Reduced-Basis Approximations and A Posteriori Error Bounds for Nonaffine and Nonlinear Partial Differential Equations: Application to Inverse Analysis

by
Nguyen Ngoc Cuong
B.Eng., HCMC University of Technology
Submitted to the HPCES Programme
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in High Performance Computation for Engineered Systems
at the
SINGAPORE-MIT ALLIANCE
June 2005

Author .....
HPCES Programme
June, 2005

Certified by .....
Anthony T. Patera
Professor of Mechanical Engineering - MIT
Thesis Supervisor

Certified by .....
Liu Gui-Rong
Associate Professor of Mechanical Engineering - NUS
Thesis Supervisor

Accepted by .....
Associate Professor Khoo Boo Cheong
Programme Co-Chair
HPCES

Accepted by .....
Professor Jaime Peraire
Programme Co-Chair
HPCES

Reduced-Basis Approximations and A Posteriori Error Bounds for Nonaffine and Nonlinear Partial Differential Equations: Application to Inverse Analysis

by

Nguyen Ngoc Cuong

Submitted to the HPCES Programme
on June, 2005, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in High Performance Computation for Engineered Systems

Abstract

Thesis Supervisor: Anthony T. Patera
Title: Professor of Mechanical Engineering - MIT

Thesis Supervisor: Liu Gui-Rong
Title: Associate Professor of Mechanical Engineering - NUS

Acknowledgments

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Engineering design and optimization require not only understanding of the principles of physics, but also application of necessary mathematical tools. Mathematically, many components/systems and processes in applied science and engineering are modeled by partial differential equations that describe the underlying physics. Typically, the quantity of primary importance is not the full field variable, but rather certain selected *outputs* defined as functionals of the field variable. Typical outputs include energies, and forces, critical stresses or strains, flowrates or pressure drops, temperature, and flux. These outputs are functions of system parameters, or *inputs*, that serve to identify a particular configuration of the component or system; these inputs typically reflect geometry, properties, and boundary conditions and loads. The input-output relationship thus encapsulates the behavior relevant to the desired engineering context. However, its evaluation demands solution of the underlying partial differential equation (PDE). Engineering design and optimization typically require thousands of input-output evaluations in real-time.

Virtually all classical numerical approaches (e.g., FEM/FDM/BEM etc.) consider effectively "dense" approximation subspaces for the underlying PDE: the computational time for a particular input is thus typically very long despite continuing advances in computer speeds and hardware capabilities. An implication of this is that: we can not address many in-operation/in-service applications in engineering design, operations, and analysis that require either real-time response or simply many queries; and hence we can not perform adaptive design and optimization of components or systems, robust parameter estimation of properties and state, or control of missions and processes.

A goal of this thesis is to remedy this deficiency and specifically to develop a computational approach that can provide output predictions that are *certifiably* as good as the classical truth approximations but literally *several order of magnitude less expensive*. Another goal of this thesis is to apply the approach for numerical analysis of inverse problems in engineering and science with special emphasis on real-time capability and robust handling of uncertainty.

1.1 Problem Definition

1.1.1 Forward Problems

We consider the "exact" (superscript e) forward problem: Given $\mu \in \mathcal{D} \subset \mathbb{R}^P$, we evaluate $s^{\rm e}(\mu) = \ell(u^{\rm e}(\mu))$, where $u^{\rm e}(\mu)$ satisfies the weak form of the μ-parametrized PDE, $a(u^{\rm e}(\mu), v; \mu) = f(v),\ \forall\, v \in X^{\rm e}$. Here μ and $\mathcal{D}$ are the input and (closed) input domain, respectively; $s^{\rm e}(\mu)$ is the output of interest; $u^{\rm e}(\mu; x)$ is our field variable; $X^{\rm e}$ is a Hilbert space defined over the physical domain $\Omega \subset \mathbb{R}^d$ with inner product $(w, v)_{X^{\rm e}}$ and associated norm $\|w\|_{X^{\rm e}} = \sqrt{(w, w)_{X^{\rm e}}}$; and $a(\cdot, \cdot; \mu)$ and $f(\cdot), \ell(\cdot)$ are $X^{\rm e}$-continuous bilinear and linear functionals, respectively.

It should be emphasized that the evaluation of input-output relationship demands solution of the parametrized PDE. In general, the PDEs are not analytically solvable, rather a classical approach like the finite element method is used to seek a weak-form solution. We henceforth introduce $X \subset X^{\rm e}$, a "truth" finite element approximation space of dimension $\mathcal{N}$. Our finite element approximation of the exact problem can then be stated as: given $\mu \in \mathcal{D}$, find

$$s(\mu) = \ell(u(\mu)) \ , \tag{1.1}$$

where $u(\mu) \in X$ satisfies a discrete weak formulation

$$a(u(\mu), v; \mu) = f(v), \quad \forall v \in X \ . \tag{1.2}$$

We assume that X is sufficiently rich that $u(\mu)$ (respectively, $s(\mu)$) is sufficiently close to

$u^e(\mu)$ (respectively, $s^e(\mu)$) for all μ in the (closed) parameter domain $\mathcal{D}$. The dimension $\mathcal{N}$ required to satisfy this condition — even with the application of appropriate (and even parameter-dependent) adaptive mesh generation/refinement strategies — is typically very large, and in particular much too large to provide real-time response in the design and optimization contexts. We shall also assume that the forward problem is strictly well-posed in the sense of Hadamard, i.e., it has a unique solution that depends continuously on data.

1.1.2 Inverse Problems

In inverse problems we are concerned with predicting the unknown parameters from the measured-observable outputs. In the context of inverse problems, our input has two components, $\mu = (\nu, \sigma)$, where $\nu \in \mathcal{D}^\nu$ is characteristic-system parameter and σ is experimental control variable. The inverse problems involve determining the true but unknown parameter ν^* from noise-free measurements $\{s(\nu^*, \sigma_k), 1 \le k \le K\}$. In practice, due to the presence of noise in measurement the experimental data is given in the form of intervals

$$\mathcal{I}(\epsilon_{\text{exp}}, \sigma_k) \equiv \left[s(\nu^*, \sigma_k) - \epsilon_{\text{exp}} \left| s(\nu^*, \sigma_k) \right| , s(\nu^*, \sigma_k) + \epsilon_{\text{exp}} \left| s(\nu^*, \sigma_k) \right| \right], k = 1, \ldots, K \ ; \quad (1.3)$$

where ϵ_{exp} is the error in measurement.

Our inverse problem formulation is thus: given experimental data $\mathcal{I}(\epsilon_{\text{exp}}, \sigma_k), k = 1, \ldots, K$, we wish to determine the region $\mathcal{P} \in \mathcal{D}^\nu$ in which the unknown parameter ν^* must reside. Towards this end, we define

$$\mathcal{P} \equiv \{\nu \in \mathcal{D}^\nu | s(\nu, \sigma_k) \in \mathcal{I}(\epsilon_{\text{exp}}, \sigma_k), 1 \le k \le K\} \quad (1.4)$$

where $s(\nu, \sigma)$ is determined by (1.1) and (1.2). Geometrically, the inverse problem formulation can be interpreted as: find a region in parameter space such that every point in this region has its image exactly in the given data set.

Unfortunately, the realization of $\mathcal{P}$ requires many *queries* of $s(\nu, \sigma)$, which in turn necessitates repeated solutions of the underlying PDE. Instead, we shall construct a

bounded "possibility region" $\mathcal{R}$ such that $\mathcal{P} \subset \mathcal{R}$. The important point is that $\mathcal{R}$ can be constructed as suitably small as $\mathcal{P}$ but very inexpensively (see Section 1.2.4 for the definition of $\mathcal{R}$ and Chapter 8 for the inverse computational method for constructing $\mathcal{R}$).

1.2 A Motivational Example

The primary focus of this thesis is on: (1) the development of real-time methods for accurate and reliable solution of the forward problems, (2) robust parameter estimation methods for very fast solution region of inverse problems characterized by parametrized PDEs, and (3) application of (1) and (2) to the adaptive design and robust optimization of engineering components or systems. To demonstrate the various aspects of the methods and illustrates the contexts in which we develop them, we consider a simple inverse scattering problem relevant to the detection of an elliptical "mine"[30, 35] and present some indicative results obtained by using the methods.

Before proceeding, we need to clarify our notation used in this section (and in much of the thesis). In the following subsection, we use a *tilde* for those variables depending on the spatial coordinates to indicate that the problem is being formulated over the original domain. Since the original domain is usually parameter-dependent, in our actual implementation, we do not solve the problem directly on the original domain, but reformulate it in terms of a fixed reference domain via a continuous geometric mapping (see Section 10.2 for further detail). In the reference domain, the corresponding variables and weak formulation will bear no tilde.

1.2.1 Problem Description

We consider the scattering of a time harmonic acoustic incident wave (pressure field) $\tilde{u}^i$ of frequency ω by a bounded object $\tilde{D}$ in n–dimensional space $\mathbb{R}^n$ $(n = 2, 3)$ having constant density ρ_D and constant sound speed c_D. We assume that the object $\tilde{D}$ is situated in a homogeneous isotropic medium with density ρ and sound speed c. The incident field is a plane wave

$$\tilde{u}^i(\tilde{x}) = e^{\mathrm{i}k\tilde{x}\cdot\tilde{d}}, \tag{1.5}$$

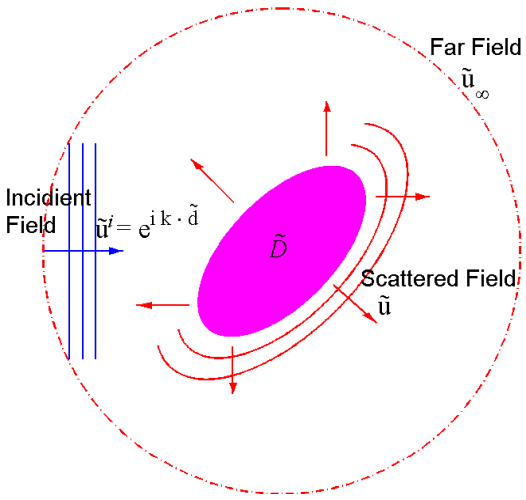


Figure 1-1: Schematic of the model inverse scattering problem: the incident field is a plane wave interacting with the object, which in turn produces the scattered field and its far field pattern.

where the wave number k is given by $k = \omega/c$, and $\tilde{d}$ is the direction of the incident field. Let $\tilde{u}$ be the scattered wave of the sound-hard object (i.e., $\rho_D/\rho \to \infty$) then the total field $\tilde{u}^t = \tilde{u}^i + \tilde{u}$ satisfies the following exterior Neumann problem [29]

$$\Delta\tilde{u}^t + k^2\tilde{u}^t = 0 \quad \text{in } \mathbb{R}^n\backslash\tilde{D}, \tag{1.6a}$$

$$\frac{\partial\tilde{u}^t}{\partial\tilde{\nu}} = 0 \qquad \text{on } \partial\tilde{D}, \tag{1.6b}$$

$$\lim_{\tilde{r}\to\infty} \tilde{r}^{(n-1)/2}\left(\frac{\partial\tilde{u}}{\partial\tilde{r}} - \mathrm{i}k\tilde{u}\right) = 0, \quad \tilde{r} = |\tilde{x}| \tag{1.6c}$$

where $\tilde{\nu}$ is the unit outward normal to $\partial\tilde{D}$. Mathematically, the Sommerfeld radiation condition (1.6c) ensures the wellposedness of the problem (1.6); physically it characterizes out-going waves [30]. Equation (1.6c) implies that the scattered wave has an asymptotic behavior of the form [33]

$$\tilde{u}(\tilde{x}) = \frac{e^{\mathrm{i}k\tilde{r}}}{\tilde{r}^{(n-1)/2}}\tilde{u}_\infty(\tilde{D}, k, \tilde{d}, \tilde{d}^s) + O\left(\frac{1}{\tilde{r}^{(n+1)/2}}\right), \tag{1.7}$$

as $|\tilde{x}| \to \infty$, where $\tilde{d}^s = \tilde{x}/|\tilde{x}|$. The function $\tilde{u}_\infty$ defined on the unit sphere $\tilde{S} \subset \mathbb{R}^n$ is known as the scattering amplitude or the far-field pattern of the scattered wave. The Green representation theorem and the asymptotic behavior of the fundamental solution

ensures a representation of the far-field pattern in the form

$$\tilde{u}_\infty(\tilde{D}, k, \tilde{d}, \tilde{d}^s) = \beta_n \int_{\partial\tilde{D}} \left\{ \tilde{u}(\tilde{x}) \frac{\partial e^{-\mathrm{i}k\tilde{d}^s\cdot\tilde{x}}}{\partial\tilde{\nu}} - \frac{\partial\tilde{u}(\tilde{x})}{\partial\tilde{\nu}} e^{-\mathrm{i}k\tilde{d}^s\cdot\tilde{x}} \right\} \tag{1.8}$$

with

$$\beta_n = \begin{cases} \frac{i}{4}\sqrt{\frac{2}{\pi k}} e^{-i\pi/4} & n = 2 \\ \frac{1}{4\pi} & n = 3\ . \end{cases} \tag{1.9}$$

The proof of (1.7) and (1.8) is given in Appendix A.

The forward problem, given the support of the object $\tilde{D}$ and the incident wave $\tilde{u}^i$, is to find the scattered wave $\tilde{u}$ and in particular the far field pattern $\tilde{u}_\infty$. Whereas, the inverse problem is to determine the support of the object $\tilde{D}$ from measurements of the far field pattern $\mathcal{I}(\epsilon_{\text{exp}}, k, \tilde{d}, \tilde{d}^s)$ with error ϵ_{exp}. In the language of our introduction, the input consists of $\tilde{D}, k, \tilde{d}, \tilde{d}^s$ in which $\tilde{D}$ is characteristic-system parameter and $k, \tilde{d}, \tilde{d}^s$ are experimental control variables, and the output is $\tilde{u}_\infty$.

In this section, we shall consider a two-dimensional scattering problem in which $\tilde{D}$ is an elliptical cross-section of an infinite cylinder. Many details including a three-dimensional inverse scattering model will be further reported in Chapters 10 and 11. The object $\tilde{D}$ is then characterized by three parameters (a, b, α), where a, b, α are the major semiaxis, minor semiaxis, and angle of the elliptical object, respectively. In this particular case, the forward problem is to calculate $\tilde{u}$ and $\tilde{u}_\infty$ for any given set of parameters $\mu \equiv (a, b, \alpha, k, \tilde{d}, \tilde{d}^s) \in \mathbb{R}^6$; and our inverse problem is:

> Given the far field data $\mathcal{I}(\epsilon_{\text{exp}}, a^*, b^*, \alpha^*, k, \tilde{d}, \tilde{d}^s)$ measured at several directions $\tilde{d}^s$ with experimental error ϵ_{exp} for one or several directions $\tilde{d}$ and wave numbers k, we wish to find the shape of the elliptical object modeled by three parameters (a^*, b^*, α^*).

1.2.2 Finite Element Discretization

Due to the complex boundary conditions and geometry, obtaining an exact solution to the continuous problem (1.6) is not easy. Instead the finite element method is used to find a good approximation to the exact solution. In the finite element method, the partial

differential equation is transformed into an integral form called the weak formulation. The weak formulation of the problem (1.6) can be derived as: find $u(\mu) \in X$ such that

$$a(u(\mu), v; \mu) = f(v; x; \mu), \quad \forall\, v \in X \ ; \tag{1.10}$$

the far-field pattern is then calculated as

$$s(\mu) = \ell(u(\mu); x; \mu) + \ell^{\mathrm{o}}(x; \mu) \ . \tag{1.11}$$

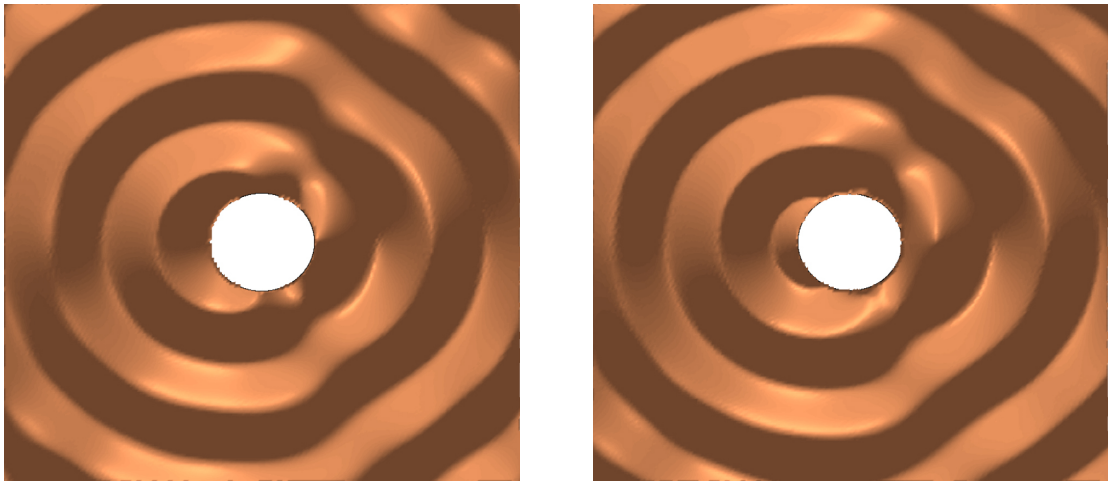
Here $a(\cdot, \cdot)$ is a parametrized bilinear form, f, ℓ, and ℓ^{o} are linear functionals, and X is a finite element "truth" approximation space; note that $u(\mu)$ is complex and X is thus a space of complex continuous functions. The precise definition of X, a, f, ℓ, and ℓ^{o} can be found in Section 10.3.

We then form the elemental matrices and vectors over each elements by representing the approximate solution as the linear combination of basis functions and substituting it into the weak formulation. Finally, by assembling elemental matrices and vectors and imposing the boundary conditions, we transform the weak formulation into a finite set of algebraic equations (see Section 2.4 for details of the finite element method)

$$\underline{A}(\mu)\, \underline{u}(\mu) = \underline{F} \ , \tag{1.12}$$

where $\underline{A}(\mu)$ is the $\mathcal{N} \times \mathcal{N}$ stiffness matrix, $\underline{F}$ is the load vector of size $\mathcal{N}$, and $\underline{u}(\mu)$ is the "complex" nodal vector of the finite element solution $u(\mu)$; here $\mathcal{N}$ is the dimension of the truth approximation space X. By solving the algebraic system of equations, we obtain nodal values from which the approximate solution $u(\mu)$ and the far-field pattern $u_\infty(\mu)$ are constructed.

As an illustrative example, we present in Figure 1-2 the scattered wave $u(\mu)$ near resonance region for $a = b = 1, \alpha = 0$ and $k = \pi$. Here the incoming incident wave is a plane wave traveling in the positive x–direction.



(a) (b)

Figure 1-2: Pressure field near resonance region (a) real part (b) imaginary part.

1.2.3 Reduced-Basis Output Bounds

Using the finite element method, we can calculate numerically the far-field pattern $s(\mu)$ for any given parameter μ. As the dimension of the truth approximation space increases, the error in the approximation decreases. We shall assume that $\mathcal{N}$ is sufficiently large such that numerical output is sufficiently close to the exact one. Unfortunately, for any reason error tolerance, the dimension $\mathcal{N}$ needed to satisfy this condition is typically extremely large, and in particular much too large to provide real-time solution of the inverse scattering problem.

Our approach is based on the reduced-basis method. The main ingredients are (*i*) rapidly uniformly convergent reduced-basis approximations — Galerkin projection onto the reduced-basis space W_N spanned by solutions of the governing partial differential equation at N (optimally) selected points in parameter space; (*ii*) *a posteriori* error estimation — relaxations of the residual equation that provide inexpensive yet sharp and rigorous bounds for the error in the outputs; and (*iii*) offline/online computational procedures — stratagems that exploit affine parametric structure to decouple the generation and projection stages of the approximation process. The operation count for the online stage — in which, given a new parameter value, we calculate the reduced-basis output $s_N(\mu)$ and associated error bound $\Delta_N^s(\mu)$ — depends only on N (typically small) and the parametric complexity of the problem.

We can thus provide output bounds $s_N^-(\mu) = s_N(\mu) - \Delta_N^s(\mu)$ and $s_N^+(\mu) = s_N(\mu) + \Delta_N^s(\mu)$ that satisfy a bound condition $s_N^-(\mu) \le s(\mu) \le s_N^+(\mu)$ and an error criterion $\Delta_N^s(\mu) \le \epsilon_{\text{tol}}^s$. Unlike the true value $s(\mu)$, these output bounds can be computed online very expensively.

1.2.4 Possibility Region

Owing to the low marginal cost, the method is ideally suited to inverse problems and parameter estimation for PDE models: rather than regularize the goodness-of-fit objective, we may instead identify all (or almost all, in the probabilistic sense) inverse solutions consistent with the available experimental data. Towards this end, we first obtain $s_N^\pm(\mu) \equiv s_N(\mu) \pm \Delta_N^s(\mu)$ by applying the reduced-basis method to the discrete problem (1.10), and thus — thanks to our rigorous output bounds — $s(\mu) \in [s_N^-(\mu), s_N^+(\mu)]$.[1] We may then define

$$\mathcal{R} \equiv \left\{ \nu \in \mathcal{D}^\nu \mid \left[s_N^-(\nu, \sigma_k), s_N^+(\nu, \sigma_k) \right] \cap \mathcal{I}(\epsilon_{\text{exp}}, \sigma_k), 1 \le k \le K \right\} \ . \tag{1.13}$$

Recall that $\nu \equiv (a, b, \alpha)$ and $\sigma \equiv (\tilde{d}^s, \tilde{d}, k)$. Clearly, we have accommodated both numerical and experimental error and uncertainty, and hence $\nu^* \in \mathcal{P} \subset \mathcal{R}$.

Central to our inverse computational method is a robust algorithm to construct $\mathcal{R}$. However, in high parametric dimension constructing $\mathcal{R}$ is numerically expensive (even with the application of the reduced-basis method) and representing $\mathcal{R}$ is geometrically difficult. It is therefore desired to have more efficient and visible geometry for representing $\mathcal{R}$ in high-dimensional parameter space. A natural choice is an ellipsoid that includes $\mathcal{R}$.

1.2.5 Indicative Results

We turn to the inverse scattering problem that will serve to illustrate the new capabilities enabled by rapid certified input-output evaluation. In particular, given experimental data in the form of intervals $\mathcal{I}(\epsilon_{\text{exp}}, \sigma_k)$ measured at several angles $\tilde{d}^s$ for several directions $\tilde{d}$ of

[1] Note for this particular example that our error estimators are not completely rigorous bounds in theoretical aspect. However, numerical results in Section 6.6.6 show that in practice the bounds are valid for all $\mu \in \mathcal{D}$ — to be rigorous, they must be provably valid — since the non-rigorous component is quite small relative to the dominant approximation error.

the fixed-frequency incident wave, we wish to determine a region $\mathcal{R} \in \mathcal{D}^{a,b,\alpha}$ in which the true — but unknown — obstacle parameters, a^*, b^* and α^*, must reside. In our numerical experiments, we use a low fixed wavenumber,[2] $k = \pi/8$, and three different directions, $\tilde{d} = \{0, \pi/4, \pi/2\}$, for the incident wave. For each direction of the incident wave, there are $I = 3$ output angles $\tilde{d}_i^s = (i-1)\pi/2, i = 1, \ldots, I$ at which the outputs are collected; hence, the number of measurements is $K = 9$. We show in Figures 1-3(a), 1-3(b), and 1-3(c) the possibility regions — more precisely, (more convenient) 3-ellipsoids that contain the possibility regions for the minor and major axes and orientation — for experimental error of 5%, 2%, and 1%.

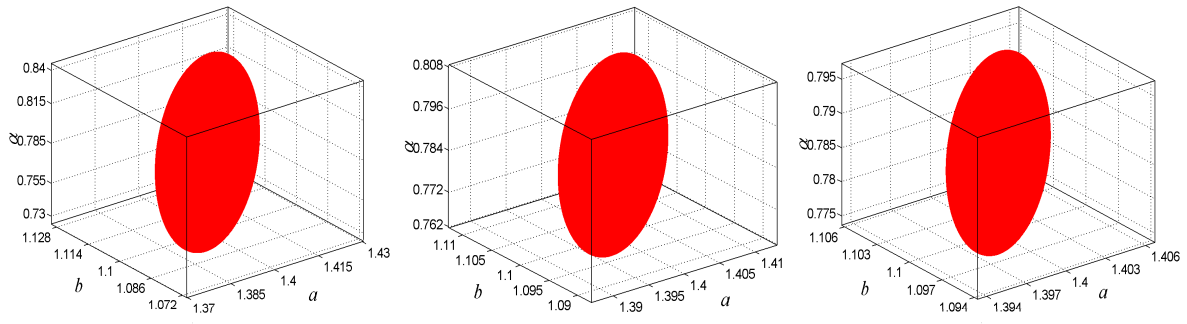


Figure 1-3: Ellipsoid containing possibility region $\mathcal{R}$ for experimental error of 5% in (a), 2% in (b), and 1% in (c). Note the change in scale in the axes: $\mathcal{R}$ shrinks as the experimental error decreases. The true parameters are $a^* = 1.4$, $b^* = 1.1$, $\alpha^* = \pi/4$.

As expected, as ϵ_{exp} decreases, $\mathcal{R}$ shrinks toward the exact (synthetic) value, $a^* = 1.4$, $b^* = 1.1$, $\alpha^* = \pi/4$. More importantly, for any finite ϵ_{exp}, $\mathcal{R}$ *rigorously captures the uncertainty* in our assessment of the unknown parameters without *a priori* assumptions.[3] The crucial new ingredient is reliable fast evaluations that permit us to conduct a much more extensive search over parameter space: for a given ϵ_{exp}, these possibility regions may be generated online in less than 285 seconds on a Pentium 1.6 GHz laptop thanks to a per forward evaluation time of only 0.008 seconds. We can thus undertake appropriate real-time actions with confidence.

[2]For low wavenumber, the inverse scattering problem is computationally easier and less susceptible in practice to scattering by particulates in the path; but, very small wavenumber can actually produce insensitive data which may cause bad recovery [21, 38, 56].

[3]In fact, all uncertainty is eliminated only in the limit of exhaustive search of the parameter space to confirm $\mathcal{R}$.

1.3 Approach

1.3.1 Reduced-Basis Methods

The reduced-basis method is a technique for accurate, reliable and real-time prediction of functional outputs of parametrized PDEs, and is particularly relevant to the efficient treatment of the forward problem 1.1-1.2. The method has been applied to a wide variety of coercive and noncoercive linear equations [93, 121, 142, 141], linear eigenvalue equations [85], semilinear elliptic equations (including incompressible Navier-Stokes) [140, 99], as well as time-dependent equations [53, 54]. In this thesis, we shall provide further extension and new development of the method for: (1) noncoercive problems in which a weaker stability statement poses seriously numerical difficulties, (2) globally nonaffine problems where differential operators do not admit either an affine decomposition or a locally nonaffine dependence, (3) nonlinear problems where highly nonlinear operators are also of our interest. In this section, we briefly review three basic components of the method.

Reduced-Basis Approximation

Recognizing that the field variable is not an arbitrary member of the truth approximation space X, and that rather than it evolves in a low-dimensional manifold induced by the parametric dependence, the reduced-basis method constructs a reduced-basis approximation space to the manifold and seeks approximations to the field variable and output in that space. Essentially, we introduce nested sample, $S_N = \{\mu_1 \in \mathcal{D}, \cdots, \mu_N \in \mathcal{D}\}, 1 \le N \le N_{\max}$ and associated Lagrangian reduced-basis space as $W_N = \text{span}\{\zeta_j \equiv u(\mu_j), 1 \le j \le N\}, 1 \le N \le N_{\max}$, where $u(\mu_j)$ is the solution to (1.2) for $\mu = \mu_j$. Next we consider a standard Galerkin projection

$$a(u_N, v; \mu) = f(v), \quad \forall\, v \in W_N \ , \tag{1.14}$$

from which an $N \times N$ linear system for the coefficients $u_{N\,j}, 1 \le j \le N$, is derived

$$\sum_{j=1}^{N} a(\zeta_j, \zeta_i; \mu) u_{N\,j} = f(\zeta_i), \quad i = 1, \ldots, N \ . \tag{1.15}$$

The reduced-basis approximations to solution and output can then be calculated as $u_N(\mu) = \sum_{i=1}^{N} u_{N\,i}\zeta_i$ and $s_N(\mu) = \ell(u_N(\mu))$, respectively.

Typically [85, 121], and in some cases provably [93], $u_N(\mu)$ (respectively, $s_N(\mu)$) converges to $u(\mu)$ (respectively, $s(\mu)$) uniformly and extremely rapidly and thus we may achieve the desired accuracy for $N \ll \mathcal{N}$. Sufficient accuracy can thus be obtained with only $N = O(10) - O(100)$ degrees of freedom.

A Posteriori Error Estimation

Despite its rapid and uniform convergence rates, without *a posteriori* error estimation the reduced-basis approximation $u_N(\mu)$ *raises* many more questions than it *answers*. Is there even a solution $u(\mu)$ near $u_N(\mu)$? This question is particularly crucial in the nonlinear context — for which in general we are guaranteed neither existence nor uniqueness. Is $|s(\mu) - s_N(\mu)| \leq \epsilon^s_{\text{tol}}$, where ϵ^s_{tol} is the maximum acceptable error? Is a crucial feasibility condition $s(\mu) \leq C$ (in, say, a constrained optimization exercise) satisfied — not just for the reduced-basis approximation, $s_N(\mu)$, but also for the "true" output, $s(\mu)$? If these questions can not be affirmatively answered, we may propose the wrong — and potentially unsafe or infeasible — action. A fourth question is also important: Is N too large, $|s(\mu) - s_N(\mu)| \ll \epsilon^s_{\text{tol}}$, with an associated steep (N^3) penalty on computational efficiency? In this case, an overly conservative approximation may jeopardize the real-time response and associated action. Do we satisfy our global "acceptable error level" condition, $|s(\mu) - s_N(\mu)| \leq \epsilon^s_{\text{tol}}$, $\forall\, \mu \in \mathcal{D}$, for (close to) the smallest possible value of N? If the answers are not affirmative, then our reduced-basis approximations are more expensive (and unstable) than necessary — and perhaps too expensive to provide real-time response.

It is therefore critical that we can *rigorously* and *sharply* bound (a posteriori) the approximation errors. In fact, in this thesis, we pay great attention to the development of procedures for obtaining inexpensive error bounds $\Delta_N(\mu)$ and $\Delta^s_N(\mu)$ such that

$$\|u(\mu) - u_N(\mu)\|_X \leq \Delta_N(\mu); \qquad |s(\mu) - s_N(\mu)| \leq \Delta^s_N(\mu)\ . \tag{1.16}$$

For efficiency, we must also require $\Delta_N(\mu)$ and $\Delta^s_N(\mu)$ are sharp bounds.

Offline–Online Computational Procedure

The remaining question we need to address is that can we calculate $s_N(\mu), \Delta_N(\mu), \Delta_N^s(\mu)$ inexpensively? To this end, we decompose the computational effort into two stages: an expensive (offline) stage performed once; and an inexpensive (online) stage performed many times. The operation count for the online stage — in which, given a new value of the input, we calculate $s_N(\mu)$, $\Delta_N(\mu)$, and $\Delta_N^s(\mu)$ — depends only on N (typically very small) and the parametric complexity of the operator. This very low marginal cost is critical in the inverse-problem context.

1.3.2 Robust Real-time Inverse Computational Method

The reduced-basis method overcomes the deficiency of the classical approaches by providing real-time prediction $s_N(\mu)$ that is *certifiably* as good as the classical truth approximation $s(\mu)$ *but literally several orders of magnitude less expensive*. The method is endowed with three basic features that account for its superiority over other competing methods. First, with regard to *accuracy*, the unform and rapid convergence of the reduced-basis approximation is facilitated by exploiting the low-dimensional structure and smoothness of the solution manifold and by choosing the optimal approximation space. Second, with regard to *reliability*, *a posteriori* error procedures for several classes of PDEs are developed to quantify the error introduced by the reduced-basis approximation. Third, with regard to *efficiency*, online complexity is independent of the dimension of the finite element truth approximation space.

These advantages are further magnified within the inverse-problem context in which thousands of output predictions are often required effectively in real-time. In particular, based on the reduce-basis method we develop a robust inverse computational method for very fast solution region of inverse problems characterized by parametrized PDEs. The essential innovations are threefold: first, we apply the reduce-basis method to the forward problem for the rapid certified evaluation of PDE input-output relations and associated rigorous error bounds; second, we incorporate the reduced-basis approximation and error bounds into the inverse problem formulation; and third, rather than strive for only one regularized inverse solution, we may instead identify all (or almost all, in the

probabilistic sense) inverse solutions consistent with the available experimental data. Ill-posedness is captured in a bounded "possibility region" that furthermore shrinks as the experimental error is decreased. Hence, not only can we rigorously accommodate numerical uncertainty, but also robustly accommodate model uncertainty. Moreover, our inverse computational method enables real-time responses in several (admittedly rather simple) contexts: nondestructive evaluation of crack and material damage in a thin plate in Chapter 9 and inverse scattering analysis of elliptical "mines" in Chapter 10.

1.4 Literature Review

1.4.1 Reduced-Basis Method

The reduced-basis method has first been introduced in the late 1970s [4, 101] for single-parameter problems in nonlinear structural analysis, further extended by Noor [102] for multi-parameter problems, and subsequently developed more broadly [45, 116, 113] to include *a priori* error analysis. Much of the earlier work focused on the efficiency and accuracy of the approach through local approximation space. Consequently, the computational gain compared to conventional numerical methods are modest. In [9, 65], global approximation space spanned by solutions of the governing partial differential equation at globally sampled points in the parameter space was suggested; accuracy and efficiency have been much improved. Nevertheless, at the time no rigorous error analysis especially *a posteriori* error estimation has been proposed to certify the approximation error.

Recently, Patera et. al. [109, 85, 93, 121, 143, 122, 104, 142, 141, 140, 99, 14, 53] have greatly developed and brought in the technique with several useful insights and new features which differ from the earlier efforts in several important ways: as stated [143] "first, we develop global approximation spaces; second, we introduce rigorous a posteriori error estimators; and third, we exploit off-line/on-line computational decompositions". In particular, Maday et. al. [93, 94] presented a first theoretical *a priori* convergence that demonstrates uniform exponential convergence of the reduced-basis approximation. Machiels et. al. [85, 86] developed the method for affine-parameter coercive elliptic linear and eigenvalue problems. In [131], Rovas analyzed the technique in great detail

and extended it more broadly for many different classes of parametrized partial differential equations including noncoercive elliptic and parabolic linear problems. Veroy [139, 143, 142] generalized the concept of bound conditioners relevant to the *a posteriori* error estimation. In her work, several bound conditioners was developed to yield rigorous and sharp error estimators. The use of the reduced-basis method for quadratically nonlinear problems — the steady incompressible Navier-Stokes equations — can be found in other work of Veroy [141, 140]. Also, Solodukhov [135] proposed several reduced-basis strategies for locally nonaffine problems and nonlinear problems.

Reduced-basis method has also found its applications in many areas such as nonlinear structural analysis [101, 45, 116], fluid flow problems [113, 65, 66], bifurcation and post-buckling analysis of composite plates [102], and nonlinear steady-state thermal analysis [100]. With regard to the current developments, Machiels et. al. [87] used the technique to find the optimal shape of a thermal fin. Ali [3] combined the method with assess-predict-optimize strategy to obtain the "best" worst case scenarios of a system under design in the presence of data uncertainty. In [51, 54], Grepl proposed new error estimation methods for linear and nonlinear time-dependent problems and applied the technique to adaptive (real-time) optimal control.

1.4.2 Model Order Reduction

Generally, there are three approaches in model-order reduction (MOR): 1) MOR algorithms based on Krylov subspace methods, 2) techniques using Karhunen-Loeve expansion (or Proper Orthogonal Decomposition), 3) methods based on Hankel norm approximants and balanced truncation. A driving force behind the development of MOR approaches is the need for efficient simulation tools for dynamical (time-varying) systems arising in circuit simulation, structural dynamics and micro-electro-mechanical systems. The basic and common idea applied by all of these approaches is a projection from high-dimensional state space to very low dimensional state space, which in turn produces the reduced-order model of the original system.

The Proper Orthogonal Decomposition (POD) has been used widely to obtain low dimensional dynamical models of many applications in engineering and science. The idea

is to start with an ensemble of data, called *snapshots*, collected from the experiment or a numerical procedure of physical systems. POD technique is then used to produce a set of basis functions from the snapshot collection and in turn implicitly captures the dominant dynamics of a system [134]. A model of reduced complexity is finally generated by the application of Galerkin projection onto a subspace spanned by these basis functions. The method has been widely used to obtain the reduced-order model of many large-scale linear dynamical systems: computational fluid dynamics [57, 130], fluid-structure interaction [37], turbo-machinery flows [148, 149], optimal control of fluid flows [84, 126]. Recently, there has been a rapidly growing number of researches into the application of POD for nonlinear systems [72], nonlinear structural dynamics [96], nonlinear MEMS devices [26]. The POD has been also used to develop reduced-order models for parametric applications such as turbomachinery flows with sampling in both time and over a range of inter-blade phase angles [43] and inverse design of transonic airfoils [22].

Over past years, a great deal of attention has been also devoted to Krylov subspace-based methods for efficient modeling, effective realization, and fast simulation of system dynamics. The basic idea is to approximate the transfer function of original systems by generating a subspace spanned by orthogonal basis functions and projecting original systems onto that subspace [55]. Owing to their robustness and low computational cost, the Krylov subspace-based methods have proved very attractive for producing reduced-order model of many large-scale linear systems and have been broadly used in engineering applications: structural dynamics [7], optimal control of fluid flows [75], circuit design [28, 50], turbomachinery [150]. A number of linear MOR techniques based on Krylov subspace have been extended to deal with weakly nonlinear problems [27, 114]. The merit idea is to be able to represent the nonlinearity with a simplified form that can be treated with standard linear or bilinear MOR procedures. The simplest form of these approaches is to linearize or bilinearize multidimensional nonlinear functions using polynomial Taylor series expansion [115]. The trajectory piecewise-linear method [129] has been proposed to effectively obtain reduce-order models for highly nonlinear systems. In this approach, the nonlinear system is represented as a combination of linear models, generated at different linearization points in the state space about the state trajectory when driven by a fixed "training" input.

Finally, we have few general remarks concerning model order reduction techniques: first, reduced-order modeling to capture parametric variation based on interpolation procedure is somehow heuristic; second, due to lack of efficient representation of nonlinearity and fast exponential growth (with the degree of the nonlinear approximation order) of computational complexity in the reduced-order model, the development of model order reduction methods for nonlinear problems remains a continuous and open task; third, although a priori error bounds to quantify the error in the model reduction have been derived but only in the linear case, a posteriori error bounds have not been adequately considered yet even for the linear case in MOR approaches; and fourth, while most MOR techniques concentrate mainly on reduced-order modeling of time-variation systems, the development of reduced-order models for parametric applications is much less common.

1.4.3 *A Posteriori* Error Estimation

A posteriori error estimation has received enormous attention in the finite-element context where choice of mesh to define the finite-element approximation spaces becomes a trade-off between computational efficiency and accuracy: A very fine mesh ensures accuracy but implies high computational cost which is prohibitive to many important applications in engineering optimization and design; on the other hand, a relatively coarse mesh guarantees efficiency but the accuracy is uncertain. To minimize the computational effort while honoring the desired accuracy, we must provide *a posteriori* rigorous, accurate error bounds for the discretization error. *A posteriori* error estimation technique for finite element discretization of partial differential equations was first introduced in the 1970s [6], and subsequently extended [74, 10, 1] to a broader class of partial differential equations. Most finite element *a posteriori* error estimation procedures developed measure the error in the energy, H^1, or L^p norms. However, more relevant to engineering purposes is the prediction of the bounds for the engineering output of interest (typically articulated as a functional of field variables). In [17, 18], Becker and Rannacher proposed a refined approach to residual-based error estimation: in essence, the residual estimators — based on *a priori* stability and approximation constants and presumed-exact mesh adjoint functions — enable rapid evaluation for adaptive refinement, but not necessarily accurate and

rigorous quantification for the quantities of interest. In [107, 110, 105, 106], Paraschivoiu et. al. introduced an *a posteriori* finite element method for the efficient computation of strict upper and lower bounds for linear-functional outputs of coercive partial differential equations. The methods are thus appropriate for both confirmation of accuracy and mesh adaptivity. The bound techniques are further extended to treat a variety of different problems — including the Helmholtz equations, the Burger equations, [111], eigenvalue problems [90], as well as nonlinear equations (incompressible Navier-Stokes) [89].

Some of ideas of a posteriori error estimation in the finite element context have been then used in the reduced-basis approximations of parametrized partial differential equations. Even though the methodologies are distinctively different, general ideas for *a posteriori* error estimation are common. In [85], [121], and [143, 141, 142], Patera et. al. introduced a family of rigorous error estimators for reduced-basis approximation of a wide variety of partial differential equations. In this thesis, we will continue this theme to develop *a posteriori* error bounds for noncoercive linear elliptic equations, nonaffine linear elliptic equations, and highly nonlinear monotonic elliptic equations.

1.4.4 Computational Approaches in Inverse Problems

Inverse problems are typically formulated as an appropriate minimization for the difference between computed outputs and measured-observable outputs. In this framework, the forward problem is taken as an additional set of constraints. One approach to the (PDE-constrained) optimization-based inverse problems combines PDE discretization techniques such as the finite element method, boundary element method, and finite volume method with optimization procedures. In solving inverse problems by this approach, the forward problem has to be solved several times, as it is required by the algorithm used to solve the optimization problem. Unfortunately, solution of the forward problem by classical numerical methods is typically time-consuming due to (very) large approximation spaces required to achieve a desired accuracy. Furthermore, the above minimization may be appropriate for inverse problems with noise-free data, but it may fail to give accurate solutions for ill-posed inverse problems whose ill-posedness comes from data uncertainty. As a result, a regularization factor and associated regularization

parameter reflecting the uncertainty are added to the minimization as a way to ensure fairly accurate solutions; this leads to minimize the so-called Tikhonov functional [137]. The basic idea here is to replace the original ill-posed inverse problem with a family of nearby well-posed problems by taking explicitly the uncertainty into the optimization problem. The solution method for the Tikhonov functional minimization and choice of regularization parameter (by which the solutions of inverse problems will be certainly affected) is an important issue. Generally, there are two general rules of thumbs namely *a priori* [39] and *a posteriori* choice [137, 112] for determining the regularization parameter.

Another approach to capture the uncertainty into the inverse formulation is by means of statistics. In [11], the authors suggested to treat the estimated parameter as a random variable with unknown distribution and then reformulate the deterministic parameter estimation problem into a problem of estimation of a random variable using sampled data from a dynamical system which depends on the parameter. In [3], the uncertain data is analyzed and incorporated into the optimization by continuously monitoring the propagation of the error via assess-predict-optimize strategy. Though, in fact, the technique was intended for optimal parametric design, it can be efficiently used to solve the inverse problems with uncertainty.

Of course, optimization techniques for solution of optimization-based inverse problems are rich. Global heuristic optimization strategies such as neural networks, simulated annealing, and genetic algorithms have powerful ability in finding optimal solutions for general nonconvex problems. Liu et. all. [79, 80, 151] developed the projection genetic algorithm which requires fewer number of generations to converge than ordinary genetic algorithms and used it for detecting cracks and assessing damage in composite material. The use of neural networks for inverse problems can be found in [152, 58]. However, the problem with these approaches is that they are heuristic by nature and computationally expensive. Therefore, gradient methods like Newton's method [8, 20, 19, 71], descent methods [60, 125], and current state-of-the-art interior-point method [24] have been employed to solve inverse problems in many cases. In [103, 104], Oliveira and Patera incorporate reduced-basis approximations and associated *a posteriori* error bounds into scaled trust-region interior-point method for the rapid reliable optimization of systems described by parametric partial differential equations. The very low marginal cost of

forward evaluation and the ability to correctly avoid non-optimal stationary points, and hence provide true (at least local) optimizers, make this method attractive.

Finally, we review particular computational methods for nondestructive testing and inverse scattering problems, which are two major applications of inverse problems and shall be considered in this thesis. With regard to nondestructive testing, Liu et. al. introduced a strip element method [77] and subsequently extended to investigate the scattering of waves by cracks and detect the cracks in anisotropic laminated plates [78, 146, 83]. The method has been also used to investigate wave scattering by rectangular flaws and assess material loss in sandwich plates [145, 79]. We also notice that various computational methods have been proposed for solving the inverse scattering problems — including the Newton-like methods [73, 19, 119, 44], linear sampling method [34, 30, 32], and point-source method [117, 118]. The linear sampling method allows to determine from measured scattering data whether or not a given point is inside a scattering object by solving a linear integral equation of the first kind, which leads to the reconstruction of the entire object. A particular advantage of the method is that it does not require *a priori* knowledge of either the boundary condition or the connectivity of the scattering obstacle. On the other hand, the method is restricted to the situation where all the data are at the same frequency and require multi-static data. In point-source method, the idea is to reconstruct the total field on parts of a domain from a finite number of measurements of the far-field pattern; the reconstructed total field and the boundary condition for a sound-soft object in the domain can be then used to find the location and shape of the object. The main impediment to the method is that the error bounds on the field reconstruction only hold at points on the exterior of the scatter, for a suitable exterior cone condition and domain of approximation. Furthermore, in locating the boundary of the scatterer from the reconstructed field, it is also required to identify points that are not on the boundary due to the error in the field reconstruction.

As a summary, there are a wide variety of techniques for solving inverse problems. However, in almost cases the inverse techniques are expensive due to the following reasons: solution of the forward problem by classical numerical approaches is typically long; associated optimization problems are usually nonlinear and nonconvex; and most importantly, inverse problems are typically ill-posed. Ill-posedness is traditionally addressed

by regularization. Though quite sophisticated, iterative regularization methods are quite expensive (often fail to achieve numerical solutions in real-time) and often need additional information and thus lose algorithmic generality (in many cases, do not well quantify uncertainty). Furthermore, in the presence of uncertainty, solution of the inverse problem should never be unique at least in terms of mathematical sense; there should be indefinite inverse solutions that are consistent with model uncertainty. However, most inverse techniques provide only one inverse solution among the universal; and hence they do not exhibit and characterize ill-posed structure of the inverse problem.

1.5 Thesis Outline

The two central themes in this thesis are the development of the reduced-basis method for parametrized partial differential equations and its application to inverse problems in engineering and science. Before proceeding with the development of reduced-basis method, we overview in the next chapter some relevant mathematical background that will be used frequently throughout the thesis. In Chapter 3, we present basic concepts of the reduced-basis method via applying the technique to a heat conduction problem which will serve to illustrate essential components and key ideas of the method. In Chapter 4, we propose an approach to the construction of rigorous and efficient lower bound — a critical component of our *a posteriori* error estimators — for the critical stability factor. In subsequent chapters, we develop reduced-basis approximations and *a posteriori* error estimators for different classes of problems: in Chapter 5 for noncoercive linear problems, in Chapter 6 for nonaffine linear problems, and in Chapter 7 for nonlinear problems. Based on the reduced-basis approximations and associated *a posteriori* error estimators of parametrized partial differential equations, we present in Chapter 8 the development of a certified real-time computational inverse method for very fast solution region of inverse problems even in the presence of significant uncertainty. In the subsequent two chapters, we apply our computational inverse method to two important applications of inverse problems: in Chapter 9 for crack detection and damage assessment of flawed materials and in Chapter 10 for inverse scattering analysis. Finally, we conclude in Chapter 11 with summary of the thesis and some suggestions for future work.

Chapter 2

Building Blocks

Before proceeding with the development of reduced-basis method, we review some relevant mathematical background ("building blocks") that will be used repeatedly in the remaining chapters of the thesis. First, we review basic elements of the functional analysis which will prove useful in deriving our weak statements of partial differential equations. Second, we review only essential concepts from differential geometry that have direct use in this thesis. Third, we review fundamental equations of linear elasticity which appear frequently in several chapters. Finally, we review the finite element method which is used for the numerical solution of continuum mechanics problems discussed in the thesis.

2.1 Review of Functional Analysis

In this section, we introduce some basic concepts of functional analysis that will be used throughout in the thesis and refer to [76, 128] for a good introduction and specific details on the topic. To begin, let $\Omega \subset \mathbb{R}^d$, $d = 1, \ldots, 3$, be an open domain with Lipschitz-continuous boundary Γ.

2.1.1 Function Spaces

Linear Spaces

Definition 1. *Let* $\mathbb{K}$ *be an* algebraic field*, where* $\mathbb{K}$ *is either* $\mathbb{R}$ *(real field) or* $\mathbb{C}$ *(complex field). A* linear vector space X *over the field* $\mathbb{K}$ *is a set of elements together with two*

operations, $u, v \in X : u + v \in X$ *(addition) and* $\alpha \in \mathbb{K}, v \in X, v \in X : \alpha v \in X$ *(scalar multiplication), if the following axioms hold*

(1) $u + v = v + u$ *(commutative) ;*

(2) $(u + v) + w = u + (v + w)$ *(associative) ;*

(3) $\exists \mathbf{0}$ *such that* $u + \mathbf{0} = u$ *for all* $u \in X$ *(null vector) ;*

(4) For each $u \in X$, $\exists - u \in X$ *such that* $u + (-u) = \mathbf{0}$ *(additive inverse vector) ;*

(5) $(\alpha\beta)u = \alpha(\beta u)$ *(associative) ;*

(6) $(\alpha + \beta)u = \alpha u + \beta u$ *(distributive) ;*

(7) $\alpha(u + v) = \alpha u + \alpha v$ *(distributive) ;*

(8) $1u = u$.

Norm

Definition 2. *A function* $\| \cdot \|_X$ *from a linear space* X *into* $\mathbb{K}$ *is called a* norm *if and only if it has the following properties*

(i) (a) $\|u\|_X \geq 0$, *and (b)* $\|u\|_X = 0$ *if and only if* $u = 0$ *;*

(ii) $\|\alpha u\|_X = |\alpha| \|u\|_X, \forall\, \alpha \in \mathbb{R}$ *;*

(iii) $\|u + v\|_X \leq \|u\|_X + \|v\|_X, \forall\, u, v \in X$ *(the triangle inequality) .*

If $\|u\|_X$ *satisfies (ia), (ii), and (iii) only, we call it a* seminorm *of the vector* u*, and denote it by* $|u|$*. A linear vector space* X *together with a norm defined on itself is a* normed space*.*

Inner Product

Definition 3. *Let* X *be a linear space over the field* $\mathbb{K}$*. An* inner product *on* X *is a scalar valued function on* $X \times X$*, whose values are denoted by* $(u, v)_X$*, that satisfies the following axioms*

(i) $(u, v)_X = \overline{(v, u)}_X, \forall\, u, v \in X$ *(symmetric);*

(ii) $(u, u)_X \geq 0, \forall\, u \in X$ *and* $(u, u)_X = 0$ *if and only if* $u = 0$ *(positive definite);*

(iii) $(u + v, w)_X = (u, w)_X + (v, w)_X, \forall\, u, v \in X$ *and* $(\alpha u, v)_X = \alpha(u, v)_X, \forall\, u, v \in X, \forall\, \alpha \in \mathbb{K}$ *(bilinear).*

A linear vector space X *on which an inner product can be defined is called an* inner product space*. (Note that one can associate a norm with every inner product by defining*

$\|u\| = \sqrt{(u,u)_X}$.)

Spaces of Continuous Functions

Definition 4. *For a nonnegative integer m, we define the set of real functions with continuous derivatives up to and including order m as*

$$C^m(\Omega) = \{v|\ D^\alpha v \text{ is uniformly continuous and bounded on } \Omega,\ 0 \le |\alpha| \le m\}\ , \quad (2.1)$$

with an associated norm

$$\|v\|_{C^m(\Omega)} = \max_{0\le|\alpha|\le m} \sup_{x\in\Omega} |D^\alpha v(x)|, \quad (2.2)$$

where α denotes n-tuple of nonnegative integers, $\alpha = (\alpha_1, \ldots, \alpha_d)$, and

$$D^\alpha = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}, \quad |\alpha| = \sum_{i=1}^{d} \alpha_i\ .$$

It is clear that $C^m(\Omega)$ defined above is a Banach space, i.e. a complete normed linear space. Also note that $C_0^m(\Omega)$ is the space of continuous, m^{th} differentiable functions with compact support, i.e. vanishing near the boundary of Ω. We shall use the subscript 0 to indicate spaces with functions of compact support.

Lebesgue Spaces

Definition 5. *For $1 \le p \le \infty$, we define the space of p^{th} integrable functions as*

$$L^p(\Omega) = \begin{cases} \left\{v|\ \displaystyle\int_\Omega |v|^p\, dx < \infty\right\},\ 1 \le p < \infty \\ \left\{v|\ \operatorname*{ess\,sup}_{x\in\Omega} |v(x)| < \infty\right\},\ p = \infty \end{cases}; \quad (2.3)$$

with an associated norm

$$\begin{aligned} \|v\|_{L^p(\Omega)} &= \left(\int_\Omega |v|^p\, dx\right)^{\frac{1}{p}},\ 1 \le p < \infty \\ \|v\|_{L^\infty(\Omega)} &= \operatorname*{ess\,sup}_{x\in\Omega} |v(x)|,\ p = \infty\ . \end{aligned} \quad (2.4)$$

These spaces are also Banach spaces. The ess sup (essential supremum) in the above definition means the smallest supremum over $\Omega \backslash B$ for all sets B of zero measure.

Hilbert Spaces

Definition 6. *For a non-negative integer m, we define the Hilbert Space $H^m(\Omega)$ as*

$$H^m(\Omega) = \left\{ v |\ D^\alpha v \in L^2(\Omega),\ \forall \alpha : |\alpha| \leq m \right\}; \tag{2.5}$$

with associated inner product

$$(w, v)_{H^m(\Omega)} = \sum_{|\alpha| \leq m} \int_\Omega D^\alpha w \cdot D^\alpha v \, dx \ , \tag{2.6}$$

and induced norm

$$\|v\|_{H^m(\Omega)} = \left(\sum_{|\alpha| \leq m} \int_\Omega |D^\alpha v|^2 \, dx \right)^{\frac{1}{2}} \ . \tag{2.7}$$

These spaces are important not only in understanding well-posedness of weak statements, but also in expressing the convergence rate of the finite element method. In addition, we introduce $H^m(\Omega)$ *semi-norm* as

$$|v|_{H^m(\Omega)} = \left(\int_\Omega |D^m v|^2 \, dx \right)^{\frac{1}{2}} \ , \tag{2.8}$$

which include only the m^{th} derivative.

Complex Hilbert Spaces

Definition 7. *For a non-negative integer m, we define the complex Hilbert Space $Z^m(\Omega)$ as*

$$Z^m(\Omega) = \left\{ v = v^R + \mathrm{i} v^I \mid v^R \in H^m(\Omega), v^I \in H^m(\Omega) \right\}; \tag{2.9}$$

with associated inner product

$$(w, v)_{Z^m(\Omega)} = \sum_{|\alpha| \leq m} \int_\Omega D^\alpha w \cdot D^\alpha \bar{v} \, dx \ , \tag{2.10}$$

and induced norm

$$\|v\|_{Z^m(\Omega)} = \left(\sum_{|\alpha| \leq m} \int_\Omega |D^\alpha v|^2 \, dx \right)^{\frac{1}{2}} . \tag{2.11}$$

Here and throughout this thesis superscript R and I denote the real and imaginary part, respectively, that is, $v^{\mathrm{R}} = \Re(v)$ and $v^{\mathrm{I}} = \Im(v)$; $\bar{v}$ and $|v|$ shall denote the complex conjugate and modulus of v, respectively.

Sobolev Spaces

Definition 8. *For $m > 0$ integer and $p > 1$, we define the Sobolev space $W^{m,p}(\Omega)$ as*

$$W^{m,p}(\Omega) = \begin{cases} \{v|\ D^\alpha v \in L^p(\Omega),\ \forall \alpha : |\alpha| \leq m\}\,,\ 1 \leq p < \infty \\ \{v|\ D^\alpha v \in L^\infty(\Omega),\ \forall \alpha : |\alpha| \leq m\}\,,\ p = \infty, \end{cases} \tag{2.12}$$

with associated norm

$$\begin{aligned} \|v\|_{W^{m,p}(\Omega)} &= \left(\sum_{|\alpha| \leq m} \int_\Omega |D^\alpha v|^p \, dx \right)^{\frac{1}{p}},\ 1 \leq p < \infty \\ \|v\|_{W^{m,\infty}(\Omega)} &= \max_{|\alpha| \leq m} \operatorname*{ess\,sup}_{x \in \Omega} |D^\alpha v(x)|,\ p = \infty. \end{aligned} \tag{2.13}$$

Note that $W^{m,2}(\Omega) = H^m(\Omega)$, our earlier Hilbert spaces; and that the Lebesgue space $L^p(\Omega)$ is a special case of $W^{m,p}(\Omega)$ for $m = 0$.

Dual Hilbert Spaces

In general, given a Hilbert space X, we can define the corresponding dual space X' as the space of all *bounded* linear functionals, $\ell(v)$, where $\ell(v)$ is bounded if $\ell(v) \leq C||v||_X, \forall v \in X$, for some positive C. The norm of $\ell(v)$ is given by

$$\|\ell\|_{X'} = \sup_{v \in X} \frac{\ell(v)}{\|v\|_z}. \tag{2.14}$$

We note that this space of functionals is a linear space; that a bounded linear functional is continuous; and that this space is also a Hilbert space and if $X = H^m(\Omega)$ we will

denote the dual space $X' = H^{-m}(\Omega)$. Generally, we have

$$H^m(\Omega) \subset \cdots \subset H^1(\Omega) \subset L^2(\Omega) \subset H^{-1}(\Omega) \subset \cdots \subset H^{-m}(\Omega).$$

Finally, from the Riesz representation theorem we know that for every $\ell \in X'$ there exists a $u_\ell \in X$ such that

$$(u_\ell, v)_X = \ell(v), \ \forall v \in X. \tag{2.15}$$

It follows that

$$\|\ell\|_{X'} = \sup_{v \in X} \frac{(u_\ell, v)_X}{\|v\|_X} = \|u_\ell\|_X. \tag{2.16}$$

2.1.2 Linear Functionals and Bilinear Forms

Linear Functionals

Let X be a linear space over the field $\mathbb{K}$ which is either $\mathbb{R}$ (real field) or $\mathbb{C}$ (complex field), a linear transformation ℓ of X into $\mathbb{K}$ is called a linear functional if and only if it satisfies

$$\ell(\alpha u + \beta v) = \alpha \ell(u) + \beta \ell(v), \quad \forall\, u, v \in X, \forall\, \alpha, \beta \in \mathbb{K}\ .$$

The set of all linear functionals on a linear space X is itself a vector space, called the *dual space* of X and denoted by X'.

Bilinear Forms

Let X and Y be two linear spaces over the field $\mathbb{K}$, an operator $a : X \times Y \to \mathbb{K}$ that maps $(u, v), u \in X, v \in Y$ into $\mathbb{K}$ is called a bilinear form if and only if it satisfies

$$a(\alpha u_1 + \beta u_2, \gamma v_1 + \lambda v_2) = \alpha\bar{\gamma} a(u_1, v_1) + \alpha\bar{\lambda} a(u_1, v_2) + \beta\bar{\gamma} a(u_2, v_1) + \beta\bar{\lambda} a(u_2, v_2)$$

for all $u_1, u_2 \in X, v_1, v_2 \in Y, \alpha, \beta, \gamma, \lambda \in \mathbb{K}$. A bilinear form $a : X \times X \to \mathbb{K}$ is said to be *symmetric* if $a(u, v) = \overline{a(v, u)}, \forall u, v \in X$.

2.1.3 Fundamental Inequalities

Cauchy-Schwarz Inequality

Let $a : X \times X \to \mathbb{K}$ be a symmetric semi-definite bilinear form. Then a satisfies the Cauchy-Schwarz inequality

$$|a(u,v)| \leq \sqrt{a(u,u)}\sqrt{a(v,v)}, \tag{2.17}$$

for all $u, v \in X$.

Hölder Inequality

If $\frac{1}{p} + \frac{1}{q} = 1, 1 < p < \infty$ then for all $u \in L^p(\Omega), v \in L^q(\Omega)$, we have

$$\|uv\|_{L^1(\Omega)} \leq \|u\|_{L^p(\Omega)}\|v\|_{L^q(\Omega)}. \tag{2.18}$$

Minkowski Inequality

If $1 \leq p \leq \infty$ then for all $u, v \in L^p(\Omega)$, we have

$$\|u \pm v\|_{L^p(\Omega)} \leq \|u\|_{L^p(\Omega)} + \|v\|_{L^p(\Omega)}. \tag{2.19}$$

Friedrichs Inequality

Let Ω be a domain with a Lipschitz boundary Γ, and let Γ_1 be its open part with a positive Lebesgue measure. Then there exists a positive constant $c > 0$, depending only on the given domain and on Γ_1 such that for every $u \in H^1(\Omega)$, we have

$$\|u\|^2_{H^1(\Omega)} \leq c \left\{ \sum_j \int_\Omega \left(\frac{\partial u}{\partial x_j} \right)^2 + \int_{\Gamma_1} |u|^2 \right\} . \tag{2.20}$$

Also, for $u \in H^2(\Omega)$, we have

$$\|u\|^2_{H^2(\Omega)} \leq c(\Omega) \left\{ \sum_{|\alpha| \leq 2} \int_\Omega |D^\alpha u|^2 + \int_\Gamma |u|^2 \right\} . \tag{2.21}$$

Note that for $u \in H_0^m(\Omega)$, the two inequalities hold without the boundary terms.

Poincaré Inequality

Let Ω be a domain with a Lipschitz boundary Γ, and let Γ be its open part with a positive Lebesgue measure. Then there exists a positive constant $c > 0$ such that for all $u \in H^m(\Omega)$

$$\|u\|^2_{H^m(\Omega)} \leq c(\Omega) \left\{ \sum_{|\alpha| \leq m} \int_\Omega |D^\alpha u|^2 + \sum_{|\alpha| < m} \left(\int_\Omega |D^\alpha u| \right)^2 \right\} . \tag{2.22}$$

2.2 Review of Differential Geometry

2.2.1 Metric Tensor and Coordinate Transformation

In an arbitrary (possibly curvilinear) three-dimensional coordinate system x^i $(i = 1, 2, 3)$, at any point A we choose three vectors $\mathbf{g}_i$ of such dimension and magnitude that the line element vector can be expressed

$$d\mathbf{s} = \sum_i \mathbf{g}_i dx^i = \mathbf{g}_i dx^i \ . \tag{2.23}$$

Here for simplicity of notation we use the summation convention: when the same Latin letter (say i) appears in a product once as a superscript and once as a subscript, that means a sum of all terms of this kind.

Now we consider a fixed point O (possibly the origin of the coordinate system) and a position vector $\mathbf{r}$ leading from O to A; the line element $d\mathbf{s}$ is the increment of $\mathbf{r}$ ($d\mathbf{s} = d\mathbf{r}$), which can be written as

$$d\mathbf{r} = \frac{\partial \mathbf{r}}{\partial x^i} dx^i \ . \tag{2.24}$$

From (2.23) and (2.24), we have

$$\mathbf{g}_i = \frac{\partial \mathbf{r}}{\partial x^i} \ . \tag{2.25}$$

The vectors $\mathbf{g}_i$ are called *covariant base vectors*. It follows from (2.23)-(2.25) that

$$ds^2 = (\mathbf{g}_i \cdot \mathbf{g}_j)\, dx^i dx^j = \left(\frac{\partial \mathbf{r}}{\partial x^i} \cdot \frac{\partial \mathbf{r}}{\partial x^j}\right) dx^i dx^j = g_{ij} dx^i dx^j \ , \tag{2.26}$$

where

$$g_{ij} = \frac{\partial \mathbf{r}}{\partial x^i} \cdot \frac{\partial \mathbf{r}}{\partial x^j} = \mathbf{g}_i \cdot \mathbf{g}_j \ . \tag{2.27}$$

The entity of the nine quantities g_{ij} defined above is call the *metric tensor*. Note that in the Cartesian coordinate system $g_{ij} = \delta_{ij}$, where δ_{ij} is the Kronecker delta symbol — $\delta_{ij} = 1$ if $i = j$, otherwise $\delta_{ij} = 0$. We next find nine quantities g^{ij} that satisfy

$$g_{ik} g^{jk} = \delta_i^j \ . \tag{2.28}$$

The entity of such nine quantities g^{ij} is call the *conjugate metric tensor*. Here δ_i^j is just another way of writing the Kronecker symbol δ_{ij}.

Now consider a new coordinate system $\bar{x}^i (i = 1, 2, 3)$ and associated base vectors $\bar{\mathbf{g}}_i$, we define a coordinate transformation from x^i to $\bar{x}^i$ by a set of transformation rules $x^i = x^i\left(\bar{x}^1, \bar{x}^2, \bar{x}^3\right), i = 1, 2, 3$. We then differentiate the relation to obtain

$$dx^i = \frac{\partial x^i}{\partial \bar{x}^j} d\bar{x}^j \ . \tag{2.29}$$

The partial derivatives are obtained from the chain rule

$$\frac{\partial}{\partial x^i} = \frac{\partial \bar{x}^j}{\partial x^i} \frac{\partial}{\partial \bar{x}^j} \ . \tag{2.30}$$

The Jacobian of the transformation is given by

$$J = \begin{vmatrix} \frac{\partial x^1}{\partial \bar{x}^1} & \frac{\partial x^1}{\partial \bar{x}^2} & \frac{\partial x^1}{\partial \bar{x}^3} \\ \frac{\partial x^2}{\partial \bar{x}^1} & \frac{\partial x^2}{\partial \bar{x}^2} & \frac{\partial x^2}{\partial \bar{x}^3} \\ \frac{\partial x^3}{\partial \bar{x}^1} & \frac{\partial x^3}{\partial \bar{x}^2} & \frac{\partial x^3}{\partial \bar{x}^3} \end{vmatrix} \ . \tag{2.31}$$

Similarly, in the new coordinate system we have $d\mathbf{r} = \bar{\mathbf{g}}_j d\bar{x}^j$. It directly follows from

(2.24), (2.25), and (2.29) that

$$\bar{\mathbf{g}}_i = \mathbf{g}_j \frac{\partial x^j}{\partial \bar{x}^i} \ . \tag{2.32}$$

2.2.2 Tangent Vectors and Normal Vectors

Curve

For a three-dimensional parametrized curve $x^i = x^i(s)$ in a generalized coordinate system with matrix tensor g_{ij} and arc length parameter s, the vector $\mathbf{T} = (T^1, T^2, T^3)$, with $T^i = \frac{dx^i}{ds}$, represents a *tangent vector* to the curve at a point P on the curve. The vector $\mathbf{T}$ is a unit vector because

$$\mathbf{T} \cdot \mathbf{T} = g_{ij} T^i T^j = g_{ij} \frac{dx^i}{ds} \frac{dx^j}{ds} = 1 \ . \tag{2.33}$$

Differentiating (2.33) with respect to s, we obtain

$$g_{ij} T^j \frac{dT^i}{ds} = 0 \ . \tag{2.34}$$

Hence, the vector $\frac{d\mathbf{T}}{ds}$ is perpendicular to the tangent vector $\mathbf{T}$. We now normalize it to get the unit normal vector $\mathbf{N}$ to the curve as

$$N^i = \frac{1}{\kappa} \frac{dT^i}{ds} \ ; \tag{2.35}$$

here κ, a scale factor called *curvature*, is determined such that $g_{ij} N^i N^j = 1$.

Surface

For our purpose here, we shall consider Cartesian frame of reference (x, y, z) with associated base vectors $\mathbf{i}_x, \mathbf{i}_y, \mathbf{i}_z$; see [47] for formulations in a generalized coordinate system. A surface in three-dimensional Euclidean space can be defined in three different ways: explicitly $z = f(x, y)$, implicitly $F(x, y, z) = 0$, or parametrically $x = x(u, v), y = y(u, v), z = z(u, v)$ which contains two independent parameters u, v called *surface coordinates*. Using the parametric form of a surface, we can define the position vector to a

point P on the surface as

$$\mathbf{r} = x(u,v)\mathbf{i}_x + y(u,v)\mathbf{i}_y + z(u,v)\mathbf{i}_z \ . \tag{2.36}$$

A square of the line element on the surface coordinates is given by

$$ds^2 = d\mathbf{r}\cdot d\mathbf{r} = \frac{\partial \mathbf{r}}{\partial u^\alpha}\frac{\partial \mathbf{r}}{\partial u^\beta}du^\alpha du^\beta = a_{\alpha\beta}du^\alpha du^\beta, \quad \alpha,\beta = 1,2 \ . \tag{2.37}$$

In differential geometry, this expression is known as the *first fundamental form*, and $a_{\alpha\beta}$ is called surface metric tensor and given by

$$a_{\alpha\beta} = \frac{\partial \mathbf{r}}{\partial u^\alpha}\frac{\partial \mathbf{r}}{\partial u^\beta}, \quad \alpha,\beta = 1,2 \ , \tag{2.38}$$

with conjugate metric tensor $a^{\alpha\beta}$ defined such that $a^{\alpha\beta}a_{\alpha\gamma} = \delta^\gamma_\alpha$.

Furthermore, the *tangent plane* to the surface at point P can be represented by two basic *tangent vectors*

$$\mathbf{T}_u = \frac{\partial \mathbf{r}}{\partial u}, \qquad \mathbf{T}_v = \frac{\partial \mathbf{r}}{\partial v} \tag{2.39}$$

from which we can construct a unit normal vector to the surface at point P as

$$\mathbf{N} = \frac{\mathbf{T}_u \times \mathbf{T}_v}{|\mathbf{T}_u \times \mathbf{T}_v|} \ . \tag{2.40}$$

If we transform from one set of curvilinear coordinates (u,v) to another set $(\bar{u},\bar{v})$ with the transformation laws $u = u(\bar{u},\bar{v}), v = v(\bar{u},\bar{v})$, we can then derive the tangent vectors for the new surface coordinates from the chain rule (2.30)

$$\frac{\partial \mathbf{r}}{\partial \bar{u}} = \frac{\partial \mathbf{r}}{\partial u}\frac{\partial u}{\partial \bar{u}} + \frac{\partial \mathbf{r}}{\partial v}\frac{\partial v}{\partial \bar{u}} \quad \text{and} \quad \frac{\partial \mathbf{r}}{\partial \bar{v}} = \frac{\partial \mathbf{r}}{\partial u}\frac{\partial u}{\partial \bar{v}} + \frac{\partial \mathbf{r}}{\partial v}\frac{\partial v}{\partial \bar{v}} \ ; \tag{2.41}$$

from which the associated normal unit vector can be readily defined.

2.2.3 Curvature

We first note from the differentiation of the unit normal vector $\mathbf{N}$ and position vector $\mathbf{r}$ to define the quadratic form

$$d\mathbf{r} \cdot d\mathbf{N} = \left(\frac{\partial \mathbf{r}}{\partial u}du + \frac{\partial \mathbf{r}}{\partial v}dv\right) \cdot \left(\frac{\partial \mathbf{N}}{\partial u}du + \frac{\partial \mathbf{N}}{\partial v}dv\right) = -b_{\alpha\beta}du^{\alpha}du^{\beta} \ . \tag{2.42}$$

In differential geometry, this equation is known as the *second fundamental form*; and $b_{\alpha\beta}$, called the *curvature tensor* of the surface, are given by

$$b_{11} = -\frac{\partial \mathbf{r}}{\partial u}\frac{\partial \mathbf{N}}{\partial u}, \quad b_{12} = -\frac{\partial \mathbf{r}}{\partial u}\frac{\partial \mathbf{N}}{\partial v} = -\frac{\partial \mathbf{r}}{\partial v}\frac{\partial \mathbf{N}}{\partial u}, \quad b_{22} = -\frac{\partial \mathbf{r}}{\partial v}\frac{\partial \mathbf{N}}{\partial v} \ , \tag{2.43}$$

from which we may derive the mixed components

$$b^{\alpha}_{\beta} = b_{\gamma\beta}a^{\gamma\beta} \ . \tag{2.44}$$

From the curvature tensor two important invariant scalar quantities can be derived. The first one is

$$\mathcal{H} = \frac{1}{2}(b^1_1 + b^2_2) \ . \tag{2.45}$$

It represents the average of the two *principle curvatures* and is called the *mean curvature*. The other invariant is the determinant

$$\mathcal{K} = \begin{vmatrix} b^1_1 & b^1_2 \\ b^2_1 & b^2_2 \end{vmatrix} = b^1_1 b^2_2 - b^1_2 b^2_1 \tag{2.46}$$

and is called the *Gaussian curvature* of the surface.

2.3 Review of Linear Elasticity

2.3.1 Strain–Displacement Relations

The displacement vector u at a point in a solid has the three components $u_i (i = 1, 2, 3)$ which are mutually orthogonal in a Cartesian coordinate system x_i and are taken to be

positive in the direction of the positive coordinate axes. Let us denote ε a strain tensor with the components ε_{ij}. Then the linearized straindisplacement relations, which form the Cauchy's infinitesimal strain tensor, are

$$\varepsilon_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right). \tag{2.47}$$

By this equation, the strain tensor is symmetric and thus consists of six components. Six strain components are required to characterize the state of strain at a point and are computed from the displacement field. However, if it is required to find three displacement components from the six components of strain, the six strain-displacement equations should possess a solution. The existence of the solution is guaranteed if the strain components satisfy the following six *compatibility conditions*

$$\frac{\partial^2 \varepsilon_{ij}}{\partial x_m \partial x_n} + \frac{\partial^2 \varepsilon_{mn}}{\partial x_i \partial x_j} = \frac{\partial^2 \varepsilon_{im}}{\partial x_j \partial x_n} + \frac{\partial^2 \varepsilon_{jn}}{\partial x_i \partial x_m} \ . \tag{2.48}$$

Although there are six conditions, only three are independent.

2.3.2 Constitutive Relations

The kinematic conditions of Section 2.3.1 are applicable to any continuum irrespective of its physical constitution. But the response of a given continuous body depends on its material. The material is introduced to the formulation through the generalized Hooke's law relates the stress tensor σ and strain tensor ε

$$\sigma_{ij} = C_{ijkl}\varepsilon_{kl} \ , \tag{2.49}$$

where C_{ijkl} depending on material properties is called elasticity tensor. Note from the symmetry of both σ_{ij} and ε_{kl} that $C_{ijkl} = C_{jikl}$ and $C_{ijkl} = C_{ijlk}$; and there are only 36 constants. When a strain-energy function exists, the number of independent constants is reduced from 36 to 21. The number of elastic constants is reduced to 13 when one plane of elastic symmetry exists, and is further reduced to 9 when three mutually orthogonal planes of elastic symmetry exist. Finally, when the material is isotropic (i.e., the material

has the same material properties in all directions), the number of independent constants reduces to 2 and the isotropic elasticity tensor has the form

$$C_{ijkl} = c_1 \delta_{ij} \delta_{kl} + c_2 \left(\delta_{ik} \delta_{jl} + \delta_{il} \delta_{jk} \right) \ ; \tag{2.50}$$

where c_1 and c_2 are the Lamé elastic constants, related to Young's modulus, E, and Poisson's ratio, ν, as follows

$$c_1 = \frac{E\nu}{(1+\nu)(1-2\nu)}, \quad c_2 = \frac{E}{2(1+\nu)} \ . \tag{2.51}$$

It can then be verified that the elasticity tensor satisfies

$$C_{ijkl} = C_{jikl} = C_{ijlk} = C_{klij} \ . \tag{2.52}$$

It thus follows from (2.47), (2.49), and (2.52) that

$$\sigma_{ij} = C_{ijkl} \frac{\partial u_k}{\partial x_l} \ . \tag{2.53}$$

2.3.3 Equations of Equilibrium/Motion

Equilibrium at a point in a solid is characterized by a relationship between stresses and body forces (forces per unit volume) b_i such as those generated by gravity. This relationship is expressed by equations of equilibrium

$$\frac{\partial \sigma_{ij}}{\partial x_j} + b_i = 0 \ . \tag{2.54}$$

Including inertial effects via D'Alembert forces gives the equations of motion

$$\frac{\partial \sigma_{ij}}{\partial x_j} + b_i = \rho \frac{\partial^2 u_i}{\partial t} \ , \tag{2.55}$$

where ρ is the material's density. When the elastic solid subjected to a harmonic loading (and harmonic body force) of frequency ω, the magnitude u of the harmonic response

$U = ue^{-i\omega t}$ satisfies

$$\frac{\partial \sigma_{ij}}{\partial x_j} + b_i + \rho\omega^2 u_i = 0 \ . \tag{2.56}$$

2.3.4 Boundary Conditions

Let Γ_D denote a part of the surface of the body on which some displacements $\bar{u}_i$ is specified. Continuity condition requires that on the surface Γ_D, the displacements u_i be equal to the specified displacements $\bar{u}_i$

$$u_i = \bar{u}_i, \quad \text{on } \Gamma_{\mathrm{D}} \ . \tag{2.57}$$

Similarly, Let Γ_N denote the part of the surface of the body on which forces are prescribed. The boundary condition requires the forces applied to Γ_N be in equilibrium with the stress components on the surface

$$\sigma_{ij}\hat{n}_j = t_i, \quad \text{on } \Gamma_{\mathrm{N}}, \tag{2.58}$$

where $\hat{n}_j$ are the components of the unit vector $\hat{n}$ normal to the surface, and t_i are specified boundary stresses (surface forces per unit area).

2.3.5 Weak Formulation

In the thesis, we shall limit our attention to only linear constitutive models. Hence, substituting (2.53) into (2.55) yields governing equations for the displacement field u as

$$\frac{\partial}{\partial x_j}\left(C_{ijkl}\frac{\partial u_k}{\partial x_l}\right) + b_i + \omega^2 u_i = 0 \quad \text{in } \Omega \ . \tag{2.59}$$

To derive the weak form of the governing equations, we introduce a function space

$$X^{\mathrm{e}} = \{v \in \left(H^1(\Omega)\right)^d \mid v_i = 0 \text{ on } \Gamma_{\mathrm{D}}\} \ , \tag{2.60}$$

and associated norm

$$||v||_{X^{\mathrm{e}}} = \left(\sum_{i=1}^{d} ||v_i||^2_{H_1(\Omega)}\right)^{1/2} . \tag{2.61}$$

Next multiplying (2.59) by a test function $v \in X^{\mathrm{e}}$ and integrating by parts we obtain

$$\int_\Omega \frac{\partial v_i}{\partial x_j} C_{ijkl} \frac{\partial u_k}{\partial x_l} - \omega^2 \int_\Omega u_i v_i - \int_\Gamma C_{ijkl} \frac{\partial u_k}{\partial x_l} \hat{n}_j v_i - \int_\Omega b_i v_i = 0 \ . \tag{2.62}$$

It thus follows from (2.58) and $v \in X^{\mathrm{e}}$ that the displacement field $u^{\mathrm{e}} \in X^{\mathrm{e}}$ satisfies

$$a(u^{\mathrm{e}}, v) = f(v) \ , \quad \forall \ v \in X^{\mathrm{e}} \ , \tag{2.63}$$

where

$$a(w, v) = \int_\Omega \frac{\partial v_i}{\partial x_j} C_{ijkl} \frac{\partial w_k}{\partial x_l} - \omega^2 w_i v_i \ , \tag{2.64}$$

$$f(v) = \int_\Omega b_i v_i + \int_{\Gamma_{\mathrm{N}}} v_i t_i \ . \tag{2.65}$$

This is the weak formulation in linear constitutive models for elastic solid subjected to a harmonic loading. In the next section, we review the finite element method which is one of the most frequently used method for numerical solution of PDEs arising in solid elasticity, fluid mechanics, heat transfer, etc.

2.4 Review of Finite Element Method

2.4.1 Weak Formulation

While the derivation of governing equations for most engineering problems is not difficult, their exact solution by analytical techniques is very hard or even impossible to find. In such cases, numerical methods are used to obtain an approximate solution. Among many possible choices, the finite element method is most frequently used to obtain an accurate approximation to the exact solution. The point of departure for the finite element method is an weighted–integral statement of a differential equation, called the weak formulation. The weak formulation (or, in short, weak form) allows for more general solution spaces and includes the natural boundary and continuity conditions of the problem. Typically, the weak form of the linear boundary value problems can be stated as: find $s^{\mathrm{e}}(\mu) = \ell(u^{\mathrm{e}}(\mu))$,

where $u^{\mathrm{e}}(\mu) \in X^{\mathrm{e}}$ is the solution of

$$a(u^{\mathrm{e}}(\mu), v; \mu) = f(v), \quad \forall\, v \in X^{\mathrm{e}} \ . \tag{2.66}$$

Here $a(\cdot, \cdot; \mu)$ is a μ-parametrized bilinear form, f is a linear functional, and X^{e} is an appropriate Hilbert space over the physical domain $\Omega \in \mathbb{R}^d$.

2.4.2 Space and Basis

In the finite element method, we seek the approximate solution over a discretized domain known as a triangulation $\mathcal{T}_h$ of the physical domain Ω: $\overline{\Omega} = \bigcup_{T_h \in \mathcal{T}_h} \overline{T}_h$, where $T_h^k, k = 1, \ldots, K$, are the elements, $x_i, i = 1, \ldots, \mathcal{N}$, are the nodes, and subscript h denoting the diameter of the triangulation $\mathcal{T}_h$ is the maximum of the longest edges of all elements. We next define a finite element "truth" approximation space $X \subset X^{\mathrm{e}}$

$$X = \{v \in X^{\mathrm{e}} \mid v|_{T_h} \in \mathbb{P}_p(T_h),\ \forall\, T_h \in \mathcal{T}_h\} \ , \tag{2.67}$$

where $\mathbb{P}_p(T_h)$ is the space of p^{th} degree polynomials over element T_h.
Furthermore, if the function space X^{e} is complex such that

$$X^{\mathrm{e}} = \{v = v^{\mathrm{R}} + \mathrm{i}v^{\mathrm{I}} \mid v^{\mathrm{R}} \in H^1(\Omega),\ v^{\mathrm{I}} \in H^1(\Omega)\} \ , \tag{2.68}$$

we must then require our truth approximation space be complexified as

$$X = \left\{v = v^{\mathrm{R}} + \mathrm{i}v^{\mathrm{I}} \in X^{\mathrm{e}} \mid v^{\mathrm{R}}|_{T_h} \in \mathbb{P}_p(T_h),\ v^{\mathrm{I}}|_{T_h} \in \mathbb{P}_p(T_h),\ \forall\, T_h \in \mathcal{T}_h\right\} \ , \tag{2.69}$$

in terms of which we define the associated inner product as

$$(w, v)_X = \int_\Omega \nabla w \nabla \bar{v} + w\bar{v} \ . \tag{2.70}$$

Recall that R and I denote the real and imaginary part, respectively; and that $\bar{v}$ and $|v|$ denote the complex conjugate and modulus of v, respectively. Note the notion of symmetry in the complex case, a bilinear form $a(w, v)$ is said to be symmetric if and only

if $a(w,v) = \overline{a(v,w)}, \forall w, v \in X$. It is clear that $(\cdot,\cdot)_X$ defined above is symmetric.

To obtain the discrete equations of the weak form, we express the field variable $u(\mu) \in X$ in terms of the nodal basic functions $\varphi_i \in X$, $\varphi_i(x_j) = \delta_{ij}$, such that

$$X = \text{span}\,\{\varphi_1, \ldots, \varphi_{\mathcal{N}}\} \ , \tag{2.71}$$

$$u(\mu) = \sum_{i=1}^{\mathcal{N}} u_i(\mu)\varphi_i, \quad \forall\, v \in X \ ; \tag{2.72}$$

here $u_i(\mu), i = 1, \ldots, \mathcal{N}$, is the nodal value of $u(\mu)$ at node x_i and is real for the real space X^{e}, otherwise complex.

Finally, we note that for complex domains involving curved boundaries or surfaces, simple triangular elements may not be sufficient. In such cases, the use of arbitrary shape elements, which are known as *isoparametric* elements, can lead to higher accuracy. However, since all problems discussed in the thesis have rather simple geometry, we shall not use isoparametric elements in our implementation of the finite element method.

2.4.3 Discrete Equations

Using the Galerkin projection on the discrete space X, we can find the approximation $u(\mu) \in X$ to $u^{\text{e}}(\mu) \in X^{\text{e}}$ from

$$a(u(\mu), v; \mu) = f(v), \ \forall\, v \in X \ . \tag{2.73}$$

We next substitute the approximation $u(\mu) = \sum_{j=1}^{\mathcal{N}} u_j(\mu)\varphi_j$ into (2.73) and take v as the basis functions $\varphi_i, i = 1, \ldots, \mathcal{N}$, to obtain the desired linear system

$$\sum_{j=1}^{\mathcal{N}} a\left(\varphi_j, \varphi_i; \mu\right) u_j(\mu) = f\left(\varphi_i\right), \quad i = 1, \ldots, \mathcal{N} \ , \tag{2.74}$$

which can be written into matrix form

$$\underline{A}(\mu)\, \underline{u}(\mu) = \underline{F} \ . \tag{2.75}$$

Here $\underline{A}(\mu)$ is an $\mathcal{N}\times\mathcal{N}$ matrix with $A_{ij}(\mu) = a\left(\varphi_j, \varphi_i; \mu\right)$, $\underline{F}$ is an vector with $F_i = f\left(\varphi_i\right)$, and $\underline{u}(\mu)$ is an vector with $u_i(\mu) = u(x_i; \mu)$, where x_i is the coordinates of the node i. The matrix $\underline{A}$ and vector $\underline{F}$ depend on the finite element mesh and type of basis functions. They can be formed via assembling elemental matrices and vectors associated with each elements T_h of $\mathcal{T}_h$.

By solving the linear system, we obtain the nodal values $\underline{u}(\mu)$ and thus $u(\mu) = \sum_{i=1}^{\mathcal{N}} u_i(\mu)\varphi_i$. Finally, the output approximation $s(\mu)$ can be calculated as

$$s(\mu) = \ell(u(\mu)) \ . \tag{2.76}$$

A complete discussion and detailed implementation of the finite element procedure can be found in most finite element method textbooks (see, for example, [15]).

2.4.4 *A Priori* Convergence

The finite element method seeks the approximate solution $u(\mu)$ (respectively, the approximate output $s(\mu)$) in the finite element "truth" approximation space X to the exact solution $u^{\mathrm{e}}(\mu)$ (respectively, the exact output $s^{\mathrm{e}}(\mu)$) of the underlying PDE. The *a priori* convergence analysis for the finite element approximation suggests that $\|u^{\mathrm{e}}(\mu) - u(\mu)\|_X$ and $|s^{\mathrm{e}}(\mu) - s(\mu)|$ will converge as h^α and h^β, respectively; here α and β are positive constants whose value depend on the specific problem, the output functional, and the regularity of force functional and domain. In general, we have $u(\mu) \to u^{\mathrm{e}}(\mu)$ and $s(\mu) \to s^{\mathrm{e}}(\mu)$, as $h \to 0$. For a particular case in which a is symmetric positive-definite, Ω and $f, (\ell = f)$ are sufficiently regular, $\|u^{\mathrm{e}}(\mu) - u(\mu)\|_X$ and $|s^{\mathrm{e}}(\mu) - s(\mu)|$ will vanish as $O(h)$ and $O(h^2)$, respectively, for $\mathbb{P}_1$ elements; it means in practice that in order to decrease $|s^{\mathrm{e}}(\mu) - s(\mu)|$ by a factor of C, we need to increase $\mathcal{N}$ roughly by the same factor for two-dimensional problems, but a factor of $C^{3/2}$ for three-dimensional problems.

As the requirements for accuracy increase, we need higher $\mathcal{N}$ to obtain accurate and reliable results; adequately converged truth approximations are thus achieved only for spaces X of very large dimension $\mathcal{N}$. For many medium or large-scale applications, $\mathcal{N}$ is typically in the order of $O(10^4)$ up to $O(10^6)$. Unfortunately, the computational complexity for solving the linear system (2.75) scales as $O(\mathcal{N}^\gamma)$, where γ depends on the

sparse structure and condition number of the stiffness matrix $\underline{A}(\mu)$. The computational time for a particular input is thus typically long; especially, in contexts where many and real-time queries of the parametrized discrete system (2.75)-(2.76) are required, the computational requirements become prohibitive.

2.4.5 Computational Complexity

Finally, we remark briefly solution methods for linear algebraic systems with specific attention to the FEM context. Typically, the FEM yields large and spare systems. Many techniques exist for solving such systems (see [95] for comprehensive discussion including algorithms, convergence analysis, preconditioning of several techniques for large linear systems). The appropriate technique depends on many factors, including the mathematical and structural properties of the matrix $\underline{A}$, the dimension of $\underline{A}$, and the number of right-hand sides. Generally, there are two classes of algorithms used to solve linear systems: direct methods and iterative methods. Direct methods obtain the solution after a finite number of arithmetic operations by performing some type of elimination procedure directly on a linear system; hence, direct methods will yield an exact solution in a finite number of steps if all calculations are exact (without truncation and round-off errors). In contrast, iterative methods define a sequence of approximations which converge to the exact solution of linear systems within some error tolerance.

The most standard direct method is the Gaussian elimination, which consists of the LU factorization and the backward substitution. The LU factorization of $\underline{A}$ generates lower and upper triangular matrices, $\underline{L}$ and $\underline{U}$, respectively, such that $\underline{A} = \underline{L}\,\underline{U}$. The backward substitution is straightforward: $\underline{L}\,\underline{w} = \underline{F}$ and $\underline{U}\,\underline{u} = \underline{w}$. Since $\underline{A}$ is sparse and banded, banded LU scheme is usually used to factorize $\underline{A}$ with (typical) cost $O(\mathcal{N}^2)$ and storage $O(\mathcal{N}^{3/2})$ for problems in $\mathbb{R}^2$. In $\mathbb{R}^3$, the order of factorization cost and storage requirement for banded LU factorization can be higher mainly due to the increasingly complicated sparse structure of the matrix.[1] In the case of symmetric positive-definite (SPD) systems, $\underline{A}$ can be factorized into $\underline{R}^T\underline{R}$, where $\underline{R}$ is upper triangular, by Cholesky

[1]Note for general domain and unstructured meshes, there are a number of heuristic methods to minimize the bandwidth. More generally, graph-based sparse matrix techniques can be applied — the edges and vertices of the matrix graph are simply the vertices and edges of the triangulation.

decomposition with a saving factor of 2 in both computational cost and storage.

Direct methods are usually preferred if the dimension of $\underline{A}$ are not too large, the spare structure is banded and well-structured, and there are a number of right-hand sides, since they are very fast and reliable in such situations. However, for general sparse matrices, the situation is considerably more complicated; in particular, the factors $\underline{L}$ and $\underline{U}$ can become extremely dense even though $\underline{A}$ is extremely spare; if pivoting is required, implementing sparse factorization can use a lot of time searching lists of numbers and creating a great deal of computational overhead. Iterative methods prove appropriate and outperform direct methods for solving general sparse and unstructured systems, especially arising from finite element discretization of three-dimensional problems.

Iterative methods start with an initial approximation $\underline{u}^0$ and construct a sequence of approximate solutions $\underline{u}^{n+1}$ to the exact solution $\underline{u}$. If converged, $\|\underline{A}\,\underline{u}^{n+1} - \underline{F}\|/\|\underline{u}^{n+1}\|$ or $\|\underline{u}^{n+1} - \underline{u}^n\|/\|\underline{u}^{n+1}\|$ becomes sufficiently small within a specific error tolerance. During the iterations, $\underline{A}$ is involved only in matrix-vector products, there is no need to store the matrix $\underline{A}$. Such methods are thus particularly useful for very large sparse systems — the matrices can be huge, sometimes involving several million unknowns. Iterative methods may be further classified into *stationary* iterative methods and *gradient* methods.

The Jacobi, Gauss-Seidel, successive over-relaxation (SOR) methods fall into the first class. The idea here is do matrix splitting $\underline{A} = \underline{M} - \underline{N}$ and write the linear system $\underline{A}\,\underline{u} = \underline{F}$ into an iterative fashion $\underline{M}\,\underline{u}^{n+1} = \underline{N}\,\underline{u}^n + \underline{F}$; here $\underline{M}$ must be nonsingular. We can further reduce the above iteration into an equivalent form, $\underline{u}^{n+1} = \underline{B}\,\underline{u}^n + \underline{C}$, where $\underline{B} = \underline{M}^{-1}\underline{N}$ and $\underline{C} = \underline{M}^{-1}\underline{F}$. A iterative scheme of this form is called *stationary iterative method* and $\underline{B}$ is the *iteration matrix* (In a non-stationary method, $\underline{B}$ varies with n). We have $\underline{M} = \underline{D}$, $\underline{N} = \underline{L} + \underline{U}$ for Jacobi method and $\underline{M} = \underline{D} - \underline{L}$, $\underline{N} = \underline{U}$ for Gauss-Seidel method, where $\underline{D}$, $\underline{L}$, $\underline{U}$ are the diagonal part, strictly negative lower triangular part, and strictly negative upper triangular part of the matrix $\underline{A}$, respectively, i.e., $\underline{A} = \underline{D} - \underline{L} - \underline{U}$. In SOR method, $\underline{M}$, $\underline{N}$, and $\underline{B}$ depend on a relaxation parameter ω; in particular, we have $\underline{B} = (\underline{D} - \omega\underline{L})^{-1}[(1-\omega)\underline{D} + \omega\underline{U}]$. Clearly, the Gauss-Seidel method is a special case of SOR method in which $\omega = 1$. The convergence and rate of convergence of the Jacobi, Gauss-Seidel, and SOR schemes depend on the spectral radius of $\underline{B}$ defined as $\rho(\underline{B}) = \max_{1\le i\le\mathcal{N}} \|\lambda_i(\underline{B})\|$, where $\lambda_i, 1 \le i \le \mathcal{N}$, are eigenvalues of $\underline{B}$. Typically, $\rho(\underline{B})$ is

large, and hence the convergence rate of the stationary iterative methods are quite slow. This observation has stimulated the development of gradient methods.

The literature of gradient methods are rich and many gradient methods have been developed over past decades; however, we shall confine our discussion to only the conjugate gradient (CG) method — one of the most important iterative methods for solving large SPD systems. The CG algorithm is given in Figure 2-1.

---

1. Set $\underline{u}^0(\text{say}) = 0$, $\underline{r}^0 = \underline{F}$, $\underline{p}^0 = \underline{r}^0$
2. **for** $n = 0, 1, \ldots,$ **until** convergence
3. $\quad \alpha^n = (\underline{r}^n)^T \underline{r}^n / (\underline{p}^n)^T \underline{A}\underline{p}^n$
4. $\quad \underline{u}^{n+1} = \underline{u}^n + \alpha^n \underline{p}^n$
5. $\quad \underline{r}^{n+1} = \underline{r}^n - \alpha^n \underline{A}\underline{p}^n$
6. $\quad \beta^n = (\underline{r}^{n+1})^T \underline{r}^{n+1} / (\underline{r}^n)^T \underline{r}^n$
7. $\quad \underline{p}^{n+1} = \underline{r}^{n+1} + \beta^n \underline{p}^n$
8. $\quad$ Test for convergence $\|\underline{r}^{n+1}\| / \|\underline{u}^{n+1}\| \leq \varepsilon$
9. **end for**

Figure 2-1: Conjugate Gradient Method for SPD systems.

The convergence rate of the CG method is given by the following estimate

$$\frac{(\underline{u} - \underline{u}^n)^T \underline{A} (\underline{u} - \underline{u}^n)}{(\underline{u})^T \underline{A}\underline{u}} \leq 2 \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^n , \tag{2.77}$$

where κ is the condition number of matrix $\underline{A}$

$$\kappa = \frac{\lambda_{\max}(\underline{A})}{\lambda_{\min}(\underline{A})} . \tag{2.78}$$

Here $\lambda_{\max}$ and $\lambda_{\min}$ refer to the maximum eigenvalue and minimum eigenvalue of $\underline{A}$. (In addition to the above result, we also obtain, at least in infinite precision, the finite termination property $\underline{u}^{\mathcal{N}} = \underline{u}$, though this is generally not much of interest.) By taking the logarithm of both sides of (2.77) and using the Taylor series for $\ln(1 + z)$ in the right-hand side, we obtain the number of iteration n_{iter} required to reduce the error by

some fixed fraction ε as

$$n_{\text{iter}} = \frac{1}{2}\sqrt{\kappa(\underline{A})}\ln\left(\frac{2}{\varepsilon}\right) \ . \tag{2.79}$$

We see that n_{iter} depends on h: as h decreases, κ increases, which in turn decreases the convergence rate. However, the dependence on h is not so strong, and is also independent of spatial dimension.

As proven in [124], the upper bound for the condition number is obtained as $\kappa(\underline{A}) \leq Ch^{-2}$ for quasi-uniform and regular meshes.[2] Hence, we have $n_{\text{iter}} \approx O(1/h) \approx O(\mathcal{N}^{1/2})$ for problems in $\mathbb{R}^2$ and $n_{\text{iter}} \approx O(1/h) \approx O(\mathcal{N}^{1/3})$ for problems in $\mathbb{R}^3$. We further observe from the CG algorithm that the work per iteration is roughly $O(\mathcal{N})$ due to the sparsity of the matrix $\underline{A}$. The complexity of the CG method is thus $O(\mathcal{N}^{3/2})$ in $\mathbb{R}^2$ and $O(\mathcal{N}^{5/3})$ in $\mathbb{R}^3$. In addition, the storage requirement for CG is only $O(\mathcal{N})$ since we only need to store the elemental matrices and the field vectors, both of which are $O(\mathcal{N})$.[3] We see that in $\mathbb{R}^2$, the CG method can be better than the banded LU factorization. In $\mathbb{R}^3$, the improvement is even more dramatic. Despite the relatively good convergence rate of the CG method, it is often of interest to improve things further by preconditioning. Especially, in the case of nonsymmetric indefinite systems and unstructured meshes, the iterative procedures are much less effective. In such cases, preconditioned iterative methods should be used to speed the convergence rate.

[2] This result is valid for any SPD second-oder elliptic PDE and any order of polynomial approximation; C depends on the polynomial order, coercivity and continuity constants, but *not* on h.

[3] Of course, with regard to the operation counts for both the computational complexity and the storage requirement, the constant in $\mathbb{R}^3$ is higher than that in $\mathbb{R}^2$.

Chapter 3

Reduced-Basis Methods: Basic Concepts

The focus in this chapter is on the computational methods that solve the direct problems very efficiently. Our approach is based on the reduced-basis method which permits rapid yet accurate and reliable evaluation of the input-output relationship induced by parametrized partial differential equations. For the purpose of illustrating essential components and key ideas of the reduced-basis method, in this chapter we choose to review the technique for coercive elliptic linear partial differential equations. In subsequent chapters, we shall develop the method for noncoercive linear and nonaffine linear elliptic equations, as well as nonlinear elliptic equations.

