Reduced Basis Method for Quantum Models of Crystalline Solids

by

George Shu Heng Pau

S.M. High Performance in Computation in Engineered System (2002)
National University of Singapore, Singapore-MIT Alliance, Singapore

B.Eng.(Hons) Mechanical Engineering (2001)
Petronas University of Technology, Malaysia

Submitted to the Department of Mechanical Engineering
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Mechanical Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2007

Author .....
Department of Mechanical Engineering
May 4, 2007

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

Accepted by .....
Lallit Anand
Professor of Mechanical Engineering
Chairman, Committee on Graduate Studies

Reduced Basis Method for Quantum Models of Crystalline Solids

by

George Shu Heng Pau

Submitted to the Department of Mechanical Engineering
on May 4, 2007, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Mechanical Engineering

Abstract

Electronic structure problems in solids usually involve repetitive determination of quantities of interest, evaluation of which requires the solution of an underlying partial differential equation. We present in this thesis the application of the reduced basis method in accurate and rapid evaluations of outputs associated with some nonlinear eigenvalue problems related to electronic structure calculations. The reduced basis method provides a systematic procedure by which efficient basis sets and computational strategies can be constructed. The essential ingredients are (*i*) rapidly convergent global reduced basis approximation spaces; (*ii*) an offline-online computational procedure to decouple the generation and projection stages of the approximation process; and (*iii*) inexpensive *a posteriori* error estimation procedure for outputs of interest.

We first propose two strategies by which we can construct efficient reduced basis approximations for vectorial eigensolutions — solutions consisting of several eigenvectors. The first strategy exploits the optimality of the Galerkin procedure to find a solution in the span of all eigenvectors at N judiciously chosen samples in the parameter space. The second strategy determines a solution in the span of N vectorial basis functions that are pre-processed to better represent the smoothness of the solution manifold induced by the parametric dependence of the solutions. We deduce from numerical results conditions in which these approximations are rapidly convergent.

For linear eigenvalue problems, we construct *a posteriori* asymptotic error estimators for our reduced basis approximations — extensions on existing work in algebraic eigenvalue problems. We further construct efficient error estimation procedures that allow efficient construction of reduced basis spaces based on the "greedy" sampling procedure. We extend our methods to nonlinear eigenvalue problems, utilizing the empirical interpolation method. We also provide a more efficient construction procedure for the empirical interpolation method.

Finally, we apply our methods to two problems in electronic structure calculations — band structure calculations and electronic ground state calculations. Band structure calculations involve approximations of linear eigenvalue problems; we demonstrate the applicability of our methods in the many query limit with several examples related to determination of spectral properties of crystalline solids. Electronic ground state energy calculations based on Density Functional Theory involve approximations of nonlinear eigenvalue problems; we demonstrate the potential of our methods within the context of geometry optimization.

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

Acknowledgments

I would first like to thank my thesis advisor, Professor Anthony T. Patera, for his support and guidance throughout my thesis work. I am most grateful for his wisdom, patience and trust.

I would also like to thank my thesis committee members, Professor Claude Le Bris, Professor Nicola Marzari, and Professor Nicolas Hadjiconstantinou for their interest, comments, and encouragement throughout my studies. In addition, I am also truly grateful to Professor Yvon Maday and Professor Eric Cancès for their help, support, and suggestions.

I am very thankful to my current and former colleagues: Simone Deparis, Sugata Sen, Phuong Huynh, Martin Grepl, Karen Veroy-Grepl, Sebastian Boyaval, Gianluigi Rozza, Luca Dede, and Ngoc Cuong Nguyen. I greatly appreciate the time spent together and the many discussions we had. I am also truly grateful to Debra Blanchard for her invaluable help, counsel, and unabating optimism.

Last but not least, I want to express my deepest gratitude to my family — my parents, my sister, my brother, my brother-in-law, and my aunt — and my strongest believer, critic and love, Ming Lee. Without their support, love and encouragement, I would not have been able to realize my dreams. To them I dedicate this thesis.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Quantum simulation provides a mean by which material behavior — ranging from microscopic properties describing the atom-atom interaction to macroscopic properties characterizing a bulk material — can be determined from *ab-initio* theoretical models. Traditionally an interpretative science, it has developed into an accurate, predictive tool widely used by engineers and scientists in the design of new materials [30, 88]. This advancement is fueled by recent theoretical advances, the vast expansion of computing resources, and the widespread availability of quantum simulation codes. Increasingly more complex systems can now be solved.

In these simulations, we usually need to repetitively determine certain outputs of interest, s, given an input parameter $\mu \in \mathcal{D}$ where $\mathcal{D} \in \mathbb{R}^P$ is the parameter space in which our input μ varies. However, these outputs are usually functionals of a field variable $u(\mu)$ obtained by solving the underlying μ-parameterized partial differential equation (PDE) derived from physical models. A model behavior is then encapsulated in an input-output relation, the evaluation of which requires solution of the underlying PDE. In electronic structure calculations for example, we may be interested in determining the ground state energy (the output $s(\mu)$) of a crystal structure given a lattice constant (the input μ). To determine this energy, we must first determine the electron wavefunctions (the field variable $u(\mu)$) based on quantum models derived from (say) Density Functional Theory.

In practice, the solution $u(\mu)$ is usually obtained through a discretization procedure such as a

finite dimensional Galerkin approximation since a closed-form analytical form is usually not available. Some computational chemistry problems afford very efficient approximations. For molecular systems in gas or liquid phase, the use of optimized atom-centered basis sets leads to small discrete algebraic problems which can then be solved efficiently. In solid state simulation, the planewave method, coupled with the pseudopotential method, is also particularly efficient. However, when we are dealing with more general problems or when we desire higher accuracy, the use of numerical methods based on more generic approximation spaces, such as the planewave method, the finite difference method and the finite element method, may be necessary. This leads to a large discrete algebraic system of which the solution can be computationally intensive, and in some cases, impractical within the many query context we are interested in.

Our goal is to develop reduced basis methods that permit the efficient evaluation of this PDE-induced input-output relation encountered in computational chemistry, especially in the many-query limit. This can be very useful in many applications. For example, molecular dynamics simulations can attain its full predictive potential through the use of a low-order *ab initio* model — the use of an empirical model or an interpolation of empirical data requires extensive prior knowledge of the problem being solved [77]. In multiscale and multiphysics simulations, or the computational material design, the rapid evaluation of the input-output relation allows efficient coupling with other components in the simulation [80, 112]. In design of nanostructures, a low-order model will allow the rapid evaluation of the dielectric response which is a function of both frequency and physical space [50, 117]. We shall describe two specific examples to further motivate our methods and the context in which they are applied.

1.1.1 Band Structure Calculations

The first application involves the study of the Schrödinger operator $\mathcal{H} = -\Delta + V_{\text{eff}}$ where V_{eff} is a periodic function such that $V_{\text{eff}}(\mathbf{x}) = V_{\text{eff}}(\mathbf{x}+\mathbf{R})$ for all Bravais lattice vector $\{\mathbf{R}_i \in \mathbb{R}^3, 1 \leq i \leq 3\}$. From the Bloch theorem (see, for example [3, 61] or Section 5.2.1 of this thesis), the eigenstates of the $\mathcal{H}$, given by $\{\psi_i,\ 1 \leq i \leq n_b\}$, have the periodicity of the Bravais lattice:

$$\psi_i(\mathbf{x};\mathbf{k}) = e^{\mathbf{ikx}} u_i(\mathbf{x};\mathbf{k}), \tag{1.1}$$

where n_b is the number of states we are interested in, $u_i(\mathbf{x}+\mathbf{R};\mathbf{k}) = u_i(\mathbf{x};\mathbf{k})$, and wave vector $\mathbf{k}$ defines the translational symmetry of the periodic potential and lies in the first Brillouin zone of the periodic structure. This leads to an equivalent parameterized eigenvalue problem: we find $u_i(\,\cdot\,;\mathbf{k})$, $1 \le i \le n_b$ and $E_i(\mathbf{k})$, $1 \le i \le n_b$ such that

$$\begin{aligned}\left(\frac{1}{2}\Delta - \mathrm{i}\mathbf{k} + \frac{1}{2}|\mathbf{k}|^2 + V_{\text{eff}}\right) u_i(\mathbf{x};\mathbf{k}) &= E_i(\mathbf{k})u_i(\mathbf{x};\mathbf{k}), \quad 1 \le i \le n_b, \\ \int_\Omega u_i(\mathbf{x};\mathbf{k})u_j(\mathbf{x};\mathbf{k}) &= \delta_{ij}, \quad 1 \le i \le j \le n_b, \end{aligned} \tag{1.2}$$

where $\Omega \in \mathbb{R}^3$ is the physical domain of our unit cell. We are thus interested in how energy levels $E_i(\mathbf{k})$, $1 \le i \le n_b$ varies with $\mathbf{k}$.

This is considered one of the most important calculations in solid state physics. The number of $\mathbf{k}$-points at which (1.2) must be evaluated, denoted by n_k, depends on the quantity and the physical system we are interested in. In some cases, n_k can be very small. For example, in determining electron densities in semiconductors, only several $\mathbf{k}$-points are sufficient — to obtain an accuracy of 1% (relative to some "truth" calculations based on larger n_k), [26] uses only 3 $\mathbf{k}$-points for HgTe and CdTe crystals; and [28] uses only 2 $\mathbf{k}$-points for C, Si and Ge crystals. In this thesis, we deal with cases where we must solve (1.2) at a large number of $\mathbf{k}$-points. For example, to accurately determine the density of states and related optical quantities of crystals such as the Fermi level and the dielectric function, we must evaluate the band energies, $E_i(\mathbf{k})$ at large number of $\mathbf{k}$-points — n_k of $O(10^3)$ are routinely used. In some extreme cases, such as the determination of the anomalous Hall conductivity [116], solutions at millions of $\mathbf{k}$-points are required in order to achieve a reasonable accuracy. Another application which requires solutions at many $\mathbf{k}$-points is the charge transport problem. Here, we must efficiently calculate scattering rates of thousands of charge carriers, each in a different $\mathbf{k}$-state. This must be repeated many times for the duration of the monte carlo simulation. Based on these examples, it is clear that an accurate and efficient numerical method for band structure calculations is clearly desirable.

1.1.2 Electronic Ground State Energy Calculations

For our second application, we are interested in determining the electronic ground state energy of a molecular system. This is central to all quantum chemistry calculations, including excited

states predictions, linear response theory calculations, conformation determinations and *ab initio* molecular dynamics simulations. For example, given a set of parameters μ of a crystalline solid, such as the locations of nuclei and geometric properties of the unit cell, we would like to determine the ground state energy of the crystalline solid, $\mathcal{E}(\mu)$.

To determine $\mathcal{E}(\mu)$ exactly, we need to solve the full Schrödinger equation, intractable for almost all real size cases. Approximation models are thus usually used in computational chemistry, of which the two main categories are Hartree-Fock (HF) type models [18, 111] and Density Functional Theory (DFT) type models [18, 32, 90]. They both involve solving a non-quadratic constrained minimization problem. In practice, we usually solve the associated Euler Lagrange equation, which is a *nonlinear eigenvalue problem.* An iterative procedure must then be used to obtain a solution — Self Consistent Field (SCF) schemes, outlined in Section 1.2, are most widely used today.

However, even with the above approximation, the computational cost of determining $\mathcal{E}(\mu)$ is still significant. In geometry optimization or *ab initio* molecular dynamics calculations, the nonlinear eigenvalue problem must be solved many times. In addition, within each iteration in a SCF scheme, we must solve a linear eigenvalue problem. Clearly, an efficient numerical method for nonlinear eigenvalue problem can greatly facilitate any quantum calculation.

We note that the methodologies, which we will describe in Section 1.3.1, are built upon solutions at several judiciously selected μ-points. Thus, they are general, and thus not restricted to either HF models or DFT models. They can very well be applicable to the full Schrödinger equation, provided we are able to obtain solutions to the full Schrödinger equation first.

1.2 Some Existing Methods

Here, we give a brief description of methods commonly used to solve eigenvalue problems encountered in computational chemistry problems. All methods involve first choosing an appropriate *finite dimensional* approximation space, followed by a solution method that solves the resulting algebraic eigenvalue problem. There exists many software packages that solve computational chemistry problems, each with a different solution strategy. Two examples of codes relevant to electronic structure calculations of extended systems are ABINIT [43] and PWSCF [6]. In subsequent chapters, we will occasionally invoke ABINIT when we discuss how the methodological details used in ABINIT affect methods we propose in this thesis.

1.2.1 Approximation Spaces

Since a close-form analytical solution is usually not available, an approximate solution to the eigenvalue problem must then be determined. This is usually achieved through an introduction of a finite dimensional approximation space $Y^{\mathcal{N}} \equiv \text{span}\ \{\phi_i, 1 \leq i \leq \mathcal{N}\}$, where $\mathcal{N}$ is the dimension of $Y^{\mathcal{N}}$. We then approximate the exact solution by a linear combination of the basis functions ϕ_i:

$$\sum_{i=1}^{\mathcal{N}} c_i \phi_i. \tag{1.3}$$

The set $\{\phi_i, 1 \leq i \leq \mathcal{N}\}$ is known as the basis set. We can define an appropriate basis set in a variety of ways, depending on problems at hand. Within the computational chemistry community however, the quest for a small basis set is extreme, driven by the need to minimize computational cost in the early days of computational chemistry when computing resources are scarce. Reproducibility of experimental results using the smallest number of basis functions trumps the need for proper numerical analysis. Within the class of local basis sets, this usually leads to a plethora of efficient, problem-dependent basis sets that frequently do not admit rigorous convergence analysis [31]. However, as computing resources become more widely available, more general basis sets with better convergence properties are increasingly used. We now describe the two main categories of basis sets — local basis sets and general basis sets — in greater details.

Local basis sets

Local basis sets are basis functions that are well localized to nucleus sites of a molecular system. They afford an efficient representation of solutions to an isolated molecular system since the wavefunctions of an isolated system are usually localized to regions close to nuclei and vanishingly small in regions far from nuclei. An early example is the Slater-type orbital (STO), given in cartesian coordinates by [35]

$$x^l y^m z^n \exp(-\zeta_{l,m,n} r), \tag{1.4}$$

where $l \in \mathbb{N}$, $m \in \mathbb{N}$, and $m \in \mathbb{N}$ are quantum numbers; $\zeta_{l,m,n}$ is a "tunable" parameter that depends on l, m and n; and $r = \sqrt{x^2 + y^2 + z^2}$. The STOs are solutions to hydrogenoic-like problem. However, due to the dependence of the exponential term on r, the evaluation of integrals involving STOs is tedious.

This leads to the introduction of the Gaussian-type orbital (GTO) of the form [35]

$$x^l y^m z^n \exp(-\zeta_{l,m,n} r^2). \tag{1.5}$$

Here, the r^2 dependence in the exponential term allows integrals involving GTO to be evaluated much more efficiently. In particular, for Hartree Fock models, computational complexity of the bielectronic integral in the exchange term can be reduced[1] from $O(\mathcal{N}^4)$ to $O(\mathcal{N}^{2.7})$ [18].

However, as these basis functions are mathematical construct, the number of GTOs required to achieve a certain accuracy is normally larger than an approximation based on STOs. There are several ways we can improve the efficiency of an approximation based on GTOs. First, we can find an optimal set of $\zeta_{l,m,n}$ that minimizes the number of GTOs required to satisfy some criteria, e.g., the lowest energy [95]. Second, we can contract the GTOs into a smaller set of orbitals that better represent wavefunction solutions from atomic calculations — the combined set of GTOs is now the new basis set, thus reducing the dimension of the resulting algebraic system of equations. Third, since we usually perform the above optimization procedures in an atomic setting, GTOs must be further facilitated with polarization functions to describe the distortion of the atomic orbitals in a molecular environment. In practice, common users do not usually perform these optimization steps; instead they choose one of many pre-optimized basis sets available in the literature and use them in their simulations [121].

It is clear that the efficiency of GTO comes at a tremendous price. First, the optimization procedure is usually nonlinear — if the number of optimization parameters considered is large, finding an optimal solution becomes increasingly more difficult. Second, when existing optimized basis sets are used, their flexibility can be limiting, since they do not provide uniform description of atoms and molecules and their properties — a particular optimized basis set was optimized in a specific environment with respect to a specific property. Third, as optimized basis sets are usually obtained in an ad-hoc manner with reliance on physical intuition, performing convergence analysis on these basis sets is not straightforward. As such, there is no simple way to evaluate *a priori* the performance of a particular basis set for a particular problem; selection of the right basis set frequently relies on prior knowledge and making an educated guess [118, 119]. Due to these reasons,

[1] An $O(\mathcal{N})$ can also be achieved with linear scaling algorithms based on Greengard and Rokhlin Fast Multipole Method [18].

the basis set used is also frequently considered as an integral part of the model [31].

While the use of more generic GTOs alleviates some of the issues identified in previous paragraph, it is accompanied by an increase in the computational cost. In addition, local basis sets are in general not suited for molecular dynamics simulations where nuclei move. This is because with local basis sets, basis functions must be recomputed with each change in locations of nuclei. In addition, the computation of ionic forces will require determining derivatives of basis functions, and integrals involving these derivatives — a process that can be very computationally expensive [48].

General basis sets

Optimized local basis sets can indeed be very efficient when applied to isolated molecular systems of which the nuclei configuration is fixed. However, this is only achieved if the correct basis set is used, perhaps chosen based on our experience with similar systems. Otherwise, for less defined problems, and for systems where the nuclei move, general basis sets may be more appropriate.

The most common basis set under this category is the Fourier basis set. They are particularly efficient for domains with periodic boundary conditions. For isolated systems, finite difference methods [33] and finite element methods [91] are increasingly used. With finite difference methods, the physical domain is divided into a grid, and solutions are found on this grid. On the other hand, finite element basis functions consist of piecewise polynomials over the physical domain. Unlike finite difference method, finite element method is variational.

General basis sets usually allow systematic convergence analysis, and provide an uniform description of the atoms and molecules and their properties. In addition, as these general basis sets are independent of locations of nuclei (assuming we are not employing adaptive meshing during implementation of finite difference methods and finite element methods), evaluation of ionic forces acting on the nuclei is straightforward with Hellmann-Feynman theorem [37]. Finite difference methods and finite element methods are also capable of handling complex geometries, making them well-suited to study of nanostructures [34].

However, as these generic basis functions do not usually bear any resemblance to the actual solutions, large number of basis functions is required to achieve a comparable level of accuracy when compared to local basis sets. In particular, to approximate wavefunctions near nucleus regions, very fine resolution is required. The resulting discrete problem is usually too large for use in the many-

query context. In addition, with HF models, the evaluation of the bielectronic integral with finite elements method planewave method will scales as $O(\mathcal{N}^4)$, where $\mathcal{N}$ is the number of basis functions. Since DFT models do not involve a bielectronic integral, general basis sets can be efficiently used to solve DFT models — the computational cost will not suffer from a $O(\mathcal{N}^4)$ dependence.

1.2.2 Solution Methods

The introduction of a finite basis set, as described in Section 1.2.1, usually leads to an algebraic eigenvalue problem. When this eigenvalue problem is linear, it can be solved efficiently based on several existing algorithms — since resulting matrices will typically be Hermitian and sparse, the Lanczos method, the implicitly restarted Lanczos method and the Jacobi-Davidson method are most common used [4]. In particular, the implicitly restarted Lanczos method is implemented in the widely used software package ARPACK [69].

However, with *ab initio* quantum models such as those based on Hartree Fock Theory and Density Functional Theory, we usually obtain nonlinear algebraic eigenvalue problems. The prevalent technique to handle this type of problems is the Self-Consistent Field (SCF) scheme. Several strategies exist in the implementation of the SCF scheme [18] and can be divided into three groups: (*i*) fixed point algorithms [100, 106, 107]; (*ii*) direct minimization methods [93]; and (*iii*) relaxed constrained algorithms [17, 19]. In this thesis, we will primarily use the Roothaan algorithm [106], one of the earliest fixed point algorithm used in the SCF scheme. We will also be using Newton iterative scheme to solve nonlinear eigenvalue problems encountered in this thesis. Further details are given in Section 4.2.2.

1.3 Computational Challenges/Thesis Objectives

The previous sections provide a background on which we shall state the challenges we hope to address in this thesis. The problems we would like to solve have the following characteristics. First, the governing equations encountered in problems we consider are eigenvalue equations (linear and nonlinear) whose solutions depend on parameters characterizing these problems. Second, we are interested in finding approximation to not just a single eigenstate, but multiple eigenstates. The solutions are thus vectorial in nature. Third, a typical calculation will require repetitive evaluation of an input-output relation $s(\mu)$, for many different sets of μ. The numerical methods used for such

evaluations must then be efficient and accurate in the many-query limit.

Existing numerical methods in computational chemistry achieve high efficiency through problem-specific basis sets. However, they have poor transferability from one problem to another. Construction of these efficient basis sets is also time-consuming, ad-hoc, and dependent on existing physical insights on the problem. The alternative, use of general basis sets, is expensive and prohibitive in the many-query context.

The primary objective of this thesis is to develop computational methods that permit accurate, and rapid evaluation of input-output relationships induced by eigenvalue equations *in the limit of many queries*. We seek to develop *general* techniques that systematically create efficient, *problem-specific* basis sets that have controllable convergence property. In particular, we desire techniques that provide (*i*) accurate approximation of relevant outputs of interest; (*ii*) inexpensive error estimator for the approximation; and (*iii*) a computational framework which allows rapid online calculation of the output approximation and associated error estimator.

Our second objective is the application of the computational methods developed to problems where significant gain in efficiency over existing methods can be achieved. In particular, we seek to use these techniques to solve representative problems in solid state calculations which involve repetitive evaluations of input-output relations and underlying eigenvalue problems. To achieve these goals, we pursue the reduced basis method.

1.3.1 Earlier Work on Reduced Basis Method

The reduced basis method recognizes that the field solution $u(\mu)$ resides on a very low-dimensional manifold induced by the parameteric dependence. Furthermore, the field variable $u(\mu)$ will often be quite regular in μ — the parametrically induced manifold is smooth — even when the field variable enjoys only limited regularity with respect to the spatial coordinate. The smoothness property can be deduced from the equation for the sensitivity derivatives; the stability and continuity properties of the partial differential operator are crucial. Clearly the latter are delicate matters in the context of eigenvalue problems. We will elaborate further issues related to sensitivity analysis in Section 2.3.2 and 2.3.7. The reduced basis method thus exploits dimension reduction afforded by the *low-dimensional* and *smooth* parametrically induced solution manifold. More precisely, rather than general basis sets consisting of, say, Fourier basis functions or finite element basis functions, the

basis set consists of solutions of the partial differential equation at N selected parameter points $\mu_i, 1 \leq i \leq N$. Then, the set of all solutions $u(\mu)$ as μ varies can be approximated very well by its projection on a finite and low dimensional vector space spanned by the $u(\mu_i)$: for sufficiently well chosen μ_i, there exist coefficients $c_i(\mu), 1 \leq i \leq N$ such that the finite sum $\sum_{i=1}^{N} c_i u(\mu_i)$ is very close to $u(\mu)$ for any μ.

Reduced basis methods were first introduced in the late 1970s to fulfill the need for more efficient parameter continuation methods in the context of nonlinear structural analysis [2, 85]. They are subsequently abstracted, analyzed, and extended to a much larger class of parameterized partial differential equations [9, 36, 67, 84, 86, 98, 104, 105], and a variety of reduced basis approximation spaces [97, 53]. Reduced basis methods are also used in solving Navier-Stokes equations and fluid dynamics problems [47, 52, 53, 54, 55, 94].

In these early methods, the approximation spaces are local in parameter space. This was due to context in which reduced basis methods were developed, and the absence of *a posteriori* error estimators and effective sampling procedures. In the more recent past the reduced basis approach and in particular associated *a posteriori* error estimation procedures have been successfully developed for (*i*) linear elliptic and parabolic PDEs that are affine in the parameter [45, 71, 75, 99] ; (*ii*) elliptic PDEs that are at most quadratically nonlinear in the first argument [82, 114, 115]; and (*iii*) general nonaffine PDEs [8, 44]. In these cases a very efficient offline-online computational strategy can be developed. The operation count for the online stage — in which, given a new parameter value, we calculate the reduced basis output and associated error bound — is *independent* of $\mathcal{N}$, the dimension of the underlying "truth" approximation.

Application of reduced-basis method to linear eigenvalue equations have also been examined previously. In [71], reduced basis approximation and rigorous *a posteriori* error bounds are developed for the approximation of the first eigenvalue. This thesis extends the theory to include evaluations of multiple eigensolutions and nonlinear eigenvalue problems. In particular, the emphasis is on applications of the method to computational chemistry problems where the underlying PDEs do not have the same nice structure that allows us to apply the methodology developed in earlier works: (*i*) the equations can contain both non-affine terms and also very nasty nonlinear terms, for example associated with an exchange-correlation term; (*ii*) the solution sought is not scalar — for each μ, we look for a set of eigensolutions; and (*iii*) the parameterizations of the

PDEs can be complex, for example due to a set of moving nuclei and periodic boundary conditions. In fact, the above three issues are the main difficulties faced by any numerical approximation of the PDEs obtained in computational chemistry and are the deciding factors when determining the appropriate numerical approach to employ.

1.4 Scope

1.4.1 Thesis contribution

In this thesis, we have developed reduced basis techniques for eigenvalue problems, in particular those encountered in computational chemistry. Contributions are made in several areas.

Approximation spaces

Constructing an efficient reduced basis space for a vectorial solution consisting of multiple eigenvectors is nontrivial. The bases in the eigensubspaces that we are approximating can have complicated, nonsmooth variations with the parameter of interest due to behavior of the problem examined and the eigensolver used to solve the problem. This may mask the actual smoothness in the variation of eigensubspaces with respect to the parameter.

We present two approaches by which we can construct a suitable reduced basis approximation space for vectorial eigensolutions. The first approach, denoted as the augmented reduced basis space, circumvents the difficulties outlined in the previous paragraph by taking the span of all eigenvectors at all sample points. Galerkin procedure then finds the best approximation in this enlarged reduced basis space, thus the term "augmented". Numerical results show that this approximation space is rapidly convergent for all problems examined.

In the second approach, denoted as the vectorial reduced basis space, we pre-process the basis such that resulting bases are more representative of the actual parametric smoothness of the eigensubspaces. Complexity of the pre-processing procedure required is dependent on the problem examined. Except for the full band structure calculation problem where variation of eigenvectors with the parameter considered is exceedingly rich, current implementations of the pre-processing procedures are adequate and lead to rapidly convergent vectorial reduced basis spaces. Furthermore, we can exploit the inherent orthogonality property of the vectorial reduced basis space to

omit orthogonality constraints from our reduced basis formulation, and yet obtain a solution that approximately satisfies these omitted orthogonality constraints. Nevertheless, the omission of the orthogonality constraints implies that the problem we solve in the reduced basis approximation differ from the original problem — this implies that any solution we obtain only approximately satisfies the original problem. This is further elaborated in Section 2.3.5.

A posteriori error estimation

For linear eigenvalue problem, we facilitate our reduced basis approximation with *a posteriori* asymptotic error bounds. In addition, the error estimation procedure admits an efficient offline-online computational decomposition, leading to rapid estimations of errors in both the augmented reduced basis approximation and the vectorial reduced basis approximation. This procedure is further incorporated into an adaptive sampling procedure [71, 82, 99] to allow the efficient construction of rapidly convergent reduced basis spaces.

Nonlinear eigenvalue problem

We extend the reduced basis method to eigenvalue problem with highly nonlinear terms. We employ the empirical interpolation method introduced in [8, 44] to approximate these nonlinear terms. We also introduce an alternative construction procedure for the empirical interpolation method. We then construct an offline-online computational procedure which remains efficient even in the presence of these nonlinear terms.

Computational chemistry

We examine applications of reduced basis methods in computational chemistry problems, first reported in [21]. In particular, we demonstrate the applicability of reduced basis methods in band structure calculations based on empirical pseudopotential models, and ground state energy calculations of crystalline solids based on Density Functional Theory. The parameterization of the problems solved are admittedly simple — it is either apparent by inspection or obtained through a simple linear geometric mapping. More complex parameterization, in particularly the parameterization of moving nuclei, is not considered in this thesis.

1.4.2 Thesis outline

In Chapter 2, we summarize the reduced basis method formulation and associated *a posteriori* error estimation for linear eigenvalue problem. In particular, we introduce specific ideas on construction of efficient approximation spaces for vectorial solutions and describe the resulting approximation, the computational procedure, and the efficient evaluation of *a posteriori* error estimators. We use an quantum harmonic oscillator problem as a numerical example.

In preparation for the treatment of nonlinear eigenvalue problems, we describe briefly the empirical interpolation procedure for parametric field in Chapter 3. This is followed by the approximation of a generic nonlinear eigenvalue problem to demonstrate how nonlinear and nonaffine terms typically encountered in quantum models can be handled within the reduced basis framework.

We then apply the methods developed in previous chapters to problems in solid state physics. In Chapter 5, the reduced basis approximation of a Hamiltonian equation with a periodic background potential allows us to efficiently determine spectral properties of crystalline solids. In Chapter 6 and 7, we look at the reduced basis approximation of Kohn Sham equations, the workhorse of Density Functional Theory. We first study the reduced basis approximation of one dimensional Kohn Sham equations in Chapter 6 to fully identify all the ingredients required for the full treatment of three dimensional Kohn Sham equations, described in Chapter 7.

Finally, in Chapter 8, we summarize our work and conclude with some suggestions for future work.

1.5 Units

We will work in atomic units (a.u.) — dimensionless units — throughout this thesis. Frequently however, we may encounter other types of units in the literature. We summarize the conversion between different units in Table 1.1 — the designations and the symbols are the commonly used names and symbols for the atomic units. In addition, all equations in this thesis are written in the dimensionless form. For example, a Schrödinger operator written as

$$-\frac{\hbar^2}{2m_e}\Delta + V_{\text{eff}} \tag{1.6}$$

Quantity	Atomic unit	Designation	Symbol	Other units	
length	1	Bohr radius	a_0	0.529177249	angstrom
				5.291772108E−11	m
mass	1	electron mass	m_e	9.1093897E−31	kg
energy	1	Hartree	E_h	27.2113961	eV
				4.35974417E−18	J
charge	1	electron charge	e	1.60217653E−19	C
angular momentum	1	Dirac constant	$\hbar$	1.05457168E−34	J

Table 1.1: Conversion between atomic units of several quantities of interest and other commonly used units in literature.

in the S.I. units is given as

$$-\frac{1}{2}\Delta + V_{\text{eff}} \tag{1.7}$$

in the atomic units.

Chapter 2

Linear Eigenvalue Problem

2.1 Introduction

The goal of this chapter is twofold. First, we give the rational behind the reduced basis method and describe the ingredients required to adequately treat an linear eigenvalue problem by the reduced basis method — the construction of the approximation space based on solutions of an linear eigenvalue problem at judiciously chosen parameter points; the exploitation of the affine parameter dependence property of functionals in the efficient offline-online computational strategy; and the construction of an optimal parameter sample set through the "greedy" sampling procedure based on *a posteriori* error estimators. Second, we extend the methodology developed in [71] to problems involving many eigenvalues and eigenvectors. We introduce and compare two approaches to constructing efficient reduced basis approximation spaces for vectorial solutions — solutions consisting of several eigenvectors.

To better explain some of the concepts in the reduced basis method, we introduce a simple numerical example — we will look at the reduced basis approximation of a harmonic oscillator equation. In spite of its simplicity, the harmonic oscillator equation governs a large number of systems in physics. In particular, harmonic oscillator equation is commonly encountered in the study of physical systems in their neighborhood of stable equilibrium positions – such as, vibrations of atoms of a molecule or a crystalline lattice.

2.2 Abstract Formulation

2.2.1 Problem Statement

Given two Hilbert spaces X and Y where $X \equiv L^2(\Omega)$ and $H_0^1(\Omega) \subset Y \subset H^1(\Omega)$, we consider the following linear eigenvalue problem: given any $\mu \in \mathcal{D}$, we are interested in finding $\hat{\boldsymbol{\lambda}}(\mu)$ where $(\hat{\mathbf{u}}(\mu) \equiv (u_1(\mu), \ldots, u_{n_b}(\mu)), \hat{\boldsymbol{\lambda}}(\mu) \equiv (\lambda_1(\mu), \ldots, \lambda_{n_b}(\mu))) \in Y^{n_b} \times \mathbb{R}^{n_b}$ satisfies

$$\begin{aligned} a(u_i(\mu), v; \mu) &= \lambda_i(\mu) m(u_i(\mu), v), \quad \forall v \in Y, \quad 1 \le i \le n_b, \\ m(u_i(\mu), u_j(\mu)) &= \delta_{ij}, \quad 1 \le i \le j \le n_b. \end{aligned} \tag{2.1}$$

Here, n_b is the number of eigensolutions we are interested in; $\lambda_1(\mu) \le \lambda_2(\mu) \le \ldots \le \lambda_{n_b}(\mu)$ where $\lambda_1(\mu)$ is the smallest eigenvalue; $\mathcal{D} \subset \mathbb{R}_+$ is our parameter domain; and Ω a bounded domain in $\mathbb{R}$. In addition, we require $a(w, v; \mu)$ and $m(w, v)$ to be continuous

$$a(w, v; \mu) \le \gamma_a \|w\|_Y \|v\|_Y, \quad \forall w, v \subset Y, \tag{2.2}$$

$$m(w, v) \le \gamma_m \|w\|_X \|v\|_X, \quad \forall w, v \subset X; \tag{2.3}$$

coercive

$$0 \le \alpha_a \equiv \inf_{w \subset Y} \frac{a(w, w; \mu)}{\|w\|_Y^2}, \tag{2.4}$$

$$0 \le \alpha_m \equiv \inf_{w \subset X} \frac{m(w, w)}{\|w\|_X^2}; \tag{2.5}$$

and symmetric, $a(w, v; \mu) = a(v, w; \mu)$, $\forall v, w \subset Y$, and $m(w, v) = m(v, w)$, $\forall v, w \subset X$. Since $a(\,\cdot\,, \cdot\,; \mu)$ and $m(\,\cdot\,, \cdot\,)$ are symmetric, we expect the eigenvalues to be real. Furthermore, since $a(\,\cdot\,, \cdot\,; \mu)$ and $m(\,\cdot\,, \cdot\,)$ are real, the eigenvectors are real as well.

Since we work with self-adjoint operators in this chapter (as well as in subsequent chapters), we do not distinguish between the algebraic multiplicity and the geometry multiplicity. We thus define multiplicity of an eigenvalue $\lambda_i(\mu)$ by $\dim(\{v |\ a(v, v; \mu) = \lambda_i(\mu) m(v, v)\})$. Then, the eigensubspace associated with $\lambda_i(\mu)$ is given by $\text{span}\{v |\ a(v, v; \mu) = \lambda_i(\mu) m(v, v)\}$. If multiplicity of $\lambda_i(\mu)$ is 1, then the corresponding eigenvector $u_i(\mu)$ is unique up to multiplication by a scalar.

2.2.2 Affine Parameter Dependence

We will further assume the functional $a(\cdot, \cdot; \mu)$ exhibits the affine parameter dependence property, i.e. $a(w, v; \mu)$ can be expressed as

$$a(w, v; \mu) = \sum_{q=1}^{Q} \Theta_q(\mu)\, a_q(w, v), \quad \forall w \in Y, \quad \forall v \in Y, \tag{2.6}$$

for some finite Q where $\Theta_q : \mathcal{D} \rightarrow \mathbb{R}, 1 \leq q \leq Q$, are smooth parameter-dependent functions, and the $a_q(w, v) : Y \times Y \rightarrow \mathbb{R}, 1 \leq q \leq Q$, are parameter-independent continuous bilinear forms. We note that $\Theta_q(\mu), 1 \leq q \leq Q$ will usually simple algebraic expressions that can be readily evaluated in $O(1)$ operations. We will exploit this property in formulating an efficient computational strategy in Section 2.3.4.

For the current problem, the functional $m(\cdot, \cdot)$ does not depend on μ. Certainly, we can generalize the problem such that m is now a function of μ as well; then we will also require m to exhibit the affine parameter dependence property.

2.2.3 "Truth" Approximation

A closed form solution to a partial differential equation is, in general, not available. More often than not, discretization methods, such as the finite element method and the planewave method, are employed to obtain a numerical approximation to the exact solution. For this purpose, we introduce a conforming approximation space $Y^{\mathcal{N}} \subset Y$ of dimension $\dim(Y^{\mathcal{N}}) = \mathcal{N}$ and associate with this space a complete set of basis functions $\phi_k^{\mathcal{N}} \in Y^{\mathcal{N}}, 1 \leq k \leq \mathcal{N}$. The inner product and the norm associated with $Y^{\mathcal{N}}$ are inherited from Y:

$$(w, v)_{Y^{\mathcal{N}}} \equiv (w, v)_Y, \quad \forall w, v \in Y^{\mathcal{N}}, \tag{2.7}$$

$$\|w\|_{Y^{\mathcal{N}}} \equiv \|w\|_Y, \quad \forall w \in Y^{\mathcal{N}}. \tag{2.8}$$

Certainly our stability and continuity conditions hold since $Y^{\mathcal{N}} \subset Y$:

$$a(w, v; \mu) \leq \gamma_a^{\mathcal{N}} \|w\|_{Y^{\mathcal{N}}} \|v\|_{Y^{\mathcal{N}}}, \quad \forall w, v \subset Y^{\mathcal{N}}, \tag{2.9}$$

$$m(w, v) \leq \gamma_m^{\mathcal{N}} \|w\|_{Y^{\mathcal{N}}} \|v\|_{Y^{\mathcal{N}}}, \quad \forall w, v \subset Y^{\mathcal{N}}; \tag{2.10}$$

and

$$0 \leq \alpha_a^{\mathcal{N}} \equiv \inf_{w \subset Y^{\mathcal{N}}} \frac{a(w,w;\mu)}{\|w\|_{Y^{\mathcal{N}}}^2}, \tag{2.11}$$

$$0 \leq \alpha_m^{\mathcal{N}} \equiv \inf_{w \subset Y^{\mathcal{N}}} \frac{m(w,w)}{\|w\|_{Y^{\mathcal{N}}}^2}. \tag{2.12}$$

We shall also require that $Y^{\mathcal{N}}$ satisfies the approximation condition

$$\max_{\mu \subset \mathcal{D}} \inf_{w \subset Y^{\mathcal{N}}} \|u(\mu) - w\|_Y \to 0 \quad \text{as} \quad \mathcal{N} \to \infty. \tag{2.13}$$

The point of departure for methods presented in this thesis is the "truth" approximation — choosing $\mathcal{N}$ large enough that the numerical approximation is sufficiently accurate that the resulting approximate solution is "indistinguishable" from the exact solution. We build our reduced basis approximation on, and measure the error in the reduced basis approximation relative to this "truth" approximation. Note that since reduced basis approximation is built upon this "truth" approximation, it cannot perform better than this "truth" approximation. Thus, large $\mathcal{N}$ should be used to obtain an accurate reduced basis approximation. Thankfully however, we see that once the reduced basis approximation has been built, the computational costs will be independent of $\mathcal{N}$.

We shall now elaborate briefly on the use of finite element approximation as the "truth" approximation. Other methods are, of course, possible: in Chapter 5, we will look at the planewave method.

