to construct 95% credible and prediction intervals for each of the states. Does the correct percentage of data lie in the prediction interval?

## Chapter 10

# Stochastic Spectral Methods

The objective of stochastic Galerkin, collocation, and discrete projection methods can be viewed from two perspectives. In one sense, they provide techniques for constructing surrogate models of the type discussed in Chapter 13 for high-dimensional problems—in the parameter space—based on low-order expansions. This interpretation lends itself to Bayesian model calibration, sensitivity analysis, design, and control implementation. From the perspective of uncertainty propagation, in which one is sampling from the posterior density, it can be useful to view them as techniques to significantly reduce the number of deterministic model solutions required to construct moments associated with a QoI. This latter goal is achieved by employing spectral expansions that exploit the smoothness, often exhibited by high-dimensional parameter spaces, to construct sampling or solution techniques with convergence rates that are significantly faster than the  $\frac{1}{\sqrt{M}}$  Monte Carlo rates for moderate parameter dimensions.

We note that all of the techniques discussed in this chapter require either mutually independent parameters or representation of the joint posterior density  $\rho_Q(q)$  for implementation. As detailed in Section 5.2, parameters in typical applications are often correlated and representations for  $\rho_Q(q)$  are usually unavailable or not feasible. If marginal distributions and a correlation matrix are provided, one can employ a Nataf transformation in combination with a Cholesky decomposition to construct a formulation with mutually independent Gaussian random variables. If this information is not available, sampling from indices of parameter chains, in the manner detailed in Remark 9.4 and illustrated in Example 9.14, may constitute the only option for constructing prediction intervals.

### 10.1 Spectral Representation of Random Processes

The goal when constructing spectral expansions is to represent random processes in a manner that exploits the smoothness often exhibited by high-dimensional parameter spaces and facilitates the construction of moments for QoI. Depending on the basis choice, these expansions are often termed polynomial chaos (PC) or gen-

eralized polynomial chaos (gPC) expansions. The term dates back to Wiener, who used it in the context of Hermite expansions employed for constructing a physical theory of chaos [262], and it is well established in the literature. In the context of uncertainty quantification, however, it is a misnomer since systems under investigation are not typically chaotic. To the degree possible, we avoid it and instead use the more descriptive terminology *spectral expansions*.

### 10.1.1 Polynomial Expansions

To motivate the expansions used to represent stochastic processes, we consider sequences  $\{Q_k(\omega)\}_{k=1}^{\infty}$  of random variables defined on the sample space  $\Omega$  of the probability space  $(\Omega, \mathcal{F}, P)$ . We let  $\mathbb{P}_k$  denote the space of polynomials with argument  $Q_i$  having degree less than or equal to  $k$  and let  $\widehat{\mathbb{P}}_k$  be the set of polynomials in  $\mathbb{P}_k$  that are orthogonal to  $\mathbb{P}_{k-1}$ . The space  $\widehat{\mathbb{P}}_k$  is sometimes termed the Wiener PC of order  $k$  when considering Gaussian variables. We note that since the random variables are functions  $Q_k : \Omega \rightarrow \mathbb{R}$ , the polynomials in  $\widehat{\mathbb{P}}_k$  can be interpreted as functionals.

As detailed in [57], second-order (finite variance) random variables  $u$  can be represented as an infinite expansion

$$\begin{aligned} u(\omega) = & u_0 \widehat{P}_0 + \sum_{i_1=1}^{\infty} u_{i_1} \widehat{P}_1(Q_{i_1}) + \sum_{i_1=1}^{\infty} \sum_{i_2=1}^{\infty} u_{i_1, i_2} \widehat{P}_2(Q_{i_1}, Q_{i_2}) \\ & + \sum_{i_1=1}^{\infty} \sum_{i_2=1}^{\infty} \sum_{i_3=1}^{\infty} u_{i_1, i_2, i_3} \widehat{P}_3(Q_{i_1}, Q_{i_2}, Q_{i_3}) + \dots \end{aligned} \quad (10.1)$$

of increasing interaction terms  $\widehat{P}_n(Q_{i_1}, \dots, Q_{i_n})$ , where  $u_{i_1}, u_{i_1, i_2}, \dots$  are real coefficients. This can be written more compactly as

$$u(Q) = \sum_{k=0}^{\infty} u_k \Psi_k(Q_1, Q_2, \dots), \quad (10.2)$$

where there is a one-to-one correspondence between the coefficients and polynomials in (10.1) and (10.2).

In practice, one typically considers a finite set of  $p$  random variables  $Q_1, \dots, Q_p$  with a limited number  $n$  of interaction terms. For example, truncation of (10.1) at second-order interactions yields

$$u(\omega) = u_0 \widehat{P}_0 + \sum_{i_1=1}^p u_{i_1} \widehat{P}_1(Q_{i_1}) + \sum_{i_1=1}^p \sum_{i_2=1}^p u_{i_1, i_2} \widehat{P}_2(Q_{i_1}, Q_{i_2}), \quad (10.3)$$

whereas truncation of (10.2) to  $K$  terms yields the expression

$$u^K(\omega) = \sum_{k=0}^K u_k \Psi_k(Q_1, Q_2, \dots, Q_p), \quad (10.4)$$

where  $K + 1 = \frac{(n+p)!}{n!p!}$ .

We now consider the representation of a random process  $u(t, x, \omega)$  that is a function of a random vector  $Q(\omega) = [Q_1(\omega), \dots, Q_p(\omega)] : \Omega \rightarrow \mathbb{R}^p$ . As in Section 5.1, we exploit the equivalence between realizations  $\omega \in \Omega$  and values of the random vector  $Q(\omega) \in \Gamma \subset \mathbb{R}^p$ , where  $\Gamma_i = Q_i(\omega)$  and  $\Gamma = \prod_{i=1}^p \Gamma_i$ , to pose the problem in the image probability space  $(\Gamma, \mathcal{B}(\Gamma), \rho_Q(q)dq)$  where  $\rho_Q(q)$  is the joint density associated with  $Q$ . For  $(t, x) \in [0, T] \times \mathcal{D}$ , we separate spatio-temporal and random dependencies to obtain the finite-dimensional representation

$$u^K(t, x, Q) = \sum_{k=0}^K u_k(t, x) \Psi_k(Q), \quad (10.5)$$

where  $u_k(t, x)$  are deterministic coefficients and  $\Psi_k(Q)$  are orthogonal polynomials that form a basis for the random component of the solution. We refer the reader to [148, 266] for details regarding the strong and weak nature of approximations and simply note that at some level, they often invoke the Cameron–Martin theorem.

To construct  $u^K$ , one must specify appropriate basis functions  $\Psi_k(Q)$  and constraints to determine the coefficients. We illustrate in Section 10.2 that stochastic Galerkin and collocation techniques can be used to specify  $u_k(t, x)$  for fairly general classes of differential equations. We focus on classes of globally defined orthogonal polynomials and refer the reader to [19] for details regarding piecewise polynomial bases.

### 10.1.2 Basis Construction for a Single Random Variable

For a single, continuous, random variable  $Q$ , we take  $\psi_k(Q)$  to be 1-D global polynomials that are orthogonal with respect to the density  $\rho_Q(q)$  and indexed so that  $\psi_0 = 1$ . It follows that

$$\mathbb{E}[\psi_0(Q)] = 1 \quad (10.6)$$

and

$$\begin{aligned} \mathbb{E}[\psi_i(Q)\psi_j(Q)] &= \int_{\Gamma} \psi_i(q)\psi_j(q)\rho_Q(q)dq \\ &= \langle \psi_i, \psi_j \rangle_{\rho} \\ &= \delta_{ij}\gamma_i, \end{aligned} \quad (10.7)$$

where  $\langle \cdot, \cdot \rangle_{\rho}$  denotes the  $L^2$  inner product on the interval  $\Gamma$  with the weight  $\rho_Q(q)$ . The normalization factor is

$$\gamma_i = \mathbb{E}[\psi_i^2(Q)] = \langle \psi_i, \psi_i \rangle_{\rho}. \quad (10.8)$$

We note that some texts employ the bracket notation  $\langle \psi_i \rangle^2$  to specify the expectation  $\mathbb{E}[\psi_i^2(Q)]$  in the image probability space  $(\Gamma, \mathcal{B}(\Gamma), \rho_Q(q)dq)$ .

#### Statistics of Orthogonal Polynomial Expansions

The orthogonal polynomial expansion (10.5) has the advantage that statistical and sensitivity analysis can be performed either analytically or with minimal

computational effort, once the coefficients have been determined. We illustrate the characterization of the mean and variance and refer the reader to [68, 241] for details regarding the use of this expansion to compute the global Sobol sensitivity indices discussed in Chapter 15.

**Property 10.1 (Mean and Variance).** For fixed values of  $t$  and  $x$ , the mean of  $u^K$  is given by

$$\begin{aligned}\mathbb{E}[u^K(t, x, Q)] &= \mathbb{E}\left[\sum_{k=0}^K u_k(t, x)\psi_k(Q)\right] \\ &= u_0(t, x)\mathbb{E}[\psi_0(Q)] + \sum_{k=1}^K u_k(t, x)\mathbb{E}[\psi_k(Q)] \\ &= u_0(t, x)\end{aligned}\tag{10.9}$$

since  $\mathbb{E}[\psi_0(Q)] = 1$  and  $\mathbb{E}[\psi_k(Q)] = 0$  for  $k > 1$ . Similarly, the variance is

$$\begin{aligned}\text{var}[u^K(t, x, Q)] &= \mathbb{E}\left[\left(u^K(t, x, Q) - \mathbb{E}[u^K(t, x, Q)]\right)^2\right] \\ &= \mathbb{E}\left[\left(\sum_{k=0}^K u_k(t, x)\psi_k(Q) - u_0(t, x)\right)^2\right] \\ &= \mathbb{E}\left[\left(\sum_{k=1}^K u_k(t, x)\psi_k(Q)\right)^2\right] \\ &= \sum_{k=1}^K u_k^2(t, x)\gamma_k,\end{aligned}\tag{10.10}$$

where the last equality results from the orthogonality of the basis functions and the definition (10.7) for  $\gamma_k$ . Higher-order moments and correlation functions can be computed in a similar manner.

### Polynomials for Normal and Uniform Densities

We now illustrate two choices for the orthogonal polynomials that are commonly used in spectral representations.

**Example 10.2 (Hermite Polynomials for  $Q \sim N(0, 1)$ ).** The pdf is

$$\rho_Q(q) = \frac{1}{\sqrt{2\pi}}e^{-q^2/2},\tag{10.11}$$

which is defined on  $\Gamma = \mathbb{R}$ . The Hermite polynomials

$$\begin{aligned}H_0(Q) &= 1, & H_1(Q) &= Q, & H_2(Q) &= Q^2 - 1, \\ H_3(Q) &= Q^3 - 3Q, & H_4(Q) &= Q^4 - 6Q^2 + 3\end{aligned}\tag{10.12}$$

constitute the natural choice for  $\psi_k(Q)$  since they are orthogonal over the real line with respect to this density. This was the basis choice for the original PC formulations.

The normalization constants can be expressed as

$$\gamma_i = \int_{\mathbb{R}} \psi_i^2(q) \rho_Q(q) dq = i!. \quad (10.13)$$

We note that the polynomials (10.12) are sometimes referred to as “probabilist” Hermite functions to differentiate them from “physicist” Hermite polynomials that are orthogonal with respect to

$$\rho_Q(q) = e^{-q^2}. \quad (10.14)$$

Unfortunately, essentially all tables that specify Gauss–Hermite quadrature points employ the weight (10.14) rather than (10.11). This necessitates that commonly tabulated quadrature points and weights be scaled for uncertainty propagation. To illustrate, we let  $w^r, x^r$  denote the quadrature weights and nodes corresponding to the density (10.14). It follows that

$$\begin{aligned} \int_{\mathbb{R}} g(q) \frac{1}{\sqrt{2\pi}} e^{-q^2/2} dq &= \frac{1}{\sqrt{\pi}} \int_{\mathbb{R}} g(\sqrt{2} \cdot x) e^{-x^2} dx \\ &\approx \sum_{r=1}^R g(\sqrt{2} \cdot x^r) \frac{w^r}{\sqrt{\pi}} \\ &= \sum_{r=1}^R g(q^r) \hat{w}^r. \end{aligned} \quad (10.15)$$

Hence the weights and nodes corresponding to the density (10.11) are

$$\hat{w}^r = \frac{w^r}{\sqrt{\pi}}, \quad q^r = \sqrt{2} \cdot x^r. \quad (10.16)$$

**Example 10.3 (Legendre Polynomials for  $Q \sim \mathcal{U}(-1, 1)$ ).** The Legendre polynomials, of which the first five are

$$\begin{aligned} P_0(Q) &= 1, \quad P_1(Q) = Q, \quad P_2(Q) = \frac{3}{2}Q^2 - \frac{1}{2}, \\ P_3(Q) &= \frac{5}{2}Q^3 - \frac{3}{2}Q, \quad P_4(Q) = \frac{35}{8}Q^4 - \frac{15}{4}Q^2 + \frac{3}{8}, \end{aligned} \quad (10.17)$$

are orthogonal on the interval  $\Gamma = [-1, 1]$  with respect to the density

$$\rho_Q(q) = \frac{1}{2}. \quad (10.18)$$

Hence they form a suitable basis for uniformly distributed random variables on  $[-1, 1]$ .

### Representation of Normal and Uniform Random Variables

The next two examples illustrate the construction of the coefficients  $u_k$  in (10.5) and representation of normal and uniform densities using Hermite and Legendre polynomials.

**Example 10.4 ( $u \sim N(\mu, \sigma^2)$ ).** The random variable  $u$  can be expressed as

$$u = \mu + \sigma Q, \quad (10.19)$$

where  $Q \sim N(0, 1)$ —we all know this from the MATLAB documentation for the normal random number function `randn.m`. Hence

$$u^K(Q) = \sum_{k=0}^K u_k \psi_k(Q) \quad (10.20)$$

provides an exact representation for  $u$  when  $\psi_k(Q)$  are Hermite polynomials and

$$u_0 = \mu, \quad u_1 = \sigma, \quad u_k = 0, \quad k > 1.$$

**Example 10.5 ( $u \sim \mathcal{U}(a, b)$ ).** Consider  $u \sim \mathcal{U}(a, b)$  with mean and variance

$$\mu = \frac{a+b}{2}, \quad \sigma^2 = \frac{(b-a)^2}{12}.$$

It is established in Exercise 4.6 that

$$u = \mu + \sqrt{3}\sigma Q,$$

where  $Q \sim \mathcal{U}(-1, 1)$ . Hence (10.20) also provides an exact representation of  $u$  where  $\psi_k(Q)$  are now Legendre polynomials defined on  $[-1, 1]$ .

#### 10.1.3 Multiple Random Variables

The fundamental concepts regarding the representation of random vectors and random processes that are functions of more than one independent random variable are analogous to the univariate case. The assumption of mutually independent random variables implies that the expectation of a product of random variables is the product of their expectations. This motivates multidimensional basis functions to be constructed as tensor products of the previously discussed univariate polynomials. These formulations are simplified through use of multi-indices.

**Definition 10.6 ( $p$ -Dimensional Multi-Index).** A  $p$ -tuple

$$\mathbf{k}' = (k_1, \dots, k_p) \in \mathbb{N}_0^p$$

of nonnegative integers is termed a  $p$ -dimensional multi-index with magnitude  $|\mathbf{k}'| = k_1 + k_2 + \dots + k_p$  and satisfying the ordering  $\mathbf{j}' \leq \mathbf{k}' \Leftrightarrow j_i \leq k_i$  for  $i = 1, \dots, p$ .

We now consider a random vector  $Q = [Q_1, \dots, Q_p]$  of mutually independent random variables and let  $\{\psi_k(Q_i)\}_{k=0}^K$  denote the univariate basis functions up to degree  $K$  in the variable  $Q_i$ . The  $p$ -variate basis functions of total degree less than or equal to  $K$  are defined to be

$$\Psi_{\mathbf{i}'}(Q) = \psi_{i_1}(Q_1), \dots, \psi_{i_p}(Q_p)$$

for  $0 \leq |\mathbf{i}'| \leq K$ . The multivariate basis functions thus satisfy the orthogonality conditions

$$\begin{aligned}\mathbb{E}[\Psi_{\mathbf{i}'}(Q)\Psi_{\mathbf{j}'}(Q)] &= \int_{\Gamma} \Psi_{\mathbf{i}'}(q)\Psi_{\mathbf{j}'}(q)\rho_Q(q)dq \\ &= \langle \Psi_{\mathbf{i}'}, \Psi_{\mathbf{j}'} \rangle_{\rho} \\ &= \delta_{\mathbf{i}'\mathbf{j}'},\end{aligned}$$

where  $\Gamma = \prod_{i=1}^p \Gamma_k$ . Similarly,  $\rho_Q(q)$  defined in (5.2) is the product of the densities associated with each random variable, whereas  $\gamma_{\mathbf{i}'} = \mathbb{E}[\Psi_{\mathbf{i}'}^2] = \gamma_{i_1} \cdots \gamma_{i_p}$  is the product of the univariate normalization constants defined in (10.8). Finally,  $\delta_{\mathbf{i}'\mathbf{j}'} = \delta_{i_1 j_1} \cdots \delta_{i_p j_p}$  denotes the extension of the Kronecker delta to  $p$  variables.

For a random process  $u(t, x, Q) : [0, T] \times \mathcal{D} \times \Gamma \rightarrow \mathbb{R}$ , we employ the expansion

$$u^K(t, x, Q) = \sum_{|\mathbf{k}'|=0}^K u_{\mathbf{k}'}(t, x)\Psi_{\mathbf{k}'}(Q),$$

which is the projection of  $u$  onto the space  $\mathbb{P}_K(Q)$  of all polynomials of  $Q \in \mathbb{R}^p$  of degree up to  $K$ .

The multi-index notation provides an elegant mechanism to specify the expansion based on multiple random variables, but it can be cumbersome to manipulate when implementing the method. Alternatively, one can employ a single index  $k$  chosen according to various conventions. A common choice is the ordering, demonstrated for  $p = 3$  in Table 10.1, that ranks multi-indices in order of increasing  $|\mathbf{k}'|$  and specifying that the first nonzero component in  $\mathbf{k}' - \mathbf{j}'$  be positive. As detailed in [266], other orderings are advantageous in certain settings. For example, componentwise orderings are employed in the HDMR expansions discussed in Section 13.5.

The expansion based on a single index is

$$u^K(t, x, Q) = \sum_{k=0}^K u_k(t, x)\Psi_k(Q).$$

To achieve order  $n$  polynomials, it follows that  $K = \frac{(n+p)!}{n!p!} - 1$ , as illustrated for  $p = 3$  and  $n = 2$  in Table 10.1.

The orthogonality of the basis functions can be exploited to obtain the representation

$$u_k(t, x) = \frac{1}{\gamma_k} \mathbb{E}[u(t, x, Q)\Psi_k(Q)] \quad (10.21)$$

for the deterministic coefficients. Whereas (10.21) is an optimal projection in an  $L^2$  sense, it is not generally useful from a computational perspective since  $u$  is unknown.

| $k$ | $ \mathbf{k}' $ | Multi-Index | Polynomial                          |
|-----|-----------------|-------------|-------------------------------------|
| 0   | 0               | (0, 0, 0)   | $\psi_0(Q_1)\psi_0(Q_2)\psi_0(Q_3)$ |
| 1   | 1               | (1, 0, 0)   | $\psi_1(Q_1)\psi_0(Q_2)\psi_0(Q_3)$ |
| 2   |                 | (0, 1, 0)   | $\psi_0(Q_1)\psi_1(Q_2)\psi_0(Q_3)$ |
| 3   |                 | (0, 0, 1)   | $\psi_0(Q_1)\psi_0(Q_2)\psi_1(Q_3)$ |
| 4   | 2               | (2, 0, 0)   | $\psi_2(Q_1)\psi_0(Q_2)\psi_0(Q_3)$ |
| 5   |                 | (1, 1, 0)   | $\psi_1(Q_1)\psi_1(Q_2)\psi_0(Q_3)$ |
| 6   |                 | (1, 0, 1)   | $\psi_1(Q_1)\psi_0(Q_2)\psi_1(Q_3)$ |
| 7   |                 | (0, 2, 0)   | $\psi_0(Q_1)\psi_2(Q_2)\psi_0(Q_3)$ |
| 8   |                 | (0, 1, 1)   | $\psi_0(Q_1)\psi_1(Q_2)\psi_1(Q_3)$ |
| 9   |                 | (0, 0, 2)   | $\psi_0(Q_1)\psi_0(Q_2)\psi_2(Q_3)$ |

**Table 10.1.** Single index, multi-index, and tensored polynomials for  $p = 3$ .

The Galerkin, collocation, and discrete projection methods provide constraints used to develop numerical algorithms for computing  $u_k(t, x)$ .

## 10.2 Galerkin, Collocation, and Discrete Projection Frameworks

The stochastic Galerkin, collocation, and discrete projection frameworks are analogous to their deterministic counterparts. In the Galerkin framework, one projects weighted residuals onto a finite-dimensional subspace spanned by appropriate basis functions to provide the constraints required to solve for the deterministic coefficients. This projection requires the construction of expectations or inner products not typically employed in deterministic codes, and hence it is *intrusive* in the sense that existing codes must be modified. For collocation methods, the constraints are provided by approximating the solution of the governing equations at a discrete set of points termed collocation points. For stochastic collocation, these points are typically values in the random variable space used to represent parameters or inputs. Because existing software and codes can be used to determine approximate solutions at the collocation points, the methods are *nonintrusive*, which is one of their primary advantages. Discrete projection relies on the use of quadrature techniques to approximate (10.21), which requires the solution of  $u(t, x, q^r)$  of the governing equations at the quadrature points  $q^r$ .

Because Galerkin and collocation methods respectively rely on projection and interpolation, their convergence analysis differs. Hence from the perspective of numerical analysis, they are typically addressed as separate topics. From the perspective of implementation, however, collocation can be interpreted as a special case of discretized Galerkin that results when basis functions satisfy the delta property  $\Psi_k(q^r) = \delta_{kr}$ . Furthermore, discrete projection methods can also be placed in the Galerkin framework with appropriate choices for basis functions and quadrature

rules. Hence we introduce collocation and discrete projection within a Galerkin framework.

As noted in Chapter 9, spectral expansions can be advantageous for problems, such as nonlinear PDEs in one or more space dimensions, that are computationally intensive. However, the complexity of such frameworks obscures the initial presentation of the stochastic frameworks. Hence to illustrate the framework, we initially develop it in Section 10.2.1 for a nonlinear scalar initial value problem while noting that other methods may be equally successful for this problem due to its simplicity. We illustrate the Galerkin, collocation, and discrete projection methods for boundary value problems and stationary PDEs in Section 10.2.2 and evolutionary PDEs in Section 10.2.3. We summarize attributes of the three methods in Section 10.2.4. We provide examples of the Galerkin method in Section 10.3 and detail sparse quadrature and collocation techniques in Chapter 11.

### 10.2.1 Scalar Initial Value Problem

We first consider the scalar ODE

$$\begin{aligned} \frac{du}{dt} &= f(t, Q, u), \quad t > 0, \\ u(0, Q) &= u_0. \end{aligned} \tag{10.22}$$

This setting allows us to illustrate attributes of the methods without the added complexity of infinite-dimensional states due to spatial dependence. We assume that  $Q = [Q_1, \dots, Q_p]$  are mutually independent random variables with range  $\Gamma \subset \mathbb{R}^p$  and joint density  $\rho_Q(q)$ , as detailed in Section 5.1.<sup>7</sup> The  $L^2$  inner product with respect to this density is denoted by  $\langle \cdot, \cdot \rangle_\rho$ . We take the quantity of interest (QoI) to be the expected state value

$$y(t) = \int_{\Gamma} u(t, q) \rho_Q(q) dq.$$

We consider solutions in the space  $L^2(0, T; Z)$  where  $Z$  is chosen to reflect the regularity of the random component of the solution. If we let  $\rho_{Q_i}(q_i)$  denote the density of the  $i^{th}$  component  $Q_i$  of  $Q$ , it is natural to consider the space  $L^2_{\rho_i}(\Gamma_i)$  of functions with finite norm

$$\|g\|_2 = \left( \int_{\Gamma_i} |g(q_i)|^2 \rho_{Q_i}(q_i) dq_i \right)^{1/2},$$

which ensures that second moments are well defined. If higher-order moments are required, one can alternatively employ spaces  $L^m_{\rho_i}(\Gamma)$ . For mutually independent random variables with finite second moments, we take

$$Z = L^2(\Gamma) = L^2_{\rho_1}(\Gamma_1) \otimes \cdots \otimes L^2_{\rho_p}(\Gamma_p). \tag{10.23}$$

---

<sup>7</sup>As discussed in Section 5.2, parameters in a number of applications are correlated, which violates the assumption of mutual independence. It is noted in Section 10.2.4 that this can constitute a limitation of stochastic spectral methods if Nataf or Rosenblatt transformations are infeasible.

To approximate  $u(t, Q)$ , we let  $\{\Psi_k\}_{k=1}^K$  be a basis for the random space and take  $Z^K = \text{span}\{\Psi_k\}_{k=1}^K \subset Z$ . The projection of  $u(t, Q)$  onto  $Z^K$  yields the representation

$$u^K(t, Q) = \sum_{k=0}^K u_k(t) \Psi_k(Q). \quad (10.24)$$

For orthogonal basis functions defined on  $\Gamma$  with density  $\rho_Q(q)$ , the generalized Fourier coefficients satisfy

$$u_k(t) = \frac{1}{\gamma_k} \int_{\Gamma} u(t, q) \Psi_k(q) \rho_Q(q) dq, \quad (10.25)$$

where  $\gamma_k = \langle \Psi_k, \Psi_k \rangle_{\rho}$ . Since we typically cannot solve (10.25) explicitly, the specification of constraints required to approximate the coefficients  $u_k(t)$  is one factor that defines solution methods. The specification of appropriate basis functions is the second issue that delineates various solution techniques.

### Stochastic Galerkin Method

As in the deterministic problem, the strategy for stochastic Galerkin methods is to project the weighted residual onto a finite-dimensional space spanned by appropriate basis functions. Here we employ polynomials that are orthogonal with respect to the density  $\rho_Q(q)$ . We thus seek approximate solutions  $u^K(t, Q)$  that satisfy

$$\begin{aligned} 0 &= \left\langle \frac{du^K}{dt} - f, \Psi_i \right\rangle_{\rho} \\ &= \int_{\Gamma} \left[ \sum_{k=0}^K \frac{du_k}{dt}(t) \Psi_k(q) - f \left( t, q, \sum_{k=0}^K u_k(t) \Psi_k(q) \right) \right] \Psi_i(q) \rho_Q(q) dq \end{aligned} \quad (10.26)$$

for all  $i \leq K$ . This is often termed the *weak stochastic model formulation*, and it is equivalent to specifying that

$$\mathbb{E} \left[ \frac{du^K(t, Q)}{dt} \Psi_i(Q) \right] = \mathbb{E} [f(t, Q, u^K) \Psi_i(Q)]. \quad (10.27)$$

Initial conditions for (10.26) or (10.27) are constructed by projecting the original initial conditions onto the space of polynomials  $Z^K$  in the manner detailed in Section 13.2.

Approximation of the inner product using a quadrature rule with points  $q^r$  and weights  $w^r$  yields

$$\sum_{r=1}^R \Psi_i(q^r) \rho_Q(q^r) w^r \left[ \sum_{k=0}^K \frac{du_k}{dt}(t) \Psi_k(q^r) - f \left( t, q^r, \sum_{k=0}^K u_k(t) \Psi_k(q^r) \right) \right] = 0, \quad (10.28)$$

which holds for  $i = 0, \dots, K$ .

For low-dimensional parameter spaces, tensored Galerkin quadrature techniques provide sufficient accuracy and efficiency. However, tensored quadrature techniques are not applicable for even moderate dimensions due to the exponential growth in the required number of points. For low to moderate dimensions, this is addressed by the sparse grid techniques discussed in Section 11.1.

### Collocation

The basic strategy for collocation is quite simple; (i) using either deterministic or stochastic methods, generate  $M$  samples  $\{q^m\}_{m=1}^M$  from the parameter space, which constitute the collocation points, and (ii) enforce

$$u(t, q^m) = u^K(t, q^m) \quad (10.29)$$

to provide the constraints necessary to solve for  $u_k(t)$  at each time. For general basis functions  $\Psi_k$ , the constraint (10.29) yields the Vandemonde matrix system

$$\begin{bmatrix} \Psi_0(q^1) & \cdots & \Psi_K(q^1) \\ \vdots & & \vdots \\ \Psi_0(q^M) & \cdots & \Psi_K(q^M) \end{bmatrix} \begin{bmatrix} u_0(t) \\ \vdots \\ u_K(t) \end{bmatrix} = \begin{bmatrix} u(t, q^1) \\ \vdots \\ u(t, q^M) \end{bmatrix}. \quad (10.30)$$

To prevent the system from being underdetermined, one would typically require  $M \geq K + 1$ . For overdetermined systems,  $M > K + 1$ , (10.30) can be solved in the least squares sense using, for example, the MATLAB command `pinv.m`.

Although very straightforward to formulate, two issues can make (10.30) difficult or impossible to implement for large parameter dimensions. The first concerns the choice of collocation points  $q^m$ . As will be illustrated in Section 11.2, uniform meshes can produce spurious oscillations and yield highly ill-conditioned collocation matrices. This can be avoided in one dimension by using nonuniformly spaced points, but care must be exercised when extending these meshes to multiple dimensions.

The second issue regards the choice of basis functions. General basis functions yield a  $M \times (K + 1)$  dense matrix where  $M$  will be extremely large for high-dimensional parameter spaces. This is avoided if one employs Lagrange polynomials that satisfy

$$L_k(q^m) = \delta_{km} \quad (10.31)$$

as basis functions  $\Psi_k(q)$ . For  $M = K + 1$  collocation points, the collocation matrix is now a  $(K + 1) \times (K + 1)$  identity matrix and

$$u_m(t) = u(t, q^m) \quad (10.32)$$

for  $m = 1, \dots, M$ .

This collocation relation can also be directly obtained from (10.28) if one employs  $\Psi_k = L_k$  and employs the quadrature points  $q^r$  as collocation points  $q^m$ . Specifically, this yields

$$\frac{du_m}{dt}(t) = f(t, q^m, u_m), \quad m = 1, \dots, M, \quad (10.33)$$

which is equivalent to specifying  $u_m(t)$  by (10.32).

**Remark 10.7.** The Lagrange basis is orthogonal only for the set of collocation points for which it was defined. Hence the relation (10.25), which relies on orthogonality with respect to  $\rho_Q(q)$ , does not generally hold.

### Discrete Projection

The strategy for discrete projection, which is also termed pseudospectral, is to approximate (10.25) by

$$u_k(t) = \frac{1}{\gamma_k} \sum_{r=1}^R u(t, q^r) \Psi_k(q^r) \rho_Q(q^r) w^r \quad (10.34)$$

to provide the time-dependent deterministic coefficients. Since  $u(t, q^r)$  is the solution of (10.22) at the  $r$  quadrature points, the computational effort is essentially equivalent to that of the collocation method.

#### 10.2.2 Boundary Value Problems and Elliptic PDEs

We now formulate the stochastic Galerkin method for boundary value problems and PDEs that exhibit spatial dependence and hence have infinite-dimensional states. To simplify the discussion, we initially focus on steady problems. The strong formulation of the deterministic model is specified as

$$\begin{aligned} \mathcal{N}(u, Q) &= F(Q) \quad , \quad x \in \mathcal{D}, \\ B(u, Q) &= G(Q) \quad , \quad x \in \partial\mathcal{D}, \end{aligned} \quad (10.35)$$

where  $u(x, Q)$  is the state,  $\mathcal{N}$  is a potentially nonlinear differential operator,  $F(Q)$  is a source term, and  $B(u, Q)$  and  $G(Q)$  are boundary operators. The spatial domain  $\mathcal{D}$  is a subset of  $\mathbb{R}^1$ ,  $\mathbb{R}^2$ , or  $\mathbb{R}^3$ . In general,  $\mathcal{N}$ ,  $F$ ,  $B$ , and  $G$  can also depend on the spatial variable  $x$ , but we suppress this dependence to simplify notation.

As a QoI, we consider the expected state value

$$y(x) = \int_{\Gamma} u(x, q) \rho_Q(q) dq \quad (10.36)$$

at points  $x \in \mathcal{D}$ . More complex QoI can be evaluated in a similar manner.

To construct a weak formulation of the deterministic problem, we let  $V$  denote a space of test functions that satisfy the essential boundary conditions. The deterministic weak formulation can then be posed as the problem of finding  $u \in V$ , which satisfies

$$\int_{\mathcal{D}} N(u, Q) S(v) dx = \int_{\mathcal{D}} F(Q) v dx \quad (10.37)$$

for all  $v \in V$ . The potentially nonlinear differential operator  $N$  and linear operator  $S$  are constructed from  $\mathcal{N}$  using integration by parts.

**Example 10.8.** Consider the heat equation

$$\begin{aligned} \alpha \frac{d^2 u}{dx^2} &= -f(x) \quad , \quad -1 < x < 1, \\ u(-1) &= u(1) = 0, \end{aligned} \tag{10.38}$$

where  $Q = \alpha$ . Here

$$\mathcal{N}(u, Q) = \alpha \frac{d^2 u}{dx^2} , \quad F(Q) = -f(x) , \quad B(u, Q) = u , \quad G(Q) = 0 \tag{10.39}$$

and the domain is  $\mathcal{D} = (-1, 1)$ . An appropriate space of test functions is  $V = H_0^1(\mathcal{D})$ , and the weak formulation is

$$\int_{\mathcal{D}} \alpha \frac{du}{dx} \frac{dv}{dx} dx = \int_{\mathcal{D}} f(x)v(x)dx, \tag{10.40}$$

which must hold for all  $v \in V$ . Hence the operators  $N$  and  $S$  are

$$N(u, Q) = \alpha \frac{du}{dx} , \quad S(v) = \frac{dv}{dx}. \tag{10.41}$$

### Stochastic Weak Formulation

For the random differential equation, we seek solutions  $u(x, Q) \in V \otimes Z$ , where  $Z$  is defined in (10.23) and  $V$  is typically a Sobolev space. The weak stochastic model formulation can be posed as follows: find  $u \in V \otimes Z$ , which satisfies

$$\int_{\Gamma} \int_{\mathcal{D}} N(u, q) S(v(x)) z(q) \rho_Q(q) dx dq = \int_{\Gamma} \int_{\mathcal{D}} F(q) v(x) z(q) \rho_Q(q) dx dq$$

for all test functions  $v \in V$ ,  $z \in Z$ .

To approximate the solution  $u(x, Q)$ , we let  $\{\phi_j(x)\}_{j=1}^J$  and  $\{\Psi_k(Q)\}_{k=0}^K$  be bases for the spatial and random spaces and take

$$V^J = \text{span}\{\phi_j\} \subset V \quad , \quad Z^K = \text{span}\{\Psi_k\} \subset Z.$$

Typical choices for  $\phi_j$  are splines, finite elements, or spectral functions, whereas we employ the spectral polynomials discussed in Section 10.1 for  $\Psi_k$ . Due to the product nature of the domain  $\mathcal{D} \times \Gamma$  and space  $V \otimes Z$ , we employ approximate solutions of the form

$$\begin{aligned} u^K(x, Q) &= \sum_{k=0}^K u_k(x) \Psi_k(Q) \\ &= \sum_{k=0}^K \sum_{j=1}^J u_{jk} \phi_j(x) \Psi_k(Q). \end{aligned}$$

### Stochastic Galerkin Method

In the Galerkin method, we project the residuals for the discrete problem onto the space of test functions to obtain

$$\begin{aligned} & \int_{\Gamma} \int_{\mathcal{D}} N \left( \sum_{k=0}^K \sum_{j=1}^J u_{jk} \phi_j(x) \Psi_k(q), q \right) S(\phi_\ell(x)) \Psi_i(q) \rho_Q(q) dx dq \\ &= \int_{\Gamma} \int_{\mathcal{D}} F(q) \phi_\ell(x) \Psi_i(q) \rho_Q(q) dx dq, \end{aligned} \quad (10.42)$$

which holds for  $\ell = 1, \dots, J$  and  $i = 0, \dots, K$ .

To determine the coefficients  $u_{jk}$ , we must approximate the integrals. In the  $p$ -dimensional parameter space, we employ  $R$  quadrature points  $q^r$  and weights  $w^r$  to obtain

$$\begin{aligned} & \sum_{r=1}^R \Psi_i(q^r) \rho_Q(q^r) w^r \int_{\mathcal{D}} N \left( \sum_{k=0}^K \sum_{j=1}^J u_{jk} \phi_j(x) \Psi_k(q^r), q^r \right) S(\phi_\ell(x)) dx \\ &= \sum_{r=1}^R \Psi_i(q^r) \rho_Q(q^r) w^r \int_{\mathcal{D}} F(q^r) \phi_\ell(x) dx \end{aligned} \quad (10.43)$$

for  $\ell = 1, \dots, J$  and  $i = 0, \dots, K$ . Standard Gaussian quadrature techniques can be employed to approximate the spatial integrals, as detailed in [95, 118, 225]. This yields a  $J(K+1) \times J(K+1)$  system that is fully coupled in the physical and parameter spaces.

We must also approximate the integral to evaluate the QoI (10.36). Whereas different quadrature rules can be employed, we simplify the discussion by employing the same quadrature points and weights. This yields

$$y(x) = \sum_{r=1}^R w^r \rho_Q(q^r) \sum_{k=0}^K \sum_{j=1}^J u_{jk} \phi_j(x) \Psi_k(q^r), \quad (10.44)$$

which is easily evaluated once the coefficients  $u_{jk}$  have been determined.

For very low parameter dimensions—e.g.,  $p \leq 5$ —one can employ tensored 1-D Gaussian techniques. For low to moderate dimensionalities, the sparse grid techniques summarized in Section 11.1 provide a reasonable balance between accuracy and efficiency.

### Collocation

To define the collocation system, we enforce

$$u(x, q^m) = u^K(x, q^m) = \sum_{j=1}^J u_{jm} \phi_j(x)$$

at  $M$  collocation points by taking the basis functions  $\Psi_k$  to be Lagrange polynomials which satisfy the collocation property  $L_k(q^m) = \delta_{km}$  at the collocation points. This

yields the  $M$  relations

$$\int_{\mathcal{D}} N \left( \sum_{j=1}^J u_{jm} \phi_j(x), q^m \right) S(\phi_\ell(x)) dx = \int_{\mathcal{D}} F(q^m) \phi_\ell(x) dx \quad (10.45)$$

for  $\ell = 1, \dots, J$ . For each collocation point  $q^m$ , solution for  $u_{jm}$  requires the solution of a  $J \times J$  system. This can be accomplished using existing codes or executable files and hence is nonintrusive. The solution of  $M$  such systems is decoupled and highly parallelizable.

We note that (10.45) can be obtained from the discretized Galerkin system (10.43) if one takes the collocation points  $q^m$  to be the quadrature points  $q^r$  and employs the basis  $\Psi_k = L_k$ . As detailed in Sections 11.1 and 11.2, we use the same sparse grid nodes for quadrature and when constructing interpolating polynomials for collocation.

To evaluate the QoI (10.36), the choice of collocation points  $q^m$  as quadrature points  $q^r$  and use of Lagrange polynomials as basis functions yield

$$\begin{aligned} y &= \sum_{r=1}^R w^r \rho_Q(q^r) \sum_{j=1}^J u_{jr} \phi_j(x) \\ &= \sum_{r=1}^R w^r \rho_Q(q^r) \hat{u}_r(x), \end{aligned} \quad (10.46)$$

where

$$\hat{u}_r(x) = \hat{u}_m(x) = \sum_{j=1}^J u_{jr} \phi_j(x) \quad (10.47)$$

has been previously computed when solving (10.45).

### Discrete Projection

For discrete projection, we employ the approximation

$$u_k(x) = \frac{1}{\gamma_k} \sum_{r=1}^R u(x, q^r) \Psi_k(q^r) \rho_Q(q^r) w^r$$

to construct the generalized Fourier coefficients. This involves essentially the same computational effort as collocation since it requires the solution of  $R$  deterministic problems of size  $J$ .

### 10.2.3 Evolutionary PDEs

The extension of the stochastic Galerkin, collocation, and discrete projection methods to an evolutionary PDE of the form

$$\begin{aligned} \frac{\partial u}{\partial t} &= \mathcal{N}(u, Q) + F(Q) \quad , \quad x \in \mathcal{D}, \quad t \in [0, \infty), \\ B(u, Q) &= G(Q) \quad , \quad x \in \partial \mathcal{D}, \quad t \in [0, \infty), \\ u(0, x, Q) &= I(Q) \quad , \quad x \in \mathcal{D}, \end{aligned}$$

can be achieved by combining the approaches detailed in Sections 10.2.1 and 10.2.2. Here  $\mathcal{N}$  again denotes a potentially nonlinear spatial differential operator,  $F$  is a source term,  $B$  and  $G$  are boundary operators,  $I$  specifies initial conditions, and  $\mathcal{D}$  is a subset of  $\mathbb{R}^1$ ,  $\mathbb{R}^2$ , or  $\mathbb{R}^3$ . To simplify notation, we suppress the dependence of these operators on  $x$  and  $t$ . The weak deterministic model formulation is

$$\int_{\mathcal{D}} \frac{\partial u}{\partial t} v dx + \int_{\mathcal{D}} N(u, Q) S(v) dx = \int_{\mathcal{D}} F(Q) v dx,$$

which holds for  $v \in V$ , where  $V$  is a space of test functions that satisfy essential boundary conditions. The weak stochastic model formulation is

$$\begin{aligned} & \int_{\Gamma} \int_{\mathcal{D}} \frac{\partial u}{\partial t} v(x) z(q) \rho_Q(q) dx dq + \int_{\Gamma} \int_{\mathcal{D}} N(u, q) S(v(x)) z(q) \rho_Q(q) dx dq \\ &= \int_{\Gamma} \int_{\mathcal{D}} F(q) v(x) z(q) \rho_Q(q) dx dq \end{aligned}$$

for  $v \in V$  and  $z \in Z$ , where  $Z$  is defined in (10.23).

The QoI is taken to be the expected value

$$y(t, x) = \int_{\Gamma} u(t, x, q) \rho_Q(q) dq.$$

To approximate  $u(t, x, q)$ , we again construct finite-dimensional subspaces  $V^J = \text{span}\{\phi_j\} \subset V$  and  $Z^K = \text{span}\{\Psi_k\} \subset Z$ , where the spatial basis functions  $\phi_j(x)$  are splines, finite elements, or spectral functions and orthogonal polynomials  $\Psi_k(q)$  are employed as a basis for the random component. The approximate solutions are then taken to be

$$u^K(t, x, Q) = \sum_{k=0}^K \sum_{j=1}^J u_{jk}(t) \phi_j(x) \Psi_k(Q).$$

### Stochastic Galerkin Method

The use of a quadrature rule with points  $q^r$  and weights  $w^r$  yields the Galerkin system

$$\begin{aligned} & \sum_{r=1}^R \Psi_i(q^r) \rho_Q(q^r) w^r \sum_{k=0}^K \sum_{j=1}^J \frac{du_{jk}}{dt} \Psi_k(q^r) \int_{\mathcal{D}} \phi_j(x) \phi_\ell(x) dx \\ &+ \sum_{r=1}^R \Psi_i(q^r) \rho_Q(q^r) w^r \int_{\mathcal{D}} N \left( \sum_{k=0}^K \sum_{j=1}^J u_{jk} \phi_j(x) \Psi_k(q^r), q^r \right) S(\phi_\ell(x)) dx \\ &= \sum_{r=1}^R \Psi_i(q^r) \rho_Q(q^r) w^r \int_{\mathcal{D}} F(q^r) \phi_\ell(x) dx, \end{aligned} \tag{10.48}$$

which holds for  $\ell = 1, \dots, J$  and  $i = 0, \dots, K$ . Once the coefficients  $u_{jk}(t)$  have been obtained by integrating the  $J(K+1)$  system, the approximated QoI is

$$y(t, x) = \sum_{r=1}^R w^r \rho_Q(q^r) \sum_{k=0}^K \sum_{j=1}^J u_{jk}(t) \phi_j(x) \Psi_k(q^r).$$

### Collocation

The basis choice  $\Psi_k(q) = L_k(q)$  which collocates at the quadrature points  $q^m = q^r$  yields the  $M = R$  evolution equations

$$\frac{du_{jr}}{dt} + \int_{\mathcal{D}} N \left( \sum_{j=1}^J u_{jr} \phi_j(x), q^r \right) S(\phi_\ell(x)) dx = \int_{\mathcal{D}} F(q^r) \phi_\ell(x) dx$$

for  $\ell = 1, \dots, J$ . The QoI is approximated by

$$y(t, x) = \sum_{r=1}^R w^r \rho_Q(q^r) \hat{u}_r(t, x),$$

where

$$\hat{u}_r(t, x) = \sum_{j=1}^J u_{jr}(t) \phi_j(x).$$

### Discrete Projection

For the discrete projection method, the generalized Fourier coefficients are approximated by the representation

$$u_k(t, x) = \frac{1}{\gamma_k} \sum_{r=1}^R u(t, x, q^r) \Psi_k(q^r) \rho_Q(q^r) w^r.$$

## 10.2.4 Attributes of the Galerkin, Collocation, and Discrete Projection Methods

### Stochastic Galerkin

- Because the Galerkin method is based on the projection of the residual onto the space of approximating polynomials, its accuracy is optimal in an  $L^2$  sense. This can reduce the number of required computations.
- For boundary value problems and PDEs in which the states are approximated by  $J$  elements, solution for  $u_{jk}$  requires the solution of a  $J(K+1) \times J(K+1)$  system that generally exhibits coupling between the spatial and parameter components. As illustrated in the examples of Section 10.3, the system can occasionally be decoupled for linear problems when the same polynomials are used to approximate parameters and states—e.g., Hermite or Legendre polynomials are used to represent Gaussian or uniform random variables.

- Quadrature rules to approximate integrals over the  $p$ -dimensional parameter space are required at two points for implementation: evaluation of the inner products in the projection and evaluation of the QoI. For low to moderate dimensions, this necessitates use of the sparse grid techniques summarized in Section 11.1.
- The stochastic Galerkin method has three primary disadvantages. First, it can be used only for densities with associated orthogonal polynomials. This precludes its use for general densities constructed using Bayesian techniques. Second, it relies on the assumption that parameters are mutually independent, which, as discussed in Section 5.2, is often not the case in applications. For applications where marginal distributions and correlation matrices are available, this can often be addressed using a Nataf transformation. Finally, a new system associated with the projection must be developed and implemented. Hence the method is *intrusive* and existing codes typically cannot be directly used to construct the coefficients. For complex applications with large legacy codes or problems for which executable files may be the only option, this disadvantage can be prohibitive.
- Convergence theory for the stochastic Galerkin method can be found in [19, 148] and the references cited therein.

### Stochastic Collocation

- Whereas the convergence analysis for collocation methods is based on interpolation theory for polynomials, implementation algorithms can be constructed by employing Lagrange polynomials as basis and test functions in the discretized Galerkin framework. By choosing the quadrature points as collocation points, one can decouple the stochastic and deterministic components of the problem. Despite this construction, however, we emphasize that collocation is an interpolation technique as compared to the Galerkin and discrete projection techniques, which are projection methods. This is manifested by the fact that the approximation space changes as the number of collocation points changes.
- The method is nonintrusive in the sense that once the  $M$  collocation points are specified, one solves  $M$  deterministic problems using existing software, including legacy codes or executable files. This decoupled solution technique is highly parallelizable and can be viewed as postprocessing. This constitutes a major advantage over the Galerkin method.
- For boundary value problems and stationary PDEs, computation of the coefficients  $u_{jm}$  and construction of  $\hat{u}_m(x)$  given by (10.47) require the solution of  $M$  deterministic problems of size  $J$ . This is in comparison to the solution of a single system of size  $J(K + 1)$  for the Galerkin method. Once  $\hat{u}_m(x)$  has been constructed, the QoI simply requires a vector multiply (10.46) rather than evaluation of the basis functions at quadrature points, as required for the Galerkin method.

- The construction of Lagrange polynomials and choice of collocation points, for low to moderate parameter dimensions, constitute the two critical aspects of the method. As discussed in Section 11.2, this is often addressed using sparse collocation techniques.
- An important feature of the collocation approach is the fact that it is applicable to general parameter distributions, including those constructed using the Bayesian model calibration techniques of Chapter 8. This is an advantage over the Galerkin and discrete projection methods where the orthogonality of the inner products depends on the compatibility between  $\rho_Q(q)$  and the basis functions  $\Psi_k(q)$ —e.g., Hermite and Legendre polynomials for normal and uniform densities. Whereas collocation does not explicitly require that parameters be mutually independent, the evaluation of the QoI requires sampling from the estimated joint density  $\rho_Q(q)$ . For correlated parameter sets, this requires that  $\rho_Q(q)$  be constructed directly since it is not the product of the marginal densities in this case. Alternatively, if marginal densities and a correlation matrix are available, one can employ a Nataf transformation, in the manner described in Section 5.2, to formulate the problem in terms of mutually independent Gaussian random variables. For applications where this information is not available, one may need to employ the sampling methods detailed in Section 9.2 and illustrated in Example 9.14.
- The accuracy of the method is dictated by the accuracy of the approximating polynomials. As detailed in Section 11.2, the interpolation error for  $p$  parameters and  $M$  collocation points is  $f - \mathcal{I}_M f = \mathcal{O}(M^{-\alpha/p})$ , where  $\mathcal{I}_M$  is the interpolation operator and  $\alpha$  depends on regularity of the solution. Hence the accuracy degrades as the dimension increases. Comparison with the convergence rate  $\mathcal{O}(M^{-1/2})$  for Monte Carlo methods illustrates that the latter is more efficient for large  $p$ . The stochastic Galerkin method is generally more accurate than stochastic collocation but requires the intrusive construction of the discrete system.
- Details regarding the theory and application of the stochastic collocation method to steady and evolutionary PDEs can be found in [18, 101, 183, 267].

## Discrete Projection

- The discrete projection method is also termed pseudospectral, nonintrusive PC, and nonintrusive spectral projection (NISP) in the literature. It shares the two primary advantages of the collocation method: it decouples the random and deterministic components of the problem, and it is nonintrusive. Once the  $R$  quadrature points are chosen, it requires  $R$  solutions of the deterministic problem which is of size  $J$  for boundary value problems and stationary PDEs. This can be achieved using existing software and is highly parallelizable.
- The method is equivalent to collocation if Lagrange polynomials are employed as basis functions,  $\Psi_k = L_k$ , and the density and weight are unity.

- For implementation with Hermite or Legendre polynomials, the assumption of mutually independent random variables is generally required to construct the density  $\rho_Q(q)$ . As discussed in Section 5.2, however, parameters are often correlated even for very simple models. This necessitates the use of a Nataf or Rosenblatt transformation to construct uncorrelated parameter sets or alternative techniques such as the sampling methods detailed in Section 9.2.
- Details regarding the construction of sparse grid quadrature techniques for low to moderate parameter dimensions are provided in Section 11.1.
- Theory for the discrete projection method can be found in [18].

## 10.3 Stochastic Galerkin Method—Examples

In this section, we illustrate aspects of the stochastic Galerkin method using very simple ODE examples in Section 10.3.1 and a 1-D boundary value problem in Section 10.3.2. In both cases, we illustrate improvements in efficiency that can be realized when the same basis functions are used to represent the parameters and state solutions for linear problems. Whereas these examples illustrate computations in a setting where they can be done explicitly, they do not represent the complexity of applications where stochastic Galerkin methods are advantageous or required. We refer the reader to [148] for details regarding the applications of the method to a 2-D heat equation approximated by finite elements as well as flow problems including the Navier–Stokes equations. An overview summarizing the application of stochastic Galerkin methods to flow in porous media, incompressible and compressible flows, reacting flows, and thermofluid flows can be found in [180, 268] and the cited references.

### 10.3.1 Scalar Initial Value Problem

**Example 10.9.** To illustrate the stochastic Galerkin method, we consider the initial value problem

$$\begin{aligned} \frac{du}{dt} &= -\alpha(\omega)u, \\ u(0, \omega) &= \bar{\beta}, \end{aligned} \tag{10.49}$$

where  $\bar{\beta}$  is deterministic and fixed. We first assume that  $\alpha \sim N(\bar{\alpha}, \sigma_\alpha^2)$  with  $\bar{\alpha} > 0$ . The QoI is the mean  $\bar{u}(t)$  and variance  $\text{var}[u(t)]$ . As illustrated in Example 10.4, the random parameter can be expressed exactly as

$$\alpha = \alpha^N = \sum_{n=0}^N \alpha_n \psi_n(Q), \quad \alpha_0 = \bar{\alpha}, \quad \alpha_1 = \sigma_\alpha, \quad \alpha_n = 0, \quad n > 1, \tag{10.50}$$

where  $\psi_n(Q)$  are Hermite polynomials defined on  $\mathbb{R}$ . Here  $Q \sim N(0, 1)$  with density  $\rho_Q(q) = \frac{1}{\sqrt{2\pi}} e^{-q^2/2}$ . The deterministic initial condition also has an exact represen-

tation in the space of Hermite polynomials since

$$\beta = \beta^N = \sum_{n=0}^N \beta_n \psi_n(Q) , \quad \beta_0 = \bar{\beta}, \quad \beta_n = 0, \quad n > 0.$$

Finally, we note that the analytic random solution is

$$u(t, Q) = \bar{\beta} e^{-(\bar{\alpha} + \sigma_\alpha Q)t}.$$

We seek approximate solutions

$$u^K(t, Q) = \sum_{k=0}^K u_k(t) \psi_k(Q)$$

subject to

$$\begin{aligned} 0 &= \left\langle \frac{du^K}{dt} + \alpha^N u^K, \psi_i \right\rangle_\rho \\ &= \int_{\mathbb{R}} \sum_{k=0}^K \frac{du_k}{dt}(t) \psi_k(q) \psi_i(q) \rho_Q(q) dq + \int_{\mathbb{R}} \alpha^N \sum_{k=0}^K u_k(t) \psi_k(q) \psi_i(q) \rho_Q(q) dq \end{aligned} \quad (10.51)$$

for  $i = 0, 1, \dots, K$ . Substitution of (10.50) yields the  $K + 1$  differential equations

$$\frac{du_i}{dt} = -\frac{1}{\gamma_i} \sum_{n=0}^N \sum_{k=0}^K \alpha_n u_k(t) e_{ink} \quad (10.52)$$

specifying the state coefficients. For Hermite basis functions, the expected values

$$\begin{aligned} \gamma_i &= \mathbb{E}[\psi_i^2(Q)] = \int_{\mathbb{R}} \psi_i^2(q) \rho_Q(q) dq, \\ e_{ink} &= \mathbb{E}[\psi_i(Q) \psi_n(Q) \psi_k(Q)] = \int_{\mathbb{R}} \psi_i(q) \psi_n(q) \psi_k(q) \rho_Q(q) dq \end{aligned} \quad (10.53)$$

have the explicit values

$$\begin{aligned} \gamma_i &= i!, \\ e_{ink} &= \begin{cases} \frac{i! n! k!}{(s-i)!(s-n)!(s-k)!}, & 2s = i + n + k \text{ is even and } s \geq i, n, k, \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \quad (10.54)$$

Since

$$u^K(0, Q) = \sum_{k=0}^K u_k(0) \psi_k(Q) = \beta = \sum_{n=0}^N \beta_n \psi_n(Q),$$

initial conditions for (10.52) are specified by

$$u_k(0) = \beta_k, \quad k = 0, \dots, K.$$

The linear coupled system of differential equations can be expressed as the vector system

$$\begin{aligned}\frac{d\vec{u}}{dt} &= A\vec{u}(t), \\ \vec{u}(0) &= \vec{\beta},\end{aligned}\tag{10.55}$$

where  $\vec{u}(t) = [u_0(t), \dots, u_K(t)]^T$  and  $\vec{\beta} = [\beta_0, \dots, \beta_K]^T = [\bar{\beta}, 0, \dots, 0]^T$ . The tridiagonal matrix  $A = [A_{ik}]$  has components

$$A_{ik} = -\frac{1}{\gamma_i} \sum_{n=0}^N \alpha_n e_{ink} = -\frac{1}{\gamma_i} [\bar{\alpha} e_{i0k} + \sigma_\alpha e_{i1k}]$$

so that

$$A = \begin{bmatrix} -\bar{\alpha} & -\sigma_\alpha & & & \\ -\sigma_\alpha & -\bar{\alpha} & -2\sigma_\alpha & & \\ & \ddots & & \ddots & \\ & & & -K\sigma_\alpha & \\ -\sigma_\alpha & & -\bar{\alpha} & & \end{bmatrix}.\tag{10.56}$$

The ODE system (10.55) can be numerically approximated using standard algorithms such as stiff solvers (e.g., `ode15s.m`) or Runge–Kutta routines (e.g., `ode45.m`).