3.1 Abstraction

3.1.1 Preliminaries

We consider the "exact" (superscript e) problem: for any $\mu \in \mathcal{D} \subset \mathbb{R}^P$, find $s^{\rm e}(\mu) = \ell(u^{\rm e}(\mu))$, where $u^{\rm e}(\mu)$ satisfies the weak form of the μ-parametrized PDE

$$a(u^{\rm e}(\mu), v; \mu) = f(v), \quad \forall\, v \in X^{\rm e}. \tag{3.1}$$

Here μ and $\mathcal{D}$ are the input and (closed) input domain, respectively; $s^{\rm e}(\mu)$ is the output of interest; $u^{\rm e}(x;\mu)$ is our field variable; $X^{\rm e}$ is a Hilbert space defined over the physical domain $\Omega \subset \mathbb{R}^d$ with inner product $(w,v)_{X^{\rm e}}$ and associated norm $\|w\|_{X^{\rm e}} = \sqrt{(w,w)_{X^{\rm e}}}$; and $a(\cdot,\cdot;\mu)$ and $f(\cdot), \ell(\cdot)$ are $X^{\rm e}$-continuous bilinear and linear functionals, respectively.

Our interest here is in second-order PDEs, and our function space $X^{\rm e}$ will thus satisfy $(H^1_0(\Omega))^\nu \subset X^{\rm e} \subset (H^1(\Omega))^\nu$, where $\nu = 1$ for a scalar field variable and $\nu = d$ for a vector field variable. Recall that $H^1(\Omega)$ (respectively, $H^1_0(\Omega)$) is the usual Hilbert space (respectively, the Hilbert space of functions that vanish on the domain boundary $\partial\Omega$) defined in Section 2.1.

In actual practice, we replace $X^{\rm e}$ with $X \subset X^{\rm e}$, a "truth" finite element approximation space of dimension $\mathcal{N}$. The inner product and norm associated with X are given by $(\cdot,\cdot)_X$ and $\|\cdot\|_X = (\cdot,\cdot)_X^{1/2}$, respectively. A typical choice for $(\cdot,\cdot)_X$ is

$$(w,v)_X = \int_\Omega \nabla w \cdot \nabla v + wv \ , \tag{3.2}$$

which is simply the standard $H^1(\Omega)$ inner product. We shall next denote by X' the dual space of X. For a $h \in X'$, the dual norm is given by

$$\|h\|_{X'} \equiv \sup_{v \in X} \frac{h(v)}{\|v\|_X} \ . \tag{3.3}$$

We shall assume that the bilinear form a is symmetric, $a(w,v;\mu) = a(v,w;\mu), \forall w, v \in X, \forall \mu \in \mathcal{D}$, and satisfies a coercivity and continuity condition

$$0 < \alpha_0 \le \alpha(\mu) \equiv \inf_{v \in X} \frac{a(v,v;\mu)}{\|v\|_X^2}, \quad \forall\, \mu \in \mathcal{D} \tag{3.4}$$

$$\sup_{v \in X} \frac{a(v,v;\mu)}{\|v\|_X^2} \equiv \gamma(\mu) < \infty, \quad \forall\, \mu \in \mathcal{D} \ . \tag{3.5}$$

Here $\alpha(\mu)$ is the coercivity constant — the minimum (generalized) singular value associated with our differential operator — and $\gamma(\mu)$ is the standard continuity constant; of course, both these "constants" depend on the parameter μ. It is then standard, by the Lax-Milgram theorem [127], to prove the existence and uniqueness for the problem (3.1)

provided that the domain Ω and functional f are sufficiently regular.

Finally, we suppose that for some finite integer Q, a may be expressed as an affine decomposition of the form

$$a(w, v; \mu) = \sum_{q=1}^{Q} \Theta^q(\mu) a^q(w, v), \tag{3.6}$$

where for $1 \leq q \leq Q$, $\Theta^q : \mathcal{D} \rightarrow \mathbb{R}$ are differentiable parameter-dependent coefficient functions and bilinear forms $a^q : X \times X \rightarrow \mathbb{R}$ are parameter-independent.

3.1.2 General Problem Statement

Our approximation of the continuous problem in the finite approximation space X can then be stated as: given $\mu \in \mathcal{D} \in \mathbb{R}^P$, we evaluate

$$s(\mu) = \ell(u(\mu)) \tag{3.7}$$

where $u(\mu) \in X$ is the solution of the discretized weak form

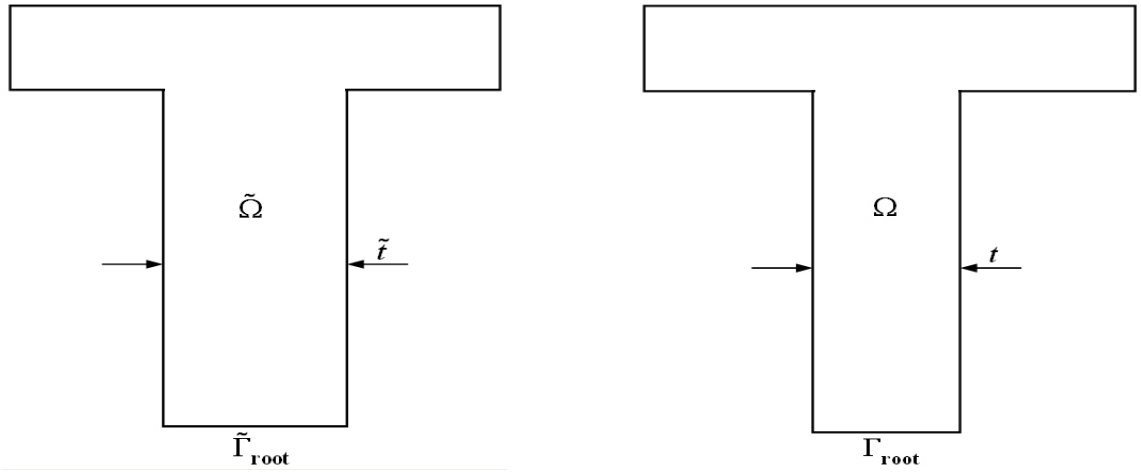
$$a(u(\mu), v; \mu) = f(v), \quad \forall v \in X \ . \tag{3.8}$$

We shall assume — hence the appellation "truth" — that X is sufficiently rich that u (respectively, s) is sufficiently close to $u^{\mathrm{e}}(\mu)$ (respectively, $s^{\mathrm{e}}(\mu)$) for all μ in the (closed) parameter domain $\mathcal{D}$. We must be certain that our formulation are *stable* and *efficient* as $\mathcal{N} \rightarrow \infty$. Unfortunately, for any reasonable error tolerance, the dimension $\mathcal{N}$ required to satisfy this condition — even with the application of appropriate (and even parameter-dependent) adaptive mesh generation/refinement strategies — is typically extremely large, and in particular much too large to provide real-time response.

3.1.3 A Model Problem

We consider heat conduction in a thermal fin of width and height unity, and thermal conductivity unity; the height of the fin post is 4/5 of the total height. The two-dimensional fin, shown in Figure 3-1(a), is characterized by a two-component param-

eter input $\mu = (\mu_1, \mu_2)$, where $\mu_1 = Bi$ and $\mu_2 = \tilde{t}/t$; μ may take on any value in a specified design set $\mathcal{D} \equiv [0.01, 1] \times [1/3, 5/3] \subset \mathbb{R}^{P=2}$. Here Bi is the Biot number, a non-dimensional heat transfer coefficient reflecting convective transport to the air at the fin surfaces; and $\tilde{t}$ is the width of the fin post.



(a) (b)

Figure 3-1: Two-dimensional thermal fin: (a) original (parameter-dependent) domain and (b) reference (parameter-independent) domain ($t = 0.3$).

The thermal fin is under a prescribed unit heat flux at the root. The steady-state temperature distribution within the fin, $\tilde{u}(\mu)$, is governed by the elliptic partial differential equation

$$-\nabla^2 \tilde{u} = 0, \quad \text{in } \tilde{\Omega} \ . \tag{3.9}$$

We now introduce a Neumann boundary condition on the fin root

$$-\nabla \tilde{u} \cdot \hat{\tilde{n}} = -1, \quad \text{on } \tilde{\Gamma}_{\text{root}}, \tag{3.10}$$

which models the heat source; and a Robin boundary condition on the remaining boundary

$$-\nabla \tilde{u} \cdot \hat{\tilde{n}} = \text{Bi}\, \tilde{u}, \quad \text{on } \partial\tilde{\Omega} \backslash \tilde{\Gamma}_{\text{root}}, \tag{3.11}$$

which models the convective heat losses; here $\partial\tilde{\Omega}$ denotes the boundary of $\tilde{\Omega}$ and $\hat{\tilde{n}}$ is the unit vector normal to the boundary.

The output considered is $s(\mu)$, the average steady-state temperature of the fin root normalized by the prescribed heat flux into the fin root

$$s(\mu) \equiv \ell(\tilde{u}(\mu)) = \int_{\tilde{\Gamma}_{\text{root}}} \tilde{u}(\mu) \ . \tag{3.12}$$

The weak formulation of (3.9), (3.10), and (3.11) is then derived as

$$\int_{\tilde{\Omega}} \nabla\tilde{u}\nabla\tilde{v} + \text{Bi} \int_{\partial\tilde{\Omega}\setminus\tilde{\Gamma}_{\text{root}}} \tilde{u}\tilde{v} = \int_{\tilde{\Gamma}_{\text{root}}} \tilde{v}, \quad \forall \tilde{v} \in H^1(\tilde{\Omega}) \ . \tag{3.13}$$

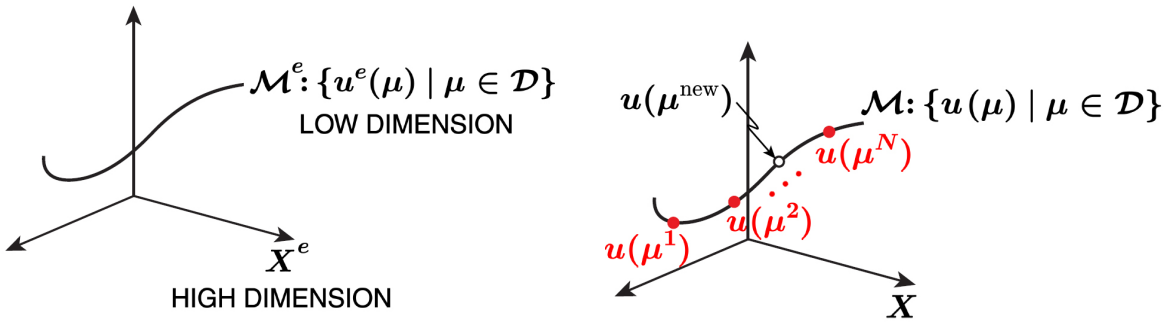
The problem statement (3.7) and (3.8) is recovered. Clearly, a is continuous, coercive, and symmetric, but not affine in the parameter yet. We now apply a continuous affine mapping from μ-dependent domain $\tilde{\Omega}$ to a fixed (μ-independent) reference domain Ω (see Figure 3-1(b)). In the reference domain, our abstract form (3.7)-(3.8) is recovered; in particular, a is affine for $Q = 5$, ℓ is "compliant" (i.e., $\ell = f$), and X is a piecewise-linear finite element approximation space of dimension $\mathcal{N} = 2977$. Note that the geometric variations are reflected, via the mapping, in the parametric coefficient functions $\Theta^q(\mu)$.

3.2 Reduced-Basis Approximation

3.2.1 Manifold of Solutions

The reduced-basis method recognizes that although the field variable $u^{\text{e}}(\mu)$ generally belongs to the infinite-dimensional space X^{e} associated with the underlying partial differential equation, in fact $u^{\text{e}}(\mu)$ resides on a very low-dimensional manifold $\mathcal{M}^{\text{e}} \equiv \{u^{\text{e}}(\mu) \,|\, \mu \in \mathcal{D}\}$ induced by the parametric dependence. For example, for a single parameter, $\mu \in \mathcal{D} \subset \mathbb{R}^{P=1}$, $u^{\text{e}}(\mu)$ will describe a one-dimensional filament that winds through X^{e} as depicted in Figure 3-2(a). The manifold containing all possible solutions of the partial differential equation induced by parametric dependence is much smaller than the function space. In the finite element method, even in the adaptive context, the approximation space X is much too general — X can approximate many functions that do not reside on the manifold of interest — and hence much too expensive. This critical observation presents a clear opportunity: we can effect significant, in many cases Draconian, dimension reduc-

tion in state space if we restrict attention to $\mathcal{M}^{\mathrm{e}}$; the field variable can then be adequately approximated by a space of dimension $N \ll \mathcal{N}$.



(a) (b)

Figure 3-2: (a) Low-dimensional manifold in which the field variable resides; and (b) approximation of the solution at μ^{new} by a linear combination of precomputed solutions.

3.2.2 Dimension Reduction

Since all solutions of the parametrized PDE live in a low-dimensional manifold, we wish to construct an approximation space to the manifold. The approximation space consists of solutions at selected points in the parameter space as shown in Figure 3-2(b). Then for any given parameter μ, we can approximate the solution $u(\mu)$ by a projection onto the approximation space. Essentially, we introduce nested samples, $S_N = \{\mu_1 \in \mathcal{D}, \cdots, \mu_N \in \mathcal{D}\}, 1 \leq N \leq N_{\max}$ and associated nested Lagrangian reduced-basis spaces as $W_N = \text{span}\{\zeta_j \equiv u(\mu_j), 1 \leq j \leq N\}, 1 \leq N \leq N_{\max}$, where $u(\mu_j)$ is the solution to (3.8) for $\mu = \mu_j$. In actual practice, the basis should be orthogonalized with respect to the inner product $(\cdot, \cdot)_X$; the algebraic systems then inherit the "conditioning" properties of the underlying PDE. The reduced-basis space W_N comprises "snapshots" on the parametrically induced manifold $\mathcal{M} \equiv \{u(\mu) \,|\, \mu \in \mathcal{D}\} \subset X$. It is clear that $\mathcal{M}$ is very *low-dimensional*; furthermore, it can be shown — we consider the equations for the sensitivity derivatives and invoke stability and continuity — that $\mathcal{M}$ is very *smooth*. We thus anticipate that $u_N(\mu) \rightarrow u(\mu)$ very rapidly, and that we may hence choose $N \ll \mathcal{N}$. Many numerical examples justify this expectation; and, in certain simple cases, exponential convergence can be proven [85, 93, 121]. We finally apply a Galerkin projection onto

W_N to obtain $u_N(\mu) \in W_N$ from

$$a(u_N(\mu), v; \mu) = f(v), \quad \forall v \in W_N \ , \tag{3.14}$$

in terms of which the reduced-basis approximation $s_N(\mu)$ to $s(\mu)$ can be evaluated as

$$s_N(\mu) = \ell(u_N(\mu)) \ . \tag{3.15}$$

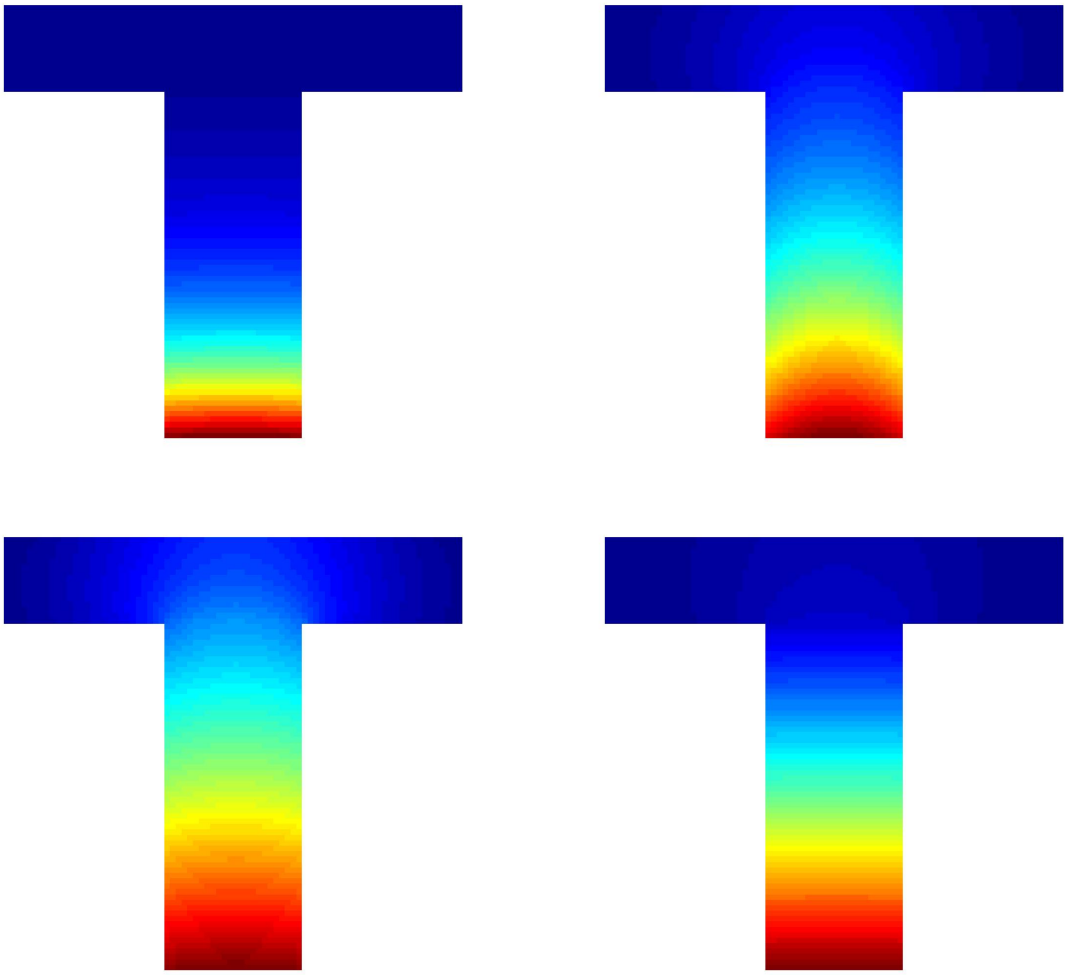


Figure 3-3: Few typical basic functions in W_N for the thermal fin problem.

An important question is that how we choose W_N so as to maximize the results

while minimizing the computational effort? An *ad hoc* or intuitive choice may not lead to satisfactory approximation even for large N. Naturally, we should find and include less smooth members of $\mathcal{M}$ into W_N because those solutions contain the highest quality information about structure of the manifold. In doing so, any information about $\mathcal{M}$ must be exploited and any corner of $\mathcal{M}$ must be explored. Of course, we can not afford the "accepted/rejected" strategy in which only a few basic solutions in W_N are selectively obtained from a large set of solutions as in POD economization procedure [134]. Our strategy is that we use inexpensive error bounds to guide us to potential candidates in $\mathcal{M}$ and an adaptive sampling procedure to explore $\mathcal{M}$. We shall discuss our way of choosing W_N in more detail shortly later.

3.2.3 *A Priori* Convergence Theory

We consider here the convergence rate of $u_N(\mu)$ and $s_N(\mu)$ to $u(\mu)$ and $s(\mu)$, respectively. In fact, it is a simple matter to show that the reduced-basis approximation $u_N(\mu)$ obtained in the reduced-basis space W_N is optimal in X-norm

$$\|u(\mu) - u_N(\mu)\|_X \leq \sqrt{\frac{\gamma(\mu)}{\alpha(\mu)}} \min_{w_N \in W_N} \|u(\mu) - w_N(\mu)\|_X \ . \tag{3.16}$$

Proof. We first note from (3.8) and (3.14)that

$$a(u(\mu) - u_N(\mu), v; \mu) = 0, \quad \forall v \in W_N \ . \tag{3.17}$$

It then follows for any $w_N = u_N + v_N \in W_N$, where $v_N \neq 0$, that

$$\begin{aligned} a(u - w_N, u - w_N; \mu) &= a(u - u_N - v_N, u - u_N - v_N; \mu) \\ &= a(u - u_N, u - u_N; \mu) - 2a(u - u_N, v_N; \mu) + a(v_N, v_N; \mu) \\ &= a(u - u_N, u - u_N; \mu) + a(v_N, v_N; \mu) \\ &> a(u - u_N, u - u_N; \mu) \ . \end{aligned} \tag{3.18}$$

Furthermore, from (3.4), (3.5), and (3.18) we have

$$\begin{aligned}
\alpha(\mu)\,\|u(\mu)-u_N(\mu)\|_X^2 &\leq a(u(\mu)-u_N(\mu),u(\mu)-u_N(\mu);\mu)\\
&\leq a(u(\mu)-w_N(\mu),u(\mu)-w_N(\mu);\mu)\\
&\leq \gamma(\mu)\min_{w_N\in W_N}\|u(\mu)-w_N\|_X^2\,, \qquad (3.19)
\end{aligned}$$

which concludes the proof. □

In a similar argument, we can also show that $s_N(\mu)$ converges optimally to $s(\mu)$ in X-norm. We show that for the compliance case $\ell = f$

$$\begin{aligned}
s(\mu)-s_N(\mu) &= \ell(u(\mu)-u_N(\mu))\\
&= a(u(\mu),u(\mu)-u_N(\mu);\mu)\\
&= a(u(\mu)-u_N(\mu),u(\mu)-u_N(\mu);\mu)\\
&\leq \gamma(\mu)\,\|u(\mu)-u_N(\mu)\|_X^2\\
&\leq \frac{\gamma^2(\mu)}{\alpha(\mu)}\min_{w_N\in W_N}\|u(\mu)-w_N(\mu)\|_X^2\,; \qquad (3.20)
\end{aligned}$$

in arriving at the above result, we use $\ell = f$ in the second equality, Galerkin orthogonality (3.17) and symmetry of a in the third equality, continuity condition in the fourth inequality, and the result (3.16) in the last equality. We see that $s_N(\mu)$ converges to $s(\mu)$ as the square of error in $u_N(\mu)$.

3.2.4 Offline-Online Computational Procedure

Of course, even though N may be small, the elements of W_N are in some sense large: $\zeta_n \equiv u(\mu_n)$ will be represented in terms of $\mathcal{N} \gg N$ truth finite element basis functions. To eliminate the $\mathcal{N}$-contamination, we must consider offline-online computational procedures. To begin, we expand our reduced-basis approximation as

$$u_N(\mu) = \sum_{j=1}^{N} u_{N\,j}(\mu)\zeta_j\ . \qquad (3.21)$$

It thus follows from (3.6) and (3.14) that the coefficients $u_{N\,j}(\mu)$, $1 \le j \le N$, satisfy the $N \times N$ linear algebraic system

$$\sum_{j=1}^{N} \left\{ \sum_{q=1}^{Q} \Theta^q(\mu)\, a^q(\zeta_j, \zeta_i) \right\} u_{N\,j}(\mu) = f(\zeta_i), \quad 1 \le i \le N \ . \tag{3.22}$$

The reduced-basis output can then be calculated as

$$s_N(\mu) = \sum_{j=1}^{N} u_{N\,j}(\mu)\, \ell(\zeta_j) \ . \tag{3.23}$$

It is clear from (3.22) that we may pursue an offline-online computational strategy to economize the output evaluation.

In the *offline* stage — performed *once* — we first solve for the ζ_i, $1 \le i \le N$; we then form *and store* $\ell(\zeta_i)$, $1 \le i \le N$, and $a^q(\zeta_j, \zeta_i)$, $1 \le i, j \le N$, $1 \le q \le Q$. In actual practice, in the offline stage we consider $N = N_{\max}$; then, in the online stage, we extract the necessary subvectors and submatrices. This will become clearer when we discuss the generation of the S_N, $1 \le N \le N_{\max}$. Note all quantities computed in the offline stage are independent of the parameter μ. Specifically, the offline computation requires N expensive finite-element solutions and $O(QN^2)$ finite-element vector inner products.

In the *online stage* — performed *many times*, for each new value of μ — we first assemble and subsequently invert the (full) $N \times N$ "stiffness matrix" $\sum_{q=1}^{Q} \Theta^q(\mu)\, a^q(\zeta_j, \zeta_i)$ in (3.22) — this yields the $u_{N\,j}(\mu)$, $1 \le j \le N$; we next perform the summation (3.23) — this yields the $s_N(\mu)$. The operation count for the online stage is respectively $O(QN^2)$ and $O(N^3)$ to assemble (recall the $a^q(\zeta_j, \zeta_i)$, $1 \le i, j \le N$, $1 \le q \le Q$, are *pre-stored*) and invert the stiffness matrix, and $O(N)$ to evaluate the output inner product (recall the $\ell(\zeta_j)$ are *pre-stored*); note that the reduced-basis stiffness matrix is, in general, *full*. The essential point is that the online complexity is *independent of* $\mathcal{N}$, the dimension of the underlying truth finite element approximation space. Since $N \ll \mathcal{N}$, we expect — and often realize — significant, orders-of-magnitude computational economies relative to classical discretization approaches.

3.2.5 Orthogonalized Basis

In forming the reduced-basis space W_N, the basis functions must be selected such that they are linearly independent to make the algebraic system (3.14) well-conditioned as possible, or at least not singular. However, the basis functions are the solutions of the parametrized partial differential equation at different configurations, they are nearly oriented in the same direction. Consequently, the associated algebraic system (3.14) is very ill-conditioned especially for large N. Typically, the condition number of the "reduced-stiffness" matrix in (3.14) grows exponentially with N. We thus need a new basis which is orthogonal and able to preserve all approximation properties of the original basis. To this end, using Gram-Schmidt orthogonalization we orthogonalize our basis with respect to the inner product associated with the Hilbert space X, $(\cdot,\cdot)_X$, and thus obtain

$$(\zeta_i, \zeta_j)_X = \delta_{ij}, \quad 1 \le i, j \le N \ . \tag{3.24}$$

Then the algebraic system (3.14) inherits the conditioning properties of the underlying PDE, as we shall now prove. We first note that for any $w_N \in W_N$, we can write $w = \sum_{i=1}^N w_{N\,i}\zeta_i$. It then follows from (3.4) and (3.24) that

$$\begin{aligned}
\sum_{i=1}^N \sum_{j=1}^N w_{N\,i} w_{N\,j} a(\zeta_i, \zeta_j; \mu) &\ge \alpha(\mu) \sum_{i=1}^N \sum_{j=1}^N w_{N\,i} w_{N\,j} (\zeta_i, \zeta_j)_X \\
&= \alpha(\mu) \sum_{i=1}^N \sum_{j=1}^N w_{N\,i} w_{N\,j} \delta_{ij} \\
&= \alpha(\mu) \sum_{i=1}^N w_{N\,i}^2 \ .
\end{aligned} \tag{3.25}$$

Similarly, we have

$$\begin{aligned}
\sum_{i=1}^N \sum_{j=1}^N w_{N\,i} w_{N\,j} a(\zeta_i, \zeta_j; \mu) &\le \gamma(\mu) \sum_{i=1}^N \sum_{j=1}^N w_{N\,i} w_{N\,j} (\zeta_i, \zeta_j)_X \\
&= \gamma(\mu) \sum_{i=1}^N \sum_{j=1}^N w_{N\,i} w_{N\,j} \delta_{ij} \\
&= \gamma(\mu) \sum_{i=1}^N w_{N\,i}^2 \ .
\end{aligned} \tag{3.26}$$

It finally follows from (3.25) and (3.26) that

$$\alpha(\mu) \leq \frac{\sum_{j=1}^{N} w_{N\,i} w_{N\,j} a(\zeta_i, \zeta_j; \mu)}{\sum_{i=1}^{N} w_{N\,i}^2} \leq \gamma(\mu), \quad \forall \underline{w}_N \in \mathbb{R}^N \ . \tag{3.27}$$

Clearly, our algebraic system in the orthogonalized basis has the same conditioning properties as the underlying PDE. In the worst case, the condition number is bounded by the ratio $\gamma(\mu)/\alpha(\mu)$, which is independent of N.

Using the thermal fin problem as typical demonstration, we present in Figure 3-4 the condition number of the reduced-stiffness matrix in the original basis and orthogonalized one as a function of N for $\mu_t = (0.1, 1.0)$. The exponential growth of the condition number of $a(\zeta_i, \zeta_j)$ in the original basis is expected; in contrast, the condition number of $a(\zeta_i, \zeta_j)$ in the orthogonalized basis increases linearly with N and begins to be saturated at $N = 4$ with value of 10.00 since for this particular test point, $\gamma(\mu_t)/\alpha(\mu_t) = 10.00$.

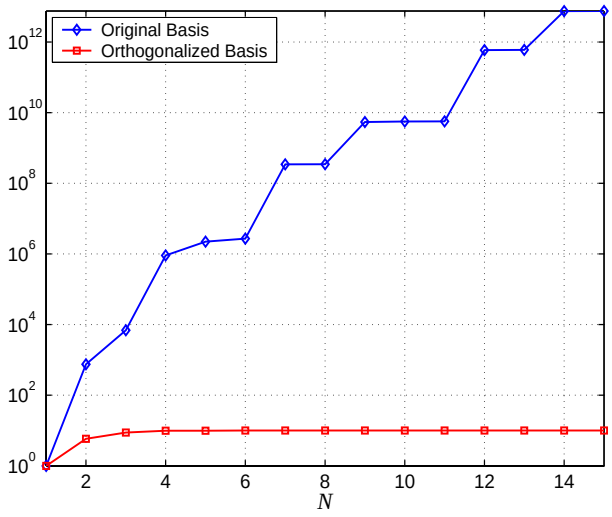


Figure 3-4: Condition number of the reduced-stifness matrix in the original and orthogonalized basis as a function of N, for the test point $\mu_t = (0.1, 1.0)$.

3.3 *A Posteriori* Error Estimation

From the previous section, we know in theory N can be chosen quite small. Nevertheless, in practice, we do not know how small N should be chosen in order for the reduced-basis method to produce desired accuracy for all parameter inputs. In fact, the reduced-basis

approximation raises many questions than it answers. Is $|s(\mu) - s_N(\mu)| \leq \varepsilon^s_{\text{tol}}$, where $\varepsilon^s_{\text{tol}}$ is the acceptable tolerance? Is N too large, $|s(\mu) - s_N(\mu)| \ll \varepsilon^s_{\text{tol}}$, with an associated steep penalty on computational efficiency? Do we satisfy the acceptable error condition $|s(\mu) - s_N(\mu)| \leq \varepsilon^s_{\text{tol}}$ for the smallest possible value of N? In short, the pre-asymptotic and essentially *ad hoc* nature of reduced-basis approximations, the strongly superlinear scaling with N of the reduced-basis complexity, and the particular needs of real-time demand rigorous *a posteriori* error estimation.

3.3.1 Error Bounds

We assume for now that we are given a positive μ-dependent lower bound $\hat{\alpha}(\mu)$ for the stability constant $\alpha(\mu)$: $\alpha(\mu) \geq \hat{\alpha}(\mu) \geq \alpha_0 > 0, \forall \mu \in \mathcal{D}$. The calculation of $\hat{\alpha}(\mu)$ will be discussed in great length in the next chapter. We next introduce the dual norm of the residual

$$\varepsilon_N(\mu) = \sup_{v \in X} \frac{r(v;\mu)}{\|v\|_X} \ , \tag{3.28}$$

where

$$r(v;\mu) = f(v) - a(u_N(\mu), v; \mu), \quad \forall v \in X \tag{3.29}$$

is the residual associated with $u_N(\mu)$. We may now define our energy error bound

$$\Delta_N(\mu) = \frac{\varepsilon_N(\mu)}{\hat{\alpha}(\mu)} \tag{3.30}$$

and the associated effectivity as

$$\eta_N(\mu) \equiv \frac{\Delta_N(\mu)}{\|u(\mu) - u_N(\mu)\|_X} \ . \tag{3.31}$$

We may also develop error bounds for the error in the output. We consider here the special "compliance" case in which $\ell = f$ and a is symmetric — more general functionals ℓ and nonsymmetric a require adjoint techniques [91, 121, 99]. We then define our output error estimator as

$$\Delta^s_N(\mu) \equiv \varepsilon^2_N(\mu)/\hat{\alpha}(\mu) \ , \tag{3.32}$$

and its corresponding effectivity as

$$\eta_N^s(\mu) \equiv \frac{\Delta_N^s(\mu)}{|s(\mu) - s_N(\mu)|} \ . \tag{3.33}$$

Note that $\Delta_N^s(\mu)$ scales as the *square* of the dual norm of the residual, $\varepsilon_N(\mu)$.

3.3.2 Rigor and Sharpness of Error Bounds

We shall prove in this section that $1 \leq \eta_N(\mu), \eta_N^s(\mu) \leq \gamma(\mu)/\hat{\alpha}(\mu), \forall N, \forall \mu \in \mathcal{D}$. Essentially, the left inequality states that $\Delta_N(\mu)$ (respectively, $\Delta_N^s(\mu)$) is a rigorous upper bound for $\|u(\mu) - u_N(\mu)\|_X$ (respectively, $|s(\mu) - s_N(\mu)|$); the right inequality states that $\Delta_N(\mu)$ (respectively, $\Delta_N^s(\mu)$) is a sharp upper bound for $\|u(\mu) - u_N(\mu)\|$ (respectively, $|s(\mu) - s_N(\mu)|$). In fact, many numerical examples [121, 142] show that $\eta_N(\mu)$ and $\eta_N^s(\mu)$ are of order unity.

Proposition 1. *For the error bounds $\Delta_N(\mu)$ and $\Delta_N^s(\mu)$ given in (3.30) and (3.32), the corresponding effectivities satisfy*

$$1 \leq \eta_N(\mu) \leq \gamma(\mu)/\hat{\alpha}(\mu), \quad \forall N, \ \forall \mu \in \mathcal{D} \ , \tag{3.34}$$

$$1 \leq \eta_N^s(\mu) \leq \gamma(\mu)/\hat{\alpha}(\mu), \quad \forall N, \ \forall \mu \in \mathcal{D} \ . \tag{3.35}$$

Proof. To begin, we note from (3.8) and (3.29) that the error $e(\mu) \equiv u(\mu) - u_N(\mu)$ satisfies

$$a(e(\mu), v; \mu) = r(v; \mu), \quad \forall \, v \in X \ . \tag{3.36}$$

Furthermore, we note from standard duality arguments that

$$\varepsilon_N(\mu) \equiv \|r(v; \mu)\|_{X'} = \|\hat{e}(\mu)\|_X \ , \tag{3.37}$$

where

$$(\hat{e}(\mu), v)_X = r(v; \mu), \quad \forall \, v \in X \ . \tag{3.38}$$

We next invoke the coercivity and continuity of the bilinear form a together with (3.36)

and (3.38) to obtain

$$\begin{aligned} \alpha(\mu)\,\|e(\mu)\|_X^2 &\leq a(e(\mu),e(\mu);\mu) \\ &= (\hat{e}(\mu),e(\mu))_X \\ &\leq \|\hat{e}(\mu)\|_X\,\|e(\mu)\|_X \ , \end{aligned} \tag{3.39}$$

and

$$\begin{aligned} \|\hat{e}(\mu)\|_X^2 &= a(e(\mu),\hat{e}(\mu);\mu) \\ &\leq a(e(\mu),e(\mu);\mu)^{1/2}a(\hat{e}(\mu),\hat{e}(\mu);\mu)^{1/2} \\ &\leq \gamma(\mu)\,\|\hat{e}(\mu)\|_X\,\|e(\mu)\|_X \ . \end{aligned} \tag{3.40}$$

Note that we have used Cauchy-Schwarz inequality in the last inequality of (3.39) and in the second inequality of (3.40). We thus conclude from (3.39) and (3.40) that

$$\alpha(\mu) \leq \frac{\|\hat{e}(\mu)\|_X}{\|e(\mu)\|_X} \leq \gamma(\mu) \ . \tag{3.41}$$

The first result immediately follows from the definition of $\eta_N(\mu)$, (3.37), and (3.41). Furthermore, we note from (3.39) and (3.41) that

$$s(\mu) - s_N(\mu) = a(e(\mu),e(\mu);\mu) \leq \|\hat{e}(\mu)\|_X\,\|e(\mu)\|_X \leq \frac{\|\hat{e}(\mu)\|_X^2}{\alpha(\mu)} \ , \tag{3.42}$$

and from (3.40) that

$$\begin{aligned} \|\hat{e}(\mu)\|_X^2 &\leq a(e(\mu),e(\mu);\mu)^{1/2}a(\hat{e}(\mu),\hat{e}(\mu);\mu)^{1/2} \\ &\leq a(e(\mu),e(\mu);\mu)^{1/2}\gamma^{1/2}(\mu)\,\|\hat{e}(\mu)\|_X \ . \end{aligned} \tag{3.43}$$

We thus conclude from (3.42) and (3.43) that

$$\alpha(\mu) \leq \frac{\|\hat{e}(\mu)\|_X^2}{s(\mu) - s_N(\mu)} \leq \gamma(\mu) \ . \tag{3.44}$$

The second result follows from the definition of $\eta_N^s(\mu)$, (3.37), and (3.44). □

The effectivity result (3.34) and (3.35) is crucial. From the left inequality, we deduce that $\Delta_N(\mu)$ (respectively, $\Delta_N^s(\mu)$) is a rigorous upper bound for the error in the solution (respectively, the error in the output) — this provides certification. From the right inequality, we deduce that $\Delta_N(\mu)$ and $\Delta_N^s(\mu)$ overestimate the true errors by at most $\gamma(\mu)/\tilde{\beta}(\mu)$, *independent of* N — this relates to efficiency: clearly an overly conservative error bound will be manifested in an unnecessarily large N and unduly expensive reduced-basis approximation. Note however that these error bounds are relative to our underlying "truth" approximation, $u(\mu) \in X$ and $s(\mu)$, not the exact solution and output, $u^{\mathrm{e}}(\mu) \in X^{\mathrm{e}}$ and $s^{\mathrm{e}}(\mu)$, respectively.

The real challenge in *a posteriori* error estimation is not the presentation of these rather classical results, but rather the development of efficient computational approaches for the evaluation of the necessary constituents. In our particular deployed context, "efficient" translates to "online complexity *independent* of $\mathcal{N}$," and "necessary constituents" translates to "dual norm of the residual, $\varepsilon_N(\mu) \equiv \|r(v;\mu)\|_{X'}$, and lower bound for the inf-sup constant, $\tilde{\beta}(\mu)$." In fact, for linear problems, the latter are rather universal — necessary and sufficient for most rigorous *a posteriori* contexts and approaches. In the nonlinear context, additional ingredients are required. We now address the former issue and leave the latter issue to be the subject of extensive study in the next chapter.

3.3.3 Offline/Online Computational Procedure

To begin, we invoke the affine assumption (3.6) to rewrite the relaxed error equation (3.38) as

$$(\hat{e}(\mu), v)_X = f(v) - \sum_{q=1}^{Q} \sum_{n=1}^{N} \Theta^q(\mu) u_{N\,n}(\mu) a^q(\zeta_n, v) \ . \tag{3.45}$$

It immediately follows from linear superposition that

$$\hat{e}(\mu) = \mathcal{C} + \sum_{q=1}^{Q} \sum_{n=1}^{N} \Theta^q(\mu) u_{N\,n}(\mu) \mathcal{L}_n^q, \quad \forall v \in X \ , \tag{3.46}$$

where $(\mathcal{C}, v)_X = f(v), \forall v \in X$, and $(\mathcal{L}_n^q, v)_X = -a^q(\zeta_n, v), \forall v \in X$, for $1 \le q \le Q, 1 \le n \le N$. Inserting the expression into (3.37), we obtain

$$\begin{aligned} \varepsilon_N^2(\mu) &= (\mathcal{C}, \mathcal{C})_X + 2\sum_{q=1}^{Q}\sum_{n=1}^{N} \Theta^q(\mu) u_{N\,n}(\mu)(\mathcal{C}, \mathcal{L}_n^q)_X \\ &\quad + \sum_{q=1}^{Q}\sum_{q'=1}^{Q}\sum_{n=1}^{N}\sum_{n'=1}^{N} \Theta^q(\mu)\Theta^{q'}(\mu) u_{N\,n}(\mu) u_{N\,n'}(\mu)(\mathcal{L}_n^q, \mathcal{L}_{n'}^{q'})_X \ . \end{aligned} \tag{3.47}$$

An efficient offline/online decomposition may now be identified.

In the offline stage — performed once — we first solve for $\mathcal{C}$ and $\mathcal{L}_n^q$, $1 \le n \le N$, $1 \le q \le Q$; we then form and store the relevant parameter-independent inner products $(\mathcal{C}, \mathcal{C})_X$, $(\mathcal{C}, \mathcal{L}_n^q)_X$, $(\mathcal{L}_n^q, \mathcal{L}_{n'}^{q'})_X$, $1 \le n, n' \le N$, $1 \le q, q' \le Q$. Note that all quantities computed in the offline stage are independent of the parameter μ.

In the online stage — performed many times, for each new value of μ — we simply evaluate the sum (3.47) in terms of the $\Theta^q(\mu), u_{N\,n}(\mu)$ and the pre-calculated and stored (parameter-independent) $(\cdot,\cdot)_X$ inner products. The operation count for the online stage is only $O(Q^2N^2)$ — again, the essential point is that the online complexity is *independent of* $\mathcal{N}$, the dimension of the underlying truth finite element approximation space. We further note that unless Q is quite large (typically, Q is $O(10)$ or less), the online cost associated with the calculation of the dual norm of the residual is commensurate with the online cost associated with the calculation of $s_N(\mu)$.

3.3.4 Bound Conditioners

We review here the simplest form of "bound conditioner" formulations developed in [139, 143] for calculating, $\hat{\alpha}(\mu)$, the lower bound of $\alpha(\mu)$. In particular, in the case of symmetric coercive operators we can often determine $\hat{\alpha}(\mu) \le \alpha(\mu)$, $\forall\, \mu \in \mathcal{D}$) "by inspection." For example, if we verify $\Theta^q(\mu) > 0$, $\forall\, \mu \in \mathcal{D}$, and $a^q(v, v) \ge 0$, $\forall\, v \in X$, $1 \le q \le Q$, then we may choose for our coercivity parameter an lower bound

$$\hat{\alpha}(\mu) = \left(\min_{q \in \{1,\ldots,Q\}} \frac{\Theta^q(\mu)}{\Theta^q(\overline{\mu})} \right) \alpha(\overline{\mu}) \ , \tag{3.48}$$

for an appropriate value $\overline{\mu} \in \mathcal{D}$, as we now prove

Proof. We invoke (3.4) and (3.6) to obtain the desired result

$$\begin{aligned}
\alpha(\mu) &\equiv \inf_{v\in X} \frac{\sum_{q=1}^{Q} \Theta^q(\mu) a^q(v,v)}{\|v\|_X^2} \\
&\geq \inf_{v\in X} \frac{\left(\min_{q\in\{1,\ldots,Q\}} \frac{\Theta^q(\mu)}{\Theta^q(\bar{\mu})}\right) \sum_{q=1}^{Q} \Theta^q(\bar{\mu}) a^q(v,v)}{\|v\|_X^2} \\
&= \left(\min_{q\in\{1,\ldots,Q\}} \frac{\Theta^q(\mu)}{\Theta^q(\bar{\mu})}\right) \alpha(\bar{\mu}), \quad \forall \mu \in \mathcal{D} \ ,
\end{aligned}$$

since $\Theta^q(\mu) > 0$, $\forall\, \mu \in \mathcal{D}$, and $a^q(v,v) \geq 0$, $\forall\, v \in X$, $1 \leq q \leq Q$. □

It can be verified that the thermal fin problem discussed in Section 3.1.3 accommodates these hypotheses. Moreover, for our choice of $(w,v)_X = \sum_{q=1}^{Q} a^q(w,v)$ and $\bar{\mu} = (1,1)$ we readily compute $\alpha(\bar{\mu}) = 1$ and $\Theta^q(\bar{\mu}) = 1$ for $1 \leq q \leq Q$. Unfortunately, these hypotheses are rather restrictive, and hence more complicated (and offline-expensive) bound conditioner recipes must be pursued [139, 143, 99]. As we shall see in subsequent chapters, the Helmholtz and elasticity problems do not admit the above hypotheses because $\Theta^q(\mu) > 0$, $\forall\, \mu \in \mathcal{D}$, and $a^q(v,v) \geq 0$, $\forall\, v \in X$, $1 \leq q \leq Q$ can not be satisfied simultaneously.

3.3.5 Sample Construction and Adaptive Online Strategy

In conclusion, we can calculate a rigorous and sharp upper bound $\Delta_N^s(\mu) = \varepsilon_N^2(\mu)/\tilde{\beta}(\mu)$ for $|s(\mu) - s_N(\mu)|$ with online complexity *independent* of $\mathcal{N}$. These inexpensive error bounds may be gainfully enlisted in the offline stage — to construct optimal samples S_N. We assume that we are given a sample S_N, and hence space W_N and associated reduced-basis approximation (procedure) $u_N(\mu)$, $\forall\, \mu \in \mathcal{D}$. We first calculate $\mu_N^* = \arg\max_{\mu\in\Xi_F} \Delta_N(\mu)$, where $\Delta_N(\mu)$ is our "online" error bound (3.30) and Ξ_F is a very (exhaustively) fine random grid over the parameter domain D of size $n_F \gg 1$; we next append μ_N^* to S_N to form S_{N+1}, and hence W_{N+1} and a procedure for $u_{N+1}(\mu)$, $\forall\, \mu \in \mathcal{D}$; we then continue this process until $\epsilon_{N_{\max}}^* \equiv \Delta_{N_{\max}}(\mu_{N_{\max}}^*) = \epsilon_{\text{tol,min}}$, where $\epsilon_{\text{tol,min}}$ is the smallest error tolerance anticipated.

Moreover, the bounds may also serve most crucially in the online stage — to choose optimal N, to confirm the desired accuracy, to establish strict feasibility, and to control sub-optimality: given any desired $\epsilon_{\text{tol}} \in [\epsilon_{\text{tol,min}}, \infty[$ and any new value of $\mu \in \mathcal{D}$ "in the

field," we first choose N from a pre-tabulated array such that $\epsilon_N^* \equiv \Delta_N(\mu_N^*) = \epsilon_{\text{tol}}$; we next calculate $u_N(\mu)$ and $\Delta_N(\mu)$, and then verify that — and if necessary, subsequently increase N *such that* — the condition $\Delta_N(\mu) \leq \epsilon_{\text{tol}}$ is indeed satisfied. We should not and do not rely on the finite sample Ξ_{F} for either rigor or sharpness. This strategy will minimize the online computational effort while simultaneously satisfying the requisite accuracy with certainty.

The crucial point is that $\Delta_N(\mu)$ is an accurate and "online-inexpensive" — order-unity effectivity and $\mathcal{N}$-independent complexity — surrogate for the true (very-expensive-to-calculate) error $\|u(\mu) - u_N(\mu)\|_X$. This surrogate permits us to (i) offline, perform a much more exhaustive and hence meaningful search for the best samples S_N and hence most rapidly *uniformly* convergent spaces W_N, and (ii) online, determine the smallest N, and hence the most efficient approximation, for which we *rigorously* achieve the desired accuracy. We may in fact view our offline sampling process as a (greedy, parameter space, "$L^\infty(\mathcal{D})$") variant of the POD economization procedure [134] in which — thanks to $\Delta_N(\mu)$ — *we need never construct* the "rejected" snapshots.

3.4 Numerical Results

We now present basic numerical results obtained with the thermal fin problem. We pursue the optimal sampling procedure described in the previous section on a regular grid Ξ_F of size $n_{\text{F}} = 1681$ to arrive at $N_{\max} = 15$ for our reduced-basis space $S_{N\,\max}$ as shown in Figure 3-5. It can be seen that nearly all of sample points lie on the boundary of $\mathcal{D}$, and that more sample points allocate on the left boundary. This is because for smaller Biot number temperature dissipates slowly so that the corresponding temperature distribution is more varying than that of large Biot number.

We next present in Table 3.1 the normalized maximum errors $\varepsilon_{N,\text{max,rel}}$ and $\varepsilon^s_{N,\text{max,rel}}$, as a function of N, for the (log) random and adaptive sampling processes (note that, in the results for the random sampling process, the sample S_N is different for each N). Here $\varepsilon_{N,\text{max,rel}}$ is the maximum over Ξ_{Test} of $\|e(\mu)\|_X / \|u(\mu)\|_X$, and $\varepsilon^s_{N,\text{max,rel}}$ is the maximum over Ξ_{Test} of $|s(\mu) - s_N(\mu)| \,/\, |s(\mu)|$, where $\Xi_{\text{Test}} \subset (\mathcal{D})^{256}$ is a regular 16×16 grid over $\mathcal{D}$. We observe that the adaptive sampling procedure yields higher accuracy even

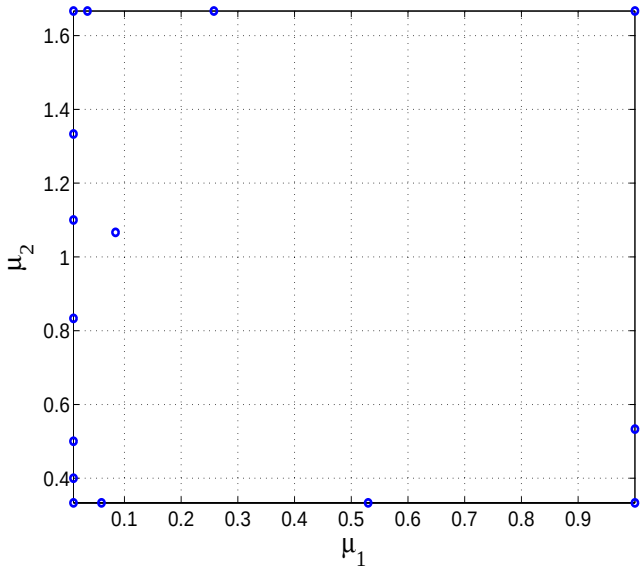


Figure 3-5: Sample $S_{N\,\max}$ from optimal sampling procedure.

with lower N as N increases; and that even these modest reductions in N can translate into measurable performance improvements especially in the context of design and optimization.

N	$\varepsilon_{N,\max,\mathrm{rel}}$ "Random"	$\varepsilon_{N,\max,\mathrm{rel}}$ "Adaptive"	$\varepsilon^s_{N,\max,\mathrm{rel}}$ "Random"	$\varepsilon^s_{N,\max,\mathrm{rel}}$ "Adaptive"
3	2.33×10^{-1}	5.55×10^{-1}	3.06×10^{-2}	2.51×10^{-1}
6	4.59×10^{-2}	1.04×10^{-1}	2.67×10^{-3}	8.55×10^{-3}
9	1.63×10^{-2}	1.58×10^{-2}	3.52×10^{-4}	2.55×10^{-4}
12	6.39×10^{-3}	2.97×10^{-3}	3.24×10^{-5}	1.20×10^{-5}
15	4.59×10^{-3}	5.42×10^{-4}	1.88×10^{-5}	2.71×10^{-7}

Table 3.1: Maximum relative errors as a function of N for random and adaptive samples.

We finally present in Table 3.2 $\Delta_{N,\max,\mathrm{rel}}$, $\eta_{N,\mathrm{ave}}$, $\Delta^s_{N,\max,\mathrm{rel}}$, and $\eta^s_{N,\mathrm{ave}}$ as a function of N. Here $\Delta_{N,\max,\mathrm{rel}}$ is the maximum over Ξ_{Test} of $\Delta_N(\mu)/\|u_N(\mu)\|$, $\eta_{N,\mathrm{ave}}$ is the average over Ξ_{Test} of $\Delta_N(\mu)/\|u(\mu)-u_N(\mu)\|$, $\Delta^s_{N,\max,\mathrm{rel}}$ is the maximum over Ξ_{Test} of $\Delta^s_N(\mu)/\|s_N(\mu)\|$, and $\eta^s_{N,\mathrm{ave}}$ is the average over Ξ_{Test} of $\Delta^s_N(\mu)/\|s(\mu)-s_N(\mu)\|$. We observe that the reduced-basis approximation converges very rapidly, and that our rigorous error bounds are in fact quite sharp as the effectivities are in $O(5)$. As expected, the error bound for the output scales as the square of the error bound for the solution.

N	$\Delta_{N,\max,\text{rel}}$	$\eta_{N,\text{ave}}$	$\Delta^s_{N,\max,\text{rel}}$	$\eta^s_{N,\text{ave}}$
3	1.07×10^{-0}	7.81	5.55×10^{-1}	8.45
6	1.43×10^{-1}	9.00	1.33×10^{-2}	9.26
9	2.35×10^{-2}	8.98	4.41×10^{-4}	9.27
12	5.96×10^{-3}	8.76	2.68×10^{-5}	6.78
15	5.99×10^{-4}	8.81	2.87×10^{-7}	3.81

Table 3.2: Error bounds and effectivities as a function of N.

3.5 Remarks

We present thus far in this chapter basic but important concepts of the reduced-basis method via applying the technique to elliptic linear problems with many restrictions on our abstract statement. More specifically, we have assumed that bilinear form $a(.,.)$ is coercive, symmetric, affine in the parameter, and that output functional is compliant $\ell = f$. However, many of these crucial concepts and observations will remain applicable to more general problems. In addition, we have not addressed the calculation of the lower bound $\hat{\alpha}(\mu)$ of the stability parameter $\alpha(\mu)$. We shall leave it as a subject of extensive study in the next chapter. In this section, we discuss briefly relaxation of these restrictions and leave the detailed and generalized developments for subsequent chapters.

3.5.1 Noncompliant Outputs and Nonsymmetric Operators

For noncompliant output ℓ ($\neq f$) or nonsymmetric operator a, we may define an error bound for the output as $\Delta^s_N(\mu) = \|\ell\|_{X'}\,\Delta_N(\mu)$, for which we clearly obtain $|s(\mu) - s_N(\mu)| = \ell(e(\mu)) \leq \|\ell\|_{X'}\|e(\mu)\|_X \leq \Delta^s_N(\mu)$. This error bound is admittedly rather crude, hence the associated effectivity may be very large. However, in numerous applications the error bound may be adequate to satisfy a specified tolerance condition thanks to the rapid convergence of our reduced-basis approximation.

In order to obtain the optimal convergence and thus recover the "square" effect, we use the adjoint techniques [121, 99]. To begin, we introduce a dual or adjoint, problem of the primal (3.8): given $\mu \in \mathcal{D}$, $\psi(\mu)$ satisfies

$$a(v, \psi(\mu); \mu) = -\ell(v), \quad \forall v \in X \ . \tag{3.49}$$

Note that if a is symmetric and $\ell = f$, which we shall denote "compliance," $\psi(\mu) = -u(\mu)$. In addition to W_N, we introduce dual reduced-basis approximation spaces $W_{N^{\rm du}}^{\rm du}$ as

$$W_{N^{\rm du}}^{\rm du} \equiv \text{span}\left\{\zeta_n^{\rm du} \equiv \psi(\mu_n^{\rm du}), 1 \le n \le N^{\rm du}\right\},$$

where $\psi(\mu_n^{\rm du}) \in X$ are the solutions to the dual problem at selected points $\mu_n^{\rm du}$, $n = 1, \ldots, N^{\rm du}$. We then apply Galerkin projection for both the primal and dual problems

$$a(u_N(\mu), v; \mu) = f(v), \quad \forall v \in W_N \ , \tag{3.50}$$

$$a(v, \psi_{N^{\rm du}}(\mu); \mu) = -\ell(v), \quad \forall v \in W_{N^{\rm du}}^{\rm du} \ , \tag{3.51}$$

in terms of which the reduced-basis approximation output can be evaluated as

$$s_N(\mu) = \ell(u_N(\mu)) - f(\psi_{N^{\rm du}}(\mu)) + a(u_N(\mu), \psi_{N^{\rm du}}(\mu); \mu) \ . \tag{3.52}$$

To show the optimal convergence rate of the approximation, we first recall the necessary result given in Section 3.2.3

$$\|u(\mu) - u_N(\mu)\|_X \le \sqrt{\frac{\gamma(\mu)}{\alpha(\mu)}} \min_{w_N \in W_N} \|u(\mu) - w_N\|_X \, , \tag{3.53}$$

$$\|\psi(\mu) - \psi_{N^{\rm du}}(\mu)\|_X \le \sqrt{\frac{\gamma(\mu)}{\alpha(\mu)}} \min_{w_{N^{\rm du}}^{\rm du} \in W_{N^{\rm du}}^{\rm du}} \left\|\psi(\mu) - w_{N^{\rm du}}^{\rm du}\right\|_X \ . \tag{3.54}$$

It then follows that

$$\begin{aligned}
|s(\mu) - s_N(\mu)| &= |\ell(u(\mu) - u_N(\mu)) + f(\psi_{N^{\rm du}}) - a(u_N(\mu), \psi_{N^{\rm du}}; \mu)| \\
&= |-a(u(\mu) - u_N(\mu), \psi(\mu); \mu) + a(u(\mu) - u_N(\mu), \psi_{N^{\rm du}}; \mu)| \\
&= |a(u(\mu) - u_N(\mu), \psi_{N^{\rm du}} - \psi(\mu); \mu)| \\
&\le \gamma(\mu) \|u(\mu) - u_N(\mu)\|_X \|\psi(\mu) - \psi_{N^{\rm du}}\|_X \\
&\le \frac{\gamma^2(\mu)}{\alpha(\mu)} \min_{w_N \in W_N} \|u(\mu) - w_N\|_X \min_{w_{N^{\rm du}}^{\rm du} \in W_{N^{\rm du}}^{\rm du}} \left\|\psi(\mu) - w_{N^{\rm du}}^{\rm du}\right\|_X
\end{aligned} \tag{3.55}$$

from the definition of the primal and dual problems, Galerkin orthogonality, continuity condition. As in the compliance case, equations (3.53), (3.54), and (3.55) together state that $u_N(\mu)$ (and $\psi_{N^{\rm du}}$) is the best approximation with respect to the X-norm, and that the error in the output converges as the product of the primal and dual errors.

To see the benefit of introducing the dual problem, we assume that $N^{\rm du}$ is in the order of $O(N)$; the online cost for solving both the primal and dual problem is thus $O(2N^3)$. Furthermore, in order for the reduced-basis formulation with the dual problem to achieve the same output error bound as the reduced-basis formulation with the dual problem does, we need to increase N by a factor of 2 or more, leading to an online cost of $O(8N^3)$ or higher.[1] As a result, the dual reduced-basis formulation typically enjoys $O(4)$ (or greater) reduction in computational effort. Note, however, that the simple crude output bound $\Delta_N^s(\mu) = \|\ell\|_{X'}\,\Delta_N(\mu)$ is very useful for cases with *many* outputs present, since adjoint techniques have a computational complexity (in both the offline and online stage) proportional to the number of outputs. A detailed formulation and theory for noncompliant problems can be found in [121, 139] upon which we extend the method for general nonaffine and noncompliant problems as described in Section 6.6.

3.5.2 Noncoercive Elliptic Problems

In noncoercive problems, the bilinear form $a(\cdot,\cdot)$ is required to satisfy the following inf-sup condition for well-posedness of problems

$$0 < \beta_0 \le \beta(\mu) \equiv \inf_{w\in X}\sup_{v\in X}\frac{a(w,v;\mu)}{\|w\|_X\|v\|_X},\ \forall\mu\in\mathcal{D}\ , \tag{3.56}$$

$$\gamma(\mu) \equiv \sup_{w\in X}\sup_{v\in X}\frac{a(w,v;\mu)}{\|w\|_X\|v\|_X}\ . \tag{3.57}$$

Here $\beta(\mu)$ is the Babuška "inf–sup" (stability) parameter — the minimum (generalized) singular value associated with our differential operator — and $\gamma(\mu)$ is the standard con-

[1] To see this, we note that the output bound is $O(\Delta_N(\mu))$ with the usual reduced-basis formulation and is $O(\Delta_N^2(\mu))$ with the primal-dual formulation (here we assume that $\|\psi(\mu)-\psi_N(\mu)\|_X$ converges like $\|u(\mu)-u_N(\mu)\|_X$). Therefore, we need to double N in order for the usual reduced-basis formulation to obtain the output bound $O(\Delta_N^2(\mu))$, if the reduced-basis approximation $u_N(\mu)$ converges exponentially; otherwise, we need to increase N by even more than a factor of 2.

tinuity constant.

Numerical difficulties arise due to noncoercivity and "weaker" stability condition in both (i) the approximation and (ii) error estimation. In (i), large and rapid variation of the field variables in both x and μ can lead to poor convergence rate. Furthermore, in the noncoercive case, standard Galerkin projection does not guarantee stability of the discrete reduced-basis system (which is another factor leading to poor approximations). However, it is possible to improve the approximation and ensure the stability by considering projections other than standard Galerkin like minimum-residual or Petrov-Galerkin projections with infimizer–supremizer enriched [91, 131]. Obviously, our adaptive sampling procedure also plays an important role in improving the convergence by ensuring good approximation properties for W_N. In (ii), the primary difficulty lies in estimation of the inf-sup parameter which is typically very small near resonances and in resonances. In particular, $\beta(\mu)$ can not typically be deduced analytically, and thus must be approximated. The developments presented in Chapter 4 can be used to obtain the necessary approximation (more specifically, a lower bound) to the inf–sup parameter. We shall leave greater discussion of noncoercive problems for Chapter 5.

3.5.3 Nonaffine Linear Elliptic Problems

Throughout this chapter we assume that $a(w, v; \mu)$ is affine in μ as given by (3.6), we then develop extremely efficient offline-online computational strategy. The online cost to evaluate $s_N(\mu)$ and $\Delta_N^s(\mu)$ is independent of $\mathcal{N}$. Unfortunately, if a is not affine in the parameter, the online complexity is no longer independent of $\mathcal{N}$. For example, for *general* $g(x; \mu)$, the bilinear form

$$a(w, v; \mu) \equiv \int_\Omega \nabla w \cdot \nabla v + \int_\Omega g(x; \mu)\, wv \tag{3.58}$$

will not admit an efficient online-offline decomposition. The difficulty here is that the nonaffine dependence of $g(x; \mu)$ on parameter μ does not allow separation of the generation and projection stages and thus leads to online $\mathcal{N}$ dependence. Consequently, the computational improvements in using the reduced-basis method relative to conventional (say) finite element approximation are modest.

In Chapter 6, we describe a technique that recovers online $\mathcal{N}$ independence even in the presence of non-affine parameter dependence. Our approach (applied to (3.58), say) is simple: we develop a "collateral" reduced-basis expansion $g_M(x;\mu)$ for $g(x;\mu)$; we then replace $g(x;\mu)$ in (3.58) with the (necessarily) affine approximation $g_M(x;\mu)$. The essential ingredients are (*i*) a "good" collateral reduced-basis approximation space, (*ii*) a stable and inexpensive interpolation procedure, and (*iii*) an effective *a posteriori* estimator to quantify the newly introduced error terms. It is perhaps only in the latter that the technique is somewhat disappointing: *the error estimators — though quite sharp and very efficient — are completely (provably) rigorous upper bounds only in certain restricted situations* [135].

3.5.4 Nonlinear Elliptic Problems

Obviously nonlinear equations do not admit the same degree of generality as linear equations. We thus present our approach to nonlinear equations for a particular nonlinear problem. In particular, we consider the following nonlinear elliptic problem

$$a(u,v;\mu) + \int_\Omega g(u;x;\mu)v = f(v), \quad \forall v \in X, \tag{3.59}$$

where as before a is a symmetric, continuous, and coercive bilinear form, f is bounded linear functional, and $g(u;x;\mu)$ is a general nonlinear function of the parameter μ, spatial coordinate x, and field variable $u(x;\mu)$. Furthermore, we need to restrict our attention to only g such that the equation (3.59) is well-posed and sufficiently stable. Even so, the nonlinearity in g creates many numerical difficulties.

It should be emphasized that the application of the reduced-basis method to quadratically nonlinear problems — the steady incompressible Navier-Stokes equations — has been considered [141, 140]. In this thesis, we shall pursue a further development of the method for highly nonlinear problems. Our approach to nonlinear elliptic problems uses the same ideas as above for nonaffine linear elliptic problems, but involves more sophisticated and expensive treatment.

Chapter 4

Lower Bounds for Stability Factors for Elliptic Problems

4.1 Introduction

In the previous chapter, we have presented various aspects of the reduced-basis method and demonstrated through the heat conduction problem the efficiency and accuracy of the technique. However, we have not addressed the calculation of the lower bound $\hat{\alpha}(\mu)$ for the stability factor $\alpha(\mu)$ — a generalized minimum singular value — which is crucial to our error estimation since the lower bound enters in the denominator of the error bounds. *Upper* bounds for minimum eigenvalues are essentially "free"; however rigorous *lower* bounds are notoriously difficult to obtain. In earlier works [121, 120, 143, 139, 142], a family of rigorous error estimators for reduced-basis approximation of a wide class of partial differential equations has been introduced; in particular, rigorous a posteriori error estimation procedures which rely critically on the existence of a *bound conditioner* – in essence, an operator preconditioner that (*i*) satisfies an additional spectral"bound" requirement, and (*ii*) admits the reduced-basis off-line/on-line computational stratagem. In this section, we shall review shortly the concept of bound conditioners upon which we construct the lower bounds and develop *a posteriori* error estimation procedures for elliptic linear problems that yield rigorous error statements for *all* N.

4.1.1 General Bound Conditioner

A new class of improved bound conditioners based on the direct approximation of the parametric dependence of the inverse of the operator (rather than the operator itself) was first introduced in [143]. In particular, the authors suggested a symmetric, continuous, and coercive bound conditioner $c : X \times X \times \mathcal{D} \to \mathbb{R}$ such that

$$c^{-1}(\,\cdot\,,\,\cdot\,;\mu) = \sum_{i \in \mathcal{I}(\mu)} \rho_i(\mu) c_i^{-1}(\,\cdot\,,\,\cdot\,)\;. \tag{4.1}$$

Here $\mathcal{D} \in \mathbb{R}^P$ is the parameter domain; X is an appropriate function space over the real field $\mathbb{R}$; $\mathcal{I}(\mu) \in \{1,\ldots,I\}$ is a parameter-dependent set of indices, where I is a finite (preferably small) integer; $c_i : X \times X \to \mathbb{R}, 1 \le i \le I$, are parameter-independent symmetric, coercive operators. The "separability" of $c^{-1}(\,\cdot\,,\,\cdot\,;\mu)$ as a sum of products of parameter-*dependent* functions $\rho_i(\mu)$ and parameter-*independent* operators c_i^{-1} allows a higher-order effectivity constructions (e.g., piecewise-linear) while simultaneously preserving online efficiency.

4.1.2 Multi-Point Bound Conditioner

When a single bound conditioner $c(\,\cdot\,,\,\cdot\,;\mu)$ is used for all $\mu \in \mathcal{D}$, we call this bound conditioner as *single-point* bound conditioner. In many cases, the effectivity bound obtained with single-point bound conditioner is quite pessimistic. The effectivity may be improved by judicious choice of *multi-point* bound conditioner. The critical observation is that using many "local" bound conditioners may lead to better approximation of the effectivity factor and thus smaller effectivity. To this end, we specify in the parameter domain a set of partitions $P_K \equiv \{\mathcal{P}_1,\ldots,\mathcal{P}_K\}$ such that $\cup_{k=1}^K \overline{\mathcal{P}}_k = \mathcal{D}$ and $\cap_{k=1}^J \mathcal{P}_k = \emptyset$, where $\overline{\mathcal{P}}_k$ is the closure of $\mathcal{P}_k$; we next associate each $\mathcal{P}_k$ with bound conditioner $c^k(\,\cdot\,,\,\cdot\,;\mu)$ and separately pursue the effectivity construction for each $c^k(\,\cdot\,,\,\cdot\,;\mu)$ on the corresponding region $\mathcal{P}_k$, $k = 1,\ldots,K$; we then select the appropriate local bound conditioner (e.g., $c^i(\cdot,\cdot;\mu)$) and the associated effectivity construction for our online calculation according to value of μ (e.g., $\mu \in \mathcal{P}_i$).

4.1.3 Stability-Factor Bound Conditioner

We consider a special case in which $I = 1$ and $K = 1$, hence $c(\cdot, \cdot; \mu) = c_1(\cdot, \cdot)/\rho(\mu)$, where $c_1(\cdot, \cdot)$ is a parameter-independent symmetric coercive operator. We shall denote $c_1(\cdot, \cdot)$ as $(\cdot, \cdot)_X$ and call it as "stability-factor" bound conditioner (in short, bound conditioner) because it can be used as the inner product to define relevant stability and continuity constants for a coercive/noncoercive operator. For a coercive operator $a(\cdot, \cdot; \mu)$, we may conveniently state the stability condition as

$$0 < \alpha_0 \leq \alpha(\mu) \equiv \inf_{v \in X} \frac{a(v, v; \mu)}{\|v\|_X^2}, \quad \forall \mu \in \mathcal{D} \ . \tag{4.2}$$

A similar stability statement is completely applicable for a noncoerive operator $a(\cdot, \cdot; \mu)$, but now in terms of the inf-sup condition

$$0 < \beta_0 \leq \beta(\mu) \equiv \inf_{w \in X} \sup_{v \in X} \frac{a(w, v; \mu)}{\|w\|_X \|v\|_X}, \quad \forall \mu \in \mathcal{D} \ . \tag{4.3}$$

A typical choice for $(\cdot, \cdot)_X$ may be

$$(w, v)_X \equiv \int_\Omega \nabla w \cdot \nabla v + \delta \int_\Omega wv \ , \tag{4.4}$$

for some appropriately pre-determined nonnegative constant $\delta \geq 0$.

In the following, we shall develop the lower bound construction for the stability factors. For simplicity of exposition, we consider the single stability-factor bound conditioner. Of course, the development can be applied to general and multi-point bound conditioners. In addition, we assume that for some finite integer Q, a may be expressed as an affine decomposition of the form

$$a(w, v; \mu) = \sum_{q=1}^{Q} \Theta^q(\mu) a^q(w, v), \quad \forall \, w, v \in X, \ \forall \, \mu \in \mathcal{D} \ , \tag{4.5}$$

where for $1 \leq q \leq Q$, $\Theta^q : \mathcal{D} \to \mathbb{R}$ are differentiable parameter-dependent functions and $a^q : X \times X \ \to \mathbb{R}$ are parameter-independent continuous forms. It is worth noting that the following lower bound formulations though developed for the real function space X

and real parametric functions $\Theta^q(\mu), 1 \le q \le Q$, can be easily generalized to the complex case in which X is the function space over the complex field $\mathbb{C}$ and $\Theta^q : \mathcal{D} \to \mathbb{C}$ are complex functions; see Appendix C for the generalization of the lower bound formulation for complex noncoercive operators.

4.2 Lower Bounds for Coercive Problems

4.2.1 Coercivity Parameter

Recall that the stability factor of the coercive operator $a(\cdot, \cdot; \mu)$ is defined as

$$\alpha(\mu) \equiv \inf_{v \in X} \frac{a(v, v; \mu)}{\|v\|_X^2}, \tag{4.6}$$

which shall be called *coercivity* parameter to distinguish it from the stability factor of noncoercive operators. We note that

Lemma 4.2.1. *If $\Theta^q(\mu), 1 \le q \le Q$, are concave function of μ and $a_q(w, w), 1 \le q \le Q$, are positive-semidefinite then the function $\alpha(\mu)$ is concave in μ.*

Proof. For any $\mu_1 \in \mathcal{D}$, $\mu_2 \in \mathcal{D}$, and $\lambda \in [0, 1]$, we have

$$\begin{aligned}
\alpha(\lambda\mu_1 + (1-\lambda)\mu_2) &= \inf_{w \in X} \frac{\sum_{q=1}^{Q} \Theta^q(\lambda\mu_1 + (1-\lambda)\mu_2) a^q(w, w)}{\|w\|_X^2} \\
&\ge \inf_{w \in X} \frac{\sum_{q=2}^{Q} \left(\lambda\Theta^q(\mu_1) + (1-\lambda)\Theta^q(\mu_2)\right) a^q(w, w)}{\|w\|_X^2} \\
&\ge \lambda \inf_{w \in X} \frac{\sum_{q=1}^{Q} \Theta^q(\mu_1) a^q(w, w)}{\|w\|_X^2} + (1-\lambda) \inf_{w \in X} \frac{\sum_{q=1}^{Q} \Theta^q(\mu_2) a^q(w, w)}{\|w\|_X^2} \\
&= \lambda\alpha(\mu_1) + (1-\lambda)\alpha(\mu_2)
\end{aligned}$$

from the concavity of $\Theta^q(\mu)$ and the positive-semidefiniteness of $a_q(\cdot, \cdot)$ for $1 \le q \le Q$. □

It is necessary to study and exploit the concavity of $\alpha(\mu)$ because if $\alpha(\mu)$ is concave we may then pursue the lower bound construction based directly on $\alpha(\mu)$ rather than the concave but more expensive intermediary $\mathcal{F}(\mu - \overline{\mu}; \overline{\mu})$. However, since the above assumptions are quite restricted, Lemma 4.2.1 is of little practical value. We henceforth opt a complicated but general construction as discussed next.

4.2.2 Lower Bound Formulation

We now consider the construction of $\hat{\alpha}(\mu)$, a lower bound for $\alpha(\mu)$. To begin, given $\overline{\mu} \in \mathcal{D}$ and $t = (t_{(1)}, \ldots, t_{(P)}) \in \mathbb{R}^P$, we introduce the bilinear form

$$\mathcal{T}(w, v; t; \overline{\mu}) = a(w, v; \overline{\mu})_X + \sum_{p=1}^{P} t_{(p)} \sum_{q=1}^{Q} \frac{\partial \Theta^q}{\partial \mu_{(p)}}(\overline{\mu}) a^q(w, v) \tag{4.7}$$

and associated Rayleigh quotient

$$\mathcal{F}(t; \overline{\mu}) = \min_{v \in X} \frac{\mathcal{T}(v, v; t; \overline{\mu})}{\|v\|_X^2} \ . \tag{4.8}$$

It is crucial to note (and we shall exploit) the property that $\mathcal{F}(t; \overline{\mu})$ is *concave* in t, and hence $\mathcal{D}^{\overline{\mu}} \equiv \left\{ \mu \in \mathbb{R}^P \mid \mathcal{F}(\mu - \overline{\mu}; \overline{\mu}) \geq 0 \right\}$ is perforce convex.

Lemma 4.2.2. *For given $\overline{\mu} \in \mathcal{D}$, the function $\mathcal{F}(t; \overline{\mu})$ is concave in t. Hence, given $t_1 < t_2$, for all $t \in [t_1, t_2]$, $\mathcal{F}(t; \overline{\mu}) \geq \min(\mathcal{F}(t_1; \overline{\mu}), \mathcal{F}(t_2; \overline{\mu}))$.*

Proof. We define $\lambda = (t_2 - t)/(t_2 - t_1) \in [0, 1]$ such that $t = \lambda t_1 + (1 - \lambda)\, t_2$. It follows from (4.7) that $\mathcal{T}(v, v; t; \overline{\mu}) = \lambda\, \mathcal{T}(v, v; t_1; \overline{\mu}) + (1 - \lambda)\, \mathcal{T}(v, v; t_2; \overline{\mu})$, and hence

$$\begin{aligned} \mathcal{F}(t; \overline{\mu}) &= \inf_{v \in X} \left(\lambda\, \mathcal{T}(v, v; t_1; \overline{\mu}) + (1 - \lambda)\, \mathcal{T}(v, v; t_2; \overline{\mu}) \right) / \|v\|_X^2 \\ &\geq \lambda\, \mathcal{F}(t_1; \overline{\mu}) + (1 - \lambda)\, \mathcal{F}(t_2; \overline{\mu}) \\ &\geq \min(\mathcal{F}(t_1; \overline{\mu}), \mathcal{F}(t_2; \overline{\mu})) \ . \end{aligned}$$

□

Next we assume that a^q are continuous in the sense that there exist positive finite constants Γ_q, $1 \leq q \leq Q$, such that

$$|a^q(w, w)| \leq \Gamma_q\, |w|_q^2, \quad \forall w \in X; \tag{4.9}$$

here $|\cdot|_q : X \to \mathbb{R}^+$ are seminorms that satisfy

$$C_X = \sup_{w \in X} \frac{\sum_{q=1}^{Q} |w|_q^2}{\|w\|_X^2}, \tag{4.10}$$

for some positive constant C_X. It is often the case that $\Theta^1(\mu) = \text{Constant}$, in which

case the $q = 1$ contribution to the sum in (4.7) and (4.10) may be discarded. (Note that C_X is typically independent of Q, since the a^q are often associated with non-overlapping subdomains of Ω.) We may then define, for $\mu \in \mathcal{D}$, $\overline{\mu} \in \mathcal{D}$,

$$\Phi(\mu, \overline{\mu}) \equiv C_X \max_{q \in \{1,\ldots,Q\}} \left(\Gamma_q \left| \Theta^q(\mu) - \Theta^q(\overline{\mu}) - \sum_{p=1}^{P} (\mu_{(p)} - \overline{\mu}_{(p)}) \frac{\partial \Theta^q}{\partial \mu_{(p)}}(\overline{\mu}) \right| \right) , \qquad (4.11)$$

where $\mu \in \mathbb{R}^P$ is denoted $(\mu_{(1)}, \ldots, \mu_{(p)})$.

In short, $\mathcal{F}(\mu - \overline{\mu}; \overline{\mu})$ represents the first-order terms in parameter expansions about $\overline{\mu}$ of $\alpha(\mu)$; and $\Phi(\mu, \overline{\mu})$ is a second-order remainder term that bounds the effect of deviation (of the operator coefficients) from linear parameter dependence.

We can now develop our lower bound $\hat{\alpha}(\mu)$. We first require a parameter sample $V_J \equiv \{\overline{\mu}_1 \in \mathcal{D}, \ldots, \overline{\mu}_J \in \mathcal{D}\}$ and associated sets of polytopes, $P_J \equiv \{\mathcal{P}^{\overline{\mu}_1} \in \mathcal{D}^{\overline{\mu}_1}, \ldots, \mathcal{P}^{\overline{\mu}_J} \in \mathcal{D}^{\overline{\mu}_J}\}$ that satisfy a "Coverage Condition,"

$$\mathcal{D} \subset \bigcup_{j=1}^{J} \mathcal{P}^{\bar{\mu}_j} , \qquad (4.12)$$

and a "Positivity Condition,"

$$\min_{\nu \in \mathcal{V}^{\bar{\mu}_j}} \mathcal{F}(\nu - \bar{\mu}_j; \bar{\mu}_j) - \max_{\mu \in \mathcal{P}^{\bar{\mu}_j}} \Phi(\mu; \bar{\mu}_j) \geq \epsilon_\alpha \alpha(\bar{\mu}_j), \ 1 \leq j \leq J \ . \qquad (4.13)$$

Here $\mathcal{V}^{\bar{\mu}_j}$ is the set of vertices associated with the polytope $\mathcal{P}^{\bar{\mu}_j}$ — for example, $\mathcal{P}^{\bar{\mu}_j}$ may be a simplex with $|\mathcal{V}^{\bar{\mu}_j}| = P+1$ vertices; and $\epsilon_\alpha \in]0, 1[$ is a prescribed accuracy constant. Our lower bound is then given by

$$\hat{\alpha}_{\mathrm{PC}}(\mu) \equiv \max_{j \in \{1,\ldots,J\} | \mu \in \mathcal{P}^{\bar{\mu}_j}} \epsilon_\alpha \alpha(\overline{\mu}_j) \ , \qquad (4.14)$$

which is a piecewise–constant approximation for $\alpha(\mu)$. However, in some cases it is advantageous to define a piecewise–linear approximation

$$\hat{\alpha}_{\mathrm{PL}}(\mu) \equiv \max_{j \in \{1,\ldots,J\} | \mu \in \mathcal{P}^{\bar{\mu}_j}} \left[(1 - \lambda(\mu)) \, \alpha(\overline{\mu}_j) + \lambda(\mu) \epsilon_\alpha \alpha(\overline{\mu}_j) \right] \ ; \qquad (4.15)$$

where $\lambda(\mu)$ is given by

$$\lambda(\mu) = \frac{|\overline{\mu}_j - \mu|}{|\overline{\mu}_j - \mu_j^e|} \ . \tag{4.16}$$

Here μ_j^e is the intersection point between the line $\overline{\mu}_j\mu$ with one of the edges of the polytope $\mathcal{P}^{\overline{\mu}_j}$; and $|\cdot|$ denotes the Euclidean length of a vector; also note that $0 \leq \lambda(\mu) \leq 1$. Finally, we introduce an index mapping $\mathcal{I} : \mathcal{D} \to \{1, \ldots, J\}$ such that for any $\mu \in \mathcal{D}$,

$$\mathcal{I}\mu = \arg \max_{j \in \{1,\ldots,J\}} \epsilon_\alpha \alpha(\overline{\mu}_j), \tag{4.17}$$

for piecewise–constant lower bound; but

$$\mathcal{I}\mu = \arg \max_{j \in \{1,\ldots,J\}} \left[(1 - \lambda(\mu))\, \alpha(\overline{\mu}_j) + \lambda(\mu) \epsilon_\alpha \alpha(\overline{\mu}_j) \right], \tag{4.18}$$

for piecewise–linear lower bound. We can readily show that

4.2.3 Bound Proof

Proposition 2. *For any V_J and P_J such that the Coverage Condition (4.12) and Positivity Condition (4.13) are satisfied, we have $\epsilon_\alpha \alpha\big(\overline{\mu}_{\mathcal{I}\mu}\big) = \hat{\alpha}_{\mathrm{PC}}(\mu) \leq \alpha(\mu)$, $\forall\, \mu \in \mathcal{D}$.*

Proof. To simplify the notation we denote $\overline{\mu}_{\mathcal{I}\mu}$ by $\overline{\mu}$ and note from (4.6) and (4.5) to express $\alpha(\mu)$ as

$$\begin{aligned}
\alpha(\mu) &= \inf_{w \in X} \frac{a(w,w;\overline{\mu}) + \sum_{q=1}^{Q} \left(\Theta^q(\mu) - \Theta^q(\overline{\mu})\right) a^q(w,w)}{\|w\|_X^2} \\
&\geq \inf_{w \in X} \frac{a(w,w;\overline{\mu}) + \sum_{p=1}^{P} \sum_{q=1}^{Q} (\mu_{(p)} - \overline{\mu}_{(p)}) \frac{\partial \Theta^q}{\partial \mu_{(p)}}(\overline{\mu}) a^q(w,w)}{\|w\|_X^2} \\
&\quad + \inf_{w \in X} \frac{\sum_{q=1}^{Q} \left(\Theta^q(\mu) - \Theta^q(\overline{\mu}) - \sum_{p=1}^{P} (\mu_{(p)} - \overline{\mu}_{(p)}) \frac{\partial \Theta^q(\overline{\mu})}{\partial \mu_{(p)}}\right) a^q(w,w)}{\|w\|_X^2} \\
&\geq \mathcal{F}(\mu - \overline{\mu}; \overline{\mu}) \\
&\quad - \sup_{w \in X} \frac{\sum_{q=1}^{Q} \left(\Theta^q(\mu) - \Theta^q(\overline{\mu}) - \sum_{p=1}^{P} (\mu_{(p)} - \overline{\mu}_{(p)}) \frac{\partial \Theta^q(\overline{\mu})}{\partial \mu_{(p)}}\right) a^q(w,w)}{\|w\|_X^2}.
\end{aligned} \tag{4.19}$$

Furthermore, from (4.9), (4.10) and (4.11) we have

$$\begin{aligned}
&\sup_{w\in X}\sum_{q=1}^{Q}\left(\Theta^q(\mu)-\Theta^q(\overline{\mu})-\sum_{p=1}^{P}(\mu_{(p)}-\overline{\mu}_{(p)})\frac{\partial\Theta^q}{\partial\mu_{(p)}}(\overline{\mu})\right)\frac{a^q(w,w)}{\|w\|_X^2}\\
&\quad\leq\ \sup_{w\in X}\sum_{q=1}^{Q}\left|\Theta^q(\mu)-\Theta^q(\overline{\mu})-\sum_{p=1}^{Q}(\mu_{(p)}-\overline{\mu}_{(p)})\frac{\partial\Theta^q}{\partial\mu_{(p)}}(\overline{\mu})\right|\frac{|a^q(w,w)|}{\|w\|_X^2}\\
&\quad\leq\ \max_{q\in\{1,\ldots,Q\}}\left(\Gamma_q\left|\Theta^q(\mu)-\Theta^q(\overline{\mu})-\sum_{p=1}^{Q}(\mu_{(p)}-\overline{\mu}_{(p)})\frac{\partial\Theta^q}{\partial\mu_{(p)}}(\overline{\mu})\right|\right)\sup_{w\in X}\frac{\sum_{q=1}^{Q}|w|_q^2}{\|w\|_X^2}\\
&\quad=\ \Phi(\mu,\overline{\mu})\ . \qquad (4.20)
\end{aligned}$$

We thus conclude from (4.19) and (4.20) that

$$\begin{aligned}
\alpha(\mu)\ &\geq\ \mathcal{F}(\mu-\bar{\mu};\bar{\mu})-\Phi(\mu,\bar{\mu})\\
&\geq\ \min_{\nu\in\mathcal{V}^{\bar{\mu}}}\mathcal{F}(\nu-\bar{\mu};\bar{\mu})-\max_{\mu\in\mathcal{P}^{\bar{\mu}}}\Phi(\mu;\bar{\mu})\\
&\geq\ \epsilon_\alpha\alpha(\bar{\mu})=\hat{\alpha}_{\mathrm{PC}}(\mu) \qquad (4.21)
\end{aligned}$$

from the construction of V_J and P_J, the definition of $\hat{\alpha}_{\mathrm{PC}}(\mu)$, and the concavity of $\mathcal{F}(\mu-\overline{\mu};\overline{\mu})$ in μ. □

In addition, there is a special case that should be exploited to enhance our lower bounds for the stability factor and also to ease computational effort. In particular, if $-\Phi(\mu,\overline{\mu})$ is concave in μ (which can be verified *a priori* for given coefficient functions $\Theta^q(\mu), 1\leq q\leq Q$), we can combine the two functions $\mathcal{F}(\mu-\overline{\mu};\overline{\mu})$ and $-\Phi(\mu,\overline{\mu})$ into the min in the Positivity Condition. Furthermore, we also obtain lower-bound property for our piecewise-linear approximation $\hat{\alpha}_{\mathrm{PL}}(\mu)$. This follows from our proof in Proposition 2 and concavity of the function $\mathcal{F}(\mu-\overline{\mu};\overline{\mu})-\Phi(\mu,\overline{\mu})$ in μ.

Corollary 4.2.3. *If $-\Phi(\mu,\overline{\mu})$ is a concave function of μ in $\mathcal{D}^{\bar{\mu}}$ for all $\bar{\mu}\in\mathcal{D}$, then the Positivity Condition is defined as*

$$\min_{\nu\in\mathcal{V}^{\bar{\mu}_j}}\{\mathcal{F}(\nu-\bar{\mu}_j;\bar{\mu}_j)-\Phi(\mu;\bar{\mu}_j)\}\geq\epsilon_\alpha\alpha(\bar{\mu}_j),\ 1\leq j\leq J\ ; \qquad (4.22)$$

and $\hat{\alpha}_{\mathrm{PL}}(\mu)$ satisfies $\hat{\alpha}_{\mathrm{PC}}(\mu)\leq\hat{\alpha}_{\mathrm{PL}}(\mu)\leq\alpha(\mu)$, $\forall\mu\in\mathcal{D}^{\bar{\mu}_{\mathcal{I}\mu}}$, where $\mathcal{I}\mu$ is determined by (4.18).

It remains to address the question for which functions $\Theta^q(\mu), 1 \le q \le Q$, $-\Phi(\mu, \overline{\mu})$ is a concave function of μ in $\mathcal{D}^{\bar{\mu}}$.

Lemma 4.2.4. *For any $\bar{\mu} \in \mathcal{D}$, if $\Theta^q(\mu), 1 \le q \le Q$, are either concave or convex in $\mathcal{D}^{\bar{\mu}}$ then the function $-\Phi(\mu, \bar{\mu})$ is concave in $\mathcal{D}^{\bar{\mu}}$.*

Proof. Equivalently, we need to show the convexity of $\Phi(\mu, \bar{\mu})$ in $\mathcal{D}^{\bar{\mu}}$. To this end, we note from convex analysis and differentiability of $\Theta^q(\mu)$ that

$$\Theta^q(\mu) \ge \Theta^q(\overline{\mu}) - \sum_{p=1}^{P} (\mu_{(p)} - \overline{\mu}_{(p)}) \frac{\partial \Theta^q}{\partial \mu_{(p)}}(\overline{\mu}), \quad \forall\, \mu \in \mathcal{D}^{\bar{\mu}}\ , \tag{4.23}$$

if $\Theta^q(\mu)$ is convex and that

$$\Theta^q(\mu) \le \Theta^q(\overline{\mu}) - \sum_{p=1}^{P} (\mu_{(p)} - \overline{\mu}_{(p)}) \frac{\partial \Theta^q}{\partial \mu_{(p)}}(\overline{\mu}), \quad \forall\, \mu \in \mathcal{D}^{\bar{\mu}}\ , \tag{4.24}$$

if $\Theta^q(\mu)$ is concave. We thus obtain

$$\begin{aligned} \Psi^q(\mu) \;&\equiv\; \left| \Theta^q(\mu) - \Theta^q(\overline{\mu}) - \sum_{p=1}^{P} (\mu_{(p)} - \overline{\mu}_{(p)}) \frac{\partial \Theta^q}{\partial \mu_{(p)}}(\overline{\mu}) \right| \\ &= \left[\begin{array}{ll} \Theta^q(\mu) - \Theta^q(\overline{\mu}) - \sum_{p=1}^{P} (\mu_{(p)} - \overline{\mu}_{(p)}) \frac{\partial \Theta^q}{\partial \mu_{(p)}}(\overline{\mu}) & \text{if } \Theta^q(\mu) \text{ is convex} \\ \Theta^q(\overline{\mu}) + \sum_{p=1}^{P} (\mu_{(p)} - \overline{\mu}_{(p)}) \frac{\partial \Theta^q}{\partial \mu_{(p)}}(\overline{\mu}) - \Theta^q(\mu) & \text{if } \Theta^q(\mu) \text{ is concave} \end{array} \right. \end{aligned} \tag{4.25}$$

which are convex functions of μ in $\mathcal{D}^{\bar{\mu}}$ for $q = 1, \ldots, Q$. Since furthermore $\Phi(\mu, \bar{\mu})$ defined in (4.11) is a pointwise maximum function, it follows from the convexity of $\Psi^q(\mu), q = 1, \ldots, Q$, that $\Phi(\mu, \bar{\mu})$ is convex in $\mathcal{D}^{\bar{\mu}}$. □

4.3 Lower Bounds for Noncoercive Problems

4.3.1 Inf-Sup Parameter

Again we consider the single stability-factor bound conditioner and also assume that the bilinear form a admits an affine decomposition (4.5). For noncoercive operators, stability

and thus well-posedness is guaranteed by an inf-sup condition

$$0 < \beta_0 \leq \beta(\mu) \equiv \inf_{w \in X} \sup_{v \in X} \frac{a(w, v; \mu)}{\|w\|_X \|v\|_X}, \quad \forall \mu \in \mathcal{D} \ . \tag{4.26}$$

Here $\beta(\mu)$ is the Babuška "inf–sup" parameter — the minimum (generalized) singular value associated with our differential operator. It shall prove convenient to write the stability parameter in terms of a supremizing operator $T^\mu : X \to X$ such that, for any w in X

$$(T^\mu w, v)_X = a(w, v; \mu), \quad \forall \, v \in X \ ; \tag{4.27}$$

it is readily shown by Riesz representation that

$$T^\mu w = \arg \sup_{v \in X} \frac{a(w, v; \mu)}{\|v\|}, \quad \forall \, v \in X \ . \tag{4.28}$$

We then define

$$\sigma(w; \mu) \equiv \frac{\|T^\mu w\|_X}{\|w\|_X} \ , \tag{4.29}$$

and note from (4.26), (4.28), (4.29) that

$$\beta(\mu) = \inf_{w \in X} \sigma(w; \mu) \ . \tag{4.30}$$

In the following, we construct a lower bound for $\beta(\mu)$ for real noncoercive problems. See Appendix C for the lower bound formulation for complex noncoercive problems in which X is the function space over the complex field $\mathbb{C}$ and $\Theta^q : \mathcal{D} \to \mathbb{C}$ are complex functions.

4.3.2 Inf-Sup Lower Bound Formulation

To begin, given $\overline{\mu} \in \mathcal{D}$ and $t = (t_{(1)}, \ldots, t_{(P)}) \in \mathbb{R}^P$, we introduce the bilinear form

$$\begin{aligned} \mathcal{T}(w, v; t; \overline{\mu}) \ &= \ (T^{\overline{\mu}} w, T^{\overline{\mu}} v)_X \\ &\quad + \sum_{p=1}^{P} t_{(p)} \left\{ \sum_{q=1}^{Q} \frac{\partial \Theta^q}{\partial \mu_{(p)}} (\overline{\mu}) \left[a^q(w, T^{\overline{\mu}} v) + a^q(v, T^{\overline{\mu}} w) \right] \right\} \end{aligned} \tag{4.31}$$

and associated Rayleigh quotient

$$\mathcal{F}(t;\overline{\mu}) = \min_{v\in X} \frac{\mathcal{T}(v,v;t;\overline{\mu})}{\|v\|_X^2} \ . \tag{4.32}$$

In a similar argument, we can prove that $\mathcal{F}(t;\overline{\mu})$ is *concave* in t; and hence $\mathcal{D}^{\overline{\mu}} \equiv \left\{\mu \in \mathbb{R}^P \mid \mathcal{F}(\mu-\overline{\mu};\overline{\mu}) \geq 0\right\}$ is perforce convex.

We next assume that a^q are continuous in the sense that there exist positive finite constants Γ_q, $1 \leq q \leq Q$, such that

$$|a^q(w,v)| \leq \Gamma_q \, |w|_q \, |v|_q \, , \quad \forall w, v \in X \ . \tag{4.33}$$

Here $|\cdot|_q : H^1(\Omega) \rightarrow \mathbb{R}^+$ are seminorms that satisfy

$$C_X = \sup_{w\in X} \frac{\sum_{q=1}^{Q} |w|_q^2}{\|w\|_X^2}, \tag{4.34}$$

for some positive parameter-independent constant C_X. We then define, for $\mu \in \mathcal{D}$, $\overline{\mu} \in \mathcal{D}$,

$$\Phi(\mu,\overline{\mu}) \equiv C_X \max_{q\in\{1,\ldots,Q\}} \left(\Gamma_q \left| \Theta^q(\mu) - \Theta^q(\overline{\mu}) - \sum_{p=1}^{P} (\mu_{(p)} - \overline{\mu}_{(p)}) \, \frac{\partial \Theta^q}{\partial \mu_{(p)}}(\overline{\mu}) \right| \right) \ . \tag{4.35}$$

In short, $\mathcal{T}(w,w;\mu-\overline{\mu};\overline{\mu})/\|w\|_X^2$ and $\mathcal{F}(\mu-\overline{\mu};\overline{\mu})$ represent the first-order terms in parameter expansions about $\overline{\mu}$ of $\sigma^2(w;\mu)$ and $\beta^2(\mu)$, respectively; and $\Phi(\mu,\overline{\mu})$ is a second-order remainder term that bounds the effect of deviation (of the operator coefficients) from linear parameter dependence.