Finite element approximation

We define our finite element space $Y^{\mathcal{N}} \equiv Y_h \subset Y$ of dimension $\mathcal{N}$ as

$$Y_h \equiv \{v \in Y \mid v|_{\mathbf{T}_h} \in \mathbb{P}_1(\mathbf{T}_h),\ \forall \mathbf{T}_h \in \mathcal{T}_h\}, \tag{2.14}$$

$$\mathbb{P}_1(\mathbf{T}_h) \equiv \text{span}\{1, x\}, \tag{2.15}$$

where $\mathcal{T}_h$ is a nonuniform "triangulation" of the domain Ω comprised of linear elements T_h, with more elements near the origin. The inner product and norm associated with Y_h are simply inherited from Y. Our finite element approximation to (2.1) is then given by: find $(\hat{\mathbf{u}}_h(\mu) \equiv$

$(u_{h,1}(\mu),\ldots,u_{h,n_b}(\mu))$, $\hat{\boldsymbol{\lambda}}_h(\mu) \equiv (\lambda_{h,1}(\mu),\ldots,\lambda_{h,n_b}(\mu))) \in Y_h^{n_b} \times \mathbb{R}^{n_b}$ such that

$$\begin{aligned} a(u_{h,i}(\mu),v;\mu) &= \lambda_{h,i}(\mu)\, m(u_{h,i}(\mu),v), \quad \forall v \in Y_h, \quad 1 \le i \le n_b, \\ m(u_{h,i}(\mu),u_{h,j}(\mu)) &= \delta_{ij}, \quad 1 \le i \le j \le n_b. \end{aligned} \tag{2.16}$$

The resulting eigenvalue problem can easily be solved by diagonalizing a $\mathcal{N} \times \mathcal{N}$ algebraic systems where $\mathcal{N}$ is chosen such that it is sufficiently large to achieve a desired accuracy for all $\mu \in \mathcal{D}$. Finally to simplify notation, we will drop the subscript h from all subsequent formulation, with the understanding that the "truth" approximation in fact refers to the finite element approximation. Thus, Y, $\hat{\mathbf{u}}$ and $\hat{\boldsymbol{\lambda}}$ shall now be understood as Y_h, $\hat{\mathbf{u}}_h$ and $\hat{\boldsymbol{\lambda}}_h$.

2.2.4 Numerical Example: Simple Harmonic Oscillator

We consider the following linear eigenvalue problem defined on $\Omega \equiv]-L,L[\subset \mathbb{R}$: given $L = 10$ and $\mu \in \mathcal{D} \equiv [1,10]$, we evaluate $(\hat{\mathbf{u}}(\mu),\hat{\boldsymbol{\lambda}}(\mu)) \in Y^{n_b} \times \mathbb{R}^{n_b}$ from (2.1) for $Y \equiv H_0^1(\Omega)$,

$$a(w,v;\mu) = a_1(w,v) + \mu^2 a_2(w,v), \tag{2.17}$$

and

$$a_1(w,v) = \frac{1}{2}\int_\Omega \frac{dw}{dx}\frac{dv}{dx}, \quad a_2(w,v) = \frac{1}{2}\int_\Omega x^2\, w\, v, \quad \text{and} \quad m(w,v) = \int_\Omega w\, v\,. \tag{2.18}$$

The $H_0^1(\Omega) \subset H^1(\Omega)$ is the usual Hilbert space of derivative square-integrable functions that vanish on the domain boundary; and $x \in \Omega$. We solve the resulting problem with the finite element method on a nonuniform mesh $\mathcal{T}$ of size $\mathcal{N} = 4200$: 4000 uniform elements in the -6.5 to 6.5 interval, 100 uniform elements in the -10 to -6.5 interval, and 100 uniform elements in the 6.5 to 10 interval.

The above description of the problem leads to a simple harmonic oscillator problem, of which the strong form is given by

$$\begin{aligned} -\frac{1}{2}\frac{d^2u_i(x;\mu)}{dx^2} + \frac{1}{2}\mu^2x^2u_i(x;\mu) &= \lambda_i(\mu)u_i(x;\mu), \quad 1 \le i \le n_b, \\ \int_\Omega u_i(x;\mu)u_j(x;\mu)dx &= \delta_{ij}, \quad 1 \le i \le j \le n_b, \end{aligned} \tag{2.19}$$

with the boundary conditions $u_i(-L) = u_i(L) = 0$, $1 \le i \le n_b$. Thus, μ is the angular frequency of the system. In the limit of $L \to \infty$, the solutions $\hat{\mathbf{u}}(\,\cdot\,;\mu)$ and $\hat{\boldsymbol{\lambda}}(\,\cdot\,;\mu)$ approach the analytical

solutions given by

$$\lambda_i(\mu) = (i+\frac{1}{2})\mu, \quad 1 \le i \le n_b, \tag{2.20}$$

$$u_i(x;\mu) = \frac{\mu^{1/4}}{\sqrt{2^i i!}} e^{-\mu x^2/2} H_i(\sqrt{\mu}x), \quad 1 \le i \le n_b. \tag{2.21}$$

where $H_i(\cdot)$ are the usual Hermite polynomials. We note that $a_2(u,v)$ is continuous as $L \to \infty$: since the behavior of u for large x is given by $e^{-\mu x^2/2}$ [10], then for $v = 1$

$$\int_0^L x^2 e^{-\mu x^2/2} dx = -\frac{L}{\mu} e^{-\mu L^2/2} + \frac{\sqrt{\pi/2}\mathrm{erf}(\sqrt{\mu/2}L)}{\mu^{3/2}}, \tag{2.22}$$

which gives $\sqrt{\frac{\pi}{2}}\frac{1}{a^{3/2}}$ as $L \to \infty$; thus $a_2(u,v)$ is bounded in the limit $L \to \infty$. We have chosen L sufficiently large that our truth approximation is close to the above analytical solutions given by (2.20) and (2.21) for $n_b \le 10$ and $\forall \mu \in \mathcal{D}$.

Lastly, we define a parameter sample set $\Xi_T \in \mathcal{D}$ which will be used in defining the error measure, and constructing our reduced basis spaces. We choose a Ξ_T consisting of 100 points uniformly distributed in $\mathcal{D}$. We show in Figure 2-1 solutions $u_i(\mu)$ for selected $\mu \in \Xi_T$ and in Figure 2-2 the variation of $\lambda_i(\mu)$ with μ, for $i = 1, 3, 5$ and 7. Although we see a linear variation in $\lambda_i(\mu)$, $1 \le i \le n_b$ with μ, eigenvectors $u_i(\mu)$, $1 \le i \le n_b$ have richer behavior, and as such an accurate approximation is not trivial.

2.3 Reduced Basis Approximation

2.3.1 Critical Observation: Dimension Reduction

In the "truth" approximation as described in Section 2.2.3, we have represented $u_i(\mu)$, $1 \le i \le n_b$ by a linear combination of $\phi_k^{\mathcal{N}} \in Y$, $1 \le k \le \mathcal{N}$ — $u_i(\mu)$ is an arbitrary member of Y. However, the solution $\hat{\mathbf{u}}(\mu)$ can in fact be localized to a much lower-dimensional manifold $\mathcal{M} \equiv \{\hat{\mathbf{u}}(\mu),\ \mu \in \mathcal{D}\}$ residing in Y^{n_b}. In the case of a single parameter, $\mathcal{M}$ can be visualized as a one-dimensional filament that winds through Y^{n_b} as sketched in Figure 2-3. Presuming that $\mathcal{M}$ is sufficiently smooth, we can then represent $\hat{\mathbf{u}}(\mu)$ by elements in span $\{\mathcal{M}\}$. The reduced basis approach explicitly recognizes this computational opportunity.

To consolidate the above argument, we introduce the notion of Kolmogorov N-width d_N [56,

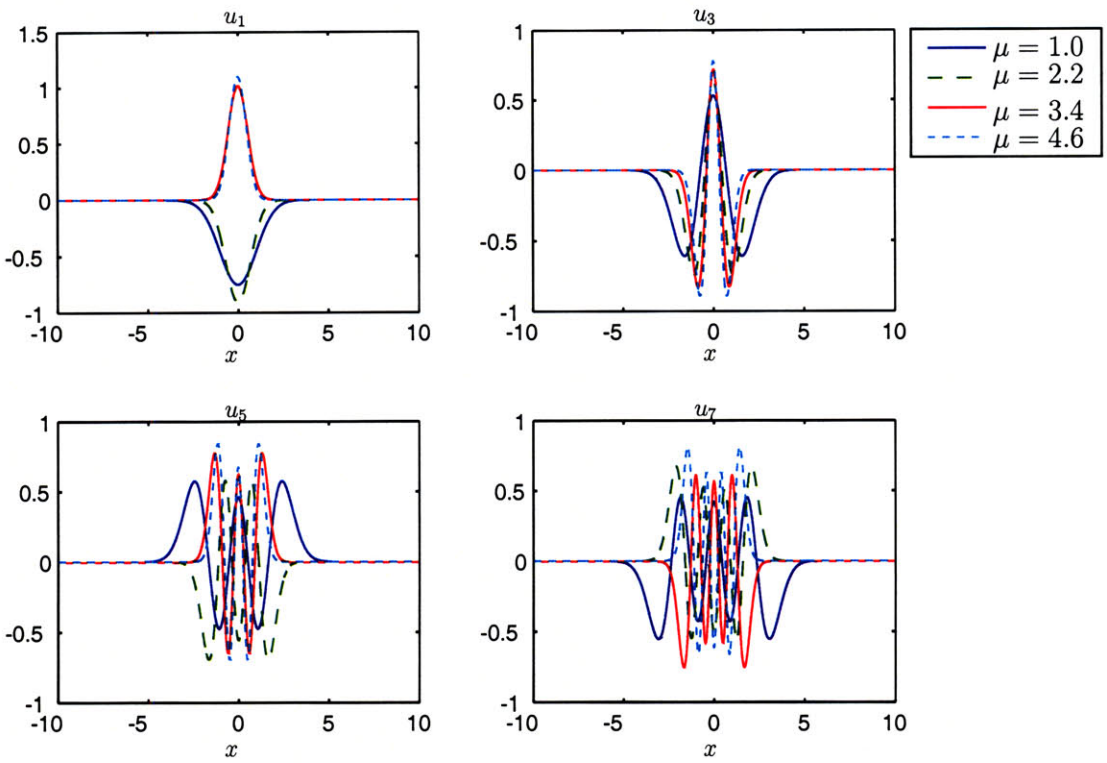

Figure 2-1: Solutions $u_1(\mu)$, $u_3(\mu)$, $u_5(\mu)$ and $u_7(\mu)$ at $\mu = 1.0$, 2.2, 3.4 and 4.6.

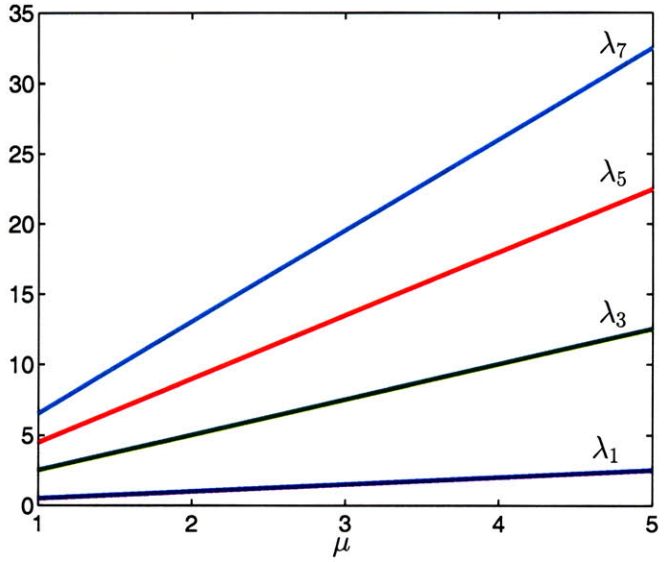

Figure 2-2: Variations of $\lambda_1(\mu)$, $\lambda_3(\mu)$, $\lambda_5(\mu)$ and $\lambda_7(\mu)$ with μ.

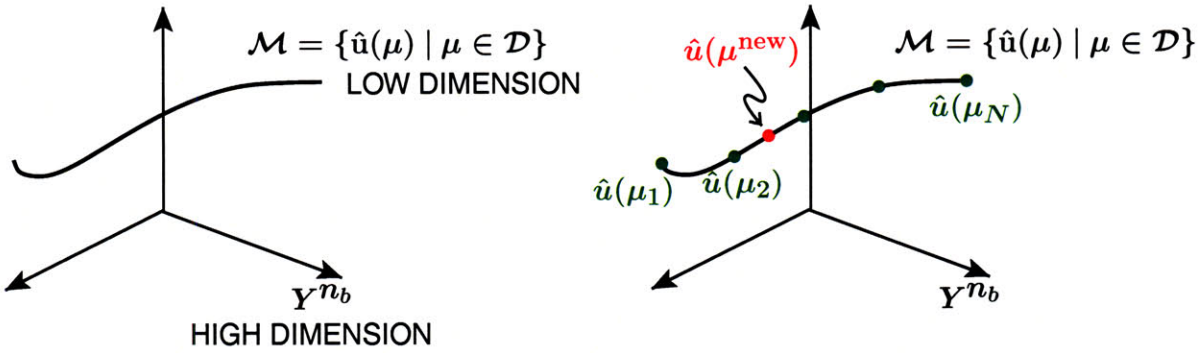

Figure 2-3: Conceptual drawing of the solution manifold $\mathcal{M}$.

63, 96]:

$$d_N(A, Y) \equiv \inf_{Y_N \subset Y} \sup_{x \in A} \inf_{y \in Y_N} \|x - y\|_Y, \tag{2.23}$$

where A is a subset of Y and Y_N an arbitrary N-dimensional subspace of Y. The Kolmogorov N-width, d_N, measures the extent to which A may be approximated by a finite dimensional space of dimension N in Y. We will have a rapidly convergent approximation if d_N approaches zero rapidly as N increases. For our case where $A \equiv \mathcal{M}$, we can attribute this to the smoothness of the solutions with respect to μ, as demonstrated for a single-parameter elliptic problem in [74]. In [72], it is further shown that d_N is almost realized if Y_N is spanned by elements in $\mathcal{M}$. The construction of $Y_N \subset \mathcal{M}$ that minimizes d_N is, however, combinatorically difficult. The reduced basis method then provides an efficient procedure by which we can construct a good surrogate to Y_N.

A naive implementation of the Lagrangian approach of reduced basis method, i.e. representing $\hat{\mathbf{u}}$ by a linear combination of $\hat{\mathbf{u}}(\mu_n)$, $1 \leq n \leq N$, is, however, doomed to fail in most cases. Suppose we are interested in a case where $n_b = 2$; $\lambda_1(\mu)$ is of multiplicity 1 for all $\mu \in \mathcal{D}$; $\lambda_2(\mu)$ is of multiplicity 1 for all $\mu \in \mathcal{D} \setminus \mu_0$; and $\lambda_2(\mu_0)$ is of multiplicity 2. This suggests that mode shapes[1] of $u_2(\mu)$ can be different for $\mu < \mu_0$ and $\mu > \mu_0$. Assuming this is the case, $\hat{\mathbf{u}}(\mu) \equiv (u_1(\mu), u_2(\mu))$ then lies on 2 separate manifolds $\mathcal{M}_1$ and $\mathcal{M}_2$. Based on the smoothness argument, sufficient basis functions from each manifold must be included before a good reduced basis approximation can be obtained — it is equivalent to having two separate reduced basis approximations. As we consider higher n_b and richer parameter domain, this discontinuity can multiply, leading to potentially large N. Such scenario is not uncommon in computational chemistry as will be shown in Chapter 5. This

[1]Here, a mode shape refers to a particular "shape". For the current example, mode shapes can be distinguished by the number of nodes, i.e., the number of times an eigenvector crosses the x axis.

thus requires new approaches to defining appropriate approximation spaces. However, we note that the purported discontinuities are often not a property of the problem but artifacts of eigenvalue solvers. We demonstrate in Section 2.3.2 that the solution manifold $\mathcal{M}$ is smooth. We may recover a smooth solution manifold with some post-processing steps.

In this chapter, we will focus on the case where $\lambda_i(\mu)$, $1 \leq i \leq n_b$ has multiplicity 1 for all $\mu \in \mathcal{D}$. We address the more difficult scenario described above in Chapter 5.

2.3.2 Parametric Derivatives

Here, we examine further the smoothness property of the solution manifold $\mathcal{M}$ — for this purpose, we examine the parametric (or sensitivity) derivatives of $\hat{\mathbf{u}}(\mu)$. We proceed formally, examining first the parametric smoothness of an arbitrary ith component of $\hat{\mathbf{u}}(\mu)$, which we denote as $u_i(\mu)$. We will also assume that $\lambda_i(\mu)$ has multiplicity 1.

To begin, we define $(\partial u/\partial\mu) : \mathcal{D} \rightarrow \mathbb{R}$ as the derivative of $u_i(x;\mu)$ with respect to the parameter μ and $(\partial\lambda_i/\partial\mu) : \mathcal{D} \rightarrow \mathbb{R}$ as the derivative of $\lambda_i(\mu)$ with respect to the parameter μ. We shall assume the functions $\Theta_q(\mu)$, $1 \leq q \leq Q$ in (2.6) are all $C^1(\mathcal{D})$ (continuously differentiable over $\mathcal{D}$). By differentiating (2.1) with respect to μ, we obtain

$$\begin{aligned} a\left(\frac{\partial u_i}{\partial\mu}(\mu), v; \mu\right) &= -\sum_{q=1}^{Q} \frac{\partial\Theta_q}{\partial\mu}(\mu) a_q(u_i(\mu), v) \\ &\quad + \lambda_i(\mu) m\left(\frac{\partial u_i}{\partial\mu}(\mu), v\right) + \frac{\partial\lambda_i}{\partial\mu}(\mu) m(u_i(\mu), v), \quad \forall v \in Y. \end{aligned} \tag{2.24}$$

Let $v = u_i(\mu) \in Y$; then

$$\begin{aligned} \frac{\partial\lambda_i}{\partial\mu}(\mu) m(u_i(\mu), u_i(\mu)) &= a\left(\frac{\partial u_i}{\partial\mu}(\mu), u_i(\mu); \mu\right) - \lambda_i(\mu) m\left(\frac{\partial u_i}{\partial\mu}(\mu), u_i(\mu)\right) \\ &\quad + \sum_{q=1}^{Q} \frac{\partial\Theta_q}{\partial\mu}(\mu) a_q(u_i(\mu), u_i(\mu)) \\ \frac{\partial\lambda_i}{\partial\mu}(\mu) &= \sum_{q=1}^{Q} \frac{\partial\Theta_q}{\partial\mu}(\mu) a_q(u_i(\mu), u_i(\mu)); \end{aligned} \tag{2.25}$$

since $m(u_i(\mu), u_i(\mu)) = 1$; and $a((\partial u_i/\partial\mu)(\mu), u_i(\mu); \mu) - \lambda_i(\mu) m((\partial u_i/\partial\mu)(\mu), u_i(\mu)) = 0$ from (2.1), and based on the symmetric property of $a(\,\cdot\,,\,\cdot\,;\mu)$ and $m(\,\cdot\,,\,\cdot\,)$.

Therefore, for any given $\mu \in \mathcal{D}$ and thus $u_i(\mu) \in Y$ and $\lambda_i(\mu) \in \mathbb{R}$, $(\partial u_i/\partial\mu)(\mu)$ satisfies

$$a\left(\frac{\partial u_i}{\partial \mu}(\mu), v; \mu\right) - \lambda_i(\mu) m\left(\frac{\partial u_i}{\partial \mu}(\mu), v\right) = -\sum_{q=1}^{Q} \frac{\partial \Theta_q}{\partial \mu}(\mu) a_q(u_i(\mu), v) + \left(\sum_{q=1}^{Q} \frac{\partial \Theta_q}{\partial \mu}(\mu) a_q(u_i(\mu), v)\right) m(u_i(\mu), v), \quad (2.26)$$

for all $v \in Y$. It directly follows from our coercivity and continuity assumptions on a and m, and differentiability assumptions on the $\Theta_q(\mu)$, $1 \leq q \leq Q$ that (2.26) admits a unique and stable solution. We can then easily bound $\|\partial u_i/\partial\mu\|_Y$. If we assume the parameter functions $\Theta_q(\mu), 1 \leq q \leq Q$ are in fact $C^\infty(\mathcal{D})$, we can proceed formally and continue this differentiation process indefinitely. Then, the parametric derivatives and well defined and bounded in Y.

We now examine the case where the multiplicity of the eigenvalues can be greater than 1. Referring to Figure 2-4,we note that $\lambda_i(\mu_0)$ and $\lambda_{i+1}(\mu_0)$ are equivalent — we yield two values of $(\partial\lambda_i/\partial\mu)(\mu_0)$ since we can use either $u_i(\mu_0)$ or $u_{i+1}(\mu_0)$ in (2.25). For a small $\Delta\mu$, we see that $\lambda_{i+1}(\mu_0 + \Delta\mu)$ is a smooth variation of $\lambda_i(\mu_0 - \Delta\mu)$; then by evaluating (2.25) based on $u_{i+1}(\mu_0 + \Delta\mu)$, we can also show that $u_{i+1}(\mu_0 + \Delta\mu)$ is a smooth variation of $u_i(\mu_0 - \Delta\mu)$. This illustrates again some of the points made in Section 2.3.1: here, the discontinuity in $u_i(\mu)$ is simply due to how $u_i(\mu)$, $1 \leq i \leq n_b$ are ordered — a factor that depends on the eigensolver. For the example in Figure 2-4, if we denote $\lambda_i(\mu)$ by the blue line and $\lambda_{i+1}(\mu)$ by the red line (instead of ordering them by their magnitudes, as indicated in Figure 2-4 by the red and blue dots), then both $\lambda_i(\mu)$ and $u_i(\mu)$ will be smooth functions of μ.

We note that the magnitude of the parametric derivatives in the Y norm will typically increase with increasing order [92] — the rate at which the magnitude of these derivatives grow is important in the development of *a priori* convergence theory. For vectorial solutions, the relative rate at which the parametric derivatives of $u_i(\mu)$, $1 \leq i \leq n_b$ — components of $\hat{\mathbf{u}}$ — grow are also important. We will examine this issue again in Section 2.3.7.

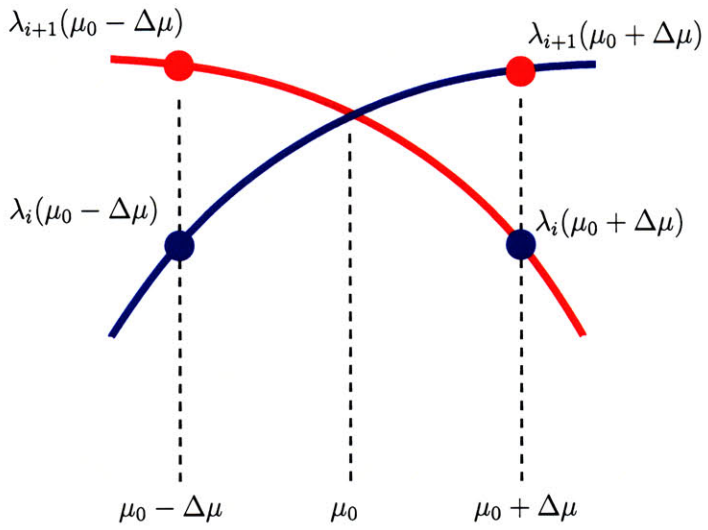

Figure 2-4: Discontinuity in $\lambda_i(\mu)$.

2.3.3 Error Measures

Projection errors

Given an approximation space W_N of dimension N, we would like to measure how well W_N approximate $\mathcal{M}$. The error measure we will use is the projection error $\varepsilon^{\text{proj}}_{N,n_b}$, given by

$$\varepsilon^{\text{proj}}_{N,n_b} = \max_{\mu \in \Xi_T} \epsilon^{\text{proj}}_{N,n_b}(\mu), \tag{2.27}$$

where Ξ_T is a test sample set;

$$\epsilon^{\text{proj}}_{N,n_b}(\mu) = \left(\frac{\sum_{i=1}^{n_b} \|u_{p,i}(\mu) - u_i(\mu)\|_Y^2}{\sum_{i=1}^{n_b} \|u_i(\mu)\|_Y^2} \right)^{1/2}; \tag{2.28}$$

and $\hat{\mathbf{u}}_p(\mu) \equiv (u_{p,1}(\mu), \ldots, u_{p,n_p}(\mu))$ is the best possible solution obtained through a projection of $\hat{\mathbf{u}}(\mu)$ onto W_N — $\hat{\mathbf{u}}_p(\mu)$ depends on how W_N is defined. We will introduce the definition of $\hat{\mathbf{u}}_p(\mu)$ when we introduce the approximation spaces that we will look at.

Reduced basis approximation errors

We denote the reduced basis approximation to $(\hat{\mathbf{u}}(\mu) \equiv (u_1(\mu), \ldots, u_{n_b}(\mu))$, $\hat{\boldsymbol{\lambda}} \equiv (\lambda_1(\mu), \ldots, \lambda_{n_b}(\mu))$ by $(\hat{\mathbf{u}}_N(\mu) \equiv (u_{N,1}(\mu), \ldots, u_{N,n_b}(\mu))$, $\hat{\boldsymbol{\lambda}}_N \equiv (\lambda_{N,1}(\mu), \ldots, \lambda_{N,n_b}(\mu))$. We then define the reduced

basis approximation errors in $\hat{\mathbf{u}}$ and $\hat{\boldsymbol{\lambda}}$ as

$$\varepsilon^u_{N,n_b} = \max_{\mu\in\Xi_T} \epsilon^u_{N,n_b}(\mu), \tag{2.29}$$

$$\varepsilon^\lambda_{N,n_b} = \max_{\mu\in\Xi_T} \epsilon^\lambda_{N,n_b}(\mu); \tag{2.30}$$

where Ξ_T is the training sample set, and[2]

$$\epsilon^u_{N,n_b}(\mu) = \frac{(\sum_{i=1}^{n_b} \|u_{N,i}(\mu) - u_i(\mu)\|_Y^2)^{1/2}}{(\sum_{i=1}^{n_b} \|u_i(\mu)\|_Y^2)^{1/2}}, \tag{2.31}$$

$$\epsilon^\lambda_{N,n_b}(\mu) = \max_{1\le i\le n_b} \left|\frac{\lambda_{N,i}(\mu) - \lambda_i(\mu)}{\lambda_i(\mu)}\right|. \tag{2.32}$$

2.3.4 Augmented Reduced Basis Space

The approximation space

We first introduce nested sample sets $S^A_N = (\mu_1, \ldots, \mu_{N_s})$, $1 \le N_s \le N_{s,\max}$. The superscript A stands for "Augmented", N_s is the number of sample points in S^A_N, and $N_{s,\max}$ is the maximum number of sample points we will use[3]. We then define the associated nested reduced-basis spaces as

$$W^A_N = \text{span}\,\{u_i(\mu_j), 1 \le i \le n_b, 1 \le j \le N_s\}, \quad 1 \le N_s \le N_{s,\max}, \tag{2.33}$$

$$= \text{span}\,\{\zeta_n, 1 \le n \le N \equiv N_s n_b\}, \quad 1 \le N_s \le N_{s,\max}. \tag{2.34}$$

Since W^A_N is the span of all n_b eigenvectors at all N_s sample points in S^A_N, the dimension of W^A_N, N, is given by $N_s \times n_b$. Further, we orthogonalize $u_i(\mu_j)$, $1 \le i \le n_b$, $1 \le j \le N_s$ to obtain a better-conditioned set of basis functions, ζ_n, $1 \le n \le N$ through the following Gram-Schmidt orthogonalization procedure in the $(\,\cdot\,,\,\cdot\,)_Y$ inner product. Figure 2-5 summarizes the orthogonalization procedure for W^A_N.

Finally, an approximation of $u_i(\mu)$ in W^A_N is given by $u_{N,i}(\mu) = \sum_{n=1}^N \alpha_{i\,n}(\mu)\zeta_n$. As an initial

[2] As mentioned in Section 2.2.1, the eigenvectors are unique up to a multiplication by scalars since the eigenvalues are distinct. Since $u_i(\mu)$ are normalized to one, this amounts to a multiplication by ± 1. We must thus remove this sign variation before computing $\epsilon^u_{N,n_b}(\mu)$. More precisely, we have

$$\epsilon^u_{N,n_b}(\mu) = \frac{(\sum_{i=1}^{n_b} \min(\|u_{N,i}(\mu) - u_i(\mu)\|_Y^2, \|u_{N,i}(\mu) + u_i(\mu)\|_Y^2))^{1/2}}{(\sum_{i=1}^{n_b} \|u_i(\mu)\|_Y^2)^{1/2}}.$$

$\zeta_i = u_i(\mu_1), 1 \le i \le n_b;$

for $j = 2: N_s$

  for $i = 1: n_b$

$$z = u_i(\mu_j) - \sum_{m=1}^{n_b(j-1)+i-1} (u_i(\mu_j), \zeta_m)_Y \, \zeta_m;$$

$$\zeta_{n_b(j-1)+i} = z/\|z\|_Y;$$

  end

end.

Figure 2-5: Orthogonalization procedure for the augmented reduced basis space, W_N^A.

n_b	N_s	N
2	9	18
3	9	27
4	8	24
5	6	30
6	5	30
7	5	35
8	4	32
9	4	36
10	4	40

Table 2.1: N_s required for $\varepsilon_{N,n_b}^{\text{proj}} < 1\text{E}-5$ and the corresponding N for W_N^A and $2 \le n_b \le 10$.

indication of the efficiency of approximating based on W_N^A, we compute the projection error given by (2.27) where the best projected solution $\hat{\mathbf{u}}_p(\mu) \equiv (u_{p,1}(\mu), \ldots, u_{p,n_p}(\mu))$ is defined as

$$u_{p,i}(\mu) = \arg \min_{w \in W_N^A} \|w - u_i(\mu)\|_Y, \quad 1 \le i \le n_b. \tag{2.35}$$

As shown in Figure 2-6, the $\varepsilon_{N,n_b}^{\text{proj}}$ is rapidly convergent with N for $2 \le n_b \le 17$. In addition, as n_b increases, the N_s required to reduce $\varepsilon_{N,n_b}^{\text{proj}}$ to below 10^{-5} decreases as shown in Table 2.1; thus N does not increase linearly with n_b. We explain this behavior in a later section when we discuss further the convergence results for our reduced basis approximation based on W_N^A.

---

[3]The choice for $N_{s,\max}$ depends on the maximum accuracy we would like to approximate $\hat{\mathbf{u}}$ and $\hat{\boldsymbol{\lambda}}$ — this will be made clear in Section 2.3.6

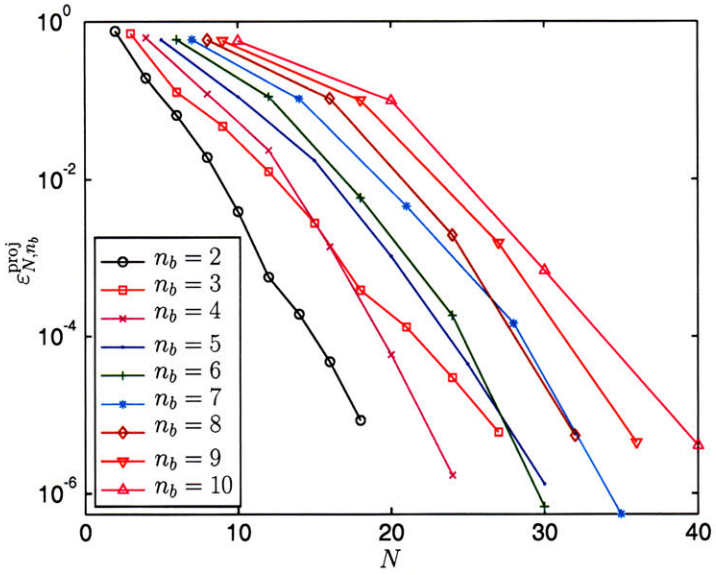

Figure 2-6: Convergence of the projection error of $\hat{\mathbf{u}}$ onto the augmented reduced basis space $(W_N^A)^{n_b}$, $\varepsilon^{\text{proj}}_{N,n_b}$ (given by (2.27)), with N for $2 \le n_b \le 10$.

The approximation

Reduced basis approximations to $\hat{\mathbf{u}}(\mu)$ and $\hat{\boldsymbol{\lambda}}(\mu)$ are obtained by a Galerkin projection onto the augmented reduced basis space: given a $\mu \in \mathcal{D}$, find $(\hat{\mathbf{u}}_N(\mu) \equiv (u_{N,1}(\mu), \ldots, u_{N,n_b}(\mu))$, $\hat{\boldsymbol{\lambda}}_N(\mu) \equiv (\lambda_{N,1}(\mu), \ldots, \lambda_{N,n_b}(\mu))) \in (W_N^A)^{n_b} \times \mathbb{R}^{n_b}$ such that

$$\begin{aligned} a(u_{N,i}(\mu), v; \mu) &= \lambda_{N,i}(\mu)\, m(u_{N,i}(\mu), v), \quad \forall v \in W_N^A, \quad 1 \le i \le n_b, \\ m(u_{N,i}(\mu), u_{N,j}(\mu)) &= \delta_{ij}, \quad 1 \le i \le j \le n_b. \end{aligned} \tag{2.36}$$

Affine parameter dependence

To develop an efficient computational procedure, we will exploit the affine parameter dependence property of $a(\cdot\,, \cdot\,; \mu)$ described in Section 2.2.2. For our numerical example, $Q = 2$, $\Theta_1(\mu) = 1$, $\Theta_2(\mu) = \mu^2$, $a_1(w, v)$ is given by (2.20) and $a_2(w, v)$ is given by (2.21). As mentioned in Section 2.2.2, the parameter functions $\Theta_q(\mu), 1 \le q \le Q$ can be readily evaluated in $O(1)$ operations. On the other hand, evaluations of $a_q(w, v), 1 \le q \le Q$ incur $O(\mathcal{N})$ operations but these functionals are parameter-independent.

To examine how the affine parameter dependence property can lead to an efficient computational strategy, we examine the reduced basis algebraic system for (2.36).

Discrete equations

We expand our reduced-basis approximation as

$$u_{N,n}(\mu) = \sum_{j=1}^{N} u_{N,n\,j}(\mu)\zeta_j. \tag{2.37}$$

Inserting this representation into (2.36) yields

$$\begin{aligned}
\left(\sum_{q=1}^{Q}\sum_{j=1}^{N}\Theta_q(\mu)A_{i,j}^{N,q}\right) u_{N,n\,j}(\mu) &= \lambda_{N,n}(\mu)M_{i,j}^{N}u_{N,n\,j}(\mu), \quad 1 \le i \le N, \quad 1 \le n \le n_b; \\
\sum_{i=1}^{N}\sum_{j=1}^{N} u_{N,n\,i}(\mu)M_{i,j}^{N}u_{N,n'\,j}(\mu) &= \delta_{n,n'}, \quad 1 \le n, n' \le n_b.
\end{aligned} \tag{2.38}$$

For our numerical example, $Q = 2$; $\Theta_1(\mu) = 1$, $\Theta_2(\mu) = \mu^2$; and $A^{N,1} \in \mathbb{R}^{N\times N}$, $A^{N,2} \in \mathbb{R}^{N\times N}$, and $M^N \in \mathbb{R}^{N\times N}$ are given by $A_{i,j}^{N,1} = a_1(\zeta_j,\zeta_i), 1 \le i,j \le N$, $A_{i,j}^{N,2} = a_2(\zeta_j,\zeta_i), 1 \le i,j \le N$, and $M_{i,j}^{N} = m(\zeta_j,\zeta_i), 1 \le i,j \le N$. To solve (2.38), we can use any eigenvalue solver.

Offline-online computational framework

We observe that we can now develop an efficient offline–online computational strategy for the rapid evaluation of $\lambda_{N,n}(\mu)$ for each μ in $\mathcal{D}$ — a strategy where the operation count in the online stage is independent of $\mathcal{N}$ and only dependent on N, which we expect to be much smaller than $\mathcal{N}$.

In the offline stage — performed once — we generate nested reduced-basis spaces $W_N^A = \{\zeta_1, \ldots, \zeta_N\}$, $1 \le N \le N_{\max}$ at the costs of $n_b\mathcal{N}^\bullet$ — the $\bullet$ denotes the actual computational complexity of the "truth" approximation, which due to sparsity should be less than 3. We then form and store $A^{N,1}$, $A^{N,2}$, and M^N at the costs of $(Q+1)N^2\mathcal{N}^2$. The storage of each matrix requires a space of $N \times N$.

In the online stage — performed many times for each new μ — we solve (2.38) for $u_{N,i}(\mu), 1 \le i \le n_b$. The reconstruction of the reduced basis system is QN^2 and solving the resulting discrete equations is of $O(N^3)$. The total operation count of the online stage is then $O(QN^2+N^3)$; we thus achieve an computational complexity that is independent of $\mathcal{N}$ and dependent only on N. The ability to calculate $a_q(\zeta_j,\zeta_i)$ offline liberates the online computation from the $O(\mathcal{N})$ complexity.

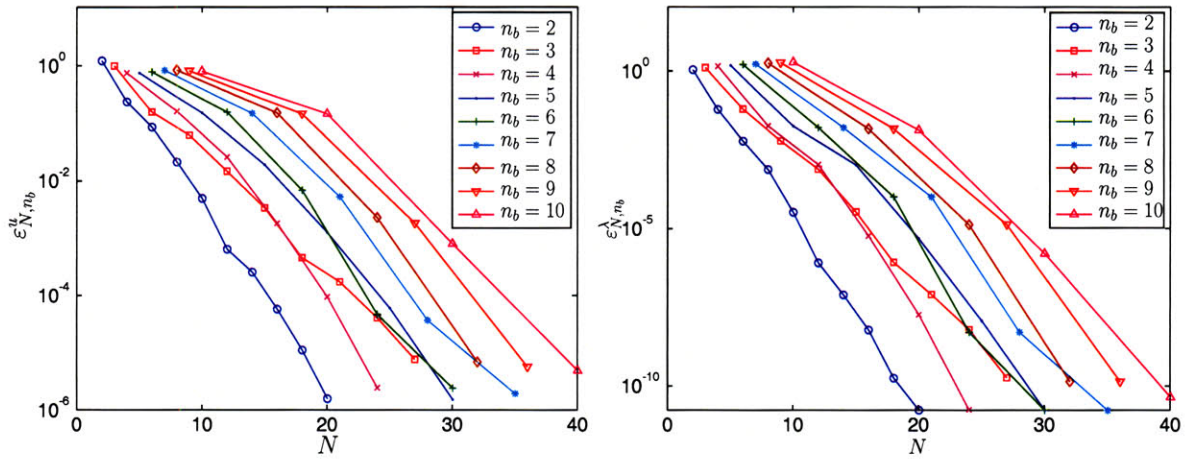

Figure 2-7: Convergence of the reduced basis error of $\hat{\mathbf{u}}_N(\mu)$, ε^u_{N,n_b} (given by (2.29)), and the reduced basis error of $\hat{\boldsymbol{\lambda}}_N(\mu)$, $\varepsilon^\lambda_{N,n_b}$ (given by (2.30)), with N for $2 \leq n_b \leq 10$ and $\hat{\mathbf{u}}_N(\mu) \in (W_N^A)^{n_b}$.

Convergence

We shall use error measures defined by (2.29) and (2.30). As shown in Figure 2-7, the reduced basis approximation is rapidly convergent — both ε^u_{N,n_b} and $\varepsilon^\lambda_{N,n_b}$ decrease rapidly with N. In addition, compared to the projection results in Figure 2-6, our reduced basis approximation has similar performance; our reduced basis approximation is close to optimal.