To evaluate the QoI, it follows from (10.9) and (10.10) that once the coefficients have been computed, the mean and variance of  $u^K(t, Q)$  are

$$\begin{aligned}\mathbb{E}[u^K(t, Q)] &= u_0(t), \\ \text{var}[u^K(t, Q)] &= \sum_{k=1}^K u_k^2(t) \gamma_k.\end{aligned}\tag{10.57}$$

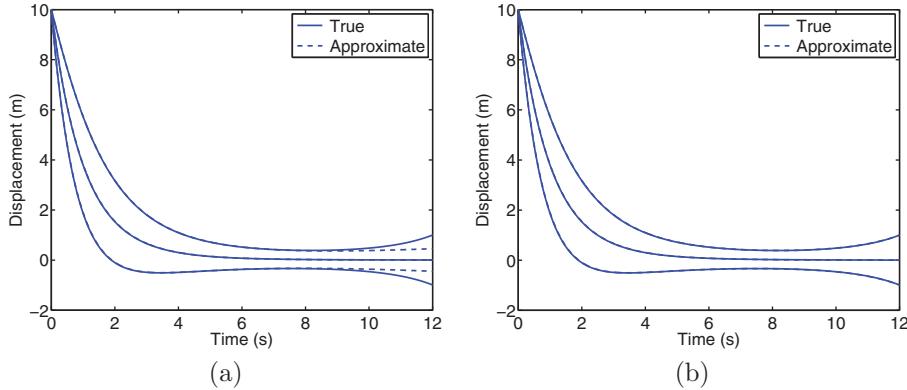
For this example, exact mean and variance values

$$\begin{aligned}\bar{u}(t) &= \int_{\mathbb{R}} \bar{\beta} e^{-(\bar{\alpha} + \sigma_\alpha q)t} \cdot \frac{1}{\sqrt{2\pi}} e^{-q^2/2} dq \\ &= \bar{\beta} e^{-\bar{\alpha}t} e^{\sigma_\alpha^2 t^2/2}\end{aligned}$$

and

$$\begin{aligned}\text{var}[u] &= \mathbb{E}[u^2(t)] - \bar{u}^2(t) \\ &= e^{-2\bar{\alpha}t} \bar{\beta}^2 \left( e^{2\sigma_\alpha^2 t^2} - e^{-\sigma_\alpha^2 t^2} \right)\end{aligned}$$

can be computed using the analytic random solution. We first note that due to the nonlinear parameter dependence,  $\bar{u}(t)$  differs from the deterministic solution  $u(t, \bar{\alpha}) = \bar{\beta} e^{-\bar{\alpha}t}$  evaluated at  $\bar{\alpha}$ . Second, both  $\bar{u}$  and  $\text{var}[u]$  exhibit unbounded growth for large  $t$ . This is due to the assumption that  $\alpha \sim N(\bar{\alpha}, \sigma_\alpha^2)$ . Despite specifying that  $\bar{\alpha} > 0$ , some realizations of  $\alpha$  will be negative, which produces exponential growth. It is shown in Example 10.11 that this is avoided if  $\alpha \sim \mathcal{U}(0, \alpha_{\max})$ .



**Figure 10.1.** True and approximate means and  $2\sigma$  credible intervals with (a)  $K = 8$  and (b)  $K = 16$ .

The true means and  $2\sigma$  credible intervals, computed using  $\bar{\alpha} = 1.0$ ,  $\sigma_\alpha = 0.25$ ,  $\bar{\beta} = 10.0$ ,  $\sigma_\beta = 2.0$ , are compared in Figure 10.1 with approximate values given by (10.57) with  $K = 8$  and  $K = 16$ . This illustrates that despite the fact that the representation for  $\alpha^N$  truncates at  $N = 2$ , a larger number of basis functions is required to quantify the exponential growth for large  $t$ .

**Example 10.10.** We now consider

$$\begin{aligned} \frac{du}{dt} &= -\alpha(\omega)u \quad , \quad t > 0, \\ u(0, \omega) &= \beta(\omega) \end{aligned} \tag{10.58}$$

with stochastic decay parameter  $\alpha \sim N(\bar{\alpha}, \sigma^2)$  and initial condition  $\beta \sim N(\bar{\beta}, \sigma_\beta^2)$ . We employ tensored Hermite basis functions

$$\Psi_k(Q) = \psi_{k_1}(Q_1)\psi_{k_2}(Q_2)$$

with the ordering summarized in Table 10.2. The state is approximated by

$$u^K(t, Q) = \sum_{|\mathbf{k}'|=0}^K u_k(t) \Psi_{k_1}(Q_1) \Psi_{k_2}(Q_2) = \sum_{k=0}^K u_k(t) \Psi_k(Q),$$

whereas the parameters have the exact representations

$$\begin{aligned} \alpha^N &= \alpha = \sum_{n=0}^N \alpha_n \Psi_n(Q) \quad , \quad \alpha_0 = \bar{\alpha}, \alpha_1 = \sigma_\alpha, \alpha_n = 0, n > 1 \\ &= \bar{\alpha} + \sigma_\alpha Q_1, \\ \beta^N &= \beta = \sum_{n=0}^N \beta_n \Psi_n(Q) \quad , \quad \beta_0 = \bar{\beta}, \beta_1 = 0, \beta_2 = \sigma_\beta, \beta_n = 0, n > 2 \\ &= \bar{\beta} + \sigma_\beta Q_2 \end{aligned}$$

since  $\Psi_0(Q) = 1$ ,  $\Psi_1(Q) = Q_1$ , and  $\Psi_2(Q) = Q_2$ .

| $k$ | $ \mathbf{k}' $ | Multi-Index | Polynomial               |
|-----|-----------------|-------------|--------------------------|
| 0   | 0               | (0, 0)      | $\psi_0(Q_1)\psi_0(Q_2)$ |
| 1   | 1               | (1, 0)      | $\psi_1(Q_1)\psi_0(Q_2)$ |
| 2   |                 | (0, 1)      | $\psi_0(Q_1)\psi_1(Q_2)$ |
| 3   | 2               | (2, 0)      | $\psi_2(Q_1)\psi_0(Q_2)$ |
| 4   |                 | (1, 1)      | $\psi_1(Q_1)\psi_1(Q_2)$ |
| 5   |                 | (0, 2)      | $\psi_0(Q_1)\psi_2(Q_2)$ |

**Table 10.2.** Single index, multi-index, and tensored polynomials for  $p = 2$ .

With the multi-index notation, the weak formulation of the model is identical to (10.51) but with integration over  $\mathbb{R}^2$ . For example, the tensored normalization constants are

$$\begin{aligned}\gamma_{\mathbf{i}'} &= \int_{\mathbb{R}^2} \Psi_{\mathbf{i}'}^2(q) \rho_Q(q) dq \\ &= \int_{\mathbb{R}} \psi_{i_1}^2(q_1) \rho_{q_1}(q_1) dq_1 \int_{\mathbb{R}} \psi_{i_2}^2(q_2) \rho_{q_2}(q_2) dq_2 \\ &= \gamma_{i_1} \gamma_{i_2}.\end{aligned}$$

In a similar manner, the 1-D relations can also be used to evaluate the expected values  $\mathbb{E}[\Psi_{\mathbf{i}'}(q)\Psi_{\mathbf{n}'}(q)\Psi_{\mathbf{k}'}(q)]$ . However, the matrix  $A$  cannot easily be constructed using tensored 1-D matrices of the form (10.56) due to the ordering of basis elements and definition in terms of the total degree  $P$ . This differs from certain finite element or spectral expansions that admit mass matrix representations  $\mathcal{M} = M \otimes M$ , where  $M$  is the 1-D mass matrix. To illustrate, we note that the matrix  $A$  for  $P = 2$  ( $K = 5$ ) is

$$A = \begin{bmatrix} -\bar{\alpha} & -\sigma_\alpha & 0 & 0 & 0 & 0 \\ -\sigma_\alpha & -\bar{\alpha} & 0 & -2\sigma_\alpha & 0 & 0 \\ 0 & 0 & -\bar{\alpha} & 0 & -\sigma_\alpha & 0 \\ 0 & -\sigma_\alpha & 0 & -\bar{\alpha} & 0 & 0 \\ 0 & 0 & -\sigma_\alpha & 0 & -\bar{\alpha} & 0 \\ 0 & 0 & 0 & 0 & 0 & -\bar{\alpha} \end{bmatrix}.$$

Hence one would typically employ exact Gauss–Hermite representations for this case rather than relying on tensored analytic 1-D relations.

The vector differential equation is again (10.55) with  $\vec{\beta} = [\bar{\beta}, 0, \sigma_\beta, 0, \dots, 0]^T$  and  $\vec{u}(t) = [u_0(t), \dots, u_K(t)]^T$ . Once the time-dependent coefficients have been determined, the mean and variance of  $u^K(t, \xi)$  are given by (10.57).

For the random solution

$$u(t, Q) = (\bar{\beta} + \sigma_\beta Q_2) e^{-(\bar{\alpha} + \sigma_\alpha Q_1)t},$$

the true mean and variance are

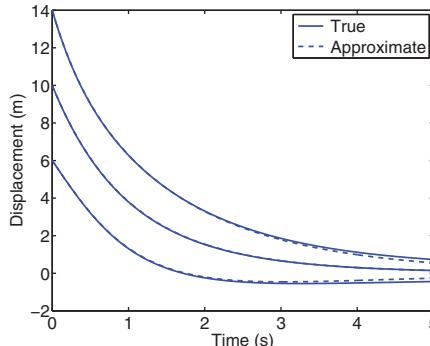
$$\begin{aligned}\bar{u}(t) &= \frac{1}{2\pi} \int_{\mathbb{R}} \int_{\mathbb{R}} (\bar{\beta} + \sigma_{\beta} q_2) e^{-(\bar{\alpha} + \sigma_{\alpha} q_1)t} e^{-q_1^2/2} e^{-q_2^2/2} dq_1 dq_2 \\ &= \bar{\beta} e^{-\bar{\alpha}t} e^{\sigma_{\alpha}^2 t^2 / 2}\end{aligned}$$

and

$$\begin{aligned}\text{var}(u) &= \mathbb{E}[u^2(t, Q)] - \bar{u}^2(t) \\ &= e^{-2\bar{\alpha}t} \left[ e^{2\sigma_{\alpha}^2 t^2} (\bar{\beta}^2 + \sigma_{\beta}^2) - \bar{\beta}^2 e^{-\sigma_{\alpha}^2 t^2} \right].\end{aligned}$$

We again note the unbounded growth of  $\bar{u}(t)$  and  $\text{var}(u)$  due to the fact that the normal density admits negative realizations of  $\alpha$ . This can be contrasted with  $\alpha \sim \mathcal{U}(0, \alpha_{\max})$  considered in Example 10.11.

The true and approximate means and  $2\sigma$ -credible bands for  $\bar{\alpha} = 1.0$ ,  $\sigma_{\alpha} = 0.25$ ,  $\bar{\beta} = 10.0$ ,  $\sigma_{\beta} = 2.0$ , and  $K = 6$  basis functions are plotted in Figure 10.2. The  $2\sigma_{\beta} = 4$  interval at  $t = 0$  reflects the variability in the initial condition. These results can be compared with those in Figure 10.1 for the single random variable  $\alpha$ . The slight discrepancy between the true and approximate means and credible intervals is due to the limited number of basis functions. The convergence of the method is illustrated in Exercise 10.3.



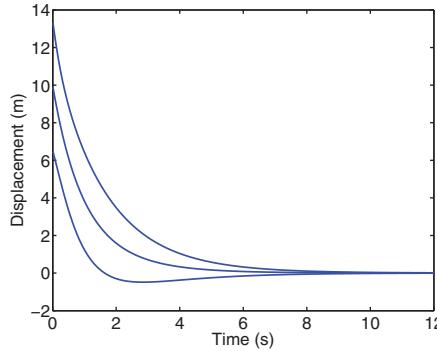
**Figure 10.2.** True and approximate means and  $2\sigma$  credible intervals obtained with  $K = 6$ .

**Example 10.11.** To provide bounded means and variances, we consider (10.58) with the assumption that  $\alpha \sim \mathcal{U}(\frac{1}{2}, \frac{3}{2})$  and  $\beta \sim \mathcal{U}(7, 13)$  are uniformly distributed with means  $\bar{\alpha} = 1$ ,  $\bar{\beta} = 10.0$  and variances  $\sigma_{\alpha}^2 = \frac{1}{12}$ ,  $\sigma_{\beta}^2 = 3$ . It follows from Example 10.5 and Exercise 4.4 that the parameters can be expressed as

$$\alpha = 1 + \frac{1}{2}Q_1, \quad \beta = 10 + 3Q_2,$$

where  $Q_1, Q_2 \sim \mathcal{U}(-1, 1)$ . The analytic solution is thus

$$u(t, Q) = (10 + 3Q_2)e^{-(1 + \frac{1}{2}Q_1)t}. \quad (10.59)$$



**Figure 10.3.** True mean and  $2\sigma$  credible interval for the solution (10.59) for uniformly distributed parameters  $\alpha$  and  $\beta$ .

It is established in Exercise 4.4 that the mean and variance of  $u$  are

$$\bar{u}(t) = \frac{20}{t} \sinh(t/2) e^{-t}$$

and

$$\text{var}[u(t)] = \left( \frac{103}{t} \sinh(t) - \frac{400}{t^2} \sinh^2(t/2) \right) e^{-2t}.$$

The mean and  $2\sigma$  credible interval are plotted in Figure 10.3. In contrast to the normally distributed parameter  $\alpha$ , which admitted negative realizations, the realizations of the uniformly distributed  $\alpha$  are guaranteed to be nonnegative. Hence  $\bar{u}$  and  $\sigma[u]$  decay for large  $t$  rather than exhibiting the exponential growth observed in Example 10.10.

Legendre polynomials now constitute the natural choice of basis functions. The construction of  $u^K(t, Q)$  using the stochastic Galerkin method is addressed in Exercise 10.2.

### 10.3.2 Boundary Value Problem

**Example 10.12.** Consider the heat equation

$$\begin{aligned} \alpha(\omega) \frac{d^2 u}{dx^2} &= -f(x) \quad , \quad -1 < x < 1, \\ u(-1) &= u(1) = 0, \end{aligned} \tag{10.60}$$

where  $\alpha \sim N(\bar{\alpha}, \sigma_\alpha^2)$ . It was shown in Example 10.9 that

$$\alpha = \alpha^N = \bar{\alpha} + \sigma_\alpha Q = \sum_{n=0}^1 \alpha_n \Psi_n(Q),$$

where  $\Psi_0(Q) = 1$  and  $\Psi_1(Q) = Q$  are the first two Hermite polynomials. Since  $Q \sim N(0, 1)$ , the density is  $\rho_Q(q) = \frac{1}{\sqrt{2\pi}} e^{-q^2/2}$ . Following Example 10.8, the

operators  $N(u, Q)$  and  $S(v)$  are

$$N(u, Q) = (\bar{\alpha} + \sigma_\alpha Q) \frac{du}{dx} , \quad S(v) = \frac{dv}{dx}$$

and the stochastic weak formulation is

$$\int_{\mathbb{R}} \int_{-1}^1 (\bar{\alpha} + \sigma_\alpha q) \frac{du}{dx} \frac{dv}{dx} z(q) \rho_Q(q) dx dq = \int_{\mathbb{R}} \int_{-1}^1 f v(x) z(q) \rho_Q(q) dx dq,$$

which must hold for all  $v \in H_0^1(-1, 1)$  and  $z \in L_\rho^2(\mathbb{R})$ .

To construct  $V^J$ , we consider a uniform partition of the interval  $[-1, 1]$  with points  $x_j = -1 + jh$ ,  $j = 0, \dots, J$  and uniform stepsize  $h = \frac{2}{J}$ . The spatial basis is

$$\phi_j(x) = \frac{1}{h} \begin{cases} x - x_{j-1} & , \quad x_{j-1} \leq x < x_j, \\ x_{j+1} - x & , \quad x_j \leq x < x_{j+1}, \\ 0 & , \quad \text{otherwise,} \end{cases}$$

where  $j = 1, \dots, J-1$  to enforce the essential boundary conditions. The spatial subspace is then  $V^J = \text{span}\{\phi_j\}_{j=1}^{J-1}$ . The random space is  $Z^K = \text{span}\{\Psi_k\}_{k=0}^K$ , where  $\Psi_k(q)$  are the Hermite polynomials, and the approximate solution is

$$u^K(x, Q) = \sum_{k=0}^K \sum_{j=1}^{J-1} u_{jk} \phi_j(x) \Psi_k(Q).$$

This yields the discretized problem

$$\begin{aligned} \int_{\mathbb{R}} (\bar{\alpha} + \sigma_\alpha q) \sum_{j=1}^{J-1} \sum_{k=0}^K u_{jk} \Psi_k(q) \left[ \int_{-1}^1 \phi'_j(x) \phi'_\ell(x) dx \right] \Psi_i(q) \rho_Q(q) dq \\ = \int_{\mathbb{R}} \left[ \int_{-1}^1 f(x) \phi'_\ell(x) dx \right] \Psi_i(q) \rho_Q(q) dq, \end{aligned}$$

which holds for  $\ell = 1, \dots, J$  and  $i = 0, \dots, K$ . With the definitions

$$\Phi_{j,\ell} = \int_{-1}^1 \phi'_j(x) \phi'_\ell(x) dx = \frac{1}{h} \begin{cases} 2 & , \quad j = \ell, \\ -1 & , \quad j = \ell - 1 \text{ or } j = \ell + 1, \\ 0 & , \quad \text{otherwise,} \end{cases}$$

$$f_\ell = \int_{-1}^1 f(x) \phi_\ell(x) dx,$$

the problem can be expressed as

$$\sum_{j=1}^{J-1} \Phi_{j,\ell} \sum_{k=0}^K u_{jk} \int_{\mathbb{R}} (\bar{\alpha} + \sigma_\alpha q) \Psi_k(q) \Psi_i(q) \rho_Q(q) dq = f_\ell \int_{\mathbb{R}} \Psi_i(q) \rho_Q(q) dq.$$

Due to the orthogonality of the Hermite polynomials, we note that

$$\int_{\mathbb{R}} \Psi_k(q) \Psi_i(q) \rho_Q(q) dq = k! \delta_{ki}.$$

Furthermore, (10.53) and (10.54) with  $n = k = 0$  yield

$$\int_{\mathbb{R}} \Psi_\ell(q) \rho_Q(q) dq = \begin{cases} 1 & , \ell = 0, \\ 0 & , \text{ otherwise,} \end{cases}$$

whereas the same expression with  $i = 1$  and  $n, k$  variable yields explicit values for

$$e_{1ki} = \int_{\mathbb{R}} q \Psi_n(q) \Psi_k(q) \rho_Q(q) dq.$$

This facilitates the construction of the  $(J - 1) \times (K + 1) \times (J - 1) \times (K + 1)$  system required to solve for the coefficients  $u_{jk}$ .

## 10.4 Discrete Projection Method—Example

**Example 10.13.** We revisit Example 9.7, where we illustrated the propagation of uncertain parameters  $Q = [m, c, k]$  through the spring model

$$\begin{aligned} m \frac{d^2 z}{dt^2} + c \frac{dz}{dt} + kz &= f_0 \cos(\omega_F t), \\ z(0) = z_0, \quad \frac{dz}{dt}(0) &= z_1 \end{aligned} \tag{10.61}$$

using the Monte Carlo sampling methods of Section 9.2 and perturbation methods of Section 9.3. We consider the response

$$y(\omega_F, Q) = \frac{1}{\sqrt{(k - m\omega_F^2)^2 + (c\omega_F)^2}}, \tag{10.62}$$

and we assume that  $Q \sim N(\bar{q}, V)$  with mean  $\bar{q} = [2.7, 0.24, 8.5]$  and covariance matrix  $V$  given by (9.19). The QoI are the mean and standard deviation of the model response for driving frequencies  $\omega_F \in [0, 2.7]$ . In this example, we illustrate the propagation of uncertainties using the discrete projection method.

The parameters are represented as

$$\begin{aligned} m &= \bar{m} \Psi_0(Q) + \sigma_m \Psi_1(Q) = \bar{m} + \sigma_m Q_1, \\ c &= \bar{c} \Psi_0(Q) + \sigma_c \Psi_2(Q) = \bar{c} + \sigma_c Q_2, \\ k &= \bar{k} \Psi_0(Q) + \sigma_k \Psi_3(Q) = \bar{k} + \sigma_k Q_3, \end{aligned}$$

where  $\Psi_k(Q) = \psi_{k_1}(Q_1) \psi_{k_2}(Q_2) \psi_{k_3}(Q_3)$  are tensored Hermite polynomials with the ordering given in Table 10.1. The approximated response is

$$y^K(\omega_F, Q) = \sum_{k=0}^K y_k(\omega_F) \Psi_k(Q),$$

where the generalized Fourier coefficients are

$$y_k(\omega_F) = \frac{1}{\gamma_k} \int_{\mathbb{R}^3} y(\omega_F, q) \Psi_k(q) \rho_Q(q) dq. \quad (10.63)$$

The density is

$$\rho_Q(q) = \left( \frac{1}{\sqrt{2\pi}} \right)^3 e^{-q_1^2/2} e^{-q_2^2/2} e^{-q_3^2/2},$$

and  $\gamma_k = \gamma_{k_1} \gamma_{k_2} \gamma_{k_3}$ . As illustrated in (10.54), the individual expected values are  $\gamma_{k_1} = k_1!$ . Once  $y_k(\omega_F)$  has been computed, it follows from Property 10.1 that

$$\begin{aligned} \bar{y}(\omega_F) &= y_0(\omega_F), \\ \text{var}[y^K(\omega_F, Q)] &= \sum_{k=1}^K y_k^2(\omega_F) \gamma_k. \end{aligned} \quad (10.64)$$

The QoI follow directly from (10.64).

The tensor product relation (11.7) can be used to approximate (10.63) since  $p = 3$  is small. This yields the approximate relation

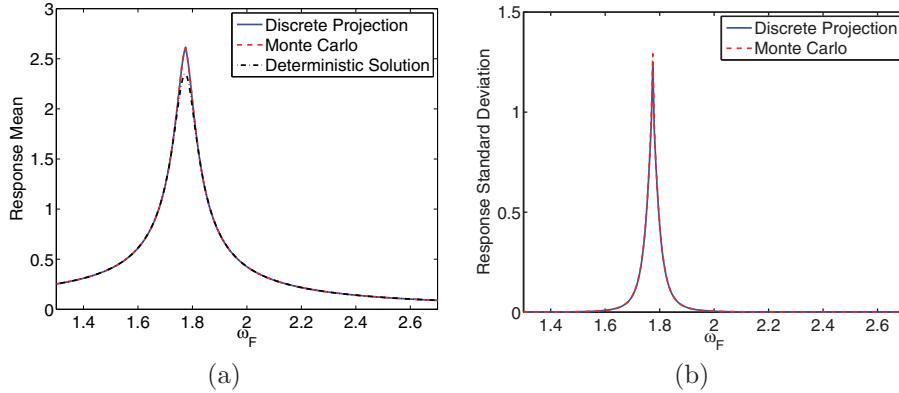
$$y_k(\omega_F) \approx \frac{1}{\gamma_k} \sum_{r_1=1}^{R_{\ell_1}} \sum_{r_2=1}^{R_{\ell_2}} \sum_{r_3=1}^{R_{\ell_3}} y(\omega_F, q^r) \Psi_k(q^r) \rho_Q(q^r) w_\ell^r,$$

where  $q^r = [q_1^{r_1}, q_2^{r_2}, q_3^{r_3}]$  and  $w_\ell^r = w_{\ell_1}^{r_1} w_{\ell_2}^{r_2} w_{\ell_3}^{r_3}$ . The quadrature points and weights can be specified using the Gauss–Hermite rule on the infinite domain or Gauss–Legendre, trapezoid, or Clenshaw–Curtis rules on a finite truncated or mapped domain.

The QoI given by (10.64) with  $K = 9$  and  $R = 10^3$  tensored Gauss–Hermite quadrature points are compared in Figure 10.4 with the Monte Carlo mean and standard deviation (9.20) obtained with  $M = 10^5$  realizations of the response  $y(\omega_F, q)$ . It is observed that the two techniques yield essentially identical QoI. We also note that the means  $\bar{y}(\omega_F)$  given by the discrete projection or sampling-based methods differ from the deterministic solution  $y(\omega_F, \bar{q})$  evaluated at  $\bar{q}$  for driving frequencies  $\omega_F$  near the natural frequency  $\omega_0 = 1.7743$  Hz. This is due to the nonlinear dependence of the solution on the parameters. Finally, we note that the total number  $R = 10^3$  of tensored quadrature points was chosen to simplify implementation and smaller numbers of points will yield comparable accuracy. The approximation of the QoI using sparse grid quadrature techniques is investigated in Exercise 11.1.

## 10.5 Stochastic Polynomial Packages

The status of comprehensive packages or toolboxes for stochastic polynomial methods is presently limited for two reasons: their use for uncertainty quantification is fairly recent, and they are intrinsically linked to the codes employed in underlying applications. There are a number of MATLAB codes being posted by researchers, and basic MATLAB toolboxes will become increasingly available as the



**Figure 10.4.** (a) Response means and (b) standard deviations computed using the discrete projection and Monte Carlo sampling methods and deterministic solution  $y(\omega_F, \bar{q})$ .

field matures. It is anticipated that these toolboxes will provide an important tool for education and research based on MATLAB application codes. For large-scale applications, the constraint of interfacing with user-supplied production codes is more stringent and dictates that toolboxes must incorporate both numerical algorithms and efficient interface mechanisms. One example is the Sandia National Laboratories toolbox DAKOTA (Design Analysis Kit for Optimization and Tera-scale Applications) [4, 5, 71]. This C++ toolkit facilitates the use of stochastic polynomial methods for uncertainty quantification—and related objectives such as optimization, sensitivity analysis, model calibration, and reduced-order model development—by providing a flexible and extendable interface between simulation codes and linear algebra, optimization, quadrature, and interpolation algorithms. The incorporation of stochastic polynomial methods in other large-scale commercial, government laboratory, and open source packages will surely grow as the field matures.

## 10.6 Exercises

**Exercise 10.1.** Consider the random differential equation

$$\begin{aligned} \frac{du}{dt} &= -\alpha(\omega)u, \quad t > 0, \\ u(0, \omega) &= \beta(\omega), \end{aligned} \tag{10.65}$$

where  $\alpha$  and  $\beta$  are uniformly distributed random variables with means  $\bar{\alpha}, \bar{\beta} > 0$  and variances  $\sigma_\alpha^2, \sigma_\beta^2$ . From (4.42), it follows that the stochastic solution is

$$u(t, Q) = (\bar{\beta} + \sqrt{3}\sigma_\beta Q_2)e^{-(\bar{\alpha} + \sqrt{3}\sigma_\alpha Q_1)t},$$

where  $Q_1, Q_2 \sim \mathcal{U}(-1, 1)$ . Show that the mean and variance of  $u$  are

$$\bar{u} = \bar{\beta} \frac{\sinh(\sqrt{3}\sigma_\alpha t)}{\sqrt{3}\sigma_\alpha t} e^{-\bar{\alpha}t}$$

and

$$\text{var}[u(t)] = \left[ (\bar{\beta}^2 + \sigma_\beta^2) \frac{\sinh(2\sqrt{3}\sigma_\alpha t)}{2\sqrt{3}\sigma_\alpha t} - \bar{\beta}^2 \frac{\sinh^2(\sqrt{3}\sigma_\alpha t)}{3\sigma_\alpha^2 t^2} \right] e^{-2\bar{\alpha}t}.$$

**Exercise 10.2.** Consider the differential equation (10.65) with  $\alpha \sim \mathcal{U}(\frac{1}{2}, \frac{3}{2})$  and  $\beta \sim \mathcal{U}(7, 13)$  as assumed in Example 10.11. Use the stochastic Galerkin method to compute the approximate solution  $u^N(t, \xi)$  using  $N = 6$  tensored Legendre basis functions. Plot the true and approximate means and  $2\sigma$  credible bands on the time interval  $[0, 5]$ .

**Exercise 10.3.** Compute the approximate mean and  $2\sigma$  credible bands for the random differential equation of Example 10.10 using a stochastic Galerkin expansion with  $N = 10$  tensored basis elements. Compare your solution with the true and approximate values, obtained with  $N = 6$ , that are plotted in Figure 10.2.

**Exercise 10.4.** Repeat Exercise 10.2 using the stochastic collocation method with various values of  $M$ . What value of  $M$  is required to achieve convergence?

**Exercise 10.5.** In Example 10.13, we illustrated the use of the discrete projection method to propagate the mean and standard deviation associated with the spring model (10.61) and response (10.62). Repeat this example using the stochastic collocation method. How do your results compare with those obtained using discrete projection and Monte Carlo sampling?

## Chapter 11

# Sparse Grid Quadrature and Interpolation Techniques

In this chapter, we discuss the tensor product and sparse grid quadrature and interpolation techniques required to implement the stochastic spectral methods of Chapter 10. This discussion necessarily focuses solely on techniques for these methods, and more general theory is cited at relevant points in the discussion.

## 11.1 Quadrature Techniques

The Galerkin, collocation, and discrete projection methods all require the approximation of integrals over the domain  $\Gamma \subset \mathbb{R}^p$ . For the Galerkin method, this occurs when projecting the residual onto the space spanned by the orthogonal polynomials (10.26), (10.42), and (10.48) and when approximating the QoI

$$y(t, x) = \mathbb{E}[u^K(t, x, Q)] = \int_{\Gamma} u^K(t, x, q) \rho_Q(q) dq.$$

Discrete projection necessitates the approximation of

$$u_k(t, x) = \frac{1}{\gamma_k} \langle u, \Psi_k \rangle_{\rho} = \frac{1}{\gamma_k} \int_{\Gamma} u(t, x, q) \Psi_k(q) \rho_Q(q) dq, \quad (11.1)$$

where  $\gamma_k = \langle \Psi_k, \Psi_k \rangle_{\rho}$ , along with the QoI.

For large parameter dimensions  $p$ , stochastic quadrature techniques are optimal and we summarize those in Section 11.1.1. In Section 11.1.2, we summarize deterministic 1-D and tensored quadrature relations. These are applicable for low dimensions and serve as a basis for constructing the sparse grid techniques discussed in Section 11.1.3.

### 11.1.1 Stochastic Quadrature Methods

The Monte Carlo method is the simplest stochastic quadrature technique. In this method, a (pseudo-)random number generator is used to sample realizations  $q^r = [q_1^r, \dots, q_p^r]$  from the joint density  $\rho_Q(q)$  constructed either experimentally or

using the statistical techniques of Chapters 7 and 8. For mutually independent components  $Q_i$ , one can independently sample from the marginal densities  $\rho_{Q_i}(q_i)$ .<sup>8</sup> For  $R$  samples, the inner product in (11.1) is approximated by

$$\langle u, \Psi_k \rangle_\rho = \frac{1}{R} \sum_{r=1}^R u(t, x, q^r) \Psi_k(q^r) \rho_Q(q^r) + \varepsilon_R,$$

where the error satisfies  $\mathbb{E}[\varepsilon_R] = 0$  and  $\varepsilon_R = \mathcal{O}(\frac{1}{\sqrt{R}})$  for large  $R$ . The advantage of the method is the fact that the convergence rate is independent of  $p$  and does not depend on the smoothness of  $u(t, x, Q)\Psi_k(Q)$ . For this reason, stochastic methods are advantageous when  $p$  is moderate to large. As detailed in Section 11.1.3, the range of  $p$  where stochastic methods become advantageous depends both on  $p$  and the regularity of  $u$ . The low convergence rate comprises the primary disadvantage of the method. As noted in Section 9.2, this convergence rate dictates that the number of simulations must be increased by a factor of 100 to gain one additional place of accuracy.

Improved convergence rates can be achieved by using more efficient stochastic sampling techniques, such as Latin hypercube sampling [170] or quasi-Monte Carlo sampling [174]. For example, quasi-Monte Carlo techniques have a convergence rate of  $\mathcal{O}(\ln R^p / R)$ . However, these techniques are still not competitive for low to moderate dimensions where tensored or sparse grid techniques are advantageous.

### 11.1.2 Deterministic Quadrature Methods: 1-D and Tensor Product Formulas

The Galerkin, collocation, and discrete projection methods require the approximation of integrals

$$I^{(p)} f = \int_{\Gamma} f(q) \rho_Q(q) dq,$$

$\Gamma \subset \mathbb{R}^p$ , by sums

$$\mathcal{Q}^{(p)} f = \sum_{r=1}^R f(q^r) w^r,$$

where the choice of quadrature points  $q^r$  and weights  $w^r$  defines the implementation algorithm and accuracy of the method. We consider first 1-D quadrature techniques since they provide the basis for quadrature in multiple dimensions.

#### 1-D Quadrature Relations

The integration and quadrature relations in one dimension are denoted by

$$I^{(1)} f = \int_{\Gamma_1} f(q) \rho_Q(q) dq \approx \sum_{r=1}^R f(q^r) w^r = \mathcal{Q}^{(1)} f,$$

where  $\Gamma_1 \subset \mathbb{R}$ .

---

<sup>8</sup>The restrictions associated with the assumption of mutually independent parameters and representations for the joint density  $\rho_Q(q)$  are discussed in Section 5.2 and Remark 9.4.

### Gaussian Quadrature Techniques

As detailed in [72, 225, 237, 240], Gaussian quadrature techniques constructed using orthogonal polynomials with densities  $\rho_Q(q)$  yield optimal accuracy for broad classes of functions. Because these polynomials are precisely the Hermite polynomials used to represent normal densities on  $(-\infty, \infty)$  and Legendre polynomials used for uniform densities on  $[-1, 1]$ , the resulting Gaussian quadrature techniques constitute a natural choice for the Galerkin and discrete collocation settings. For uniform and normal densities, this yields the Gauss–Legendre and Gauss–Hermite relations

$$\begin{aligned} I^{(1)} f &= \frac{1}{2} \int_{-1}^1 f(q) dq \approx \frac{1}{2} \sum_{r=1}^R f(q^r) w^r, \\ I^{(1)} f &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} f(q) e^{-q^2/2} dq \approx \sum_{r=1}^R f(q^r) w^r, \end{aligned}$$

where the nodes and weights for the Gauss–Legendre rule are summarized in Table 11.1. As noted in Example 10.2, most tables specify the points and weights for Gauss–Hermite quadrature based on the “physicist” Hermite polynomials which are orthogonal with respect to the density  $\rho_Q(q) = e^{-q^2}$ . These reported nodes and weights can be transformed for the density  $\rho_Q(q) = \frac{1}{\sqrt{2\pi}} e^{-q^2/2}$  using the relation (10.16).

Gauss–Legendre quadrature relations integrate polynomials of order  $2R - 1$  exactly. However, Table 11.1 illustrates that nodes are not nested in the sense that the set of nodes at level  $R$  contains those at level  $R - 1$ . This motivates the construction of nested quadrature techniques.

### Nested Quadrature Techniques

We consider nested nodal structures where the number of nodes or quadrature points at each level is denoted by  $R^\ell$ , and  $q_\ell^r$  and  $w_\ell^r$  denote the points and weights

| $r$ | Nodes $q^r$  | Weights $w^r$  |
|-----|--|--|
| 1   | 0  | 2  |
| 2   | $\pm \frac{1}{\sqrt{3}}$   | 1  |
| 3   | 0<br>$\pm \sqrt{\frac{3}{5}}$  | $\frac{8}{9}$<br>$\frac{5}{9}$                               |
| 4   | $\pm \frac{\sqrt{15+2\sqrt{30}}}{\sqrt{35}}$<br>$\pm \frac{\sqrt{15-2\sqrt{30}}}{\sqrt{35}}$ | $\frac{49}{6(18+\sqrt{30})}$<br>$\frac{49}{6(18-\sqrt{30})}$ |

**Table 11.1.** Nodes and weights for Gauss–Legendre quadrature on  $[-1, 1]$ .

at each level. The 1-D quadrature rule is thus

$$\mathcal{Q}_\ell^{(1)} f = \sum_{r=1}^{R_\ell} f(q_\ell^r) w_\ell^r.$$

To simplify the discussion, we consider the case of a uniform density on the interval  $[0, 1]$ . For a random variable  $q$  defined on  $(a, b)$  with the density  $\rho_Q(q)$ , one can map the parameter space to the cdf, which is defined on  $[0, 1]$ , using the bijective mapping

$$x(q) = F(q) = \int_a^q \rho_Q(\zeta) d\zeta \in [0, 1] \quad (11.2)$$

for  $q \in (a, b)$ . For  $f(q)$ , one then has the transformed integral and quadrature relations

$$I^{(1)} f = \int_0^1 f(F^{-1}(x)) dx$$

and

$$\mathcal{Q}_\ell^{(1)} f = \sum_{r=1}^{R_\ell} f(q_\ell^r) w_\ell^r,$$

where  $q_\ell^r = F^{-1}(x_\ell^r)$  and  $w_\ell^r$  is the weight on  $[0, 1]$  for the  $\ell^{th}$ -level. Without loss of generality, we thus consider  $f$  to be posed on  $[0, 1]$  with uniform measure.

The simplest nested quadrature rule is the composite trapezoid formula

$$\mathcal{Q}_\ell^{(1)} f = \frac{h_\ell}{2} \left[ f(0) + f(1) + 2 \sum_{r=1}^{R_\ell-2} f(q_\ell^r) \right], \quad (11.3)$$

where

$$h_\ell = \frac{1}{2^{\ell-1}} \quad , \quad R_\ell = 2^{\ell-1} + 1 \quad , \quad q_\ell^r = r h_\ell = \frac{r}{2^{\ell-1}} \text{ for } r = 0, \dots, R_\ell. \quad (11.4)$$

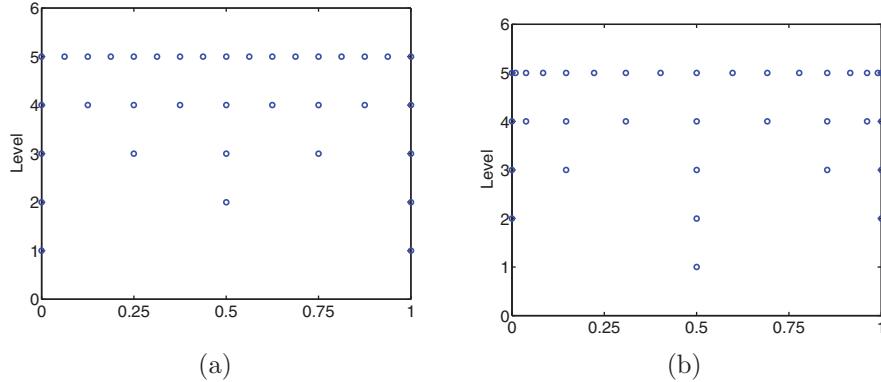
The weights are thus  $[\frac{h_\ell}{2}, h_\ell, \dots, h_\ell, \frac{h_\ell}{2}]$ .

The nodes for levels  $\ell = 1$  through  $\ell = 5$  are shown in Figure 11.1(a) to illustrate the nested structure. Whereas the use of equally spaced nodes simplifies the grid construction, it is illustrated in Example 11.6 that it can produce spurious oscillations when employed for interpolation. This is avoided by the Clenshaw–Curtis and Fejér rules, which we summarize next.

The Clenshaw–Curtis and Fejér nodes are the extrema of Chebyshev polynomials which are typically defined on the interval  $[-1, 1]$ . When mapped to  $[0, 1]$ , the Clenshaw–Curtis nodes are given by

$$q_\ell^r = \frac{1}{2} \left[ 1 - \cos \frac{\pi(r-1)}{R_\ell-1} \right] , \quad r = 1, \dots, R_\ell, \quad (11.5)$$

where  $R_1 = 1$  and  $R_\ell = 2^{\ell-1} + 1$  for  $\ell > 1$ . As illustrated in Figure 11.1(b), the resulting nodes are unequally spaced and cluster at the endpoints of the interval.



**Figure 11.1.** Quadrature points for levels  $\ell = 1 - 5$  for (a) the trapezoid formula (11.3) and (b) Clenshaw–Curtis points (11.5).

The boundary nodes are neglected in the Fejér rule, which is advantageous if mapping random variables with unbounded domains, e.g., normal random variables. Details regarding the construction of nodes and weights at a given level are provided in [148]. We refer the reader to [249] for an overview of Clenshaw–Curtis and Fejér quadrature rules and details illustrating their performance compared to Gauss quadrature.

The 1-D trapezoid and Clenshaw–Curtis rules both satisfy the error bound

$$\left| \mathcal{Q}_1^{(1)} f - I^{(1)} f \right| = \mathcal{O}(R_\ell^{-\alpha}) \quad (11.6)$$

for  $f \in C^\alpha[0, 1]$ . This forms the basis for the tensor product and sparse grid error formulas.

### Tensor Product Formulation

We now address the approximation of integrals

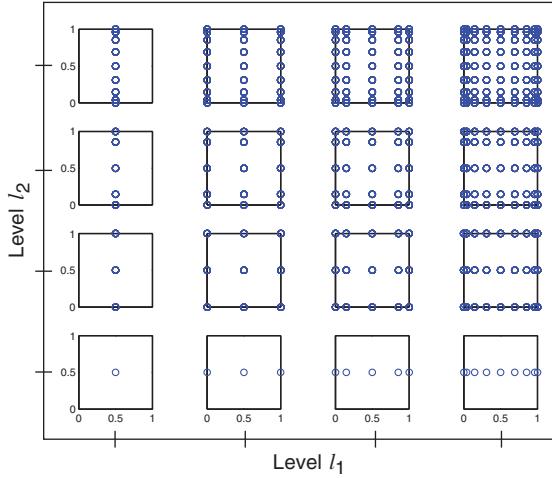
$$I^{(p)} f = \int_{\Gamma} f(q) \rho_Q(q) dq$$

on the  $p$ -dimensional hypercube  $\Gamma = [0, 1]^p$ . If we let  $\mathcal{Q}_{\ell_i}$  denote the quadrature rule for the  $i^{th}$  integration direction, then a tensor product rule is defined by

$$\begin{aligned} \mathcal{Q}_\ell^{(p)} f &= \left( \mathcal{Q}_{\ell_1}^{(1)} \otimes \cdots \otimes \mathcal{Q}_{\ell_p}^{(1)} \right) f \\ &\equiv \sum_{r_1=1}^{R_{\ell_1}} \cdots \sum_{r_p=1}^{R_{\ell_p}} f(q_1^{r_1}, \dots, q_p^{r_p}) w_{\ell_1}^{r_1} \cdots w_{\ell_p}^{r_p}. \end{aligned} \quad (11.7)$$

The total number of quadrature points is

$$R = \prod_{i=1}^p R_{\ell_i}$$



**Figure 11.2.** Tensor product of Clenshaw–Curtis points for  $p = 2$  and four levels of  $\ell_1$  and  $\ell_2$ .

or

$$R = (R_\ell)^p$$

if the same number of points is used in each direction, as illustrated for tensored Clenshaw–Curtis grids in Figure 11.2 for  $p = 2$ . This exponential growth in the number of required nodes precludes the use of tensored quadrature relations for all but low-dimensional problems—e.g.,  $p$  ranging from 5 to 8.

The curse of dimensionality is also manifested as a significant reduction in the convergence rate of error bounds as the dimension grows. For  $R = (R_\ell)^p$  quadrature points, the error for the  $p$ -dimensional Clenshaw–Curtis quadrature rule satisfies

$$\left| I^{(p)} f - \mathcal{Q}_\ell^{(p)} f \right| = \mathcal{O}(R_\ell^{-\alpha/p}) \quad (11.8)$$

for functions  $f$  in the space

$$C^\alpha([0, 1]^p) = \left\{ f : [0, 1]^p \rightarrow \mathbb{R} \mid \max_{|\mathbf{k}'| \leq \alpha} \left\| \frac{\partial^{|\mathbf{k}'|} f}{\partial q_1^{k_1} \cdots \partial q_p^{k_p}} \right\|_\infty < \infty \right\} \quad (11.9)$$

with bounded derivatives up to order  $\alpha$ . Here  $\mathbf{k}' = (k_1, \dots, k_p) \in \mathbb{N}^p$  is a multi-index with  $|\mathbf{k}'| = \sum_{i=1}^p k_i$ . The exponential growth in the required number of quadrature points and the diminished convergence rate motivates the use of sparse grid quadrature techniques for problems with moderate dimensionality.

### 11.1.3 Sparse Grid Construction

To motivate ideas underlying sparse grid construction, we consider products of monomials, as illustrated in Figure 11.3. The accuracy of quadrature rules is typically quantified in terms of the degree  $R$  that can be integrated exactly. To specify

| $R$ |       |        |          |        |       |
|-----|-------|--------|----------|--------|-------|
| 0   |       |        |          | 1      |       |
| 1   | $x$   |        |          | $y$    |       |
| 2   | $x^2$ | $xy$   |          | $y^2$  |       |
| 3   | $x^3$ | $x^2y$ | $xy^2$   | $y^3$  |       |
| 4   | $x^4$ | $x^3y$ | $x^2y^2$ | $xy^3$ | $y^4$ |

**Figure 11.3.** Products of monomials up to degree 4.

monomials of degree four, one need only consider the five terms listed at  $R = 4$ . This is in contrast to a tensor product which will have 25 terms that include monomials such as  $x^4y$  that are higher accuracy than necessary. The goal with sparse grid methods, which were originally proposed by Smolyak in [226], is to construct grids and weights that yield the same accuracy as tensor products but with a significantly reduced number of required points.

There are two formulations for sparse grids that differ slightly in the hierarchy of employed spaces. We describe the form detailed in [93] which is formulated in terms of the levels  $\ell$  and summarize the second formulation in Remark 11.3. In that remark, we note that the two typically yield the same grids and weights and hence have the same accuracy and efficiency.

To construct a sparse grid quadrature rule, we start with the 1-D relation

$$\mathcal{Q}_\ell^{(1)} f = \sum_{r=1}^{R_\ell} f(q_\ell^r) w_\ell^r, \quad (11.10)$$

where  $q_\ell^r, w_\ell^r$  are the nodes and weights for the  $\ell^{th}$  nested level, and define the difference relations

$$\Delta_\ell^{(1)} f = \left( \mathcal{Q}_\ell^{(1)} - \mathcal{Q}_{\ell-1}^{(1)} \right) f$$

with  $\mathcal{Q}_0^{(1)} f \equiv 0$ . We note that  $\Delta_\ell^{(1)} f$  is also a quadrature formula in which the nodes are the same as those for  $\mathcal{Q}_\ell^{(1)} f$  and the weights are the difference between those for the  $\ell$  and  $\ell - 1$  levels. The 1-D set of nodal points is denoted by

$$\Theta_\ell^{(1)} = \left\{ q_\ell^1, \dots, q_\ell^{R_\ell} \right\}. \quad (11.11)$$

**Example 11.1.** Consider the composite trapezoid formula defined by (11.3) and (11.4). For  $\ell = 2$ , the points and weights are  $\Theta_2^{(1)} = \{0, \frac{1}{2}, 1\}$  and  $[\frac{1}{4}, \frac{1}{2}, \frac{1}{4}]$ , whereas they are  $\Theta_1^{(1)} = \{0, 1\}$  and  $[\frac{1}{2}, \frac{1}{2}]$  for  $\ell = 1$ . Thus

$$\Delta_2^{(1)} f = -\frac{1}{4}f(0) + \frac{1}{2}f(1/2) - \frac{1}{4}f(1).$$

Hence the nodes  $\Theta_2^{(1)} = \{0, \frac{1}{2}, 1\}$  for the difference rule are the same as those for  $\mathcal{Q}_2^{(1)}$ , but the weights  $[-\frac{1}{4}, \frac{1}{2}, -\frac{1}{4}]$  reflect the difference  $\mathcal{Q}_2^{(1)} - \mathcal{Q}_1^{(1)}$ . We note

that unlike Newton–Cotes and Gauss quadrature formulas, negative weights are permissible for sparse grid quadrature relations.

The sparse quadrature formula at level  $\ell$  is then given by

$$\mathcal{Q}_\ell^{(p)} f = \sum_{|\ell'| \leq \ell+p-1} \left( \Delta_{\ell_1}^{(1)} \otimes \cdots \otimes \Delta_{\ell_p}^{(1)} \right) f, \quad (11.12)$$

where  $\ell' = (\ell_1, \dots, \ell_p) \in \mathbb{N}^p$  is a multi-index with  $|\ell'| = \sum_{i=1}^p \ell_i$ . This is in contrast to the tensor product formulation (11.7), which can be expressed as

$$\mathcal{Q}_\ell^{(p)} f = \sum_{\max \ell' \leq \ell} \left( \Delta_{\ell_1}^{(1)} \otimes \cdots \otimes \Delta_{\ell_p}^{(1)} \right) f, \quad (11.13)$$

where  $\max \ell' \equiv \max\{\ell_1, \dots, \ell_p\}$ . The notation  $|\ell'|_1$  and  $|\ell'|_\infty$  is often used to express the limits used in (11.12) and (11.13). These formulas can both be formulated as

$$\mathcal{Q}_\ell^{(p)} f = \sum_{\ell' \in \mathbb{I}(\ell)} \left( \Delta_{\ell_1}^{(1)} \otimes \cdots \otimes \Delta_{\ell_p}^{(1)} \right) f, \quad (11.14)$$

where  $\mathbb{I}(\ell)$  is a multi-index set that is a function of the level  $\ell$ . For the sparse and full grid formulas (11.12) and (11.13), the respective multi-index sets are

$$\mathbb{I}(\ell) = \left\{ \ell' \in \mathbb{N}^p \mid \sum_{i=1}^p \ell_i \leq \ell + p - 1 \right\}$$

and

$$\mathbb{I}(\ell) = \left\{ \ell' \in \mathbb{N}^p \mid \ell_i \leq \ell, i = 1, \dots, p \right\}.$$

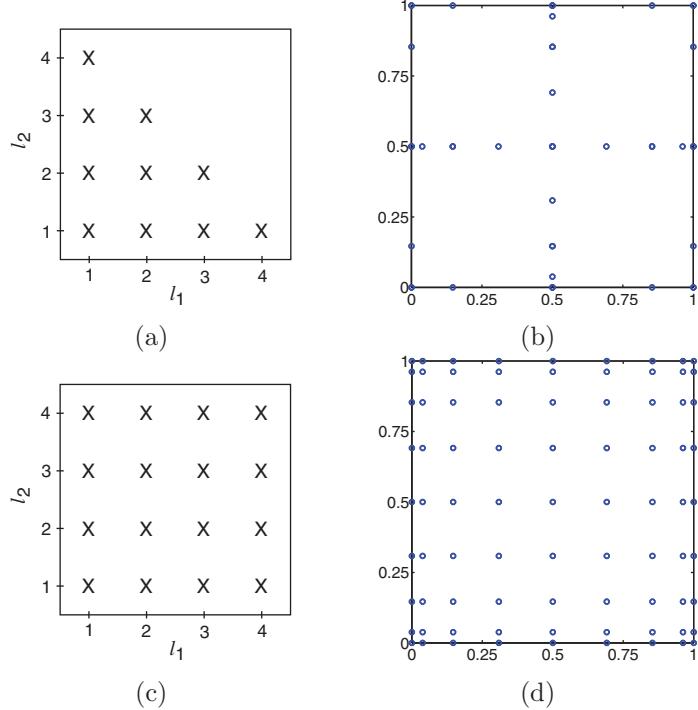
The nodal set for the sparse grid is

$$\Theta_\ell^{(p)} = \bigcup_{|\ell'| \leq \ell+p-1} \Theta_{\ell_1}^{(1)} \times \cdots \times \Theta_{\ell_p}^{(1)}. \quad (11.15)$$

**Example 11.2.** Consider the Clenshaw–Curtis rule with nodes specified by (11.5) on the interval  $[0, 1]$  so that  $\Theta_1^{(1)} = \{\frac{1}{2}\}$  and  $\Theta_2^{(1)} = \{0, \frac{1}{2}, 1\}$ . For  $p = 2$ , the multi-index is  $\ell' = (\ell_1, \ell_2)$ , so  $|\ell'| = \ell_1 + \ell_2$ . For  $\ell = 4$ , the sparse grid nodal set is

$$\begin{aligned} \Theta_4^{(2)} &= \left( \Theta_1^{(1)} \times \Theta_1^{(1)} \right) \quad (\ell_1 = 1, \ell_2 = 1) \\ &\cup \left( \Theta_1^{(1)} \times \Theta_2^{(1)} \right) \cup \left( \Theta_2^{(1)} \times \Theta_1^{(1)} \right) \\ &\cup \left( \Theta_1^{(1)} \times \Theta_3^{(1)} \right) \cup \left( \Theta_2^{(1)} \times \Theta_2^{(1)} \right) \cup \left( \Theta_1^{(1)} \times \Theta_3^{(1)} \right) \\ &\cup \left( \Theta_1^{(1)} \times \Theta_4^{(1)} \right) \cup \left( \Theta_2^{(1)} \times \Theta_3^{(1)} \right) \cup \left( \Theta_3^{(1)} \times \Theta_2^{(1)} \right) \cup \left( \Theta_4^{(1)} \times \Theta_1^{(1)} \right), \end{aligned}$$

which utilizes the indices depicted in Figure 11.4(a). Combination of the points shown in Figure 11.2 using this index pattern yields the sparse Clenshaw–Curtis



**Figure 11.4.** (a) Indices for the union of the 1-D grids  $\Theta_{\ell_1}^{(1)} \times \Theta_{\ell_2}^{(1)}$  and (b) the sparse Clenshaw-Curtis grid resulting from the application of these formulas to the points in Figure 11.2. (c) Indices for the tensor product formulation (11.13) and (d) the full tensor product grid.

grid shown in Figure 11.4(b). The indices for the full tensor product formulation (11.13) are depicted in Figure 11.4(c), which yields the full tensor product grid shown in Figure 11.4(d). We note that for this case, the sparse grid has 29 points, whereas there are  $9^2 = 81$  points in the full grid.

**Remark 11.3.** The sparse grid formula (11.12) and nodal set are often expressed using equivalent formulations or formulations based on slightly different hierarchies of spaces. An equivalent formulation for the sparse grid quadrature formula (11.12), based on  $\mathcal{Q}_\ell^{(1)}$  rather than  $\Delta_\ell^{(1)}$ , is

$$\mathcal{Q}_\ell^{(p)} f = \sum_{\ell \leq |\ell'| \leq \ell+p-1} (-1)^{\ell+p-|\ell'|-1} \cdot \binom{p-1}{|\ell'|-1} \cdot \left( \mathcal{Q}_{\ell_1}^{(1)} \otimes \cdots \otimes \mathcal{Q}_{\ell_p}^{(1)} \right) f. \quad (11.16)$$

Alternatively, many authors define the sparse quadrature rule as

$$\mathcal{A}(q, p)f = \sum_{|\mathbf{k}'| \leq q} \left( \Delta_{k_1}^{(1)} \otimes \cdots \otimes \Delta_{k_p}^{(1)} \right) f \quad (11.17)$$

or, equivalently,

$$\mathcal{A}(q, p)f = \sum_{q-p+1 \leq |\mathbf{k}'| \leq q} (-1)^{q-|\mathbf{k}'|} \cdot \binom{p-1}{q-|\mathbf{k}'|} \cdot \left( \mathcal{Q}_{\ell_1}^{(1)} \otimes \cdots \otimes \mathcal{Q}_{\ell_p}^{(1)} \right) f \quad (11.18)$$

for integers  $q \geq p$ . The sparse grid is defined as

$$\mathcal{H}(q, p) = \bigcup_{q-p+1 \leq |\mathbf{k}'| \leq q} \Theta_{k_1}^{(1)} \times \cdots \times \Theta_{k_p}^{(1)}. \quad (11.19)$$

We first note that if  $q = \ell + p - 1$ , then the two formulations yield the same nodes and have the same accuracy. In this case, the formulation (11.17)–(11.19) utilizes fewer low-level hierarchies and, at first glance, would appear to be more efficient to implement. As detailed in [93], however, one typically omits previously employed points in the nested 1-D relations which will render the two formulations equally expensive to implement. It is established in [47] that the relation  $|\ell'| \leq \ell + p - 1$  optimizes a cost-benefit ratio which motivates the formulations (11.12), (11.11), and (11.16) based on the number of levels  $\ell$ . Either formulation can be used for the Galerkin, collocation, and discrete projection methods used for uncertainty propagation.

We compile in Table 11.2 the number of points in full and sparse grids for various levels and dimensions. This illustrates the immense computational savings that can be realized with sparse grids and demonstrates how they significantly increase the dimensionality of integrals that can be accurately approximated.

Furthermore, it is shown in [185] that if we let  $\mathcal{R}$  denote the number of sparse grid nodes used by  $\mathcal{A}(q, p)$ , then the quadrature error satisfies

$$\|I^{(p)}f - \mathcal{A}(q, p)f\| = \mathcal{O}\left(\mathcal{R}^{-\alpha} \log(\mathcal{R})^{(p-1)(\alpha+1)}\right) \quad (11.20)$$

| $p$ | $R_\ell$ | Sparse Grid $\mathcal{R}$ | Tensored Grid $R = (R_\ell)^p$ |
|-----|----------|---------------------------|--------------------------------|
| 2   | 5        | 13                        | 25                             |
|     | 9        | 29                        | 81                             |
| 5   | 5        | 61                        | 3125                           |
|     | 9        | 241                       | 59,049                         |
| 10  | 5        | 221                       | 9,765,625                      |
|     | 9        | 1581                      | $> 3 \times 10^9$              |
| 50  | 5        | 5101                      | $> 8 \times 10^{34}$           |
|     | 9        | 171,901                   | $> 5 \times 10^{47}$           |
| 100 | 5        | 20,201                    | $> 7 \times 10^{69}$           |
|     | 9        | 1,353,801                 | $> 2 \times 10^{95}$           |

**Table 11.2.** Number of sparse grid and tensored Clenshaw–Curtis quadrature points with  $R_\ell = 5$  and 9 nodes in each direction for dimensions  $p = 2$  to 100.

for  $f \in C^\alpha([0, 1]^p)$  defined in (11.9). Comparison with the tensor product error (11.8) illustrates that the reduced number of sparse grid nodes  $\mathcal{R}$  versus  $R = (R_\ell)^p$  helps push back the onset of the curse of dimensionality. However, for sufficiently large  $p$ , the term  $\mathcal{R}^{-\alpha}$  ceases to dominate and the convergence rate is no longer competitive with the rate  $\mathcal{O}(R^{-1/2})$  exhibited by Monte Carlo methods. For certain problems, the efficiency of sparse grid quadrature techniques can be extended by adaptive grid construction, which accommodates anisotropic behavior often associated with various components  $q_i$  of the parameter vector.