We now require a parameter sample $V_J \equiv \{\overline{\mu}_1 \in \mathcal{D}, \ldots, \overline{\mu}_J \in \mathcal{D}\}$ and associated sets of polytopes, $P_J \equiv \{\mathcal{P}^{\overline{\mu}_1} \in \mathcal{D}^{\overline{\mu}_1}, \ldots, \mathcal{P}^{\overline{\mu}_J} \in \mathcal{D}^{\overline{\mu}_J}\}$ that satisfy a "Coverage Condition,"

$$\mathcal{D} \subset \bigcup_{j=1}^{J} \mathcal{P}^{\overline{\mu}_j} \ , \tag{4.36}$$

and a "Positivity Condition,"

$$\min_{\nu \in \mathcal{V}^{\bar{\mu}_j}} \sqrt{\mathcal{F}(\nu - \bar{\mu}_j; \bar{\mu}_j)} - \max_{\mu \in \mathcal{P}^{\bar{\mu}_j}} \Phi(\mu; \bar{\mu}_j) \geq \epsilon_\beta \beta(\bar{\mu}_j), \ 1 \leq j \leq J \ . \tag{4.37}$$

Here $\mathcal{V}^{\bar{\mu}_j}$ is the set of vertices associated with the polytope $\mathcal{P}^{\bar{\mu}_j}$; and $\epsilon_\beta \in]0,1[$ is a prescribed accuracy constant. Our lower bound is then given by

$$\hat{\beta}_{\text{PC}}(\mu) \equiv \max_{j \in \{1,\ldots,J\} | \mu \in \mathcal{P}^{\bar{\mu}_j}} \epsilon_\beta \beta(\overline{\mu}_j). \tag{4.38}$$

which is a piecewise–constant approximation for $\beta(\mu)$. However, in some cases it is possible to define a piecewise–linear approximation

$$\hat{\beta}_{\text{PL}}(\mu) \equiv \max_{j \in \{1,\ldots,J\} | \mu \in \mathcal{P}^{\bar{\mu}_j}} \left[(1 - \lambda(\mu)) \, \beta(\overline{\mu}_j) + \lambda(\mu) \epsilon_\beta \beta(\overline{\mu}_j) \right] \ , \tag{4.39}$$

where as before $\lambda(\mu)$ is given by

$$\lambda(\mu) = \frac{|\overline{\mu}_j - \mu|}{|\overline{\mu}_j - \mu_j^e|} \ . \tag{4.40}$$

Here μ_j^e is the intersection point between the line $\overline{\mu}_j \mu$ with one of the edges of the polytope $\mathcal{P}^{\bar{\mu}_j}$. We finally introduce an index mapping $\mathcal{I} : \mathcal{D} \to \{1, \ldots, J\}$ such that for any $\mu \in \mathcal{D}$,

$$\mathcal{I}\mu = \arg \max_{j \in \{1,\ldots,J\}} \epsilon_\beta \beta(\overline{\mu}_j), \tag{4.41}$$

for piecewise–constant lower bound; but

$$\mathcal{I}\mu = \arg \max_{j \in \{1,\ldots,J\}} \left[(1 - \lambda(\mu)) \, \beta(\overline{\mu}_j) + \lambda(\mu) \epsilon_\beta \beta(\overline{\mu}_j) \right], \tag{4.42}$$

for piecewise–linear lower bound. We can readily demonstrate that

4.3.3 Bound Proof

Proposition 3. *For any V_J and P_J such that the Coverage Condition (4.36) and Positivity Condition (4.37) are satisfied, we have $\epsilon_\beta \beta\big(\overline{\mu}_{\mathcal{I}\mu}\big) = \hat{\beta}_{\text{PC}}(\mu) \leq \beta(\mu)$, $\forall \, \mu \in \mathcal{D}^{\overline{\mu}_{\mathcal{I}\mu}}$.*

Proof. To simplify the notation we denote $\overline{\mu}_{\mathcal{I}\mu}$ by $\overline{\mu}$ and express $T^\mu w = T^{\overline{\mu}} w + (T^\mu w -$

$T^{\overline{\mu}}w$) as in (4.29) to obtain

$$\sigma^2(w;\mu) = \{\|T^{\overline{\mu}}w\|_X^2 + \|T^{\mu}w - T^{\overline{\mu}}w\|_X^2 + 2(T^{\mu}w - T^{\overline{\mu}}w, T^{\overline{\mu}}w)_X\}/\|w\|_X^2 \ . \tag{4.43}$$

We next note from (4.5) and (4.27) that, for $t = \mu - \overline{\mu}$,

$$\begin{aligned}(T^{\mu}w - T^{\overline{\mu}}w, T^{\overline{\mu}}w)_X &= \sum_{q=1}^{Q}(\Theta^q(\mu) - \Theta^q(\overline{\mu}))a^q(w, T^{\overline{\mu}}w) \\ &= \sum_{p=1}^{P}\sum_{q=1}^{Q} t_{(p)}\frac{\partial\Theta^q}{\partial\mu_{(p)}}(\overline{\mu})a^q(w, T^{\overline{\mu}}w) \\ &+ \sum_{q=1}^{Q}\left(\Theta^q(\mu) - \Theta^q(\overline{\mu}) - \sum_{p=1}^{P} t_{(p)}\frac{\partial\Theta^q(\overline{\mu})}{\partial\mu_{(p)}}\right)a^q(w, T^{\overline{\mu}}w) \end{aligned} \tag{4.44}$$

Furthermore, it follows from (4.33), (4.34), (4.35), triangle inequality, and Cauchy-Schwarz inequality that

$$\begin{aligned}&\left|\sum_{q=1}^{Q}\left(\Theta^q(\mu) - \Theta^q(\overline{\mu}) - \sum_{p=1}^{P}(\mu_{(p)} - \overline{\mu}_{(p)})\frac{\partial\Theta^q}{\partial\mu_{(p)}}(\overline{\mu})\right)a^q(w, T^{\overline{\mu}}w)\right| \\ &\quad\leq \max_{q\in\{1,\ldots,Q\}}\left(\Gamma_q\left|\Theta^q(\mu) - \Theta^q(\overline{\mu}) - \sum_{p=1}^{P}(\mu_{(p)} - \overline{\mu}_{(p)})\frac{\partial\Theta^q}{\partial\mu_{(p)}}(\overline{\mu})\right|\right)\sum_{q=1}^{Q}|w|_q|T^{\overline{\mu}}w|_q \\ &\quad\leq \Phi(\mu,\overline{\mu})||w||_X||T^{\overline{\mu}}w||_X \\ &\quad\leq \Phi(\mu,\overline{\mu})||w||_X\left(||T^{\mu}w||_X + ||T^{\overline{\mu}}w - T^{\mu}w||_X\right) \\ &\quad\leq \Phi(\mu,\overline{\mu})||w||_X||T^{\mu}w||_X + \frac{1}{2}||T^{\overline{\mu}}w - T^{\mu}w||_X^2 + \frac{1}{2}\Phi^2(t;\overline{\mu})||w||_X^2. \end{aligned} \tag{4.45}$$

We thus conclude from (4.31), (4.43), (4.44), and (4.45) that

$$\sigma^2(w;\mu) \geq \frac{\mathcal{T}(w,w;t;\overline{\mu})}{\|w\|_X^2} - 2\Phi(\mu,\overline{\mu})\,\sigma(w;\mu) - \Phi^2(t;\overline{\mu}) \ . \tag{4.46}$$

We now solve the quadratic inequality to obtain

$$\sigma(w;\mu) \geq \left[\mathcal{T}(w,w;\mu-\overline{\mu};\mu)/\|w\|_X^2\right]^{\frac{1}{2}} - \Phi(\mu,\overline{\mu}) \ . \tag{4.47}$$

It immediately follows from (4.30, (4.32) and (4.47) that

$$\beta(\mu) \geq \sqrt{\mathcal{F}(\mu-\overline{\mu};\overline{\mu})} - \Phi(\mu,\overline{\mu}) \ . \tag{4.48}$$

The desired result finally follows from the construction of V_J and P_J, the definition of $\hat{\beta}_{\mathrm{PC}}(\mu)$ and $\hat{\beta}_{\mathrm{PL}}(\mu)$, and the concavity of $\mathcal{F}(\mu-\overline{\mu};\overline{\mu})$ in μ. □

In fact, we can prove Proposition 3 by following a straightforward (simpler) route

$$\begin{aligned}
\beta(\mu) &\geq \inf_{w\in X} \frac{a(w,T^{\overline{\mu}}w;\mu)}{\|w\|_X\|T^{\overline{\mu}}w\|_X} \\
&= \inf_{w\in X} \frac{a(w,T^{\overline{\mu}}w;\bar{\mu}) + a(w,T^{\overline{\mu}}w;\mu) - a(w,T^{\overline{\mu}}w;\bar{\mu})}{\|w\|_X\|T^{\overline{\mu}}w\|_X} \\
&= \inf_{w\in X} \left\{ \frac{(T^{\overline{\mu}}w,T^{\overline{\mu}}w)_X + \sum_{p=1}^{P}\sum_{q=1}^{Q}(\mu_{(p)}-\bar{\mu}_{(p)})\frac{\partial\Theta^q}{\partial\mu_{(p)}}(\overline{\mu})a^q(w,T^{\overline{\mu}}w)}{\|w\|_X\|T^{\overline{\mu}}w\|_X} \right. \\
&\quad \left. + \frac{\sum_{q=1}^{Q}\left(\Theta^q(\mu)-\Theta^q(\overline{\mu})-\sum_{p=1}^{P}(\mu_{(p)}-\bar{\mu}_{(p)})\frac{\partial\Theta^q(\overline{\mu})}{\partial\mu_{(p)}}\right)a^q(w,T^{\overline{\mu}}w)}{\|w\|_X\|T^{\overline{\mu}}w\|_X} \right\} \\
&\geq \inf_{w\in X} \frac{(T^{\overline{\mu}}w,T^{\overline{\mu}}w)_X + \sum_{p=1}^{P}\sum_{q=1}^{Q}(\mu_{(p)}-\bar{\mu}_{(p)})\frac{\partial\Theta^q}{\partial\mu_{(p)}}(\overline{\mu})a^q(w,T^{\overline{\mu}}w)}{\|w\|_X\|T^{\overline{\mu}}w\|_X} \\
&\quad - \sup_{w\in X} \frac{\sum_{q=1}^{Q}\left(\Theta^q(\mu)-\Theta^q(\overline{\mu})-\sum_{p=1}^{P}(\mu_{(p)}-\bar{\mu}_{(p)})\frac{\partial\Theta^q(\overline{\mu})}{\partial\mu_{(p)}}\right)a^q(w,T^{\overline{\mu}}w)}{\|w\|_X\|T^{\overline{\mu}}w\|_X} \\
&\geq \sqrt{\inf_{w\in X} \frac{\left((T^{\overline{\mu}}w,T^{\overline{\mu}}w)_X + \sum_{p=1}^{P}\sum_{q=1}^{Q}(\mu_{(p)}-\bar{\mu}_{(p)})\frac{\partial\Theta^q}{\partial\mu_{(p)}}(\overline{\mu})a^q(w,T^{\overline{\mu}}w)\right)^2}{\|w\|_X^2\|T^{\overline{\mu}}w\|_X^2}} \\
&\quad - \max_{q\in\{1,\ldots,Q\}}\left(\Gamma_q\left|\Theta^q(\mu)-\Theta^q(\overline{\mu})-\sum_{p=1}^{P}(\mu-\overline{\mu})_{(p)}\frac{\partial\Theta^q(\overline{\mu})}{\partial\mu_{(p)}}\right|\right)\sup_{w\in X}\frac{\sum_{q=1}^{Q}|w|_q|T^{\overline{\mu}}w|_q}{\|w\|_X\|T^{\overline{\mu}}w\|_X} \\
&\geq \left[\mathcal{T}(w,w;\mu-\overline{\mu};\mu)/\|w\|_X^2\right]^{\frac{1}{2}} - \Phi(\mu,\overline{\mu})
\end{aligned}$$

from (4.5), (4.27), (4.31), (4.43), (4.44), (4.45), and Cauchy-Schwarz inequality.

By a similar argument as for the coercive case, we can also state

Corollary 4.3.1. *If $-\Phi(\mu,\overline{\mu})$ is a concave function of μ in $\mathcal{D}^{\bar{\mu}}$ for all $\bar{\mu}$ in $\mathcal{D}$, then the Positivity Condition is defined as*

$$\min_{\nu\in\mathcal{V}^{\bar{\mu}_j}}\{\sqrt{\mathcal{F}(\nu-\bar{\mu}_j;\bar{\mu}_j)} - \Phi(\mu;\bar{\mu}_j)\} \geq \epsilon_\beta\beta(\bar{\mu}_j),\ 1\leq j\leq J\ ; \tag{4.49}$$

and $\hat{\beta}_{\rm PL}(\mu)$ *satisfies* $\hat{\beta}_{\rm PC}(\mu) \leq \hat{\beta}_{\rm PL}(\mu) \leq \beta(\mu), \forall \mu \in \mathcal{D}^{\overline{\mu}_{\mathcal{I}_\mu}}$, *where* $\mathcal{I}\mu$ *is determined by (4.42).*

In Lemma 4.2.4, we demonstrate that if $\Theta^q(\mu), q = 1, \ldots, Q$, are either convex or concave in $\mathcal{D}^{\bar{\mu}}$ then $-\Phi(\mu, \overline{\mu})$ is concave in $\mathcal{D}^{\bar{\mu}}$.

4.3.4 Discrete Eigenvalue Problems

We address the numerical calculation of the inf-sup parameter $\beta(\mu)$ and the Rayleigh quotient $\mathcal{F}(\mu - \bar{\mu}; \bar{\mu})$. To begin, we denote by $\underline{A}(\mu)$, $\underline{A}^q$, $\underline{C}$ the finite element matrices associated with $a(\cdot, \cdot; \mu)$, $a^q(\cdot, \cdot)$, $(\cdot, \cdot)_X$, respectively. We introduce the discrete eigenproblem: Given $\mu \in \mathcal{D}$, find the minimum eigenmode $(\underline{\chi}_{\rm min}(\mu), \lambda_{\rm min}(\mu))$ such that

$$(\underline{A}(\mu))^T \underline{C}^{-1} \underline{A}(\mu) \underline{\chi}_{\rm min}(\mu) = \lambda_{\rm min}(\mu) \underline{C} \underline{\chi}_{\rm min}(\mu) \ , \tag{4.50}$$

$$(\underline{\chi}_{\rm min}(\mu))^T \underline{C} \underline{\chi}_{\rm min}(\mu) = 1 \ . \tag{4.51}$$

The discrete value of $\beta(\mu)$ is then $\sqrt{\lambda_{\rm min}(\mu)}$.

The computation of $\mathcal{F}$ involving a more complex eigenvalue problem is rather complicated and more expensive. In particular, we first write the matrix form $\underline{\mathcal{T}}(\mu - \bar{\mu}; \bar{\mu})$ of the bilinear form $\mathcal{T}(\cdot, \cdot; \mu - \bar{\mu}; \bar{\mu})$ in (4.31) as

$$\begin{aligned} \underline{\mathcal{T}}(\mu - \bar{\mu}; \bar{\mu}) &= (\underline{A}(\overline{\mu}))^T \underline{C}^{-1} \underline{A}(\overline{\mu}) \\ &+ \sum_{p=1}^{P} (\mu - \bar{\mu})_{(p)} \left\{ \sum_{q=1}^{Q} \frac{\partial \Theta^q(\overline{\mu})}{\partial \mu_{(p)}} \left[(\underline{A}^q)^T \underline{C}^{-1} \underline{A}(\overline{\mu}) + (\underline{A}(\overline{\mu}))^T \underline{C}^{-1} \underline{A}^q \right] \right\} . \end{aligned}$$

Next we introduce the second discrete eigenproblem: Given a pair $(\mu \in \mathcal{D}, \bar{\mu} \in \mathcal{D})$, find $\underline{\Psi}_{\rm min}(\mu - \bar{\mu}; \bar{\mu}) \in \mathbb{R}^{\mathcal{N}}$, $\rho_{\rm min}(\mu - \bar{\mu}; \bar{\mu}) \in \mathbb{R}$ such that

$$\underline{\mathcal{T}}(\mu - \bar{\mu}; \bar{\mu}) \, \underline{\Psi}_{\rm min}(\mu - \bar{\mu}; \bar{\mu}) = \rho_{\rm min}(\mu - \bar{\mu}; \bar{\mu}) \underline{C} \, \underline{\Psi}_{\rm min}(\mu - \bar{\mu}; \bar{\mu}) \ , \tag{4.52}$$

$$(\underline{\Psi}_{\rm min}(\mu - \bar{\mu}; \bar{\mu}))^T \, \underline{C} \, \underline{\Psi}_{\rm min}(\mu - \bar{\mu}; \bar{\mu}) = 1 \ . \tag{4.53}$$

And $\mathcal{F}(\mu - \bar{\mu}; \bar{\mu})$ is essentially the minimum eigenvalue $\rho_{\rm min}(\mu - \bar{\mu}; \bar{\mu})$.

We see that the two discrete eigenproblems involve the invert matrix $\underline{C}^{-1}$. However,

in our actual implementation, the two eigenproblems can be addressed efficiently by the Lanczos procedure without calculating $\underline{C}^{-1}$ explicitly (see Appendix B for the Lanczos procedure). Take the first eigenproblem for example, during the Lanczos procedure we often compute $\underline{w}(\mu) = (\underline{A}(\mu))^T \underline{C}^{-1} \underline{A}(\mu)\underline{v}$ for some $\underline{v}$ and do this as follows: solve the linear system $\underline{C}\underline{y}(\mu) = \underline{A}(\mu)\underline{v}$ for $\underline{y}(\mu)$ and simply set $\underline{w}(\mu) = (\underline{A}(\mu))^T \underline{y}(\mu)$.

4.4 Choice of Bound Conditioner and Seminorms

In this section, we shall give a general guideline how to select appropriate bound conditioner $(\cdot;\cdot)_X$ and seminorms $|\cdot|_q$ such that the associated constants $\Gamma_q, 1 \le q \le Q$, and C_X are small. For simplicity of exposition, we confine our demonstration to two-dimensional problems. The results for three-dimensional problems can be similarly derived. It shall prove useful to have a summation convention that repeated subscript indices imply summation, and unless otherwise indicated, subscript indices take on integers 1 through 2.

4.4.1 Poisson Problems

We are concerned with defining bound conditioner and seminorms for the following bilinear form

$$a(w, v; \mu) = \int_\Omega C_{ij}(\mu) \frac{\partial v}{\partial x_i} \frac{\partial w}{\partial x_j} + D(\mu) v w \ . \tag{4.54}$$

Here $C_{ij}(\mu)$ and $D(\mu)$ are parameter-dependent coefficient functions; note that we permit negative value of $D(\mu)$ and in such case arrive at the noncoercive operator. More generally, we consider inhomogeneous physical domain Ω which consists of R non-overlapping homogeneous subdomains Ω^r such that $\overline{\Omega} = \bigcup_{r=1}^R \overline{\Omega}^r$ ($\overline{\Omega}$ denotes the closure of Ω). The bilinear form a is thus given by

$$a(w, v; \mu) = \sum_{r=1}^R \int_{\Omega^r} C_{ij}^r(\mu) \frac{\partial v}{\partial x_i} \frac{\partial w}{\partial x_j} + D^r(\mu) v w \ . \tag{4.55}$$

Assuming that the tensors $C_{ij}^r(\mu), 1 \le r \le R$, are symmetric, we next rearrange (4.55) to obtain the desired form (4.5) with

$$a^{q(1,r)}(w,v) = \int_{\Omega^r} \frac{\partial w}{\partial x_1}\frac{\partial v}{\partial x_2} + \frac{\partial w}{\partial x_2}\frac{\partial v}{\partial x_1}, \quad \Theta^{q(1,r)}(\mu) = C_{12}^r(\mu) = C_{21}^r(\mu) \ ,$$

$$a^{q(2,r)}(w,v) = \int_{\Omega^r} \frac{\partial w}{\partial x_1}\frac{\partial v}{\partial x_1}, \quad \Theta^{q(2,r)}(\mu) = C_{11}^r(\mu) \ ,$$

$$a^{q(3,r)} = \int_{\Omega^r} \frac{\partial w}{\partial x_2}\frac{\partial v}{\partial x_2}, \quad \Theta^{q(3,r)}(\mu) = C_{22}^r(\mu) \ ,$$

$$a^{q(4,r)}(w,v) = \int_{\Omega^r} wv, \quad \Theta^{q(4,r)}(\mu) = D^r(\mu) \ ,$$

for $q : \{1,\dots,4\} \times \{1,\dots,R\} \to \{1,\dots,Q\}$. We then define associated seminorms

$$|w|^2_{q(1,r)} = \int_{\Omega^r} \left(\frac{\partial w}{\partial x_1}\right)^2 + \left(\frac{\partial w}{\partial x_2}\right)^2, \qquad |w|^2_{q(2,r)} = \int_{\Omega^r} \left(\frac{\partial w}{\partial x_1}\right)^2, \tag{4.56}$$

$$|w|^2_{q(3,r)} = \int_{\Omega^r} \left(\frac{\partial w}{\partial x_2}\right)^2, \qquad |w|^2_{q(4,r)} = \int_{\Omega^r} w^2 \ . \tag{4.57}$$

By using Cauchy-Schwarz inequality, we find $\Gamma_q = 1, 1 \le q \le Q$ as follows

$$\begin{aligned} a^{q(1,r)}(w,v) &= \int_{\Omega^r} \frac{\partial w}{\partial x_1}\frac{\partial v}{\partial x_2} + \frac{\partial w}{\partial x_2}\frac{\partial v}{\partial x_1} \\ &\le \sqrt{\int_{\Omega^r} \left(\frac{\partial w}{\partial x_1}\right)^2}\sqrt{\int_{\Omega^r} \left(\frac{\partial v}{\partial x_2}\right)^2} + \sqrt{\int_{\Omega^r} \left(\frac{\partial w}{\partial x_2}\right)^2}\sqrt{\int_{\Omega^r} \left(\frac{\partial v}{\partial x_1}\right)^2} \\ &\le \sqrt{\int_{\Omega^r} \left(\frac{\partial w}{\partial x_1}\right)^2 + \left(\frac{\partial w}{\partial x_2}\right)^2}\sqrt{\int_{\Omega^r} \left(\frac{\partial v}{\partial x_1}\right)^2 + \left(\frac{\partial v}{\partial x_2}\right)^2} = |w|_{q(1,r)}\,|v|_{q(1,r)} \end{aligned}$$

for $1 \le r \le R$; furthermore, the remaining bilinear forms are positive-semidefinite and thus satisfy $a^q(w,v) \le \sqrt{a^q(w,w)}\sqrt{a^q(v,v)} = |w|_q|v|_q$.

Finally, we define our bound conditioner as

$$(w,v)_X = \int_{\Omega} \frac{\partial w}{\partial x_1}\frac{\partial v}{\partial x_1} + \frac{\partial w}{\partial x_2}\frac{\partial v}{\partial x_2} + wv \ , \tag{4.58}$$

which is simply the standard $H^1(\Omega)$ inner product. We thus obtain

$$C_X = \sup_{w \in X} \frac{\sum_{q(1,r)}^{Q} |w|^2_{q(1,r)}}{\|w\|^2_X} = \sup_{w \in X} \frac{\int_\Omega 2\,|\nabla w|^2 + w^2}{\int_\Omega |\nabla w|^2 + w^2} \leq 2\ . \tag{4.59}$$

4.4.2 Elasticity Problems

We consider here *plane elasticity* problems. A similar derivation can be easily carried out for more general cases including three-dimensional elasticity problems. In particular, we wish to choose bound conditioner and seminorms for the following elasticity operator

$$a(w, v; \mu) = \sum_{r=1}^{R} \int_{\Omega^r} \frac{\partial v_i}{\partial x_j} C^r_{ijk\ell}(\mu) \frac{\partial w_k}{\partial x_\ell} + D^r_i(\mu) v_i w_i\ . \tag{4.60}$$

Here Ω consists of R non-overlapping homogeneous subdomains Ω^r such that $\overline{\Omega} = \bigcup_{r=1}^{R} \overline{\Omega}^r$; $C^r_{ijk\ell}(\mu)$ is the elasticity tensor and $D^r_i(\mu)$ is related to frequency and material quantity such as density; both of them are parameter-dependent and $D^r_i(\mu)$ can be negative. We assume that the tensors $C^r_{ijk\ell}(\mu)$ are symmetric such that a comprises the following parameter-independent bilinear forms

$$a^{q(1,r)}(w,v) = c^r_1 \int_{\Omega^r} \left(\frac{\partial v_1}{\partial x_1}\frac{\partial w_2}{\partial x_2} + \frac{\partial v_2}{\partial x_2}\frac{\partial w_1}{\partial x_1} \right), \quad a^{q(2,r)}(w,v) = c^r_2 \int_{\Omega^r} \left(\frac{\partial v_1}{\partial x_2}\frac{\partial w_2}{\partial x_1} + \frac{\partial v_2}{\partial x_1}\frac{\partial w_1}{\partial x_2} \right),$$

$$a^{q(3,r)}(w,v) = c^r_3 \int_{\Omega^r} \left(\frac{\partial v_1}{\partial x_1}\frac{\partial w_1}{\partial x_1} \right), \qquad a^{q(4,r)}(w,v) = c^r_4 \int_{\Omega^r} \left(\frac{\partial v_2}{\partial x_1}\frac{\partial w_2}{\partial x_1} \right),$$

$$a^{q(5,r)}(w,v) = c^r_5 \int_{\Omega^r} \left(\frac{\partial v_2}{\partial x_2}\frac{\partial w_2}{\partial x_2} \right), \qquad a^{q(6,r)}(w,v) = c^r_6 \int_{\Omega^r} \left(\frac{\partial v_1}{\partial x_2}\frac{\partial w_1}{\partial x_2} \right),$$

$$a^{q(7,r)}(w,v) = c^r_7 \int_{\Omega^r} w_1 v_1, \qquad a^{q(8,r)}(w,v) = c^r_8 \int_{\Omega^r} w_2 v_2\ ,$$

for $q : \{1,\ldots,8\} \times \{1,\ldots,R\} \to \{1,\ldots,Q\}$, where $c^r_1, \ldots, c^r_8$ are positive constants. We next introduce associated seminorms

$$|w|^2_{q(1,r)} = c^r_1 \int_{\Omega^r} \left(\frac{\partial w_1}{\partial x_1} \right)^2 + \left(\frac{\partial w_2}{\partial x_2} \right)^2, \quad |w|^2_{q(2,r)} = c^r_2 \int_{\Omega^r} \left(\frac{\partial w_2}{\partial x_1} \right)^2 + \left(\frac{\partial w_1}{\partial x_2} \right)^2, \tag{4.61}$$

$$|w|^2_{q(3,r)} = c_3^r \int_{\Omega^r} \left(\frac{\partial w_1}{\partial x_1}\right)^2, \quad |w|^2_{q(4,r)} = c_4^r \int_{\Omega^r} \left(\frac{\partial w_2}{\partial x_1}\right)^2, \tag{4.62}$$

$$|w|^2_{q(5,r)} = c_5^r \int_{\Omega^r} \left(\frac{\partial w_1}{\partial x_2}\right)^2, \quad |w|^2_{q(6,r)} = c_6^r \int_{\Omega^r} \left(\frac{\partial w_2}{\partial x_2}\right)^2, \tag{4.63}$$

$$|w|^2_{q(7,r)} = c_7^r \int_{\Omega^r} w_1^2, \quad |w|^2_{q(8,r)} = c_8^r \int_{\Omega^r} w_2^2 \ . \tag{4.64}$$

Again by using Cauchy-Schwarz inequality, we obtain $\Gamma_q = 1, 1 \le q \le Q$, as follows

$$\begin{aligned}
a^{q(1,r)}(w,v) &= c_1^r \int_{\Omega^r} \left(\frac{\partial v_1}{\partial x_1}\frac{\partial w_2}{\partial x_2} + \frac{\partial v_2}{\partial x_2}\frac{\partial w_1}{\partial x_1}\right) \\
&\le c_1^r \sqrt{\int_{\Omega^r} \left(\frac{\partial v_1}{\partial x_1}\right)^2} \sqrt{\int_{\Omega^r} \left(\frac{\partial w_2}{\partial x_2}\right)^2} + c_1^r \sqrt{\int_{\Omega^r} \left(\frac{\partial v_2}{\partial x_2}\right)^2} \sqrt{\int_{\Omega^r} \left(\frac{\partial w_1}{\partial x_1}\right)^2} \\
&\le c_1^r \sqrt{\int_{\Omega^r} \left(\frac{\partial v_1}{\partial x_1}\right)^2 + \left(\frac{\partial v_2}{\partial x_2}\right)^2} \sqrt{\int_{\Omega^r} \left(\frac{\partial w_1}{\partial x_1}\right)^2 + \left(\frac{\partial w_2}{\partial x_2}\right)^2} \\
&= |w|_{q(1,r)} |v|_{q(1,r)} \ ,
\end{aligned}$$

$$\begin{aligned}
a^{q(2,r)}(w,v) &= c_2^r \int_{\Omega^r} \left(\frac{\partial v_1}{\partial x_2}\frac{\partial w_2}{\partial x_1} + \frac{\partial v_2}{\partial x_1}\frac{\partial w_1}{\partial x_2}\right) \\
&\le c_2^r \sqrt{\int_{\Omega^r} \left(\frac{\partial v_1}{\partial x_2}\right)^2} \sqrt{\int_{\Omega^r} \left(\frac{\partial w_2}{\partial x_1}\right)^2} + c_2^r \sqrt{\int_{\Omega^r} \left(\frac{\partial v_2}{\partial x_1}\right)^2} \sqrt{\int_{\Omega^r} \left(\frac{\partial w_1}{\partial x_2}\right)^2} \\
&\le c_2^r \sqrt{\int_{\Omega^r} \left(\frac{\partial v_1}{\partial x_2}\right)^2 + \left(\frac{\partial v_2}{\partial x_1}\right)^2} \sqrt{\int_{\Omega^r} \left(\frac{\partial w_1}{\partial x_2}\right)^2 + \left(\frac{\partial w_2}{\partial x_1}\right)^2} \\
&= |w|_{q(2,r)} |v|_{q(2,r)} \ ,
\end{aligned}$$

for $1 \le r \le R$; the other bilinear forms are positive-semidefinite and thus satisfy $a^q(w,v) \le \sqrt{a^q(w,w)}\sqrt{a^q(v,v)} = |w|_q |v|_q$.

Finally, we define our bound conditioner as

$$\begin{aligned}
(w,v)_X &= \sum_{r=1}^{R} \int_{\Omega^r} c_3^r \frac{\partial v_1}{\partial x_1}\frac{\partial w_1}{\partial x_1} + c_4^r \frac{\partial v_2}{\partial x_1}\frac{\partial w_2}{\partial x_1} + c_5^r \frac{\partial v_1}{\partial x_2}\frac{\partial w_1}{\partial x_2} \\
&\quad + c_6^r \frac{\partial v_2}{\partial x_2}\frac{\partial w_2}{\partial x_2} + c_7^r w_1 v_1 + c_8^r w_2 v_2 \ .
\end{aligned} \tag{4.65}$$

The associated parameter-independent continuity constant is thus bounded by

$$C_X = \sup_{w \in X} \frac{\sum_{q=1}^{Q} |w|_q^2}{\|w\|_X^2} \leq \max_{1 \leq r \leq R} \left\{ \frac{c_1^r + c_3^r}{c_3^r}, \frac{c_2^r + c_4^r}{c_4^r}, \frac{c_1^r + c_5^r}{c_5^r}, \frac{c_2^r + c_6^r}{c_6^r} \right\}. \tag{4.66}$$

In the case of an isotropic medium, c_1^r and c_2^r are typically smaller than $c_3^r, \ldots, c_6^r, 1 \leq r \leq R$, the continuity constant C_X will thus be less than 2. Note also that C_X may be deduced analytically for simple cases; however, for most problems, it can be computed sharply numerically as a maximum eigenvalue of an eigenproblem.

4.4.3 Remarks

In conclusion, we may select seminorms and define bound conditioner such that $\Gamma_q = 1, 1 \leq q \leq Q$ and $C_X = O(1)$. However, there are certain special structure of the bilinear form a that can be exploited to obtain tighter bound for C_X. In particular, we observe that if a is affine such that

$$a(w, v; \mu) = \Theta^1 a^1(w, v) + \sum_{q=2}^{Q} \Theta^q(\mu) a^q(w, v) \; ; \tag{4.67}$$

where Θ^1 is a "constant", then the $q = 1$ contribution to the sum in (4.31) and (4.34) may be discarded. We may therefore obtain a sharper value of C_X. For example, when a geometric affine transformation involving only dilation and translation from a μ-dependent domain to a fixed reference domain is applied, it appears that the Θ^q associated with the "cross terms" $a^q(w, v)$ (e.g., $a^{q(1,r)}(w, v)$ in Poisson problems and $a^{q(1,r)}(w, v), a^{q(2,r)}(w, v)$ in Elasticity problems) are independent of μ. In such case, we sum these bilinear forms into a^1 and need only to define seminorms for the remaining bilinear forms. This leads to $C_X = 1$, since $(w, w)_X = \sum_{q=2}^{Q} a^q(w, w)$. In fact, several numerical examples in this and subsequent chapters support it. This observation is so important that we formally state

Corollary 4.4.1. (Dilation-Translation Corollary) *In the above definition of bound conditioner and seminorms, if the coefficient functions $\Theta^q, q = 1, \ldots, Q_C < Q$, associated with the cross terms $a^q(\cdot; \cdot), q = 1, \ldots, Q_C$, are parameter-independent, then C_X is unity.*

4.5 Lower Bound Construction

In this section, we discuss the construction of our lower bounds for noncoercive operators. Application of the following development to coercive operators is straightforward.

4.5.1 Offline/Online Computational Procedure

We now turn to the offline/online decomposition. The *offline* stage comprises two parts: the *generation* of a set of points and polytopes/vertices, $\bar{\mu}_j$ and $\mathcal{P}^{\bar{\mu}_j}$, $\mathcal{V}^{\bar{\mu}_j}$, $1 \le j \le J$; and the *verification* that (4.36) (trivial) and (4.37) (nontrivial) are indeed satisfied. We first focus on verification. To verify (4.37), the essential observation is that the expensive terms — "truth" eigenproblems associated with $\mathcal{F}$, (4.32), and β, (4.30) — are limited to a finite set of *vertices*,

$$J + \sum_{j=1}^{J} |\mathcal{V}^{\bar{\mu}_j}|$$

in total; only for the extremely inexpensive — and typically algebraically very simple — $\Phi(\mu; \bar{\mu}_j)$ terms must we consider maximization over the *polytopes*. The dominant computational cost is thus $\sum_{j=1}^{J} |\mathcal{V}^{\bar{\mu}_j}|$ $\mathcal{F}$-solves and J β-solves. Next, we create a search/look-up table of size $J \times P$ which has row j storing $\bar{\mu}_j$ while column p storing the p^{th} component of the vector $\bar{\mu}_j$ and is ordered such that $\bar{\mu}_j \le \bar{\mu}_i$ for $j \le i$;[1] furthermore, we assign each $\bar{\mu}_j$ a list I_j containing indices of its "neighbors" (i.e., if $i \in I_j$ then μ_i is neighboring to μ_j). The generation is rather complicated and left for the next subsection.

Fortunately, the *online* stage (4.38)-(4.39) is very simple: for a given new parameter μ, we conduct a binary chop search (with cost $\log J$) for an index j such that $\mu \in [\bar{\mu}_j, \bar{\mu}_{j+1}]$ and then check (with cost polynomial in P) all possible polytopes $\mathcal{P}_i, i \in I_j \cup I_{j+1}$, which contain the parameter μ.

[1] By placing the most significant weight to the first component and the least significant weight to the last component of a vector, we can compare two "vectors" in the same way as two numbers. For example, the vector (2, 9, 1) is greater than (2, 8, 12) since their first components are equal and the second component of the first vector is greater than the second component of the second vector.

4.5.2 Generation Algorithm

The offline eigenvalue problems (4.30) and (4.32) can be rather nasty due to the generalized nature of our singular value (note $T^{\overline{\mu}}$ involves the *inverse* Laplacian) and the presence of a continuous component to the spectrum. However, effective computational strategies can be developed by making use of inexpensive surrogates for $\beta(\mu)$ and in particular $\mathcal{F}(\mu - \overline{\mu}; \overline{\mu})$. We assume that we may compute efficiently accurate surrogates $\tilde{\beta}(\mu)$ for $\beta(\mu)$ and $\tilde{\mathcal{F}}(\mu - \overline{\mu}; \overline{\mu})$ for $\mathcal{F}(\mu - \overline{\mu}; \overline{\mu})$. To form V_J and P_J for prescribed ϵ_β such that the coverage condition is satisfied, we exploit a maximal-polytopes construction based on directional binary chop. For simplicity of exposition, we suppose that $\Phi(\mu, \bar{\mu}) = 0$. Now assume that we are given $V_{J'}$ and $P_{J'}$, we next choose a new point $\bar{\mu}_{J'+1} \in \mathcal{D}$ such that $\bar{\mu}_{J'+1} \notin \cup_{j=1}^{J'} \mathcal{P}^{\bar{\mu}_j}$; we then find the next vertex tuple $\mathcal{V}^{\bar{\mu}_{J'+1}} \equiv \{\mu_i' \in \mathcal{D}^{\bar{\mu}_{J'+1}}, 1 \le i \le |\mathcal{V}^{\bar{\mu}_{J'+1}}|\}$ and the associated polytope $\mathcal{P}^{\bar{\mu}_{J'+1}}$ by using binary chop algorithm to solve $|\mathcal{V}^{\bar{\mu}_{J'+1}}|$ nonlinear algebraic equations $\sqrt{\tilde{\mathcal{F}}(\mu_i' - \bar{\mu}_{J'+1}; \bar{\mu}_{J'+1})} = \epsilon_\beta \tilde{\beta}(\bar{\mu}_{J'+1})$ for vertex points $\mu_i', i = 1, \ldots, |\mathcal{V}^{\bar{\mu}_{J'+1}}|$, respectively;[2] we continue this process until the Coverage Condition $\cup_{j=1}^{J} \mathcal{P}^{\bar{\mu}_j} = \mathcal{D}$ is satisfied. Note that all vertex tuples $\mathcal{V}^{\bar{\mu}_j}$ consist of vertex points satisfying $\min_{\nu \in \mathcal{V}^{\bar{\mu}_j}} \sqrt{\tilde{\mathcal{F}}(\nu - \bar{\mu}_j; \bar{\mu}_j)} = \epsilon_\beta \tilde{\beta}(\bar{\mu}_j)$ exactly, which will in turn lead to maximal polytopes; and hence J is as small as possible.

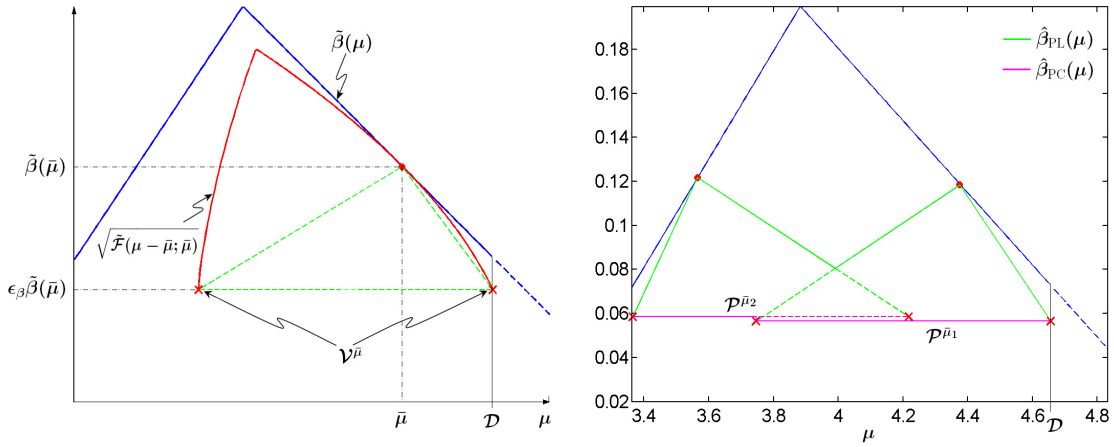
For our choice of surrogates, the reduced-basis approximation $\beta_N(\mu)$ to $\beta(\mu)$ and $\mathcal{F}_N(\mu - \overline{\mu}; \overline{\mu})$ to $\mathcal{F}(\mu - \overline{\mu}; \overline{\mu})$ are particularly relevant; thanks to the rapid uniform convergence of the reduced-basis approximation, N can be chosen quite small to achieve extremely inexpensive yet accurate surrogates [85].

4.5.3 A Simple Demonstration

As a simple demonstration, we apply lower bound construction to the Helmholtz-elasticity crack example described in Section 4.6.1 in which the crack location b and crack length L are fixed, and only the frequency squared ω^2 is permitted to vary in $\mathcal{D}$. It can be verified for this particular instantiation that $P = 1$, $\mu \equiv \omega^2$, and that $Q = 2$, $\Theta^1(\mu) = 1$, $\Theta^2(\mu) = -\omega^2$, $a^1(w, v)$ is the sum of the first seven bilinear forms in Table 4.1, $a^2(w, v)$ is the sum of the last three bilinear forms in Table 4.1. Clearly, we have $\Phi(\mu, \bar{\mu}) = 0$.

[2] Note that the concavity of $\mathcal{F}(\mu - \bar{\mu}; \bar{\mu})$ (and hence $\tilde{\mathcal{F}}(\mu - \bar{\mu}; \bar{\mu})$) allows us to perform very efficient binary search for the roots of these equations.

Furthermore for bound conditioner $(w,v)_X = a^1(w,v) + a^2(w,v)$ and seminorms $|w|_1^2 = a^1(w,w)$, $|w|_2^2 = a^2(w,w)$, we readily obtain $\Gamma_1 = 1$, $\Gamma_2 = 1$, and $C_X = 1$.



(a) (b)

Figure 4-1: A simple demonstration: (a) construction of $\mathcal{V}^{\bar{\mu}}$ and $\mathcal{P}^{\bar{\mu}}$ for a given $\bar{\mu}$ and (b) set of polytopes P_J and associated lower bounds $\hat{\beta}_{\rm PC}(\mu)$, $\hat{\beta}_{\rm PL}(\mu)$.

Now for a given $\bar{\mu} \equiv \bar{\mu}_1$, we find $\mathcal{V}^{\bar{\mu}}$ and $\mathcal{P}^{\bar{\mu}}$ by using binary chop algorithm to solve $\sqrt{\tilde{\mathcal{F}}(\mu-\bar{\mu};\bar{\mu})} = \epsilon_\beta\tilde{\beta}(\bar{\mu})$. Since $\tilde{\mathcal{F}}(\mu-\bar{\mu};\bar{\mu})$ is concave, the equation has two roots (represented by the cross points) which form $\mathcal{V}^{\bar{\mu}}$ as shown in Figure 4-1(a). We next choose a second point $\mu_2 \notin \mathcal{P}^{\bar{\mu}_1}$ and similarly construct $\mathcal{V}^{\bar{\mu}_2}$ and $\mathcal{P}^{\bar{\mu}_2}$. As shown in Figure 4-1(b), the two polytopes $\mathcal{P}^{\bar{\mu}_1}$ and $\mathcal{P}^{\bar{\mu}_2}$ are overlapped and satisfy $\mathcal{D} \subset \mathcal{P}^{\bar{\mu}_1} \cup \mathcal{P}^{\bar{\mu}_2}$; hence the generation stage is done. The verification is simple: we first obtain $\beta(\bar{\mu}_1), \beta(\bar{\mu}_2)$ and $\mathcal{F}(\mu'-\bar{\mu}_1;\bar{\mu}_1)$ for $\mu' \in \mathcal{V}^{\bar{\mu}_1}$, $\mathcal{F}(\mu'-\bar{\mu}_2;\bar{\mu}_2)$ for $\mu' \in \mathcal{V}^{\bar{\mu}_2}$ and then verify that (4.37) is indeed satisfied for $\epsilon_\beta = 0.48$. The piecewise-constant approximation $\hat{\beta}_{\rm PC}(\mu)$ and piecewise-linear approximation $\hat{\beta}_{\rm PL}(\mu)$ to $\beta(\mu)$ are also presented in Figure 4-1(b).

Note however that we use reduced-basis surrogates to generate the necessary set of points and polytopes; hence in the verification stage the Positivity Condition may not be respected for the prescribed ϵ_β which is used during the generation. In this case, we need to adjust ϵ_β. This may result in a slightly different new value of ϵ_β since the reduced-basis surrogates are generally very accurate.

4.6 Numerical Examples

4.6.1 Helmholtz-Elasticity Crack Problem

We consider a two-dimensional thin plate with a horizontal crack at the (say) interface of two lamina with a horizontal crack at the (say) interface of two lamina: the (original) domain $\tilde{\Omega}(b, L) \subset \mathbb{R}^2$ is defined as $[0, 2]\times[0, 1]\backslash\tilde{\Gamma}_{\rm C}$, where $\tilde{\Gamma}_{\rm C} \equiv \{\tilde{x}_1 \in [b-L/2, b+L/2], \tilde{x}_2 = 1/2\}$ defines the idealized crack. The crack surface is modeled extremely simplistically as a stress-free boundary. The left surface of the plate $\tilde{\Gamma}_{\rm D}$ is secured; the top and bottom boundaries $\tilde{\Gamma}_{\rm N}$ are stress-free; and the right boundary $\tilde{\Gamma}_{\rm F}$ is subject to a vertical oscillatory uniform force of frequency ω. Our parameter is thus $\mu \equiv (\mu_{(1)}, \mu_{(2)}, \mu_{(3)}) = (\omega^2, b, L)$. We model the plate as plane-stress linear isotropic elastic with (scaled) density unity, Young's modulus unity, and Poisson ratio 0.25. The governing equations for the displacement field $\tilde{u}(\tilde{x}; \mu) \in \tilde{X}(\mu)$ are thus

$$\begin{aligned} \frac{\partial\tilde{\sigma}_{11}}{\partial\tilde{x}_1} + \frac{\partial\tilde{\sigma}_{12}}{\partial\tilde{x}_2} + \omega^2\tilde{u}_1^2 = 0 \\ \frac{\partial\tilde{\sigma}_{12}}{\partial\tilde{x}_1} + \frac{\partial\tilde{\sigma}_{22}}{\partial\tilde{x}_2} + \omega^2\tilde{u}_2^2 = 0 \end{aligned} \tag{4.68}$$

$$\tilde{\varepsilon}_{11} = \frac{\partial\tilde{u}_1}{\partial\tilde{x}_1}, \quad \tilde{\varepsilon}_{22} = \frac{\partial\tilde{u}_2}{\partial\tilde{x}_2}, \quad 2\tilde{\varepsilon}_{12} = \left(\frac{\partial\tilde{u}_1}{\partial\tilde{x}_2} + \frac{\partial\tilde{u}_2}{\partial\tilde{x}_1}\right), \tag{4.69}$$

$$\begin{Bmatrix} \tilde{\sigma}_{11} \\ \tilde{\sigma}_{22} \\ \tilde{\sigma}_{12} \end{Bmatrix} = \begin{bmatrix} c_{11} & c_{12} & 0 \\ c_{12} & c_{22} & 0 \\ 0 & 0 & c_{66} \end{bmatrix} \begin{Bmatrix} \tilde{\varepsilon}_{11} \\ \tilde{\varepsilon}_{22} \\ \tilde{\varepsilon}_{12} \end{Bmatrix} \tag{4.70}$$

where the constitutive constants are given by

$$c_{11} = \frac{1}{1-\nu^2}, \qquad c_{22} = c_{11}, \qquad c_{12} = \frac{\nu}{1-\nu^2}, \qquad c_{66} = \frac{1}{2(1+\nu)} \ .$$

The boundary conditions on the (secured) left edge are

$$\tilde{u}_1 = \tilde{u}_2 = 0, \quad \text{on } \tilde{\Gamma}_{\rm D} \ . \tag{4.71}$$

The boundary conditions on the top and bottom boundaries and the crack surface are

$$\begin{aligned} \tilde{\sigma}_{11}\hat{\tilde{n}}_1 + \tilde{\sigma}_{12}\hat{\tilde{n}}_2 = 0 \quad \text{on } \tilde{\Gamma}_{\mathrm{N}} \cup \tilde{\Gamma}_{\mathrm{C}} \ , \\ \tilde{\sigma}_{12}\hat{\tilde{n}}_1 + \tilde{\sigma}_{22}\hat{\tilde{n}}_2 = 0 \quad \text{on } \tilde{\Gamma}_{\mathrm{N}} \cup \tilde{\Gamma}_{\mathrm{C}} \ . \end{aligned} \tag{4.72}$$

The boundary conditions on the right edge are

$$\begin{aligned} \tilde{\sigma}_{11}\hat{\tilde{n}}_1 + \tilde{\sigma}_{12}\hat{\tilde{n}}_2 = 0 \quad \text{on } \tilde{\Gamma}_{\mathrm{F}} \ , \\ \tilde{\sigma}_{12}\hat{\tilde{n}}_1 + \tilde{\sigma}_{22}\hat{\tilde{n}}_2 = 1 \quad \text{on } \tilde{\Gamma}_{\mathrm{F}} \ . \end{aligned} \tag{4.73}$$

Here $\hat{\tilde{n}}$ is the unit outward normal to the boundary. We now introduce $\tilde{X}(\mu)$ — a quadratic finite element truth approximation subspace (of dimension $\mathcal{N} = 14{,}662$) of $X^{\mathrm{e}}(\mu) = \{\tilde{v} \in (H^1(\tilde{\Omega}(b, L)))^2 \,|\, \tilde{v}|_{\tilde{\Gamma}_{\mathrm{F}}} = 0\}$. The weak formulation can then be derived as

$$\tilde{a}(\tilde{u}(\mu), \tilde{v}; \mu) = \tilde{f}(\tilde{v}), \quad \forall\, \tilde{v} \in \tilde{X}(\mu) \tag{4.74}$$

where

$$\begin{aligned} \tilde{a}(w, v; \mu) \ = \ & c_{12} \int_{\tilde{\Omega}} \left(\frac{\partial v_1}{\partial \tilde{x}_1} \frac{\partial w_2}{\partial \tilde{x}_2} + \frac{\partial v_2}{\partial \tilde{x}_2} \frac{\partial w_1}{\partial \tilde{x}_1} \right) + c_{66} \int_{\tilde{\Omega}} \left(\frac{\partial v_1}{\partial \tilde{x}_2} \frac{\partial w_2}{\partial \tilde{x}_1} + \frac{\partial v_2}{\partial \tilde{x}_1} \frac{\partial w_1}{\partial \tilde{x}_2} \right) \\ & + c_{11} \int_{\tilde{\Omega}} \left(\frac{\partial v_1}{\partial \tilde{x}_1} \frac{\partial w_1}{\partial \tilde{x}_1} \right) + c_{66} \int_{\tilde{\Omega}} \left(\frac{\partial v_2}{\partial \tilde{x}_1} \frac{\partial w_2}{\partial \tilde{x}_1} \right) \\ & + c_{22} \int_{\tilde{\Omega}} \left(\frac{\partial v_2}{\partial \tilde{x}_2} \frac{\partial w_2}{\partial \tilde{x}_2} \right) + c_{66} \int_{\tilde{\Omega}} \left(\frac{\partial v_1}{\partial \tilde{x}_2} \frac{\partial w_1}{\partial \tilde{x}_2} \right) - \omega^2 \int_{\tilde{\Omega}} w_1 v_1 + w_2 v_2, \end{aligned} \tag{4.75}$$

$$\tilde{f}(v) = \int_{\tilde{\Gamma}_{\mathrm{F}}} v_2 \ . \tag{4.76}$$

We now define three subdomains $\Omega_1 \equiv]0, b_r - L_r/2[\times]0, 1[$, $\Omega_2 \equiv]b_r - L_r/2, b_r + L_r/2[\times]0, 1[$, $\Omega_3 \equiv]b_r + L_r/2, 2[\times]0, 1[$ and a reference domain Ω as $\overline{\Omega} = \overline{\Omega}_1 \cup \overline{\Omega}_2 \cup \overline{\Omega}_3$; clearly, Ω is corresponding to the geometry $b = b_{\mathrm{r}} = 1.0$ and $L = L_{\mathrm{r}} = 0.2$. We then map $\tilde{\Omega}(b, L) \to \Omega \equiv \tilde{\Omega}(b_{\mathrm{r}}, L_{\mathrm{r}})$ by a continuous piecewise-affine (in fact, piecewise-dilation-in-x_1) transformation (details of the problem formulation in terms of the reference domain can be found in Section 9.2.) This new problem can now be cast precisely in the desired form $a(u, v; \mu) = f(v), \forall v \in X$, in which Ω, X, and $(w, v)_X$ are independent of the

q	$\Theta^q(\mu)$	$a^q(w,v)$
1	1	$c_{12}\int_\Omega \left(\frac{\partial v_1}{\partial x_1}\frac{\partial w_2}{\partial x_2}+\frac{\partial v_2}{\partial x_2}\frac{\partial w_1}{\partial x_1}\right)+c_{66}\int_\Omega\left(\frac{\partial v_1}{\partial x_2}\frac{\partial w_2}{\partial x_1}+\frac{\partial v_2}{\partial x_1}\frac{\partial w_1}{\partial x_2}\right)$
2	$\frac{b_r-L_r/2}{b-L/2}$	$c_{11}\int_{\Omega_1}\left(\frac{\partial v_1}{\partial x_1}\frac{\partial w_1}{\partial x_1}\right)+c_{66}\int_{\Omega_1}\left(\frac{\partial v_2}{\partial x_1}\frac{\partial w_2}{\partial x_1}\right)$
3	$\frac{L_r}{L}$	$c_{11}\int_{\Omega_2}\left(\frac{\partial v_1}{\partial x_1}\frac{\partial w_1}{\partial x_1}\right)+c_{66}\int_{\Omega_2}\left(\frac{\partial v_2}{\partial x_1}\frac{\partial w_2}{\partial x_1}\right)$
4	$\frac{2-b_r-L_r/2}{2-b-L/2}$	$c_{11}\int_{\Omega_3}\left(\frac{\partial v_1}{\partial x_1}\frac{\partial w_1}{\partial x_1}\right)+c_{66}\int_{\Omega_3}\left(\frac{\partial v_2}{\partial x_1}\frac{\partial w_2}{\partial x_1}\right)$
5	$\frac{b-L/2}{b_r-L_r/2}$	$c_{22}\int_{\Omega_1}\left(\frac{\partial v_2}{\partial x_2}\frac{\partial w_2}{\partial x_2}\right)+c_{66}\int_{\Omega_1}\left(\frac{\partial v_1}{\partial x_2}\frac{\partial w_1}{\partial x_2}\right)$
6	$\frac{L}{L_r}$	$c_{22}\int_{\Omega_2}\left(\frac{\partial v_2}{\partial x_2}\frac{\partial w_2}{\partial x_2}\right)+c_{66}\int_{\Omega_2}\left(\frac{\partial v_1}{\partial x_2}\frac{\partial w_1}{\partial x_2}\right)$
7	$\frac{2-b-L/2}{2-b_r-L_r/2}$	$c_{22}\int_{\Omega_3}\left(\frac{\partial v_2}{\partial x_2}\frac{\partial w_2}{\partial x_2}\right)+c_{66}\int_{\Omega_3}\left(\frac{\partial v_1}{\partial x_2}\frac{\partial w_1}{\partial x_2}\right)$
8	$-\omega^2\frac{b-L/2}{b_r-L_r/2}$	$\int_{\Omega_1} w_1v_1+w_2v_2$
9	$-\omega^2\frac{L}{L_r}$	$\int_{\Omega_2} w_1v_1+w_2v_2$
10	$-\omega^2\frac{2-b-L/2}{2-b_r-L_r/2}$	$\int_{\Omega_3} w_1v_1+w_2v_2$

Table 4.1: Parametric functions $\Theta^q(\mu)$ and parameter-independent bilinear forms $a^q(w,v)$ for the two-dimensional crack problem.

parameter μ. In particular, our bilinear form a is affine for $Q = 10$ as shown in Table 4.1.

4.6.2 A Coercive Case: Equilibrium Elasticity

As an illustrative example of coercive problems, we consider the Helmholtz-elasticity crack example above for $\mu = (\omega^2 = 0, b \in [0.9, 1.1], L \in [0.15, 0.25])$. The elasticity operator becomes coercive for zero frequency. Our affine assumption (4.5) thus applies for $Q = 7$, where $\Theta^q(\mu)$ and $a^q, 1 \le q \le 7$, are the first seven entries in Table 4.1 and convex in $\mathcal{D} \equiv [0.9, 1.1] \times [0.15, 0.25]$; furthermore, since $\Theta^1(\mu) = 1$ we may choose our bound conditioner $(w, v)_X = \sum_{q=1}^{Q} a^q(w, v)$ and seminorms $|w|_q^2 = a^q(w, w)$, $2 \le q \le Q$. It thus follows that $\Gamma_q = 1, 2 \le q \le Q$, and (numerically computed) $C_X = 1.9430$.

We present in Figure 4-2 $\alpha(\mu)$ and $\hat{\alpha}(\mu;\overline{\mu}) \equiv \mathcal{F}(\mu;\overline{\mu}) - \Phi(\mu,\overline{\mu})$ for $\overline{\mu} = (0, 1.0, 0.2)$ as a function of μ. We find that a sample $E_{J=1}$ suffices to satisfy our Positivity and Coverage Conditions for $\epsilon_\alpha = 0.38$. The value of J is equal to 1 since (i) $\alpha(\mu)$ is highly smooth in μ — generally the case for coercive operators, (ii) $\mathcal{F}$ correctly captures the first-order information, and (iii) the more pessimistic bounds (e.g., C_X) appear only to second order. We further observe that both $\alpha(\mu)$ and $\hat{\alpha}(\mu;\overline{\mu})$ are concave in μ and that $\hat{\alpha}(\mu;\overline{\mu})$ is a strict lower bound of $\alpha(\mu)$. The concavity of $\hat{\alpha}(\mu;\overline{\mu})$ follows from the

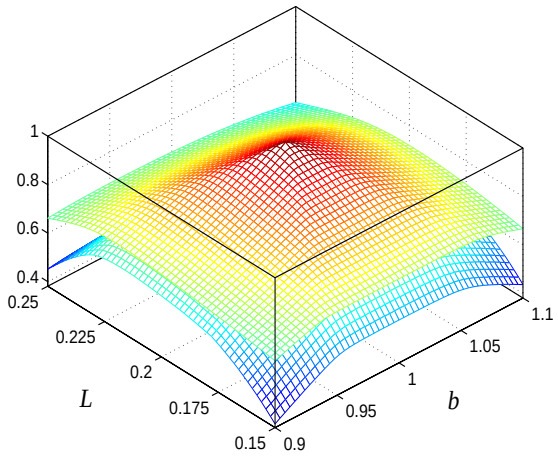


Figure 4-2: $\alpha(\mu)$ (upper surface) and $\hat{\alpha}(\mu;\overline{\mu})$ (lower surface) as a function of μ.

concavity of $\mathcal{F}(\mu;\overline{\mu})$ and $-\Phi(\mu,\overline{\mu})$, since $\Theta^q, 1 \le q \le Q$, are convex in $\mathcal{D}$.

4.6.3 A Noncoercive Case: Helmholtz Elasticity

As an illustrative example of noncoercive problems, we consider the Helmholtz-elasticity crack example above for $\mu = (\omega^2 = 4.0, b \in [0.9, 1.1], L \in [0.15, 0.25])$. For positive constant frequency, our affine assumption (4.5) applies for $Q = 10$, where $\Theta^q(\mu)$ and $a^q, 1 \le q \le Q$, given in Table 4.1 (with $\omega^2 = 4.0$) are convex in $\mathcal{D} \equiv [0.9, 1.1] \times [0.15, 0.25]$. We now define bound conditioner $(w, v)_X = \sum_{q=2}^{Q} a^q(w, v)$; thanks to the Dirichlet condition at $\tilde{x}_1 = 0$, $(\cdot,\cdot)_X$ is appropriately coercive. We observe that $\Theta^1(\mu) = 1$ and we can thus disregard the $q = 1$ term in our continuity bound. We may then choose $|v|_q^2 = a^q(v, v)$, $2 \le q \le Q$, since the $a^q(\cdot,\cdot)$ are positive semi-definite; it thus follows from the Cauchy-Schwarz inequality that $\Gamma^q = 1$, $2 \le q \le Q$. Furthermore, from (4.34), we directly obtain $C_X = 1$.

We show in Figure 4-3 $\beta^2(\mu)$ and $\mathcal{F}(\mu - \bar{\mu};\bar{\mu})$ for $\bar{\mu} = (4.0, 1.0, 0.2)$. We observe that (in this particular case, even without $\Phi(\mu;\bar{\mu})$), $\mathcal{F}(\mu - \bar{\mu};\bar{\mu})$ is a lower bound for $\beta^2(\mu)$; that $\mathcal{F}(\mu - \bar{\mu};\mu)$ is concave; and that $\mathcal{F}(\mu - \bar{\mu};\mu)$ is tangent to $\beta^2(\mu)$ at $\mu = \bar{\mu}$.

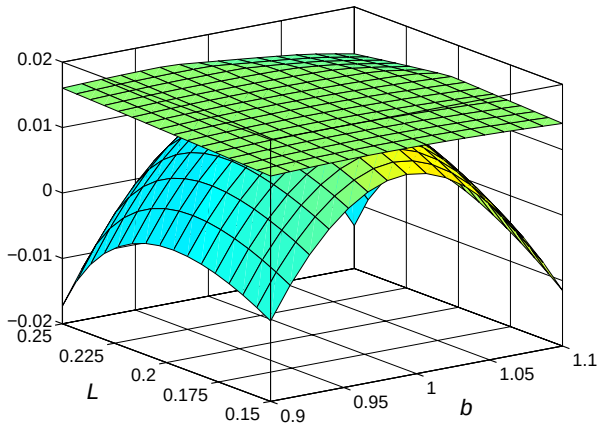


Figure 4-3: $\beta^2(\mu)$ and $\mathcal{F}(\mu - \bar{\mu}; \bar{\mu})$ for $\bar{\mu} = (4, 1, 0.2)$ as a function of (b, L); $\omega^2 = 4.0$.

4.6.4 A Noncoercive Case: Damping and Resonance

We consider here the particular case: $P = 1$, $Q = 2$, $\Theta^1(\mu) = C$ (a constant function), $\Theta^2(\mu) = \mu$, $a(w, v; \mu) = Ca^1(w, v) + \mu a^2(w, v)$ and X is a complex function space; we further suppose that $\mathcal{D}$ is convex. Given any $\overline{\mu} \in \mathcal{D}$, we introduce $\mathcal{T}(w, v; \mu; \overline{\mu}) \equiv (T^{\overline{\mu}} w, T^{\overline{\mu}} v)_X + (\mu - \overline{\mu})[a^2(w, T^{\overline{\mu}} v) + \overline{a^2(v, T^{\overline{\mu}} w)}]$ and $\mathcal{D}^{\overline{\mu}} \equiv \{\mu \in \mathcal{D} \,|\, \mathcal{T}(v, v; \mu; \overline{\mu}) \geq 0\}$; it can be easily verified that $\mathcal{T}(\cdot, \cdot; \mu; \overline{\mu})$ is symmetric: $\mathcal{T}(w, v; \mu; \overline{\mu}) = \overline{(T^{\overline{\mu}} v, T^{\overline{\mu}} w)_X} + (\mu - \overline{\mu})\overline{[a^2(v, T^{\overline{\mu}} w) + \overline{a^2(w, T^{\overline{\mu}} v)}]} = \overline{\mathcal{T}(v, w; \mu; \overline{\mu})}$ due to the symmetry of $(\cdot, \cdot)_X$ and $a^2(w, T^{\overline{\mu}} v) + \overline{a^2(v, T^{\overline{\mu}} w)}$ since $a^2(w, T^{\overline{\mu}} v)$ is a complex conjugate of $\overline{a^2(v, T^{\overline{\mu}} w)}$. We may then define

$$\hat{\beta}(\mu; \overline{\mu}) \equiv \sqrt{\inf_{w \in X} \mathcal{T}(w, w; \mu; \overline{\mu}) / \|w\|_X^2}, \quad \forall\, \mu \in \mathcal{D}^{\overline{\mu}}\,, \tag{4.77}$$

and note from $\Phi(\mu, \overline{\mu}) = 0$ that our function $\hat{\beta}(\mu; \overline{\mu})$ enjoys three properties: (i) $\beta(\mu) \geq \hat{\beta}(\mu; \overline{\mu}) \geq 0$, $\forall\, \mu \in \mathcal{D}^{\overline{\mu}}$; (ii) $\hat{\beta}(\mu; \overline{\mu})$ is concave in μ over the convex domain $\mathcal{D}^{\overline{\mu}}$; and (iii) $\hat{\beta}(\mu; \overline{\mu})$ is tangent to $\beta(\mu)$ at $\mu = \overline{\mu}$. To make property (iii) rigorous we must in general consider non-smooth analysis and also possibly a continuous spectrum as $\mathcal{N} \to \infty$. We now prove those properties of $\hat{\beta}(\mu; \overline{\mu})$ and refer to Appendix C for detailed formulation of the inf-sup lower bounds for general complex noncoercive problems. The concavity of

$\hat{\beta}(\mu;\overline{\mu})$ follows from Lemma 4.2.2. The lower bound property of $\hat{\beta}(\mu;\overline{\mu})$ is proven below

$$
\begin{aligned}
\beta^2(\mu) &= \inf_{w\in X} \frac{(T^{\overline{\mu}}w + T^{\mu}w - T^{\overline{\mu}}w, T^{\overline{\mu}}w + T^{\mu}w - T^{\overline{\mu}}w)_X}{\|w\|_X^2} \\
&= \inf_{w\in X} \frac{\|T^{\overline{\mu}}w\|_X^2 + (T^{\mu}w - T^{\overline{\mu}}w, T^{\overline{\mu}}w)_X + \overline{(T^{\mu}w - T^{\overline{\mu}}w, T^{\overline{\mu}}w)_X} + \|T^{\mu}w - T^{\overline{\mu}}w\|_X^2}{\|w\|_X^2} \\
&= \inf_{w\in X} \left\{ \frac{\|T^{\overline{\mu}}w\|_X^2 + \sum_{q=1}^{Q} \left(\Theta^q(\mu) - \Theta^q(\overline{\mu})\right) a^q(w, T^{\overline{\mu}}w)}{\|w\|_X^2} + \right. \\
&\qquad \left. \frac{\sum_{q=1}^{Q} \overline{\left(\Theta^q(\mu) - \Theta^q(\overline{\mu})\right) a^q(w, T^{\overline{\mu}}w)} + \|T^{\mu}w - T^{\overline{\mu}}w\|_X^2}{\|w\|_X^2} \right\} \\
&= \inf_{w\in X} \frac{\|T^{\overline{\mu}}w\|_X^2 + (\mu - \overline{\mu}) \left(a^2(w, T^{\overline{\mu}}w) + \overline{a^2(w, T^{\overline{\mu}}w)}\right) + \|T^{\mu}w - T^{\overline{\mu}}w\|_X^2}{\|w\|_X^2} \\
&\geq \inf_{w\in X} \frac{\mathcal{T}(w, w; \mu; \overline{\mu})}{\|w\|_X^2} \equiv \hat{\beta}^2(\mu;\overline{\mu}), \quad \forall\, \mu \in \mathcal{D}^{\overline{\mu}}.
\end{aligned}
$$

Furthermore, it follows from the above result that we have

$$
\frac{d\beta^2(\mu)}{d\mu} = \frac{d\hat{\beta}^2(\mu;\overline{\mu})}{d\mu} = \inf_{w\in X} \frac{a^2(w, T^{\overline{\mu}}w) + \overline{a^2(w, T^{\overline{\mu}}w)}}{\|w\|_X^2}
$$

at $\mu = \overline{\mu}$, which means property (iii) .

Specifically, we consider the Helmholtz-elasticity crack example described in Section 4.6.1 for $\mu = (\omega^2 \in [2.5, 5.0],\ b = 1.0,\ L = 0.2)$ — only ω^2 is permitted to vary — and material damping coefficient d_m (note that both third mode and fourth mode resonances are within our frequency range). Since the problem is a complex boundary value problem, our quadratic finite element truth approximation space of dimension $\mathcal{N} = 14{,}662$ is complexified such that

$$
X = \left\{ v = v^{\mathrm{R}} + \mathrm{i}v^{\mathrm{I}} \in X^{\mathrm{e}} \mid v^{\mathrm{R}}|_{T_h} \in \mathbb{P}_2(T_h),\ v^{\mathrm{I}}|_{T_h} \in \mathbb{P}_2(T_h),\ \forall\, T_h \in \mathcal{T}_h \right\} , \tag{4.78}
$$

where $\mathbb{P}_2(T_h)$ is the space of second-order polynomials over element T_h and X^{e} is a complex function space defined as

$$
X^{\mathrm{e}} = \left\{ v = v^{\mathrm{R}} + \mathrm{i}v^{\mathrm{I}} \mid v^{\mathrm{R}} \in \left(H^1(\Omega)\right)^2,\ v^{\mathrm{I}} \in \left(H^1(\Omega)\right)^2, v^{\mathrm{R}}|_{x_1=0} = 0, v^{\mathrm{I}}|_{x_1=0} = 0 \right\} . \tag{4.79}
$$

Recall that R and I denote the real and imaginary part, respectively; and that $\bar{v}$ denotes the complex conjugate of v, and $|v|$ the modulus of v. By a simple "hysteretic" Kelvin model [13] for complex Young's modulus, our bilinear form is given by

$$a(w, v; \mu) = \Theta^1(\mu) a^1(w, v) + \Theta^2(\mu) a^2(w, v) \ , \tag{4.80}$$

where the parameter-dependent functions are

$$\Theta^1(\mu) = 1 + \mathrm{i} d_m, \qquad \Theta^2(\mu) = -\omega^2 = -\mu \ , \tag{4.81}$$

and the parameter-independent bilinear forms are

$$\begin{aligned} a^1(w, v) \quad = \quad & c_{12} \int_\Omega \left(\frac{\partial w_1}{\partial x_1} \frac{\partial \bar{v}_2}{\partial x_2} + \frac{\partial w_2}{\partial x_2} \frac{\partial \bar{v}_1}{\partial x_1} \right) + c_{66} \int_\Omega \left(\frac{\partial w_1}{\partial x_2} \frac{\partial \bar{v}_2}{\partial x_1} + \frac{\partial w_2}{\partial x_1} \frac{\partial \bar{v}_1}{\partial x_2} \right) \\ & + c_{11} \int_\Omega \left(\frac{\partial w_1}{\partial x_1} \frac{\partial \bar{v}_1}{\partial x_1} \right) + c_{22} \int_\Omega \left(\frac{\partial w_2}{\partial x_2} \frac{\partial \bar{v}_2}{\partial x_2} \right) \\ & + c_{66} \int_\Omega \left(\frac{\partial w_2}{\partial x_1} \frac{\partial \bar{v}_2}{\partial x_1} \right) + c_{66} \int_\Omega \left(\frac{\partial w_1}{\partial x_2} \frac{\partial \bar{v}_1}{\partial x_2} \right) \end{aligned} \tag{4.82}$$

$$a^2(w, v) = \int_\Omega w_1 \bar{v}_1 + w_2 \bar{v}_2 \ . \tag{4.83}$$

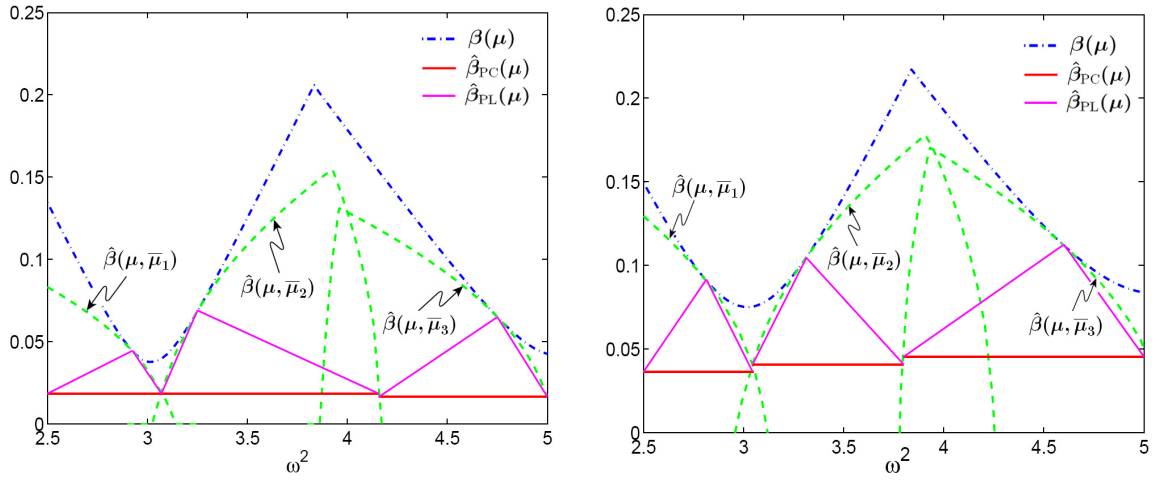
Note that the $a^1(w, v)$ and $a^2(w, v)$ are symmetric positive-semidefinite. We furthermore define our bound conditioner $(\cdot, \cdot)_X$ as

$$(w, v)_X = a^1(w, v) + a^2(w, v) \tag{4.84}$$

which is a μ-independent continuous coercive symmetric bilinear form.

We present in Figure 4-4 $\beta(\mu)$, $\hat{\beta}(\mu; \overline{\mu}_j)$ for $\mu \in \mathcal{D}^{\overline{\mu}_j}, 1 \le j \le J$, $\hat{\beta}_{\mathrm{PC}}(\mu)$, and $\hat{\beta}_{\mathrm{PL}}(\mu)$ for material damping coefficient of 0.05 and 0.1. We find that a sample $E_{J=3}$ suffices to satisfy our Positivity and Coverage Conditions with $\epsilon_\beta = 0.32$ for $d_m = 0.05$ and with $\epsilon_\beta = 0.4$ for $d_m = 0.1$. Unlike the previous example $\beta(\mu)$ is not concave (or convex) or even quasi-concave, and hence $\hat{\beta}(\mu; \overline{\mu})$ is a necessary intermediary in the construction (in fact, constructive proof) of our lower bound. We further observe that the damping coefficient has a strong "shift-up" effect on our inf-sup parameter and lower bounds especially near

resonance region: increasing d_m tends to move the curve $\beta(\mu)$ up.



(a) (b)

Figure 4-4: Plots of $\beta(\mu)$; $\hat{\beta}(\mu;\overline{\mu}_1)$, $\hat{\beta}(\mu;\overline{\mu}_2)$, $\hat{\beta}(\mu;\overline{\mu}_3)$ for $\mu \in \mathcal{D}^{\overline{\mu}_j}, 1 \le j \le J$; and our lower bounds $\hat{\beta}_{\rm PC}(\mu)$ and $\hat{\beta}_{\rm PL}(\mu)$: (a) $d_m = 0.05$ and (b) $d_m = 0.1$.

4.6.5 A Noncoercive Case: Infinite Domain

We consider the Helmholtz equation $\Delta^2 u + k^2 u = 0$ in $\Omega \subset \mathbb{R}^3$, $\frac{\partial u}{\partial \hat{n}} = 1$ on $\Gamma_{\rm N}$, and $\frac{\partial u}{\partial \hat{n}} = \left(\mathrm{i}k - \frac{1}{R}\right) u$ on $\Gamma_{\rm R}$; here Ω is bounded by a inner unit sphere and outer sphere of radius R; $\Gamma_{\rm N}$ is the surface of the unit sphere; $\Gamma_{\rm R}$ is the surface of the outer sphere; and $\hat{n}$ is the unit outward normal to the boundary. Our parameter is $\mu = k \in \mathcal{D} \equiv [0.1, 1.5]$, where k is a wave number. The exact solution is given by $u^{\rm e}(r) = \frac{e^{\mathrm{i}k(r-1)}}{r(\mathrm{i}k-1)}$, where r is a distance from the origin. We further note for large R that the "exact" Robin condition can be approximated by an "inexact" boundary condition, $\frac{\partial u}{\partial \hat{n}} = \mathrm{i}ku$ on $\Gamma_{\rm R}$. In this example, we investigate the behavior of the inf-sup parameter $\beta(\mu)$ and the lower bound $\hat{\beta}(\mu)$ for a large variation of radius R for both exact and inexact conditions. This study give us a better understanding into the effect of domain truncation and boundary condition approximation on numerical solutions and reduced-basis formulation of the inverse scattering problems discussed in Chapter 10.

By invoking the symmetry of the problem, we can simplify it into a one-dimensional problem: $\frac{\partial u}{\partial r}\left(r^2 \frac{\partial u}{\partial r}\right) + k^2 r^2 u = 0$ in $\Omega \equiv]1, R[$, $\frac{\partial u}{\partial r} = 1$ at $r = 1$, and $\frac{\partial u}{\partial r} = \left(\mathrm{i}k - \frac{1}{R}\right) u$

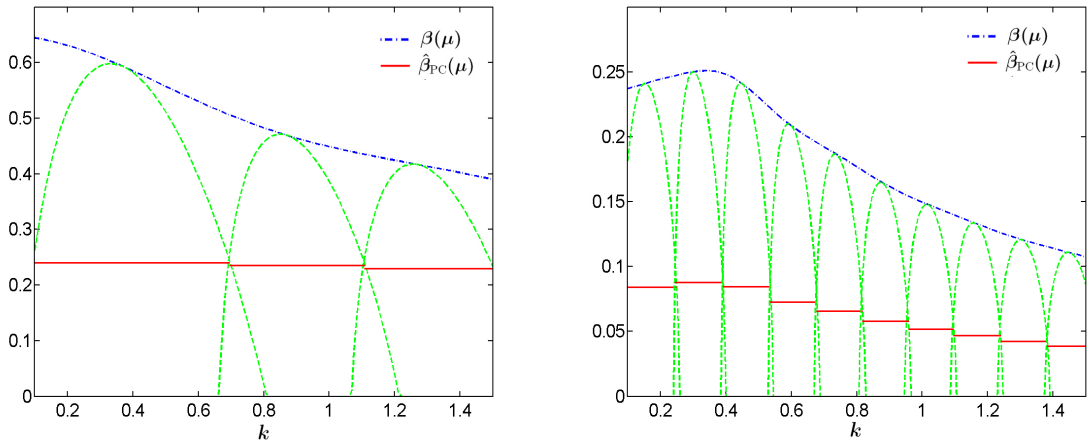
at $r = R$; and the "inexact" boundary condition is given by $\frac{\partial u}{\partial r} = \mathrm{i}ku$ at $r = R$. It is then a simple matter to show that: $Q = 3$, $a^1(w,v) = \int_\Omega r^2 \frac{\partial w}{\partial r}\frac{\partial \bar{v}}{\partial r}$, $a^2(w,v) = \int_\Omega r^2 w\bar{v}$, and $a^3(w,v) = R^2 w(R)\bar{v}(R)$; furthermore we have $\Theta^1(\mu) = 1$, $\Theta^2(\mu) = -\mu^2$, $\Theta^3(\mu) = -\mathrm{i}\mu + \frac{1}{R}$ for exact Robin condition, but $\Theta^1(\mu) = 1$, $\Theta^2(\mu) = -\mu^2$, $\Theta^3(\mu) = -\mathrm{i}\mu$ for approximate Robin condition. We next choose bound conditioner $(w,v)_X \equiv \int_\Omega r^2 \frac{\partial w}{\partial r}\frac{\partial \bar{v}}{\partial r} + \frac{1}{R}\int_\Omega r^2 w\bar{v}$,[3] and seminorms $|w|_1^2 = a^1(w,w)$, $|w|_2^2 = a^2(w,w)$, $|w|_3^2 = a^3(w,w)$. We readily calculate $\Gamma_1 = 1$, $\Gamma_2 = 1$, $\Gamma_3 = 1$; note however that the constant C_X depends on R — $C_X = 3.35$ for $R = 3$ and $C_X = 10.03$ for $R = 10$.

We present $\beta(\mu)$, $\hat{\beta}_{\mathrm{PC}}(\mu)$, and $\hat{\beta}(\mu;\overline{\mu}_j)$, $1 \le j \le J$, for exact and approximate Robin conditions in Figure 4-5 and in Figure 4-6, respectively, where

$$\hat{\beta}(\mu;\overline{\mu}) \equiv \sqrt{\max(\mathcal{F}(\mu-\overline{\mu};\overline{\mu}), \Phi^2(\mu,\overline{\mu}))} - \Phi(\mu,\overline{\mu}) \ . \tag{4.85}$$

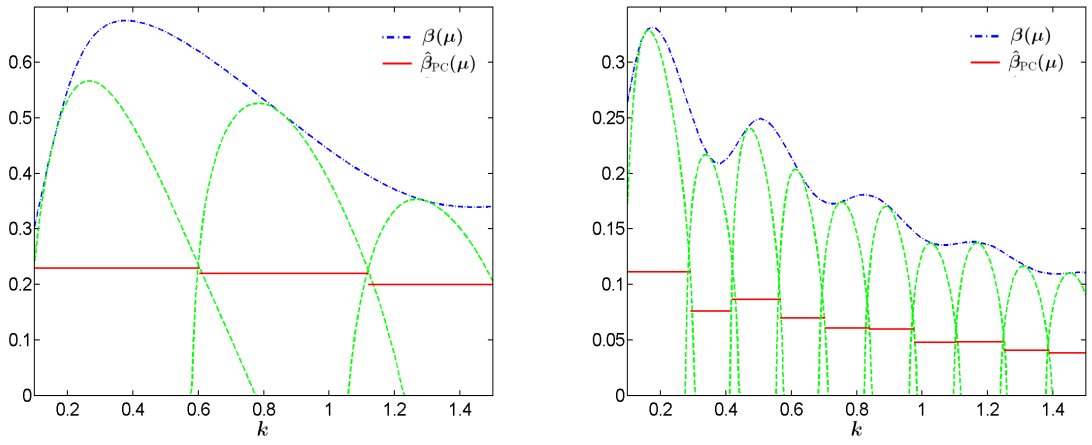
We observe in both cases that J increases with R — $J = 3$ for $R = 3$ and $J = 10$ for $R = 10$. Clearly, increasing R has strong effect on $\beta(\mu)$ and $\hat{\beta}(\mu;\overline{\mu}_j)$, as R increases $\beta(\mu)$ is smaller while $\hat{\beta}(\mu;\overline{\mu}_j)$ decreases even more rapidly. This is because (i) C_X is quite large and grows rapidly with R and (ii) $\mathcal{F}(\mu-\overline{\mu};\overline{\mu})$ decreases with $\mu-\overline{\mu}$ more rapidly as R increases. Particularly, we observe that the C_X term dominates $\mathcal{F}$ in causing the large J for $R = 3$, but the $\mathcal{F}$ function is a primary cause for the large J for $R = 10$. Note also that for approximate Robin condition there is not only outgoing, but incoming wave in the solution. This is reflected by the oscillation of the associated inf-sup parameter.

[3] The $1/R$ scaling factor in $\int_\Omega r^2 wv$ will increase smoothness and magnitude of the inf-sup parameter $\beta(\mu)$, albeit at the large value of C_X.



(a) (b)

Figure 4-5: Plots of $\beta(\mu)$; $\hat{\beta}_{\rm PC}(\mu)$; and $\hat{\beta}(\mu;\overline{\mu}_j)$, $1 \leq j \leq J$, for exact Robin Condition: (a) $R = 3, J = 3$ and (b) $R = 10, J = 10$.



(a) (b)

Figure 4-6: Plots of $\beta(\mu)$; $\hat{\beta}_{\rm PC}(\mu)$ and $\hat{\beta}(\mu;\overline{\mu}_j)$, $1 \leq j \leq J$, for approximate Robin Condition: (a) $R = 3, J = 3$ and (b) $R = 10, J = 10$.

Chapter 5

A Posteriori Error Estimation for Noncoercive Elliptic Problems

5.1 Abstraction

5.1.1 Preliminaries

We consider the "exact" (superscript e) problem: Given $\mu \in \mathcal{D} \subset \mathbb{R}^P$, we evaluate $s^{\rm e}(\mu) = \ell(u^{\rm e}(\mu))$, where $u^{\rm e}(\mu)$ satisfies the weak form of the μ-parametrized PDE

$$a(u^{\rm e}(\mu), v; \mu) = f(v), \quad \forall\, v \in X^{\rm e} \ . \tag{5.1}$$

Here μ and $\mathcal{D}$ are the input and (closed) input domain, respectively; $u^{\rm e}(x;\mu)$ is field variable; $X^{\rm e}$ is a Hilbert space with inner product $(w,v)_{X^{\rm e}}$ and associated norm $\|w\| = \sqrt{(w,w)_{X^{\rm e}}}$; and $a(\cdot,\cdot;\mu)$ and $f(\cdot), \ell(\cdot)$ are $X^{\rm e}$-continuous bilinear and linear functionals, respectively. (We may also consider complex-valued fields and spaces.) Our interest here is in second-order PDEs, and our function space $X^{\rm e}$ will thus satisfy $(H_0^1(\Omega))^\nu \subset X^{\rm e} \subset (H^1(\Omega))^\nu$, where $\Omega \subset \mathbb{R}^d$ is our spatial domain, a point of which is denoted x, and $\nu = 1$ for a scalar field variable and $\nu = d$ for a vector field variable.

We now introduce X (typically, $X \subset X^{\rm e}$), a "truth" finite element approximation space of dimension $\mathcal{N}$. The inner product and norm associated with X are given by

$(\cdot,\cdot)_X$ and $\|\cdot\|_X = (\cdot,\cdot)_X^{1/2}$, respectively. A typical choice for $(\cdot,\cdot)_X$ is

$$(w,v)_X = \int_\Omega \nabla w \cdot \nabla v + wv \ , \tag{5.2}$$

which is simply the standard $H^1(\Omega)$ inner product. We shall denote by X' the dual space of X. For a $h \in X'$, the dual norm is given by

$$\|h\|_{X'} \equiv \sup_{v\in X} \frac{h(v)}{\|v\|_X} \ . \tag{5.3}$$

In this chapter, we continue to assume that our output functional is compliant, $\ell = f$, and that a is symmetric, $a(w,v;\mu) = a(v,w;\mu), \forall w, v \in X$. This assumption will be readily relaxed in the next chapter.

We shall also make two crucial hypotheses. The first hypothesis is related to well-posedness, and is often verified only *a posteriori*. We assume that a satisfies a continuity and inf-sup condition for all $\mu \in \mathcal{D}$, as we now state more precisely. It shall prove convenient to state our hypotheses by introducing a supremizing operator $T^\mu : X \to X$ such that, for any w in X

$$(T^\mu w, v)_X = a(w,v;\mu), \quad \forall\, v \in X \ . \tag{5.4}$$

We then define

$$\sigma(w;\mu) \equiv \frac{\|T^\mu w\|_X}{\|w\|_X} \ , \tag{5.5}$$

and note that

$$\beta(\mu) \equiv \inf_{w\in X} \sup_{v\in X} \frac{a(w,v;\mu)}{\|w\|_X \|v\|_X} = \inf_{w\in X} \sigma(w;\mu) \tag{5.6}$$

$$\gamma(\mu) \equiv \sup_{w\in X} \sup_{v\in X} \frac{a(w,v;\mu)}{\|w\|_X \|v\|_X} = \sup_{w\in X} \sigma(w;\mu) \ . \tag{5.7}$$

Here $\beta(\mu)$ is the Babuška "inf-sup" (stability) constant and $\gamma(\mu)$ is the standard continuity constant; of course, both these "constants" depend on the parameter μ. Our first hypothesis is then: $0 < \beta_0 \le \beta(\mu)$ and $\gamma(\mu) \le \gamma_0 < \infty$, $\forall\, \mu \in \mathcal{D}$.

The second hypothesis is related primarily to numerical efficiency, and is typically

verified *a priori*. We assume that for some finite integer Q, a may be expressed as an affine decomposition of the form

$$a(w, v; \mu) = \sum_{q=1}^{Q} \Theta^q(\mu) a^q(w, v), \quad \forall\, w, v \in X, \forall\, \mu \in \mathcal{D}\ , \tag{5.8}$$

where for $1 \le q \le Q$, $\Theta^q : \mathcal{D} \to \mathbb{R}$ are differentiable parameter-dependent coefficient functions and bilinear forms $a^q : X \times X \to \mathbb{R}$ are parameter-independent. This hypothesis is quite restricted and will be relaxed in the next chapter.

Finally, it directly follows from (5.4) and (5.8) that, for any $w \in X$, $T^\mu w \in X$ may be expressed as

$$T^\mu w = \sum_{q=1}^{Q} \Theta^q(\mu)\, T^q w\ , \tag{5.9}$$

where, for any $w \in X$, $T^q w$, $1 \le q \le Q$, is given by

$$(T^q w, v)_X = a^q(w, v), \quad \forall\, v \in X\ . \tag{5.10}$$

Note that the operators $T^q \colon X \to X$ are independent of the parameter μ.

5.1.2 General Problem Statement

Our truth finite-element approximation to the continuous problem (5.1) is stated as: Given $\mu \in \mathcal{D}$, we evaluate

$$s(\mu) = \ell(u(\mu)), \tag{5.11}$$

where the finite element approximation $u(\mu) \in X$ is the solution of

$$a(u(\mu), v; \mu) = f(v), \quad \forall\, v \in X\ . \tag{5.12}$$

In essence, $u(\mu) \in X$ is a calculable surrogate for $u^{\mathrm{e}}(\mu)$ upon which we will build our RB approximation and with respect to which we will evaluate the RB error; $u(\mu)$ shall also serve as the "classical alternative" relative to which we will assess the efficiency of our approach. We assume that $\|u^{\mathrm{e}}(\mu) - u(\mu)\|$ is suitably small and hence that $\mathcal{N}$ is typically very large: our formulation must be both *stable* and *efficient* as $\mathcal{N} \to \infty$.

5.1.3 A Model Problem

Our model problem is the Helmholtz-Elasticity Crack example described thoroughly in Section 4.6.1. Recall that the input is $\mu \equiv (\mu^1, \mu^2, \mu^3) = (\omega^2, b, L)$, where ω is the frequency of oscillatory uniform force applied at the right edge, b is the crack location, and L is the crack length. The weak form for the displacement field $\tilde{u}(\tilde{x};\mu) \in \tilde{X}(\mu)$ is

$$\tilde{a}(\tilde{u}(\mu), \tilde{v}; \mu) = \tilde{f}(\tilde{v}), \quad \forall\, \tilde{v} \in \tilde{X}(\mu) \tag{5.13}$$

where $\tilde{X}(\mu)$ is a quadratic finite element truth approximation subspace (of dimension $\mathcal{N} = 14{,}662$) of $X^{\mathrm{e}}(\mu) = \{\tilde{v} \in (H^1(\tilde{\Omega}(b,L)))^2 \,|\, \tilde{v}|_{\tilde{x}_1=0} = 0\}$, and

$$\begin{aligned}
\tilde{a}(w, v; \mu) \;=\; & c_{12} \int_{\tilde{\Omega}} \left(\frac{\partial v_1}{\partial \tilde{x}_1} \frac{\partial w_2}{\partial \tilde{x}_2} + \frac{\partial v_2}{\partial \tilde{x}_2} \frac{\partial w_1}{\partial \tilde{x}_1} \right) + c_{66} \int_{\tilde{\Omega}} \left(\frac{\partial v_1}{\partial \tilde{x}_2} \frac{\partial w_2}{\partial \tilde{x}_1} + \frac{\partial v_2}{\partial \tilde{x}_1} \frac{\partial w_1}{\partial \tilde{x}_2} \right) \\
& + c_{11} \int_{\tilde{\Omega}} \left(\frac{\partial v_1}{\partial \tilde{x}_1} \frac{\partial w_1}{\partial \tilde{x}_1} \right) + c_{66} \int_{\tilde{\Omega}} \left(\frac{\partial v_2}{\partial \tilde{x}_1} \frac{\partial w_2}{\partial \tilde{x}_1} \right) \\
& + c_{22} \int_{\tilde{\Omega}} \left(\frac{\partial v_2}{\partial \tilde{x}_2} \frac{\partial w_2}{\partial \tilde{x}_2} \right) + c_{66} \int_{\tilde{\Omega}} \left(\frac{\partial v_1}{\partial \tilde{x}_2} \frac{\partial w_1}{\partial \tilde{x}_2} \right) - \omega^2 \int_{\tilde{\Omega}} w_1 v_1 + w_2 v_2
\end{aligned} \tag{5.14}$$

$$\tilde{f}(v) = \int_{\tilde{\Gamma}_{\mathrm{F}}} v_2 \; . \tag{5.15}$$

The output is the (oscillatory) amplitude of the average vertical displacement on the right edge of the plate, $s(\mu) = \tilde{\ell}(\tilde{u}(\mu))$ with $\tilde{\ell} = \tilde{f}$; we are thus "in compliance".

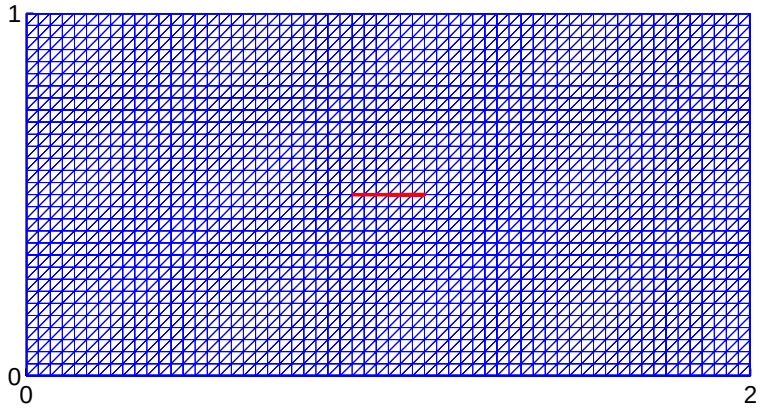


Figure 5-1: Quadratic triangular finite element mesh on the reference domain with the crack in red. Note that each element has six nodes.

By using a continuous piecewise-affine (in fact, piecewise-dilation-in-x_1) transformation to map the original domain $\tilde{\Omega}(b, L)$ to the reference domain $\Omega \equiv \tilde{\Omega}(b_{\rm r}, L_{\rm r})$ with $b_{\rm r} = 1.0$ and $L_{\rm r} = 0.2$, we arrive at the desired form (5.12) in which Ω, X, and $(\cdot,\cdot)_X$ are independent of the parameter μ, a is affine for $Q = 10$ as given in Table 4.1, and $f(v) = \int_{\tilde{\Gamma}_{\rm F}} v_2$. Furthermore, we use a regular quadratic triangular mesh for X as shown in Figure 5-1. (No crack-tip element is needed as the output of interest is on the right edge — far from the crack tips.)

5.2 Reduced-Basis Approximation

In this section we review briefly the reduced-basis approximation since many details has been already discussed in Chapter 3. Moreover, we shall also discuss approximation approaches other than Galerkin projection, in particular the Petrov-Galerkin projection, which can be advantageous for noncoercive problems.

5.2.1 Galerkin Approximation

In the "Lagrangian" [116] reduced-basis approach, the field variable $u(\mu)$ is approximated by (typically) Galerkin projection onto a space spanned by solutions of the governing PDE at N selected points in parameter space. We introduce nested parameter samples $S_N \equiv \{\mu_1 \in \mathcal{D}, \cdots, \mu_N \in \mathcal{D}\}, 1 \le N \le N_{\max}$ and associated nested reduced-basis spaces $W_N \equiv \text{span}\{\zeta_j \equiv u(\mu_j), 1 \le j \le N\}, 1 \le N \le N_{\max}$, where $u(\mu_j)$ is the solution to (5.12) for $\mu = \mu_j$. We next apply Galerkin projection onto W_N to obtain $u_N(\mu) \in W_N$ from

$$a(u_N(\mu), v; \mu) = f(v), \quad \forall\, v \in W_N \ , \tag{5.16}$$

in terms of which the reduced-basis approximation to $s(\mu)$ is then calculated as

$$s_N(\mu) = \ell(u_N(\mu)) \ . \tag{5.17}$$

However, Galerkin projection does not guarantee stability of the discrete reduced-basis system. More sophisticated minimum-residual [91, 131] and in particular Petrov-Galerkin

[92, 131] approaches restore (guaranteed) stability, albeit at some additional complexity.

5.2.2 Petrov-Galerkin Approximation

In addition to the primal problem, the Petrov-Galerkin approach shall require the dual problem: find $\psi(\mu) \in X$ such that

$$a(v, \psi(\mu); \mu) = -\ell(v), \quad \forall\, v \in X\ . \tag{5.18}$$

Note that the dual problem is useful to the noncompliance case in which a is nonsymmetric or $\ell \neq f$. In the compliance case, symmetric a and $\ell = f$, the dual problem becomes unnecessary since $\psi(\mu) = -u(\mu)$.

We can now introduce sample $S^{\rm pr}_{N_1} = \{\mu^{\rm pr}_1 \in \mathcal{D}, \cdots, \mu^{\rm pr}_{N_1} \in \mathcal{D}\}$ and associated Lagrangian space $W^{\rm pr}_{N_1} = \text{span}\{u(\mu^{\rm pr}_j), \forall \mu^{\rm pr}_j \in S^{\rm pr}_{N_1}\}$. Similarly, we select sample $S^{\rm du}_{N_2} = \{\mu^{\rm du}_1 \in \mathcal{D}, \cdots, \mu^{\rm du}_{N_2} \in \mathcal{D}\}$, possibly different from the ones above, and form associated dual space $W^{\rm du}_{N_2} = \text{span}\{\psi(\mu^{\rm du}_j),\ \forall \mu^{\rm du}_j \in S^{\rm du}_{N_2}\}$. We then define the *infimizing* space as

$$\begin{aligned} W_N &= W^{\rm pr}_{N_1} + W^{\rm du}_{N_2} \\ &= \text{span}\{u(\mu^{\rm pr}_i), \psi(\mu^{\rm du}_j),\ \forall \mu^{\rm pr}_i \in S^{\rm pr}_{N_1},\ \forall \mu^{\rm du}_j \in S^{\rm du}_{N_2}\} \\ &\equiv \text{span}\{\zeta_1, \ldots, \zeta_N\}. \end{aligned} \tag{5.19}$$

The dimension of our reduced-basis approximation is thus $N = N_1 + N_2$.

The Petrov-Galerkin will also need supremizing space. To this end, we compute $T^q\zeta_n$ from (5.10) for $1 \le n \le N$ and $1 \le q \le Q$, and define the *supremizing* space as

$$V_N \equiv \text{span}\left\{\sum_{q=1}^{Q} \Theta^q(\mu) T^q \zeta_n,\ n = 1, \ldots, N\right\}. \tag{5.20}$$

We make a few observations: first, while the infimizing space W_N effects good approximation, the supremizing space V_N is crucial for stability of the reduced-basis approximation; second, the supremizing space is related to infimizing space through the choice of ζ_i; third, unlike earlier definitions of reduced-basis spaces, the supremizing space is now parameter-dependent — this will require modifications of the offline/online computational procedure;

and fourth, even though we need NQ functions, the $T^q\zeta_n$, the supremizing space has dimension N. See [131] for greater details including the important proof of *good* behavior of the discrete inf-sup parameter essential to both approximation and stability.

With the defined infimizing space W_N and supremizing space V_N, we can readily obtain $u_N(\mu) \in W_N$ and $\psi_N(\mu) \in W_N$ from

$$a(u_N(\mu), v; \mu) = f(v), \quad \forall\, v \in V_N \ ; \tag{5.21}$$

$$a(v, \psi_N(\mu); \mu) = -\ell(v), \quad \forall\, v \in V_N \ ; \tag{5.22}$$

which are Petrov-Galerkin projections onto W_N for the primal and dual problems, respectively. Our output approximation is then given by

$$s_N(\mu) = \ell(u_N(\mu)) - f(\psi_N(\mu)) + a(\psi_N(\mu), u_N(\mu); \mu) \ ; \tag{5.23}$$

the additional adjoint terms will improve the accuracy [88, 108].

Finally, we have two important remarks. First, there are significant computational and conditioning advantages associated with a "segregated" approach in which we introduce separate primal $W_N^{\rm pr} = \text{span}\{u(\mu_j^{\rm pr}), 1 \le j \le N\}$ and dual $W_N^{\rm du} = \text{span}\{\psi(\mu_j^{\rm du}), 1 \le j \le N\}$ approximation spaces for $u(\mu)$ and $\psi(\mu)$, respectively. Particularly, if a is symmetric and $\ell = f$, then there will probably be degeneracy in the spaces W_N and V_N and ill-conditioning in our reduced-basis systems for the "nonsegregated" approach described above; but this is usually not the case for the segregated approach. Second, there is another simple reduced-basis approximation that can work very well for the output accuracy: we introduce $W_N = \text{span}\{u(\mu_j^{\rm pr}), 1 \le j \le N\}$ and $V_N = \text{span}\{\psi(\mu_j^{\rm du}), 1 \le j \le N\}$, and evaluate $s_N(\mu) = \ell(u_N(\mu))$, where $u_N(\mu) \in W_N$ satisfies $a(u_N(\mu), v; \mu) = f(v), \forall v \in V_N$. This simple approach may lead to high accuracy for the output approximation, albeit at the loss of stability.

It should be clear that we include the Petrov-Galerkin projection mainly for the sake of completeness and will only use the Galerkin projection for all numerical examples in the thesis.