In Table 2.2, for a tolerance criteria of $\varepsilon^\lambda_{N,n_b} \leq 1\text{E}{-}10$, we see N_s decreases with n_b. However, for a coarser tolerance criteria of $\varepsilon^\lambda_{N,n_b} \leq 1\text{E}{-}2$, N_s remains approximately constant for all n_b. This suggests two things. First, N_s must be above some critical value of N_s in order to obtain a reasonable approximation — here N_s must be greater than 3 in order to get an accuracy of $\varepsilon^\lambda_{N,n_b} \leq 1\text{E}{-}2$ for all n_b examined. Second, for N_s greater than this critical value, incremental improvement in the solutions can be obtained either through inclusion of solutions at more μ points or higher eigenmodes or both. Functions approximation based on the eigenmodes is a common technique in spectral methods, for example expansion in Fourier modes or eigenfunctions of a suitable Sturm Liouville problem [23]. The second observation, coupled with the observation that $i^* \equiv \arg\max_{1 \leq i \leq n_b} \left| \frac{\lambda_{N,i}(\mu) - \lambda_i(\mu)}{\lambda_i(\mu)} \right|$ is always 1 for all $\mu \in \Xi_T$ thus explain the decrease of N_s as the tolerance criteria is tightened.

We note that even if the eigenvalues are not distinct, for example in the case illustrated in Figure 2-4, the augmented reduced basis approximation will not break down. In fact, for the

n_b	N_s		
	$\varepsilon^{\lambda}_{N,n_b} < 1\text{E}{-2}$	$\varepsilon^{\lambda}_{N,n_b} < 1\text{E}{-4}$	$\varepsilon^{\lambda}_{N,n_b} < 1\text{E}{-10}$
2	3	5	9
3	3	5	9
4	3	4	8
5	3	4	6
6	3	4	5
7	3	4	5
8	3	3	4
9	3	3	4
10	3	3	4

Table 2.2: N_s required to reduce the reduced basis error of $\hat{\boldsymbol{\lambda}}_N$, $\varepsilon^{\lambda}_{N,n_b}$(given by (2.30)), to below $1\text{E}{-2}$, $1\text{E}{-4}$ and $1\text{E}{-10}$ for $2 \leq n_b \leq 10$ and $(\hat{\mathbf{u}}_N(\mu), \hat{\boldsymbol{\lambda}}_N(\mu)) \in (W_N^A)^{n_b} \times \mathbb{R}^{n_b}$.

particular case illustrated in Figure 2-4, should the eigenvectors at $\mu_0 \pm \Delta\mu$ be included in W_N^A, the discontinuity does not affect the approximation at all — the Galerkin procedure will find the right linear combination of the basis functions in W_N^A. For example, to approximate $u_i(\mu)$ for $\mu_0 < \mu < \mu_0 + \Delta\mu$, the Galerkin procedure will choose a linear combination involving $u_i(\mu_0 + \Delta\mu)$ and $u_{i+1}(\mu_0 - \Delta\mu)$, instead of a linear combination involving $u_i(\mu_0 + \Delta\mu)$ and $u_i(\mu_0 - \Delta\mu)$.

2.3.5 Vectorial Reduced Basis Space

Here, we look at the strict interpretation of Lagrangian reduced basis approximation: expressing $\hat{\mathbf{u}}(\mu)$ as a linear combination of selected solutions on the solution manifold $\mathcal{M}$. We shall demonstrate that this is possible only when the components of basis solutions are first preprocessed to recover a smooth solution manifold. In addition, under certain circumstances, the resulting approximation space can be very economical as it exploits (through the reduced basis space) the inherent orthogonality properties between the solutions $u_i(\mu)$, $1 \leq i \leq n_b$ for a given μ, and their common smoothness.

The approximation space

We first introduce nested sample sets $S_N^V = (\mu_1, \ldots, \mu_N)$, $1 \leq N \leq N_{\max}$, where N is the number of sample points in S_N^V, and $N_{\max}$ is the maximum number of sample points we will use. We then

define associated nested reduced-basis spaces as

$$W_N^V = \text{span } \{\hat{\mathbf{u}}(\mu_n), 1 \le n \le N\}, \quad 1 \le N \le N_{\max}, \tag{2.39}$$

$$= \text{span } \{\hat{\boldsymbol{\zeta}}_n, 1 \le n \le N\}, \quad 1 \le N \le N_{\max}; \tag{2.40}$$

where $\hat{\mathbf{u}}(\mu_n) \equiv (u_1(\mu_n), \ldots, u_{n_b}(\mu_n))$ are solutions of (2.1) at $\mu = \mu_n$; and $\hat{\boldsymbol{\zeta}} \equiv (\zeta_1, \ldots, \zeta_{n_b})$ are basis functions obtained after $\hat{\mathbf{u}}(\mu_n), 1 \le n \le N$ are preprocessed. We note that dimensions of the sample set S_N^V and the approximation space W_N^V are the same — both are of dimension N. This is because each basis of W_N^V consists of n_b components, and as such is vectorial in nature. The superscript V in W_N^V then stands for "Vectorial". We now describe the two preprocessing steps required to obtain a well-conditioned approximation space.

Pre-processing

The first preprocessing step is the alignment procedure. For this particular problem, due to the one-dimensional nature of the problem and multiplicity of 1 for all $\lambda_i(\mu)$, $1 \le i \le n_b$, $\forall \mu \in \mathcal{D}$, the procedure is simple: we only need to remove the sign variation in the eigenfunctions. Given a pre-sorted space $U_N \equiv \{\hat{\boldsymbol{\zeta}}_n^s, 1 \le n \le N\}$ where $\hat{\boldsymbol{\zeta}}_n^s$, $1 \le n \le N$ are the sorted basis functions of $\hat{\mathbf{u}}_n$, $1 \le n \le N$, we wish to add $\hat{\mathbf{u}}(\mu_{N+1})$ to U_N to form U_{N+1}. We first select a $\hat{\boldsymbol{\zeta}}_n^s \in U_N$ such that $\mu_n \in S_N^V$ is closest to μ_{N+1}. For $i = 1, \ldots, n_b$, we determine Y norms $\|\cdot\|_Y$ of differences between $\zeta_{n,i}^s$ and $u_i(\mu_{N+1})$, and between $\zeta_{n,i}^s$ and $-u_i(\mu_{N+1})$; the smaller of two then determines whether $\zeta_{N+1,i}^s = u_i(\mu_{N+1})$ or $-u_i(\mu_{N+1})$. The result is $\hat{\boldsymbol{\zeta}}^s$ that varies smoothly with μ, as shown in Figure 2-9. Figure 2-8 summarizes the above procedure for $N \le N_{\max}$.

The second preprocessing step is the pseudo-orthogonalization of the basis functions $\hat{\boldsymbol{\zeta}}_n^s$, $1 \le n \le N$ in the inner product $(\cdot, \cdot)_V$ given by

$$(\hat{\mathbf{w}}, \hat{\mathbf{w}})_V = \sum_{i=1}^{n_b} (w_i, w_i)_Y, \tag{2.41}$$

where $\hat{\mathbf{w}} \equiv (w_1, \ldots, w_{n_b})$ and the norm $\|\cdot\|_V$ is defined as

$$\|\hat{\mathbf{w}}\|_V^2 = \frac{1}{n_b} \sum_{i=1}^{n_b} \|w_i\|_Y^2. \tag{2.42}$$

$U_1 = \{\hat{\boldsymbol{\zeta}}_1^s \equiv \hat{\mathbf{u}}(\mu_1)\};$

for $N = 2 : N_{\max}$

$n^* = \arg \min_{1 \le n \le N-1} |\mu_n - \mu_N|;$

for $i = 1 : n_b$

$e^+ = \|\zeta_{n^*,i}^s + u_i(\mu_N)\|_Y;$

$e^- = \|\zeta_{n^*,i}^s - u_i(\mu_N)\|_Y;$

if $e^- > e^+$

$\zeta_{N,i}^s = -u_i(\mu_N);$

else

$\zeta_{N,i}^s = u_i(\mu_N);$

end

end

$U_N = U_{N-1} \cup \hat{\boldsymbol{\zeta}}_N^s;$

end.

Figure 2-8: The alignment procedure for problems with eigenvectors with non-degenerate eigenvalues.

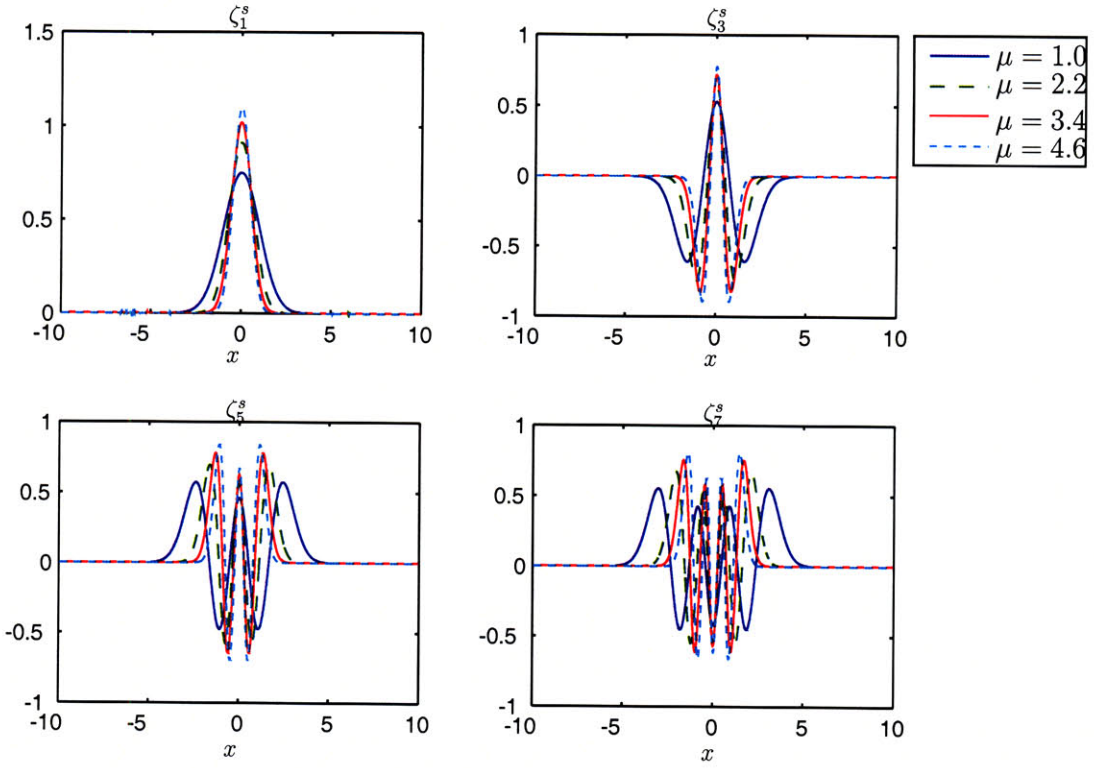

Figure 2-9: Solutions ζ_1^s, ζ_3^s, ζ_5^s and ζ_7^s at $\mu = 1.0$, 2.2, 3.4 and 4.6.

Again, the subscript V serves as a reminder that we are dealing with vectorial solutions. The orthogonalization procedure then proceeds as follow: given a reduced basis space $W_N^V = \text{span}\ \{\hat{\boldsymbol{\zeta}}_j, 1 \leq j \leq N\}$, we would like to add a new basis $\hat{\boldsymbol{\zeta}}_{N+1}^s$ to W_N^V. We first compute $\hat{\mathbf{b}} = \hat{\boldsymbol{\zeta}}_{N+1}^s - \sum_{n=1}^N \alpha_n^* \hat{\boldsymbol{\zeta}}_n$, where $\alpha^* \in \mathbb{R}^N$ is given by $\arg\min_{\alpha \in \mathbb{R}^N} \sum_{i=1}^{n_b} \|\zeta_{N+1,i}^s - \sum_{j=1}^N \alpha_j \zeta_{j,i}\|_Y^2$. Determining α^* is equivalent to solving the following algebraic equations:

$$\sum_{i=1}^{n_b} \sum_{n=1}^{N-1} (\zeta_{m,i}, \zeta_{n,i})_Y \alpha_n^* = \sum_{j=1}^{n_b} (\zeta_{m,j}, \zeta_{N,j}^s)_Y, \quad 1 \leq m \leq N-1. \tag{2.43}$$

The new, pseudo-orthogonalized basis function is then given by $\hat{\boldsymbol{\zeta}}_{N+1} = \hat{\mathbf{b}}/(\frac{1}{n_b} \sum_{i=1}^{n_b} \|b_i\|_Y^2)^{1/2}$ and $W_{N+1}^V = W_N^V + \text{span}\ \{\hat{\boldsymbol{\zeta}}_{N+1}\}$. The normalization is consistent with the norm defined in (2.42). Figure 2-10 summarizes the above orthogonalization procedure. We note that components of any two basis functions are not orthogonal, i.e. $m(\zeta_{n,i}, \zeta_{m,j}) \neq 0$ for $i \neq j$ if $n \neq m$. However, the $\hat{\boldsymbol{\zeta}}_n$, $1 \leq n \leq N$ are orthogonal in the $\|\cdot\|_V$. Thus, the term "pseudo-orthogonalization" is used to distinguish the current procedure from the orthogonalization procedure given by Figure 2-5 for

$W_1^V \equiv \text{span}\ \{\hat{\zeta}_1 \equiv \hat{\zeta}_1^s\}$;

for $N = 2 : N_{\max}$

$$\alpha^* = \arg \min_{\alpha \in \mathbb{R}^{N-1}} \sum_{i=1}^{n_b} \left\| \zeta_{N,i}^s - \sum_{j=1}^{N-1} \alpha_j \zeta_{j,i} \right\|_Y^2 ;$$

$$\hat{\mathrm{b}} = \hat{\zeta}_{N+1}^s - \sum_{j=1}^{N-1} \alpha_j^* \hat{\zeta}_j ;$$

$$\hat{\zeta}_N = \hat{\mathrm{b}} \left(\frac{1}{n_b} \sum_{i=1}^{n_b} \|b_i\|_Y^2 \right)^{-1/2} ;$$

$$W_N^V = W_{N-1}^V + \text{span}\ \{\hat{\zeta}_N\};$$

end.

Figure 2-10: The pseudo-orthogonalization procedure for W_N^V.

augmented reduced basis space, W_N^A.

These two preprocessing steps will lead to smaller N and better stability in the solution method. We now determine the efficiency of W_N^V by examining the projection error $\varepsilon_{N,n_b}^{\text{proj}}$ given by (2.27). The best projection of $\hat{\mathbf{u}}(\mu)$ onto W_N^V is given by $\hat{\mathbf{u}}_p(\mu) = \sum_{j=1}^N \beta_j^*(\mu) \hat{\zeta}_j$ where

$$\beta^*(\mu) = \arg \min_{\beta \in \mathbb{R}^N} \sum_{i=1}^{n_b} \left\| \sum_{j=1}^{N} \beta_j \zeta_{i,j} - u_i(\mu) \right\|_Y^2 , \tag{2.44}$$

where $\beta^*(\mu) \in \mathbb{R}^N$. To determine $\beta^*(\mu)$ for a given $\hat{\mathbf{u}}(\mu)$, we solve the following algebraic equation:

$$\sum_{i=1}^{n_b} \sum_{n=1}^{N} (\zeta_{m,i}, \zeta_{n,i})_Y \beta_n^*(\mu) = \sum_{j=1}^{n_b} (\zeta_{m,j}, u_j(\mu))_Y, \quad 1 \le m \le N. \tag{2.45}$$

We observe that the projection error $\varepsilon_{N,n_b}^{\text{proj}}$ is rapidly convergent as shown in Figure 2-11. When we compare this result to Figure 2-6, it suggests that the vectorial reduced basis space can be more efficient than the augmented reduced basis space. For example, for $n_b = 10$ and an accuracy of $\varepsilon_{N,n_b}^{\text{proj}} \le 1\text{E}{-5}$, the required dimension of the vectorial reduced basis space is 13 while the required dimension of the augmented reduced basis space is 40, as shown in Table 2.1. Thus, the required dimension of the augmented reduced basis space is more than double that required by the vectorial

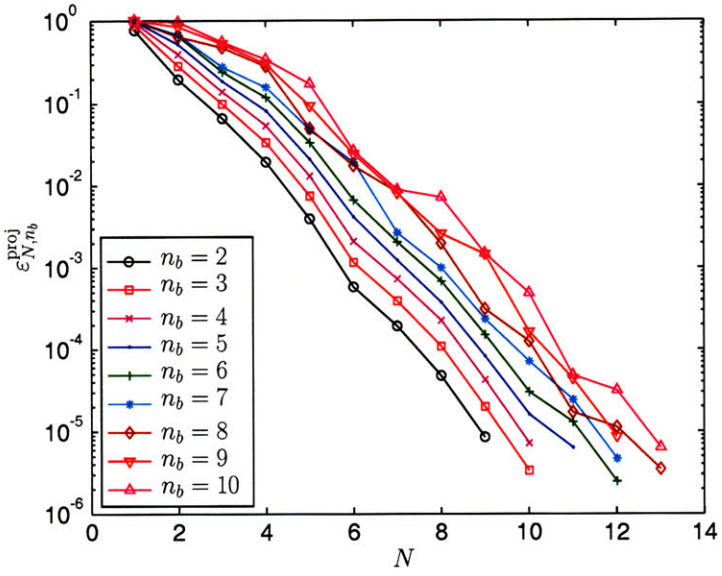

Figure 2-11: Convergence of the projection error of $\hat{\mathbf{u}}$ onto W_N^V, $\varepsilon^{\text{proj}}_{N,n_b}$ (given by (2.27)), with N for $2 \leq n_b \leq 10$.

reduced basis space. This difference in the dimension is particularly significant in the large n_b limit, and can play a role in deciding the appropriate reduced basis space to use.

The approximation

Reduced basis approximations to $\hat{\mathbf{u}}(\mu)$ and $\hat{\boldsymbol{\lambda}}(\mu)$ are obtained by first solving the following equations: given a μ, find $(\hat{\mathbf{u}}_N(\mu) \equiv (u_{N,1}(\mu), \ldots, u_{N,n_b}(\mu)), \hat{\boldsymbol{\lambda}}_N^L \equiv (\lambda_{N,1}^L(\mu), \ldots, \lambda_{N,n_b}^L(\mu))) \in W_N^V \times \mathbb{R}^{n_b}$ such that

$$\begin{aligned} \sum_{i=1}^{n_b} a(u_{N,i}(\mu), v_i; \mu) &= \sum_{j=1}^{n_b} \lambda_{N,j}^L(\mu) m(u_{N,j}(\mu), v_j), \quad \forall \hat{\mathbf{v}} \equiv (v_1, \ldots, v_{n_b}) \in W_N^V, \\ m(u_{N,i}(\mu), u_{N,i}(\mu)) &= 1, \quad 1 \leq i \leq n_b. \end{aligned} \tag{2.46}$$

We then compute $\hat{\boldsymbol{\lambda}}_N(\mu)$, the reduced basis approximation to $\hat{\boldsymbol{\lambda}}(\mu)$, from the Rayleigh Quotient given by

$$\lambda_{N,n}(\mu) = \sum_{i=1}^{N} \sum_{j=1}^{N} u_{N,i}(\mu) \left(\sum_{q=1}^{Q} \Theta_q(\mu) A_{i,j}^{N,q} \right) u_{N,j}(\mu), \quad 1 \leq n \leq n_b. \tag{2.47}$$

Note that in (2.46), we have only imposed the constraints $\int_\Omega u_{N,i}^2(\mu) = 1$, $1 \leq i \leq n_b$. We remind that in the original problem (2.1), the constraints are $\int_\Omega u_i(\mu) u_j(\mu) = \delta_{ij}, 1 \leq i < j \leq n_b$.

Thus, $\hat{\boldsymbol{\lambda}}_N^L$ is the set of Lagrange multipliers for the reduced set of constraints while $\hat{\boldsymbol{\lambda}}_N$ is the set of Lagrange multipliers corresponding to the full set of constraints associated with (2.1). As such, $\hat{\boldsymbol{\lambda}}_N \neq \hat{\boldsymbol{\lambda}}_N^L$, and must be computed from (2.47).

Hypothesis 2.1. *For sufficiently large $N \ll \mathcal{N}$, $\hat{\mathbf{u}}_N(\mu) \equiv (u_{N,1}(\mu), \ldots, u_{N,n_b}(\mu))$ that satisfies (2.46) also approximately satisfies the orthogonality constraints, i.e.*

$$m(u_{N,i}(\mu), u_{N,j}(\mu)) \approx 0, \quad 1 \leq i < j \leq n_b. \tag{2.48}$$

Hypothesis 2.1 thus assumes that the orthogonality of the components in $\hat{\mathbf{u}}_N$ will be approximately satisfied by construction (implicit to our space W_N^V) in (2.46). At present, we cannot prove that solutions $\hat{\mathbf{u}}_N(\mu)$ that satisfy (2.46) will always be good approximations of $\hat{\mathbf{u}}(\mu)$. However, we can demonstrate that if the approximation error in $\hat{\mathbf{u}}_N(\mu)$ is small, Hypothesis 2.1 holds — we shall show that $\int_\Omega u_{N,i}(\mu)\, u_{N,j}(\mu)$ is always bounded by the approximation error in $\hat{\mathbf{u}}_N(\mu)$ for $1 \leq i < j \leq n_b$.

Proposition 2.1. *Let $(w,v)_{L^2} = m(w,v) = \int_\Omega wv$ and $\|\cdot\|_{L^2} = \sqrt{(\cdot\,,\cdot)_{L^2}}$; then for $i \neq j$*

$$(u_{N,i}(\mu), u_{N,j}(\mu))_{L^2} \leq \sum_{n=1}^{n_b} \|u_n(\mu) - u_{N,n}(\mu)\|_{L^2}. \tag{2.49}$$

Proof. From the definition of $(\cdot\,,\cdot)_{L^2}$, we have

$$\begin{aligned}
(u_{N,i}(\mu), u_{N,j}(\mu))_{L^2} &= (u_{N,i}(\mu) - u_i(\mu), u_{N,j}(\mu))_{L^2} + (u_i(\mu), u_{N,j}(\mu) - u_j(\mu))_{L^2} \\
&\quad + (u_i(\mu), u_j(\mu))_{L^2} \\
&\leq \|u_{N,i}(\mu) - u_i(\mu)\|_{L^2} + \|u_{N,j}(\mu) - u_j(\mu)\|_{L^2},
\end{aligned} \tag{2.50}$$

since $\|u_i(\mu)\|_{L^2} = \|u_{N,j}(\mu)\|_{L^2} = 1$ and $(u_i(\mu), u_j(\mu))_{L^2} = 0$, for all $1 \leq i < j \leq n_b$. Then, (2.49) follows. □

From above, we can conclude that as $\hat{\mathbf{u}}_N \to \hat{\mathbf{u}}$, $\int_\Omega u_{N,i}\, u_{N,j} \to 0$, for $1 \leq i < j \leq n_b$. However, note that for N sufficiently large, we can represent any member of our (finite-dimensional) truth approximation space, presuming the linear independence of the snapshots. But, clearly

(2.46) is not equivalent to (2.1) due to the absence of orthogonality constraints. In the limit that $N = \mathcal{N}$, we see that (2.46) will give $(u_1(\mu), \lambda_1(\mu))$ for all n_b solutions we seek [65]. Thus, to be consistent and perhaps more accurate, we should systematically add in orthogonality constraints as N increases so that (2.46) approaches (2.1) — we introduce the following problem: find $(\hat{\mathbf{u}}_N(\mu) \equiv (u_{N,1}(\mu), \ldots, u_{N,n_b}(\mu)), \hat{\boldsymbol{\lambda}}^L \equiv (\lambda^L_{N,1}(\mu), \ldots, \lambda^L_{N,n_b}(\mu))) \in W^V_N \times \mathbb{R}^{n_b}$ such that

$$\begin{aligned}
\sum_{i=1}^{n_b} a(u_{N,i}(\mu), v_i; \mu) &= \sum_{i'=1}^{n_b} \lambda^L_{N,i'}(\mu) m(u_{N,i'}(\mu), v_{i'}), \quad \forall \hat{\mathbf{v}} \equiv (v_1, \ldots, v_{n_b}) \in W^V_N, \\
m(u_{N,i}(\mu), u_{N,i}(\mu)) &= 1, \quad 1 \le i \le n_b, \\
m(u_{N,i}(\mu), u_{N,j}(\mu)) &= 0, \quad (i,j) \in I_o,
\end{aligned} \tag{2.51}$$

where I_o consists of n_o pairs of indices (i,j) that denote the orthogonality constraints $m(u_{N,i}(\mu), u_{N,j}(\mu)) = 0$ we choose to enforce. We have not determined how I_o should best be selected. For our current example, the order by which the orthogonality constraints are included is determined by the difference between the two indices i and j — constraints with the smallest $|i-j|$ are included first.

Since (2.46) and (2.51) are both constrained optimization problem, there must be sufficient degree of freedom to obtain a good solution — N must thus be greater than the number of constraints we impose. The minimum number of constraints we impose is n_b, corresponding to the n_b normality constraints. Thus, $N > n_b$ to get any meaningful results. For (2.51), the number of orthogonality constraints we can add is constrained by N and total number of orthogonality constraints in (2.1). The maximum number of orthogonality constraints must then be less than $\min(N - n_b, \frac{1}{2} n_b(n_b - 1))$.

We shall further compare (2.46) and (2.51) in a later section numerically.

Discrete equation

We expand our reduced-basis approximation as

$$\hat{\mathbf{u}}_N(\mu) = \sum_{j=1}^{N} u_{N\,j}(\mu) \hat{\boldsymbol{\zeta}}_j. \tag{2.52}$$

Inserting this representation into (2.46) yields

$$\begin{aligned}\left(\sum_{q=1}^{Q}\sum_{j=1}^{N}\Theta_q(\mu)A_{i,j}^{N,q}\right)u_{N\,j}(\mu) &= \sum_{n=1}^{n_b}\lambda_{N,n}^{L}(\mu)M_{i,j}^{N,n,n}u_{N\,j}(\mu), \quad 1\le i\le N,\\ \sum_{i=1}^{N}\sum_{j=1}^{N}u_{N\,i}(\mu)M_{i,j}^{N,n,n}u_{N\,j}(\mu) &= 1, \quad 1\le n\le n_b, \end{aligned} \tag{2.53}$$

where Q and $\Theta_q(\mu)$ are similar to that defined in Section 2.3.4: $Q=2$, $\Theta_1(\mu)=1$, and $\Theta_2(\mu)=\mu^2$. But, $A^{N,1}\in\mathbb{R}^{N\times N}$, $A^{N,2}\in\mathbb{R}^{N\times N}$, and $M^{N,n,n'}\in\mathbb{R}^{N\times N}$, $1\le n\le n'\le n_b$ are given by $A_{i,j}^{N,1}=\sum_{n=1}^{n_b}a_1(\zeta_{n,j},\zeta_{n,i}), 1\le i,j\le N$, $A_{i,j}^{N,2}=\sum_{n=1}^{n_b}a_2(\zeta_{n,j},\zeta_{n,i}), 1\le i,j\le N$, and $M_{i,j}^{N,n,n'}=m(\zeta_{n,j},\zeta_{n',i}), 1\le i,j\le N$, respectively. The discrete equations for (2.51) can be obtained analogously.

It is clear that the above discrete equations admit the offline-online computational decomposition detailed in Section 2.3.4. We can first precompute $A^{N,1}$, $A^{N,2}$ and $M^{N,n,n'}$, $1\le n\le n'\le n_b$ in the offline stage at a cost dependent on $\mathcal{N}$. Then, during the online stage, we reconstruct our reduced basis matrices at a cost of $O(QN^2)$ and solve the resulting discrete system at a cost of $O(N^3)$, independent of $\mathcal{N}$.

To solve (2.53) however, diagonalization cannot be used. The system of discrete equations does not correspond to a typical algebraic eigenvalue problem: we are looking for a vector $\{u_{N\,i}(\mu)\}_{i=1}^{N}$ $\in\mathbb{R}^N$ but n_b $\lambda_{N,n}^{L}(\mu)$; and we only have normality constraints. Instead we use the Newton iterative scheme to solve (2.53): in each Newton iteration and given a current iterate $\bar{u}_{N\,j}(\mu), 1\le j\le N$, and $\bar{\lambda}_{N,n}^{L}(\mu)$, $1\le n\le n_b$, we must find an increment $\delta u_{N\,j}(\mu), 1\le j\le N$, and $\delta\lambda_{N,n}^{L}(\mu)$, $1\le n\le n_b$ such that

$$\begin{aligned}&\sum_{j=1}^{N}\left(\sum_{q=1}^{Q}A_{i,j}^{N,q}+\sum_{n=1}^{n_b}\bar{\lambda}_{N,n}^{L}(\mu)M_{i,j}^{N,n,n}\right)\delta u_{N\,j}(\mu)\\ &\qquad-\sum_{n'=1}^{n_b}\delta\lambda_{N,n'}^{L}(\mu)\sum_{k=1}^{N}M_{i,k}^{N,n',n'}\bar{u}_{N\,k}(\mu) =\\ &\qquad\qquad-\sum_{q=1}^{Q}\sum_{j=1}^{N}\Theta_q(\mu)A_{i,j}^{N,q}\bar{u}_{N\,j}(\mu)\\ &\qquad\qquad+\sum_{n=1}^{n_b}\bar{\lambda}_{N,n}^{L}(\mu)\sum_{j=1}^{N}M_{i,j}^{N,n,n}\bar{u}_{N\,j}(\mu), \quad 1\le i\le N\ ; \end{aligned} \tag{2.54}$$

and

$$2\sum_{i=1}^{N}\sum_{j=1}^{N}\bar{u}_{N\,i}(\mu)M_{i,j}^{N,n,n}\delta u_{N\,j}(\mu) = -\sum_{i=1}^{N}\sum_{j=1}^{N}\bar{u}_{N\,i}(\mu)M_{i,j}^{N,n,n}\bar{u}_{N\,j}(\mu) + 1, \quad 1 \le n \le n_b. \tag{2.55}$$

Convergence

We shall use the error measures defined by (2.29) and (2.30). In addition, we introduce the error measure, $\varepsilon_{N,n_b}^{\text{ortho}}$ given by

$$\varepsilon_{N,n_b}^{\text{ortho}} = \max_{\mu\in\Xi_T} \epsilon_{N,n_b}^{\text{ortho}}(\mu), \tag{2.56}$$

where

$$\epsilon_{N,n_b}^{\text{ortho}}(\mu) = \max_{1\le i<j\le n_b} \int_\Omega u_{N,i}(\mu)\, u_{N,j}(\mu). \tag{2.57}$$

The $\varepsilon_{N,n_b}^{\text{ortho}}$ measures how well orthogonality constraints are satisfied. As shown in Figure 2-12, the approximation is rapidly convergent. We only require $N = 14$ to reach an error of $\varepsilon_N^u < 1\,\text{E}{-5}$ and $\varepsilon_N^\lambda < 1\,\text{E}{-10}$ for $n_b = 10$. In addition, from Figure 2-13, we see that orthogonality constraints are also increasingly satisfied as N increases, *without* the imposition of these orthogonality constraints. This implies that the minimization procedure and our approximation space W_N^V have some intrinsic level of orthogonality "built-in". In addition, from (2.49), we obtain

$$\varepsilon_{N,n_b}^{\text{ortho}} \le \max_{\mu\in\Xi_T} \sum_{i=1}^{n_b} \|u_i(\mu) - u_{N,i}(\mu)\|_{L^2}. \tag{2.58}$$

This implies that as $\hat{\mathbf{u}}_N(\mu) \to \hat{\mathbf{u}}(\mu)$, $\varepsilon_{N,n_b}^{\text{ortho}} \to 0$ and since ε_{N,n_b}^u decreases with increasing N, so does $\varepsilon_{N,n_b}^{\text{ortho}}$.

We now examine (2.51), where we add n_0 orthogonality constraints to our problem. We first note that n_0 must be less than $N-n_b$. We also observe that the addition of orthogonality constraints significantly affects the convergence of the Newton's method; we usually cannot get a converged solution unless our initial solution is sufficiently close to a solution of (2.51) and N is sufficiently large. We thus proceed as follows: we first solve (2.46); the solution to (2.46) is then used as initial solution to solve (2.51). Table 2.3 compares the results for (2.46) and (2.51), where $\min(\frac{1}{2}n_b(n_b - 1), N - n_b - 1)$ orthogonality constraints are added; the orthogonality constraints we enforced are chosen in an ad-hoc manner currently. We note that for this comparison, the approximation based

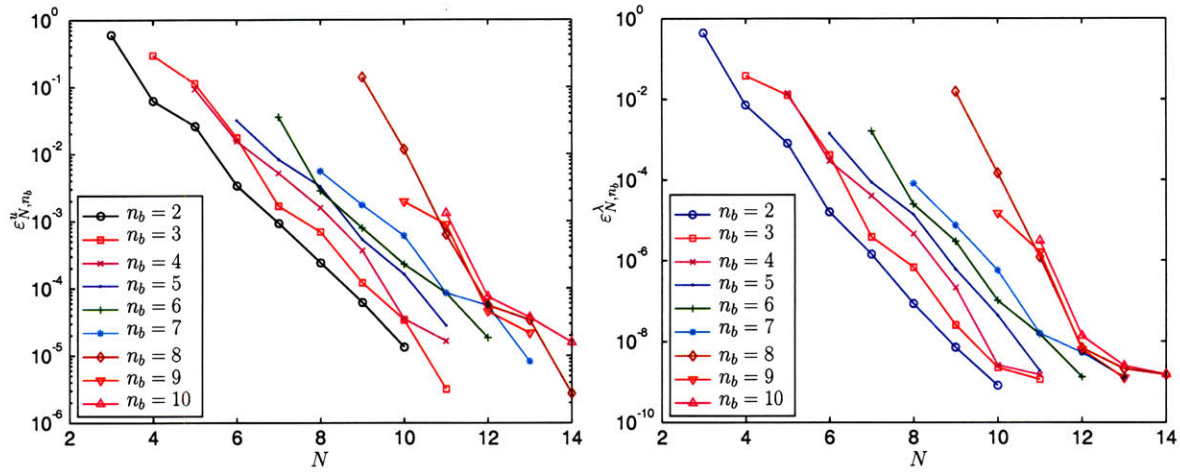

Figure 2-12: Convergence of reduced basis error of $\hat{\mathbf{u}}_N$, ε^u_{N,n_b} (given by (2.29)), and reduced basis error of $\hat{\boldsymbol{\lambda}}_N$, $\varepsilon^\lambda_{N,n_b}$ (given by (2.30)), with N for $2 \leq n_b \leq 10$ and $\hat{\mathbf{u}}_N \in W^V_N$.

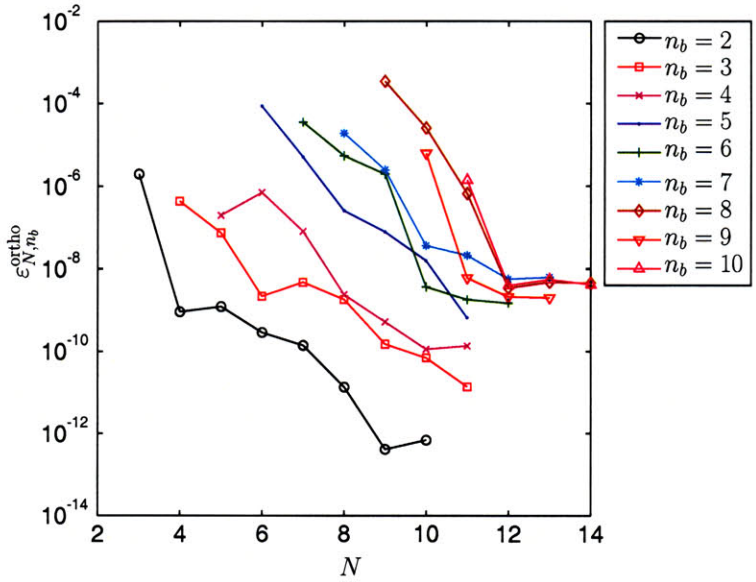

Figure 2-13: Convergence of the orthogonality error in $\hat{\mathbf{u}}_N$, $\varepsilon^{\text{ortho}}_{N,n_b}$ (given by (2.56)), with N for $2 \leq n_b \leq 10$ and $\hat{\mathbf{u}}_N \in W^V_N$.

		without orthogonality constraints			with n_o orthogonality constraints		
n_b	N	ε^u_{N,n_b}	$\varepsilon^\lambda_{N,n_b}$	$\varepsilon^{\text{ortho}}_{N,n_b}$	ε^u_{N,n_b}	$\varepsilon^\lambda_{N,n_b}$	$\varepsilon^{\text{ortho}}_{N,n_b}$
2	10	1.3413E-5	8.1434E-10	6.8842E-13	6.3422E-6	3.1233E-10	3.7683E-14
3	11	3.2212E-6	1.1695E-9	1.3791E-11	4.7347E-6	5.3670E-10	6.1706E-12
4	11	1.6699E-5	1.5185E-9	1.3757E-10	5.7550E-6	7.4413E-10	1.9507E-11
5	11	2.7995E-5	1.8848E-9	6.4663E-10	5.5269E-6	7.5138E-10	2.2638E-10
6	12	1.8356E-5	1.3228E-9	1.4639E-9	3.5315E-6	5.8200E-10	4.4746E-10
7	13	8.1863E-6	1.3172E-9	6.1119E-9	1.0366E-5	6.8658E-10	6.0989E-10
8	13	2.7685E-6	1.5352E-9	4.5632E-9	2.8111E-6	5.5693E-10	4.8658E-10
9	14	2.1823E-5	1.2614E-9	2.0083E-9	3.2631E-6	6.6143E-10	5.4145E-10
10	14	1.5930E-5	1.5533E-9	4.2163E-9	2.0259E-6	6.0658E-10	4.9733E-10

Table 2.3: Comparing the additional iteration with orthogonality constraints. The number of orthogonality constraints, n_o, is given by $\min(\frac{n_b}{2}(n_b-1), N-n_b-1)$; the condition $N-n_b-1$ is set to allow at least one degree of freedom in the optimization procedure.

on (2.46) are already very good. The addition of orthogonality constraints led to further uniform decrease in $\varepsilon^\lambda_{N,n_b}$ and $\varepsilon^{\text{ortho}}_{N,n_b}$ for all n_b, although the change in ε^u_{N,n_b} is less uniform. For the case $n_b = 2$ and 3 where all orthogonality constraints are added, $\varepsilon^{\text{ortho}}_{N,n_b}$ are very small. This shows that the systematic inclusion of orthogonality constraints can improve our approximations. A better solution method in the online stage may perhaps alleviate the limitations described above.

2.3.6 Construction of Samples

So far, we have not mentioned how the nested reduced basis sample sets S^A_N or S^V_N are chosen. A sample set must be well-chosen in order to obtain a rapidly convergent reduced basis approximation, and a well-conditioned reduced basis discrete system. In particular, we seek a sampling procedure that ensures "maximally independent" snapshots. We shall use the "greedy" adaptive sampling procedure outlined in [82, 99, 115].

We now describe the construction of S^V_N (and thus W^V_N) based on the "greedy" sampling procedure — S^A_N (and W^A_N) can be constructed in a similar manner. We first assume that we are given a sample S^V_N and hence reduced-basis space W^V_N, and the associated reduced-basis approximation (procedure to determine) $\hat{\mathbf{u}}_N(\mu)$ and $\hat{\boldsymbol{\lambda}}_N(\mu)$, $\forall \mu \in \mathcal{D}$. Then, for a suitably fine grid Ξ_T over the parameter space $\mathcal{D}$, we determine $\mu^*_{N+1} = \arg\max_{\mu\in\Xi_T} \epsilon^\bullet_N(\mu)$, where $\epsilon^\bullet_N(\mu)$ is an error measure of the approximation based on W^V_N. Then we append μ^*_{N+1} to S^V_N to form S^V_{N+1} and hence W^V_{N+1}. The procedure is repeated until $\varepsilon_{\max} = \epsilon^\bullet_N(\mu^*_{N+1})$ is below ε_{tol}, a tolerance we desire. This tolerance ε_{tol} determines the size of $N_{\max}$. Figure 2-14 summarizes the "greedy" sampling procedure.