### Adaptive Sparse Grids

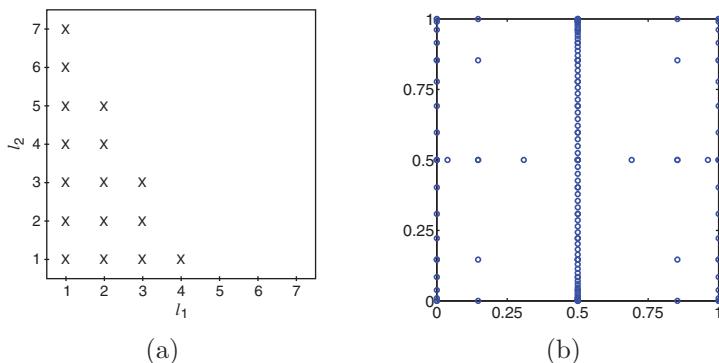
In the quadrature formula (11.14), the multi-index set  $\mathbb{I}(\ell)$  defines the structure of the grid. To allow grid adaptation, one can modify  $\mathbb{I}(\ell)$  to weight various dimensions in accordance with their influence—e.g., the influence of random variables in Karhunen–Loëve expansions for certain random fields diminishes as the index increases. To construct an anisotropic grid with variable accuracy for certain random variables, one can employ the multi-index set

$$\mathbb{I}(\ell) = \left\{ \ell' \in \mathbb{N}^p \mid \ell' \cdot \mathbf{a} = \sum_{i=1}^p a_i \ell_i \leq \ell + p - 1 \right\}, \quad (11.21)$$

where  $\mathbf{a} \in \mathbb{R}_+^p$  is a vector of weights. A sparse grid created using a weighting vector that varies the influence of  $\ell_1$  is illustrated in Figure 11.5.

This strategy has two limitations: it is often difficult to prescribe  $\mathbf{a}$  based on a priori knowledge of the model, and the flexibility of the method to yield general grid coarsening strategies is limited by the prespecified structure of the multi-index set  $\mathbb{I}$ . These issues are addressed by strategies that provide adaptation through sequential construction of the multi-index set. We refer the reader to [148] for details and analysis regarding more general adaptation strategies.

**Remark 11.4.** The adaptation discussed here pertains to anisotropies in the influence of various random parameters. This differs from  $h$  and  $p$  adaptation used to refine the accuracy of finite element or Galerkin approximations.



**Figure 11.5.** (a) Indices and (b) the resulting adaptive sparse grid.

**Remark 11.5.** The tensor product and sparse grid quadrature techniques discussed here are distinct from cubature rules, which are specifically optimized for multidimensional integration and hence are not based on combinations of 1-D quadrature rules [239]. Further attributes of cubature rules in the context of stochastic spectral methods are discussed in [4].

## 11.2 Interpolating Polynomials for Collocation

The collocation techniques described in Section 10.2 rely on the construction of interpolating polynomials  $L_k$  that satisfy the property  $L_k(q^m) = \delta_{km}$  at the points  $q^m$ . Although the collocation algorithms were constructed in the context of Galerkin methods, the convergence theory relies on properties of the interpolating polynomials and the choice of collocation points rather than the theory of  $L^2$  projections. We note that the choice of points is fundamental to interpolation theory and poor choices can significantly degrade the method's accuracy.

The objective for interpolation can be broadly summarized as follows. We assume that we have a process  $u(q)$  that is a function of  $p$  inputs  $q = [q_1, \dots, q_p]$ . In the case of experiments, the true process is typically unknown and we simply have a set of  $M$  measurements or realizations  $u^m = u(q^m)$ ,  $m = 1, \dots, M$ , corresponding to  $M$  values of the input vector. For mathematical models, the construction of solutions  $u(q)$  is often computationally expensive, so we wish to infer the behavior of  $u$  for various parameter values  $q$  based on a computationally tractable number  $M$  of computed solutions  $u(q^m)$ . In both cases, we seek polynomials  $u^M(q) \in \mathbb{P}_{M-1}$  of degree  $M$  that satisfy

$$u^M(q^m) = u^m = u(q^m), \quad m = 1, \dots, M, \quad (11.22)$$

at the specified set of interpolation points and accurately approximate  $u(q)$  for remaining parameter values  $q \in \Gamma$ . Since the determination of values  $u^m = u(q^m)$  is often computationally or experimentally expensive, we seek polynomials and interpolation points that optimize the accuracy of the approximation  $u^M(q) \approx u(q)$ , where the dimensionality  $p$  of  $q$  is often relatively large, with the smallest possible number of interpolation points.

In Section 11.2.1, we describe the construction of 1-D Lagrange polynomials and indicate one choice for specifying nonuniform interpolation points. We summarize tensor and sparse grid techniques for  $p$ -dimensional interpolation in Sections 11.2.2 and 11.2.3.

### 11.2.1 1-D Interpolation

We consider  $M_1$  pairs  $(q^m, u^m)$  where  $u^m = u(q^m)$  are solutions computed with realizations  $q^m \in \mathbb{R}^1$ . Throughout this discussion, we assume that the points  $q^m$  are distinct. To approximate  $u(q)$ , we seek polynomials  $u^{M_1}(q) \in \mathbb{P}_{M_1-1}$  that satisfy

$$u^{M_1}(q^m) = u^m, \quad m = 1, \dots, M_1. \quad (11.23)$$

Such polynomials can be uniquely specified as

$$u^{M_1}(q) = \sum_{m=1}^{M_1} u^m L_m(q), \quad (11.24)$$

where  $L_m(q)$  are Lagrange interpolating polynomials defined by

$$\begin{aligned} L_m(q) &= \prod_{\substack{j=0 \\ j \neq m}}^{M_1} \frac{q - q^j}{q^m - q^j} \\ &= \frac{(q - q^1) \cdots (q - q^{m-1})(q - q^{m+1}) \cdots (q - q^{M_1})}{(q^m - q^1) \cdots (q^m - q^{m-1})(q^m - q^{m+1}) \cdots (q^m - q^{M_1})} \end{aligned} \quad (11.25)$$

for  $m = 1, \dots, M_1$ . By construction, the Lagrange polynomials satisfy

$$L_m(q^n) = \delta_{mn}, \quad 1 \leq m, n \leq M_1, \quad (11.26)$$

which ensures that  $u^{M_1}(q^m) = u^m$  for all points. This approach has the advantage that representations accommodating new points  $q^{M_1+k}$  can be constructed by simply appending terms to (11.25). We denote the 1-D interpolating polynomial based on  $M_1$  distinct points by

$$\mathcal{I}^{(1)} u(q) = u^{M_1}(q). \quad (11.27)$$

The construction of  $M_1 - 1$  degree polynomials that interpolate for  $M_1$  pairs only requires that the interpolation points be distinct. However, the next example illustrates that the accuracy of the interpolating polynomial can be highly dependent on the choice of interpolation points.

**Example 11.6.** Consider the Runge function

$$f(q) = \frac{1}{1 + 25q^2}$$

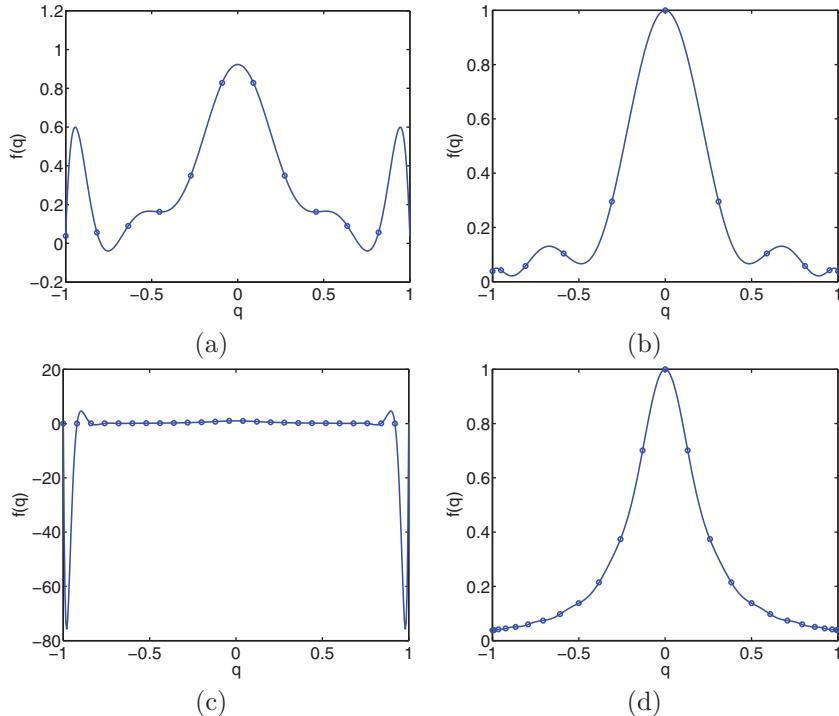
on the interval  $[-1, 1]$ . This classical example illustrates spurious oscillations, which are one of the difficulties associated with interpolation using high degree polynomials. This is often termed Runge's phenomenon, and it can be especially pronounced when using uniformly distributed points; for example,

$$q^j = -1 + (j - 1) \frac{2}{M_1}, \quad j = 1, \dots, M_1 + 1,$$

as illustrated in Figure 11.6(a) and (c). Details regarding the theoretical basis for this phenomenon can be found in [70, 248].

The use of nonuniformly spaced collocation points that cluster near the ends of the interval can help to mitigate this problem. The abscissae

$$q^j = -\cos \frac{\pi(j - 1)}{M_1 - 1}, \quad j = 1, \dots, M_1, \quad (11.28)$$



**Figure 11.6.** Interpolation using  $M_1 = 11$  and  $M_1 = 25$  uniformly spaced points (a), (c) and Chebyshev points (b), (d).

of the Chebyshev polynomials constitute one choice that accomplishes this goal. Figure 11.6(b) and (d) illustrate that whereas some oscillations occur using  $M_1 = 11$  Chebyshev points, the interpolation using  $M_1 = 25$  points is devoid of oscillations.

An alternative strategy, to improve conditioning and improve spurious oscillations, even using uniform grids, is to employ lower-order or piecewise polynomials (splines). The improved stability often outweighs the corresponding reduction in accuracy.

The 1-D interpolation error satisfies the bound

$$\|u - \mathcal{I}^{(1)}u\|_\infty \leq E_{M_1-1}(u) \cdot (1 + \Lambda_{M_1}), \quad (11.29)$$

where  $E_{M_1-1}$  and  $\Lambda_{M_1}$  respectively denote the approximation error and Lebesgue constant. It is shown in [29] that

$$\Lambda_{M_1} \leq \frac{2}{\pi} \log(M_1 - 1) + 1, \quad M_1 \geq 2,$$

for the nodes (11.28). This yields the bound

$$\|u - \mathcal{I}^{(1)}u\|_\infty \leq c_\alpha \log(M_1) M_1^{-\alpha} \quad (11.30)$$

for  $u \in C^\alpha[-1, 1]$ . This is analogous to the 1-D quadrature bound (11.6).

### 11.2.2 Multidimensional Interpolation Using Tensor Products

To construct interpolation formulas for parameters  $Q = [Q_1, \dots, Q_p]$ , we employ tensor products of the 1-D interpolating polynomials and collocation points. This is analogous to the approach used in Section 11.1.2 for multidimensional quadrature, and additional details regarding the general approach can be found in that section. We again assume that the mapping (11.2) can be used to transform parameter components defined on  $(a, b)$  to the interval  $[0, 1]$ , and we consider interpolation on the hypercube  $[0, 1]^p$ .

In a manner analogous to (11.5), we define the interpolation points at the  $\ell^{\text{th}}$  nested level to be

$$q_\ell^m = \frac{1}{2} \left[ 1 - \cos \frac{\pi(m-1)}{M_\ell - 1} \right], \quad m = 1, \dots, M_\ell, \quad (11.31)$$

where  $M_1 = 1$  and  $M_\ell = 2^{\ell-1} + 1$ ,  $\ell > 1$ , for the Clenshaw–Curtis points. The choice (11.31) yields nested points, which facilitates implementation and minimizes oscillations in the manner illustrated in Example 11.6. The 1-D grid of interpolation points is denoted by

$$\Theta_\ell^{(1)} = \left\{ q_\ell^1, \dots, q_\ell^{M_\ell} \right\}. \quad (11.32)$$

The multidimensional, tensor product, interpolation grid is

$$\Theta_\ell^{(p)} = \bigcup_{\max \ell' \leq \ell} \Theta_{\ell_1}^{(1)} \times \cdots \times \Theta_{\ell_p}^{(1)}, \quad (11.33)$$

which has

$$M = (M_\ell)^p \quad (11.34)$$

nodes if the same number of points are used in each direction. The exponential growth in the number of interpolation points is the first manifestation of the curse of dimensionality.

A tensor product of the 1-D relations (11.27) yields the  $p$ -dimensional interpolation formula

$$\begin{aligned} \mathcal{I}^{(p)} u(q_1, \dots, q_p) &= \left( \mathcal{I}^{(1)} \otimes \cdots \otimes \mathcal{I}^{(1)} \right) u \\ &\equiv \sum_{m_1=1}^{M_{\ell_1}} \cdots \sum_{m_p=1}^{M_{\ell_p}} u(q_1^{m_1}, \dots, q_p^{m_p}) L_{m_1}(q_1) \cdots L_{m_p}(q_p), \end{aligned} \quad (11.35)$$

which requires that the model be evaluated at the  $M = M_{\ell_1} \cdots M_{\ell_p}$  interpolation points. This will be prohibitive for all but low dimensions—e.g., this requires  $p \leq 5$  to 8.

For  $M = (M_\ell)^p$  distinct collocation points, the interpolation error satisfies

$$\|u - \mathcal{I}^{(p)} u\|_\infty = \mathcal{O}(M^{-\alpha/p}) \quad (11.36)$$

for  $u \in C^\alpha([0, 1]^p)$  defined in (11.9). We note that this is an algebraic rather than exponential convergence rate. Derivations of these bounds in the context of stochastic collocation are provided in [18].

### 11.2.3 Sparse Grid Interpolation

The development of sparse grid interpolation techniques is analogous to the sparse grid quadrature techniques summarized in Section 11.1.3, and sparse grids constructed using Clenshaw–Curtis, Fejér, or Gauss points can serve as either quadrature or interpolation points.

We let

$$\mathcal{I}_\ell^{(1)} u = \sum_{m=1}^{M_\ell} u^m L_m(q)$$

denote the 1-D operator that interpolates at the  $M_\ell$  points  $\Theta_\ell^{(1)} = \{q_\ell^1, \dots, q_\ell^{M_\ell}\}$  in the  $\ell^{th}$  nested level, where, for specificity, we define  $q_\ell^m$  to be zeros of the Chebyshev polynomials (11.32). We also define the difference relations

$$\Delta_\ell^{(1)} u = (\mathcal{I}_\ell^{(1)} - \mathcal{I}_{\ell-1}^{(1)}) u \quad , \quad \mathcal{I}_0^{(1)} u = 0,$$

which are also interpolation formulas. The sparse grid interpolation formula at level  $\ell$  is then

$$\mathcal{I}_\ell^{(p)} u = \sum_{|\ell'| \leq \ell+p-1} (\Delta_{\ell_1}^{(1)} \otimes \cdots \otimes \Delta_{\ell_p}^{(1)}) u.$$

This is equivalent to the formulation

$$\mathcal{I}_\ell^{(p)} u = \sum_{\ell \leq |\ell'| \leq \ell+p-1} (-1)^{\ell+p-|\ell'|-1} \cdot \binom{p-1}{|\ell'|-l} \cdot (\mathcal{I}_{\ell_1}^{(1)} \otimes \cdots \otimes \mathcal{I}_{\ell_p}^{(1)}) u.$$

The nodal set for the sparse grid is

$$\Theta_\ell^{(p)} = \bigcup_{|\ell'| \leq \ell+p-1} \Theta_{\ell_1}^{(1)} \times \cdots \times \Theta_{\ell_p}^{(1)}$$

as compared with the full tensor product interpolation grid (11.33). Further details regarding the full and sparse grid constructs can be found in Section 11.1.3. Adaptive sparse grid interpolation techniques can be constructed in a manner analogous to that outlined for quadrature.

As noted in Remark 11.3, many authors define the sparse interpolation as (11.16) or (11.17) and the sparse grid as (11.19). For implementation of the stochastic collocation method, the two formulations are essentially equivalent.

## 11.3 Sparse Grid Software

The Sparse Grid Interpolation Toolbox provides MATLAB software for initial investigation of sparse grid quadrature and interpolation routines. We caution the reader that the “Clenshaw–Curtis” points designated in this toolbox are equally spaced and are actually Newton–Cotes points. To specify the more standard Clenshaw–Curtis nodes (11.5), one must use the designation “Chebyshev.” Sparse grid capabilities are also available in the Sandia National Laboratories toolbox DAKOTA [4, 5, 71].

## 11.4 Exercises

**Exercise 11.1.** Consider the spring model

$$m \frac{d^2z}{dt^2} + c \frac{dz}{dt} + kz = f_0 \cos(\omega_F t),$$

$$z(0) = z_0, \quad \frac{dz}{dt}(0) = z_1$$

and steady state response

$$y(\omega_F, Q) = \frac{1}{\sqrt{(k - m\omega_F^2)^2 + (c\omega_F)^2}},$$

where  $Q = [m, c, k]$ ; see Examples 9.7 and 10.13. The QoI are the mean and standard deviation for driving frequencies  $\omega_F \in [0, 2.7]$ . Assume that  $Q \sim N(\bar{q}, V)$  with  $\bar{q} = [2.7, 0.24, 8.5]$  and  $V$  given by (9.19). Use the sparse grid techniques of Section 11.1.3 to approximate the integrals in (10.63). Compare the resulting QoI given by (10.64) with those given by (9.20) with  $M = 10^5$  Monte Carlo samples. How does the number  $R$  of sparse quadrature points compare with the  $R = 10^3$  tensored points used in Example 10.13?

## Chapter 12

# Prediction in the Presence of Model Discrepancy

*Essentially, all models are wrong, but some are useful,* George E.P. Box

In Section 7.1, we considered statistical models of the form

$$\begin{aligned}\Upsilon_i &= f(t_i, q) + \delta(t_i) + \varepsilon_i, \\ \Upsilon_i &= f(x_i, q) + \delta(x_i) + \varepsilon_i\end{aligned}$$

for evolution and stationary processes with model responses  $f(t_i, q)$  or  $f(x_i, q)$ . Here  $\Upsilon_i$  is a random variable whose realizations  $v_i$  are measurements from an experiment. Model errors or discrepancies are represented by the terms  $\delta(t_i)$  or  $\delta(x_i)$ , which do not depend on the physical parameters  $q$ , and  $\varepsilon_i$  denotes measurement errors. For the discussion in Chapters 7 and 8, we combined the model and measurement errors into a single random variable  $\varepsilon_i$  which we assumed was iid, and, for some analysis, we required  $\varepsilon_i \sim N(0, \sigma^2)$ . In this chapter, we discuss issues that arise when the assumption of iid combined errors is rendered invalid by structured, correlated, or biased model discrepancy terms  $\delta(t_i)$  or  $\delta(x_i)$ . Whereas there are statistical and mathematical techniques to quantify discrepancies for certain problems, general methods are lacking, due in part to the problem-dependent nature of the issue. The development of robust techniques to quantify model errors for broad classes of problems constitutes an active research area.

The next three examples illustrate differing types of model discrepancy. The ramifications of unaccommodated model errors are discussed in Section 12.1. In Sections 12.2 and 12.3, we discuss techniques to quantify  $\delta(t_i)$  or  $\delta(x_i)$ , and we indicate unresolved research issues and future research directions in Section 12.4.

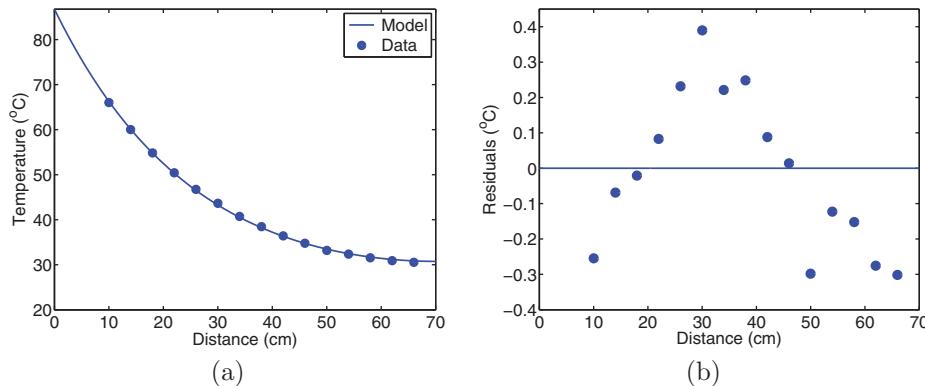
**Example 12.1.** Consider the model

$$\begin{aligned}\frac{d^2T_s}{dx^2} &= \frac{2(a+b)}{ab} \frac{h}{k} [T_s(x) - T_{amb}], \\ \frac{dT_s}{dx}(0) &= \frac{\Phi}{k}, \quad \frac{dT_s}{dx}(L) = \frac{h}{k} [T_{amb} - T_s(L)].\end{aligned}\tag{12.1}$$

As detailed in Example 3.5, this model quantifies the steady state temperature of an uninsulated rod with source heat flux  $\Phi$  at  $x = 0$  and ambient air temperature  $T_{amb}$ . The thermal conductivity  $k$  for the aluminum and copper rods employed in experiments is well documented, so we treat it as known. The parameters are taken to be  $q = [\Phi, h]$ , where  $h$  is the convective heat transfer coefficient.

In Example 7.16, we employed nonlinear least squares to estimate the parameters using the data compiled in Table 3.2 from the aluminum rod. The residuals in Figure 7.3(b) exhibit no discernible pattern, thus motivating the hypothesis that the combined errors  $\varepsilon_i$  are iid.

A similar least squares fit to the copper rod data in Table 3.3 yields the parameter estimates  $\Phi = -9.93$  and  $h = 0.00143$  along with the fit and residuals plotted in Figure 12.1. The residuals indicate that the combined errors in this case are clearly not iid due to unaccommodated model errors  $\delta(x_i)$ . We illustrate in Section 12.2 that algebraic or statistical techniques can be used to quantify this form of model discrepancy.

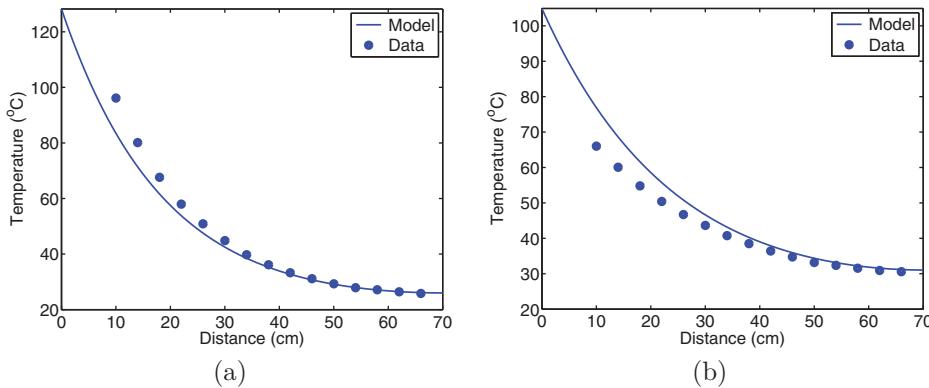


**Figure 12.1.** (a) *Model fit and (b) residuals for the copper rod.*

**Example 12.2.** In Examples 7.16 and 12.1, we respectively obtained the parameter estimates  $\Phi = -18.41$ ,  $h = 0.00191$  and  $\Phi = -9.93$ ,  $h = 0.00143$  using data from aluminum and copper rods. However,  $\Phi$  and  $h$  are material independent, and any material dependency should be accommodated by the thermal conductivity values  $k = 2.37 \frac{W}{cm \cdot C}$  and  $k = 4.01 \frac{W}{cm \cdot C}$  for aluminum and copper.

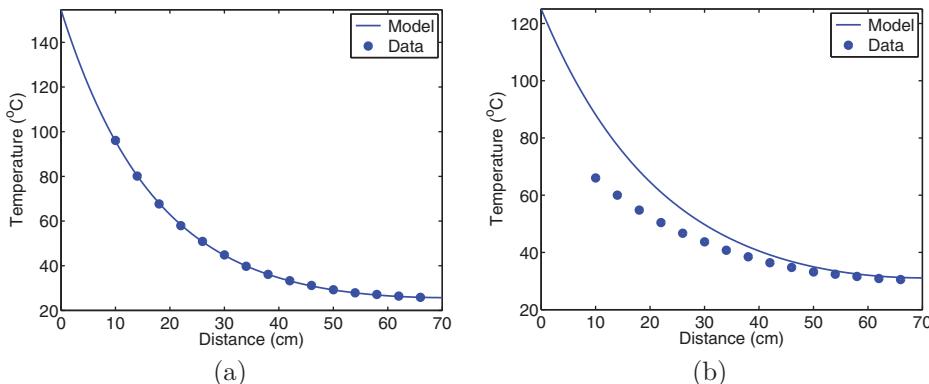
To test the validity of the model to characterize the temperature distribution for both materials using one parameter set, we simultaneously perform a least squares fit to both data sets. The parameter estimates  $\Phi = -13.75$  and  $h = 0.00166$  yield the model fits shown in Figure 12.2. Because the errors produce opposite biases, algebraic or statistical models for  $\delta(x_i)$  will be ineffective in this case.

To illustrate the ramification of unaccommodated discrepancies  $\delta(x_i)$  on the predictive capabilities of the model, we employ the solution with parameters obtained in Example 7.16 through a fit to the aluminum data to predict the temperature distribution for the copper rod. Specifically, we compare the model solution



**Figure 12.2.** Model fit for the (a) aluminum and (b) copper rod using the simultaneously optimized parameters  $\Phi = -13.75$ ,  $h = 0.00166$ .

obtained with  $\Phi = -18.41$ ,  $h = 0.00191$  and the copper conductivity value  $k = 4.01$  with the copper data in Figure 12.3. Due to the model discrepancy, the prediction is highly inaccurate. We illustrate in Example 12.4 that this must be addressed by incorporating physics that is neglected in the model (12.1).



**Figure 12.3.** (a) Model fit to the aluminum data and (b) prediction for the copper rod.

**Example 12.3.** Here we consider the experimental data and model discussed in Example 3.7 for a thin cantilever beam driven by a voltage spike applied to surface-mounted piezoelectric patches. Data consists of temporal measurements  $v_i$  collected using a proximity sensor at  $\bar{x} = 128$  mm. For a beam of length  $L$ , a weak model formulation for the transverse displacements  $w(t, x)$  is provided by the Euler–Bernoulli

equation

$$\int_0^L \left[ \rho(x) \frac{\partial^2 w}{\partial t^2} + \gamma \frac{\partial w}{\partial t} \right] \phi dx + \int_0^L \left[ YI(x) \frac{\partial^2 w}{\partial x^2} + cI(x) \frac{\partial^3 w}{\partial x^2 \partial t} \right] \phi'' dx \\ = k_p V(t) \int_{x_1}^{x_2} \phi'' dx,$$

which holds for all test functions  $\phi \in V = \{\phi \in H^2(0, L) \mid \phi(0) = \phi'(0) = 0\}$ . The density, stiffness, and damping relations

$$\rho(x) = \rho h b + \rho_p h_p b_p \chi_p(x), \quad YI(x) = YI + Y_p I_p \chi_p(x), \\ cI(x) = cI + c_p I_p \chi_p(x)$$

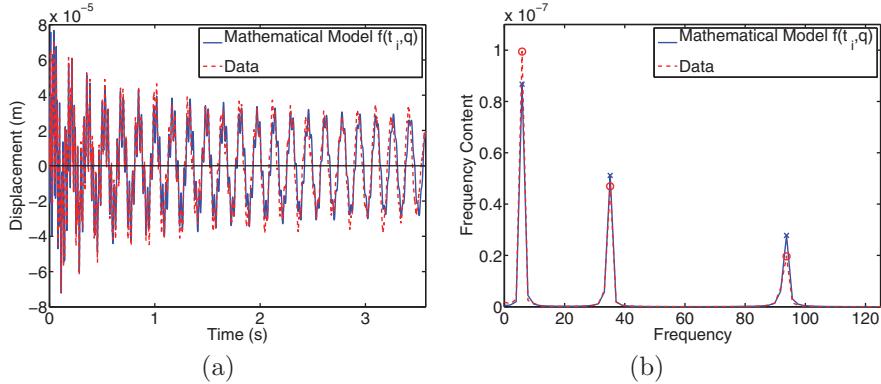
reflect the differing geometry and material properties in the region covered by the patch. The material parameters and constants are defined in Example 3.7. The reported density  $2700 \text{ kg/m}^3$  for aluminum yields  $\tilde{\rho}_b = \rho h b = 0.08775$ , which is fixed to ensure that the remaining parameters are identifiable. The parameter set is

$$q = [\tilde{\rho}_b, \gamma, YI_p, YI_b, cI_b, cI_p, k_p],$$

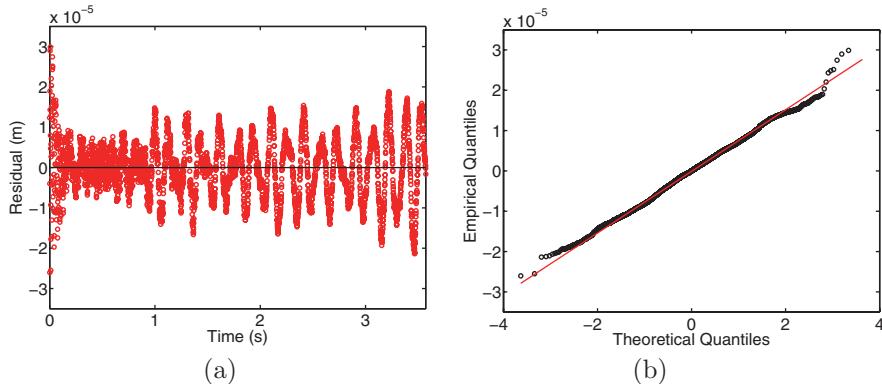
where  $\tilde{\rho}_p = \rho_p h_p b_p$ ,  $YI_p = Y_p I_p$ ,  $YI_b = YI$ ,  $cI_p = C_p I_p$ , and  $cI_b = C_b I_b$ . The model response is the displacement  $y(t_i, q) = w(t_i, \bar{x}, q)$  evaluated at the point  $\bar{x} = 128 \text{ mm}$ .

Parameter distributions were constructed using the delayed rejection adaptive Metropolis (DRAM) algorithm, discussed in Section 8.6, applied to the first second of data. The optimized fit on the time interval  $[0, 1]$  and prediction for  $[1, 3.573]$  are compared with the data in Figure 12.4(a), and the frequencies for the entire time interval are plotted in Figure 12.4(b). The residuals and Q-Q plot of the residuals are plotted in Figure 12.5.

It is observed that even though the model is accurately quantifying the behavior of the device, the residuals are heteroscedastic and exhibit clearly discernible



**Figure 12.4.** Model fit for  $t \in [0, 1]$  and predictions for  $t \in [1, 3.572]$  in the (a) time and (b) frequency domains.



**Figure 12.5.** (a) *Residuals* and (b) *Q-Q plot* of the residuals for  $[0, 3.572]$ .

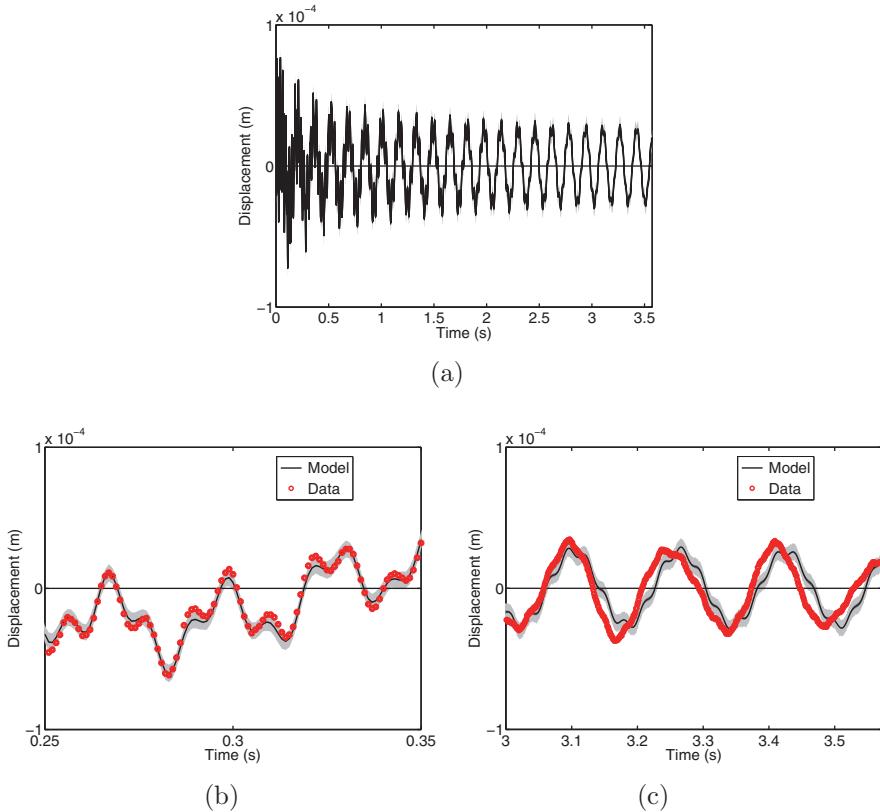
periodic behavior. The non-Gaussian behavior of the residuals is quantified by the Q-Q plot. In combination, this illustrates that despite the accurate model fit, the combined measurement and model errors are neither Gaussian nor iid.

To statistically quantify the predictive capabilities of the model, we employ the techniques of Chapter 9 to construct prediction intervals which are plotted in Figure 12.6. For the time interval  $[0.25, 0.35]$ , which is included in the fitting interval  $[0, 1]$ , the prediction interval overlaps the correct percentage of the data. In the truly predictive region  $[3, 3.572]$ , the prediction interval includes essentially *none* of the data due to the incorrect modeling of errors. This is due to unaccommodated discrepancy terms  $\delta(t_i)$  and the resulting non-iid behavior of the combined error  $\delta(t_i) + \varepsilon_i$ .

## 12.1 Effects of Unaccommodated Model Discrepancy

Depending on the magnitude and nature of model errors or discrepancy terms  $\delta(t_i)$  or  $\delta(x_i)$ , their neglect can negatively impact model calibration and prediction in four ways.

- It can diminish the validity of estimates  $q$  for physical parameters since all parameters are treated as tuning parameters whose optimized values may attempt to compensate for missing physics.
- Highly correlated or heteroscedastic errors invalidate hypotheses associated with the employed likelihoods, which can diminish the accuracy of sampling distributions or estimated parameter distributions.
- As illustrated in Example 12.3, the neglect of model discrepancies can produce inaccurate prediction intervals or intervals that are significantly larger than those constructed for models that incorporate  $\delta(t_i)$  in a statistically consistent manner.
- Unaccommodated model discrepancy terms can yield highly inaccurate extrapolatory predictions when inputs other than those used for calibration are employed in the model. This is illustrated in Example 12.2.



**Figure 12.6.** Prediction intervals and data for the time intervals (a)  $[0, 3.572]$ , (b)  $[0.25, 0.35]$ , and (c)  $[3, 3.572]$ .

In combination, neglect of model errors can negatively impact all three components of predictive estimation, as defined in Chapter 1. It limits the accuracy of both parameter estimates and their uncertainties during model calibration as well as the accuracy of predictions and prediction intervals. The resulting inaccuracy can be especially pronounced when predictions require extrapolation from the calibration regime, as is often the case for time-dependent problems. Finally, it can shrink validation regimes, which further limits the model's predictive capabilities.

We illustrate two techniques to quantify model discrepancy terms  $\delta(x_i)$  or  $\delta(t_i)$ : incorporation of unmodeled physical or biological properties, and statistical or algebraic techniques to quantify model errors. The first technique yields superior extrapolatory predictive capabilities and may be the only option, but it is problem-dependent and often very difficult to achieve. The second approach is more general but typically yields less accurate extrapolatory predictions unless one imposes stringent prior information for both the physical parameters and model discrepancy terms.

## 12.2 Incorporation of Missing Physical Mechanisms

**Example 12.4.** In Example 12.2, we illustrated that algebraic or statistical techniques will be ineffective for quantifying the discrepancy  $\delta(x_i)$  associated with the model (12.1) when considering rods of different materials. This necessitates reformulating the model to incorporate the neglected physics which produces the conflicting biases illustrated in Figure 12.2.

We employ three physical mechanisms when constructing the model (12.1) detailed in Example 3.5: conservation of energy, Newton's law of cooling along the exposed surface, and flux inputs at the source. The first two reflect well-founded physical principles, so we focus on the source boundary condition.

Flux balance analogous to Newton's law yields the Robin boundary condition

$$k \frac{dT_s}{dx}(0) - \eta T_s(0) = -\eta T_{source}, \quad (12.2)$$

where  $\eta$ , with units of  $\frac{W}{cm^2 \cdot C}$ , is a second heat transfer coefficient and  $T_{source}$  is a source term. We note that the original source condition

$$\frac{dT_s}{dx}(0) = \frac{1}{k} \Phi$$

is an approximation of (12.2) if one takes  $\Phi = -\eta T_{source}$  and neglects the term  $\eta T_s(0)$ . The solution to the modified model

$$\begin{aligned} \frac{d^2T_s}{dx^2} &= \frac{2(a+b)}{ab} \frac{h}{k} [T_s(x) - T_{amb}], \\ \frac{dT_s}{dx}(0) - \frac{\eta}{k} T_s(0) &= -\frac{\eta}{k} T_{source} \quad , \quad \frac{dT_s}{dx}(L) + \frac{h}{k} T_s(L) = \frac{h}{k} T_{amb} \end{aligned} \quad (12.3)$$

is

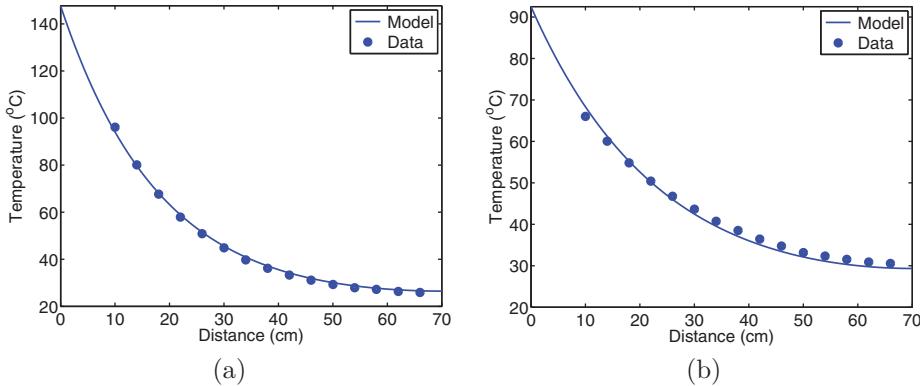
$$T_s(x, q) = c_1(q)e^{-\gamma x} + c_2(q)e^{\gamma x} + T_{amb}, \quad (12.4)$$

where  $\gamma = \sqrt{\frac{2(a+b)h}{abk}}$  and

$$\begin{aligned} c_1(q) &= \frac{\eta(T_{amb} - T_{source})}{\eta - k\gamma} \left[ \frac{e^{\gamma L}(h + k\gamma)}{e^{-\gamma L}(h - k\gamma) + e^{\gamma L}\Gamma(h + k\gamma)} \right], \\ c_2(q) &= \frac{\eta(T_{source} - T_{amb})}{\eta - k\gamma} + c_1(q)\Gamma. \end{aligned}$$

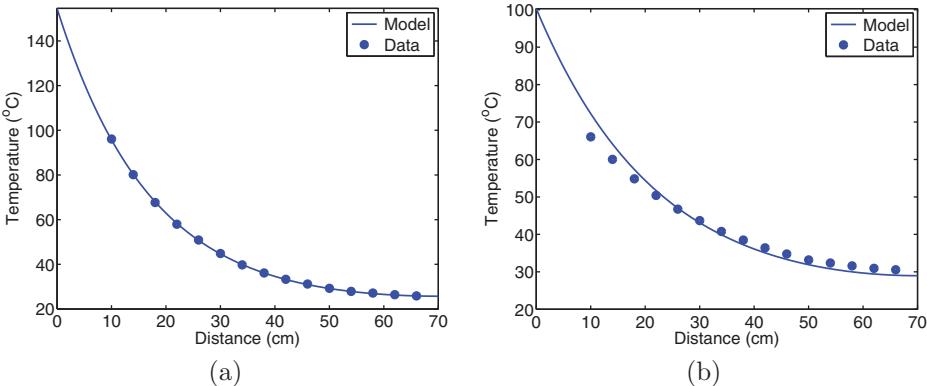
Here  $\Gamma = \frac{k\gamma + \eta}{k\gamma - \eta}$  and the parameter set is taken to be  $q = [T_{source}, h, \eta]$ . We again suppress the parameter dependence of  $\gamma$  and  $\Gamma$  to clarify the notation.

We repeat Example 12.2 with the new model (12.3). A least squares fit to the two data sets using the thermal conductivity values  $k = 2.37$  and  $k = 4.01$  yields the parameter estimates  $T_{source} = -49.08^\circ C$ ,  $h = 0.00172$ ,  $\eta = -0.0841$  and the fits shown in Figure 12.7. Comparison with the fits shown in Figure 12.2 for the original model (12.1) shows a marked improvement in the accuracy.



**Figure 12.7.** Fit of the model (12.3) for the (a) aluminum and (b) copper rods.

To illustrate the predictive capability of the extended model, we again estimate  $q = [T_{\text{source}}, h, \eta]$  using the aluminum data with  $k = 2.37$  and use the calibrated model to predict the copper temperature distribution by employing the copper thermal conductivity value  $k = 4.01$ . The results in Figure 12.8 illustrate that the prediction is now reasonable as compared with the original prediction illustrated in Figure 12.3. One can ascertain the uncertainty in this prediction by simulating with thermal conductivity values sampled from the ranges  $2.04\text{--}2.50 \frac{W}{cm \cdot C}$  and  $3.53\text{--}4.01 \frac{W}{cm \cdot C}$  reported for aluminum and copper.



**Figure 12.8.** (a) Fit of the model (12.3) to the aluminum data and (b) prediction for the copper rod.

**Remark 12.5.** A solution that is commonly proposed in the literature, to address model discrepancies, is to incorporate physical mechanisms missing in the original model. In the problem illustrated in Examples 12.2 and 12.4, this was essentially the only solution and, when this approach is possible, it generally yields prediction intervals that are tighter than those obtained using phenomenological algebraic or

statistical relations to quantify  $\delta(x_i)$  or  $\delta(t_i)$ . For Example 12.3, one could employ the more comprehensive Timoshenko model with loss mechanisms incorporated in the boundary conditions at  $x = 0$ . However, this will significantly complicate implementation, and there is no guarantee that the new model will have more physical parameters and yield iid residuals with reduced prediction intervals. More generally, the inclusion of missing physics is often not a reasonable solution since the physics would have been initially incorporated if physically or computationally feasible. This motivates the consideration of algebraic or statistical techniques to quantify model discrepancies  $\delta(x_i)$  or  $\delta(t_i)$ .

## 12.3 Techniques to Quantify Model Errors

To simplify the discussion, we focus on the spatial statistical model

$$\Upsilon_i = f(x_i, q) + \delta(x_i) + \varepsilon_i, \quad x \in \mathbb{R}^1,$$

and note that time-dependent problems or problems with  $x \in \mathbb{R}^2$  or  $\mathbb{R}^3$  can be addressed in an analogous manner.

### 12.3.1 Polynomial Models

For simplicity, we consider the quadratic discrepancy representation

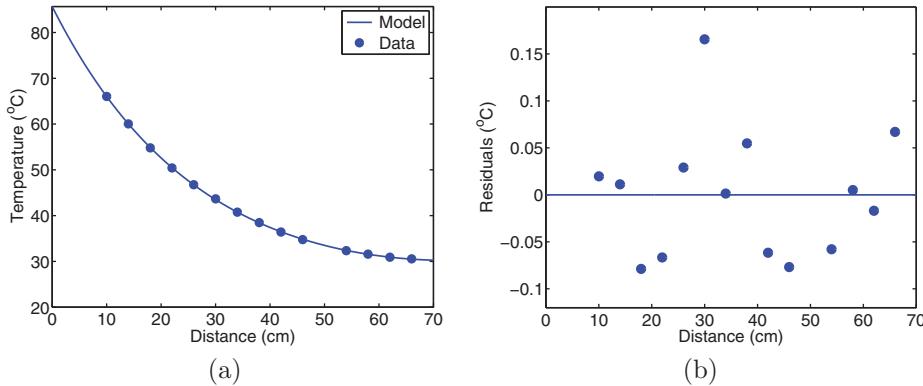
$$\delta(x_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 \quad (12.5)$$

and note that higher-order polynomial representations can be constructed in a similar manner. Here  $\beta_0, \beta_1$ , and  $\beta_2$  are unknown hyperparameters which we estimate, along with the model parameters, using the frequentist or Bayesian techniques detailed in Chapters 7 and 8. The augmented parameter set is thus  $q_{aug} = [q, \beta_0, \beta_1, \beta_2]$ , and, as detailed in Section 12.4, the identifiability of the augmented parameter set and balance between  $f(x_i, q)$  and  $\delta(x_i)$  constitute critical issues when quantifying model discrepancy in this manner. Finally, readers are referred to Section 13.1.1, and particularly (13.9), for analogous construction of quadratic response surface representations employed as surrogate models.

**Example 12.6.** In Example 12.1, we showed that the original heat model (12.1) accurately fit the copper rod data compiled in Table 3.3 but yielded the correlated residuals shown in Figure 12.1. Hence the combined errors  $\delta(x_i) + \varepsilon_i$  are not iid. Here we show that this can be addressed by modeling the discrepancy terms using the quadratic representation (12.5). The complete model is thus

$$\Upsilon_i = f(x_i, q) + \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i, \quad (12.6)$$

where  $f(x_i, q) = T_s(x_i, q)$  is given by (3.22) and  $q_{aug} = [\Phi, h, \beta_0, \beta_1, \beta_2]$ . A least squares fit to the copper data yields the parameter estimates  $q_{aug} = [-9.49, 0.00135, -5.67 \times 10^{-5}, 2.94 \times 10^{-3}, -2.37 \times 10^{-4}]$  and the fit and residuals shown in Figure 12.9. The residuals no longer exhibit a discernible pattern, thus motivating the



**Figure 12.9.** (a) Fit of the combined model  $f(x_i, q) + \delta(x_i)$  and (b) residuals for the copper rod.

assumption that the measurement errors are iid. Furthermore, the estimates for  $\Phi$  and  $h$  are close to the values  $\Phi = -9.93$  and  $h = 0.00143$  obtained in Example 12.1, which indicates that the physical model dominates the algebraic discrepancy term  $\delta(x_i)$ .

The fit of the model (12.6), where  $f(x_i, q) = T_s(x_i, q)$  is the solution to extended heat model (12.3), is explored in Exercise 12.1. The construction of densities for  $q_{aug}$ , using the Bayesian techniques of Chapter 8, is investigated in Exercise 12.3.

### 12.3.2 Kriging or Gaussian Process Models

Polynomial discrepancy representations are intuitive and direct to implement, but they do not provide mechanisms to incorporate known or assumed distributions for  $\delta(x_i)$ . In a manner analogous to that described in Section 13.1.1 for surrogate model construction, kriging or Gaussian process representations provide this capability.

For model evaluations at  $n$  data points  $x = [x_1, \dots, x_n]$ , the model discrepancy term

$$\delta(x, \mu, \sigma^2, \theta) = [\delta(x_1, \mu, \sigma^2, \theta), \dots, \delta(x_n, \mu, \sigma^2, \theta)]$$

is assumed to be from a multivariate normal distribution,

$$\delta \sim N(\mu I, \sigma^2 R), \quad (12.7)$$

where  $I = [1, \dots, 1]$  is a unit  $n$ -vector and  $R$  is a symmetric, positive definite matrix with elements

$$R_{ij} = \exp(-\theta|x_i - x_j|). \quad (12.8)$$

We note that the correlation function (12.8) is the univariate version of the  $q$ -variate representation (13.13), illustrated in Figure 13.4, that is employed for surrogate model construction. The hyperparameters  $\sigma, \mu$ , and  $\theta$  are combined with the physical parameters  $q$  to form the augmented parameter set  $q_{aug} = [q, \sigma, \mu, \theta]$  to be estimated. For spatial processes, this is often termed a kriging representation in

reference to its geophysical origin. The manner in which an interpolatory kriging or Gaussian process model provides uncertainty bounds is illustrated in Figure 13.3. The use of a Gaussian process model to quantify the model discrepancy term discussed in Example 12.1 is pursued in Exercise 12.2.

Whereas this approach is commonly considered due to its generality and statistical attributes, it often exhibits limitations in applications. Representative issues are discussed in the next section.

## 12.4 Issues Pertaining to Model Discrepancy Representations

The phenomenological nature of the polynomial and kriging, or Gaussian process, model discrepancy representations (12.5) and (12.7) has the advantage that its use requires no knowledge of the underlying physical or biological process quantified by the model  $f(x_i, q)$ . Hence these representations can be employed for a fairly broad range of applications. The kriging or Gaussian process representations have the additional advantage that they provide a natural mechanism for incorporating prior knowledge about the process, *provided that it is Gaussian*. However, the fact that the representations are phenomenological also imbues them with inherent limitations, which we discuss in this section. This also renders the topic of quantifying model discrepancy terms for general processes an open and active research area.

### 12.4.1 Parameter Identifiability

In Chapter 6, we discussed parameter selection techniques to isolate identifiable or influential subspaces of model parameters or inputs. We used these techniques to reduce the number of parameters to those which could be uniquely determined from responses as required for OLS algorithms. We noted in Section 8.5 that Bayesian techniques could, in some cases, be used to construct densities for nonidentifiable parameters if one has sufficiently informative priors.

These observations also apply to the augmented parameter set  $q_{aug} = [q, q_{dis}]$ , where  $q_{dis} = [\beta_0, \beta_1, \beta_2]$  or  $q_{dis} = [\mu, \sigma, \theta]$  in the model

$$\Upsilon_i = f(x_i, q) + \delta(x_i, q_{dis}) + \varepsilon_i.$$

The addition of the discrepancy terms  $\delta(x_i, q_{dis})$  can yield an unidentifiable augmented parameter set  $q_{aug}$  for problems where  $q$  is identifiable. In some cases, this can be addressed using one of the following techniques.

- Employ sufficiently informative priors for the parameters and discrepancy function to permit Bayesian inference of  $q_{aug}$ . In some cases, this can be achieved by restricting the admissible parameter space based on a priori information. For parameters with little to no prior information, this may not be feasible.
- Iteratively solve for  $\delta$  and  $\delta_{dis}$  while keeping one parameter set fixed. This can be employed if  $q$  is identifiable but  $q_{aug}$  is not. This approach is suboptimal and may not yield a global minimum.

### 12.4.2 Confounding of Physical and Phenomenological Components during Extrapolation

In Examples 12.2 and 12.3, we illustrated model predictions that require extrapolation beyond the calibration domain. In the first example, this involved predictions using different physical coefficients, whereas in the second, it involved predictions in time. For these examples, unaccommodated model discrepancy terms yielded inaccurate predictions in the first case and inaccurate predictions intervals in the second.

For the statistical model

$$\Upsilon_i = f(x_i, q) + \delta(x_i, q_{dis}) + \varepsilon_i,$$

it is the physical or biological model  $f(x_i, q)$  which propagates the information required for extrapolatory predictions. The role of the phenomenological discrepancy function  $\delta(x_i, q_{dis})$  is to ensure that physical parameters  $q$  are estimated in a statistically consistent manner so that predictions and prediction intervals are accurate. Because this term is phenomenological, it has essentially no predictive capabilities outside the calibration domain unless stringent prior information is imposed on the admissible space of functions used to construct  $\delta(x_i, q_{dis})$ . There are a number of difficulties that arise when attempting to specify a class of functions for  $\delta(x_i, q_{dis})$  with physically reasonable prior information.

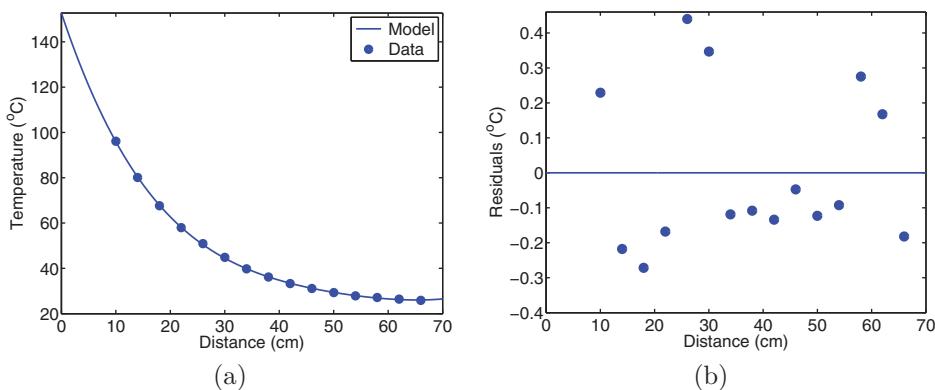
- We illustrated in Example 12.2 that no class of functions could be used to construct  $\delta(x_i, q_{dis})$  for different materials due to opposing biases in the residuals. We demonstrated in Example 12.4 that additional physics had to be incorporated to improve predictions for materials not used during model calibration.
- If insufficient prior structure is imposed on  $\delta(x_i, q_{dis})$  during model calibration, the hyperparameters can be chosen so that  $\delta(x_i, q_{dis})$  overly compensates for physical or biological mechanisms. The combined model  $f(x_i, q) + \delta(x_i, q_{dis})$  can provide accurate predictions in the calibration domain but have minimal predictive capability outside this domain.

To illustrate, consider the steady state heat model (12.3) with the non-physical value  $\eta = 0$ . From (12.4), this yields  $T_s(x_i, q) = T_{amb}$ , so the statistical model is

$$\Upsilon_i = T_{amb} + \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 + \varepsilon_i \quad (12.9)$$

if one employs a quartic discrepancy relation. A least squares fit to determine the hyperparameters, using the aluminum data in Table 3.2, yields the accurate fit illustrated in Figure 12.10. However, the model has no ability to predict the heat distribution for the copper rod. To avoid this scenario, one must impose prior information on the representation for  $\delta(x_i, q_{dis})$  that prohibits the choice  $\eta = 0$ .

- The Gaussian process representation (12.7) will be ineffective for quantifying the model discrepancy term  $\delta(t_i, q_{dis})$  for the time-dependent beam equation



**Figure 12.10.** Fit of the phenomenological model (12.9) to the aluminum data.

in Example 12.3. The quantification of  $\delta(t_i, q_{dis})$  for extrapolatory predictions of time-dependent processes constitutes an active research area.

- Gaussian process representations will not be effective for highly non-Gaussian random processes. The development of statistical representations for such problems is also an active research area.

## 12.5 Notes and References

The analysis of model discrepancy as a source of uncertainty in predictive simulation models was initiated by Kennedy and O'Hagan [134], who referred to it as *model inadequacy*. Statistical models that incorporate model discrepancy terms have been employed in applications that include environment [13], hydrology [205], climate [218, 234], and engineering [12, 108]. The use of this framework to construct a rigorous validation procedure for simulation models is addressed in [31]. Despite this activity, the limitations detailed in Section 12.4 indicate a number of open research issues that remain to be addressed. The primary issue concerns the construction of the phenomenological discrepancy term  $\delta(x_i, q_{dis})$  so that it does not confound the physical model  $f(x_i, q)$  when making extrapolatory predictions outside the calibration domain. As detailed in Section 12.4.2 and [43], this necessitates that stringent prior information be imposed on the admissible space of functions used to construct  $\delta(x_i, q_{dis})$ . For some applications, control theory may be used to construct  $\delta(x_i)$  [202] and quantification of model discrepancies in this manner constitutes a current research direction.

## 12.6 Exercises

**Exercise 12.1.** Consider the steady state heat model (12.1) with the parameter set  $q = [\Phi, h]$  and the copper rod data compiled in Table 3.3. Using the thermal

conductivity value  $k = 4.01 \frac{W}{cm \cdot C}$  for copper in the solution (3.22), repeat the analysis of Example 12.6 to fit the combined physical plus quadratic discrepancy model to the data. Your augmented parameter set is  $q_{aug} = [\Phi, h, \beta_0, \beta_1, \beta_2]$ . You can use the MATLAB routines `fminsearch.m` or `lsqnonlin.m` to implement the least squares fit. Do your residuals appear to be iid?

**Exercise 12.2.** Repeat Exercise 12.1 using the Gaussian process discrepancy representation (12.7) rather than the quadratic model (12.5). How do the physical parameters in the two cases compare?

**Exercise 12.3.** Consider the model

$$\Upsilon_i = f(x_i, q) + \delta(x_i) + \varepsilon_i,$$

where  $f(x_i, q) = T_s(x_i, q)$  is the solution to the steady state heat equation (12.1) and  $\delta(x_i)$  is represented by the quadratic relation (12.5). You can use the thermal conductivity value  $k = 2.37 \frac{W}{cm \cdot C}$  for aluminum.

Use the DRAM algorithm discussed in Section 8.6 to construct densities for the augmented parameter set  $q_{aug} = [\Phi, h, \beta_0, \beta_1, \beta_2]$ . When constructing characteristic functions for noninformative priors, you should enforce the positive or negative behavior of physical parameters. How do the means of your posterior distributions compare with the least squares estimates determined in Exercise 12.1?