5.2.3 *A Priori* Convergence Theory

We shall demonstrate the optimal convergence rate of $u_N(\mu) \to u(\mu)$ and $s(\mu) \to s_N(\mu)$ for the Galerkin projection (see [131] for convergence results in the case of Petrov-Galekin). To begin, we introduce the operator $T_N^\mu : W_N \to W_N$ such that, for any $w_N \in W_N$,

$$(T_N^\mu w_N, v_N)_X = a(w_N, v_N; \mu), \quad \forall\, v_N \in W_N \ .$$

We then define $\beta_N(\mu) \in \mathbb{R}$ as

$$\beta_N(\mu) \equiv \inf_{w_N \in W_N} \sup_{v_N \in W_N} \frac{a(w_N, v_N; \mu)}{\|w_N\|_X \|v_N\|_X} \ , \tag{5.24}$$

and note that

$$\beta_N(\mu) = \inf_{w_N \in W_N} \frac{\|T_N^\mu w_N\|_X}{\|w_N\|_X} \ .$$

It thus follows that

$$\beta_N(\mu)\|w_N\|_X \|T_N^\mu w_N\|_X \le a(w_N, T_N^\mu w_N; \mu), \ \forall\, w_N \in W_N \ . \tag{5.25}$$

We now demonstrate that if $\beta_N(\mu) \ge \beta_0 > 0, \ \forall \mu \in \mathcal{D}$, then $u_N(\mu)$ is optimal in the X-norm

$$\|u(\mu) - u_N(\mu)\|_X \le \left(1 + \frac{\gamma_0}{\beta_0}\right) \min_{w_N \in W_N} \|u(\mu) - w_N\|_X \ . \tag{5.26}$$

Proof. We first note from (5.12) and (5.16) that

$$a(u(\mu) - u_N(\mu), v; \mu) = 0, \quad \forall\, v \in W_N \ . \tag{5.27}$$

It thus follows for any $w_N \in W_N$ that

$$\begin{aligned}
\beta_N(\mu)\|w_N - u_N\|_X \|T_N^\mu(w_N - u_N)\|_X &\le a(w_N - u_N, T_N^\mu(w_N - u_N); \mu) \\
&= a(w_N - u + u - u_N, T_N^\mu(w_N - u_N); \mu) \\
&= a(w_N - u, T_N^\mu(w_N - u_N); \mu) \\
&\qquad + a(u - u_N, T_N^\mu(w_N - u_N); \mu) \\
&\le \gamma(\mu)\|u - w_N\|_X \|T_N^\mu(w_N - u_N)\|_X \ .
\end{aligned} \tag{5.28}$$

The desired result immediately follows from (5.28), the triangle inequality, and our hypothesis on $\beta_N(\mu)$. □

In the compliance case $\ell = f$, we may further show for any $w_N \in W_N$ that

$$\begin{aligned} |s(\mu) - s_N(\mu)| &= |a(u(\mu) - u_N(\mu), u(\mu); \mu)| \\ &= |a(u(\mu) - u_N(\mu), u(\mu) - w_N; \mu)| \\ &\leq \gamma(\mu) \|u(\mu) - u_N(\mu)\|_X \|u(\mu) - w_N\|_X \\ &\leq \gamma_0 \left(1 + \frac{\gamma_0}{\beta_0}\right) \min_{w_N \in W_N} \|u(\mu) - w_N\|_X^2 ; \end{aligned} \tag{5.29}$$

from symmetry of a, Galerkin orthogonality (5.27), continuity condition, and (5.26). Note that $s_N(\mu)$ converges to $s(\mu)$ as the square of error in the field variable.

5.3 *A Posteriori* Error Estimation

5.3.1 Objective

We wish to develop *a posteriori* error bounds $\Delta_N(\mu)$ and $\Delta_N^s(\mu)$ such that

$$\|u(\mu) - u_N(\mu)\|_X \leq \Delta_N(\mu) \ , \tag{5.30}$$

and

$$|s(\mu) - s_N(\mu)| \leq \Delta_N^s(\mu) \ . \tag{5.31}$$

It shall prove convenient to introduce the notion of effectivity, defined (here) as

$$\eta_N(\mu) \equiv \frac{\Delta_N(\mu)}{\|u(\mu) - u_N(\mu)\|_X}, \qquad \eta_N^s(\mu) \equiv \frac{\Delta_N^s(\mu)}{|s(\mu) - s_N(\mu)|} \ . \tag{5.32}$$

Our certainty requirement (5.30) and (5.31) may be stated as $\eta_N(\mu) \geq 1$ and $\eta_N^s(\mu) \geq 1$, $\forall\, \mu \in \mathcal{D}^\mu$. However, for efficiency, we must also require $\eta_N(\mu) \leq C_\eta$ and $\eta_N^s(\mu) \leq C_\eta$, where $C_\eta \geq 1$ is a constant independent of N and μ; preferably, C_η is close to unity, thus ensuring that we choose the *smallest* N — and hence most economical — reduced-basis approximation consistent with the specified error tolerance.

5.3.2 Error Bounds

We assume that we may calculate μ-dependent lower bound $\hat{\beta}(\mu)$ for the inf-sup parameter $\beta(\mu)$: $\beta(\mu) \geq \hat{\beta}(\mu) \geq \beta_0 > 0, \forall \mu \in \mathcal{D}$. The calculation of $\hat{\beta}(\mu)$ has been extensively studied in the previous chapter. We next introduce the dual norm of the residual

$$\varepsilon_N(\mu) = \sup_{v \in X} \frac{r(v;\mu)}{\|v\|_X} , \tag{5.33}$$

where

$$r(v;\mu) = f(v) - a(u_N(\mu), v; \mu), \quad \forall\, v \in X \tag{5.34}$$

is the residual associated with $u_N(\mu)$.

We can now define our energy error bound

$$\Delta_N(\mu) \equiv \frac{\varepsilon_N(\mu)}{\hat{\beta}(\mu)} , \tag{5.35}$$

and output error bound

$$\Delta_N^s(\mu) \equiv \varepsilon_N^2(\mu)/\hat{\beta}(\mu) . \tag{5.36}$$

We shall prove that $\Delta_N(\mu)$ and $\Delta_N^s(\mu)$ are rigorous and sharp bounds for $\|u(\mu) - u_N(\mu)\|_X$ and $|s(\mu) - s_N(\mu)|$, respectively.

5.3.3 Bounding Properties

Proposition 4. *For the error bounds $\Delta_N(\mu)$ of (5.35) and $\Delta_N^s(\mu)$ of (5.36), the corresponding effectivities satisfy*

$$1 \leq \eta_N(\mu) \leq \frac{\gamma(\mu)}{\hat{\beta}(\mu)}, \quad \forall\, \mu \in \mathcal{D} , \tag{5.37}$$

$$1 \leq \eta_N^s(\mu), \quad \forall \mu \in \mathcal{D} . \tag{5.38}$$

Proof. We first note from (5.12) and (5.34) that the error $e(\mu) \equiv u(\mu) - u_N(\mu)$ satisfies

$$a(e(\mu), v; \mu) = r(v; \mu), \quad \forall\, v \in X, \tag{5.39}$$

Furthermore, from standard duality argument we have

$$\varepsilon_N(\mu) = \|\hat{e}(\mu)\|_X \ , \tag{5.40}$$

where

$$(\hat{e}(\mu), v)_X = r(v; \mu), \quad \forall\, v \in X \ . \tag{5.41}$$

It then follows from (5.4), (5.39), and (5.41) that

$$\|\hat{e}(\mu)\|_X = \|T^\mu e(\mu)\|_X \ . \tag{5.42}$$

In addition, from (5.5) we know that

$$\|e(\mu)\|_X = \frac{\|T^\mu e(\mu)\|_X}{\sigma(e(\mu); \mu)} \ . \tag{5.43}$$

It thus follows from (5.32), (5.35), (5.40), (5.42), and (5.43) that

$$\eta_N(\mu) = \frac{\sigma(e(\mu); \mu)}{\hat{\beta}(\mu)} \ ; \tag{5.44}$$

this proves the desired result (5.37) since $\gamma(\mu) \geq \sigma(e(\mu); \mu) \geq \beta(\mu) \geq \hat{\beta}(\mu)$.

Finally, it follows from symmetry of a, compliance of ℓ, (5.12), Galerkin orthogonality, (5.39), and the result (5.37) that

$$\begin{aligned} |s(\mu) - s_N(\mu)| &= |a(e(\mu), u(\mu); \mu)| \\ &= |a(e(\mu), e(\mu); \mu)| \\ &= |r(e(\mu); \mu)| \\ &\leq \|r\|_{X'} \|e(\mu)\|_X \\ &\leq \frac{\|\hat{e}(\mu)\|_X^2}{\hat{\beta}(\mu)}. \end{aligned}$$

This concludes the proof. □

5.3.4 Offline/Online Computational Procedure

It remains to develop associated offline-online computational procedure for the efficient evaluation of ε_N. To begin, we note from our reduced-basis approximation $u_N(\mu) = \sum_{n=1}^N u_{N\,n}(\mu)\,\zeta_n$ and affine assumption (5.8) that $r(v;\mu)$ may be expressed as

$$r(v;\mu) = f(v) - \sum_{q=1}^{Q}\sum_{n=1}^{N} \Theta^q(\mu) u_{N\,n}(\mu)\, a^q(\zeta_n, v), \quad \forall\, v \in X. \tag{5.45}$$

It thus follows from (5.41) and (5.45) that $\hat{e}(\mu) \in X$ satisfies

$$(\hat{e}(\mu), v)_X = f(v) - \sum_{q=1}^{Q}\sum_{n=1}^{N} \Theta^q(\mu)\, u_{N\,n}(\mu)\, a^q(\zeta_n, v), \quad \forall\, v \in X. \tag{5.46}$$

The critical observation is that the right-hand side of (5.46) is a sum of products of parameter-dependent functions and parameter-independent linear functionals. In particular, it follows from linear superposition that we may write $\hat{e}(\mu) \in X$ as

$$\hat{e}(\mu) = \mathcal{C} + \sum_{q=1}^{Q}\sum_{n=1}^{N} \Theta^q(\mu)\, u_{N\,n}(\mu)\, \mathcal{L}_n^q\ , \tag{5.47}$$

where $(\mathcal{C}, v)_X = f(v)$, $\forall\, v \in X$, and $(\mathcal{L}_n^q, v)_X = -a^q(\zeta_n, v)$, $\forall\, v \in X$, $1 \le n \le N$, $1 \le q \le Q$; note that the latter are simple parameter-independent (scalar or vector) Poisson, or Poisson-like, problems. It thus follows that

$$\begin{aligned} \|\hat{e}(\mu)\|_X^2 \;=\; (\mathcal{C},\mathcal{C})_X + \sum_{q=1}^{Q}\sum_{n=1}^{N} \Theta^q(\mu)\, u_{N\,n}(\mu) \Bigg\{ & 2(\mathcal{C}, \mathcal{L}_n^q)_X \\ & + \sum_{q'=1}^{Q}\sum_{n'=1}^{N} \Theta^{q'}(\mu)\, u_{N\,n'}(\mu)\, (\mathcal{L}_n^q, \mathcal{L}_{n'}^{q'})_X \Bigg\}\ . \end{aligned} \tag{5.48}$$

The expression (5.48) is the sum of products of parameter-dependent (simple, known) functions and parameter-independent inner products. The offline-online decomposition is now clear.

In the offline stage — performed once — we first solve for $\mathcal{C}$ and $\mathcal{L}_n^q$, $1 \le n \le N$, $1 \le q \le Q$; we then evaluate and save the relevant parameter-independent inner products $(\mathcal{C}, \mathcal{C})_X$, $(\mathcal{C}, \mathcal{L}_n^q)_X$, $(\mathcal{L}_n^q, \mathcal{L}_{n'}^{q'})_X$, $1 \le n, n' \le N$, $1 \le q, q' \le Q$. Note that all quantities computed in the offline stage are independent of the parameter μ.

In the online stage — performed many times, for each new value of μ "in the field" — we simply evaluate the sum (5.48) in terms of the $\Theta^q(\mu), u_{N\,n}(\mu)$ and the precalculated and stored (parameter-independent) $(\cdot,\cdot)_X$ inner products. The operation count for the online stage is only $O(Q^2N^2)$ — again, the essential point is that the online complexity is *independent of* $\mathcal{N}$, the dimension of the underlying truth finite element approximation space. We further note that unless Q is quite large, the online cost associated with the calculation of the dual norm of the residual is commensurate with the online cost associated with the calculation of $s_N(\mu)$.

5.4 Numerical Results

In this section, we shall present and discuss several numerical results for our model problem. We consider the parameter domain $\mathcal{D} \equiv [3.2, 4.8] \times [0.9, 1.1] \times [0.15, 0.25]$. Note that $\mathcal{D}$ does not contain any resonances, and hence $\beta(\mu)$ is bounded away from zero; however, $\omega^2 = 3.2$ and $\omega^2 = 4.8$ are in fact quite close to corresponding natural frequencies, and hence the problem is distinctly non-coercive.

Recall that our affine assumption is applied for $Q = 10$, and the $\Theta^q(\mu)$, $a^q(w, v), 1 \le q \le Q$, were summarized in Table 4.1. We define $(w, v)_X = \sum_{q=2}^{Q} a^q(w, v)$ for our bound conditioner; thanks to the Dirichlet conditions at $x_1 = 0$, $(\cdot,\cdot)_X$ is appropriately coercive. We further observe that $\Theta^1(\mu) = 1(\Gamma_1 = 0)$ and we can thus disregard the $q = 1$ term in our continuity bound. We may then choose $|v|_q^2 = a^q(v, v)$, $2 \le q \le Q$, since the $a^q(\cdot,\cdot)$ are positive semi-definite; it thus follows from the Cauchy-Schwarz inequality that $\Gamma^q = 1$, $2 \le q \le Q$. Furthermore, from (4.34), we directly obtain $C_X = 1$. We readily perform piecewise-constant construction of the inf-sup lower bounds: we can cover $\mathcal{D}$ (for $\bar{\epsilon}_\beta = 0.2$) such that (4.36) and (4.37) are satisfied with only $J = 84$ polytopes; in this particular case the $\mathcal{P}^{\bar{\mu}_j}$, $1 \le j \le J$, are hexahedrons such that $|\mathcal{V}^{\mu_j}| = 8$, $1 \le j \le J$.

Armed with the inf-sup lower bounds, we can now pursue the adaptive sampling

strategy described in Section 3.3.5: for $\epsilon_{\text{tol, min}} = 10^{-3}$ and $n_{\text{F}} = 729$ we obtain $N_{\max} = 32$ (as shown in Figure 5-2) such that $\epsilon_{N_{\max}} \equiv \Delta_{N_{\max}}(\mu^{\text{pr}}_{N_{\max}}) = 9.03 \times 10^{-4}$. We observe that more sample points lie at the two ends of the frequency range, $\omega^2 = 3.2$ and $\omega^2 = 4.8$. This is because $\omega^2 = 3.2$ and $\omega^2 = 4.8$ are quite close to corresponding natural frequencies, at which the solutions vary greatly and the inf-sup parameter decreases rapidly to zero.

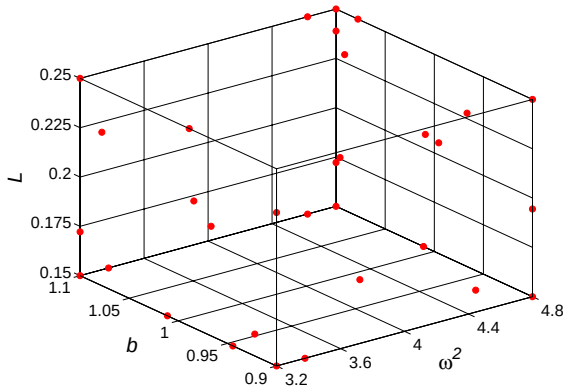
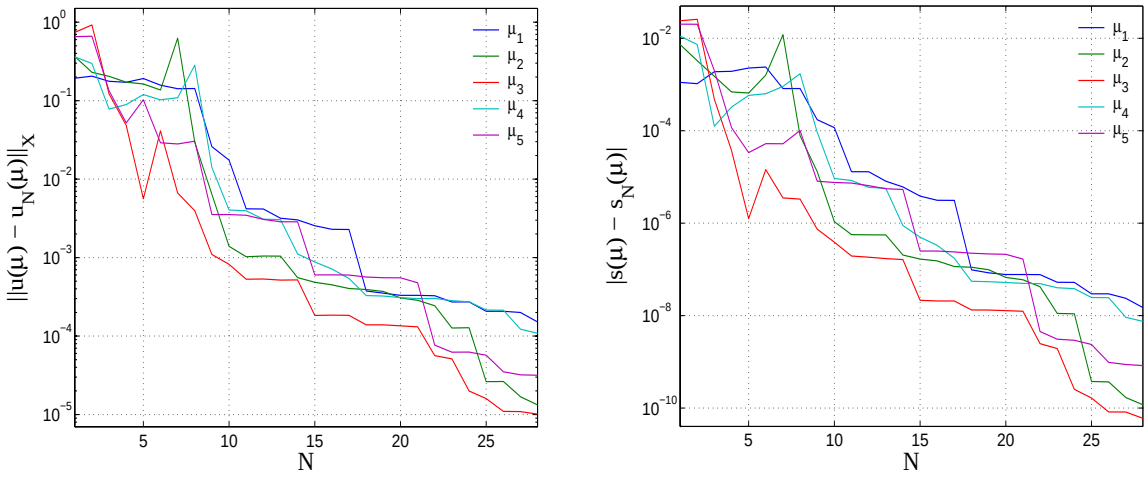


Figure 5-2: Sample $S_{N_{\max}}$ obtained with the adaptive sampling procedure for $N_{\max} = 32$.



(a) (b)

Figure 5-3: Convergence for the reduced-basis approximations at test points: (a) error in the solution and (b) error in the output.

We next present in Figure 5-3 the error in the output and the error in the solution as a function of N for five random test points. We observe that initially for small values of N (less than 10) the errors are quite significant, oscillating, and not reduced by increasing

N. This is because for small values of N the basis functions included in the reduced-basis space have no good approximation properties for the solutions at the test points. As we further increase N we see that the errors decrease rapidly with N; that the convergence rate is quite similar for all test points; and that the error in the output is square of the error in the solution (note that the "square" effect is typically true for the compliance case — here the model problem is as such).

We furthermore present in Table 5.1 $\Delta_{N,\max,\text{rel}}$, $\eta_{N,\text{ave}}$, $\Delta^s_{N,\max}$, and $\eta^s_{N,\text{ave}}$ as a function of N. Here $\Delta_{N,\max,\text{rel}}$ is the maximum over Ξ_{Test} of $\Delta_N(\mu)/\|u_{\max}\|_X$, $\eta_{N,\text{ave}}$ is the average over Ξ_{Test} of $\Delta_N(\mu)/\|u(\mu) - u_N(\mu)\|_X$, $\Delta^s_{N,\max,\text{rel}}$ is the maximum over Ξ_{Test} of $\Delta^s_N(\mu)/|s_{\max}|$, and $\eta^s_{N,\text{ave}}$ is the average over Ξ_{Test} of $\Delta^s_N(\mu)/|s(\mu) - s_N(\mu)|$. Here $\Xi_{\text{Test}} \in (\mathcal{D})^{343}$ is a random sample of size 343; $\|u_{\max}\|_X \equiv \max_{\mu\in\Xi_{\text{Test}}} \|u(\mu)\|_X$ and $|s_{\max}| \equiv \max_{\mu\in\Xi_{\text{Test}}} |s(\mu)|$. We observe that the reduced-basis approximation converges very rapidly, and that our rigorous error bounds are in fact quite sharp. The effectivities are not quite $O(1)$ primarily due to the relatively crude piecewise-constant inf-sup lower bound. Effectivities $O(10)$ are acceptable within the reduced-basis context: thanks to the very rapid convergence rates, the "unnecessary" increase in N — to achieve a given error tolerance — is proportionately very small.

N	$\Delta_{N,\max,\text{rel}}$	$\eta_{N,\text{ave}}$	$\Delta^s_{N,\max,\text{rel}}$	$\eta^s_{N,\text{ave}}$
12	1.54×10^{-1}	13.41	3.31×10^{-2}	15.93
16	3.40×10^{-2}	12.24	2.13×10^{-3}	14.86
20	1.58×10^{-2}	13.22	4.50×10^{-4}	15.44
24	5.91×10^{-3}	12.56	4.81×10^{-5}	14.45
28	2.42×10^{-3}	12.44	9.98×10^{-6}	14.53

Table 5.1: Effectivities for the model problem.

Turning now to the computational effort, we present in Table 5.2 the time ratio normalized to the running time of computing $s_N(\mu)$ for $N = 20$ (recall that for $N \geq 20$, $|\Delta^s_N(\mu)/s_N(\mu)| \leq 4.50 \times 10^{-4}$). We achieve computational savings of $O(500)$: N is very small thanks to the good convergence properties of S_N and hence W_N; and the marginal cost to evaluate $s_N(\mu)$ and $\Delta^s_N(\mu)$ depends only on N, *not* on $\mathcal{N}$ thanks to the offline-online decomposition. We emphasize that the reduced-basis entry does *not* include the

extensive offline computations — and is thus only meaningful in the real-time or many-query contexts. As illustrated in Chapter 9, the significant reduction in computational time enables the deployed/real-time Assess-Act scenario in which we Assess all possible crack parameters consistent with experimental measurements through robust parameter estimation procedures and subsequently Act upon our earlier crack assessments through adaptive optimization procedures to provide an intermediate and fail-safe action.

	Online Time	Online Time	Time
N	$s_N(\mu)$	$\Delta_N^s(\mu)$	$s(\mu)$ ($\mathcal{N} = 14{,}662$)
12	0.65	0.84	
16	0.9	0.94	
20	1.0	1.05	882
24	1.23	1.29	
28	1.45	1.54	

Table 5.2: Time savings per online evaluation.

5.5 Additional Example: Material Damage Model

5.5.1 Problem Description

We consider a two-dimensional sandwich plate with a rectangular flaw at the core layer of three lamina: the (original) domain $\tilde{\Omega} \subset \mathbb{R}^2$ is defined as $[0, 2] \times [0, 1]$; the thickness of core layer is 0.8 while that of two face layers is 0.1; the left surface of the plate, $\tilde{\Gamma}_{\rm D}$, is secured; the top and bottom boundaries, $\tilde{\Gamma}_{\rm N}$, are stress-free; and the right boundary, $\tilde{\Gamma}_{\rm F}$, is subject to a vertical oscillatory uniform force of frequency $\tilde{\omega}$. To simplify the problem, we assume that the flaw is throughout the thickness of the core. The flaw length is denoted by L and the distance from the center of the flaw to the left surface is denoted by b. The rectangular flaw is considered as a damaged zone in which density of the material remains the same but the elastic constants are reduced by a factor δ (damage factor). Note that $\delta = 1$ indicates no flaw while $\delta = 0$ means a void in the sandwich plate.

We model the plate as plane-stress linear elastic lamina structure in which material properties of the two face layers and core layer are shown in Table 5.3. We introduce

nondimensional quantities $\omega^2 = (\tilde{\omega})^2\tilde{\rho}^c/\tilde{E}^c$, $E^c = \tilde{E}^c/\tilde{E}^c$, $E^f = \tilde{E}^f/\tilde{E}^c$, $\rho^c = \tilde{\rho}^c/\tilde{\rho}^c$, and $\rho^f = \tilde{\rho}^f/\tilde{\rho}^c$, where $\tilde{E}^c$ and $\tilde{\rho}^c$ are the Young's modulus and density of the core layer and $\tilde{E}^f$ and $\tilde{\rho}^f$ are the Young's modulus and density of the face layer. Our input is $\mu \equiv (\mu_{(1)}, \mu_{(2)}, \mu_{(3)}, \mu_{(4)}) = (\omega^2, b, L, \delta) \in \mathcal{D}^\omega \times \mathcal{D}^{b,L,\delta}$, where $\mathcal{D}^{b,L,\delta} \equiv [0.9, 1.1] \times [0.5, 0.7] \times [0.4, 0.6]$; our output is the (oscillatory) amplitude of the average vertical displacement on the right edge of the plate.

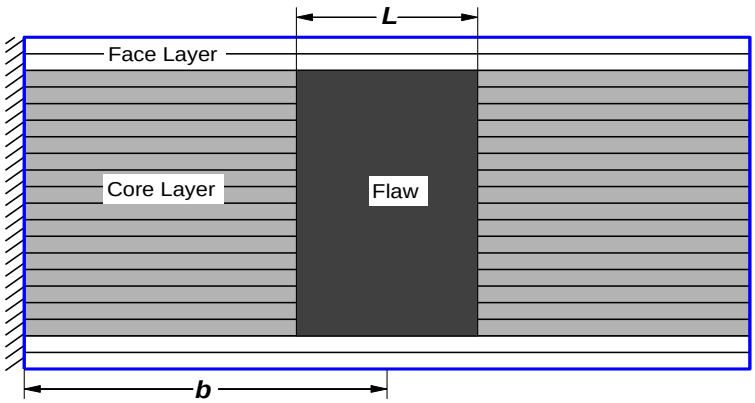


Figure 5-4: Rectangular flaw in a sandwich plate.

	Young's modulus (N/m^2)	Density (kg/m^3)	Poisson ratio
Face layers	1.67×10^{10}	1760	0.3
Core layer	0.013×10^{10}	130	0.3

Table 5.3: Material properties of core layer and face layers.

The governing equations for the displacement field $\tilde{u}(\tilde{x}; \mu) \in \tilde{X}(\mu)$ are thus

$$\frac{\partial \tilde{\sigma}_{11}}{\partial \tilde{x}_1} + \frac{\partial \tilde{\sigma}_{12}}{\partial \tilde{x}_2} + \rho\omega^2 \tilde{u}_1^2 = 0$$
$$\frac{\partial \tilde{\sigma}_{12}}{\partial \tilde{x}_1} + \frac{\partial \tilde{\sigma}_{22}}{\partial \tilde{x}_2} + \rho\omega^2 \tilde{u}_2^2 = 0$$

$$\tilde{\varepsilon}_{11} = \frac{\partial \tilde{u}_1}{\partial \tilde{x}_1}, \quad \tilde{\varepsilon}_{22} = \frac{\partial \tilde{u}_2}{\partial \tilde{x}_2}, \quad 2\tilde{\varepsilon}_{12} = \left(\frac{\partial \tilde{u}_1}{\partial \tilde{x}_2} + \frac{\partial \tilde{u}_2}{\partial \tilde{x}_1} \right),$$

$$\left\{ \begin{array}{c} \tilde{\sigma}_{11} \\ \tilde{\sigma}_{22} \\ \tilde{\sigma}_{12} \end{array} \right\} = \left[\begin{array}{ccc} c_{11} & c_{12} & 0 \\ c_{12} & c_{22} & 0 \\ 0 & 0 & c_{66} \end{array} \right] \left\{ \begin{array}{c} \tilde{\varepsilon}_{11} \\ \tilde{\varepsilon}_{22} \\ \tilde{\varepsilon}_{12} \end{array} \right\}$$

where the constitutive constants are given by

$$c_{11} = \frac{E}{1-\nu^2}, \qquad c_{22} = c_{11}, \qquad c_{12} = \frac{E\nu}{1-\nu^2}, \qquad c_{66} = \frac{E}{2(1+\nu)} \ .$$

Note importantly that, in the above equations, the density ρ and Young's modulus E are different for face layers, core layer, and damage zone; in particular, we have $\rho = \rho^{\mathrm{f}}$, $E = E^{\mathrm{f}}$ in face layers, $\rho = \rho^{\mathrm{c}}$, $E = E^{\mathrm{c}}$ in core layer, and $\rho = \rho^{\mathrm{c}}$, $E = \delta E^{\mathrm{c}}$ in damage zone. The boundary conditions on the (secured) left edge are

$$\tilde{u}_1 = \tilde{u}_2 = 0, \quad \text{on } \tilde{\Gamma}_{\mathrm{D}} \ .$$

The boundary conditions on the top and bottom boundaries are

$$\begin{aligned} \tilde{\sigma}_{11}\hat{\tilde{n}}_1 + \tilde{\sigma}_{12}\hat{\tilde{n}}_2 = 0 \quad \text{on } \tilde{\Gamma}_{\mathrm{N}} \ , \\ \tilde{\sigma}_{12}\hat{\tilde{n}}_1 + \tilde{\sigma}_{22}\hat{\tilde{n}}_2 = 0 \quad \text{on } \tilde{\Gamma}_{\mathrm{N}} \ . \end{aligned}$$

The boundary conditions on the right edge are

$$\begin{aligned} \tilde{\sigma}_{11}\hat{\tilde{n}}_1 + \tilde{\sigma}_{12}\hat{\tilde{n}}_2 = 0 \quad \text{on } \tilde{\Gamma}_{\mathrm{F}} \ , \\ \tilde{\sigma}_{12}\hat{\tilde{n}}_1 + \tilde{\sigma}_{22}\hat{\tilde{n}}_2 = 1 \quad \text{on } \tilde{\Gamma}_{\mathrm{F}} \ . \end{aligned}$$

We now introduce $\tilde{X}(\mu)$ — a quadratic finite element truth approximation subspace (of dimension $\mathcal{N} = 14{,}640$) of $X^{\mathrm{e}}(\mu) = \{\tilde{v} \in (H^1(\tilde{\Omega}(b,L)))^2 \,|\, \tilde{v}|_{\tilde{\Gamma}_{\mathrm{F}}} = 0\}$. The weak formulation can then be derived as

$$\tilde{a}(\tilde{u}(\mu), \tilde{v}; \mu) = \tilde{f}(\tilde{v}), \quad \forall\, \tilde{v} \in \tilde{X}(\mu) \tag{5.49}$$

where

$$\begin{aligned}\tilde{a}(w,v;\mu) \;=\; & c_{12}\int_{\tilde{\Omega}}\left(\frac{\partial v_1}{\partial \tilde{x}_1}\frac{\partial w_2}{\partial \tilde{x}_2}+\frac{\partial v_2}{\partial \tilde{x}_2}\frac{\partial w_1}{\partial \tilde{x}_1}\right)+c_{66}\int_{\tilde{\Omega}}\left(\frac{\partial v_1}{\partial \tilde{x}_2}\frac{\partial w_2}{\partial \tilde{x}_1}+\frac{\partial v_2}{\partial \tilde{x}_1}\frac{\partial w_1}{\partial \tilde{x}_2}\right)\\ & +c_{11}\int_{\tilde{\Omega}}\left(\frac{\partial v_1}{\partial \tilde{x}_1}\frac{\partial w_1}{\partial \tilde{x}_1}\right)+c_{66}\int_{\tilde{\Omega}}\left(\frac{\partial v_2}{\partial \tilde{x}_1}\frac{\partial w_2}{\partial \tilde{x}_1}\right)\\ & +c_{22}\int_{\tilde{\Omega}}\left(\frac{\partial v_2}{\partial \tilde{x}_2}\frac{\partial w_2}{\partial \tilde{x}_2}\right)+c_{66}\int_{\tilde{\Omega}}\left(\frac{\partial v_1}{\partial \tilde{x}_2}\frac{\partial w_1}{\partial \tilde{x}_2}\right)-\omega^2\int_{\tilde{\Omega}}\rho(w_1v_1+w_2v_2)\ ,\end{aligned}$$

$$\tilde{f}(v)=\int_{\tilde{\Gamma}_{\rm F}} v_2\ .$$

The output is given by $\tilde{s}(\mu)=\tilde{\ell}(\tilde{u}(\mu))$, where $\tilde{\ell}(v)=\tilde{f}(v)$; we are thus "in compliance."

We now define a reference domain corresponding to the geometry $b=b_r=1$ and $L=L_r=0.5$. We then map $\tilde{\Omega}(b,L)\rightarrow\Omega\equiv\tilde{\Omega}(b_r,L_r)$ by a continuous piecewise-affine (in fact, piecewise-dilation-in-x_1) transformation. We define three subdomains, $\Omega_1\equiv]0,b_r-L_r/2[\times]0,1[$, $\Omega_2\equiv]b_r-L_r/2,b_r+L_r/2[\times]0,1[$, $\Omega_3\equiv]b_r+L_r/2,2[\times]0,1[$, such that $\bar{\Omega}=\bar{\Omega}_1\cup\bar{\Omega}_2\cup\bar{\Omega}_3$; in addition, we define $\Omega_1^{\rm c}\equiv]0,b_r-L_r/2[\times]0.1,0.9[$, $\Omega_1^{\rm f}\equiv\Omega_1\backslash\Omega_1^{\rm c}$, $\Omega^{\rm d}\equiv]b_r-L_r/2,b_r+L_r/2[\times]0.1,0.9[$, $\Omega_2^{\rm f}\equiv\Omega_2\backslash\Omega^{\rm d}$, $\Omega_2^{\rm c}\equiv]b_r+L_r/2,2[\times]0.1,0.9[$, $\Omega_2^{\rm f}\equiv\Omega_3\backslash\Omega_2^{\rm c}$.

We thus arrive at the desired form (5.12) in which $f(v)=\int_{\tilde{\Gamma}_{\rm F}}v_2$ and the bilinear form a is expressed as an affine sum for $Q=13$; the $\Theta^q(\mu)$, $a^q(w,v)$, $1\le q\le 13$, are given in Table 5.4. For plane stress and a linear isotropic solid, the constitutive constants in Table 5.4 are given by

$$c_{11}^{\rm c}=\frac{E^{\rm c}}{1-\nu^2},\qquad c_{22}^{\rm c}=c_{11}^{\rm c},\qquad c_{12}^{\rm c}=\frac{E^{\rm c}\nu}{1-\nu^2},\qquad c_{66}^{\rm c}=\frac{E^{\rm c}}{2(1+\nu)},$$

$$c_{11}^{\rm f}=\frac{E^{\rm f}}{1-\nu^2},\qquad c_{22}^{\rm f}=c_{11}^{\rm f},\qquad c_{12}^{\rm f}=\frac{E^{\rm f}\nu}{1-\nu^2},\qquad c_{66}^{\rm f}=\frac{E^{\rm f}}{2(1+\nu)},$$

where $\nu=0.3$ is the Poisson ratio and the normalized Young's modulus $E^{\rm c}$ and $E^{\rm f}$ are introduced earlier.

q	$\Theta^q(\mu)$	$a^q(w,v)$
1	1	$\sum_{r=1}^{2}\left\{c_{12}^{\mathrm{c}}\int_{\Omega_r^{\mathrm{c}}}\left(\frac{\partial v_1}{\partial x_1}\frac{\partial w_2}{\partial x_2}+\frac{\partial v_2}{\partial x_2}\frac{\partial w_1}{\partial x_1}\right)+c_{66}^{\mathrm{c}}\int_{\Omega_r^{\mathrm{c}}}\left(\frac{\partial v_1}{\partial x_2}\frac{\partial w_2}{\partial x_1}+\frac{\partial v_2}{\partial x_1}\frac{\partial w_1}{\partial x_2}\right)\right\}$ $+\sum_{r=1}^{3}\left\{c_{12}^{\mathrm{f}}\int_{\Omega_r^{\mathrm{f}}}\left(\frac{\partial v_1}{\partial x_1}\frac{\partial w_2}{\partial x_2}+\frac{\partial v_2}{\partial x_2}\frac{\partial w_1}{\partial x_1}\right)+c_{66}^{\mathrm{f}}\int_{\Omega_r^{\mathrm{f}}}\left(\frac{\partial v_1}{\partial x_2}\frac{\partial w_2}{\partial x_1}+\frac{\partial v_2}{\partial x_1}\frac{\partial w_1}{\partial x_2}\right)\right\}$
2	δ	$c_{12}^{\mathrm{c}}\int_{\Omega^{\mathrm{d}}}\left(\frac{\partial v_1}{\partial x_1}\frac{\partial w_2}{\partial x_2}+\frac{\partial v_2}{\partial x_2}\frac{\partial w_1}{\partial x_1}\right)+c_{66}^{\mathrm{c}}\int_{\Omega^{\mathrm{d}}}\left(\frac{\partial v_1}{\partial x_2}\frac{\partial w_2}{\partial x_1}+\frac{\partial v_2}{\partial x_1}\frac{\partial w_1}{\partial x_2}\right)$
3	$\frac{b_{\mathrm{r}}-L_{\mathrm{r}}/2}{b-L/2}$	$c_{11}^{\mathrm{c}}\int_{\Omega_1^{\mathrm{c}}}\left(\frac{\partial v_1}{\partial x_1}\frac{\partial w_1}{\partial x_1}\right)+c_{66}^{\mathrm{c}}\int_{\Omega_1^{\mathrm{c}}}\left(\frac{\partial v_2}{\partial x_1}\frac{\partial w_2}{\partial x_1}\right)$ $+c_{11}^{\mathrm{f}}\int_{\Omega_1^{\mathrm{f}}}\left(\frac{\partial v_1}{\partial x_1}\frac{\partial w_1}{\partial x_1}\right)+c_{66}^{\mathrm{f}}\int_{\Omega_1^{\mathrm{f}}}\left(\frac{\partial v_2}{\partial x_1}\frac{\partial w_2}{\partial x_1}\right)$
4	$\frac{L_{\mathrm{r}}}{L}$	$c_{11}^{\mathrm{f}}\int_{\Omega_2^{\mathrm{f}}}\left(\frac{\partial v_1}{\partial x_1}\frac{\partial w_1}{\partial x_1}\right)+c_{66}^{\mathrm{f}}\int_{\Omega_2^{\mathrm{f}}}\left(\frac{\partial v_2}{\partial x_1}\frac{\partial w_2}{\partial x_1}\right)$
5	$\delta\frac{L_{\mathrm{r}}}{L}$	$c_{11}^{\mathrm{c}}\int_{\Omega^{\mathrm{d}}}\left(\frac{\partial v_1}{\partial x_1}\frac{\partial w_1}{\partial x_1}\right)+c_{66}^{\mathrm{c}}\int_{\Omega^{\mathrm{d}}}\left(\frac{\partial v_2}{\partial x_1}\frac{\partial w_2}{\partial x_1}\right)$
6	$\frac{2-b_{\mathrm{r}}-L_{\mathrm{r}}/2}{2-b-L/2}$	$c_{11}^{\mathrm{c}}\int_{\Omega_2^{\mathrm{c}}}\left(\frac{\partial v_1}{\partial x_1}\frac{\partial w_1}{\partial x_1}\right)+c_{66}^{\mathrm{c}}\int_{\Omega_2^{\mathrm{c}}}\left(\frac{\partial v_2}{\partial x_1}\frac{\partial w_2}{\partial x_1}\right)$ $+c_{11}^{\mathrm{f}}\int_{\Omega_3^{\mathrm{f}}}\left(\frac{\partial v_1}{\partial x_1}\frac{\partial w_1}{\partial x_1}\right)+c_{66}^{\mathrm{f}}\int_{\Omega_3^{\mathrm{f}}}\left(\frac{\partial v_2}{\partial x_1}\frac{\partial w_2}{\partial x_1}\right)$
7	$\frac{b-L/2}{b_{\mathrm{r}}-L_{\mathrm{r}}/2}$	$c_{22}^{\mathrm{c}}\int_{\Omega_1^{\mathrm{c}}}\left(\frac{\partial v_2}{\partial x_2}\frac{\partial w_2}{\partial x_2}\right)+c_{66}^{\mathrm{c}}\int_{\Omega_1^{\mathrm{c}}}\left(\frac{\partial v_1}{\partial x_2}\frac{\partial w_1}{\partial x_2}\right)$ $+c_{22}^{\mathrm{f}}\int_{\Omega_1^{\mathrm{f}}}\left(\frac{\partial v_2}{\partial x_2}\frac{\partial w_2}{\partial x_2}\right)+c_{66}^{\mathrm{f}}\int_{\Omega_1^{\mathrm{f}}}\left(\frac{\partial v_1}{\partial x_2}\frac{\partial w_1}{\partial x_2}\right)$
8	$\frac{L}{L_{\mathrm{r}}}$	$c_{22}^{\mathrm{f}}\int_{\Omega_2^{\mathrm{f}}}\left(\frac{\partial v_2}{\partial x_2}\frac{\partial w_2}{\partial x_2}\right)+c_{66}^{\mathrm{f}}\int_{\Omega_2^{\mathrm{f}}}\left(\frac{\partial v_1}{\partial x_2}\frac{\partial w_1}{\partial x_2}\right)$
9	$\delta\frac{L}{L_{\mathrm{r}}}$	$c_{22}^{\mathrm{c}}\int_{\Omega^{\mathrm{d}}}\left(\frac{\partial v_2}{\partial x_2}\frac{\partial w_2}{\partial x_2}\right)+c_{66}^{\mathrm{c}}\int_{\Omega^{\mathrm{d}}}\left(\frac{\partial v_1}{\partial x_2}\frac{\partial w_1}{\partial x_2}\right)$
10	$\frac{2-b-L/2}{2-b_{\mathrm{r}}-L_{\mathrm{r}}/2}$	$c_{22}^{\mathrm{c}}\int_{\Omega_2^{\mathrm{c}}}\left(\frac{\partial v_2}{\partial x_2}\frac{\partial w_2}{\partial x_2}\right)+c_{66}^{\mathrm{c}}\int_{\Omega_2^{\mathrm{c}}}\left(\frac{\partial v_1}{\partial x_2}\frac{\partial w_1}{\partial x_2}\right)$ $+c_{22}^{\mathrm{f}}\int_{\Omega_3^{\mathrm{f}}}\left(\frac{\partial v_2}{\partial x_2}\frac{\partial w_2}{\partial x_2}\right)+c_{66}^{\mathrm{f}}\int_{\Omega_3^{\mathrm{f}}}\left(\frac{\partial v_1}{\partial x_2}\frac{\partial w_1}{\partial x_2}\right)$
11	$-\omega^2\frac{b-L/2}{b_{\mathrm{r}}-L_{\mathrm{r}}/2}$	$\int_{\Omega_1^{\mathrm{c}}}w_1v_1+w_2v_2+\rho^{\mathrm{f}}\int_{\Omega_1^{\mathrm{f}}}w_1v_1+w_2v_2$
12	$-\omega^2\frac{L}{L_{\mathrm{r}}}$	$\int_{\Omega^{\mathrm{d}}}w_1v_1+w_2v_2+\rho^{\mathrm{f}}\int_{\Omega_2^{\mathrm{f}}}w_1v_1+w_2v_2$
13	$-\omega^2\frac{2-b-L/2}{2-b_{\mathrm{r}}-L_{\mathrm{r}}/2}$	$\int_{\Omega_2^{\mathrm{c}}}w_1v_1+w_2v_2+\rho^{\mathrm{f}}\int_{\Omega_3^{\mathrm{f}}}w_1v_1+w_2v_2$

Table 5.4: Parametric functions $\Theta^q(\mu)$ and parameter-independent bilinear forms $a^q(w,v)$ for the two-dimensional damage material problem.

5.5.2 Numerical Results

We first show in Figure 5-5 the finite element mesh on which our quadratic truth approximation subspace of dimension $\mathcal{N} = 14{,}640$ is defined.

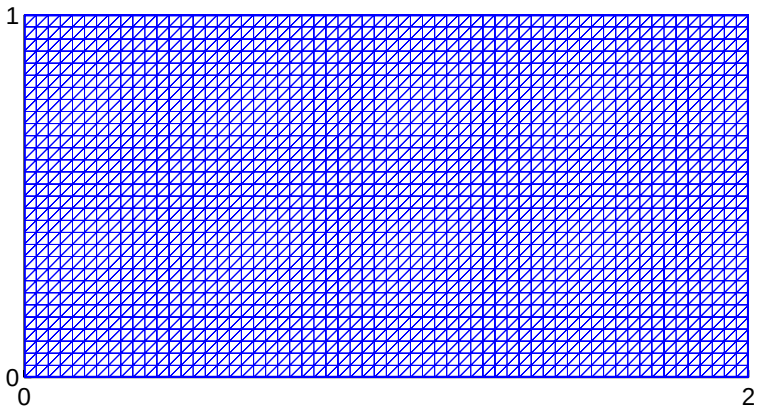


Figure 5-5: Quaratic triangular finite element mesh on the reference domain. Note that each element has six nodes.

Next we define our inner product-cum-bound conditioner as $(w, v)_X \equiv \sum_{q=3}^{Q} a^q(w, v)$; thanks to the Dirichlet conditions at $x_1 = 0$ (and also the $w_i v_i$ term), $(\cdot, \cdot)_X$ is appropriately coercive. We now observe that $\Theta(\mu) = 1$ ($\Gamma^1 = 0$) and we can thus disregard the $q = 1$ term in our continuity bounds. We may then choose

$$|v|_2^2 = c_{12}^{\mathrm{c}} \int_{\Omega^{\mathrm{d}}} \left(\frac{\partial v_1}{\partial x_1}\right)^2 + \left(\frac{\partial v_2}{\partial x_2}\right)^2 + c_{66}^{\mathrm{c}} \int_{\Omega^{\mathrm{d}}} \left(\frac{\partial v_1}{\partial x_2}\right)^2 + \left(\frac{\partial v_2}{\partial x_1}\right)^2$$

and $|v|_q^2 = a^q(v, v)$, $3 \le q \le Q$, since the $a^q(\cdot, \cdot), 3 \le q \le Q$, are positive semi-definite; it thus follows from the Cauchy-Schwarz inequality that $\Gamma^q = 1$, $2 \le q \le Q$; furthermore, from (4.34), we calculate $C_X = 1.0000$ numerically.

We shall consider three different frequencies and associated reduced-basis models: Model I for $\omega_1^2 = 0.58$, Model II for $\omega_2^2 = 1.53$, and model III for $\omega_3^2 = 2.95$; these frequencies are in fact quite close to the corresponding resonance modes, and hence the problem is distinctly noncoercive. We henceforth perform piecewise-constant construction of the inf-sup lower bounds for each model: we can cover $\mathcal{D}^{b,L,\delta}$ (for $\bar{\epsilon}_\beta = 0.5$) with $J^{\mathrm{I}} = 133$ polytopes, $J^{\mathrm{II}} = 169$ polytopes, and $J^{\mathrm{III}} = 196$ polytopes such that the Coverage and Positivity conditions are satisfied; here the $\mathcal{P}^{\bar{\mu}_j}$, $1 \le j \le J$, are hexahedrons such

that $|\mathcal{V}^{\mu_j}| = 8$, $1 \le j \le J$. Armed with the piecewise constant lower bounds, we pursue the adaptive sampling strategy: for $n_{\rm F} = 729$ we obtain $N^{\rm I}_{\rm max} = 50$, $N^{\rm II}_{\rm max} = 50$, and $N^{\rm III}_{\rm max} = 50$.

We next show the convergence of the reduced-basis approximation. We present in Tables 5.5, 5.6, and 5.7 $\Delta_{N,\rm max,rel}$, $\eta_{N,\rm ave}$, $\Delta^s_{N,\rm max,rel}$, and $\eta^s_{N,\rm ave}$ as a function of N for three models. Here $\Delta_{N,\rm max,rel}$ is the maximum over $\Xi_{\rm Test}$ of $\Delta_N(\mu)/\|u(\mu)\|_X$, $\eta_{N,\rm ave}$ is the average over $\Xi_{\rm Test}$ of $\Delta_N(\mu)/\|u(\mu) - u_N(\mu)\|_X$, $\Delta^s_{N,\rm max,rel}$ is the maximum over $\Xi_{\rm Test}$ of $\Delta^s_N(\mu)/|s(\mu)|$, and $\eta^s_{N,\rm ave}$ is the average over $\Xi_{\rm Test}$ of $\Delta^s_N(\mu)/|s(\mu) - s_N(\mu)|$; where $\Xi_{\rm Test} \in (\mathcal{D}^{\rm I})^{343}$ is a random parameter sample of size 343. We observe that the reduced-basis approximation converges very rapidly, and that our rigorous error bounds are moderately sharp. The effectivities are not quite $O(1)$ primarily due to the high-mode frequencies near resonances; but note that, thanks to the rapid convergence of the reduced-basis approximation, $O(10)$ effectivities do not significantly affect efficiency — the induced increase in reduced-basis dimension N is quite modest. Note also that $\Delta_{N,\rm max,rel}$ is not really the square order of $\Delta^s_{N,\rm max,rel}$ due to our different choice of the denominator in the respective normalization; however, the error in the output and output error bound do in fact converge quadratically with respect to the error norm and energy error bound, respectively.

Finally, we note that the total Online computational time on a Pentium 1.6GHz processor to compute $s_N(\mu)$ and $\Delta^s_N(\mu)$ to a relative error of 10^{-4} (with $N = 40$) is less than 1/279 times the Total Time to directly calculate the truth output $s(\mu) = \ell(u(\mu))$. Clearly, the savings will be even larger for problems with more complex geometry and solution structure in particular in three space dimensions. Nevertheless, even for our current very modest example, the computational economies are very significant.

N	$\Delta_{N,\rm max,rel}$	$\eta_{N,\rm ave}$	$\Delta^s_{N,\rm max,rel}$	$\eta^s_{N,\rm ave}$
10	4.33×10^{-1}	11.25	3.06×10^{-0}	15.08
20	8.02×10^{-3}	11.33	7.74×10^{-4}	12.76
30	2.43×10^{-3}	8.03	6.93×10^{-5}	12.82
40	1.49×10^{-3}	9.09	2.56×10^{-5}	13.90
50	9.81×10^{-4}	9.05	1.16×10^{-5}	12.50

Table 5.5: Convergence and effectivities for Model I.

N	$\Delta_{N,\max,\mathrm{rel}}$	$\eta_{N,\mathrm{ave}}$	$\Delta^s_{N,\max,\mathrm{rel}}$	$\eta^s_{N,\mathrm{ave}}$
10	1.56×10^{-1}	12.37	1.44×10^{-1}	35.74
20	5.23×10^{-2}	12.04	1.63×10^{-2}	30.13
30	1.71×10^{-2}	12.72	2.10×10^{-3}	27.76
40	5.44×10^{-3}	13.14	1.71×10^{-4}	25.91
50	1.05×10^{-3}	11.37	8.36×10^{-6}	17.71

Table 5.6: Convergence and effectivities for Model II.

N	$\Delta_{N,\max,\mathrm{rel}}$	$\eta_{N,\mathrm{ave}}$	$\Delta^s_{N,\max,\mathrm{rel}}$	$\eta^s_{N,\mathrm{ave}}$
10	6.17×10^{-0}	10.06	$8.49 \times 10^{+2}$	21.41
20	1.90×10^{-2}	10.35	4.92×10^{-3}	19.31
30	4.13×10^{-3}	9.07	2.38×10^{-4}	19.88
40	1.34×10^{-3}	10.42	2.42×10^{-5}	19.94
50	8.83×10^{-4}	11.09	1.14×10^{-5}	18.28

Table 5.7: Convergence and effectivities for Model III.

Chapter 6

An Empirical Interpolation Method for Nonaffine Elliptic Problems

Thus far we have developed the reduced-basis method for parametrized partial differential equations with *affine parameter* dependence. The affine assumption allows us to develop extremely efficient offline-online computational strategy; the online cost to evaluate $s_N(\mu)$ and $\Delta_N^s(\mu)$ depends only on N and Q, *not* on $\mathcal{N}$ — the dimension of the truth approximation space. Unfortunately, if the differential operator is not affine in the parameter, the online complexity is no longer independent of $\mathcal{N}$. This is because operators of *nonaffine parameter* dependence do not accommodate the separation of the generation and projection stages during the online computation.

In this chapter we describe a technique that recovers online $\mathcal{N}$ independence even in the presence of non-affine parameter dependence. In our approach, we replace non-affine functions of the parameter and spatial coordinate with collateral reduced-basis expansions. The essential ingredients of the approach are (i) good collateral reduced-basis samples and spaces, (ii) a stable and inexpensive online interpolation procedure by which to determine the collateral reduced-basis coefficients (as a function of the parameter), and (iii) an effective *a posteriori* error bounds with which to quantify the effect of the newly introduced truncation.

6.1 Abstraction

6.1.1 Preliminaries

We consider the "exact" (superscript e) problem: for any $\mu \in \mathcal{D} \subset \mathbb{R}^P$, find $s^{\rm e}(\mu) = \ell(u^{\rm e}(\mu))$, where $u^{\rm e}(\mu)$ satisfies the weak form of the μ-parametrized PDE

$$a(u^{\rm e}(\mu), v; g(x;\mu)) = f(v; h(x;\mu)),\ \forall\, v \in X^{\rm e}\ . \tag{6.1}$$

Here μ and $\mathcal{D}$ are the input and input domain; $a(\cdot,\cdot;g(x;\mu))$ is a $X^{\rm e}$-continuous linear operator; and $f(\cdot;h(x;\mu))$ and $\ell(\cdot)$ are $X^{\rm e}$-continuous linear functionals. Note that a and f depend on $g(x;\mu)$ and $h(x;\mu)$; we further assume that these functions are continuous in the closed domain $\bar{\Omega}$ and sufficiently smooth with respect to all μ in $\mathcal{D}$. We shall suppose that a is of the form

$$a(w, v; g(x;\mu)) = a_0(w, v) + a_1(w, v, g(x;\mu)), \tag{6.2}$$

where $a_0(\cdot,\cdot)$ is a continuous (and, for simplicity, parameter-independent) bilinear form and $a_1(\cdot,\cdot,g(\cdot))$ is a trilinear form. For simplicity of exposition, we assume here that $h(x;\mu) = g(x;\mu)$.

We consider here second-order PDEs; and hence $(H_0^1(\Omega))^\nu \subset X^{\rm e} \subset (H^1(\Omega))^\nu$, where $\nu = 1$ for a scalar field variable and $\nu = d$ for a vector field variable. In actual practice, we replace $X^{\rm e}$ with $X \subset X^{\rm e}$, a "truth" approximation space of dimension $\mathcal{N}$. The inner product and norm associated with X are given by $(\cdot,\cdot)_X$ and $\|\cdot\|_X = (\cdot,\cdot)_X^{1/2}$, respectively.

We shall assume that a satisfies a coercivity and continuity condition

$$0 < \alpha_0 \le \alpha(\mu) \equiv \inf_{w\in X} \frac{a(w, w; g(x;\mu))}{\|w\|_X^2}, \quad \forall\, \mu \in \mathcal{D}, \tag{6.3}$$

$$\gamma(\mu) \equiv \sup_{w\in X} \frac{a(w, w; g(x;\mu))}{\|w\|_X^2} \le \gamma_0 < \infty, \quad \forall\, \mu \in \mathcal{D}; \tag{6.4}$$

here $\alpha(\mu)$ and $\gamma(\mu)$ are the coercivity constant and the standard continuity constant, respectively. (We (plausibly) suppose that $\alpha_0,\ \gamma_0$ may be chosen independent of $\mathcal{N}$.)

Finally, we assume that the trilinear form a_1 satisfies

$$a_1(w, v, z) \leq \gamma_1 \|w\|_X \, \|v\|_X \, \|z\|_{L^\infty(\Omega)}, \quad \forall \, w, v \in X. \tag{6.5}$$

It is then standard, given that $g(\cdot; \mu) \in L^\infty(\Omega)$, to prove existence and uniqueness.

6.1.2 General Problem Statement

Our approximation of the continuous problem in the finite approximation subspace X can then be stated as: given $\mu \in \mathcal{D} \in \mathbb{R}^P$, we evaluate

$$s(\mu) = \ell(u(\mu)) \ , \tag{6.6}$$

where $u(\mu) \in X$ is the solution of the discretized weak formulation

$$a(u(\mu), v; g(x; \mu)) = f(v; g(x; \mu)), \quad \forall v \in X \ . \tag{6.7}$$

We shall assume — hence the appellation "truth" — that X is sufficiently rich that u (respectively, s) is sufficiently close to $u^{\rm e}(\mu)$ (respectively, $s^{\rm e}(\mu)$) for all μ in the (closed) parameter domain $\mathcal{D}$. The reduced-basis approximation shall be built upon our reference finite element approximation, and the reduced-basis error will thus be evaluated with respect to $u(\mu) \in X$. Typically, $\mathcal{N}$, the dimension of X, will be very large; our formulation must be both stable and computationally efficient as $\mathcal{N} \to \infty$.

6.1.3 A Model Problem

We consider the following model problem: the input is $\mu = (\mu_{(1)}, \mu_{(2)}) \in \mathcal{D} \equiv [-1, -0.01]^2$; the spatial domain is the unit square $\Omega =]0, 1[^2 \in \mathbb{R}^2$; our piecewise-linear finite element approximation space $X = H^1_0 \equiv \{v \in H^1(\Omega) \mid v|_{\partial\Omega} = 0\}$ has dimension $\mathcal{N} = 2601$; the field variable $u(\mu)$ satisfies (6.7) with

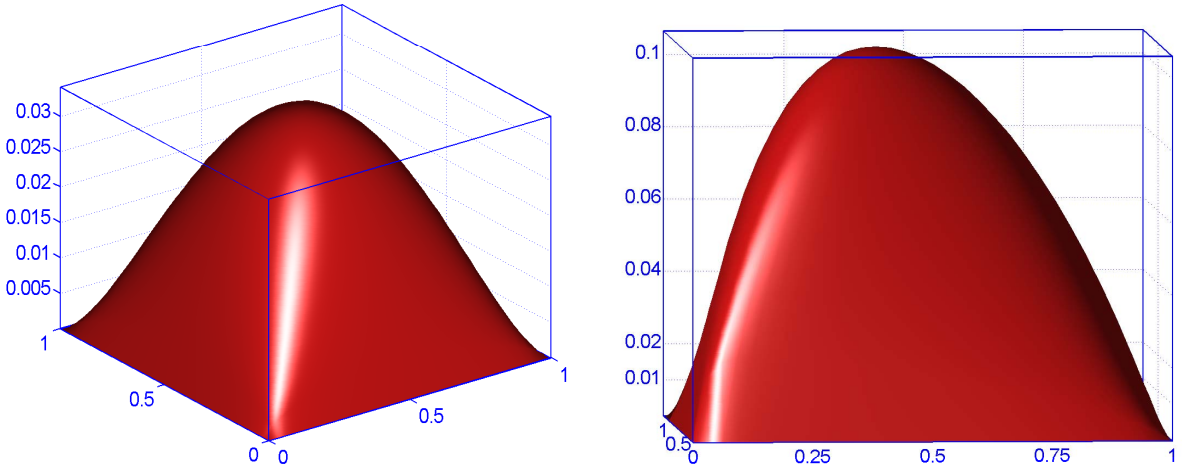
$$a_0(w, v) = \int_\Omega \nabla w \cdot \nabla v, \quad a_1(w, v, z) = \int_\Omega z \ w \ v, \quad f(v; z) = \int_\Omega z \ v, \tag{6.8}$$

$$g(x;\mu) = \frac{1}{\sqrt{(x_{(1)} - \mu_{(1)})^2 + (x_{(2)} - \mu_{(2)})^2}} \; ; \tag{6.9}$$

and the output $s(\mu)$ is evaluated from (6.6) with

$$\ell(v) = \int_\Omega v \; . \tag{6.10}$$

We give in Figure 6-1 the solutions corresponding to smallest parameter value and largest parameter value obtained with a piecewise-linear finite element approximation of $\mathcal{N} = 2601$. It should be noted that the solution develops a boundary layer in the vicinity of $x = (0, 0)$ for μ near the "corner" $(-0.01, -0.01)$. We further observe that the peak of the solution at the largest parameter value is much higher than that of the solution at the smallest parameter value.



(a) (b)

Figure 6-1: Numerical solutions at typical parameter points: (a) $\mu = (-1, -1)$ and (b) $\mu = (-0.01, -0.01)$.

6.2 Empirical Interpolation Method

6.2.1 Function Approximation Problem

We consider the problem of approximating a given μ-dependent function $g(\,\cdot\,;\mu) \in L^\infty(\Omega)$, $\forall \mu \in \mathcal{D}$, of sufficient regularity by a linear combination of known basis functions. To this end, we assume for now that we are given nested samples $S_M^g = \{\mu_1^g \in \mathcal{D}, \ldots, \mu_M^g \in \mathcal{D}\}$ and associated nested approximation spaces $W_M^g = \text{span}\,\{\xi_m \equiv g(x;\mu_m^g), 1 \le m \le M\}$, $1 \le M \le M_{\max}$. In essence, W_M^g comprises basis functions on the parametrically induced manifold $\mathcal{M}^g \equiv \{g(\,\cdot\,;\mu) \mid \mu \in \mathcal{D}\}$. Our approximation to $g(\cdot;\mu) \in \mathcal{M}^g$ is then $g_M(\cdot;\mu) \in W_M^g$. As for the reduced-basis approximation the critical observation is that: since the manifold $\mathcal{M}^g$ is low-dimensional and very smooth in μ, we may thus anticipate that $g_M(\cdot;\mu) \to g(\cdot;\mu)$ very rapidly, and that we may hence choose M very small. But as for the reduced-basis approximation, it gives rise to two questions: an immediate question is how to choose S_M^g so as to ensure *good* approximation properties for W_M^g; equally important question is how to obtain *good* approximation g_M *efficiently*. In the following, we shall address our choice of S_M^g and develop an efficient approximation approach for computing such $g_M(\cdot;\mu)$.

6.2.2 Coefficient-Function Approximation Procedure

To begin, we choose μ_1^g, and define $S_1^g = \{\mu_1^g\}$, $\xi_1 \equiv g(x;\mu_1^g)$, and $W_1^g = \text{span}\,\{\xi_1\}$; we assume that $\xi_1 \neq 0$. Then, for $M \ge 2$, we set $\mu_M^g = \arg\max_{\mu\in\Xi^g} \inf_{z\in W_{M-1}^g} \|g(x;\mu) - z\|_{L^\infty(\Omega)}$, where Ξ^g is a suitably fine parameter sample over $\mathcal{D}$. We then set $S_M^g = S_{M-1}^g \cup \mu_M^g$, $\xi_M = g(x;\mu_M^g)$, and $W_M^g = \text{span}\,\{\xi_m,\ 1 \le m \le M\}$. It should be noted that our coefficient-function approximation is required to be consistent with the truth approximation of the underlying PDE, the "vector" $g(x;\mu)$ is thus in fact the interpolant of the "function" $g(\cdot;\mu)$ on the finite element truth mesh; and hence, $\inf_{z\in W_{M-1}^g} \|g(x;\mu) - z\|_{L^\infty(\Omega)}$ is simply a *standard linear program*.

Before we proceed, we note that the evaluation of $\varepsilon_M^*(\mu)$, $1 \le M \le M_{\max}$, requires the solution of a linear program for *each* parameter sample in Ξ^g; the computational cost involved thus depends strongly on the size of Ξ^g as well as on $M_{\max}$. Fortunately, we

can avoid solving the costly linear program by simply replacing the $L^\infty(\Omega)$-norm in our best approximation by the $L^2(\Omega)$-norm — our next sample point is thus based on $\mu_M^g = \arg\max_{\mu\in\Xi^g} \inf_{z\in W_{M-1}^g} \|g(x;\mu)-z\|_{L^2(\Omega)}$ — which is relatively inexpensive to evaluate; the computational cost is $O(M\mathcal{N}) + O(M^3)$. Although the following analysis is not rigorous for this alternative (or "surrogate") construction of S_M^g, we in fact obtain very similar convergence results in practice [52]. Hence, the $L^2(\Omega)$-based construction is extremely useful for problems with *many* parameters and large dimensional truth approximation. However, we shall consider only the $L^\infty(\Omega)$-based construction, because (*i*) the following analysis remains valid with our choice and (*ii*) the linear program cost is affordable for all numerical examples in the thesis.

Lemma 6.2.1. *Suppose that $M_{\max}$ is chosen such that the dimension of $\mathcal{M}^g$ exceeds $M_{\max}$, then the space W_M^g is of dimension M.*

Proof. We first introduce the best approximation

$$g_M^*(\,\cdot\,;\mu) \equiv \arg\min_{z\in W_M^g} \|g(\,\cdot\,;\mu) - z\|_{L^\infty(\Omega)}\ , \tag{6.11}$$

and the associated error

$$\varepsilon_M^*(\mu) \equiv \|g(\,\cdot\,;\mu) - g_M^*(\,\cdot\,;\mu)\|_{L^\infty(\Omega)}\ . \tag{6.12}$$

It directly follows from our hypothesis on $M_{\max}$ that $\varepsilon_0 \equiv \varepsilon_{M_{\max}}^*(\mu_{M_{\max}+1}^g) > 0$; our "arg max" construction then implies $\varepsilon_{M-1}^*(\mu_M^g) \geq \varepsilon_0$, $2 \leq M \leq M_{\max}$, since $\varepsilon_{M-1}^*(\mu_M^g) \geq \varepsilon_{M-1}^*(\mu_{M+1}^g) \geq \varepsilon_M^*(\mu_{M+1}^g)$. We now prove lemma 6.2.1 by induction. Clearly, $\dim(W_1^g) = 1$. Assume $\dim(W_{M-1}^g) = M-1$; then if $\dim(W_M^g) \neq M$, we have $g(\,\cdot\,;\mu_M^g) \in W_{M-1}^g$ and thus $\varepsilon_{M-1}^*(\mu_M^g) = 0$; however, the latter contradicts $\varepsilon_{M-1}^*(\mu_M^g) \geq \varepsilon_0 > 0$. □

We now construct nested sets of interpolation points $T_M = \{t_1, \ldots, t_M\}$, $1 \leq M \leq M_{\max}$. We first set $t_1 = \arg\operatorname{ess\,sup}_{x\in\Omega} |\xi_1(x)|$, $q_1 = \xi_1(x)/\xi_1(t_1)$, $B_{11}^1 = 1$. Then for $M = 2, \ldots, M_{\max}$, we solve the linear system $\sum_{j=1}^{M-1} \sigma_j^{M-1} q_j(t_i) = \xi_M(t_i)$, $1 \leq i \leq M-1$, and set $r_M(x) = \xi_M(x) - \sum_{j=1}^{M-1} \sigma_j^{M-1} q_j(x)$, $t_M = \arg\operatorname{ess\,sup}_{x\in\Omega} |r_M(x)|$, $q_M(x) = r_M(x)/r_M(t_M)$, and $B_{ij}^M = q_j(t_i)$, $1 \leq i,j \leq M$. It remains to demonstrate

Lemma 6.2.2. *The construction of the interpolation points is well-defined, and the functions* $\{q_1, \ldots, q_M\}$ *form a basis for* W_M^g.

Proof. We shall proceed by induction. Clearly, we have $W_1^g = \text{span}\,\{q_1\}$. Next we assume $W_{M-1}^g = \text{span}\,\{q_1, \ldots, q_{M-1}\}$; if (i) $|r_M(t_M))| > 0$ and (ii) B^{M-1} is invertible, then our construction may proceed and we may form $W_M^g = \text{span}\,\{q_1, \ldots, q_M\}$. To prove (i), we observe that $|r_M(t_M)| \geq \varepsilon_{M-1}^*(\mu_M^g) \geq \varepsilon_0 > 0$ since $\varepsilon_{M-1}^*(\mu_M^g)$ is the error associated with the best approximation. To prove (ii), we just note by the construction procedure that

$$B_{i\,j}^M = q_j(t_i) = r_j(t_i)/r_j(t_j) = 0, \quad \text{for } i < j$$

since $r_j(t_i) = 0, 1 \leq i \leq j-1, 2 \leq j \leq M$; that

$$B_{i\,j}^M = q_j(t_i) = 1, \quad \text{for } i = j$$

since $q_i(t_i) = r_i(t_i)/r_i(t_i), 1 \leq i \leq M$; and that

$$\left|B_{i\,j}^M\right| = |q_j(t_i)| \leq 1, \quad \text{for } i > j$$

since $t_i = \arg\,\text{ess sup}_{x\in\Omega}\,|r_i(x)|, 1 \leq i \leq M$. Hence, B^{M-1} is lower triangular with unity diagonal. □

Lemma 6.2.3. *For any* M*-tuple* $(\alpha_i)_{i=1,\ldots,M}$ *of real numbers, there exists a unique element* $w \in W_M^g$ *such that* $\forall i, 1 \leq i \leq M, w(t_i) = \alpha_i$.

Proof. Since the functions $\{q_1, \ldots, q_M\}$ form a basis for W_M^g (Lemma 6.2.2), any member of W_M^g can be expressed as $w = \sum_{j=1}^M q_j(x)\,\kappa_j$. Recalling that B^M is invertible, we may now consider the particular function w corresponding to the choice of coefficients κ_j, $1 \leq j \leq M$, such that $\sum_{j=1}^M B_{ij}^M \kappa_j = \alpha_i$, $1 \leq i \leq M$; but since $B_{ij}^M = q_j(t_i)$, $w(t_i) = \sum_{j=1}^M q_j(t_i)\,\kappa_j = \sum_{j=1}^M B_{ij}^M\,\kappa_j = \alpha_i$, $1 \leq i \leq M$, which hence proves existence. To prove uniqueness, we need only consider two possible candidates and again invoke the invertibility of B^M. □

It remains to develop an *efficient* procedure for obtaining a *good* collateral reduced-basis expansion $g_M(\cdot;\mu)$. Based on the approximation space W_M^g and set of interpolation

points T_M, we can readily construct an approximation to $g(;\mu)$. Indeed, our coefficient function approximation is the interpolant of g over T_M as provided for Lemma 6.2.3:

$$g_M(x;\mu) = \sum_{m=1}^{M} \varphi_{M\,m}(\mu)\, q_m(x) \ , \tag{6.13}$$

where $\varphi_M(\mu) \in \mathbb{R}^M$ is the solution of

$$\sum_{j=1}^{M} B^M_{i\,j}\, \varphi_{M\,j}(\mu) = g(t_i;\mu),\ 1 \le i \le M \ ; \tag{6.14}$$

note that $g_M(t_i;\mu) = g(t_i;\mu), 1 \le i \le M$. We define the associated error as

$$\varepsilon_M(\mu) \equiv \|g(\,\cdot\,;\mu) - g_M(\,\cdot\,;\mu)\|_{L^\infty(\Omega)}. \tag{6.15}$$

6.3 Error Analyses for the Empirical Interpolation

6.3.1 *A Priori* Framework

To begin, we define a "Lebesgue constant" [123]

$$\Lambda_M = \sup_{x\in\Omega} \sum_{m=1}^{M} |V^M_m(x)| \ . \tag{6.16}$$

Here the $V^M_m(x) \in W^g_M$ are characteristic functions satisfying $V^M_m(t_n) = \delta_{mn}$, the existence and uniqueness of which is guaranteed by Lemma 6.2.3. It can be shown that

Lemma 6.3.1. *The characteristic functions V^M_m are a basis for W^g_M. And the two bases $q_m,\ 1 \le m \le M$, and $V^M_m,\ 1 \le m \le M$, are related by*

$$q_i(x) = \sum_{j=1}^{M} B^M_{j\,i}\, V^M_j(x), \quad 1 \le i \le M \ . \tag{6.17}$$

Proof. We first consider $x = t_n,\ 1 \le n \le M$, and note that $\sum_{j=1}^M B^M_{j\,i}\, V^M_j(t_n) = \sum_{j=1}^M B^M_{j\,i}\, \delta_{jn} = B^M_{n\,i} = q_i(t_n),\ 1 \le i \le M$; it thus follows from Lemma 6.2.3 that (6.17) holds. It further follows from Lemma 6.2.2 and from Lemma 6.2.3 that any $w \in$

W_M^g can be uniquely expressed as $w = \sum_{i=1}^M \kappa_i q_i(x) = \sum_{i=1}^M \kappa_i (\sum_{j=1}^M B_{j\,i}^M V_j^M(x)) = \sum_{j=1}^M (\sum_{i=1}^M \kappa_i B_{j\,i}^M) V_j^M(x) = \sum_{j=1}^M \alpha_j V_j^M(x)$, where $\alpha_j = w(t_j)$, $1 \le j \le M$; thus the V_j^M, $1 \le j \le M$, form a ("nodal") basis for W_M^g. □

We further observe that Λ_M depends on W_M^g and T_M, but not on μ nor on our choice of basis for W_M^g. We can further prove

Lemma 6.3.2. *The interpolation error $\varepsilon_M(\mu)$ satisfies $\varepsilon_M(\mu) \le \varepsilon_M^*(\mu)(1+\Lambda_M)$, $\forall \mu \in \mathcal{D}$.*

Proof. We first define the error function for $g_M^*(x;\mu)$ as

$$\begin{aligned} e_M^*(x;\mu) &\equiv g(x;\mu) - g_M^*(x;\mu) \\ &= (g(x;\mu) - g_M(x;\mu)) + (g_M(x;\mu) - g_M^*(x;\mu)) \ . \end{aligned} \tag{6.18}$$

Since $g_M(x;\mu) \in W_M^g, g_M^*(x;\mu) \in W_M^g$ there exists $\kappa(\mu) \in \mathbb{R}^M$ such that

$$g_M(x;\mu) - g_M^*(x;\mu) = \sum_{m=1}^M \kappa_m(\mu)\, q_m(x) \ . \tag{6.19}$$

It then follows from (6.18) and (6.19) that

$$\begin{aligned} e_M^*(t_i;\mu) &= (g(t_i;\mu) - g_M(t_i;\mu)) + \sum_{m=1}^M \kappa_m(\mu)\, q_m(t_i) \\ &= \sum_{m=1}^M B_{i\,m}^M \kappa_m(\mu), \quad 1 \le i \le M \ ; \end{aligned} \tag{6.20}$$

here we invoke (6.13) and (6.14) to arrive at the second equality. The desired result

immediately follows

$$\begin{aligned}
\varepsilon_M(\mu) - \varepsilon_M^*(\mu) &= \|g(\,\cdot\,;\mu) - g_M(\,\cdot\,;\mu)\|_{L^\infty(\Omega)} - \|g(\,\cdot\,;\mu) - g_M^*(\,\cdot\,;\mu)\|_{L^\infty(\Omega)} \\
&\le \|g_M(\,\cdot\,;\mu) - g_M^*(\,\cdot\,;\mu)\|_{L^\infty(\Omega)} \\
&= \|\sum_{m=1}^{M} \kappa_m(\mu)\, q_m(x)\|_{L^\infty(\Omega)} \\
&= \|\sum_{k=1}^{M}\sum_{m=1}^{M} B_{k\,m}^M\, \kappa_m(\mu)\, V_k^M(x)\|_{L^\infty(\Omega)} \\
&= \|\sum_{i=1}^{M} e_M^*(t_i;\mu)\, V_i^M(x)\|_{L^\infty(\Omega)} \\
&\le \varepsilon_M^*(\mu)\, \Lambda_M
\end{aligned}$$

from triangle inequality, (6.19), (6.17), (6.20), and $|e_M^*(t_i;\mu)| \le \varepsilon_M^*(\mu)$, $1 \le i \le M$. □

We can further show

Proposition 5. *The Lebesgue constant Λ_M satisfies $\Lambda_M \le 2^M - 1$.*

Proof. We first recall two crucial properties of the matrix B^M: (i) B^M is lower triangular with unity diagonal — $q_m(t_m) = 1$, $1 \le m \le M$, and (ii) all entries of B^M are of modulus no greater than unity — $\|q_m\|_{L^\infty(\Omega)} \le 1$, $1 \le m \le M$. Hence, from (6.17) we can write

$$\begin{aligned}
|V_m^M(x)| &= \left| q_m(x) - \sum_{i=m+1}^{M} B_{i\,m}^M V_i^M(x) \right| \\
&\le |q_m(x)| + \sum_{i=m+1}^{M} |V_i^M(x)| \\
&\le 1 + \sum_{i=m+1}^{M} |V_i^M(x)| \qquad (6.21)
\end{aligned}$$

for $m = 1, \ldots, M-1$. It follows that, starting from $|V_M^M(x)| = |q_M(x)| \le 1$, we can deduce $|V_{M+1-m}^M(x)| \le 1 + |V_M^M(x)| + \ldots + |V_{M+2-m}^M(x)| \le 2^{m-1}$, $2 \le m \le M$, and thus have $\sum_{m=1}^{M} |V_m^M(x)| \le 2^M - 1$. □

Proposition 5 is very pessimistic and of little practical value (though $\varepsilon_M^*(\mu)$ does often converge sufficiently rapidly that $\varepsilon_M^*(\mu)\, 2^M \to 0$ as $M \to \infty$); this is not surprising given

analogous results in the theory of polynomial interpolation [123]. However, Proposition 5 does provide some notion of stability.

6.3.2 *A Posteriori* Estimators

Given a coefficient function approximation $g_M(x;\mu)$ for $M \leq M_{\max} - 1$, we define

$$\mathcal{E}_M(x;\mu) \equiv \hat{\varepsilon}_M(\mu)\, q_{M+1}(x)\ , \tag{6.22}$$

where

$$\hat{\varepsilon}_M(\mu) \equiv |g(t_{M+1};\mu) - g_M(t_{M+1};\mu)|\ . \tag{6.23}$$

In general, $\varepsilon_M(\mu) \geq \hat{\varepsilon}_M(\mu)$, since $\varepsilon_M(\mu) = ||g(\cdot;\mu) - g_M(\cdot;\mu)||_{L^\infty(\Omega)} \geq |g(x;\mu) - g_M(x;\mu)|$ for all $x \in \Omega$, and thus also for $x = t_{M+1}$. However, we can prove

Proposition 6. *If $g(\,\cdot\,;\mu) \in W^g_{M+1}$, then (i) $g(x;\mu) - g_M(x;\mu) = \pm\mathcal{E}_M(x;\mu)$ (either $\mathcal{E}_M(x;\mu)$ or $-\mathcal{E}_M(x;\mu)$), and (ii) $\|g(\,\cdot\,;\mu) - g_M(\,\cdot\,;\mu)\|_{L^\infty(\Omega)} = \hat{\varepsilon}_M(\mu)$.*

Proof. By our assumption $g(\,\cdot\,;\mu) \in W^g_{M+1}$, there exists $\kappa(\mu) \in \mathbb{R}^{M+1}$ such that $g(x;\mu) - g_M(x;\mu) = \sum_{m=1}^{M+1} \kappa_m(\mu)\, q_m(x)$. We now consider $x = t_i, 1 \leq i \leq M+1$, and arrive at

$$\sum_{m=1}^{M+1} \kappa_m(\mu)\, q_m(t_i) = g(t_i;\mu) - g_M(t_i;\mu), \quad 1 \leq i \leq M+1\ . \tag{6.24}$$

We next note from (6.13) and (6.14) that

$$g(t_i;\mu) - g_M(t_i;\mu) = 0, \quad 1 \leq i \leq M\ . \tag{6.25}$$

Therefore, $\kappa_m(\mu) = 0$, $1 \leq m \leq M$, since the matrix $q_m(t_i)$ is lower triangular, and $\kappa_{M+1}(\mu) = g(t_{M+1};\mu) - g_M(t_{M+1};\mu)$, since $q_{M+1}(t_{M+1}) = 1$; this concludes the proof of (i). The proof of (ii) then directly follows from $\|q_{M+1}\|_{L^\infty(\Omega)} = 1$. □

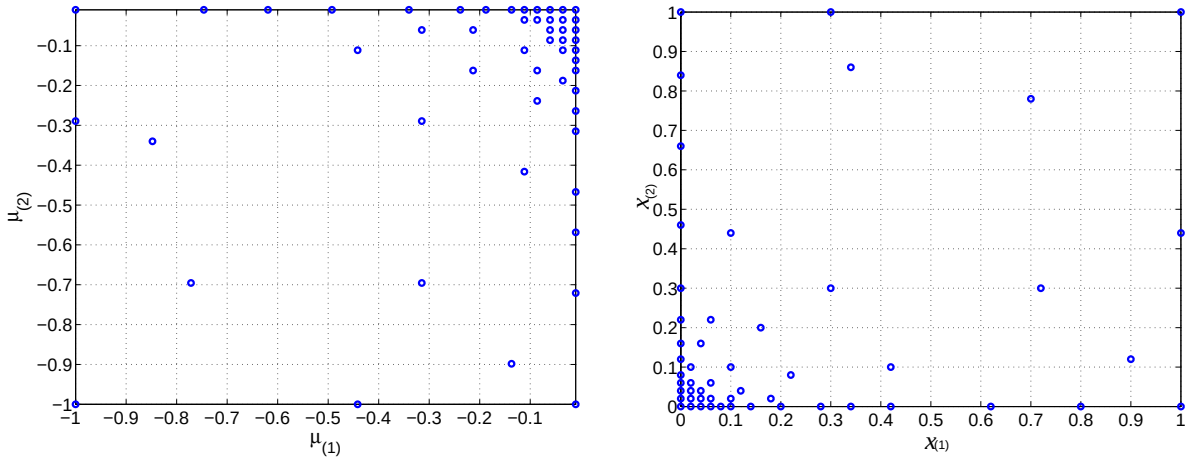
Of course, in general $g(\,\cdot\,;\mu) \notin W^g_{M+1}$, and hence our estimator $\hat{\varepsilon}_M(\mu)$ is not quite a rigorous upper bound; however, if $\varepsilon_M(\mu) \to 0$ very fast, we expect that the effectivity

$$\eta_M(\mu) \equiv \frac{\hat{\varepsilon}_M(\mu)}{\varepsilon_M(\mu)}\,, \tag{6.26}$$

shall be close to unity. Furthermore, the estimator is very inexpensive – *one additional evaluation* of $g(\,\cdot\,;\mu)$ at a single point in Ω.

6.3.3 Numerical Results

We consider the nonaffine function $G(x;\mu) \equiv \left((x_{(1)}-\mu_{(1)})^2 + (x_{(2)}-\mu_{(2)})^2\right)^{-1/2}$ for $x \in \Omega \equiv]0,1[^2$ and $\mu \in \mathcal{D} \equiv [-1,-0.01]^2$. We choose for Ξ^g a deterministic grid of 40×40 parameter points over $\mathcal{D}$ and we take $\mu_1^g = (-0.01,-0.01)$. Next we then pursue the empirical interpolation procedure described in Section 6.2 to construct S_M^g, W_M^g, T_M, and B^M, $1 \le M \le M_{\max}$, for $M_{\max} = 52$. We present in Figure 6-2 $S_{M_{\max}}^g$ and $T_{M_{\max}}$. It is not surprising from the given form of $G(x;\mu)$ that the sample points are distributed mostly around the corner $(-0.01,-0.01)$ of the parameter domain; and that the interpolation points are allocated mainly around the corner $(0.00, 0.00)$ of the physical domain.



(a) (b)

Figure 6-2: (a) Parameter sample set S_M^g, $M_{\max} = 52$, and (b) Interpolation points t_m, $1 \le m \le M_{\max}$, for the nonaffine function (6.9).

We now introduce a regular parameter test sample Ξ_{Test}^g of size $Q_{\text{Test}} = 225$, and define $\varepsilon_{M,\max}^* = \max_{\mu\in\Xi_{\text{Test}}^g} \varepsilon_M^*(\mu)$, $\overline{\rho}_M = Q_{\text{Test}}^{-1} \sum_{\mu\in\Xi_{\text{Test}}^g} \left(\varepsilon_M(\mu)/(\varepsilon_M^*(\mu)(1+\Lambda_M))\right)$, $\overline{\eta}_M = Q_{\text{Test}}^{-1} \sum_{\mu\in\Xi_{\text{Test}}^g} \eta_M(\mu)$, $\varkappa_M$ as condition number of B^M; here $\eta_M(\mu)$ is the effectivity defined in (6.26). We present in Table 6.1 these quantities as a function of M ($M_{\max} = 52$). We observe that $\varepsilon_{M,\max}^*$ converges rapidly with M; that the Lebesgue constant provides a reasonably sharp measure of the interpolation-induced error; that the Lebesgue constant grows very slowly — $\varepsilon_M(\mu)$ is *only slightly larger that the min max result* $\varepsilon_M^*(\mu)$; that

the error estimator effectivity is reasonably close to unity; and that B^M is quite well-conditioned for our choice of basis (For the non-orthogonalized basis $\xi_m,\ 1 \le m \le M$, the condition number of B^M will grow exponentially with M.) These results are expected since the given function $G(x;\mu)$ is quite regular and smooth in the parameter μ.

M	$\varepsilon^*_{M,\max}$	$\overline{\rho}_M$	Λ_M	$\overline{\eta}_M$	$\varkappa_M$
8	8.30 E−02	0.68	1.76	0.17	3.65
16	4.22 E−03	0.67	2.63	0.10	6.08
24	2.68 E−04	0.49	4.42	0.28	9.19
32	5.64 E−05	0.48	5.15	0.20	12.86
40	3.66 E−06	0.54	4.98	0.60	18.37
48	6.08 E−07	0.37	7.43	0.29	20.41

Table 6.1: $\varepsilon^*_{M,\max}$, $\overline{\rho}_M$, Λ_M, $\overline{\eta}_M$, and $\varkappa_M$ as a function of M.

6.4 Reduced-Basis Approximation

6.4.1 Discrete Equations

We begin with motivating the need for the empirical interpolation approach in dealing with nonaffine problems. Specifcally, we introduce nested samples, $S_N = \{\mu^u_1 \in \mathcal{D}, \cdots, \mu^u_N \in \mathcal{D}\}, 1 \le N \le N_{\max}$ and associated nested Lagrangian reduced-basis spaces as $W_N = \text{span}\{\zeta_j \equiv u(\mu^u_j), 1 \le j \le N\}, 1 \le N \le N_{\max}$, where $u(\mu^u_j)$ is the solution to (6.7) for $\mu = \mu^u_j$. Were we to follow the classical recipe, our reduced-basis approximation would then be: for a given $\mu \in \mathcal{D}$, we evaluate $s_N(\mu) = \ell(u_N(\mu))$, where $u_N(\mu) \in W_N$ is the solution of

$$a_0(u_N(\mu), v) + a_1(u_N(\mu), v, g(x;\mu)) = f(v; g(x;\mu)), \quad \forall v \in W_N\ . \tag{6.27}$$

If we now express $u_N(\mu) = \sum_{j=1}^N u_{N\,j}(\mu)\zeta_j$ and choose a test function $v = \zeta_n, 1 \le n \le N,$, we obtain the $N \times N$ linear algebraic system

$$\sum_{j=1}^{N} \left(a_0(\zeta_i, \zeta_j) + a_1(\zeta_i, \zeta_j, g(x;\mu))\right) u_{N\,j}(\mu) = f(\zeta_i; g(x;\mu)), \quad 1 \le i \le N. \tag{6.28}$$

We observe that while $a_0(\zeta_i, \zeta_j)$ is parameter-independent and can thus be pre-computed offline, $f(\zeta_i; g(x;\mu))$ and $a_1(\zeta_i, \zeta_j, g(x;\mu))$ depend on $g(x;\mu)$ and must thus be evaluated online for every new parameter value μ. The operation count for the online stage will thus scale as $O(N^2\mathcal{N})$, where $\mathcal{N}$ is the dimension of the underlying truth finite element approximation space: the reduction in marginal cost gain obtained in moving from the truth finite element approximation space to the reduced-basis space will be quite modest regardless of the dimension reduction.

To recover online $\mathcal{N}$ independence, we replace $g(x;\mu)$ by $g_M(x;\mu) = \sum_{m=1}^{M} \varphi_{M\,m}(\mu) q_m$ which is a coefficient-function approximation defined in Section 6.2 and analyzed in Section 6.3. We thus construct nested samples $S_M^g = \{\mu_1^g \in \mathcal{D}, \cdots, \mu_M^g \in \mathcal{D}\}, 1 \le M \le M_{\max}^g$, associated nested approximation spaces $W_M^g = \text{span}\{\xi_m \equiv g(\mu_m^g), 1 \le m \le M\}, 1 \le M \le M_{\max}$, and nested sets of interpolation points $T_M = \{t_1, \ldots, t_M\}, 1 \le M \le M_{\max}$ following the procedure of Section 6.2. Our reduced-basis approximation is now: Given $\mu \in \mathcal{D}$, we evaluate $s_{N,M}(\mu) = \ell(u_{N,M}(\mu))$, where $u_{N,M}(\mu) \in W_N$ is the solution of

$$a_0(u_{N,M}(\mu), v) + a_1(u_{N,M}(\mu), v, g_M(x;\mu)) = f(v; g_M(x;\mu)), \quad \forall v \in W_N \ . \tag{6.29}$$

It thus follow from $u_{N,M}(\mu) = \sum_{j=1}^{N} u_{N,M\,j}(\mu)\zeta_j$ and trilinearity of a_1 that the $u_{N,M\,j}, 1 \le j \le N$, satisfies the $N \times N$ linear algebraic system

$$\sum_{j=1}^{N} \left(a_0(\zeta_j, \zeta_i) + \sum_{m=1}^{M} \varphi_{M\,m}(\mu) a_1(\zeta_j, \zeta_i, q_m) \right) u_{N,M\,j}(\mu) = \sum_{m=1}^{M} \varphi_{M\,m}(\mu) f(\zeta_i; q_m), \tag{6.30}$$

for $i = 1, \cdots, N$; here $\varphi_{M\,m}(\mu)$, $1 \le m \le M$, is determined from (6.14). We recover the online $\mathcal{N}$-independence: the quantities $a_0(\zeta_i, \zeta_j)$, $a_1(\zeta_i, \zeta_j, q_m)$, and $f(\zeta_i; q_m)$ are all *parameter independent* and can thus be pre-computed offline as discribed in Section 6.4.3.

6.4.2 A *Priori* Theory

We consider here the convergence rate of $u_{N,M}(\mu) \to u(\mu)$. In fact, it is a simple matter to demonstrate the optimality of $u_{N,M}(\mu)$ in

Proposition 7. *For $\varepsilon_M(\mu)$ of (6.15) satisfying $\varepsilon_M(\mu) \le \frac{1}{2}\frac{\alpha(\mu)}{\phi_2(\mu)}$, we have*

$$\begin{aligned} \|u(\mu) - u_{N,M}(\mu)\|_X &\le \left(1 + \frac{\gamma(\mu)}{\alpha(\mu)}\right) \inf_{w_N \in W_N} \|u(\mu) - w_N\|_X \\ &\quad + \varepsilon_M(\mu) \left(\frac{\phi_1(\mu)\alpha(\mu) + 2\phi_2(\mu)\phi_3(\mu)}{\alpha^2(\mu)}\right) ; \end{aligned} \tag{6.31}$$

here $\phi_1(\mu)$, $\phi_2(\mu)$, and $\phi_3(\mu)$ are given by

$$\phi_1(\mu) = \frac{1}{\varepsilon_M(\mu)} \sup_{v \in X} \frac{f(v; g(\cdot;\mu) - g_M(\cdot;\mu))}{|v\|_X} , \tag{6.32}$$

$$\phi_2(\mu) = \frac{1}{\varepsilon_M(\mu)} \sup_{w \in X} \sup_{v \in X} \frac{a(w, v; g(\cdot;\mu) - g_M(\cdot;\mu))}{\|w\|_X \|v\|_X} , \tag{6.33}$$

$$\phi_3(\mu) = \sup_{v \in X} \frac{f(v; g_M(\cdot;\mu))}{|v\|_X} . \tag{6.34}$$

Proof. For any $w_N = u_{N,M}(\mu) + v_N \in W_N$, we have

$$\begin{aligned} \alpha(\mu)\|w_N - u_{N,M}\|_X^2 &\le a(w_N - u_{N,M}, w_N - u_{N,M}; g(\cdot;\mu)) \\ &= a(w_N - u, v_N; g(\cdot;\mu)) + a(u - u_{N,M}, v_N; g(\cdot;\mu)) \\ &\le \gamma(\mu)\|w_N - u\|_X \|v_N\|_X + a(u - u_{N,M}, v_N; g(\cdot;\mu)) . \end{aligned} \tag{6.35}$$

It follows from (6.1), (6.29), and (6.32)-(6.34) that the second term can be bounded by

$$\begin{aligned} a(u - u_{N,M}, v_N; g(\cdot;\mu)) &= f(v_N; g(\cdot;\mu)) - a(u_{N,M}, v_N; g(\cdot;\mu)) \\ &= f(v_N; g(\cdot;\mu) - g_M(\cdot;\mu)) - a(u_{N,M}, v_N; g(\cdot;\mu) - g_M(\cdot;\mu)) \\ &\le \varepsilon_M(\mu)\phi_1(\mu)\|v_N\|_X + \varepsilon_M(\mu)\phi_2(\mu)\|v_N\|_X \|u_{N,M}\|_X \\ &\le \varepsilon_M(\mu) \left(\frac{\phi_1(\mu)\alpha(\mu) + 2\phi_2(\mu)\phi_3(\mu)}{\alpha(\mu)}\right) \|v_N\|_X, \end{aligned} \tag{6.36}$$

where the last inequality derives from

$$\begin{aligned}\alpha(\mu)\|u_{N,M}(\mu)\|_X^2 &\leq a(u_{N,M}(\mu), u_{N,M}(\mu); g(x;\mu)) \\ &= f(u_{N,M}(\mu); g_M(x;\mu)) + a(u_{N,M}(\mu), u_{N,M}(\mu); g(x;\mu) - g_M(x;\mu)) \\ &\leq \phi_3(\mu)\|u_{N,M}(\mu)\|_X + \varepsilon_M(\mu)\phi_2(\mu)\|u_{N,M}(\mu)\|_X^2, \end{aligned} \tag{6.37}$$

and our hypothesis on $\varepsilon_M(\mu)$. It then follows from (6.35) and (6.36) that $\forall\, w_N \in W_N$,

$$\|w_N - u_{N,M}(\mu)\|_X \leq \frac{\gamma(\mu)}{\alpha(\mu)}\|w_N - u(\mu)\|_X + \varepsilon_M(\mu)\left(\frac{\phi_1(\mu)\alpha(\mu) + 2\phi_2(\mu)\phi_3(\mu)}{\alpha^2(\mu)}\right) . \tag{6.38}$$

The result follows from (6.38) and the triangle inequality. (Note for a affine $\phi_1(\mu) = \phi_2(\mu) = 0, \forall \mu \in \mathcal{D}$, we recover the optimality result for affine linear problems [121].) □

As regards the best approximation, we note that W_N comprises "snapshots" on the parametrically induced manifold $\mathcal{M}^u \equiv \{u(\mu) \,|\, \forall\, \mu \in \mathcal{D}\} \subset X$. The critical observation is that $\mathcal{M}^u$ is very *low-dimensional*; and that $\mathcal{M}^u$ is *smooth* under our hypotheses on stability and continuity. We thus expect that the best approximation will converge to $u(\mu)$ very rapidly, and hence that N may be chosen small. (This is proven for a particularly simple case in [93].)

6.4.3 Offline/Online Computational Procedure

The theoretical and numerical results of Sections 6.3 and 6.3.3 suggest that M may also be chosen small. We now develop offline-online computational procedures that exploit this dimension reduction provided by the reduced-basis method [9, 65, 85, 121] and our empirical interpolation method.

In the offline stage — performed only *once* — we first construct nested approximation spaces $W_M^g = \{q_1, \ldots, q_M\}$ and nested sets of interpolation points T_M, $1 \leq M \leq M_{\max}$; we then solve for the ζ_n, $1 \leq n \leq N$; we finally form and store B^M, $a_0(\zeta_j, \zeta_i)$, $a_1(\zeta_j, \zeta_i, q_m)$, and $f(\zeta_i; q_m)$, $\ell(\zeta_i)$, $1 \leq i, j \leq N$, $1 \leq m \leq M_{\max} - 1$. Note that all quantities computed in the offline stage are independent of the parameter μ.

In the online stage — performed many times for each new μ — we first compute $\varphi_M(\mu)$ from (6.14) at cost $O(M^2)$ by multiplying the pre-computed inverse matrix $(B^M)^{-1}$ with

the vector $g(t_i;\mu), 1 \le i \le M$; we then assemble and invert the (full) $N \times N$ reduced-basis stiffness matrix $a_0(\zeta_j,\zeta_i) + \sum_{m=1}^{M} \varphi_{M\,m}(\mu) a_1(\zeta_j,\zeta_i,q_m)$ to obtain $u_{N,M\,j}$, $1 \le j \le N$, at cost $O(N^2M) + O(N^3)$; we finally evaluate the reduced-basis output $s_{N,M}(\mu)$ at cost $O(N)$. The operation count for the online stage is thus only $O(M^2 + N^2M + N^3)$; the online complexity is *independent* of $\mathcal{N}$, the dimension of the underlying "truth" finite element approximation space. Since N, $M \ll \mathcal{N}$ we expect significant computational savings in the online stage relative to classical discretization and solution approaches (and relative to standard reduced-basis approaches built upon (6.28)).

6.5 A Posteriori Error Estimation

6.5.1 Error Bounds

We assume that we may calculate $\hat{\alpha}(\mu)$ such that $\alpha(\mu) \ge \hat{\alpha}(\mu), \forall \mu \in \mathcal{D}$ as discussed in Chapter 4. We then define an error bound $\Delta_{N,M}(\mu)$ for $\|u(\mu) - u_{N,M}(\mu)\|_X$ as

$$\begin{aligned}\Delta_{N,M}(\mu) \;=\; & \frac{1}{\hat{\alpha}(\mu)} \sup_{v\in X} \frac{r(v;g_M(x;\mu))}{\|v\|_X} \\ & + \frac{\hat{\varepsilon}_M}{\hat{\alpha}(\mu)} \sup_{v\in X} \frac{f(v;q_{M+1}(x)) - a_1(u_{N,M}(\mu),v,q_{M+1}(x))}{\|v\|_X}\;,\end{aligned} \tag{6.39}$$

and an output error bound $\Delta^s_{N,M}(\mu)$ for $|s(\mu) - s_{N,M}(\mu)|$ as

$$\Delta^s_{N,M}(\mu) = \sup_{v\in X} \frac{\ell(v)}{\|v\|_X} \Delta_{N,M}(\mu)\;. \tag{6.40}$$

Here $r(v;g_M(x;\mu))$ is the residual associated with $u_{N,M}(\mu)$ and $g_M(x;\mu)$

$$r(v;g_M(x;\mu)) = f(v;g_M(x;\mu)) - a(u_{N,M}(\mu),v;g_M(x;\mu)), \quad \forall\, v \in X\;. \tag{6.41}$$

For our purposes here, we shall focus on the energy error bound, $\Delta_{N,M}(\mu)$, rather than $\Delta^s_{N,M}(\mu)$; the latter may be significantly improved by the introduction of adjoint techniques described in Section 6.6. We can readily prove

Proposition 8. *Suppose that $g(x;\mu) \in W^g_{M+1}$, then for the error bounds $\Delta_{N,M}(\mu)$ of (6.39) and $\Delta^s_{N,M}(\mu)$ of (6.40), the corresponding effectivities satisfy $1 \le \eta_{N,M}(\mu), \forall \mu \in \mathcal{D}$*

and $1 \leq \eta^s_{N,M}(\mu), \forall \mu \in \mathcal{D}$.