Given S_1^V, W_1^V;

Repeat $N = 2, \ldots$

$$\mu_N^* = \arg\max_{\mu \in \Xi_T} \epsilon_{N-1}^\bullet(\mu);$$
$$\varepsilon_{\max} = \epsilon_N^\bullet(\mu_N^*);$$
$$S_N^V = S_{N-1}^V \cup \mu_N^*;$$
$$W_N^V = W_{N-1}^V + \text{span}\ \{\hat{\mathbf{u}}_N(\mu_N^*)\};$$

until $\varepsilon_{\max} \leq \varepsilon_{\text{tol}}$.

Figure 2-14: The "greedy" sampling procedure to construct an optimal S_N^V.

We may define $\epsilon_N^\bullet(\mu)$ in several ways. For example, we may use error measures listed in Section 2.3.3 — the projection error $\epsilon_{N,n_b}^{\text{proj}}(\mu)$ given by (2.28), the reduced basis approximation error in $\hat{\mathbf{u}}$, $\epsilon_{N,n_b}^u(\mu)$ given by (2.31), or the reduced basis approximation error in $\hat{\boldsymbol{\lambda}}$, $\epsilon_{N,n_b}^\lambda(\mu)$ given by (2.32). However, evaluations of $\epsilon_{N,n_b}^{\text{proj}}(\mu)$ and $\epsilon_{N,n_b}^u(\mu)$ are in fact expensive since the computational cost is of order $O(\mathcal{N})$. In addition, "truth" solutions must be evaluated for all $\mu \in \Xi_T$. However, if *a posteriori* error estimator is available, a more efficient procedure is possible [82, 99]. This will be elaborated in Section 2.4.

2.3.7 Comparison of the Reduced Basis Spaces

We shall make a comparison between the augmented reduced basis approximation and the vectorial reduced basis approximation from three aspects: (*i*) the dimension of the spaces; (*ii*) the computational complexity of the online stage; and (*iii*) the computational complexity of the offline stage.

From Table 2.4, it is clear that the dimension of reduced basis spaces, N, increases with n_b — but both scales less than linearly with n_b. However, the nonlinear effect is more significant in W_N^V compared to W_N^A for this numerical example. We consider the N required for $\varepsilon_{N,M}^\lambda \leq 2\text{E}{-}9$. For W_N^A, N scales as $N_s n_b$ where N_s decreases with n_b when higher accuracy is sought, as shown in Table 2.2. This is due to spectral effects resulting from inclusion of higher eigenvectors, as explained in Section 2.3.4. On the other hand, the dimension of W_N^V increases only very weakly with n_b — when n_b increases from 2 to 10, N required only increases by 4. This is because W_N^V more closely

n_b	N	
	W_N^A	W_N^V
2	18	10
3	27	11
4	24	11
5	30	11
6	30	12
7	35	13
8	32	14
9	36	13
10	40	14

Table 2.4: Comparison between the augmented reduced basis approximation and the vectorial reduced basis approximation based on N required to reduce $\varepsilon_{N,M}^{\lambda}$ to below $2\text{E}{-}9$ for $2 \leq n_b \leq 10$.

approximate the solution manifold $\mathcal{M} \equiv \{\hat{\mathbf{u}}(\mu), \mu \in \mathcal{D}\}$, resulting in a more efficient representation.

During the online stage, the vectorial reduced basis approximation requires the use of Newton iterative scheme. The current implementation of the scheme is not as efficient as the eigenvalue solver used in the augmented reduced basis approximation. As such, although the required dimension of W_N^V for a particular accuracy is smaller, the total computational cost can be higher, and as such less efficiency. There are certainly room for improvement in the design of the online solution method.

For the offline stage, construction of the space W_N^V is again less efficient than W_N^A. First, we need to perform N-solve in the former but only N_s-solve in the later, where $N_s < N$ in general. In addition, in more general case, the preprocessing steps required for W_N^V can be time-consuming.

Let us touch on the issue related to smoothness of the solution manifold $\mathcal{M}$ and the parametric derivatives of $\hat{\mathbf{u}}$ again. With the augmented reduced basis approximation, the approximations of u_i, $1 \leq i \leq n_b$ are independent of one another once W_N^A is defined. As such, the approximation of a particular u_i is not affected by the rate at which magnitudes of the parameteric derivatives (in the Y norm) of u_j, $j \neq i$ grow. On the other hand, for the vectorial reduced basis approximation, the approximations of u_i, $1 \leq i \leq n_b$ are coupled — therefore good approximation can only be obtained if magnitudes of the parametric derivatives of u_i, $1 \leq i \leq n_b$ grow at a similar rate. A vectorial reduced basis approximation is rapidly convergent if all components of $\hat{\mathbf{u}}(\mu)$ have similar smoothness property. We will not examine this issue in this thesis — for W_N^V, obtaining a set of basis functions that is $C^0(\mathcal{D})$ is an overriding issue at present.

2.4 *A Posteriori* Error Estimation

A posteriori error estimation procedures are well-developed for algebraic eigenvalue problems [4, 51, 89] and approximation of eigenvalue problems based on, say, finite element method [1, 83]. Simple error estimates for a computed eigenvalue can be determined from the residual vector. However, these error estimates usually do not provide rigorous bounds. Within the reduced basis context, asymptotic error bounds are first formulated for reduced basis approximation of symmetric positive definite eigenvalue problem in [71]. In addition, [71] provides a very efficient procedure by which these bounds can be computed through the offline-online computational framework.

In previous work on reduced basis approximation of partial differential equations [45, 75, 99, 82, 114, 115], significant emphasis is placed on obtaining inexpensive and sharp error bounds for our output of interest. Absent such rigorous error bounds, we cannot provide a certificate for our reduced basis approximation and must rely on prior calculations to justify the accuracy of a reduced basis approximation.

There are no existing rigorous error bounds for reduced basis approximation of eigenvalue problems. However, non-rigorous error bounds can still be very useful. In particular, they play an important role in the "greedy" adaptive sampling procedure outlined in Section 2.3.6. They provide an efficient alternative to computing the actual errors ε^u_{N,n_b} and $\varepsilon^\lambda_{N,n_b}$, which requires determination the eigensolutions at all sample points in the training sample set Ξ_T. For this particular purpose, an asymptotic error bound may be sufficient since its main purpose is to serve as a guide in the construction of the reduced basis sample set, and not as an certification of the result. The goal of this section is to construct asymptotic *a posteriori* error bounds for reduced basis approximation of n_b eigensolutions to linear eigenvalue problems, specialized to the current numerical example. The development parallels that for algebraic eigenvalue problems.

2.4.1 Derivation

For $i = 1, \ldots, n_b$, we define the residual as

$$R_i(v;\mu) = a(u_{N,i}(\mu), v;\mu) - \lambda_{N,i}(\mu) m(u_{N,i}(\mu), v), \tag{2.59}$$

for $\forall v \in Y$. We further define a reconstructed error $\hat{e}_i$ in Y, such that

$$\hat{a}(\hat{e}_i, v) = R_i(v;\mu), \quad \forall v \in Y, \tag{2.60}$$

where

$$\hat{a}(w, v) = a_1(w, v) + a_2(w, v); \tag{2.61}$$

$$\|R_i(\,\cdot\,;\mu)\| \equiv \sup_{v\in Y} \frac{R_i(v;\mu)}{\hat{a}(v,v)^{1/2}} = \hat{a}(\hat{e}_i, \hat{e}_i)^{1/2}; \tag{2.62}$$

and $\|\,\cdot\,\| = \hat{a}(\,\cdot\,,\,\cdot\,)^{1/2}$.

Hypothesis 2.2. *Assuming our reduced-basis approximation is convergent in the sense that*

$$\lambda_{N,i}(\mu) \to \lambda_i(\mu), \quad 1 \le i \le n_b, \quad as \quad N \to \infty. \tag{2.63}$$

Then, for sufficiently large N,

$$i = \arg \min_{1\le j\le \mathcal{N}} \left|1 - \frac{\lambda_{N,i}(\mu)}{\lambda_j(\mu)}\right|. \tag{2.64}$$

Proposition 2.2. *Assume our reduced-basis approximation is convergent in the sense that*

$$\lambda_{N,i}(\mu) \to \lambda_i(\mu), \quad 1 \le i \le n_b, \quad as \quad N \to \infty. \tag{2.65}$$

Then, for sufficiently large N,

$$\left|\frac{\lambda_i(\mu) - \lambda_{N,i}(\mu)}{\lambda_i(\mu)}\right| \le \frac{\|R_i(\,\cdot\,;\mu)\|}{(\lambda_{N\,i}(\mu))^{1/2}}, \quad 1 \le i \le n_b. \tag{2.66}$$

In addition,

$$\|u_{N,i}(\mu) - u_i(\mu)\| \le \frac{\|R_i(\,\cdot\,;\mu)\|}{d_i}, \quad 1 \le i \le n_b. \tag{2.67}$$

and

$$|\lambda_{N,i}(\mu) - \lambda_i(\mu)| \le \frac{\|R_i(\,\cdot\,;\mu)\|^2}{d_i^2}, \quad 1 \le i \le n_b. \tag{2.68}$$

where $d_i = \min_{j\neq i} \left|\frac{\lambda_{N,j}(\mu)-\lambda_{N,i}(\mu)}{\lambda_{N,j}(\mu)}\right|$.

Proof. For $i = 1, \ldots, n_b$, we define $\tilde{e}_i \in Y$ as

$$a(\tilde{e}_i, v; \mu) = R_i(v; \mu), \quad \forall v \in Y; \tag{2.69}$$

$$|||R_i(\,\cdot\,; \mu)||| \equiv \sup_{v \in Y} \frac{R_i(v; \mu)}{a(v, v; \mu)^{1/2}} = a(\tilde{e}_i, \tilde{e}_i; \mu)^{1/2}; \tag{2.70}$$

and $|||\,\cdot\,||| = a(\,\cdot\,,\,\cdot\,; \mu)^{1/2}$. We note that

$$\begin{aligned} a(v, v; \mu) &= a_1(v, v) + \mu^2 a_2(v, v) \\ &\leq a_1(v, v) + a_2(v, v); \end{aligned} \tag{2.71}$$

since $\mu \geq 1$. In addition, $a_1(v, v)$ and $a_2(v, v)$ are both symmetric positive definite. Therefore

$$0 \leq \hat{a}(v, v) \leq a(v, v; \mu), \tag{2.72}$$

and

$$|||R_i(\,\cdot\,; \mu)||| \leq \|R_i(\,\cdot\,; \mu)\|. \tag{2.73}$$

Let $u_{N,i}(\mu) = \sum_{j=1}^{\mathcal{N}} \alpha_j u_j(\mu)$ and $\tilde{e}_i = \sum_{j=1}^{\mathcal{N}} \beta_j u_j(\mu)$. From (2.69),

$$\begin{aligned} a\left(\sum_{j'=1}^{\mathcal{N}} \beta_{j'} u_{j'}(\mu), v; \mu\right) &= a\left(\sum_{j=1}^{\mathcal{N}} \alpha_j u_j(\mu), v; \mu\right) - \lambda_{N,i}(\mu) m\left(\sum_{j=1}^{\mathcal{N}} \alpha_j u_j(\mu), v\right) \\ \sum_{j'=1}^{\mathcal{N}} \beta_{j'} \lambda_{j'}(\mu) m(u_{j'}(\mu), v; \mu) &= \sum_{j=1}^{\mathcal{N}} \alpha_j (\lambda_j(\mu) - \lambda_{N,i}(\mu)) m(u_j(\mu), v; \mu) \\ \beta_j &= \alpha_j \left(\frac{\lambda_j(\mu) - \lambda_{N,i}(\mu)}{\lambda_j(\mu)}\right). \end{aligned} \tag{2.74}$$

Then,

$$\begin{aligned} |||R_i(\,\cdot\,; \mu)|||^2 &= a(\tilde{e}_i, \tilde{e}_i; \mu) \\ &= \sum_{j=1}^{\mathcal{N}} \beta_j^2 \lambda_j(\mu) m(u_j(\mu), u_j(\mu)) \\ &= \sum_{j=1}^{\mathcal{N}} \left(\frac{\lambda_j(\mu) - \lambda_{N,i}(\mu)}{\lambda_j(\mu)}\right)^2 \alpha_j^2 \lambda_j(\mu). \end{aligned} \tag{2.75}$$

Dividing by $\lambda_{N,i}(\mu)$, we have

$$\begin{aligned}
\frac{|||R_i(\cdot;\mu)|||^2}{\lambda_{N,i}(\mu)} &= \frac{1}{a(u_{N,i},u_{N,i};\mu)}\sum_{j=1}^{\mathcal{N}}\left(\frac{\lambda_j(\mu)-\lambda_{N,i}(\mu)}{\lambda_j(\mu)}\right)^2\alpha_j^2\lambda_j(\mu), \\
&\geq \min_{1\leq j\leq\mathcal{N}}\left(\frac{\lambda_j(\mu)-\lambda_{N,i}(\mu)}{\lambda_j(\mu)}\right)^2\frac{\sum_{j=1}^{\mathcal{N}}\alpha_j^2\lambda_j(\mu)}{\sum_{j'=1}^{\mathcal{N}}\alpha_{j'}^2\lambda_{j'}(\mu)} \\
&= \min_{1\leq j\leq\mathcal{N}}\left(\frac{\lambda_j(\mu)-\lambda_{N,i}(\mu)}{\lambda_j(\mu)}\right)^2 \qquad (2.76)
\end{aligned}$$

Based on Hypothesis 2.2, we have $i=\arg\min_j\left|\frac{\lambda_j(\mu)-\lambda_{N,i}(\mu)}{\lambda_j(\mu)}\right|$ and

$$\begin{aligned}
\left|\frac{\lambda_i(\mu)-\lambda_{N,i}(\mu)}{\lambda_i(\mu)}\right| &\leq \frac{|||R_i(\cdot;\mu)|||}{(\lambda_{N,i}(\mu))^{1/2}} \\
&\leq \frac{\|R_i(\cdot;\mu)\|}{(\lambda_{N,i}(\mu))^{1/2}}, \qquad (2.77)
\end{aligned}$$

from (2.73). This proves (2.66).

To prove (2.67), we first note that

$$u_{N,i}(\mu)-u_i(\mu)=\sum_{j\neq i}\alpha_j u_j(\mu)+(\alpha_i-1)u_i(\mu), \qquad (2.78)$$

which leads to

$$\begin{aligned}
|||u_{N,i}(\mu)-u_i(\mu)|||^2 &= a\left(\sum_{j\neq i}\alpha_j u_j(\mu)+(\alpha_i-1)u_i(\mu),\sum_{j\neq i}\alpha_j u_j(\mu)+(\alpha_i-1)u_i(\mu);\mu\right) \\
&= a\left(\sum_{j\neq i}\alpha_j u_j(\mu),\sum_{j\neq i}\alpha_j u_j(\mu);\mu\right)+a((\alpha_i-1)u_i(\mu),(\alpha_i-1)u_i(\mu);\mu\Big) \\
&\quad + a\left(\sum_{j\neq i}\alpha_j u_j(\mu),(\alpha_i-1)u_i(\mu);\mu\right)+a\left((\alpha_i-1)u_i(\mu),\sum_{j\neq i}\alpha_j u_j(\mu);\mu\right) \\
&= \sum_{j\neq i}\alpha_j^2\lambda_j(\mu)+(\alpha_i-1)^2\lambda_i(\mu). \qquad (2.79)
\end{aligned}$$

In addition, from (2.78),

$$\begin{aligned}\|u_{N,i}(\mu)-u_i(\mu)\|_{L^2}^2 &= m(u_{N,i}(\mu)-u_i(\mu),u_{N,i}(\mu)-u_i(\mu)) \\ &= m\left(\sum_{j\neq i}\alpha_j u_j(\mu)+(\alpha_i-1)u_i(\mu),\sum_{j\neq i}\alpha_j u_j(\mu)+(\alpha_i-1)u_i(\mu)\right) \\ &= 2(1-\alpha_i); \end{aligned} \tag{2.80}$$

and from Poincaré-Friedrichs inequality [87], we have

$$\begin{aligned}|||u_{N,i}(\mu)-u_i(\mu)|||^2 &\geq C\|u_{N,i}(\mu)-u_i(\mu)\|_{L^2}^2 \\ &= 2C(1-\alpha_i), \end{aligned} \tag{2.81}$$

where C is a constant. From (2.75), we also have

$$\begin{aligned}|||R_i(\,\cdot\,;\mu)|||^2 &\geq \sum_{j\neq i}\alpha_j^2\left(\frac{\lambda_j(\mu)-\lambda_{N,i}(\mu)}{\lambda_j(\mu)}\right)^2\lambda_j(\mu) \\ &\geq \min_{j\neq i}\left(\frac{\lambda_j(\mu)-\lambda_{N\,i}(\mu)}{\lambda_j(\mu)}\right)^2\sum_{j\neq i}\alpha_j^2\lambda_j(\mu). \end{aligned} \tag{2.82}$$

Let $\tilde{d}_i \equiv \min_{j\neq i}\left|\frac{\lambda_j(\mu)-\lambda_{N,i}(\mu)}{\lambda_j(\mu)}\right|$. Then

$$|||u_{N,i}(\mu)-u_i(\mu)|||^2-\frac{\lambda_i(\mu)}{4C^2}|||u_{N,i}(\mu)-u_i(\mu)|||^4 \leq \frac{|||R_i(\,\cdot\,;\mu)|||^2}{\tilde{d}_i^2}. \tag{2.83}$$

By solving for $|||u_{N,i}(\mu)-u_i(\mu)|||^2$ and expanding the square root term, we obtain

$$\begin{aligned}|||u_{N,i}(\mu)-u_i(\mu)|||^2 &\leq \frac{|||R_i(\,\cdot\,;\mu)|||^2}{\tilde{d}_i^2} \\ &\leq \frac{\|R_i(\,\cdot\,;\mu)\|^2}{\tilde{d}_i^2}, \end{aligned} \tag{2.84}$$

based on (2.73), after ignoring the higher-order term involving $|||\,\cdot\,|||^4$. Finally, in the asymptotic limit of (2.65), we can approximate $\tilde{d}_i$ by

$$\tilde{d}_i \approx \min_{j\neq i}\left|\frac{\lambda_{N,j}(\mu)-\lambda_{N,i}(\mu)}{\lambda_{N,j}(\mu)}\right| \equiv d_i. \tag{2.85}$$

This proves (2.67).

To prove (2.68), we note that

$$\begin{aligned}\lambda_{N,i}(\mu) - \lambda_i(\mu) &= a(u_{N,i}(\mu), u_{N,i}(\mu); \mu) - a(u_i(\mu), u_i(\mu); \mu) \\ &= \sum_{j=1}^{\mathcal{N}} \alpha_j^2 \lambda_j(\mu) - \lambda_i(\mu) \\ &= \sum_{j \neq i} \alpha_j^2 \lambda_j(\mu) - (1 - \alpha_i^2)\lambda_i(\mu). \end{aligned} \tag{2.86}$$

Substituting (2.79) into (2.86), we get

$$\begin{aligned}\lambda_{N,i}(\mu) - \lambda_i(\mu) &= |||u_{N,i}(\mu) - u_i(\mu)|||^2 - (\alpha_i - 1)^2 \lambda_i(\mu) - (1 - \alpha_i^2)\lambda_i(\mu) \\ &\leq |||u_{N,i}(\mu) - u_i(\mu)|||^2; \end{aligned} \tag{2.87}$$

since $1 - \alpha_i^2 = \sum_{j \neq i} \alpha_j^2 \geq 0$ and $\lambda_i > 0$. From (2.84), this proves (2.68). □

2.4.2 Offline-online Computational Framework

Augmented reduced basis approximation

We can also construct very efficient offline-online computational strategies for the evaluation of our error estimators. From (2.60) and our reduced basis approximation, we have

$$\hat{a}(\hat{e}_i, v) = \sum_{q=1}^{Q} \Theta_q(\mu) a_q(u_{N,i}(\mu), v) - \lambda_{N,i}(\mu) m(u_{N,i}(\mu), v), \quad v \in Y, \quad 1 \leq i \leq n_b. \tag{2.88}$$

It then follows from linear superposition that

$$\hat{e}_i(\mu) = \sum_{q=1}^{Q} \sum_{n=1}^{N} \Theta_q(\mu) u_{N,i\,n}(\mu) \xi_n^q - \lambda_{N,i}(\mu) \sum_{n=1}^{N} u_{N,i\,n}(\mu) \xi_n^0, \tag{2.89}$$

where

$$\hat{a}(\xi_n^q, v) = a_q(\zeta_n, v), \quad v \in Y, \quad 1 \leq n \leq N, \quad 1 \leq q \leq Q, \tag{2.90}$$

$$\hat{a}(\xi_n^0, v) = m(\zeta_n, v), \quad v \in Y, \quad 1 \leq n \leq N. \tag{2.91}$$

Then, $\|R_i(\,\cdot\,;\mu)\|$ is given by

$$\begin{aligned}
\|R_i(\,\cdot\,;\mu)\|^2 &= \hat{a}(\hat{e}_i,\hat{e}_i) \\
&= \sum_{n=1}^{N}\sum_{n'=1}^{N}\sum_{q=1}^{Q}\sum_{q'=1}^{Q} u_{N,i\,n}(\mu)u_{N,i\,n'}(\mu)\Theta_q(\mu)\Theta_{q'}(\mu)\hat{A}^{q,q'}_{n,n'} \\
&\quad + \sum_{n=1}^{N}\sum_{n'=1}^{N} \lambda^2_{N,i}(\mu)u_{N,i\,n}(\mu)u_{N,i\,n'}(\mu)\hat{A}^{0,0}_{n,n'} \\
&\quad + \sum_{n=1}^{N}\sum_{n'=1}^{N}\sum_{q=1}^{Q} u_{N,i\,n}(\mu)\lambda_{N,i}(\mu)\Theta_q(\mu)\hat{A}^{q,0}_{n,n'}; \qquad (2.92)
\end{aligned}$$

where $\hat{A}^{q,q'} \in \mathbb{R}^N \times \mathbb{R}^N$, $0 \le q,q' \le Q$ are given by $\hat{A}^{q,q'}_{n,n'} = \hat{a}(\xi^q_n,\xi^{q'}_{n'})$, $0 \le q,q' \le Q$, $1 \le n,n' \le N$. We now see that the dual norm of the residual is the sum of products of parameter-dependent functions and parameter-independent functionals. The offline-online decomposition is now clear.

In the offline stage, we compute ξ^q_n, $0 \le q \le Q$, $1 \le n \le N$, based on (2.88) at the cost of $O((Q+1)N\mathcal{N}^\bullet)$, where the $\bullet$ denotes computational complexity of the linear solver used to obtain ξ^q_n. We then evaluate $\hat{A}^q$ and $\hat{M}$ at the cost of $O((Q+1)N^2\mathcal{N}^2)$. We store the matrices $\hat{A}^q$ and $\hat{M}$ at a total cost of $(Q+1)N^2$.

In the online stage, we simply evaluate the sum (2.89) for a given $u_{N,i}(\mu)$ and $\lambda_{N,i}(\mu)$, $1 \le i \le n_b$. The operation count is only $O(n_bQ^2N^2)$. The online complexity is thus independent of $\mathcal{N}$. Unless Q is large, the online cost to compute the error estimator is then a fraction of the cost required to obtain $u_{N,i}(\mu)$ and $\lambda_{N,i}(\mu)$.

Vectorial reduced basis approximation

Our point of departure is again (2.88). However, $\hat{e}_i(\mu)$ is defined as

$$\hat{e}_i(\mu) = \sum_{q=1}^{Q}\sum_{n=1}^{N}\Theta_q(\mu)u_{N\,n}(\mu)\xi^q_{i\,n} - \lambda_{N,i}(\mu)\sum_{n=1}^{N}u_{N\,n}(\mu)\xi^0_{i\,n}, \qquad (2.93)$$

where

$$\hat{a}(\xi^q_{i\,n},v) = a_q(\zeta_{i\,n},v), \quad v \in Y, \quad 1 \le i \le n_b, \quad 1 \le n \le N, \quad 1 \le q \le Q, \qquad (2.94)$$

$$\hat{a}(\xi^0_{i\,n},v) = m(\zeta_{i\,n},v), \quad v \in Y, \quad 1 \le i \le n_b, \quad 1 \le n \le N. \qquad (2.95)$$

Then, $\|R_i(\,\cdot\,;\mu)\|$ is given by

$$\begin{aligned}
\|R_i(\,\cdot\,;\mu)\|^2 &= \hat{a}(\hat{e}_i,\hat{e}_i) \\
&= \sum_{n=1}^{N}\sum_{n'=1}^{N}\sum_{q=1}^{Q}\sum_{q'=1}^{Q} u_{N,i\,n}(\mu)u_{N,i\,n'}(\mu)\Theta_q(\mu)\Theta_{q'}(\mu)\hat{A}_{n,n'}^{q,q',i} \\
&\quad + \sum_{n=1}^{N}\sum_{n'=1}^{N} \lambda_{N,i}^2(\mu)u_{N,i\,n}(\mu)u_{N,i\,n'}(\mu)\hat{A}_{n,n'}^{0,0,i} \\
&\quad + \sum_{n=1}^{N}\sum_{n'=1}^{N}\sum_{q=1}^{Q} u_{N,i\,n}(\mu)\lambda_{N,i}(\mu)\Theta_q(\mu)\hat{A}_{n,n'}^{q,0,i}; \qquad (2.96)
\end{aligned}$$

where $\hat{A}^{q,q',i} \in \mathbb{R}^N \times \mathbb{R}^N$, $0 \le q,q' \le Q$, $1 \le i \le n_b$ are given by $\hat{A}_{n,n'}^{q,q',i} = \hat{a}(\xi_{i\,n}^q, \xi_{i\,n'}^q)$, $0 \le q,q' \le Q$, $1 \le n,n' \le N$. Again, we see that the dual norm of the residual is the sum of products of parameter-dependent functions and parameter-independent functionals. The offline-online decomposition then follows closely that of augmented reduced basis approximation.

In the offline stage, we compute $(\xi_i^q)_n$, $1 \le q \le Q$, $1 \le i \le n_b$, $1 \le n \le N$ and $(\xi_i^0)_n$ $1 \le i \le n_b$, $1 \le n \le N$ at the cost of $O((Q+1)n_bN\mathcal{N}^\bullet)$. We then evaluate $\hat{A}^{q,i}$ and $\hat{M}^i$ at the cost of $O((Q+1)n_bN^2\mathcal{N}^2)$. We then store the matrices $\hat{A}^{q,i}$ and $\hat{M}^i$.

In the online stage, we simply evaluate the sum (2.93) for a given $\hat{\mathbf{u}}_N$ and $\hat{\boldsymbol{\lambda}}_N$. The operation count is only $O(n_bQ^2N^2)$. The online complexity is thus independent of $\mathcal{N}$. Again, unless Q is large, the online cost to compute the error estimator is then a fraction of the costs required to obtain $\hat{\mathbf{u}}_N$ and $\hat{\boldsymbol{\lambda}}_N$.

2.4.3 Numerical Results

We define our error estimator $\Delta_{N,n_b}^u(\mu)$ and $\Delta_{N,n_b,\bullet}^\lambda(\mu)$ as

$$\Delta_{N,n_b}^u(\mu) = \left(\sum_{i=1}^{n_b}\frac{\|R_i(\,\cdot\,;\mu)\|^2}{d_i^2}\right)^{1/2}\left(\frac{1}{\sum_{j=1}^{n_b}\|u_{N,j}(\mu)\|_Y^2}\right)^{1/2}, \qquad (2.97)$$

$$\Delta_{N,n_b,1}^\lambda(\mu) = \max_{1\le i\le n_b}\frac{\|R_i(\,\cdot\,;\mu)\|}{(\lambda_{N,i}(\mu))^{1/2}}, \qquad (2.98)$$

$$\Delta_{N,n_b,2}^\lambda(\mu) = \max_{1\le i\le n_b}\frac{\|R_i(\,\cdot\,;\mu)\|^2}{d_i^2\lambda_{N,i}(\mu)}; \qquad (2.99)$$

and the following effectivity measures:

$$\eta^u_{N,n_b}(\mu) = \frac{\Delta^u_{N,n_b}(\mu)}{\epsilon^u_{N,n_b}(\mu)}, \tag{2.100}$$

$$\eta^\lambda_{N,n_b,1}(\mu) = \frac{\Delta^\lambda_{N,n_b,1}(\mu)}{\epsilon^\lambda_{N,n_b}(\mu)}, \tag{2.101}$$

$$\eta^\lambda_{N,n_b,2}(\mu) = \frac{\Delta^\lambda_{N,n_b,2}(\mu)}{\epsilon^\lambda_{N,n_b}(\mu)}. \tag{2.102}$$

We note that $\eta^\lambda_{N,n_b,1}(\mu)$ will diverge as N increases since $|\lambda_{N,i}(\mu) - \lambda_i(\mu)|$ is of $O(\|u_{N,i}(\mu) - u_i(\mu)\|^2_Y) \approx O(\|R_i(\cdot;\mu)\|^2))$ and $\Delta^\lambda_{N,n_b,1}(\mu)$ is of $O(\|R_i(\cdot;\mu)\|)$. It is also obvious from the fact that since both $\Delta^\lambda_{N,n_b,1}(\mu)$ and $\Delta^\lambda_{N,n_b,2}(\mu)$ are bounds and $\Delta^\lambda_{N,n_b,2}(\mu)$ is approximately $(\Delta^\lambda_{N,n_b,1}(\mu))^2$, $\eta^\lambda_{N,n_b,1}(\mu)$ will diverge. We further note that $\eta^u_{N,n_b}(\mu)$ and $\eta^\lambda_{N,n_b,2}(\mu)$ deviate from the actual effectivities by a factor of $(\sum_{i=1}^{n_b} \|u_{N,i}(\mu)\|^2_Y / \sum_{i=1}^{n_b} \|u_i(\mu)\|^2_Y)^{1/2}$ and $\lambda_{N,i}/\lambda_i$ respectively. Since $(\sum_{i=1}^{n_b} \|u_{N,i}(\mu)\|^2_Y / \sum_{i=1}^{n_b} \|u_i(\mu)\|^2_Y)^{1/2}$ and $\lambda_{N,i}/\lambda_i$ can be greater than 1, $\eta^u_{N,n_b}(\mu)$ and $\eta^\lambda_{N,n_b,2}(\mu)$ may thus be less than 1 when bounds (2.67) and (2.68) are sharp, especially in the large N limit.

We first look at results for the augmented reduced basis approximation. Figure (2-15) – (2-17) shows how $\bar{\eta}^u_{N,n_b}$, $\bar{\eta}^\lambda_{N,n_b,1}$ and $\bar{\eta}^\lambda_{N,n_b,2}$ vary with N for different n_b, where $\bar{\eta}^u_{N,n_b}$, $\bar{\eta}^\lambda_{N,n_b,1}$ and $\bar{\eta}^\lambda_{N,n_b,2}$ are averages of $\eta^u_{N,n_b}(\mu)$, $\eta^\lambda_{N,n_b,1}(\mu)$ and $\eta^\lambda_{N,n_b,2}(\mu)$ over the sample set Ξ_T given in Section 2.2.4. From these figures, we can conclude that $\Delta^\lambda_{N,n_b,2}(\mu)$ is a more effective bound for λ_i compared to $\Delta^\lambda_{N,n_b,1}(\mu)$. In addition, the effectivities are in general very good — it is of $O(10)$ for $\bar{\eta}^u_{N,n_b}$ and $O(10^2)$ for $\bar{\eta}^\lambda_{N,n_b,2}$. Lastly, we note that the error estimators presented here are asymptotic bounds for the actual errors; they are thus not rigorous, especially for small N.

For the vectorial reduced basis approximation, we obtain results that parallel that obtained for the augmented reduced basis approximation as shown in Figure (2-18) – (2-20). The error bound $\Delta^\lambda_{N,n_b,2}(\mu)$ is again a better bound than $\Delta^\lambda_{N,n_b,1}(\mu)$. In addition, we notice that $\bar{\eta}^u_{N,n_b}$ for the vectorial case is better than that obtained in the augmented case.

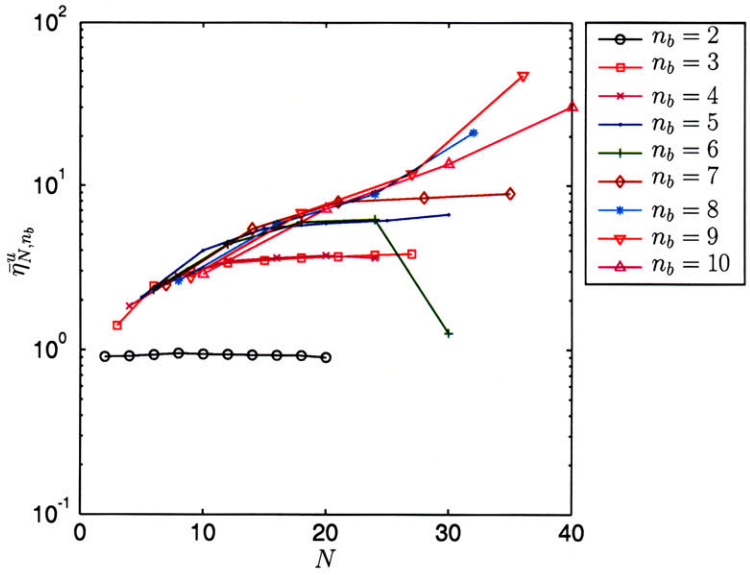

Figure 2-15: Variations of the average effectivity of Δ^u_{N,n_b}, $\bar{\eta}^u_{N,n_b}$ with N for $2 \leq n_b \leq 10$ and $\hat{\mathbf{u}} \in (W_N^A)^{n_b}$.

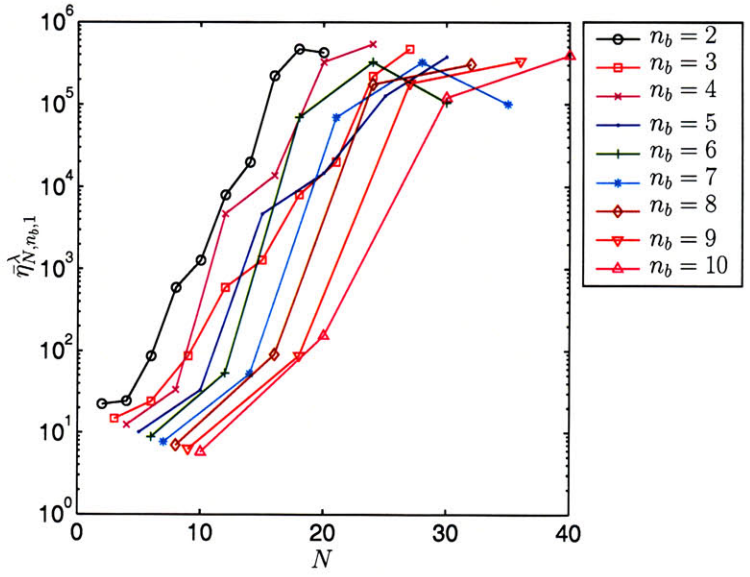

Figure 2-16: Variations of the average effectivity of $\Delta^\lambda_{N,n_b,1}$, $\bar{\eta}^\lambda_{N,n_b,1}$, with N for $2 \leq n_b \leq 10$ and $\hat{\mathbf{u}} \in (W_N^A)^{n_b}$.

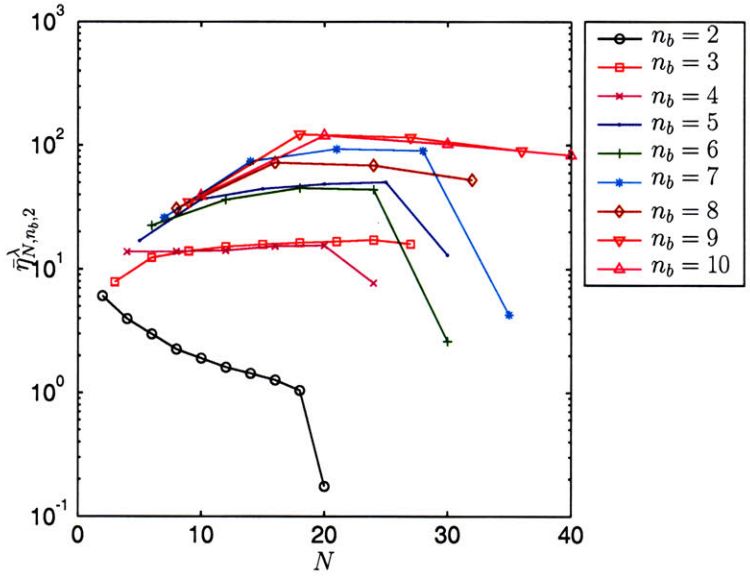

Figure 2-17: Variations of the average effectivity of $\Delta^{\lambda}_{N,n_b,2}$, $\bar{\eta}^{\lambda}_{N,n_b,2}$, with N for $2 \leq n_b \leq 10$ and $\hat{\mathbf{u}} \in (W_N^A)^{n_b}$.

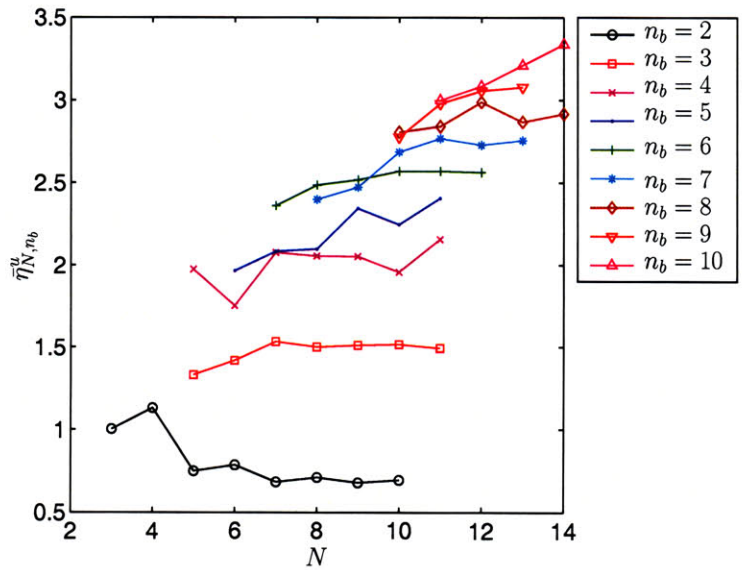

Figure 2-18: Variations of the average effectivity of Δ^{u}_{N,n_b}, $\bar{\eta}^{u}_{N,n_b}$ with N for $2 \leq n_b \leq 10$ and $\hat{\mathbf{u}} \in W_N^V$.

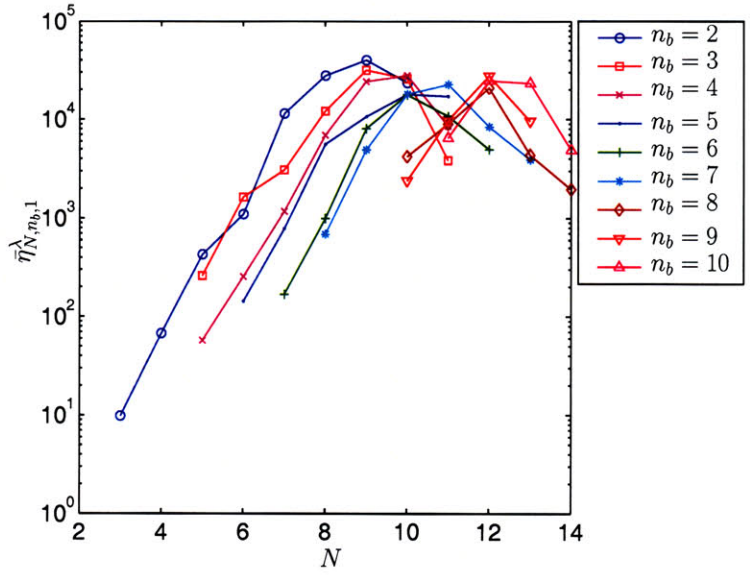

Figure 2-19: Variations of the average effectivity of $\Delta^{\lambda}_{N,n_b,1}$, $\bar{\eta}^{\lambda}_{N,n_b,1}$, with N for $2 \leq n_b \leq 10$ and $\hat{\mathbf{u}} \in W_N^V$.