## Chapter 13

# Surrogate Models

*Everything should be made as simple as possible, but not simpler,* Albert Einstein

The equations of atmospheric physics (2.7), hydrology model (2.12), neutron transport equations (2.14), and nuclear thermal-hydraulic equations (2.16) and (2.17) are complex PDEs that have numerous parameters and can require minutes to days to perform a single forward simulation. Hence the Markov chain techniques of Chapter 8 and uncertainty propagation techniques of Chapter 9, which can require thousands to millions of realizations, will often be infeasible for these models. Furthermore, this will be the case for many complex models quantifying distributed, nonlinear, multiscale, multiphysics, or coupled biological phenomena. This motivates the development of surrogate models to facilitate optimization, uncertainty quantification, and control design.

We consider two frameworks when developing surrogate models. To motivate the first, we consider the algebraic model

$$A(q)\phi = s(q) \quad (13.1)$$

with observations

$$y = \mathcal{C}^T(q)\phi. \quad (13.2)$$

A motivating application is provided in Example 3.6, where a model of this form arises from the discretization of a differential equation quantifying neutron diffusion. There  $\phi = [\varphi_1, \dots, \varphi_{N-1}]^T$ , where  $\varphi_i \approx \varphi(x_i)$  approximates the flux at  $x_i$ , and the parameters are  $q = [D, A_a, S, A_d]$ , where  $D$ ,  $A_a$ ,  $S$ , and  $A_d$  respectively denote the diffusion coefficient, a macroscopic absorption cross-section, a constant source, and the detector cross-section. The matrix  $A(q)$  and vectors  $s(q)$  and  $\mathcal{C}(q)$  are defined in (3.29) and (3.32).

The observation or response can thus be expressed as

$$y = f(q), \quad (13.3)$$

where the nonlinear function  $f$  is given by

$$f(q) = \mathcal{C}^T(q)A^{-1}(q)s(q).$$

For fine-scale discretizations or discretizations of 2-D or 3-D models, the numerical expense of evaluating  $f$  motivates the construction of inexpensive surrogates that retain the essential physics.

Second, we consider the evolutionary PDE

$$\begin{aligned}\frac{\partial u}{\partial t} &= \mathcal{N}(u, q) + F(q), \quad x \in \mathcal{D}, \quad t \in [0, \infty), \\ B(u, q) &= G(q), \quad x \in \partial\mathcal{D}, \quad t \in [0, \infty), \\ u(0, x, q) &= I(q), \quad x \in \mathcal{D},\end{aligned}\tag{13.4}$$

discussed in Section 10.2.3. The corresponding weak formulation is

$$\int_{\mathcal{D}} \frac{\partial u}{\partial t} v dx + \int_{\mathcal{D}} N(u, q) S(v) dx = \int_{\mathcal{D}} F(q) v dx,\tag{13.5}$$

which holds for all  $v \in V$ , where  $V$  is an appropriate space of test functions. Here  $\mathcal{N}$  and  $N$  are potentially nonlinear spatial differential operators,  $S$  is a linear operator that results from integration by parts,  $F$  is a source term,  $B$  and  $G$  are boundary operators,  $I$  specifies initial conditions, and  $\mathcal{D}$  is a subset of  $\mathbb{R}^1$ ,  $\mathbb{R}^2$ , or  $\mathbb{R}^3$ . The response is taken to be observations of the state at  $(t, x)$  so that

$$y = u(t, x, q).\tag{13.6}$$

In general, one must solve (13.4) or (13.5) numerically, which can be extremely expensive for nonlinear operators with 2-D or 3-D spatial domains. For such applications, surrogate models are required for design, uncertainty quantification, and control implementation.

For all models, one first seeks to reduce the number of parameters or inputs to be estimated and propagated to only those that are identifiable or influential in the sense defined in Definitions 6.1 and 6.2. Noninfluential parameters are then fixed at nominal values for subsequent model calibration and uncertainty propagation. This is critical for two reasons: (i) only identifiable parameters and their uncertainties can be estimated using the frequentist model calibration techniques of Chapter 7 or Bayesian techniques of Chapter 8 with noninformative priors, and (ii) the parameter space must be reduced for models such as those arising in neutron transport or systems biology where  $p$  can be on the order of millions. Parameter selection techniques are detailed in Chapter 6.

The objective of all surrogate models is to construct representations that quantify the primary features of the high-fidelity model while providing the computational efficiency required for Bayesian model calibration, uncertainty quantification, design, and control implementation. Surrogate models can be broadly categorized in three classes: regression or interpolation-based models, projection-based models, and hierarchical models. As detailed in Section 13.1, the first class of models is constructed by treating the original model as a black box from which samples are drawn to construct efficient input-output relations based on interpolation or regression theory. These models are often termed *data-fit models*, *response surface models*, *emulators*, *meta-models*, or *approximation models*, and construction techniques include stochastic collocation, polynomial approximations, radial basis functions, and

Gaussian process or kriging representations. It is important to note that these methods are generally nonintrusive, which facilitates their use for large-scale applications and general purpose software packages. The second class of models, commonly termed *reduced-order models*, is constructed by projecting states and distributed parameters onto low-order subspaces in the manner detailed in Section 13.2. We detail eigenfunction or modal expansions, proper orthogonal decomposition (POD), and high-dimensional model representation (HDMR) methods. Hierarchical surrogates are based on techniques such as coarser grids, relaxed tolerances, or simplified physical or biological assumptions. We refer the reader to [77, 87] and the included references for details regarding this latter class of surrogate models.

## 13.1 Regression or Interpolation-Based Models

Certain principles underlying data-fit models—also termed response surface models, emulators, meta-models, or approximation models—are illustrated in Figure 13.1. The high-fidelity image is analogous to a highly resolved simulation code or fully characterized physical phenomenon from which samples are drawn to construct the underresolved images. Despite the limited number of sampled pixels, the partial images include enough information so that the brain is able to construct a surrogate model and identify the image. This is due in part to its ability to enforce structure through symmetry and prior knowledge of faces.

### 13.1.1 Algebraic Models

We consider first the construction of an emulator for the algebraic model (13.3) where  $q \in \Gamma \subset \mathbb{R}^p$ . In general,  $f(q)$  can be output from a high-fidelity simulation code that is computationally expensive or a physical process that can be measured for various values of  $q$ . Response data consists of  $M$  realizations or measurements

$$y_m = f(q^m), \quad m = 1, \dots, M, \quad (13.7)$$

generated by  $M$  realizations of the parameter or input vector. We first note that the techniques used to sample  $q$  are critical for the accuracy of the emulator—e.g.,  $M$  evaluations of the same parameter vector would yield a terrible emulator. Examples of appropriate sampling strategies include the Monte Carlo, Latin hypercube, and sparse grid techniques detailed in Chapter 11. Second, we note that for high-fidelity



**Figure 13.1.** High-fidelity image and poorly resolved sampled images.

simulations,  $f(q)$  is treated as a black box, so sampling is nonintrusive in the sense that it does not require modification of existing software packages. This permits the direct use of legacy codes or executable files, which is a significant advantage for many problems.

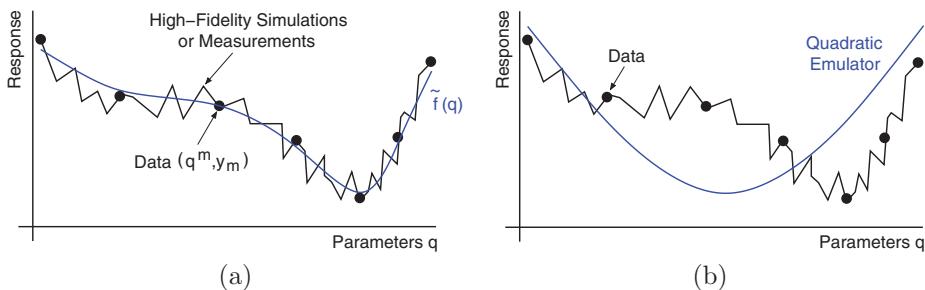
The objective is to use the sampled data  $(q^m, y_m)$  to construct an emulator  $\tilde{f}(q)$  that approximates  $f(q)$  with sufficient accuracy while providing the efficiency to repeatedly predict responses for new values of  $q$  as required for optimization, uncertainty quantification, or control implementation. Whereas there exist a number of techniques for constructing emulators, we focus on quadratic response surface models, kriging or Gaussian process models, and radial basis formulations. Emulators based on stochastic polynomial expansions are discussed in Section 13.1.2 in the context of the evolutionary PDE (13.4).

To motivate the statistical framework used to construct emulators, consider the response depicted in Figure 13.2(a). This represents high-fidelity simulations or experiments that have been resolved at a very fine scale along with sampled data  $(q^m, y_m)$ . The fine-scale resolution may be idealized and, for applications such as numerical or experimental resolution of turbulent flows, cannot typically be achieved for operations requiring numerous simulations.

To accommodate unresolved fine-scale behavior, as well as experimental measurement errors, we consider the deterministic response to be a realization of a random process quantified by the statistical model

$$Y_m = \tilde{f}(q^m) + \varepsilon_m \quad , \quad m = 1, \dots, M, \quad (13.8)$$

where  $Y_m$  are random variables with realizations  $y_m$ . We denote the vector of observations by  $\mathbf{y}_s = [y_1, \dots, y_M]^T$ . Here  $\varepsilon = [\varepsilon_1, \dots, \varepsilon_M]^T$  is a random vector associated with unresolved fine-scale behavior or measurement errors. We assume that  $\varepsilon_m$  are iid and normally distributed with mean 0 and true but unknown variance  $\sigma_0^2$  so that  $\varepsilon_m \sim N(0, \sigma_0^2)$ . We note that the use of the statistical model (13.8) does not imply that high-fidelity simulations introduce random behavior but rather that it provides a framework for constructing interpolation and regression models when fine-scale behavior is difficult or impossible to incorporate.



**Figure 13.2.** (a) High-fidelity simulations or measurements, data, and emulator  $\tilde{f}(q)$  and (b) quadratic emulator.

### Quadratic Response Surface Model

The construction of a polynomial response surface model can be posed in the linear regression framework detailed in Chapter 7. To balance efficiency and accuracy, we illustrate the quadratic emulator

$$\tilde{f}(q, \beta) = \beta_0 + \sum_{i=1}^p \beta_i q_i + \sum_{i=1}^p \beta_{ii} q_i^2 + \sum_{i=1}^p \sum_{j>i}^p \beta_{ij} q_i q_j, \quad (13.9)$$

where  $q = [q_1, \dots, q_p]^T$  is fixed and known and  $\beta_0, \beta_i, \beta_{ii}$ , and  $\beta_{ij}$  are unknown deterministic parameters for which we will construct an estimator. Since there are  $P = \frac{(p+1)(p+2)}{2}$  coefficients, we need  $M > P$  samples from our high-fidelity simulation code or experimental measurements.

The linear regression theory of Chapter 7 yields the least squares estimate

$$\beta = [X^T X]^{-1} X^T \mathbf{y}_s, \quad (13.10)$$

where the design matrix is

$$X = \begin{bmatrix} 1 & q_1^1 & \cdots & q_p^1 & (q_1^1)^2 & \cdots & (q_p^1)^2 & q_1^1 q_2^1 & \cdots & q_{p-1}^1 q_p^1 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1 & q_1^M & \cdots & q_p^M & (q_1^M)^2 & \cdots & (q_p^M)^2 & q_1^M q_2^M & \cdots & q_{p-1}^M q_p^M \end{bmatrix} \in \mathbb{R}^{M \times p}.$$

Once  $\beta$  has been constructed, (13.9) can be used to predict responses for new values of  $q$ .

Quadratic emulators are popular surrogates for large-scale optimization problems since they provide analytic values for optima. However, they can have limited accuracy for high-fidelity models such as that illustrated in Figure 13.2(b), which motivates consideration of higher-order polynomial, Gaussian process, or radial basis representations.

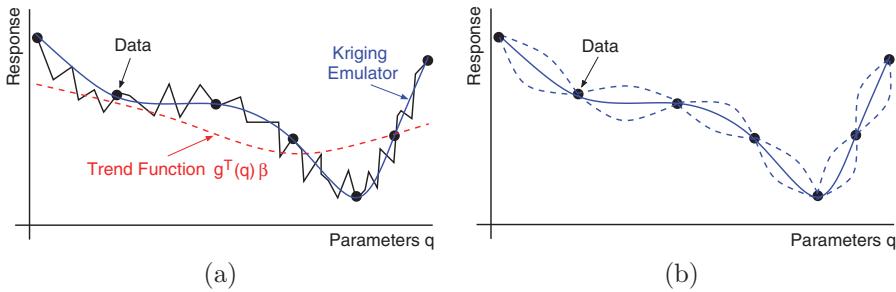
### Kriging Model

Kriging estimation—which is also termed Gaussian process regression—originated in the geophysical sciences with the work of South African mining engineer Danie Krige. We again consider the statistical model (13.8) where the kriging emulator

$$\tilde{f}(q, \beta) = g^T(q)\beta + Z(q) \quad (13.11)$$

is comprised of a deterministic trend function  $g^T(q)\beta$  and a Gaussian process error model  $Z(q)$ , as illustrated in Figure 13.3(a). In ordinary kriging, the trend function is assumed constant, so  $g^T(q)\beta = \beta_0$ , whereas universal kriging assumes a polynomial representation  $g^T(q)\beta = \sum_{k=0}^K \beta_k g_k(q)$  with coefficients  $\beta = [\beta_0, \dots, \beta_K]$  determined using least squares regression. To simplify the discussion, we consider ordinary kriging, where we construct an estimator  $\hat{\beta}_0$  for  $\beta_0$ , and refer readers to [212] for details regarding universal kriging.

The Gaussian process ensures that, in the absence of measurement errors, the emulator interpolates—and hence has zero uncertainty—at the sample points  $q^m$ ;



**Figure 13.3.** (a) Interpolatory kriging emulator in the absence of measurement error and (b) uncertainty bounds.

that is,  $\tilde{f}(q^m, \beta_0) = y_m$ . In the absence of measurement noise,  $Z(\cdot)$  is assumed to be a stationary random process, as defined in Definition 4.45, with zero mean, variance  $\sigma^2$ , and nonzero covariance

$$\text{cov}[Z(q^i), Z(q^j)] = \sigma^2 R(q^i, q^j) + \sigma_0^2 \delta(q^i - q^j), \quad (13.12)$$

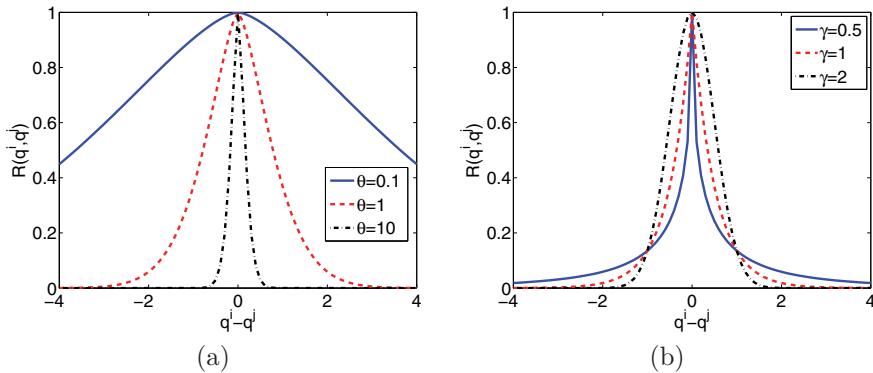
where

$$\delta(q^i - q^j) = \begin{cases} 1, & q^i - q^j = 0, \\ 0, & \text{else} \end{cases}$$

and

$$R(q^i, q^j) = \exp\left(-\sum_{k=1}^p \theta_k |q_k^i - q_k^j|^{\gamma_k}\right), \quad 0 < \gamma_k \leq 2, \quad \theta_k > 0, \quad (13.13)$$

is the correlation function. The hyperparameters  $\theta_k, \gamma_k, k = 1, \dots, p$ , can be tuned to achieve varying degrees of correlation, as illustrated in Figure 13.4. The ratio  $\eta =$



**Figure 13.4.** Correlation function  $R$  given by (13.13) for  $q \in \mathbb{R}^1$ : (a) fixed  $\gamma = 1.5$  and (b) fixed  $\theta = 2$ .

$\sigma^2/\sigma_0^2$  between the unadjusted variance and variance  $\sigma_0^2$  of the measurement noise is sometimes termed the *nugget*. Alternative choices for the correlation function are provided in [212].

As detailed in [212], the kriging prediction  $\tilde{f}(q, \beta_0)$ , for new values of  $q$ , is given by

$$\tilde{f}(q, \beta_0) = \beta_0 + r^T(q)\mathcal{R}^{-1}[\mathbf{y}_s - \beta_0\mathbf{1}], \quad (13.14)$$

where

$$\beta_0(\theta, \gamma) = [\mathbf{1}^T \mathcal{R}^{-1} \mathbf{1}]^{-1} \mathbf{1}^T \mathcal{R}^{-1} \mathbf{y}_s \quad (13.15)$$

is the least squares estimate for  $\beta_0$ . Here  $\mathbf{1} = [1, \dots, 1]^T \in \mathbb{R}^M$  and  $\mathcal{R}$  is an  $M \times M$  correlation matrix defined componentwise by  $\mathcal{R}_{ij} = R(q^i, q^j)$ . The  $M \times 1$  vector  $r(q)$ , with components  $r_i(q) = R(q^i, q)$ , quantifies the Gaussian process correlations between values at the design sites and the new input  $q$ . We note that the second term in (13.14) employs the data to adjust the mean and provide a best linear unbiased estimate of the true or high-fidelity function.

Additionally, the theory provides mean squared errors for predictions which can guide sample point refinement. In the absence of measurement noise  $\varepsilon$ , the adjusted kriging emulator variance is

$$\text{var}[\tilde{f}(q, \hat{\beta}_0)] = s^2 \left[ 1 - r^T \mathcal{R} r + \frac{(r^T \mathcal{R}^{-1} \mathbf{1} - 1)^2}{\mathbf{1}^T \mathcal{R}^{-1} \mathbf{1}} \right], \quad (13.16)$$

where the MLE of the unadjusted variance is

$$s^2 = \frac{1}{M} [\mathbf{y}_s - \beta_0(\theta, \gamma) \mathbf{1}]^T \mathcal{R}^{-1} [\mathbf{y}_s - \beta_0(\theta, \gamma) \mathbf{1}]. \quad (13.17)$$

The use of the variance estimate to construct confidence intervals for predictions is illustrated in Figure 13.3(b).

To construct  $s^2$  and hence  $\text{var}[\tilde{f}(q, \hat{\beta}_0)]$ , one must first estimate the hyperparameters  $\theta = [\theta_1, \dots, \theta_p]$  and  $\gamma = [\gamma_1, \dots, \gamma_p]$ . One approach is to estimate them using the MCMC methods detailed in Chapter 8. A more common approach for kriging models is to estimate them by maximizing an appropriate likelihood function. Based on the assumption that the sampled data are drawn from a Gaussian process, the likelihood function is

$$L(\beta_0, s^2 | \theta, \gamma) = \frac{1}{\sqrt{2\pi(s^2)^M |\mathcal{R}|}} \exp\left(-\frac{[\mathbf{y}_s - \beta_0 \mathbf{1}]^T \mathcal{R}^{-1} [\mathbf{y}_s - \beta_0 \mathbf{1}]}{2s^2}\right)$$

so that the log-likelihood is

$$\begin{aligned} \ell(\beta_0, s^2 | \theta, \gamma) &= -\frac{M}{2} \ln(2\pi) - \frac{M}{2} \ln(s^2) - \frac{1}{2} \ln |\mathcal{R}| \\ &\quad - \frac{1}{2s^2} [\mathbf{y}_s - \beta_0 \mathbf{1}]^T \mathcal{R}^{-1} [\mathbf{y}_s - \beta_0 \mathbf{1}]. \end{aligned} \quad (13.18)$$

The necessary conditions  $\frac{\partial \ell}{\partial s^2} = 0$  and  $\frac{\partial \ell}{\partial \beta_0} = 0$  respectively yield (13.17) and (13.15). We then substitute (13.15) into (13.18) to obtain

$$\ell(\theta, \gamma) = -\frac{M}{2} [\ln(2\pi) + 1] - \frac{M}{2} \ln s^2(\theta, \gamma) - \frac{1}{2} \ln |\mathcal{R}(\theta, \gamma)|. \quad (13.19)$$

Whereas we cannot maximize (13.19) analytically, one can readily do so using various optimization routines.

Issues associated with Gaussian process or kriging representations include ill-conditioning of  $\mathcal{R}$ , exponential growth in the number of hyperparameters for large parameter dimensions  $p$ , multiple local maxima, and ridges near maximum values. The final two issues motivate the use of global optimization routines. Ill-conditioning of  $\mathcal{R}$  is addressed through attention to sampling algorithms and adding nuggets  $\eta$  to the diagonal of  $\mathcal{R}$  until  $\mathcal{R} + \eta I$  is no longer ill-conditioned. However, this latter strategy smooths the kriging model so that it approximates rather than interpolates the sampled data. The curse of dimensionality for large  $p$  can dictate that other techniques be employed to construct surrogate models for high-dimensional problems.

Details regarding tuning strategies for the hyperparameters can be found in [247], whereas extensions of the kriging formulation that include gradient information are detailed in [85].

### Radial Basis Functions

As an alternative to the polynomial basis expansion (13.9), one can employ the representation

$$\tilde{f}(q) = \sum_{m=1}^M f_m \Psi_m(q) + P(q), \quad (13.20)$$

where  $\Psi_m(q) = \psi(\|q^m - q\|)$  are radial basis functions defined in terms of the Euclidean distance between  $q^m$  and  $q$ , and  $P(q)$  is a global trend function. For this discussion, we take  $P(q) = \beta_0$  as we did for the kriging model. For  $r_m = \|q^m - q\|$ , specific choices for  $\psi$  include

$$\psi(r_m) = \begin{cases} e^{r_m/2\sigma^2} & , \text{ Gaussian}, \\ r_m^n, n = 1, 2, 3 & , \text{ Power law}, \\ r_m^2 \ln r_m & , \text{ Thin plate spline}. \end{cases} \quad (13.21)$$

To specify the coefficients  $f_m$ , we enforce the interpolation condition

$$\tilde{f}(q^m, \beta_0) = y_m \quad , m = 1, \dots, M,$$

along with the constraint

$$\sum_{m=1}^M f_m = 0,$$

which results from the inclusion of the trend function. This yields the radial basis function prediction

$$\tilde{f}(q, \beta_0) = \beta_0 + \Psi^T(q) \Phi^{-1} [\mathbf{y}_s - \beta_0 \mathbf{1}], \quad (13.22)$$

where  $\beta_0 = [\mathbf{1}^T \Phi^{-1} \mathbf{1}]^{-1} \mathbf{1}^T \Phi^{-1} \mathbf{y}_s$  and  $\Phi_{mk} = \Psi_k(q^m) = \psi(\|q^k - q^m\|)$ .

A comparison of (13.22) and (13.14) illustrates that the radial basis function expansion and kriging model have essentially the same form. The difference lies in

the fact that the radial basis function model is formulated in terms of the gram matrix and vector, constructed in terms of the basis functions, whereas the kriging model employs the correlation matrix and vector. The kriging framework also has the advantage that it provides variance estimates for the prediction.

### 13.1.2 Evolutionary PDEs

Here we consider the evolutionary PDE (13.4) or (13.5) with state observations (13.6). Due to the computational expense of solving PDEs with potentially nonlinear operators on 2-D or 3-D spatial domains, emulators are required for Bayesian model calibration and uncertainty propagation. In fact, we have already constructed such emulators in Section 10.2.3, and we simply highlight here the manner in which those constructs achieve the objectives of surrogate models.

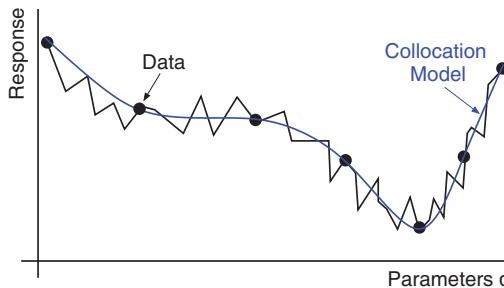
Here the surrogate model is taken to be

$$\tilde{u}(t, x, q) = \sum_{k=0}^K \sum_{j=1}^J u_{jk}(t) \phi_j(x) \Psi_k(q), \quad (13.23)$$

where  $\phi_j(x)$  are finite element or spectral basis functions and the input basis functions  $\Psi_k(q) = L_k(q)$  are Lagrange polynomials that collocate at the  $M = R$  quadrature points  $q^m = q^r$ ; that is,  $\Psi_k(q^m) = \delta_{km}$  in the manner shown in Figure 13.5. We note that the input basis representation (13.23) is analogous to the algebraic expansions (13.9) and (13.20) with Lagrange polynomials used instead of quadratic polynomials or radial basis functions. The coefficients  $u_{jm}(t) = u_{jr}(t)$  are determined by solving the  $MJ$  evolution equations

$$\frac{du_{jr}}{dt} + \int_{\mathcal{D}} N \left( \sum_{j=1}^J u_{jr} \phi_j(x), q^r \right) S(\phi_\ell(x)) dx = \int_{\mathcal{D}} F(q^r) \phi_\ell(x) dx \quad (13.24)$$

for  $\ell = 1, \dots, J$ . The points  $q^m = q^r$  are specified using the Monte Carlo or sparse grid techniques detailed in Chapter 11.



**Figure 13.5.** Collocation of the surrogate model (13.23) through data generated using a high-fidelity model.

The surrogate model (13.23) is thus constructed using solutions of the high-fidelity simulation code with the number of required solutions minimized through the use of sparse grid techniques for moderate parameter dimensions. The use of this emulator for uncertainty propagation is detailed in Chapters 9 and 10.

**Remark 13.1.** This is a surrogate in the parameter space but not in the spatial domain. For 2-D or 3-D spatial domains  $\mathcal{D}$ , the spatial discretization level  $J$  can easily be on the order of millions or higher, which can render the solution of (13.24) prohibitively expensive, especially for nonlinear operators  $\mathcal{N}$  and  $N$ . This is addressed for certain problems by the projection-based reduced-order methods.

## 13.2 Projection-Based Models

Reduced-order models are constructed by projecting high-dimensional states  $u$  and parameters  $q$  onto low-dimensional subspaces in the manner depicted in Figure 13.6. This is in contrast to interpolation or regression-based data-driven models.

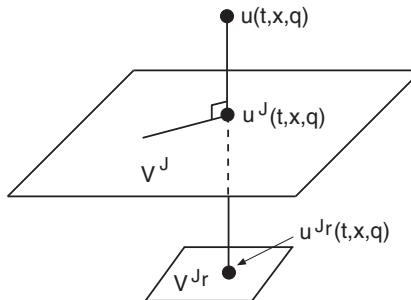
To motivate, we again consider the evolutionary PDE

$$\begin{aligned} \frac{\partial u}{\partial t} &= \mathcal{N}(u, q) + F(q) \quad , \quad x \in \mathcal{D}, \quad t \in [0, \infty), \\ B(u, q) &= G(q) \quad , \quad x \in \partial \mathcal{D}, \quad t \in [0, \infty), \\ u(0, x, q) &= u_0(x, q) \quad , \quad x \in \mathcal{D}, \end{aligned} \tag{13.25}$$

which has the deterministic weak formulation

$$\int_{\mathcal{D}} \frac{\partial u}{\partial t} v dx + \int_{\mathcal{D}} N(u, q) S(v) dx = \int_{\mathcal{D}} F(q) v dx , \quad v \in V. \tag{13.26}$$

Here  $q \in \Gamma \subset \mathbb{R}^p$  and  $V$  is an appropriate space of test functions. To approximate the state  $u(t, x, q)$ , we project the problem onto the finite-dimensional subspace  $V^J = \text{span}\{\phi_j\} \subset V$ , where  $\phi_j(x)$  are finite elements or spectral basis functions,



**Figure 13.6.** Projection of the infinite-dimensional problem (13.26) onto the finite-dimensional subspace  $V^J$  and reduced-order subspace  $V^{J_r}$ .

and consider the approximate solution

$$u^J(t, x, q) = \sum_{j=1}^J u_{jk}(t) \phi_j(x), \quad (13.27)$$

which must satisfy

$$\int_{\mathcal{D}} \frac{\partial u^J}{\partial t} \phi_\ell dx + \int_{\mathcal{D}} N(u^J, q) S(\phi_\ell) dx = \int_{\mathcal{D}} F(q) \phi_\ell dx \quad (13.28)$$

for  $\ell = 1, \dots, J$ . The difficulty is that for  $\mathcal{D} \subset \mathbb{R}^2$  or  $\mathbb{R}^3$ ,  $J$  can easily be on the order of  $10^6$ – $10^8$ , which, when combined with the potentially nonlinear spatial differential operator  $N$ , can yield simulations that take hours to days to complete.

The goal with reduced-order methods is to project (13.26) onto a significantly lower-dimensional subspace  $V^{J_r} = \text{span}\{\phi_j^r\} \subset V$ ,  $J_r \ll J$ , so that the solution of

$$\int_{\mathcal{D}} \frac{\partial u^{J_r}}{\partial t} \phi_\ell^r dx + \int_{\mathcal{D}} N(u^{J_r}, q) S(\phi_\ell^r) dx = \int_{\mathcal{D}} F(q) \phi_\ell^r dx, \quad \ell = 1, \dots, J_r, \quad (13.29)$$

with reduced-order solutions

$$\tilde{u}(t, x, q) = u^{J_r}(t, x, q) = \sum_{j=1}^{J_r} u_j(t) \phi_j^r(x) \quad (13.30)$$

is sufficiently efficient to permit optimization, uncertainty quantification, and model-based control implementation.

### Initial Conditions

To construct initial conditions for the reduced-order model, it is necessary to project  $u(0, x, q) = u_0(x, q)$  onto  $V^{J_r}$ . To this end, we assume that for each  $t$ ,  $u(t, x)$  is an element in the Hilbert space  $X$  having an inner product  $\langle \cdot, \cdot \rangle$  and norm  $\|\cdot\|$ , as detailed in Definition A.3. The semidiscretization of (13.25) can then be expressed as

$$\begin{aligned} \frac{du^{J_r}}{dt} &= P^{J_r} \mathcal{N}(P^{J_r} u^{J_r}) + P^{J_r} F, \\ u^{J_r}(0, x) &= P^{J_r} u_0(x), \end{aligned} \quad (13.31)$$

where  $P^{J_r}$  is a projection operator from  $X$  onto  $V^{J_r}$  and we suppress parameter dependence.

For Galerkin approximation, the projection operator  $P^{J_r} : X \rightarrow V^{J_r}$  is given by

$$P^{J_r} f = \sum_{j=1}^{J_r} \eta_j \phi_j^r, \quad (13.32)$$

where  $f$  is an arbitrary element in  $X$ . The vector  $\eta = [\eta_1, \dots, \eta_{J_r}]^T$  is given by

$$\eta = M^{-1} \mathcal{F}$$

and

$$M_{ij} = \langle \phi_i^r, \phi_j^r \rangle, \quad \mathcal{F}_i = \langle f, \phi_i^r \rangle$$

for  $i, j = 1, \dots, J_r$ . We note that  $M$  is typically termed the mass matrix. This representation for  $P^{J_r}$  results from the orthogonality property

$$\langle P^{J_r} f - f, g \rangle = 0,$$

which holds for all  $g \in V^{J_r}$ . Details regarding the matrix representation  $L^{J_r}$  for projected linear operators  $L$  can be found in Section II.2 of [25].

### Reduced-Order States and Snapshot Sets

We focus on the construction of reduced-order states when the number of states is significantly larger than the number of parameters,  $J \gg p$ . For models with large parameter dimensionality, one would first apply the parameter selection techniques of Chapter 6 to reduce the dimension of the parameter space.

The crux of reduced-order methods is the construction of a reduced-order basis  $\{\phi_j^r\}$  and projection of the problem onto the space  $V^{J_r} = \text{span}\{\phi_j^r\} \subset V^J$ . We focus on three techniques to construct reduced-order basis functions.

- Eigenfunctions or modes have long been employed as reduced-order basis functions for engineering and science applications, including thermal, structural, acoustic, and fluids problems. We provide an example illustrating this approach in Section 13.3.
- Snapshot-based methods are discussed in Section 13.4. In this approach, high-fidelity numerical codes or experiments are evaluated at several independent variable or parameter values to construct a snapshot set from which a coherent set of reduced-order basis functions is constructed. We focus on proper orthogonal decomposition (POD) methods, which are closely related to Karhunen–Loëve expansions, principal component analysis (PCA), and singular value decompositions (SVDs). As detailed in [48], centroidal Voronoi tesselations (CVTs) constitute an alternative snapshot-based method.
- In the high-dimensional model representation (HDMR) techniques discussed in Section 13.5, reduced-order basis functions are constructed by quantifying first- and second-order interactions between input parameters. In ANOVA-HDMR, this is accomplished by integrating over the parameter space, whereas cut-HDMR representations are constructed by evaluating responses or states at references points  $\bar{q}$ . The relation between HDMR techniques and variance-based global sensitivity analysis is detailed in Chapter 15.

All three techniques produce global basis functions which in turn yield dense systems. This is in contrast to the original systems, which are generally sparse, and necessitates that  $J_r$  be truly small so that solution of the dense system is sufficiently efficient for optimization, uncertainty quantification, or control implementation.

### 13.3 Eigenfunction or Modal Expansions

Eigenfunction or modal expansions constitute a common choice of reduced-order model that has long been employed for simulations and control design in engineering and scientific applications. They are typically constructed from solution or approximate solution of the original problem using separation of variables.

**Example 13.2.** To illustrate, consider the heat equation

$$\begin{aligned}\frac{\partial T}{\partial t} &= \alpha \frac{\partial^2 T}{\partial x^2}, & 0 < x < L, \quad t > 0, \\ T(t, 0) &= T(t, L) = 0, & t > 0, \\ T(0, x) &= T_0(x), & 0 < x < L,\end{aligned}$$

which has the solution

$$T(t, x) = \sum_{j=1}^{\infty} \gamma_j e^{-\alpha \lambda_j^2 t} \sin(\lambda_j x)$$

with eigenvalues  $\lambda_j = \frac{j\pi}{L}$ , eigenfunctions  $X_j(x) = \sin(\lambda_j x)$ , and coefficients

$$\gamma_j = \frac{2}{L} \int_0^L T_0(x) \sin(\lambda_j x) dx.$$

The choice of eigenfunctions as reduced basis functions,  $\phi_j^r(x) = \sin(\frac{j\pi x}{L})$ , thus yields the reduced-order model

$$\tilde{T}(t, x) = \sum_{j=1}^{J_r} T_j(t) \sin\left(\frac{j\pi x}{L}\right),$$

where  $u_j(t)$  are determined by solving

$$\int_0^L \frac{\partial \tilde{T}}{\partial t} \phi_i^r dx + \alpha \int_0^L \frac{\partial \tilde{T}}{\partial x} \frac{d\phi_i^r}{dx} dx = 0$$

for  $i = 1, \dots, J_r$ . This yields the vector system

$$\dot{\mathbb{M}}\vec{T} + \mathbb{K}\vec{T} = 0, \quad T(0) = T_0, \tag{13.33}$$

where  $\vec{T}(t) = [T_1(t), \dots, T_{J_r}(t)]^T$ ,  $\mathbb{M} = \frac{L}{2}I$ , and  $\mathbb{K}$  is a  $J_r \times J_r$  diagonal matrix with diagonal elements  $\frac{L}{2}\{(\frac{\pi}{L})^2, (\frac{2\pi}{L})^2, \dots, (\frac{J_r\pi}{L})^2\}$ . The initial condition vector is  $T_0 = [\gamma_1, \dots, \gamma_{J_r}]^T$ . We note that (13.33) reduces to solving the  $J_r$  independent equations

$$\begin{aligned}\frac{dT_j}{dt} &= -(j\pi/L)^2 T_j, \\ T_j(0) &= \gamma_j\end{aligned}\tag{13.34}$$

for  $j = 1, \dots, J_r$ . As detailed in [98], the error  $T(t, x) - \tilde{T}(t, x)$ , for this example, converges more rapidly than  $e^{-J_r^2 t}$  as  $J_r \rightarrow \infty$  for any  $t$ . Hence very few reduced-order basis functions are required to achieve high accuracy.

For PDEs arising in applications such as those detailed in Chapters 1 and 3, one typically cannot construct analytic eigenfunction relations but instead must seek numerical approximations. For many structural and fluid applications, the use of analytic or numerical eigenfunction expansions is well established for system representation, optimization, and control design, and their use for uncertainty quantification will certainly grow as the field matures.

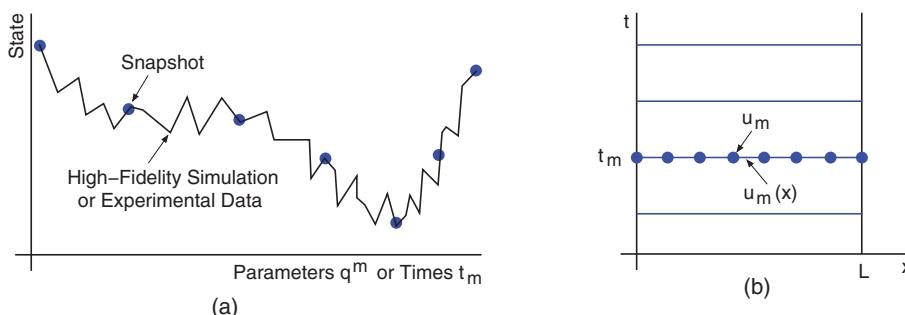
### 13.4 Snapshot-Based Methods including POD

Reduced-order methods based on high-fidelity simulations or experimental measurements are often categorized as Lagrange, Hermite, or Taylor basis methods.

- Lagrange basis methods employ numerical or experimental state solutions evaluated at various parameter values  $q^m$  or independent parameter values such as times  $t_m$ .
- In Hermite methods, one employs both state solutions and their derivatives with respect to parameters—their sensitivities—to construct reduced basis functions. As with Lagrange methods, the state solutions and sensitivities are evaluated at a variety of values for parameters or independent variables.
- Taylor bases are constructed by evaluating the state, sensitivities, and higher derivatives with respect to parameters at a fixed set of parameters and independent variables. This approach is complicated by the complexity of constructing higher-order sensitivities and the fact that their number grows quickly as the order increases.

We focus primarily on Lagrange basis methods.

**Definition 13.3 (Snapshot Set).** The set of numerical or experimental solutions generated at several independent variable values or using various parameter values is termed a snapshot set; see Figure 13.7. For example,  $u^k(t_m, x, q)$  and  $u^k(t, x, q^m)$  are temporal and parameter snapshot sets for the evolution model (13.26).



**Figure 13.7.** (a) Numerical or experimental snapshots and (b) snapshots  $u_m(x)$  and  $u_m \in \mathbb{R}^n$  at times  $t_m$ .

It is clear that the techniques used to sample parameters or determine evaluation values for independent parameters are critical for constructing a basis that is sufficiently *rich* in the sense that it incorporates all expected system dynamics. This is notably critical for two important uses of models: *prediction based on extrapolation* and *control design*. The latter is especially challenging since the application of feedback can introduce dynamics not present in open loop responses. In both cases, one seeks inputs—e.g., impulsive forcing for time-dependent problems—that excite the widest possible range of dynamics and sampling strategies that adequately incorporate this information. Reasonable parameter sampling strategies include greedy sampling [154] as well as the sparse grid and Monte Carlo techniques detailed in Chapter 11. The evaluation at randomly chosen  $q^m$  is also related to the nonlinear parameter selection techniques discussed in Section 6.2.

Whereas there is some theory to guide the choice of snapshot locations [141], the choice of independent variable values at which to generate snapshots often relies on expert knowledge of the application and can be more of an art than an exact science. Statistical theory regarding the *design of experiments* has been used to guide samples for certain applications, but comprehensive algorithms to guide snapshot construction are still lacking, and this constitutes an active area of research.

Both numerical and experimental snapshot sets can contain significant amounts of redundant information from which the basic structure must be extracted when constructing reduced basis functions. This can be accomplished using the POD techniques discussed here or CVTs [48].

### POD with Distributed Observations

We consider first  $M$  distributed observations  $\{u_m(x)\}_{m=1}^M$  numerically or experimentally determined for all  $x$  in a domain  $\mathcal{D}$ . For example,  $u_m(x)$  could represent solutions of the evolutionary PDE (13.4) at times  $t_m$ , as illustrated in Figure 13.7(b). In general,  $x$  can be a generic independent variable.

Since one is typically interested in deviations about a mean value, the first step is to construct a modified snapshot set

$$v_m = u_m - \bar{u}, \quad m = 1, \dots, M, \quad (13.35)$$

where

$$\bar{u} = \langle u \rangle = \frac{1}{M} \sum_{m=1}^M u_m(x)$$

is the average of the ensemble set. The POD technique provides an algorithm for extracting a compressed description of the behavior encapsulated in the redundant set (13.35). The procedure is closely related to the Karhunen–Loëve expansion detailed in Section 10.2.1, PCA, and the SVD discussed in Section 6.1.

To quantify the behavior encapsulated in the modified snapshot set  $\{v_m\}_{m=1}^M$ , the POD technique extracts from this set a coherent structure which has the largest mean square projection onto the set of observations. This is achieved by seeking

basis functions of the form

$$\phi(x) = \sum_{m=1}^M a_m v_m(x), \quad (13.36)$$

where the coefficients  $a_m$  are chosen to maximize

$$\frac{1}{M} \sum_{m=1}^M |\langle v_m, \phi \rangle|^2 \quad \text{subject to } \langle \phi, \phi \rangle = \|\phi\|^2 = 1. \quad (13.37)$$

Here  $\langle \cdot, \cdot \rangle$  and  $\| \cdot \|$  denote the usual  $L^2$  inner product and norm on  $\mathcal{D}$ .

We follow the procedure in [161] and define

$$\begin{aligned} C(x, y) &= \frac{1}{M} \sum_{m=1}^M v_m(x)v_m(y), \\ R\phi &= \int_{\mathcal{D}} C(x, y)\phi(y)dy \end{aligned} \quad (13.38)$$

for  $\phi \in L^2(\mathcal{D})$ . As developed in Exercise 13.1, it follows that

$$\langle R\phi, \phi \rangle = \frac{1}{M} \sum_{m=1}^M |\langle v_m, \phi \rangle|^2 \quad (13.39)$$

and

$$\langle R\phi, \psi \rangle = \langle \phi, R\psi \rangle$$

for all  $\phi, \psi \in L^2(\mathcal{D})$ . Because  $R$  is a symmetric, nonnegative operator on  $L^2(\mathcal{D})$ , the problem of maximizing (13.37) is equivalent to finding the largest eigenvalue  $\lambda$  of

$$R\phi = \lambda\phi \quad \text{subject to } \|\phi\| = 1 \quad (13.40)$$

or

$$\int_{\mathcal{D}} C(x, y)\phi(y)dy = \lambda\phi \quad \text{with } \|\phi\| = 1. \quad (13.41)$$

The substitution of (13.36) for  $\phi$  into (13.41), with  $C$  given by (13.38), yields

$$\sum_{m=1}^M \left[ \sum_{k=1}^M \left( \frac{1}{M} \int_{\mathcal{D}} v_m(y)v_k(y)dy \right) a_k \right] v_m(x) = \sum_{m=1}^M \lambda a_m v_m(x),$$

which can be expressed as the eigenvalue problem

$$KV = \lambda V, \quad (13.42)$$

where

$$K_{mk} = \frac{1}{M} \int_{\mathcal{D}} v_m(x)v_k(x)dx \quad (13.43)$$

and  $V = [a_1, a_2, \dots, a_M]^T$ . Because  $K$  is a nonnegative self-adjoint matrix, it has a complete set of orthogonal eigenvectors

$$V_1 = \begin{bmatrix} a_1^1 \\ \vdots \\ a_M^1 \end{bmatrix}, V_2 = \begin{bmatrix} a_1^2 \\ \vdots \\ a_M^2 \end{bmatrix}, \dots, V_M = \begin{bmatrix} a_1^M \\ \vdots \\ a_M^M \end{bmatrix}$$

with corresponding eigenvalues  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_M \geq 0$ .

The relation (13.37) is maximized by

$$\phi_1(x) = \sum_{m=1}^M a_m^1 v_m(x),$$

where  $a_m^1$  are the elements of the eigenvector  $V_1$  corresponding to the largest eigenvalue  $\lambda_1$ . The remaining functions

$$\phi_j(x) = \sum_{m=1}^M a_m^j v_m(x)$$

are constructed using elements of subsequent eigenvectors  $V_j$  corresponding to the  $j^{th}$  ordered, decreasing eigenvalue. These functions are orthogonal but not orthonormal. It is established in Exercise 13.2 that the functions

$$\phi_j^r(x) = \sum_{m=1}^M \frac{1}{\sqrt{M\lambda_j}} a_m^j v_m(x) , \quad j = 1, \dots, M, \quad (13.44)$$

form an orthonormal basis set. To construct a reduced-order basis set, one employs the first  $J^r$  basis functions where  $J^r \ll M$ .

**Remark 13.4.** The relation (13.41) is precisely the integral equation (5.6) that is solved to construct the Karhunen–Loève representation (5.5) for a correlated, second-order random field  $\alpha(x, \omega)$ . This illustrates one of the ties between the POD and Karhunen–Loève techniques.

### POD with Discrete Observations

The assumption of distributed measurements  $u_m(x)$ ,  $x \in \mathcal{D}$ , facilitates discussion in terms of the familiar  $L^2$  inner product but is rarely achievable in practice. Instead, one typically measures each snapshot at  $n$  points. For example, this would be the case if one has  $n$  spatial sensors to measure flow for  $M$  input parameter values or at  $M$  different times, as illustrated in Figure 13.7(b).

We consider  $M$  snapshots  $u_m \in \mathbb{R}^n$  along with the modified snapshot set  $v_m = u_m - \bar{u}$ , where  $\bar{u}$  again denotes the average of the ensemble set. We summarize the construction of the POD basis functions for the case  $M < n$  and note that the theory parallels that for distributed observations with the Euclidean dot product and norm used instead of the  $L^2$  inner product and norm.

We first construct the  $n \times M$  snapshot matrix

$$A = [v_1, \dots, v_M], \quad (13.45)$$

having rank  $r \leq \min\{n, M\}$ , and  $M \times M$  matrix

$$K = \frac{1}{M} A^T A.$$

Since  $K_{mk} = \frac{1}{M} v_m^T v_k$ , this is the discrete version of the matrix defined in (13.43). The POD basis  $\{\phi_j^r\}_{j=1}^{J_r}$ ,  $J_r \in \{1, \dots, r\}$ , is computed by solving the  $M \times M$  eigenvalue problem

$$KV_j = \lambda_j V_j \quad (13.46)$$

and taking

$$\phi_j^r = \frac{1}{\sqrt{M\lambda_j}} AV_j \quad (13.47)$$

for  $j = 1, \dots, J_r$ . The eigenvalues are taken in decreasing order. We note that the eigenvalue problem (13.46) corresponds to (13.42), whereas the POD basis function  $\phi_j^r$  of (13.47) is the discrete version of  $\phi_j^r(x)$  defined in (13.44).

To relate the POD basis to the SVD, we consider the factorization

$$A = U\Sigma V^T,$$

where  $U = [U_1, \dots, U_n] \in \mathbb{R}^{n \times n}$  and  $V \in \mathbb{R}^{M \times M}$  are orthogonal matrices and  $\Sigma \in \mathbb{R}^{n \times M}$  has the form

$$\Sigma = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix},$$

where  $D = \text{diag}(\sigma_1, \dots, \sigma_r) \in \mathbb{R}^{r \times r}$ . The singular values  $\sigma_j$  are related to the eigenvalues  $\lambda_j$  of  $K$  by the relation  $\sigma_j^2 = M\lambda_j$ . It is established in Definition 1.2 and Theorem 1.1 of [257] that the POD basis functions are the first  $J_r$  left singular values  $U_j$  of  $A$ ; that is,  $\sigma_j^r = U_j$ .

One then employs the reduced representation

$$\tilde{u} = \sum_{j=1}^{J_r} u_j \phi_j^r,$$

where  $J_r \ll M$ . From SVD theory, it follows that the error incurred by using the reduced basis set is

$$E = \sum_{j=J^r+1}^M \sigma_j^2 = M \sum_{j=J^r+1}^M \lambda_j.$$

To achieve a relative error less than  $\delta$ , one chooses the smallest  $J^r$  so that

$$\sum_{j=1}^{J_r} \sigma_j^2 / \sum_{j=1}^M \sigma_j^2 = \sum_{j=1}^{J_r} \lambda_j / \sum_{j=1}^M \lambda_j \geq 1 - \delta.$$

For historical reasons, this technique is often referred to as the *method of snapshots* when  $M < n$ . If  $M \geq n$ , the POD basis is directly computed by solving the eigenvalue problem

$$AA^T U_j = \lambda_j U_j, \quad j = 1, \dots, J_r,$$

and taking  $\phi_j^r = U_j$ .

The reader is referred to [257] for theory, error analysis, and examples illustrating the relation between POD, SVD, and balanced truncation.

## 13.5 High-Dimensional Model Representation (HDMR) Techniques

Consider the nonlinear algebraic model

$$y = f(q) \tag{13.48}$$

of (13.3), where  $q \in \Gamma \subset \mathbb{R}^p$  and  $f$  denotes the output of a high-fidelity simulation code. For this discussion, we assume that the support of each parameter  $q_i$  has been mapped to the interval  $[0, 1]$ —e.g., using the bijective mapping (11.2)—so that  $\Gamma = [0, 1]^p$  is a  $p$ -dimensional hypercube. HDMR techniques for general domains and densities are detailed in Section 15.1.2. In general,  $f(t, x, q)$  can also depend on independent variables such as space and time, but we suppress these dependencies to simplify notation. We note that (13.48) can result from the algebraic representation of a discrete application or the finite element, finite difference, or finite volume approximation of an ODE or PDE.

To construct an HDMR or Sobol representation, for the response  $f(q)$ , one employs the finite, hierarchical expansion

$$\begin{aligned} f(q) = f_0 + \sum_{i=1}^p f_i(q_i) + \sum_{1 \leq i < j \leq p} f_{ij}(q_i, q_j) + \cdots \\ + \sum_{1 \leq i_1 < \dots < i_s \leq p} f_{i_1, \dots, i_s}(q_{i_1}, \dots, q_{i_s}) + \cdots + f_{1, 2, \dots, p}(q_1, \dots, q_p). \end{aligned} \tag{13.49}$$

We establish in Remark 15.3 that the constant function  $f_0$  is the mean response of  $f$ , whereas the first-order univariate functions  $f_i(q_i)$  represent independent contributions due to the individual parameters. The bivariate functions  $f_{ij}(q_i, q_j)$  quantify the interactions of  $q_i$  and  $q_j$  on the response  $y$  with similar interpretations for higher-order interaction terms. The final term  $f_{1, 2, \dots, p}(q_1, \dots, q_p)$  quantifies unincorporated high-order residual effects, thus ensuring that the expansion (13.49) provides an *exact* representation for  $f(q)$ . This is in contrast to the polynomial expansion (10.1) or (10.2), which requires an infinite number of terms to ensure that it is exact.

In practice, one typically employs the approximate expansion

$$f(q) \approx f_0 + \sum_{i=1}^p f_i(q_i) + \sum_{1 \leq i < j \leq p} f_{ij}(q_i, q_j) \tag{13.50}$$

based on the assumption that higher-order interaction terms have a negligible effect on the response. Whereas this assumption is reasonable for a number of applications, its feasibility for specific problems must be ascertained using further physical or numerical analysis. For example, it will not be valid if  $y$  is discontinuous with respect to  $q$ . We note that this decomposition can be interpreted as iteratively fitting along coordinate-aligned subspaces starting with the 0-D subspace associated with  $f_0$ .

**Remark 13.5.** In the context of the projection-based framework discussed in Section 13.2, the terms  $f_0$ ,  $f_i(q_i)$ , and  $f_{ij}(q_i, q_j)$  constitute the reduced-order basis functions  $\phi_j^r$  and the generalized Fourier coefficients are unity. We detail the nature of the projections in the context of specific HDMR representations.

**Remark 13.6.** The representation (13.49) is analogous to the many-body expansions employed in molecular physics to quantify the energy due to atoms in a molecule. The truncated expansion (13.50) represents the case when higher-order interactions have a negligible effect on the energy.

**Remark 13.7.** The motivation for truncating HDMR expansions in high dimensions is related to the concentration of measure phenomenon, which, in its simplest form, states that every Lipschitz function is accurately approximated by a constant function if the dimension  $p$  is sufficiently large; e.g., see page 7 of [47].

The representations (13.49) or (13.50) are not unique, and hence additional structure must be imposed to construct the components  $f_i(q_i)$ ,  $f_{ij}(q_i, q_j)$  and higher-order terms. As proven in [200, 227], each term  $f_{i_1, \dots, i_s}(q_{i_1}, \dots, q_{i_s})$ ,  $s = 0, \dots, p$ , where  $f_0$  corresponds to  $s = 0$ , is uniquely specified by minimizing the functional

$$\int_{\Gamma} \left[ f(q) - \left( f_0 + \sum_{i=1}^p f_i(q_i) + \dots + \sum_{1 \leq i_1 < \dots < i_s \leq p} f_{i_1, \dots, i_s}(q_{i_1}, \dots, q_{i_s}) \right) \right]^2 d\mu(q) \quad (13.51)$$

subject to

$$\int_{[0,1]} f_{i_1, \dots, i_s}(q_{i_1}, \dots, q_{i_s}) dq_{i_k} = 0 \quad (13.52)$$

for  $k = 1, \dots, s$ . The measure  $d\mu(q)$  defines a projection operator from  $\Gamma$  onto subspaces defined by the individual components of the expansion, and hence it defines the particular form of the expansion—readers unfamiliar with measure theory can interpret  $d\mu(q)$  as  $w dq$ , where  $w = [w_{i_1}, \dots, w_{i_s}]$  is a weight and  $dq = [dq_1, \dots, dq_p]$ . The constraint (13.52) ensures that functions  $f_{i_1, \dots, i_r}(q_{i_1}, \dots, q_{i_r})$  and  $f_{i_1, \dots, i_s}(q_{i_1}, \dots, q_{i_s})$  are orthogonal, that is,

$$\int_{\Gamma} f_{i_1, \dots, i_r}(q_{i_1}, \dots, q_{i_r}) f_{i_1, \dots, i_s}(q_{i_1}, \dots, q_{i_s}) d\mu(q) = 0, \quad (13.53)$$

when at least one index differs in the sets  $\{i_1, \dots, i_r\}$  and  $\{i_1, \dots, i_s\}$ .

### 13.5.1 ANOVA-HDMR

We first consider the case when  $d\mu(q)$  is the Lebesgue measure on  $\Gamma$  so that  $d\mu(q) = \prod_{i=1}^p dq_i$ . As detailed in [162, 200, 227], the zeroth-, first-, and second-order terms in this case are

$$\begin{aligned} f_0 &= \int_{\Gamma} f(q) dq, \\ f_i(q_i) &= \int_{\Gamma^{p-1}} f(q) dq_{\sim i} - f_0, \\ f_{ij}(q_i, q_j) &= \int_{\Gamma^{p-2}} f(q) dq_{\sim \{ij\}} - f_i(q_i) - f_j(q_j) - f_0, \end{aligned} \quad (13.54)$$

where  $\Gamma^{p-1} = [0, 1]^{p-1}$  and  $\Gamma^{p-2} = [0, 1]^{p-2}$ . The notation  $q_{\sim \{ij\}}$  denotes the vector having the components of  $q$  except those in the set  $\{ij\}$ ; hence  $dq_{\sim i} = dq_1 \cdots dq_{i-1} dq_{i+1} \cdots dq_p$ .

This is the representation employed in analysis of variance (ANOVA) statistical techniques to determine the variance components of the response. This is central to the global sensitivity analysis detailed in Chapter 15.

As illustrated in the next example, one must include the density  $\rho_Q(q)$  in the representations (13.54) if it is not uniform on  $[0, 1]^p$ . HDMR expansions for general densities are detailed in Section 15.1.2.

**Example 13.8.** To illustrate the analytic construction of  $f_0$ ,  $f_i(q_i)$ , and  $f_{ij}(q_i, q_j)$ , consider the Ishigami function

$$f(q_1, q_2, q_3) = \sin q_1 + a \sin^2 q_2 + b q_3^4 \sin q_1$$

with the density

$$\rho_{Q_i}(q_i) = \begin{cases} \frac{1}{2\pi} & , -\pi \leq q_i \leq \pi, \\ 0 & , \text{else}, \end{cases}$$

as proposed in [122] and detailed in [215]. As established in Exercise 13.3,

$$\begin{aligned} f_0 &= \frac{a}{2}, \\ f_1(q_1) &= \left(1 + \frac{1}{5}b\pi^4\right) \sin q_1 , \quad f_2(q_2) = a \sin^2 q_2 - \frac{a}{2} , \quad f_3(q_3) = 0, \\ f_{12}(q_1, q_2) &= 0 , \quad q_{13}(q_1, q_3) = \left(bq_3^4 - \frac{1}{5}b\pi^4\right) \sin q_1 , \quad f_{23}(q_2, q_3) = 0, \end{aligned} \quad (13.55)$$

and the residual term is  $f_{123}(q_1, q_2, q_3) = 0$ . Hence the representation (13.50) is exact. The orthogonality of the terms is established in Exercise 13.4.

Whereas the representation (13.54) facilitates sensitivity analysis and uncertainty quantification, the construction of  $f_0$  generally requires numerical integration over all of  $\Gamma$ , while one must integrate over all variables except  $q_i$  when constructing

$f_i(q_i)$ . For moderate parameter dimensions, this can be achieved using the sparse grid or Monte Carlo techniques of Chapter 11. The latter yields random sampling (RS)-HDMR techniques. Alternatively, one can circumvent the difficulties associated with high-dimensional quadrature by employing cut-HDMR representations based on values of  $f$  evaluated at nominal points  $\bar{q} \in \Gamma$ .