Proof. We first note from (6.7) and (6.41) that $e(\mu) \equiv u(\mu) - u_{N,M}(\mu)$ satisfies

$$a(e(\mu), v; g(\cdot;\mu)) = r(v; g_M(x;\mu)) + f(v; g(\cdot;\mu) - g_M(\cdot;\mu)) \\ - a_1(u_{N,M}(\mu), v, g(\cdot;\mu) - g_M(\cdot;\mu)), \quad \forall\, v \in X\ . \quad (6.42)$$

The first result immediately follows from

$$\begin{aligned}
\|e(\mu)\|_X &\leq \frac{1}{\alpha(\mu)} \frac{a(e(\mu), e(\mu); g(\cdot;\mu))}{\|e(\mu)\|_X} \\
&\leq \frac{1}{\hat{\alpha}(\mu)} \left\{ \frac{r(e(\mu); g_M(\cdot;\mu)) + f(e(\mu); g(\cdot;\mu) - g_M(\cdot;\mu))}{\|e(\mu)\|_X} \right. \\
&\qquad \left. - \frac{a_1(u_{N,M}(\mu), e(\mu), g(\cdot;\mu) - g_M(\cdot;\mu))}{\|e(\mu)\|_X} \right\} \\
&\leq \frac{1}{\hat{\alpha}(\mu)} \sup_{v \in X} \left\{ \frac{r(v; g_M(\cdot;\mu)) + f(v; g(\cdot;\mu) - g_M(\cdot;\mu))}{\|e(\mu)\|_X} \right. \\
&\qquad \left. - \frac{a_1(u_{N,M}(\mu), v, g(\cdot;\mu) - g_M(\cdot;\mu))}{\|v\|_X} \right\} \\
&\leq \frac{1}{\hat{\alpha}(\mu)} \left(\sup_{v \in X} \frac{r(v; g_M(\cdot;\mu))}{\|v\|_X} + \hat{\varepsilon}_M \sup_{v \in X} \frac{f(v; q_{M+1}) - a_1(u_{N,M}(\mu), v, q_{M+1})}{\|v\|_X} \right)
\end{aligned}$$

where we have used a-coercivity in the first step, (6.42) in the second step, and our assumption $g(\cdot;\mu) \in W^g_{M+1}$ and Proposition 6 in the last step. Furthermore, we have

$$|s(\mu) - s_{N,M}(\mu)| = |\ell(e_{N,M}(\mu))| \leq \sup_{v \in X} \frac{\ell(v)}{\|v\|_X} \|e_{N,M}(\mu)\|_X \leq \Delta^s_{N,M}(\mu)\ .$$

This concludes the proof. □

In general, $g(x;\mu) \in W^g_{M+1}$ is not satisfied and our error bounds may thus not be completely rigorous due to the second term in (6.39), since $\hat{\varepsilon}_M(\mu)$ is indeed a *lower bound* surrogate for $\varepsilon_M(\mu)$. Therefore, for rigor of the energy error bound (and thus the output bound) M should be chosen sufficiently large such that the "Safety Condition"

$$\frac{\Delta_{N,M,\mathrm{n}}(\mu)}{\Delta_{N,M}(\mu)} \leq 1/2 \quad (6.43)$$

satisfies, where

$$\Delta_{N,M,\mathrm{n}}(\mu) = \frac{\hat{\varepsilon}_M(\mu)}{\hat{\alpha}(\mu)} \sup_{v\in X}[\{f(v;q_{M+1}) - a_1(u_{N,M}(\mu),v,q_{M+1})\}/\|v\|_X] \ . \tag{6.44}$$

This implies that ε_M should be roughly $O(\|r(v;g_M(x;\mu))\|_{X'})$ since $\sup_{v\in X}[\{f(v;q_{M+1}) - a_1(u_{N,M}(\mu),v,q_{M+1})\}/\|v\|_X]$ are actually $O(1)$ due to $\|q_{M+1}\|_{L^\infty(\Omega)} = 1$; note also that rather than $1/2$, more conservative choices will be even safer. If M is chosen too small the nonrigorous component $\Delta_{N,M,\mathrm{n}}(\mu)$ will dominate, we may thus risk to obtain nonrigorous bounds. Of course, M should not also be chosen too large, since the online complexity scale as $O(M^2N^2 + N^3)$ as discussed follows.

6.5.2 Offline/Online Computational Procedure

It remains to develop the offline-online computational procedure for the efficient calculation of $\Delta_{N,M}(\mu)$ and $\Delta^s_{N,M}(\mu)$. To begin, we invoke duality arguments to obtain

$$\sup_{v\in X} \frac{r(v;g_M(x;\mu))}{\|v\|_X} = \|\hat{e}_{N,M}(\mu)\|_X \ , \tag{6.45}$$

where $\hat{e}_{N,M}(\mu)$ is the solution of

$$(\hat{e}_{N,M}(\mu),v)_X = r(v;g_M(x;\mu)), \quad \forall v \in X \ . \tag{6.46}$$

We next note from our reduced-basis approximation $u_{N,M\,n}(\mu) = \sum_{n=1}^N u_{N,M\,n}(\mu)\,\zeta_n$ and coefficient-function approximation $g_M(;\mu) = \sum_{m=1}^M \varphi_{M\,m}(\mu)\,q_m$ to expand

$$\begin{aligned} r(v;g_M(x;\mu)) \ &= \ \sum_{m=1}^M \varphi_{M\,m}(\mu)\,f(v;q_m) - \sum_{n=1}^N u_{N,M\,n}(\mu) a_0(\zeta_n,v) \\ &\quad - \sum_{m=1}^M \sum_{n=1}^N \varphi_{M\,m}(\mu) u_{N,M\,n}(\mu)\,a_1(\zeta_n,v,q_m), \quad \forall\, v \in X. \end{aligned} \tag{6.47}$$

It follows from (6.46)-(6.47) and linear superposition that we may write $\hat{e}_{N,M}(\mu) \in X$ as

$$\hat{e}_{N,M}(\mu) = \sum_{k=1}^{M+N} \sigma_k(\mu)\mathcal{C}_k + \sum_{m=1}^M \sum_{n=1}^N \varphi_{M\,m}(\mu)\,u_{N,M\,n}(\mu)\,\mathcal{L}_{mn} \ , \tag{6.48}$$

where $\sigma_k(\mu) = \varphi_{M\,k}(\mu), (\mathcal{C}_k, v)_X = f(v; q_k), \forall v \in X, 1 \le k \le M$ and $\sigma_{k+M}(\mu) = u_{N,M\,k}(\mu)$, $(\mathcal{C}_{k+M}, v)_X = -a_0(\zeta_k; v), \forall v \in X, 1 \le k \le N$, and $(\mathcal{L}_{mn}, v)_X = -a_1(\zeta_n, v, q_m)$, $\forall\, v \in X$, $1 \le n \le N$, $1 \le m \le M$; note that the latter are simple parameter-independent (scalar or vector) Poisson, or Poisson-like, problems. It thus follows that

$$\begin{aligned}\|(\hat{e}_{N,M}(\mu)\|_X^2 \;=\; & \sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{m'=1}^{M}\sum_{n'=1}^{N} \varphi_{M\,m}(\mu)\, u_{N,M\,n}(\mu)\varphi_{M\,m'}(\mu)\, u_{N,M\,n'}(\mu)\, (\mathcal{L}_{mn}, \mathcal{L}_{m'n'})_X \\ & + 2\sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{k=1}^{M+N} \varphi_{M\,m}(\mu)\, u_{N,M\,n}(\mu)\sigma_k(\mu)(\mathcal{C}_k, \mathcal{L}_{mn})_X \\ & + \sum_{k=1}^{M+N}\sum_{k'=1}^{M+N} \sigma_k(\mu)\sigma_{k'}(\mathcal{C}_k, \mathcal{C}_{k'})_X \;. \end{aligned} \tag{6.49}$$

Similarly, we have

$$\begin{aligned}\sup_{v\in X} \frac{f(v; q_{M+1}(x)) - a_1(u_{N,M}(\mu), v, q_{M+1}(x))}{\|v\|_X} \;=\; & (\mathcal{Z}_0, \mathcal{Z}_0)_X + \sum_{n=1}^{N} u_{N,M\,n}(\mu)(\mathcal{Z}_0, \mathcal{Z}_n)_X \\ & + \sum_{n=1}^{N}\sum_{n'=1}^{N} u_{N,M\,n}(\mu)u_{N,M\,n'}(\mu)\, (\mathcal{Z}_n, \mathcal{Z}_{n'})_X \end{aligned} \tag{6.50}$$

where $(\mathcal{Z}_0, v)_X = f(v; q_{M+1}(x))$, $(\mathcal{Z}_n, v)_X = -a_1(\zeta_n, v, q_{M+1}(x))$, $\forall\, v \in X$, $1 \le n \le N$. The offline-online decomposition may now be identified.

In the offline stage — performed only once — we first solve for $\mathcal{C}_k, \mathcal{L}_{mn}, \mathcal{Z}_0$, and $\mathcal{Z}_n$, $1 \le k \le M+N, 1 \le n \le N, 1 \le m \le M$; we then form and store the associated parameter-independent inner products $(\mathcal{C}_k, \mathcal{C}_{k'})_X, (\mathcal{C}_k, \mathcal{L}_{mn})_X$, $(\mathcal{L}_{mn}, \mathcal{L}_{m'n'})_X, (\mathcal{Z}_0, \mathcal{Z}_0)_X$, $(\mathcal{Z}_0, \mathcal{Z}_n)_X, (\mathcal{Z}_n, \mathcal{Z}_{n'})_X$, $1 \le n, n' \le N, 1 \le m, m' \le M$, $1 \le k, k' \le M+N$. This requires $1 + M + 2N + MN$ (expensive) finite element solutions and $1 + N + N^2 + (M+N)^2 + NM(M+N) + M^2N^2$ finite-element-vector inner products. Note that all quantities computed offline are independent of the parameter μ.

In the online stage — performed many times for each new μ — we simply evaluate the two sums (6.49) and (6.50) in terms of $\varphi_{M\,m}(\mu), u_{N,M\,n}(\mu)$ and the precomputed inner products. The operation count for the online stage is only $O(M^2N^2)$; again the online complexity is independent of $\mathcal{N}$. Note however that if M is the same order of N, the online cost for calculating the error bounds is one degree higher than the online cost for

evaluating $s_{N,M}(\mu)$.

6.5.3 Sample Construction and Adaptive Online Strategy

Our error estimation procedures also allow us to pursue (i) more rational constructions of our parameter sample S_N and (ii) efficient execution of the online stage in which we can choose minimal N and M such that the error criterion $\|u(\mu) - u_N(\mu)\|_X \equiv \|e(\mu)\|_X \leq \epsilon_{\text{tol}}$ and the Safety Condition (6.43) are satisfied. We denote the smallest error tolerance anticipated as $\epsilon_{\text{tol, min}}$ — this must be determined *a priori* offline; we then permit $\epsilon_{\text{tol}} \in [\epsilon_{\text{tol, min}}, \infty[$ to be specified online. In addition to the random sample Ξ^g of size $n_{\text{G}} \gg 1$, we introduce $\Xi^u \in \mathcal{D}^{n_{\text{F}}}$, a very fine random sample over the parameter domain $\mathcal{D}$ of size $n_{\text{F}} \gg 1$.

We first consider the offline stage. We set $M = M_{\max} - 1$, $N = 1$, and choose an initial (random) sample set $S_1 = \{\mu_1\}$ and hence space W_1. We then calculate $\mu^*_{N+1} = \arg\max_{\mu \in \Xi^u} \Delta_{N,M}(\mu)$; here $\Delta_{N,M}(\mu)$ is our "online" error bound (6.39) that, in the limit of $n_{\text{F}} \to \infty$ queries, may be evaluated (on average) at cost $O(N^2M^2 + N^3)$. We next append μ^*_{N+1} to S_N to form S_{N+1}, and hence W_{N+1}. We continue this process until $N = N_{\max}$ such that $\epsilon^*_{N_{\max}} = \epsilon_{\text{tol,min}}$, where $\epsilon^*_N \equiv \Delta_{N,M}(\mu^*_N)$, $1 \leq N \leq N_{\max}$. In addition, we compute and store $\varepsilon^*_M \equiv \arg\max_{\mu \in \Xi^u} \varepsilon_M(\mu)$ and $\|\hat{e}_{N,M_{\max}-1}(\mu^*_N)\|_X$ for all $M \in [1, M_{\max}]$ and $N \in [1, N_{\max}]$.

In the online stage, given any desired $\epsilon_{\text{tol}} \in [\epsilon_{\text{tol, min}}, \infty[$ and any new μ, we first choose N from a pre-tabulated array such that $\epsilon^*_N \, (\equiv \Delta_{N,M}(\mu^*_N)) = \epsilon_{\text{tol}}$ and choose M accordingly from another pre-tabulated array such that $\varepsilon^*_M \approx \|\hat{e}_{N,M_{\max}-1}(\mu^*_N)\|_X$. We next calculate $u_{N,M}(\mu)$ and $\Delta_{N,M}(\mu)$ totally in $O(M^2N^2 + N^3)$ operations, and verify that $\Delta_{N,M}(\mu) \leq \epsilon_{\text{tol}}$ is indeed satisfied. If the condition is not yet satisfied we increment $M := M + M^+$ (say, $M^+ = 1$) until either $\Delta_{N,M}(\mu) \leq \epsilon_{\text{tol}}$, $\Delta_{N,M,\text{n}}(\mu)/\Delta_{N,M}(\mu) \leq 1/2$ or $\Delta_{N,M}(\mu)$ does not further decrease;[1] in the latter case, we subsequently increase N while ensuring $\Delta_{N,M,\text{n}}(\mu)/\Delta_{N,M}(\mu) \leq 1/2$ until $\Delta_{N,M}(\mu) \leq \epsilon_{\text{tol}}$. This strategy will provide not only online efficiency but also the requisite rigor and accuracy with certainty. (We should

[1] We should increase M first because (i) our sample construction would ensure $\Delta_{N,M}(\mu) \leq \epsilon_{\text{tol}}, \forall \mu \in \mathcal{D}$ (in the limit of $n_{\text{F}} \to \infty$) for the chosen N and $M = M_{\max} - 1$, and (ii) the online cost grows faster with N than with M.

not and do not rely on the finite sample Ξ^u for either rigor or sharpness.)

6.5.4 Numerical Results

We readily apply our approach to the model problem described in Section 6.1.3. It should be mentioned that the problem is coercive and that we choose bound conditioner, $(w,v)_X = \int_\Omega \nabla w \cdot \nabla v$. It thus follows that $\alpha(\mu) \equiv \inf_{v\in X}\{a(v,v,g(x;\mu))/||v||_X^2\} > 1$; and hence $\hat{\alpha}(\mu) = 1$ is a valid lower bound for $\alpha(\mu), \forall \mu \in \mathcal{D}$. The sample set S_N and associated reduced-basis space W_N are developed based on an adaptive sampling procedure 6.5.3: for $n_{\rm F} = 1600$ and $\epsilon_{\rm tol,min} = 2 \times 10^{-5}$, we obtain $N_{\max} = 20$.

We now introduce a parameter sample $\Xi_{\rm Test} \subset (\mathcal{D})^{225}$ of size 225 (in fact, a regular 15×15 grid over $\mathcal{D}$), and define $\varepsilon_{N,M,\max,\rm rel} = \max_{\mu\in\Xi_{\rm Test}} \|e_{N,M}(\mu)\|_X/\|u_{\max}\|_X$ and $\varepsilon^s_{N,M,\max,\rm rel} = \max_{\mu\in\Xi_{\rm Test}} |s(\mu) - s_{N,M}(\mu)|/|s_{\max}|$; here $\|u_{\max}\|_X = \max_{\mu\in\Xi_{\rm Test}} \|u(\mu)\|_X$ and $|s_{\max}| = \max_{\mu\in\Xi_{\rm Test}} |s(\mu)|$. We present in Figure 6-3 $\varepsilon_{N,M,\max,\rm rel}$ and $\varepsilon^s_{N,M,\max,\rm rel}$ as a function of N and M. We observe the reduced-basis approximations converge very rapidly. Note the "plateau" in the curves for M fixed and the "drops" in the $N \to \infty$ asymptotes as M is increased, reflecting the trade-off between the reduced-basis approximation and coefficient-function approximation contribution to the error: for fixed M the error in our coefficient function approximation $g_M(x;\mu)$ to $g(x;\mu)$ will ultimately dominate for large N; increasing M renders the coefficient function approximation more accurate, which in turn leads to the drops in the error. Note further the separation points in the convergence plot reflecting the balanced contribution of the reduced-basis approximation and coefficient-function approximation to the error: increasing either N or M have very small effect on the error, and the error can only be reduced by increasing both N and M.

We furthermore present in Table 6.2 $\Delta_{N,M,\max,\rm rel}$, $\bar{\eta}_{N,M}$, $\Delta^s_{N,M,\max,\rm rel}$, and $\bar{\eta}^s_{N,M}$ as a function of N and M. Here $\Delta_{N,M,\max,\rm rel}$ is the maximum over $\Xi_{\rm Test}$ of $\Delta_{N,M}(\mu)/\|u_{\max}\|_X$, $\bar{\eta}_{N,M}$ is the average over $\Xi_{\rm Test}$ of $\Delta_{N,M}(\mu)/\|e(\mu)\|_X$, $\Delta^s_{N,M,\max,\rm rel}$ is the maximum over $\Xi_{\rm Test}$ of $\Delta^s_{N,M}(\mu)/|s_{\max}|$, and $\bar{\eta}^s_{N,M}$ is the average over $\Xi_{\rm Test}$ of $\Delta^s_{N,M}(\mu)/|s(\mu) - s_{N,M}(\mu)|$. We observe that the reduced-basis approximation — in particular, for the solution — converges very rapidly, and that the energy error bound is quite sharp as its effectivities are in order of $O(1)$. However, the effectivities for the output estimate are large and thus

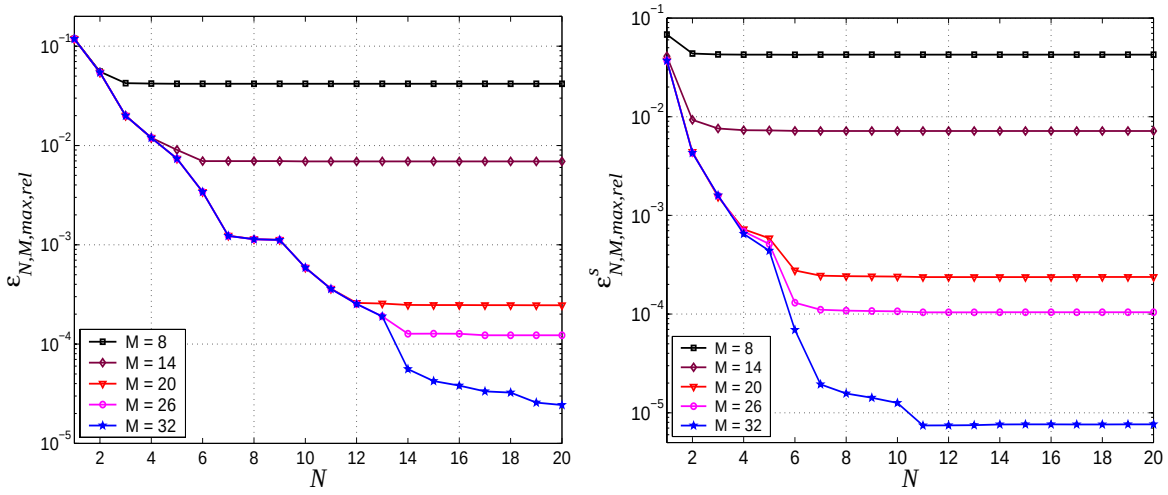


Figure 6-3: Convergence of the reduced-basis approximations for the model problem.

our output bounds are not sharp — we will further discuss this issue in the next section.

N	M	$\Delta_{N,M,\text{max,rel}}$	$\bar{\eta}_{N,M}$	$\Delta^s_{N,M,\text{max,rel}}$	$\bar{\eta}^s_{N,M}$
4	15	1.35 E – 02	1.16	1.43 E – 02	11.32
8	20	1.23 E – 03	1.01	1.30 E – 03	13.41
12	25	2.77 E – 04	1.08	2.92 E – 04	17.28
16	30	3.93 E – 05	1.00	4.15 E – 05	20.40

Table 6.2: Effectivities for the model problem.

In general, $g(x;\mu) \in W^g_{M+1}$ is not satisfied and our error estimators may thus not be completely rigorous upper bounds. We may thus investigate the relative contribution of the rigorous and non-rigorous components to the error bound $\Delta_{N,M}(\mu)$. For this purpose we define $\Delta_{N,M,\text{ave}}$ as the average over Ξ_{Test} of $\Delta_{N,M}(\mu)$ and $\Delta_{N,M,\text{ave,n}}$ as the average over Ξ_{Test} of $\Delta_{N,M,\text{n}}(\mu)$. We present in Figure 6-4 the ratio $\Delta_{N,M,\text{ave,n}}/\Delta_{N,M,\text{ave}}$ as a function of N and M. We observe that the ratio increases with N, but decreases with M. This is because $\Delta_{N,M}(\mu)$ converges faster with N but slower with M than $\Delta_{N,M,\text{n}}(\mu)$. We can now understand our adaptive online strategy more clearly by looking at the graph 6-4. For example, in the online stage, we already choose $N = 12$ from a pre-tabulated, our online adaptivity will be likely to give $M = 20$. This is because, for $M < 20$, the Safety Condition (possibly the error criterion as well) is not satisfied; at $M = 20$, the Safety Condition is satisfied as seen from the graph; increasing M above 20 will not improve the error bounds too much, but online complexity increases. Note further that for $M = 26$,

the nonrigorous component $\Delta_{N,M,\mathrm{n}}$ is almost less than 10 percentage of $\Delta_{N,M}(\mu)$ for all N ($\leq N_{\max}$) — its contribution to the error bound is very small.

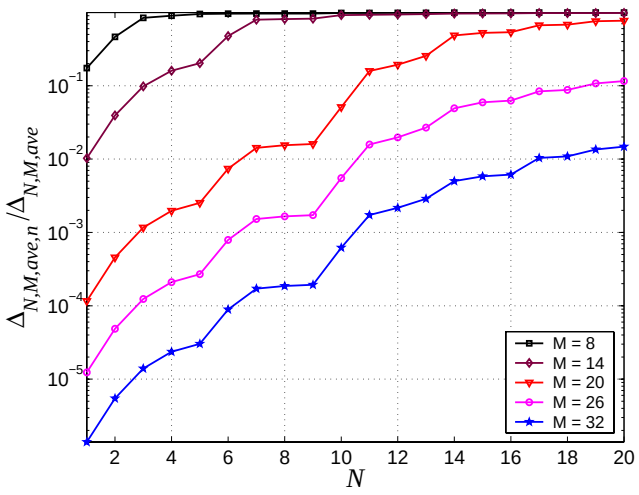


Figure 6-4: $\Delta_{N,M,\mathrm{ave,n}}/\Delta_{N,M,\mathrm{ave}}$ as a function of N and M.

Finally, we present in Table 6.3 the online computational times to calculate $s_{N,M}(\mu)$ and $\Delta^s_{N,M}(\mu)$ as a function of (N, M). The values are normalized with respect to the computational time for the direct calculation of the truth approximation output $s(\mu) = \ell(u(\mu))$. We achieve significant computational savings: for an accuracy of close to 0.1 percent ($N = 8$, $M = 20$) in the output bound, the online saving is more than a factor of 100. We also note that the time to calculate $\Delta^s_{N,M}(\mu)$ exceeds that of calculating $s_N(\mu)$ considerably — this is due to the higher computational cost, $O(M^2N^2)$, to evaluate $\Delta_{N,M}(\mu)$. Hence, although the theory suggests to choose M large so that the nonrigorous component due to the coefficient function approximation does not dominate the rigorous component, we should choose M as small as possible to retain the computational efficiency. Our online adaptive strategy is thus necessary for providing rigor, accuracy and efficiency.

6.5.5 Remark on Noncoercive Case

We close this section with a short discussion on the application of our approach to noncoercive problems. We have presented the approach to a coercive case in which the lower bound of the stability factor, $\hat{\alpha}(\mu)$, may be deduced analytically as shown in the numerical example. For noncoercive case, analytical deduction is generally not easy; and in

N	M	$s_{N,M}(\mu)$	$\Delta^s_{N,M}(\mu)$	$s(\mu)$
4	15	2.39 E−04	3.77 E−03	1
8	20	4.33 E−04	6.40 E−03	1
12	25	5.41 E−03	9.90 E−03	1
16	30	6.93 E−03	1.41 E−02	1

Table 6.3: Online computational times (normalized with respect to the time to solve for $s(\mu)$) for the model problem.

most cases the inf-sup lower bound $\hat{\beta}(\mu)$ to $\beta(\mu)$ must typically be constructed, where

$$\beta(\mu) \equiv \inf_{w \in X} \sup_{v \in X} \frac{a(w, v, g(x;\mu))}{\|w\|_X \|v\|_X} \ . \tag{6.51}$$

The primary difficulty here lies in the nonaffine dependence of $a(w, v, g(x;\mu))$ on μ, because our lower bound construction described in Chapter 4 is only valid for affine operators and thus no longer applicable to this case. To resolve the difficulty, we first construct a lower bound for the inf-sup parameter $\beta_M(\mu)$ associated with the approximated bilinear form; we must then add an additional inf-sup correction β^c due to the operator perturbation. Specifically, we define

$$\beta_M(\mu) \equiv \inf_{w \in X} \sup_{v \in X} \frac{a_0(w, v) + a_1(w, v, g_M(x;\mu))}{\|w\|_X \|v\|_X} \ , \tag{6.52}$$

$$\gamma^c_M \equiv \sup_{w \in X} \sup_{v \in X} \frac{a_1\left(w, v, q_{M+1}(x)\right)}{\|w\|_X \|v\|_X} \ , \tag{6.53}$$

and introduce a supremizer

$$v^*(w) = \arg\sup_{v \in X} \frac{a_0(w, v) + a_1(w, v, g_M(x;\mu))}{\|v\|_X} \ . \tag{6.54}$$

It follows from (6.52) and (6.54) that

$$\beta_M(\mu) = \inf_{w \in X} \frac{a_0(w, v^*(w)) + a_1(w, v^*(w), g_M(x;\mu))}{\|w\|_X \|v^*(w)\|_X} \ . \tag{6.55}$$

We now continue to assume that $g(x;\mu) \in W^g_{M+1}$. It immediately follows from (6.51)-(6.55) that

$$\begin{aligned}
\beta(\mu) &= \inf_{w\in X} \sup_{v\in X} \frac{a_0(w,v) + a_1\left(w,v,g_M(x;\mu) \pm \hat{\varepsilon}_M(\mu) q_{M+1}(x)\right)}{\|w\|_X \|v\|_X} \\
&= \inf_{w\in X} \sup_{v\in X} \frac{a_0(w,v) + a_1(w,v,g_M(x;\mu)) \pm \hat{\varepsilon}_M(\mu) a_1\left(w,v,q_{M+1}(x)\right)}{\|w\|_X \|v\|_X} \\
&\geq \inf_{w\in X} \frac{a_0(w,v^*(w)) + a_1(w,v^*(w),g_M(x;\mu)) \pm \hat{\varepsilon}_M(\mu) a_1\left(w,v^*(w),q_{M+1}(x)\right)}{\|w\|_X \|v^*(w)\|_X} \\
&\geq \inf_{w\in X} \frac{a_0(w,v^*(w)) + a_1(w,v^*(w),g_M(x;\mu))}{\|w\|_X \|v^*(w)\|_X} - \hat{\varepsilon}_M(\mu) \sup_{w\in X} \frac{a_1\left(w,v^*(w),q_{M+1}(x)\right)}{\|w\|_X \|v^*(w)\|_X} \\
&\geq \beta_M(\mu) - \hat{\varepsilon}_M(\mu)\gamma^{\mathrm{c}}_M \ . \qquad (6.56)
\end{aligned}$$

Since $a_1(w,v,g_M(x;\mu))$ is affine in parameter, we can construct the lower bound $\hat{\beta}_M(\mu)$ for $\beta_M(\mu)$ by the method developed in Chapter 4. Note further that the inf-sup correction γ^{c}_M is independent of the parameter and can thus be computed offline. Hence, given any new μ we obtain the lower bound for $\beta(\mu)$ as $\hat{\beta}(\mu) = \hat{\beta}_M(\mu) - \hat{\varepsilon}_M(\mu)\gamma^{\mathrm{c}}_M$. Of course, our remark is also applicable to coercive problems in which the lower bound for the stability factor can not be found by inspection.

6.6 Adjoint Techniques

We consider here an alternative reduced-basis approximation and *a posteriori* error estimation procedure relevant to noncompliant output functional ℓ ($\neq f$). In particular, we shall employ a "primal-dual" formulation well-suited to good approximation and error characterization of the output. As a generalization of our abstract formulation in Section 6.1, we define the primal problem as in (6.7) and also introduce an associated dual, or adjoint, problem: given $\mu \in \mathcal{D}$, $\psi(\mu) \in X$ satisfies

$$a(v,\psi(\mu);g(x;\mu)) = -\ell(v), \quad \forall\ v \in X. \qquad (6.57)$$

6.6.1 Important Theoretical Observation

Before proceeding with our development, we give a theoretical explanation for why the reduced-basis formulation described in Sections 6.4 and 6.5 can result in slower output convergence for the noncompliance case than the primal-dual formulation discussed in this section. We first need to prove an intermediate result

Proposition 9. *Suppose that $g(x;\mu) \in W^g_{M+1}$, then, for all $w_N \in W_N$, we have*

$$\begin{aligned} s(\mu) - s_{N,M}(\mu) &= -a(u(\mu) - u_{N,M}(\mu), \psi(\mu) - w_N; g(\cdot;\mu)) \\ &\quad \pm \varepsilon_M(\mu)\{a(u_{N,M}(\mu), w_N; q_{M+1}) - f(w_N; q_{M+1})\}. \end{aligned} \tag{6.58}$$

Proof. We invoke (6.6),(6.7), (6.57), our hypothesis $g(x;\mu) \in W^g_{M+1}$ and Proposition 6 to obtain the desired result

$$\begin{aligned} s(\mu) - s_{N,M}(\mu) &= \ell(u(\mu) - u_{N,M}(\mu)) \\ &= -a(u(\mu) - u_{N,M}(\mu), \psi(\mu) - w_N; g(\cdot;\mu)) \\ &\quad - a(u(\mu) - u_{N,M}(\mu), w_N; g(\cdot;\mu)), \\ &= -a(u(\mu) - u_{N,M}(\mu), \psi(\mu) - w_N; g(\cdot;\mu)) - f(w_N; g(\cdot;\mu) - g_M(\cdot;\mu)) \\ &\quad + a(u_{N,M}(\mu), w_N; g(\cdot;\mu) - g_M(\cdot;\mu)), \\ &= -a(u(\mu) - u_{N,M}(\mu), \psi(\mu) - w_N; g(\cdot;\mu)) \\ &\quad \pm \varepsilon_M(\mu)\{a(u_{N,M}(\mu), w_N; q_{M+1}) - f(w_N; q_{M+1})\}, \end{aligned} \tag{6.59}$$

for all $w_N \in W_N$. □

We readily interpret the theoretical implications of Proposition 9. If $\varepsilon_M(\mu)$ is sufficiently small such that the second term of the output error can be ignored for any $w_N \in W_N$, we consider two cases of the dual solution $\psi(\mu)$ to the primal approximation space W_N. In the first case, $\psi(\mu) - w_N$ is large for all $w_N \in W_N$, the first term of the output error is thus also large compared to $a(u(\mu) - u_{N,M}(\mu), \psi(\mu) - \psi_{N,M}(\mu); g(\cdot;\mu))$ which results from the primal-dual formulation; here $\psi_{N,M}(\mu)$ is the reduced-basis approxima-

tion for $\psi(\mu)$. The reduced-basis output without dual correction will thus converge slower than that with dual correction. In the second case, $\psi(\mu) - w_N$ is small for some $w_N \in W_N$, the actual output error is thus much smaller than $\Delta^s_{N,M}(\mu) \equiv \sup_{v \in X}[\ell(v)/\|v\|_X]\Delta_{N,M}(\mu)$ since this error bound is essentially based on (6.58) for $w_N = 0$.[2] This will result in large output effectivity as already observed in Table 6.2. The primal-dual formulation can significantly improve the poor output effectivity due to this effect; however, as we shall point out, the output effectivity can still be large because the primal-dual formulation does not capture the "correlation" between the primal error and dual error into the output bound. Of course, if $\varepsilon_M(\mu)$ is large such that the second term in (6.58) dominates, we can not improve the output convergence even with the introduction of the adjoint techniques.

Numerical results in [52] also confirmed our theoretical claim: by using the adjoint techniques, significant improvement for both the output approximation and output effectivity has been observed. This improvement can then translate to online efficiency relative to the usual reduced-basis formulation. In particular, we observe that the online cost for solving either the primal problem or the dual problem is typically $O(N^3)$ under some reasonable assumption on the order of M [52], the online complexity for the primal-dual formulation is thus $O(2N^3)$; we further assume that $\varepsilon_M(\mu)$ is in order of $O(a(u(\mu) - u_{N,M}(\mu), \psi(\mu) - \psi_{N,M}(\mu); g_M(\cdot;\mu)))$; thus in order to obtain the same output bound for the usual reduced-basis formulation without the dual problem we would need to increase N by a factor of 2 or even more, leading to an online cost of $O(8N^3)$ or higher (Section 3.5.1 provides clarification for our claim). As a result, the dual reduced-basis formulation typically enjoys $O(4)$ (or greater) reduction in computational effort. However, the simple crude output bound $\Delta^s_N(\mu) = \|\ell\|_{X'}\,\Delta_N(\mu)$ is still very useful for cases with *many* outputs present, since adjoint techniques have a computational complexity (in both the offline and online stage) proportional to the number of outputs.

In this section, we shall not discuss the primal-dual formulation and theory for this set of problems, as the detail was given in [52]. Instead, we will develop a primal-dual formulation for another set of problems relevant to the inverse scattering problems described in Chapter 10.

[2] To see this, we need only note from (6.57) that $\sup_{v \in X} \frac{\ell(v)}{\|v\|_X} = \sup_{v \in X} \frac{a(v,\psi(\mu);g(\cdot;\mu))}{\|v\|_X}$.

6.6.2 Problem Statement

We consider the following problem: Given $\mu \in \mathcal{D}$, we evaluate

$$s(\mu) = \ell(\overline{u}(\mu); h(x;\mu)) \ , \tag{6.60}$$

where $u(\mu)$ is the solution of

$$a(u, v; \mu) = f(v; g(x;\mu)), \quad \forall\, v \in X \ . \tag{6.61}$$

Here $a(\cdot,\cdot;\mu)$ and $f(\cdot; g(x;\mu)), \ell(\cdot; h(x;\mu))$ are continuous *complex* bilinear form and linear functionals, respectively; the nonaffine *complex* functions, $g(x;\mu)$ and $h(x;\mu)$, are assumed to be continuous in the closed domain $\bar{\Omega}$ and sufficiently smooth with respect to μ; and X is a *complexified* truth approximation space. We further assume that a satisfies a continuity and inf-sup condition in terms of a supremizing operator $T^\mu : X \to X$: in particular, for any w in X,

$$(T^\mu w, v)_X = a(w, v; \mu), \quad \forall\, v \in X \ ; \tag{6.62}$$

we then define

$$\sigma(w;\mu) \equiv \frac{\|T^\mu w\|_X}{\|w\|_X} \ , \tag{6.63}$$

and require that

$$0 < \beta_0 \le \beta(\mu) \equiv \inf_{w\in X} \sup_{v\in X} \frac{|a(w, v;\mu)|}{\|w\|_X \|v\|_X} = \inf_{w\in X} \sigma(w;\mu), \quad \forall\, \mu \in \mathcal{D} \ , \tag{6.64}$$

$$\gamma(\mu) \equiv \sup_{w\in X} \sup_{v\in X} \frac{|a(w, v;\mu)|}{\|w\|_X \|v\|_X} = \sup_{w\in X} \sigma(w;\mu) < \infty, \quad \forall\, \mu \in \mathcal{D} \ . \tag{6.65}$$

Finally, we assume that for some finite integer Q, a may be expressed

$$a(w, v;\mu) = \sum_{q=1}^{Q} \Theta^q(\mu) a^q(w, v), \quad \forall\, w, v \in X,\ \forall\, \mu \in \mathcal{D}; \tag{6.66}$$

where for $1 \le q \le Q$, $\Theta^q : \mathcal{D} \to \mathbb{C}$ are differentiable complex parameter-dependent coefficient functions and bilinear forms $a^q : X \times X \to \mathbb{C}$ are parameter-independent.

6.6.3 Reduced-Basis Approximation

Discrete Equations

We first develop a primal-dual reduced-basis approximation that can significantly improve accuracy of the output approximation. To begin, we introduce a dual, or adjoint, problem: given $\mu \in \mathcal{D}$, $\psi(\mu) \in X$ satisfies

$$a(v, \psi(\mu); \mu) = -\ell(\bar{v}; h(x; \mu)), \quad \forall\ v \in X. \tag{6.67}$$

We next construct nested parameter samples $S^g_{M^g} = \{\mu^g_1 \in \mathcal{D}, \cdots, \mu^g_{M^g} \in \mathcal{D}\}, 1 \le M^g \le M^g_{\max}$, $S^h_{M^h} = \{\mu^h_1 \in \mathcal{D}, \cdots, \mu^h_{M^h} \in \mathcal{D}\}, 1 \le M^h \le M^h_{\max}$, associated nested approximation spaces $W^g_{M^g} = \text{span}\{q^g_1, \ldots, q^g_{M^g}\}, 1 \le M^g \le M^g_{\max}$, $W^h_{M^h} = \text{span}\{q^h_1, \ldots, q^h_{M^h}\}$, $1 \le M^h \le M^h_{\max}$, and the nested sets of interpolation points $T^g_M = \{t^g_1, \ldots, t^g_{M^g}\}, 1 \le M^g \le M^g_{\max}$, $T^h_M = \{t^h_1, \ldots, t^h_{M^h}\}, 1 \le M^h \le M^h_{\max}$ following the procedure of Section 6.2. For the primal problem, (6.61), we introduce nested parameter samples $S_N \equiv \{\mu^{\text{pr}}_1 \in \mathcal{D}, \ldots, \mu^{\text{pr}}_N \in \mathcal{D}\}$ and associated nested reduced-basis spaces $W_N \equiv \text{span}\ \{\zeta_n \equiv u(\mu^{\text{pr}}_n), 1 \le n \le N\}$ for $1 \le N \le N_{\max}$; similarly, for the dual problem (6.67), we define corresponding samples $S^{\text{du}}_{N^{\text{du}}} \equiv \{\mu^{\text{du}}_1 \in \mathcal{D}, \ldots, \mu^{\text{du}}_{N^{\text{du}}} \in \mathcal{D}\}$ and reduced-basis approximation spaces $W^{\text{du}}_{N^{\text{du}}} \equiv \text{span}\ \{\zeta^{\text{du}}_n \equiv \psi(\mu^{\text{du}}_n),\ 1 \le n \le N^{\text{du}}\}$ for $1 \le N^{\text{du}} \le N^{\text{du}}_{\max}$. Our reduced-basis approximation is thus: given $\mu \in \mathcal{D}$, we evaluate

$$s_N(\mu) = \ell(\overline{u}_N(\mu); h_{M^h}(x; \mu)) - r^{\text{pr}}(\psi_{N^{\text{du}}}(\mu); g_{M^g}(x; \mu))\ , \tag{6.68}$$

where $u_N(\mu) \in W_N$ and $\psi_{N^{\text{du}}}(\mu) \in W^{\text{du}}_{N^{\text{du}}}$ satisfy

$$a(u_N(\mu), v; \mu) = f(v; g_{M^g}(x; \mu)), \quad \forall\ v \in W_N\ , \tag{6.69}$$

$$a(v, \psi_{N^{\text{du}}}(\mu); \mu) = -\ell(\overline{v}; h_{M^h}(x; \mu)), \quad \forall\ v \in W^{\text{du}}_{N^{\text{du}}}\ . \tag{6.70}$$

Here $r^{\rm pr}(v; g_{M^g}(x;\mu))$ is the residual associated with the primal problem

$$r^{\rm pr}(v; g_{M^g}(x;\mu)) = f(v; g_{M^g}(x;\mu)) - a(u_N(\mu), v; \mu), \quad \forall\ v \in X\ . \tag{6.71}$$

Recall that $g_{M^g}(x;\mu)$ and $h_{M^h}(x;\mu)$ — the coefficient-function approximations for $g(x;\mu)$ and $h(x;\mu)$, respectively — are given by

$$g_{M^g}(x;\mu) = \sum_{m=1}^{M^g} \varphi^g_{M^g\,m}(\mu) q^g_m\ , \qquad h_{M^h}(x;\mu) = \sum_{m=1}^{M^h} \varphi^h_{M^h\,m}(\mu) q^h_m\ ; \tag{6.72}$$

where $\sum_{j=1}^{M^g} q^g_j(t^g_i)\ \varphi^g_{M^g\,j}(\mu) = g(t^g_i;\mu), 1 \le i \le M^g$, and $\sum_{j=1}^{M^h} q^h_j(t^h_i)\ \varphi^h_{M^h\,j}(\mu) = h(t^h_i;\mu), 1 \le i \le M^h$.

Offline-Online Computational Procedure

We expand our reduced-basis approximations as

$$u_N(\mu) = \sum_{j=1}^{N} u_{N\,j}(\mu)\zeta_j\ , \qquad \psi_{N^{\rm du}}(\mu) = \sum_{j=1}^{N^{\rm du}} \psi_{N^{\rm du}\,j}(\mu)\zeta^{\rm du}_j\ . \tag{6.73}$$

It then follows from (6.66), (6.68), (6.71), (6.72), and (6.73) that

$$\begin{aligned} s_N(\mu) \;=\; & \sum_{m=1}^{M^h}\sum_{j=1}^{N} \varphi^h_{M^h\,m}(\mu) u_{N\,j}(\mu)\ \ell(\overline{\zeta}_j; q^h_m) - \sum_{m=1}^{M^g}\sum_{j=1}^{N^{\rm du}} \varphi^g_{M^g\,m}(\mu)\overline{\psi}_{N^{\rm du}\,j}(\mu) f(\zeta^{\rm du}_j; q^g_m) \\ & + \sum_{j=1}^{N}\sum_{j'=1}^{N^{\rm du}}\sum_{q=1}^{Q} u_{N\,j}(\mu)\overline{\psi}_{N^{\rm du}\,j'}(\mu)\Theta^q(\mu) a^q(\zeta_j, \zeta^{\rm du}_{j'})\ , \end{aligned} \tag{6.74}$$

where the coefficients $u_{N\,j}(\mu)$, $1 \le j \le N$, and $\psi_{N^{\rm du}\,j}$, $1 \le j \le N^{\rm du}$, satisfy the $N \times N$ and $N^{\rm du} \times N^{\rm du}$ linear algebraic systems

$$\sum_{j=1}^{N}\left\{\sum_{q=1}^{Q}\Theta^q(\mu)\ a^q(\zeta_j,\zeta_i)\right\} u_{N\,j}(\mu) = \textstyle\sum_{m=1}^{M^g} \varphi^g_{M^g\,m}(\mu) f(\zeta_i; q^g_m), \quad 1 \le i \le N\ , \tag{6.75}$$

$$\sum_{j=1}^{N^{\rm du}}\left\{\sum_{q=1}^{Q}\Theta^q(\mu)\ a^q(\zeta^{\rm du}_i,\zeta^{\rm du}_j)\right\} \overline{\psi}_{N^{\rm du}\,j}(\mu) = -\textstyle\sum_{m=1}^{M^h} \varphi^h_{M^h\,m}(\mu)\ell(\overline{\zeta}^{\rm du}_i; q^h_m), 1 \le i \le N^{\rm du}\ . \tag{6.76}$$

The offline-online decomposition is now clear. For simplicity below we assume that $N^{\rm du} = N$ and $M^h = M^g \equiv M$.

In the offline stage — performed *once* — we first generate nested approximation spaces $W_M^g = \{q_1^g, \ldots, q_M^g\}$, $W_M^h = \{q_1^h, \ldots, q_M^h\}$; we then solve for the ζ_i, $\zeta_i^{\rm du}$, $1 \le i \le N$; we finally form *and store* $\ell(\overline{\zeta}_i; q_m^h)$, $f(\zeta_i; q_m^g)$, $\ell(\overline{\zeta}_i^{\rm du}; q_m^h)$, and $f(\zeta_i^{\rm du}; q_m^g)$, $1 \le i \le N$, $1 \le m \le M$, and $a^q(\zeta_j, \zeta_i)$, $a^q(\zeta_i^{\rm du}, \zeta_j^{\rm du})$, $a^q(\zeta_i, \zeta_j^{\rm du})$, $1 \le i, j \le N$, $1 \le q \le Q$. This requires $2N$ (expensive) finite element solutions, linear optimization cost for constructing W_M^g and W_M^h, and $4MN + 3QN^2$ finite-element-vector inner products. Note all quantities computed in the offline stage are independent of the parameter μ.

In the online stage — performed *many times*, for each new value of μ — we first solve for the $\varphi^g_{M^g\,m}(\mu), \varphi^h_{M^h\,m}(\mu)$, $1 \le m \le M$; we then assemble and subsequently invert the $N \times N$ "stiffness matrices" $\sum_{q=1}^Q \Theta^q(\mu)\, a^q(\zeta_j, \zeta_i)$ of (6.75) and $\sum_{q=1}^Q \Theta^q(\mu)\, a^q(\zeta_i^{\rm du}, \zeta_j^{\rm du})$ of (6.76) — this yields the $u_{N\,j}(\mu)$, $\psi_{N^{\rm du}\,j}(\mu)$, $1 \le j \le N$; we finally perform the summation (6.74) — this yields the $s_N(\mu)$. The operation count for the online stage is respectively $O(M^2)$ to calculate the $\varphi^g_{M^g\,m}(\mu), \varphi^h_{M^h\,m}(\mu)$, $1 \le m \le M$, $O(QN^2)$ and $O(N^3)$ to assemble and invert the stiffness matrices, and $O(MN) + O(QN^2)$ to evaluate the output. The essential point is that the online complexity is *independent of* $\mathcal{N}$, the dimension of the underlying truth finite element approximation space. Since $M^g, M^h, N, N^{\rm du} \ll \mathcal{N}$, we expect — and often realize — significant, orders-of-magnitude computational economies relative to classical discretization approaches.

6.6.4 *A Posteriori* Error Estimators

Error Bounds

We assume that we are privy to a lower bound for the inf-sup parameter, $\hat{\beta}(\mu)$, such that $\beta(\mu) \ge \hat{\beta}(\mu) \ge \epsilon_\beta \beta(\mu)$, $\forall\, \mu \in \mathcal{D}$, where $\epsilon_\beta \in]0, 1[$. The construction of $\hat{\beta}(\mu)$ is described in detail in Section 4.3. We then introduce an energy error bound for the primal

$$\Delta_N(\mu) = \frac{1}{\hat{\beta}(\mu)} \left(\sup_{v \in X} \frac{r^{\rm pr}(v; g_{M^g})}{\|v\|_X} + \hat{\varepsilon}^g_{M^g} \sup_{v \in X} \frac{f(v; q^g_{M^g+1})}{\|v\|_X} \right) , \tag{6.77}$$

and an energy error bound for the dual

$$\Delta^{\rm du}_{N^{\rm du}}(\mu) = \frac{1}{\hat{\beta}(\mu)} \left(\sup_{v \in X} \frac{r^{\rm du}(v; h_{M^h})}{\|v\|_X} + \hat{\varepsilon}^h_{M^h} \sup_{v \in X} \frac{\ell(v; q^h_{M^h+1})}{\|v\|_X} \right) . \tag{6.78}$$

Here $r^{\rm du}(v; h_{M^h}(x;\mu))$ is the residual associated with the dual problem

$$r^{\rm du}(v; h_{M^h}(x;\mu)) = -\ell(\overline{v}; h_{M^h}(x;\mu)) - a(v, \psi_{N^{\rm du}}(\mu); \mu), \quad \forall\, v \in X\ . \tag{6.79}$$

We can then state

Proposition 10. *Suppose that $g(x;\mu) \in W^g_{M^g+1}$ and $h(x;\mu) \in W^h_{M^h+1}$, then the energy error bounds satisfy $\|u(\mu) - u_N(\mu)\|_X \le \Delta_N(\mu)$, $\|\psi(\mu) - \psi_{N^{\rm du}}(\mu)\|_X \le \Delta^{\rm du}_{N^{\rm du}}(\mu)$, $\forall \mu \in \mathcal{D}$.*

Proof. We consider only the primal since the dual result can be derived by a similar route. We first note from (6.61) and (6.71) that the error $e(\mu) \equiv u(\mu) - u_N(\mu)$ satisfies

$$a(e(\mu), v; \mu) = r^{\rm pr}(v; g_{M^g}(x;\mu)) + f(v; g(x;\mu) - g_{M^g}(x;\mu)), \quad \forall\, v \in X\ . \tag{6.80}$$

It then follows (6.62) and (6.80) that

$$(T^\mu e(\mu), v)_X = r^{\rm pr}(v; g_{M^g}(x;\mu)) + f(v; g(x;\mu) - g_{M^g}(x;\mu)), \quad \forall\, v \in X\ . \tag{6.81}$$

The desired result immediately follows

$$\begin{aligned} \|e(\mu)\|_X &\le \frac{\|T^\mu e(\mu)\|_X}{\beta(\mu)} \\ &\le \frac{1}{\hat{\beta}(\mu)} \frac{r^{\rm pr}(T^\mu e(\mu); g_{M^g}(x;\mu)) + f(T^\mu e(\mu); g(x;\mu) - g_{M^g}(x;\mu))}{\|T^\mu e(\mu)\|_X} \\ &\le \frac{1}{\hat{\beta}(\mu)} \left(\sup_{v \in X} \frac{r^{\rm pr}(v; g_{M^g}(x;\mu))}{\|v\|_X} + \hat{\varepsilon}^g_{M^g} \sup_{v \in X} \frac{f(v; q^g_{M^g+1})}{\|v\|_X} \right) , \end{aligned} \tag{6.82}$$

from (6.63) and (6.64) in the first inequality, (6.81) in the second inequality, and our assumption $g(x;\mu) \in W^g_{M^g+1}$ in the last inequality. □

We may also define an output error bound for the error in the output as

$$\Delta_N^s(\mu) = \hat{\varepsilon}_{M^h}^h(\mu)|\ell(\overline{u}_N(\mu); q_{M^h+1}^h)| + \hat{\varepsilon}_{M^g}^g(\mu)|f(\psi_{N^{\rm du}}(\mu); q_{M^g+1}^g)| + \hat{\beta}(\mu)\Delta_N(\mu)\Delta_{N^{\rm du}}^{\rm du}(\mu) \tag{6.83}$$

for which we readily demonstrate

Proposition 11. *Suppose that $g(x;\mu) \in W_{M^g+1}^g$ and $h(x;\mu) \in W_{M^h+1}^h$, then the output error bound satisfies $|s(\mu) - s_N(\mu)| \le \Delta_N^s(\mu)$.*

Proof. We first note from (6.61), (6.67), (6.69), (6.70), and (6.71) to express

$$\begin{aligned}
s(\mu) - s_N(\mu) &= \ell(\overline{u}; h) - \ell(\overline{u}_N; h_{M^h}) + r^{\rm pr}(\psi_{N^{\rm du}}; g_{M^g}) \\
&= \ell(\overline{u}_N; h - h_{M^h}) - a(u, \psi; \mu) + a(u_N, \psi; \mu) + r^{\rm pr}(\psi_{N^{\rm du}}; g_{M^g}) \\
&= \ell(\overline{u}_N; h - h_{M^h}) - f(\psi; g) + f(\psi; g_{M^g}) \\
&\qquad - f(\psi; g_{M^g}) + a(u_N, \psi; \mu) + r^{\rm pr}(\psi_{N^{\rm du}}; g_{M^g}) \\
&= \ell(\overline{u}_N; h - h_{M^h}) - f(\psi_{N^{\rm du}}; g - g_{M^g}) \\
&\qquad - f(\psi - \psi_{N^{\rm du}}; g - g_{M^g}) - r^{\rm pr}(\psi - \psi_{N^{\rm du}}; g_{M^g}) \ .
\end{aligned} \tag{6.84}$$

It thus follows from $g(x;\mu) \in W_{M^g+1}^g$, $h(x;\mu) \in W_{M^h+1}^h$, and Proposition 10 that

$$\begin{aligned}
|s(\mu) - s_N(\mu)| &\le \hat{\varepsilon}_{M^h}^h(\mu)|\ell(\overline{u}_N; q_{M^h+1}^h)| + \hat{\varepsilon}_{M^g}^g(\mu)|f(\psi_{N^{\rm du}}; q_{M^g+1}^g)| \\
&\qquad + \hat{\varepsilon}_{M^g}^g(\mu)|f(\psi - \psi_{N^{\rm du}}; q_{M^g+1}^g)| + |r^{\rm pr}(\psi - \psi_{N^{\rm du}}; g_{M^g})| \\
&\le \hat{\varepsilon}_{M^h}^h(\mu)|\ell(\overline{u}_N; q_{M^h+1}^h)| + \hat{\varepsilon}_{M^g}^g(\mu)|f(\psi_{N^{\rm du}}; q_{M^g+1}^g)| \\
&\qquad + \hat{\varepsilon}_{M^g}^g(\mu)\Delta_{N^{\rm du}}^{\rm du}(\mu) \sup_{v \in X} \frac{f\left(v; q_{M^g+1}^g\right)}{\|v\|_X} + \sup_{v \in X} \frac{r^{\rm pr}(v; g_{M^g})}{\|v\|_X} \Delta_{N^{\rm du}}^{\rm du}(\mu) \\
&= \Delta_N^s(\mu) \ .
\end{aligned}$$

This concludes the proof. □

Equation (6.83) suggests that for rigor of the output bound M^g and M^h should be chosen such that

$$\frac{\Delta_{N,\rm n}^s(\mu)}{\Delta_N^s(\mu)} \le 1/2 \ , \tag{6.85}$$

where $\Delta^s_{N,\mathrm{n}}(\mu)$ is the nonrigorous component in the output bound

$$\Delta^s_{N,\mathrm{n}}(\mu) = \hat{\varepsilon}^h_{M^h}(\mu)|\ell(\overline{u}_N(\mu); q^h_{M^h+1})| + \hat{\varepsilon}^g_{M^g}(\mu)|f(\psi_{N^{\mathrm{du}}}(\mu); q^g_{M^g+1})| \ . \tag{6.86}$$

From the perspective of computational efficiency, M^g and M^h should be chosen so that the ratio $\Delta^s_{N,\mathrm{n}}(\mu)/\Delta^s_N(\mu)$ is as close to 1/2 as possible; that is, $\varepsilon^g_{M^g}(\mu)$ and $\varepsilon^h_{M^h}(\mu)$ are roughly $O(\hat{\beta}(\mu)\Delta_N(\mu)\Delta^{\mathrm{du}}_{N^{\mathrm{du}}}(\mu))$ since $|f(\psi_{N^{\mathrm{du}}}(\mu), q^g_{M^g+1})|$ and $|\ell(\overline{u}_N(\mu), q^h_{M^h+1})|$ are actually $O(1)$ due to $\|q^g_{M^g+1}\|_{L^\infty(\Omega)} = \|q^h_{M^h+1}\|_{L^\infty(\Omega)} = 1$. If M^g and M^h are chosen too large the last term in (6.83) dominates, we may thus obtain rigorous bounds but at the expense of unnecessarily increasing the computational cost. On the other hand, we may risk to obtain nonrigorous bounds if M^g or M^h are chosen too small.

Moreover, if $\varepsilon^g_{M^g}(\mu)$ and $\varepsilon^h_{M^h}(\mu)$ are sufficiently small, we then obtain the same result as for the affine linear case [99], $|s(\mu) - s_N(\mu)| \approx a(u(\mu) - u_N(\mu), \psi(\mu) - \psi_{N^{\mathrm{du}}}) = r^{\mathrm{pr}}(\psi(\mu) - \psi_{N^{\mathrm{du}}}(\mu); g_{M^g}(x;\mu))$ and $\Delta^s_{N,M}(\mu) \approx \hat{\beta}(\mu)\Delta_N(\mu)\Delta^{\mathrm{du}}_{N^{\mathrm{du}}}(\mu)$: the output error (and output error bound) vanishes as the *product* of the primal and dual error (bounds), and hence much more rapidly than either the primal or dual error. From the perspective of computational efficiency, a good choice is $\Delta_N(\mu) \approx \Delta^{\mathrm{du}}_{N^{\mathrm{du}}}(\mu)$; the latter also ensures that the bound (6.83) will be quite sharp. However, the output effectivity can still be rather large because the "correlation" between the primal and dual errors is not captured into the output error bound which is constructed by using either the Cauchy-Schwarz inequality for $a(u(\mu) - u_N(\mu), \psi(\mu) - \psi_{N^{\mathrm{du}}})$ or the Riesz representation for $r^{\mathrm{pr}}(\psi(\mu) - \psi_{N^{\mathrm{du}}}(\mu); g_{M^g}(x;\mu))$.

Offline-Online Computational Procedure

We consider only the primal residual; the dual residual admits a similar treatment. To begin, we note from standard duality arguments that

$$\|r^{\mathrm{pr}}(v; g_{M^g}(x;\mu))\|_{X'} = \|\hat{e}(\mu)\|_X \ , \tag{6.87}$$

where $\hat{e}(\mu) \in X$ satisfies

$$(\hat{e}(\mu), v)_X = r^{\mathrm{pr}}(v; g_{M^g}(x;\mu)), \quad \forall\, v \in X \ . \tag{6.88}$$

We next observe from our reduced-basis representation (6.73) and affine assumption (6.66) that $r^{\rm pr}(v; g_{M^g}(x;\mu))$ may be expressed as

$$r^{\rm pr}(v; g_{M^g}(x;\mu)) = \sum_{m=1}^{M^g} \varphi^g_{M^g\,m}(\mu) f(v; q^g_m) - \sum_{q=1}^{Q}\sum_{n=1}^{N} u_{N\,n}(\mu)\Theta^q(\mu) a^q(\zeta_j, v), \quad \forall v \in X. \tag{6.89}$$

It thus follows from (6.88) and (6.89) that $\hat{e}(\mu) \in X$ satisfies

$$(\hat{e}(\mu), v)_X = \sum_{m=1}^{M^g} \varphi^g_{M^g\,m}(\mu) f(v; q^g_m) - \sum_{q=1}^{Q}\sum_{n=1}^{N} u_{N\,n}(\mu)\Theta^q(\mu) a^q(\zeta_j, v), \quad \forall v \in X. \tag{6.90}$$

The critical observation is that the right-hand side of (6.90) is a sum of products of parameter-dependent functions and parameter-independent linear functionals. In particular, it follows from linear superposition that we may write $\hat{e}(\mu) \in X$ as

$$\hat{e}(\mu) = \sum_{m=1}^{M^g} \varphi^g_{M^g\,m}(\mu)\, \mathcal{C}_m + \sum_{q=1}^{Q}\sum_{n=1}^{N} u_{N\,n}(\mu)\Theta^q(\mu)\, \mathcal{L}^q_n \ ,$$

for $\mathcal{C}_m \in X$ satisfying $(\mathcal{C}_m, v)_X = f(v; q^g_m)$, $\forall\, v \in X$, $1 \le m \le M^g$, and $\mathcal{L}^q_n \in X$ satisfying $(\mathcal{L}^q_n, v)_X = -a^q(\zeta_n, v)$, $\forall\, v \in X$, $1 \le n \le N$, $1 \le q \le Q$. It thus follows that

$$\begin{aligned}
\|\hat{e}(\mu)\|^2_X \;=\; & \sum_{m=1}^{M^g}\sum_{m'=1}^{M^g} \varphi^g_{M^g\,m}(\mu)\overline{\varphi}^g_{M^g\,m'}(\mu)(\mathcal{C}_m, \mathcal{C}_{m'})_X + \sum_{q=1}^{Q}\sum_{n=1}^{N} \Theta^q(\mu) u_{N\,n}(\mu) \times \\
& \left\{ \sum_{m=1}^{M^g} \overline{\varphi}^g_{M^g\,m}(\mu)(\mathcal{L}^q_n, \mathcal{C}_m)_X + \sum_{q'=1}^{Q}\sum_{n'=1}^{N} \overline{\Theta}^{q'}(\mu)\overline{u}_{N\,n'}(\mu)(\mathcal{L}^q_n, \mathcal{L}^{q'}_{n'})_X \right\} \\
& + \sum_{q=1}^{Q}\sum_{n=1}^{N}\sum_{m=1}^{M^g} \overline{\Theta}^q(\mu)\overline{u}_{N\,n}(\mu)\varphi^g_{M^g\,m}(\mu)(\mathcal{C}_m, \mathcal{L}^q_n)_X \ .
\end{aligned} \tag{6.91}$$

Furthermore, we invoke our reduced-basis representation (6.73) to write

$$\ell(\overline{u}_N(\mu); q^h_{M^h+1}) = \sum_{n=1}^{N} u_{N\,n}(\mu)\ell(\overline{\zeta}_n; q^h_{M^h+1}) \ , \tag{6.92}$$

$$f(\psi_{N^{\rm du}}(\mu); q^g_{M^g+1}) = \sum_{n=1}^{N^{\rm du}} \overline{\psi}_{N^{\rm du}\,n}(\mu) f(\zeta^{\rm du}_n; q^g_{M^g+1}) \ . \tag{6.93}$$

Note that $\ell(\overline{\zeta}_n; q^h_{M^h+1})$ and $f(\zeta^{\rm du}_n; q^g_{M^g+1})$ have been already formed and stored in the previous section. The offline-online decomposition may now be identified. For simplicity below we assume that $N^{\rm du} = N$ and $M^h = M^g \equiv M$.

In the offline stage — performed once — we first solve for $\mathcal{C}_m$, $1 \le m \le M$, and $\mathcal{L}^q_n$, $1 \le n \le N$, $1 \le q \le Q$; we then evaluate and save the relevant parameter-independent inner products $(\mathcal{C}_m, \mathcal{C}_{m'})_X$, $(\mathcal{C}_m, \mathcal{L}^q_n)_X$, $(\mathcal{L}^q_n, \mathcal{L}^{q'}_{n'})_X$, $1 \le m, m' \le M$, $1 \le n, n' \le N$, $1 \le q, q' \le Q$. This requires $M + QN$ finite element solutions and $M^2 + MQN + Q^2N^2$ finite-element inner products. Note that all quantities computed in the offline stage are independent of the parameter μ.

In the online stage — performed many times, for each new value of μ "in the field" — we simply evaluate the sums (6.91), (6.92), and (6.93) in terms of the $\Theta^q(\mu), u_{N\,n}(\mu)$ and the precalculated and stored (parameter-independent) $(\cdot,\cdot)_X$ inner products. The operation count for the online stage is only $O(M^2 + MQN + Q^2N^2)$ — again, the essential point is that the online complexity is *independent of* $\mathcal{N}$, the dimension of the underlying truth finite element approximation space. We further note that, unless M and Q are quite large, the online cost associated with the calculation of the dual norm of the residual is commensurate with the online cost associated with the calculation of $s_N(\mu)$.

6.6.5 A Forward Scattering Problem

We apply the above primal-dual reduced-basis formulation for the forward scattering problem described in Section 10.3. We briefly mention that the problem in the original domain $\tilde{\Omega}$ is reformulated in terms of a reference domain Ω corresponding to the geometry bounded by a unit circle ∂D and a square $\Gamma \equiv [-5,5] \times [-5,5]$ as shown in Figure 6-5; and that the mapped problem can be cast in the desired form (6.60)-(6.61) in which $\mu = (\mu_1, \mu_2, \mu_3, \mu_4, \mu_5, \mu_6) \in \mathcal{D} \equiv [\pi/16, 3\pi/16] \times [1/3, 1] \times [0, \pi] \times [0, 2\pi] \times [0, 2\pi] \times [\pi/8, \pi/8] \subset \mathbb{R}^6$, the bilinear form is affine for $Q = 5$ as shown in Table 6.4, and the force and output functionals are given below

$$f(v; g(x;\mu)) = -\int_{\partial D} \bar{v} g(x;\mu), \qquad \ell(v; h(x;\mu)) = -\frac{\rm i}{4}\sqrt{\frac{2}{\pi k}} e^{-{\rm i}\pi/4} \int_{\partial D} \overline{v} h(x;\mu)\ , \quad (6.94)$$

$$g(x;\mu) = \mathrm{i}\mu_1\left((\mu_2 x_1\cos\mu_3 - x_2\sin\mu_3)\cos\mu_4 + (\mu_2 x_1\sin\mu_3 + x_2\cos\mu_3)\sin\mu_4\right)\times e^{\mathrm{i}\mu_1((x_1\cos\mu_3-\mu_2 x_2\sin\mu_3)\cos\mu_4+(x_1\sin\mu_3+\mu_2 x_2\cos\mu_3)\sin\mu_4)},$$

$$h(x;\mu) = \mathrm{i}\mu_1\left((\mu_2 x_1\cos\mu_3 - x_2\sin\mu_3)\cos\mu_5 + (\mu_2 x_1\sin\mu_3 + x_2\cos\mu_3)\sin\mu_5\right)\times e^{-\mathrm{i}\mu_1((x_1\cos\mu_3-\mu_2 x_2\sin\mu_3)\cos\mu_5+(x_1\sin\mu_3+\mu_2 x_2\cos\mu_3)\sin\mu_5)}.$$

q	$\Theta^q(\mu)$	$a^q(w,v)$
1	μ_2	$\int_\Omega \frac{\partial w}{\partial x_1}\frac{\partial \bar{v}}{\partial x_1}$
2	$\frac{1}{\mu_2}$	$\int_\Omega \frac{\partial w}{\partial x_2}\frac{\partial \bar{v}}{\partial x_2}$
3	$-\mu_1^2\mu_2$	$\int_\Omega w\bar{v}$
4	$-\mathrm{i}\mu_1$	$\int_{\Gamma_1} w\bar{v} + \int_{\Gamma_3} w\bar{v}$
5	$-\mathrm{i}\mu_1\mu_2$	$\int_{\Gamma_2} w\bar{v} + \int_{\Gamma_4} w\bar{v}$

Table 6.4: Parametric functions $\Theta^q(\mu)$ and parameter-independent bilinear forms $a^q(w,v)$ for the forward scattering problem.

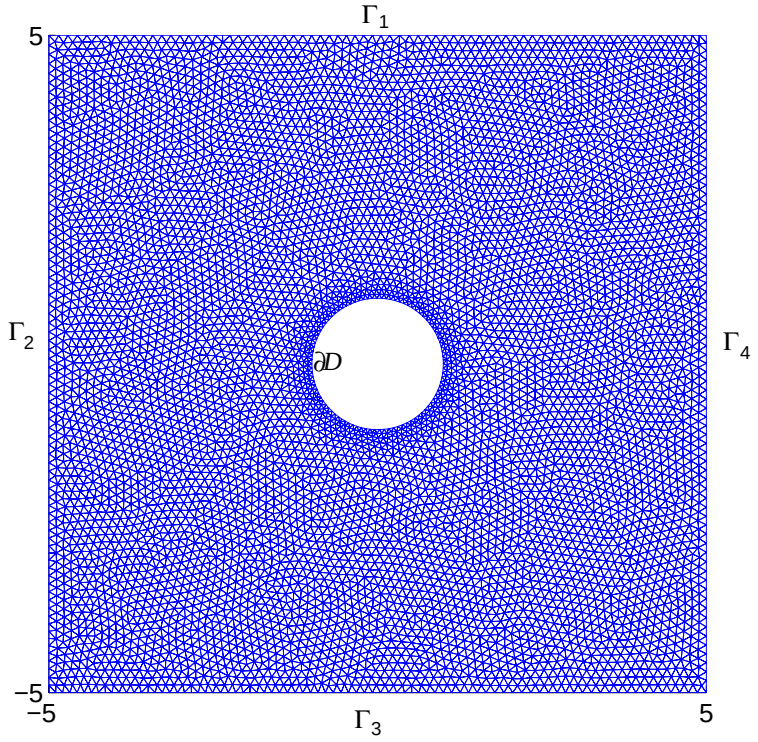


Figure 6-5: Linear triangular finite element mesh on the reference domain.

It should be noted that the forward scattering problem is a complex boundary value

problem, our piecewise-linear finite element approximation space of dimension $\mathcal{N} = 6{,}863$ is thus complexified such that

$$X = \left\{ v = v^{\mathrm{R}} + \mathrm{i}v^{\mathrm{I}} \in X^{\mathrm{e}} \mid v^{\mathrm{R}}|_{T_h} \in \mathbb{P}_1(T_h),\ v^{\mathrm{I}}|_{T_h} \in \mathbb{P}_1(T_h),\ \forall\, T_h \in \mathcal{T}_h \right\} \ , \tag{6.95}$$

where X^{e} is a complex function space defined as

$$X^{\mathrm{e}} = \{ v = v^{\mathrm{R}} + \mathrm{i}v^{\mathrm{I}} \mid v^{\mathrm{R}} \in H^1(\Omega),\ v^{\mathrm{I}} \in H^1(\Omega) \} \ . \tag{6.96}$$

The associated inner product is defined as

$$(w, v)_X = \int_\Omega \nabla w \nabla \bar{v} + w\bar{v} \ . \tag{6.97}$$

Here R and I denote the real and imaginary part, respectively; and $\bar{v}$ denotes the complex conjugate of v, and $|v|$ the modulus of v. Figure 6-5 shows the triangulation $\mathcal{T}_h$.

6.6.6 Numerical Results

We pursue the empirical interpolation procedure described in Section 6.2 to construct $S^g_{M^g}$, $W^g_{M^g}$, $T^g_{M^g}$, $1 \le M^g \le M^g_{\max}$, for $M^g_{\max} = 21$, and $S^h_{M^h}$, $W^h_{M^h}$, $T^h_{M^h}$, $1 \le M^h \le M^h_{\max}$, for $M^h_{\max} = 21$. We present in Table 6.5 $\varepsilon^g_{M^g,\max}$ for different values of M^g and $\varepsilon^h_{M^h,\max}$ for different values of M^h, where $\varepsilon^g_{M^g,\max} = \max_{\mu \in \Xi^g_{\mathrm{Test}}} \varepsilon^g_{M^g}(\mu)$ and $\varepsilon^h_{M^h,\max} = \max_{\mu \in \Xi^h_{\mathrm{Test}}} \varepsilon^h_{M^h}(\mu)$, and $\Xi^g_{\mathrm{Test}} = \Xi^h_{\mathrm{Test}} \subset (\mathcal{D})^{256}$ is a regular parameter grid of size 256. We observe that the coefficient function approximations converge very rapidly. This is expected as both the functions $g(\cdot;\mu)$ and $h(\cdot;\mu)$ defined on ∂D are smooth and regular in the parameter μ.

$M^g = M^h$	$\varepsilon^g_{M^g,\max}$	$\varepsilon^h_{M^h,\max}$
4	2.65×10^{-01}	4.32×10^{-01}
8	2.18×10^{-02}	2.19×10^{-02}
12	3.42×10^{-04}	3.53×10^{-04}
16	1.34×10^{-05}	3.37×10^{-05}
20	8.73×10^{-07}	1.36×10^{-06}

Table 6.5: $\varepsilon^g_{M^g,\max}$ as a function of M^g and $\varepsilon^h_{M^h,\max}$ as a function of M^h.

We next consider the piecewise-constant construction for the inf-sup lower bounds. For this purpose, we choose $(w;v)_X = \int_\Omega \nabla w \cdot \nabla \bar{v} + w\bar{v}$, $|v|_q^2 = a^q(v,v)$, $1 \le q \le q$, since the $a^q(\cdot,\cdot)$ are positive semi-definite; it thus follows from the Cauchy-Schwarz inequality that $\Gamma^q = 1$, $1 \le q \le Q$, and the numerically calculated $C_X = 1.0000$. We can cover the parameter space of the bilinear form a (for $\bar{\epsilon}_\beta = 0.5$) with $J = 20$ polytopes;[3] here the $\mathcal{P}^{\bar{\mu}_j}$, $1 \le j \le J$, are quadrilaterals such that $|\mathcal{V}^{\mu_j}| = 4$, $1 \le j \le J$. Armed with the inf-sup lower bounds, we can pursue the adaptive sampling strategy to arrive at $N_{\max} = N^{\mathrm{du}}_{\max} = 60$ for $n_{\mathrm{F}} = 1024$.

N	$\Delta_{N,\max,\mathrm{rel}}$	$\eta_{N,\mathrm{ave}}$	$\Delta^{\mathrm{du}}_{N^{\mathrm{du}},\max,\mathrm{rel}}$	$\eta^{\mathrm{du}}_{N^{\mathrm{du}},\mathrm{ave}}$	$\Delta^{s}_{N,\max,\mathrm{rel}}$	$\eta^{s}_{N,\mathrm{ave}}$
10	3.50×10^{-00}	29.15	2.40×10^{-00}	30.45	7.87×10^{-00}	47.49
20	1.83×10^{-00}	28.15	6.97×10^{-01}	30.88	6.67×10^{-01}	50.41
30	1.97×10^{-01}	25.68	4.85×10^{-01}	28.82	6.23×10^{-02}	92.96
40	8.54×10^{-02}	30.86	1.28×10^{-01}	29.24	1.07×10^{-02}	92.62
50	2.85×10^{-02}	27.92	3.98×10^{-02}	29.48	8.58×10^{-04}	99.33
60	1.52×10^{-03}	27.91	1.47×10^{-02}	29.36	1.90×10^{-04}	68.65

Table 6.6: Convergence and effectivities for the forward scattering problem obtained with $M^g = M^h = 20$.

We readily present basic numerical results and take $N^{\mathrm{du}} = N$ for this purpose. We show in Table 6.6 $\Delta_{N,\max,\mathrm{rel}}$, $\eta_{N,\mathrm{ave}}$,$\Delta^{\mathrm{du}}_{N^{\mathrm{du}},\max,\mathrm{rel}}$, $\eta^{\mathrm{du}}_{N^{\mathrm{du}},\mathrm{ave}}$, $\Delta^{s}_{N,\max,\mathrm{rel}}$, and $\eta^{s}_{N,\mathrm{ave}}$ as a function of N. Here $\Delta_{N,\max,\mathrm{rel}}$ is the maximum over Ξ_{Test} of $\Delta_N(\mu)/\|u(\mu)\|_X$, $\eta_{N,\mathrm{ave}}$ is the average over Ξ_{Test} of $\Delta_N(\mu)/\|u(\mu) - u_N(\mu)\|_X$, $\Delta^{\mathrm{du}}_{N^{\mathrm{du}},\max,\mathrm{rel}}$ is the maximum over Ξ_{Test} of $\Delta^{\mathrm{du}}_{N^{\mathrm{du}}}(\mu)/\|\psi(\mu)\|_X$, $\eta^{\mathrm{du}}_{N^{\mathrm{du}},\mathrm{ave}}$ is the average over Ξ_{Test} of $\Delta^{\mathrm{du}}_{N^{\mathrm{du}}}(\mu)/\|\psi(\mu) - \psi_{N^{\mathrm{du}}}(\mu)\|_X$, $\Delta^{s}_{N,\max,\mathrm{rel}}$ is the maximum over Ξ_{Test} of $\Delta^{s}_N(\mu)/|s(\mu) - s_N(\mu)|$, and $\eta^{s}_{N,\mathrm{ave}}$ is the average over Ξ_{Test} of $\Delta^{s}_N(\mu)/|s(\mu) - s_N(\mu)|$, where $\Xi_{\mathrm{Test}} \subset (\mathcal{D})^{256}$ is a regular parameter grid of size 256. We observe that the reduced-basis approximation converges very rapidly; that our error bounds are fairly sharp; and that the output error (and output error bound) vanishes as the product of the primal and dual error (bounds) since $\varepsilon^g_{M^g,\max}$ and $\varepsilon^h_{M^h,\max}$ are very small for $M^g = M^h = 20$. The output effectivity is quite large primarily due to the fact that correlation between the primal error and dual error is not captured into the output error bound. However, effectivities $O(100)$ are readily acceptable within the

[3] Although the problem has six-component parameter, $\mu = (\mu_{(1)}, \ldots, \mu_{(6)})$, but a depends only on $\mu_{(1)}$ and $\mu_{(2)}$; hence its parameter space is two-dimensional. Note further that no inf-sup correction is required since a is affine in the parameter.

reduced-basis context: thanks to the very rapid convergence rates, the "unnecessary" increase in N and $N^{\rm du}$ — to achieve a given error tolerance — is proportionately very small.

Next we look at the relative contribution of the rigorous and nonrigorous components to the error bounds $\Delta_N(\mu)$, $\Delta^{\rm du}_{N^{\rm du}}(\mu)$, $\Delta^s_N(\mu)$. We provide in Table 6.7 $\Delta_{N,\rm ave,n}/\Delta_{N,\rm ave}$, $\Delta^{\rm du}_{N^{\rm du},\rm ave,n}/\Delta^{\rm du}_{N^{\rm du},\rm ave}$, and $\Delta^s_{N,\rm ave,n}/\Delta^s_{N,\rm ave}$ as a function of N. Here $\Delta_{N,\rm ave}$ is the average over $\Xi_{\rm Test}$ of $\Delta_N(\mu)$; $\Delta_{N,\rm ave,n}$ is the average over $\Xi_{\rm Test}$ of $\frac{\hat\varepsilon^g_{M^g}}{\hat\beta(\mu)}\sup_{v\in X}[f(v;q^g_{M^g+1})/\|v\|_X]$; $\Delta^{\rm du}_{N^{\rm du},\rm ave}$ is the average over $\Xi_{\rm Test}$ of $\Delta^{\rm du}_{N^{\rm du}}(\mu)$; $\Delta^{\rm du}_{N^{\rm du},\rm ave,n}$ is the average over $\Xi_{\rm Test}$ of $\frac{\hat\varepsilon^h_{M^h}}{\hat\beta(\mu)}\sup_{v\in X}[\ell(v;q^h_{M^h+1})/\|v\|_X]$; $\Delta^s_{N,\rm ave}$ is the average over $\Xi_{\rm Test}$ of $\Delta^s_N(\mu)$; $\Delta^s_{N,\rm ave,n}$ is the average over $\Xi_{\rm Test}$ of $\Delta^s_{N,\rm n}$. As expected, the ratios increase with N, but still much less than unity; and thus, in the error bounds, the rigorous components strongly dominate the nonrigorous components.

N	$\Delta_{N,\rm ave,n}/\Delta_{N,\rm ave}$	$\Delta^{\rm du}_{N^{\rm du},\rm ave,n}/\Delta^{\rm du}_{N^{\rm du},\rm ave}$	$\Delta^s_{N,\rm ave,n}/\Delta^s_{N,\rm ave}$
10	1.34×10^{-06}	1.44×10^{-06}	1.84×10^{-06}
20	4.69×10^{-06}	6.09×10^{-06}	1.28×10^{-05}
30	1.73×10^{-05}	1.78×10^{-05}	8.46×10^{-05}
40	5.96×10^{-05}	5.29×10^{-05}	7.28×10^{-04}
50	1.41×10^{-04}	1.32×10^{-04}	4.07×10^{-03}
60	3.46×10^{-04}	3.21×10^{-04}	2.34×10^{-02}

Table 6.7: Relative contribution of the non-rigorous components to the error bounds as a function of N for $M^g = M^h = 20$.

Turning now to computational effort, for (say) $N = 30$ and any given μ (say, $a = b = 1$, $\alpha = 0$, $k = \pi/8$, $\tilde d = (1,0)$, $\tilde d^s = (1,0)$) — for which the error in the reduced-basis output $s_N(\mu)$ relative to the truth approximation $s(\mu)$ is *certifiably* less than $\Delta^s_N(\mu)$ $(= 2.29\times10^{-5})$ — the Online Time (marginal cost) to compute both $s_N(\mu)$ and $\Delta^s_N(\mu)$ is less than 1/122 the Total Time to directly calculate the truth result $s(\mu) = \ell(u(\mu))$. Clearly, the savings will be even larger for problems with more complex geometry and solution structure in particular in three space dimensions. Nevertheless, even for our current very modest example, the computational economies are very significant.

Chapter 7

An Empirical Interpolation Method for Nonlinear Elliptic Problems

In this chapter, we extend the technique developed in Chapter 6 to nonlinear elliptic problems in which g is a nonaffine *nonlinear* function of the parameter μ, spatial coordinate x, *and* field variable u — we hence treat certain classes of nonlinear problems. The nonlinear dependence of g on u introduces new numerical difficulties (and a new opportunity) for our approach: first, our greedy choice of basis functions ensures good approximation properties, but it is quite expensive in the nonlinear case; second, since u is not known in advance, it is difficult to generate an explicitly affine approximation for $g(u; x; \mu)$; and third, it is challenging to ensure that the online complexity remains independent of $\mathcal{N}$ even in the presence of highly nonlinear terms. We shall address most of these concerns in this chapter and leave some for future research.

Our approach to nonlinear elliptic problems is based on the ideas described in Chapter 6: we first apply the empirical interpolation method to build a collateral reduced-basis expansion for $g(u; x; \mu)$; we then approximate $g(u_{N,M}(x;\mu); x; \mu)$ — as required in our reduced-basis projection for $u_{N,M}(\mu)$ — by $g_M^{u_{N,M}}(x;\mu) = \sum_{m=1}^{M} \varphi_{M\,m}(\mu) q_m(x)$; we finally construct an efficient offline-online computational procedure to rapidly evaluate the reduced-basis approximation $u_{N,M}(\mu)$ and $s_{N,M}(\mu)$ to $u(\mu)$ and $s(\mu)$ and associated *a posteriori* error bounds $\Delta_{N,M}(\mu)$ and $\Delta^s_{N,M}(\mu)$.

7.1 Abstraction

7.1.1 Weak Statement

Of course, nonlinear equations do not admit the same degree of generality as linear equations. We thus present our approach to nonlinear equations for a particular nonlinear problem. In particular, we consider the following "exact" (superscript e) problem: for any $\mu \in \mathcal{D} \subset \mathbb{R}^P$, find $s^{\rm e}(\mu) = \ell(u^{\rm e}(\mu))$, where $u^{\rm e}(\mu) \in X^{\rm e}$ satisfies the weak form of the μ-parametrized nonlinear PDE

$$a^L(u^{\rm e}(\mu), v) + \int_\Omega g(u^{\rm e}; x; \mu)v = f(v), \ \forall\, v \in X^{\rm e}. \tag{7.1}$$

Here $g(u^{\rm e}; x; \mu)$ is a general nonaffine nonlinear function of the parameter μ, spatial coordinate x, and field variable $u^{\rm e}(x; \mu)$; $a^L(\cdot, \cdot)$ and $f(\cdot), \ell(\cdot)$ are $X^{\rm e}$-continuous bounded bilinear and linear functionals, respectively; these forms are assumed to be parameter-independent for the sake of simplicity.

We next introduce $X \subset X^{\rm e}$, a reference finite element approximation space of dimension $\mathcal{N}$. The truth finite element approximation is then found by (say) Galerkin projection: Given $\mu \in \mathcal{D} \in \mathbb{R}^P$, we evaluate

$$s(\mu) = \ell(u(\mu)) \tag{7.2}$$

where $u(\mu) \in X$ is the solution of the discretized weak formulation

$$a^L(u(\mu), v) + \int_\Omega g(u; x; \mu)v = f(v), \quad \forall v \in X \ . \tag{7.3}$$

We assume that $\|u^{\rm e}(\mu) - u(\mu)\|_X$ is suitably small and hence that $\mathcal{N}$ will typically be very large.

We shall make the following assumptions. First, we assume that the bilinear form $a^L(\cdot, \cdot) : X \times X \to \mathbb{R}$ is symmetric, $a^L(w, v) = a^L(v, w), \forall\, w, v \in X$. We shall also make two crucial hypotheses related to well-posedness. Our first hypothesis is that the bilinear

form a^L satisfies a stability and continuity condition

$$0 < \alpha \equiv \inf_{v \in X} \frac{a^L(v,v)}{\|v\|_X^2} \; ; \tag{7.4}$$

$$\sup_{v \in X} \frac{a^L(v,v)}{\|v\|_X^2} \equiv \gamma < \infty \; . \tag{7.5}$$

For the second hypothesis we require that g be a monotonically increasing function of its first argument and be of such nonlinearity that the equation (7.3) is well-posed and sufficiently stable.

Finally, we note that under the above assumptions if solution of the problem (7.3) exists then it is unique: Suppose that (7.3) has two solution, u_1 and u_2, this implies

$$a^L(u_1 - u_2, v) + \int_\Omega \left(g(u_1; x; \mu) - g(u_2; x; \mu)\right) v = 0, \quad \forall\, v \in X \; .$$

By choosing $v = u_1 - u_2$, we obtain

$$a^L(u_1 - u_2, u_1 - u_2) + \int_\Omega \left(g(u_1; x; \mu) - g(u_2; x; \mu)\right) (u_1 - u_2) = 0, \quad \forall\, v \in X \; ;$$

it thus follows from the coercivity of a^L and monotonicity of g that $u_1 = u_2$. For proof of existence, we refer to [52].

7.1.2 A Model Problem

We consider the following model problem $-\nabla^2 u + \mu_{(1)} \frac{e^{\mu_{(2)} u} - 1}{\mu_{(2)}} = 10^2 \sin(2\pi x_{(1)}) \cos(2\pi x_{(2)})$ in a domain $\Omega =]0,1[^2$ with a homogeneous Dirichlet condition on boundary $\partial\Omega$, where $\mu = (\mu_{(1)}, \mu_{(2)}) \in \mathcal{D}^\mu \equiv [0.01, 10]^2$. The output of interest is the average of the product of field variable and force over the physical domain. The weak formulation is then stated as: given $\mu \in \mathcal{D}^\mu$, find $s(\mu) = \int_\Omega f(x) u(\mu)$, where $u(\mu) \in X = H_0^1(\Omega) \equiv \{v \in H_1(\Omega) \mid v|_{\partial\Omega} = 0\}$ is the solution of

$$\int_\Omega \nabla u \cdot \nabla v + \int_\Omega \mu_{(1)} \frac{e^{\mu_{(2)} u} - 1}{\mu_{(2)}} v = 100 \int_\Omega \sin(2\pi x_{(1)}) \, \cos(2\pi x_{(2)}) \, v, \quad \forall v \in X \; . \tag{7.6}$$

Our abstract statement (7.2) and (7.3) is then obtained for

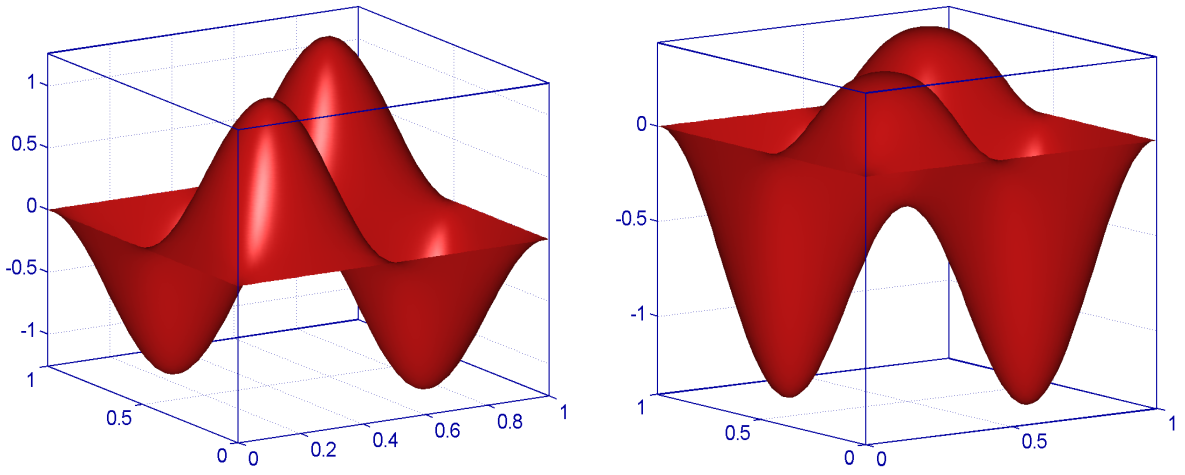
$$a^L(w,v) = \int_\Omega \nabla w \cdot \nabla v, \quad f(v) = 100 \int_\Omega \sin(2\pi x_{(1)})\, \cos(2\pi x_{(2)})\, v, \quad \ell(v) = \int_\Omega v, \quad (7.7)$$

and

$$g(u;\mu) = \mu_{(1)} \frac{e^{\mu_{(2)} u} - 1}{\mu_{(2)}} \ . \quad (7.8)$$

Our model problem is well-posed as proven in [52]. Note also that $\mu_{(1)}$ controls the strength of the sink term and $\mu_{(2)}$ controls the strength of the nonlinearity.

We give in Figure 7-1 two typical solutions obtained with a piecewise-linear finite element approximation space X of dimension $\mathcal{N} = 2601$. We see for $\mu = (0.01, 0.01)$ that the solution has two negative peaks and two positive peaks with the same height (this solution is very similar to that of the linear problem in which $g(u;\mu)$ is absent). However, due to the exponential nonlinearity, as μ increases the negative peaks remain largely unchanged while the positive peaks get rectified as shown in Figure 7-1(b) for $\mu = (10, 10)$. This is because the exponential function $\mu_{(1)} e^{\mu_{(2)} u}$ in $g(u;\mu)$ sinks the positive part of $u(\mu)$, but has no effect on the negative part of $u(\mu)$ as μ increases.



(a) (b)

Figure 7-1: Numerical solutions at typical parameter points: (a) $\mu = (0.01, 0.01)$ and (b) $\mu = (10, 10)$.

7.2 Coefficient–Approximation Procedure

Given a continuous non-affine nonlinear function $g(u;x;\mu) \in L^\infty(\Omega)$ of sufficient regularity, we seek to approximate $g(w;x;\mu)$ for any given $w \in X$ by a collateral reduced-basis expansion $g_M^w(x;\mu)$ of an approximation space W_M^g spanned by basis functions at M selected points in the parameter space. Specifically, we choose μ_1^g, and define $S_1^g = \{\mu_1^g\}$, $\xi_1 \equiv g(u;x;\mu_1^g)$, and $W_1^g = \text{span}\,\{\xi_1\}$; we assume that $\xi_1 \neq 0$. Then, for $M \geq 2$, we set $\mu_M^g = \arg\max_{\mu\in\Xi^g} \inf_{z\in W_{M-1}^g} \|g(\,\cdot;\cdot\,;\mu) - z\|_{L^\infty(\Omega)}$, where Ξ^g is a suitably fine parameter sample over $\mathcal{D}$ of size J^g. We then set $S_M^g = S_{M-1}^g \cup \mu_M^g$, $\xi_M = g(u;x;\mu_M^g)$, and $W_M^g = \text{span}\,\{\xi_m,\ 1 \leq m \leq M\}$ for $M \leq M_{\max}$. Note that since w is in the finite element approximation space X, $g(w;x;\mu)$ is really the interpolant of $g(w^{\rm e};x;\mu), w^{\rm e} \in X^{\rm e}$, on the finite element "truth" mesh.

Next, we construct nested sets of interpolation points $T_M = \{t_1,\ldots,t_M\}$, $1 \leq M \leq M_{\max}$. We first set $t_1 = \arg\,\text{ess sup}_{x\in\Omega} |\xi_1(x)|$, $q_1 = \xi_1(x)/\xi_1(t_1)$, $B_{11}^1 = 1$. Then for $M = 2,\ldots,M_{\max}$, we solve the linear system $\sum_{j=1}^{M-1} \sigma_j^{M-1} q_j(t_i) = \xi_M(t_i)$, $1 \leq i \leq M-1$, and set $r_M(x) = \xi_M(x) - \sum_{j=1}^{M-1} \sigma_j^{M-1} q_j(x)$, $t_M = \arg\,\text{ess sup}_{x\in\Omega} |r_M(x)|$, $q_M(x) = r_M(x)/r_M(t_M)$, and $B_{ij}^M = q_j(t_i)$, $1 \leq i,j \leq M$.

Finally, our coefficient-function approximation to $g(w;x;\mu)$ is the interpolant of g over T_M as defined from Lemma 6.2.3: $g_M^w(x;\mu) = \sum_{m=1}^M \varphi_{M\,m}(\mu) q_m(x)$, where $\varphi_M \in \mathbb{R}^M$ is found from $\sum_{j=1}^M B_{ij}^M \varphi_{M\,j}(\mu) = g(w(t_i);t_i;\mu)$, $1 \leq i \leq M$. Moreover, we define the interpolation error as $\varepsilon_M(\mu) \equiv \|g(w,x;\mu) - g_M^w(x;\mu)\|_{L^\infty(\Omega)}$ and calculate the associated error estimator as $\hat{\varepsilon}_M(\mu) \equiv |g(w,t_{M+1};\mu) - g_M^w(t_{M+1};\mu)|$.

7.3 Reduced-Basis Approximation

7.3.1 Discrete Equations

We first motivate the need for incorporating the empirical interpolation procedure into the reduced-basis method to treat nonlinear equations. The most significant numerical difficulty here is in finding an efficient representation of the nonlinear terms. To understand the implications, we consider a Galerkin projection directly on the nonlinear equation (7.3). Towards this end, we introduce nested samples, $S_N = \{\mu_1^u \in$

$\mathcal{D}, \cdots, \mu_N^u \in \mathcal{D}\}, 1 \le N \le N_{\max}$ and associated nested Lagrangian reduced-basis spaces as $W_N = \text{span}\{\zeta_j \equiv u(\mu_j^u), 1 \le j \le N\}, 1 \le N \le N_{\max}$, where $u(\mu_j^u)$ is the solution to (7.3) for $\mu = \mu_j^u$. Our reduced-basis approximation is then: for a given $\mu \in \mathcal{D}$, we evaluate $s_N(\mu) = \ell(u_N(\mu))$, where $u_N(\mu) \in W_N$ satisfies

$$a^L(u_N(\mu), v) + \int_\Omega g(u_N(\mu); x; \mu)v = f(v), \quad \forall v \in W_N. \tag{7.9}$$

To obtain $u_N(\mu)$ and $s_N(\mu) = \ell(u_N(\mu))$, we may apply a Newton iterative scheme: given a current iterate $\bar{u}_N(\mu) = \sum_{j=1}^N \bar{u}_{N\,j}(\mu)\zeta_j$, find an increment $\delta u_N \in W_N$ such that

$$a^L(\delta u_N, v) + \int_\Omega g_1(\bar{u}_N(\mu); x; \mu)\delta u_N v = r(v; g(\bar{u}_N(\mu); x; \mu)), \quad \forall\, v \in W_N\ ; \tag{7.10}$$

here $r(v; g(\bar{u}_N(\mu); x; \mu)) = f(v) - a^L(\bar{u}_N(\mu), v) + \int_\Omega g(\bar{u}_N(\mu); x; \mu)v, \forall\, v \in X$, is the usual residual; and g_u is the partial derivative with respect to its first argument.

The associated algebraic system is thus

$$\begin{aligned}\sum_{j=1}^N \Big\{a^L(\zeta_j, \zeta_i) + \int_\Omega g_1\Big(\sum_{n=1}^N \bar{u}_{N\,n}(\mu)\zeta_n; x; \mu\Big)\zeta_j\zeta_i\Big\}\delta u_{N\,j} \\ = f(\zeta_i) - a^L\Big(\sum_{n=1}^N \bar{u}_{N\,n}(\mu)\zeta_n, \zeta_i\Big) - \int_\Omega g\Big(\sum_{n=1}^N \bar{u}_{N\,n}(\mu)\zeta_n; x; \mu\Big)\zeta_i, \quad 1 \le i \le N.\end{aligned} \tag{7.11}$$

Observe that if g is a low-order (at most quadratically) polynomial nonlinearity of u, we can then develop an efficient offline-online procedure by resolving the nonlinear terms, $g\big(\sum_{n=1}^N \bar{u}_{N\,n}(\mu)\zeta_n; x; \mu\big)$ and $g_1\big(\sum_{n=1}^N \bar{u}_{N\,n}(\mu)\zeta_n; x; \mu\big)$, into their power series. Unfortunately, this useful trick can not be applied to high-order polynomial and non-polynomial nonlinearities; and hence the quantities $\int_\Omega g_1\Big(\sum_{n=1}^N \bar{u}_{N\,n}(\mu)\zeta_n; x; \mu\Big)\zeta_j\zeta_i$ and $\int_\Omega g\big(\sum_{n=1}^N \bar{u}_{N\,n}(\mu)\zeta_n; x; \mu\big)\zeta_i$ must be evaluated *online* at every Newton iteration with $\mathcal{N}$-dependent cost. The operation count for the on-line stage will thus scale in some order with $\mathcal{N}$ — the dimension of the truth finite element approximation space: the computational advantage relative to classical approaches using advanced iterative techniques is not obvious; and, in any event, real-time response may not be guaranteed.

In the reduced-basis approach, we seek an online evaluation cost that depends only

on the dimension of reduced-basis approximation spaces and the parametric complexity of the problems, *not* on $\mathcal{N}$. To achieve this goal, we develop a collateral reduced-basis expansion for the nonlinear terms by using the empirical interpolation method. We henceforth construct nested samples $S_M^g = \{\mu_1^g \in \mathcal{D}, \cdots, \mu_M^g \in \mathcal{D}\}, 1 \le M \le M_{\max}$, associated nested approximation spaces $W_M^g = \text{span}\{\xi_m \equiv g(u(\mu_m^g); x; \mu_m^g), 1 \le m \le M\} = \text{span}\{q_1, \ldots, q_M\}, 1 \le M \le M_{\max}$, and nested sets of interpolation points $T_M = \{t_1, \ldots, t_M\}$, $1 \le M \le M_{\max}$ following the procedure of Section 7.2. Then for any given $w \in X$ and M, we may approximate $g(w; x; \mu)$ by $g_M^w(x; \mu) = \sum_{m=1}^M \varphi_{M\,m}(\mu) q_m(x)$, where $\sum_{j=1}^M B_{i\,j}^M \varphi_{M\,j}(\mu) = g(w(t_i); t_i; \mu)$, $1 \le i \le M$; note that although this "composed" interpolant is defined for general $w \in X$, we expect good approximation only for w (very) close to the manifold $\mathcal{M}^u \equiv \{u(\mu) | \mu \in \mathcal{D}\}$ as which W_M^g is constructed. The construction of W_M^g requires solutions of the underlying nonlinear PDE (7.3) at all parameter points in Ξ^g — the associated computational cost may be very high. (In contrast, in the linear nonaffine case, g was only a function of the spatial coordinate x and the parameter μ; the construction of W_M^g involved only the evaluation of the function $g(x; \mu)$, $\forall \mu \in \mathcal{D}$, and *not* the solution of the PDE itself.) Hence, at present our approach to nonlinear problems is not very effective in high-dimensional parameter spaces.

We may now approximate $g(u_{N,M}; x; \mu)$ — as required in our reduced-basis projection for $u_{N,M}(\mu)$ — by $g_M^{u_{N,M}}(x; \mu)$. Our reduced-basis approximation is thus: Given $\mu \in \mathcal{D}$, we evaluate

$$s_{N,M}(\mu) = \ell(u_{N,M}(\mu)) \tag{7.12}$$

where $u_{N,M}(\mu) \in W_N$ satisfies

$$a^L(u_{N,M}(\mu), v; \mu) + \int_\Omega g_M^{u_{N,M}}(x; \mu) v = f(v), \quad \forall\, v \in W_N\ . \tag{7.13}$$

The parameter sample S_N and associated reduced-basis space W_N are constructed using the adaptive sampling procedure described in Section 6.5.3.

7.3.2 Offline/Online Computational Procedure

The most significant new issue is efficient calculation of the nonlinear term $g_M^{u_{N,M}}(x;\mu)$. Note that we can not directly solve (7.13) as in the linear case since $g_M^{u_{N,M}}(x;\mu)$ is, in fact, a nonlinear function of the $u_{N,M\,j}(\mu), 1 \le j \le N$. To see this more clearly, we expand our reduced-basis approximation and coefficient-function approximation as

$$u_{N,M}(\mu) = \sum_{j=1}^{N} u_{N,M\,j}(\mu)\zeta_j \ , \qquad g_M^{u_{N,M}}(x;\mu) = \sum_{m=1}^{M} \varphi_{M\,m}(\mu) q_m \ . \tag{7.14}$$

Inserting these representations into (7.13) yields

$$\sum_{j=1}^{N} A_{i\,j}^N u_{N,M\,j}(\mu) + \sum_{m=1}^{M} C_{i\,m}^{N,M} \varphi_{M\,m}(\mu) = F_{N\,i}, \quad 1 \le i \le N \ ; \tag{7.15}$$

where $A^N \in \mathbb{R}^{N\times N}, C^{N,M} \in \mathbb{R}^{N\times M}, F_N \in \mathbb{R}^N$ are given by $A_{i\,j}^N = a^L(\zeta_j,\zeta_i), 1 \le i,j \le N$, $C_{i\,m}^{N,M} = \int_\Omega q_m \zeta_i, 1 \le i \le N, 1 \le m \le M$, and $F_{N\,i} = f(\zeta_i), 1 \le i \le N$, respectively. Furthermore, $\varphi_M(\mu) \in \mathbb{R}^M$ is given by

$$\begin{aligned} \sum_{k=1}^{M} B_{m\,k}^M \varphi_{M\,k}(\mu) &= g(u_{N,M}(t_m;\mu);t_m;\mu), \quad 1 \le m \le M \\ &= g\Big(\sum_{n=1}^{N} u_{N,M\,n}(\mu)\zeta_n(t_m);t_m;\mu\Big), \quad 1 \le m \le M \ . \end{aligned} \tag{7.16}$$

We then substitute $\varphi_M(\mu)$ from (7.16) into (7.15) to obtain the following nonlinear algebraic system

$$\sum_{j=1}^{N} A_{i\,j}^N u_{N,M\,j}(\mu) + \sum_{m=1}^{M} D_{i\,m}^{N,M} g\Big(\sum_{n=1}^{N} \zeta_n(t_m) u_{N,M\,n}(\mu);t_m;\mu\Big) = F_{N\,i}, \quad 1 \le i \le N \ , \tag{7.17}$$

where $D^{N,M} = C^{N,M}(B^M)^{-1} \in \mathbb{R}^{N\times M}$.