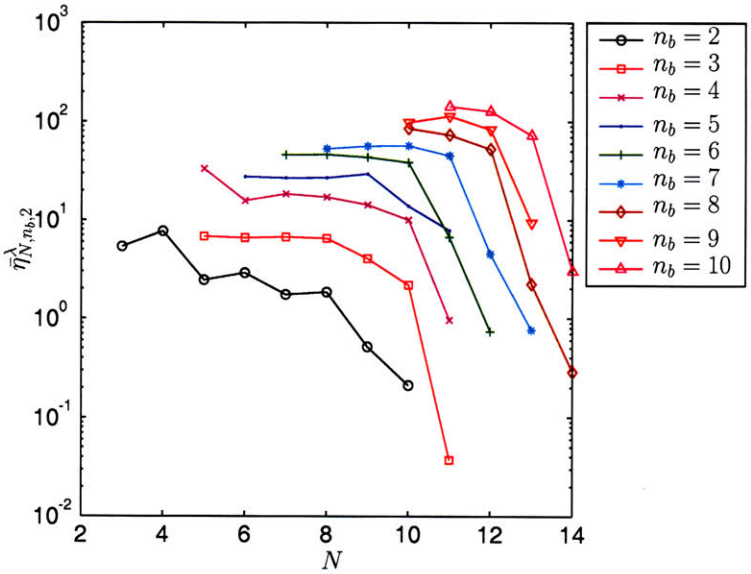

Figure 2-20: Variations of the average effectivity of $\Delta^{\lambda}_{N,n_b,2}$, $\bar{\eta}^{\lambda}_{N,n_b,2}$, with N for $2 \leq n_b \leq 10$ and $\hat{\mathbf{u}} \in W_N^V$.

Chapter 3

Empirical Interpolation Method

3.1 Introduction

In this chapter, we outline the empirical interpolation method detailed in [8, 44] and generalized in [73]. This procedure proves to be particularly useful in the efficient reduced basis approximation of nonlinear elliptic and parabolic equations [8, 44]. We shall demonstrate the same for a nonlinear eigenvalue problem in Chapter 4.

We will examine the approximation of parametric function using the empirical interpolation procedure. However, the procedure is applicable to a wide variety of problems. In [73], we show that this procedure compares favorably to some standard results in classical algebraic polynomial approximations of some typical geometries. In addition, the method is very versatile: it can be used for nonstandard geometries and non-polynomial spaces and the accuracy of the approximation can be easily controlled by using *a posteriori* error estimator. It also has applications in the approximation of solutions of partial differential equations beyond the reduced basis context, for example in modal analysis and dynamic simulations.

3.2 Problem Formulation

3.2.1 Problem Statement

We consider the approximation of the parameter dependent function $g(x;\mu) \in C^0(\mathcal{D};X)$ of sufficient regularity where $X \equiv L^\infty(\Omega) \cap C^0(\Omega)$; $\Omega \in \mathbb{R}^d$ is our spatial domain of dimension d; $L^\infty(\Omega) \equiv$

$\{v|\ \text{ess sup}_{x\in\Omega}|v(x)| \leq \infty\}$; and $C^0(\mathcal{D};X) \equiv \{v(\,\cdot\,;\mu) \in X, \forall \mu \in \mathcal{D}\}$; and $\mathcal{D} \in \mathbb{R}^p$ is our parameter domain of dimension p. We also associate X with a inner product $(w,v)_X$ and norm $\|\cdot\|_X = (\cdot,\cdot)_X^{1/2}$.

In particular, given only partial knowledge on $g(\,\cdot\,;\mu)$ over Ω, we would like to approximate $g(\,\cdot\,;\mu)$ by a collateral reduced basis expansion $g_M(\,\cdot\,;\mu)$ in the approximation space

$$W_M^g = \text{span}\ \{g(x;\mu_m^g), 1 \leq m \leq M\} \tag{3.1}$$

induced by a parameter sample

$$S_M^g = \{\mu_m^g \in \mathcal{D}, 1 \leq m \leq M\}, \tag{3.2}$$

where M is the dimension of S_M^g.

3.2.2 Critical Observation

The rational for the above proposal is similar to that of reduced basis approximation. We define the *Kolmogorov M-width* of $\mathcal{U} \equiv \{g(\,\cdot\,;\mu), \mu \in \mathcal{D}\}$ as [56, 63, 96]

$$d_M(\mathcal{U},X) = \inf_{X_M} \sup_{x\in\mathcal{U}} \inf_{y\in X_M} \|x-y\|_X \tag{3.3}$$

where X_M is a M-dimensional subspace of X. In Section 2.3.1, we have noted that the approximation will be rapidly convergent if d_M approaches zero rapidly as M increases. We again expect this to hold in our case based on the regularity of the solutions $g(\,\cdot\,;\mu)$ with respect to μ. In [72], it is further shown that d_M is almost realized if X_M is spanned by elements in $\mathcal{U}$. Exponential convergence is also achieved when analyticity exists in the parameter dependency.

However, there are two difficulties related to determination of X_M and ultimately our approximation based on (3.3). First, determination of X_M based on (3.3) is combinatorically difficult. Second, assuming that X is provided with a scalar product, the best fit of an element $g \in \mathcal{U}$ in some finite dimensional space X_M that leads to a solution closest to d_M is then given by the orthogonal projection onto X_M. This is the basis for the optimality of reduced basis approximation. However, in many cases, this is a costly process and the knowledge of $g(\,\cdot\,;\mu)$ over the entire domain Ω is required. Thus, when the elements in $\mathcal{U}$ are continuous, the interpolation is a tool that provide a

cheap surrogate to the evaluation of the orthogonal projection.

Within the context of the reduced basis approximation, an interpolation scheme is needed when dealing with nonaffine and nonlinear PDEs which do not admit an efficient, i.e., online $\mathcal{N}$ independent, computational decomposition. This will be further discussed in Chapter 4.

3.2.3 Empirical Interpolation Method

The empirical interpolation procedure seeks to construct

$$S_M^g = \{\mu_m^g \in \mathcal{D}, 1 \leq m \leq M\}, \tag{3.4}$$

and the associated approximation space

$$\begin{aligned} W_M^g &= \text{span } \{g(x;\mu_m^g), 1 \leq m \leq M\} \\ &= \text{span } \{q_m, 1 \leq m \leq M\}, \end{aligned} \tag{3.5}$$

that will act as a good surrogate to X_M, based on the "greedy" sampling procedure outlined in [82, 99, 115]. In addition, the procedure construct a set of interpolation points,

$$T_M^g \equiv \{t_m, 1 \leq m \leq M\}, \tag{3.6}$$

that allows us to construct an approximation to $g(\,\cdot\,;\mu)$ based on the interpolant of g over T_M^g:

$$g_M(\,\cdot\,;\mu) = \sum_{j=1}^{M} \alpha_{M\,j}(\mu) q_j(\,\cdot\,)\ , \tag{3.7}$$

where $\alpha_M(\mu) \in \mathbb{R}^M$ is given by

$$\sum_{j=1}^{M} q_j(t_i)\alpha_{M\,j}(\mu) = g(t_i;\mu), \quad i = 1,\ldots,M\ . \tag{3.8}$$

If we define $B^M \in \mathbb{R}^M \times \mathbb{R}^M$ as

$$B_{ij}^M = q_j(t_i), \tag{3.9}$$

we can more directly define an interpolation operator $\mathcal{I}_M$ in W_M^g:

$$(\mathcal{I}_M g)(\,\cdot\,;\mu) = \sum_{i=1}^{M} g(t_i;\mu) h_i^M(\,\cdot\,), \tag{3.10}$$

where

$$h_i^M(\,\cdot\,) = \sum_{j=1}^{M} q_j(\,\cdot\,)(B^M)_{ji}^{-1}. \tag{3.11}$$

We then have $g_M(\,\cdot\,;\mu) \equiv (\mathcal{I}_M g)(\,\cdot\,;\mu)$. We note that since by construction $(\mathcal{I}_M g)(t_j;\mu) = g(t_j;\mu)$, $1 \le j \le M$, we obtain

$$h_i^M(t_j) = \delta_{ij}, 1 \le i, j \le M, \tag{3.12}$$

from (3.10).

The interpolation error, $\varepsilon_M^g(\mu)$, is defined as

$$\varepsilon_M^g(\mu) = \|g(\,\cdot\,;\mu) - g_M(\,\cdot\,;\mu)\|_{L^\infty(\Omega)}\,. \tag{3.13}$$

In practice, we usually know only solutions $g(\cdot;\mu)$ and $g_M(\cdot;\mu)$ at finite points in Ω, for example the vertices of a mesh resulting from a triangulation $\mathcal{T}$. By assuming that the functions are piecewise linear over elements $T \in \mathcal{T}$, we approximate $\varepsilon_M^g(\mu)$ by the maximum of $|g(\cdot;\mu) - g_M(\cdot;\mu)|$ evaluated at these finite points.

The construction procedure for the empirical interpolation method starts with a large sample set $U \equiv \{g(\,\cdot\,;\mu), \mu \in \Xi_T \subset \mathcal{D}\} \subset \mathcal{U}$. We assume U is representative of the entire set $\mathcal{U}$ in the sense that $\sup_{x\in\mathcal{U}} \inf_{y\in \text{span }\{U\}} \|x - y\|_X$ is much smaller than the approximation we envision through the interpolation process. We now construct, for $M_{\max} < \dim$ (span $\{U\}$), nested sample sets S_M^g, $1 \le M \le M_{\max}$, nested approximation spaces W_M^g, $1 \le M \le M_{\max}$, and nested interpolation points T_M^g, $1 \le M \le M_{\max}$. We choose our first sample point μ_1^g to be $\arg\max_{\mu\in\Xi_T} \|g(\,\cdot\,;\mu)\|_{L^\infty(\Omega)}$. In addition, we set $t_1 = \arg\max_{x\in\Omega} |g(x;\mu_1^g)|$, $q_1 = g(x;\mu_1^g)/g(t_1;\mu_1^g)$, and $B_{11}^1 = 1$. Then for $M = 2, \ldots, M_{\max}$, we determine

$$\mu_M^g = \arg\max_{\mu\in\Xi_T} \varepsilon_{M-1}^g(\mu)\,. \tag{3.14}$$

In addition, we define

$$r_M(x) = g(x;\mu_M^g) - g_{M-1}(x;\mu_M^g)\ , \tag{3.15}$$

$$t_M = \arg\max_{x\in\Omega} |r_M(x)|\ . \tag{3.16}$$

We then set

$$q_M(x) = r_M(x)/r_M(t_M), \tag{3.17}$$

$$B_{ij}^M = q_j(t_i), 1 \le i,j \le M. \tag{3.18}$$

Finally,

$$S_M^g = S_{M-1}^g \cup \mu_M^g, \tag{3.19}$$

$$W_M^g = W_{M-1}^g + \text{span}\,\{q_M\}, \tag{3.20}$$

$$T_M^g = T_{M-1}^g \cup t_M. \tag{3.21}$$

We repeat this procedure until $\max_{\mu\in\Xi_T} \varepsilon_M^g(\mu)$ is below $\varepsilon_{\text{tol}}^g$, a tolerance criteria we define. Then, $M_{\max}$ is the dimension S_M^g for which this tolerance criteria is satisfied.

The construction procedure described above differ slightly from that proposed earlier in [8, 44]. In the current construction, we use directly the interpolation error $\varepsilon_M^g(\mu)$ as the error measure. In [8, 44] however, the best fit error as defined by $\inf_{z\in W_M^g} \|g(\cdot;\mu)-z\|_{L^\infty(\Omega)}$ is used — this involves a standard linear program, which can be very expensive. It remains for us to demonstrate that the current construction of T_M^g and W_M^g is well-posed.

Lemma 3.1. *Suppose that $M_{\max}$ is chosen such that $M_{\max} < \dim\,(\text{span}\{U\})$, then the space W_M^g is of dimension M.*

Proof. It directly follows from our hypothesis on $M_{\max} < \dim\,(\text{span}\,\{U\})$ that $\varepsilon_M^g(\mu_{M+1}^g) = \max_{\mu\in\mathcal{D}} \varepsilon_M^g(\mu) > 0$ for any $M < M_{\max}$. We now prove lemma 3.1 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 $\varepsilon_{M-1}^g(\mu_M^g) = 0$; however, the latter contradicts $\varepsilon_{M-1}^g(\mu_M^g) > 0$. □

Lemma 3.2. *The construction of the interpolation points T_M^g is well-defined, and the functions $\{q_1, \ldots, q_M\}$ form a basis for W_M^g.*

Proof. We proceed by induction. Clearly, we have $W_1^g = \text{span}\,\{q_1\}$. Assuming $W_{M-1}^g = \text{span}\,\{q_1, \ldots, q_{M-1}\}$, the procedure is well-defined if (i) $|r_M(t_M))| > 0$ and (ii) B^{M-1} is invertible — we may form $W_M^g = \text{span}\,\{q_1, \ldots, q_M\}$. To prove (i), we observe that $|r_M(t_M)| = \varepsilon_{M-1}^g(\mu_M^g) > 0$. To prove (ii), we just note by the construction procedure that $B_{ij}^{M-1} = r_j(t_i)/r_j(t_j) = 0$ for $i < j$ since $g(t_i; \mu_j^g) = (\mathcal{I}_{j-1}g)(t_i; \mu_j^g)$ for $i < j$; that $B_{ij}^{M-1} = r_j(t_i)/r_j(t_j) = 1$ for $i = j$; and that $\left|B_{ij}^{M-1}\right| = |r_j(t_i)/r_j(t_j)| \leq 1$ 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 3.3. *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. Let $w = \sum_{j=1}^M \kappa_j q_j(x) \in W_M^g$, where $\kappa \in \mathbb{R}^M$ is the solution of $\sum_{j=1}^M B_{ij}^M \kappa_j = \alpha_i, 1 \leq i \leq M$. Clearly, we have $w(t_i) = \sum_{j=1}^M \kappa_j\, q_j(t_i) = \alpha_i$. Furthermore, it follows from the invertibility of B^M that κ is unique. □

Theorem 3.1. *Assume that there exists a Banach space $\mathcal{Y}$ such that $\mathcal{U} \subset \mathcal{Y} \subset L^\infty(\Omega)$, and that there exists a sequence of finite dimensional spaces*

$$W_1^g \subset W_2^g \subset \cdots \subset W_M^g \subset \cdots \subset \text{span}\{\mathcal{U}\}, \quad \text{and,} \quad \dim W_M^g = M, \tag{3.22}$$

such that there exists $c > 0$ and $\alpha > \log(4)$ with

$$\inf_{v\in W_M^g} \|g(\,\cdot\,;\mu) - v\|_{\mathcal{Y}} \leq ce^{-\alpha M}, \quad \forall g(\,\cdot\,;\mu) \in \mathcal{U}. \tag{3.23}$$

Then,

$$\varepsilon_M^g \leq ce^{-(\alpha M - \log(4))}. \tag{3.24}$$

The proof of Theorem 3.1 can be found in [73]. Theorem 3.1 states that, under the reasonable condition that the reduced space allows an exponential convergence (actually even faster convergence is observed most of the times, as explained in [16]), the empirical interpolation procedure :

(i) proposes a discrete space (spanned by the chosen $g(\,\cdot\,;\mu_m^g)$) where the best fit is good, and (ii) provides a set of interpolation points that leads to a convergent interpolant.

3.2.4 Error Analysis

A priori stability: Lebesgue constant

To begin, we define a Lebesgue constant Λ_M as [101]

$$\Lambda_M = \sup_{x\in\Omega} \sum_{i=1}^{M} |h_i^M(x)|. \tag{3.25}$$

We observe that Λ_M depends on W_M^g and T_M but not on μ or our choice of basis for W_M^g. We can further prove that [44]

Lemma 3.4. *For any $g \in \mathcal{U}$, the interpolation error satisfies*

$$\|g(\,\cdot\,;\mu) - (\mathcal{I}_M g)(\,\cdot\,)\|_{L^\infty(\Omega)} \le (1+\Lambda_M) \inf_{v\in W_M^g} \|g(\,\cdot\,;\mu) - v(\,\cdot\,)\|_{L^\infty(\Omega)}. \tag{3.26}$$

Furthermore, $\Lambda_M \le 2^M - 1$.

Lemma 3.4 is very pessimistic and of little practical value; it nevertheless provides some notion of stability. In most cases, $\inf_{v\in W_M^g} \|g(\,\cdot\,;\mu) - v(\,\cdot\,)\|_{L^\infty(\Omega)}$ decreases sufficiently rapidly that $2^M \inf_{v\in W_M^g} \|g(\,\cdot\,;\mu) - v(\,\cdot\,)\|_{L^\infty(\Omega)} \to 0$ as $M \to \infty$.

A posteriori error estimation

Given an approximation $g_M(x;\mu)$ for $M \le M_{\max}-1$, we define $\hat{\varepsilon}_M^g(\mu) = |g(t_{M+1};\mu) - g_M(t_{M+1};\mu)|$. We can then prove the following [44]:

Proposition 3.1. *If $g(\,\cdot\,;\mu) \in W_{M+1}^g$, then (i) $g(x;\mu) - g_M(x;\mu) = \pm\hat{\varepsilon}_M^g(\mu) q_{M+1}(x)$, and (ii) $\hat{\varepsilon}_M^g(\mu) = \|g(\,\cdot\,;\mu) - g_M(\,\cdot\,;\mu)\|_{L^\infty(\Omega)}$.*

Proof. By our assumption $g(\cdot;\mu) \in W_{M+1}^g$, there exists $\kappa(\mu) \in \mathbb{R}^{M+1}$ such that $g(\cdot;\mu) - g_M(\cdot;\mu) = \sum_{j=1}^{M+1} \kappa_j(\mu)\, q_j(\,\cdot\,)$. We now consider $x = t_i, 1 \le i \le M+1$, and arrive at

$$\sum_{j=1}^{M+1} \kappa_j(\mu)\, q_j(t_i) = g(t_i;\mu) - g_M(t_i;\mu), \quad 1 \le i \le M+1\,. \tag{3.27}$$

It thus follows that $\kappa_j(\mu) = 0$, $1 \le j \le M$, since $g(t_i;\mu) - g_M(t_i;\mu) = 0, 1 \le i \le M$ and the matrix $q_j(t_i)(= B_{ij}^M)$ is lower triangular, and that $\kappa_{M+1}(\mu) = g(t_{M+1};\mu) - g_M(t_{M+1};\mu)$ since $q_{M+1}(t_{M+1}) = 1$; this concludes the proof of (i). The proof of (ii) then directly follows from $\|q_{M+1}\|_{L^\infty(\Omega)} = 1$. □

In general, $g(x;\mu) \notin W_{M+1}^g$ and hence

$$\varepsilon_M^g(\mu) \ge \hat{\varepsilon}_M^g(\mu); \tag{3.28}$$

$\hat{\varepsilon}_M^g(\mu)$ is then a lower bound to $\varepsilon_M^g(\mu)$. However, if $\varepsilon_M^g(\mu) \to 0$ very fast, we expect the effectivity,

$$\eta_M^g(\mu) = \frac{\hat{\varepsilon}_M^g(\mu)}{\varepsilon_M^g(\mu)} \tag{3.29}$$

to be close to unity. In addition, to determine our error estimator $\hat{\varepsilon}_M^g(\mu)$, we only need to determine t_{M+1}. This is very inexpensive — we only need to do an additional iteration of the empirical interpolation procedure.

3.3 Numerical Example

Here, we look at an example unrelated to the reduced basis approximation — application of empirical interpolation method within the reduced basis framework will be explored in Chapter 4. We shall look at approximation of a parameter-dependent function $g(\mathbf{x};\mu)$ given by

$$g(\mathbf{x};\mu) = \int_\Omega \ell(\mathbf{x}';\mu) f(\mathbf{x},\mathbf{x}') d\mathbf{x}', \tag{3.30}$$

where $\Omega \subset \mathbb{R}^2$, $\ell \in L_\infty(\Omega)$, and $f \in L_\infty(\Omega)$. For a discretization of Ω into $\mathcal{N}$ points, a full evaluation the convoluted function (3.32) for every new μ will required $O(\mathcal{N}^2)$ operations. However, if for a given Ξ_T, we construct an approximation space W_M^g and the associated interpolation points T_M^g, we will only required $\mathcal{O}(M\mathcal{N})$ operations — we only evaluate the integral at M magic points.

As an example, we consider a domain $\Omega \equiv [-0.5, 0.5] \times [-0.5, 0.5] \subset \mathbb{R}^2$, $\mu \in [1, 10]$, $\mathbf{x} \equiv (x, y)$, $\ell(\mathbf{x};\mu) = \sin(2\pi\mu|\mathbf{x}|)$, and $f(\mathbf{x},\mathbf{y}) = \frac{50}{\pi}\exp(-50|\mathbf{x}-\mathbf{y}|^2)$. We construct our approximation based on the sample set $U^g(\equiv \{g(\,\cdot\,;\mu), \mu \in \Xi_T \subset \mathcal{D} \equiv [1, 10]\}$, where Ξ_T consists of 100 μ-points distributed uniformly in $\mathcal{D}$. This may have applications in areas such as animation where μ represent

M	$\max_{\mu\in\Xi_T} \varepsilon_M^g(\mu)$	Λ_M	$\max_{\mu\in\Xi_T} \hat{\varepsilon}_M^g(\mu)$	$\bar{\eta}_M^g$	$\max_{\mu\in\Xi_T} \eta_M^g(\mu)$	$\min_{\mu\in\Xi_T} \eta_M^g(\mu)$
2	8.5969 E−1	1.5054	8.5969 E−1	0.01	1.00	2.45 E−4
4	2.7407 E−1	1.9762	2.7407 E−1	0.20	1.00	2.13 E−2
6	9.7983 E−2	3.1647	9.7983 E−2	0.42	1.00	5.09 E−2
8	6.0001 E−2	3.8942	6.0001 E−2	0.36	1.00	9.29 E−3
10	3.8791 E−3	3.2768	3.8791 E−3	0.73	1.00	1.80 E−1
12	6.4367 E−4	4.5580	6.4367 E−4	0.53	1.00	2.78 E−2
14	6.3474 E−5	5.9444	6.3474 E−5	0.97	1.00	3.30 E−1
16	8.1714 E−7	4.5135	8.1714 E−7	1.00	1.00	1.00

Table 3.1: Comparison between the error estimate and the actual error, for $g(\,\cdot\,;\mu) = \int_\Omega \ell(\mathbf{x}';\mu)\, f(\,\cdot\,,\mathbf{x}')\, d\mathbf{x}'$, where $\ell(\mathbf{x};\mu) = \sin(2\pi\mu|\mathbf{x}|)$, $f(\mathbf{x},\mathbf{y}) = \frac{50}{\pi}\exp(-50|\mathbf{x}-\mathbf{y}|^2)$, $\Omega \equiv [-0.5,0.5]\times[-0.5,0.5] \subset \mathbb{R}^2$, and $\mu \in [1,10]$. The $\bar{\eta}_M^g$ is the mean of $\eta_M^g(\mu)$ for $\mu \in \mathcal{D}$.

temporal variables, or the regeneration of 3D tomographic data sets where μ represent spatial variables.

Table 3.1 shows that the error $\max_{\mu\in\Xi_T} \varepsilon_M^g(\mu)$ decreases monotonically and the Lebesgue constants are generally small for all M. Thus, the approximation leads to fast evaluation of g with minimal loss of accuracy. The error estimator, $\hat{\varepsilon}_M^g(\mu)$, provides a good estimate to actual interpolation error, $\varepsilon_M^g(\mu)$ — the average effectivities, $\bar{\eta}_M^g$ are close to 1, especially for larger M. In addition $\max_{\mu\in\Xi_T} \eta_M^g(\mu)$ is always close to 1 while $\min_{\mu\in\Xi_T} \eta_M^g(\mu)$ approaches 1 as M increases.

We note that $\max_{\mu\in\Xi_T} \hat{\varepsilon}_M^g(\mu)$ is always equal to $\max_{\mu\in\Xi_T} \varepsilon_M^g(\mu)$. This is due to our construction procedure that uses $\varepsilon_M^g(\mu)$ as the error measure: from (3.14) and (3.16), we have

$$\begin{aligned}
\max_{\mu\in\Xi_T} \varepsilon_M^g(\mu) &= \max_{\mu\in\Xi_T} \|g(\,\cdot\,;\mu) - g_M(\,\cdot\,;\mu)\|_{L^\infty(\Omega)} \\
&= g(\,\cdot\,;\mu_{M+1}^g) - g_M(\,\cdot\,;\mu_{M+1}^g)\|_{L^\infty(\Omega)} \\
&= \|r_{M+1}(\,\cdot\,)\|_{L^\infty(\Omega)} \\
&= |r_{M+1}(t_{M+1})| \\
&= |g(t_{M+1};\mu_{M+1}^g) - g_M(t_{M+1};\mu_{M+1}^g)| \\
&= \max_{\mu\in\Xi_T} |g(t_{M+1};\mu) - g_M(t_{M+1};\mu)| \\
&= \max_{\mu\in\Xi_T} \hat{\varepsilon}_M^g(\mu). \qquad (3.31)
\end{aligned}$$

The above is not necessarily true for the construction procedure given in [8, 44].

We now relax slightly the parametric smoothness requirement of the empirical interpolation

method on admissible functions. We consider functions $\tilde{g}(\,\cdot\,;\mu)$ that are piecewise continuous in $\mathcal{D}$. We first note that if $W_M^{\tilde{g}}$ contains sufficient solutions from all smooth segments of the solution manifold that are disjointed, the best fit error $\inf_{v\in W_M^{\tilde{g}}} \|\tilde{g}(\,\cdot\,;\mu)-v(\,\cdot\,)\|_{L^\infty(\Omega)}$ will be small. In other words, the best fit error does not require $\tilde{g}(\,\cdot\,;\mu)$ to be continuous in $\mathcal{D}$ although this smoothness property will allow more rapid convergence of the best fit error with M. Since the inequality (3.26) always bounds the interpolation error $\|\tilde{g}(\,\cdot\,;\mu) - (\mathcal{I}_M\tilde{g})(\,\cdot\,)\|_{L^\infty(\Omega)}$ by the best fit error, the approximation based on the empirical interpolation method will be close to the best fit approximation, provided that the Lebesgue constant Λ_M remains small. In addition, we note that Λ_M only depends on $T_M^{\tilde{g}}$ and has no explicit dependence on the smoothness requirement. Thus, we expect the empirical interpolation approximation will be reasonably good even in the case where the functions $\tilde{g}(\,\cdot\,;\mu)$ are only piecewise continuous in $\mathcal{D}$.

We demonstrate this observation based on the following numerical example: we define $\tilde{g}(\mathbf{x};\mu)$ as

$$\tilde{g}(\mathbf{x};\mu) = \int_\Omega \tilde{\ell}(\mathbf{x}';\mu) f(\mathbf{x},\mathbf{x}')d\mathbf{x}', \tag{3.32}$$

where $\mathbf{x}\in\Omega\equiv[-0.5,0.5]\times[-0.5,0.5]\subset\mathbb{R}^2$; $\mu\in[1,10]$;

$$\tilde{\ell}(\mathbf{x};\mu) = \begin{cases} \sin(2\pi\mu|\mathbf{x}|), & \mu\le 3, \\ \cos(2\pi\mu|\mathbf{x}|), & \mu > 3; \end{cases} \tag{3.33}$$

and $f(\mathbf{x},\mathbf{y}) = \frac{50}{\pi}\exp(-50|\mathbf{x}-\mathbf{y}|^2)$. Figure 3-1 shows the piecewise continuous property of $\tilde{g}(\,\cdot\,;\mu)$ with respect to μ. From Table 3.2, we observe that Λ_M is $O(10)$ for $M\le 16$, the approximation error $\max_{\mu\in\Xi_T}\varepsilon_M^{\tilde{g}}(\mu)$ decreases with M, and the average effectivity $\bar{\eta}_M^{\tilde{g}}$ approaches 1 as M increases. Nevertheless, compare to the results in Table 3.1, we observe a slower convergence in the approximation — with $M=16$, the approximation error in $\tilde{g}$ is of $O(10^{-5})$ but the approximation error in g is of $O(10^{-7})$. Thus, the smoothness property affects the convergence rate of the approximation.

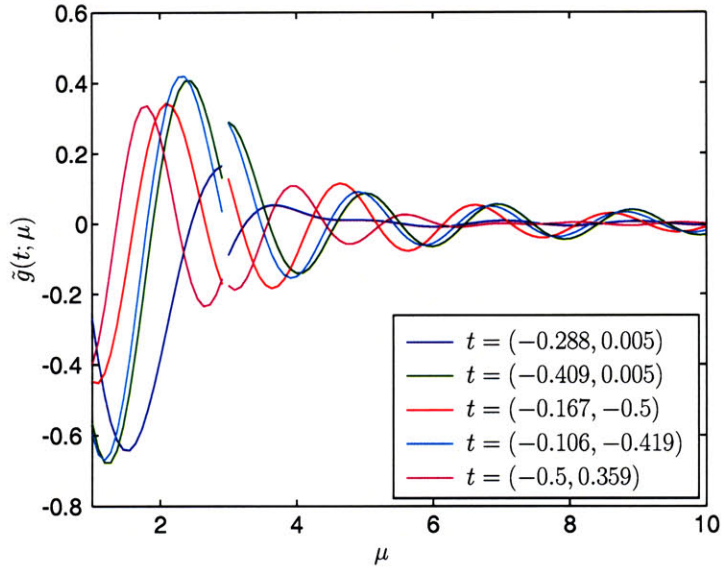

Figure 3-1: Discontinuity in $\tilde{g}(t;\mu)$ at $\mu = 3$ for some selected values of t — $\tilde{g}(\,\cdot\,;\mu)$ is piecewise continuous between $1 \leq \mu \leq 3$ and $3 < \mu \leq 10$.

M	$\max_{\mu\in\Xi_T} \varepsilon_M^{\tilde{g}}(\mu)$	Λ_M	$\max_{\mu\in\Xi_T} \hat{\varepsilon}_M^{\tilde{g}}(\mu)$	$\bar{\eta}_M^{\tilde{g}}$	$\max_{\mu\in\Xi_T} \eta_M^{\tilde{g}}(\mu)$	$\min_{\mu\in\Xi_T} \eta_M^{\tilde{g}}(\mu)$
2	4.3379 E−1	1.4484	4.3379 E−1	0.05	1.00	1.8201 E−3
4	3.1848 E−1	2.0227	3.1848 E−1	0.07	1.00	2.8063 E−3
6	1.0347 E−1	2.4519	1.0347 E−1	0.49	1.00	1.9018 E−2
8	2.6034 E−2	3.3224	2.6034 E−2	0.41	1.00	3.5207 E−2
10	6.0917 E−3	4.0432	6.0917 E−3	0.81	1.00	1.4041 E−1
12	1.0721 E−3	5.1275	1.0721 E−3	0.41	1.00	9.6578 E−3
14	2.7326 E−4	5.7885	2.7326 E−4	0.91	1.00	1.6602 E−1
16	1.2224 E−5	6.0140	1.2224 E−5	0.95	1.00	2.7575 E−1

Table 3.2: Comparison between the error estimate and the actual error, for $\tilde{g}(\,\cdot\,;\mu) = \int_\Omega \tilde{\ell}(\mathbf{x}';\mu)\, f(\,\cdot\,,\mathbf{x}')\, d\mathbf{x}'$, where $\tilde{\ell}(\mathbf{x};\mu) = \sin(2\pi\mu|\mathbf{x}|)$ for $\mu \leq 3$ and $\cos(2\pi\mu|\mathbf{x}|)$ otherwise; $f(\mathbf{x},\mathbf{y}) = \frac{50}{\pi}\exp(-50|\mathbf{x}-\mathbf{y}|^2)$; $\Omega \equiv [-0.5, 0.5] \times [-0.5, 0.5] \subset \mathbb{R}^2$; and $\mu \in [1, 10]$. The $\bar{\eta}_M^{\tilde{g}}$ is the mean of $\eta_M^{\tilde{g}}(\mu)$ for $\mu \in \mathcal{D}$.

Chapter 4

Nonlinear Model Problem

4.1 Introduction

We shall now consider the extension of the reduced basis method to a nonlinear eigenvalue problem. The reduced basis method developed in Chapter 2 and the associated *a posteriori* error estimation procedure rely on the affine parameter dependence property of the problem to arrive at an efficient offline-online computational procedure. When this property is absent, the computational strategy breaks down, leading to an online computational cost that depends on $\mathcal{N}$.

Recently, [8, 44] incorporated the empirical interpolation procedure described in Chapter 3 within the reduced basis framework to successfully treat any elliptic problems with nonaffine parameter dependence and recover the efficiency of the offline-online computational strategy. We shall here apply the technique to a nonlinear eigenvalue problem. We consider only the determination of the smallest eigenvalue and the associated eigenvector. It serves as an introduction to the next two chapters, where we will consider the reduced basis approximation of Kohn Sham equations with vectorial eigensolutions. Additionally, the numerical example examined involves a nonlinearity term similar to that encountered in Thomas Fermi model [12, 70], a prototypical model typically used for mathematical purposes as its mathematical features are similar to that of Density Functional Theory models. The methodology developed can also applied to models based on the Orbital Free Density Functional Theory [24, 39].

4.2 Abstract Formulation

4.2.1 Problem Statement

Given two Hilbert spaces X and Y where $X \equiv L^2(\Omega)$ and $H_0^1(\Omega) \subset Y \subset H^1(\Omega)$, we consider the following nonlinear eigenvalue problem: given any $\mu \in \mathcal{D}$, we are interested in finding $\lambda(\mu)$ where $(u(\mu), \lambda(\mu)) \in Y \times \mathbb{R}$ satisfies

$$\begin{aligned} a_1(u(\mu), v) + \mu^2 \int_\Omega g(u(\mu); \cdot; \mu) v &= \lambda(\mu) m(u(\mu), v), \quad \forall v \in Y, \\ m(u(\mu), u(\mu)) &= 1; \end{aligned} \tag{4.1}$$

where[1] $\mathcal{D} \subset \mathbb{R}_+$ is our parameter domain; Ω is a bounded domain in $\mathbb{R}$; and $a_1(w, v)$ and $m(w, v)$ continuous

$$a_1(w, v) \leq \gamma_a \|w\|_Y \|v\|_Y, \quad \forall w, v \subset Y, \tag{4.2}$$

$$m(w, v) \leq \gamma_m \|w\|_X \|v\|_X, \quad \forall w, v \subset X; \tag{4.3}$$

coercive

$$0 \leq \alpha_a \equiv \inf_{w \subset Y} \frac{a_1(w, w)}{\|w\|_Y^2}, \tag{4.4}$$

$$0 \leq \alpha_m \equiv \inf_{w \subset X} \frac{m(w, w)}{\|w\|_X^2}; \tag{4.5}$$

and symmetric, $a_1(w, v) = a_1(v, w)$, $\forall v, w \subset Y$, and $m(w, v) = m(v, w)$, $\forall v, w \subset X$. Here, g is a general Cnonaffine nonlinear function of the field solution $u(x; \mu)$, spatial coordinate x, and parameter μ. We require $g : \mathbb{R} \times \Omega \times \mathcal{D} \rightarrow \mathbb{R}$ to be continuous in its arguments, monotonically increasing in the first argument, and $g(z; x; \mu) \geq 0$, $\forall z \in \mathbb{R}$, $\forall x \in \Omega$, $\forall \mu \in \mathcal{D}$.

As the focus of this chapter is on the nonlinear term g, we assume that a_1 and m are parameter-independent; parameter dependence can however be readily admitted. In anticipation of messier equations to be encountered in Chapter 6 and 7, we rewrite (4.1) in a more convenient form: find

[1] The μ^2 in (4.1) is used to obtain a sufficiently interesting problem for application of reduced basis approximation.

$\mathbf{u}(\mu) \equiv (u(\mu), \lambda(\mu)) \in \mathcal{Y} \equiv (Y \times \mathbb{R})$ such that

$$\mathcal{A}(\mathbf{u}(\mu), \mathbf{v}; \mu) = 0, \quad \forall \mathbf{v} \in \mathcal{Y}, \tag{4.6}$$

where

$$\mathcal{A}(\mathbf{w} \equiv (w, \sigma), \mathbf{v} \equiv (v, \varphi); \mu) \equiv a_1(w, v) + \mu^2 a^{\mathrm{nl}}(w, v; \mu) - \sigma m(w, v) + \varphi(m(w, w) - 1), \tag{4.7}$$

and $a^{\mathrm{nl}}(w, v; \mu) = \int_\Omega g(w; \cdot; \mu)\, v$. Given $\mathbf{z} \equiv (z, \vartheta) \in \mathcal{Y}$, we shall also define the derivative bilinear form $d\mathcal{A}(\,\cdot\,, \,\cdot\,; \mu; \mathbf{z})$: $\mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ as

$$\begin{aligned} &d\mathcal{A}(\mathbf{w} \equiv (w, \sigma), \mathbf{v} \equiv (v, \varphi); \mu; \mathbf{z} \equiv (z, \vartheta)) \quad \equiv \\ &\quad a_1(w, v) + \mu^2 \int_\Omega g'(z; x; \mu) w\, v - \sigma m(w, v) - \sigma m(z, v) - \vartheta m(w, v) + \varphi(2m(w, z)), \end{aligned} \tag{4.8}$$

such that

$$\mathcal{A}(\mathbf{z} + \mathbf{w}, \mathbf{v}; \mu) = \mathcal{A}(\mathbf{z}, \mathbf{v}; \mu) + d\mathcal{A}(\mathbf{w}, \mathbf{v}; \mu; \mathbf{z}) + \frac{\mu^2}{2} \int_\Omega g''(z; x; \mu) w^2 v - \sigma m(w, v) + \varphi m(w, w), \tag{4.9}$$

where

$$g'(w; x; \mu) = \frac{\partial g(s; \cdot; \cdot)}{\partial s}(w; x; \mu), \quad g''(w; x; \mu) = \frac{\partial^2 g(s; \cdot; \cdot)}{\partial s^2}(w; x; \mu). \tag{4.10}$$

4.2.2 "Truth" Approximation

We proceed by first developing a "truth" approximation based on the finite element approximation. We define a finite element space $Y_h \subset Y$ of dimension $\mathcal{N}$ as

$$Y_h \equiv \{v \in Y \mid v|_{\mathbf{T}_h} \in \mathbb{P}_1(\mathbf{T}_h), \ \forall \mathbf{T}_h \in \mathcal{T}_h\}, \tag{4.11}$$

$$\mathbb{P}_1(\mathbf{T}_h) \equiv \mathrm{span}\{1, x\}, \tag{4.12}$$

where $\mathcal{T}_h$ is a uniform "triangulation" of the domain Ω comprised of linear elements T_h. The inner product and norm associated with Y_h are simply inherited from Y. The finite element

approximation to (4.6) is then given by: find $\mathbf{u}_h(\mu) = (u_h(\mu), \lambda_h(\mu)) \in \mathcal{Y} \equiv Y_h \times \mathbb{R}$ such that

$$\mathcal{A}_h(\mathbf{u}_h(\mu), \mathbf{v}_h; \mu) = 0, \quad \forall \mathbf{v} \in \mathcal{Y}_h, \tag{4.13}$$

where we have replace the nonlinear term $a^{\text{nl}}(w, v; \mu)$ by a quadrature sum given by $a^{\text{nl,quad}}(w, v; \mu)$ $= \sum_{\text{quad}} g(w; \cdot\,; \mu)\ v(\cdot)$. We similarly replace the nonlinear terms in the derivative form of $\mathcal{A}_h$, $d\mathcal{A}_h$ by quadrature sums.