### 13.5.2 RS-HDMR

Random sampling (RS)-HDMR techniques employ the Monte Carlo techniques of Section 11.1.1 to approximate the high-dimensional integrals in the expressions (13.54). This yields the approximate relation

$$f_0 = \frac{1}{R} \sum_{r=1}^R f(\bar{q}^r) \quad (13.56)$$

for the mean with analogous relations for  $f_i(q_i)$  and  $f_{ij}(q_i, q_j)$ . As detailed in [152], however, the required number of random samples increases exponentially as the order of component functions grows, thus prohibiting this direct approach for many problems.

To reduce the sampling effort, one can represent the first- and second-order interaction terms as

$$\begin{aligned} f_i(q_i) &= \sum_{k=1}^K \alpha_k^i \Psi_k(q_i), \\ f_{ij}(q_i, q_j) &= \sum_{k=1}^K \sum_{\ell=1}^L \beta_{k\ell}^{ij} \Psi_{k\ell}(q_i, q_j), \end{aligned} \quad (13.57)$$

where  $\Psi_k, \Psi_{k\ell}$  are Legendre polynomials scaled to the interval  $[0, 1]$ . It is illustrated in [152] that substitution of the expression

$$f(q) \approx f_0 + \sum_{i=1}^p \sum_{k=1}^K \alpha_k^i \Psi_k(q_i) + \sum_{1 \leq i < j \leq p} \sum_{k=1}^K \sum_{\ell=1}^L \beta_{k\ell}^{ij} \Psi_{k\ell}(q_i, q_j) \quad (13.58)$$

into the minimization problem (13.51) yields the equivalent minimization problems

$$\begin{aligned} \min_{\alpha_k^i} \int_0^1 \left[ f_i(q_i) - \sum_{k=1}^K \alpha_k^i \Psi_k(q_i) \right]^2 dq_i, \\ \min_{\beta_{k\ell}^{ij}} \int_0^1 \int_0^1 \left[ f_{ij}(q_i, q_j) - \sum_{k=1}^K \sum_{\ell=1}^L \beta_{k\ell}^{ij} \Psi_{k\ell}(q_i, q_j) \right]^2 dq_i dq_j \end{aligned}$$

for the coefficients.

The coefficients  $\alpha = [\alpha_1^i, \dots, \alpha_K^i]^T$  are specified by the solution of the matrix system  $A\alpha = b$ , where

$$A_{k\ell} = \int_0^1 \Psi_k(q_i) \Psi_\ell(q_i) dq_i = \gamma_k \quad , \quad k, \ell = 1, \dots, K,$$

with  $\gamma_k = \frac{1}{2k+1} \delta_{k\ell}$ . The components of  $b$  are given by

$$\begin{aligned} b_k &= \int_0^1 f_i(q_i) \Psi_k(q_i) dq_i \\ &= \int_0^1 \left[ \int_{\Gamma^{p-1}} f(q) \prod_{j \neq i} dq_j - f_0 \right] \Psi_k(q_i) dq_i \\ &= \int_{\Gamma} f(q) \Psi_k(q_i) dq, \end{aligned}$$

where the second and third equalities respectively follow from (13.54) and the fact that  $\int_0^1 f_0 \Psi_k(q) dq_i = 0$  for Legendre polynomials with  $k \geq 1$ . The components of  $b$  can then be approximated by

$$b_k \approx \frac{1}{R} \sum_{r=1}^R f(q^r) \Psi_k(q^r),$$

which employs the same samples  $f(q^r)$  used in (13.56) to approximate  $f_0$ . The coefficients  $\alpha_k^i$  are thus approximated by

$$\alpha_k^i \approx \frac{1}{\gamma_k} \frac{1}{R} \sum_{r=1}^R f(q^r) \Psi_k(q_i^r). \quad (13.59)$$

If one employs the tensor product basis functions

$$\Psi_{k\ell}(q_i, q_j) = \Psi_k(q_i) \Psi_\ell(q_j),$$

the coefficients  $\beta_{k\ell}^{ij}$  are specified in a similar manner by

$$\beta_{k\ell}^{ij} \approx \frac{1}{\gamma_{k\ell}} \frac{1}{R} \sum_{r=1}^R f(q^r) \Psi_k(q_i^r) \Psi_\ell(q_j^r), \quad (13.60)$$

where  $\gamma_{k\ell} = \gamma_k \gamma_\ell$ . It is established in Exercise 13.5 that (13.59) and (13.60) are precisely the relations yielded by the discrete projection methods discussed in Chapter 10. The only difference is the ordering of basis functions used to construct the multinomial basis functions.

The connection between RS-HDMR, formulated using orthogonal polynomials, and discrete projection indicates the manner in which this technique can be used to construct projection-based reduced-order models. The use of HDMR expansions for global sensitivity analysis is detailed in Chapter 15.

### 13.5.3 Cut-HDMR

Cut-HDMR avoids the issues associated with high-dimensional quadrature by employing a Dirac measure  $d\mu(\bar{q}) = \prod_{i=1}^p \delta(q_i - \bar{q}_i) dq_i$  when minimizing the functional

(13.51). This yields a representation based on evaluation of  $f(\bar{q})$  at the reference point  $\bar{q} = [\bar{q}_1, \dots, \bar{q}_p]$ . The components are

$$\begin{aligned} f_0 &= f(\bar{q}), \\ f_i(q_i) &= f(q)|_{q=\bar{q}\setminus q_i} - f_0, \\ f_{ij}(q_i, q_j) &= f(q)|_{q=\bar{q}\setminus\{q_i, q_j\}} - f_i(q_i) - f_j(q_j) - f_0, \end{aligned} \quad (13.61)$$

where the notation  $q = \bar{q} \setminus q_i$  indicates that the components of  $q$  other than  $q_i$  are set equal to those of the reference point; that is,

$$f(q)|_{q=\bar{q}\setminus q_i} = f(\bar{q}_1, \dots, \bar{q}_{i-1}, q_i, \bar{q}_{i+1}, \dots, \bar{q}_p).$$

The first-order projection in this case is defined by

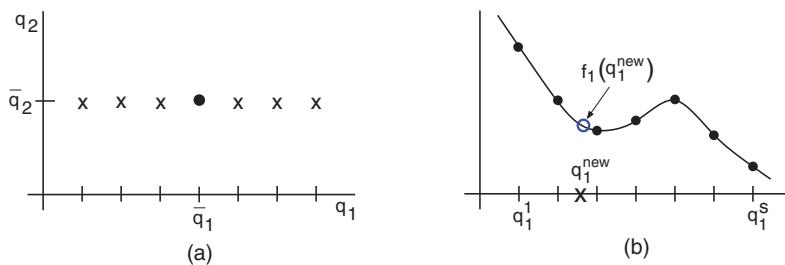
$$P^i f(q_i) = f(q)|_{q=\bar{q}\setminus q_i},$$

and higher-order projections are defined in an analogous manner. The name results from the property that component functions are defined along cut lines, planes, and hyperplanes through the reference point  $\bar{q}$ , as illustrated in Figure 13.8. This approach is also termed anchored HDMR.

The choice of reference or anchor points  $\bar{q}$  is critical for cut-HDMR, especially when only first- and second-order terms are employed. As detailed in [90], choices include the mean value so that  $f(\bar{q}) \approx \bar{f}$ , the centroid in uniform parameter spaces, or the centroid of sparse quadrature grids. Alternatively, one can employ multiple anchor points [152] or anchor points specified through the Morris screening techniques detailed in Chapter 15.

To employ the cut-HDMR representation as a surrogate model, one evaluates the component functions  $f_i(q_i^m)$  and  $f_{ij}(q_i^m, q_j^m)$  at  $s$  discrete values  $\{q_i^m\}_{m=1}^s$  along cut lines and planes through  $\bar{q}$ , as illustrated in Figure 13.8. The cost to construct the zeroth-, first-, and second-order terms is

$$1 + p(s-1) + \frac{p(p-1)(s-1)^2}{2},$$



**Figure 13.8.** (a) Reference point  $(\bar{q}_1, \bar{q}_2)$  and cut points  $q_i^m$ , and (b) interpolating function  $f_1(q_1)$ .

which exhibits polynomial growth in  $p$  and  $s$  as compared with the exponential growth rate  $s^p$  associated with tensor product sampling.

The first- and second-order interaction terms for arbitrary points  $q_i^{new}, q_j^{new}$  can then be represented as

$$\begin{aligned} f_i(q_i^{new}) &= \sum_{m=1}^s L_m(q_i^{new}) f(q) \Big|_{q=\bar{q} \setminus q_i^m} - f_0, \\ f_{ij}(q_i^{new}, q_j^{new}) &= \sum_{m=1}^s \sum_{n=1}^s L_{mn}(q_i^{new}, q_j^{new}) f(q) \Big|_{q=\bar{q} \setminus \{q_i^m, q_j^n\}} \\ &\quad - f_i(q_i^{new}) - f_j(q_j^{new}) - f_0, \end{aligned} \tag{13.62}$$

where

$$L_m(q_i) = \prod_{\substack{k=1 \\ k \neq m}}^s \frac{q_i - q_i^k}{q_i^m - q_i^k}$$

is the 1-D Lagrange interpolating polynomial that satisfies

$$L_m(q_i^k) = \delta_{mk} \quad , \quad 1 \leq m, k \leq s.$$

Similarly,  $L_{mn}(q_i, q_j) = L_m(q_i) \otimes L_n(q_j)$  is the 2-D tensor product interpolating polynomial. A 1-D interpolating function constructed in this manner is illustrated in Figure 13.8(b).

**Example 13.9.** We again consider the function

$$f(q_1, q_2, q_3) = \sin q_1 + a \sin^2 q_2 + b q_3^4 \sin q_1$$

of Example 13.8 where  $q \in [-\pi, \pi]^3$ . For the centroid reference point  $\bar{q} = [0, 0, 0]$ , the cut-HDMR components are

$$\begin{aligned} f_0 &= 0, \\ f_1(q_1) &= \sin q_1 \quad , \quad f_2(q_2) = a \sin^2 q_2 \quad , \quad f_3(q_3) = 0, \\ f_{12}(q_1, q_2) &= 0 \quad , \quad q_{13}(q_1, q_3) = b q_3^4 \sin q_1 \quad , \quad f_{23}(q_2, q_3) = 0, \end{aligned} \tag{13.63}$$

which differ from the ANOVA-HDMR components in (13.55). For example, the value  $f_0 = 0$  differs greatly from the actual mean  $f_0 = \frac{a}{2}$ . However, the representation

$$f(q) = f_0 + \sum_{i=1}^3 f_i(q_i) + \sum_{i \leq i < j \leq 3} f_{ij}(q_1, q_2)$$

is still exact. This illustrates the inaccuracy in the statistical interpretation (15.15) that can occur when truncated cut-HDMR expansions with a single reference point are employed for highly nonlinear functions.

### 13.5.4 ANOVA-HDMR Based on Cut-HDMR Expansions

It is illustrated in Chapter 15 that ANOVA-HDMR expansion can be used to construct Sobol indices for global sensitivity analysis. However, this comes at the cost of high-dimensional integrals, which necessitates random sampling techniques or representations based on Legendre polynomials. Cut-HDMR techniques are significantly more efficient but do not directly yield expectation or variance relations since they replace integration with single point evaluation. Here we construct an ANOVA-HDMR expansion based on cut-HDMR terms which can be used for global sensitivity analysis. To simplify the discussion, we consider uniform densities on the unit hypercube  $\Gamma = [0, 1]^p$ . The extension to iid random variables with general densities is illustrated in Section 15.1.2.

For an arbitrary parameter  $q = [q_1, \dots, q_p]^T$  and anchor point  $\bar{q} = [\bar{q}_1, \dots, \bar{q}_p]^T$ , the second-order ANOVA-HDMR based on the cut-HDMR expansion

$$f^{cut}(q) = f_0^{cut} + \sum_{i=1}^p f_i^{cut}(q_i) + \sum_{1 \leq i < j \leq p} f_{ij}^{cut}(q_i, q_j) \quad (13.64)$$

is

$$f^{\text{ANOVA}}(q) = f_0^{\text{ANOVA}} + \sum_{k=1}^p f_k^{\text{ANOVA}}(q_k) + \sum_{1 \leq k < \ell \leq p} f_{k\ell}^{\text{ANOVA}}(q_k, q_\ell), \quad (13.65)$$

where

$$\begin{aligned} f_0^{\text{ANOVA}} &= \int_{\Gamma} f^{cut}(q) dq, \\ f_k^{\text{ANOVA}}(q_k) &= \int_{\Gamma^{p-1}} f^{cut}(q) dq_{\sim k} - f_0^{\text{ANOVA}}, \\ f_{k\ell}^{\text{ANOVA}}(q_k, q_\ell) &= \int_{\Gamma^{p-2}} f^{cut}(q) dq_{\sim \{k\ell\}} - f_k^{\text{ANOVA}}(q_k) - f_\ell^{\text{ANOVA}}(q_\ell) - f_0^{\text{ANOVA}}. \end{aligned}$$

To facilitate implementation, we employ the relations (13.62) to represent the first- and second-order cut-HDMR functions in terms of Lagrange expansions. We also employ the notation

$$S_1(q) = \sum_{i=1}^p f_i^{cut}(q_i) \quad , \quad S_2(q) = \sum_{1 \leq i < j \leq p} f_{ij}^{cut}(q_i, q_j)$$

to simplify the discussion of these sums.

The constant ANOVA term  $f_0 = \mathbb{E}(Y)$  has the representation

$$f_0^{\text{ANOVA}} = f_0^{cut} + I_1 + I_2, \quad (13.66)$$

where

$$\begin{aligned} I_1 &= \int_{\Gamma} S_1(q) dq \\ &= \sum_{i=1}^p \int_0^1 \left[ \sum_{m=1}^s L_m(q_i) f(q)|_{q=\bar{q} \setminus q_i^m} - f_0^{cut} \right] dq_i \end{aligned}$$

involves approximation of  $p$  1-D integrals and

$$I_2 = \sum_{1 \leq i < j \leq p} \int_{\Gamma} f_{ij}^{cut}(q_i, q_j) dq_i dq_j$$

requires 2-D quadrature. This can be accomplished using Gauss-Legendre quadrature techniques where the points  $q_i^m$  are chosen to be the roots of the Legendre polynomials.

The first-order terms are

$$\begin{aligned} f_k^{\text{ANOVA}}(q_k) &= \int_{\Gamma^{p-1}} [f_0^{cut} + S_1(q) + S_2(q)] dq_{\sim k} - f_0^{\text{ANOVA}} \\ &= \int_{\Gamma^{p-1}} [S_1(q) + S_2(q)] dq_{\sim k} - \int_{\Gamma} [S_1(q) + S_2(q)] dq. \end{aligned}$$

It is established in Exercise 13.6 that

$$\int_{\Gamma^{p-1}} S_1(q) dq_{\sim k} - \int_{\Gamma} S_1(q) dq = f_k^{cut}(q_k) - \int_0^1 f_k^{cut}(q_k) dq_k \quad (13.67)$$

and

$$\begin{aligned} \int_{\Gamma^{p-1}} S_2(q) dq_{\sim k} - \int_{\Gamma} S_2(q) dq &= \sum_{1 \leq i < k} \left[ \int_0^1 f_{ik}^{cut}(q_i, q_k) dq_i - \int_0^1 \int_0^1 f_{ik}^{cut}(q_i, q_k) dq_i dq_k \right] \\ &\quad + \sum_{k < j \leq p} \left[ \int_0^1 f_{kj}^{cut}(q_k, q_j) dq_j - \int_0^1 \int_0^1 f_{kj}^{cut}(q_k, q_j) dq_k dq_j \right] \end{aligned} \quad (13.68)$$

so that

$$\begin{aligned} f_k^{\text{ANOVA}}(q_k) &= f_k^{cut}(q_k) - \int_0^1 f_k^{cut}(q_k) dq_k \\ &\quad + \sum_{1 \leq i < k} \left[ \int_0^1 f_{ik}^{cut}(q_i, q_k) dq_i - \int_0^1 \int_0^1 f_{ik}^{cut}(q_i, q_k) dq_i dq_k \right] \\ &\quad + \sum_{k < j \leq p} \left[ \int_0^1 f_{kj}^{cut}(q_k, q_j) dq_j - \int_0^1 \int_0^1 f_{kj}^{cut}(q_k, q_j) dq_k dq_j \right]. \end{aligned} \quad (13.69)$$

Similarly, it is established in Exercise 13.7 that

$$\begin{aligned} f_{kl}^{\text{ANOVA}}(q_k, q_l) &= f_{kl}^{cut}(q_k, q_l) - \int_0^1 f_{kl}^{cut}(q_k, q_l) dq_k \\ &\quad - \int_0^1 f_{kl}^{cut}(q_k, q_l) dq_l + \int_0^1 \int_0^1 f_{kl}^{cut}(q_k, q_l) dq_k dq_l. \end{aligned} \quad (13.70)$$

The relations (13.66), (13.69), and (13.70) can be employed in the manner described in Chapter 15 to construct Sobol global sensitivity indices.

**Remark 13.10.** The accuracy of the expansion (13.65) is limited by the accuracy of the cut-HDMR (13.64). Hence while it is advantageous for constructing the Sobol global sensitivity indices discussed in Chapter 15, it will not provide the full accuracy attained using  $f_0$ ,  $f_i(q_i)$ , and  $f_{ij}(q_i, q_j)$  given by (13.54) or the Legendre polynomial representations detailed in Section 13.5.2. This is illustrated in Exercise 13.8.

## 13.6 Surrogate-Based Bayesian Model Calibration

In Chapter 8, we detailed techniques for constructing input densities through either the direct application of Bayes' relation

$$\pi(v|q) = \frac{\pi(v|q)\pi_0(q)}{\int_{\mathbb{R}^p} \pi(v|q)\pi_0(q)dq}, \quad (13.71)$$

where  $\pi_0(q)$ ,  $\pi(v|q)$ , and  $\pi(q|v)$  respectively denote the prior density, likelihood, and posterior density, or by constructing Markov chains whose stationary distribution is the posterior density. Based on the assumption that measurement errors are iid and normally distributed,  $\varepsilon_i \sim N(0, \sigma^2)$ , we employ the likelihood function

$$\pi(v|q) = L(q, \sigma^2|v) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-SS_q/2\sigma^2}, \quad (13.72)$$

where

$$SS_q = \sum_{i=1}^n [v_i - f_i(q)]^2 \quad (13.73)$$

is the sum of squares error. Here  $v_i$  and  $f_i(q)$  respectively denote the data and corresponding model response; i.e.,  $f_i(q) = f(t_i, q)$  or  $f_i(q) = f(x_i, q)$  for evolutionary or stationary processes. The often insurmountable difficulty associated with approximating the integral in (13.71) or running sufficiently long MCMC chains is the computational expense associated with constructing the required number of model solutions  $f_i(q)$  for complex, high-fidelity models such as multiphysics or multidimensional nonlinear PDEs. Here we discuss the use of surrogate models for Bayesian model calibration.

The most direct approach is to replace high-fidelity evaluations  $f_i(q)$  with the highly efficient surrogate evaluations  $\tilde{f}_i(q)$  and employ the surrogate likelihood function

$$\tilde{\pi}(v|q) = \tilde{L}(q, \sigma^2|v) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\tilde{SS}_q/2\sigma^2}, \quad (13.74)$$

where

$$\tilde{SS}_q = \sum_{i=1}^n [v_i - \tilde{f}_i(q)]^2. \quad (13.75)$$

In this case, the surrogate samples from the prior in a manner analogous to that discussed in Chapter 10 for sampling from the posterior to propagate input uncertainties through the model. Furthermore, sampling from the prior is typically easier

than sampling from the posterior in the sense that the assumption of mutually independent parameters is always satisfied when using noninformative priors defined on the admissible parameter space. Hence one can represent the joint prior as

$$\pi_0(q) = \prod_{i=1}^p \pi_{0_i}(q_i),$$

where  $\pi_{0_i}$  is the marginal prior density for  $Q_i$ . As detailed in Section 5.2, this is in contrast to the posterior density, which typically cannot be represented as a product of the marginal densities due to correlation among parameters.

This approach is detailed in the context of response surface models in [20] and Gaussian process models in [134]. Convergence analysis and examples illustrating the use of stochastic Galerkin and collocation representations for efficient Bayesian model calibration are reported in [167, 168].

To illustrate, consider the reduced-order model

$$\tilde{f}_i(q) = u^{J_r}(t_i, x_i, q) = \sum_{j=1}^{J_r} u_j(t_i) \phi_j^r(x_i) \quad (13.76)$$

given by (13.30). Once the coefficients  $u_j(t)$  have been determined by solving (13.29), one can evaluate the surrogate likelihood (13.74) for quadrature values  $q^r$  required to evaluate (13.71) or chain values  $q^k$  generated by MCMC algorithms. This permits significant sampling with a computational cost that is a fraction of that required for the high-fidelity simulations of (13.28).

## 13.7 Notes and References

The discussion of regression or interpolation-based models is necessarily brief, and readers are referred to [39, 67, 236] for details regarding the statistical theory of response surface functions and kriging models and [84, 85, 89] for discussion illustrating the use of surrogate models for engineering design and optimization. Further details regarding the construction of surrogate models based on Gaussian process representations can be found in [199]. The use of response surface models, Gaussian process models, and stochastic collocation techniques to construct surrogate likelihoods for Bayesian model calibration is detailed in [20, 45, 108, 133, 134, 168, 198]. Software for constructing Gaussian process and kriging models is available in the Sandia package DAKOTA [4, 5]. Response surface models can also be constructed using the MATLAB SUMO toolbox.

There is a large literature on POD representations for control applications, and readers are referred to [17, 140, 141, 160, 161, 195, 203, 257] and the included references for details and error analysis regarding POD and adaptive POD. A general discussion of POD is provided in [142]. The more recent use of POD representations to facilitate surrogate likelihood implementation and uncertainty propagation is illustrated in [46, 87, 88, 154]. This includes the use of greedy sampling algorithms to adequately cover large-dimensional sample spaces.

HDMR expansions have proven advantageous both as surrogate models [8, 9, 152, 153, 162, 200, 201] and for the structure they provide for global sensitivity analysis using variance-based Sobol indices [177, 215, 217]. We detail this latter role in Chapter 15.

## 13.8 Exercises

**Exercise 13.1.** For  $R : L^2(\mathcal{D}) \rightarrow L^2(\mathcal{D})$  given by (13.38) and  $\phi$  expanded in the manner (13.36), show that

$$\langle R\phi, \phi \rangle = \frac{1}{M} \sum_{i=1}^M |\langle v_i, \phi \rangle|^2$$

and

$$\langle R\phi, \psi \rangle = \langle \phi, R\psi \rangle$$

for all  $\phi, \psi \in L^2(\mathcal{D})$ . This establishes that  $R$  is a nonnegative and symmetric operator on  $L^2(\mathcal{D})$ .

**Exercise 13.2.** For the basis functions  $\phi_j^r$  defined in (13.44), show that

$$\langle \phi_j^r, \phi_k^r \rangle = \delta_{jk}.$$

This establishes that they constitute an orthonormal basis set.

**Exercise 13.3.** For  $f(q) = \sin q_1 + a \sin^2 q_2 + b q_3^4 \sin q_1$  considered in Example 13.8, use the relations (13.54) to establish the relations (13.55) for the ANOVA-HDMR interaction terms.

**Exercise 13.4.** Show that the terms  $f_0$ ,  $f_i(q_i)$ , and  $f_{ij}(q_i, q_j)$  constructed in Exercise 13.3 are orthogonal.

**Exercise 13.5.** Consider the spring model

$$\begin{aligned} m \frac{d^2 z}{dt^2} + c \frac{dz}{dt} + kz &= f_0 \cos(\omega_F t), \\ z(0) &= z_0, \quad \frac{dz}{dt}(0) = z_1 \end{aligned}$$

with the parameters  $Q = [m, c, k]$ . As illustrated in Examples 3.2 and 9.7, the response

$$y(\omega_F, Q) = \frac{1}{\sqrt{(k - m\omega_F^2)^2 + (c\omega_F)^2}}$$

quantifies the relative magnitude of the displacement as a function of the driving frequency  $\omega_F$ . In Section 10.4, we illustrated the use of the discrete projection method to propagate means and variances in the response.

Show that the RS-HDMR techniques of Section 13.5.2 yield the same relations as discrete projection when Legendre polynomials are used to represent the first-

and second-order interactions. The only difference is the ordering of basis functions used to construct the multinomial basis functions.

**Exercise 13.6.** Establish the relations (13.67) and (13.68) for  $\int_{\Gamma^{p-1}} S_1(q)dq_{\sim k} - \int_{\Gamma} S_1(q)dq$  and  $\int_{\Gamma^{p-1}} S_2(q)dq_{\sim k} - \int_{\Gamma} S_2(q)dq$ .

**Exercise 13.7.** Establish the relation (13.70) for  $f_{k\ell}^{\text{ANOVA}}(q_k, q_\ell)$ .

**Exercise 13.8.** Use the expressions (13.66), (13.69), and (13.70) to compute the ANOVA-HDMR components for the function

$$f(q_1, q_2, q_3) = \sin q_1 + a \sin^2 q_2 + b q_3^4 \sin q_1$$

based on the cut-HDMR components (13.63). Note that they differ from the true ANOVA-HDMR components (13.55) and hence will be inaccurate if used for global sensitivity analysis based on Sobol indices. This illustrates Remark 13.10, which notes that the accuracy of ANOVA components computed in this manner will be limited by the accuracy of the cut-HDMR terms.

## Chapter 14

# Local Sensitivity Analysis

The term *sensitivity analysis* has different connotations in various modeling communities, and, even in mathematics, its meaning has evolved significantly in the last 25 years. The objective of sensitivity analysis can be broadly viewed as quantifying the relative contributions due to individual parameters or inputs and determining how variations in parameters affect measured responses. The reasons for sensitivity analysis include the following:

- ascertain whether the model is robust or overly fragile with regard to various parameters;
- determine whether the model can be simplified by fixing insensitive parameters;
- specify regimes in the parameter space that optimally impact responses or their uncertainties;
- guide experimental design to determine measurement regimes that have the greatest impact on parameter or response sensitivity.

The methods for sensitivity analysis are typically classified as local or global.

**Definition 14.1 (Local Sensitivity Analysis).** Local sensitivity analysis focuses on the variability of the response when parameters or inputs are perturbed about a nominal value. This is typically achieved, and often defined, by the derivative  $\frac{\partial y}{\partial q_i}$  of the response with respect to the individual parameters. This technique is often employed to determine insensitive parameters that can be fixed in models since their variation minimally influences outputs. Local sensitivity analysis is also central to optimization, adjoint methods, and model calibration. Whereas local theory constitutes the majority of sensitivity analysis in the literature, the examples in Chapter 15 illustrate limitations of this approach when investigating the global behavior of nonlinearly parameterized problems.

There are two differing concepts of global sensitivity analysis. The first, which is detailed in [50], focuses on determining all critical points for a system—e.g.,

bifurcations—and performing local sensitivity analysis about these points. The second is more statistical in nature and focuses on the relation between input and output uncertainties for a model. Due to its direct ties to uncertainty quantification, we focus on the latter.

**Definition 14.2 (Global Sensitivity Analysis).** One objective of global sensitivity analysis is to ascertain how uncertainty in model outputs can be apportioned to uncertainties in model inputs, taken either singly or in combination, when considered over the entire range of input values. This analysis is focused solely on properties of the model and does not rely on experimental data. Hence global sensitivity analysis of this form complements uncertainty quantification, which focuses on the determination of input and response distributions or confidence intervals based on measured data. As detailed in Chapter 15, global sensitivity analysis techniques can be broadly categorized as regression, variance, or screening-based methods.

We employed local sensitivity analysis for several facets of model calibration and uncertainty quantification. In Chapter 7, we illustrated that the sensitivity matrix was required to construct the Fisher information matrix for nonlinearly parameterized models. Similar analysis in Chapter 8 illustrated the role played by the sensitivity equations for formulating the covariance matrix used to construct initial proposal densities for MCMC routines. It was illustrated in Chapter 9 that the product of the sensitivity matrix, covariance matrix, and transpose of the sensitivity matrix provided the response covariance matrix for linear perturbations. Local sensitivity analysis is also central to the parameter selection techniques of Chapter 6.

In Section 7.3.1, we noted three techniques that could be used to construct local sensitivities: finite difference approximations, solution of sensitivity equations, or automatic differentiation. The following example illustrates these techniques and motivates the more general local sensitivity approach detailed in this chapter.

**Example 14.3.** Consider the nonlinear spring model

$$\begin{aligned} \frac{d^2z}{dt^2} + C \left( \frac{dz}{dt} \right)^2 + K z &= 0, \\ z(0) = 2, \quad \frac{dz}{dt}(0) &= 0 \end{aligned} \tag{14.1}$$

with the responses

$$y(t) = [1 \ 0] \begin{bmatrix} z \\ \dot{z} \end{bmatrix} = z(t) \tag{14.2}$$

or

$$y = \int_0^{t_f} \gamma(t) z(t) dt. \tag{14.3}$$

For the latter case,  $\gamma$  acts as a filter or weight over the time interval of interest.

For the response (14.2), the parameters are  $q = [K, C]$ . To construct sensitivity equations, one differentiates (14.1) with respect to each parameter and switches the order of integration to obtain

$$\begin{aligned} \frac{d^2 z_K}{dt^2} + 2C \frac{dz}{dt} \frac{dz_K}{dt} + K z_K &= -z \quad , \quad z_K(0) = 0, \quad \frac{dz_K}{dt}(0) = 0, \\ \frac{d^2 z_C}{dt^2} + 2C \frac{dz}{dt} \frac{dz_C}{dt} + K z_C &= -\left(\frac{dz}{dt}\right)^2 \quad , \quad z_C(0) = 0, \quad \frac{dz_C}{dt}(0) = 0, \end{aligned} \quad (14.4)$$

where  $z_K(t) \equiv \frac{\partial z}{\partial K}$  and  $z_C(t) \equiv \frac{\partial z}{\partial C}$ . Because there is no analytic solution, one must use numerical routines to approximate  $z(t)$ ,  $z_K(t)$ , and  $z_C(t)$ . The response sensitivities  $\frac{\partial y}{\partial K}$  and  $\frac{\partial y}{\partial C}$  are then given by (14.2).

Alternatively, one can approximate the response sensitivities using the finite difference relations

$$\frac{\partial y}{\partial K}(t) \approx \frac{z(t, K + h_K, C) - z(t, K, C)}{h_K}, \quad \frac{\partial y}{\partial C}(t) \approx \frac{z(t, K, C + h_C) - z(t, K, C)}{h_C},$$

which are specific cases of (7.35). As illustrated in Exercise 7.2, the accuracy of this approach is highly dependent on the choices of  $h_K$  and  $h_C$ , which must be scaled according to the magnitude of parameters. Hence while this approach is the simplest conceptually, it is often the least effective.

For evolution problems of this type, automatic differentiation (AD) software constitutes a third option for computing sensitivities. This approach exploits the fact that all computer programs can be decomposed into a combination of elementary arithmetic operations—e.g., addition, subtraction, multiplication, and division—and function evaluations—e.g., exponentials, logarithms, cosines, and sines. Automatic differentiation (AD) software computes derivatives of arbitrary order by applying the chain rule to these operations. AD routines have been developed for MATLAB, C/C++, Fortran, and Python and have been incorporated into packages such as Trilinos. We emphasize that automatic differentiation is fundamentally different from *symbolic differentiation*, where algorithms manipulate the mathematical model expressions.

Now consider the model (14.1) with the response (14.3) and parameter set  $q = [K, C, \gamma(t)]$ . Sensitivity equation and AD techniques cannot be directly employed to construct  $\frac{\partial y}{\partial \gamma}$ , and finite difference approximations will have limited accuracy. Furthermore, these techniques do not provide the capability for quantifying the sensitivity of  $y$  when all three parameters are simultaneously varied. This motivates the more general local sensitivity analysis techniques detailed in this chapter.

We discuss techniques, based on the Gâteaux variations, for sensitivity analysis that are applicable to evolution and stationary differential equations, integral equations, and algebraic systems. Furthermore, the techniques can be used to quantify the local response sensitivity when multiple inputs are perturbed simultaneously.

To simplify the analysis, we focus on systems that exhibit a linear state dependence which permits the decoupling of state and adjoint equations. As noted in Section 14.3, similar analysis can be applied to nonlinear state-dependent systems, but the coupling between state and adjoint equations complicates their solution.

To motivate attributes and issues associated with the forward sensitivity analysis procedure (FSAP) and adjoint sensitivity analysis procedure (ASAP), we consider the neutron transport models detailed in Example 3.6. In Section 14.2, we summarize and illustrate the functional analytic theory of [50], which is applicable to a broad class of linear and nonlinear state-dependent models.

## 14.1 Motivating Examples—Neutron Diffusion

To illustrate the FSAP and ASAP, we consider the neutron diffusion equation

$$\begin{aligned} A_a \varphi - D \frac{d^2 \varphi}{dx^2} &= S \quad , \quad x \in (-a, a), \\ \varphi(\pm a) &= 0 \end{aligned} \tag{14.5}$$

with a response

$$y = A_d \varphi(b) \tag{14.6}$$

measured using a detector of width  $A_d$  located at  $x = b$ . As detailed in Example 3.6,  $\varphi$ ,  $D$ ,  $A_a$ , and  $S$  respectively denote the neutron flux, a diffusion coefficient, a macroscopic absorption cross-section, and a constant distributed source. The parameter set is

$$q = [A_a, D, S, A_d], \tag{14.7}$$

and it is assumed that we know nominal values  $\bar{q}$  and variations  $\delta q$ . For example, one might take  $\bar{q}$  to be mean values and  $\delta q$  to be one standard deviation for each of the components. Other choices are possible, so, throughout this chapter,  $\bar{q}$  should be interpreted as nominal rather than solely as mean values of  $q$ . For the nominal parameter values, the solution to (14.5) is

$$\bar{\varphi}(x) = \frac{\bar{S}}{\bar{A}_a} \left( 1 - \frac{\cosh(xk)}{\cosh(ak)} \right) , \quad k = \sqrt{\bar{A}_a / \bar{D}}. \tag{14.8}$$

### 14.1.1 Matrix System

We illustrated in Example 3.6 that a central difference Taylor approximation for the second derivative yields the observed matrix system

$$\begin{aligned} A(q)\phi &= s(q), \\ y &= \mathcal{C}^T(q)\phi = \mathcal{C}^T(q)A^{-1}(q)s(q), \end{aligned}$$

where  $\phi = [\varphi_1, \dots, \varphi_{N-1}]^T$  with  $\varphi_i \approx \varphi(x_i)$  and  $A(q), \mathcal{C}(q), s(q)$  are given in (3.29) and (3.32). The mean parameter values  $\bar{q}$  yield nominal values  $\bar{A} = A(\bar{q})$ ,  $\bar{s} = s(\bar{q})$ , and  $\bar{\mathcal{C}} = \mathcal{C}(\bar{q})$  as well as the nominal solution  $\bar{\phi} = \bar{A}^{-1}\bar{s}$  and response  $\bar{y} = \bar{\mathcal{C}}^T\bar{\phi}$ . Furthermore, parameter variations  $\delta q$  produce corresponding variations  $\delta A$ ,  $\delta s$ ,  $\delta \mathcal{C}$ ,  $\delta \phi$ , and  $\delta y$  in the system components, solution, and response.

We now illustrate the FSAP and ASAP to quantify the effect of the parameter variations  $\delta q$  on perturbations  $\delta \phi$  and  $\delta y$  in the solution and response.

### Forward Sensitivity Analysis Procedure (FSAP)

From

$$y(q) = \mathcal{C}^T(q)A^{-1}(q)s(q),$$

we can formally express  $\delta y$  as

$$\delta y = \frac{\partial y}{\partial A_a}\delta A_a + \frac{\partial y}{\partial D}\delta D + \frac{\partial y}{\partial S}\delta S + \frac{\partial y}{\partial A_d}\delta A_d.$$

The difficulty is that computation of the sensitivities  $\frac{\partial y}{\partial q_k}$  requires differentiation of  $A^{-1}(q)$ , which typically does not have a closed-form representation.

Instead, we use the Gâteaux variation, defined in Appendix A, to directly construct forward sensitivity equations and response perturbations  $\delta y$ . We recall that the Gâteaux variation is a functional analytic generalization of the directional derivative from multivariable calculus. Hence it quantifies responses of a functional, evaluated at a point in the vector space, in various directions.

From Definition A.7, the Gâteaux variation of  $y(q) - \mathcal{C}^T(q)\phi(q) = 0$ , evaluated at the nominal parameter and response values  $\bar{q}$  and  $\bar{y}$ , is

$$\left\{ \frac{d}{d\varepsilon} [(\bar{y} + \varepsilon\delta y) - (\bar{\mathcal{C}}^T + \varepsilon\delta\mathcal{C}^T)(\bar{\phi} + \varepsilon\delta\phi)] \right\}_{\varepsilon=0} = 0, \quad (14.9)$$

which yields the response perturbation relation

$$\delta y = \bar{\mathcal{C}}^T\delta\phi + \delta\mathcal{C}^T\bar{\phi}. \quad (14.10)$$

The terms  $\bar{\mathcal{C}}$  and  $\delta\mathcal{C}$  can be computed from (3.32) using known values for  $\bar{A}_d$  and  $\delta A_d$ . Similarly,  $\bar{\phi} = A^{-1}(\bar{q})s(\bar{q})$  can be computed using nominal parameter values. To compute  $\delta\phi$ , we apply the Gâteaux variation to  $A(q)\phi - s(q) = 0$  to obtain

$$\left\{ \frac{d}{d\varepsilon} [(\bar{A} + \varepsilon\delta A)(\bar{\phi} + \varepsilon\delta\phi) - (\bar{s} + \varepsilon\delta s)] \right\}_{\varepsilon=0} = 0.$$

This yields the sensitivity equations

$$\bar{A}\delta\phi = \delta s - \delta A\bar{\phi} \quad (14.11)$$

that are solved to obtain  $\delta\phi$ . In summary, one solves the nominal system  $\bar{A}\phi = \bar{s}$  and sensitivity system (14.11) in the FSAP to obtain  $\bar{\phi}$  and  $\delta\phi$  and hence compute the perturbation response (14.10).

The disadvantage of this approach is that (14.11) must be re-solved to accommodate information resulting from new data. For example, changes in  $\delta D, \delta S, \delta A_a$ , or  $\delta A_d$ , due to additional measurements or further Bayesian analysis, produce corresponding changes in  $\delta A$  and  $\delta s$ . For large systems, the repeated solution of (14.11), to incorporate new information, is often prohibitive and is avoided in the ASAP.

### Adjoint Sensitivity Analysis Procedure (ASAP): Perturbation Approach

We will illustrate the ASAP from two perspectives. For this example, the first is more fundamental, and it is in the spirit of the general functional analytic approach summarized in Section 14.2. The second illustrates the variational perspective, which, for general problems, requires fewer results from functional analysis and can provide a more natural framework for incorporating constraints.

We begin by defining the adjoint sensitivity equation

$$\bar{A}^T \psi = \bar{\mathcal{C}}, \quad (14.12)$$

where  $\psi$  is termed the adjoint function and  $\bar{A}^T$  is the transpose of  $\bar{A}$ . We will motivate (14.12) in subsequent discussion.

We then consider the dot or inner product

$$\langle \psi, \delta s - \delta A \bar{\phi} \rangle = \langle \psi, \bar{A} \delta \phi \rangle \quad (14.13)$$

obtained by multiplying (14.11) by  $\psi^T$ . From the bilinear identity

$$\langle v, z \rangle = v^T z = z^T v = \langle z, v \rangle,$$

it follows that

$$\langle \psi, \bar{A} \delta \phi \rangle = \psi^T \bar{A} \delta \phi = \delta \phi^T \bar{A}^T \psi = \langle \delta \phi, \bar{A}^T \psi \rangle. \quad (14.14)$$

The combination of (14.12)–(14.14) yields

$$\langle \psi, \delta s - \delta A \bar{\phi} \rangle = \langle \delta \phi, \bar{\mathcal{C}} \rangle = \bar{\mathcal{C}}^T \delta \phi. \quad (14.15)$$

The final term is precisely the component of the response perturbation (14.10) that is unknown, so we substitute (14.15) into that relation to obtain

$$\delta y = \delta \mathcal{C}^T \bar{\phi} + \psi^T [\delta s - \delta A \bar{\phi}]. \quad (14.16)$$

To employ (14.16), one must solve the adjoint equation (14.12) for  $\psi$ , but only once since it depends on nominal parameter values. Variations  $\delta y$  due to updated values of  $\delta \mathcal{C}$ ,  $\delta f$ , and  $\delta A$  are then easily computed since (14.16) requires only vector multiplication rather than re-solving linear systems.

**Multiple Responses  $\nu$ .** For this example,  $y = \mathcal{C}^T \phi$  is a single response. In general, we will have  $\nu$  responses or measurements. Note that this can be achieved by specifying a  $\nu \times (N - 1)$  observation matrix  $\mathcal{C}^T$ . The perturbation response relation (14.10) is unchanged, and the product of the  $i^{th}$  rows of  $\bar{\mathcal{C}}^T$  and  $\delta \mathcal{C}^T$  with  $\delta \phi$  and  $\bar{\phi}$  specify the  $i^{th}$  row  $[\delta y]_i$ . For the FSAP, the sensitivity equations (14.11) used to specify  $\delta \phi$  are also unchanged, and  $p$  parameter updates necessitate that (14.11) be solved  $p$  times.

For the ASAP, the transposed adjoint solution  $\psi^T$  is now a  $\nu \times (N - 1)$  matrix whose  $i^{th}$  row is the solution of

$$\bar{A}^T \psi = \bar{\mathcal{C}}_i,$$

where  $\bar{\mathcal{C}}_i$  is the  $i^{th}$  column of  $\bar{\mathcal{C}}$ . Hence the determination of  $\delta R$  using (14.16) requires the solution of  $\nu$  linear systems.

In conclusion, for linear or quasi-linear systems in which adjoints can be efficiently constructed or approximated, the ASAP is more efficient than the FSAP when the number of parameters  $p$  exceeds the number of responses  $\nu$ .

### Adjoint Sensitivity Analysis Procedure (ASAP): Variational Approach

Response perturbations based on adjoint solutions can also be constructed using a variational approach. As illustrated in subsequent examples, this approach requires less functional analysis and will be natural to readers familiar with variational calculus, optimal control, or sensitivity-based design. Furthermore, it can prove advantageous for problems that involve multiple constraints.

For  $A\phi = s$  with the response  $y = \mathcal{C}^T\phi$ , we consider the augmented or unconstrained response functional

$$\begin{aligned}\tilde{y} &= y - \psi^T [A\phi - s], \\ &= [\mathcal{C}^T - \psi^T A] \phi + \psi^T s,\end{aligned}\tag{14.17}$$

where  $\psi^T$  is a Lagrange multiplier. Taking Gâteaux variations of (14.17) about nominal values  $\bar{A}$ ,  $\bar{\mathcal{C}}$ , and  $\bar{s}$ , in the manner illustrated in (14.9), yields the perturbations

$$\delta\tilde{y} = \delta\mathcal{C}^T\phi - \psi^T [\delta A\phi - \delta s] - \delta\psi^T (\bar{A}\phi - \bar{s}) + (\bar{\mathcal{C}}^T - \psi^T \bar{A}) \delta\phi.$$

To eliminate the dependence on solutions variations  $\delta\phi$ , we enforce the relations

$$\psi^T \bar{A} = \bar{\mathcal{C}}^T,$$

which is the transpose of the adjoint sensitivity equation (14.12). We then employ the exact solution  $\bar{\phi} = \bar{A}^{-1}\bar{s}$  to obtain

$$\delta\tilde{y} = \delta\mathcal{C}^T\bar{\phi} + \psi^T[\delta s - \delta A\bar{\phi}].$$

Because we use the exact solution  $\bar{\phi}$ , it follows that  $\delta\tilde{y} = \delta y$ , as illustrated through comparison with (14.16).

#### 14.1.2 Boundary Value Problem

We now return to the boundary value problem (14.5), which has the solution (14.8) and response (14.6). The Gâteaux variation of the response is

$$\delta y = \delta A_d\varphi(b) + \bar{A}_d\delta\varphi(b),$$

where  $\delta A_d$ ,  $\varphi(b)$ , and  $\bar{A}_d$  are known and  $\delta\varphi(b)$  is unknown.

### Forward Sensitivity Analysis Procedure (FSAP)

To specify variations  $\delta\varphi$  in the solution, we need to construct the forward sensitivity equations. We take Gâteaux variations

$$\left\{ \frac{d}{d\varepsilon} \left[ (\bar{A}_a + \varepsilon\delta A_a)(\bar{\varphi} + \varepsilon\delta\varphi) - (\bar{D} + \varepsilon\delta D)\frac{d^2}{dx^2}(\bar{\varphi} + \varepsilon\delta\varphi) - (\bar{S} + \varepsilon\delta S) \right] \right\}_{\varepsilon=0} = 0,$$

$$\left\{ \frac{d}{d\varepsilon}(\bar{\varphi} + \varepsilon\delta\varphi)(\pm a) \right\}_{\varepsilon=0} = 0$$

of the system (14.5) to obtain the sensitivity equations

$$\begin{aligned} \bar{A}_a\delta\varphi - \bar{D}\frac{d^2\delta\varphi}{dx^2} &= \delta S - \delta A_a\bar{\varphi} - \delta D\frac{d^2\bar{\varphi}}{dx^2}, \\ \delta\varphi(\pm a) &= 0. \end{aligned} \quad (14.18)$$

It is shown in [50] that the solution to (14.18) is

$$\begin{aligned} \delta\varphi(x) &= C_1 [\cosh(xk) - \cosh(ak)] \\ &\quad + C_2 [x \sinh(xk) \cosh(ak) - a \sinh(ak) \cosh(xk)], \end{aligned}$$

where

$$C_1 = \frac{\delta A_a \bar{S}/\bar{A}_a - \delta S}{\bar{A}_a \cosh(ak)} , \quad C_2 = \frac{(\delta D/\bar{D} - \delta A_a/\bar{A}_a)\bar{S}}{2 \cosh^2(ak) \sqrt{\bar{D}\bar{A}_a}} , \quad k = \sqrt{\bar{A}_a/\bar{D}}.$$

We note that, in general, one cannot construct analytic solutions to the sensitivity equations and instead must use numerical approximations. For  $p$  parameters, this would require  $p$  numerical solutions to incorporate variations for each parameter.

### Adjoint Sensitivity Analysis Procedure (ASAP): Variational Approach

We consider the augmented response functional

$$\tilde{y} = y + \int_{-a}^a \left[ A_a \varphi - D \frac{d^2 \varphi}{dx^2} - S \right] \psi dx, \quad (14.19)$$

where  $\psi \in L^2(-a, a)$  is a Lagrange multiplier. Since  $y$  can be expressed as

$$y = A_d \varphi(b) = \int_{-a}^a A_d \varphi \delta(x - b) dx, \quad (14.20)$$

where  $\delta(x - b)$  is the Dirac density evaluated at  $b$ , it follows that

$$\tilde{y} = \int_{-a}^a \left[ \mathcal{H}(\varphi, q, \psi) - D \frac{d^2 \varphi}{dx^2} \psi \right] dx, \quad (14.21)$$

where

$$\mathcal{H}(\varphi, q, \psi) \equiv A_d \varphi \delta(x - b) + [A_a \varphi - S] \psi \quad (14.22)$$

is the Hamiltonian.

We take the Gâteaux variation of (14.21) to obtain

$$\begin{aligned}\delta\tilde{y} &= \int_{-a}^a [\mathcal{H}_\varphi\delta\varphi + \mathcal{H}_q\delta q + (\mathcal{H}_\psi - \bar{D}\varphi'')\delta\psi - \psi\varphi''\delta D - \bar{D}\psi\delta(\varphi'')] dx \\ &= \int_{-a}^a [(\mathcal{H}_\varphi - (\bar{D}\psi)'')\delta\varphi + \mathcal{H}_q\delta q + (\mathcal{H}_\psi - \bar{D}\varphi'')\delta\psi - \psi\varphi''\delta D] dx \\ &\quad - \bar{D}\psi(\delta\varphi)'|_{-a}^a + \bar{D}\psi'\delta\varphi|_{-a}^a,\end{aligned}$$

where  $\mathcal{H}_\varphi, \mathcal{H}_q, \mathcal{H}_\psi$  are partial Gâteaux derivatives with respect to  $\varphi, q, \psi$ , evaluated at nominal parameter and solution values, and we have used the property that  $\delta(\varphi'') = (\delta\varphi)''$ , as developed in Exercise 14.2. We note that because  $\varphi(\pm a) = 0$ , variations in the solution also satisfy  $\delta\varphi(\pm a) = 0$ , so the final term vanishes. Second, we let  $\psi$  satisfy the adjoint boundary value problem

$$\begin{aligned}\mathcal{H}_\varphi - \bar{D}\psi'' &= 0 & \Rightarrow & \bar{D}\frac{d^2\psi}{dx^2} - \bar{A}_a\psi = \bar{A}_d\delta(x - b) \\ \psi(\pm a) &= 0 & & \psi(\pm a) = 0\end{aligned}\tag{14.23}$$

and specify  $\bar{\varphi}$  as a solution of  $\bar{D}\varphi'' = \mathcal{H}_\psi$  to obtain

$$\begin{aligned}\delta\tilde{y} &= \int_{-a}^a [\mathcal{H}_q\delta q - \bar{\varphi}''\psi\delta D] dx \\ &= \int_{-a}^a [\mathcal{H}_{A_a}\delta A_a + (H_D - \bar{\varphi}''\psi)\delta D + \mathcal{H}_{A_d}\delta A_d + \mathcal{H}_s\delta S] dx \\ &= \bar{\varphi}(b)\delta A_d + \int_{-a}^a (\delta A_a\bar{\varphi} - \delta S - \delta D\bar{\varphi}'')\psi dx.\end{aligned}\tag{14.24}$$

We note that once (14.23) has been solved for  $\psi$ , solutions to (14.24) can be efficiently specified in terms of the solution  $\bar{\varphi}$  given by (14.8) and specified parameter variations  $\delta q = [\delta A_a, \delta D, \delta S, \delta A_d]$ .

We caution the reader that care must be exercised if employing the integral response representation (14.20) in the functional analytic framework of Section 14.2. In that framework, the Riesz representation theory is invoked to represent bounded linear functionals in terms of the Hilbert space inner products. The difficulty is that  $\delta$  is not bounded in  $L^2$  but rather in the dual space  $H^{-1}$  of  $H_0^1$ . Since  $H_0^1 \hookrightarrow L^2 \hookrightarrow H^{-1}$  is a Gelfand triple with dense and compact embeddings [265], density arguments and use of the duality product, rather than the  $L^2$  inner product, can be used to formulate the response so that it rigorously fits in the functional analytic framework.

## 14.2 Functional Analytic Framework for FSAP and ASAP

We summarize here the functional analytic FSAP and ASAP framework of [50] for the general linear model

$$\begin{aligned} L(q)u &= F(q(\chi)) \quad , \quad \chi \in \Omega, \\ B(q)u &= G(q) \quad , \quad \chi \in \partial\Omega, \end{aligned} \tag{14.25}$$

discussed in Section 3.3.1. Potential spatial or temporal dependence of the parameters  $q(\chi) = [q_1(\chi), \dots, q_p(\chi)]^T$  and states  $u(\chi) = [u_1(\chi), \dots, u_N(\chi)]^T$  is indicated by  $\chi = [x, t] \in \mathcal{J} \times \mathcal{T} \equiv \Omega$ , where  $\mathcal{J}$  is a subset of  $\mathbb{R}^1, \mathbb{R}^2$ , or  $\mathbb{R}^3$  and  $\mathcal{T}$  is a subset of  $\mathbb{R}^1$ . Here  $L(q) = [L_1(q), \dots, L_N(q)]^T$  is a vector of operators that depend linearly on  $u$  and typically nonlinearly on  $q$  and  $G(q), B(q)$  are operators associated with initial or boundary conditions. The response or observation is represented by

$$y = \mathcal{R}(u, q) = \mathcal{R}(e) , \quad e = [u, q] \tag{14.26}$$

in the Hilbert space  $H = H_u \times H_q$ , where  $H_u$  and  $H_q = Q$  are state and parameter spaces. Similarly, the sources  $F$  are assumed to be elements in the Hilbert space  $H_F$ . For differential operators,  $\text{dom}(L)$  is typically defined to be a dense subspace of  $H_u$ . Examples of the operators and spaces for the models in Section 3.1 are provided in Section 3.3.1.

As in Section 14.1, we assume that we know nominal (typically mean) parameter values  $\bar{q}$  and variations or perturbations  $\delta q = h_q$ . Nominal solution values  $\bar{u}$  are computed by solving (14.25) with  $q = \bar{q}$ . Perturbations of the solution are denoted by  $\delta u = h_u$ , and the combined perturbations of  $e$  about  $\bar{e}$  are denoted by  $h = [h_u, h_q]$ .

### Response Perturbations

In general, the Gâteaux variation

$$\delta\mathcal{R}(\bar{e}; h) = \lim_{\varepsilon \rightarrow 0} \frac{\mathcal{R}(\bar{e} + \varepsilon h) - \mathcal{R}(\bar{e})}{\varepsilon} = \frac{d}{d\varepsilon} \mathcal{R}(\bar{e} + \varepsilon h) \Big|_{\varepsilon} = 0$$

need not exhibit a linear dependence on  $h$  and hence  $h_u$ . To facilitate the discussion, we assume that  $\delta\mathcal{R}$  satisfies the necessary and sufficient linearity and continuity conditions detailed in [50], so it can be expressed as the Gâteaux differential

$$\delta\mathcal{R}(\bar{e}; h) = \mathcal{R}'_u(\bar{e})h_u + \mathcal{R}'_q(\bar{e})h_q, \tag{14.27}$$

where  $\mathcal{R}'_u(\bar{e})$  and  $\mathcal{R}'_q(\bar{e})$  denote the Gâteaux partial derivatives with respect to  $u$  and  $q$ . The term  $\mathcal{R}'_q(\bar{e})h_q$  depends on known perturbations  $h_q$  and is thus termed the *direct effect*. Because  $\mathcal{R}'_u(\bar{e})h_u$  depends on solution variations  $h_u = \delta u$ , which have yet to be computed, it is termed the *indirect effect*. The FSAP and ASAP represent two techniques to compute  $h_u$ .

### Forward Sensitivity Analysis Procedure (FSAP)

To construct sensitivity equations specifying  $h_u$ , we take the Gâteaux variations of (14.25) at  $\bar{e}$  in the direction  $h$  to obtain

$$\begin{aligned} 0 &= \left\{ \frac{d}{d\varepsilon} [L(\bar{q} + \varepsilon h_q)(\bar{u} + \varepsilon h_u) - (F + \varepsilon \delta F)] \right\}_{\varepsilon=0} \\ &= \{L(\bar{q} + \varepsilon h_q)h_u + L'_q(\bar{q} + \varepsilon h_q)(\bar{u} + \varepsilon h_u)h_q - \delta F(\bar{q}; h_q)\}_{\varepsilon=0} \\ &= L(\bar{q})h_u + [L'_q(\bar{q})\bar{u}]h_q - \delta F(\bar{q}; h_q), \end{aligned}$$

where  $L'_q(\bar{q})$  denotes the partial Gâteaux derivative of  $L$  at  $\bar{q}$ . Similar treatment of the boundary conditions and enforcement of the stationary condition for Gâteaux variations yield the forward sensitivity equations

$$\begin{aligned} L(\bar{q})h_u + [L'_q(\bar{q})\bar{u}]h_q &= \delta F(\bar{q}; h_q) \quad , \quad \chi \in \Omega, \\ B(\bar{q})h_u + [B'_q(\bar{q})\bar{u}]h_q &= \delta G(\bar{q}; h_q) \quad , \quad \chi \in \partial\Omega. \end{aligned} \tag{14.28}$$

It was illustrated in the examples of Section 14.1 that whereas the solution of these equations yields the sensitivities  $h_u = \delta u$  necessary to compute the indirect effect  $\mathcal{R}'_u(\bar{e})h_u$ , they must be resolved for each computed or updated parameter response  $q_i$ ,  $i = 1, \dots, p$ . If the number of parameters  $p$  is larger than the number of responses  $\nu$ , it is more efficient to employ the ASAP, provided the adjoints can be constructed or approximated in an efficient manner.

### Adjoint Sensitivity Analysis Procedure (ASAP)

We consider the case when there is a single observation so that  $\nu = 1$  and  $\mathcal{R}$  is a functional. We refer the reader to [50] for the general case of nonlinear response operators.

The Riesz Representation Theorem A.5 guarantees that every bounded linear functional in a Hilbert space can be uniquely represented in terms of the inner product. Since  $\mathcal{R}'_u(\bar{e})h_u$  is linear with regard to solution perturbations  $h_u$ , it thus follows that there exists a unique element  $\nabla_u \mathcal{R}(\bar{e}) \in H_u$  such that

$$\mathcal{R}'_u(\bar{e})h_u = \langle \nabla_u \mathcal{R}(\bar{e}), h_u \rangle_u \tag{14.29}$$

for all  $h_u \in H_u$ .

We know from the examples of Section 14.1 that the adjoint solution plays a fundamental role in the ASAP. From (A.4) of Appendix A, it follows that  $L$  and its formal adjoint  $L^*$  are related by

$$\langle L(\bar{q})h_u, \psi \rangle_F = \langle L^*(\bar{q})\psi, h_u \rangle_u + P(h_u, \psi)|_{\partial\Omega}, \tag{14.30}$$

where  $P(h_u, \psi)|_{\partial\Omega}$  is a bilinear form, evaluated at  $\bar{q}$ , that arises through integration by parts in the case of differential operators  $L$ . For unbounded operators (e.g., differential operators), the specification of the domain, including boundary

conditions, is necessary to fully define the operator. For differential operators, we consider adjoint boundary conditions

$$B^*(\psi; \bar{q}) = G^*(\bar{q}), \quad \chi \in \partial\Omega, \quad (14.31)$$

that are constructed to satisfy the following requirements.