To solve (7.17) for $u_{N,M\,j}(\mu), 1 \le j \le N$, we may again apply a Newton iterative scheme: given a current iterate $\bar{u}_{N,M\,j}(\mu), 1 \le j \le N$, we must find an increment

$\delta u_{N,M\,j}, 1 \le j \le N$, such that

$$\sum_{j=1}^{N} \big(A^N_{i\,j} + \bar{E}^N_{i\,j}\big)\delta u_{N,M\,j}(\mu) = F_{N\,i} - \sum_{j=1}^{N} A^N_{i\,j}\bar{u}_{N,M\,j}(\mu) - \sum_{m=1}^{M} D^{N,M}_{i\,m}\, g\Big(\sum_{n=1}^{N} \zeta_n(t_m)\bar{u}_{N,M\,n}(\mu); t_m; \mu\Big), \quad 1 \le i \le N\ ; \quad (7.18)$$

here $\bar{E}^N \in \mathbb{R}^{N\times N}$ must be calculated at every Newton iteration as

$$\bar{E}^N_{i\,j} = \sum_{m=1}^{M} D^{N,M}_{i\,m}\, g_1\Big(\sum_{n=1}^{N} \zeta_n(t_m)\bar{u}_{N,M\,n}(\mu); t_m; \mu\Big)\zeta_j(t_m), \quad 1 \le i, j \le N\ . \quad (7.19)$$

Finally, the output can be evaluated as

$$s_{N,M}(\mu) = \sum_{j=1}^{N} u_{N,M\,j}(\mu) L_{N\,j}\ , \quad (7.20)$$

where $L_N \in \mathbb{R}^N$ is the output vector with entries $L_{N\,j} = \ell(\zeta_j), 1 \le j \le N$. We observe that we can now develop an efficient offline-online procedure for the rapid evaluation of $s_{N,M}(\mu)$ for each μ in $\mathcal{D}$.

In the offline stage — performed once — we first generate nested reduced-basis approximation spaces $W_N = \{\zeta_1, \ldots, \zeta_N\}, 1 \le N \le N_{\max}$, nested approximation spaces $W^g_M = \{q_1, \ldots, q_M\}, 1 \le M \le M_{\max}$, and nested sets of interpolation points $T_M = \{t_1, \ldots, t_M\}$; we then form and store $A^N, B^M, D^{N,M}, F_N$, and L_N.

In the online stage — performed many times for each new μ — we solve (7.17) for $u_{N,M\,j}(\mu), 1 \le j \le N$ and evaluate $s_{N,M}(\mu)$ from (7.20). The operation count of the online stage is essentially the predominant Newton update component: at each Newton iteration, we first assemble the right-hand side and compute $\bar{E}^N$ at cost $O(MN^2)$ — note we perform the sum in the parenthesis of (7.19) before performing the outer sum; we then form and invert the left-hand side (Jacobian) at cost $O(N^3)$. The online complexity depends only on N, M, and number of Newton iterations; we thus recover online $\mathcal{N}$ independence.

7.3.3 Implementation Issues

At this point we need to comment on an important issue concerning the actual numerical implementation of our proposed method which, if not addressed properly, can lead to erroneous results. To begin, we consider a Newton iterative scheme to solve the truth finite element approximation (7.3) for $u(\mu)$: given a current iterate $\bar{u}(\mu)$, find an increment $\delta u(\mu) \in X$ such that

$$a^L(\delta u(\mu), v) + \int_\Omega g_u(\bar{u}(\mu); x; \mu)\delta u(\mu)v = f(v) - a^L(\bar{u}(\mu), v) + \int_\Omega g(\bar{u}(\mu); x; \mu)v, \ \forall v \in X \ . \tag{7.21}$$

We must then require the numerical integration of terms of the form $\int_\Omega g(\bar{u}; x; \mu)v$ (and $\int_\Omega g_u(\bar{u}(\mu); x; \mu)wv$) — which (usually) have to be evaluated by Gaussian quadrature:

$$\int_\Omega v \ g(\bar{u}(x;\mu); x; \mu) \approx \sum_{j=1}^{\mathcal{N}_{\rm QP}} \omega_j \, v(x_j^{\rm QP}) \ g(\bar{u}(x_j^{\rm QP}; \mu), x_j^{\rm QP}; \mu), \tag{7.22}$$

where the ω_j are the elemental Gauss-Legendre quadrature weights, $x_j^{\rm QP}$ are the corresponding elemental quadrature points, and $\mathcal{N}_{\rm QP}$ is the total number of quadrature points. Similarly, the reduced-basis approximation procedure requires, during the offline stage, the evaluation of (say) $\int_\Omega \zeta_i q_m$. For consistency, the term should be evaluated using the same quadrature rule that was used to develop the "truth" finite element approximation

$$\int_\Omega \zeta_i \ q_m \approx \sum_{j=1}^{\mathcal{N}_{\rm QP}} \omega_j \, \zeta_i(x_j^{\rm QP}) \ q_m(x_j^{\rm QP}); \tag{7.23}$$

absent this consistency, $u_{N,M}(\mu)$ will not converge to $u(\mu)$ as $N, M \to \infty$.

From the construction of the interpolation points t_i, $1 \le i \le M_{\max}$, we note that the q_m, $1 \le m \le M_{\max}$, can be written as a linear combination of the basis function $\xi_i = g(u; x; \mu_i^g)$, $1 \le i \le M_{\max}$, which we obtained from our greedy adaptive procedure in Section 7.2. It is easy to show that $\xi_i = T_{im} q_m$, $1 \le i, m \le M_{\max}$, where $T \in \mathbb{R}^{M_{\max} \times M_{\max}}$ is the corresponding transformation matrix . Unfortunately, it turns out that T is badly conditioned and the resulting q_m required in (7.23) susceptible to large round-off errors. To avoid this problem we thus have to follow a different route: while generating the basis

functions $\xi_i = g(u(\mu_i^g); x; \mu_i^g) \in \mathbb{R}^{\mathcal{N}}$, $1 \le i \le M_{\max}$, we also generate a corresponding set of functions ξ_i^{QP} evaluated at the quadrature points x_j^{QP}, $1 \le j \le \mathcal{N}_{\text{QP}}$, that is $\xi_i^{\text{QP}} = g(u(\mu_i^g); x_j^{\text{QP}}; \mu_i^g)$, $1 \le j \le \mathcal{N}_{\text{QP}}$, $1 \le i \le M_{\max}$. Next, we construct the set of interpolation points t_i and basis functions q_i from the ξ_i according to Section 7.2. During this procedure we also evaluate q_m^{QP} by starting with $q_1^{\text{QP}} = \xi_1^{\text{QP}}(x)/\xi_1^{\text{QP}}(t_1)$ and then setting $r_M^{\text{QP}}(x) = \xi_M^{\text{QP}}(x) - \sum_{i=1}^{M-1} \sigma_i^{M-1} q_i^{\text{QP}}(x)$, $q_M^{\text{QP}}(x) = r_M^{\text{QP}}(x)/r_M^{\text{QP}}(t_M)$, $2 \le M \le M_{\max}$, where the σ_i^{M-1} are determined during the construction of the q_i.

Note that q_m^{QP} is simply the "basis" function corresponding to q_m, but evaluated at the quadrature points. Given the q_m^{QP}, we can then directly evaluate the integral $\int_\Omega \zeta_i \, q_m$ by Gauss Quadrature

$$\int_\Omega \zeta_i \; q_m \approx \sum_{j=1}^{\mathcal{N}_{\text{QP}}} \; \omega_j \; \zeta_i(x_j^{\text{QP}}) \; q_m^{\text{QP}}(x_j^{\text{QP}}) \; . \tag{7.24}$$

Using this approach during the numerical implementation we can avoid the round-off errors that resulted from the conditioning problems of the transformation matrix T.

7.4 A Posteriori Error Estimation

7.4.1 Error Bounds

We first assume we are given α and $\hat{\varepsilon}_M(\mu)$ which are the coercivity parameter of a^L and the *a posteriori* error estimator for the error in our coefficient function approximation, respectively.[1] We may now introduce our error bound $\Delta_{N,M}(\mu)$ for $\|u(\mu) - u_{N,M}(\mu)\|_X$,

$$\Delta_{N,M}(\mu) = \frac{1}{\alpha} \left(\sup_{v \in X} \frac{r(v; g_M^{u_{N,M}}(x;\mu))}{\|v\|_X} + \hat{\varepsilon}_M(\mu) \sup_{v \in X} \frac{\int_\Omega q_{M+1} v}{\|v\|_X} \right) \; , \tag{7.25}$$

where $r(v; g_M^{u_{N,M}}(x;\mu))$, the residual associated with $u_{N,M}(\mu)$ and $g_M^{u_{N,M}}(x;\mu)$, is given by

$$r(v; g_M^{u_{N,M}}(x;\mu)) = f(v) - a^L(u_{N,M}(\mu), v) - \int_\Omega g_M^{u_{N,M}}(x;\mu) v, \quad \forall \, v \in X \; . \tag{7.26}$$

[1] Note if a^L is parameter-dependent we will then require $\hat{\alpha}(\mu)$ — a lower bound for $\alpha(\mu)$.

We next define the error bound for the error in the output as

$$\Delta^s_{N,M}(\mu) = \sup_{v\in X} \frac{\ell(v)}{\|v\|_X} \Delta_{N,M}(\mu) \ . \tag{7.27}$$

Proposition 12. *Suppose that $g(u_{N,M}(\mu); x; \mu) \in W^g_{M+1}$, we then have*

$$\|u(\mu) - u_{N,M}(\mu)\|_X \le \Delta_{N,M}(\mu), \quad |s(\mu) - s_{N,M}(\mu)| \le \Delta^s_{N,M}(\mu), \quad \forall\, \mu \in \mathcal{D} \ . \tag{7.28}$$

Proof. We first note from (7.3) and (7.26) that $e_{N,M}(\mu) \equiv u(\mu) - u_{N,M}(\mu)$ satisfies

$$a^L(e_{N,M}(\mu), v) + \int_\Omega \big(g(u(\mu); x; \mu) - g(u_{N,M}(\mu); x; \mu)\big)v = r(v; g_M^{u_{N,M}}(x;\mu)) + \int_\Omega \big(g_M^{u_{N,M}}(x;\mu) - g(u_{N,M}(\mu); x; \mu)\big)v, \quad \forall\, v \in X.$$

We next choose $v = e(\mu)$ and invoke the monotonicity of g to obtain

$$a^L(e(\mu), e(\mu)) \le r(e(\mu); g_M^{u_{N,M}}(x;\mu)) + \int_\Omega \big(g_M^{u_{N,M}}(x;\mu) - g(u_{N,M}(\mu); x; \mu)\big)e(\mu) \ . \tag{7.29}$$

It follows from (7.29), a-coercivity, and our assumption $g(u_{N,M}(\mu); x; \mu) \in W^g_{M+1}$ that

$$\begin{aligned}
\|e(\mu)\|_X &\le \frac{1}{\alpha}\left(\frac{r(e(\mu); g_M^{u_{N,M}}(x;\mu)) + \int_\Omega \big(g_M^{u_{N,M}}(x;\mu) - g(u_{N,M}(\mu); x; \mu)\big)e(\mu)}{\|e(\mu)\|_X}\right) \\
&\le \frac{1}{\alpha}\left(\sup_{v\in X} \frac{r(v; g_M^{u_{N,M}}(x;\mu))}{\|v\|_X} + \sup_{v\in X} \frac{\int_\Omega \big(g_M^{u_{N,M}}(x;\mu) - g(u_{N,M}(\mu); x; \mu)\big)v}{\|v\|_X}\right) \\
&= \frac{1}{\alpha}\left(\sup_{v\in X} \frac{r(v; g_M^{u_{N,M}}(x;\mu))}{\|v\|_X} + \hat{\varepsilon}_M(\mu) \sup_{v\in X} \frac{\int_\Omega q_{M+1}(x)v}{\|v\|_X}\right) \ .
\end{aligned}$$

Finally, it follows from the continuity of ℓ that $|s(\mu) - s_{N,M}(\mu)| \le \|\ell\|_{X'} \|e_{N,M}(\mu)\|_X \le \|\ell\|_{X'} \Delta_{N,M}(\mu)$. This concludes the proof. □

Our hypothesis $g(u_{N,M}(\mu); x; \mu) \in W^g_{M+1}$ is obviously unlikely since W^g_M is constructed upon $g(u; x; \mu)$, and hence our error bound $\Delta_{N,M}(\mu)$ is not completely rigorous; however, if $u_{N,M}(\mu) \to u(\mu)$ very fast we expect that the effectivity $\eta_{N,M}(\mu) \equiv \Delta_{N,M}(\mu)/\|u(\mu) - u_{N,M}(\mu)\|_X$ is close to (and above) unity.

7.4.2 Offline/Online Computational Procedure

It remains to develop the offline-online computational procedure for the efficient calculation of our error bounds $\Delta_{N,M}(\mu)$ and $\Delta^s_{N,M}(\mu)$. To begin, we note from standard duality arguments that

$$\sup_{v\in X} \frac{r(v; g_M^{u_{N,M}}(x;\mu))}{\|v\|_X} = \|\hat{e}_{N,M}(\mu)\|_X \ , \tag{7.30}$$

where $\hat{e}_{N,M}(\mu)$ is given by

$$(\hat{e}_{N,M}(\mu), v)_X = r(v; g_M^{u_{N,M}}(x;\mu)), \quad \forall\, v \in X \ . \tag{7.31}$$

We next substitute $u_{N,M}(\mu) = \sum_{n=1}^N u_{N\,n}(\mu)\,\zeta_n$ and $g_M^{u_{N,M}}(x;\mu) = \sum_{m=1}^M \varphi_{M\,m}(\mu)\,q_m(x)$ into (7.26) to expand $r(v; g_M^{u_{N,M}}(x;\mu))$ as

$$r(v; g_M^{u_{N,M}}(x;\mu)) = f(v) - \sum_{n=1}^N u_{N,M\,n}(\mu)\,a^L(\zeta_n, v) - \sum_{m=1}^M \varphi_{M\,m}(\mu) \int_\Omega q_m(x) v, \quad \forall\, v \in X. \tag{7.32}$$

It then follows from (7.31)-(7.32) and linear superposition that we may express $\hat{e}_{N,M}(\mu) \in X$ as

$$\hat{e}_{N,M}(\mu) = \mathcal{C} + \sum_{j=1}^{N+M} \sigma_j(\mu)\,\mathcal{L}_j \ , \tag{7.33}$$

where $(\mathcal{C}, v)_X = f(v)$, $\forall\, v \in X$; $\sigma_n(\mu) = u_{N,M\,n}(\mu)$, $(\mathcal{L}_n, v)_X = -a^L(\zeta_n, v)$, $\forall\, v \in X$, for $1 \le n \le N$; $\sigma_{m+N}(\mu) = \varphi_m(\mu)$, $(\mathcal{L}_{m+N}, v)_X = -\int_\Omega q_m(x) v$, $\forall\, v \in X$, for $1 \le m \le M$. It thus follows that

$$\|\hat{e}_{N,M}(\mu)\|_X^2 \;=\; (\mathcal{C},\mathcal{C})_X + 2\sum_{j=1}^{N+M} (\mathcal{C},\mathcal{L}_j)_X + \sum_{j=1}^{N+M}\sum_{j'=1}^{N+M} \sigma_j(\mu)\sigma_{j'}(\mu)(\mathcal{L}_j,\mathcal{L}_{j'})_X \ . \tag{7.34}$$

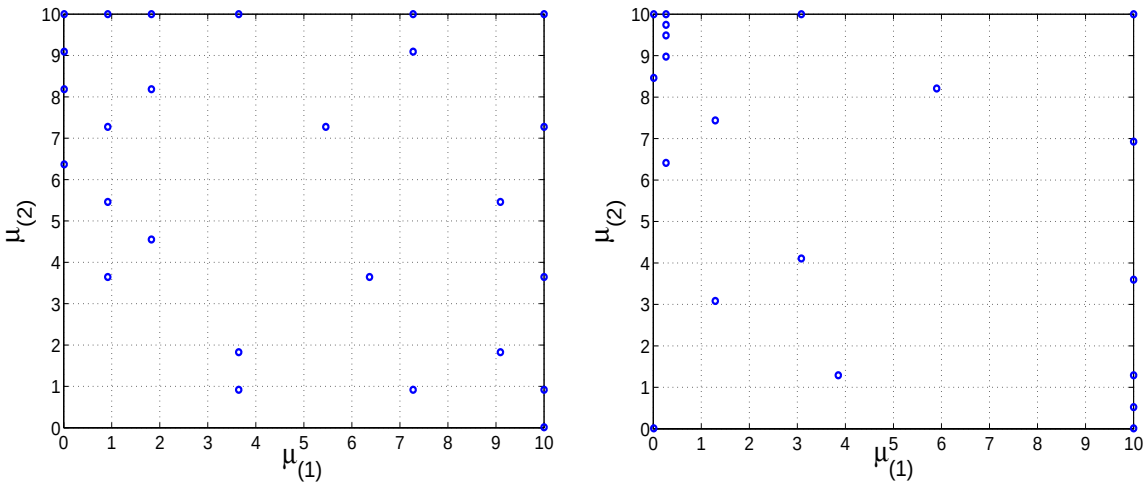
Finally, by invoking duality arguments we calculate $\sup_{v\in X}[\int_\Omega q_{M+1}(x)v/\|v\|_X] = \|\mathcal{Z}\|_X$, where $(\mathcal{Z}, v)_X = \int_\Omega q_{M+1}(x)v, \forall\, v \in X$. The offline-online decomposition is now clear.

In the offline stage — performed only once — we first solve for $\mathcal{C}$, $\mathcal{Z}$, and $\mathcal{L}_j$, $1 \le j \le N+M$; we then form and store the associated parameter-independent inner products $(\mathcal{Z},\mathcal{Z})_X$, $(\mathcal{C},\mathcal{C})_X$, $(\mathcal{C},\mathcal{L}_j)_X$, $(\mathcal{L}_j,\mathcal{L}_{j'})_X$, $1 \le j, j' \le N+M$. Note that these inner products computed offline are independent of the parameter μ.

In the online stage — performed many times for each new μ — we simply evaluate the sum (7.34) in terms of $\varphi_{M\,m}(\mu), u_{N,M\,n}(\mu)$ — at cost $O((N+M)^2)$. The online cost is independent of $\mathcal{N}$. We further note that unless M is very large, the online cost associated with the calculation of the error bounds is much less expensive than the online cost associated with the calculation of $s_{N,M}(\mu)$.

7.5 Numerical Results

In this section, we apply our approach to the model problem described in Section 7.1.2 and present associated numerical results. We first choose bound conditioner $(w,v)_X = \int_\Omega \nabla w \cdot \nabla v$ and thus obtain $\alpha = 1$. We then introduce Ξ^g — a regular grid of 144 parameter points — up on which S^g_M, W^g_M, T_M, and B^M, $1 \le M \le M_{\max}$, are constructed for $M_{\max} = 26$ by the procedure of Section 7.2. We readily pursue the adaptive sampling procedure described in Section 6.5.3 to construct the nested samples S_N: for $\epsilon_{\text{tol,min}} = 10^{-5}$ and $n_F = 1600$, we obtain $N_{\max} = 20$. We present in Figure 7-2 the two samples $S^g_{M_{\max}}$ and $S_{N_{\max}}$. As expected from the form of nonlinearity, both the samples are distributed mainly around the two "corners" $(10, 0.01)$ and $(0.01, 10)$. .



(a) (b)

Figure 7-2: Parameter sample set: (a) $S^g_{M_{\max}}$ and (b) $S_{N_{\max}}$.

We now introduce a parameter test sample Ξ_{Test} of size 225 (in fact, a regular

15×15 grid over $\mathcal{D}$), and define $\varepsilon_{N,M,\text{max,rel}} = \max_{\mu \in \Xi_{\text{Test}}} \|e_{N,M}(\mu)\|_X / \|u(\mu)\|_X$ and $\varepsilon^s_{N,M,\text{max,rel}} = \max_{\mu \in \Xi_{\text{Test}}} |s_{N,M}(\mu)|/|s(\mu)|$. We present in Figure 7-3 $\varepsilon_{N,M,\text{max,rel}}$ and $\varepsilon^s_{N,M,\text{max,rel}}$ as a function of N and M. We observe very rapid convergence of the reduced-basis approximations. Furthermore, the errors behave very similarly as in the linear elliptic example of the previous chapter: the errors initially decrease, but then maintain persistently plateau with N for a particular value of M; increasing M effectively brings the error curves down.

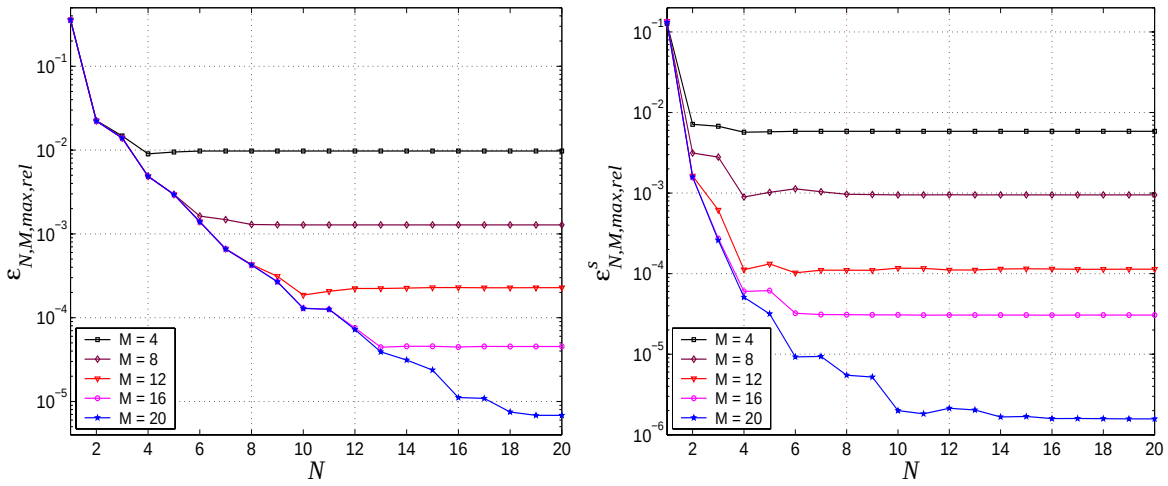


Figure 7-3: Convergence of the reduced-basis approximations for the model problem.

N	M	$\Delta_{N,M,\text{max,rel}}$	$\bar{\eta}_{N,M}$	$\Delta^s_{N,M,\text{max,rel}}$	$\bar{\eta}^s_{N,M}$
3	4	1.05 E – 01	3.28	1.08 E – 01	32.33
6	8	7.07 E – 03	1.74	7.40 E – 03	28.94
9	12	9.33 E – 04	1.84	9.78 E – 04	24.09
12	16	9.44 E – 05	2.15	9.66 E – 05	16.61
15	20	2.60 E – 05	1.27	2.60 E – 05	55.35
18	24	9.27 E – 06	1.08	9.53 E – 06	53.12

Table 7.1: Effectivities for the model problem.

We further show in Table 7.1 $\Delta_{N,M,\text{max,rel}}$, $\bar{\eta}_{N,M}$, $\Delta^s_{N,M,\text{max,rel}}$, and $\bar{\eta}^s_{N,M}$ as a function of N and M. Here $\Delta_{N,M,\text{max,rel}}$ is the maximum over Ξ_{Test} of $\Delta_{N,M}(\mu)/\|u_{N,M}(\mu)\|_X$, $\bar{\eta}_{N,M}$ is the average over Ξ_{Test} of $\Delta_{N,M}(\mu)/\|e_{N,M}(\mu)\|$, $\Delta^s_{N,M,\text{max,rel}}$ is the maximum over Ξ_{Test} of $\Delta^s_{N,M}(\mu)/|s_{N,M}(\mu)|$, and $\bar{\eta}^s_{N,M}$ is the average over Ξ_{Test} of $\Delta^s_{N,M}(\mu)/|s(\mu) - s_{N,M}(\mu)|$. We observe that $\Delta_{N,M,\text{max,rel}}$ converges very rapidly, and the associated effectivities are $O(1)$; hence our energy error bound is very close to the true error. However, the output

bound does not perform that well as the output effectivities are quite high primarily due to the relatively crude output bounds by using the dual norm of the output functional.

Next we look at the relative contribution of the rigorous and non-rigorous components to the energy error bound $\Delta_{N,M}(\mu)$. We display in Figure 7-4 the ratio $\Delta_{N,M,\text{ave,n}}/\Delta_{N,M,\text{ave}}$ as a function of N and M. Here $\Delta_{N,M,\text{ave}}$ is the average over Ξ_{Test} of $\Delta_{N,M}(\mu)$; $\Delta_{N,M,\text{ave,n}}$ is the average over Ξ_{Test} of $\frac{\hat{\varepsilon}_M(\mu}{)}\alpha \sup_{v\in X}[\int_\Omega q_{M+1}v/\|v\|_X]$ — note this quantity is nonrigorous due to the lower bound property of $\hat{\varepsilon}_M(\mu)$ as a surrogate for $\varepsilon_M(\mu)$. We see that very similar as in the linear elliptic example of the previous chapter, the ratio tends to increase with N, but decrease with M.

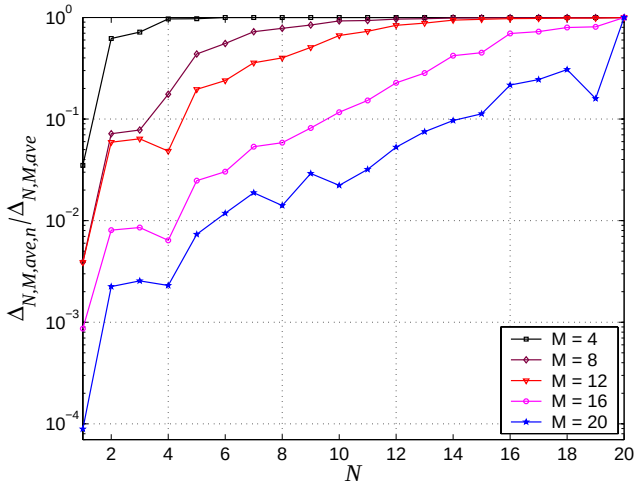


Figure 7-4: $\Delta_{N,M,\text{ave,n}}/\Delta_{N,M,\text{ave}}$ as a function of N and M.

Finally, we present in Table 7.2 the online computational times to calculate $s_{N,M}(\mu)$ and $\Delta^s_{N,M}(\mu)$ as a function of (N, M). The values are normalized with respect to the computational time for the direct calculation of the truth approximation output $s(\mu) = \ell(u(\mu))$. The computational savings are much larger in the nonlinear case: for an accuracy of less than 0.1 percent ($N = 9$, $M = 12$) in the output bound, we observe the online factor of improvement of $O(5000)$. This is mainly because solution of the "truth" approximation (7.3) involving the matrix assembly of the nonlinear terms and Newton update iterates is computationally expensive. Note also that the time to calculate $s_N(\mu)$ considerably exceeds that of calculating $\Delta^s_{N,M}(\mu)$ — this is due to the higher computational cost, $O(MN^2 + N^3)$ at each Newton iteration, to solve for $u_{N,M}(\mu)$. The online cost thus grows faster with N than with M, but can be controlled tightly by our online

adaptive strategy.

N	M	$s_{N,M}(\mu)$	$\Delta^s_{N,M}(\mu)$	$s(\mu)$
3	4	5.83 E – 05	6.07 E – 06	1
6	8	1.33 E – 04	1.09 E – 05	1
9	12	2.31 E – 04	1.21 E – 05	1
12	16	3.72 E – 04	1.33 E – 05	1
15	20	5.48 E – 04	1.58 E – 05	1
18	24	7.40 E – 04	2.32 E – 05	1

Table 7.2: Online computational times (normalized with respect to the time to solve for $s(\mu)$) for the model problem.

However, the offline computations in the nonlinear case are also more extensive primarily due to the sampling procedure for S^g_M. Nevertheless, at the present, the technique can be gainfully employed in many practical applications that willingly accept the extensive offline cost in exchange for the real-time responses in the design, optimization, control, characterization contexts.

Chapter 8

A Real-Time Robust Parameter Estimation Method

8.1 Introduction

Inverse problem has received enormous attention due to its practical importance in many engineering and science areas such as nondestructive evaluation, computer tomography and imaging, geophysics, biology, medicine and life science. Formally defined, inverse problems are concerned with determining unknown causes from a desired or an observed effect. In this thesis, we shall view inverse problems as follows: the *forward problem* is to evaluate the PDE-induced input-output relationship (which in turn demands solution of the underlying partial differential equation); in contrast, the *inverse problem* is concerned with deducing the inputs from the measured-observable outputs.

In Chapters 3 through 7, we develop the method for rapid and reliable evaluation of the PDE-induced input-output relationship for various classes of PDEs: linear coercive and noncoercive equations, nonaffine elliptic equations, as well as highly nonlinear monotonic elliptic equations. The three essential components are *reduced-basis approximations*, *a posteriori error estimation*, and *offline-online computational procedures*. Thanks to the rapid convergence of the reduced-basis approximation and the offline/online computational stratagem we can, in fact, enable real-time prediction of the outputs; and, thanks to our a posteriori error estimators, we can associate rigorous certificates of fidelity to

our (very fast) output predictions. The method is thus ideally suited to the inverse-problem context in which thousands of output predictions are often required effectively in real-time.

The wide range of applications has stimulated the development of various solution techniques for inverse problems. However, the inverse problem is typically ill-posed; in almost cases the techniques are quite expensive and do not well quantify uncertainty. Ill-posedness is traditionally addressed by regularization. Unfortunately, though adaptive regularization techniques are quite sophisticated, the ultimate prediction is nevertheless affected by the *a priori* assumptions — in ways that are difficult to quantify in a robust fashion.

Our approach promises significant improvements. In particular, based on the reduce-basis method we develop a *robust* inverse computational method for very *fast solution region* of inverse problems characterized by parametrized PDEs. The essential innovations are threefold. First, we apply the reduce-basis method to the forward problem for the rapid certified evaluation of PDE input-output relations and associated rigorous error bounds. Second, we incorporate the reduced-basis approximation and error bounds into the inverse problem formulation. Third, rather than regularize the goodness-of-fit objective, we may instead identify all (or almost all, in the probabilistic sense) inverse solutions consistent with the available experimental data. Ill-posedness is captured in a bounded "possibility region" that furthermore shrinks as the experimental error is decreased.

We further extend our inverse method to an "Analyze-Assess-Act" approach for the adaptive design and robust optimization of engineering systems. In the Analyze stage we analyze system characteristics to determine which ranges of experimental control variables may produce sensitive data. In the Assess stage we pursue robust parameter estimation procedures that map measured-observable outputs to (all) possible system-characteristic inputs. In the subsequent Act stage we pursue adaptive design and robust optimization procedures that map mission-objective outputs to best control-variable inputs. The essential mathematical ingredients of our approach are twofold. First, we employ reduced-basis approximations and associated *a posteriori* error estimation to provide extremely rapid output bounds for the output of interest. Second, we employ a combination of advanced optimization procedures and less sophisticated probabilistic and enumerative techniques:

techniques which incorporate our reduced-basis output bounds for efficient minimization of objective functions with strict adherence to constraints.

8.2 Problem Definition

8.2.1 Forward Problems

A mathematical formulation of the PDE-induced input-output relationship is stated as: For given $\mu \in \mathcal{D} \subset \mathbb{R}^P$, we evaluate $s^{\mathrm{e}}(\mu) = \ell(u^{\mathrm{e}}(\mu))$, where the field variable $u^{\mathrm{e}}(\mu)$ satisfies a μ-parametrized partial differential equation that describes the underlying physics, $\mathfrak{g}(u^{\mathrm{e}}(\mu), v; \mu) = 0,\ \forall\ v \in X^{\mathrm{e}}$. Here $\mathcal{D}$ is the parameter domain in which our P-tuple input μ resides; X^{e} is an appropriate Hilbert space defined over spatial domain $\Omega \in \mathbb{R}^d$; $f(\cdot), \ell(\cdot)$ are X^{e}-continuous linear functionals; and $\mathfrak{g}$ is the weak form of the underlying partial differential equation. In the linear case, we typically have

$$\mathfrak{g}(w, v; \mu) \equiv a(w, v; \mu) - f(v)\ , \tag{8.1}$$

where $a(\cdot, \cdot; \mu)$ is a continuous bilinear form. In the nonlinear case, $\mathfrak{g}$ typically includes general nonlinear functions of the field variable $u^{\mathrm{e}}(\mu)$ and/or its partial derivatives. We assume explicitly here that the forward problem is well-posed in the Hadamard sense. This essentially requires the solution exist, be unique, and depend continuously on data.

Recall that the PDE may not be analytically solvable; rather classical numerical approaches like the finite element analysis are used. In the finite element method, we first introduce a piecewise-polynomial "truth" approximation subspace X ($\subset X^{\mathrm{e}}$) of dimension $\mathcal{N}$. The "truth" finite element approximation is then found by (say) Galerkin projection: Given $\mu \in \mathcal{D} \subset \mathbb{R}^P$, we evaluate

$$s(\mu) = \ell(u(\mu))\ ; \tag{8.2}$$

where $u(\mu) \in X$ is the solution of

$$\mathfrak{g}(u(\mu), v; \mu) = 0, \quad \forall\, v \in X\ . \tag{8.3}$$

For accurate numerical solution of the underlying PDE, we shall assume — hence the appellation "truth" — that X is sufficiently rich that $u(\mu)$ (respectively, $s(\mu)$) is sufficiently close to $u^{\mathrm{e}}(\mu)$ (respectively, $s^{\mathrm{e}}(\mu)$) for all μ in the (closed) parameter domain $\mathcal{D}$. Unfortunately, for any reasonable error tolerance, the dimension $\mathcal{N}$ required to satisfy this condition — even with the application of appropriate (and even parameter-dependent) adaptive mesh generation/refinement strategies — is typically extremely large, and in particular much too large to provide real-time solution of inverse problems.

For clarification of the following discussion, let us denote the input-output relationship by an explicit mapping $F : \mu \in \mathcal{D} \rightarrow s \in \mathbb{R}$, hence $s(\mu) = F(\mu)$. The mapping F encompassing the mathematical formulation of the forward problem is called *forward operator*; and hence evaluating F is equivalent to solving (8.3) and (8.2).

8.2.2 Inverse Problems

As mentioned earlier, in inverse problems we are concerned with predicting the unknown input parameters from the measured/observable outputs. In the inverse-problem context, our input has two components, $\mu = (\nu, \sigma)$. The first component $\nu = (\nu^1, \ldots, \nu^M) \in \mathcal{D}^\nu \subset \mathbb{R}^M$ comprises system-characteristic parameters that must be identified; here $\mathcal{D}^\nu$ is the associated domain of interest. The second component σ consists of experimental control variables that are used to obtain the experimental data. Inverse problems may then involve identifying the "exact" but unknown parameter ν^* from

$$\{\nu \in \mathcal{D}^\nu \mid F(\nu, \sigma_k) = s(\nu^*, \sigma_k), 1 \leq k \leq K\}, \tag{8.4}$$

where $s(\nu^*, \sigma_k), 1 \leq k \leq K$ are the "noise-free" data, and K is the number of measurements. For convenience below, let us denote $\mathbf{F}(\nu) = (F(\nu, \sigma_1), \ldots, F(\nu, \sigma_K))$ and $\mathbf{s}(\nu^*) = (s(\nu^*, \sigma_1), \ldots, s(\nu^*, \sigma_K))$, and thus rewrite (8.4) as

$$\mathbf{F}(\nu) = \mathbf{s}(\nu^*) \ . \tag{8.5}$$

Neither existence nor uniqueness of a solution to (8.5) are guaranteed. The notion of *parameter identifiability* is important since parameter identifiability is concerned with

the question whether the parameters can be uniquely identified from knowledge about the outputs, assuming perfect data. The parameter ν^* is called *identifiable* if $\mathbf{F}(\nu) = \mathbf{s}(\nu^*)$ implies $\nu = \nu^*$, if otherwise non-identifiable. A problem is called *parameter identifiable* if every $\nu^* \in \mathcal{D}^\nu$ is identifiable, i.e., the mapping $\mathbf{F}$ is one-to-one. The identifiability of ν^* depends not only on the governing equation, but also on the outputs and the number of observations/measurements. Furthermore, a problem may be parameter identifiable when considering the analytic forward operator $F^{\mathrm{e}}(\mu)$ and analytic output data $\mathbf{s}^{\mathrm{e}}(\nu^*)$, but loses this property for the discretized inverse problem (8.5). We refer to [12] for detailed description of the identifiability concept in inverse problems.

Clearly, under assumptions that the exact data $\mathbf{s}(\nu^*)$ is attainable and that the underlying model is correct, solution of (8.5) does exist, but need not be necessarily unique. Typically, solution of the problem (8.5) is found by solving the minimization problem

$$\nu^* = \arg\min_{\nu \in \mathcal{D}^\nu} ||\mathbf{F}(\nu) - \mathbf{s}(\nu^*)|| \; ; \tag{8.6}$$

here $\|\cdot\|$ denotes the Euclidean norm. In actual practice, due to errors in the measurement the exact data are not known precisely, and only the perturbed experimental data, $\{s^\delta(\nu^*, \sigma_k), 1 \leq k \leq K\}$, satisfying

$$\left| \frac{s^\delta(\nu^*, \sigma_k) - s(\nu^*, \sigma_k)}{s(\nu^*, \sigma_k)} \right| \leq \epsilon_{\mathrm{exp}}, \quad 1 \leq k \leq K \; , \tag{8.7}$$

are available; here ϵ_{exp} is the experimental error in the data. Note from (8.7) that the norm of noise in measurements is bounded by

$$\|\mathbf{s}^\delta(\nu^*) - \mathbf{s}(\nu^*)\| \leq \delta(\epsilon_{\mathrm{exp}}) \; , \tag{8.8}$$

where δ known as the noise level is a function of the experimental error ϵ_{exp}, and $\mathbf{s}^\delta(\nu^*) = (s^\delta(\nu^*, \sigma_1), \ldots, s^\delta(\nu^*, \sigma_K))$. The output least-squares formulation (8.6) is thus replaced by the minimization problem

$$\nu^\delta = \arg\min_{\nu \in \mathcal{D}^\nu} ||\mathbf{F}(\nu) - \mathbf{s}^\delta(\nu^*)|| \; , \tag{8.9}$$

with the noise estimate (8.7) or more conveniently (8.8).

Our approach to treat inverse problems in the presence of uncertainty is different from traditional approaches in the sense that: rather than strive for only one regularized solution, we may identify all (or almost all, in the probabilistic sense) inverse solutions consistent with the available experimental data. In particular, instead of the vector $\mathbf{s}^{\delta}(\nu^*)$, the experimental data is given in the form of intervals

$$\mathcal{I}(\epsilon_{\text{exp}}, \sigma_k) \equiv \left[s(\nu^*, \sigma_k) - \epsilon_{\text{exp}} \left| s(\nu^*, \sigma_k) \right|, s(\nu^*, \sigma_k) + \epsilon_{\text{exp}} \left| s(\nu^*, \sigma_k) \right| \right], \ 1 \leq k \leq K. \quad (8.10)$$

Our inverse problem formulation is thus proposed as: Given experimental data in the form of intervals $\mathcal{I}(\epsilon_{\text{exp}}, \sigma_k), 1 \leq k \leq K$, we wish to determine a region $\mathcal{P} \in \mathcal{D}^{\nu}$ in which all possible inverse solutions must reside. Towards this end, we define

$$\mathcal{P} \equiv \{ \nu \in \mathcal{D}^{\nu} | F(\nu, \sigma_k) \in \mathcal{I}(\epsilon_{\text{exp}}, \sigma_k), 1 \leq k \leq K \} \ . \quad (8.11)$$

Here we use $\mathcal{I}(\epsilon_{\text{exp}}, \sigma_k)$ defined by (8.10) in our numerical experiments. Of course, in practice, we are not privy to $s(\nu^*, \sigma_k), 1 \leq k \leq K$; in such cases, the experimental data must be replaced with $\mathcal{I}(\epsilon_{\text{exp}}, \sigma_k) \equiv [s_{\min}(\sigma_k), s_{\max}(\sigma_k)] \supset s(\nu^*, \sigma_k), 1 \leq k \leq K$, where $s_{\min}(\sigma_k)$ and $s_{\max}(\sigma_k)$ are determined by manipulating the original experimental data set by means of statistical and error analysis [3, 139].

The crucial observation is the following

Proposition 13. *All possible solutions ν^* of the problem (1.4) reside in $\mathcal{P}$.*

Proof. We first note from (8.10) that $s(\nu^*, \sigma_k) \in \mathcal{I}(\epsilon_{\text{exp}}, \sigma_k), 1 \leq k \leq K$. The result follows from (8.5) and the definition of $\mathcal{P}$ in (8.11). □

Clearly, we have accommodated experimental error and model uncertainty. The size of $\mathcal{P}$ depends on the level of noise in experimental data and the particular problem involved and in turn reflects how severely ill-posed the inverse problem is. Also note importantly that the optimization formulation (8.9) yields only one particular solution $\nu^{\delta} \in \mathcal{P}$, since $s^{\delta}(\nu^*, \sigma_k) \in \mathcal{I}(\epsilon_{\text{exp}}, \sigma_k), 1 \leq k \leq K$.

Before proceeding with the development of the computational method for determining $\mathcal{P}$ (more precisely, another region that includes $\mathcal{P}$ and can be constructed very inexpen-

sively), we give in the next section a short discussion of computational approaches that can treat inverse problems with uncertain data.

8.3 Computational Approaches for Inverse Problems

8.3.1 Regularization Methods

In inverse problems, one must address ill-posedness namely existence, uniqueness, and continuous dependence of the solution on data. Under our assumptions of well-posedness of the forward problem and of boundedness of the output functional, and if we consider "zero" noise level ($\epsilon_{\text{exp}} = 0\%$), the solution of (8.9) does exist and depend continuously on data provided that mathematical model agrees with physical phenomena. But the inverse problem is still possibly ill-posed since multiplicity of solutions may exist. One practical way to counter the problem of multiple solutions is to obtain more data. One can then hope that fairly accurate results can be obtained by using "over-determined" data. In practice, due to the presence of noise in data, the use of over-determined data often leads to bad solutions. Consequently, in face of uncertainty, posing inverse problems as an optimization problem in (8.9) may work for some problems, but may not for others. As a result, *a priori* regularization hypotheses accommodating the uncertainty are usually made, and iterative regularization algorithms are often pursued [137, 39, 63].

Before proceeding with our discussion, it is worthwhile to point out under what circumstances regularization methods are appropriate for solving inverse problems. To begin, we note from (8.6) and (8.9) that

$$\mathbf{F}(\nu^*) - \mathbf{F}(\nu^\delta) = \mathbf{s}(\nu^*) - \mathbf{s}^\delta(\nu^*) \ . \tag{8.12}$$

If $\mathbf{F}$ is linear, we can then write (8.12) as

$$\mathbf{F}(\nu^* - \nu^\delta) = \mathbf{s}(\nu^*) - \mathbf{s}^\delta(\nu^*) \ . \tag{8.13}$$

Assuming $\mathbf{F}$ is invertible and if the inverse of $\mathbf{F}$ is bounded such that a small error in data leads to small error in the solution, ν^δ can then be considered a good approximation

for ν^*. However, if the inverse is unbounded, a very small error in data can be magnified by the inverse itself so that the error in the solution is unacceptably large when solving (8.9) directly for ν^δ.

In the nonlinear case, by replacing $\mathbf{F}(\nu^*)$ and $\mathbf{F}(\nu^\delta)$ with a first-order approximation around a point ν_0, we can obtain

$$\mathbf{F}'(\nu_0)(\nu^* - \nu^\delta) = \mathbf{s}^\delta(\nu^*) - \mathbf{s}(\nu^*) \; ; \tag{8.14}$$

here $\mathbf{F}'(\nu)$ is the Fréchet derivative of $\mathbf{F}$ at ν. Although it is not always true, but ill-posedness of a nonlinear problem is frequently characterized via its linearization [40]. Therefore, if a nonlinear inverse problem is solved by methods of linearization, its degree of ill-posedness is typically represented by the unboundedness of the inverse of $\mathbf{F}'(\nu)$. Regularization methods prove very appropriate for inverse problems whose ill-posedness comes from the unbounded property of the inverse of the operator or its Fréchet derivative.

In Tikhonov regularization, a *regularization parameter* α and associated *regularization operator* R reflecting the uncertainty are added to the problem (8.9) as a way to ensure fairly accurate solutions. This leads to the minimization problem

$$\nu^\delta = \arg\min_{\nu \in \mathcal{D}^\nu} ||\mathbf{F}(\nu) - \mathbf{s}^\delta(\nu^*)|| + \alpha(\delta) \, \|R(\nu)\| \; . \tag{8.15}$$

The basic idea is to replace the original ill-posed inverse problem with a family of well-posed problems by taking explicitly the uncertainty into the optimization problem. For a positive α, minimizers always exist under the well-posedness assumptions on the forward problems but need not be unique. In [41], it was shown that ν^δ depends continuously on the data and converges towards a solution ν^* of (8.6) in a set-valued sense as $\alpha(\delta) \to 0$ and $\delta^2/\alpha(\delta) \to 0$ as δ tends to zero. The solution method for (8.15) and the choice of α and R (by which the solution of inverse problems will be certainly affected) have inspired the development of many regularization techniques. Typically [42, 40], R is set to either ν or $\nu - \nu^a$, where ν^a is some *a priori* estimate of the desired solution ν^* of (8.6). For choosing the regularization parameter, there are two general rules of thumbs namely *a priori* and *a posteriori* choice. In *a priori* choice [40], the parameter is defined a function of only noise level δ such as $\alpha = O(\delta^{\frac{2}{2\rho+1}})$ for some $\rho \in [1/2, 1]$. Although simple and

less computationally expensive, the *a priori* choice does not guarantee the best accuracy for the solution of (8.15) since the optimal value of α at which the best accuracy is obtained is generally unknown. As a consequence, several *a posteriori* strategies have been introduced to provide the best possible solution. In the *a posteriori* choice, α is decided by the generalized discrepancy principle [137, 112] such that

$$\left\|\mathbf{F}(\nu^\delta) - \mathbf{s}^\delta(\nu^*)\right\| = \delta(\epsilon_{\text{exp}}) \tag{8.16}$$

holds; see [133] for *a posteriori* strategy that yields optimal convergence rate.

Tikhonov regularization methods have been very successful in solving linear ill-posed problems. However, in the nonlinear case, solving the problem (8.15) with adherence to (8.16) is quite complicated and requires high computational efforts. Since furthermore the nonlinearity and nonconvexity of the problem (8.15) could make gradient methods fail if the problem is ill-posed, iterative regularization methods prove an attractive alternative.

A starting point for our discussion of iterative regularization methods is the Newton's method for solving (8.9)

$$\nu_{n+1}^\delta = \nu_n^\delta + \mathbf{F}'(\nu_n^\delta)^{-1}\left(\mathbf{s}^\delta(\nu^*) - \mathbf{F}(\nu_n^\delta)\right); \tag{8.17}$$

here $\mathbf{F}'(\nu)$ is the Fréchet derivative of $\mathbf{F}$ at ν. The stability and convergence of the Newton scheme (8.17) strongly depend on smoothness of $\mathbf{F}$ and invertibility of $\mathbf{F}'$. Even if $\mathbf{F}$ is well-posed and $\mathbf{F}'$ is invertible, the inverse of $\mathbf{F}'$ is usually unbounded for ill-posed problems. Therefore, some regularization technique has to be applied since (8.17) means to solve a linear ill-posed problem at every iteration. The idea is to replace the possibly unbounded inverse $\mathbf{F}'(\nu_n^\delta)^{-1}$ in (8.17) with a bounded operator $\mathbf{G}_{\alpha_n}$

$$\nu_{n+1}^\delta = \nu_n^\delta + \mathbf{G}_{\alpha_n}(\mathbf{F}'(\nu_n^\delta))\left(\mathbf{s}^\delta(\nu^*) - \mathbf{F}(\nu_n^\delta)\right) \ , \tag{8.18}$$

where $\{\alpha_n\}$ is a sequence of regularization parameters. There are several choices for $\mathbf{G}_{\alpha_n}$ leading to several iterative regularization schemes (see [23] and references therein

for detail). Typically, the Levenberg-Marquardt method [59] defines

$$\mathbf{G}_{\alpha_n} = (\mathbf{F}'(\nu_n^\delta)^*\mathbf{F}'(\nu_n^\delta) + \alpha_n\mathbf{I})^{-1}\mathbf{F}'(\nu_n^\delta)^* \ ; \tag{8.19}$$

here $\mathbf{F}'(\nu)^*$ is the adjoint of $\mathbf{F}'(\nu)$. This method is essentially the Tikhonov regularization applied to the linearized problem (8.17). The parameter α_n is chosen from the Morozov's discrepancy principle

$$\|\mathbf{s}^\delta(\nu^*) - \mathbf{F}(\nu_n^\delta) - \mathbf{F}'(\nu_n^\delta)(\nu_{n+1}^\delta - \nu_n^\delta)\| = \rho\|\mathbf{s}^\delta(\nu^*) - \mathbf{F}(\nu_n^\delta)\| \tag{8.20}$$

with some positive $\rho < 1$. The convergence and stability of the scheme was given in [59] (Theorem 2.3) under the assumptions that $\mathbf{F}'$ is locally bounded and that

$$\|\mathbf{F}(\nu) - \mathbf{F}(\tilde{\nu}) - \mathbf{F}'(\nu - \tilde{\nu})\| \leq C\|\mathbf{F}(\nu) - \mathbf{F}(\tilde{\nu})\|\|\nu - \tilde{\nu}\| \ , \tag{8.21}$$

for all ν and $\tilde{\nu}$ in a ball $\mathcal{B}$ around ν^* and some fixed $C > 0$. However, the convergence rate result has not been established.

The iteratively regularized Gauss-Newton method [8] suggested augmenting (8.18) with additional stabilization term

$$\nu_{n+1}^\delta = \nu_n^\delta + \mathbf{G}_{\alpha_n}\left(\mathbf{s}^\delta(\nu^*) - \mathbf{F}(\nu_n^\delta)\right) - \left(\mathbf{F}'(\nu_n^\delta)^*\mathbf{F}'(\nu_n^\delta) + \alpha_n\mathbf{I}\right)^{-1}\alpha_n\left(\nu_n^\delta - \nu^a\right) \ ; \tag{8.22}$$

where α_k is the *a priori* chosen sequence satisfying

$$\alpha_n > 0, \quad 1 \leq \frac{\alpha_n}{\alpha_{n+1}} \leq r, \quad \lim_{n\to\infty}\alpha_n = 0 \tag{8.23}$$

for some constant $r > 1$. The convergence and convergence rate of this method have been analyzed [20] under certain conditions on $\mathbf{F}$ and $\nu^a - \nu^*$. These conditions are in fact quite restricted and difficult to verify for many important inverse problems arising in medical imaging and nondetructive testing; see [67] for the improved scheme yielding higher rates of convergence even under weaker conditions.

Alternatively, the methods of steepest descent are also used to solve nonlinear ill-posed

problems. A typical scheme of this kind is the Landweber iteration

$$\nu_{n+1}^{\delta} = \nu_n^{\delta} + \mathbf{F}'(\nu_n^{\delta}) \left(\mathbf{s}^{\delta}(\nu^*) - \mathbf{F}(\nu_n^{\delta}) \right) \ . \tag{8.24}$$

In [60], the convergence analysis of the Landweber iteration is given under the following assumptions: for a ball $\mathcal{B}_\rho(\nu_0)$ of radius ρ around the initial guess ν_0, $\mathbf{F}$ is required to satisfy

$$\|\mathbf{F}'(\nu)\| \leq 1, \quad \nu \in \mathcal{B}_\rho(\nu_0), \tag{8.25}$$

$$\|\mathbf{F}(\tilde{\nu}) - \mathbf{F}(\nu) - \mathbf{F}'(\tilde{\nu} - \nu)\| \leq \eta \|\mathbf{F}(\nu) - \mathbf{F}(\tilde{\nu})\|, \quad \nu, \tilde{\nu} \in \mathcal{B}_\rho(\nu_0) \tag{8.26}$$

for some positive $\eta < 1/2$; in addition to (8.25) and (8.26), there are other (even more restricted) conditions that $\mathbf{F}$ and ν^* must meet. Despite slower convergence, the much simpler Landweber iteration are still useful in many situations (in which the evaluation of $\mathbf{G}_{\alpha_n}$ is computationally prohibitive), since a single step in (8.26) is much less expensive than in (8.18) and (8.22). Further discussion and other variants of the Landweber iteration can be also found in [36, 132].

8.3.2 Statistical Methods

Bayesian statistical methods have been applied to both linear and nonlinear inverse problems. In Bayesian approach [46, 147, 68, 16, 98], the experimental data, $\mathbf{s}^{\delta}(\nu^*) = (s^{\delta}(\nu^*, \sigma_1), \ldots, s^{\delta}(\nu^*, \sigma_K))$, is treated as a random variable with probability density function $p(\mathbf{s}^{\delta}(\nu^*)|\nu)$. Typically, the noise in measurement is normally distributed with zero mean and standard deviation λ, the probability density for the data in this case is given by

$$p(\mathbf{s}^{\delta}(\nu^*) \mid \nu) = \left(\frac{1}{2\pi\lambda^2} \right)^{K/2} \exp\{-\|\mathbf{F}(\nu) - \mathbf{s}^{\delta}(\nu^*)\|^2 / 2\lambda^2\} \ . \tag{8.27}$$

Bayesian approach also treats the unknown parameter ν^* as random variable; hence, *a prior* distribution $\Pi(\nu^*)$ is needed to express initial uncertainty about the unknown model parameter ν^*. For a simple Gaussian prior of mean ν^{a} and standard deviation τ, we have

the a priori distribution of the form

$$\Pi(\nu) = \exp\{-\|\nu - \nu^{\mathrm{a}}\|^2/2\tau^2\} \ . \tag{8.28}$$

To infer the information about ν^*, we compute the posterior density function, $\Pi(\nu|\mathbf{s}^\delta(\nu^*))$, which is the probability distribution of ν on the given data $\mathbf{s}^\delta(\nu^*)$. Bayes's theorem states that the posterior density function (PDF) is proportional to the product of the probability density of the data and the a priori distribution

$$\Pi(\nu \mid \mathbf{s}^\delta(\nu^*)) = \frac{p(\mathbf{s}^\delta(\nu^*) \mid \nu)\Pi(\nu)}{\int_{\mathcal{D}^\nu} p(\mathbf{s}^\delta(\nu^*) \mid \nu)\Pi(\nu)d\nu} \ . \tag{8.29}$$

It thus follows that

$$\Pi(\nu \mid \mathbf{s}^\delta(\nu^*)) = \frac{\exp\{-\|\mathbf{F}(\nu) - \mathbf{s}^\delta(\nu^*)\|^2/2\lambda^2\}\exp\{-\|\nu - \nu^{\mathrm{a}}\|^2/2\tau^2\}}{\int_{\mathcal{D}^\nu}\exp\{-\|\mathbf{F}(\nu) - \mathbf{s}^\delta(\nu^*)\|^2/2\lambda^2\}\exp\{-\|\nu - \nu^{\mathrm{a}}\|^2/2\tau^2\}d\nu} \ . \tag{8.30}$$

We now maximize to obtain the maximum a posteriori (MAP) estimate for the unknown parameter ν^* as

$$\nu^\delta_{\mathrm{MAP}} = \arg\max_{\nu\mathcal{D}^\nu} \Pi(\nu \mid \mathbf{s}^\delta(\nu^*)) \ . \tag{8.31}$$

By noting that maximization of the PDF is equivalent to minimization of the negative log of the PDE and that the denominator of (8.30) is constant, we arrive at

$$\nu^\delta_{\mathrm{MAP}} = \arg\min_{\nu\mathcal{D}^\nu} \|\mathbf{F}(\nu) - \mathbf{s}^\delta(\nu^*)\|^2 + \frac{\lambda^2}{\tau^2}\|\nu - \nu^{\mathrm{a}}\|^2 \ . \tag{8.32}$$

This expression reveals a connection between the Tikhonov regularization and Bayesian approach; in particular, the regularization parameter α is related to the covariance scale factor τ^2 by $\alpha = \lambda^2/\tau^2$. See [46] for an excellent discussion of the relationship of the Bayesian approach to Tikhonov regularization. Selection of the regularization parameter in Bayesian framework — more precisely, the covariance scale factor τ^2 — is also a critical issue, as it substantially affects the quality of the MAP estimate. Bayesian framework provides an automatic procedure to choose τ^2 based on the noise level and priori distribution models. Specifically, by treating τ^2 as a random variable, an optimal distribution of τ^2 can be obtained during iteration process [46, 144]. Nevertheless,

the selection procedure often involving the Monte-Carlo simulation is computationally extensive.

Generally, there are two solution methods in Bayesian estimation. The first one is the Markov Chain Monte Carlo (MCMC) simulation to solve the maximization (8.31) by exploring the posteriori density function $\Pi(\nu \mid \mathbf{s}^\delta(\nu^*))$. The basic idea is to draw a large set of samples $\{\nu_i^\delta\}_{i=1}^L$ from $\Pi(\nu \mid \mathbf{s}^\delta(\nu^*))$ based on an appropriate proposal distribution $q(\nu|\nu_i)$ that ensures the convergence of the irreducible Markov chain to the MAP estimate ν_{MAP}^δ. The set $\{\nu_i^\delta\}_{i=1}^L$ can then be used to construct the MAP estimate as $\nu_{\text{MAP}}^\delta = \arg\max_{i\in\{1,\dots,L\}} \Pi(\nu_i^\delta \mid \mathbf{s}^\delta(\nu^*))$. The efficiency of an MCMC algorithm depends on the proposal distribution; careful design of the proposal distribution can improve the convergence speed. We refer to [97, 68] for details of this method. The second solution method is any optimization procedure to solve the minimization (8.32), which is essentially the Tikhonov regularization in Bayesian framework.

While the expense of extensive computations (mostly due to the need of using Markov Chain Monte Carlo for the expectation and the regularization parameters) is a major disadvantage of the Bayesian approach to inverse problems, there are few advantages of this approach. First, the Bayesian approach provides a more robust and integrated analysis of inverse problems by including the statistics of the measurement error, the maximum likelihood of the unknown parameter, and the a priori distribution about the unknown parameter into the statistical process. (Note however that, in the case of nonlinear inverse problems the single most likely solution ν_{MAP}^δ may not be a good estimate for the true unknown parameter ν^* [48].) Second, not only a point estimate of the unknown parameter, but also summary statistics such as the probability distribution and variance of the estimate which give a measure of uncertainty in the estimate can be obtained at the same time.

8.3.3 Assess-Predict-Optimize Strategy

The idea of finding all possible inverse solutions consistent with the experimental data is not new. In fact, it was proposed in [3, 139] as one of the three components of the Assess-Predict-Optimize (APO) Strategy. The APO has been used in the design contexts

for Assessment, Prediction, and Optimization of evolving systems (under design) under changing environmental conditions and dynamic objectives.
In the Assess component, the experimental data is first given in the form of K intervals

$$I^k = [s_{\min}(\sigma_k), s_{\max}(\sigma_k)], \quad k = 1, \ldots, K \ , \tag{8.33}$$

where $s_{\min}(\sigma_k)$ and $s_{\max}(\sigma_k)$ are determined by manipulating the original experimental data set using statistical and error analysis. The set B of all possible parameter values ν consistent with the experimental data is then defined as

$$B = \{\nu \in \mathcal{D}^\nu \mid s(\nu, \sigma_k) \in I^k, k = 1, \ldots, K\}. \tag{8.34}$$

In the Predict component, by applying the reduced-basis output bounds a conservative approximation to B, $\hat{B}$, can be defined such that

$$B \subset \hat{B}. \tag{8.35}$$

Note that the definition of $\hat{B}$ is identical that of the region $\mathcal{R}$ described in Section 8.4.1.
In the Optimize component, the goal is to find the optimal "design" parameter θ^* that minimizes a design objective $f(\theta)$ while strictly meeting a dynamic requirement, for example, $\max_{\mu \in \hat{B}} g(\theta, \mu) \leq C$. Mathematically, we look for

$$\begin{aligned} \theta^* \quad = \quad & \arg \min_{\theta \in \mathcal{D}^\theta} f(\theta) \\ & \text{s.t.} \ \max_{\mu \in \hat{B}} g(\theta, \mu) \leq C \ ; \end{aligned} \tag{8.36}$$

here g is a dynamic-constraint function, and constant C is an allowable limit. For rapid and reliable solution of the optimization problem (8.36), we refer to the work of Oliveira and Patera [103, 104] in which the authors developed the reduced-basis output bounds, the derivative and Hessian of the output bounds and incorporated them into a trust-region sequential quadratic programing implementation of interior-point methods to obtain very fast global (at least local) minimizers.

Though, in fact, the technique has been intended for optimal parametric design, it

can be efficiently used to solve the inverse problems with uncertainty: the method not only addresses both experimental and numerical errors rigorously and effectively, but also reduces the model complexity significantly by applying the reduced-basis approximation and associated error bounds to the original model. However, no direct construction of $\hat{B}$ was proposed; instead, in the inner optimization problem, the feasible domain $\hat{B}$ was replaced with the parameter domain $\mathcal{D}^{\mu}$ and associated set of relevant constraints. As a consequence, the optimization procedure to solve the two-level optimization (8.36) was very costly even with the incorporation of the reduced-basis output bounds. Our aim of this chapter is to remedy this problem by providing an efficient construction of $\hat{B}$ and thus render the APO strategy more useful.

8.3.4 Remarks

We see that solution of inverse problems amounts to solving some kind of optimization problems. Of course, optimization techniques for solution of optimization-based inverse problems are rich. Global heuristic optimization strategies such as neural networks, MCMC simulation, simulated annealing, and genetic algorithms have powerful ability in finding globally optimal solutions for general nonconvex problems [79, 80, 151, 152, 58]. However, the problem with these methods is that they are heuristic by nature and computationally expensive. Therefore, gradient methods like Newton's method [8, 20, 19, 71], descent methods [60, 125], and current state-of-the-art interior-point method [24, 104] have been employed to solve inverse problems in many cases. Unfortunately, the objective is usually nonlinear and nonconvex, leading to the presence of multiple local minima which can not easily be bypassed by local optimization strategies. Moreover, it is more much difficult and expensive to obtain gradient $F'(\nu)$ and Hessian $F''(\nu)$ of the forward operator F, which could further restricts the use of gradient optimization procedures for inverse problems. The choice of optimization methods for solving inverse problems depends on many factors such as the convexity/nonconvexity of objective, the accessibility to the gradient $F'(\nu)$ and Hessian $F''(\nu)$, the particular problem involved, and the availability of computational resources.

In summary, there are a wide variety of techniques for solving inverse problems. How-

ever, in almost cases the inverse techniques are expensive due to the following reasons: solution of the forward problem by classical numerical approaches is typically long; associated optimization problems are usually nonlinear and nonconvex; and most importantly, inverse problems are typically ill-posed. Ill-posedness is traditionally addressed by regularization methods or Bayesian statistical approach. Though quite sophisticated, regularization and Bayesian methods are quite expensive (often fail to achieve numerical solutions in real-time) and often need additional information and thus lose algorithmic generality (in many cases, do not well quantify uncertainty). Furthermore, in the presence of uncertainty, solution of the inverse problem should never be unique at least in terms of mathematical sense; there should be indefinite inverse solutions that are consistent with model uncertainty. However, most inverse techniques provide only one inverse solution among the universal; and hence they do not exhibit and characterize ill-posed structure of the inverse problem.

8.4 A Robust Parameter Estimation Method

In this section, we aim to develop a robust inverse computational method for very fast solution region of many inverse problems in PDEs. The essential components are: (*i*) reduced-inverse model — application of the reduced-basis method to the forward problem for effecting significant reduction in computational expense, and incorporation of very fast output bounds into the inverse problem formulation for defining a possibility region that contains (all) inverse solutions consistent with the available experimental data; (*ii*) robust inverse algorithm — efficient construction of the possibility region by conducting a binary chop at different angles to map out its boundary; (*iii*) ellipsoid of the possibility region — introduction of the small ellipsoid containing the possibility region by solving an appropriate convex quaratic minimization.

8.4.1 Reduced Inverse Problem Formulation

Identifying the very high dimensionality and complexity of the inverse problem formulation (8.11) originated by the need for solving the forward problem, we first apply the reduced-basis method to obtain the output approximation $s_N(\nu, \sigma)$ and associated error

bound $\Delta_N^s(\nu,\sigma)$. We then introduce $s_N^{\pm}(\nu,\sigma) \equiv s_N(\nu,\sigma) \pm \Delta_N^s(\nu,\sigma)$, and recall that — thanks to our rigorous bounds — $s(\nu,\sigma) \in [s_N^-(\nu,\sigma), s_N^+(\nu,\sigma)]$.[1] We finally define

$$\mathcal{R} \equiv \left\{ \nu \in \mathcal{D}^\nu \mid \left[s_N^-(\nu,\sigma_k), s_N^+(\nu,\sigma_k) \right] \cap \mathcal{I}(\epsilon_{\text{exp}},\sigma_k) \neq \emptyset, 1 \leq k \leq K \right\} . \tag{8.37}$$

The remarkable result is that

Proposition 14. *The region $\mathcal{R}$ is a superset of $\mathcal{P}$, i.e., $\mathcal{P} \subset \mathcal{R}$; and hence $\nu^* \in \mathcal{R}$.*

Proof. For any ν in $\mathcal{P}$, we have $F(\nu,\sigma_k) \in \mathcal{I}(\epsilon_{\text{exp}},\sigma_k)$; furthermore, we also have $F(\nu,\sigma) \in [s_N^-(\nu,\sigma), s_N^+(\nu,\sigma)]$. It thus follows that $[s_N^-(\nu,\sigma_k), s_N^+(\nu,\sigma_k)] \cap \mathcal{I}(\epsilon_{\text{exp}},\sigma_k) \neq \emptyset, 1 \leq k \leq K$; and hence ν in $\mathcal{R}$. □

Let us now make a few important remarks: First, by introducing $\mathcal{R}$ we have not only accommodated model uncertainty (within our model assumptions) but also numerical error. Second, unlike the inverse problem formulation (8.11), the complexity of our reduced inverse model (8.37) is *independent* of $\mathcal{N}$ — the dimension of the underlying truth finite element approximation space. Third, $\mathcal{R}$ is almost indistinguishable from $\mathcal{P}$ if the error bound $\Delta_N^s(\mu)$ is very small compared to the experimental error ϵ_{exp} — this is typically observed given the rapid convergence of reduced-basis approximations and the rigor and sharpness of error bounds as demonstrated in the earlier chapters. And fourth, in the absence of measurement and numerical errors ($\epsilon_{\text{exp}} = \Delta_N^s(\mu) = 0$), the possibility region for an "identifiable" inverse problem is just the unique parameter point ν^*, i.e., $\mathcal{R} \equiv \nu^*$. In practice, it is unlikely to find such $\mathcal{R}$ due to the numerical error and computational expense; however, we can numerically test and confirm this behavior. We simply decrease the measurement error gradually and plot the possibility region for each error level. We will use this as a regular test when discussing numerical results in the next two chapters.

8.4.2 Construction of the Possibility Region

Of course, it is not possible to find all points in $\mathcal{R}$, and hence the idea is to construct the boundary of $\mathcal{R}$. Towards this end, we first find one point ν_c in $\mathcal{R}$ which is called the

[1] We do note that in nonaffine and nonlinear case our *a posteriori* error estimators — though quite sharp and efficient — are completely rigorous upper bounds only in certain restricted situations.

initial center; next for a chosen direction d_j from the initial center ν_c we conduct a binary chop to find the associated boundary point, ν_j, of $\mathcal{R}$; we repeat the second step for J different directions to obtain a discrete set of J points $R_J = \{\nu_1, \ldots, \nu_J\}$ representing the boundary of $\mathcal{R}$. The algorithm is given below

---

1. Set $R_J =$ and find $\nu_c \in \mathcal{R}$;
2. **For** $j = 1 : J$
3. Set $\nu_i = \nu_c$ and choose a direction d_j;
4. Find λ such that $\nu_o = \nu_c + \lambda d_j \notin \mathcal{R}$;
5. **Repeat**
6. Set $\nu_j = (\nu_i + \nu_o)/2$;
7. **If** $\nu_j \in \mathcal{R}$ **Then** $\nu_i = \nu_j$ **Else** $\nu_o = \nu_j$;
8. **Until** $\{||\nu_o - \nu_i||$ is sufficiently small.$\}$
9. $R_J = R_J \cup \{\nu_j\}$;
10. **End For**

Figure 8-1: Robust algorithm for constructing the solution region $\mathcal{R}$.

In essence, we move from the center ν_c toward the boundary of $\mathcal{R}$ by subsequently halving the distance between the inner point ν_i and the outer point ν_o. Note from (8.37) that $\nu \in \mathcal{R}$ if and only if ν resides in $\mathcal{D}^\nu$ and satisfies

$$
\begin{aligned}
s_N(\nu, \sigma_k) + \Delta_N^s(\nu, \sigma_k) &\geq s(\nu^*, \sigma_k) - \varepsilon_{\text{exp}} |s(\nu^*, \sigma_k)|, \quad k = 1, \ldots, K \\
s_N(\nu, \sigma_k) - \Delta_N^s(\nu, \sigma_k) &\leq s(\nu^*, \sigma_k) + \varepsilon_{\text{exp}} |s(\nu^*, \sigma_k)|, \quad k = 1, \ldots, K \ .
\end{aligned}
$$

To find the initial center, we propose to solve the following minimization

$$
\begin{aligned}
\text{(ICP)} \qquad & \text{minimize}_\nu \ \|\mathbf{s}_N(\nu) - \mathbf{s}(\nu^*)\| \\
& |s_N(\nu, \sigma_k) - s(\nu^*, \sigma_k)| \leq \Delta_N^s(\nu, \sigma_k), \quad k = 1, \ldots, K \\
& \nu \in \mathcal{D}^\nu,
\end{aligned}
$$

for the minimizer ν_{min}, where $\mathbf{s}_N(\nu) = (s_N(\nu, \sigma_1), \ldots, s_N(\nu, \sigma_K))$.[2] We can demonstrate

Proposition 15. *Minimizer of the ICP problem exists and resides in* $\mathcal{R}$.

[2] In practical contexts, since the exact data $\mathbf{s}(\nu^*)$ is not accessible, we should replace $\mathbf{s}(\nu^*)$ in the ICP problem with $\mathbf{s}^c(\varepsilon_{\text{exp}}) = (s^c(\varepsilon_{\text{exp}}, \sigma_1), \ldots, s^c(\varepsilon_{\text{exp}}, \sigma_K))$, where $s^c(\varepsilon_{\text{exp}}, \sigma_k)$ is the midpoint of the interval $\mathcal{I}(\varepsilon_{\text{exp}}, \sigma_k)$.

Proof. It is clear that ν^* satisfies the constraints; hence the feasible region is nonempty. This shows the existence of $\nu_{\min}$. We further note that if $\nu_{\min} \equiv \nu^*$ then $\nu_{\min} \in \mathcal{R}$ by Proposition 14; otherwise, we have $s(\nu^*, \sigma_k) \in [s_N^-(\nu_{\min}, \sigma_k), s_N^+(\nu_{\min}, \sigma_k)]$ by the constraints on $\nu_{\min}$ and $s(\nu^*, \sigma_k) \in \mathcal{I}(\epsilon_{\exp}, \sigma_k)$, and hence $[s_N^-(\nu_{\min}, \sigma_k), s_N^+(\nu_{\min}, \sigma_k)] \cap \mathcal{I}(\epsilon_{\exp}, \sigma_k) \neq \emptyset, 1 \leq k \leq K$. This proves $\nu_{\min} \in \mathcal{R}$. □

Solution of the ICP problem is certainly not easy due to its constraints. In actual practice, we solve the bound-constrained minimization problem instead

$$\nu_{\min}^b = \arg\min_{\nu \in \mathcal{D}^\nu} \|\mathbf{s}_N(\nu) - \mathbf{s}(\nu^*)\| \tag{8.38}$$

and see if $\nu_{\min}^b$ is in $\mathcal{R}$. Furthermore, it is not necessary to solve the problem for the minimizer; rather than we make use of the search mechanism provided by optimization procedures to obtain a necessary point $\nu_c \in \mathcal{R}$. The essential observation is that during iterative optimization process, the current iterate $\bar{\nu}_{\min}^b$ may satisfy $\bar{\nu}_{\min}^b \in \mathcal{R}$ at some early stage even before the minimizer $\nu_{\min}^b$ is actually found. Hence, instead of the minimizers $\nu_{\min}$ or $\nu_{\min}^b$, ν_c is in fact any iterate $\bar{\nu}_{\min}^b$ residing in $\mathcal{R}$. If there is no such point found by this "trickery", we turn back to solve the ICP problem by the technique proposed in [104].

There are a few issues facing by our construction algorithm. First, the solution region $\mathcal{R}$ may not be completely constructed if it is not "star-shaped" with respect to ν_i.[3] To remedy this problem, we may restart the algorithm with another or more initial centers to map out the missing boundary of $\mathcal{R}$. Second, $\mathcal{R}$ may be non-connected. We may need to perform extensive search for multiple initial centers resided in different non-connected subregions and construct the non-connected region $\mathcal{R}$ with these initial centers. Thirdly, in high dimensional space, constructing $\mathcal{R}$ is numerically expensive and representing it by a discrete set of points is geometrically difficult. A continuous region like the smallest ellipsoid or more conservatively the smallest box containing $\mathcal{R}$ is needed. The advantages are that the ellipsoid or box is geometrically visible in higher than three dimension and is much less expensive to be formed.

[3]Note by definition that the region U is called star-shaped if there is a point $p \in U$ such that line segment $\overline{pq}$ is contained in U for all $q \in U$; we then say U is star-shaped with respect to p.

8.4.3 Bounding Ellipsoid of The Possibility Region

Now given a discrete set of J points $R_J = \{\nu_j, \ldots, \nu_J\}$ representing the boundary of $\mathcal{R}$, the smallest volume ellipsoid $\mathcal{E}(B, \nu_0)$ containing R_J is found from the following minimization

$$\begin{aligned} \text{(MSE)} \qquad & \text{minimize}_{B,\nu_0} \quad -\ln(\det(B)) \\ & (\nu_j - \nu_0)B(\nu_j - \nu_0) \leq 1, \quad j = 1, \ldots, J \\ & B \text{ is SPD} . \end{aligned}$$

By factoring $B = A^2$ and letting $y = -A\nu_0$, we can transform the MSE problem into a simpler convex minimization

$$\begin{aligned} \text{(CMP)} \quad & \text{minimize}_{A,y} \quad -\ln(\det(A)) \\ & \|A\nu_j + y\| \leq 1, \quad j = 1, \ldots, J \\ & A \text{ is SPD} . \end{aligned}$$

This problem can be solved efficiently by methods of semi-definite programming. We refer to [138] for a detailed description of the primal-dual path-following algorithm which is used here for the solution of the CMP problem.

8.4.4 Bounding Box of the Possibility Region

The smallest ellipsoid $\mathcal{E}$ constructed on the finite set R_J is in some sense not conservative, i.e., may not include entirely the continuous region $\mathcal{R}$. To address this potential issue, we introduce the smallest box bounding the solution region $\mathcal{R}$ as

$$\mathcal{B} \equiv \prod_{m=1}^{M} \left[\nu_{(m)}^{\min}, \nu_{(m)}^{\max}\right] = \prod_{m=1}^{M} \left[\nu_{(m)}^{\min}, \nu_{(m)}^{\min} + \Delta\nu_{(m)}\right] , \tag{8.39}$$

where for $m = 1, \ldots, M$, $\Delta\nu_{(m)} = \nu_{(m)}^{\max} - \nu_{(m)}^{\min}$ denotes the m^{th} length of the bounding box $\mathcal{B}$ and

$$\nu_{(m)}^{\min} = \min_{\nu \in \mathcal{R}} \nu_{(m)}, \qquad \nu_{(m)}^{\max} = \max_{\nu \in \mathcal{R}} \nu_{(m)} ; \tag{8.40}$$

which can be expressed more explicitly as

$$
\begin{aligned}
\text{(MIP)} \qquad & \text{minimize}_{\nu}\ \nu_{(m)} \\
& s_N(\nu,\sigma_k) + \Delta_N^s(\nu,\sigma_k) \geq s(\nu^*,\sigma_k) - \varepsilon_{\exp}|s(\nu^*,\sigma_k)|, \quad k = 1,\ldots,K \\
& s_N(\nu,\sigma_k) - \Delta_N^s(\nu,\sigma_k) \leq s(\nu^*,\sigma_k) + \varepsilon_{\exp}|s(\nu^*,\sigma_k)|, \quad k = 1,\ldots,K \\
& \nu \in \mathcal{D}^{\nu}\ ,
\end{aligned}
$$

$$
\begin{aligned}
\text{(MAP)} \qquad & \text{maximize}_{\nu}\ \nu_{(m)} \\
& s_N(\nu,\sigma_k) + \Delta_N^s(\nu,\sigma_k) \geq s(\nu^*,\sigma_k) - \varepsilon_{\exp}|s(\nu^*,\sigma_k)|, \quad k = 1,\ldots,K \\
& s_N(\nu,\sigma_k) - \Delta_N^s(\nu,\sigma_k) \leq s(\nu^*,\sigma_k) + \varepsilon_{\exp}|s(\nu^*,\sigma_k)|, \quad k = 1,\ldots,K \\
& \nu \in \mathcal{D}^{\nu}\ .
\end{aligned}
$$

Solution method for these minimization and maximization problems have been discussed in [104] in which the authors developed the gradient and Hessian of $s_N(\nu,\sigma)$ and $\Delta_N^s(\nu,\sigma)$ and incorporated them into a trust-region sequential quadratic programing implementation of interior-point methods to obtain global (at least local) optimizers. However, the monotonicity of the objectives allows us to pursue a simple descent derivative-free strategy for solution of these problems. The idea is to find and follow feasible descent directions until there is no such direction found.[4] The following algorithm guarantees (at least local) optimal solutions for the m^{th} MIP or MAP problem.

1. Starting with the center ν_i and a feasible descent direction $d^0 = (0,\ldots,d^0_{(m)},\ldots,0)$ where $d^0_{(m)} = -1$ for the m^{th} MIP problem and $d^0_{(m)} = 1$ for the m^{th} MAP problem, we conduct a binary chop to find the associated boundary point ν_b^0.

2. If there is no feasible descent direction at ν_b^0, then ν_b^0 solves the m^{th} MIP/MAP problem. Otherwise, we set $k = 1$.

3. Starting from the boundary point ν_b^{k-1}, we find a feasible descent direction d^k and conduct a binary chop along d^k to find a new boundary point ν_b^k.

4. We set $k = k + 1$ and repeat step (3) until there is no feasible descent direction

[4] Note that a direction d is said to be feasible at a point $\nu \in \mathcal{R}$ if there exists a small $\delta > 0$ such that $\nu + \delta d \in \mathcal{R}$ and is said to be descent if the objective is decreased with respect to minimization or increased with respect to maximization when traveling along that direction.

found at the current boundary point ν_b^k. The point ν_b^k is thus the solution of the m^{th} MIP/MAP problem.

The algorithm correctly finds the global optima in the case of convex region $\mathcal{R}$. Since however $\mathcal{R}$ is usually nonconvex, multi-start strategy in which multiple-local optima are sought from multiple initial centers should be effectively used to find the global optimizer.

We emphasize that any fast forward solver other than the reduced-basis output bound methods can be used to construct the solution possibility region $\mathcal{R}$, the bounding ellipsoid $\mathcal{E}$, or the bounding box $\mathcal{B}$. However, the reliable fast evaluations provided by the reduced-basis output bound methods permit us to conduct a much more extensive search over parameter space. More importantly, $\mathcal{R}$ *rigorously captures the uncertainty* due to both the numerical approximation and experimental measurement in our prediction of the unknown parameter without *a priori* regularization hypotheses. Of course, our search over possible parameters will never be truly exhaustive, and hence there may be small undiscovered "pockets of possibility"; nevertheless, we have certainly reduced the uncertainty relative to more conventional approaches. Needless to say, our procedure can also only *characterize* the unknown parameters within our selected low-dimensional parametrization; but, more general null hypotheses can be constructed to *detect* model deviation.

8.5 Analyze-Asses-Act Approach

The inverse problem is to predict the true but "unknown" parameter ν^* from experimental measurements $\mathcal{I}(\epsilon_{\text{exp}}, \sigma_k)$ (with experimental error ϵ_{exp}) corresponding to several values of experimental control variable $\sigma_k, 1 \leq k \leq K$. In practice, more than just the inverse problem, we often face the following questions: What values of experimental control variable σ should be used to produce sensitive experimental data that are useful to the prediction of all possible unknown parameters? Can we provide solutions of the inverse problem effectively in real-time even with significant noise in experimental measurements and how we deal with this uncertainty? How we use our inverse solutions meaningfully and in particular how we act upon them to tackle engineering design and optimization

problems? To address these questions in a reliable, robust, real-time fashion we employ the Analyze-Assess-Act approach.

In particular, we extend our inverse computational method for the adaptive design and robust optimization of critical components and systems. The essential innovations are threefold. The first innovation addresses pre-experimental phase (the first question): application of the reduced-basis approximation to analyze system characteristics to determine which ranges of experimental control variable may produce sensitive data. The second innovation addresses numerical efficiency and fidelity, as well as model uncertainty (the second question): application of our robust parameter estimation method to identify (all) system configurations consistent with the available experimental data. The third innovation addresses real-time and uncertain decision problems (the third question): efficient and reliable minimization of mission objectives over the configuration possibility region to provide an intermediate and fail-safe action.

Our discussion here is merely a proof of concept; many further improvements and more efficient algorithmic implementation are possible and will leave for future work.

8.5.1 Analyze Stage

In the Analyze stage, we aim to address the first question. Poor choice of experimental control variable may lead to unacceptable (or even wrong) prediction, while careful choice will substantially improve the result. To begin, we assume that we are given a number of experimental control variable values $\Pi_I = \{\sigma_i, 1 \le i \le I\}$.[5] Next we pick a "nominal" point $\bar{\nu}$ and solve the forward problem to simulate the associated "numerical" data $\mathcal{I}(\epsilon_{\text{exp}}, \sigma_i) = [s(\bar{\nu}, \sigma_i) - \epsilon_{\text{exp}}|s(\bar{\nu}, \sigma_i)|, s(\bar{\nu}, \sigma_i) + \epsilon_{\text{exp}}|s(\bar{\nu}, \sigma_i)|], 1 \le i \le I$. We then apply the inverse algorithm to obtain a set of possibility regions $\mathcal{R}_i, 1 \le i \le I$,

$$\mathcal{R}_i = \Big\{ \nu \in \mathcal{D}^\nu \,|\, s_N(\nu, \sigma_i) \subset \mathcal{I}(\epsilon_{\text{exp}}, \sigma_i) \Big\}, \quad 1 \le i \le I \ . \tag{8.41}$$

[5] In practice, the set Π_I can be obtained from many sources including knowledge of the problem, pre-experimental analysis, modal analysis, and engineers' experiences.

We finally choose in Π_I a smallest subset, $\Pi_K = \{\sigma_k, 1 \leq k \leq K\}$, that satisfies an "Intersection" Condition

$$\bigcap_{\{k \mid \sigma_k \in \Pi_K\}} \mathcal{R}_k = \bigcap_{i=1}^{I} \mathcal{R}_i \ , \tag{8.42}$$

here the k-th element in Π_K may not necessarily be the k-th element in Π_I. It is important to note that in constructing the above possibility regions, we require only the reduced-basis approximation $s_N(\mu)$, the associated offline is thus not computationally extensive. Therefore, Π_I is allowed to be very large so that $\bigcap_{i=1}^{I} \mathcal{R}_i$ is suitably small.

We emphasize that since we use the synthetic numerical data associated with the particular parameter $\bar{\nu}$ to perform the "pre-analysis", our choice of experimental control variable is thus particularly good to the prediction of unknown parameters near the nominal point $\bar{\nu}$. More generally, our Analyze stage can accept synthetic numerical data from many different nominal points so that the resulting set of experimental control variable is useful to the prediction of not one but all possible unknown parameters.

8.5.2 Assess Stage

In our attempt to address the second question, we consider the Assess stage: Given experimental measurements, $\mathcal{I}_(\epsilon_{\text{exp}}, \sigma_k), 1 \leq k \leq K$, we wish to determine a region $\mathcal{P} \in \mathcal{D}^{\nu}$ in which the true — but unknown — parameter, ν^*, must reside. Essentially, the Assess stage is the inverse problem formulation (8.11) and can thus be addressed efficiently by our inverse computational method in which a region $\mathcal{R}$ is constructed very inexpensively such that $\nu^* \in \mathcal{P} \subset \mathcal{R}$.

8.5.3 Act Stage

We finally consider the Act stage as a way to address the last question. We presume here that our objective is the "real-time" verification of a "safety" demand about whether $s(\nu^*, \bar{\sigma})$ exceeds a specified value s_{max}, where $\bar{\sigma}$ is a specific value of the "design" variable. (For simplicity, we use σ as both the design variable and the experimental control variable; in actual practice, the design variable can be different from the experimental control variable.) Of course, in practice, we will not be privy to ν^*. To address this difficulty we

first define

$$s_{\mathcal{R}}^{+} = \max_{\nu \in \mathcal{R}} s_N^{+}(\nu, \bar{\sigma}) \ , \tag{8.43}$$

where $s_N^{+}(\nu, \bar{\sigma}) = s_N(\nu, \bar{\sigma}) + \Delta_N^{s}(\bar{\sigma}, \nu)$; our corresponding "go/no-go" criterion is then given by $s_{\mathcal{R}}^{+} \leq s_{\max}$. It is readily observed that $s_{\mathcal{R}}^{+}$ rigorously accommodates both experimental and numerical uncertainty — $s(\nu^*, \bar{\sigma}) \leq s_{\mathcal{R}}^{+}$ — and that the associated go/no-go discriminator is hence *fail-safe*.

Needless to say, depending on particular applications and specific targets, other optimization statements (such as, in the APO strategy, bilevel optimization problems) over the possibility region $\mathcal{R}$ (more precisely, the ellipsoid containing $\mathcal{R}$) with additional constraints are also possible.

Chapter 9

Nondestructive Evaluation

9.1 Introduction

Nondestructive evaluation has played a significant role in the structural health monitoring of aeronautical, mechanical, and industrial systems (e.g., aging aircraft, oil and gas pipelines, and nuclear power plant etc.). There are several theoretical, computational and/or experimental techniques [49, 81, 83, 79, 82, 2, 136, 25] devoted to the assessment and characterization of fatigue cracks and regions of material loss in manufactured components. However, in almost all cases, the techniques are expensive due to the presence of uncertainty and number of computational tasks required.

Our particular interest — or certainly the best way to motivate our approach — is in "deployed" systems: components or processes that are in service, in operation, or in the field. For example, we may be interested in assessment, evolution, and accommodation of a crack in a critical component of an in-service jet engine. Typical computational tasks include pre-experimental sensitivity analysis, robust parameter estimation (inverse problems), and adaptive design (optimization problems): in the first task — for example, selection of good exciting frequencies — we must determine appropriate values of experimental control parameters used to obtain experimental data; in the second task — for example, assessment of current crack length and location — we must deduce inputs representing system characteristics based on outputs reflecting measured observables; in the third task — for example, prescription of allowable load to meet safety demands and economic/time constraints — we must deduce inputs representing control variables

based on outputs reflecting current process objectives. These demanding activities must support an action in the presence of continually evolving environmental and mission parameters. The computational requirements are thus formidable: the entire computation must be *real-time*, since the action must be *immediate*; the entire computation must be *robust* since the action must be *safe* and *feasible*.

In this chapter, we apply the robust real-time parameter estimation method developed in the previous chapter for deployed components/systems arising in nondestructive testing. In particular, the method is employed to permit rapid and reliable characterization of crack and damage in a two-dimensional thin plate even in the presence of significant experimental errors. Numerical results are also presented throughout to test the method and confirm its advantages over traditional approaches.

9.2 Formulation of the Helmholtz-Elasticity

Inverse analysis based on the Helmholtz-elasticity PDE can gainfully serve in nondestructive evaluation, including crack characterization [64, 81, 83] and damage assessment [79, 82]. In this section, we first introduce the governing equations of the linear Helmholtz-Elasticity problem; we then reformulate the problem in terms of a reference (parameter-independent) domain. In this and the following sections, our notation is that repeated physical indices imply summation, and that, unless otherwise indicated, indices take on the values 1 through d, where d is the dimensionality of the problem. Furthermore, we use a tilde to indicate a general dependence on the parameter μ (e.g., $\tilde{\Omega} \equiv \Omega(\mu)$, or $\tilde{u} \equiv u(\mu)$) particularly when formulating the problem in an original (parameter-dependent) domain.

9.2.1 Governing Equations

We consider an elastic body $\tilde{\Omega} \in \mathbb{R}^d$ with (scaled) density unity subject to an oscillatory force of frequency $\tilde{\omega}$. We recall in Section 2.3 that under the assumption that the displacement gradients are small compared to unity, the equations governing the dynamical response of the linear elastic body are expressed as

$$\frac{\partial \tilde{\sigma}_{ij}}{\partial \tilde{x}_j} + \tilde{b}_i + \tilde{\omega}^2 \tilde{u}_i = 0 \quad \text{in } \tilde{\Omega} \,, \tag{9.1}$$

$$\tilde{\sigma}_{ij} = \tilde{C}_{ijkl}\tilde{\varepsilon}_{kl} \ , \tag{9.2}$$

$$\tilde{\varepsilon}_{kl} = \frac{1}{2}\left(\frac{\partial \tilde{u}_k}{\partial \tilde{x}_l} + \frac{\partial \tilde{u}_l}{\partial \tilde{x}_k}\right) . \tag{9.3}$$

For simplicity we consider isotropic materials, though our methods are in fact applicable to general anisotropic and nonlinear materials, $\tilde{C}_{ijkl}(\tilde{u}; \tilde{x}; \mu)$. The isotropic elasticity tensor thus has the form

$$\tilde{C}_{ijkl} = \tilde{c}_1 \delta_{ij}\delta_{kl} + \tilde{c}_2 \left(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}\right) \ ; \tag{9.4}$$

where $\tilde{c}_1$ and $\tilde{c}_2$ are the Lame elastic constants, related to Young's modulus, $\tilde{E}$, and Poisson's ratio, $\tilde{\nu}$, as follows

$$\tilde{c}_1 = \frac{\tilde{E}\tilde{\nu}}{(1+\tilde{\nu})(1-2\tilde{\nu})} \ , \qquad \tilde{c}_2 = \frac{\tilde{E}}{2(1+\tilde{\nu})} \ . \tag{9.5}$$

Due to the symmetry of $\tilde{\sigma}_{ij}$, $\tilde{\varepsilon}_{kl}$ and isotropy, the elasticity tensor satisfies

$$\tilde{C}_{ijkl} = \tilde{C}_{jikl} = \tilde{C}_{ijlk} = \tilde{C}_{klij} \ . \tag{9.6}$$

It thus follows from (9.2), (9.3), and (9.6) that

$$\sigma_{ij} = C_{ijkl}\frac{\partial u_k}{\partial x_l} \ . \tag{9.7}$$

Substituting (9.7) into (9.1) yields governing equations for the displacement $\tilde{u}$ as

$$\frac{\partial}{\partial \tilde{x}_j}\left(\tilde{C}_{ijkl}\frac{\partial \tilde{u}_k}{\partial \tilde{x}_l}\right) + \tilde{b}_i + \tilde{\omega}^2 \tilde{u}_i = 0 \quad \text{in } \tilde{\Omega} \ . \tag{9.8}$$

The displacement and traction boundary conditions are given by

$$\tilde{u}_i = 0 \ , \quad \text{on } \tilde{\Gamma}_{\mathrm{D}} \ , \tag{9.9}$$

and

$$\tilde{C}_{ijkl}\frac{\partial \tilde{u}_k}{\partial \tilde{x}_l}\,\tilde{\mathrm{e}}^n_j = \tilde{t}_i \ , \quad \text{on } \tilde{\Gamma}_{\mathrm{N}} \ , \tag{9.10}$$

where $\tilde{\mathrm{e}}^n$ is the unit normal vector on the boundary $\tilde{\Gamma}$; $\tilde{\Gamma}_D$ and $\tilde{\Gamma}_N$ are (disjoint) portions of the boundary; and $\tilde{t}_i$ are specified boundary stresses. Note that we consider homogeneous Dirichlet conditions for the sake of simplicity.

9.2.2 Weak Formulation

To derive the weak form of the governing equations, we first introduce a function space

$$\tilde{X} = \{\tilde{v} \in \left(H^1(\tilde{\Omega})\right)^d \ | \ \tilde{v}_i = 0 \text{ on } \tilde{\Gamma}_{\mathrm{D}}\} \ , \tag{9.11}$$

and associated norm

$$||\tilde{v}||_{\tilde{X}} = \left(\sum_{i=1}^{d} ||\tilde{v}_i||^2_{H_1(\tilde{\Omega})}\right)^{1/2} . \tag{9.12}$$

Next multiplying (9.8) by a test function $\tilde{v} \in \tilde{X}$ and integrating by parts we obtain

$$\int_{\tilde{\Omega}} \frac{\partial \tilde{v}_i}{\partial \tilde{x}_j}\tilde{C}_{ijkl}\frac{\partial \tilde{u}_k}{\partial \tilde{x}_l} - \tilde{\omega}^2 \int_{\tilde{\Omega}} \tilde{u}_i\tilde{v}_i - \int_{\tilde{\Gamma}} \tilde{C}_{ijkl}\frac{\partial \tilde{u}_k}{\partial \tilde{x}_l}\tilde{\mathrm{e}}^n_j\tilde{v}_i - \int_{\tilde{\Omega}} \tilde{b}_i\tilde{v}_i = 0 \ . \tag{9.13}$$

It thus follows from (9.10) and $\tilde{v} \in \tilde{X}$ that the displacement field $\tilde{u} \in \tilde{X}$ satisfies

$$\tilde{a}(\tilde{u}, \tilde{v}) = \tilde{f}(\tilde{v}) \ , \quad \forall \ \tilde{v} \in \tilde{X} \ , \tag{9.14}$$

where

$$\tilde{a}(\tilde{w}, \tilde{v}) = \int_{\tilde{\Omega}} \frac{\partial \tilde{v}_i}{\partial \tilde{x}_j}\tilde{C}_{ijkl}\frac{\partial \tilde{w}_k}{\partial \tilde{x}_l} - \tilde{\omega}^2\tilde{w}_i\tilde{v}_i \ ; \tag{9.15}$$

$$\tilde{f}(\tilde{v}) = \int_{\tilde{\Omega}} \tilde{b}_i\tilde{v}_i + \int_{\tilde{\Gamma}_{\mathrm{N}}} \tilde{v}_i\tilde{t}_i \ . \tag{9.16}$$

Now we generalize the results to inhomogeneous bodies $\tilde{\Omega}$ consisting of $\tilde{R}$ homogeneous subdomains $\tilde{\Omega}^{\tilde{r}}$ such that

$$\overline{\tilde{\Omega}} = \bigcup_{\tilde{r}=1}^{\tilde{R}} \overline{\tilde{\Omega}}^{\tilde{r}} \ ; \tag{9.17}$$

here $\overline{\tilde{\Omega}}$ is the closure of $\tilde{\Omega}$. By using similar arguments and taking into account additional displacement and traction continuity conditions at the interfaces between the $\tilde{\Omega}^{\tilde{r}}$, $1 \leq \tilde{r} \leq \tilde{R}$, we arrive at the weak formluation (9.14) in which

$$\tilde{a}(\tilde{w}, \tilde{v}) = \sum_{\tilde{r}=1}^{\tilde{R}} \int_{\tilde{\Omega}^{\tilde{r}}} \frac{\partial \tilde{v}_i}{\partial \tilde{x}_j} \tilde{C}^{\tilde{r}}_{ijkl} \frac{\partial \tilde{w}_k}{\partial \tilde{x}_l} - \tilde{\omega}^2 \tilde{w}_i \tilde{v}_i \ , \tag{9.18}$$

$$\tilde{f}(\tilde{v}) = \sum_{\tilde{r}=1}^{\tilde{R}} \int_{\tilde{\Omega}^{\tilde{r}}} \tilde{b}^{\tilde{r}}_i \tilde{v}_i + \int_{\tilde{\Gamma}^{\tilde{r}}_{\rm N}} \tilde{v}_i \tilde{t}^{\tilde{r}}_i \ ; \tag{9.19}$$

here $\tilde{C}^{\tilde{r}}_{ijkl}$ is the elasticity tensor in $\tilde{\Omega}^{\tilde{r}}$, and $\tilde{\Gamma}^{\tilde{r}}_{\rm N}$ is the section of $\tilde{\Gamma}_{\rm N}$ in $\overline{\tilde{\Omega}}^{\tilde{r}}$.