To solve (4.13), we will examine two solution methods — the Newton iterative scheme, and the fixed point method. The fixed point method is part of the Self Consistent Field (SCF) schemes frequently used to solve computational chemistry problems [18], as mentioned in Section 1.2.2. We describe these two methods next, and provide a comparison between them in Section 4.2.3.

Newton iterative scheme

Ignoring the higher order terms after the $d\mathcal{A}(\,\hat{\mathbf{u}}\,,\,\hat{\mathbf{v}}\,;\mu;\mathbf{z})$ term in (4.9), we begin with an initial guess $\mathbf{u}_h^0$ and construct the sequence $\mathbf{u}_h^k \in \mathcal{Y}_h$ by

$$d\mathcal{A}_h(\mathbf{u}_h^{k+1} - \mathbf{u}_h^k, \mathbf{v}; \mu; \mathbf{u}_h^k) = -\mathcal{A}_h(\mathbf{u}_h^k, \mathbf{v}; \mu), \quad \forall \mathbf{v} \in \mathcal{Y}_h; \tag{4.14}$$

at each step k, we solve a linear differential problem associated with the linear operator $\text{d}\mathcal{A}$. We repeat this procedure until convergence criteria given by

$$\|\mathcal{A}(\mathbf{u}_h^{k_{\max}}(\mu), \mathbf{v}_h)\|_2 \leq \varepsilon_{\text{tol}}, \quad \forall \mathbf{v}_h \in \mathcal{Y}_h, \tag{4.15}$$

is satisfied. Here, ε_{tol} is a preselected tolerance level, and $k_{\max}$ is the maximum iterations required to satisfy the convergence criteria (4.15).

Fixed point method

We will use the Roothaan algorithm [106] which has been critically analyzed in [20]. We first rewrite $g(w; x; \mu)$ as $p(w; x; \mu)w$. Then, starting from an initial guess $\mathbf{u}_h^0(\mu)$, we construct the sequence $\mathbf{u}_h^k(\mu) \equiv (u_h^k(\mu), \lambda_h^k(\mu)) \in \mathcal{Y}_h$ where

$$u_h^k(\mu) = \alpha u_h^{k-1}(\mu) + (1 - \alpha) u_h^f(\mu), \tag{4.16}$$

α is an adjustable parameter to improve convergence, and $u_h^f(\mu)$ satisfies

$$\begin{aligned} a_1(u_h^f(\mu), v) + \mu^2 \int_\Omega p(u_h^{k-1}(\mu)) u_h^f v &= \lambda_h^k(\mu) m(u_h^f(\mu), v), \quad \forall v \in Y_h, \\ m(u_h^f(\mu), u_h^f(\mu)) &= 1. \end{aligned} \tag{4.17}$$

Therefore, at each step k, we must solve a linear algebraic eigenvalue problem of dimension $\mathcal{N}$. The procedure is repeated until convergence criteria (4.15) is satisfied.

As was done in Chapter 2, the reduced basis approximation will be built upon this "truth" approximation and we will drop the subscript h from all subsequent formulation, with the understanding that the "truth" approximation in fact refers to the finite element approximation. Thus, $\mathcal{Y}$, $\mathcal{A}$, $\delta\mathcal{A}$, u and λ shall now be understood as $\mathcal{Y}_h$, $\mathcal{A}_h$, $\delta\mathcal{A}_h$, u_h and λ_h.

4.2.3 Numerical Example

We consider the following nonlinear eigenvalue problem defined on $\Omega \equiv]0,1[\subset \mathbb{R}$: given $\mu \in \mathcal{D} \equiv [1,10]$, we evaluate $\mathbf{u} \in \mathcal{Y}$ from (4.6) where $Y \equiv H_0^1(\Omega)$,

$$a_1(w,v) = \frac{1}{2}\int_\Omega \frac{dw}{dx}\frac{dv}{dx}\,dx, \quad m(w,v) = \int_\Omega w\,v\,dx, \tag{4.18}$$

and the nonlinear function

$$g(w;x;\mu) = p(w)w, \quad p(w) = |w|^{7/3}. \tag{4.19}$$

The $H_0^1(\Omega) \subset H^1(\Omega)$ is the usual Hilbert space of derivative square-integrable functions that vanish on the domain boundary. The derivative of $g(w;x;\mu)$ with respect to w is positive and given by

$$\begin{aligned} g'(w;x;\mu) &= \frac{7}{3}|w|^{4/3}\mathrm{sgn}(w)w + |w|^{7/3} \\ &= \frac{10}{3}|w|^{7/3}. \end{aligned} \tag{4.20}$$

We solve the resulting problem using the finite element method with $\mathcal{N} = 400$, employing both the Newton iterative scheme and the fixed point method. Table 4.1 shows that the Newton iterative scheme is clearly more efficient than the fixed point method for this particular example.

	Newton scheme		fixed point method		
μ	$k_{\max}$	time, s	α	$k_{\max}$	time,s
1	4	0.4	0	9	0.8
10	6	0.4	0.95	388	32.9

Table 4.1: Comparison of the computational cost of the fixed point method and the Newton iterative scheme in "truth" approximation.

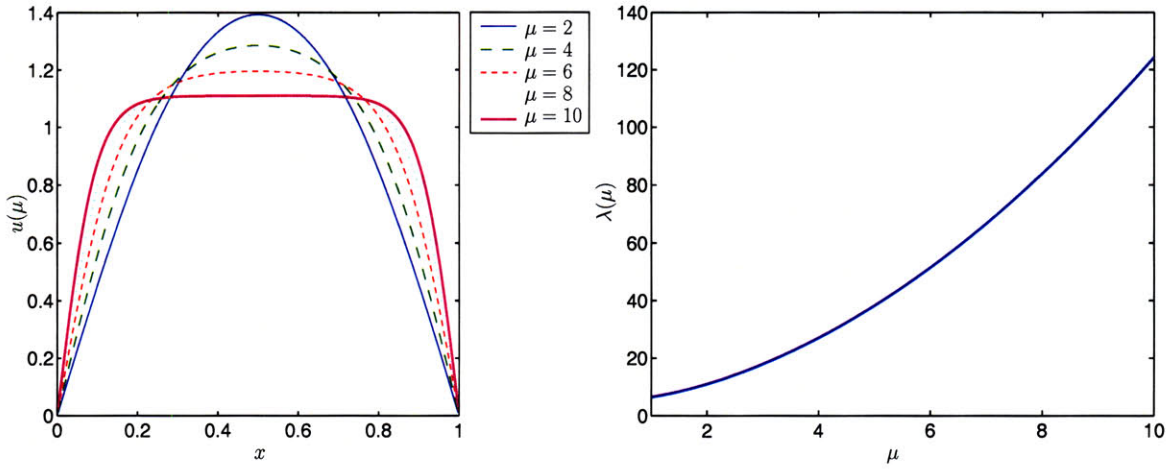

Figure 4-1: Variation of $u(\mu)$ and $\lambda(\mu)$ with μ for $\mu \in \mathcal{D} \equiv [1, 10]$.

For $\mu = 1$, Newton iterative scheme converges in half the time required by fixed point method. This is expected since the Newton iterative scheme converges quadratically while the fixed point method only converges linearly. In addition, as μ increases, α, the mixing parameter in (4.16), must be sufficiently large for the fixed point method to converge, severely affecting the convergence rate of the fixed point method. For $\mu = 10$, we have used $\alpha = 0.95$, and as a result, the computational cost of the Newton iterative scheme outperforms the fixed point method by a factor of 80.

For this example, it is clear that the method of choice for our "truth" approximation is the Newton iterative scheme. Nevertheless, for more complicated problems, for example those encountered in Chapter 6 and 7 where we must solve for several eigensolutions, the advantage of the Newton iterative scheme over the fixed point method is not as straightforward. In particular, the addition of the orthonormality constraints may lead to an ill-conditioned Jacobian matrix; continuation scheme coupled with preconditioner for the Jacobian matrix may be required to achieve convergence.

The resulting solutions $u(\mu)$ and $\lambda(\mu)$ for some selected μ are shown in Figure 4-1. We notice nontrivial variation of $u(\mu)$ with μ.

4.3 Reduced Basis Approximation

4.3.1 Classical Approach

The space and the approximation

We first introduce nested sample sets $S_N^u = (\mu_1, \ldots, \mu_N)$, $1 \le N \le N_{\max}$ and define the associated nested reduced-basis spaces as

$$\begin{aligned} W_N^u &= \text{span } \{u(\mu_j), 1 \le j \le N\}, \quad 1 \le N \le N_{\max}, \\ &= \text{span } \{\zeta_j, 1 \le j \le N\}, \quad 1 \le N \le N_{\max}; \end{aligned} \tag{4.21}$$

where $u(\mu_j)$ are solutions of (4.6) at $\mu = \mu_j$; and ζ_j, $1 \le j \le N$ are basis functions obtained after $u(\mu_j)$, $1 \le j \le N$ are orthogonalized.

The approximations to $u(\mu)$ and $\lambda(\mu)$ are then obtained by a standard Galerkin projection: given a μ, find $\mathbf{u}_N \equiv (u_N(\mu), \lambda_N(\mu)) \in \mathcal{Y}_N \equiv W_N^u \times \mathbb{R}$ such that

$$\mathcal{A}(\mathbf{u}_N, \mathbf{v}_N; \mu) = 0, \quad \forall \mathbf{v}_N \in \mathcal{Y}_N, \tag{4.22}$$

where $\mathcal{A}$ is as defined in (4.7).

Discrete Equations

We now demonstrate how the nonlinearity in $a^{\text{nl}}(\cdot\,, \cdot\,; \mu)$ affects the efficiency of the offline-online computational strategy. We expand our reduced-basis approximation as

$$u_N(\mu) = \sum_{j=1}^{N} u_{N\,j}(\mu)\zeta_j. \tag{4.23}$$

Inserting this representation into (4.22) and choose as test functions $v = \zeta_i$, $1 \le i \le N$ yields

$$\sum_{j=1}^{N} u_{N\,j}(\mu) a_1(\zeta_j, \zeta_i) + \mu^2 \int_\Omega g\left(\sum_{j'=1}^{N} u_{N\,j'}(\mu)\zeta_{j'}\right)\zeta_i \;=\; \lambda_N(\mu)\sum_{j=1}^{N} u_{N\,j}(\mu) m(\zeta_j, \zeta_i), \quad 1 \le i \le N, \tag{4.24}$$

$$\sum_{j=1}^{N}\sum_{i=1}^{N} u_{N\,j}(\mu) u_{N\,i}(\mu) m(\zeta_j, \zeta_i) \;=\; 1. \tag{4.25}$$

If we now apply a Newton iterative scheme to solve (4.24)–(4.25), then in each Newton iteration and given a current iterate $\bar{u}_{N\,j}(\mu), 1 \le j \le N$, and $\bar{\lambda}_N(\mu)$ we must find an increment $\delta u_{N\,j}(\mu), 1 \le j \le N$, and $\delta\lambda_N(\mu)$ such that

$$\begin{aligned}\sum_{j=1}^{N} \delta u_{N\,j}(\mu)\Big\{ a_1(\zeta_j, \zeta_i) + \mu^2 \int_\Omega g'\left(\sum_{j'=1}^{N} \bar{u}_{N\,j'}(\mu)\zeta_{j'}\right)\zeta_j\,\zeta_i \\ - \bar{\lambda}_N(\mu) m(\zeta_j, \zeta_i)\Big\} - \delta\lambda_N(\mu)\sum_{j=1}^{N} \bar{u}_{N\,j}(\mu) m(\zeta_j, \zeta_i) \;=\; \\ - \sum_{j=1}^{N} \bar{u}_{N\,j}(\mu) a_1(\zeta_j, \zeta_i) - \mu^2 \int_\Omega g\left(\sum_{j'=1}^{N} \bar{u}_{N\,j'}(\mu)\zeta_{j'}\right)\zeta_i \\ + \bar{\lambda}(\mu)\sum_{j=1}^{N} \bar{u}_{N\,j}(\mu) m(\zeta_j, \zeta_i), \quad 1 \le i \le N\end{aligned} \tag{4.26}$$

and

$$2\sum_{j=1}^{N}\sum_{i=1}^{N} \delta u_{N\,j}(\mu)\bar{u}_{N\,i}(\mu) m(\zeta_j, \zeta_i) = -\sum_{j=1}^{N}\sum_{i=1}^{N} \bar{u}_{N\,j}(\mu)\bar{u}_{N\,i}(\mu) m(\zeta_j, \zeta_i) + 1. \tag{4.27}$$

If g and g' are quadratic polynomial nonlinearities in u, an expansion into their power series will recover the affine parameter dependence, allowing an efficient offline-online computational decomposition. Unfortunately, for higher polynomial nonlinearity, such expansion leads to complex offline construction of the required matrices and potentially large online computational cost, while for non-polynomial nonlinearities, it does not exist. As such, the evaluations of $\int_\Omega g\left(\sum_{j'=1}^{N} \bar{u}_{N\,j'}(\mu)\zeta_{j'}\right)\zeta_i$ and $\int_\Omega g'\left(\sum_{j'=1}^{N} \bar{u}_{N\,j'}(\mu)\zeta_{j'}\right)\zeta_j\,\zeta_i$, in general, incur a cost of $O(\mathcal{N})$.

4.3.2 Empirical Interpolation Method

To obtain online computational cost that is independent of $\mathcal{N}$, we replace $g(w;\,\cdot\,;\mu)$ by $g_M^w(\,\cdot\,;\mu)$, an approximation based on the empirical interpolation method described in Chapter 3. We first introduce nested sample sets

$$S_M^g = \{\mu_m^g \in \mathcal{D}, 1 \le m \le M\}, \quad 1 \le M \le M_{\max}, \tag{4.28}$$

the associated approximation spaces,

$$\begin{aligned} W_M^g &= \text{span}\ \{g(x;\mu_m^g), 1 \le m \le M\} \quad 1 \le M \le M_{\max} \\ &= \text{span}\ \{q_m^g, 1 \le m \le M\}, \quad 1 \le M \le M_{\max}, \end{aligned} \tag{4.29}$$

and the interpolation points,

$$T_M^g \equiv \{t_m^g, 1 \le m \le M\}, \quad 1 \le M \le M_{\max}, \tag{4.30}$$

where $M_{\max}$ is determined by the maximum accuracy to which we would like to approximate $g(\cdot\,;\cdot\,;\mu)$. The construction procedure is detailed in Section 3.2.3. The approximation of $g(w;\,\cdot\,;\mu)$, $g_M^w(\,\cdot\,;\mu)$, is then given by

$$g_M^w(\,\cdot\,;\mu) = \sum_{j=1}^{M} \alpha_{M\,j}(\mu) q_j^g(\,\cdot\,)\ , \tag{4.31}$$

where $\alpha_M(\mu) \in \mathbb{R}^M$ is given by

$$\sum_{j=1}^{M} q_j^g(t_i^g)\alpha_{M\,j}(\mu) = g(w(t_i^g);t_i^g;\mu), \quad i = 1,\ldots,M\ . \tag{4.32}$$

Although this "composed" interpolant is defined for general $w \in Y$, we expect good approximation only for w (very) close to the manifold $\mathcal{M} \equiv \{u(\mu), \mu \in \mathcal{D}\}$ on which W_M^g is constructed.

Our reduced basis approximation is thus: given $\mu \in \mathcal{D}$, find $\mathbf{u}_{N,M}(\mu) \equiv (u_{N,M}(\mu), \lambda_{N,M}(\mu)) \in \mathcal{Y}_N$ such that

$$\mathcal{A}_M(\mathbf{u}_{N,M}(\mu), \mathbf{v}_{N,M};\mu) = 0, \quad \forall \mathbf{v}_{N,M} \in \mathcal{Y}_N, \tag{4.33}$$

where

$$\mathcal{A}_M(\mathbf{w} \equiv (w,\sigma), \mathbf{v} \equiv (v,\varphi);\mu) \equiv a_1(w,v) + \mu^2 a^{\mathrm{nl,M}}(w,v;\mu) - \sigma m(w,v) + \varphi(m(w,w)-1), \quad (4.34)$$

and

$$a^{\mathrm{nl,M}}(w,v;\mu) \equiv \int_\Omega g_M^w(\,\cdot\,;\mu)\, v. \quad (4.35)$$

The subscript N,M emphasizes the fact that the solution $\mathbf{u}_{N,M}(\mu)$ is affected by two discretization procedures — this will be further elaborated in Section 4.3.4.

4.3.3 Offline-online Computational Framework

We now demonstrate how the incorporation of the empirical interpolation procedure into the reduced basis framework leads to an efficient offline-online computational strategy. We give the implementation details of the Newton iterative scheme and the fixed point method within this offline-online computational framework.

Newton iterative scheme

As was done in the classical reduced basis approximation, we first expand our reduced basis approximation as

$$u_{N,M}(\mu) = \sum_{j=1}^{N} u_{N,M\,j}(\mu)\zeta_j. \quad (4.36)$$

In addition, we expand our empirical interpolation approximation for $g(u_{N,M}(\mu);x;\mu)$ as

$$g_M^{u_{N,M}}(x;\mu) = \sum_{m=1}^{M} \varrho_{M\,m}(\mu) q_m^g(x)\ , \quad (4.37)$$

where $\varrho_M(\mu) \in \mathbb{R}^M$ is given by

$$\begin{aligned} \sum_{k=1}^{M} B_{m,k}^M \varrho_{M,k}(\mu) &= g(u_{N,M}(t_m^g;\mu);t_m^g;\mu), \quad 1 \le m \le M \\ &= g\left(\sum_{n=1}^{N} u_{N,M\,n}(\mu)\zeta_n(t_m^g);t_m^g;\mu\right), \quad 1 \le m \le M\ ; \end{aligned} \quad (4.38)$$

and $B^M \in \mathbb{R}^M \times \mathbb{R}^M$ is given by $B^M_{m,k} = q^g_m(t^g_k)$, $1 \le m, k \le M$. We note that $\{q^g_m, 1 \le m \le M\}$ is pre-constructed offline based on the empirical interpolation method, but $\varrho_M(\mu)$ is computed online given any $\mu \in \mathcal{D}$.

Inserting the above representations (4.36) and (4.37) into (4.34), we obtain the following discrete equations

$$\sum_{j=1}^{N} A^N_{i,j} u_{N,M\,j}(\mu) + \mu^2 \sum_{m=1}^{M} C^{N,M}_{i,m} \varrho_{M\,m}(\mu) = \lambda_{N,M}(\mu) \sum_{j=1}^{N} M^N_{i,j} u_{N,M\,j}(\mu), \quad 1 \le i \le N \tag{4.39}$$

$$\sum_{i=1}^{N}\sum_{j=1}^{N} u_{N,M\,i}(\mu) M^N_{i,j} u_{N,M\,j}(\mu) = 1\ ; \tag{4.40}$$

where $A^N \in \mathbb{R}^{N\times N}$, $M^N \in \mathbb{R}^{N\times N}$, $C^{N,M} \in \mathbb{R}^{N\times M}$ are given by $A^N_{i,j} = a_1(\zeta_j, \zeta_i)$, $1 \le i, j \le N$, $M^N_{i,j} = m(\zeta_j, \zeta_i)$, $1 \le i, j \le N$, and $C^{N,M}_{i\,m} = \int_\Omega q^g_m \zeta_i$, $1 \le i \le N, 1 \le m \le M$, respectively.

We then substitute $\varrho_M(\mu)$ from (4.38) into (4.39) to obtain the following

$$\sum_{j=1}^{N} A^N_{i,j} u_{N,M\,j}(\mu) + \mu^2 \sum_{m=1}^{M} D^{N,M}_{i,m}\, g\left(\sum_{n=1}^{N} \zeta_n(t^g_m) u_{N,M\,n}(\mu); t^g_m; \mu\right) = \lambda_{N,M}(\mu) \sum_{j=1}^{N} M^N_{i,j} u_{N,M\,j}(\mu), \quad 1 \le i \le N\ , \tag{4.41}$$

where $D^{N,M} = C^{N,M}(B^M)^{-1} \in \mathbb{R}^{N\times M}$.

We now solve (4.41) for $u_{N,M\,j}(\mu), 1 \le j \le N$ by Newton iterative scheme: given a current iterate $\bar{u}_{N,M\,j}(\mu), 1 \le j \le N$, and $\bar{\lambda}_{N,M}(\mu)$ we must find an increment $\delta u_{N,M\,j}(\mu), 1 \le j \le N$, and $\delta\lambda_{N,M}(\mu)$ such that

$$\begin{aligned}\sum_{j=1}^{N} \left(A^N_{i\,j} + \mu^2 \bar{E}^N_{i\,j} + \bar{\lambda}_{N,M}(\mu) M^N_{i\,j}\right) \delta u_{N,M\,j}(\mu) - \delta\lambda_{N,M}(\mu) \sum_{k=1}^{N} M^N_{i\,k} \bar{u}_{N,M\,k}(\mu) = \\ - \sum_{j=1}^{N} A^N_{i,j} \bar{u}_{N,M\,j}(\mu) - \mu^2 \sum_{m=1}^{M} D^{N,M}_{i,m}\, g\left(\sum_{n=1}^{N} \zeta_n(t^g_m) \bar{u}_{N,M\,n}(\mu); t^g_m; \mu\right) \\ + \bar{\lambda}_{N,M}(\mu) \sum_{j=1}^{N} M^N_{i,j} \bar{u}_{N,M\,j}(\mu), \quad 1 \le i \le N\ ;\end{aligned} \tag{4.42}$$

and

$$2\sum_{i=1}^{N}\sum_{j=1}^{N}\bar{u}_{N,M\,i}(\mu)M^N_{i,j}\delta u_{N,M\,j}(\mu) = -\sum_{i=1}^{N}\sum_{j=1}^{N}\bar{u}_{N,M\,i}(\mu)M^N_{i,j}\bar{u}_{N,M\,j}(\mu) + 1. \tag{4.43}$$

Here $\bar{E}^N \in \mathbb{R}^{N\times N}$ must be calculated at every Newton iteration as

$$\bar{E}^N_{i,j} = \sum_{m=1}^{M} D^{N,M}_{i,m} g'\left(\sum_{n=1}^{N}\zeta_n(t^g_m)\bar{u}_{N,M\,n}(\mu); t^g_m; \mu\right)\zeta_j(t^g_m), \quad 1 \le i,j \le N\ , \tag{4.44}$$

where $g'(w;t;\mu)$ is the first derivative of g with respect to w. We can now develop an efficient offline-online procedure for the rapid evaluation of $\lambda_{N,M}(\mu)$ for each μ in $\mathcal{D}$.

In the offline stage — performed once — we generate nested reduced-basis spaces $W^u_N = \{\zeta_1,\ldots,\zeta_N\}, 1 \le N \le N_{\max}$, nested approximation spaces $W^g_M = \{q^g_1,\ldots,q^g_M\}, 1 \le M \le M_{\max}$, and nested sets of interpolation points $T^g_M = \{t^g_1,\ldots,t^g_M\}, 1 \le M \le M_{\max}$; we then form and store A^N, M^N, B^M, and $D^{N,M}$.

In the online stage — performed many times for each new μ — we solve (4.44) for $u_{N,M\,j}(\mu), 1 \le j \le N$. 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 that we perform the sum in the parenthesis of (4.44) before performing the outer sum — the evaluation of $\sum_{n=1}^{N}\zeta_n(t^g_m)\bar{u}_{N,M\,n}(\mu)$ is of $O(MN)$ while the evaluation of the outer summation is only of $O(M)$. 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.

Fixed point method

Similarly, we first expand our reduced basis approximation as

$$u_{N,M}(\mu) = \sum_{j=1}^{N} u_{N,M\,j}(\mu)\zeta_j. \tag{4.45}$$

Instead of constructing an empirical interpolation approximation for $g(u_{N,M}(\mu);\,\cdot\,;\mu)$, we construct an empirical interpolation approximation for $p(u_{N,M}(\mu);\,\cdot\,;\mu) = |u_{N,M}(\mu)|^{7/3}$. We first construct nested sample sets $S^p_M \equiv \{\mu^p_1,\ldots,\mu^p_M\}$, $1 \le M \le M_{\max}$, nested approximation spaces $W^p_M \equiv \text{span}\ \{q^p_1,\ldots,q^p_M\}$, $1 \le M \le M_{\max}$, and nested interpolation points $T^p_M \equiv \{t^p_1,\ldots,t^p_M\}$,

$1 \leq M \leq M_{\max}$. We then expand our empirical interpolation approximation for $p(u_{N,M}(\mu); x; \mu)$ as

$$p_M^{u_{N,M}}(x;\mu) = \sum_{m=1}^{M} \beta_{M,m}(\mu) q_m^p(x) \ , \tag{4.46}$$

where $\beta_M(\mu) \in \mathbb{R}^M$ is given by

$$\begin{aligned} \sum_{k=1}^{M} B_{m,k}^{M,p} \beta_{M\,k}(\mu) &= p(u_{N,M}(t_m^p;\mu); t_m^p; \mu), \quad 1 \leq m \leq M \\ &= p\left(\sum_{n=1}^{N} u_{N,M\,n}(\mu)\zeta_n(t_m^p); t_m^p; \mu\right), \quad 1 \leq m \leq M \ ; \end{aligned} \tag{4.47}$$

and $B^{M,p} \in \mathbb{R}^M \times \mathbb{R}^M$ is given by $B_{m,k}^{M,p} = q_m^p(t_k^p)$, $1 \leq m, k \leq M$. We note again that $\{q_m^p, 1 \leq m \leq M\}$ is pre-constructed offline based on the empirical interpolation method, but $\beta_M(\mu)$ is computed online given any $\mu \in \mathcal{D}$. Inserting the above representations (4.45) and (4.46) into (4.34), we obtain the following discrete equations

$$\sum_{j=1}^{N} A_{i,j}^N u_{N,M\,j}(\mu) + \mu^2 \left(\sum_{m=1}^{M} C^{N,p,m} \beta_{M\,m}(\mu)\right) u_{N,M\,j}(\mu) \quad =$$

$$\lambda_{N,M}(\mu) \sum_{j=1}^{N} M_{i,j}^N u_{N,M\,j}(\mu), \quad 1 \leq i \leq N \ ; \tag{4.48}$$

$$\sum_{i=1}^{N}\sum_{j=1}^{N} u_{N,M\,i}(\mu) M_{i,j}^N u_{N,M\,j}(\mu) = 1 \ ; \tag{4.49}$$

where $A^N \in \mathbb{R}^{N\times N}$, $M^N \in \mathbb{R}^{N\times N}$, $C^{N,p,m} \in \mathbb{R}^{N\times N}$, $1 \leq m \leq M$ are given by $A_{i,j}^N = a_1(\zeta_j, \zeta_i)$, $1 \leq i,j \leq N$, $M_{i,j}^N = m(\zeta_j, \zeta_i)$, $1 \leq i,j \leq N$, and $C_{i,j}^{N,p,m} = \int_\Omega q_m^p \zeta_j \zeta_i$, $1 \leq i,j \leq N$, $1, \leq m \leq M$ respectively.

The offline-online computational decomposition is then clear. In the offline stage — performed once — we generate nested reduced-basis spaces $W_N^u = \{\zeta_1, \ldots, \zeta_N\}, 1 \leq N \leq N_{\max}$, nested approximation spaces $W_M^p = \{q_1^p, \ldots, q_M^p\}, 1 \leq M \leq M_{\max}$, and nested sets of interpolation points $T_M^p = \{t_1^p, \ldots, t_M^p\}, 1 \leq M \leq M_{\max}$; we then form and store $A^N, M^N, B^{M,p}$, and $C^{N,p,m}$, $1 \leq m \leq M_{\max}$.

In the online stage — performed many times for each new μ — we solve (4.48) – (4.49) for $u_{N,M\,j}(\mu), 1 \leq j \leq N$. We begin with an initial guess $u_{N,M}^0(\mu)$, and construct a sequence of

$u^k_{N,M}(\mu)$ given by

$$u^k_{N,M}(\mu) = \alpha u^{k-1}_{N,M}(\mu) + (1-\alpha) u^f_{N,M}(\mu), \tag{4.50}$$

where $u^f_{N,M}(\mu)$ satisfies

$$\sum_{j=1}^{N} A^N_{i,j} u^f_{N,M\,j}(\mu) + \mu^2 \left(\sum_{m=1}^{M} C^{N,p,m} \beta^k_{M\,m}(\mu) \right) u^f_{N,M\,j}(\mu) \;=\; \lambda^k_{N,M}(\mu) \sum_{j=1}^{N} M^N_{i,j} u^f_{N,M\,j}(\mu), \quad 1 \le i \le N\,; \tag{4.51}$$

$$\sum_{i=1}^{N}\sum_{j=1}^{N} u^f_{N,M\,i}(\mu) M^N_{i,j} u^f_{N,M\,j}(\mu) \;=\; 1\,; \tag{4.52}$$

and $\beta^k_{M\,m}(\mu)$ is given by

$$\sum_{\ell=1}^{M} B^{M,p}_{m,k} \beta^k_{M\,\ell}(\mu) \;=\; p\Big(\sum_{n=1}^{N} u^{k-1}_{N,M\,n}(\mu) \zeta_n(t^p_m); t^p_m; \mu \Big), \quad 1 \le m \le M\,. \tag{4.53}$$

The operation count of the online stage is essentially the cost of solving linear eigenvalue problems: at each fixed point iteration, we first determine $\beta^k_M(\mu)$ at a cost of $O(M^2)$ and assemble $\sum_{m=1}^{M} \tilde{C}^{N,M,p,m} \beta^{k-1}_{M\,m}(\mu)$ at a cost of $O(MN^2)$; we then solve the resulting linear algebraic eigenvalue problem at a cost of $O(N^3)$. The online complexity depends only on N, M and number of fixed point iterations; we thus recover online $\mathcal{N}$ independence.

Comparison between Newton iterative scheme and fixed point method

We first note that the computational cost of a single Newton iteration and a single fixed point iteration is similar. However, from Table 4.2, it is clear that the number iterations required to achieve convergence is much higher for the fixed point method than for the Newton iterative scheme, resulting in a much higher computational cost. The result is similar to that observed in Section 4.2.3. In particular, for higher μ, the mixing parameter α in (4.50) has to be large in order to achieve any convergence. For $\mu = 10$, the Newton iterative scheme is 70 times faster than the fixed point method.

The choice for our online solution method for this example is clear — the Newton iterative scheme is more efficient that fixed point method. However, for the same reasons given in Sec-

	Newton scheme		fixed point method		
μ	$k_{\max}$	time,s	α	$k_{\max}$	time, s
2	4	0.001	0.75	58	0.01
8	5	0.001	0.95	273	0.07

Table 4.2: Comparison of computational cost of the fixed point method and the Newton iterative scheme in reduced basis approximation. Here $N = 6$ and $M = 12$.

tion 4.2.3, this may not be true for more difficult problems, such as those encountered in Section 6 and 7.

4.3.4 Convergence

We first define an error measure $\varepsilon^{\lambda}_{N,M}$ as

$$\varepsilon^{\lambda}_{N,M} = \max_{\mu \in \Xi_T} \left| \frac{\lambda_{N,M}(\mu) - \lambda(\mu)}{\lambda(\mu)} \right|, \tag{4.54}$$

where $\Xi_T \subset \mathcal{D}$ is a test sample set of size 100 with the sample points uniformly distributed in $\mathcal{D}$. We now evaluate $\varepsilon^{\lambda}_{N,M}$ for different N and M. From Figure 4-2, we observe that the reduced basis approximation converges very rapidly with N and M. In addition, the quality of our reduced basis approximation depends on N and M in a strongly coupled manner: for a fixed value of M, the error decreases monotonically with N for $N \leq N_M$, where N_M is such that there is no appreciable change in $\varepsilon^{\lambda}_{N,M}$ for $N > N_M$. However, when M is increased, the achievable $\varepsilon^{\lambda}_{N,M}$ decreases further as N_M increases; this strongly suggests that the reduced-basis error may obviously be degraded by a poor initial choice of M; we can always choose M large enough that the incurred error does not affect our desired accuracy. The choice for the sample set is discussed next.

4.3.5 Construction of Samples

We must now construct two sample sets S^g_M and S^u_N and the associated approximation spaces, W^g_M and W^u_N. We first construct S^g_M, W^g_M and T^g_M (defined in Section 4.3.2) based on the algorithm given in Section 3.2.3. It is however an expensive process. Before we can apply the adaptive sampling procedure, we must first compute and store the solution $u(\mu)$ for all $\mu \in \Xi_T$. For our example, this is equivalent to performing 100 $O(\mathcal{N}^{\bullet})$ solves and saving 100 solutions each of length $\mathcal{N}$. This inefficiency can be attributed to the lack of an *a posteriori* error estimator that does not

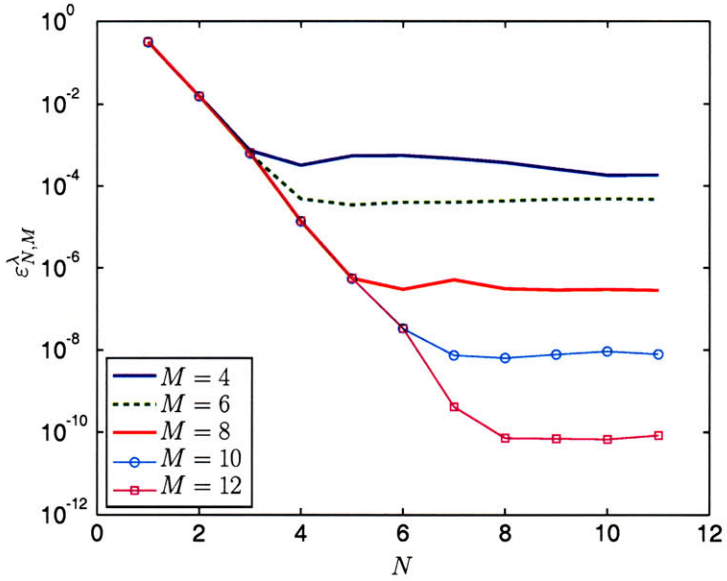

Figure 4-2: Variations of the reduced basis error in $\lambda_{N,M}$, $\varepsilon^{\lambda}_{N,M}$ (given by (4.54)), with N for $M = 4, 6, 8, 10$ and 12.

require knowing the approximation of $g(\,\cdot\,;\,\cdot\,;\mu)$ beforehand.

Once we have constructed the approximation for $g(\cdot;\cdot;\mu)$, we can proceed with the construction S^u_N and W^u_N based on the procedure outlined in Chapter 2. Note that the accuracy that can be achieved is limited by the accuracy of $g_M^{u_{N,M}}(\,\cdot\,;\mu)$; W^g_M must initially be constructed such that the accuracy in $g_M^{u_{N,M}}(\,\cdot\,;\mu)$ is equivalent the desired accuracy of $\lambda_{N,M}(\mu)$. In addition, since we have solutions at all $\mu \in \Xi_T$, we can use the exact error in $\lambda_{N,M}(\mu)$ given by $\frac{|\lambda_{N,M}(\mu)-\lambda(\mu)|}{|\lambda(\mu)|}$ as our error measure when applying the adaptive sampling algorithm. We also do not have an *a posteriori* error estimator for nonlinear eigenvalue problem at this point.

Chapter 5

Band Structure Calculations

5.1 Introduction

We now consider the application of the reduced basis method to the calculation of the band structure of a periodic solid, a problem commonly encountered in solid state physics. This problem can arise from a semi-empirical model of crystalline solids [29], within the inner loop of the self-consistent field algorithm when solving *ab initio* models [18, 37], or as a post-processing step following a calculation based on *ab initio* models [109, 116].

The method used here is similar to that developed in Chapter 2, although with an increase in complexity due to less well-behaved solutions and a richer parameter domain. We shall describe how we handle the resulting difficulties. We shall demonstrate how the significant improvement in the computational efficiency can lead to rapid determination of three quantities relevant to study of crystalline solids, namely the integrated density of states, the joint density of states and the dielectric function. In what follows, we utilize the empirical pseudopotential model [29] in our calculations. However, we can easily extend the methodology to any other model.

5.2 Abstract Formulation

5.2.1 Preliminaries

We give some basic materials needed to understand the problem we intend to solve. The materials are adapted from [3, 60].

Crystal Lattices

A three dimensional Bravais lattice consists of all points with position vectors $\mathbf{R}$ defined by

$$\mathbf{R} = \sum_{i=1}^{3} \ell_i \mathbf{a}_i, \quad \ell_i \in \mathbb{Z}. \tag{5.1}$$

where $\mathbf{a}_1$, $\mathbf{a}_2$ and $\mathbf{a}_3$ are three independent vectors in $\mathbb{R}^3$ and are known as the primitive vectors — $\mathbf{a}_1$, $\mathbf{a}_2$ and $\mathbf{a}_3$ span the Bravais lattice. A primitive unit cell Ω is a volume of space that when translated through all the vectors in a Bravais lattice, fills the whole space without either overlapping itself or leaving voids. There is no unique way of defining Ω but an obvious choice is given by that enclosed by the primitive vectors: any points $\mathbf{x} \in \Omega$ can be represented by

$$\mathbf{x} = \sum_{i=1}^{3} x_i \mathbf{a}_i, \tag{5.2}$$

where $0 \leq x_i \leq 1$, $i = 1$, 2 and 3.

A crystal is described by its underlying Bravais lattice and the locations of atoms within the primitive cell Ω given by the basis vectors $\tau_1, \ldots, \tau_{n_\tau}$ where n_τ is number of atoms in Ω. In Figure 5-1 we show a unit cell of diamond structure — a face center cubic Bravais lattice with two basis vectors: given a lattice length a, the primitive lattice vectors and the the basis vectors are defined as follows:

$$\mathbf{a}_2 = \frac{a}{2}(0,1,1), \quad \mathbf{a}_2 = \frac{a}{2}(1,0,1), \quad \mathbf{a}_3 = \frac{a}{2}(1,1,0); \tag{5.3}$$

$$\tau_1 = -\sum_{i=1}^{3} \frac{1}{8}\mathbf{a}_i, \quad \tau_2 = \sum_{i=1}^{3} \frac{1}{8}\mathbf{a}_i. \tag{5.4}$$

The reciprocal space is the dual of the discrete linear space spanned by the Bravais lattices $\mathbf{R}$. It is spanned by the reciprocal lattice vectors given by

$$\mathbf{G} = \sum_{i=1}^{3} m_i \mathbf{b}_i, \quad m_i \in \mathbb{Z}, \tag{5.5}$$

where $\mathbf{b}_1$, $\mathbf{b}_2$ and $\mathbf{b}_3$ are independent vectors in $\mathbb{R}^3$ defined by

$$\mathbf{a}_i \cdot \mathbf{b}_j = 2\pi\delta_{ij}, \quad 1 \leq i, j \leq 3. \tag{5.6}$$

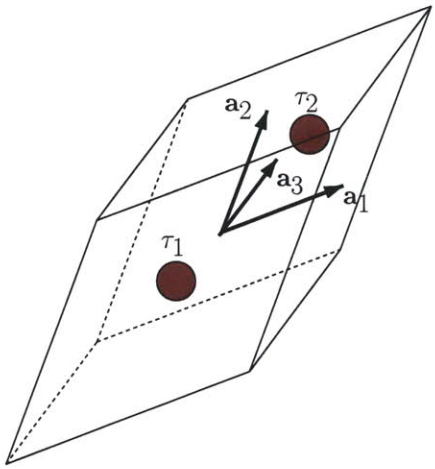

Figure 5-1: The unit cell of diamond structure, The primitive lattice vectors are given by $\mathbf{a}_1 = a(0,1,1)$,$\mathbf{a}_2 = a(1,0,1)$ and $\mathbf{a}_3 = a(1,1,0)$. In addition, the basis vectors of the nuclei are given by $\tau_1 = -\sum_{i=1}^{3} \mathbf{a}_i/8$, and $\tau_2 = -\sum_{i=1}^{3} \mathbf{a}_i/8$.