### Adjoint Boundary and Initial Condition Requirements

- (i) The boundary conditions (14.31) do not include  $h_u, h_q$ , or Gâteaux derivatives with respect to  $q$ .
- (ii) All terms containing the unknown variations  $h_u = \delta u$  must vanish when the sensitivity and adjoint boundary conditions are substituted into  $P(h_u, \psi)|_{\partial\Omega}$ . The boundary terms that remain are denoted by  $\tilde{P}(h_q, \psi; q)$ .

We are now ready to complete the procedure. Since  $\nabla_u \mathcal{R}(\bar{e})$  is unique, we impose the condition

$$L^*(\bar{q})\psi = \nabla_u \mathcal{R}(\bar{e}), \quad (14.32)$$

along with (14.31), to obtain an adjoint system and reduced boundary relations  $\tilde{P}(h_q, \psi; q)$ . By employing the sensitivity equations (14.28), it follows that

$$\langle L(\bar{q})h_u, \psi \rangle_F = -\langle [L'_q(\bar{q})\bar{u}]h_q, \psi \rangle_F + \langle \delta F(\bar{q}; h_q), \psi \rangle_F, \quad (14.33)$$

where the right-hand side is independent of  $h_u = \delta u$ . Combination of (14.29), (14.30), (14.32), and (14.33) thus yields

$$\mathcal{R}'_u(\bar{e})h_u = \langle \delta F(\bar{q}; h_q) - [L'_q(\bar{q})\bar{u}]h_q, \psi \rangle_F - \tilde{P}(h_q, \psi; q) \quad (14.34)$$

so that response perturbations are

$$\delta \mathcal{R}(\bar{e}; h) = \mathcal{R}'_q(\bar{e})h_q + \langle \delta F(\bar{q}; h_q) - [L'_q(\bar{q})\bar{u}]h_q, \psi \rangle_F - \tilde{P}(h_q, \psi; q). \quad (14.35)$$

Once  $\psi$  has been determined by solving (14.32), all of the terms in (14.35) are known so that response uncertainties  $\delta \mathcal{R}$  can be computed directly.

We now illustrate the functional analytic FSAP and ASAP in examples.

#### 14.2.1 Neutron Diffusion—Matrix Systems

We revisit the model of Section 14.1.1 that had the response

$$y = \mathcal{R}(\phi, q) = \mathcal{C}^T(q)\phi$$

subject to the matrix constraint

$$A(q)\phi = s(q).$$

As detailed in Section 14.1.1 and Example A.11,  $L = A$  and  $L^* = A^T$  for this example. Furthermore,  $u = \phi$  and  $H_F = \mathbb{R}^N$  with the Euclidean dot product.

The relation (14.33) yields

$$\langle \bar{A}\delta\phi, \psi \rangle = \langle \delta s - \delta A\bar{\phi}, \psi \rangle$$

since  $L(\bar{q})h_u = A(\bar{q})\phi$  and  $[L'_q(\bar{q})\bar{u}]h_q = A'(\bar{q})\delta q\bar{\phi} = \delta A\bar{\phi}$ . This is precisely (14.13). The adjoint constraint (14.32) yields

$$\bar{A}^T\psi = \bar{C},$$

which is (14.12). Finally, the response perturbation relation (14.35) yields

$$\delta y = \delta \mathcal{R} = \delta C^T \bar{\phi} + \psi^T [\delta s - \delta A\bar{\phi}]$$

since there are no boundary terms  $\tilde{P}$ . This is exactly the response perturbation (14.16). We point out that these relations are also identical to those obtained using the variational adjoint approach.

### 14.2.2 Spring Model

To illustrate the functional analytic framework in the context of a differential operator, we revisit the spring model

$$\begin{aligned} \frac{d^2z}{dt^2} + Kz &= 0, \\ z(0) = z_0, \quad \frac{dz}{dt}(0) = z_1 \end{aligned} \tag{14.36}$$

of Example 14.3 with the response or observation

$$y = \mathcal{R}(z, q) = \int_0^{t_f} \gamma(t)z(t)dt. \tag{14.37}$$

Here  $u = z$  is considered in  $H_u = L^2(0, t_f)$  with the standard inner product. The parameters are

$$q = [K, z_0, z_1, \gamma] \in (0, \infty) \times \mathbb{R} \times \mathbb{R} \times L^2(0, t_f). \tag{14.38}$$

With the definitions

$$L(q)u = \ddot{z} + Kz, \quad F(q) = 0, \quad B(q)u = \begin{bmatrix} z \\ \dot{z} \end{bmatrix}, \quad G(q) = \begin{bmatrix} z_0 \\ z_1 \end{bmatrix}, \tag{14.39}$$

the model can be posed in the general operator framework (14.27). Finally, we take  $H_F = H_u$ .

#### Response Perturbations

The Gâteaux differential of (14.37), evaluated at  $\bar{y}(t)$  and  $\bar{z}(t)$ , is

$$\delta \mathcal{R} = \int_0^{t_f} \bar{\gamma}(t)\delta z(t)dt + \int_0^{t_f} \delta \gamma(t)\bar{z}(t)dt. \tag{14.40}$$

The nominal solution is

$$\bar{z}(t) = \bar{z}_0 \cos \bar{\omega}_0 t + \frac{\bar{z}_1}{\bar{\omega}_0} \sin \bar{\omega}_0 t, \quad (14.41)$$

where  $\bar{\omega}_0 = \bar{K}^{1/2}$ . The goal in the FSAP and ASAP is to determine the solution perturbations  $\delta z(t)$ .

### Forward Sensitivity Analysis Procedure (FSAP)

Taking Gâteaux variations of (14.36) yields the sensitivity equations

$$\begin{aligned} \frac{d^2 \delta z}{dt^2} + \bar{K} \delta z &= -\delta K \bar{z}, \\ \delta z(0) &= \delta z_0, \quad \frac{d \delta z}{dt}(0) = \delta z_1. \end{aligned} \quad (14.42)$$

Note that these relations can be obtained from (14.28) using the operator definitions (14.39) since

$$L(\bar{q})h_u + [L'_q(\bar{q})\bar{u}]h_q = \frac{d^2 \delta z}{dt^2} + \bar{K} \delta z + \delta K \bar{z}.$$

By employing the nominal solution  $\bar{z}$  given by (14.41), we obtain the analytic sensitivity solution

$$\delta z(t) = \left[ -\frac{\bar{z}_0}{2\bar{\omega}_0} t \sin \bar{\omega}_0 t + \frac{\bar{z}_1}{2\bar{\omega}_0^2} \cos \bar{\omega}_0 t \right] \delta K + \cos \bar{\omega}_0 t \delta z_0 + \frac{1}{\bar{\omega}_0} \sin \bar{\omega}_0 t \delta z_1. \quad (14.43)$$

This can be employed in (14.40) to obtain the response perturbations or uncertainties. The individual sensitivities are thus

$$\begin{aligned} \frac{\partial \bar{z}}{\partial \bar{K}} &= -\frac{\bar{z}_0}{2\bar{\omega}_0} t \sin \bar{\omega}_0 t + \frac{\bar{z}_1}{2\bar{\omega}_0^2} \cos \bar{\omega}_0 t, \\ \frac{\partial \bar{z}}{\partial \bar{z}_0} &= \cos \bar{\omega}_0 t, \\ \frac{\partial \bar{z}}{\partial \bar{z}_1} &= \frac{1}{\bar{\omega}_0} \sin \bar{\omega}_0 t, \end{aligned}$$

which are precisely the relations that result from differentiating (14.41) with respect to the parameters.

### Adjoint Sensitivity Analysis Procedure (ASAP): Variational Approach

To form an unconstrained response, we employ the augmented functional

$$\begin{aligned} \tilde{y} &= \tilde{\mathcal{R}} = \mathcal{R}(z, q) - \int_0^{t_f} \psi [\ddot{z} + K z] dt \\ &= \int_0^{t_f} [\mathcal{H}(q, \psi, z) - \psi \dot{z}] dt, \end{aligned}$$

where the Hamiltonian is

$$\mathcal{H}(q, \psi, z) = (\gamma - K\psi)z.$$

We note that subtraction of the constraint facilitates comparison with the functional analytic approach. However, one can just as easily add the constraint which yields the same final result.

The Gâteaux variation is thus

$$\begin{aligned}\tilde{\mathcal{R}} &= \int_0^{t_f} [\mathcal{H}_z \delta z + \mathcal{H}_q \delta q + (\mathcal{H}_\psi - \ddot{z}) \delta \psi - \psi \ddot{\delta z}] dt \\ &= \int_0^{t_f} [(\mathcal{H}_z - \ddot{\psi}) \delta z + \mathcal{H}_q \delta q + (\mathcal{H}_\psi - \ddot{z}) \delta \psi] dt + [-\psi \dot{\delta z} + \dot{\psi} \delta z] \Big|_0^{t_f},\end{aligned}\quad (14.44)$$

where  $\delta z(0) = \delta z_0$ ,  $\dot{\delta z}(0) = \delta z_1$  are unknown and  $\mathcal{H}_z$ ,  $\mathcal{H}_q$ ,  $\mathcal{H}_\psi$  are partial Gâteaux derivatives of  $\mathcal{H}$  with respect to  $z$ ,  $q$ , and  $\psi$ . We have also used the property that  $\frac{d\delta z}{dt} = \delta \frac{dz}{dt} = \delta \dot{z}$ , as developed in Exercise 14.1.

To eliminate the unknown boundary terms at  $t_f$ , we enforce the conditions  $\psi(t_f) = \dot{\psi}(t_f) = 0$  in the adjoint problem

$$\begin{array}{ll}\ddot{\psi} - \mathcal{H}_z = 0 & \ddot{\psi} + \bar{K}\psi = \bar{\gamma} \\ \psi(t_f) = \dot{\psi}(t_f) = 0 & \psi(t_f) = \dot{\psi}(t_f) = 0.\end{array}\quad (14.45)$$

Furthermore, we employ the solution  $\bar{z}$  to the state relation  $\ddot{z} - \mathcal{H}_\psi = 0$  to obtain

$$\delta \tilde{\mathcal{R}} = \int_0^{t_f} [\bar{z} \delta \gamma - \bar{z} \psi \delta K] dt + \psi(0) \delta z_1 - \dot{\psi}(0) \delta z_0. \quad (14.46)$$

Finally, we note that by employing the nominal solution  $z = \bar{z}$ , it follows that  $\delta \tilde{\mathcal{R}} = \delta \mathcal{R}$ . By comparing with (14.40), we see that

$$\int_0^{t_f} \bar{\gamma}(t) \delta z(t) dt = - \int_0^{t_f} \bar{z}(t) \psi(t) \delta K dt + \psi(0) \delta z_1 - \dot{\psi}(0) \delta z_0$$

in the response perturbation. The advantage of the adjoint formulation (14.46) is that the adjoint problem (14.45) must be solved only once, whereas  $\delta z$  given by (14.43) must be recomputed for each change in  $\delta z_0$ ,  $\delta z_1$ , or  $\delta K$ .

### Adjoint Sensitivity Analysis Procedure (ASAP): Perturbation Approach

To illustrate the functional analytic approach, we note from the operator definitions (14.39) that the adjoint relation (14.30) yields

$$\int_0^{t_f} (\ddot{\delta z} + K \delta z) \psi dt = \int_0^{t_f} (\ddot{\psi} + K\psi) \delta z + [\psi \dot{\delta z} - \dot{\psi} \delta z] \Big|_0^{t_f}. \quad (14.47)$$

To eliminate the unknown terms at  $t_f$ , we enforce the adjoint boundary conditions

$$\psi(t_f) = \dot{\psi}(t_f) = 0$$

indicated in operator format in (14.31). The remaining boundary terms are thus  $\tilde{P}(\delta q, \psi) = -\psi(0)\delta z_1 + \dot{\psi}(0)\delta z_0$ .

To specify the right-hand side of the adjoint equation, we note that the indirect effect to the perturbation response is

$$\mathcal{R}'_u(\bar{e})h_u = \int_0^{t_f} \bar{\gamma}(t)\delta z(t)dt$$

so that the adjoint system is

$$\begin{aligned} \ddot{\psi} + K\psi &= \gamma, \\ \psi(t_f) &= \dot{\psi}(t_f) = 0. \end{aligned} \tag{14.48}$$

By using the sensitivity equation (14.42) to replace the left-hand side of (14.47), it follows that

$$\int_0^{t_f} \bar{\gamma}(t)\delta z(t)dt = - \int_0^{t_f} \delta K \bar{z}(t)\psi(t)dt + \psi(0)\delta z_1 - \dot{\psi}(0)\delta z_0$$

so that

$$\delta \mathcal{R} = \int_0^{t_f} [\bar{z}\delta\gamma - \bar{z}\psi\delta K] dt + \psi(0)\delta z_1 - \dot{\psi}(0)\delta z_0, \tag{14.49}$$

which is precisely the relation (14.46) obtained through variational analysis. We note that (14.49) is given by (14.35) with

$$R'_q(\bar{e})h_q - \langle [L'_q(\bar{q})\bar{u}]h_q, \psi \rangle_F = \int_0^{t_f} \delta\gamma\bar{z}dt - \int_0^{t_f} \delta K \bar{z}\psi dt.$$

**Remark 14.4.** It is observed that if the response  $\mathcal{R}$  is posed in terms of the inner product for  $H_u$ , as is the case in this example, then the two approaches are actually the same. The difference arises for problems in which the Riesz representation theorem must be invoked to represent response functionals, such as point evaluations, that are not initially posed in terms of the Hilbert space inner product.

### 14.3 Notes and References

The FSAP and perturbation approach to the ASAP follow closely the development in [50, 53], and the reader is referred to those references for additional details and applications of the techniques. To simplify the discussion, we have focused solely on linear problems where the adjoint problem can be solved independently from the forward problem. The reader is again referred to [50, 53] for sensitivity analysis of nonlinear systems where the forward and adjoint problems are coupled, which requires simultaneous solution techniques.

## 14.4 Exercises

**Exercise 14.1.** Use the definition of the Gâteaux variation to show that  $(\delta\varphi)' = \delta(\varphi')$  and that  $(\delta\varphi)'' = \delta(\varphi'')$ .

**Exercise 14.2.** Consider the initial value problem

$$\begin{aligned}\frac{dz}{dt} &= az^2 + bu, \\ z(0) &= z_0\end{aligned}$$

and response

$$y = \mathcal{R}(z, u, q) = \int_0^{t_f} [kz^2(t) + ru^2(t)]dt,$$

where  $a, b$ , and  $z_0$  are scalars and  $q = [a, b, z_0]$ . You can treat  $k$  and  $r$  as known, fixed design parameters.

Determine the adjoint equation, along with an appropriate boundary condition, and specify the response variation  $\delta\mathcal{R}$ .

**Exercise 14.3.** Exercises 7.1–7.5 illustrate additional facets of local sensitivity analysis.

## Chapter 15

# Global Sensitivity Analysis

As detailed in Definition 14.2, one objective of global sensitivity analysis is to quantify how uncertainties in model outputs can be apportioned to uncertainties in model inputs that are considered over the entire range of input values. Unlike local sensitivity analysis, where inputs are varied about a nominal value, uncertainties due to combinations of parameters throughout the admissible parameter space are considered in global sensitivity analysis. In both cases, analysis focuses solely on properties of the model and does not rely on experimental data. The determination of parameter and output uncertainties using experimental data constitutes the complementary processes of model calibration and uncertainty propagation.

Global sensitivity analysis is often used to determine noninfluential parameters in nonlinearly parameterized models, which can be fixed for subsequent model calibration or uncertainty propagation. Hence these techniques provide the basis for the parameter selection methods of Section 6.2.

The differences between global and local sensitivity analysis and the manner in which global sensitivity is evaluated are illustrated in the next example.

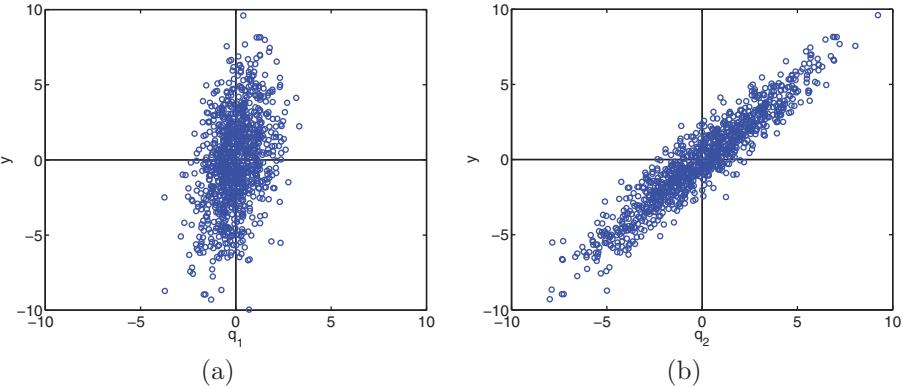
**Example 15.1.** Consider the linear portfolio model

$$Y = c_1 Q_1 + c_2 Q_2 \quad (15.1)$$

discussed in [215]. Here  $Q_1$  and  $Q_2$  represent hedged portfolios and  $c_1$  and  $c_2$  are the amounts invested in each portfolio. For example, each portfolio could be comprised of options and stocks with associated risks. We initially assume that the average return from each portfolio is zero and that they are independent and normally distributed with  $Q_1 \sim N(0, \sigma_1^2)$  and  $Q_2 \sim N(0, \sigma_2^2)$ , where  $\sigma_1 = 1$  and  $\sigma_2 = 3$ . In this example, we take  $c_1$  and  $c_2$  to be constant with  $c_1 = 2$ ,  $c_2 = 1$ . The random variable  $Y$  is the return for the investment, although it is often termed the risk if it is negative. From (4.18), it follows that  $Y \sim N(0, \sigma_Y^2)$ , where

$$\sigma_Y^2 = c_1^2 \sigma_1^2 + c_2^2 \sigma_2^2 = 13. \quad (15.2)$$

Since  $\sigma_2 > \sigma_1$ , the second portfolio is said to be more volatile than the first. To illustrate the effects of portfolio uncertainty on the return, scatterplots obtained



**Figure 15.1.** Scatterplots of  $y$  versus (a)  $q_1$  and (b)  $q_2$  constructed using 1000 joint realizations.

with 1000 joint realizations  $q_1$ ,  $q_2$ , and  $y$  are plotted in Figure 15.1. These plots indicate that  $Q_2$  has more influence on  $Y$  than  $Q_1$  since the realizations  $(q_2, y)$  clearly reflect the trend of the model (15.1), whereas the realizations  $(q_1, y)$  and  $(-q_1, y)$  are nearly identical and there is no clear trend for the nearly uniform scatterplot. From a global perspective,  $Y$  is thus considered more sensitive to  $Q_2$  than  $Q_1$ .

In contrast, the local derivative relation

$$s_i \equiv \frac{\partial Y}{\partial Q_i} \quad (15.3)$$

yields

$$s_1 = 2 > s_2 = 1,$$

which reflects the amounts invested in the two portfolios rather than the effects of their volatility on the return. Hence the local technique does not accommodate the nonlinear uncertainty structure over the global admissible parameter space  $\mathbb{Q} = \mathbb{R}^2$ .

Alternatively, one can employ the sigma-normalized relations

$$S_i^\sigma \equiv \frac{\sigma_i}{\sigma_Y} \frac{\partial Y}{\partial Q_i} = c_i \frac{\sigma_i}{\sigma_Y}, \quad (15.4)$$

which are hybrid local-global in nature since  $\sigma_i$  incorporates variability over the range of input values. Here

$$S_1^\sigma = \frac{2}{\sqrt{13}} < \frac{3}{\sqrt{13}} = S_2^\sigma, \quad (15.5)$$

which is consistent with the scatterplot information in Figure 15.1. From the definition, it follows immediately that

$$(S_1^\sigma)^2 + (S_2^\sigma)^2 = 1$$

so that each squared relation  $(S_i^\sigma)^2$  quantifies the contribution of that individual factor to the variance of the output or QoI. We note that the relations (15.4) constitute one technique recommended for global sensitivity analysis by the 1999 and 2000 Intergovernmental Panel for Climate Change (IPCC) [115]; see also Section 2.2.

The model (15.1) is monotone and additive in the sense defined next.

**Definition 15.2.** A model  $Y = f(Q_1, \dots, Q_p)$  is additive if it can be expressed as  $Y = \sum_{i=1}^p f_i(Q_i)$ .

We now discuss the construction of global sensitivity measures that can be applied to nonmonotone models or models with parameter interactions.

## 15.1 Variance-Based Methods

### 15.1.1 Sobol Decomposition for Uniform Densities

Consider the scalar-valued, nonlinear model

$$Y = f(Q), \quad (15.6)$$

where  $Q = [Q_1, \dots, Q_p] \in \Gamma \subset \mathbb{R}^p$ . We initially assume that the random variables are independent and uniformly distributed on  $[0, 1]$  so that

$$Q_i \sim \mathcal{U}(0, 1) \quad , \quad \Gamma = [0, 1]^p.$$

The case of general densities is discussed in Section 15.1.2. We consider the second-order HDMR or Sobol expansion

$$f(q) = f_0 + \sum_{i=1}^p f_i(q_i) + \sum_{1 \leq i < j \leq p} f_{ij}(q_i, q_j) \quad (15.7)$$

discussed in Section 13.5 subject to the condition

$$\int_0^1 f_i(q_i) dq_i = \int_0^1 f_{ij}(q_i, q_j) dq_i = \int_0^1 f_{ij}(q_i, q_j) dq_j = 0, \quad (15.8)$$

which ensures that the functions are orthogonal in the sense that

$$\int_\Gamma f_i(q_i) f_j(q_j) dq_i dq_j = \int_\Gamma f_i(q_i) f_{ij}(q_i, q_j) dq_i dq_j = 0 \quad (15.9)$$

for  $i, j = 1, \dots, p$ . As detailed in [200, 227], the zeroth-, first-, and second-order terms can then be expressed as

$$\begin{aligned} f_0 &= \int_\Gamma f(q) dq, \\ f_i(q_i) &= \int_{\Gamma^{p-1}} f(q) dq_{\sim i} - f_0, \\ f_{ij}(q_i, q_j) &= \int_{\Gamma^{p-2}} f(q) dq_{\sim \{ij\}} - f_i(q_i) - f_j(q_j) - f_0, \end{aligned} \quad (15.10)$$

where  $\Gamma^{p-1} = [0, 1]^{p-1}$  and  $\Gamma^{p-2} = [0, 1]^{p-2}$ . Recall that the notation  $q_{\sim i}$  denotes the vector having all the components of  $q$  except those in the set  $i$ ; for example,

$$q_{\sim i} = [q_1, \dots, q_{i-1}, q_{i+1}, \dots, q_p]. \quad (15.11)$$

The expressions (15.10) are precisely those in (13.54) for ANOVA-HDMR.

The total variance  $D$  of the response  $Y$  is given by

$$D = \text{var}(Y) = \int_{\Gamma} f^2(q)dq - f_0^2 \quad (15.12)$$

since  $f_0 = \mathbb{E}(Y)$ , as detailed in Remark 15.3. By employing the expansion (15.7) and enforcing the orthonormality conditions (15.8) and (15.9), the total variance can be expressed as

$$D = \sum_{i=1}^p D_i + \sum_{1 \leq i < j \leq p} D_{ij},$$

where the partial variances are

$$\begin{aligned} D_i &= \int_0^1 f_i^2(q_i)dq_i, \\ D_{ij} &= \int_0^1 \int_0^1 f_{ij}^2(q_i, q_j)dq_idq_j. \end{aligned} \quad (15.13)$$

The Sobol indices are defined to be

$$S_i = \frac{D_i}{D} \quad , \quad S_{ij} = \frac{D_{ij}}{D} \quad , \quad i, j = 1, \dots, p,$$

so, by definition, they satisfy

$$\sum_{i=1}^p S_i + \sum_{1 \leq i < j \leq p} S_{ij} = 1.$$

The terms  $S_i$  are often termed the *importance measures* or *first-order sensitivity indices*, and large values of  $S_i$  indicate parameters that strongly influence the response variance. Similarly,  $S_{ij}$  account for the influence of interaction terms.

Because the number of first- and second-order Sobol indices is  $p + \frac{p(p-1)}{2}$ , their analysis quickly becomes untenable for large parameter dimensions. This motivates the consideration of *total sensitivity indices*

$$S_{T_i} = S_i + \sum_{j=1}^p S_{ij}, \quad (15.14)$$

which quantify the total effect of the parameter  $Q_i$  on the response  $Y$ .

**Remark 15.3.** The expansion terms, partial variances, and Sobol indices all have expectation or variance interpretations. To set notation, we let

$$\begin{aligned}\mathbb{E}(Y|q_i) &= \int_{\Gamma^{p-1}} f(q)dq_{\sim i}, \\ \mathbb{E}(Y|q_i, q_j) &= \int_{\Gamma^{p-2}} f(q)dq_{\sim \{ij\}}\end{aligned}$$

denote the expected responses when the components  $q_i$  and  $q_i, q_j$  are fixed. From (15.10), it follows that

$$\begin{aligned}f_0 &= \mathbb{E}(Y), \\ f_i(q_i) &= \mathbb{E}(Y|q_i) - f_0, \\ f_{ij}(q_i, q_j) &= \mathbb{E}(Y|q_i, q_j) - f_i(q_i) - f_j(q_j) - f_0.\end{aligned}\tag{15.15}$$

Since

$$\mathbb{E}[\mathbb{E}(Y|q_i)] = \int_0^1 \left[ \int_{\Gamma^{p-1}} f(q)dq_{\sim i} \right] dq_i = f_0,\tag{15.16}$$

it follows that

$$D_i = \text{var}[\mathbb{E}(Y|q_i)]\tag{15.17}$$

and hence

$$S_i = \frac{\text{var}[\mathbb{E}(Y|q_i)]}{\text{var}(Y)}.\tag{15.18}$$

Similarly, one can show that

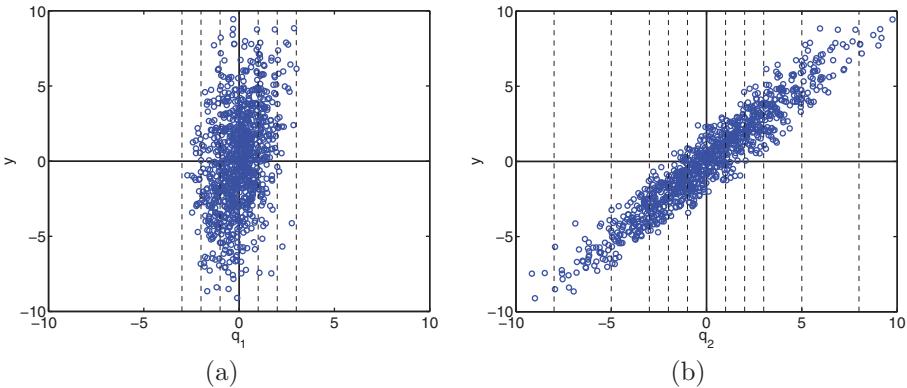
$$D_{ij} = \text{var}[\mathbb{E}(Y|q_i, q_j)] - \text{var}[\mathbb{E}(Y|q_i)] - \text{var}[\mathbb{E}(Y|q_j)],$$

which yields a variance interpretation for  $S_{ij}$ . Finally, the total sensitivity index has the interpretation

$$S_{T_i} = 1 - \frac{\text{var}[\mathbb{E}(Y|q_{\sim i})]}{\text{var}(Y)} = \frac{\mathbb{E}[\text{var}(Y|q_{\sim i})]}{\text{var}(Y)}.\tag{15.19}$$

The interpretation of  $\mathbb{E}(Y|q_i)$  and  $\text{var}[\mathbb{E}(Y|q_i)]$  is further illustrated in Figure 15.2 using scatterplots for the linear portfolio model (15.1). The conditional expectations for fixed  $q_1$  and  $q_2$  are the average values of  $Y$  along vertical slices. The partial variances  $D_i$  quantify the variability of these average values. For this example,  $D_2$  is significantly larger than  $D_1$ , thus quantifying the structure observed in the figures. The Sobol indices quantify this relative influence on the scale  $[0, 1]$ .

**Remark 15.4.** From (15.19), we note that if  $S_{T_i} \approx 0$ , then  $\mathbb{E}[\text{var}(Y|q_{\sim i})] \approx 0$ , which, by the nonnegativity of the variance operator, implies that  $\text{var}(Y|q_{\sim i}) \approx 0$  for any admissible value of  $q_{\sim i}$ . The condition  $S_{T_i} \approx 0$  thus implies that  $Q_i$  is noninfluential and can be fixed in subsequent model calibration and uncertainty quantification. The use of the total sensitivity index to reduce model complexity in this manner is an important aspect of global sensitivity analysis.



**Figure 15.2.** Response values for fixed (a)  $q_1$  and (b)  $q_2$  used to construct the expectations  $\mathbb{E}(Y|q_i)$  and variances  $\text{var}[\mathbb{E}(Y|q_i)]$ .

### 15.1.2 Sobol Decomposition for General Densities

Here we consider the nonlinear model (15.6) where  $Q_1, \dots, Q_p$  are considered to be iid random variables with ranges  $\Gamma_k$  and densities  $\rho_{Q_k}(q_k)$ . The range and joint density for  $Q$  are then

$$\Gamma = \prod_{k=1}^p \Gamma_k \quad , \quad \rho_Q(q) = \prod_{k=1}^p \rho_{Q_k}(q_k).$$

Independence is required to expand  $\rho_Q(q)$  as a product of marginal densities. The assumption of identical distributions is made solely to simplify notation, and these relations can be easily extended to accommodate differing distributions.

The complete Sobol decomposition of  $f$  is then

$$f(q) = \sum_{\mathbf{i}' \subseteq \{1, 2, \dots, p\}} f_{\mathbf{i}'}(q_{\mathbf{i}'}), \quad (15.20)$$

where  $\mathbf{i}' = \{i_1, \dots, i_s\}$  is a set of integers with cardinality  $s$ ,  $q_{\mathbf{i}'} = [q_{i_1}, \dots, q_{i_s}]$ , and  $f_\emptyset = f_0$ . Each of the functions  $f_{\mathbf{i}'}$ , except  $f_0$ , is assumed to satisfy

$$\int_{\Gamma_k} f_{\mathbf{i}'}(q_{\mathbf{i}'}) \rho_{Q_k}(q_k) dq_k = 0 \quad (15.21)$$

for any  $q_k$  and all  $\mathbf{i}' \subseteq \{1, 2, \dots, p\}$  that include  $k$ ; this is the generalization of (15.8). This ensures that the components are orthogonal in the sense that

$$\int_{\Gamma} f_{\mathbf{i}'}(q_{\mathbf{i}'}) f_{\ell'}(q_{\ell'}) \rho_Q(q) dq \quad \text{for all } \mathbf{i}' \neq \ell',$$

for which (15.9) is a special case. With assumption (15.21), the Sobol or HDMR

decomposition is unique and the component functions are given by

$$f_{\mathbf{i}'}(q_{\mathbf{i}'}) = \int_{\Gamma^{p-s}} f(q) \rho_Q(q_{\sim \mathbf{i}'}) dq_{\sim \mathbf{i}'} - \sum_{\ell' \subset \mathbf{i}' \atop \ell' \neq \mathbf{i}'} f_{\ell'}(q_{\ell'}),$$

where  $q_{\sim i}$  is defined in (15.11).

In a manner analogous to (15.12) and (15.13), the total and conditional or partial variances are defined by

$$D = \int_{\Gamma} f^2(q) \rho_Q(q) dq - f_0^2$$

and

$$\begin{aligned} D_{\mathbf{i}'} &= \int_{\Gamma^s} f_{\mathbf{i}'}^2(q_{\mathbf{i}'}) \rho_Q(q_{\mathbf{i}'}) dq_{\mathbf{i}'} \\ &= \text{var}[\mathbb{E}(Y|q_{\mathbf{i}'})] - \sum_{\ell' \subset \mathbf{i}' \atop \ell' \neq \mathbf{i}', \ell' \neq \emptyset} D_{\ell'}. \end{aligned}$$

Due to the orthogonality of the functions, the total variance can be expressed as

$$D = \sum_{\substack{\mathbf{i}' \subseteq \{1, 2, \dots, p\} \\ \mathbf{i}' \neq \emptyset}} D_{\mathbf{i}'}.$$

The Sobol indices are defined to be

$$S_{\mathbf{i}'} = \frac{D_{\mathbf{i}'}}{D}$$

so that

$$\sum_{\substack{\mathbf{i}' \subseteq \{1, 2, \dots, p\} \\ \mathbf{i}' \neq \emptyset}} S_{\mathbf{i}'} = 1.$$

The total sensitivity indices

$$S_{T_{\mathbf{i}'}} \equiv \sum_{k \ni \mathbf{i}'} S_k$$

quantify the sensitivity of the variances of  $Y$  with respect to  $Q_{\mathbf{i}'}$  along with its interaction with all other inputs. To illustrate for  $p = 3$ , we note that

$$S_{T_2} = S_{\{2\}} + S_{\{1, 2\}} + S_{\{2, 3\}} + S_{\{1, 2, 3\}}.$$

**Remark 15.5.** The expansion (15.20) is complete in the sense that it includes all interaction terms through order  $p$ . For the following examples, we truncate at second-order terms since, as discussed in Section 13.5, HDMR techniques are efficient only if high-order interactions are negligible.

**Remark 15.6.** Variance-based indices are advantageous over regression and correlation-based indices since they do not require linearity or monotonicity. For this reason, they are sometimes referred to as *model-free* methods. However, the assumption that parameters are mutually independent is typically required to ensure that  $\rho_Q(q)$  can be expressed as a product of marginal densities. As detailed in Section 5.2, parameters in physical problems are typically correlated, which violates this assumption. If sufficient statistical information is available, one can employ Nataf or Rosenblatt transformations to reformulate the problem in terms of independent parameters. However, obtaining the required marginal and correlation information can be difficult for complex problems.

**Example 15.7.** We revisit the additive portfolio model of Example 15.1, where

$$\rho_{Q_1}(q_1) = \frac{1}{\sigma_1 \sqrt{2\pi}} e^{-q_1^2/2\sigma_1^2}, \quad \rho_{Q_2}(q_2) = \frac{1}{\sigma_2 \sqrt{2\pi}} e^{-q_2^2/2\sigma_2^2},$$

and  $\rho_Q(q) = \rho_{Q_1}(q_1)\rho_{Q_2}(q_2)$ . Here

$$\begin{aligned} f_0 &= 0, \\ f_1(q_1) &= \int_{\mathbb{R}} \rho_{Q_2}(q_2)[c_1 q_1 + c_2 q_2] dq_2 = c_1 q_1, \\ f_2(q_2) &= c_2 q_2, \end{aligned}$$

and  $f_{ij}(q_i, q_j) = 0$ . The partial variances are

$$\begin{aligned} D_i &= \int_{\mathbb{R}} c_i^2 q_i^2 \rho_{Q_i}(q_i) dq_i = c_i^2 \sigma_i^2, \\ D_{ij} &= 0, \end{aligned}$$

and the total variance is

$$D = c_1^2 \sigma_1^2 + c_2^2 \sigma_2^2,$$

as noted in (15.2). The Sobol indices are

$$S_i = \frac{c_i^2 \sigma_i^2}{c_1^2 \sigma_1^2 + c_2^2 \sigma_2^2}, \quad S_{ij} = 0$$

so that  $S_1 = \frac{4}{13}$  and  $S_2 = \frac{9}{13}$  since  $c_1 = 2$ ,  $c_2 = 1$ ,  $\sigma_1 = 1$ , and  $\sigma_2 = 3$ . For this additive example, we thus see that  $S_i = (S_i^\sigma)^2$ , where  $S_i^\sigma$  is defined in (15.4).

**Example 15.8.** We now consider the model

$$Y = Q_3 Q_1 + Q_4 Q_2,$$

where  $Q = [Q_1, Q_2, Q_3, Q_4]$  is a random vector and

$$\begin{aligned} Q_1 &\sim N(0, \sigma_1^2), \quad Q_2 \sim N(0, \sigma_2^2), \\ Q_3 &\sim N(c_1, \sigma_3^2), \quad Q_4 \sim N(c_2, \sigma_4^2). \end{aligned}$$

This is a generalization of the portfolio model (15.1), where the amounts are normally distributed about  $c_1 = 2$  and  $c_2 = 1$ . In this case

$$f_0 = 0,$$

$$f_1(q_1) = \int_{\mathbb{R}^3} \rho_{Q_2}(q_2) \rho_{Q_3}(q_3) \rho_{Q_4}(q_4) [q_3 q_1 + q_4 q_2] dq_2 dq_3 dq_4 = 0,$$

$$f_2(q_2) = f_3(q_3) = f_4(q_4) = 0,$$

$$f_{13}(q_1, q_3) = \int_{\mathbb{R}^2} \rho_{Q_2}(q_2) \rho_{Q_4}(q_4) [q_3 q_1 + q_4 q_2] dq_2 dq_4 = q_3 q_1,$$

$$f_{24}(q_2, q_4) = q_4 q_2 , f_{12}(q_1, q_2) = f_{23}(q_2, q_3) = 0$$

and

$$D_i = 0,$$

$$D_{13} = \int_{\mathbb{R}^2} \rho_{Q_1}(q_1) \rho_{Q_3}(q_3) q_3^2 q_1^2 dq_1 dq_3 = \sigma_1^2 \sigma_3^2 , D_{24} = \sigma_2^2 \sigma_4^2$$

so that

$$S_i = 0,$$

$$S_{13} = \frac{\sigma_1^2 \sigma_3^2}{\sigma_1^2 \sigma_3^2 + \sigma_2^2 \sigma_4^2} , S_{24} = \frac{\sigma_2^2 \sigma_4^2}{\sigma_1^2 \sigma_3^2 + \sigma_2^2 \sigma_4^2}.$$

**Example 15.9.** Consider the Ishigami function

$$f(q) = \sin q_1 + a \sin^2 q_2 + b q_3^4 \sin q_1 \quad (15.22)$$

discussed in Example 13.8. It is established in Exercise 15.2 that the total and partial variances are

$$\begin{aligned} D &= \frac{a^2}{8} + \frac{b\pi^4}{5} + \frac{b^2\pi^8}{18} + \frac{1}{2}, \\ D_1 &= \frac{b\pi^4}{5} + \frac{b^2\pi^8}{50} + \frac{1}{2} , D_2 = \frac{a^2}{8} , D_3 = 0, \\ D_{12} &= 0 , D_{13} = \frac{b^2\pi^4}{18} - \frac{b^2\pi^8}{50} , D_{23} = D_{123} = 0. \end{aligned} \quad (15.23)$$

The Sobol indices  $S_i$  and  $S_{ij}$  and total Sobol indices  $S_{T_i}$  provide comprehensive measures for quantifying the influence of parameter uncertainty on the variance of the response. However, their computation can be prohibitively expensive for large parameter dimensions since they require the approximation of integrals up to dimension  $p$ . This has motivated significant research in two directions: techniques to efficiently compute the partial variances and Sobol indices in high dimensions, and screening algorithms that approximate sensitivity information using a linearization of the model. The algorithm summarized in Section 15.1.3 is widely employed to construct first-order sensitivity indices. Additionally, evaluation techniques based on the stochastic polynomial or cut-HDMR expansions discussed in Sections 13.5.2 and 13.5.4 constitute two recently employed techniques. We discuss Morris screening algorithms in Section 15.2.

### 15.1.3 Algorithm to Compute Sensitivity Indices

The computation of  $D_i$  and  $S_i$  given by (15.17) and (15.18) requires the approximation of  $\text{var}[\mathbb{E}(Y|q_i)]$ . If one uses  $M$  Monte Carlo evaluations to approximate the conditional mean  $\mathbb{E}(Y|q_i)$  for fixed  $q_i$  and repeats the procedure  $M$  times to approximate the variance, a total of  $M^2$  evaluations will be required to evaluate a single sensitivity index. For large parameter dimensions  $p$ , this brute-force approach is clearly prohibitive. The following algorithm of Saltelli [214], which is based on Sobol's original approach [227], reduces the number of required function evaluations to  $M(p+2)$ .

**Algorithm 15.10.** 1. Create two  $M \times p$  sample matrices

$$A = \begin{bmatrix} q_1^1 & \cdots & q_i^1 & \cdots & q_p^1 \\ \vdots & & \vdots & & \vdots \\ q_1^M & \cdots & q_i^M & \cdots & q_p^M \end{bmatrix}, \quad B = \begin{bmatrix} \hat{q}_1^1 & \cdots & \hat{q}_i^1 & \cdots & \hat{q}_p^1 \\ \vdots & & \vdots & & \vdots \\ \hat{q}_1^M & \cdots & \hat{q}_i^M & \cdots & \hat{q}_p^M \end{bmatrix},$$

where  $q_i^j$  and  $\hat{q}_i^j$  are quasi-random numbers drawn from the respective densities.

2. Create  $M \times p$  matrices

$$C_i = \begin{bmatrix} \hat{q}_1^1 & \cdots & q_i^1 & \cdots & \hat{q}_p^1 \\ \vdots & & \vdots & & \vdots \\ \hat{q}_1^M & \cdots & q_i^M & \cdots & \hat{q}_p^M \end{bmatrix},$$

which are identical to  $B$  with the exception that the  $i^{th}$  column is taken from  $A$ .

3. Compute  $M \times 1$  vectors of model outputs

$$y_A = f(A), \quad y_B = f(B), \quad y_{C_i} = f(C_i)$$

by evaluating the model at the input values in  $A$ ,  $B$ , and  $C_i$ . The evaluation of  $y_A$  and  $y_B$  requires  $2M$  model evaluations, whereas the evaluation of  $y_{C_i}$ ,  $i = 1, \dots, p$ , requires  $pM$  evaluations. Hence the total number of model evaluations is  $M(p+2)$ .

4. The estimates for the first-order sensitivity indices are

$$S_i = \frac{\text{var}[\mathbb{E}(Y|q_i)]}{\text{var}(Y)} = \frac{\frac{1}{M} y_A^T y_{C_i} - f_0^2}{\frac{1}{M} y_A^T y_A - f_0^2} = \frac{\frac{1}{M} \sum_{j=1}^M y_A^j y_{C_i}^j - f_0^2}{\frac{1}{M} \sum_{j=1}^M (y_A^j)^2 - f_0^2}, \quad (15.24)$$

where the mean is approximated by

$$f_0^2 = \left( \frac{1}{M} \sum_{j=1}^M y_A^j \right) \left( \frac{1}{M} \sum_{j=1}^M y_B^j \right). \quad (15.25)$$

The estimates for the total effects indices are

$$S_{T_i} = 1 - \frac{\text{var}[\mathbb{E}(Y|q_{\sim i})]}{\text{var}(Y)} = 1 - \frac{\frac{1}{M} y_B^T y_{C_i} - f_0^2}{\frac{1}{M} y_A^T y_A - f_0^2} = 1 - \frac{\frac{1}{M} \sum_{j=1}^M y_B^j y_{C_i}^j - f_0^2}{\frac{1}{M} \sum_{j=1}^M (y_A^j)^2 - f_0^2}. \quad (15.26)$$

The intuition for the algorithm is the following. In the scalar product  $y_A^T y_{C_i}$ , the response computed from values in  $A$  is multiplied by values for which all parameters except  $q_i$  have been resampled. If  $q_i$  is influential, then large (or small) values of  $y_A$  will be correspondingly multiplied by large (or small) values of  $y_{C_i}$  yielding a large value of  $S_i$ . If  $q_i$  is not influential, large and small values of  $y_A$  and  $y_{C_i}$  will occur more randomly and  $S_i$  will be small.

Details regarding the derivation of the algorithm and modifications to improve its accuracy are provided in [155, 214, 215].

## 15.2 Morris Screening

Screening methods provide an alternative to variance-based methods for identifying critical inputs to high-dimensional input spaces or models whose computational expense prohibits construction of Sobol indices. Screening methods generally provide the capability to rank parameters according to their importance but, unlike variance-based methods, they typically do not quantify how much more important one parameter is than another.

As detailed in the review paper [216], screening methods are in the class of *One factor At a Time (OAT)* methods in which one measures the variation in outputs as inputs are varied individually. The Morris algorithm [175] partially eliminates the local nature of OAT methods, which is one of their main limitations, by averaging over local derivative approximations to provide more global sensitivity measures. The goal of Morris screening is to identify those inputs or parameters—collectively termed *factors* in the literature—that are (i) negligible, (ii) linear and additive, or (iii) nonlinear or comprised of interactions between inputs.

We again consider the model

$$y = f(q), \quad q = [q_1, \dots, q_p].$$

We assume that each input term has been scaled to the interval  $[0, 1]$ , but, as noted in Remark 15.12, other scalings can be employed to facilitate computations.

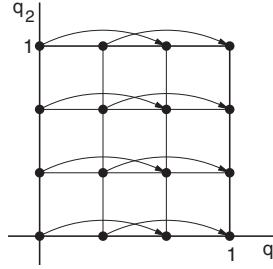
The concept of Morris screening is very simple; one averages coarse local sensitivity approximations, often termed *elementary effects*, over the input space to provide a measure of global sensitivity. Hence it is based on a linearization of the model. To construct the elementary effect, one partitions  $[0, 1]$  into  $\ell$ -levels, which, as illustrated in Figure 15.3, restricts each input to  $\ell$  values. The elementary effect associated with the  $i^{th}$  input is then defined by the difference quotient

$$d_i(q) = \frac{f(q_1, \dots, q_{i-1}, q_i + \Delta, q_{i+1}, \dots, q_p) - f(q)}{\Delta} = \frac{f(q + \Delta e_i) - f(q)}{\Delta}, \quad (15.27)$$

where the stepsize  $\Delta$  is chosen from the set

$$\Delta \in \left\{ \frac{1}{\ell-1}, \dots, 1 - \frac{1}{\ell-1} \right\}. \quad (15.28)$$

As illustrated in Figure 15.3, possible stepsizes for  $\ell = 4$  are  $\Delta \in \{\frac{1}{3}, \frac{2}{3}\}$ . If one denotes the set of gridpoints by  $\Gamma_\ell$ , the definition (15.27) holds for any input vector  $q$  such that  $q + \Delta e_i \in \Gamma_\ell$ , where  $e_i$  is a vector of zeros with one in the  $i^{th}$  component.



**Figure 15.3.** Four-level grid ( $\ell = 4$ ) for  $q = [q_1, q_2]$  with  $\Delta = \frac{2}{3}$ .

**Remark 15.11.** Due to the magnitude of  $\Delta$ , the elementary effect is a very coarse approximation of the local sensitivity. It can be used to rank the relative importance of inputs but not to resolve fine-scale gradient or local sensitivity behavior unless one employs significantly smaller stepsizes  $\Delta$ .

**Remark 15.12.** Rather than scaling parameters to the unit hypercube  $\Gamma = [0, 1]^p$ , some authors scale so that  $0 \leq q_i \leq \ell - 1$  [56]. This yields  $\Delta \in \{1, \dots, \ell - 1\}$  so that parameters are evaluated at integer values.

**Remark 15.13.** It is illustrated in Example 15.16 that the unscaled elementary effect  $d_i(q)$  given by (15.27) yields an incorrect classification of parameters for the linear model (15.1) since it lacks a mechanism to incorporate the variability of parameters. This motivates the use of the scaled elementary effect

$$d_i^\sigma(q) = \frac{f(q + \Delta e_i) - f(q)}{\Delta} \cdot \frac{\sigma_i}{\sigma_Y}, \quad (15.29)$$

where  $\sigma_i$  and  $\sigma_Y$  are the standard deviations of the parameter  $Q_i$  and response  $Y$  [223]. This is analogous to the sigma-normalized sensitivity relation (15.4), which is hybrid local-global in nature.

The elementary effects  $d_i(q)$  quantify the approximate, large scale, local sensitivity behavior at the point  $q$ . To provide a pseudoglobal sensitivity measure, one approximates the mean and variance of the finite-dimensional distribution  $G_i$  associated with each  $|d_i(q)|$  that is constructed by randomly sampling  $q$  from admissible points in  $\Gamma_\ell$ . As detailed in [215], the choice of the distribution associated with  $|d_i(q)|$  rather than  $d_i(q)$  avoids Type II errors, which can occur when the distribution has both positive and negative elements.

For  $r$  sample points, the sensitivity measures for  $q_i$  are taken to be the sampling mean and variance

$$\begin{aligned} \mu_i^* &= \frac{1}{r} \sum_{j=1}^r |d_i^j(q)|, \\ \sigma_i^2 &= \frac{1}{r-1} \sum_{j=1}^r (d_i^j(q) - \mu_i)^2, \quad , \quad \mu_i = \frac{1}{r} \sum_{j=1}^r d_i^j(q), \end{aligned} \quad (15.30)$$

where

$$d_i^j = \frac{f(q^j + \Delta e_i) - f(q^j)}{\Delta} \quad (15.31)$$

is the elementary effect associated with the  $i^{th}$  parameter and  $j^{th}$  sample. The mean quantifies the individual effect of the input on the output, whereas the variance estimates the combined effects of the input due to nonlinearities or interactions with other inputs. The latter interpretation is motivated by the observation that large variances indicate a strong dependence on neighboring input values.

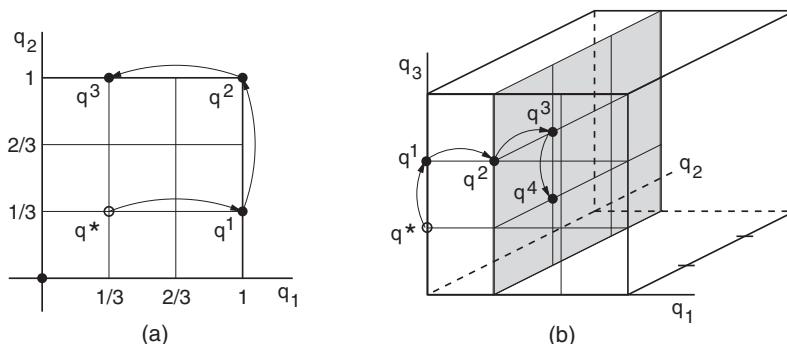
It is illustrated in Examples 15.17 and 15.18 that the ordering of  $\mu_i^*$  and  $\sigma_i$  can be used to rank parameters according to their relative importance in a manner similar to that provided by the total sensitivity indices  $S_{T_i}$ . This can be used to determine noninfluential parameters that can be fixed in subsequent model calibration, sensitivity analysis, and uncertainty quantification to reduce model complexity.

While  $\mu_i^*$  and  $\sigma_i$  can be used to screen the effects of individual parameters, they do not quantify the magnitude of parameter interactions. Second-order interactions can be screened by approximating cross derivatives in the manner detailed in [59, 69].

Implementation issues include the choice of  $\ell$ , choice of  $\Delta$ , and strategies to optimally sample  $r$  elementary effects from  $G_i$ . The choices of  $\ell$  and  $r$  are linked in the sense that larger values of both yield improved accuracy. As detailed in [175], taking  $\ell$  to be even and choosing  $\Delta = \frac{\ell}{2(\ell-1)}$  has the advantage that it guarantees equal probability sampling from the distributions  $G_i$ . This motivated our choices of  $\ell = 4$  and  $\Delta = \frac{2}{3}$  in Figure 15.3 and Example 15.14. We next discuss a strategy for efficiently sampling elementary effects from  $G_i$ .

### Morris Sampling Strategy

Since the computation of each elementary effect requires two model evaluations, naive sampling would require  $2pr$  model evaluations to construct  $\mu_i^*$  and  $\sigma_i$  using  $r$  sample points. In the Morris sampling algorithm, one employs neighbors—in the manner illustrated in Figure 15.4—to reduce the number of model evaluations



**Figure 15.4.** (a) Random initial vector  $q^*$  and model evaluations required to construct  $B^*$  for Example 15.14. (b) Example trajectory when  $p = 3$ .

required to construct  $p$  elementary effects from  $2p$  to  $p+1$ . In this manner  $\mu_i^*$  and  $\sigma_i$  can be constructed with  $(p+1)r$  samples.

To construct trajectories in which neighbors differ in only one component, as required for the difference formulas (15.27) and (15.31), one employs a  $(p+1) \times p$  sampling matrix  $B^*$  comprised of  $p+1$  model realizations with the elements in the  $i^{th}$  row representing the parameter values used in the  $i^{th}$  evaluation. This matrix is constructed so that for every column,  $j = 1, \dots, p$ , there are two rows that differ only in their  $j^{th}$  component. By subtracting the elements in consecutive rows, one can evaluate the  $p$  elementary effects associated with the random initial point  $q^*$ .

A deterministic  $B^*$  is given by

$$B^* = J_{p+1,p} q^* + \Delta B,$$

where  $B$  is taken to be a  $(p+1) \times p$  strictly lower triangular matrix of ones,  $J_{p+1,p}$  is a  $(p+1) \times p$  matrix of ones, and  $\Delta$  is the stepsize from (15.28). The difficulty is that elementary effects constructed in this manner are not randomly selected.

To obtain random samples from  $G_i$ , one employs the orientation matrix

$$B^* = \left( J_{p+1,1} q^* + \frac{\Delta}{2} [(2B - J_{p+1,p}) D^* + J_{p+1,p}] \right) P^*,$$

where  $D^*$  is a  $p \times p$  diagonal matrix whose elements are randomly chosen from the set  $\{-1, 1\}$ , and the  $p \times p$  matrix  $P^*$  is constructed by randomly permuting the columns of a  $p \times p$  identity matrix.

**Example 15.14.** Take  $p = 2$ ,  $\ell = 4$ , and  $\Delta = \frac{2}{3}$ , as depicted in Figure 15.4(a). The seed value  $q^* = [\frac{1}{3}, \frac{1}{3}]$  and choices

$$D^* = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \quad P^* = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}, \quad J = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}$$

yield

$$B^* = \begin{bmatrix} 1 & 1/3 \\ 1 & 1 \\ 1/3 & 1 \end{bmatrix}.$$

Hence one employs model evaluations at  $q^1 = [1, \frac{1}{3}]$ ,  $q^2 = [1, 1]$ , and  $q^3 = [\frac{1}{3}, 1]$  to construct two elementary effects  $d_1$  and  $d_2$  associated with  $q^*$ . We note that  $q^*$  seeds the algorithm but is not employed as a sample point. A similar trajectory for  $p = 3$  is plotted in Figure 15.4(b).

The final issue concerns the generation of the seed parameter values  $q^*$ . In Morris' algorithm,  $q^*$  was generated randomly from values in the parameter space. However, this can lead to nonoptimal coverage of the parameter space for high-dimensional problems. This is addressed in [60] by an algorithm that optimizes the distance between initial points to ensure that they best cover the input space. The determination of optimal distance metrics remains an open research topic.

**Remark 15.15.** The only point at which the parameter densities play a role is in the generation of random seed values  $q^*$ . For nonuniform densities, samples can be

mapped to  $[0, 1]$  using an appropriate transformation. If joint densities are available, one can relax the requirement that parameters be mutually independent.

**Example 15.16.** We revisit the linear portfolio model

$$Y = c_1 Q_1 + c_2 Q_2$$

of Example 15.1 where  $c_1 = 2$ ,  $c_2 = 1$  and  $Q_1 \sim N(0, 1)$ ,  $Q_2 \sim N(0, 9)$ . For any choice of  $\ell$ ,  $r$ , and  $\Delta$ , the elementary effects are  $d_i = \pm c_i$  so that  $|d_i|$  is the local sensitivity  $s_i = \frac{\partial Y}{\partial Q_i}$  defined in (15.3). Hence the means of  $G_i$  are  $\mu_1^* = 2$  and  $\mu_2^* = 1$ , which reflect the local behavior rather than the uncertainty associated with  $Q_1$  and  $Q_2$ . We note that  $\mu_i$  can attain values between  $-2$  and  $2$  since  $d_i = \pm c_i$ . This illustrates an advantage of considering the distribution  $G_i$  associated with  $|d_i|$  and a limitation of this sampling strategy using unscaled elementary effects  $d_i$  for global sensitivity analysis. Alternatively, use of the sigma-normalized elementary effect  $d_i^\sigma$  given by (15.29) yields  $\mu_1^* = \frac{2}{\sqrt{13}}$  and  $\mu_2^* = \frac{3}{\sqrt{13}}$ , which are the same as the sigma-normalized sensitivity values obtained in (15.5). These values are consistent with the scatterplot information shown in Figure 15.1.