9.2.3 Reference Domain Formulation

We further partition the subdomains $\tilde{\Omega}^{\tilde{r}}$, $\tilde{r} = 1, \ldots, \tilde{R}$, into a total of R subdomains $\tilde{\Omega}^r$, $r = 1, \ldots, R$. We then map each subdomain $\tilde{\Omega}^r$ to a pre-defined reference subdomain Ω^r via a one-to-one continuous (assumed to exist) transformation $\mathcal{G}^r(\tilde{x}; \mu)$: for any $\tilde{x} \in \tilde{\Omega}^r$, its image $x \in \Omega^r$ is given by

$$x = \mathcal{G}^r(\tilde{x}; \mu) \ . \tag{9.20}$$

We further assume that the corresponding inverse mapping $(\mathcal{G}^r)^{-1}$ is also one-to-one and continuous such that for any $x \in \Omega^r$, there is uniquely $\tilde{x} \in \tilde{\Omega}^r$ where

$$\tilde{x} = (\mathcal{G}^r)^{-1}(x; \mu) \ . \tag{9.21}$$

A reference domain Ω can then be defined as $\overline{\Omega} = \bigcup_{r=1}^{R} \overline{\Omega}^r$; and hence for any $\tilde{x} \in \tilde{\Omega}$, its image $x \in \Omega$ is given by

$$x = \mathcal{G}(\tilde{x}; \mu) \ . \tag{9.22}$$

where $\mathcal{G}(\tilde{x}; \mu) : \tilde{\Omega} \rightarrow \Omega$, a compose of the $\mathcal{G}^r(\tilde{x}; \mu)$, is also a one-to-one continuous mapping. We can thus write for $1 \leq r \leq R$,

$$\frac{\partial}{\partial \tilde{x}_i} = \frac{\partial x_j}{\partial \tilde{x}_i} \frac{\partial}{\partial x_j} = \frac{\partial \mathcal{G}^r_j(\tilde{x}; \mu)}{\partial \tilde{x}_i} \frac{\partial}{\partial x_j} = G^r_{ji}(x; \mu) \frac{\partial}{\partial x_j} \ ; \tag{9.23}$$

for $\tilde{x} \in \tilde{\Omega}^r$, and

$$d\tilde{\Omega}^r = J^r(x;\mu)\, d\Omega^r, \quad d\tilde{\Gamma}^r = J_s^r(x;\mu)\, d\Gamma^r \ . \tag{9.24}$$

Here $G_{ji}^r(x;\mu)$ is obtained by substituting $\tilde{x}$ from (9.21) into $\partial \mathcal{G}_j^r(\tilde{x};\mu)/\partial \tilde{x}_i$; $J^r(x;\mu)$ is the Jacobian of the transformation $\mathcal{G}^r : \tilde{\Omega}^r \to \Omega^r$; and $J_s^r(x;\mu)$ is determined by

$$J_s^r(x;\mu) = \begin{vmatrix} \frac{\partial \tilde{y}^r}{\partial y^r} & \frac{\partial \tilde{y}^r}{\partial z^r} \\ \frac{\partial \tilde{z}^r}{\partial y^r} & \frac{\partial \tilde{z}^r}{\partial z^r} \end{vmatrix} , \tag{9.25}$$

where $(\tilde{y}^r, \tilde{z}^r)$ and (y^r, z^r) — functions of spatial coordinate x and the parameter μ — are surface coordinates associated with $\tilde{\Gamma}^r$ and Γ^r, respectively. See Section 2.2 for the definitions of the above quantities.

We now define a function space X in terms of the reference domain Ω as

$$X = \{v \in \left(H^1(\Omega)\right)^d \mid v_i = 0 \text{ on } \Gamma_{\mathrm{D}}\} \ ; \tag{9.26}$$

clearly, for any function $\tilde{w} \in \tilde{X}$, there is a unique function $w \in X$ such that $w(x) = \tilde{w}(\mathcal{G}^{-1}(x;\mu))$, and vice versa. It thus follows that the displacement field $u \in X$ corresponding to $\tilde{u} \in \tilde{X}$ satisfies

$$a(u,v) = f(v), \quad \forall\ v \in X \ , \tag{9.27}$$

where

$$a(w,v) = \sum_{r=1}^{R} \int_{\Omega^r} \left\{ \frac{\partial v_i}{\partial x_j} C_{ijk\ell}^r(x;\mu) \frac{\partial w_k}{\partial x_\ell} - \omega^2 w_i v_i \right\} J^r(x;\mu), \tag{9.28}$$

$$f(v) = \sum_{r=1}^{R} \int_{\Omega^r} b_i^r v_i J^r(x;\mu) + \int_{\Gamma_{\mathrm{N}}^r} v_i t_i^r J_s^r(x;\mu) \ ; \tag{9.29}$$

here $C_{ijk\ell}^r(x;\mu)$, the elasticity tensor in the reference domain, is given by

$$C_{ijk\ell}^r(x;\mu) = G_{jj'}^r(x;\mu) \tilde{C}_{ijk\ell}^r G_{\ell\ell'}^r(x;\mu) \ . \tag{9.30}$$

Finally, we observe that when the geometric mappings $\mathcal{G}^r(\tilde{x};\mu), r = 1, \ldots, R$, are

affine such that

$$\mathcal{G}^r(\tilde{x};\mu) = G^r(\mu)\tilde{x} + g^r(\mu) \ ; \tag{9.31}$$

our bilinear form a is affine in μ, since G^r and J^r depend only on μ, not on x.

9.3 The Inverse Crack Problem

9.3.1 Problem Description

We revisit the two-dimensional thin plate with a horizontal crack described thoroughly in Sections 4.6.1 and 5.1.3. Recall that our input is $\mu \equiv (\mu_1, \mu_2, \mu_3) = (\omega^2, b, L)$, where ω is the frequency of oscillatory uniform force applied at the right edge, b is the crack location, and L is the crack length. The forward problem is that for any input parameter μ, we evaluate the output $s(\mu)$ which is the (oscillatory) amplitude of the average vertical displacement on the right edge of the plate. The inverse problem is to predict the true but "unknown" crack parameter $(b^*, L^*) \in \mathcal{D}^{b,L}$ from experimental data

$$\mathcal{I}(\epsilon_{\text{exp}}, \omega_k^2) = [s(\omega_k^2, b^*, L^*) - \epsilon_{\text{exp}}|s(\omega_k^2, b^*, L^*)|, s(\omega_k^2, b^*, L^*) + \epsilon_{\text{exp}}|s(\omega_k^2, b^*, L^*)|], 1 \le k \le K \ .$$

Recall that ϵ_{exp} is experimental error, and K is number of measurements.

More broadly and practically, we shall focus our attention on the following questions: What value of frequencies should be used to obtain sensitive experimental data that yields good prediction of all possible unknown crack parameters in consideration? Can we provide rapid predictions even in facing significant error in experimental measurements and how we deal with this uncertainty? Can the cracked thin plate withstand an in-service steady force such that the deflection does not exceed a specified value? To address these questions in a real-time yet reliable and robust fashion, we employ the Analyze-Assess-Act approach developed in the previous chapter.

9.3.2 Analyze Stage

We first perform modal analysis to select a set of candidate frequencies. In particular, we display in Figure 9-1 natural frequencies in the first six modes as a function of b and

L. We observe that the natural frequencies in the first three modes are invariant with b and L, which indicates that frequencies in the range of these modes may not be a good choice. We begin to see some variation from the fourth mode onward. We may hence suggest $\Pi_I = \{2.8, 3.2, 4.8\}$ which are a set of frequencies squared in the frequency region between the third mode and the fifth mode.

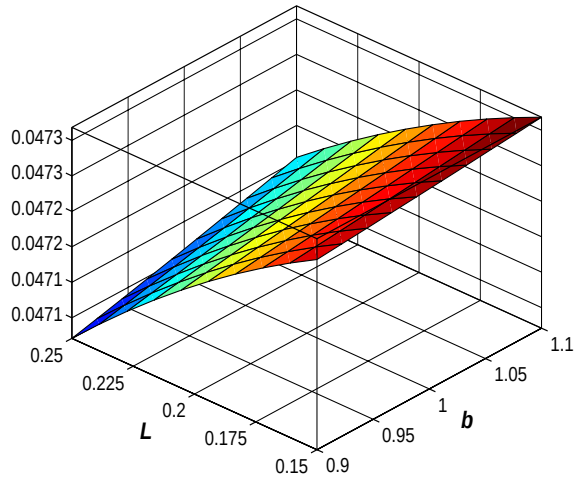
We next consider a "nominal" point $(\bar{b}, \bar{L}) = (1.0, 2.0)$ and present in Figure 9-2 possibility regions $\mathcal{R}_i$, $1 \le i \le I$, associated with Π_I. We see that two subsets $\Pi_{K_1} = \{2.8, 4.8\}$ or $\Pi_{K_2} = \{3.2, 4.8\}$ are equally good because their intersection regions $\bigcap_{\{k \,|\, \omega_k^2 \in \Pi_{K_1}\}} \mathcal{R}_k$ and $\bigcap_{\{k \,|\, \omega_k^2 \in \Pi_{K_2}\}} \mathcal{R}_k$ are very small and almost coincide with $\bigcap_{i=1}^{3} \mathcal{R}_i$ which is the shaded region. However, we will choose the second subset for illustrating the subsequent Assess and Act stages.

As an additional note, we observe that no frequency alone can identify well the unknown parameter (b^*, L^*); and that only good choice and good combination of frequencies result in good prediction (for example, the subset $\Pi_{K_3} = \{2.8, 3.2\}$ gives unacceptably large possibility region, while its counterparts, Π_{K_1} and Π_{K_2}, produce reasonably small regions).

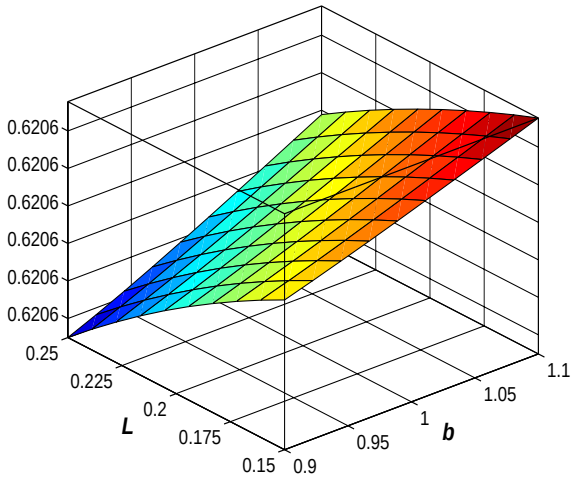
9.3.3 Assess Stage

Having determined the appropriate frequencies, we employ the inverse algorithm introduced in Chapter 8 to construct $\mathcal{R}$ defined by 8.37. Here we could develop two different reduced-basis models for each of frequencies $\omega_1^2 = 3.2$ and $\omega_1^2 = 4.8$ over a smaller parameter domain $\mathcal{D}^{b,L} \equiv [0.9, 1.1] \times [0.15, 0.25]$ to achieve more economical online cost. However, we shall reuse the reduced-basis model that was developed in Chapter 5 for the problem with the parameter domain $\mathcal{D} \equiv (\omega^2 \in [3.2, 4.8]) \times (b \in [0.9, 1.1]) \times (L \in [0.15, 0.25])$, because small error tolerance ϵ_{tol} can be satisfied with very small N (recall that $N_{\max} = 32$). In Figure 9-3 we plot $\mathcal{R}$ and $\mathcal{E}$ for $\epsilon_{\text{exp}} = 0.5\%, 1\%, 5\%$ for two test cases $(b^*, L^*) = (1.0, 0.2)$ and $(b^*, L^*) = (1.05, 0.17)$. We furthermore tabulate in Table 9.1 the half lengths of $\mathcal{B}$ relative to the exact (synthesis) value (b^*, L^*).

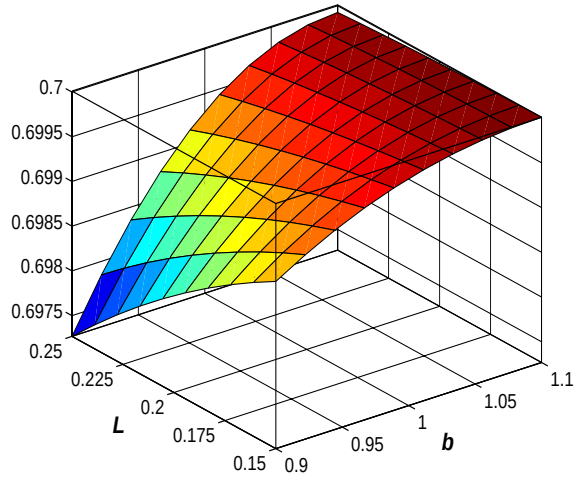
As expected, as ϵ_{exp} decreases, $\mathcal{R}$ shrinks towards the exact (synthetic) value, b^*, L^*. We further observe that the half lengths of $\mathcal{B}$ relative to the exact value are about order



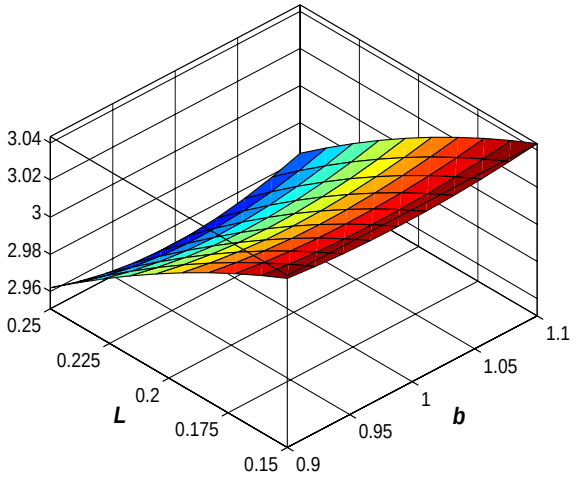
(a) First mode



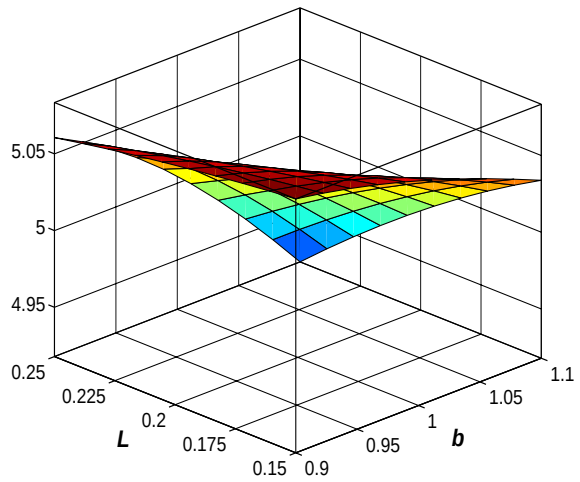
(b) Second mode



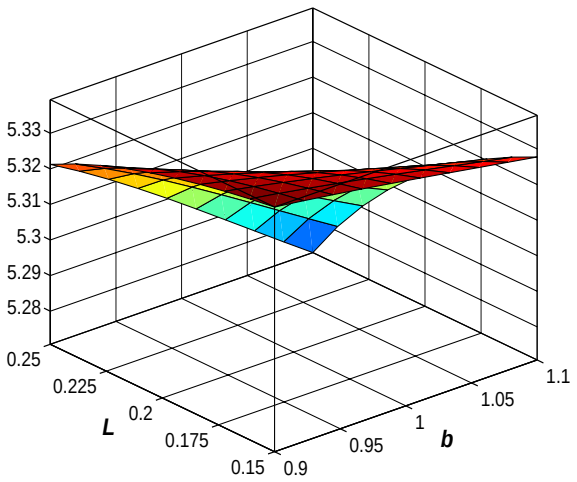
(c) Third mode



(d) Fourth mode



(e) Fifth mode



(f) Sixth mode

Figure 9-1: Natural frequencies of the cracked thin plate as a function of b and L. The vertical axis in the graphs is the natural frequency squared.

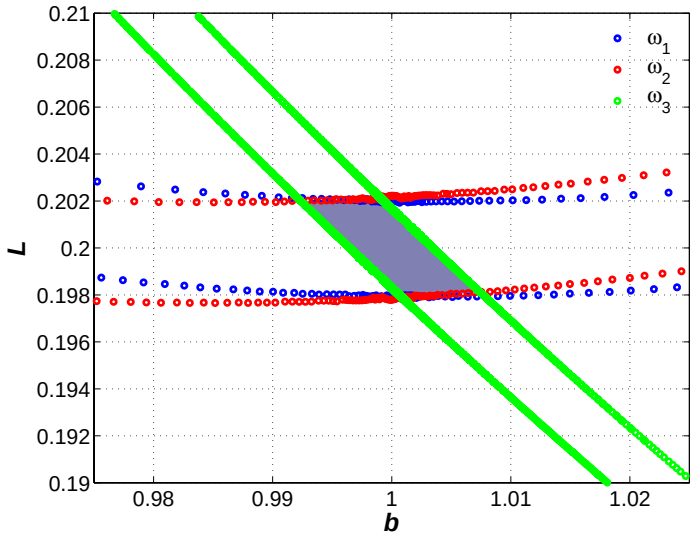
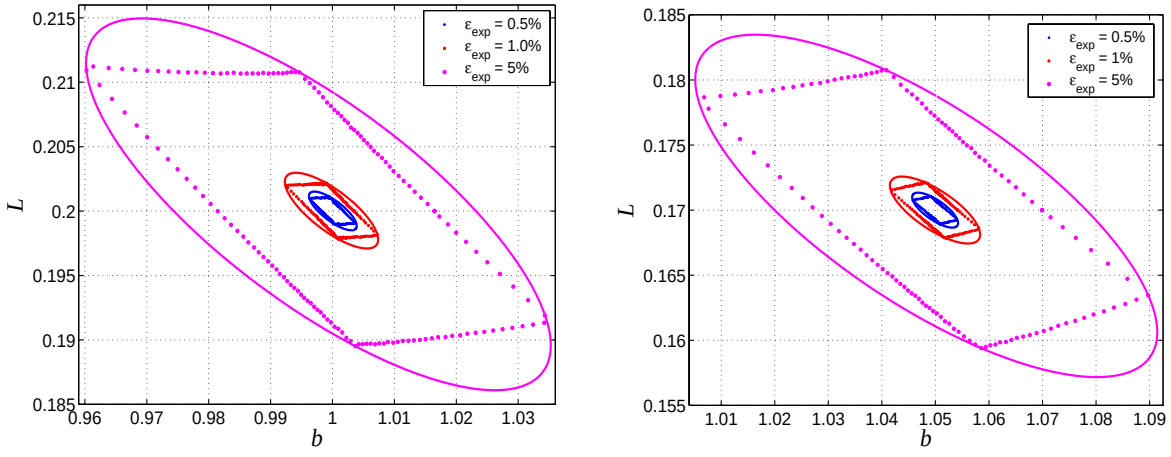


Figure 9-2: Possibility regions $\mathcal{R}_i$ for $\omega_1^2 = 2.8$, $\omega_2^2 = 3.2$, $\omega_3^2 = 4.8$ and $\epsilon_{\text{exp}} = 1.0\%$.



(a) (b)

Figure 9-3: Crack parameter regions $\mathcal{R}$ and $\mathcal{E}$ obtained with $N = 24$: (a) $(b^*, L^*) = (1.0, 0.2)$ and (b) $(b^*, L^*) = (1.05, 0.17)$.

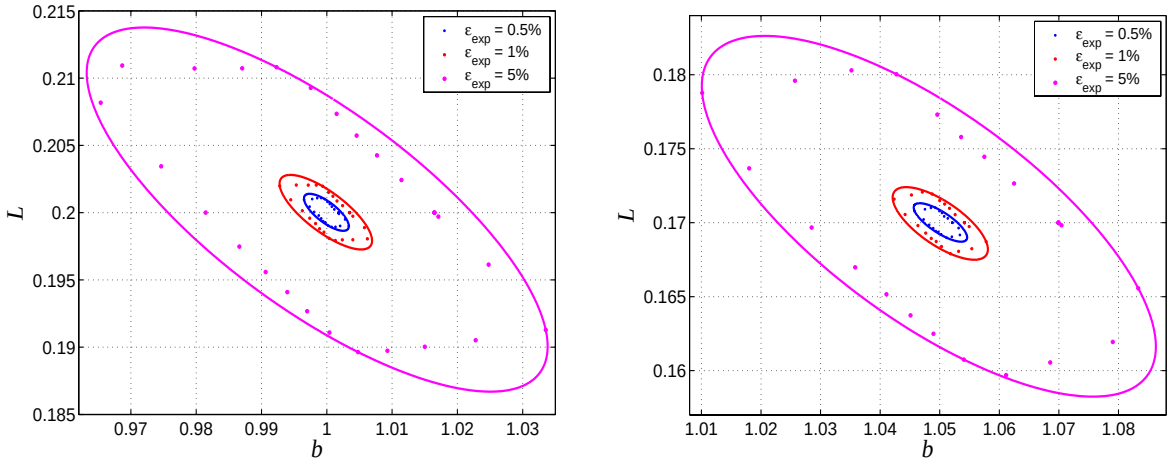
	Test case (a)		Test case (b)	
ϵ_{exp}	$0.5\Delta b/b^*$	$0.5\Delta L/L^*$	$0.5\Delta b/b^*$	$0.5\Delta L/L^*$
5.0%	3.70%	5.42%	3.95%	6.27%
1.0%	0.74%	1.09%	0.78%	1.26%
0.5%	0.37%	0.55%	0.39%	0.63%

Table 9.1: The half lengths of the bounding box $\mathcal{B}$ relative to b^*, L^* for $\epsilon_{\text{exp}} = 5.0\%, 1\%, 0.5\%$.

of ε_{exp} and increases linearly with ε_{exp}. The problem is thus linearly ill-posed for the two test cases. The crucial capability here is reliable and fast evaluations that permit us to conduct a much more extensive search over parameter space: for a given ϵ_{exp}, $\mathcal{R}$ requires 2650 forward evaluations and is generated online in less than 38 seconds on a Pentium 1.6 GHz, while $\mathcal{B}$ can be found less than 11 seconds for 740 forward evaluations. More importantly, for any finite ϵ_{exp}, $\mathcal{R}$ *rigorously captures both the model uncertainty and numerical error* in our assessment of the crack parameters without *a priori* regularization hypotheses. Furthermore, since the relative output bound is much less than experimental error of 0.5% (recall in Section 5.4 that for $N = 24$ our relative output bound is essentially less than 5.0×10^{-5}), the experimental error greatly dominates the numerical error in our numerical results; we could hence even achieve faster parameter estimation response — at little cost in precision — by decreasing N to balance the experimental and numerical error. And $\mathcal{R}$ is almost indistinguishable from $\mathcal{P}$; however, the latter is about 350 times more expensive than the former.

In addition, we present in Figure 9-4 $\mathcal{E}$ obtained with much coarser region $\mathcal{R}$ for the same test cases. We observe that although R_J in Figure 9-4 is considerably fewer than that in Figure 9-3, the ellipses in 9-4 are slightly smaller than those in Figure 9-3; hence, they do not cover entirely the possibility region $\mathcal{R}$ and thus solution region $\mathcal{P}$. Note however that the number of forward evaluations and time to construct $\mathcal{E}$ now is dropped by approximately a factor of 4 to 274 and 12.75 seconds, respectively.

We see that for this particular example that good results have been obtained for different unknown parameters even with only one nominal point used for the Analyze stage. Nevertheless, our search over possible crack parameters will never be truly exhaustive, and hence there may be small undiscovered "pockets of possibility" in $\mathcal{D}^{b,L}$; however, we have certainly reduced the uncertainty relative to more conventional approaches. Needless to say, the procedure can also only *characterize* cracks within our selected low-dimensional parametrization; however, more general null hypotheses (for future work) can be constructed to *detect* model deviation.



(a) (b)

Figure 9-4: Crack parameter regions $\mathcal{R}$ and $\mathcal{E}$ obtained with $N = 24$: (a) $(b^*, L^*) = (1.0, 0.2)$ and (b) $(b^*, L^*) = (1.05, 0.17)$.

9.3.4 Act Stage

Finally, we consider the Act stage. We presume here that the component must withstand an in-service steady force (normalized to unity) such that the deflection $s(0, b^*, L^*)$ in the "next mission" does not exceed a specified value $s_{\max}$ $(= 0.95)$; of course, in practice, we will not be privy to (b^*, L^*). To address this difficulty we first define

$$s^+_{\mathcal{R}} = \max_{(b,L)\in\mathcal{R}} s^+_N(0, b, L) \ , \tag{9.32}$$

where $s^+_N(0, b, L) = s_N(0, b, L) + \Delta^s_N(0, b, L)$; our corresponding "go/no-go" criterion is then given by $s^+_{\mathcal{R}} \le s_{\max}$. It is readily observed that $s^+_{\mathcal{R}}$ rigorously accommodates both experimental (crack) and numerical uncertainty — $s(0, b^*, L^*) \le s^+_{\mathcal{R}}$ — and that the associated go/no-go discriminator is hence *fail-safe*.

Before presenting numerical results, we note that the Act stage is essentially steady linear elasticity — $\omega^2 = 0$ — and the problem is thus coercive and relatively easy; we shall thus omit the detail (indeed, for this coercive problem, we need only $N_{\max} = 6$ for $\epsilon_{\text{tol,min}} = 10^{-4}$). To be clear in our notation, we shall rename N by N^{I} in Assess stage and by N^{II} in Act stage. Our primary objective is to obtain $s^+_{\mathcal{R}}$ defined by (9.32), which is always an upper bound of $s(0, b^*, L^*)$; and hence, even under significant uncertainty we

can still provide real-time actions with some confidence. We tabulate in the Table 9.2 the ratio, $[s^+_{\mathcal{R}} - s(0, b^*, L^*)]/s(0, b^*, L^*)$, as a function of $N^{\rm I}$ and $\epsilon_{\rm exp}$ for $(b^*, L^*) = (1.0, 0.2)$ and $N^{\rm II} = 6$. We observe that as $\epsilon_{\rm exp}$ tends to zero and $N^{\rm I}$ increases, $s^+_{\mathcal{R}}$ will tend to $s(0, b^*, L^*)$, and thus we may control the sub-optimality of our "Act" decision.

$N^{\rm I}$	5.0%	1.0%	0.5%
12	1.19×10^{-3}	4.44×10^{-4}	3.25×10^{-4}
18	4.20×10^{-4}	8.07×10^{-5}	4.16×10^{-5}
24	4.07×10^{-4}	7.40×10^{-5}	3.70×10^{-5}

Table 9.2: $[s^+_{\mathcal{R}} - s(0, b^*, L^*)]/s(0, b^*, L^*)$ as a function of $N^{\rm I}$ and $\epsilon_{\rm exp}$ for $(b^*, L^*) = (1.0, 0.2)$.

In conclusion, we achieve very fast Analyze-Assess-Act calculation: Π_K may be obtained from the set Π_I in less than 31 seconds, $\mathcal{R}$ may be generated online in less than 38 seconds, and $s^+_{\mathcal{R}}$ may be computed online less than 0.93 seconds on a Pentium 1.6 GHz laptop. Hence, in real-time, we can Analyze the component to facilitate sensitive experimental data, Assess the current state of the crack and subsequently Act to ensure the safety (or optimality) of the next "sortie."

9.4 Additional Application: Material Damage

In this section, we apply our Analyze-Assess-Act approach to the rapid and reliable characterization of the location, size and type of damage in materials. The characteristics of damage in structures play a key role in defining preemptive actions in order to improve reliability and reduce life-cycle costs. It serves crucially in the structural health monitoring of aeronautical, mechanical, civil, and electrical systems. Our particular example is the prediction of the location, size and severity factor of damage in sandwich plates.

9.4.1 Problem Description

We revisit the two-dimensional thin plate with a rectangular damaged zone described thoroughly in Section 5.5. Recall that our input is $\mu \equiv (\omega^2, b, L, \delta) \in \mathcal{D}^{\omega} \times \mathcal{D}^{b,L,\delta}$, where

$\mathcal{D}^{b,L,\delta} \equiv [0.9, 1.1] \times [0.5, 0.7] \times [0.4, 0.6]$; and our output $s(\mu)$ is the (oscillatory) amplitude of the average vertical displacement on the right edge of the plate. The forward problem is that for any input parameter μ, we evaluate the output $s(\mu)$. The inverse problem is to predict the true but "unknown" damage parameter $(b^*, L^*, \delta^*) \in \mathcal{D}^{b,L,\delta}$ from experimental measurements $\mathcal{I}(\epsilon_{\text{exp}}, \omega_k^2), 1 \le k \le K$. with experimental error ϵ_{exp}. The primary goal in this example is to demonstrate new capabilities (in dealing with inverse problems) enabled by our robust parameter estimation method; and our focus is thus on the Assess stage.

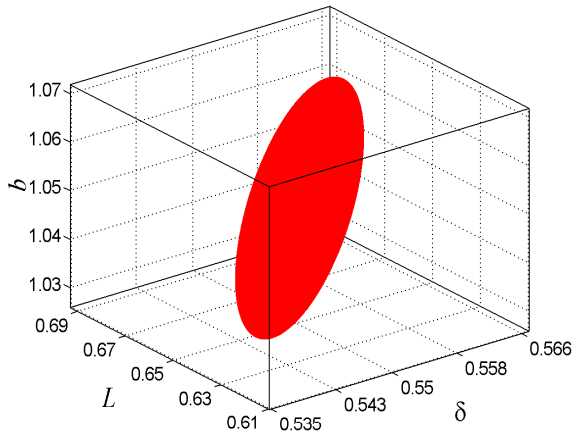
9.4.2 Numerical Results

We directly consider the Assess stage with the given set $\Pi_K = \{\omega_1^2 = 0.58, \omega_2^2 = 1.53, \omega_3^2 = 2.95\}$ which is indeed obtained by pursuing the Analyze stage. We henceforth need three different reduced-basis models for each of frequencies square: Model I for $\omega_1^2 = 0.58$, Model II for $\omega_2^2 = 1.53$, and model III for $\omega_3^2 = 2.95$. Recall that these reduced-basis models were already developed in Section 5.5 (see the section for details of the reduced-basis formulation for these models and associated numerical results).

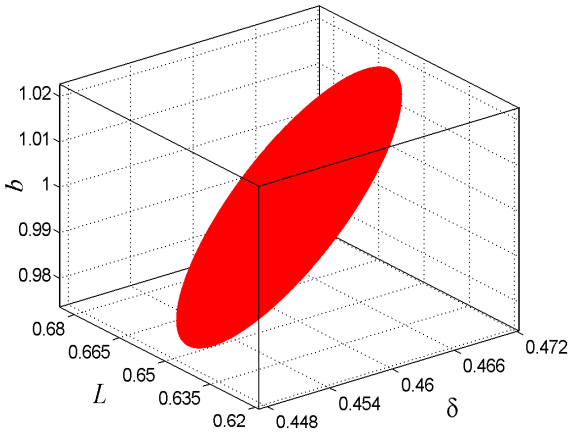
We present in Figure 9-5 the possibility regions for the location and length and damage factor of the flaw — more precisely, ellipsoids that contain the possibility regions for the location and length and damage factor — for experimental error of 2%, 1%, and 0.5% for Case 1 ($b^* = 1.05, L^* = 0.65, \delta^* = 0.55$) and Case 2 ($b^* = 1.00, L^* = 0.65, \delta^* = 0.46$). We observe that as ϵ_{exp} decreases, $\mathcal{E}$ shrinks towards the exact (synthetic) value (b^*, L^*, δ^*). The bounding ellipsoid $\mathcal{E}$ constructed from 450 boundary points requires roughly 12150 forward evaluations and 336 seconds in online.

	Test Case 1			Test Case 2		
ϵ_{exp}	$0.5\Delta b/b^*$	$0.5\Delta L/L^*$	$0.5\Delta\delta/\delta^*$	$0.5\Delta b/b^*$	$0.5\Delta L/L^*$	$0.5\Delta\delta/\delta^*$
2.0%	2.16%	5.80%	2.33%	2.37%	4.49%	2.54%
1.0%	1.07%	3.04%	1.41%	1.19%	2.44%	1.32%
0.5%	0.55%	1.39%	0.59%	0.60%	1.29%	0.71%

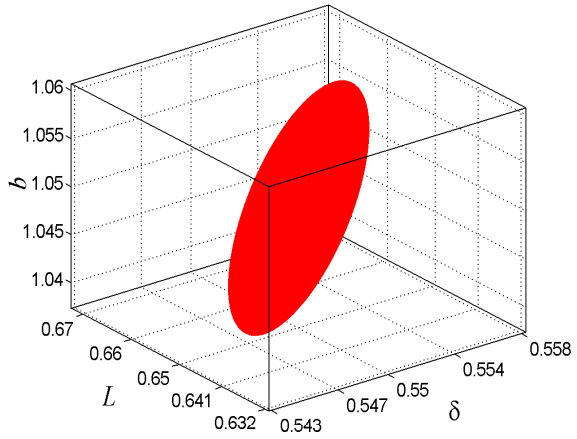
Table 9.3: The half lengths of the bounding box $\mathcal{B}$ relative to b^*, L^*, δ^* as a function of ϵ_{exp} for two test cases.



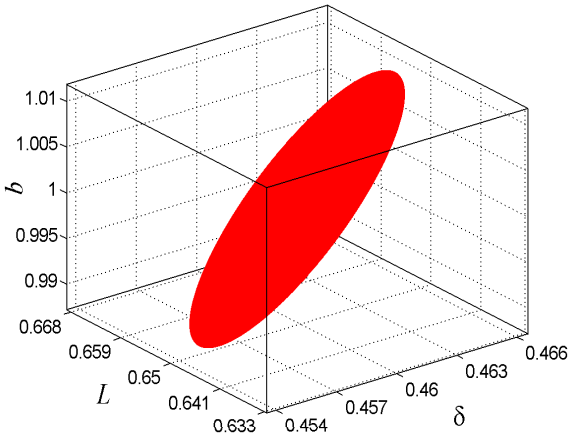
(a) $\epsilon_{\text{exp}} = 2.0\%$



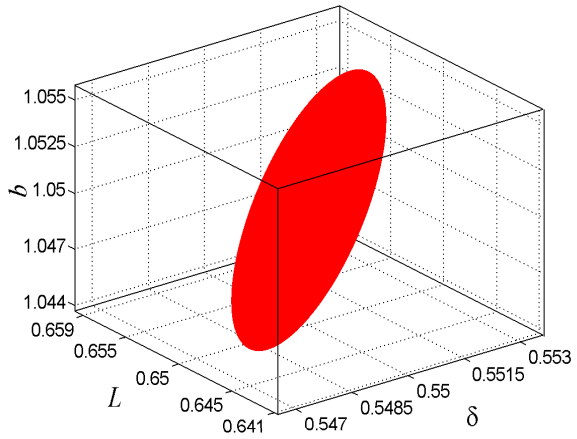
(b) $\epsilon_{\text{exp}} = 2.0\%$



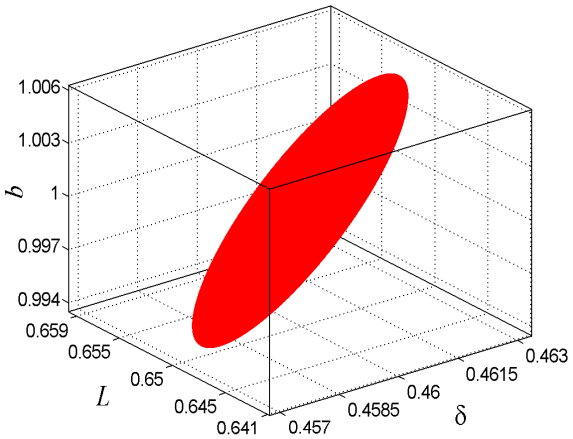
(c) $\epsilon_{\text{exp}} = 1.0\%$



(d) $\epsilon_{\text{exp}} = 1.0\%$



(e) $\epsilon_{\text{exp}} = 0.5\%$



(f) $\epsilon_{\text{exp}} = 0.5\%$

Figure 9-5: Ellipsoids containing possibility regions obtained with $N = 40$ for $b^* = 1.05, L^* = 0.65, \delta^* = 0.55$ in (a), (c), and (e) and for $b^* = 1.00, L^* = 0.65, \delta^* = 0.46$ in (b), (d), and (f). Note the change in scale in the axes: $\mathcal{E}$ shrinks as the experimental error decreases.

We also show in Table 9.3 the half lengths of the box bounding the ellipsoid relative to the exact value. Unlike $\mathcal{E}$, $\mathcal{B}$ needs approximately 2980 forward evaluations. In both test cases, the relative half lengths for b^* and δ^* are slightly larger than ε_{exp} while the relative half length for L^* is more than twice as ε_{exp}. Furthermore, the half-lengths of $\mathcal{B}$ appear to increase linearly with ε_{exp}; hence, for these test cases, the problem is linearly ill-posed. Note further that our relative output bound is much less than experimental error of 0.5% (recall in Section 5.5 that for $N = 40$ the relative output bound is less than 2.56×10^{-5} for model I, 1.71×10^{-4} for model II, 2.42×10^{-5} for model III). Therefore, the results are largely indistinguishable from those based on the direct finite element calculation of $s(\mu)$. However, the latter is much (280 times) more expensive than our method. In short, $\mathcal{B}$ — obtained online in (less than) 78 seconds on a Pentium 1.6 GHz thanks to a per forward evaluation of 0.0261 seconds – quantifies uncertainty in both the numerical approximation and experimental data without *a priori* hypotheses; we can thus provide appropriate real-time actions with some confidence.

Rather than a picture, another (or even better) way to see the possibility region is by means of numbers. In particular, we tabulate in Table 9.4 the center, half-lengths, and directions of the ellipsoids for different values of ϵ_{exp}. We see that the center is closer to the exact unknown parameter (b^*, L^*, δ^*), the half-lengths get smaller, and the directions change slightly as ϵ_{exp} decreases; indeed, these ellipsoids shrink toward the synthesis value $(1.00, 0.6, 0.5)$. The key point here is that we can see the possibility region and how it changes very clearly even in higher than three-dimensional space where it is not possible with picture.

Since our primary goal in this problem is the nondestructive assessment of material damage in a sandwich, we shall not pursue the Act stage. However, we do identify that there are two immediate situations that can be well tackled by our Act stage: (1) a real-time query in which we verify whether the structure can withstand a steady static force such that the deflection does not exceed a specific value (robust optimization); (2) a design problem in which a shim is designed to have its weight minimized while effectively strengthening the structure and maintaining the deflection at the desired level (adaptive design) [139]. In parallel to Section 9.3.4, these subproblems can be formulated as ap-

ϵ_{exp}	Center	Half-lengths	Directions		
2.0%	1.0006	0.0444	0.2826	0.3345	-0.8990
	0.5988	0.0202	0.9492	0.0373	0.3123
	0.4966	0.0060	-0.1380	0.9417	0.3069
1.0%	0.9998	0.0216	0.2817	0.3578	-0.8903
	0.5998	0.0098	0.9487	0.0353	0.3143
	0.4999	0.0030	-0.1439	0.9331	0.3295
0.5%	0.9999	0.0110	0.2790	0.3693	-0.8864
	0.5999	0.0048	0.9482	0.0400	0.3151
	0.5000	0.0015	-0.1518	0.9284	0.3390

Table 9.4: The center, half-lengths, and directions of $\mathcal{E}$ for $(b^*, L^*, \delta^*) = (1.00, 0.60, 0.50)$ as ϵ_{exp} decreases.

propriate optimization problems over the possibility region (or preferably the ellipsoid containing it) constructed by the Assess stage.

9.5 Chapter Summary

In this chapter we have applied the "Analyze-Assess-Act" approach developed in the previous chapter to the assessment of crack and damage in a two-dimensional thin plate. Although characterized by simple physical model and geometry, these problems and related numerical results show strong advantages of our approach over traditional methods in two key aspects. First, as regards the computational expense and uncertainty in the model, our approach is more efficient and robust: (1) real-time and reliable evaluation of functional outputs associated with the PDEs of continuum mechanics rather than time-consuming calculation by use of classical numerical approaches; and (2) robust and efficient identification of all (or almost all, in the probabilistic sense) inverse solutions consistent with the available experimental data without *a priori* regularization hypotheses rather than only one regularized inverse solution with *a priori* assumptions. Second, as regards the practical application and implementation for real-life engineering problems, our approach is more practical and effective: (1) a systematic way rather than a "trial and error" method to the selection of experimental control parameters; and (2) use of many frequencies and single-sensor measurement rather than one frequency and multiple-sensor measurements.

Chapter 10

Inverse Scattering Analysis

10.1 Introduction

Inverse scattering problem has attracted enormous interest due to its wide range of practical applications in engineering and science such as medicine, geophysics, defense science. In particular, inverse scattering problems arise in medical imaging, detection of mines, underwater surveillance, and target acquisition. In all of the abovementioned areas the common goal is to recover characteristics (like geometric measures, material properties, boundary conditions, etc.) of an interrogated object from experimental data (far-field pattern measured at distributed sensor locations) obtained by sending incident waves at the object.

The wide range of applications has stimulated the development of different solution methods for inverse scattering problems. However, the task of solving inverse scattering problems is admittedly difficult for two reasons. First, due to the incomplete and noisy data the inverse scattering problems are typically ill-posed and numerically ill-conditioned — the existence, uniqueness, and stability of the solution are not simultaneously ensured. Uniqueness theorems [70, 61, 69, 30], which are crucial for both the theoretical study and the implementation of numerical algorithms, have been long investigated often under the assumption of complete and accurate data. Restoring stability is not less important especially in practical contexts where error in measurements is intrinsically present. In order to restore stability some kind of *a priori* information is needed and regularization methods [31, 30, 62, 119, 44] making use of the available *a priori* information are often

used. Though quite sophisticated, iterative regularization methods often need additional information and thus lose algorithmic generality. Second, since the inverse scattering problems are inherently nonlinear, most methods are computationally expensive and often fail to have the numerical solution in real-time.

In this chapter we apply our robust inverse computational method developed in Chapter 8 for rapid and reliable solution of inverse scattering problems. The key and unique components of our method are reduced-basis approximations and associated *a posteriori* error estimation procedures. As we shall see, the inverse scattering problems are generally nonaffine and noncompliant. We thus need to develop a coefficient-function approximation and integrate it into a "dual-primal" formulation as described thoroughly in Section 6.6. The coefficient-function approximation and dual-primal consideration simultaneously ensure online $\mathcal{N}$ independence (efficiency), and more rapidly convergent reduced-basis approximations and better error bounds (accuracy). These advantages are further leveraged within the inverse scattering context: we can, in fact, achieve robust but efficient construction of a bounded possibility region that captures all (or almost all, in the probabilistic sense) model parametrizations consistent with the available experimental data. Numerical results for a simple two-dimensional inverse scattering problem are also presented to demonstrate these capabilities.

10.2 Formulation of the Inverse Scattering Problems

In this section we first introduce the governing equations of the inverse scattering problem. We then reformulate the problem in terms of a reference (parameter-independent) domain. In this and the following sections, our notation is that repeated physical indices imply summation, and that, unless otherwise indicated, indices take on the values 1 through n, where n is the dimensionality of the problem. Furthermore, we use a tilde to indicate a general dependence on the parameter μ (e.g., $\tilde{\Omega} \equiv \Omega(\mu)$, or $\tilde{u} \equiv u(\mu)$) particularly when formulating the problem in an original (parameter-dependent) domain.

10.2.1 Governing Equations

We consider the scattering of a time harmonic acoustic incident wave $\tilde{u}^i$ of frequency ω by a bounded object $\tilde{D} \in \mathbb{R}^n\,(n = 2, 3)$ having constant density ρ_D and constant sound speed c_D. We assume that the object $\tilde{D}$ is situated in a homogeneous isotropic medium with constant density ρ and constant sound speed c. The corresponding wave numbers are given by $k_D = \omega/c_D$ and $k = \omega/\tilde{c}$. Let $\tilde{u}$ be the scattered wave, then the total field $\tilde{u}^t = \tilde{u}^i + \tilde{u}$ and the transmitted wave $\tilde{v}$ satisfy the acoustic transmission problem

$$\Delta\tilde{u}^t + k^2\tilde{u}^t = 0 \quad \text{in } \mathbb{R}^n\backslash\tilde{D}, \tag{10.1a}$$

$$\Delta\tilde{v} + k_D^2\tilde{v} = 0 \quad \text{in } \tilde{D}, \tag{10.1b}$$

$$\tilde{u}^t = \tilde{v} \quad \text{on } \partial\tilde{D}, \tag{10.1c}$$

$$\frac{1}{\tilde{\rho}}\frac{\partial\tilde{u}^t}{\partial\tilde{\nu}} = \frac{1}{\tilde{\rho}_D}\frac{\partial\tilde{v}}{\partial\tilde{\nu}} \quad \text{on } \partial\tilde{D}, \tag{10.1d}$$

$$\lim_{\tilde{r}\to\infty} \tilde{r}^{(n-1)/2}\left(\frac{\partial\tilde{u}}{\partial\tilde{r}} - \mathrm{i}k\tilde{u}\right) = 0, \quad \tilde{r} = |\tilde{x}|\ ; \tag{10.1e}$$

where $\tilde{\nu}$ denotes the unit outward normal to the boundary and the incident field $\tilde{u}^i$ is a plane wave moving in direction $\tilde{d}$, i.e.,

$$\tilde{u}^i(\tilde{x}) = e^{\mathrm{i}k\tilde{x}\cdot\tilde{d}}, \quad |\tilde{d}| = 1\ . \tag{10.2}$$

Note that the continuity of the waves and the normal velocity across $\partial\tilde{D}$ leads to the transmission conditions (10.1c) and (10.1d), and that the scattered field satisfies the Sommerfeld radiation condition (10.1e). Mathematically, the Sommerfeld condition ensures the well-posedness of the problem (10.1); physically it characterizes out-going waves [30].

We shall consider a special case of the above transmission problem. In particular, if the object is sound-hard, i.e., $\rho_D/\rho \to \infty$, we are led to the exterior Neumann problem [29]

$$\Delta\tilde{u} + k^2\tilde{u} = 0 \quad \text{in } \mathbb{R}^n\backslash\tilde{D}, \tag{10.3a}$$

$$\frac{\partial}{\partial\tilde{\nu}}\left(\tilde{u} + \tilde{u}^i\right) = 0 \quad \text{on } \partial\tilde{D}, \tag{10.3b}$$

$$\lim_{\tilde{r}\to\infty} \tilde{r}^{(n-1)/2}\left(\frac{\partial\tilde{u}}{\partial\tilde{r}} - \mathrm{i}k\tilde{u}\right) = 0, \quad \tilde{r} = |\tilde{x}|\ . \tag{10.3c}$$

Equation (10.3c) implies that the scattered field behaves asymptotically like an outgoing spherical wave

$$\tilde{u}(\tilde{x}) = \frac{e^{\mathrm{i}k\tilde{r}}}{\tilde{r}^{(n-1)/2}}\tilde{u}_\infty(\tilde{D}, \tilde{d}^s, \tilde{d}, k) + O\left(\frac{1}{\tilde{r}^{(n+1)/2}}\right) \tag{10.4}$$

as $|\tilde{x}| \to \infty$, where $\tilde{d}^s = \tilde{x}/|\tilde{x}|$. The function $\tilde{u}_\infty$ defined on the unit sphere $\tilde{S} \subset \mathbb{R}^n$ is known as the scattering amplitude or the far-field pattern of the scattered wave. The Green representation theorem and the asymptotic behavior of the fundamental solution ensures a representation of the far-field pattern in the form

$$\tilde{u}_\infty(\tilde{D}, \tilde{d}^s, \tilde{d}, k) = \tilde{\beta}_n \int_{\partial\tilde{D}} \tilde{u}(\tilde{x}) \frac{\partial e^{-\mathrm{i}k\tilde{d}^s\cdot\tilde{x}}}{\partial\tilde{\nu}} - \frac{\partial\tilde{u}(\tilde{x})}{\partial\tilde{\nu}} e^{-\mathrm{i}k\tilde{d}^s\cdot\tilde{x}} \, , \tag{10.5}$$

with

$$\tilde{\beta}_n = \begin{cases} \frac{\mathrm{i}}{4}\sqrt{\frac{2}{\pi k}}e^{-\mathrm{i}\pi/4} & n = 2 \\ \frac{1}{4\pi} & n = 3. \end{cases} \tag{10.6}$$

The proof of (10.4) and (10.5) can be found in Appendix A.

We can now be more explicit what we mean by the acoustic inverse scattering problem. Given specific geometry $\tilde{D}$ of the object and the incident wave $\tilde{u}^i$, the forward problem is to find the scattered wave $\tilde{u}$ and in particular the far field pattern $\tilde{u}_\infty$. In contrast, the inverse problem is to determine the unknown geometry $\tilde{D}^*$ from the knowledge of the far field data $\mathcal{I}(\epsilon_{exp}, \tilde{d}^s, \tilde{d}, k)$ measured on the unit sphere $\tilde{S} := \{\tilde{x} \ : \ |\tilde{x}| = 1\}$ for one or several directions $\tilde{d}$ and wave numbers k with measurement error ϵ_{exp}.

10.2.2 Radiation Boundary Conditions

Since the problem is posed over indefinite domain, before attempting to numerically solve the problem, it is required to replace the indefinite domain with an artificial closed boundary $\tilde{\Gamma}$ enclosing the object. A boundary condition is then introduced on $\tilde{\Gamma}$ in such a way that the resulting boundary-value problem is well-posed and its solution approximates well the restriction of $\tilde{u}$ to the bounded domain $\tilde{\Omega}$ limited by $\partial\tilde{D}$ and $\tilde{\Gamma}$. From numerical point of view, there are two classes of such a boundary condition that can be used: (1) exact conditions give exactly the restriction of $\tilde{u}$ if no further approximation is made; (2) approximate (also called radiation or absorbing) boundary conditions only

yield an approximation of this restriction. Because exact conditions can only be derived for certain special cases, several approximate conditions have been developed (See [5] for generalized results on various radiation conditions including the most accurate one, namely, the second-order Bayliss-Turkel radiation condition).

For simplicity, we consider simple first-order complex-valued Robin condition; higher order approximation of the Sommerfeld radiation condition requires more sophisticated implementation and will be considered in future research. As a result, we have the direct acoustic scattering problem (10.3) being replaced with

$$\Delta\tilde{u} + k^2\tilde{u} = 0 \qquad \text{in} \quad \tilde{\Omega}, \tag{10.7a}$$

$$\frac{\partial\tilde{u}}{\partial\tilde{\nu}} = -\frac{\partial\tilde{u}^i}{\partial\tilde{\nu}} \qquad \text{on} \quad \partial\tilde{D}, \tag{10.7b}$$

$$\frac{\partial\tilde{u}}{\partial\tilde{\nu}} - \mathrm{i}k\tilde{u} + \tilde{\mathcal{H}}\tilde{u} = 0, \quad \text{on} \quad \tilde{\Gamma}\ ; \tag{10.7c}$$

where $\tilde{\mathcal{H}}$ is the mean curvature of $\tilde{\Gamma}$ (see Section 2.2 for the definition of mean curvature).

10.2.3 Weak Formulation

To derive the weak form of the problem (10.7), we first introduce a *complex* function space

$$\tilde{X} = \{\tilde{v} = \tilde{v}^{\mathrm{R}} + \mathrm{i}\tilde{v}^{\mathrm{I}} \mid \tilde{v}^{\mathrm{R}} \in H^1(\tilde{\Omega}), \tilde{v}^{\mathrm{I}} \in H^1(\tilde{\Omega})\}\ , \tag{10.8}$$

and associated inner product

$$(\tilde{w}, \tilde{v})_{\tilde{X}} = \int_{\tilde{\Omega}} \nabla\tilde{w} \cdot \nabla\bar{\tilde{v}} + \tilde{w}\bar{\tilde{v}}\ . \tag{10.9}$$

Here superscripts R and I denote the real and imaginary part, respectively; and $\bar{\tilde{v}}$ denotes the complex conjugate of $\tilde{v}$, and $|\tilde{v}|$ the modulus of $\tilde{v}$. Our weak formulation of the direct scattering problem is then: find $\tilde{u} \in \tilde{X}$ such that

$$\tilde{a}(\tilde{u}, \tilde{v}) = \tilde{f}(\tilde{v}), \quad \forall\tilde{v} \in \tilde{X}; \tag{10.10}$$

here the forms are given by

$$\tilde{a}(\tilde{w},\tilde{v}) = \int_{\tilde{\Omega}} \nabla\tilde{w}.\nabla\bar{\tilde{v}} - k^2\tilde{w}\bar{\tilde{v}} - \int_{\tilde{\Gamma}} \left(\mathrm{i}k - \tilde{\mathcal{H}}\right)\tilde{w}\bar{\tilde{v}} \ , \tag{10.11}$$

$$\tilde{f}(\tilde{v}) = \int_{\partial\tilde{D}} -\mathrm{i}k\tilde{d}\cdot\tilde{\nu}e^{\mathrm{i}k\tilde{x}\cdot\tilde{d}}\bar{\tilde{v}} \ . \tag{10.12}$$

Finally, it follows from (10.2), (10.7b), and (10.5) that the far field pattern is given by

$$\tilde{u}_\infty(\tilde{D},\tilde{d}^s,\tilde{d},k) = \beta_n \int_{\partial\tilde{D}} -\tilde{u}(\tilde{x})\mathrm{i}k\tilde{d}^s\cdot\tilde{\nu}e^{-\mathrm{i}k\tilde{d}^s\cdot\tilde{x}} + \mathrm{i}k\tilde{d}\cdot\tilde{\nu}e^{\mathrm{i}k\tilde{x}\cdot\tilde{d}}e^{-\mathrm{i}k\tilde{d}^s\cdot\tilde{x}} \ , \tag{10.13}$$

10.2.4 Reference Domain Formulation

We now define a reference domain Ω. We then map $\tilde{\Omega} \to \Omega$ by a one-to-one continuous transformation $\mathcal{G}(\tilde{x};\mu)$: for any $\tilde{x} \in \tilde{\Omega}$, its image $x \in \Omega$ is given by

$$x = \mathcal{G}(\tilde{x};\mu) \ . \tag{10.14}$$

We further assume that the corresponding inverse mapping $\mathcal{G}^{-1}$ is also one-to-one and continuous such that for any $x \in \Omega$, there is uniquely $\tilde{x} \in \tilde{\Omega}$, where

$$\tilde{x} = \mathcal{G}^{-1}(x;\mu) \ . \tag{10.15}$$

We can thus write

$$\frac{\partial}{\partial\tilde{x}_i} = \frac{\partial x_j}{\partial\tilde{x}_i}\frac{\partial}{\partial x_j} = \frac{\partial\mathcal{G}_j(\tilde{x};\mu)}{\partial\tilde{x}_i}\frac{\partial}{\partial x_j} = G_{ji}(x;\mu)\frac{\partial}{\partial x_j}, \tag{10.16}$$

$$d\tilde{\Omega} = J(x;\mu)\ d\Omega \ , \tag{10.17}$$

$$d\partial\tilde{D} = J_d(x;\mu)\ d\partial D \ , \tag{10.18}$$

$$d\tilde{\Gamma} = J_s(x;\mu)\ d\Gamma \ . \tag{10.19}$$

Here $G_{ji}(x;\mu)$ is obtained by substituting $\tilde{x}$ from (10.15) into $\partial\mathcal{G}_j(\tilde{x};\mu)/\partial\tilde{x}_i$; $J(x;\mu)$ is the Jacobian of the transformation; $J_d(x;\mu)$ is given by

$$J_d(x;\mu) = \begin{vmatrix} \frac{\partial\tilde{y}}{\partial y} & \frac{\partial\tilde{y}}{\partial z} \\ \frac{\partial\tilde{z}}{\partial y} & \frac{\partial\tilde{z}}{\partial z} \end{vmatrix} , \tag{10.20}$$

where $(\tilde{y},\tilde{z})$ and (y,z) are surface coordinates associated with $\partial\tilde{D}$ and ∂D, respectively; and $J_s(x;\mu)$ is similarly determined. See Section 2.2 for the definitions and formulas of these quantities.

Next we define the function space X in terms of the reference domain Ω as

$$X = \{v^{\mathrm{R}} + \mathrm{i}v^{\mathrm{I}} \mid v^{\mathrm{R}} \in H^1(\Omega), v^{\mathrm{I}} \in H^1(\Omega)\} , \tag{10.21}$$

in terms of which we introduce an appropriate bound conditioner

$$(w,v)_X = \int_\Omega \nabla w \cdot \nabla\bar{v} + w\bar{v} . \tag{10.22}$$

It then follows that the scattered wave $u \in X$ corresponding to $\tilde{u} \in \tilde{X}$ satisfies

$$a(u,v;x;\mu) = f(v;x;\mu), \quad \forall\ v \in X ; \tag{10.23}$$

where

$$a(w,v;x;\mu) = \int_\Omega \left(G_{ji}(x;\mu)\frac{\partial w}{\partial x_j} \cdot G_{ki}(x;\mu)\frac{\partial\bar{v}}{\partial x_k} - k^2 w\bar{v} \right) J(x;\mu)$$

$$- \int_\Gamma \left(\mathrm{i}k - \tilde{\mathcal{H}}(x;\mu) \right) w\bar{v}J_s(x;\mu) , \tag{10.24}$$

$$f(v;x;\mu) = -\int_{\partial D} \bar{v}\mathrm{i}k\tilde{d}\cdot\tilde{\nu}e^{\mathrm{i}k\tilde{x}\cdot\tilde{d}}J_d(x;\mu) . \tag{10.25}$$

The far-field pattern is then calculated as

$$u_\infty(\mu) = \ell(\overline{u}(\mu);x;\mu) + \ell^{\mathrm{o}}(x;\mu) ; \tag{10.26}$$

where ℓ takes $\bar{u}(\mu) \in X$ as its argument, but ℓ^o does not as shown below

$$\ell(v; x; \mu) = -\beta_n \int_{\partial D} \overline{v} \mathrm{i} k \tilde{d}^s \cdot \tilde{\nu} e^{-\mathrm{i} k \tilde{d}^s \cdot \tilde{x}} J_d(x; \mu), \tag{10.27}$$

$$\ell^o(x; \mu) = \beta_n \int_{\partial D} \mathrm{i} k \tilde{d} \cdot \tilde{\nu} e^{\mathrm{i} k \tilde{x} \cdot \tilde{d}} e^{-\mathrm{i} k \tilde{d}^s \cdot \tilde{x}} J_d(x; \mu) \ . \tag{10.28}$$

Finally, we note that when the geometric mapping $\mathcal{G}(\tilde{x}; \mu)$ is affine such that

$$\mathcal{G}(\tilde{x}; \mu) = G(\mu)\tilde{x} + g(\mu) \ , \tag{10.29}$$

and $J_s(x; \mu)$ and $\tilde{\mathcal{H}}(x; \mu)$ are all together affine, our bilinear form a is an affine operator.

10.2.5 Problems of Current Consideration

In this thesis, we shall only consider the inverse scattering problems in which $\tilde{D}$ is a simple geometry such as an ellipsoid (in future work, we shall apply our robust parameter estimation method to more complex curved geometries); and hence a is affine for an appropriate consideration of the truncated domain $\tilde{\Omega}$. In this case, the direct scattering problem can be restated more generally as: given $\mu \equiv (\tilde{D}, \tilde{d}^s, \tilde{d}, k) \in \mathcal{D}$, find $u_\infty(\mu) \equiv s(\mu) + \ell^o(x; \mu)$ with

$$s(\mu) = \ell(\overline{u}(\mu); h(x; \mu)) \ , \tag{10.30}$$

where $u(\mu)$ satisfies

$$a(u, v; \mu) = f(v; g(x; \mu)), \quad \forall\, v \in X \ . \tag{10.31}$$

Here $f(v; g(x; \mu))$ and $\ell(v; h(x; \mu))$ are given by

$$f(v; g(x; \mu)) \equiv -\int_{\partial D} \overline{v} g(x; \mu), \qquad \ell(v; h(x; \mu)) \equiv -\beta_n \int_{\partial D} \overline{v} h(x; \mu) \ , \tag{10.32}$$

where $g(x; \mu)$ and $h(x; \mu)$ are nonaffine functions of coordinate x and parameter μ and can be found by referring to (10.25) and (10.27)

$$g(x; \mu) = \mathrm{i} k \tilde{d} \cdot \tilde{\nu} e^{\mathrm{i} k \tilde{x} \cdot \tilde{d}} J_d(x; \mu) \ , \qquad h(x; \mu) = \mathrm{i} k \tilde{d}^s \cdot \tilde{\nu} e^{-\mathrm{i} k \tilde{d}^s \cdot \tilde{x}} J_d(x; \mu) \ . \tag{10.33}$$

Moreover, a is expressed as an affine sum of the form

$$a(w, v; \mu) = \sum_{q=1}^{Q} \Theta^q(\mu) a^q(w, v) \ , \tag{10.34}$$

for $\Theta^q : \mathcal{D} \to \mathbb{R}$ and $a^q : X \times X \to \mathbb{R}$, $1 \le q \le Q$.

We further assume that a satisfies a continuity and inf-sup condition

$$0 < \beta(\mu) \equiv \inf_{w \in X} \sup_{v \in X} \frac{a(w, v; \mu)}{\|w\|_X \|v\|_X} \ , \tag{10.35}$$

$$\sup_{w \in X} \sup_{v \in X} \frac{a(w, v; \mu)}{\|w\|_X \|v\|_X} \equiv \gamma(\mu) < \infty \ . \tag{10.36}$$

Here $\beta(\mu)$ is the Babuška "inf-sup" constant and $\gamma(\mu)$ is the standard continuity constant.

Because f (respectively, ℓ) depends on a general nonaffine function g (respectively, h) of μ and x, and $\ell \neq f$, such problem is nonaffine and noncompliant. Fortunately, effective and efficient reduced-basis formulation has been developed in Section 6.6 to treat this class of problems. In particular, by constructing the coefficient-function approximation and incorporating it into the primal-dual reduced-basis formulation, we can recover online independence and fast output convergence and better output effectivity.

10.3 A Simple Inverse Scattering Problem

10.3.1 Problem Description

We consider the scattering of a time harmonic acoustic incident wave $\tilde{u}^i(\tilde{x}) = e^{\mathrm{i}k\tilde{x}\cdot\tilde{d}}$ moving in direction $\tilde{d}$ by an infinite cylinder with bounded cross section $\tilde{D}$, where k is the wave number of the incident plane wave $\tilde{u}^i$. As a simple demonstration we consider an two-dimensional ellipse of unknown major semiaxis a (half length of the major axis), unknown minor semiaxis b (half length of the minor axis), and unknown orientation α for $\tilde{D}$. The input μ thus consists of $k, \tilde{d}, \tilde{d}^s, a, b,$ and α in which $(a, b, \alpha) \in \mathcal{D}^{a,b,\alpha} \equiv \times[0.5, 1.5] \times [0.5, 1.5] \times [0, \pi]$ are characteristic-system parameters and $(k, \tilde{d}, \tilde{d}^s) \in \mathcal{D}^{k,\tilde{d},\tilde{d}^s} \equiv [\pi/8, \pi/8] \times [0, 2\pi] \times [0, 2\pi]$ are experimental control parameters; and the output of interest

is the far-field pattern $\tilde{u}_\infty$. More specifically, we define $\mu \equiv (\mu_1, \mu_2, \mu_3, \mu_4, \mu_5, \mu_6) \in \mathcal{D} \subset \mathbb{R}^6$ as: $\mu_1 = ka, \mu_2 = b/a, \mu_3 = \alpha$, μ_4 is radian angle of incident wave (i.e., $\tilde{d} = (\cos\mu_4, \sin\mu_4)$), μ_5 is radian angle of the far-field pattern (i.e., $\tilde{d}^s = (\cos\mu_5, \sin\mu_5)$), and $\mu_6 = k$; note that the first five parameters are nondimensional.

In addition, the truncated domain $\tilde{\Omega}$ is bounded by the elliptical boundary $\partial\tilde{D}$ and an artificial boundary $\tilde{\Gamma}$. Here $\tilde{\Gamma}$ is an oblique rectangle of size $10a \times 10b$ which has the same orientation as the ellipse and is scaled with the major and minor semiaxes as shown in Figure 10-1(a). We define a reference domain corresponding to the geometry bounded by a unit circle ∂D and a square of size 10×10 as shown in Figure 10-1(b). We then map $\tilde{\Omega}(a, b, \alpha) \to \Omega$ via a continuous piecewise-affine transformation. The geometric mapping is simply rotation and scaling as given below

$$G(\mu) = \begin{bmatrix} \cos\alpha/a & \sin\alpha/a \\ -\sin\alpha/b & \cos\alpha/b \end{bmatrix}, \qquad g(\mu) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{10.37}$$

Furthermore, we have

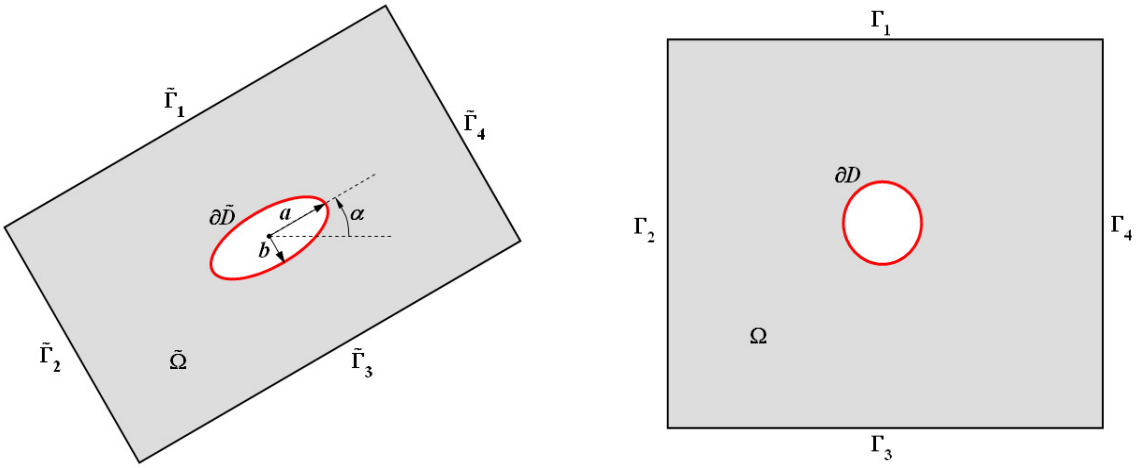
$$J(x;\mu) = \det G^{-1}(\mu), \qquad J_s(x;\mu) = |G^{-1}(\mu)\tau|, \qquad J_d(x;\mu) = |G^{-1}(\mu)\tau^o| \; ; \tag{10.38}$$

here $\tau = (\tau_1, \tau_2)$ and $\tau^o = (\tau_1^o, \tau_2^o)$ are the unit vectors tangent to the boundary Γ and ∂D, respectively. Note further that the mean curvature $\tilde{\mathcal{H}}(\tilde{x};\mu)$ is zero for the chosen boundary $\tilde{\Gamma}$ except for the corner points. We can thus ignore this term in our formulation of the direct scattering problem.

The problem can now be recast precisely in the desired abstract form (10.31), in which Ω, X defined in (10.21), and $(w;v)_X$ defined in (10.22) are independent of the parameter μ; and our affine assumption (10.34) applies for $Q = 5$. We summarize the $\Theta^q(\mu), a^q(w,v), 1 \le q \le Q$, in Table 10.1.

To derive the explicit form for $g(x;\mu)$, $h(x;\mu)$, and $\ell^o(x;\mu)$, we first need the unit normal vector to the elliptic boundary $\partial\tilde{D}$ in the reference coordinate

$$\tilde{\nu}_1(x;\mu) = (bx_1\cos\alpha - ax_2\sin\alpha) \,/\, \sqrt{b^2x_1^2 + a^2x_2^2} \; , \tag{10.39}$$



(a) (b)

Figure 10-1: Two-dimensional scattering problem: (a) original (parameter-dependent) domain and (b) reference domain.

q	$\Theta^q(\mu)$	$a^q(w,v)$
1	μ_2	$\int_\Omega \frac{\partial w}{\partial x_1}\frac{\partial \bar{v}}{\partial x_1}$
2	$\frac{1}{\mu_2}$	$\int_\Omega \frac{\partial w}{\partial x_2}\frac{\partial \bar{v}}{\partial x_2}$
3	$-\mu_1^2\mu_2$	$\int_\Omega w\bar{v}$
4	$-\mathrm{i}\mu_1$	$\int_{\Gamma_1} w\bar{v} + \int_{\Gamma_3} w\bar{v}$
5	$-\mathrm{i}\mu_1\mu_2$	$\int_{\Gamma_2} w\bar{v} + \int_{\Gamma_4} w\bar{v}$

Table 10.1: Parametric functions $\Theta^q(\mu)$ and parameter-independent bilinear forms $a^q(w,v)$ for the two-dimensional inverse scattering problem.

$$\tilde{\nu}_2(x;\mu) = (bx_1 \sin\alpha + ax_2 \cos\alpha) \,/\sqrt{b^2x_1^2 + a^2x_2^2}\ . \tag{10.40}$$

Since ∂D is the unit circle, from (10.38) we have

$$J_d(x;\mu) = \sqrt{b^2x_1^2 + a^2x_2^2}\ . \tag{10.41}$$

It finally follows from (10.28), (10.33), and (10.37)-(10.41) that

$$\begin{aligned} g(x;\mu) &= \mathrm{i}\mu_1\left((\mu_2 x_1 \cos\mu_3 - x_2 \sin\mu_3)\cos\mu_4 + (\mu_2 x_1 \sin\mu_3 + x_2\cos\mu_3)\sin\mu_4\right) \\ &\qquad e^{\mathrm{i}\mu_1((x_1\cos\mu_3 - \mu_2 x_2 \sin\mu_3)\cos\mu_4 + (x_1\sin\mu_3 + \mu_2 x_2\cos\mu_3)\sin\mu_4)}\ , \end{aligned} \tag{10.42}$$

$$\begin{aligned} h(x;\mu) &= \mathrm{i}\mu_1\left((\mu_2 x_1 \cos\mu_3 - x_2 \sin\mu_3)\cos\mu_5 + (\mu_2 x_1 \sin\mu_3 + x_2\cos\mu_3)\sin\mu_5\right) \\ &\qquad e^{-\mathrm{i}\mu_1((x_1\cos\mu_3 - \mu_2 x_2 \sin\mu_3)\cos\mu_5 + (x_1\sin\mu_3 + \mu_2 x_2\cos\mu_3)\sin\mu_5)}\ , \end{aligned} \tag{10.43}$$

$$\begin{aligned} \ell^o(x;\mu) &= \beta_n \int_{\partial D} \mathrm{i}\mu_1\left((\mu_2 x_1 \cos\mu_3 - x_2 \sin\mu_3)\cos\mu_4 + (\mu_2 x_1 \sin\mu_3 + x_2\cos\mu_3)\sin\mu_4\right) \\ &\qquad e^{\mathrm{i}\mu_1((x_1\cos\mu_3 - \mu_2 x_2 \sin\mu_3)\cos\mu_4 + (x_1\sin\mu_3 + \mu_2 x_2\cos\mu_3)\sin\mu_4)} \\ &\qquad e^{-\mathrm{i}\mu_1((x_1\cos\mu_3 - \mu_2 x_2 \sin\mu_3)\cos\mu_5 + (x_1\sin\mu_3 + \mu_2 x_2\cos\mu_3)\sin\mu_5)}\ . \end{aligned} \tag{10.44}$$

10.3.2 Numerical results

We first show in Figure (10-2) a few FEM solutions for slightly different parameters. We observe that changing only one component of the parameter results in a dramatical change in both solution structure and magnitude. This will create approximation difficulty in the reduced-basis method, and thus it may require N large to achieve sufficient accuracy. Recall that the reduced-basis approximation and associated *a posteriori* error estimators for the direct scattering problem were developed in Section 6.6; also see the section for related numerical results including the convergence and effectivities, rigorousness of our error bounds, as well as computational savings relative to the finite element method.

We can now turn to the inverse problem that illustrates the new capabilities enabled by rapid certified input-output evaluation. In particular, given limited aperture far-field data in the form of intervals $\mathcal{I}(\epsilon_{\text{exp}}, k, \tilde{d}, \tilde{d}^s)$ obtained at several angles $\tilde{d}^s$ for several directions $\tilde{d}$ and fixed wave number k, we wish to determine a region $\mathcal{R} \in \mathcal{D}^{a,b,\alpha}$ in which the true but unknown parameter, (a^*, b^*, α^*), must reside; recall that ϵ_{exp} is experimental

error. In our numerical experiments, we set the wave number fixed to $k = \pi/8$ and use three different directions $\{0, \pi/4, \pi/2\}$ for the incident wave. For each direction of the incident wave, there are I angles, $\{(i-1)\pi/2, 1 \leq i \leq I\}$, at which the far-field data are obtained; hence, the number of measurements is $K = 3 \times I$. We tabulate in Table 10.2 the half lengths of the bounding box $\mathcal{B}$ relative to the exact value as a function of the experimental error and the number of measurements for $a^* = 1.35, b^* = 1.15, \alpha^* = \pi/2$.

	ϵ_{exp}	$K = 3$	$K = 6$	$K = 9$	$K = 12$
$0.5\Delta a/a^*$	5.0	2.59	2.50	2.44	2.44
	2.0	1.04	1.01	0.98	0.98
	1.0	0.52	0.50	0.49	0.49
$0.5\Delta b/b^*$	5.0	18.70	2.44	2.40	2.40
	2.0	9.72	0.97	0.97	0.96
	1.0	3.83	0.49	0.49	0.48
$0.5\Delta\alpha/\alpha^*$	5.0	31.57	4.18	2.27	2.24
	2.0	16.94	1.82	0.82	0.82
	1.0	6.34	0.89	0.40	0.40

Table 10.2: The half lengths of $\mathcal{B}$ relative to $a^* = 1.35, b^* = 1.15, \alpha^* = \pi/2$ vary with ε_{exp} and K. Note that the results shown in the table are percentage values.

We observe that as ϵ_{exp} decreases and K increases, $\mathcal{B}$ shrinks toward (a^*, b^*, α^*) and that the number of measurements have strongly different impact on the identification of the unknown parameters. For $K = 3$, the relative half-length of a^* is smaller than the experimental error, but those of b^* and α^* are significantly larger than the experimental error. As K increases to 6, $\mathcal{B}$ shrinks very rapidly in the order of $O(10)$ along the b-axis and α-axis. Further increasing K to 9 leads to the improvement for only α^*. Meanwhile, the bounding boxes for $K = 9$ and $K = 12$ are almost the same; this implies that the experimental error dominates at $K = 9$ at which $\mathcal{B}$ no longer shrinks with K increasing. Hence, for this particular instantiation, we should use $K = 9$ at which the relative half-lengths are less than one-half of the experimental error.

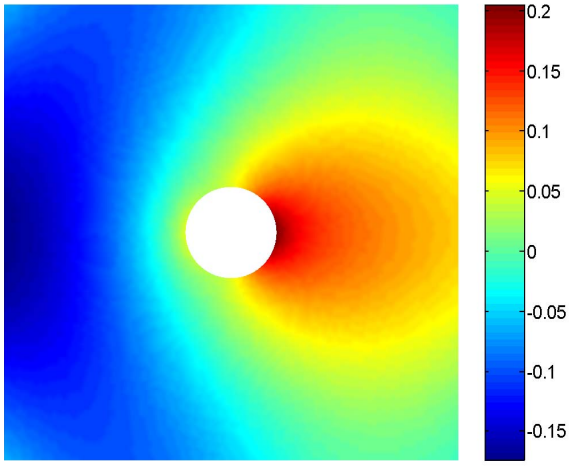
We perform another test for the case of $a^* = 1.2, b^* = 0.8, \alpha^* = 3\pi/4$ and show the results in Table 10.3. We see again that $\mathcal{B}$ begins to saturate at $K = 9$. The implication is that the ill-posedness of the problem also depends on the number of measurements. In both cases, the relative half-lengths are increasing linearly with the experimental error for

sufficiently large K ($K \geq 6$); the problem is thus linearly ill-posed in both test cases. Most importantly, these results not only quantify robustly uncertainty in both the numerical approximation and measurements, but also are obtained within $9 \times K$ seconds on a Pentium 1.6 GHz laptop thanks to a per forward evaluation time of only 0.008 seconds and yield significantly computational savings for our method compared to conventional approaches.

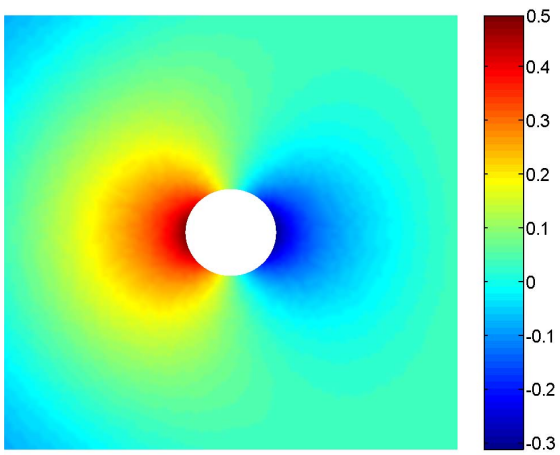
	ϵ_{exp}	$K = 3$	$K = 6$	$K = 9$	$K = 12$
$0.5\Delta a/a^*$	5.0	7.18	2.26	2.30	2.30
	2.0	3.28	0.90	0.92	0.92
	1.0	2.15	0.45	0.46	0.46
$0.5\Delta b/b^*$	5.0	17.28	6.46	4.51	4.51
	2.0	11.09	2.58	1.75	1.75
	1.0	6.24	1.29	0.88	0.88
$0.5\Delta\alpha/\alpha^*$	5.0	11.04	2.69	2.42	2.20
	2.0	6.48	1.08	0.96	0.88
	1.0	3.74	0.54	0.48	0.44

Table 10.3: The half lengths of $\mathcal{B}$ relative to $a^* = 1.2, b^* = 0.8, \alpha^* = 3\pi/4$ vary with ε_{exp} and K. Note that the results shown in the table are percentage values.

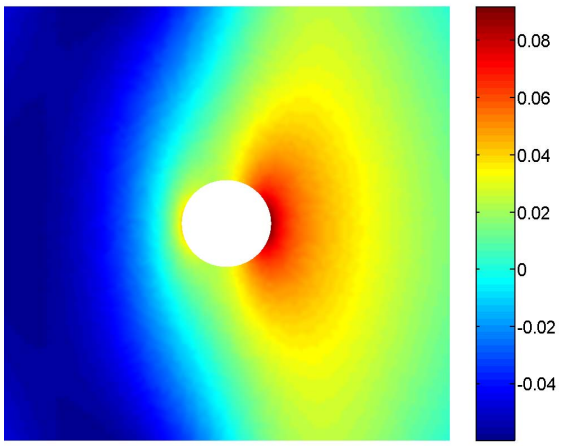
In addition, we plot $\mathcal{E}$ in Figure 10-3 and tabulate $\mathcal{B}$ in Table 10.4 for $a^* = 0.85, b^* = 0.65, \alpha^* = \pi/4$ at values of $\epsilon_{\text{exp}} = 5\%, 2\%, 1\%$ and $K = 6, 9$; note that $\mathcal{E}$ (constructed from 450 points) and $\mathcal{B}$ require roughly $4150 \times K$ and $1060 \times K$ forward evaluations, respectively. We see that $\mathcal{E}$ is not only more expensive, but also less conservative than $\mathcal{B}$ since $\mathcal{E}$ does not include entirely $\mathcal{R}$. It should also be noted that these results are obtained with $N = 40$ for which the largest relative output bounds are about than 1.0% (see Table 6.6). Therefore, the results here are largely indistinguishable from those obtained by using the finite element method, but yield a factor of $O(100)$ in computational savings for our method. Of course, our search over all possible parameters will never be truly exhaustive, and hence there may be undiscovered "pockets of possibility" in $\mathcal{D}^{a,b,\alpha}$ if $\mathcal{R}$ is non-connected. However, we have certainly been able to characterize ill-posed structure of the inverse scattering problem and reduce the uncertainty to a certain degree. (All uncertainty is eliminated only in the limit of exhaustive search of the parameter space to confirm $\mathcal{B}$.)



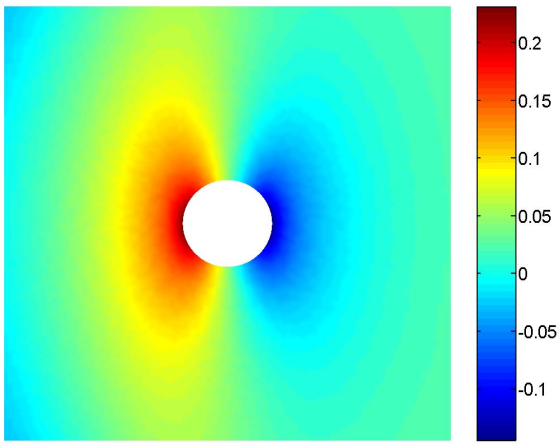
(a) Real part



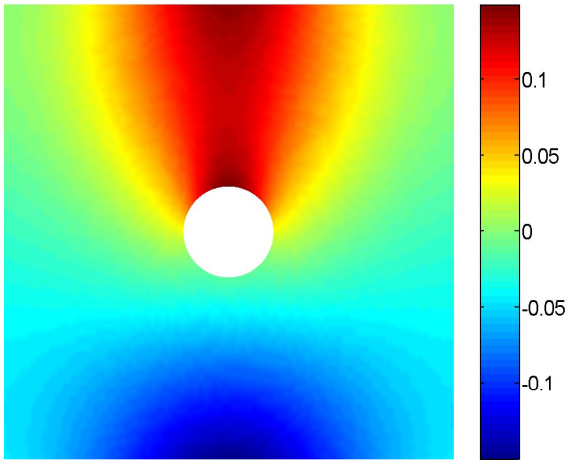
(b) Imaginary part



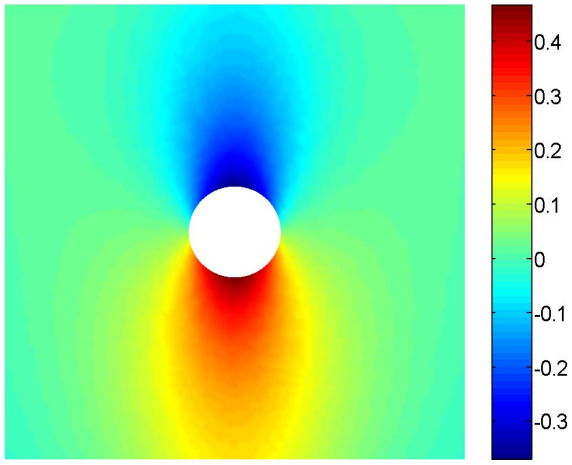
(c) Real part



(d) Imaginary part

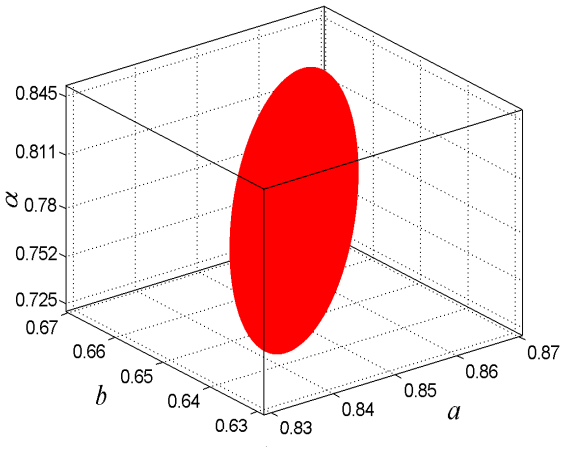


(e) Real part

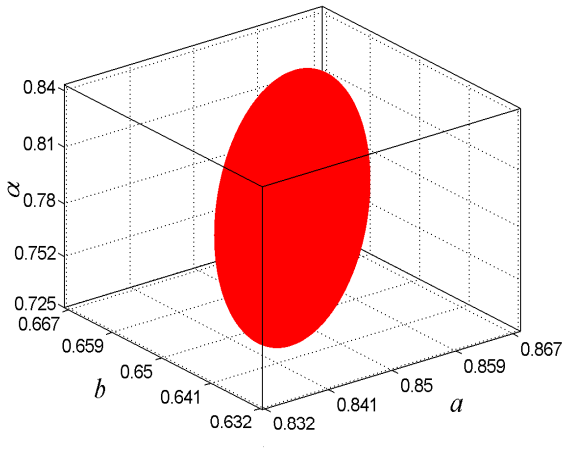


(f) Imaginary part

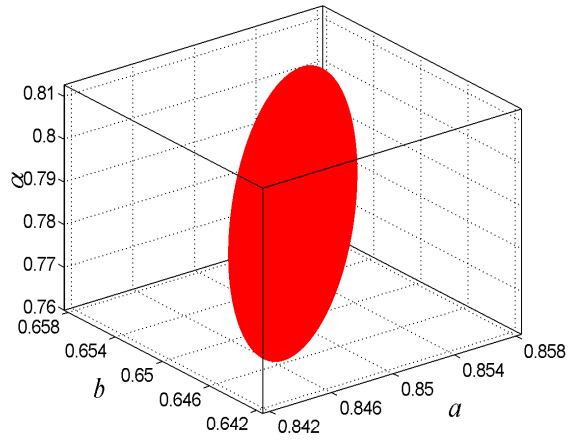
Figure 10-2: FEM solutions for $ka = \pi/8$, $b/a = 1$, $\alpha = 0$, and $\tilde{d} = (1, 0)$ in (a) and (b); for $ka = \pi/8$, $b/a = 1/2$, $\alpha = 0$, and $\tilde{d} = (1, 0)$ in (c) and (d); and for $ka = \pi/8$, $b/a = 1/2$, $\alpha = 0$, and $\tilde{d} = (0, 1)$ in (e) and (f). Note here that $\mathcal{N} = 6{,}863$.



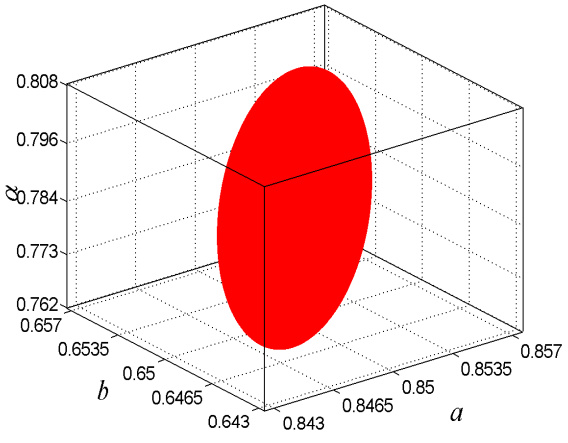
(a) $\epsilon_{\text{exp}} = 5.0\%$



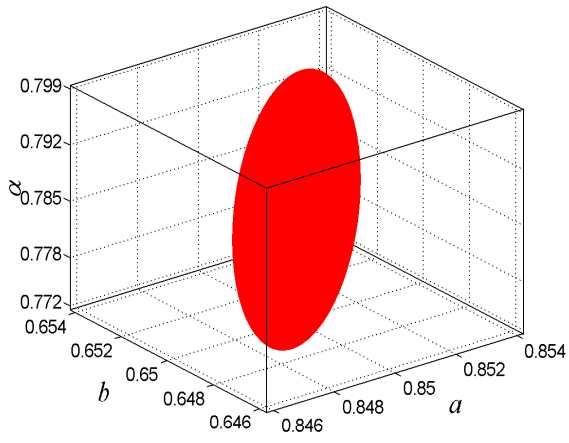
(b) $\epsilon_{\text{exp}} = 5.0\%$



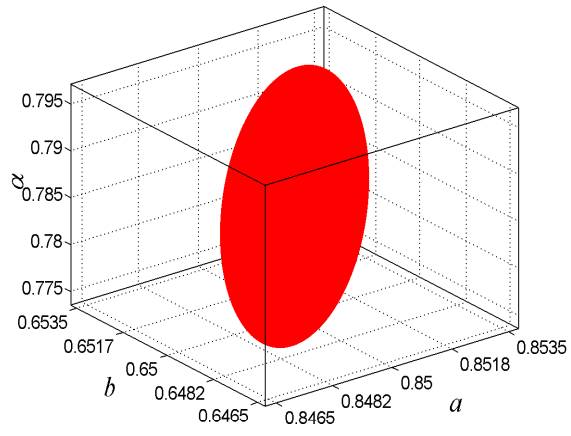
(c) $\epsilon_{\text{exp}} = 2.0\%$



(d) $\epsilon_{\text{exp}} = 2.0\%$



(e) $\epsilon_{\text{exp}} = 1.0\%$



(f) $\epsilon_{\text{exp}} = 1.0\%$

Figure 10-3: Ellipsoids containing possibility regions obtained with $N = 50$ for $a^* = 0.85$, $b^* = 0.65$, $\alpha^* = \pi/4$ for: $K = 6$ in (a), (c), (e) and $K = 9$ in (b), (d), (f). Note the change in scale in the axes: $\mathcal{R}$ shrinks as the experimental error decreases and the number of measurements increases.

ϵ_{exp}	$K = 6$	$K = 9$
5.0%	$[0.8288, \quad 0.8706] \times$ $[0.6343, \quad 0.6653] \times$ $[0.7200, \quad 0.8509]$	$[0.8306, \quad 0.8688] \times$ $[0.6343, \quad 0.6653] \times$ $[0.7249, \quad 0.8466]$
2.0%	$[0.8416, \quad 0.8583] \times$ $[0.6437, \quad 0.6562] \times$ $[0.7582, \quad 0.8127]$	$[0.8422, \quad 0.8577] \times$ $[0.6437, \quad 0.6562] \times$ $[0.7593, \quad 0.8109]$
1.0%	$[0.8458, \quad 0.8542] \times$ $[0.6468, \quad 0.6532] \times$ $[0.7715, \quad 0.7994]$	$[0.8462, \quad 0.8539] \times$ $[0.6468, \quad 0.6532] \times$ $[0.7725, \quad 0.7979]$

Table 10.4: $\mathcal{B}$ for different values of ε_{exp} and K. The true parameters are $a^* = 0.85, b^* = 0.65, \alpha^* = \pi/4$.

10.4 Chapter Summary

In this chapter, by applying our inverse method for a simple two-dimensional inverse scattering problem, we have once again demonstrated the robustness and efficiency of the method. Even though the object geometry is simple and number of parameters $O(5)$ is quite small, this example shows that not only results can be obtained essentially in real-time; but also numerical and experimental errors can be addressed rigorously and robustly. Furthermore, our method favors the use of several incident waves and limited-aperture far-field data more than one incident wave and full-aperture far-field data. Of course, the former is of more practical use than the latter, since placement of sensors on the entire unit sphere seems quite impractical.

Although this example is encouraging, it is not entirely satisfactory. Our vision for the method is in three-dimensional inverse scattering problems with many more parameters. Such problems bring new opportunities and exciting challenges. On one hand, the savings will be even much greater for problems with more complex geometry and physical modeling. In this regard, it is important to note that the online complexity is independent of the dimension of the underlying truth approximation space; and hence approximations, error bounds, and computational complexity are asymptotically invariant as the numerical (or physical/engineering) fidelity of the models is increased. On the other hand, these problems will often require very high dimension of the truth approximation space

and large number of parameters. This leads to many numerical difficulties: (1) exploring high-dimensional parameter space by greedy strategies and enumeration techniques might be impossible; (2) although the online cost is low, the offline cost is prohibitively high; (3) the inverse computational method is not yet sufficiently efficient since the associated inverse algorithms are not very effective in high-dimensional parameter space. Several recommendations to improve the efficiency and thus broaden the reach of our methods will be given in the final (next) chapter.

Chapter 11

Conclusions

In this final chapter, the theoretical developments and numerical results of the previous ten chapters are summarized. Suggestions are also provided for further improvement and extensions of the work in this thesis.

11.1 Summary

The central themes of this thesis have been the development of the reduced-basis approximations and a posteriori error bounds for different classes of parametrized partial differential equations and their application to inverse analysis in engineering and science.

We began with introducing basic but very important concepts of the reduced-basis approach, laying out a solid foundation for several subsequent chapters. The essential components of the approach are (*i*) rapidly uniformly convergent reduced-basis approximations — Galerkin projection onto the reduced-basis space W_N spanned by solutions of the governing partial differential equation at N (optimally) selected points in parameter space; (*ii*) *a posteriori* error estimation — relaxations of the residual equation that provide inexpensive yet sharp and rigorous bounds for the error in the outputs; and (*iii*) offline/online computational procedures — stratagems that exploit affine parameter dependence to decouple the generation and projection stages of the approximation process. The operation count for the online stage — in which, given a new parameter value, we calculate the output and associated error bound — depends only on N (typically small) and the parametric complexity of the problem. The method is thus ideally suited

to robust parameter estimation and adaptive design, as well as system optimization and real-time control. Furthermore, we also brought in additional ingredients: orthogonalized basis to reduce greatly the condition number of the reduced-stiffness matrix, adaptive online strategy to control tightly the growth of N while strictly satisfying the required accuracy, and sampling procedure to select optimally the approximation basis.

We further developed a very promising method to the construction of rigorous and efficient (online-inexpensive) lower bound for the critical stability factor — a generalized minimum singular value — that appears in the denominator of our *a posteriori* error bounds. The lower bound construction is applicable to linear coercive and noncoerive problems, as well as nonlinear problems. The method exploits an intermediate first-order approximation of the stability factor around a linearization point $\bar{\mu}$, which allows us to construct piecewise constant or linear lower bounds for the stability factor. Several numerical examples were presented to confirm the theoretical results and demonstrate that our lower bound construction has worked well even for strongly noncoercive case.

Until recently, the reduced-basis methods could only treat partial differential equations $g(w, v; \mu)$ that are (i) affine — more generally, affine in *functions of* μ — in μ, and (ii) at most quadratically nonlinear in the first argument. Both of these restrictions can be addressed by the "empirical interpolation" method developed (in collaboration with Professor Yvon Maday of University Paris VI) in this thesis. By replacing non-affine functions of the parameter and spatial coordinate with collateral reduced-basis expansions, we proposed an efficient reduced-basis technique that recovers online $\mathcal{N}$ independent calculation of the reduced-basis approximations and *a posteriori* error estimators for *non-affine* elliptic problems. The essential ingredients of the approach are (i) good collateral reduced-basis samples and spaces, (ii) a stable and inexpensive online interpolation procedure by which to determine the collateral reduced-basis coefficients (as a function of the parameter), and (iii) an effective *a posteriori* error bounds to quantify the newly introduced error terms. Numerical examples were presented along with the theoretical developments to test and confirm the theoretical results and illustrate various aspects of the method.

In addition, we extended the technique to treat nonlinear elliptic problems in which g consists of general nonaffine nonlinear functions of the parameter μ, spatial coordinate

x, and field variable u. By applying the empirical interpolation method to construct a collateral reduced-basis expansion for a general non-affine nonlinear function and incorporating it into the reduced-basis approximation and *a posteriori* error estimation procedure, we recovered online $\mathcal{N}$ independence even in the presence of highly nonlinear terms. Our theoretical claim was numerically confirmed by a particular problem in which the nonlinear term is an exponent function of the field variable.

Based on the reduced-basis approximation and *a posteriori* error estimation methods developed (in this thesis) for coercive and noncoercive linear elliptic equations, nonaffine elliptic equations, as well as nonlinear elliptic equations, we proposed a *robust* parameter estimation method for very *fast solution region* of inverse problems characterized by partial differential equations even in the presence of significant uncertainty. The essential innovations are threefold. The first innovation is the application of the reduce-basis techniques to the forward problem for obtaining reduced-basis approximation $s_N(\mu)$ and associated rigorous error bound $\Delta_N^s(\mu)$ of the PDE-induced output $s(\mu)$. The second innovation is the incorporation of our (very fast) lower bounds and upper bounds for the true output $s(\mu)$ — $s_N(\mu) - \Delta_N^s(\mu)$ and $s_N(\mu) + \Delta_N^s(\mu)$, respectively — into the inverse problem formulation. The third innovation is the identification of all (or almost all, in the probabilistic sense) inverse solutions consistent with the available experimental data. Ill-posedness is captured in a bounded "possibility region" that furthermore shrinks as the experimental error is decreased. The configuration possibility region may then serve in subsequent robust optimization and adaptive design studies.

Finally, we applied our robust parameter estimation method to two major areas in inverse problems: nondestructive evaluation in which crack and damage of flawed materials are identified and inverse scattering problems in which unknown buried objects ("mines") are recovered. These inverse problems though characterized by simple physical model and geometry present a promising prospect: not only numerical results can be obtained merely in seconds on a serial computer with at least $O(100)$ savings in computational time; but also numerical and (some) model uncertainties can be accommodated rigorously and robustly. These examples also show strong advantages of our approach over other computational approaches for inverse problems. First, as regards the computational expense and numerical fidelity, our approach is more efficient and reliable:

real-time and certified evaluation of functional outputs associated with the PDEs of continuum mechanics as opposed to time-consuming calculation by use of classical numerical methods. Second, as regards the model uncertainty and ill-posedness, our approach is more robust and able to exhibit/characterize ill-posed structure of the inverse problems: efficient construction of the solution region containing (all) inverse solutions consistent with the available experimental data without *a priori* regularization hypotheses as opposed to only one regularized inverse solution with *a priori* assumptions.

11.2 Suggestions for future work

There are still many aspects of this work which must still be investigated and improved. We indicate here several suggestions for future work in the hope that ongoing algorithmic and theoretical progresses to improve the efficiency and broaden the reach of the work in this thesis will continue.

First suggestion related to parametric complexity: How many parameters P can we consider — for P how large are our techniques still viable? It is undeniably the case that ultimately we should anticipate exponential scaling (of both N and certainly J) as P increases, with a concomitant unacceptable increase certainly in offline but also perhaps in online computational effort. Fortunately, for smaller P, the growth in N is rather modest, as (good) sampling procedures will automatically identify the more interesting regions of parameter space. Unfortunately, the growth in J — the number of polytopes required to cover the parameter domain of the differential operator — is more problematic: the number of eigenproblem solves is proportional to J and the discrete eigenproblems (4.50) and (4.52) can be very expensive to solve due to the generalized nature of the singular value and the presence of a continuous component to the spectrum. It is thus necessary to have more efficient construction and verification procedures for our inf-sup lower bound samples: fewer polytope coverings, inexpensive construction of the polytopes (lower cost per polytope), and more efficient eigenvalue techniques.

Second suggestion related to our empirical interpolation method and reduced-basis treatment of nonaffine elliptic problems: the regularity requirements and the $L^\infty(\Omega)$ norm used in the theoretical analysis are perhaps too strong and thus limit the scope

of the method; the theoretical worst-case Lebesgue constant $O(2^M)$ is very pessimistic relative to the numerically observed $O(10)$ Lebesgue constant; and the error estimators — though quite sharp and efficient (only one additional evaluation) — are completely rigorous upper bounds only in very restricted situations. In [52], by simply replacing the L^∞-norm by the L^2-norm in the coefficient-function procedure, we can avoid solving the costly linear program and still obtain (equally) good approximation. But rigorous theoretical framework for the weaker regularity and norm remains an open issue for further investigation.

Third suggestion related to the reduced-basis treatment of nonlinear elliptic problems: the greedy sample construction demanding solutions of the nonlinear PDEs over the sample Ξ^g is very expensive; the assumption of monotonicity is essential for the stability of the reduced-basis approximation and critical to the current development of *a posteriori* error estimation, but also restricts the application of our approach to a broader class of PDEs. It is important to note that reduced-basis treatment of weakly nonlinear non-monotonic equations has been considered [141, 140]. It can thus be hopeful that with combination of the theory in [140] and ideas presented in the thesis, it is possible to treat certain highly nonlinear nonmonotonic equations.