For the fcc Bravais lattice defined by the primitive lattice vectors in (5.3), the reciprocal lattice vectors are given by

$$\mathbf{b}_2 = \frac{2\pi}{a}(-1,1,1), \quad \mathbf{b}_2 = \frac{2\pi}{a}(1,-1,1), \quad \mathbf{b}_3 = \frac{2\pi}{a}(1,1,-1). \tag{5.7}$$

The first Brillouin zone (BZ) is the primitive cell with the full symmetry of the reciprocal lattice vectors. The first Brillouin zone about a lattice point is the region of space that is closer to that point than to any other lattice point. The first Brillouin zone of fcc Bravais lattice is shown in Figure 5-2.

Bloch's theorem

Theorem 5.1. *The eigenstates $\psi_i(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^3$ of the one-electron Schrödinger equation given by*

$$-\frac{1}{2}\Delta\psi_i(\mathbf{x}) + V_{\text{eff}}(\mathbf{x})\psi_i(\mathbf{x}) = E_i\psi_i(\mathbf{x}), \quad i = 1,2,\ldots, \tag{5.8}$$

$$\int_\Omega \psi_i(\mathbf{x})\psi_j(\mathbf{x}) = \delta_{ij}, \quad i = 1,2,\ldots, \quad j = 1,2,\ldots, \tag{5.9}$$

where $V_{\text{eff}}(\mathbf{x}+\mathbf{R}) = V_{\text{eff}}(\mathbf{x})$ for all $\mathbf{R}$ in a Bravais lattice, can be chosen to have the following form:

$$\psi_i(\mathbf{x};\mathbf{k}) = e^{\mathbf{i}\mathbf{k}\mathbf{x}}u_i(\mathbf{x};\mathbf{k}), \quad u_i(\mathbf{x}+\mathbf{R};\mathbf{k}) = u_i(\mathbf{x};\mathbf{k}), \tag{5.10}$$

where the Bloch wavevectors, $\mathbf{k} \in \mathbb{R}^3$, belongs to the first Brillouin zone BZ due to the periodicity of the Bravais lattice [3].

The form (5.10) leads to a parameterized from of (5.11):

$$\left(-\frac{1}{2}\Delta - \mathrm{i}\mathbf{k}\nabla + \frac{|\mathbf{k}|^2}{2} + V_{\text{eff}}(\mathbf{x})\right) u_i(\mathbf{x};\mathbf{k}) = E_i(\mathbf{k}) u_i(\mathbf{x};\mathbf{k}), \quad i = 1, 2, \ldots, \tag{5.11}$$

$$\int_\Omega u_i(\mathbf{x};\mathbf{k}) u_j(\mathbf{x};\mathbf{k}) = \delta_{ij}, \quad i = 1, 2, \ldots, \quad j = 1, 2, \ldots, \tag{5.12}$$

with the periodic conditions $u_i(\mathbf{x} + \mathbf{R};\mathbf{k}) = u_i(\mathbf{x};\mathbf{k})$; and $\mathbf{x} \in \mathbb{R}^3$. In addition, due to symmetry of the crystal, the wavevector $\mathbf{k}$ can be further confined to a subset of the first Brillouin zone, called the irreducible Brillouin zone. The irreducible Brillouin zone of fcc structure is shown in Figure 5-2 — it is a polyhedron with vertices at L $\equiv \frac{2\pi}{a}(1/2, 1/2, 1/2)$, $\Gamma \equiv \frac{2\pi}{a}(0, 0, 0)$, X $\equiv \frac{2\pi}{a}(1, 0, 0)$, W $\equiv \frac{2\pi}{a}(3/4, 3/4, 0)$, K $\equiv \frac{2\pi}{a}(1, 0, 1/2)$, and U $\equiv \frac{2\pi}{a}(1, 1/4, 1/4)$. These points are known as the high-symmetry points because they exhibit many symmetric properties of the crystal[1]. The edges of the polyhedron are similarly known as the high-symmetry lines. The band energies $E_i(\mathbf{k})$ and wavefunctions $u_i(\mathbf{k})$ at the high-symmetry points and along the high-symmetry lines are important because they provide a qualitative picture of how the band energies and the wavefunctions vary in the whole Brillouin zone — a plot of band energies along the high-symmetry lines, like the one shown in Figure 5-5, is typical in the literature.

5.2.2 Problem Statement

Consider a crystal structure defined by the Bravais lattice vectors $\{\mathbf{a}_i \in \mathbb{R}^3, 1 \le i \le 3\}$ and the basis vectors $\boldsymbol{\tau} \equiv (\tau_1, \ldots, \tau_{n_\tau})$. For any given $\mathbf{k} \equiv (k_1, k_2, k_3) \in \mathcal{D}$, we would like to find the band energies, $E_i(\mathbf{k})$, $1 \le i \le n_b$, given by

$$E_i(\mathbf{k}) = \lambda_i(\mathbf{k}) + \frac{1}{2}|\mathbf{k}|^2; \tag{5.13}$$

[1]Since we have not exploited the symmetric properties of the crystal in our approximation except in defining the parameter domain in which $\mathbf{k}$ lies, we shall not give a discussion on symmetry groups and symmetry operations; interested readers should refer to [3, 61, 60].

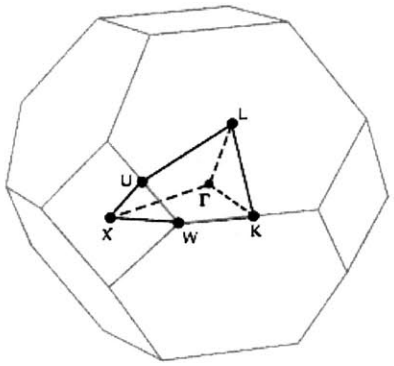

Figure 5-2: First Brillouin zone and the irreducible Brillouin zone (shaded) of face center cubic structure. The irreducible Brillouin zone is a polyhedron with high symmetry points at the vertices: $\mathrm{L} \equiv \frac{2\pi}{a}(1/2, 1/2, 1/2)$, $\Gamma \equiv \frac{2\pi}{a}(0, 0, 0)$, $\mathrm{X} \equiv \frac{2\pi}{a}(1, 0, 0)$, $\mathrm{W} \equiv \frac{2\pi}{a}(3/4, 3/4, 0)$, $\mathrm{K} \equiv \frac{2\pi}{a}(1, 0, 1/2)$, and $\mathrm{U} \equiv \frac{2\pi}{a}(1, 1/4, 1/4)$. Taken from [59].

where $\mathcal{D} \subset \mathbb{R}^3$ is a bounded domain given by the irreducible Brillouin zone of the Bravais lattice; and $(\hat{\mathbf{u}}(\mathbf{k}), \hat{\boldsymbol{\lambda}}(\mathbf{k})) \in (Y^{n_b} \times \mathbb{R}^{n_b})$ satisfies[2]

$$\begin{aligned} \left(-\frac{1}{2}\Delta - \mathrm{i}\mathbf{k}\nabla + V_{\text{eff}}(\mathbf{x}; \boldsymbol{\tau})\right) u_i(\mathbf{k}) &= \lambda_i(\mathbf{k}) u_i(\mathbf{k}), \qquad 1 \le i \le n_b, \\ \int_\Omega u_i(\mathbf{k}) u_j(\mathbf{k}) &= \delta_{ij}, \qquad 1 \le i \le j \le n_b. \end{aligned} \tag{5.14}$$

Here, $\hat{\mathbf{u}}(\mathbf{k}) \equiv (u_1(\mathbf{k}), \ldots, u_{n_b}(\mathbf{k}))$; $\hat{\lambda}(\mathbf{k}) \equiv (\lambda_1(\mathbf{k}), \ldots, \lambda_{n_b}(\mathbf{k}))$; $Y \equiv H^1_{\text{per}}(\Omega)$ is the space of $\{\mathbf{a}_i, 1 \le i \le 3\}$-periodic complex functions in $H^1(\mathbb{R}^3)$; Ω is the primitive unit cell as defined in Section 5.2.1; $\mathbf{x}$ is a point in Ω and defined by (5.2); $V_{\text{eff}}(\,\cdot\,; \boldsymbol{\tau}) \in Y$ is a real periodic function dependent on $\boldsymbol{\tau}$; and n_b is the number of band energies we are interested in. Components in $\hat{\boldsymbol{\lambda}}(\mathbf{k})$ are arranged such that $\lambda_1(\mathbf{k}) \le \lambda_2(\mathbf{k}) \le \ldots \le \lambda_{n_b}(\mathbf{k})$. We prove $\lambda_i(\mathbf{k})$, $1 \le i \le n_b$ are real in Section 5.2.3. We also note that (5.14) is a linear eigenvalue problem, and as such the procedure outlined in Chapter 2 is directly extensible to the current problem.

[2]The solution $\hat{\mathbf{u}}$ and $\hat{\boldsymbol{\lambda}}$ is of course dependent on $\boldsymbol{\tau}$ as well, in addition to $\mathbf{k}$. However, in this Chapter, $\boldsymbol{\tau}$ is considered constant, and the only parameter we are interested in is $\mathbf{k}$. As such, to simplify notation, we only show the dependence of $\hat{\mathbf{u}}$ and $\hat{\boldsymbol{\lambda}}$ on $\mathbf{k}$.

5.2.3 Parameterized Weak Form

The parameterized weak form of (5.14) is given by: for a given $\mathbf{k} \in \mathcal{D}$, find $(\hat{\mathbf{u}}(\mathbf{k}), \hat{\boldsymbol{\lambda}}(\mathbf{k})) \in (Y^{n_b} \times \mathbb{R}^{n_b})$ that satisfies

$$a(u_i(\mathbf{k}), v; V_{\text{eff}}; \mathbf{k}) = \lambda_i(\mathbf{k}) m(u_i(\mathbf{k}), v), \quad \forall v \in Y, \quad 1 \le i \le n_b, \tag{5.15}$$

$$m(u_i(\mathbf{k}), u_j(\mathbf{k})) = \delta_{ij}, \quad 1 \le i \le j \le n_b, \tag{5.16}$$

where

$$a(w, v; t; \mathbf{k}) \equiv \frac{1}{2}\int_\Omega \nabla w \nabla v^* + \int_\Omega t\, w\, v^* - \mathrm{i}\sum_{j=1}^{3} \int_\Omega \frac{\partial w}{\partial x_j} v^*, \tag{5.17}$$

$$m(w, v) \equiv \int_\Omega w\, v^*, \tag{5.18}$$

for any $w \in Y$, $v \in Y$ and $t \in Y$. Here $*$ denotes complex conjugation. We note that the functional form a is affine with respect to the parameter $\mathbf{k}$ — we can express $a(\,\cdot\,, \cdot\,; \cdot\,; \mathbf{k})$ as

$$a(w, v; t; \mathbf{k}) = a_1(w, v; t) + \sum_{j=1}^{3} k_j a_{2,j}(w, v), \tag{5.19}$$

where the $\mathbf{k}$-independent forms $a_1(w, v; t)$, and $a_{2,j}(w, v)$, $1 \le j \le 3$ are given by

$$a_1(w, v; t) \equiv \frac{1}{2}\int_\Omega \nabla w \nabla v^* + \int_\Omega t\, w\, v^*, \tag{5.20}$$

$$a_{2,j}(w, v) \equiv -\mathrm{i}\int_\Omega \frac{\partial w}{\partial x_j} v^*. \tag{5.21}$$

It is also clear that $a(w, v; t; \mathbf{k})$ and $m(w, v)$ are continuous

$$|a(w, v; t; \mathbf{k})| \le \gamma_a \|w\|_Y \|v\|_Y, \quad \forall w, v \subset Y, \tag{5.22}$$

$$|m(w, v)| \le \gamma_m \|w\|_Y \|v\|_Y, \quad \forall w, v \subset Y; \tag{5.23}$$

and coercive

$$0 \leq \alpha_a \equiv \inf_{w \subset Y} \frac{|a(w,w;t;\mathbf{k})|}{\|w\|_Y^2}, \tag{5.24}$$

$$0 \leq \alpha_m \equiv \inf_{w \subset Y} \frac{|m(w,w)|}{\|w\|_Y^2}. \tag{5.25}$$

Clearly $m(w,v)$ is hermitian: $m(w,v) = m^*(v,w)$, $\forall v, w \subset Y$. We now show that $a(w,v;t;\mathbf{k})$ is also hermitian: clearly $a_1(w,v) = a_1^*(v,w)$, $\forall v, w \subset Y$; for $i = 1, \ldots, 3$

$$\begin{aligned} a_{2,j}(w,v) &= -\mathrm{i} \int_\Omega \frac{\partial w}{\partial x_j} v^* \\ &= \mathrm{i} \int_\Omega w \frac{\partial v^*}{\partial x_j} \\ &= a_{2,j}^*(v,w), \end{aligned} \tag{5.26}$$

$\forall v, w \subset Y$; thus $a(v,w;t;\mathbf{k}) = a^*(v,w;t;\mathbf{k})$, $\forall v, w \subset Y$. The problem (5.16) is then well posed. In addition the hermitian of $a(\,\cdot\,,\,\cdot\,;\,\cdot\,;\mathbf{k})$ proves that $\lambda_i(\mathbf{k})$, $1 \leq j \leq n_b$ are real for all $\mathbf{k} \in \mathcal{D}$.

5.2.4 Numerical Example

We consider the band structure calculation of a diamond structure of silicon based on the empirical pseudopotential model in [29]. The diamond structure is used as an example in Section 5.2.1; we repeat the key information here. The Bravais lattice vectors are defined by

$$\mathbf{a}_1 = \frac{a}{2}(0,1,1), \quad \mathbf{a}_2 = \frac{a}{2}(1,0,1), \quad \mathbf{a}_3 = \frac{a}{2}(1,1,0), \tag{5.27}$$

where a is the lattice length, and the unit cell is as shown in Figure 5-1. The basis vectors defining the locations of the two nuclei are given by $\boldsymbol{\tau} = (-\tau_0, \tau_0)$. The model in [29] is only valid for

$$a = 10.32, \quad \text{and} \quad \tau_0 = \sum_{i=1}^{3} \frac{1}{8} \mathbf{a}_i, \tag{5.28}$$

where $\mathbf{a}_i$, $1 \leq i \leq 3$ are given by (5.27). Our parameter, $\mathbf{k}$, lies in the domain $\mathcal{D}$ given by the irreducible Brillouin zone of the fcc structure defined by the polyhedron with vertices at L $\equiv \frac{2\pi}{a}(1/2,1/2,1/2)$, $\Gamma \equiv \frac{2\pi}{a}(0,0,0)$, X $\equiv \frac{2\pi}{a}(1,0,0)$, W $\equiv \frac{2\pi}{a}(3/4,3/4,0)$, K $\equiv \frac{2\pi}{a}(1,0,1/2)$, and U

$\equiv \frac{2\pi}{a}(1, 1/4, 1/4)$, as shown in Figure 5-2.

The effective potential V_{eff}(as defined in [29]) is given by

$$V_{\text{eff}}(\mathbf{x}; \boldsymbol{\tau}(\tau_0)) = \sum_{\mathbf{G}} S(\mathbf{G}; \tau_0) V(\mathbf{G}) e^{\mathrm{i}\mathbf{G}\mathbf{x}}, \tag{5.29}$$

where $\mathbf{G}$ is defined in (5.5); $S(\mathbf{G}; \tau_0) = \cos \mathbf{G}\tau_0$; and $V(\mathbf{G})$ is given by

$$V(\mathbf{G}) = \begin{cases} -0.21, & |\mathbf{G}|^2 = 3\left(\frac{2\pi}{a}\right)^2, \\ 0.04, & |\mathbf{G}|^2 = 8\left(\frac{2\pi}{a}\right)^2, \\ 0.08, & |\mathbf{G}|^2 = 11\left(\frac{2\pi}{a}\right)^2, \\ 0, & \text{otherwise.} \end{cases} \tag{5.30}$$

As shown in Figure 5-3, $V_{\text{eff}}(\cdot; \boldsymbol{\tau})$ is smooth and is represented with just 44 Fourier modes.

Note that the eigenvalue problem given by (5.16) and (5.16) is not derived from *ab initio* theory that is based on the minimization of the total energy of the system. Rather, the pseudopotential function V_{eff} given by (5.29) is constructed such that it reproduces the experimental results — reflectivity and photoemission measurements. It is thus only suitable for calculating the band structure of the crystalline silicon and interpreting results from optical experiments [29]. In addition, since the experimental results on which the V_{eff} are built on are for the parameters given in (5.28), the model is only valid for those values. As such, any attempt to use the model in dynamical simulation, for example, forces on nuclei, will be non-physical.

The following derivation of the internuclear force is thus purely a numerical experiment that allows us to make comparison between different approximation spaces. For any given $\tau_0 \equiv (\tau, \tau, \tau)$, we can then define the force $F(\tau; \mathbf{k})$ by

$$\begin{aligned} F(\tau; \mathbf{k}) &= \sqrt{3} \sum_{i=1}^{n_v} \frac{\partial}{\partial \tau} a(u_i(\mathbf{k}), u_i(\mathbf{k}); V_{\text{eff}}(\,\cdot\,; \boldsymbol{\tau}(\tau)); \mathbf{k}) \\ &= \sqrt{3} \sum_{i=1}^{n_v} \int_{\Omega} u_i^*(\mathbf{k}) u_i(\mathbf{k}) \frac{\partial V_{\text{eff}}(\,\cdot\,; \boldsymbol{\tau}(\tau))}{\partial \tau}, \end{aligned} \tag{5.31}$$

based on Hellmann-Feynman theorem [37]. Here, n_v, the number of valence electrons, is 4. We

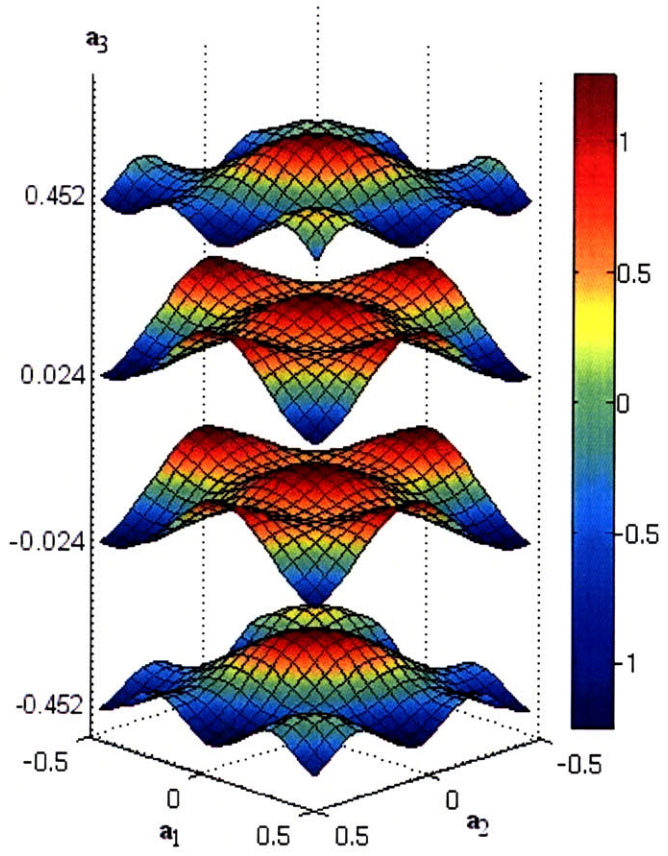

Figure 5-3: $V_{\text{eff}}(\mathbf{x};\boldsymbol{\tau})$ along $\mathbf{a}_1$-$\mathbf{a}_2$ plane for $-0.452\mathbf{a}_3$, $-0.024\mathbf{a}_3$, $0.024\mathbf{a}_3$ and $0.452\mathbf{a}_3$.

define a bilinear form $a_3(w, v; t)$ as

$$a_3(w, v; t) = \sqrt{3} \int_\Omega t\, w\, v^*; \tag{5.32}$$

and thus (5.31) can be expressed as

$$F(\tau; \mathbf{k}) = \sum_{i=1}^{n_v} a_3 \left(u_i(\mathbf{k}), u_i(\mathbf{k}); \frac{\partial V_{\text{eff}}(\,\cdot\,; \boldsymbol{\tau}(\tau))}{\partial \tau} \right). \tag{5.33}$$

5.2.5 "Truth" Approximation

We now consider the approximation of (5.16) by the planewave method. We define our Fourier approximation space $Y^{\mathcal{N}} \subset Y$ of dimension $\mathcal{N}$ as

$$Y^{\mathcal{N}} \equiv \text{span} \left\{ \varphi_{\mathbf{G}} \equiv e^{i\mathbf{G}\mathbf{x}}, \mathbf{x} \in \Omega, \frac{1}{2}|\mathbf{G}|^2 \leq E_{\text{cut}} \right\} \tag{5.34}$$

where $\mathbf{G}$ is defined by (5.5); and E_{cut} is a user-defined cutoff kinetic energy of the planewaves — $\mathcal{N}$ is then the number of reciprocal lattice vectors that satisfy the inequality $\frac{1}{2}|\mathbf{G}|^2 \leq E_{\text{cut}}$. The inner product and norm associated with $Y^{\mathcal{N}}$ are simply inherited from Y. Our planewave approximation to (5.16) is then given by: for a given $\mathbf{k} \in \mathcal{D}$, find $(\hat{\mathbf{u}}^{\mathcal{N}}(\mathbf{k}), \hat{\boldsymbol{\lambda}}^{\mathcal{N}}(\mathbf{k})) \in ((Y^{\mathcal{N}})^{n_b} \times \mathbb{R}^{n_b})$ that satisfies

$$\begin{aligned} a(u_i^{\mathcal{N}}(\mathbf{k}), v; V_{\text{eff}}; \mathbf{k}) &= \lambda_i^{\mathcal{N}}(\mathbf{k}) m(u_i^{\mathcal{N}}(\mathbf{k}), v), \quad \forall v \in Y^{\mathcal{N}}, \quad 1 \leq i \leq n_b, \\ m(u_i^{\mathcal{N}}(\mathbf{k}), u_j^{\mathcal{N}}(\mathbf{k})) &= \delta_{ij}, \quad 1 \leq i \leq j \leq n_b. \end{aligned} \tag{5.35}$$

We expand the planewave approximation as

$$u_i^{\mathcal{N}}(\mathbf{k}) = \sum_{\mathbf{G}} u_{i\,\mathbf{G}}^{\mathcal{N}}(\mathbf{k}) \varphi_{\mathbf{G}}, \quad 1 \leq n \leq n_b, \quad \forall \varphi_{\mathbf{G}} \in Y^{\mathcal{N}}, \tag{5.36}$$

and insert this representation into (5.35) to obtain

$$\begin{aligned} \sum_{\mathbf{G}} a(\varphi_{\mathbf{G}'}, \varphi_{\mathbf{G}}; V_{\text{eff}}; \mathbf{k}) u_{n\,\mathbf{G}}^{\mathcal{N}}(\mathbf{k}) &= \lambda_n^{\mathcal{N}}(\mathbf{k}) u_{n\,\mathbf{G}}^{\mathcal{N}}(\mathbf{k}) \delta_{\mathbf{G},\mathbf{G}'}, \quad \forall \varphi_{\mathbf{G}'} \in Y^{\mathcal{N}}, \quad 1 \leq n \leq n_b; \\ \sum_{\mathbf{G}} u_{n\,\mathbf{G}}^{\mathcal{N}}(\mathbf{k}) u_{n'\,\mathbf{G}}^{\mathcal{N}}(\mathbf{k}) &= \delta_{n,n'}, \quad 1 \leq n, n' \leq n_b; \end{aligned} \tag{5.37}$$

since $m(\varphi_{\mathbf{G}'}, \varphi_{\mathbf{G}}) = \delta_{\mathbf{G},\mathbf{G}'}$. The above then gives a $\mathcal{N} \times \mathcal{N}$ algebraic system which can then be diagonalized to obtain the desired solutions. We now determine convergence of the solutions with respect to number of planewaves, $\mathcal{N}$; this provides a benchmark for subsequent comparison with reduced basis approximation. We first define our error measures are

$$\varepsilon_{\mathcal{N},n_b}^{\lambda} = \max_{\mathbf{k} \in \Xi_T} \epsilon_{\mathcal{N},n_b}^{\lambda}(\mathbf{k}), \tag{5.38}$$

$$\varepsilon_{\mathcal{N}}^{F} = \max_{\mathbf{k} \in \Xi_T} \epsilon_{\mathcal{N}}^{F}(\mathbf{k}); \tag{5.39}$$

where

$$\epsilon^{\lambda}_{\mathcal{N},n_b}(\mathbf{k}) = \max_{1\le i\le n_b} \left| \frac{\lambda_i^{\mathcal{N}}(\mathbf{k}) - \lambda_i(\mathbf{k})}{\lambda_i(\mathbf{k})} \right|, \tag{5.40}$$

$$\epsilon^{F}_{\mathcal{N}}(\mathbf{k}) = \left| F_{\mathcal{N}}(\tau;\mathbf{k}) - F(\tau;\mathbf{k}) \right|, \tag{5.41}$$

and $\Xi_T \subset \mathcal{D}$. We shall choose $\Xi_T = \Xi_0 \equiv \{\mathbf{k}_0 \equiv \frac{2\pi}{a}(0.6223, 0.2953, 0)\}$, where $\mathbf{k}_0$ is a Baldereschi mean value point given in [5]. From Figure 5-4, we see that $\epsilon^{\lambda}_{\mathcal{N},n_b}$ for $n_b = 20$ converges to machine precision at $\mathcal{N} = 1807$. From Table 5.1, we see that for a typical desired accuracy [78] given by $\epsilon^{\lambda}_{\mathcal{N},n_b} < 0.01$ and $\epsilon^{F}_{\mathcal{N}} < 5\,\mathrm{E}{-4}$, $\mathcal{N}$ required is 137.

We now define our "truth" approximation — we take solutions evaluated with sufficiently large $\mathcal{N}$, denoted here by $\mathcal{N}_t$, as the "truth" solutions. From the previous paragraph, since $\hat{\boldsymbol{\lambda}}^{\mathcal{N}}$ converges to machine precision at $\mathcal{N} = 1807$, we shall take $\mathcal{N}_t = 1807$. In addition, for any $\mathcal{N} < \mathcal{N}_t$, we consider that a planewave approximation; we will compare the accuracy and efficiency of solutions based on $\mathcal{N} < \mathcal{N}_t$ with the reduced basis approximation. To simplify the notation, we drop the subscript $\mathcal{N}_t$ from all subsequent formulations, with the understanding that the "truth"approximation in fact refers to the planewave approximation with $\mathcal{N} = \mathcal{N}_t$. Thus, Y, $\hat{\mathbf{u}}$, and $\hat{\boldsymbol{\lambda}}$ shall now be understood as $Y^{\mathcal{N}_t}$, $\hat{\mathbf{u}}^{\mathcal{N}_t}$, and $\hat{\boldsymbol{\lambda}}^{\mathcal{N}_t}$.

Figure 5-5 shows the variations of the band energies $E_i(\mathbf{k})$ $1 \le i \le 10$ with $\mathbf{k}$. It is clear that the degeneracy property of $E_i(\mathbf{k})$ varies with both i and $\mathbf{k}$ in a complicated manner, in contrast to the well-separated bands encountered in Chapter 2. In Figure 5-6, we show a sample of the solutions given by $\mathrm{Re}(u_i(\mathbf{k}))$, $1 \le i \le 4$ at several $\mathbf{k}$-points along the midplane of the simulation cell. The relations between $u_i(\mathbf{k})$ with $\mathbf{k}$ are nontrivial. In particular, we observe that the behavior of $u_i(\mathbf{k})$ becomes richer as i increases. It is also clear that the raw outputs from an eigenvalue solver do not exhibit the smoothness property required for an efficient vectorial reduced basis approximation.

5.2.6 Existing Approaches to Many k-points Calculations

Most approaches focus less on reducing the computational cost of a single $\mathbf{k}$-point calculation, and more on minimizing the number of $\mathbf{k}$-points used to evaluate a quantity of interest — typically an

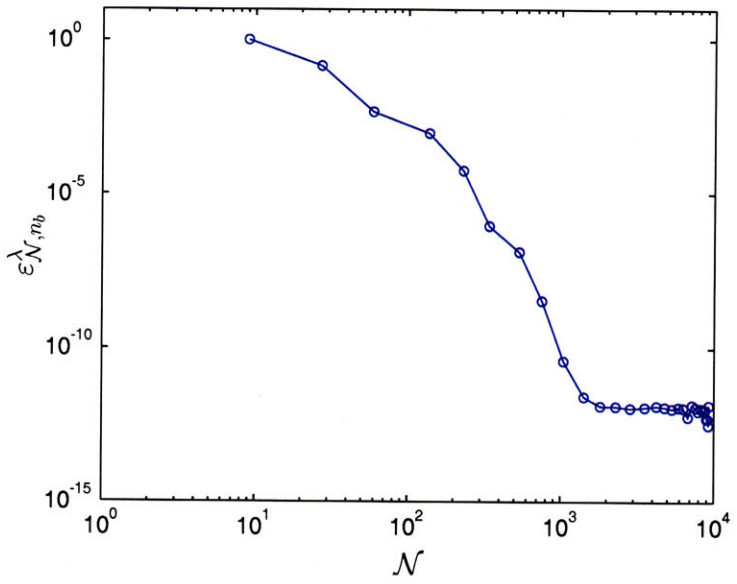

Figure 5-4: Convergence of the planewave approximation error in $\hat{\boldsymbol{\lambda}}^{\mathcal{N}}$, $\varepsilon^{\lambda}_{\mathcal{N},n_b}$ (given by (5.38)), with $\mathcal{N}$ for $n_b = 20$.

$\mathcal{N}$	$\varepsilon^{\lambda}_{\mathcal{N},n_b}$	$\varepsilon^{F}_{\mathcal{N}}$
27	1.03	2.54E−3
59	1.45E−1	7.72E−5
65	6.87E−2	1.19E−5
113	2.32E−2	4.92E−4
137	4.76E−3	2.64E−7
169	1.95E−3	1.20E−5
181	1.48E−3	4.49E−6
229	8.43E−4	2.98E−6
259	4.69E−4	9.36E−7
283	1.87E−4	2.50E−6
331	6.13E−5	9.36E−7

Table 5.1: Convergence of $\varepsilon^{\lambda}_{\mathcal{N},n_b}$ (given by (5.38)) and $\varepsilon^{F}_{\mathcal{N}}$ (given by (5.39)) with $\mathcal{N}$ for $\Xi_T = \Xi_0$ and $n_v = 4$.

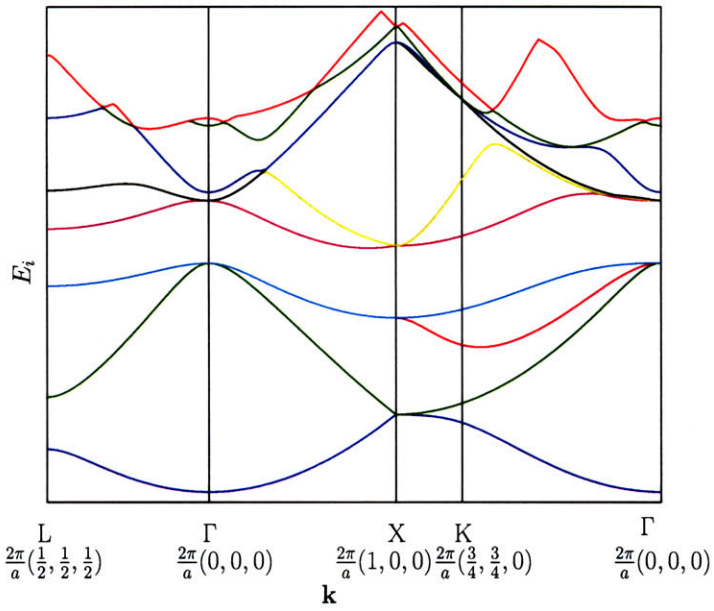

Figure 5-5: Variations of the band energies, E_i, $1 \leq i \leq 10$ along symmetry lines in $\mathcal{D}$, the irreducible Brillouin zone. Here L, Γ, X, and K are special symmetry points of the fcc crystal structure defined in Figure 5-2; $\mathbf{k}$ varies linearly between the points in the above plot.

average of $\mathbf{k}$-dependent function, $f(\mathbf{k})$ (or more precisely $f(\hat{\mathbf{u}}(\mathbf{k}))$ over the parameter space $\mathcal{D}$ [37]:

$$\bar{f} = \frac{1}{|\mathcal{D}|} \int_{\mathcal{D}} f(\mathbf{k}) d\mathbf{k}. \tag{5.42}$$

Several schemes have been proposed, in particular the special point technique [5, 27, 81], and the tetrahedron method [14]. The special point technique approximates $\bar{f}$ by $\sum_{i=1}^{n_k} w_i f(\mathbf{k}_i)$ where $\{\mathbf{k}_i\}_{i=1}^{n_k}$ is the set of judiciously selected $\mathbf{k}$-points; n_k is the total number of special points used; and w_i is the weight associated with a particular $\mathbf{k}_i$. The tetrahedron method first divides the Brillouin zone into a set of tetrahedra, and (5.42) is evaluated by approximating the function $f(\mathbf{k})$ as piecewise linear within each tetrahedron.

The n_k required to approximate $\bar{f}$ to a desired accuracy is of course dependent on the smoothness of the function $f(\mathbf{k})$. In [49], the authors showed that the special point technique leads to faster convergence than the tetrahedron method when $f(\mathbf{k}) \in C^\infty(\mathbb{R}^3)$ has sufficient smoothness. In practice, n_k can be very small: in [28], two $\mathbf{k}$-points are sufficient to give good predictions of lattice dynamics properties of C, Si, and Ge. In [76], it is suggested that 10 $\mathbf{k}$-points are sufficient for almost all modern calculations involving total energy calculations of insulators where (5.42) enters

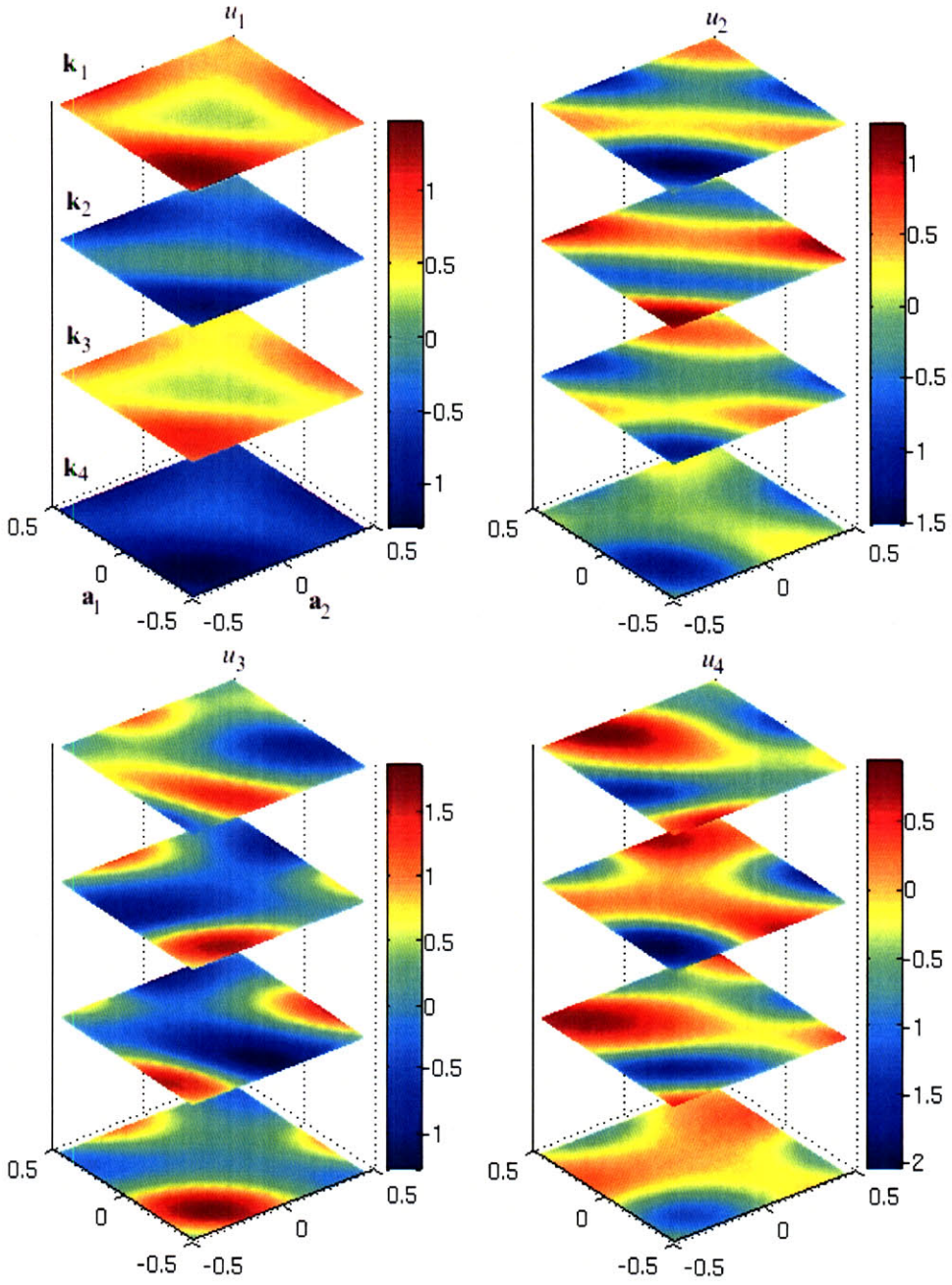

Figure 5-6: Solutions $\mathrm{Re}(u_i(\mathbf{k}_n))$, $1 \leq i \leq 4$ on the $(\mathbf{a}_1, \mathbf{a}_2)$ plane cutting the origin for $\mathbf{k}_1 = \frac{2\pi}{a}(0,0,0)$, $\mathbf{k}_2 = \frac{2\pi}{a}(0.50, 0.07, 0.21)$, $\mathbf{k}_3 = \frac{2\pi}{a}(0.50, 0.43, 0.29)$ and $\mathbf{k}_4 = \frac{2\pi}{a}(0.93, 0, 0.07)$. The color maps correspond to the magnitude and the 4 layers correspond to different $\mathbf{k}$-points, with the top being $\mathbf{k}_1$ and bottom $\mathbf{k}_4$.

as an evaluation of the electron density — $\rho = \frac{1}{|\mathcal{D}|}\int_{\mathcal{D}} \sum_{i=1}^{n_v} (u(\mathbf{k}))^2 d\mathbf{k}$.

However, when $f(\mathbf{k})$ is not smooth, tetrahedron method is more appropriate as it allows refinement in regions where smoothness is lacking. For example, to accurately evaluate the electron density in metals, finer tetrahedron mesh can be constructed near the Fermi surface where there is a sharp variation in the electron density. In addition, evaluation of the density of states, and other related optical properties also involve integration of functions with limited smoothness. As an example on how large n_k needs to be, [41] found that 4000 **k**-points is needed to sufficiently resolve the van Hove singularities [113] in the density of states.

Thus, there exists an opportunity to improve the average cost of a single **k**-point, especially when n_k required is large. The use of the Wannier representation of the wavefunctions is recently proposed in [79, 110]. The efficient Wannier representation, coupled with the Slater-Koster interpolation scheme [109] leads to rapid evaluation of band energies for any given $\mathbf{k} \in \mathcal{D}$. This approach is used in [116] to evaluate the anomalous Hall conductivity, which requires evaluations of functions at millions of **k**-points.

Here, we would like to propose another approach — the reduced basis method.