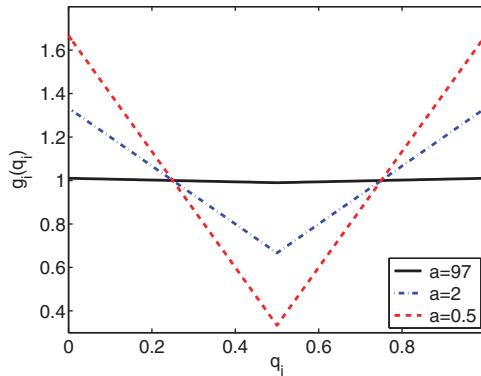
**Example 15.17.** Consider the function

$$Y = \prod_{i=1}^p g_i(Q_i) \quad , \quad g_i(Q_i) = \frac{|4Q_i - 2| + a_i}{1 + a_i}, \quad (15.32)$$

attributed to Sobol' [227], where  $a_i \geq 0$  are fixed, deterministic coefficients. Since

$$1 - \frac{1}{1 + a_i} \leq g_i(q_i) \leq 1 + \frac{1}{1 + a_i},$$

the coefficients  $a_i$  determine the relative importance of the random parameter  $Q_i$ , as illustrated in Figure 15.5. This function is widely employed as a test case for global



**Figure 15.5.** Component functions  $g_i(q_i)$  for  $a = 0.5$ ,  $a = 2$ , and  $a = 97$ .

|           | $Q_1$                | $Q_2$                | $Q_3$                | $Q_4$                | $Q_5$                | $Q_6$                |
|-----------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|
| $a_i$     | 78                   | 12                   | 0.5                  | 2                    | 97                   | 33                   |
| $D_i$     | $5.3 \times 10^{-5}$ | $2.0 \times 10^{-3}$ | $1.5 \times 10^{-1}$ | $3.7 \times 10^{-2}$ | $3.5 \times 10^{-5}$ | $2.9 \times 10^{-4}$ |
| $S_i$     | $2.8 \times 10^{-4}$ | $1.0 \times 10^{-2}$ | $7.7 \times 10^{-1}$ | $1.9 \times 10^{-1}$ | $1.8 \times 10^{-4}$ | $1.5 \times 10^{-3}$ |
| $S_{T_i}$ | $3.3 \times 10^{-4}$ | $1.2 \times 10^{-2}$ | $8.0 \times 10^{-1}$ | $2.2 \times 10^{-1}$ | $2.1 \times 10^{-4}$ | $1.8 \times 10^{-3}$ |

**Table 15.1.** Coefficients, first-order partial variances, and Sobol indices for  $p = 6$ .

sensitivity analysis since it is strongly nonlinear, nonmonotonic, and has nonzero interaction terms by construction. Furthermore, the partial and total variances have the explicit representations

$$\begin{aligned} D_i &= \text{var}[\mathbb{E}(Y|q_i)] = \frac{1}{3(1+a_i)^2}, \\ D_{ij} &= \text{var}[\mathbb{E}(Y|q_i, q_j)] - D_i - D_j = D_i D_j, \\ D &= \text{var}(Y) = -1 + \prod_{i=1}^p (1+D_i) \end{aligned} \quad (15.33)$$

for  $Q_i \sim \mathcal{U}(0, 1)$ ,  $i = 1, \dots, p$ . This yields explicit formulas for the Sobol indices  $S_i = \frac{D_i}{D}$ ,  $S_{ij} = \frac{D_{ij}}{D}$  and total sensitivity  $S_{T_i}$  given by (15.14).

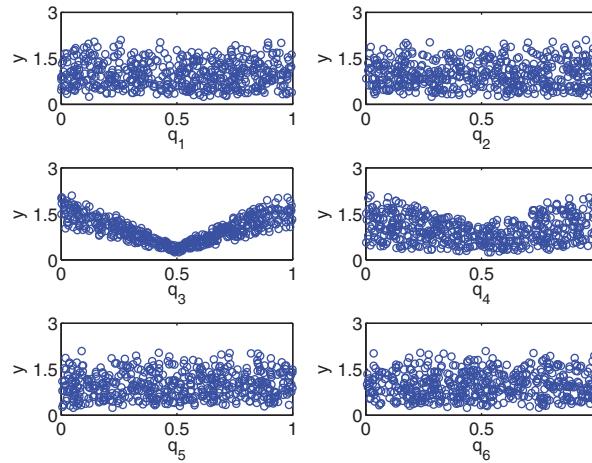
The first-order partial variances and Sobol indices for six coefficients  $a_i$  are summarized in Table 15.1. We then took  $\ell = 4$ ,  $\Delta = \frac{2}{3}$ , and  $r = 4$  and employed the sampled trajectories compiled in Table 3.2 of [215] to obtain the sensitivity measures  $\mu, \mu^*$ , and  $\sigma$  summarized in Table 15.2. Finally, scatterplots obtained with 500 joint realizations  $q_1, \dots, q_6, y$  are plotted in Figure 15.6.

For this choice of coefficients  $a_i$ ,  $\sum_{i=1}^6 S_i = 0.97$  and  $\sum_{i=1}^6 S_{T_i} = 1.03$ , so second-order interactions are negligible. This is due to the fact that the large values of  $a_i$  yield the small reported partial variances  $D_i$  which in turn produce negligible cross partial variances  $D_{ij} = D_i D_j$ . The effect of significant interaction terms due to small values of  $a_i$  is illustrated in Exercise 15.4.

The scatterplots in Figure 15.6 reveal that the parameter  $Q_3$  is most influential since the realizations  $(q_3, y)$  clearly exhibit the underlying functional behavior of the model (15.32). The parameter  $Q_4$  is the second most influential, whereas the remaining parameters have significantly less influence, as illustrated by the observation that their scatterplots are nearly uniform. This trend is reflected in the partial

|            | $Q_1$  | $Q_2$  | $Q_3$  | $Q_4$  | $Q_5$ | $Q_6$  |
|------------|--------|--------|--------|--------|-------|--------|
| $\mu_i$    | -0.006 | -0.078 | -0.130 | -0.004 | 0.012 | -0.004 |
| $\mu_i^*$  | 0.056  | 0.277  | 1.760  | 1.185  | 0.035 | 0.099  |
| $\sigma_i$ | 0.064  | 0.321  | 2.049  | 1.370  | 0.041 | 0.122  |

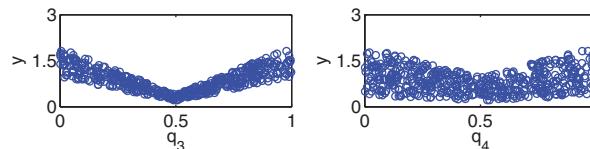
**Table 15.2.** Estimated Morris sensitivity measures obtained with  $\ell = 4$ ,  $\Delta = \frac{2}{3}$ , and  $r = 4$  from [215].



**Figure 15.6.** Scatterplots of  $y$  versus  $q_1$  through  $q_6$  constructed using 500 joint realizations.

variances, Sobol indices, and Morris sensitivity measures, where it is observed that  $D_3, D_4, S_3, S_4, \mu_3, \mu_4$  and  $\sigma_3, \sigma_4$  are dominant over the remaining sensitivity indices.

Because  $Q_1, Q_2, Q_5, Q_6$  are relatively insensitive, their values can be fixed for subsequent model calibration and sensitivity analysis. To illustrate, the scatterplots obtained with  $q_1 = q_2 = q_5 = q_6 = \frac{1}{2}$  and  $Q_3 \sim \mathcal{U}(0, 1), Q_4 \sim \mathcal{U}(0, 1)$  are plotted in Figure 15.7. Comparison with Figure 15.6 illustrates that the effect of uncertainty in  $Q_3$  and  $Q_4$  on  $Y$  are nearly the same in the two cases. This illustrates the manner in which global sensitivity analysis can reduce the number of critical random parameters without significantly diminishing the model accuracy.



**Figure 15.7.** Scatterplots of  $y$  versus  $q_3$  and  $q_4$  with  $q_1 = q_2 = q_5 = q_6 = \frac{1}{2}$ .

## 15.3 Time- or Space-Dependent Responses

So far in this chapter, we have assumed that the nonlinear model  $Y = f(Q)$  is scalar-valued and a function only of the parameters. However, several of the models in Chapters 2 and 3 are also functions of time, space, or other independent variables. We summarize here issues associated with global sensitivity analysis for time- and space-dependent responses

$$Y(t) = f(t, u, Q) \quad (15.34)$$

or

$$Y(x) = f(x, u, Q), \quad (15.35)$$

where  $u$  is the state and  $x \in \mathbb{R}^1, \mathbb{R}^2$ , or  $\mathbb{R}^3$ .

### Time-Dependent Response

We consider first the time-dependent case. The most direct approach is to construct a set of sensitivity indices  $\{S_{T_i}(t_j)\}$  or  $\{\mu_i^*(t_j)\}$ ,  $i = 1, \dots, p$ , at time points  $t_j$  of interest to quantify the influence of parameters throughout the time interval. Sensitivity measures constructed in this manner can indicate whether the relative influence of parameters changes as a function of time. Alternatively, the time-dependent response can be integrated to obtain a scalar-valued response

$$Y = \int_{t_0}^{t_f} f(t, u, Q) dt \quad (15.36)$$

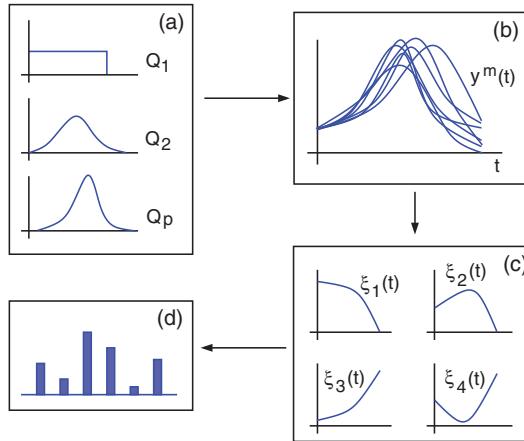
if the objective is to identify most influential parameters for the entire time interval. As illustrated in Example 15.18, one might employ this approach if the goal is to reduce the model complexity by fixing noninfluential parameters.

For certain applications, one may seek to extract salient features of the time-dependent output and quantify how they depend on parameter uncertainties. One approach is to combine functional principal component analysis (fPCA) with the variance-based or screening techniques of Sections 15.1 or 15.2 to identify those parameters and their interactions that most strongly influence a dynamical model. We summarize the methodology of [242] and refer readers to that reference for details illustrating the approach for an insulin signaling model.

The global sensitivity analysis approach based on fPCA is depicted in Figure 15.8. One first generates  $M$  parameter values  $\{q_i^m\}_{m=1}^M$  from respective densities  $\rho_{Q_i}(q_i)$  using the sampling techniques detailed in Sections 15.1 or 15.2. For example, this would entail quasi-random number generation if using Algorithm 15.10 or the generation of seed values  $q^*$  if using Morris screening. Based on the assumption of mutually independent parameters, this yields  $M$  parameter vectors  $q^m = [q_1^m, \dots, q_p^m]$ . For each parameter vector, one evaluates the model to obtain  $M$  time-dependent responses  $\{y^m(t)\}$ . Functional principal component analysis (fPCA) is then used to construct a low number  $M_{pc} \ll M$  of functions  $\xi_k(t)$  and coefficients  $y_{mk}$  so that each response can be represented as

$$y^m(t) = \sum_{k=1}^{M_{pc}} y_{mk} \xi_k(t) \quad , \quad m = 1, \dots, M.$$

Because the coefficient  $y_{mk}$  quantifies the degree to which the component  $\xi_k(t)$  contributes to the response  $y^m(t)$ , determination of how parameters influence each coefficient  $y_{mk}$  provides a measure of how they will influence the model behavior described by  $\xi_k(t)$ . The techniques of Sections 15.1 and 15.2 are then used to determine those parameters that have the most influence on the fPCA coefficients so that noninfluential parameters can be fixed to reduce model complexity.



**Figure 15.8.** (a) Sample from marginal densities  $\rho_{Q_i}(q_i)$  to construct  $M$  parameter vectors  $q^m = [q_1^m, \dots, q_p^m]$  and (b) responses  $y^m(t) = f(t, u, q^m)$ ,  $m = 1, \dots, M$ . (c) Functional principal components  $\xi_k(t)$  and (d) sensitivity of coefficients  $y_{mk}$  to model parameters.

Details comparing the performance of Sobol indices and Morris screening in this framework are illustrated for a complex insulin signaling model in [242]. Those results demonstrate that for this application, the Morris screening approach provides qualitative sensitivity measures that are consistent with the Sobol indices but at substantially less computational cost—approximately 15 minutes for Morris screening compared with approximately 1.06 days for computing Sobol indices.

### Spatially Dependent Response

Techniques for global sensitivity analysis for spatially varying responses (15.35), or responses that are functions of other independent variables, are similar to those for time-varying responses; hence we simply provide references. Techniques for general functional outputs are discussed in [58], whereas the analytic and numerical computation of spatially varying Sobol indices is detailed in [166]. The paper [155] provides an overview of techniques for global sensitivity analysis with spatially varying models along with a review of applications.

**Example 15.18.** To illustrate the global sensitivity analysis of a time-dependent model, we consider the SIR disease model

$$\begin{aligned} \frac{dS}{dt} &= \delta N - \delta S - \gamma k I S \quad , \quad S(0) = S_0, \\ \frac{dI}{dt} &= \gamma k I S - (r + \delta) I \quad , \quad I(0) = I_0, \\ \frac{dR}{dt} &= r I - \delta R \quad , \quad R(0) = R_0, \end{aligned} \tag{15.37}$$

where  $S(t)$ ,  $I(t)$ , and  $R(t)$  are the number of susceptible, infectious, and recovered individuals in a population of size  $N$ . As noted in Example 3.4, the parameters are

$$q = [\gamma, k, r, \delta],$$

where  $\gamma$ ,  $k$ , and  $r$  respectively denote the infection coefficient, the interaction coefficient which quantifies the probability that an individual comes in contact with others, and the recovery rate. Whereas  $\gamma$  and  $k$  both influence the disease dynamics, they differ in the sense that  $\gamma$  is a property of the disease, while  $k$  reflects the degree of personal contact. Hence  $\gamma$  is difficult to control, while  $k$  can be controlled via policies such as isolation or quarantine. Finally, the birth and death rates are assumed to be equal with both denoted by  $\delta$ .

We take the parameter distributions to be

$$\gamma \sim \mathcal{U}(0, 1), \quad k \sim \text{Beta}(\alpha, \beta), \quad r \sim \mathcal{U}(0, 1), \quad \delta \sim \mathcal{U}(0, 1), \quad (15.38)$$

where the beta distribution is defined in Definition 4.16. The choice of uniform distributions for  $\gamma$ ,  $r$ , and  $\delta$  is made to reflect limited prior knowledge about the parameters. As illustrated in Figure 15.9, the beta distribution can be tuned, through the choice of  $\alpha$  and  $\beta$ , to quantify the degree to which individuals interact with others in the population. We consider two cases, Beta(2, 7) and Beta(0.2, 15), which reflect large and limited degrees of interaction. The initial values are taken to be  $S_0 = 900$ ,  $R_0 = 0$ , and  $I_0 = 100$  so that  $N = 1000$ . The scalar response is taken to be

$$y = \int_0^5 R(t, q) dt,$$

where  $R(t, q)$  is computed by numerically integrating the coupled system (15.37).

### Case 1: Large Degree of Interactions Beta(2,7)

The choice  $\alpha = 2$ ,  $\beta = 7$  models the case when individuals have a large probability of interacting with up to half the population and a low probability of meeting with everyone.

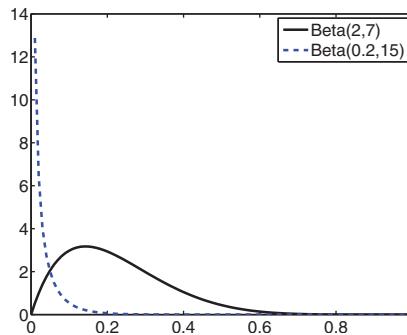


Figure 15.9. Beta distributions Beta(2, 7) and Beta(0.2, 15).

|                          | $\gamma$ | $k$     | $r$    | $\delta$ |
|--------------------------|----------|---------|--------|----------|
| $S_i$                    | 0.0997   | 0.0312  | 0.7901 | 0.1750   |
| $S_{T_i}$                | -0.0637  | -0.0541 | 0.5634 | 0.2029   |
| $\mu_i^* (\times 10^3)$  | 0.2532   | 0.2812  | 2.0184 | 1.2328   |
| $\sigma_i (\times 10^3)$ | 0.9539   | 1.6245  | 6.6748 | 3.9886   |

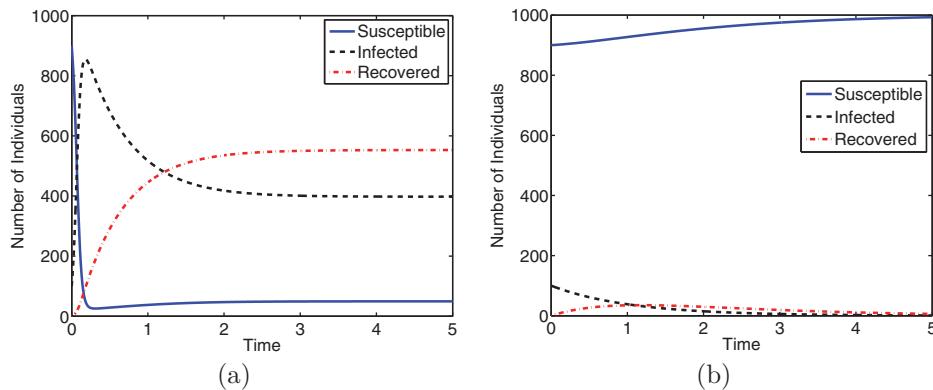
**Table 15.3.** Sobol indices and Morris sensitivity measures for  $k \sim \text{Beta}(2, 7)$ .

The first-order Sobol indices  $S_i$  and  $S_{T_i}$ , computed using the relations (15.24) and (15.25) with  $M = 1000$ , are summarized in Table 15.3 along with the Morris measures  $\mu_i^*$  and  $\sigma_i$  given by (15.30) with  $r = 20$  and  $\ell = 40$ . The total indices and absolute means both indicate that for this case, the recovery rate has primary influence, the birth-death rate  $\delta$  has secondary influence, and the infection and contact coefficients  $\gamma$  and  $k$  have negligible influence. The negative values for  $S_{T_1}$  and  $S_{T_2}$  reflect the approximate nature of the expressions (15.24) and (15.25) for small or moderate sample sizes  $M$ .

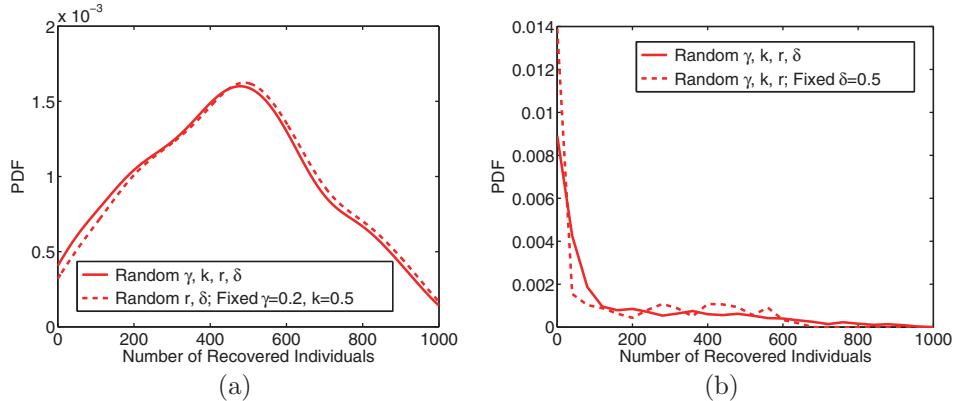
A representative realization of trajectories is plotted in Figure 15.10(a). Because of the large rate of interactions, the number of susceptible individuals rapidly diminishes with commensurate growth in the number of infected. The recovery is slower due to the magnitude of the rate  $r$ . The influence of the birth rate  $\delta$  is easily observed in  $S(t)$  after 0.5 seconds.

To illustrate the use of the model for uncertainty quantification, the density for  $R(t_f)$  at  $t_f = 5$  s was constructed from the  $M = 1000$  realizations of the model solution and is plotted in Figure 15.11(a). The uncertainty in parameters induces significant uncertainty in this response.

The parameters  $\gamma$  and  $k$  are relatively noninfluential in this case because the number of contacts ensures rapid infection. Hence we can fix these parameters



**Figure 15.10.** Dynamics of  $S(t)$ ,  $I(t)$ , and  $R(t)$  for a single realization of parameters from (15.38) with (a)  $k \sim \text{Beta}(2, 7)$  and (b)  $k \sim \text{Beta}(0.2, 15)$ .



**Figure 15.11.** Densities for  $R(t_f)$  at  $t_f = 5$  with (a)  $k \sim \text{Beta}(2, 7)$  and (b)  $k \sim \text{Beta}(0.2, 15)$ ; — all parameters random, - - noninfluential parameters fixed.

without significantly changing the distribution of responses. To illustrate, the density obtained with random  $r$  and  $\delta$  and  $\gamma = 0.5$ ,  $k = 0.2$  fixed is also plotted in Figure 15.11(a). It is observed that there is little difference between this density and that obtained with random  $\gamma, k, r, \delta$ .

### Case 2: Limited Interactions Beta(0.2,15)

The mechanism in the model that we can directly control is the interaction coefficient  $k$ . As illustrated in Figure 15.9, the choice  $k \sim \text{Beta}(0.2, 15)$  has a high probability of limited interactions, so this enforces isolation of infected individuals.

The sensitivity indices and Morris sensitivity measures in Table 15.4 indicate that  $k$  is now the most influential parameter and the birth-death rate  $\delta$  is the least influential. The trajectory in Figure 15.10(b) illustrates a typical realization in which the disease does not spread and the number of infected diminishes due to the death rate  $\delta$ . This illustrates that if the infection coefficient is not too high, isolation of infected individuals can effectively deter the disease spread.

The densities for  $R(t_f)$  at  $t_f = 5$ , plotted in Figure 15.11(b), illustrate that the percentage of the population with a high recovery rate is small since few were

|                          | $\gamma$ | $k$    | $r$    | $\delta$ |
|--------------------------|----------|--------|--------|----------|
| $S_i$                    | 0.0711   | 0.6233 | 0.1450 | 0.0910   |
| $S_{T_i}$                | 0.1428   | 0.7026 | 0.1801 | 0.0248   |
| $\mu_i^* (\times 10^3)$  | 3.4296   | 4.6425 | 2.9326 | 1.7947   |
| $\sigma_i (\times 10^3)$ | 6.0782   | 9.3842 | 3.5811 | 2.0125   |

**Table 15.4.** Sobol indices and Morris sensitivity measures for  $k \sim \text{Beta}(0.2, 15)$ .

infected. Furthermore, similar results are obtained when  $\delta$  is fixed at  $\frac{1}{2}$  as motivated by the results in Table 15.4. It is established in Exercise 15.5 that one can obtain a similar density with both  $\gamma$  and  $\delta$  fixed.

## 15.4 Notes and References

An important objective of global sensitivity analysis is to ascertain noninfluential parameters which can be fixed to reduce the complexity of models for subsequent model calibration and uncertainty propagation. Hence it can be used to guide which parameters are included in the HDMR surrogate models discussed in Section 13.5 or for parameter selection as detailed in Section 6.2.

However, as illustrated in Example 15.18, *the relative influence of parameters can be a function of their distribution, which is often unknown and to be determined using Bayesian model calibration techniques.* A conservative approach might be to treat parameters as uniform during this initial sensitivity analysis to reflect poorly known prior knowledge. However, this can yield poor results if parameters are incorrectly fixed in surrogate or full-order models used to estimate parameter densities. This can be partially addressed by employing alternative algorithms to construct Morris seed values  $q^*$  or using local or global sensitivity techniques which do not utilize parameter distributions. However, theoretical algorithmic solutions to this problem in the context of global sensitivity analysis are generally lacking and constitute an area of current research.

The variance-based and Morris screening techniques each have advantages and disadvantages. The Sobol indices  $S_i$ ,  $S_{ij}$ , and  $S_{Ti}$  provide a comprehensive measure that quantifies the influence of parameter uncertainties on the response variance, but their computational cost can be extensive or prohibitive. Morris screening techniques can be used to rank the relative influence of parameters at a fraction of their computational cost, but, as illustrated in Example 15.16, they can yield incorrect results even for linear problems if global effects are not correctly incorporated. Furthermore, they provide qualitative rather than quantitative measures of each parameter's influence. Both methods generally rely on the assumption that parameters are mutually independent, although this assumption is easily relaxed for the screening methods by choosing appropriate methods to generate seed values  $q^*$ .

The gradient-based methods of [1, 21, 66] fall within the same general framework as Morris screening in that local, linear sensitivities are evaluated at random input values to provide pseudoglobal sensitivity measures. Details regarding these approaches are provided in Section 6.2.2.

Readers are referred to [215, 216, 217] for an overview of issues pertaining to global sensitivity analysis. This includes regression-based sensitivity analysis methods, which we did not cover in this chapter. The use of stochastic polynomial methods to construct surrogate models that permit the analytic computation of Sobol indices based on the expansion coefficient is detailed in [68, 241]. This includes a comparison of the methods for the Ishigami function (15.22), the Sobol function (15.32), and a finite element model for soil mass. The computation of Sobol sensitivity indices based on the cut-HDMR and RS-HDMR expansions of

Section 13.5 is detailed in [153, 177] and illustrated in the context of a finite element model for a masonry wall. Finally, readers are referred to [65, 165] for details illustrating global sensitivity analysis and uncertainty quantification for biological models.

## 15.5 Exercises

**Exercise 15.1.** Use the relation (15.16) to show that the first-order partial variances  $D_i$  have the probabilistic interpretation (15.17).

**Exercise 15.2.** For the Ishigami function

$$f(q) = \sin q_1 + a \sin^2 q_2 + b q_3^4 \sin q_1$$

discussed in Examples 13.8 and 15.9, establish the total and partial variance relations (15.23). For  $a = b = 0.1$ , apply Morris screening with various levels  $\ell$  and stepsizes  $\Delta$  and compare the parameter rankings with the analytic relations.

**Exercise 15.3.** Consider the Sobol function

$$Y = \prod_{i=1}^p \frac{|4Q_i - 2| + a_i}{1 + a_i}$$

with  $Q_i \sim \mathcal{U}(0, 1)$ . Establish the variance relations (15.33).

**Exercise 15.4.** Consider the function

$$Y = \prod_{i=1}^p \frac{|4Q_i - 2| + a_i}{1 + a_i},$$

discussed in Example 15.17, with  $p = 6$  and  $a = [0.2, 0.3, 0.2, 0.1, 0.4, 0.05]$ . Compute the Sobol indices  $S_i$  and  $S_{T_i}$ , and show that the second-order effects are significant for this parameter set.

**Exercise 15.5.** For the SIR model (15.37) in Example 15.18, compute the Sobol indices  $S_i$ ,  $S_{T_i}$  and Morris measures  $\mu_i^*$ ,  $\sigma_i$  for  $k \sim \text{Beta}(2, 7)$  and  $k \sim \text{Beta}(0.2, 15)$  and compare both the results and the required computational times. For  $k \sim \text{Beta}(0.2, 15)$ , compare the density for  $R(t_f)$  at  $t_f = 5$  obtained with random  $\gamma, k, r, \delta$  with that obtained with  $\gamma$  and  $\delta$  fixed at 0.5.

## Appendix A

# Concepts from Functional Analysis

We summarize here concepts from functional analysis that are employed in the text. The discussion is necessarily selective, and readers are referred to cited references for additional theory and details.

### Functionals, Dual Spaces, and Hilbert Spaces

Throughout this discussion, we take  $X$  to be a vector space,  $Y$  is a normed space, and  $T : \text{dom}(T) \subset X \rightarrow Y$  is a possibly nonlinear operator. For the case  $Y = \mathbb{R}$ , the operator is a real-valued functional which we denote by  $J$ .

**Definition A.1 (Dual Space  $X^*$ ).** Let  $X$  be a normed space. The set  $\mathcal{L}(X, \mathbb{R})$  of all bounded linear functionals on  $X$  also constitutes a normed space with the norm defined by

$$\|J\| = \sup_{\substack{x \in X \\ x \neq 0}} \frac{|J(x)|}{\|x\|} = \sup_{\substack{x \in X \\ \|x\|=1}} |J(x)|. \quad (\text{A.1})$$

This is termed the dual space  $X^*$  of  $X$ .

**Example A.2.** Let  $X = C[a, b]$  denote the space of continuous functions on the interval  $I = [a, b]$ , and define  $J : X \rightarrow \mathbb{R}$  by

$$J(x) = \int_a^b x(t) dt.$$

The linearity of  $J$  follows directly from the linearity of integration. Furthermore,

$$|J(x)| = \left| \int_a^b x(t) dt \right| \leq (b-a) \max_{t \in I} |x(t)| = (b-a)\|x\|$$

so that  $\|J\| \leq b-a$ . If we consider  $x = x_0 = 1$ , it follows that

$$\|J\| \geq \frac{|J(x_0)|}{\|x_0\|} = |J(x_0)| = \int_a^b dt = b-a$$

so that  $\|J\| = b-a$ .

The norm on a vector space generalizes the concept of the length for a vector. Similarly, we will typically impose inner products which provide measures of orthogonality as an extension of the Euclidean dot product. Specifically, we will generally consider operations defined on Hilbert spaces.

**Definition A.3 (Hilbert Space).** An inner product space is a vector space  $X$  with an associated inner product  $\langle \cdot, \cdot \rangle$ . A complete inner product space is termed a Hilbert space. We note that the inner product defines a norm

$$\|x\| = \sqrt{\langle x, x \rangle}$$

and metric

$$d(x, y) = \|x - y\|$$

on  $X$ .

**Example A.4.** Consider an inner product space  $X$ , and define the functional

$$\ell(x) = \langle u, x \rangle, \quad (\text{A.2})$$

where  $u \in X$  is arbitrary and fixed. The linearity of  $\ell$  is obvious, and it follows from the Cauchy–Schwarz inequality that

$$|\ell(x)| \leq \|u\| \|x\|$$

so that  $\|\ell\| \leq \|u\|$ . Furthermore,  $|\ell(u)| = \langle u, u \rangle = \|u\|^2$  so that  $\|\ell\| \geq \frac{|\langle u, u \rangle|}{\|u\|} = \|u\|$  and hence  $\|\ell\| = \|u\|$ .

Given a Hilbert space, we can always construct a bounded linear functional in this manner. One of the “big” theorems of functional analysis is the Riesz representation theorem, which establishes that in a Hilbert space, the converse is also true. The reader is referred to [139] for proofs.

**Theorem A.5 (Riesz Representation Theorem).** To every bounded linear functional  $\ell$  on a Hilbert space  $H$ , there corresponds a uniquely defined element  $v$  such that

$$\ell(u) = \langle u, v \rangle$$

for all  $u \in H$ . Furthermore, the norm of  $\ell$  satisfies

$$\|\ell\| = \|v\|.$$

**Remark A.6.** The Riesz representation theorem establishes that Hilbert spaces are self-dual in the sense that for each  $\ell \in H^*$ , there exists a unique  $v \in H$  such that

$$\ell(u) = (\ell, u) = \langle u, v \rangle$$

for all  $u \in H$ . Here the symbol  $(\cdot, \cdot)$  denotes the pairing of  $H^*$  and  $H$  and  $(\ell, u)$  is the real number  $\ell(u)$ . The mapping  $\ell \mapsto v$  is a linear isomorphism of  $H^*$  onto  $H$ .

## Gâteaux and Fréchet Derivatives

The local sensitivity analysis of Chapter 14 relies heavily on the concepts of Gâteaux variations and differentials, which we summarize here. Additional details can be found in [50, 204].

**Definition A.7 (Gâteaux Variation, Differential, and Derivative).** For  $T : \text{dom}(T) \subset X \rightarrow Y$ , consider  $x \in \text{dom}(T)$  and arbitrary  $\eta \in X$ . If the limit

$$\delta T(x; \eta) = \lim_{\varepsilon \rightarrow 0} \frac{1}{\varepsilon} [T(x + \varepsilon\eta) - T(x)]$$

exists for each  $\eta \in X$ , then  $\delta T(x; \eta)$  is termed the Gâteaux variation of  $T$  at  $x$  with increment or perturbation  $\eta$ .

For functionals  $J$ , the Gâteaux variation, when it exists, is

$$\delta J(x; \eta) = \frac{d}{d\varepsilon} J(x + \varepsilon\eta) \Big|_{\varepsilon=0}.$$

Note that for each fixed  $x \in \text{dom}(T)$ ,  $\delta J(x; \eta)$  is a functional with respect to  $\eta \in X$ . We also note that  $\delta J(x; \eta)$  is neither necessarily linear nor continuous with respect to  $\eta$ , so it may not map  $X$  to  $X^*$ . If  $\delta J(x; \eta)$  is linear and continuous with respect to  $\eta$ , and hence  $\delta J(x; \eta) = DJ(x)\eta$ , then  $\delta J : X \rightarrow X^*$  is termed the Gâteaux differential and  $DJ(x) : X \rightarrow \mathbb{R}$  is the Gâteaux derivative of  $J$  at  $x$ . Details regarding necessary and sufficient conditions to establish that  $\delta J(x; \eta)$  is linear and continuous in  $\eta$  can be found in [50].

**Example A.8.** Let  $J : \mathbb{R}^2 \rightarrow \mathbb{R}$  by

$$J(x) = \begin{cases} x_1^2(1 + 1/x_2) & , x_2 \neq 0 \\ 0 & , x_2 = 0. \end{cases}$$

Here

$$\lim_{\varepsilon \rightarrow 0} \frac{1}{\varepsilon} [J(x + \varepsilon\eta) - J(x)] = \frac{\eta_1^2}{\eta_2}$$

so that  $\delta J(x; \eta)$  is not linear with respect to  $\eta$ .

We note that the Gâteaux variation generalizes the concept of the directional derivative from calculus and the first variation in the calculus of variations. Because the Gâteaux variation and derivative require no norm on  $X$ , they cannot be directly used to establish continuity. This stronger concept is provided by the Fréchet derivative, which generalizes the calculus concept of differentiability.

**Definition A.9 (Fréchet Differential and Derivative).**  $T$  is said to be Fréchet differentiable at  $x \in \text{dom}(T)$  in the normed space  $X$  if for each  $\eta \in X$ , there exists  $\delta T(x; \eta) = DJ(x)\eta \in Y$ , which is linear, is continuous with respect to  $\eta$ , and satisfies

$$\lim_{\|\eta\| \rightarrow 0} \frac{\|T(x + \eta) - T(x) - \delta T(x; \eta)\|}{\|\eta\|} = 0.$$

When it exists,  $\delta T(x; \eta)$  is termed the Fréchet differential of  $T$  at  $x$  with increment  $\eta$  and  $DJ(x)$  is termed the Fréchet derivative. Further relations between the Gâteaux and Fréchet derivatives can be found in [50].

### Adjoint Operators

The use of adjoint operators is central to the adjoint sensitivity analysis procedure (ASAP) detailed in Chapter 14. We focus on Hilbert space adjoints and refer the reader to [50, 139] for the theory of adjoint operators defined on Banach spaces. The text [145] is recommended for additional historical perspective and development of adjoint operators for differential operators with various boundary conditions. We refer the reader to [107] for a description of adjoint operators in the context of design and sensitivity analysis for structural systems and [22] for theory and examples illustrating the role of adjoints in PDEs, parameter estimation, and distributed control theory. Finally, the text [235] provides a very nice introduction to adjoint operators and their relation to Green's functions for differential equations.

**Definition A.10 (Hilbert Space Adjoint).** Let  $H_1$  and  $H_2$  be Hilbert spaces with associated inner products  $\langle \cdot, \cdot \rangle_{H_1}$  and  $\langle \cdot, \cdot \rangle_{H_2}$ , and let  $L : H_1 \rightarrow H_2$  be a bounded linear operator. The Hilbert space adjoint operator  $L^*$  of  $L$  is the bounded linear operator  $L^* : H_2 \rightarrow H_1$  such that for all  $u \in H_1$  and  $v \in H_2$ ,

$$\langle Lu, v \rangle_{H_2} = \langle u, L^*v \rangle_{H_1}.$$

The norm is  $\|L^*\| = \|L\|$ .

As detailed in [139], the uniqueness of the adjoint operator results from the Riesz representation theory, which establishes that the dual  $H^*$  of a Hilbert space can be canonically identified with  $H$  in the sense detailed in Remark A.6. If  $L : H \rightarrow H$  is a bounded linear operator,  $L$  is said to be self-adjoint if  $L = L^*$ .

**Example A.11.** Let  $A : \mathbb{R}^n \rightarrow \mathbb{R}^m$  be a matrix, and consider the usual Euclidean dot product  $\langle x, y \rangle = x^T y$ . Since

$$\langle Ax, y \rangle = (Ax)^T y = x^T A^T y = \langle x, A^T y \rangle,$$

it follows that  $A^* = A^T$ , so the adjoint of a matrix is its transpose. Note that the first dot product is an inner product for  $\mathbb{R}^m$ , whereas the second is for  $\mathbb{R}^n$ . Similar analysis holds for linear operators  $L : \mathbb{R}^n \rightarrow \mathbb{R}^m$  since they can always be represented in terms of a unique  $n \times m$  matrix.

The difficulty is that differential operators are unbounded and the specification of the domain is critical. The following definition classifies adjoints for unbounded linear operators.

**Definition A.12.** Let  $H$  be a Hilbert space with inner product  $\langle \cdot, \cdot \rangle$ , and let  $L : \text{dom}(L) \rightarrow H$  be an unbounded linear operator with  $\text{dom}(L)$  dense in  $H$ . We define  $\text{dom}(L^*)$  to be the set of elements  $v$  such that there exists  $w$  so that  $\langle Lu, v \rangle = \langle u, w \rangle$

for all  $u \in \text{dom}(L)$ . The adjoint operator  $L^*$  with domain  $\text{dom}(L^*)$  is that which satisfies

$$\langle Lu, v \rangle = \langle u, L^*v \rangle \quad (\text{A.3})$$

for all  $u \in \text{dom}(L)$  and  $v \in \text{dom}(L^*)$ . A densely defined operator  $L : H \rightarrow H$  is self-adjoint if  $L = L^*$  and  $\text{dom}(L^*) = \text{dom}(L)$ .

As will be illustrated in examples, the adjoint of a differential operator necessarily requires specification of initial or boundary conditions. This is in contrast to the *formal adjoint*, which simply describes the coefficients of the differential operator.

To motivate the approach used to construct adjoints for differential operators, we take  $H = L^2(\Omega)$  so that

$$\langle u, v \rangle = \int_{\Omega} u(x)v(x)dx.$$

The bilinear identity (A.3) can then be interpreted as a generalized Green's identity or successive integration by parts with adjoint boundary conditions chosen to ensure that boundary terms vanish.

Since differential operators  $L$  are not defined for all  $L^2$  functions, the domains  $\text{dom}(L)$  must be chosen as subsets of sufficiently smooth functions that satisfy initial or boundary conditions. Due to their completeness and approximation properties, subsets of the Sobolev spaces  $H^k(\Omega) = W^{k,2}(\Omega)$  provide a natural setting for analysis since they are continuously and densely embedded in  $L^2(\Omega)$  and have trace properties that facilitate boundary analysis. Readers are referred to [6, 78, 265] for details regarding Sobolev spaces in the context of differential equations.

**Example A.13.** To illustrate the construction of the adjoint operator and boundary conditions for a boundary value problem, consider

$$\begin{aligned} Lu &\equiv u''(x) = f(x), \quad 0 < x < 1, \\ B_1 u &\equiv u'(0) - u(1) = 0, \\ B_2 u &\equiv u'(1) = 0. \end{aligned}$$

The linear differential operator  $L$  is thus  $L = \frac{d^2}{dx^2}$ , and  $B_1$  and  $B_2$  are boundary functionals that map sufficiently smooth functions  $u$  to numbers  $B_1u$  and  $B_2u$ . Successive integration by parts yields

$$\int_0^1 u''v dx = \int_0^1 uv'' dx + P(u, v)|_0^1,$$

where the bilinear form  $P$ , sometimes referred to as the conjunct of  $u$  and  $v$ , is given by

$$P(u, v)|_0^1 = u'(1)v(1) - u'(0)v(0) - u(1)v'(1) + u(0)v'(0).$$

The choice  $L^* = \frac{d^2}{dx^2}$  as the *formal adjoint* yields

$$\langle Lu, v \rangle - \langle u, L^*v \rangle = P(u, v)|_0^1.$$

To achieve the bilinear identity (A.3), adjoint boundary conditions are chosen to ensure that  $P(u, v)|_0^1 = 0$ . Since  $u'(0) = u(1)$  and  $u'(1) = 0$ ,

$$P(u, v)|_0^1 = -u(1)[v(0) + v'(1)] + u(0)v'(0),$$

where  $u(0)$  and  $u(1)$  are arbitrary. One choice for the adjoint boundary conditions is

$$B_1^*v = v'(0) = 0, \quad B_2^*v = v(0) + v'(1) = 0.$$

We note that this choice is not unique, so we typically choose the minimal conditions required to eliminate boundary terms.

Although  $L = L^*$ , the boundary conditions differ so that  $\text{dom}(L^*) \neq \text{dom}(L)$ , and hence the operator is not self-adjoint. This can be quantified by noting that appropriate choices for the domains are

$$\begin{aligned} \text{dom}(L) &= \{u \in H^2(0, 1) \mid u'(0) - u(1) = 0, u'(1) = 0\}, \\ \text{dom}(L^*) &= \{v \in H^2(0, 1) \mid v'(0) = 0, v(0) + v'(1) = 0\}. \end{aligned}$$

#### Example A.14.

$$\begin{aligned} u''(x) &= f(x), \quad 0 < x < 1, \\ u(0) &= u'(0) = 0. \end{aligned}$$

Since

$$\int_0^1 (u''v - uv'')dx = u'(1)v(1) - u(1)v'(1),$$

appropriate adjoint conditions are  $v(1) = v'(1) = 0$ . For initial value problems, the adjoint equations will always have final time conditions, which is known to readers familiar with the adjoint or co-state relations arising in optimal control. This is further illustrated in Section 14.2.2.

To motivate the construction of adjoints for partial differential operators, we consider the Laplacian  $\Delta$  in  $\mathbb{R}^n$  which satisfies Green's second identity

$$\begin{aligned} \int_{\Omega} (v\Delta u - u\Delta v)dx &= \int_{\partial\Omega} n \cdot (v\nabla u - u\nabla v)dS \\ &= \int_{\partial\Omega} \left( v\frac{\partial u}{\partial n} - u\frac{\partial v}{\partial n} \right) dS. \end{aligned}$$

Hence the Laplacian is formally self-adjoint, and adjoint conditions are chosen to ensure that boundary terms vanish. For an arbitrary linear differential operator  $L$ , the corresponding relation is

$$\int_{\Omega} (vLu - uL^*v)dx = \int_{\partial\Omega} n \cdot P(u, v)dS, \tag{A.4}$$

where  $P(u, v)$  is dependent on the form of the differential operator.

**Example A.15.** Consider the diffusion operator

$$L = \frac{\partial}{\partial t} - \left( \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} + \frac{\partial^2}{\partial x_3^2} \right)$$

for  $x \in \Omega_x \subset \mathbb{R}^3$  and  $t \in [0, t_f]$ . The space-time domain is  $\Omega = \Omega_x \times [0, t_f]$  with boundary  $\partial\Omega$ . Green's formula yields

$$\int_{\Omega} (vLu - uL^*v) dxdt = \int_{\partial\Omega} n \cdot (e_t uv + u \nabla_x v - v \nabla_x u) dS, \quad (\text{A.5})$$

where the formal adjoint operator is

$$L^* = -\frac{\partial}{\partial t} - \left( \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} + \frac{\partial^2}{\partial x_3^2} \right),$$

$e_t$  is a unit vector in time, and  $\nabla_x$  indicates the gradient with respect to spatial variables. The bilinear form  $P$  is thus

$$P(u, v) = e_t uv + u \nabla_x v - v \nabla_x u.$$

As detailed in [235], the fact that  $\Omega$  is a cylinder in space-time can be exploited to simplify (A.5) to

$$\int_0^{t_f} dt \int_{\Omega_x} (vLu - uL^*v) dx = \int_{\Omega_x} uv|_0^{t_f} dx + \int_0^{t_f} dt \int_{\partial\Omega_x} \left( u \frac{\partial v}{\partial n_x} - v \frac{\partial u}{\partial n_x} \right) dS_x.$$

## A.1 Exercises

**Exercise A.1.** Consider the differential equation

$$(a_2(x)u')' + a_0(x)u = f(x)$$

so that  $L = \frac{d}{dx}(a_2 \frac{d}{dx}) + a_0 = D(a_2 D) + a_0$ , where  $D \equiv \frac{d}{dx}$  in  $\mathbb{R}^1$ . Take the boundary functionals to be

$$B_1 u = \alpha_{11} u(a) + \alpha_{12} u'(a), \quad B_2 u = \beta_{21} u(b) + \beta_{22} u'(b).$$

Show that  $L$  is self-adjoint.

**Exercise A.2.** Consider the fourth-order operator

$$L = D^2(a_2 D^2) + D(a_1 D) + a_0,$$

where  $a_0(x)$ ,  $a_1(x)$ , and  $a_2(x)$  are arbitrary functions, and define the boundary operators to be

$$B_1 u = u(a), \quad B_2 u = u''(a), \quad B_3 u = u(b), \quad B_4 u = u''(b).$$

Note that these are typically termed pinned boundary conditions. Determine the adjoint operator  $L^*$  and adjoint boundary conditions. Is  $L$  self-adjoint?



# Bibliography

- [1] H.S. Abdel-Khalik, Y. Bang, and C. Wang, “Overview of hybrid subspace methods for uncertainty quantification, sensitivity analysis,” *Annals of Nuclear Engineering*, 52, pp. 28–46, 2013.
- [2] B.M. Adams, H.T. Banks, M. Davidian, H-D. Kwon, H.T. Tran, S.N. Wynne, and E.S. Rosenberg, “HIV dynamics: Modeling, data analysis, and optimal treatment protocols,” *Journal of Computational and Applied Mathematics*, 184, pp. 10–49, 2005.
- [3] B.M. Adams, H.T. Banks, M. Davidian, and E.S. Rosenberg, “Estimation and prediction with HIV treatment interruption data,” *Bulletin of Mathematical Biology*, 69(2), pp. 563–584, 2007.
- [4] B.M. Adams, M.S. Ebeida, M.S. Eldred, J.D. Jakeman, L.P. Swiler, W.J. Bohnhoff, K.R. Dalbey, J.P. Eddy, K.T. Hu, D.M. Vigil, L.E. Bauman, and P.D. Hough, “Dakota, a multilevel parallel object-oriented framework for design optimization, parameter estimation, uncertainty quantification, and sensitivity analysis,” Version 5.3 Theory Manual, Sandia Technical Report SAND2011-9106; <http://dakota.sandia.gov/publications.html>.
- [5] B.M. Adams, M.S. Ebeida, M.S. Eldred, J.D. Jakeman, L.P. Swiler, W.J. Bohnhoff, K.R. Dalbey, J.P. Eddy, K.T. Hu, D.M. Vigil, L.E. Bauman, and P.D. Hough, “Dakota, a multilevel parallel object-oriented framework for design optimization, parameter estimation, uncertainty quantification, and sensitivity analysis,” Version 5.3 User’s Manual, Sandia Technical Report SAND2011-9106; <http://dakota.sandia.gov/publications.html>.
- [6] R.A. Adams and J.J.F. Fournier, *Sobolev Spaces*, Second Edition, Elsevier, Oxford, UK, 2003.
- [7] J. Albert and J. Bennett, *Curve Ball: Baseball, Statistics and the Role of Chance in the Game*, Springer-Verlag, Copernicus Books, New York, 2003.
- [8] Ö.F. Alış and H. Rabitz, “Efficient implementation of high dimensional model representations,” *Journal of Mathematical Chemistry*, 29(2), pp. 127–142, 2001.

- [9] D. Allaire and K. Willcox, “Surrogate modeling for uncertainty assessment with application to aviation environmental system models,” *AIAA Journal*, 48(8), pp. 1791–1803, 2010.
- [10] R.M. Anderson and R.M. May, *Infectious Diseases of Humans*, Oxford University Press, Oxford, UK, 1991.
- [11] C. Andrieu and J. Thoms, “A tutorial on adaptive MCMC,” *Statistics and Computing*, 18, pp. 343–373, 2008.
- [12] W. Apley, J. Liu, and W. Chen, “Understanding the effects of model uncertainty in robust design with computer experiments,” *Journal of Mechanical Design*, 128, pp. 945–958, 2006.
- [13] G.B. Arhonditsis, D. Papantou, W. Zhang, G. Perhar, E. Massos, and M. Shi, “Bayesian calibration of mechanistic aquatic biogeochemical models and benefits for environmental management,” *Journal of Marine Systems*, 73, pp. 8–30, 2008.
- [14] O. Arino, D. Axelrod, and M. Kimmel, Eds., *Advances in Mathematical Population Dynamics – Molecules, Cells and Man*, Series in Mathematical Biology and Medicine, Vol. 6, World Scientific, Singapore, 1997.
- [15] R.C. Aster, B. Borchers, and C.H. Thurber, *Parameter Estimation and Inverse Problems*, Second Edition, Academic Press, Elsevier, Amsterdam, 2013.
- [16] K.J. Åström and P. Eykhoff, “System identification – A survey,” *Automatica*, 7, pp. 123–162, 1971.
- [17] J.A. Atwell and B.B. King, “Reduced order controllers for spatially distributed systems via proper orthogonal decomposition,” *SIAM Journal on Scientific Computing*, 26(1), pp. 128–151, 2004.
- [18] I. Babuška, F. Nobile, and R. Tempone, “A stochastic collocation method for elliptic partial differential equations with random input data,” *SIAM Review*, 52(2), pp. 317–355, 2010.
- [19] I. Babuška, R. Tempone, and G.E. Zouraris, “Galerkin finite element approximations of stochastic elliptic partial differential equations,” *SIAM Journal on Numerical Analysis*, 42(2), pp. 800–825, 2004.
- [20] S. Balakrishnan, A. Roy, M.G. Ierapetritou, G.P. Flach, and P.G. Georgopoulos, “Uncertainty reduction and characterization for complex environmental fate and transport models: An empirical Bayesian framework incorporating the stochastic response surface method,” *Water Resources Research*, 39(12), pp. 1350–1362, 2003.
- [21] Y. Bang, H.S. Abdel-Khalik, and J.M. Hite, “Hybrid reduced order modeling applied to nonlinear models,” *International Journal for Numerical Methods in Engineering*, 91, pp. 929–949, 2012.

- [22] H.T. Banks, *A Functional Analysis Framework for Modeling, Estimation and Control in Science and Engineering*, CRC Press, Taylor and Francis Group, Boca Raton, FL, 2012.
- [23] H.T. Banks, A. Cintrón-Arias, and F. Kappel, “Parameter selection methods in inverse problem formulation,” in *Mathematical Model Development and Validation in Physiology: Application to the Cardiovascular and Respiratory Systems*, Lecture Notes in Mathematics, Mathematical Biosciences Subseries, Springer-Verlag, 2012, to appear.
- [24] H.T. Banks, M. Davidian, J.R. Samuels, Jr., and K.L. Sutton, “An inverse problem statistical methodology summary,” in *Statistical Estimation Approaches in Epidemiology*, G. Chowell, M. Hyman, N. Hengartner, L.M.A. Bettencourt, and C. Castillo-Chavez, Eds., Springer, Berlin, pp. 249–302, 2009.
- [25] H.T. Banks and K. Kunisch, *Estimation Techniques for Distributed Parameter Systems*, Birkhäuser, Boston, 1989.
- [26] H.T. Banks and H.T. Tran, *Mathematical and Experimental Modeling of Physical and Biological Processes*, Chapman and Hall/CRC Press, Boca Raton, FL, 2009.
- [27] J.M. Bardsley, “MCMC-based image reconstruction with uncertainty quantification,” *SIAM Journal on Scientific Computing*, 34(3), pp. A1316–A1332, 2012.
- [28] R.G. Barry and R.J. Chorley, *Atmosphere, Weather and Climate*, Eighth Edition, Routledge, Taylor and Francis, London, UK, 2003.
- [29] V. Barthelmann, E. Novak, and K. Ritter, “High dimensional polynomial interpolation on sparse grids,” *Advances in Computational Mathematics*, 12, pp. 273–388, 2000.
- [30] M.J. Bayarri and J.O. Berger, “The interplay of Bayesian and frequentist analysis,” *Statistical Science*, 19(1), pp. 58–80, 2004.
- [31] M.J. Bayarri, J.O. Berger, R. Paulo, J. Sacks, J.A. Cafeo, J. Cavendish, C-H. Lin, and J. Tu, “A framework for validation of computer models,” *Technometrics*, 49(2), pp. 138–154, 2007.
- [32] T. Bedford and R. Cooke, *Probabilistic Risk Analysis: Foundations and Methods*, Cambridge University Press, Cambridge, UK, 2003.
- [33] R. Bellman and K.J. Åström, “On structural identifiability,” *Mathematical Biosciences*, 7(3-4), pp. 329–339, 1970.
- [34] J.O. Berger, *Statistical Decision Theory and Bayesian Analysis*, Second Edition, Springer-Verlag, New York, 1985.

- [35] J.O. Berger, “The case for objective Bayesian analysis,” *Bayesian Analysis*, 1(3), pp. 385–402, 2006.
- [36] P. Billingsley, *Probability and Measure*, Third Edition, John Wiley and Sons, New York, 1995.
- [37] A.W. Bowman and A. Azzalini, *Applied Smoothing Techniques for Data Analysis*, Oxford University Press, New York, 1997.
- [38] G.E.P. Box and N.R. Draper, *Empirical Model-Building and Response Surfaces*, John Wiley and Sons, New York, 1987.
- [39] G.E.P. Box and N.R. Draper, *Response Surfaces, Mixtures, and Ridge Analysis*, Second Edition, John Wiley and Sons, Hoboken, NJ, 2007.
- [40] F. Brauer and C. Castillo-Chavez, *Mathematical Models for Communicable Diseases*, SIAM, Philadelphia, 2013.
- [41] R. Brookmeyer, “Measuring the HIV/AIDS epidemic: Approaches and challenges,” *Epidemiologic Reviews*, 32, pp. 26–37, 2010.
- [42] S.P. Brooks and G.O. Roberts, “Convergence assessment techniques for Markov chain Monte Carlo,” *Statistics and Computing*, 8(4), pp. 319–335, 1998.
- [43] J. Brynjarsdóttir and A. O’Hagan, “Learning about physical parameters: The importance of model discrepancy,” *SIAM/ASA Journal on Uncertainty Quantification*, submitted.
- [44] R.C. Buchanan and R. Buddemeier, *Kansas Ground Water*, Kansas Geological Survey, Educational Series 10, 1993; available at <http://www.kgs.ku.edu/Publications/Bulletins/ED10/index.html>.
- [45] T. Bui-Thanh, O. Ghattas, and D. Higdon, “Adaptive Hessian-based non-stationary Gaussian process response surface method for probability density approximation with application to Bayesian solution of large-scale inverse problems,” *SIAM Journal on Scientific Computing*, 34(6), pp. A2837–A2871, 2012.
- [46] T. Bui-Thanh, K. Willcox, and O. Ghattas, “Model reduction for large-scale systems with high-dimensional parametric input space,” *SIAM Journal on Scientific Computing*, 30(6), pp. 3270–3288, 2008.
- [47] H-J. Bungartz and M. Griebel, “Sparse grids,” *Acta Numerica*, 13, pp. 147–269, 2004.
- [48] J. Burkardt, M. Gunzburger, and H.-C. Lee, “Centroidal Voronoi tessellation-based reduced-order modeling of complex systems,” *SIAM Journal on Scientific Computing*, 28(2), pp. 459–484, 2006.

- [49] W.J. Burroughs, *Climate Change: A Multidisciplinary Approach*, Second Edition, Cambridge University Press, New York, 2007.
- [50] D.G. Cacuci, *Sensitivity and Uncertainty Analysis: Theory*, Chapman and Hall/CRC, Boca Raton, FL, 2003.
- [51] D.G. Cacuci, Ed., *Handbook of Nuclear Engineering*, Springer-Verlag, New York, 2010.
- [52] D.G. Cacuci and M. M. Ionescu-Bujor, “Best-estimate model calibration and prediction through experimental data assimilation — I: Mathematical framework,” *Nuclear Science and Engineering*, 165, pp. 18–44, 2010.
- [53] D.G. Cacuci, M. Ionescu-Bujor, and I.M. Navon, *Sensitivity and Uncertainty Analysis: Application to Large-Scale Systems*, Chapman and Hall/CRC, Boca Raton, FL, 2005.
- [54] D.G. Cacuci, I.M. Navon, and M. Ionescu-Bujor, *Computational Methods for Data Evaluation and Assimilation*, Chapman and Hall/CRC, Boca Raton, FL, 2013.
- [55] D.S. Callaway and A.S. Perelson, “HIV-1 infection and low steady state viral loads,” *Bulletin of Mathematical Biology*, 64, pp. 29–64, 2001.
- [56] D. Calvetti and E. Somersalo, *Introduction to Bayesian Scientific Computing: Ten Lectures on Subjective Computing*, Springer, New York, 2007.
- [57] R.H. Cameron and W.T. Martin, “The orthogonal development of nonlinear functionals in series of Fourier-Hermite functionals,” *Annals of Mathematics*, 48(2), pp. 385–392, 1947.
- [58] K. Campbell, M.D. McKay, and B.J. Williams, “Sensitivity analysis when model outputs are functions,” *Reliability Engineering and System Safety*, 91, pp. 1468–1472, 2006.
- [59] F. Campolongo and R. Braddock, “The use of graph theory in the sensitivity analysis of the model output: A second order screening method,” *Reliability Engineering and System Safety*, 64, pp. 1–12, 1999.
- [60] F. Campolongo, J. Cariboni, and A. Saltelli, “An effective screening design for sensitivity analysis of large models,” *Environmental Modelling and Software*, 22, pp. 1509–1518, 2007.
- [61] C. Canuto, M.Y. Hussaini, A. Quarteroni, and T.A. Zhang, *Spectral Methods in Fluid Dynamics*, Springer-Verlag, Berlin, 1988.
- [62] G. Casella and R.L. Berger, *Statistical Inference*, Duxbury Press, Belmont, CA, 1990.
- [63] S. Chandrasekaran and I.C.F. Ipsen, “On rank-revealing factorisations,” *SIAM Journal on Matrix Analysis and Applications*, 15(2), pp. 592–622, 1994.