Fourth suggestion related to our inverse computational method, as the method is new many improvements are possible and indeed necessary: exploration of the search space using probabilistic and enumeration techniques is not effective in high-dimensional parameter space, hence advanced optimization procedures like interior point methods must be considered; construction of the ellipsoid containing the possibility region with linear program is still a heuristic approach, hence rigorous but equally efficient construction is required; the method can only characterize the solution region within the selected low-dimensional parametrization, hence more general null hypotheses are needed to detect model deviation; sensor deployment and sensitivity analysis to facilitate better design and optimized control of the system should be also considered. Furthermore, the proposed "Analyze-Assess-Act" approach is merely a proof of concepts: the Analyze stage is still heuristic not algorithmic yet; the Act stage requires to solve some optimization problems over an ellipsoidal feasible region, hence optimization procedures exploiting this feature should be developed to reduce computational time.

Final suggestion related to application of this work to engineering design, optimization, and analysis: to be of any practical value, our methods must be applied to solve real-life problems, for example, in (1) nondestructive evaluation of materials and structures relevant to the structural health monitoring of aeronautical and mechanical systems (e.g., aging aircraft, oil pipelines, and nuclear power plant), and in (2) inverse scattering and tomography relevant to medical imaging (e.g., of tumor), unexploded ordnance detection (e.g., of mines), underwater surveillance (e.g., of submarines), and tomographic scans (e.g., of biological tissues). These practical large-scale applications bring many new opportunities and exciting challenges. On one hand, the savings will be even much greater for problems with more complex geometry and physical modeling. On the other hand, these problems often require very high dimension of the "truth" approximation space associated with the underlying PDE and large number of parameters. This leads to many numerical difficulties: exploring high-dimensional parameter space by greedy strategies and enumeration techniques might be impossible; although the online cost is low, the offline cost is prohibitively high; and the inverse computational method is not yet satisfactorily effective as mentioned earlier. The treatment of these challenging problems will certainly require both theoretical and algorithmic progress on our methods as described above. To understand the implications more clearly, we consider a particular application (our last example).

11.3 Three-Dimensional Inverse Scattering Problem

We apply our methods to the three-dimensional inverse problems described thoroughly in Appendix D. Recall that the problem has the parameter input of 11 components $(a, b, c, \alpha, \beta, \gamma, k, \tilde{d}, \tilde{d}^s)$ and the piecewise-linear finite element approximation space of dimension $\mathcal{N} = 10{,}839$. However, for purpose of indicating specific directions in future work, we shall not undertake the full-scale model, but consider a simpler model in which $b = c$, $\beta = \gamma = 0$, and the incident direction and output in the plane, and fixed wave number $k = \pi/4$. Our parameter is thus $\mu = (\mu_{(1)}, \ldots, \mu_{(5)}) \in \mathcal{D} \subset \mathbb{R}^5$, where $\mu_{(1)} = a, \mu_{(2)} = b, \mu_{(3)} = \alpha$, $\mu_{(4)}$ such that $\tilde{d} = (\cos\mu_{(4)}, \sin\mu_{(4)}, 0)$, $\mu_{(5)}$ such that $\tilde{d}^s = (\cos\mu_{(5)}, \sin\mu_{(5)}, 0)$, and $\mathcal{D} \times [0.5, 1.5] \times [0.5, 1.5] \times [0, \pi] \times [0, \pi] \times [0, \pi]$.

We first note that since our first-order Robin condition is rather crude, the domain is truncated at a large distance as shown in Figure 11-1 (and $\mathcal{N}$ is thus also large) to ensure accuracy of the finite element solutions and outputs. *Future research must consider second-order radiation conditions* [5] to reduce substantially the size of domain and the dimension of the finite element approximation space.

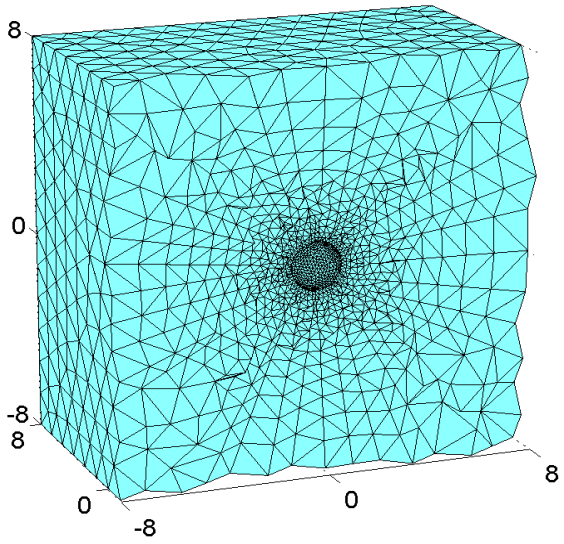


Figure 11-1: Finite element mesh on the (truncated) reference domain Ω.

We next pursue the empirical interpolation procedure described in Section 6.2 to construct $S^g_{M^g}$, $W^g_{M^g}$, $T^g_{M^g}$, $1 \le M^g \le M^g_{\max}$, for $M^g_{\max} = 39$, and $S^h_{M^h}$, $W^h_{M^h}$, $T^h_{M^h}$, $1 \le M^h \le M^h_{\max}$, for $M^h_{\max} = 39$. We next consider the piecewise-constant construction for the inf-sup lower bounds: we can cover the parameter space of the bilinear form with $J = 36$ polytopes for $\bar{\epsilon}_\beta = 0.5$;[1] here the $\mathcal{P}^{\bar{\mu}_j}$, $1 \le j \le J$, are quadrilaterals such that $|\mathcal{V}^{\mu_j}| = 4$, $1 \le j \le J$. Armed with the inf-sup lower bounds, we can pursue the adaptive sampling strategy to arrive at $N_{\max} = N^{\text{du}}_{\max} = 80$ on a grid Ξ^F of $n_{\text{F}} = 8^4 = 4096$. Would we use the full-scale model and sample along each dimension with eight intervals, then $n_{\text{F}} = 8^9 = 134{,}217{,}728$, since both the primal and dual problems have parameter space of 9 dimensions. In this case, our adaptive sampling procedure would take 1242 days to reach to $N_{\max} = 80$ for an average online evaluation time of 0.01

[1] Note that the bilinear form depends only on $\mu_{(1)}$ and $\mu_{(2)}$; hence its parameter space is two-dimensional.

seconds. Furthermore, our inf-sup lower bound construction would suffer as well due to high-dimensional parameter space, very high dimension of the truth approximation space, and expensive generalized eigenproblems (4.50) and (4.52). In any event, treatment of many tens of truly independent parameters by the global methods described in this thesis is not practicable; in such cases, *more local approaches must be pursued.*[2]

We now tabulate in Table 11.1 $\Delta_{N,\max,\mathrm{rel}}$, $\eta_{N,\mathrm{ave}}$,$\Delta^{\mathrm{du}}_{N^{\mathrm{du}},\max,\mathrm{rel}}$, $\eta^{\mathrm{du}}_{N^{\mathrm{du}},\mathrm{ave}}$, $\Delta^{s}_{N,\max,\mathrm{rel}}$, and $\eta^{s}_{N,\mathrm{ave}}$ as a function of N for $M^g = M^h = 38$. Here $\Delta_{N,\max,\mathrm{rel}}$ is the maximum over Ξ_{Test} of $\Delta_N(\mu)/\|u(\mu)\|_X$, $\eta_{N,\mathrm{ave}}$ is the average over Ξ_{Test} of $\Delta_N(\mu)/\|u(\mu) - u_N(\mu)\|_X$, $\Delta^{\mathrm{du}}_{N^{\mathrm{du}},\max,\mathrm{rel}}$ is the maximum over Ξ_{Test} of $\Delta^{\mathrm{du}}_{N^{\mathrm{du}}}(\mu)/\|\psi(\mu)\|_X$, $\eta^{\mathrm{du}}_{N^{\mathrm{du}},\mathrm{ave}}$ is the average over Ξ_{Test} of $\Delta^{\mathrm{du}}_{N^{\mathrm{du}}}(\mu)/\|\psi(\mu) - \psi_{N^{\mathrm{du}}}(\mu)\|_X$, $\Delta^{s}_{N,\max,\mathrm{rel}}$ is the maximum over Ξ_{Test} of $\Delta^{s}_{N}(\mu)/|s(\mu) - s_N(\mu)|$, and $\eta^{s}_{N,\mathrm{ave}}$ is the average over Ξ_{Test} of $\Delta^{s}_{N}(\mu)/|s(\mu) - s_N(\mu)|$, where $\Xi_{\mathrm{Test}} \subset (\mathcal{D})^{223}$ is a random parameter grid of size 223. We observe that the reduced-basis approximations converge quite fast, but still slower than those in the two-dimensional inverse scattering problem as shown in Section 6.6.6, although the two problems have the same parametric dimension. However, we do realize online factors of improvement of $O(1000)$: for an accuracy close to 0.1 percent ($N = 60$), the total Online computational time on a Pentium 1.6GHz processor to compute $s_N(\mu)$ and $\Delta^{s}_{N}(\mu)$ is less than 1/1517 times the Total Time to directly calculate the truth output $s(\mu)$.

N	$\Delta_{N,\max,\mathrm{rel}}$	$\eta_{N,\mathrm{ave}}$	$\Delta^{\mathrm{du}}_{N^{\mathrm{du}},\max,\mathrm{rel}}$	$\eta^{\mathrm{du}}_{N^{\mathrm{du}},\mathrm{ave}}$	$\Delta^{s}_{N,\max,\mathrm{rel}}$	$\eta^{s}_{N,\mathrm{ave}}$
10	1.86 E−00	12.72	1.42 E−00	10.87	1.44 E−00	16.76
20	9.66 E−01	12.80	6.13 E−01	11.21	3.18 E−01	18.38
30	3.68 E−01	13.34	3.42 E−01	12.33	7.33 E−02	20.69
40	1.83 E−01	13.97	1.78 E−01	12.78	2.07 E−02	17.41
50	1.23 E−01	15.15	1.10 E−01	13.91	7.40 E−03	16.68
60	5.68 E−02	17.41	4.75 E−02	16.82	2.04 E−03	22.90
70	3.70 E−02	19.19	3.07 E−02	17.71	6.79 E−04	21.19
80	1.81 E−02	19.71	1.38 E−02	19.27	1.77 E−04	27.44

Table 11.1: Relative error bounds and effectivities as a function of N for $M^g = M^h = 38$.

Finally, we find a region $\mathcal{R} \in \mathcal{D}^{a,b,\alpha}$ in which the true but unknown parameter, (a^*, b^*, α^*), must reside from the far-field data superposed with the error ϵ_{exp}. To obtain

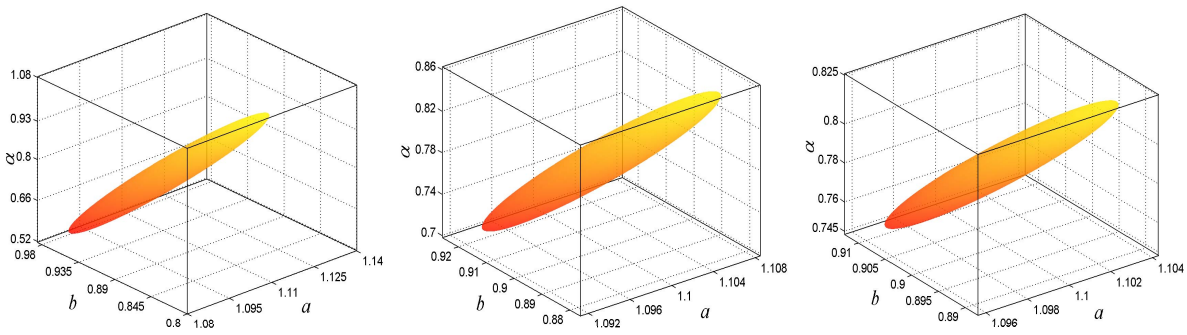
[2]We do note that at least some problems with ostensibly many parameters in fact involve *highly coupled or correlated* parameters: certain classes of shape optimization certainly fall into this category. In these situations, global progress can be made.

the experimental data, we use three different directions $\mu_{(4)} = \{0, \pi/4, \pi/2\}$ for the incident wave. For each direction of the incident wave, there are I angles, $\mu_{(5)} = \{\pi(i-1)/I, 1 \leq i \leq I\}$, at which the far-field data are obtained. We display in Figure 11-2 the ellipsoids containing the possibility regions — for experimental error of 5%, 2%, and 1% and number of measurements of 3 ($I = 1$), 6 ($I = 2$), and 9 ($I = 3$); here the ellipsoids are constructed from the corresponding sets of 450 region boundary points obtained by using our inverse algorithm described in Section 8.4.2. We also present in Table 11.2 the half lengths of $\mathcal{R}$ — more precicely, the half lengths of the box containing $\mathcal{R}$ — relative to the exact (synthetic) value $a^* = 1.1, b^* = 0.9, \alpha^* = \pi/4$.

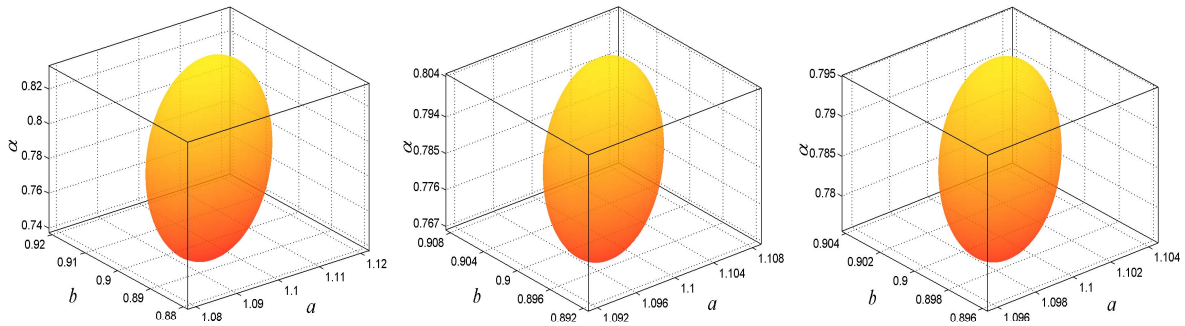
	ϵ_{exp}	$K = 3$	$K = 6$	$K = 9$
$0.5\Delta a/a^*$	5.0%	2.72%	1.99%	1.54%
	2.0%	0.78%	0.81%	0.62%
	1.0%	0.40%	0.41%	0.31%
$0.5\Delta b/b^*$	5.0%	10.15%	2.35%	1.71%
	2.0%	2.75%	0.95%	0.69%
	1.0%	1.39%	0.47%	0.35%
$0.5\Delta\alpha/\alpha^*$	5.0%	35.92%	6.11%	6.78%
	2.0%	10.48%	2.45%	2.72%
	1.0%	5.25%	1.25%	1.35%

Table 11.2: The half lengths of the box containing $\mathcal{R}$ relative to a^*, b^*, α^* as a function of experimental error ϵ_{exp} and number of measurements K.

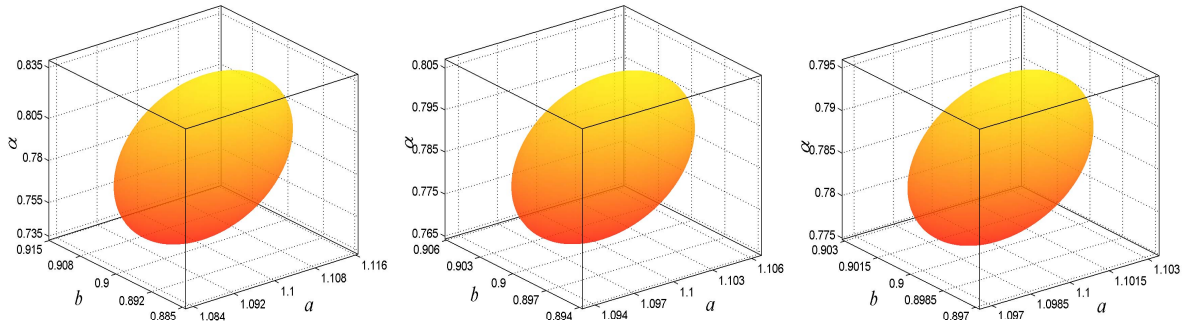
We see that as ϵ_{exp} decreases and K increases, $\mathcal{R}$ shrinks toward (a^*, b^*, α^*). The results are indeed largely indistinguishable from the finite element method, since the relative output bound for $N = 60$ is considerably less than 1.0%. More importantly, these ellipsoids not only quantify robustly uncertainty in both the numerical approximation and expermental error, but also are obtained online within 342 seconds on a Pentium 1.6 GHz thanks to a "per forward evaluation time" of only 0.0448 seconds. However, if we consider the full-scale model, the construction of an ellipsoidal possibility region for $(a^*, b^*, c^*, \alpha^*, \beta^*, \gamma^*)$ by our inverse computational method can be much more computationally extensive, but still viable. Of course, treatment of many more parameters by our simple enumeration techniques is not practicable; in such cases, *more rigorous inverse techniques and efficient optimization procedures must be required.*



(a) $\epsilon_{\text{exp}} = 5.0\%$ (b) $\epsilon_{\text{exp}} = 2.0\%$ (c) $\epsilon_{\text{exp}} = 1.0\%$



(d) $\epsilon_{\text{exp}} = 5.0\%$ (e) $\epsilon_{\text{exp}} = 2.0\%$ (f) $\epsilon_{\text{exp}} = 1.0\%$



(g) $\epsilon_{\text{exp}} = 5.0\%$ (h) $\epsilon_{\text{exp}} = 2.0\%$ (i) $\epsilon_{\text{exp}} = 1.0\%$

Figure 11-2: Ellipsoids containing possibility regions obtained with $N = 60$ for $a^* = 1.1$, $b^* = 0.9$, $\alpha^* = \pi/4$ for: $K = 3$ in (a), (b), (c); $K = 6$ in (d), (e), (f); and $K = 9$ in (g), (h), (i). Note the change in scale in the axes: $\mathcal{R}$ shrinks as the experimental error decreases and the number of measurements increases.

Appendix A

Asymptotic Behavior of the Scattered Field

We consider the Helmholtz equation with the Sommerfeld radiation condition

$$\Delta\tilde{u} + k^2\tilde{u} = 0 \qquad \text{in } \mathbb{R}^n\backslash\tilde{D}, \tag{A.1a}$$

$$\lim_{\tilde{r}\to\infty} \tilde{r}^{(n-1)/2}\left(\frac{\partial\tilde{u}}{\partial\tilde{r}} - \mathrm{i}k\tilde{u}\right) = 0, \quad \tilde{r} = |\tilde{x}|\ . \tag{A.1b}$$

We shall prove that solution $\tilde{u}$ to the problem (A.1) has the asymptotic behavior of an outgoing spherical wave

$$\tilde{u}(\tilde{x}) = \frac{e^{\mathrm{i}k\tilde{r}}}{\tilde{r}^{(n-1)/2}}\tilde{u}_\infty(\tilde{D}, \tilde{d}^s, \tilde{d}, k) + O\left(\frac{1}{\tilde{r}^{(n+1)/2}}\right), \quad |\tilde{x}| \to \infty, \tag{A.2}$$

uniformly in all directions $\tilde{d}^s = \tilde{x}/|\tilde{x}|$ where the function $\tilde{u}_\infty$ defined on the unit sphere $\tilde{S} \subset \mathbb{R}^n$ is known as the far-field pattern of the scattered wave $\tilde{u}$ and is given by

$$\tilde{u}_\infty(\tilde{D}, \tilde{d}^s, \tilde{d}, k) = \tilde{\beta}_n \int_{\partial\tilde{D}} \tilde{u}(\tilde{x})\frac{\partial e^{-\mathrm{i}k\tilde{d}^s\cdot\tilde{x}}}{\partial\tilde{\nu}} - \frac{\partial\tilde{u}(\tilde{x})}{\partial\tilde{\nu}}e^{-\mathrm{i}k\tilde{d}^s\cdot\tilde{x}}\ , \tag{A.3}$$

with

$$\tilde{\beta}_n = \begin{cases} \frac{\mathrm{i}}{4}\sqrt{\frac{2}{\pi k}}e^{-\mathrm{i}\pi/4} & n = 2 \\ \frac{1}{4\pi} & n = 3. \end{cases} \tag{A.4}$$

Recall that $\tilde{\nu}$ is the unit normal to the boundary $\partial\tilde{D}$ and directed into the exterior of $\tilde{D}$.

We now introduce some relevant mathematics needed for our proof. First, we need Green's integral theorems: Let $\tilde{\Omega}$ be a bounded domain of class C^1 and let $\tilde{\nu}$ denote the unit normal vector to the boundary $\partial\tilde{\Omega}$ directed into the exterior of $\tilde{\Omega}$; then for $\tilde{u} \in C^1(\tilde{\Omega})$ and $\tilde{v} \in C^2(\tilde{\Omega})$, we have Green's first theorem

$$\int_{\tilde{\Omega}} \tilde{u}\Delta\tilde{v} + \nabla\tilde{u}\nabla\tilde{v} = \int_{\partial\tilde{\Omega}} \tilde{u}\frac{\partial\tilde{v}}{\partial\tilde{\nu}} \ , \tag{A.5}$$

and for $\tilde{u}, \tilde{v} \in C^2(\tilde{\Omega})$ we have Green's second theorem

$$\int_{\tilde{\Omega}} \tilde{u}\Delta\tilde{v} - \tilde{v}\Delta\tilde{u} = \int_{\partial\tilde{\Omega}} \tilde{u}\frac{\partial\tilde{v}}{\partial\tilde{\nu}} - \tilde{v}\frac{\partial\tilde{u}}{\partial\tilde{\nu}} \ . \tag{A.6}$$

Second, we need the *fundamental solution*[1] to the Helmholtz equation (A.1a) defined by

$$\Phi(\tilde{x},\tilde{y}) = \begin{cases} \frac{\mathrm{i}}{4}H_0^{(1)}(k|\tilde{x}-\tilde{y}|) & n=2 \\ \frac{1}{4\pi}\frac{e^{\mathrm{i}k|\tilde{x}-\tilde{y}|}}{|\tilde{x}-\tilde{y}|} & n=3 \end{cases} \tag{A.7}$$

where $H_0^{(1)}$ is the Hankel function of first kind of zero order. We note that $\Phi(\tilde{x},\tilde{y})$ has the following asymptotic behavior

$$\frac{\partial\Phi(\tilde{x},\tilde{y})}{\partial\tilde{\nu}} - \mathrm{i}k\Phi(\tilde{x},\tilde{y}) = O\left(\frac{1}{\tilde{r}^{(n+1)/2}}\right), \quad |\tilde{x}| \to \infty \ . \tag{A.8}$$

This can be derived from

$$\frac{e^{\mathrm{i}k|\tilde{x}-\tilde{y}|}}{|\tilde{x}-\tilde{y}|} = \frac{e^{\mathrm{i}k|\tilde{x}|}}{|\tilde{x}|}\left\{e^{-\mathrm{i}k\tilde{d}^s\cdot\tilde{y}} + O\left(\frac{1}{|\tilde{x}|}\right)\right\} \tag{A.9}$$

$$\frac{\partial}{\partial\tilde{\nu}(\tilde{y})}\frac{e^{\mathrm{i}k|\tilde{x}-\tilde{y}|}}{|\tilde{x}-\tilde{y}|} = \frac{e^{\mathrm{i}k|\tilde{x}|}}{|\tilde{x}|}\left\{\frac{\partial e^{-\mathrm{i}k\tilde{d}^s\cdot\tilde{y}}}{\partial\tilde{\nu}(\tilde{y})} + O\left(\frac{1}{|\tilde{x}|}\right)\right\} \tag{A.10}$$

[1]The fundamental solution is in fact the Green function for the Helmholtz equation and plays an important role in theoretical analysis and numerical computation of solutions to the direct scattering problem (A.1).

$$H_0^{(1)}(k|\tilde{x}-\tilde{y}|) = \sqrt{\frac{2}{\pi k|\tilde{x}-\tilde{y}|}}e^{\mathrm{i}(k|\tilde{x}-\tilde{y}|-\pi/4)} = \sqrt{\frac{2}{\pi k}}e^{-\mathrm{i}\pi/4}\frac{e^{\mathrm{i}k|\tilde{x}|}}{\sqrt{|\tilde{x}|}}\left\{e^{-\mathrm{i}k\tilde{d}^s\cdot\tilde{y}} + O\left(\frac{1}{|\tilde{x}|}\right)\right\} \tag{A.11}$$

$$\frac{\partial}{\partial\tilde{\nu}(\tilde{y})}H_0^{(1)}(k|\tilde{x}-\tilde{y}|) = \sqrt{\frac{2}{\pi k}}e^{-\mathrm{i}\pi/4}\frac{e^{\mathrm{i}k|\tilde{x}|}}{\sqrt{|\tilde{x}|}}\left\{\frac{\partial e^{-\mathrm{i}k\tilde{d}^s\cdot\tilde{y}}}{\partial\tilde{\nu}(\tilde{y})} + O\left(\frac{1}{|\tilde{x}|}\right)\right\} \tag{A.12}$$

as $|\tilde{x}| \to \infty$, since

$$|\tilde{x}-\tilde{y}| = \sqrt{\tilde{x}^2 - 2\tilde{x}\cdot\tilde{y}+\tilde{y}^2} = |\tilde{x}| - \tilde{d}^s\cdot\tilde{y}\ .$$

We can readily prove the important result (A.2) and (A.3) as follows

Proof. We follow the proof given in [33] (Theorems 2.1, 2.4 and 2.5). Let $\tilde{S}_{\tilde{r}}$ denote the sphere of radius $\tilde{r}$ and center at the origin, we note from the radiation condition (A.1b) that

$$\int_{\tilde{S}_{\tilde{r}}}\left|\frac{\partial\tilde{u}}{\partial\tilde{\nu}} - ik\tilde{u}\right|^2 = \int_{\tilde{S}_{\tilde{r}}}\left\{\left|\frac{\partial\tilde{u}}{\partial\tilde{\nu}}\right|^2 + k^2|\tilde{u}|^2 + 2k\Im\left(\tilde{u}\frac{\partial\bar{\tilde{u}}}{\partial\tilde{\nu}}\right)\right\} \to 0, \quad \tilde{r}\to\infty\ , \tag{A.13}$$

where $\tilde{\nu}$ is the unit outward normal to $\tilde{S}_{\tilde{r}}$. We next take $\tilde{r}$ large enough such that $\tilde{D}$ is contained in $\tilde{S}_{\tilde{r}}$ and apply the Green's theorem (A.5) in the domain $\tilde{\Omega}_{\tilde{r}} \equiv \{\tilde{y}\in\mathbb{R}^n\backslash\bar{\tilde{D}}\,|\,|\tilde{y}|<\tilde{r}\}$ to get

$$\int_{\tilde{S}_{\tilde{r}}}\tilde{u}\frac{\partial\bar{\tilde{u}}}{\partial\tilde{\nu}} = \int_{\partial\tilde{D}}\tilde{u}\frac{\partial\bar{\tilde{u}}}{\partial\tilde{\nu}} + \int_{\tilde{\Omega}_{\tilde{r}}}\tilde{u}\Delta\tilde{u} + \int_{\tilde{\Omega}_{\tilde{r}}}|\nabla\tilde{u}|^2 = \int_{\partial\tilde{D}}\tilde{u}\frac{\partial\bar{\tilde{u}}}{\partial\tilde{\nu}} - k^2\int_{\tilde{\Omega}_{\tilde{r}}}|\tilde{u}|^2 + \int_{\tilde{\Omega}_{\tilde{r}}}|\nabla\tilde{u}|^2. \tag{A.14}$$

We then insert the imaginary part of the last equation into (A.13) to obtain

$$\lim_{\tilde{r}\to\infty}\int_{\tilde{S}_{\tilde{r}}}\left\{\left|\frac{\partial\tilde{u}}{\partial\tilde{\nu}}\right|^2 + k^2|\tilde{u}|^2\right\} = -2k\Im\left(\int_{\partial\tilde{D}}\tilde{u}\frac{\partial\bar{\tilde{u}}}{\partial\tilde{\nu}}\right)\ . \tag{A.15}$$

Since both terms on the left hand side of are nonnegative and their sum tends to a finite limit, they must be individually bounded as $\tilde{r}\to\infty$, i.e., we have

$$\int_{\tilde{S}_{\tilde{r}}}|\tilde{u}|^2 = O(1), \quad \tilde{r}\to\infty\ . \tag{A.16}$$

It thus follows from (A.8), (A.16) and Cauchy-Schwarz inequality that

$$\int_{\tilde{S}_{\tilde{r}}} \tilde{u}(\tilde{y}) \left(\frac{\partial \Phi(\tilde{x},\tilde{y})}{\partial \tilde{\nu}(\tilde{y})} - \mathrm{i}k\Phi(\tilde{x},\tilde{y}) \right) \to 0, \quad \tilde{r} \to \infty \ . \tag{A.17}$$

Furthermore, from the radiation condition (A.1b) for u and $\Phi(\tilde{x},\tilde{y}) = O(1/\tilde{r}^{(n-1)/2})$ for $\tilde{y} \in \tilde{S}_{\tilde{r}}$, we have

$$\int_{\tilde{S}_{\tilde{r}}} \Phi(\tilde{x},\tilde{y}) \left(\frac{\partial \tilde{u}}{\partial \tilde{\nu}}(\tilde{y}) - \mathrm{i}k\tilde{u}(\tilde{y}) \right) \to 0, \quad \tilde{r} \to \infty \ . \tag{A.18}$$

Subtracting (A.17) from (A.18) yields

$$\int_{\tilde{S}_{\tilde{r}}} \left(\tilde{u}(\tilde{y}) \frac{\partial \Phi(\tilde{x},\tilde{y})}{\partial \tilde{\nu}(\tilde{y})} - \frac{\partial \tilde{u}}{\partial \tilde{\nu}}(\tilde{y}) \Phi(\tilde{x},\tilde{y}) \right) \to 0, \quad \tilde{r} \to \infty \ . \tag{A.19}$$

We now circumscribe an arbitrary point $\tilde{x} \in \tilde{\Omega}_{\tilde{r}}$ with an infinitesimal sphere $\tilde{S}(\tilde{x},\tilde{\rho}) \equiv \{\tilde{y} \in \mathbb{R}^n \,|\, |\tilde{x}-\tilde{y}| = \tilde{\rho}\}$ and direct the normal $\tilde{\nu}$ into the interior of $\tilde{S}(\tilde{x},\tilde{\rho})$. We apply the Green's theorem (A.6) to the function $\tilde{u}$ and $\Phi(\tilde{x};\cdot)$ in the domain $\tilde{\Omega}_{\tilde{\rho}} \equiv \{\tilde{y} \in \tilde{\Omega}_{\tilde{r}} \,|\, |\tilde{x}-\tilde{y}| > \tilde{\rho}\}$ to obtain

$$\begin{aligned}
&\int_{\partial \tilde{D}} \left(\tilde{u}(\tilde{y}) \frac{\partial \Phi(\tilde{x},\tilde{y})}{\partial \tilde{\nu}(\tilde{y})} - \frac{\partial \tilde{u}}{\partial \tilde{\nu}}(\tilde{y}) \Phi(\tilde{x},\tilde{y}) \right) + \int_{\tilde{S}_{\tilde{r}} \cup \tilde{S}(\tilde{x},\tilde{\rho})} \left(\frac{\partial \tilde{u}}{\partial \tilde{\nu}}(\tilde{y}) \Phi(\tilde{x},\tilde{y}) - \tilde{u}(\tilde{y}) \frac{\partial \Phi(\tilde{x},\tilde{y})}{\partial \tilde{\nu}(\tilde{y})} \right) \\
&\quad = \int_{\tilde{\Omega}_{\tilde{\rho}}} \{\Phi(\tilde{x},\tilde{y}) \Delta \tilde{u} - \tilde{u} \Delta \Phi(\tilde{x},\tilde{y})\} \\
&\quad = \int_{\tilde{\Omega}_{\tilde{\rho}}} \{\Delta \tilde{u} + k^2 \tilde{u}\} \Phi(\tilde{x},\tilde{y}) \\
&\quad = 0 \ .
\end{aligned} \tag{A.20}$$

Since on $\tilde{S}(\tilde{x},\tilde{\rho})$ we have

$$\Phi(\tilde{x},\tilde{y}) = \frac{e^{\mathrm{i}k\tilde{\rho}}}{4\pi\tilde{\rho}}, \qquad \frac{\partial \Phi(\tilde{x},\tilde{y})}{\partial \tilde{y}} = \left(\frac{1}{\tilde{\rho}} - \mathrm{i}k \right) \frac{e^{\mathrm{i}k\tilde{\rho}}}{4\pi\tilde{\rho}} \tilde{\nu}(\tilde{y}) \ ,$$

it then follows from the mean value theorem that

$$\lim_{\tilde{\rho} \to 0} \int_{\tilde{S}(\tilde{x},\tilde{\rho})} \left(\tilde{u}(\tilde{y}) \frac{\partial \Phi(\tilde{x},\tilde{y})}{\partial \tilde{\nu}(\tilde{y})} - \frac{\partial \tilde{u}}{\partial \tilde{\nu}}(\tilde{y}) \Phi(\tilde{x},\tilde{y}) \right) = \tilde{u}(\tilde{x}) \ . \tag{A.21}$$

We thus conclude from (A.19)-(A.21) by passing to the limit $\tilde{r} \to \infty$ and $\tilde{\rho} \to 0$ that

$$\tilde{u}(\tilde{x}) = \int_{\partial \tilde{D}} \left(\tilde{u}(\tilde{y}) \frac{\partial \Phi(\tilde{x}, \tilde{y})}{\partial \tilde{\nu}(\tilde{y})} - \frac{\partial \tilde{u}}{\partial \tilde{\nu}}(\tilde{y}) \Phi(\tilde{x}, \tilde{y}) \right), \quad \tilde{x} = \mathbb{R}^n \backslash \bar{\tilde{D}} \ . \tag{A.22}$$

Finally, inserting the asymptotic representation of $\Phi(\tilde{x}, \tilde{y})$ and $\frac{\partial \Phi(\tilde{x}, \tilde{y})}{\partial \tilde{\nu}(\tilde{y})}$ from (A.9)-(A.12) into (A.22) yields the desired result (A.2) and (A.3). □

Appendix B

Lanczos Algorithm for Generalized Hermitian Eigenvalue Problems

We consider a generalized Hermitian eigenvalue problem (GHEP)

$$Ax = \lambda Bx \tag{B.1}$$

where $A \in \mathbb{R}^{\mathcal{N}\times\mathcal{N}}$ and $B \in \mathbb{R}^{\mathcal{N}\times\mathcal{N}}$ are Hermitian matrices, i.e., $A^H = A$ and $B^H = B$.

Since we are interested in the minimum eigenmode $(\lambda_{\min}, x_{\min})$ and the maximum eigenmode $(\lambda_{\max}, x_{\max})$ of the eigenproblem (B.1), Lanczos method is most suitable for this task. Because these extreme eigenvalues are often (not always) isolated from the rest in the spectrum, Lanczos method can give rapid convergence rate for these eigenvalues. However, the convergence rate can be slow in some cases due to the generalized nature of the eigenvalues and the presence of a continuous component (if any) to the spectrum.

We give in Figure B-1 the Lanczos algorithm and a short description as follows. Step 4. is the computation of the mutual orthogonalized bases $V_\ell = \{v_1, \ldots, v_\ell\}$ and $W_\ell = \{w_1, \ldots, w_\ell\}$, $W_\ell^H V_\ell = I$. Steps 5., to 9., are the computation of the residual vector r. In step 12. we update the tridiagonal matrix H_ℓ from $H_{\ell-1}$. In steps 13. and 14. we compute the approximate eigenvalues Λ_ℓ and approximate eigenvectors X_ℓ. In step 15. we check for convergence. We see that the Lanczos iteration works by replacing the eigenproblem (B.1) with a much simpler eigenproblem (associated with H_ℓ) that approximates well (certain part of) the spectrum.

1. Start with normalized random vector q
2. Set $w_0 = 0$, $r = Bq$, $\beta_0 = \sqrt{|q^H r|}$
3. **for** $\ell = 1, 2, \ldots,$ **until** convergence
4. $\quad w_\ell = r/\beta_{\ell-1}$, $v_\ell = q/\beta_{\ell-1}$
5. $\quad r = Av_\ell$, $\alpha_\ell = v_\ell^H r$
6. $\quad r = r - \beta_{\ell-1} w_{\ell-1} - \alpha_\ell w_\ell$
7. $\quad$ **for** $i = 1, \ldots, \ell$
8. $\quad\quad r = r - \frac{v_i^H r}{w_i^H v_i} w_i$
9. $\quad$ **end for**
10. $\quad$ Solve system $Bq = r$ for q
11. $\quad \beta_\ell = \sqrt{|q^H r|}$
12. $\quad H_\ell = \begin{pmatrix} \alpha_1 & \beta_1 & & & \\ \beta_1 & \alpha_2 & \beta_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \beta_{\ell-1} \\ & & & \beta_{\ell-1} & \alpha_\ell \end{pmatrix}$
13. $\quad$ Compute approximate eigenvalues Λ_ℓ from $H_\ell = S_\ell \Lambda_\ell S_\ell^H$
14. $\quad$ Compute approximate eigenvectors $X_\ell = V_\ell S_\ell$
15. $\quad$ Test for convergence
16. **end for**

Figure B-1: Lanczos Algorithm for GHEP.

Appendix C

Inf-Sup Lower Bound Formulation for Complex Noncoercive Problems

C.1 Inf-Sup Parameter

We consider the lower bound construction for the inf-sup parameter

$$\beta(\mu) \equiv \inf_{v \in X} \sup_{w \in X} \frac{|a(w, v; \mu)|}{\|w\|_X \|v\|_X} \ . \tag{C.1}$$

Here $a : X \times X \times \mathcal{D} \to \mathbb{C}$ is a parametrized complex noncoercive bilinear form; $\mathcal{D} \in \mathbb{R}^P$ is the parameter domain; and X is a complex function space over the complex field $\mathbb{C}$. We assume that for some finite integer Q, a may be expressed as an affine decomposition of the form

$$a(w, v; \mu) = \sum_{q=1}^{Q} \Theta^q(\mu) a^q(w, v), \quad \forall\, w, v \in X,\ \forall\, \mu \in \mathcal{D} \ , \tag{C.2}$$

where for $1 \le q \le Q$, $\Theta^q : \mathcal{D} \to \mathbb{R}$ are differentiable complex parameter-dependent functions and $a^q : X \times X \to \mathbb{R}$ are parameter-independent continuous forms.

Next we introduce the supremizing operator $T^\mu : X \to X$, for any given $\mu \in \mathcal{D}$ and any w in X

$$(T^\mu w, v) = a(w, v; \mu), \quad \forall\, v \in X \ ; \tag{C.3}$$

it is readily shown by Riesz representation that

$$T^\mu w = \arg\sup_{v\in X} \frac{a(w,v;\mu)}{\|v\|}, \quad \forall\, v \in X \ . \tag{C.4}$$

It follows from (C.1) and (C.4) that our inf-sup parameter $\beta(\mu)$ is expressed as

$$\beta(\mu) = \inf_{v\in X} \frac{\|T^\mu v\|}{\|v\|} = \inf_{v\in X} \sqrt{\frac{b(v,v;\mu)}{\|v\|^2}} \ , \tag{C.5}$$

where the bilinear form $b(\cdot,\cdot;\mu)$ is given by

$$b(w,v;\mu) = (T^\mu w, T^\mu v), \quad \forall\, w,v \in X \ . \tag{C.6}$$

It is a simple matter to show that $b(\cdot,\cdot;\mu)$ is symmetric positive-definite: $b(w,v;\mu) = (T^\mu w, T^\mu v) = \overline{(T^\mu v, T^\mu w)} = \overline{b(v,w;\mu)}, \forall\, w,v \in X$ and $b(v,v;\mu) = (T^\mu v, T^\mu v) > 0, \forall\, v \in X, v \neq 0$, from the symmetric positive-definiteness of $(\cdot,\cdot)$ and $T^\mu v \neq 0, \forall\, v \neq 0$. Furthermore, it follows from (C.2) and (C.3) that, for any $w \in X$, $T^\mu w$ can be expressed as

$$T^\mu w = \sum_{q=1}^{Q} \Theta^q(\mu)\, T^q w \ , \tag{C.7}$$

where, for any $w \in X$, $T^q w$, $1 \le q \le Q$, are given by

$$(T^q w, v)_X = a^q(w,v), \quad \forall\, v \in X \ . \tag{C.8}$$

Note that the operators $T^q \colon X \to X$ are independent of the parameter μ.

We now introduce the eigenproblem: Given $\mu \in \mathcal{D}^\mu$, find the minimum eigenmode $\chi_{\min}(\mu) \in X$, $\lambda_{\min}(\mu) \in \mathbb{R}$ such that

$$b(\chi_{\min}(\mu), v; \mu) = \lambda_{\min}(\mu)(\chi_{\min}(\mu), v), \ \forall\, v \in X \ , \tag{C.9}$$

$$\|\chi_{\min}(\mu)\| = 1 \ , \tag{C.10}$$

which can be rewritten as

$$a(\chi_{\min}(\mu), T^{\mu}v; \mu) = \lambda_{\min}(\mu)(\chi_{\min}(\mu), v), \ \forall\, v \in X \ , \tag{C.11}$$

$$\|\chi_{\min}(\mu)\| = 1 \ . \tag{C.12}$$

It thus follows from (C.5) and (C.9) that $\beta(\mu) = \sqrt{\lambda_{\min}(\mu)}$. Since furthermore $b(\cdot, \cdot; \mu)$ is symmetric positive-definite, $\beta(\mu)$ is real and positive.

C.2 Inf-Sup Lower Bound Formulation

We consider the construction of $\hat{\beta}(\mu)$, a lower bound for $\beta(\mu)$. To begin, given $\overline{\mu} \in \mathcal{D}$ and $t = (t_{(1)}, \ldots, t_{(P)}) \in \mathbb{R}^P$, we introduce the bilinear form

$$\begin{aligned} \mathcal{T}(w, v; t; \overline{\mu}) \ &= \ (T^{\overline{\mu}}w, T^{\overline{\mu}}v)_X + \sum_{p=1}^{P} t_{(p)} \times \\ &\left\{ \sum_{q=1}^{Q} \frac{\partial \Theta^q}{\partial \mu_{(p)}}(\overline{\mu}) a^q(w, T^{\overline{\mu}}v) + \sum_{q=1}^{Q} \frac{\partial \overline{\Theta}^q}{\partial \mu_{(p)}}(\overline{\mu}) \overline{a^q(v, T^{\overline{\mu}}w)} \right\} \end{aligned} \tag{C.13}$$

and associated Rayleigh quotient

$$\mathcal{F}(t; \overline{\mu}) = \min_{v \in X} \frac{\mathcal{T}(v, v; t; \overline{\mu})}{\|v\|_X^2} \ . \tag{C.14}$$

It is readily shown that $\mathcal{F}(t; \overline{\mu})$ is *concave* in t; and hence $\mathcal{D}^{\overline{\mu}} \equiv \left\{ \mu \in \mathbb{R}^P \mid \mathcal{F}(\mu - \overline{\mu}; \overline{\mu}) \geq 0 \right\}$ is perforce convex. Note also that $t(\cdot, \cdot; t; \overline{\mu})$ is symmetric since $b(\cdot, \cdot; \mu)$ is symmetric and $\sum_{q=1}^{Q} \frac{\partial \Theta^q}{\partial \mu_{(p)}}(\overline{\mu}) a^q(w, T^{\overline{\mu}}v)$ is a complex conjugate transpose of $\sum_{q=1}^{Q} \frac{\partial \overline{\Theta}^q}{\partial \mu_{(p)}}(\overline{\mu}) \overline{a^q(v, T^{\overline{\mu}}w)}$; hence, $\mathcal{F}(\mu - \overline{\mu}; \overline{\mu})$ is real and positive for all $\mu \in \mathcal{D}^{\overline{\mu}}$. (In general $\mathcal{F}(\mu - \overline{\mu}; \overline{\mu})$ is real, but negative for $\mu \notin \mathcal{D}^{\overline{\mu}}$; the restriction of $\mathcal{F}(\mu - \overline{\mu}; \overline{\mu})$ on $\mathcal{D}^{\overline{\mu}}$ is thus necessary for our inf-sup lower bound construction.)

We next assume that a^q are continuous in the sense that there exist positive finite constants Γ_q, $1 \leq q \leq Q$, such that

$$|a^q(w, v)| \leq \Gamma_q \, |w|_q \, |v|_q \, , \quad \forall w, v \in X \ . \tag{C.15}$$

Here $|\cdot|_q : H^1(\Omega) \to \mathbb{R}^+$ are seminorms that satisfy

$$C_X = \sup_{w \in X} \frac{\sum_{q=1}^{Q} |w|_q^2}{\|w\|_X^2}, \tag{C.16}$$

for some positive parameter-independent constant C_X. We then define, for $\mu \in \mathcal{D}$, $\overline{\mu} \in \mathcal{D}$,

$$\Phi(\mu, \overline{\mu}) \equiv C_X \max_{q \in \{1,\ldots,Q\}} \left(\Gamma_q \left| \Theta^q(\mu) - \Theta^q(\overline{\mu}) - \sum_{p=1}^{P} (\mu_{(p)} - \overline{\mu}_{(p)}) \frac{\partial \Theta^q}{\partial \mu_{(p)}}(\overline{\mu}) \right| \right) . \tag{C.17}$$

In short, $\mathcal{T}(w, w; \mu - \overline{\mu}; \overline{\mu})/\|w\|_X^2$ and $\mathcal{F}(\mu - \overline{\mu}; \overline{\mu})$ represent the first-order terms in parameter expansions about $\overline{\mu}$ of $\sigma^2(w; \mu)$ and $\beta^2(\mu)$, respectively; and $\Phi(\mu, \overline{\mu})$ is a second-order remainder term that bounds the effect of deviation (of the operator coefficients) from linear parameter dependence.

We now require a parameter sample $V_J \equiv \{\overline{\mu}_1 \in \mathcal{D}, \ldots, \overline{\mu}_J \in \mathcal{D}\}$ and associated sets of polytopes, $P_J \equiv \{\mathcal{P}^{\overline{\mu}_1} \in \mathcal{D}^{\overline{\mu}_1}, \ldots, \mathcal{P}^{\overline{\mu}_J} \in \mathcal{D}^{\overline{\mu}_J}\}$ that satisfy a "Coverage Condition,"

$$\mathcal{D} \subset \bigcup_{j=1}^{J} \mathcal{P}^{\bar{\mu}_j} , \tag{C.18}$$

and a "Positivity Condition,"

$$\min_{\nu \in \mathcal{V}^{\bar{\mu}_j}} \sqrt{\mathcal{F}(\nu - \bar{\mu}_j; \bar{\mu}_j)} - \max_{\mu \in \mathcal{P}^{\bar{\mu}_j}} \Phi(\mu; \bar{\mu}_j) \geq \epsilon_\beta \beta(\bar{\mu}_j), \ 1 \leq j \leq J . \tag{C.19}$$

Here $\mathcal{V}^{\bar{\mu}_j}$ is the set of vertices associated with the polytope $\mathcal{P}^{\bar{\mu}_j}$; and $\epsilon_\beta \in]0, 1[$ is a prescribed accuracy constant. Our lower bound is then given by

$$\hat{\beta}_{\mathrm{PC}}(\mu) \equiv \max_{j \in \{1,\ldots,J\} | \mu \in \mathcal{P}^{\bar{\mu}_j}} \epsilon_\beta \beta(\overline{\mu}_j). \tag{C.20}$$

which is a piecewise–constant approximation for $\beta(\mu)$. We finally introduce an index mapping $\mathcal{I} : \mathcal{D} \to \{1, \ldots, J\}$ such that for any $\mu \in \mathcal{D}$,

$$\mathcal{I}\mu = \arg \max_{j \in \{1,\ldots,J\}} \epsilon_\beta \beta(\overline{\mu}_j), \tag{C.21}$$

for piecewise–constant lower bound. We can readily demonstrate that

C.3 Bound Proof

Proposition 16. *For any V_J and P_J such that the Coverage Condition (C.18) and Positivity Condition (C.19) are satisfied, we have $\epsilon_\beta \beta(\overline{\mu}_{\mathcal{I}_\mu}) = \hat{\beta}_{\mathrm{PC}}(\mu) \leq \beta(\mu)$, $\forall\, \mu \in \mathcal{D}^{\overline{\mu}_{\mathcal{I}_\mu}}$.*

Proof. We first note from (C.1), (C.2), (C.3), (C.13), (C.15), (C.16), (C.17) and Cauchy-Schwarz inequality that

$$
\begin{aligned}
\beta(\mu) &\geq \inf_{w\in X} \frac{|a(w, T^{\overline{\mu}} w; \mu)|}{\|w\|_X \|T^{\overline{\mu}} w\|_X} \\
&= \inf_{w\in X} \frac{|a(w, T^{\overline{\mu}} w; \bar{\mu}) + a(w, T^{\overline{\mu}} w; \mu) - a(w, T^{\overline{\mu}} w; \bar{\mu})|}{\|w\|_X \|T^{\overline{\mu}} w\|_X} \\
&= \inf_{w\in X} \left| \frac{\|T^{\overline{\mu}} w\|_X^2 + \sum_{p=1}^P \sum_{q=1}^Q (\mu_{(p)} - \bar{\mu}_{(p)}) \frac{\partial \Theta^q}{\partial \mu_{(p)}}(\overline{\mu}) a^q(w, T^{\overline{\mu}} w)}{\|w\|_X \|T^{\overline{\mu}} w\|_X} \right. \\
&\quad \left. + \frac{\sum_{q=1}^Q \left(\Theta^q(\mu) - \Theta^q(\overline{\mu}) - \sum_{p=1}^P (\mu_{(p)} - \bar{\mu}_{(p)}) \frac{\partial \Theta^q(\overline{\mu})}{\partial \mu_{(p)}} \right) a^q(w, T^{\overline{\mu}} w)}{\|w\|_X \|T^{\overline{\mu}} w\|_X} \right| \\
&\geq \inf_{w\in X} \frac{\left| \|T^{\overline{\mu}} w\|_X^2 + \sum_{p=1}^P \sum_{q=1}^Q (\mu_{(p)} - \bar{\mu}_{(p)}) \frac{\partial \Theta^q}{\partial \mu_{(p)}}(\overline{\mu}) a^q(w, T^{\overline{\mu}} w) \right|}{\|w\|_X \|T^{\overline{\mu}} w\|_X} \\
&\quad - \sup_{w\in X} \frac{\left| \sum_{q=1}^Q \left(\Theta^q(\mu) - \Theta^q(\overline{\mu}) - \sum_{p=1}^P (\mu_{(p)} - \bar{\mu}_{(p)}) \frac{\partial \Theta^q(\overline{\mu})}{\partial \mu_{(p)}} \right) a^q(w, T^{\overline{\mu}} w) \right|}{\|w\|_X \|T^{\overline{\mu}} w\|_X} \\
&\geq \sqrt{\inf_{w\in X} \frac{\left| \|T^{\overline{\mu}} w\|_X^2 + \sum_{p=1}^P \sum_{q=1}^Q (\mu_{(p)} - \bar{\mu}_{(p)}) \frac{\partial \Theta^q}{\partial \mu_{(p)}}(\overline{\mu}) a^q(w, T^{\overline{\mu}} w) \right|^2}{\|w\|_X^2 \|T^{\overline{\mu}} w\|_X^2}} \\
&\quad - \max_{q\in\{1,\ldots,Q\}} \left(\Gamma_q \left| \Theta^q(\mu) - \Theta^q(\overline{\mu}) - \sum_{p=1}^P (\mu - \overline{\mu})_{(p)} \frac{\partial \Theta^q(\overline{\mu})}{\partial \mu_{(p)}} \right| \right) \sup_{w\in X} \frac{\sum_{q=1}^Q |w|_q |T^{\overline{\mu}} w|_q}{\|w\|_X \|T^{\overline{\mu}} w\|_X} \\
&\geq \left[\mathcal{T}(w, w; \mu - \overline{\mu}; \mu) / \|w\|_X^2 \right]^{\frac{1}{2}} - \Phi(\mu, \overline{\mu})
\end{aligned}
$$

where the last inequality derives from the following inequality

$$
|A + B|^2 \geq (A + \Re(B))^2 \geq (A^2 + 2A\Re(B))^2 = A^2 + A(B + \overline{B})
$$

for $A = \|T^{\overline{\mu}} w\|_X^2$ real and $B = \sum_{p=1}^P \sum_{q=1}^Q (\mu_{(p)} - \bar{\mu}_{(p)}) \frac{\partial \Theta^q}{\partial \mu_{(p)}}(\overline{\mu}) a^q(w, T^{\overline{\mu}} w)$ complex.

It thus follows that

$$\beta(\mu) \geq \sqrt{\mathcal{F}(\mu - \overline{\mu}; \overline{\mu})} - \Phi(\mu, \overline{\mu}) \ . \tag{C.22}$$

The desired result finally follows from the construction of V_J and P_J, the definition of $\hat{\beta}_{\rm PC}(\mu)$ and $\hat{\beta}_{\rm PL}(\mu)$, and the concavity of $\mathcal{F}(\mu - \overline{\mu}; \overline{\mu})$ in μ. □

Of course, if $-\Phi(\mu, \overline{\mu})$ is a concave function of μ, we can develop a (better) piecewise-linear lower bound $\hat{\beta}_{\rm PL}(\mu)$.

C.4 Discrete Eigenvalue Problems

We give a short discussion of the numerical calculation of the inf-sup parameter $\beta(\mu)$ and the Rayleigh quotient $\mathcal{F}(\mu - \bar{\mu}; \bar{\mu})$. To begin, we denote by $\underline{A}(\mu)$, $\underline{A}^q$, $\underline{C}$ the finite element matrices associated with $a(\cdot, \cdot; \mu)$, $a^q(\cdot, \cdot)$, $(\cdot, \cdot)_X$, respectively. We introduce the discrete eigenproblem: Given $\mu \in \mathcal{D}$, find the minimum eigenmode $(\underline{\chi}_{\rm min}(\mu), \lambda_{\rm min}(\mu))$ such that

$$(\underline{A}(\mu))^H \underline{C}^{-1} \underline{A}(\mu) \underline{\chi}_{\rm min}(\mu) = \lambda_{\rm min}(\mu) \underline{C} \underline{\chi}_{\rm min}(\mu) \ , \tag{C.23}$$

$$(\underline{\chi}_{\rm min}(\mu))^H \underline{C} \underline{\chi}_{\rm min}(\mu) = 1 \ . \tag{C.24}$$

The discrete value of $\beta(\mu)$ is then $\sqrt{\lambda_{\rm min}(\mu)}$. Note the notion of the "H" Hermitian transpose in the complex case, the Hermitian transpose of a complex quantity (such as a complex number, complex vector, or complex matrix) is in fact the complex conjugate transpose of itself.

The computation of $\mathcal{F}$ involving a more complex eigenvalue problem is rather complicated and more expensive. In particular, we first write the matrix form $\underline{\mathcal{T}}(\mu - \bar{\mu}; \bar{\mu})$ of the bilinear form $\mathcal{T}(\cdot, \cdot; \mu - \bar{\mu}; \bar{\mu})$ in (C.13) as

$$\begin{aligned} \underline{\mathcal{T}}(\mu - \bar{\mu}; \bar{\mu}) = \ & (\underline{A}(\overline{\mu}))^H \underline{C}^{-1} \underline{A}(\overline{\mu}) + \sum_{p=1}^{P} (\mu - \bar{\mu})_{(p)} \\ & \left\{ \sum_{q=1}^{Q} \frac{\partial \Theta^q(\overline{\mu})}{\partial \mu_{(p)}} (\underline{A}^q)^H \underline{C}^{-1} \underline{A}(\overline{\mu}) + \sum_{q=1}^{Q} \frac{\partial \overline{\Theta}^q(\overline{\mu})}{\partial \mu_{(p)}} (\underline{A}(\overline{\mu}))^H \underline{C}^{-1} \underline{A}^q \right\} . \end{aligned}$$

Next we introduce the second discrete eigenproblem: Given a pair $(\mu \in \mathcal{D}, \bar{\mu} \in \mathcal{D})$, find $\underline{\Psi}_{\min}(\mu - \bar{\mu}; \bar{\mu}) \in \mathbb{R}^{\mathcal{N}}$, $\rho_{\min}(\mu - \bar{\mu}; \bar{\mu}) \in \mathbb{R}$ such that

$$\underline{\mathcal{T}}(\mu - \bar{\mu}; \bar{\mu})\, \underline{\Psi}_{\min}(\mu - \bar{\mu}; \bar{\mu}) \;=\; \rho_{\min}(\mu - \bar{\mu}; \bar{\mu}) \underline{C}\, \underline{\Psi}_{\min}(\mu - \bar{\mu}; \bar{\mu}) \;, \quad \text{(C.25)}$$

$$(\underline{\Psi}_{\min}(\mu - \bar{\mu}; \bar{\mu}))^H\, \underline{C}\, \underline{\Psi}_{\min}(\mu - \bar{\mu}; \bar{\mu}) \;=\; 1 \;. \quad \text{(C.26)}$$

Note that the matrix $\underline{\mathcal{T}}(\mu - \bar{\mu}; \bar{\mu})$ is symmetric (more precisely, conjugate symmetric) and that $\mathcal{F}_{\min}(\mu - \bar{\mu}; \bar{\mu})$ is essentially the square root of the minimum eigenvalue, i.e., we have $\mathcal{F}_{\min}(\mu - \bar{\mu}; \bar{\mu}) = \sqrt{\rho_{\min}(\mu - \bar{\mu}; \bar{\mu})}, \forall\, \mu \in \mathcal{D}^{\bar{\mu}}$. The eigenproblem (C.25)-(C.26) can be solved by using the Lanczos procedure without calculating $\underline{C}^{-1}$ explicitly. During the Lanczos procedure we often compute $\underline{w}(\mu) = \underline{T}(\mu; \bar{\mu})\underline{v}$ for some $\underline{v}$ and do this as follows: first solve the linear algebraic systems

$$\underline{C}\underline{y}^0 \;=\; \underline{A}(\overline{\mu})\underline{v}$$

and

$$\underline{C}\underline{y}^q = \underline{A}^q\underline{v}$$

for $\underline{y}^0$ and $\underline{y}^q, 1 \le q \le Q$, respectively, and then simply set

$$\underline{w}(\mu) \;=\; (\underline{A}(\overline{\mu}))^H\underline{y}^0 - \sum_{p=1}^{P}(\mu - \bar{\mu})_{(p)} \left[\sum_{q=1}^{Q} \frac{\partial \Theta^q(\overline{\mu})}{\partial \mu_{(p)}} (\underline{A}^q)^H \underline{y}^0 + \sum_{q=1}^{Q} \frac{\partial \overline{\Theta}^q(\overline{\mu})}{\partial \mu_{(p)}} (\underline{A}(\overline{\mu}))^H \underline{y}^q\right] \;.$$

Appendix D

Three-Dimensional Inverse Scattering Example

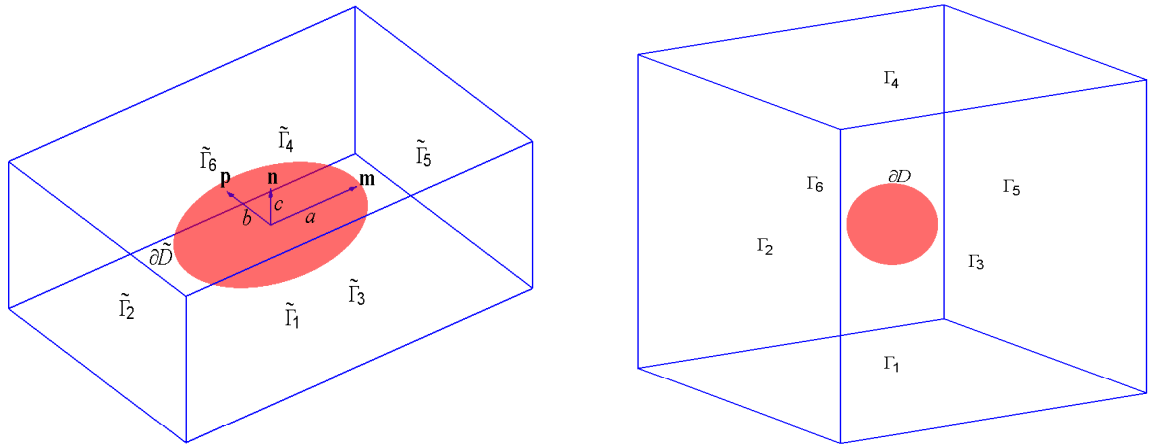
D.1 Problem Description

In this example we seek to identify a three-dimensional ellipsoid $\tilde{D}$ of three unknown semiaxes (a, b, c) and three unknown orientations $(\mathbf{m}, \mathbf{n}, \mathbf{p})$ (which together describe the size and shape of the ellipsoid) from the experimental data given in the form of intervals. The experimental data is the far field pattern $\tilde{u}_\infty$ measured at several angles $\tilde{d}^s$ with experimental error ϵ_{exp} for one or several incident directions $\tilde{d}$ and wave numbers k. Hence, the input μ consists of $(a, b, c, \mathbf{m}, \mathbf{n}, \mathbf{p}, k, \tilde{d}, \tilde{d}^s)$; and the output is $\tilde{u}_\infty(\mu)$.

D.2 Domain truncation and Mapping

The truncated domain $\tilde{\Omega}$ is bounded by the ellipsoid and an artificial boundary $\tilde{\Gamma}$. Here $\tilde{\Gamma}$ is an oblique box of size $16a \times 16b \times 16c$ which has the same orientation as the ellipsoid and is scaled with the three semiaxes as shown in Figure D-1(a). Hence, the mean curvature $\tilde{\mathcal{H}}(\cdot; \mu)$ is zero for the chosen boundary $\tilde{\Gamma}$ except for the corner points and can thus be ignored in our formulation of the direct scattering problem. Furthermore, we define a reference domain Ω corresponding to the geometry bounded by the unit sphere and a perpendicular box of size $16 \times 16 \times 16$ as shown in Figure D-1(b).

We now map $\tilde{\Omega}(a, b, c, \mathbf{m}, \mathbf{n}, \mathbf{p}) \to \Omega$ via a continuous piecewise-affine transformation.



(a) (b)

Figure D-1: Three-dimensional scattering problem: (a) original (parameter-dependent) domain and (b) reference domain.

This can be done in two steps. In the first step, we map the three orientations $(\mathbf{m}, \mathbf{n}, \mathbf{p})$ to the three Cartesian unit vectors $(\mathbf{i}_x, \mathbf{i}_y, \mathbf{i}_z)$ by: first rotating $(\mathbf{m}, \mathbf{n}, \mathbf{p})$ about the z-axis an angle $-\alpha$ and about the y-axis an angle β so that $\mathbf{m}$ coincides with $\mathbf{i}_x$ and $\mathbf{n}$ is in the yz plan; and then rotating the resulting orientations $(\mathbf{m}, \mathbf{n}, \mathbf{p})$ about x-axis an angle $-\gamma$ so that $\mathbf{n}$ coincides with $\mathbf{i}_y$. The rotation transformation is thus given by

$$R(\mu) = R_x(-\gamma)R_y(\beta)R_z(-\alpha) \tag{D.1}$$

where

$$R_x(\phi) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\phi & -\sin\phi \\ 0 & \sin\phi & \cos\phi \end{bmatrix}, \quad R_y(\phi) = \begin{bmatrix} \cos\phi & 0 & \sin\phi \\ 0 & 1 & 0 \\ -\sin\phi & 0 & \cos\phi \end{bmatrix}, \tag{D.2}$$

$$R_z(\phi) = \begin{bmatrix} \cos\phi & -\sin\phi & 0 \\ \sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{D.3}$$

In essence, we have transformed the original domain with the orientated ellipsoid to another domain with a perpendicular ellipsoid. In the second step, we map the resulting

domain with the perpendicular ellipsoid to the reference domain with the unit sphere by the stretching transformation

$$S(\mu) = \begin{bmatrix} \frac{1}{a} & 0 & 0 \\ 0 & \frac{1}{b} & 0 \\ 0 & 0 & \frac{1}{c} \end{bmatrix}. \tag{D.4}$$

It finally follows from (D.1)-(D.4) that the geometric transformation that map $\tilde{\Omega}$ to Ω is given by

$$G(\mu) = S(\mu)R(\mu) = \begin{bmatrix} \frac{\cos\alpha\cos\beta}{a} & \frac{\sin\alpha\cos\beta}{a} & \frac{\sin\beta}{a} \\ \frac{-\sin\alpha\cos\gamma-\cos\alpha\sin\beta\sin\gamma}{b} & \frac{\cos\alpha\cos\gamma-\sin\alpha\sin\beta\sin\gamma}{b} & \frac{\cos\beta\sin\gamma}{b} \\ \frac{\sin\alpha\sin\gamma-\cos\alpha\sin\beta\cos\gamma}{c} & \frac{-\cos\alpha\sin\gamma-\sin\alpha\sin\beta\cos\gamma}{c} & \frac{\cos\beta\cos\gamma}{c} \end{bmatrix}. \tag{D.5}$$

We see that the orientations of the ellipsoid can be described more conveniently by three angles (α, β, γ) than by three vectors $(\mathbf{m}, \mathbf{n}, \mathbf{p})$. From now on, we shall thus use $(a, b, c, \alpha, \beta, \gamma, k, \tilde{d}, \tilde{d}^s)$ for the input.

D.3 Forms in Reference Domain

To begin, we note that the geometric transformation $\tilde{\Omega} \rightarrow \Omega$ is affine, the problem can thus be recast precisely in the desired abstract form (10.31), in which Ω, X defined in (10.21), and $(w; v)_X$ defined in (10.22) are independent of the parameter μ; furthermore, our affine assumption applies for $Q = 7$. We summarize the $\Theta^q(\mu), a^q(w, v), 1 \leq q \leq Q$, in Table D.3. We can then choose $|v|_q^2 = a^q(v, v)$, $1 \leq q \leq Q$, since the $a^q(\cdot, \cdot)$ are positive semi-definite; it thus follows that $C_X = 1.0000$ and that $\Gamma^q = 1$, $1 \leq q \leq Q$, by the Cauchy-Schwarz inequality .

To derive the explicit form for $g(x; \mu)$, $h(x; \mu)$ and $\ell^o(x; \mu)$, we first note from (10.20) that

$$J_d(x; \mu) = \sqrt{b^2c^2x_1^2 + a^2c^2x_2^2 + a^2b^2x_3^2}\,, \tag{D.6}$$

since ∂D is the unit sphere. Furthermore, it can be easily shown that the unit normal

q	$\Theta^q(\mu)$	$a^q(w,v)$
1	$\frac{bc}{a}$	$\int_\Omega \frac{\partial w}{\partial x}\frac{\partial \bar{v}}{\partial x}$
2	$\frac{ca}{b}$	$\int_\Omega \frac{\partial w}{\partial y}\frac{\partial \bar{v}}{\partial x_2}$
3	$\frac{ab}{c}$	$\int_\Omega \frac{\partial w}{\partial x_3}\frac{\partial \bar{v}}{\partial x_2}$
4	$-k^2abc$	$\int_\Omega w\bar{v}$
5	$-\mathrm{i}kab$	$\int_{\Gamma_1} w\bar{v} + \int_{\Gamma_4} w\bar{v}$
6	$-\mathrm{i}kbc$	$\int_{\Gamma_2} w\bar{v} + \int_{\Gamma_5} w\bar{v}$
7	$-\mathrm{i}kca$	$\int_{\Gamma_3} w\bar{v} + \int_{\Gamma_6} w\bar{v}$

Table D.1: Parametric functions $\Theta^q(\mu)$ and parameter-independent bilinear forms $a^q(w,v)$ for the three-dimensional inverse scattering problem.

vector $\tilde{\nu}$ to the boundary $\partial\tilde{D}$ can be expressed in terms of the reference coordinate as

$$\tilde{\nu}(x;\mu) = \frac{\tilde{v}(x;\mu)}{J_d(x;\mu)} \ , \tag{D.7}$$

where the three components of $v(x;\mu)$ are given by

$$\begin{aligned}
\tilde{v}_1(x;\mu) &= bcx_1\cos\alpha\cos\beta - acx_2\left(\sin\alpha\cos\gamma + \cos\alpha\sin\beta\sin\gamma\right) \ , \\
&\quad + abx_3\left(\sin\alpha\sin\gamma - \cos\alpha\sin\beta\cos\gamma\right) &\text{(D.8)}\\
\tilde{v}_2(x;\mu) &= bcx_1\sin\alpha\cos\beta + acx_2\left(\cos\alpha\cos\gamma - \sin\alpha\sin\beta\sin\gamma\right) \ , \\
&\quad - abx_3\left(\cos\alpha\sin\gamma + \sin\alpha\sin\beta\cos\gamma\right) &\text{(D.9)}\\
\tilde{v}_3(x;\mu) &= bcx_1\sin\beta + acx_2\cos\beta\sin\gamma + abx_3\cos\beta\cos\gamma \ . &\text{(D.10)}
\end{aligned}$$

It finally follows from (10.28), (10.33), and (D.5)-(D.10) that

$$\begin{aligned}
g(x;\mu) &= \mathrm{i}k\tilde{d}\cdot\tilde{v}e^{\mathrm{i}k\tilde{d}\cdot(G^{-1}(\mu)x)} \ , &\text{(D.11)}\\
h(x;\mu) &= \mathrm{i}k\tilde{d}^s\cdot\tilde{v}e^{-\mathrm{i}k\tilde{d}^s\cdot(G^{-1}(\mu)x)} \ , &\text{(D.12)}\\
\ell^o(x;\mu) &= \beta_n\int_{\partial D}\mathrm{i}k\tilde{d}\cdot\tilde{v}e^{\mathrm{i}k\tilde{d}\cdot(G^{-1}(\mu)x)}e^{-\mathrm{i}k\tilde{d}^s\cdot(G^{-1}(\mu)x)} \ , &\text{(D.13)}
\end{aligned}$$

which are functions of the coordinate x and the parameter μ.

Bibliography

[1] M. Ainsworth and J. T. Oden. A posteriori error estimation in finite element analysis. *Comp. Meth. Appl. Mech. Engrg.*, 142:1–88, 1997.

[2] G. Alessandrini and L. Rondi. Stable determination of a crack in a planar inhomogeneous conductor. *SIAM J. MATH. ANAL.*, 30(2):326–340, 1998.

[3] S. Ali. *Real-time Optimal Parametric Design using the Assess-Predict-Optimize Strategy.* PhD thesis, Singapore-MIT Alliance, Nanyang Technological University, Singapore, 2003. In progress.

[4] B. O. Almroth, P. Stern, and F. A. Brogan. Automatic choice of global shape functions in structural analysis. *AIAA Journal*, 16:525–528, May 1978.

[5] X. Antoine, H. Barucq, and A. Bendali. Bayliss-turkel-like radiation conditions on surfaces of arbitrary shape. *Journal of Mathematical Analysis and Applications*, 229:184–221, 1999.

[6] I. Babuška and W.C. Rheinboldt. *A Posteriori* error estimates for the finite element method. *Int. J. Numer. Meth. Engrg.*, 12:1597–1615, 1978.

[7] Z. Bai. Krylov subspace techniques for reduced-order modeling of large-scale dynamical systems. *Applied Numerical Mathematics*, 43:9–44, 2002.

[8] A. B. Bakushinskii. The problem of the convergence of the iteratively regularized gauss-newton method. *Comput. Math. Math. Phys.*, 32:1353–1359, 1992.

[9] E. Balmes. Parametric families of reduced finite element models: Theory and applications. *Mechanical Systems and Signal Processing*, 10(4):381–394, 1996.

[10] R. E. Bank and A. Weiser. Some a posteriori error estimators for elliptic partial differential equations. *Math. Comput.*, 44(170):283–301, 1985.

[11] H. T. Banks and K. L. Bihari. Modelling and estimating uncertainty in parameter estimation. *Inverse Problems*, 17:95–111, 2001.

[12] H. T. Banks and K. Kunisch. *Estimation Techniques for Distributed Parameter Systems*. Birkhauser, Boston, 1989.

[13] E. Barkanov. Transient response analysis of structures made from viscoelasticity materials. *Int. J. Numer. Meth. Engng.*, 44:393–403, 1999.

[14] M. Barrault, N. C. Nguyen, Y. Maday, and A. T. Patera. An "empirical interpolation" method: Application to efficient reduced-basis discretization of partial differential equations. *C. R. Acad. Sci. Paris, Série I*, 2004. Submitted.

[15] K.J. Bathe. *Finite Element Procedure*. Prentice-Hall, Inc., 1996.

[16] A. Baussard, Denis Prémel, and O. Venard. A bayesian approach for solving inverse scattering from microwave laboratory-controlled data. *Inverse Problems*, 17:1659–1669, 2001.

[17] R. Becker and R. Rannacher. A feedback approach to error control in finite element method: Basic analysis and examples. *East - West J. Numer. Math.*, 4:237–264, 1996.

[18] R. Becker and R. Rannacher. Weighted a posteriori error control in fe methods. Preprint, october 1996.

[19] P. M. Van Den Berg and R. E. Kleinman. Gradient methods in inverse acoustic and electromagnetic scattering. In P.G. Ciarlet and J.L. Lions, editors, *Mathematics and its Applications, Vol. 92*, Large-Scale Optimization with Applications (Part 1), pages 173–194. Springer-Verlag, New York, 1997.

[20] B. Blaschke, A. Neubauer, and O. Scherzer. On the convergence rates for the iteratively regularized gauss-newton method. *IMA Journal of Numerical Analysis*, 17:421–436, 1997.

[21] B. Borden. Mathematical problems in radar inverse scattering. *Inverse Problems*, 18:R1–R28, 2002.

[22] T.T. Bui, M. Damodaran, and K. Wilcox. Proper orthogonal decomposition extensions for parametric applications in transonic aerodynamics (AIAA Paper 2003-4213). In *Proceedings of the 15th AIAA Computational Fluid Dynamics Conference*, June 2003.

[23] M. Burger and B. Kaltenbacher. Regularizing newton-kaczmarz methods for nonlinear ill-posed problems. 2004. SFB-Report 04-17.

[24] C. Byrne. Block-iterative interior point optimization methods for image reconstruction from limited data. *Inverse Problems*, 16:1405–1419, 2000.

[25] G. Chen, H. Mu, D. Pommerenke, and J. L. Drewniak. Damage detection of reinforced concrete beams with novel distributed crack/strain sensors. *Structural Health Monitoring*, 3:225–243, 2004.

[26] J. Chen and S-M. Kang. Model-order reduction of nonlinear mems devices through arclength-based karhunen-love decomposition. In *Proceeding of the IEEE international Symposium on Circuits and Systems*, volume 2, pages 457–460, 2001.

[27] Y. Chen and J. White. A quadratic method for nonlinear model order reduction. In *Proceeding of the international Conference on Modeling and Simulation of Microsystems*, pages 477–480, 2000.

[28] M. Chou and J. White. Efficient formulation and model-order reduction for the transient simulation of three-dimensional vlsi interconnect. In *IEEE Transactions On Computer-Aided Design of Integrated Circuit and Systems*, volume 16, pages 1454–1476, 1997.

[29] D. Colton. Inverse acoustic and electromagnetic scattering theory. *Inverse Problems*, 47:67–110, 2003.

[30] D. Colton, J. Coyle, and P. Monk. Recent developments in inverse acoustic scattering theory. *SIAM Review*, 42:369–414, 2000.

[31] D. Colton, K. Giebermann, and P. Monk. A regularized sampling method for solving three dimensional inverse scattering problems. *SIAM J. Sci. Comput.*, 21:2316–2330, 2000.

[32] D. Colton, H. Haddar, and P. Monk. The linear sampling method for solving the electromagnetic inverse scattering problem. *SIAM J. Sci. Comput.*, 24:719–731, 2002.

[33] D. Colton and R. Kress. *Inverse Acoustic and Electromagnetic Scattering Theory*. Springer, 1998.

[34] D. Colton and P. Monk. A linear sampling method for the detection of leukemia using microwaves. *SIAM J. Appl. Math.*, 58:926–941, 1998.

[35] Tie Jun Cui, Yao Qin, Gong-Li Wang, and Weng Cho Chew. Low-frequency detection of two-dimensional buried objects using high-order extended born approximations. *Inverse Problems*, 20:S41–S62, 2004.

[36] P. Deuflhard, H. W. Engl, and O. Scherzer. A convergence analysis of iterative methods for the solution of nonlinear ill-posed problems under affinely invariant conditions. *Inverse Problems*, 14:1081–1106, 1998.

[37] E. Dowell and K. Hall. Modeling of fluid-structure interaction. *Annual Review of Fluid Mechanics*, 33:445–490, 2001.

[38] B. Duchêne, A. Joisel, and M. Lambert. Nonlinear inversions of immersed objects using laboratory-controlled data. *Inverse Problems*, 20:S81–S98, 2004.

[39] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Kluwer Academic, Dordrecht, 1996.

[40] H. W. Engl and P. Kugler. Nonlinear inverse problems: Theoretical aspects and some industrial applications. In *Multidisciplinary Methods for Analysis, Optimization and Control of Complex Systems*. Springer Verlag, 2004. To appear.

[41] H. W. Engl, K. Kunisch, and A. Neubauer. Convergence rates for tikhonov regularization of nonlinear ill-posed problems. *Inverse Problems*, 5:523–540, 1989.

[42] H. W. Engl and J. Zou. A new approach to convergence rate analysis of tikhonov regularization for parameter identification in heat conduction. *Inverse Problems*, 16:1907–1923, 2000.

[43] B. Epureanu, E. Dowell, and K. Hall. A parametric analysis of reduced order models of potential flows in turbomachinery using proper orthogonal decomposition. In *Proceedings of ASME TURBO EXPO 2001*, pages 2001–GT–0434, New Orleans, Louisiana, June 2001.

[44] C. Farhat, R. Tezaur, and R. Djellouli. On the solution of three-dimensional inverse obstacle acoustic scattering problems by a regularized newton method. *Inverse Problems*, 18:1229–1246, 2002.

[45] J. P. Fink and W. C. Rheinboldt. On the error behavior of the reduced basis technique for nonlinear finite element approximations. *Z. Angew. Math. Mech.*, 63:21–28, 1983.

[46] B. G. Fitzpatrick. Bayesian analysis in inverse problems. *Inverse Problems*, 7:675–702, 1991.

[47] W. Flugge. *Tensor Analysis and Continuum Mechanics*. Springer-Verlag, New York, 1972.

[48] C. Fox and G. Nicholls. Statistical estimation of the parameters of a pde. *Canadian Applied Mathematics Quarterly*, 10:277–306, 2002.

[49] A. Friedman and M. Vogelius. Determining cracks by boundary measurements. *Indiana Univ. Math. J.*, 38:527–556, 1989.

[50] F.Wang and J. White. Automatic model order reduction of a microdevice using the arnoldi approach. In *Proceedings of the International Mechanical Engineering Congress and Exposition*, pages 527–530, 1998.

[51] M. Grepl. *Reduced-Basis Approximations for Time-Dependent Partial Differential Equations: Application to Optimal Control*. PhD thesis, Massachusetts Institute of Technology, 2005. In progress.

[52] M. A. Grepl, Y. Maday, N. C. Nguyen, and A. T. Patera. Efficient approximation and *a posteriori* error estimation for reduced-basis treatment of nonaffine and nonlinear partial differential equations. *M2AN Math. Model. Numer. Anal.*, 2005. Working Paper.

[53] M. A. Grepl, N. C. Nguyen, K.Veroy A. T. Patera, and G. R. Liu. Certified rapid solution of parametrized partial differential equations for real-time applications. In *Proceedings of the 2nd Sandia Workshop of PDE-Constrained Optimization: Towards Real-Time and On-Line PDE-Constrained Optimization*, SIAM Computational Science and Engineering Book Series, 2004. submitted for consideration.

[54] M. A. Grepl and A. T. Patera. Reduced-basis approximation for time-dependent parametrized partial differential equations. *M2AN Math. Model. Numer. Anal.*, 2004. Submitted.

[55] E.J. Grimme. *Krylov Projection Methods for Model Reduction*. PhD thesis, University of Illinois at Urbana-Champain, 1997.

[56] M. Gustafsson. Multi-static synthetic aperture radar and inverse scattering. 2004. Techniqual Report.

[57] K. Hall, J. Thomas, and E. Dowell. Proper orthogonal decomposition technique for transonic unsteady aerodynamic flows. *AIAA*, 38:1853–1862, 2000.

[58] X. Han, D. Xu, and G. R. Liu. A computational inverse technique for material characterization of a functionally graded cylinder using a progressive neural network. *Neurocomputing*, 51:341–360, 2003.

[59] M. Hanke. A regularizing levenberg-marquardt scheme with applications to inverse groundwater filtration problems. *Inverse Problems*, 13:79–95, 1997.

[60] M. Hanke, A. Neubauer, and O. Scherzer. A convergence analysis of the landweber iteration for nonlinear ill-posed problems. *Numeriches Mathematik*, 72:21–37, 1995.

[61] F. Hettlich. On the uniqueness of the inverse conductive scattering problem for the helmholtz equation. *Inverse Problems*, 10:129–144, 1994.

[62] T. Hohage. On the numerical solution of a three-dimensional inverse medium scattering problem. *Inverse Problems*, 17:1743–1763, 2001.

[63] V. Isakov. *Inverse problems for partial differential equations*, volume 127 of *Applied mathematical sciences*. Springer, 1997.

[64] S. I. Ishak, G. R. Liu, S. P. Lim, and H. M. Shang. Locating and sizing of delamination in composite laminates using computational and experimental methods. *Composite Part B*, 32:287–298, 2001.

[65] K. Ito and S. S. Ravindran. A reduced-order method for simulation and control of fluid flows. *Journal of Computational Physics*, 143(2):403–425, July 1998.

[66] K. Ito and S. S. Ravindran. Reduced basis method for optimal control of unsteady viscous flows. *International Journal of Computational Fluid Dynamics*, 15(2):97–113, 2001.

[67] Q. N. Jin. On the iteratively regularized gauss-newton method for solving nonlinear ill-posed problems. *Mathematics of Computation*, 69(232):1603–1623, 2000.

[68] J. P. Kaipio, V. Kolehmainen, E. Somersalo, and M. Vauhkonen. Statistical inversion and monte carlo sampling methods in electrical impedance tomography. *Inverse Problems*, 16:1487–1522, 2000.

[69] A. Kirsch. Uniqueness theorems in inverse scattering theory for periodic structures. *Inverse Problems*, 10:145–152, 1994.

[70] A. Kirsch and R. Kress. Uniqueness in inverse obstacle scattering. *Inverse Problems*, 9:285–299, 1993.

[71] A. D. Klose and A. H. Hielscher. Quasi-newton methods in optical tomographic image reconstruction. *Inverse Problems*, 19:387–409, 2003.

[72] M.E. Kowalski and J-M. Jin. Karhunen-love based model order reduction of nonlinear systems. In *Proceeding of the International Conference on Modeling and Simulation of Microsystems*, volume 1, pages 552–555, 2002.

[73] R. Kress and W. Rundell. A quasi-newton method in inverse obstacle scattering. *Inverse Problems*, 10:1145–1157, 1994.

[74] P. Ladeveze and D. Leguillon. Error estimation procedures in the finite element method and applications. *SIAM J. Numer. Anal.*, 20:485–509, 1983.

[75] G. Lassaux and K. Wilcox. Model reduction for active control design using multi-point arnoldi methods (AIAA Paper 2003-0616). In *Proceedings of the 41st Aerospace Sciences Meeting and Exibit*, January 2003.

[76] Peter D. Lax. *Functional Analysis*. New York, Wiley, 2001.

[77] G. R. Liu and J. D. Achenbach. A strip element method for stress analysis of anisotropic linearly elastic solid. *ASME J. Applied Mechanics*, 61:270–277, 1994.

[78] G. R. Liu and J. D. Achenbach. Strip element method to analyze wave scattering by cracks in anisotropic laminated plates. *ASME J. Applied Mechanics*, 62:607–613, 1995.

[79] G. R. Liu and S. C. Chen. Flaw detection in sandwich plates based on time-harmonic response using genetic algorithm. *Comput. Methods Appl. Mech. Engrg.*, 190:5505–5514, 2001.

[80] G. R. Liu, X. Han, and K. Y. Lam. A combined genetic algorithm and nonlinear least squares method for material characterization using elastic waves. *Comput. Methods Appl. Mech. Engrg.*, 191:1909–1921, 2002.

[81] G. R. Liu and K. Y. Lam. Characterization of a horizontal crack in anisotropic laminated plates. *International Journal of Solids and Structures*, 31:2965–2977, 1994.

[82] G. R. Liu, K. Y. Lam, and J. Tani. Characterization of flaws in sandwich plates: Numerical experiment. *JSME International Journal*, 38:554–562, 1995.

[83] G. R. Liu, Z. C. Xi, K. Y. Lam, and H. M. Shang. A strip element method for analyzing wave scattering by a crack in an immersed composite laminate. *Journal of Applied Mechanics*, 66:898–903, 1999.

[84] H.V. Ly and H.T. Tran. Modeling and control of physical processes using proper orthogonal decomposition. *Journal of Mathematical and Computer Modeling*, 1999.

[85] L. Machiels, Y. Maday, I. B. Oliveira, A. T. Patera, and D.V. Rovas. Output bounds for reduced-basis approximations of symmetric positive definite eigenvalue problems. *C. R. Acad. Sci. Paris, Série I*, 331(2):153–158, July 2000.

[86] L. Machiels, Y. Maday, and A. T. Patera. Output bounds for reduced-order approximations of elliptic partial differential equations. *Comp. Meth. Appl. Mech. Engrg.*, 190(26-27):3413–3426, 2001.

[87] L. Machiels, Y. Maday, A. T. Patera, and D. V. Rovas. Blackbox reduced-basis output bound methods for shape optimization. In *Proceedings* 12^{th} *International Domain Decomposition Conference*, pages 429–436, Chiba, Japan, 2000.

[88] L. Machiels, A. T. Patera, J. Peraire, and Y. Maday. A general framework for finite element a posteriori error control: Application to linear and nonlinear convection-dominated problems. In *ICFD Conference on numerical methods for fluid dynamics*, Oxford, England, 1998.

[89] L. Machiels, J. Peraire, and A. T. Patera. A posteriori finite element output bounds for the incompressible Navier-Stokes equations; Application to a natural convection problem. *Journal of Computational Physics*, 172:401–425, 2001.

[90] Y. Maday, A. T. Patera, and J. Peraire. A general formulation for a posteriori bounds for output functionals of partial differential equations; Application to the eigenvalue problem. *C. R. Acad. Sci. Paris, Série I*, 328:823–828, 1999.

[91] Y. Maday, A. T. Patera, and D. V. Rovas. A blackbox reduced-basis output bound method for noncoercive linear problems. In D. Cioranescu and J.-L. Lions, editors, *Nonlinear Partial Differential Equations and Their Applications, Collége de France Seminar Volume XIV*, pages 533–569. Elsevier Science B.V., 2002.

[92] Y. Maday, A. T. Patera, and D.V. Rovas. Petrov-Galerkin reduced-basis approximations to noncoercive linear partial differential equations. In progress.

[93] Y. Maday, A. T. Patera, and G. Turinici. Global *a priori* convergence theory for reduced-basis approximation of single-parameter symmetric coercive elliptic partial differential equations. *C. R. Acad. Sci. Paris, Série I*, 335(3):289–294, 2002.

[94] Y. Maday, A. T. Patera, and G. Turinici. A priori convergence theory for reduced–basis approximations of single-parameter elliptic partial differential equations. *Journal of Scientific Computing*, 17(1–4):437–446, December 2002.

[95] G. Meurant. *Computer solution of large linear systems*. Elsevier, 1999.

[96] M. Meyer and H.G. Matthies. Efficient model reduction in nonlinear dynamics using the karhunen-love expansion and dual-weighted-residual methods. *Computational Mechanics*, 31:179–191, 2003.

[97] K. Mosegaard and M. Sambridge. Monte carlo analysis of inverse problems. *Inverse Problems*, 18:R29–R54, 2002.

[98] K. Mosegaard and A. Tarantola. Probabilistic approach to inverse problems. *International Handbook of Earthquake and Engineering Seismology, Part A*, pages 237–265, 2002.

[99] N. C. Nguyen, K. Veroy, and A. T. Patera. Certified real-time solution of parametrized partial differential equations. In *Handbook of Materials Modeling*. Kluwer Academic Publishing, 2004. To appear.

[100] A. K. Noor, C. D. Balch, and M. A. Shibut. Reduction methods for non-linear steady-state thermal analysis. *Int. J. Num. Meth. Engrg.*, 20:1323–1348, 1984.

[101] A. K. Noor and J. M. Peters. Reduced basis technique for nonlinear analysis of structures. *AIAA Journal*, 18(4):455–462, April 1980.

[102] A. K. Noor and J. M. Peters. Multiple-parameter reduced basis technique for bifurcation and post-buckling analysis of composite plates. *Int. J. Num. Meth. Engrg.*, 19:1783–1803, 1983.

[103] I. B. Oliveira and A. T. Patera. Reliable real-time optimization of nonconvex systems described by parametrized partial differential equations. In *Proceedings Singapore-MIT Alliance Symposium*, January 2003.

[104] I. B. Oliveira and A. T. Patera. Reduced-basis techniques for rapid reliable optimization of systems described by parametric partial differential equations. *Optimization and Engineering*, 2004. To appear.

[105] M. Paraschivoiu and A. T. Patera. A hierarchical duality approach to bounds for the outputs of partial differential equations. *Comp. Meth. Appl. Mech. Engrg.*, 158(3-4):389–407, June 1998.

[106] M. Paraschivoiu, J. Peraire, Y. Maday, and A. T. Patera. Fast bounds for outputs of partial differential equations. In J. Borgaard, J. Burns, E. Cliff, and S. Schreck, editors, *Computational methods for optimal design and control*, pages 323–360. Birkhäuser, 1998.

[107] M. Paraschivoiu, J. Peraire, and A. T. Patera. *A Posteriori* finite element bounds for linear-functional outputs of elliptic partial differential equations. *Comp. Meth. Appl. Mech. Engrg.*, 150:289–312, 1997.

[108] A. T. Patera and E. M. Rønquist. A general output bound result: Application to discretization and iteration error estimation and control. *Math. Models Methods Appl. Sci.*, 11(4):685–712, 2001.

[109] A. T. Patera, D. Rovas, and L. Machiels. Reduced–basis output–bound methods for elliptic partial differential equations. *SIAG/OPT Views-and-News*, 11(1), April 2000.

[110] J. Peraire and A. T. Patera. Bounds for linear-functional outputs of coercive partial differential equations: Local indicators and adaptive refinement. In P. Ladeveze and J.T. Oden, editors, *Proceedings of the Workshop on New Advances in Adaptive Computational Methods in Mechanics*. Elsevier, 1997.

[111] J. Peraire and A. T. Patera. Asymptotic a Posteriori finite element bounds for the outputs of noncoercive problems: the helhmoltz and burger equations. *Comp. Meth. Appl. Mech. Engrg.*, 171:77–86, 1999.

[112] S. Pereverzev and E. Schock. Morozov's discrepancy principle for tikhonov regularization of severely ill-posed problems in finite-dimensional subspaces. *Numer. Funct. Anal. Optim.*, 21:901–916, 2000.

[113] J. S. Peterson. The reduced basis method for incompressible viscous flow calculations. *SIAM J. Sci. Stat. Comput.*, 10(4):777–786, July 1989.

[114] J.R. Phillips. Projection frameworks for model reduction of weakly nonlinear systems. In *Proceeding of the 37th ACM/IEEE Design Automation Conference*, pages 184–189, 2000.

[115] J.R. Phillips. Projection-based approaches for model reduction of weakly nonlinear systems, time-varying systems. In *IEEE Transactions On Computer-Aided Design of Integrated Circuit and Systems*, volume 22, pages 171–187, 2003.

[116] T. A. Porsching. Estimation of the error in the reduced basis method solution of nonlinear equations. *Mathematics of Computation*, 45(172):487–496, October 1985.

[117] R. Potthast. A fast new method to solve inverse scattering problems. *Inverse Problems*, 12:731–742, 1996.

[118] R. Potthast. A point-source method for inverse acoustic and electromagnetic obstacle scattering problems. *IMA J. Appl. Math.*, 61:119–140, 1998.

[119] R. Potthast. On the convergence of a new newton-type method in inverse scattering. *Inverse Problems*, 17:1145–1157, 2001.

[120] C. Prud'homme and A. T. Patera. Reduced-basis output bounds for approximately parametrized elliptic coercive partial differential equations. *Computing and Visualization in Science*, 2003. Accepted.

[121] C. Prud'homme, D. Rovas, K. Veroy, Y. Maday, A. T. Patera, and G. Turinici. Reliable real-time solution of parametrized partial differential equations: Reduced-basis output bound methods. *Journal of Fluids Engineering*, 124(1):70–80, March 2002.

[122] Christophe Prud'homme, Dimitrios V. Rovas, Karen Veroy, and Anthony T. Patera. A mathematical and computational framework for reliable real-time solution of parametrized partial differential equations. *M2AN Math. Model. Numer. Anal.*, 36(5):747–771, 2002. Programming.

[123] A. Quarteroni, R. Sacco, and F. Saleri. *Numerical Mathematics*, volume 37 of *Texts in Applied Mathematics*. Springer, New York, 1991.

[124] A. Quarteroni and A. Valli. *Numerical Approximation of Partial Differential Equations*. Springer, 2nd edition, 1997.

[125] R. Ramlau. A steepest descent algorithm for the global minimization of the tikhonov functional. *Inverse Problems*, 18:381–403, 2002.

[126] S. S. Ravindaran. A reduced-order approach for optimal control of fluids using proper orthogonal decomposition. *Int. J. Numer. Meth. Fluids*, 34:425–448, 2000.

[127] J. N. Reddy. *Applied Functional Analysis and Variational Methods in Engineering*. McGraw-Hill, 1987.

[128] J.N. Reddy. *Applied Functional Analysis and Variational Methods in Engineering*. McGraw-Hill, 1986.

[129] M. Rewienski and J. White. A trajectory piecewise-linear approach to model order reduction and fast simulation of nonlinear circuits and micromachined devices. In *IEEE Transactions On Computer-Aided Design of Integrated Circuit and Systems*, volume 22, pages 155–170, 2003.

[130] M. Romanowski. Reduced order unsteady aerodynamic and aeroelastic models using karhunen-loeve eigenmode (AIAA Paper 96-194). 1996.

[131] D.V. Rovas. *Reduced-Basis Output Bound Methods for Parametrized Partial Differential Equations*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, October 2002.

[132] O. Scherzer. A modified landweber iteration for solving parameter estimation problems. *Applied Mathematics and Optimization*, 38:45–68, 1998.

[133] O. Scherzer, H. W. Engl, and K. Kunisch. Convergence rates for tikhonov regularization of nonlinear ill-posed problems. *SIAM J. Numer. Anal.*, 30(6):1796–1838, 1993.

[134] L. Sirovich. Turbulence and the dynamics of coherent structures, part 1: Coherent structures. *Quarterly of Applied Mathematics*, 45(3):561–571, October 1987.

[135] Y. Solodukhov. *Reduced-Basis Methods Applied to Locally Non-Affine Problems*. PhD thesis, Massachusetts Institute of Technology, 2004. In progress.

[136] Z. Su, L. Ye, X. Bu, X. Wang, and Y. W. Mai. Quantitative assessment of damage in a structural beam based on wave propagation by impact excitation. *Structural Health Monitoring*, 2:27–40, 2003.

[137] A.N. Tikhonov, A.V. Goncharsky, V.V. Stepanov, and A.G. Yagola. *Numerical Methods for the Solution of Ill-Posed Problems*. Kluwer Academic, Dordrecht, 1995.

[138] Kim-Chuan Toh. Primal-dual path-following algorithms for determinant maximization problems with linear matrix inequalities. *Comput. Optim. Appl.*, 14(3):309–330, 1999.

[139] K. Veroy. *Reduced-Basis Methods Applied to Problems in Elasticity: Analysis and Applications*. PhD thesis, Massachusetts Institute of Technology, 2003. In progress.

[140] K. Veroy and A. T. Patera. Certified real-time solution of the parametrized steady incompressible navier-stokes equations; rigorous reduced-basis *a posteriori* error bounds. *Submitted to International Journal for Numerical Methods in Fluids*, 2004. (Special Issue — Proceedings for 2004 ICFD Conference on Numerical Methods for Fluid Dynamics, Oxford).

[141] K. Veroy, C. Prud'homme, and A. T. Patera. Reduced-basis approximation of the viscous Burgers equation: Rigorous *a posteriori* error bounds. *C. R. Acad. Sci. Paris, Série I*, 337(9):619–624, November 2003.

[142] K. Veroy, C. Prud'homme, D. V. Rovas, and A. T. Patera. *A Posteriori* error bounds for reduced-basis approximation of parametrized noncoercive and nonlinear elliptic partial differential equations (AIAA Paper 2003-3847). In *Proceedings of the 16th AIAA Computational Fluid Dynamics Conference*, June 2003.

[143] K. Veroy, D. Rovas, and A. T. Patera. *A Posteriori* error estimation for reduced-basis approximation of parametrized elliptic coercive partial differential equations: "Convex inverse" bound conditioners. *Control, Optimisation and Calculus of Variations*, 8:1007–1028, June 2002. Special Volume: A tribute to J.-L. Lions.

[144] J. Wang and N. Zabaras. Hierarchical bayesian models for inverse problems in heat conduction. *Inverse Problems*, 21:183–206, 2005.

[145] Y. Y. Wang, K. Y. Lam, and G. R. Liu. Detection of flaws in sandwich plates. *Composite Structures*, 34:409–418, 1996.

[146] Y. Y. Wang, K. Y. Lam, and G. R. Liu. Wave scattering of interior vertical cracks in plates and the detection of the crack. *Engineering Fracture Mechanics*, 59(1):1–16, 1998.

[147] A.T. Watson, J.G. Wade, and R.E. Ewing. Parameter and system identification for fluid flow in underground reservoirs. In H.W. Engl and J. McLaughlin, editors, *Inverse Problems and Optimal Design in Industry*. Teubner, Stuttgart, 1994.

[148] K. Willcox and J. Peraire. Application of model order reduction to compressor aeroelastic models. In *Proceedings of ASME International Gas Turbine and Aeroengine Congress*, pages 2000–GT–0377, Munich, Germany, 2000.

[149] K. Willcox and J. Peraire. Application of reduced-order aerodynamic modeling to the analysis of structural uncertainty in bladed disks. In *Proceedings of ASME International Gas Turbine and Aeroengine Congress*, Amsterdam, The Netherlands, June 2002.

[150] K. Willcox, J. Peraire, and J. White. An arnoldi approach for generation of reduced-order models for turbomachinery. *Computers and Fluids*, 31(3):369–389, 2002.

[151] Y. G. Xu, G. R. Liu, and Z. P. Wu. Damage detection for composite plates using lamb waves and projection genetic algorithm. *AIAA Journal*, 191(9):1860–1866, 2002.

[152] Y. G. Xu, G. R. Liu, Z. P. Wu, and X. M. Huang. Adaptive multilayer perceptron networks for detection of cracks in anisotropic laminated plates. *International Journal of Solids and Structures*, 38:5625–5645, 2001.