- [64] E.K.P. Chong and S.H. Źak, *An Introduction to Optimization*, John Wiley and Sons, New York, 1996.
- [65] R. Confalonieri, G. Bellocchi, S. Bregaglio, M. Donatelli, and M. Acutis, “Comparison of sensitivity analysis techniques: A case study with the rice model WARM,” *Ecological Modelling*, 221, pp. 1897–1906, 2010.
- [66] P.G. Constantine, E. Dow, and Q. Wang, “Active subspace methods in theory and practice: Applications to kriging surfaces,” *SIAM Journal on Scientific Computing*, 36, pp. A1500–A1524, 2014.
- [67] N.A.C. Cressie, *Statistics for Spatial Data*, Revised Edition, John Wiley and Sons, New York, 1993.
- [68] T. Crestaux, O. Le Maître, and J.-M. Martinez, “Polynomial chaos expansion for sensitivity analysis, *Reliability Engineering and System Safety*, 94, pp. 1161–1172, 2009.
- [69] R.A. Cropp and R.D. Braddock, “The new Morris method: An efficient second-order screening method,” *Reliability Engineering and System Safety*, 78, pp. 77–83, 2002.
- [70] G. Dahlquist and Å. Björck, *Numerical Methods*, Translated by N. Anderson, Prentice-Hall, Englewood Cliffs, NJ, 1974.
- [71] DAKOTA (Design Analysis Kit for Optimization and Terascale Applications); <http://dakota.sandia.gov/index.html>.
- [72] P.J. Davis and P. Rabinowitz, *Numerical Integration*, Blaisdell, Waltham, MA, 1967.
- [73] G.E. Delury, Ed., *The World Almanac and Book of Facts*, Newspaper Enterprise Association of United Media, New York, 1995.
- [74] A. Der Kiureghian and P.-L. Liu, “Structural reliability under incomplete information,” *Journal of Engineering Mechanics*, 112(1), pp. 85–104, 1986.
- [75] A. Doucet, J.F.G. de Freitas, and N.J. Gordon, “An introduction to sequential Monte Carlo methods,” in *Sequential Monte Carlo Methods in Practice*, A. Doucet, J.F.G. de Freitas, and N.J. Gordon, Eds., Springer-Verlag, New York, pp. 3–14, 2001.
- [76] J.J. Duderstadt and L.J. Hamilton, *Nuclear Reactor Analysis*, John Wiley and Sons, New York, 1976.
- [77] M.S. Eldred and D.M. Dunlavy, “Formulations for surrogate-based optimization with data fit, multifidelity, and reduced-order models,” AIAA Paper 2006-7117, in Proceedings of the 11th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, Portsmouth, VA, 2006.

- [78] L.C. Evans, *Partial Differential Equations*, American Mathematical Society, Providence RI, 1998.
- [79] G. Evensen, *Data Assimilation: The Ensemble Kalman Filter*, Second Edition, Springer-Verlag, Berlin, 2009.
- [80] J. Fan and J. Lv, “A selective overview of variable selection in high dimensional feature space,” *Statistica Sinica*, 20(1), pp. 101–148, 2010.
- [81] W. Feller, *An Introduction to Probability and Its Applications, Volume I*, Wiley, New York, 1968.
- [82] W. Feller, *An Introduction to Probability and Its Applications, Volume II*, Wiley, New York, 1971.
- [83] A.L. Flint, L.E. Flint, G.S. Bodvarsson, E.M. Kwicklis, and J. Fabryka-Martin, “Evolution of the conceptual model of unsaturated zone hydrology at Yucca Mountain, Nevada,” *Journal of Hydrology*, 247(1-2), pp. 1–30, 2001.
- [84] A.I.J. Forrester, A. Sóbester, and A.J. Keane, “Multi-fidelity optimization via surrogate modelling,” *Proceedings of the Royal Society A*, 463, pp. 3251–3269, 2007.
- [85] A.I.J. Forrester, A. Sóbester, and A.J. Keane, *Engineering Design via Surrogate Modelling: A Practical Guide*, Progress in Astronautics and Aeronautics, Vol. 226, John Wiley and Sons, Chichester, UK, 2008.
- [86] L.A. Frakes, *Climates throughout Geologic Time*, Elsevier, Amsterdam, 1979.
- [87] M. Frangos, Y. Marzouk, K. Willcox, and B. van Bloemen Waanders, “Surrogate and reduced-order modeling: A comparison of approaches for large-scale statistical inverse problems,” in *Large-Scale Inverse Problems and Quantification of Uncertainty*, L. Biegler et al., Eds., John Wiley and Sons, Chichester, UK, pp. 123–150, 2011.
- [88] D. Galbally, K. Fidkowski, K. Willcox, and O. Ghattas, “Non-linear model reduction for uncertainty quantification in large-scale inverse problems,” *International Journal for Numerical Methods in Engineering*, 81, pp. 1581–1608, 2010.
- [89] S.E. Gano, J.E. Renaud, J.D. Martin, and T.W. Simpson, “Update strategies for kriging models used in variable fidelity optimization,” *Structural and Multidisciplinary Optimization*, 32, pp. 287–298, 2006.
- [90] Z. Gao and J.S. Hesthaven, “On ANOVA expansions and strategies for choosing the anchor point,” *Applied Mathematics and Computation*, 217, pp. 3274–3285, 2010.
- [91] T.C. Gard, *Introduction to Stochastic Differential Equations*, Marcel Dekker, New York, 1988.

- [92] A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin, *Bayesian Data Analysis*, Second Edition, Chapman and Hall/CRC, Boca Raton, FL, 2004.
- [93] T. Gerstner and M. Griebel, “Numerical integration using sparse grids,” *Numerical Algorithms*, 18, pp. 209–232, 1998.
- [94] R.G. Ghanem and P.D. Spanos, *Stochastic Finite Elements: A Spectral Approach*, Revised Edition, Dover, Mineola, NY, 2003.
- [95] M.S. Gockenbach, *Understanding and Implementing the Finite Element Method*, SIAM, Philadelphia, 2006.
- [96] M.A. Golberg and H.A. Cho, *Introduction to Regression Analysis*, WIT Press, Southampton, UK, 2004.
- [97] G.H. Golub and C.F. Van Loan, *Matrix Computations*, Second Edition, The Johns Hopkins University Press, Baltimore, MD, 1989.
- [98] D. Gottlieb and S.A. Orszag, *Numerical Analysis of Spectral Methods: Theory and Applications*, SIAM, Philadelphia, 1977.
- [99] G.R. Grimmett and D.R. Stirzaker, *Probability and Random Processes*, Second Edition, Oxford University Press, Oxford, UK, 1992.
- [100] M. Gu and S.C. Eisenstat, “Efficient algorithms for computing strong rank-revealing QR factorization,” *SIAM Journal on Scientific Computing*, 17(4), pp. 848–869, 1996.
- [101] M. Gunzburger, C.G. Webster, and G. Zhang, “An adaptive wavelet stochastic collocation method for irregular solutions of stochastic partial differential equations,” ORNL Report ORNL/TM-2012/186.
- [102] H. Haario, M. Laine, A. Mira, and E. Saksman, “DRAM: Efficient adaptive MCMC,” *Statistics and Computing*, 16(4), pp. 339–354, 2006.
- [103] H. Haario, E. Saksman, and J. Tamminen, “An adaptive Metropolis algorithm,” *Bernoulli*, 7(2), pp. 223–242, 2001.
- [104] G. Hald, *Statistical Theory with Engineering Applications*, John Wiley and Sons, New York, 1952.
- [105] N. Halko, P.G. Martinsson, and J.A. Tropp, “Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions,” *SIAM Review*, 53(2), pp. 217–288, 2011.
- [106] J. Hansen et al., “A Pinatubo climate modeling investigation,” in *The Mount Pinatubo Eruption: Effects on the Atmosphere and Climate*, G. Fiocco, D. Fua, and G. Visconti, Eds., NATO ASI Series, Vol. I42, Springer-Verlag, Heidelberg, pp. 233–272, 1996.

- [107] E.J. Haug, K.K. Choi, and V. Komkov, *Design Sensitivity Analysis of Structural Systems*, Academic Press, Orlando, FL, 1986.
- [108] D. Higdon, J. Gattiker, B. Williams, and M. Rightley, “Computer model calibration using high-dimensional output,” *Journal of the American Statistical Association*, 103(no. 482), pp. 570–583, 2008.
- [109] M.C. Hill, D. Kavetski, M. Clark, M. Ye, and D. Lu, “Uncertainty quantification for environmental models,” *SIAM News*, 45(9), Philadelphia, 2012.
- [110] M.C. Hill and C.R. Tiedman, *Effective Groundwater Model Calibration: With Analysis of Data, Sensitivities, Predictions, and Uncertainty*, John Wiley and Sons, Hoboken, NJ, 2007.
- [111] R.G. Hills and T.G. Trucano, *Statistical Validation of Engineering and Scientific Models: Background*, Sandia Report SAND99-1256, 1999.
- [112] R.V. Hogg and A.T. Craig, *Introduction to Mathematical Statistics*, Fourth Edition, Macmillan, New York, 1978.
- [113] J.R. Holton, *An Introduction to Dynamic Meteorology*, Third Edition, Academic Press, San Diego, CA, 1992.
- [114] F.C. Hoppensteadt and C.S. Peskin, *Modeling and Simulation in Medicine and the Life Sciences*, Second Edition, Springer, New York, 2002.
- [115] J.T. Houghton, Y. Ding, D.J. Griggs, M. Noguer, P.J. van der Linden, X. Dai, K. Maskell, and C.A. Johnson, Eds., *Climate Change 2001: The Scientific Basis*, Cambridge University Press, Cambridge, UK, 2001.
- [116] Z. Hu, R.C. Smith, N. Burch, M. Hays, and W.S. Oates, “A modeling and uncertainty quantification framework for a flexible structure with macro-fiber composite (MFC) actuators operating in hysteretic regimes,” *Journal of Intelligent Material Systems and Structures*, to appear.
- [117] C. Huber et al., “Isotope calibrated Greenland temperature record over Marine Isotope Stage 3 and its relation to CH<sub>4</sub>,” *Earth and Planetary Science Letters*, 243, pp. 504–519, 2006.
- [118] T.J.R. Hughes, *The Finite Element Method: Linear Static and Dynamic Finite Element Analysis*, Dover, Mineola, NY, 2000.
- [119] I.C.F. Ipsen, *Numerical Matrix Analysis: Linear Systems and Least Squares*, SIAM, Philadelphia, 2009.
- [120] I.C.F. Ipsen, C.T. Kelley, and S.R. Pope, “Rank-deficient nonlinear least squares problems and subset selection,” *SIAM Journal on Numerical Analysis*, 49(3), pp. 1244–1266, 2011.
- [121] D.L. Isaacson and R.W. Madsen, *Markov Chains: Theory and Applications*, John Wiley and Sons, New York, 1976.

- [122] M. Ishii and T. Hibiki, *Thermo-Fluid Dynamics of Two-Phase Flow*, Second Edition, Springer Science and Business Media, New York, 2006.
- [123] M.Z. Jacobson, *Fundamentals of Atmospheric Modeling*, Cambridge University Press, Cambridge, UK, 1999.
- [124] H. Järvinen, M. Laine, A. Solonen, and H. Haario, “Ensemble prediction and parameter estimation system: The concept,” *Quarterly Journal of the Royal Meteorological Society*, 138, pp. 281–288, 2012 Part B.
- [125] H. Järvinen, P. Räisänen, M. Laine, J. Tamminen, A. Ilin, E. Oja, A. Solonen, and H. Haario, “Estimation of ECHAM5 climate model closure parameters with adaptive MCMC,” *Atmospheric Chemistry and Physics*, 10(20), pp. 9993–10002, 2010.
- [126] P.W. Jones and P. Smith, *Stochastic Processes: An Introduction*, Second Edition, CRC Press, Boca Raton, FL, 2010.
- [127] J. Jouzel et al., “Orbital and millennial Antarctic climate variability over the past 800,000 years,” *Science*, 317, pp. 793–796, 2007.
- [128] J. Kaipio and E. Somersalo, *Statistical and Computational Inverse Problems*, Springer, New York, 2005.
- [129] E. Kalnay, *Atmospheric Modeling, Data Assimilation and Predictability*, Cambridge University Press, Cambridge, UK, 2003.
- [130] S. Karlin, *A First Course in Stochastic Processes*, Academic Press, New York, 1968.
- [131] C.T. Kelley, *Iterative Methods for Optimization*, SIAM, Philadelphia, 1999.
- [132] C.T. Kelley, *Implicit Filtering*, SIAM, Philadelphia, 2011.
- [133] M.C. Kennedy and A. O’Hagan, “Predicting the output from a complex computer code when fast approximations are available,” *Biometrika*, 87, pp. 1–13, 2000.
- [134] M.C. Kennedy and A. O’Hagan, “Bayesian calibration of computer models,” *Journal of the Royal Statistical Society. Series B*, 63(3), pp. 425–464, 2001.
- [135] S.P. Kenny, L.G. Crespo, and D.P. Giesy, *UQTools: The Uncertainty Quantification Toolbox – Introduction and Tutorial*, NASA Technical Report NASA/TM-2012-217561.
- [136] J.T. Kiehl and V. Ramanathan, Eds., *Frontiers of Climate Modeling*, Cambridge University Press, Cambridge, UK, 2006.
- [137] J.T. Kiehl and K.E. Trenberth, “Earth’s annual global mean energy budget,” *Bulletin of the American Meteorological Society*, 78(2), pp. 197–208, 1997.

- [138] P.E. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations*, Springer-Verlag, Berlin, 1992.
- [139] E. Kreyszig, *Introductory Functional Analysis with Applications*, John Wiley and Sons, New York, 1978.
- [140] K. Kunisch and S. Volkwein, “Control of Burgers’ equation by reduced order approach using proper orthogonal decomposition,” *Journal of Optimization Theory and Applications*, 102, pp. 345–371, 1999.
- [141] K. Kunisch and S. Volkwein, “Optimal snapshot location for computing POD basis functions,” *ESAIM: Mathematical Modelling and Numerical Analysis*, 44, pp. 509–529, 2010.
- [142] J.N. Kutz, *Data-Driven Modeling & Scientific Computation: Methods for Complex Systems & Big Data*, Oxford University Press, Oxford, UK, 2013.
- [143] M. Laine, A. Solonen, H. Haario, and H. Järvinen, “Ensemble prediction and parameter estimation system: The method,” *Quarterly Journal of the Royal Meteorological Society*, 138, pp. 289–297, 2012 Part B.
- [144] E. Laloy and J.A. Vrugt, “High-dimensional posterior exploration of hydrologic models using multiple-try DREAM<sub>(ZS)</sub> and high-performance computing,” *Water Resources Research*, 48(1), W01526, doi:10.1029/2011WR010608, 2012.
- [145] C. Lanczos, *Linear Differential Operators*, SIAM Classics in Applied Mathematics, Philadelphia, 1996.
- [146] S. Lang, *Real and Functional Analysis*, Springer-Verlag, New York, 1993.
- [147] W.K.M. Lau and D.E. Waliser, Eds., *Intraseasonal Variability in the Atmosphere-Ocean Climate System*, Second Edition, Springer-Verlag, Berlin, 2012.
- [148] O.P. Le Maître and O.M. Knio, *Spectral Methods for Uncertainty Quantification*, Springer, Dordrecht, 2010.
- [149] M. Leroux, *Global Warming — Myth or Reality? The Erring Ways of Climatology*, Spring and Praxis, Chichester, UK, 2005.
- [150] M. Leutbecher and T.N. Palmer, “Ensemble forecasting,” *Journal of Computational Physics*, 227, pp. 3515–3539, 2008.
- [151] E.E. Lewis and W.F. Miller, Jr., *Computational Methods of Neutron Transport*, American Nuclear Society, La Grange Park, IL, 1993.
- [152] G. Li, S.-W. Wang, and H. Rabitz, “Practical approaches to construct RS-HDMR component functions,” *The Journal of Physical Chemistry A*, 106, pp. 8721–8733, 2002.

- [153] G. Li, S.-W. Wang, H. Rabitz, S. Wang, and P. Jaffé, “Global uncertainty assessments by high dimensional model representations (HDMR), *Chemical Engineering Science*, 57, pp. 4445–4460, 2002.
- [154] C. Lieberman, K. Willcox, and O. Ghattas, “Parameter and state model reduction for large-scale statistical inverse problems,” *SIAM Journal on Scientific Computing*, 32, pp. 2523–2542, 2010.
- [155] L. Lilburne and S. Tarantola, “Sensitivity analysis of spatial models,” *International Journal of Geographical Information Science*, 23(2), pp. 151–168, 2009.
- [156] G. Lillacci and M. Khammash, “Parameter estimation and model selection in computational biology,” *PLoS Computational Biology*, 6(3):e1000696-1–17.DOI: 10.1371/journal.pcbi.1000696, 2010.
- [157] G. Lillacci and M. Khammash, “A distribution-matching method for parameter estimation and model selection in computational biology,” *International Journal of Robust and Nonlinear Control*, 22, pp. 1065–1081, 2012.
- [158] R.S. Lipster and A.N. Shiryaev, *Statistics of Random Processes. I. General Theory*, Second Edition, Translated by A.B. Aries, Springer, Berlin, 2001.
- [159] A.J. Lotka, *Elements of Physical Biology*, Williams and Wilkins, Baltimore, MD, 1925.
- [160] H.V. Ly and H.T. Tran, “Modeling and control of physical processes using proper orthogonal decomposition,” *Mathematical and Computer Modelling*, 33, pp. 223–236, 2001.
- [161] H.V. Ly and H.T. Tran, “Proper orthogonal decomposition for flow calculations and optimal control in a horizontal CVD reactor,” *Quarterly of Applied Mathematics*, 60(4), pp. 631–656, 2002.
- [162] X. Ma and N. Zabaras, “An adaptive high-dimensional stochastic model representation technique for the solution of stochastic partial differential equations,” *Journal of Computational Physics*, 229, pp. 3884–3915, 2010.
- [163] A.J. Majda, “Challenges in climate science and contemporary applied mathematics,” *Communications on Pure and Applied Mathematics*, 65(7), pp. 920–948, 2012.
- [164] M.E. Mann, R.S. Bradley, and M.K. Hughes, “Northern hemisphere temperatures during the past millennium: Inferences, uncertainties and limitations,” *Geophysical Research Letters*, 26, pp. 759–762, 1999.
- [165] S. Marino, I.B. Hogue, C.J. Ray, and D.E. Kirschner, “A methodology for performing global uncertainty and sensitivity analysis in systems biology,” *Journal of Theoretical Biology*, 254, pp. 178–196, 2008.

- [166] A. Marrel, B. Iooss, M. Jullien, B. Laurent, and E. Volkova, “Global sensitivity analysis for models with spatially dependent outputs,” *Environmetrics*, 22, pp. 383–397, 2011.
- [167] Y.M. Marzouk, H.N. Najm, and L.A. Rahn, “Stochastic spectral methods for efficient Bayesian solution of inverse problems,” *Journal of Computational Physics*, 224(2), pp. 560–586, 2007.
- [168] Y.M. Marzouk and D. Xiu, “A stochastic collocation approach to Bayesian inference in inverse problems,” *Communications in Computational Physics*, 6(4), pp. 826–847, 2009.
- [169] K. McGuffie and A. Henderson-Sellers, *A Climate Modelling Primer*, Second Edition, John Wiley and Sons, Chichester, UK, 1997.
- [170] M. McKay, W. Conover, and R. Beckman, “A comparison of three methods for selecting values of input variables in the analysis of output from a computer code,” *Technometrics*, 21, pp. 239–245, 1979.
- [171] W. Mendenhall, R.L. Scheaffer, and D.D. Wackerly, *Mathematical Statistics with Applications*, Second Edition, Duxbury Press, Boston, MA, 1981.
- [172] M.W. Moncrieff, M.V. Shapiro, J.M. Slingo, and F. Molteni, “Collaborative research at the intersection of weather and climate,” *World Meteorological Organization Bulletin*, 56(3), pp. 204–211, 2007.
- [173] J.M. Moran and M.D. Morgan, *Meteorology: The Atmosphere and Science of Weather*, Fifth Edition, Prentice-Hall, Upper Saddle River, NJ, 1997.
- [174] W. Morokoff and R. Caflisch, “Quasi-Monte Carlo integration,” *Journal of Computational Physics*, 122(2), pp. 218–230, 1995.
- [175] M.D. Morris, “Factorial sampling plans for preliminary computational experiments,” *Technometrics*, 33(2), pp. 161–174, 1991.
- [176] J.L. Mueller and S. Siltanen, *Linear and Nonlinear Inverse Problems with Practical Applications*, SIAM, Philadelphia, 2012.
- [177] D. Mukherjee, B.N. Rao, and A.M. Prasad, “Global sensitivity analysis of unreinforced masonry structure using high dimensional model representation,” *Engineering Structures*, 33, pp. 1316–1325, 2011.
- [178] J.M. Murphy, D.M.H. Sexton, D.N. Barnett, G.S. Jones, M.J. Webb, M. Collins, and D.A. Stainforth, “Quantification of modelling uncertainties in a large ensemble of climate change simulations,” *Letters to Nature*, 430, pp. 768–772, 2004.
- [179] J.D. Murray, *Mathematical Biology II: Spatial Models and Biomedical Applications*, Third Edition, Springer-Verlag, New York, 2003.

- [180] H.N. Najm, “Uncertainty quantification and polynomial chaos techniques in computational fluid dynamics,” *Annual Review of Fluid Mechanics*, 41, pp. 35–52, 2009.
- [181] National Research Council of the National Academies, *Mathematics and 21st Century Biology*, The National Academies Press, Washington, DC, 2005.
- [182] I.M. Navon, “Data assimilation for numerical weather prediction: A review,” in *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications*, S.K. Park and L. Xu, Eds., Springer-Verlag, Berlin, pp. 21–65, 2009.
- [183] F. Nobile, R. Tempone, and C.G. Webster, “A sparse grid stochastic collocation method for partial differential equations with random input data,” *SIAM Journal on Numerical Analysis*, 46(5), pp. 2309–2345, 2008.
- [184] J.R. Norris, *Markov Chains*, Cambridge University Press, Cambridge, UK, 1997.
- [185] E. Novak and K. Ritter, “High dimensional integration of smooth functions over cubes,” *Numerische Mathematik*, 75, pp. 79–97, 1996.
- [186] B. Øksendal, *Stochastic Differential Equations: An Introduction with Applications*, Sixth Edition, Springer, Berlin, 2007.
- [187] S.G. Osborn, A. Vengosh, N.R. Warner, and R.B. Jackson, “Methane contamination of drinking water accompanying gas-well drilling and hydraulic fracturing,” *Proceedings of the National Academy of Sciences*, 108(20), pp. 8172–8176, 2011.
- [188] R.K. Pachauri and A. Reisinger, Eds., Core Writing Team, *Climate Change 2007: Synthesis Report. Contributions of Working Groups I, II and III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, IPCC, Geneva, Switzerland, 2008.
- [189] T. Palmer and R. Hagedom, Eds., *Predictability of Weather and Climate*, Cambridge University Press, New York, 2006.
- [190] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.
- [191] A.S. Perelson and P.W. Nelson, Mathematical analysis of HIV-1 dynamics in vivo,” *SIAM Review*, 41, pp. 3–44, 1999.
- [192] K.K. Phoon, H.W. Huang, and S.T. Quek, “Simulation of strongly non-Gaussian processes using Karhunen–Loeve expansion,” *Probabilistic Engineering Mechanics*, 20, pp. 188–198, 2005.
- [193] R.A. Pielke, Sr., *Mesoscale Meteorological Modeling*, Second Edition, Academic Press, San Diego, CA, 2002.

- [194] G.F. Pinder and M.A. Celia, *Subsurface Hydrology*, John Wiley and Sons, Hoboken, NJ, 2006.
- [195] S. Pitchaiah and A. Armaou, “Output feedback control of distributed parameter systems using adaptive proper orthogonal decomposition,” *Industrial and Engineering Chemistry Research*, 49, pp. 10496–10509, 2010.
- [196] T.D. Potter and B.R. Colman, Eds., *Handbook of Weather, Climate and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements*, Wiley-Interscience, Hoboken, NJ, 2003.
- [197] W.C. Proctor, *Elements of High-Order Predictive Model Calibration Algorithms with Applications to Large-Scale Reactor Physics Systems*, Ph.D. Dissertation, North Carolina State University, Raleigh, NC, 2012.
- [198] P.Z.G. Qian and C.F.J. Wu, “Bayesian hierarchical modeling for integrating low-accuracy and high-accuracy experiments,” *Technometrics*, 50(2), pp. 192–204, 2008.
- [199] Z. Qian, C.C. Seepersad, V.R. Joseph, J.K. Allen, and C.F.J. Wu, “Building surrogate models based on detailed and approximate simulations,” *Journal of Mechanical Design*, 128(4), pp. 668–677, 2005.
- [200] H. Rabitz and Ö.F. Alış, “General foundations of high-dimensional model representations,” *Journal of Mathematical Chemistry*, 25, pp. 197–233, 1999.
- [201] H. Rabitz, Ö.F. Alış, J. Shorter, and K. Shim, “Efficient input-output model representations,” *Computer Physics Communications*, 117, pp. 11–20, 1999.
- [202] J.R. Raol, G. Girija, and J. Singh, *Modelling and Parameter Estimation in Dynamic Systems*, The Institution of Electrical Engineers, London, UK, 2004.
- [203] M. Rathinam and L.R. Petzold, “A new look at proper orthogonal decomposition,” *SIAM Journal on Numerical Analysis*, 41(5), pp. 1893–1925, 2003.
- [204] B.D. Reddy, *Introductory Functional Analysis with Applications to Boundary Value Problems and Finite Elements*, Springer-Verlag, New York, 1998.
- [205] P. Reichert and J. Mieleitner, “Analyzing input and structural uncertainty of nonlinear dynamic models with stochastic, time-dependent parameters,” *Water Resources Research*, 45, W10402, 2009.
- [206] *RELAP5-3D<sup>®</sup> Code Manual Volume I: Code Structure, System Models and Solution Methods*, Idaho National Laboratory, INEEL-EXT-98-00834, Revision 2.4, June 2005.
- [207] M. Renardy and R.C. Rogers, *An Introduction to Partial Differential Equations*, Springer-Verlag, New York, 1993.
- [208] G.O. Roberts and J.S. Rosenthal, “Examples of adaptive MCMC,” *Journal of Computational and Graphical Statistics*, 18(2), pp. 349–367, 2009.

- [209] E.H. Roseboom, Jr., “Disposal of high-level nuclear waste above the water table in arid regions,” Geological Survey Circular 903, Alexandria, VA, 1983.
- [210] M. Rosenblatt, “Remarks on a multivariate transformation,” *Annals of Mathematical Statistics*, 23(3), pp. 470–472, 1952.
- [211] T.M. Russi, *Uncertainty Quantification with Experimental Data and Complex System Models*, Ph.D. Dissertation, UC Berkeley, Berkeley, CA, 2010.
- [212] J. Sacks, W.J. Welch, T.J. Mitchell, and H.P. Wynn, “Design and analysis of computer experiments,” *Statistical Science*, 4, pp. 409–423, 1989.
- [213] M.L. Salby, *Physics of the Atmosphere and Climate*, Cambridge University Press, New York, 2012.
- [214] A. Saltelli, “Making best use of model evaluations to compute sensitivity indices,” *Computer Physics Communications*, 145, pp. 280–297, 2002.
- [215] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola, *Global Sensitivity Analysis: The Primer*, John Wiley and Sons, Chichester, UK, 2008.
- [216] A. Saltelli, M. Ratto, S. Tarantola, and F. Campolongo, “Sensitivity analysis practices: Strategies for model-based inference,” *Reliability Engineering and System Safety*, 91, pp. 1109–1125, 2006.
- [217] A. Saltelli, S. Tarantola, F. Campolongo, and M. Ratto, *Sensitivity Analysis in Practice*, John Wiley and Sons, Chichester, UK, 2004.
- [218] B. Sansó and C. Forest, “Statistical calibration of climate system properties,” *Journal of the Royal Statistical Society C*, 58, pp. 485–503, 2009.
- [219] G.A.F. Seber and C.J. Wild, *Nonlinear Regression*, John Wiley and Sons, Hoboken, NJ, 2003.
- [220] E. Seneta, *Non-negative Matrices*, Halsted Press, New York, 1973.
- [221] S.J. Sheather and M.C. Jones, “A reliable data-based bandwidth selection method for kernel density estimation,” *Journal of the Royal Statistical Society B*, 53(3), pp. 683–690, 1991.
- [222] *Simulation-Based Engineering Science: Revolutionizing Engineering Science through Simulation*, Report of the NSF Blue Ribbon Panel on Simulation-Based Engineering Science, February 2006.
- [223] G. Sin and K. Gernaey, “Improving the Morris method for sensitivity analysis by scaling the elementary effects,” Proceedings of the 19th European Symposium on Computer Aided Process Engineering – ESCAPE19, Eds. J. Jezowski and J. Thullie, Oxford, UK, pp. 925–930, 2009.

- [224] D.S. Sivia with J. Skilling, *Data Analysis: A Bayesian Tutorial*, Second Edition, Oxford University Press, Oxford, UK, 2006.
- [225] R.C. Smith, *Smart Material Systems: Model Development*, SIAM, Philadelphia, 2005.
- [226] S. Smolyak, “Quadrature and interpolation formulas for tensor products of certain classes of functions,” *Doklady Akademii Nauk SSSR*, 4, pp. 240–243, 1963.
- [227] I.M. Sobol’, “Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates,” *Mathematics and Computers in Simulation*, 55, pp. 271–280, 2001.
- [228] S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor, and H.L. Miller, Eds., *Climate Change 2007: Working Group I: The Physical Science Basis*, Cambridge University Press, Cambridge, UK, 2007.
- [229] A. Solonen, *Monte Carlo Methods in Parameter Estimation of Nonlinear Models*, MS Thesis, Lappeenranta University of Technology, Lappeenranta, Finland, 2006.
- [230] A. Solonen, P. Ollinaho, M. Laine, H. Haario, J. Tamminen, and H. Järvinen, “Efficient MCMC for climate model parameter estimation: Parallel adaptive chains and early rejection,” *Bayesian Analysis*, 7(3), pp. 715–736, 2012.
- [231] T.T. Soong, *Random Differential Equations in Science and Engineering*, Academic Press, New York, 1973.
- [232] J.C. Spall, *Introduction to Stochastic Search and Optimization*, John Wiley and Sons, Hoboken, NJ, 2003.
- [233] W.M. Stacey, *Nuclear Reactor Physics*, John Wiley and Sons, New York, 2001.
- [234] D.A. Stainforth, M.R. Allen, E.R. Tredger, and L.A. Smith, “Confidence, uncertainty and decision-supported relevance in climate predictions,” *Philosophical Transactions of the Royal Society A*, 365, pp. 2145–2161, 2007.
- [235] I. Stakgold, *Green’s Functions and Boundary Value Problems*, Second Edition, John Wiley and Sons, New York, 1998.
- [236] M. Stein, *Interpolation of Spatial Data: Some Theory for Kriging*, Springer-Verlag, New York, 1999.
- [237] J. Stoer and R. Bulirsch, *Introduction to Numerical Analysis*, Second Edition, Springer-Verlag, New York, 1993.
- [238] M. Strong, J.E. Oakley, and J. Chilcott, “Managing structural uncertainty in health economic decision models: A discrepancy approach,” *Journal of the Royal Statistical Society C*, 61, pp. 25–45, 2012.

- [239] A. Stroud, *Approximate Calculation of Multiple Integrals*, Prentice–Hall, Englewood Cliffs, NJ, 1971.
- [240] A. Stroud and D. Secrest, *Gaussian Quadrature Formula*, Prentice–Hall, New York, 1966.
- [241] B. Sudret, “Global sensitivity analysis using polynomial chaos expansions,” *Reliability Engineering and System Safety*, 93, pp. 964–979, 2008.
- [242] T. Sumner, E. Shephard, and I.D.L. Bogle, “A methodology for global-sensitivity analysis of time-dependent outputs in systems biology modelling,” *Journal of the Royal Society Interface*, 9, pp. 2156–2166, 2012.
- [243] L.P. Swiler, B.M. Adams, and M.S. Eldred, “Model calibration under uncertainty: Matching distribution data,” Sandia Technical Report SAND Report 2008-0632A; AIAA-2008-5944 in Proc. 12th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, Victoria, British Columbia, September 10–12, 2008.
- [244] A. Tarantola, *Inverse Problem Theory and Methods for Model Parameter Estimation*, SIAM, Philadelphia, 2005.
- [245] D.M Tartakovsky, “Assessment and management of risk in subsurface hydrology: A review and perspective,” *Advances in Water Resources*, 51, pp. 247–260, 2013.
- [246] C.J.F. Ter Braak, “A Markov chain Monte Carlo version of the genetic algorithm differential evolution: Easy Bayesian computing for real parameter spaces,” *Statistics and Computing*, 16(3), pp. 239–249, 2006.
- [247] D.J.J. Toal, N.W. Bressloff, and J. Kean, “Kriging hyperparameter tuning strategies,” *AIAA Journal*, 46(5), pp. 1240–1252, 2008.
- [248] L.N. Trefethen, *Spectral Methods in MATLAB*, SIAM, Philadelphia, 2000.
- [249] L.N. Trefethen, “Is Gauss Quadrature Better than Clenshaw–Curtis?”, *SIAM Review*, 50(1), pp. 67–87, 2008.
- [250] US Nuclear Regulatory Commission, *Safety Evaluation Report Related to Disposal of High-Level Radioactive Wastes in a Geological Repository at Yucca Mountain, Nevada*, NUREG-1949, Volume 2, Washington, DC, 2010.
- [251] P. Valko and M.J. Economides, *Hydraulic Fracture Mechanics*, John Wiley and Sons, Chichester, UK, 1995.
- [252] G.K. Vallis, *Atmospheric and Oceanic Fluid Dynamics: Fundamentals and Large-Scale Circulation*, Cambridge University Press, Cambridge, UK, 2006.

- [253] P.M.J. Van den Hoff, J.F.M. Van Doren, and S.G. Douma, “Identification of parameters in large scale physical model structures, for the purpose of model-based operations,” in *Model-Based Control: Bridging Rigorous Theory and Advanced Technology*, P.M.J. Van den Hoff, C. Scherer, and P.S.C. Heuberger, Eds., Springer, Dordrecht, pp. 125–143, 2009.
- [254] E. Vanmarcke, *Random Fields: Analysis and Synthesis*, Revised and Expanded New Edition, World Scientific, Hackensack, NJ, 2010.
- [255] M. Vihola, “Robust adaptive Metropolis algorithm with coerced acceptance rate,” *Statistics and Computing*, 22(5), pp. 997–1008, 2012.
- [256] C.R. Vogel, *Computational Methods for Inverse Problems*, SIAM, Philadelphia, 2002.
- [257] S. Volkwein, “Model reduction using proper orthogonal decomposition,” Preprint; <http://www.math.uni-konstanz.de/numerik/personen/volkwein/teaching/POD-Vorlesung.pdf>.
- [258] V. Volterra, “Fluctuations in the abundance of species considered mathematically,” *Nature*, 119, pp. 558–560, 1926.
- [259] J.A. Vrugt and C.J.F. Ter Braak, “DREAM<sub>(D)</sub>: An adaptive Markov Chain Monte Carlo simulation algorithm to solve discrete, noncontinuous, and combinatorial posterior parameter estimation problems,” *Hydrology and Earth System Sciences*, 15, pp. 3701–3713, 2011.
- [260] J.A. Vrugt, C.J.F. Ter Braak, M.P. Clark, J.M. Hyman, and B.A. Robinson, “Treatment of input uncertainty in hydrological modeling: Doing hydrology backward with Markov chain Monte Carlo, simulation,” *Water Research Resources*, 44(12), W00B09, doi:10.1029/2007WR006720, 2008.
- [261] J.A. Vrugt, C.J.F. Ter Braak, C.G.H. Diks, B.A. Robinson, J.M. Hyman, and D. Higdon, “Accelerating Markov chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling,” *International Journal of Nonlinear Sciences and Numerical Simulation*, 10(3), pp. 273–290, 2009.
- [262] N. Wiener, “The homogeneous chaos,” *American Journal of Mathematics*, 60, pp. 897–936, 1938.
- [263] D.S. Wilks, *Statistical Methods in the Atmospheric Sciences: An Introduction*, Academic Press, San Diego, CA, 1995.
- [264] M.L. Williams, “Perturbation theory for nuclear reactor analysis,” in *CRC Handbook of Nuclear Reactors Calculation*, Volume 3, Y. Ronen, Ed., CRC Press, Boca Raton, FL, pp. 63–188, 1986.
- [265] J. Wloka, *Partial Differential Equations*, Second Edition, Translated by C.B. and M.J. Thomas, Cambridge University Press, Cambridge, UK, 1992.

- [266] D. Xiu, *Numerical Methods for Stochastic Computations: A Spectral Method Approach*, Princeton University Press, Princeton, NJ, 2010.
- [267] D. Xiu and J.S. Hesthaven, “High-order collocation methods for differential equations with random inputs,” *SIAM Journal on Scientific Computing*, 27(3), pp. 1118–1139, 2005.
- [268] D. Xiu and G.E. Karniadakis, “Modeling uncertainty in flow simulations via generalized polynomial chaos,” *Journal of Computational Physics*, 187, pp. 137–167, 2003.
- [269] C.H. Yew, *Mechanics of Hydraulic Fracturing*, Gulf, Houston, TX, 1997.

# Index

## Symbols

4D-VAR, 18

## A

- Acceptance probability, 159
  - delayed rejection, 173
- Acceptance ratio, 170
- Active subspace method, 113
- Additive model, 323
- Adjoint
  - formal, 349, 351
  - of unbounded operator, 348–349
  - example, 349–350
- Adjoint boundary conditions, 314
- Adjoint Hilbert space, 348
- Adjoint matrix, 348
- Adjoint sensitivity analysis
  - procedure (ASAP)
  - approach perturbation, 308–309, 317–318
  - approach variational, 309–311, 316–317
  - examples
    - algebraic problem, 308–309
    - boundary value problem, 310–311
    - ODE, 316–318
    - functional analysis, 313–314
- Adjoint sensitivity equation
  - algebraic problem, 308
- Aerosols, 11
- Aleatoric uncertainty, 7
- Analysis of variance (ANOVA), 291
- ANOVA-HDMR, *see*

High-dimensional model representation (HDMR),  
ANOVA-

Anthropogenic climate forces, 25–28

Asymptotic sampling distribution, 139

Atmospheric physics

- conservation relations, 13–15
- phenomenological models, 15

Autocorrelation, 170

Automatic differentiation (AD), 145, 305

Autoregressive models, 89

## B

Bayes' formula, 100

Bayes' theorem of inverse problems, 156

Bayesian inference, 100–104

empirical, 100

Beta distribution, 74

- conjugate prior, 103
- example, 339–342

Bilinear form

ASAP, 313

to construct adjoint, 349–350

Binomial distribution, 84, 101–102

Binomial model, 103

Biological systems, 44–47

HIV model, 47–50, 54–55

uncertainties, 45–50

Burn-in, *see* Metropolis algorithm,  
convergence

**C**

- Central limit theorem, 86–87, 139  
 Chebyshev nodes, 252  
 Chi-squared distribution, 72  
 Cholesky decomposition, 160  
 Clenshaw–Curtis  
     nodes, 242, 254  
     quadrature errors, 243–244  
 Climate, 21–33  
     aerosol emission, 28  
     boundary value problem, 21  
     deforestation, 28  
     energy budget, 21  
     equations of atmospheric  
       physics, 14  
     greenhouse effect, 25  
     greenhouse gases, 25–28  
       uncertainties, 27–28  
       water vapor, 29  
     ice albedo effects, 29  
     segment length curse, 27  
     solar radiation, 23  
     uncertainties, 27–28, 30–32  
     volcanic effects, 24  
 Climate debate, 33  
 Climate forces, 22–29  
     anthropogenic, 25–28  
     feedback mechanisms, 28–29  
     natural, 23  
 Climate models, 21–22, 29–32  
 Climate questions, 22  
 Climate scenarios, 30  
 Climate simulation codes, 29–30  
 Collocation method, *see* Stochastic  
     collocation method  
 Conditional pdf, *see* Probability  
     density function  
 Confidence interval, 80–82  
     for parameters, 139–142, 146,  
       152  
     interpretation, 99  
     versus prediction interval,  
       197–200  
 Conjugacy, 103  
 Conjugate prior, *see* Prior  
     distribution

**Conservation relations**

- atmospheric, 13–14  
 neutron transport, 40  
 subsurface hydrology, 35  
 thermal-hydraulic, 41–42

**Convergence**

- almost sure, 85  
 in distribution, 85–86, 139  
 in probability, 85, 139

**Correlation**

- Nataf and Rosenblatt  
     transformations, 108–109  
     versus identifiability, 125–127

**Correlation coefficient**, 77, 125**Correlation function**, 276**Covariance**, 77

- Covariance matrix**  
     chain, 172–173  
     definition, 78  
     estimate, 162  
     in proposal distribution, 160  
     parameter estimation, 136, 145,  
       152

**Credible interval**, 100**Cubature rules**, 250**Cumulative distribution function**  
     (cdf), 68**joint**, 76**Cut-HDMR**, *see* High-dimensional  
     model representation  
     (HDMR), cut-**D**

- DAKOTA**, *see* Design Analysis Kit  
     for Optimization and  
     Terascale Applications  
**Data-fit model**, *see* Surrogate model,  
     regression, interpolation  
**Delayed rejection adaptive**  
     Metropolis, *see* DRAM  
**Design Analysis Kit for Optimization**  
     and Terascale Applications  
     (DAKOTA), 236  
**GP**, kriging models, 299  
**Detailed balance condition**, 96,  
       168–171

- delayed rejection, 174  
DiffeRential Evolution Adaptive Metropolis, *see* DREAM  
Direct effect, 312  
Discrete projection, *see* Stochastic discrete projection  
Distribution, 70  
    beta, 74  
    binomial, 84, 101–102  
    chi-squared, 72  
    gamma, 73  
    inverse chi-squared, 74  
    inverse-gamma, 73–74  
        conjugate prior, 163  
    multivariate normal, 78  
    normal, 70–71  
    proposal, 160  
    sampling, 80  
    Student’s *t*-, 72–73  
    uniform, 71–72  
DRAM, 172–180  
    algorithm, 175–176  
    examples  
        heat model, 176–179  
        HIV model, 179–180  
    software, 175  
DREAM, 181–183  
Dual space, 345
- E**
- Elementary effect, 125, 331  
Emulator, *see* Surrogate model, regression, interpolation  
Energy budget, 21  
Ensemble forecasts, 19  
Epistemic uncertainty, 8  
Errors  
    measurement and model, 133  
    variance estimator, 136–137, 142, 146  
        Bayesian, 163  
Estimate  
    for covariance matrix, 162  
    for parameters, 135, 142, 146, 151  
    maximum a posteriori, 157
- maximum likelihood, 83–85  
OLS, 82, 135  
point and interval, 79–82  
realization of estimator, 80  
Estimator  
    confidence interval, 80–82  
    consistent, 86  
    definition, 80  
    error variance, 136–137, 142, 146  
    for parameters, 135, 142, 146, 151  
    interval, 80  
    maximum likelihood, 83–85  
    OLS, 82, 135, 142, 146  
    unbiased, 80  
Evolution processes, *see* Models  
Expectation, 70  
Explanatory variables, 132
- F**
- Factors, 331  
Fejér nodes, 242  
Fisher information matrix, 164  
Forward sensitivity analysis  
    procedure (FSAP)  
examples  
    algebraic problem, 306  
    boundary value problem, 310  
    ODE, 316  
    functional analysis, 313  
Fracking, 34  
Fréchet differential, derivative, 347–348  
Functional, 345  
Functional principal component analysis (fPCA), 338
- G**
- Gâteaux differential, derivative, 347  
Gâteaux variation  
    algebraic problem, 307  
    boundary value problem, 310  
    definition, 347  
    response, 310, 312  
    to construct sensitivity equations, 313

- Galerkin method, *see* Stochastic Galerkin method
- Gamma distribution, 73
- Gauss–Hermite quadrature, *see* Quadrature rule, Gauss–Hermite
- Gaussian process (GP)
- as surrogate model, 275–278
  - definition, 89
  - for model discrepancy, 266
- Gelfand triple, 311
- Generalized Fourier coefficients, 216
- Global sensitivity analysis, *see* Sensitivity analysis, global
- Greenhouse effect, 25
- Greenhouse gases, 25–28
- H**
- Hamiltonian, 310
- Hermite basis method, 284
- Hermite polynomials, 210–211
- High-dimensional model
- representation (HDMR), 289–298
  - ANOVA-, 290–292
  - based on cut-, 296–298
  - cut-, 293–295
  - RS-, 292–293
  - second-order expansion, 323
- Hilbert space, 346
- Human immunodeficiency virus (HIV) model, *see* Models
- Hyperparameters, 103, 163, 277
- for model discrepancy, 265, 267
- I**
- Identifiable parameter subspace
- definition, 113–114
  - example, 53, 56
  - for model discrepancy, 267
  - relation to range, 116–117
  - versus correlation, 125–127
- Importance measures, 324
- Independent and identically distributed (iid) random variables, 79
- Influential parameters, 114
- Initial condition, projection, 281
- Inner product space, 346
- Inputs, 3
- uncertainties, 6
- Interpolation
- 1-D, 250–252
  - Chebyshev nodes, 252
  - error bound, 252
  - sparse grid, 254
  - tensor product, 253
- Interval estimator, 80–82
- Intrusive methods, 214
- Inverse chi-squared distribution, 74
- Inverse-gamma distribution, 73–74, 163
- conjugate prior, 104
- Inverse transform sampling, 76
- Inverse uncertainty quantification, 6, 132
- Irreducible uncertainty, *see* Aleatoric uncertainty
- Ishigami function, 329
- J**
- Jeffreys prior, 164
- Jumping distribution, *see* Proposal distribution
- K**
- Karhunen–Loëve expansion, 109–112
- relation to POD, 287
- Kernel density estimation (kde), 75–76
- Kriging model, 275–278
- for model discrepancy, 266
- Kronecker delta, multiple variables, 213
- L**
- Lagrange basis method, 284
- Lagrange polynomial, 251
- for cut-HDMR, 295
- Law of large numbers, 86, 139
- Least squares, *see* Ordinary least squares
- Lebesgue constant, 252

- Legendre polynomials, 211  
Likelihood function  
    Bayesian, 101, 155–156, 161  
    definition, 83–84  
    surrogate, 298  
Linear regression, 134–141  
Local sensitivity analysis, *see*  
    Sensitivity analysis, local
- M**
- Marginal pdf, *see* Probability density function  
Markov chain, 90–96  
    definition, 94  
    detailed balance, 96  
    homogeneous, 91  
    irreducible, 93  
    parameter density, 159–162  
    periodic, 94  
    stationary distribution, 93  
Markov chain Monte Carlo (MCMC), 159  
Matrix  
    Cholesky decomposition, 160  
    idempotent, 137  
    null space, 116  
    positive, 94  
    QR factorization, 118  
        in random algorithm, 119  
    range, 116  
    row-stochastic, 91  
    SVD, 117–118  
        in random algorithm, 119  
    trace properties, 137  
Maximum a posteriori estimate, 157  
Maximum likelihood estimate (MLE), 84  
Maximum likelihood estimator, 83–85  
Mean, 70  
Measurement errors, 133  
Meta-model, *see* Surrogate model, regression, interpolation  
Method of snapshots, 289  
Metropolis algorithm, 159–165  
    acceptance ratio, 170, 173  
    adaptive, 172–173  
    convergence, 168–171, 174  
    delayed rejection, 173–174  
    examples  
        heat model, 176–179  
        HIV model, 179–180  
        spring model, 165–167  
    mixing, 161, 173  
    random walk, 160  
    scaled parameters, 176  
    using surrogate, 298  
Metropolis–Hastings algorithm, 165  
Model calibration, 8, 82  
Model discrepancy, 133  
    bias, 9  
    effects, 261  
    issues, 267–269  
    quantification, 265–267  
    relation to epistemic  
        uncertainties, 8  
Model errors, *see* Model discrepancy  
Models  
    abstract framework  
        linear, 63–65  
        nonlinear, 65  
    algebraic, 62  
    atmospheric physics, 14  
    autoregressive (AR), 89  
    beam, 58–60  
    Burgers' equation, 60  
    evolution processes, 61–62  
    exponential processes, 51–52  
    groundwater flow, 35–36  
    heat, 55–57  
    HIV, 47–50, 54–55  
    neutron, 40, 57–58  
    portfolio, 321, 328–329, 335  
    simple harmonic oscillator, 52–54  
    SIR, 55  
    stationary processes, 62  
    thermal-hydraulic, 41–42  
Morris screening, 331–337  
    elementary effect, 125, 331  
        scaled, 332  
    factors, 331

- for parameter selection, 124–125  
 sampling strategy, 333–335  
 sensitivity measures, 332
- M**  
**Multi-index**  
 definition, 212  
 sparse grids, 246
- Multivariate normal distribution, 78
- N**  
 Nataf transformation, 109
- Noninfluential parameters**, *see* Parameters
- Noninformative prior, 100
- Nonintrusive methods, 214
- Normal distribution, 70–71
- Nuclear reactor  
 CASL, 37  
 design, 36–39  
 light water reactors, 37–39  
 models, 39–42  
 neutron, 39–41, 57–58  
 simulation packages, 41, 42  
 thermal-hydraulic, 41–42
- QoI, 43  
 uncertainties, 42–43
- Nugget, in Kriging model, 277
- Null space, *see* Matrix
- Numerical errors and uncertainties, 7
- Numerical weather prediction  
 (NWP), *see* Weather, models
- O**
- Optimization routines, 143
- Ordinary least squares (OLS)  
 estimate, scaled, 143  
 estimator, 82, 135, 142, 146, 151  
 functional, 135, 142  
 scaled, 143
- Orthogonal complement, 113
- Orthogonal polynomials  
 Hermite polynomials, 210–211  
 Legendre polynomials, 211
- P**
- Parameter estimation, 6, 132
- Bayesian perspective, 155–158  
 frequentist perspective, 133–134
- Parameter selection  
 linear problems  
 deterministic, 117–118  
 random algorithms, 119–122  
 nonlinear problems  
 linearization-based, 123–125  
 variance-based, 122–123
- Parameters  
 as random variables, 107–112  
 correlated, 108  
 finite-dimensional  
 representation, 109–112  
 mutually independent, 107  
 confidence interval, 139–142,  
 146, 152  
 estimator and estimate,  
 135–136, 142, 146, 151  
 identifiability, 113–114  
 Bayesian algorithms, 171–172  
 relation to range, 116–117  
 versus correlation, 125–127  
 in Markov chain, 159  
 influential, 114  
 polynomial representation  
 multiple variables, 212–214  
 single variable, 209  
 relation to inputs, 3  
 sampling distribution, 138–139,  
 145, 151
- Perron–Frobenius theorem, 94
- POD, *see* Proper orthogonal decomposition
- Point estimate, 79
- Polynomial chaos (PC), 208
- Posterior density, 100–101, 156  
 based on conjugate, 163
- Prediction interval, 197–203  
 definition, 199  
 extrapolation, 199–200  
 for uncertainty quantification,  
 201–202  
 versus confidence interval,  
 197–200
- Predictive estimation, 8–10

- Predictive science, 1  
Prior distribution  
  conjugate, 103–104, 163  
  Jeffreys, 164  
  noninformative, 100  
  parameter estimation, 156  
Probabilistic risk assessment, 44  
Probability density function (pdf),  
  69–70  
  conditional, 78–79  
  marginal, 78  
Probability mass function, 70  
Probability space, 67  
  image space, 107, 209  
Propagation of moments, 194  
Proper orthogonal decomposition  
  (POD), 285–289  
  relation to SVD, 288  
Proposal distribution, 160  
  delayed rejection, 173
- Q**
- QR factorization, 118  
  in random algorithm, 119  
Quadrature rule  
  Gauss–Hermite, 211, 241  
  Gauss–Legendre, 241  
  nested, 241–243  
    Clenshaw–Curtis, 242  
    composite trapezoid, 242  
    error bound, 243  
  sparse grid, 244–250  
    adaptive, 249  
    Clenshaw–Curtis, 246  
    error, 248  
    multi-index, 246  
    nodal set, 246  
  stochastic, 239–240  
  tensor product, 243–244  
    Clenshaw–Curtis error, 244  
Quantile-quantile (Q–Q) plot, 74  
Quantity of interest (QoI), 4  
  climate models, 22, 31  
  evolutionary PDE, 222  
  HIV model, 50  
  nuclear reactor models, 43
- ODE, 215  
stationary problems, 218  
weather models, 3, 19, 21
- R**
- Radial basis functions, 278  
Random differential equation, 96  
Random field, 89  
Random process, 87–90  
  correlated and uncorrelated, 111  
  definition, 88  
  Gaussian, 89  
  in Kriging model, 277  
  polynomial expansion, 208–209  
  second-order, 88  
  spectral representation, 207–208  
  state space, 90  
  stationary, 89  
Random range algorithm, 119  
Random variable  
  definition, 68  
  iid, 79  
  independent, 77  
  multiple, 76  
  normal, representation, 212  
  polynomial representation  
    mean and variance, 210  
    multiple variables, 212–214  
    single variable, 209  
  S-valued, 90  
  uncorrelated, 77  
  uniform, representation, 212  
Random vector, 76  
Random walk Metropolis, *see*  
  Metropolis algorithm  
Realization, 68  
Reduced-order model, *see* Surrogate  
  model, projection-based  
Regression, *see* Linear regression  
Regressor variables, 132  
Response surface model, *see*  
  Surrogate model,  
    regression, interpolation  
Riesz representation theorem, 346  
  for ASAP, 311  
Rosenblatt transformation, 109

- RS-HDMR, *see* High-dimensional model representation (HDMR), RS-
- Runge function, 251
- S**
- Sample mean, 80, 173
- Sample variance, 80
- Sampling distribution, 80
- asymptotic, 139
  - parameter, 138–139, 145, 151
- Sandwich relation, 193
- Screening techniques, *see* Morris screening
- Sensitivity analysis
- examples
    - Ishigami function, 329
    - neutron diffusion, 306–311, 314–315
    - portfolio model, 321–323, 328–329, 335
    - SIR disease model, 339–342
    - Sobol function, 335–337
    - spring model, 304–305, 315–318
  - global
    - definition, 304
    - for parameter selection, 123
    - Morris screening, 331–337
    - Sobol indices, 324–329
    - time- or space-dependent, 338–342
  - local
    - ASAP, 308–311, 313
    - definition, 303
    - derivative relation, 322
    - for parameter selection, 123–125
    - for uncertainty quantification, 192
    - FSAP, 306, 310, 313
    - sigma-normalized, 322
- Sensitivity equations
- algebraic problem, 307
  - boundary value problem, 310
  - general model, 313
- Sensitivity indices
- local, 322
  - Morris, 332
  - Sobol, 324
  - computational algorithm, 330–331
  - general densities, 327
- Sensitivity matrix
- parameter estimation, 144, 152
- Singular value decomposition (SVD), 117–118
- in random algorithm, 119
  - relation to POD, 288
- Snapshot set, 284
- Sobol function, 335
- Sobol indices, 324
- computational algorithm, 330–331
  - for parameter selection, 123
  - general densities, 327
  - statistical properties, 325
  - total sensitivity indices, 324
- Sobol representation, 289, 323
- general densities, 326–329
- Sparse grid
- interpolation, 254
  - quadrature, 244–250
  - adaptive, 249
  - Clenshaw–Curtis, 246
  - error, 248
  - multi-index, 246
  - nodal set, 246
- Standard deviation, 70
- Standard error, 141, 152
- State space, 90
- Stationary distribution, 93, 168–171
- Stationary processes, *see* Models
- Statistic, 80
- Statistical inference
- Bayesian, 100–104
  - frequentist vs. Bayesian, 98–100
  - goals, 98
- Statistical model, 133, 142
- Statistical uncertainty, *see* Aleatoric uncertainty
- Stochastic collocation method

- as surrogate model, 279–280  
attributes, 224–225  
evolutionary PDE, 223  
ODE, 217  
stationary problems, 220–221
- Stochastic differential equation (SDE), 97
- Stochastic discrete projection  
    attributes, 226  
    evolutionary PDE, 223  
    examples, 234–235  
    ODE, 218  
    stationary problems, 221
- Stochastic Galerkin method  
    attributes, 223–224  
    evolutionary PDE, 222  
    examples, 226–234  
    ODE, 216  
    stationary problems, 220
- Stochastic polynomial packages, 235
- Stochastic process, *see* Random process
- Stochastic weak model formulation, 216, 219, 233
- Student's *t*-distribution, 72–73
- Subset selection, 113
- Subsurface hydrology, 33–36  
    hydraulic fracturing, 33  
    models, 35–36  
    uncertainties, 35–36  
    Yucca mountain, 33
- Sum of squares error, 156  
    with surrogate, 298
- Surrogate model  
    for model calibration, 298–299  
    projection-based, 280–282  
        eigenfunction, 283–284  
        HDMR, 289–298  
        POD, 284–289  
    regression, interpolation, 273–280  
        kriging, GP, 275–278  
        quadratic, 275  
        radial basis function, 278–279  
    stochastic collocation, 279–280
- Systematic uncertainty, *see* Epistemic uncertainty
- T**
- Taylor basis method, 284
- U**
- Unbounded operator, adjoint, 348–349
- Uncertainties  
    aleatoric, 7  
    epistemic, 8  
    sources  
        experimental, 5  
        models and inputs, 5–7  
        numerical, 7
- Uncertainty propagation  
    linear models, 188–191  
    perturbation methods, 192–197  
    sampling methods, 191–192
- Unidentifiable parameter subspace,  
    *see* Identifiable parameter subspace
- Uniform distribution, 71–72
- V**
- Validation, 4
- Validation regime, 8
- Vandemonde system, 217
- Variance  
    definition, 70  
    partial and total, 324  
    for parameter selection, 123  
    general densities, 327  
    statistical properties, 325
- Verification, 4
- W**
- Weak model formulation  
    deterministic, 218  
    stochastic, 216, 219, 233
- Weather, 2–3, 11–20  
    ensemble forecasts, 19  
    equations of atmospheric physics, 14  
    Katrina, 19  
    models

- 4D-VAR, 18  
closure relations, 15  
conservation relations, 13–15  
data assimilation, 16–18  
primitive equations, 14  
numerical models, 15–16  
ECMWF, UK Met, 16, 18  
uncertainties, 18–19  
Wiener PC, 209