5.3 Reduced Basis Method

5.3.1 Augmented Reduced Basis Space

The approximation space

We first introduce nested sample sets $S_N^A = (\mathbf{k}_1, \ldots, \mathbf{k}_{N_s})$, $1 \le N_s \le N_{s,\max}$ and define the associated nested reduced-basis spaces as

$$\begin{aligned} W_N^A &= \text{span}\,\{u_i(\mathbf{k}_j), 1 \le i \le n_b, 1 \le j \le N_s\}, \quad 1 \le N_s \le N_{s,\max}, \\ &= \text{span}\,\{\zeta_n, 1 \le n \le N \equiv N_s n_b\}, \quad 1 \le N_s \le N_{s,\max}; \end{aligned} \tag{5.43}$$

where $(u_1(\mathbf{k}_n), \ldots, u_{n_b}(\mathbf{k}_j))$ are the solutions of (5.16) at $\mathbf{k} = \mathbf{k}_j$; and ζ_n, $1 \le n \le N$ are basis functions obtained after $u_i(\mathbf{k}_j)$, $1 \le i \le n_b$, $1 \le j \le N_s$ are orthogonalized. An approximation of $u_i(\mathbf{k})$ in W_N^A is then given by $u_{N,i}(\mathbf{k}) = \sum_{n=1}^{N} \alpha_{i\,n}(\mathbf{k})\zeta_n$. The reduced basis spaces are constructed based on the adaptive sampling procedure outlined in Section 2.3.6 and for each n_b, we construct a

different set of hierarchical reduced basis spaces. Here, n_b is specified according to the applications that we look at. For example, for studying ground state properties, $n_b = n_v = 4$ is sufficient. For studying optical properties, n_b may need to be as high as 10.

The approximation

The reduced basis approximation to $(\hat{\mathbf{u}}(\mathbf{k}), \hat{\lambda}(\mathbf{k}))$ is given by: for a given $\mathbf{k} \subset \mathcal{D}$, find $(\hat{\mathbf{u}}_N(\mathbf{k}), \hat{\boldsymbol{\lambda}}_N(\mathbf{k})) \in ((W_N^A)^{n_b} \times \mathbb{R}^{n_b})$ such that

$$\begin{aligned} a(u_{N,i}(\mathbf{k}), v; V_{\text{eff}}; \mathbf{k}) &= \lambda_{N,i}(\mathbf{k}) m(u_{N,i}(\mathbf{k}), v), \quad 1 \le i \le n_b, \quad \forall v \in W_N^A \\ m(u_{N,i}(\mathbf{k}), u_{N,j}(\mathbf{k})) &= \delta_{ij}, \quad 1 \le i \le j \le n_b. \end{aligned} \tag{5.44}$$

In addition, the approximation to $F(\tau; \mathbf{k})$ is given by

$$F_N(\tau; \mathbf{k}) = \sum_{i=1}^{n_v} a_3 \left(u_{N,i}(\mathbf{k}), u_{N,i}(\mathbf{k}); \frac{\partial V_{\text{eff}}(\cdot; \boldsymbol{\tau}(\tau))}{\partial \tau} \right). \tag{5.45}$$

Discrete Equations

We expand our reduced-basis approximation as

$$u_{N,n}(\mathbf{k}) = \sum_{j=1}^{N} u_{N,n\,j}(\mathbf{k}) \zeta_j, \quad 1 \le n \le n_b, \tag{5.46}$$

and insert this representation into (5.44) to obtain

$$\begin{aligned} \left(\sum_{j=1}^{N} A_{i,j}^{N,1} + \sum_{l=1}^{3} k_j A_{i,j}^{N,2,l} \right) u_{N,n\,j}(\mathbf{k}) &= \lambda_{N,n}(\mathbf{k}) M_{i,j}^N u_{N,n\,j}(\mathbf{k}), \quad 1 \le i \le N, \quad 1 \le n \le n_b; \\ \sum_{i=1}^{N} \sum_{j=1}^{N} u_{N,n\,i}(\mathbf{k}) M_{i,j}^N u_{N,n'\,j}(\mathbf{k}) &= \delta_{n,n'}, \quad 1 \le n, n' \le n_b; \end{aligned} \tag{5.47}$$

where $A^{N,1} \in \mathbb{C}^{N \times N}$, $A^{N,2,l} \in \mathbb{C}^{N \times N}$, $1 \le l \le 3$, and $M^N \in \mathbb{C}^{N \times N}$ are given by $A_{i,j}^{N,1} = a_1(\zeta_j, \zeta_i), 1 \le i, j \le N$, $A_{i,j}^{N,2,l} = -\mathrm{i} a_{2,l}(\zeta_j, \zeta_i), 1 \le i, j \le N$, and $M_{i,j}^N = m(\zeta_j, \zeta_i), 1 \le i, j \le N$. Then, (5.47) can be solved using any eigenvalue solver.

Due to the affine parameter dependence property of $a(w, v; t; \mathbf{k})$ expressed by (5.19), offline-

online computational strategy can be applied to (5.47). In the offline stage, we compute the solutions for $\mathbf{k} \in S_N^A$ at a cost of $O(\mathcal{N}_t^\bullet)$ where $\bullet$ denotes the complexity of the solution method used in our "truth" approximation; and $A^{N,1}$, $A^{N,2,l}$, $1 \le l \le 3$, and M^N at a cost of order $O(\mathcal{N}_t N^2)$. During the online stage, we reconstruct the reduced basis matrices at a cost of $O(4N^2)$ and solve the resulting algebraic eigenvalue problems at a cost of $O(N^3)$, leading to a online computational complexity independent of $\mathcal{N}$.

Convergence

For our convergence analysis, we introduce a test sample $\Xi_{\mathbf{k}}$ consisting of 488 k-points distributed uniformly in $\mathcal{D}$. We will also define the following error measures:

$$\varepsilon_{N,n_b}^{\lambda} = \max_{\mathbf{k}\in\Xi_T} \epsilon_{N,n_b}^{\lambda}(\mathbf{k}), \tag{5.48}$$

$$\varepsilon_N^F = \max_{\mathbf{k}\in\Xi_T} \epsilon_N^F(\mathbf{k}), \tag{5.49}$$

where

$$\epsilon_{N,n_b}^{\lambda}(\mathbf{k}) = \max_{1\le i\le n_b} \left| \frac{\lambda_{N,i}(\mathbf{k}) - \lambda_i(\mathbf{k})}{\lambda_i(\mathbf{k})} \right|, \tag{5.50}$$

$$\epsilon_N^F(\mathbf{k}) = \left| F_N(\tau;\mathbf{k}) - F(\tau;\mathbf{k}) \right|. \tag{5.51}$$

For $\Xi_T = \Xi_{\mathbf{k}}$, we observe that the behavior of W_N^A is consistent with the results from Section 2.3.4: (*i*) from Figure 5-7, we have a rapidly convergent reduced basis approximation as demonstrated by the convergence of $\varepsilon_{N,n_b}^{\lambda}$ at different n_b; (*ii*) from Table 5.2, N_s must be above a critical value for reasonable approximation for all n_b; and (*iii*) N_s decreases with increasing n_b as the desired accuracy is increased.

We now perform the convergence analysis of the reduced basis approximation for $\Xi_T = \Xi_0$ and $n_b = n_v = 4$ so that a direct comparison with the planewave method can be made. Table 5.3 shows that we require 16 basis functions to achieve the convergence criteria given by $\varepsilon_{\mathcal{N},n_b}^{\lambda} < 0.01$ and $\varepsilon_{\mathcal{N}}^F < 5\mathrm{E}{-4}$. From Table 5.1, we require $\mathcal{N} = 137$ to achieve the same convergence criteria — the dimension of the reduced basis space is then close to 1/10 of the number of planewaves required. In addition, the computational time for reduced basis approximation based on $N = 16$ is 0.005s; for $\mathcal{N} = 137$, it is 0.07s. We again see a computational saving of $O(10)$.

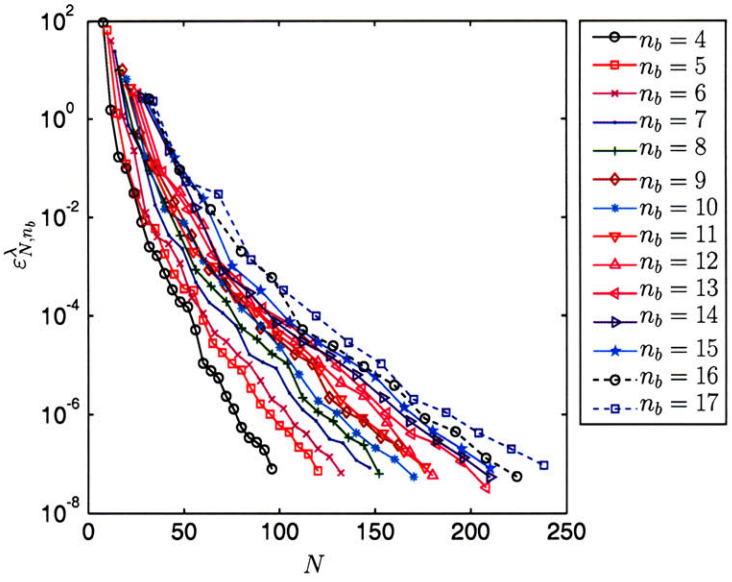

Figure 5-7: Convergence of the reduced basis error in $\hat{\lambda}_N$, $\varepsilon^{\lambda}_{N,n_b}$ (given by (5.48)) with N for $4 \leq n_b \leq 17$ and $\hat{\mathbf{u}} \in (W_N^A)^{n_b}$.

	N_s		
n_b	$\varepsilon^{\lambda}_{N,n_b} < 1\text{E}{-2}$	$\varepsilon^{\lambda}_{N,n_b} < 1\text{E}{-4}$	$\varepsilon^{\lambda}_{N,n_b} < 1\text{E}{-7}$
4	7	14	24
5	6	12	24
6	6	11	22
7	6	11	21
8	6	10	19
9	6	10	18
10	5	9	17
11	5	9	16
12	5	8	15
13	5	8	16
14	5	7	15
15	5	7	14
16	5	7	14
17	5	7	14

Table 5.2: N_s required to reduce the reduced basis error in $\hat{\lambda}_N$, $\varepsilon^{\lambda}_{N,n_b}$ (given by (5.48)) to below $1\text{E}{-2}$, $1\text{E}{-4}$ and $1\text{E}{-7}$ for $4 \leq n_b \leq 17$ and $\hat{\mathbf{u}} \in (W_N^A)^{n_b}$.

N	$\varepsilon^{\lambda}_{N,n_b}$	ε^{F}_{N}
8	5.42E−1	1.38E−4
16	5.75E−3	4.71E−4
24	1.72E−3	3.50E−4
32	3.58E−4	5.56E−5
40	3.56E−5	5.88E−5
48	2.05E−5	1.89E−5

Table 5.3: Convergence of $\varepsilon^{\lambda}_{N,n_b}$ (given by (5.48)) and ε^{F}_{N} (given by (5.49)) with N for $n_b = 4$, $\Xi_T = \Xi_0$ and $\hat{\mathbf{u}} \in (W^A_N)^{n_b}$.

5.3.2 Vectorial Reduced Basis Space

The approximation space

We first introduce nested sample sets $S^V_N = (\mathbf{k}_1, \ldots, \mathbf{k}_N)$, $1 \le N \le N_{\max}$ and define the associated nested reduced-basis spaces as

$$W^V_N = \text{span}\ \{\hat{\mathbf{u}}(\mathbf{k}_n), 1 \le n \le N\}, \quad 1 \le N \le N_{\max}, \tag{5.52}$$

$$= \text{span}\ \{\hat{\boldsymbol{\zeta}}_n, 1 \le n \le N\}, \quad 1 \le N \le N_{\max}; \tag{5.53}$$

where $\hat{\mathbf{u}}(\mathbf{k}_n) \equiv (u_1(\mathbf{k}_n), \ldots, u_{n_b}(\mathbf{k}_n))$ are the solutions of (5.16) at $\mathbf{k} = \mathbf{k}_n$; and $\hat{\boldsymbol{\zeta}}_n \equiv (\zeta_{n,1}, \ldots, \zeta_{n,n_b})$, $1 \le n \le N$ are basis functions obtained after $\hat{\mathbf{u}}(\mathbf{k}_n), 1 \le n \le N$ are preprocessed — sorted, aligned and orthogonalized.

Sort, align and orthogonalize

The goal of the sort and alignment procedure is to obtain $U_N \equiv \{\hat{\boldsymbol{\zeta}}^s_n, 1 \le n \le N\}$ — $\hat{\boldsymbol{\zeta}}^s_n$ is a unitary transformation of $\hat{\mathbf{u}}(k_n)$ — such that $\zeta^s_{n,j}$, $1 \le n \le N$ vary smoothly for $1 \le j \le n_b$. It must take into account the following issues:

1. *Mode shapes of the components of* $\hat{\mathbf{u}}$ *vary with* $\mathbf{k}$. For a given $\mathbf{k}$, $\lambda_i(\mathbf{k})$, $1 \le i \le n_b$ are usually ordered in ascending order, from which we obtain a corresponding order in $u_i(\mathbf{k})$, $1 \le i \le n_b$. The multiplicity of $\lambda_i(\mathbf{k})$, however varies with $\mathbf{k}$, leading to merging and branching of the eigenvalue manifolds as shown in Figure 5-5. As such, mode shape of a particular $u_i(\mathbf{k})$ before and after an intersection point may change. We must rearrange locations of $u_i(\mathbf{k})$, $1 \le i \le n_b$ in $\hat{\mathbf{u}}(\mathbf{k})$ according to some reference set of mode shapes.

2. *The solutions $u_j(\mathbf{k})$, $1 \leq j \leq n_b$ for all $\mathbf{k}$ have arbitrary phase shift ϕ* — a consequence of the periodic boundary conditions. For different $\mathbf{k}$, the eigenvalue solver will not necessarily give the same ϕ. We must remove this random relative phase shift.

3. *Multiplicity of eigenvalues can be greater than 1 and varies with $\mathbf{k}$.* We must then deal with eigensubspaces instead of eigenvectors. When aligning a component of a particular eigensubspace to a reference mode shape, we must consider a linear combination of the bases in that eigensubspace.

Suppose we are given a pre-sorted set $U_N \equiv \{\hat{\zeta}^s_n, 1 \leq n \leq N\}$ where $\hat{\zeta}^s_n$, $1 \leq n \leq N$ are the sorted basis functions of $\{\hat{\mathbf{u}}(\mathbf{k}_n),\ 1 \leq n \leq N\}$, and $W^V_N = \text{span}\{U_N\}$. Then given $\hat{\mathbf{u}}(\mathbf{k}_{N+1})$, we would like to find $\hat{\zeta}^s_{N+1}$ based on $\hat{\mathbf{u}}(\mathbf{k}_{N+1})$ such that $\zeta^s_{N+1,i}$ is a smooth variation of $\zeta^s_{n,i}$, $1 \leq n \leq N+1$ for $1 \leq i \leq n_b$. We assume $\hat{\mathbf{u}}(\mathbf{k}_{N+1})$ is of size $n_{b,\max} \geq n_b$ and contains eigenvectors of similar mode shapes to all components of all basis functions in U_N. We now describe the algorithm used to achieve this goal — it is based on finding an unitary transformation that minimizes the projection error between $\zeta^s_{n,i}$, $1 \leq n \leq N$ for $1 \leq i \leq n_b$.

The first step is to subdivide $\hat{\mathbf{u}}(\mathbf{k}_{N+1})$ into eigensubspaces, each associated with an unique eigenvalue. However, numerical errors introduced by the numerical method used make defining an unique eigenvalue difficult — in particular, to what precision should we consider two eigenvalues as equivalent?

We shall use the accuracy to which we would like to approximate the eigenvalues as the gauge; we denoted this by $\epsilon^\lambda_{\text{tol}}$. We group $\lambda_i(\mathbf{k}_{N+1})$, $1 \leq i \leq n_b$, into p number of distinct clusters, $P^\lambda_j = \{\lambda^j_1, \ldots, \lambda^j_{n^p_j}\}$, $1 \leq j \leq p$ based on the following criteria:

$$\max_{1 \leq q < q' \leq n^p_j}, |\lambda^j_q - \lambda^j_{q'}| \leq \epsilon^\lambda_{\text{tol}}, \tag{5.54}$$

where n^p_j is the size of P^λ_j. We further define $P^u_j = \{u^j_1, \ldots, u^j_{n^p_j}\}$ as

$$P^u_j = \{w \ | a(w, v; V_{\text{eff}}; \mathbf{k}) = \vartheta m(w, v), v \in Y, \vartheta \in P^\lambda_j\}, \tag{5.55}$$

and dim $(P^u_j) = n^p_j$. This clustering procedure is well-defined — we shall show that given two different eigenvalues $\lambda_i(\mathbf{k})$ and $\lambda_{i'}(\mathbf{k})$ for which the difference is less than $\epsilon^\lambda_{\text{tol}}$, an approximation

based on the eigensubspace (5.55) will always yield an eigenvalue with an accuracy $\epsilon_{\text{tol}}^{\lambda}$. For an eigensolution given by $(\tilde{u}(\mathbf{k}), \tilde{\lambda}(\mathbf{k}))$ where $\tilde{u}(\mathbf{k}) \in \text{span}\ \{P_j^u\}$, we can represent $\tilde{u}(\mathbf{k})$ by

$$\tilde{u}(\mathbf{k}) = \sum_{q=1}^{n_j^p} \alpha_q(\mathbf{k}) u_q^j, \tag{5.56}$$

and since $m(\tilde{u}(\mathbf{k}), \tilde{u}(\mathbf{k})) = 1$, we have $\sum_{q=1}^{n_j^p} \alpha_q^2(\mathbf{k}) = 1$. Then

$$\begin{aligned} \tilde{\lambda}(\mathbf{k}) - a_1(\tilde{u}(\mathbf{k}), \tilde{u}(\mathbf{k}); V_{\text{eff}}; \mathbf{k}) &= \lambda_i(\mathbf{k}) - \sum_{q=1}^{n_j^p} \sum_{q'=1}^{n_j^p} a(u_q^j, u_{q'}^j; V_{\text{eff}}; \mathbf{k}) \\ &= \tilde{\lambda}(\mathbf{k}) - \sum_{q=1}^{n_j^p} \alpha_q^2(\mathbf{k}) \lambda_q^j, \end{aligned} \tag{5.57}$$

since $m(u_q^j, u_{q'}^j) = \delta_{qq'}$. Then since $\tilde{\lambda}(\mathbf{k}) - a(\tilde{u}(\mathbf{k}), \tilde{u}(\mathbf{k}); V_{\text{eff}}, \mathbf{k}) = 0$, we have

$$\tilde{\lambda}(\mathbf{k}) = \sum_{q=1}^{n_j^p} \alpha_q^2(\mathbf{k}) \lambda_q^j; \tag{5.58}$$

and from $\sum_{q=1}^{n_j^p} \alpha_q^2(\mathbf{k}) = 1$, we conclude that$\tilde{\lambda}(\mathbf{k})$ is accurate up to $\epsilon_{\text{tol}}^{\lambda}$.

The next step is to find a reference basis in U_N — we select a $\hat{\zeta}_{n^*}^s \in U_N$ for which $n^* = \arg\min_{1 \le n \le N} |\mathbf{k}_n - \mathbf{k}_{N+1}|$. The assumption is that the "difference" between $\hat{\mathbf{u}}(\mathbf{k}_{n^*})$ and $\hat{\mathbf{u}}(\mathbf{k}_{N+1})$ will be the smallest based on the smoothness argument. We then associate components of $\hat{\zeta}_{n^*}^s$ to P_j^u, $1 \le j \le p$ defined earlier such that $\zeta_{n^*,i}^s$ associated with P_j^u form a subspace closest to span $\{P_j^u\}$. To achieve this, we first define $e_{i,j} = \min_{\chi \in \text{span}\{P_j^u\}} \|\zeta_{n^*,i}^s - \chi\|_Y$, $1 \le i \le n_b$, $1 \le j \le p$ — a correlation matrix between $\hat{\zeta}_{n^*}^s$ and subspaces given by span $\{P_j^u\}$. We then associate $\zeta_{n^*,i}^s$ to P_j^u for which $e_{i,j}$ is the smallest, with the constraint that the number of $\zeta_{n^*,i}^s$ associate to a particular P_j^u must not exceed n_j^p. Figure 5-8 shows the iterative procedure used for this purpose.

The final step involves determining components in span $\{P_j^u\}$ — given by $\zeta_{N+1,i}^s$ — that are closest to $\zeta_{n^*,i}^s$ associated with P_j^u. We find the projection of $\zeta_{n^*,i}^s$ onto span $\{P_j^u\}$, with the constraints that $\zeta_{N+1,i}^s$ must be normalized and orthogonal to $\zeta_{N+1,j\neq i}^s$. We summarize the procedure in Figure 5-9. For the (unsorted) solutions $u_i(\mathbf{k}_n)$, $1 \le i \le 4$ shown in Figure 5-6, we shown the

```
Initialize I = {i, 1 ≤ i ≤ n_b};
Initialize J = {j, 1 ≤ j ≤ p};
Initialize I_j = { },   j = 1, ..., p;
for i = 1 : n_b
        j* = arg min_{j∈J} e_{i,j};
        I_{j*} = I_{j*} ∪ i;
end
while J is not empty
        j* = arg min_{j∈J} min_{i∈I} e_{i,j};
        J = J\j*;
        if dim (I_{j*}) > n^P_{j*}
                Initialize I_temp = { };
                while dim (I_temp) ≤ n^P_{j*}
                        i* = arg min_{i∈I_{j*}} e_{i,j*};
                        I_temp = I_temp ∪ i*;
                        I = I\i*;
                        I_{j*} = I_{j*}\i*;
                end
                for i ∈ I_{j*}
                        j+ = arg min_{j∈J} e_{i,j};
                        I_{j+} = I_{j+} ∪ i;
                end
                I_{j*} = I_temp;
        else
                I = I\I_{j*};
        end
end.
```

Figure 5-8: The algorithm to associate $\zeta^s_{n^*,i}$, $1 \leq i \leq n_b$ to P^u_j, $1 \leq j \leq p$.

```
for j = 1 : p
        Initialize I_temp = { };
        for i ∈ I_j
                { ζ^s_{N+1,i} = min_{χ∈span{P^u_j}} ‖χ − ζ^s_{n*,i}‖_Y,
                { s.t. m(χ, χ) = 1,
                {      m(ζ^s_{N+1,k}, χ) = 0,    k ∈ I_temp;
            I_temp = I_temp ∪ i;
        end
end.
```

Figure 5-9: The algorithm to determine $\zeta^s_{N+1,i}$, $1 \le i \le n_b$.

resulting sorted basis functions $\zeta^s_{n,i}$, $1 \le i \le 4$ in Figure 5-10. We see that $\zeta^s_{n,i}$, $1 \le i \le 3$ appear to be well sorted and aligned (by the crude standard of a visual inspection), but variation in $\zeta^s_{n,4}$ does not appear to be "smooth". We return to this point later in the section.

Finally, $\hat{\boldsymbol{\zeta}}^s_{N+1} \equiv (\zeta^s_{N+1,1}, \ldots, \zeta^s_{N+1,n_b})$ is pseudo-orthogonalized to obtain $\hat{\boldsymbol{\zeta}}_{N+1}$ according to the procedure in Section 2.3.5 and $W^V_{N+1} = W^V_N + \text{span}\,\{\hat{\boldsymbol{\zeta}}_{N+1}\}$.

An approximation of $\hat{\mathbf{u}}(\mathbf{k})$ in W^V_N is then given by $\hat{\mathbf{u}}_N(\mathbf{k}) = \sum_{n=1}^N \alpha_n(\mathbf{k})\hat{\boldsymbol{\zeta}}_n$ — the components of $\hat{\mathbf{u}}_N(\mathbf{k})$ is given by $u_{N\,i}(\mathbf{k}) = \sum_{n=1}^N \alpha_n(\mathbf{k})\zeta_{n,i}$, $1 \le i \le n_b$. Again W^V_N can be constructed based on the adaptive sampling procedure and for each n_b, we construct a separate set of nested reduced basis spaces.

The Approximation

Here, the reduced basis approximation to $(\hat{\mathbf{u}}(\mathbf{k}), \hat{\lambda}(\mathbf{k}))$ is given by: for a given $\mathbf{k} \in \mathcal{D}$, find $(\hat{\mathbf{u}}_N(\mathbf{k}), \hat{\boldsymbol{\lambda}}_N(\mathbf{k})) \in (W^V_N \times \mathbb{R}^{n_b})$ such that

$$\begin{aligned} \sum_{i=1}^{n_b} a(u_{N,i}(\mathbf{k}), v_i; V_{\text{eff}}; \mathbf{k}) &= \sum_{i=1}^{n_b} \lambda^L_{N,i}(\mathbf{k}) m(u_{N,i}(\mathbf{k}), v_i), \quad \forall \hat{v} \equiv (v_1, \ldots, v_{n_b}) \in W^V_N \\ m(u_{N,i}(\mathbf{k}), u_{N,i}(\mathbf{k})) &= 1, \quad 1 \le i \le n_b. \end{aligned} \tag{5.59}$$

Similar to Section 2.3.5, $\lambda_{N,i}(\mathbf{k}) \neq \lambda^L_{N,i}(\mathbf{k})$ — we recover $\lambda_{N,i}(\mathbf{k})$ from the Rayleigh Quotient:

$$\lambda_{N,i}(\mathbf{k}) = a(u_{N,i}(\mathbf{k}), v_i; V_{\text{eff}}; \mathbf{k}). \tag{5.60}$$

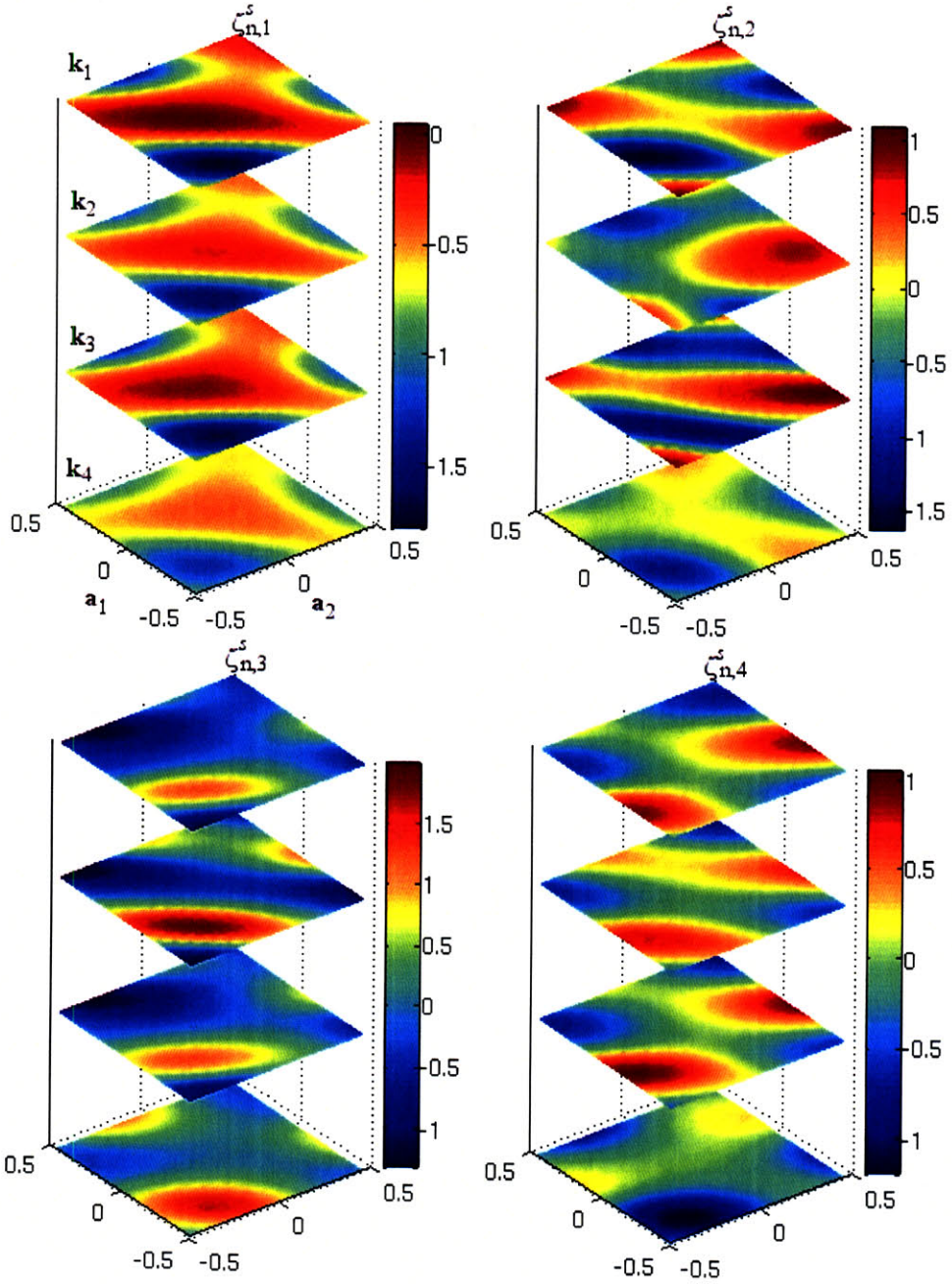

Figure 5-10: Solutions $\mathrm{Re}(\zeta^s_{n,i})$, $1 \leq i \leq 4$ on the $(\mathbf{a}_1, \mathbf{a}_2)$ plane cutting the origin corresponding to sorted $u_i(\mathbf{k}_n)$, $1 \leq i \leq 4$ for $\mathbf{k}_1 = \frac{2\pi}{a}(0,0,0)$, $\mathbf{k}_2 = \frac{2\pi}{a}(0.50, 0.07, 0.21)$, $\mathbf{k}_3 = \frac{2\pi}{a}(0.50, 0.43, 0.29)$ and $\mathbf{k}_4 = \frac{2\pi}{a}(0.93, 0, 0.07)$. The color maps correspond to the magnitude and the 4 layers correspond to different $\mathbf{k}$-points, with the top being $\mathbf{k}_1$ and bottom $\mathbf{k}_4$.

In addition, we again enforce only the normality constraints in (5.59) based on the Hypothesis 2.1. The above problem gives meaningful solutions only for $N > n_b$.

Discrete Equations

We expand our reduced-basis approximation as

$$\hat{\mathbf{u}}_N(\mathbf{k}) = \sum_{j=1}^{N} u_{N\,j}(\mathbf{k})\hat{\boldsymbol{\zeta}}_j. \tag{5.61}$$

Inserting this representation into (5.59) yields

$$\begin{aligned} \sum_{j=1}^{N}\left(A_{i,j}^{N,1} + \sum_{l=1}^{3} k_l A_{i,j}^{N,2,l}\right) u_{N\,j}(\mathbf{k}) &= \sum_{j=1}^{N}\sum_{n=1}^{n_b} \lambda_{N,n}^{L}(\mathbf{k}) M_{i,j}^{N,n,n} u_{N\,j}(\mathbf{k}), \quad 1 \le i \le N; \\ \sum_{i=1}^{N}\sum_{j=1}^{N} u_{N\,i}(\mathbf{k}) M_{i,j}^{N,n,n} u_{N\,j}(\mathbf{k}) &= 1, \quad 1 \le n \le n_b; \end{aligned} \tag{5.62}$$

where $A^{N,1} \in \mathbb{C}^{N\times N}$, $A^{N,2,l} \in \mathbb{C}^{N\times N}$, $1 \le l \le 3$, and $M^{N,l,l'} \in \mathbb{C}^{N\times N}$, $1 \le l \le l' \le n_b$ are given by $A_{i,j}^{N,1} = \sum_{n=1}^{n_b} a_1(\zeta_{n,j}, \zeta_{n,i}), 1 \le i,j \le N$, $A_{i,j}^{N,2,l} = \sum_{n=1}^{n_b} a_{2,l}(\zeta_{n,j}, \zeta_{n,i}), 1 \le i,j \le N$, and $M_{i,j}^{N,l,l'} = m(\zeta_{l,j}, \zeta_{l',i}), 1 \le i,j \le N$, respectively.

We note the affine parametric dependence property holds and thus the offline-online computational decomposition can be applied. To solve (5.62), we use the Newton iterative scheme outlined in Section 2.3.5.

Convergence

From Figure 5-11, it is clear that the performance of vectorial reduced basis space is significantly poorer than augmented reduced basis space. For $\Xi_T = \Xi_{\mathbf{k}}$, we require approximately 250 basis functions to achieve an accuracy of $1\,\mathrm{E}{-4}$ for $n_b = 8$. This is in contrary to the results obtained in Section 2.3.5 where we show that vectorial reduced basis space can be very efficient. The behavior of $\hat{\boldsymbol{\zeta}}_n^s$ offers a possible explanation. In Figure 5-10, we show a sample of the solutions given by $\mathrm{Re}(\zeta_{n,i}^s)$, $1 \le i \le 4$ at several $\mathbf{k}$-points along the midplane of the simulation cell. For $\zeta_{n,1}^s$, the pre-processing steps have performed as intended. However, for $\zeta_{n,4}^s$, it is clear from visual inspection that it does not vary smoothly with $\mathbf{k}$. This indicates our pre-processing steps are not sufficiently robust to

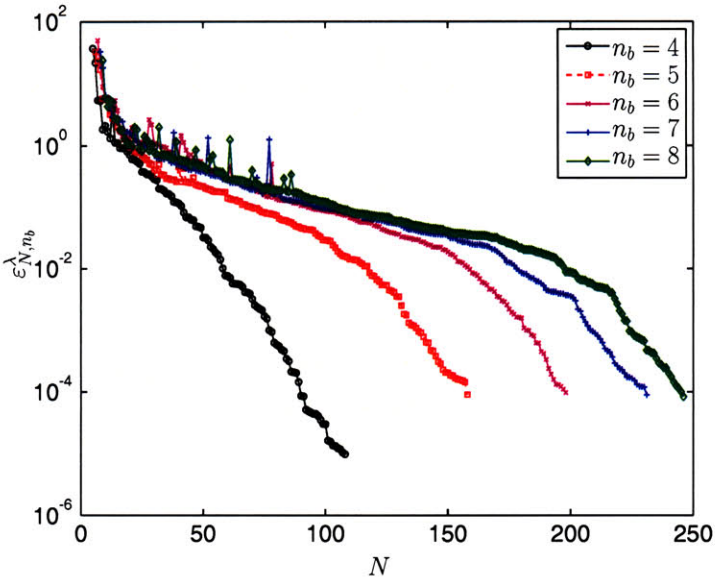

Figure 5-11: Convergence of the reduced basis error in $\hat{\boldsymbol{\lambda}}_N$, $\varepsilon^{\lambda}_{N,n_b}$ (given by (5.48)), with N for $4 \leq n_b \leq 8$ and $\hat{\mathbf{u}} \in W_N^V$.

recover the smooth solution manifold. The failure to satisfy this smoothness requirement explains the poor performance of the vectorial reduced basis approximation for this particular problem.

For $\Xi_T = \Xi_0$, we note that we we require $N = 44$ to the convergence criteria of Section 5.2.5 for $n_b = 4$, as shown in Table 5.4. This is significantly poorer than the augmented reduced basis space ($N = 16$). The dimension is even comparable to that required by the planewave method, indicating that this approximation is not competitive, even compared to the planewave method.

5.3.3 Comparison of the Reduced Basis Spaces

We shall compare the augmented reduced basis approximation and the vectorial reduced basis approximation from three different aspects: (*i*) the dimension of the resulting reduced basis spaces; (*ii*) the computational complexity of the online stage; and (*iii*) the computational complexity of the offline stage.

From the results of the previous two sections, it is clear that dimension of W_N^A required to achieve certain accuracy is much smaller than dimension of W_N^V. This suggests that in cases where variation of $\hat{\mathbf{u}}(\mathbf{k})$ with respect to $\mathbf{k}$ is large or the preprocessing step is unable to obtain a reasonably smooth variation with respect to the parameter, it is more efficient to use the formulation based on the augmented reduced basis space. Indeed, the augmented reduced basis space fully exploits

N	$\varepsilon^{\lambda}_{N,n_b}$	ε^{F}_{N}
8	2.25	2.45E−2
16	4.93E−1	3.29E−3
24	2.38	1.35E−2
32	1.03E−1	1.51E−3
40	3.36E−2	3.88E−4
48	1.34E−3	4.83E−5
56	7.39E−4	7.50E−5
64	4.00E−4	8.34E−6
72	1.80E−4	2.41E−5
76	8.01E−5	6.95E−6

Table 5.4: Convergence of $\varepsilon^{\lambda}_{N,n_b}$ (given by (5.48)) and ε^{F}_{N} (given by (5.49)) with N for $n_b = 4$, $\Xi_T = \Xi_0$ and $\hat{\mathbf{u}} \in W^V_N$.

the optimality of the Galerkin method — the difficulties outlined in Section 5.3.2 are automatically resolved by the projection step. On the other hand, vectorial reduced basis approximation relies on the "goodness" of the pre-constructed reduced basis space — a poorly implemented construction process will certainly not be optimal.

For the online stage, the larger N required for vectorial reduced basis approximation is clearly less efficient than augmented reduced basis approximation. In addition, the current implementation of Newton iterative scheme is not as efficient as the eigenvalue solver. This further degrade the performance of augmented reduced basis space compared to augmented reduced basis space. During the offline stage, construction of the space W^V_N is also less efficient than W^A_N — we need to perform N-solve in the former but only N_s-solve in the later, which for this example, N_s is much smaller than N. This is in addition to the extensive preprocessing step required for vectorial reduced basis approximation.

We can thus conclude that for the current problem, the augmented reduced basis approximation is a better approximation.

5.3.4 Comparison of Augmented Reduced Basis Space with Planewave Basis Set

We first consider only the computational cost at the online stage. Our first comparison is based on the problem given in Section 5.2.5 where we would like to determine the approximation of $\hat{\boldsymbol{\lambda}}$ and F to an relative error of 0.01 and absolute error of 1E−4 respectively for $\mathbf{k} = \frac{2\pi}{a}(0.6223, 0.2953, 0)$. The

	W_N^A		Planewave		
N	time, s	$\varepsilon^{\lambda}_{N,n_b}$	$\mathcal{N}$	time, s	$\varepsilon^{\lambda}_{\mathcal{N},n_b}$
8	0.0006	6.07E-2			
16	0.005	5.27E-2	59	0.04	1.91E-2
24	0.005	1.07E-3	113	0.07	3.62E-3
32	0.007	6.75E-4	169	0.1	4.17E-4
40	0.01	2.11E-7	531	0.5	1.49E-7
48	0.02	6.75E-8	645	0.7	5.54E-8
56	0.02	2.88E-8	749	0.9	9.25E-9

Table 5.5: Comparison of the computational cost of reduced basis method and planewave method required to achieve similar level of accuracy. Comparison is made based on $\mathbf{k} = \frac{2\pi}{a}(0.2, 0.2, 0.2)$.

smallest N required for the augmented reduced basis approximation is 16 while for approximation based on planewave basis set, it is 197. We have reduced the size of the problem by approximately a factor of 10. In Table 5.5, we compare the computational cost required by reduced basis method and planewave method to achieve a similar level of approximation error for an arbitrary $\mathbf{k}$-point. Here, we choose $\mathbf{k}$ to be $\frac{2\pi}{a}(0.2, 0.2, 0.2)$. We observe that the computational saving achieved ranges from a factor of 10 to 40. For both, we use the eigenvalue solvers in MATLAB: for the reduced basis approximation, we use full eigensolver command `eig` while for the planewave approximation, we use sparse eigensolver command `eigs`. In addition, for the planewave approximation, approximately 50% of the time is spent in construction of the discrete matrices — if these matrices are pre-computed as with reduced basis method, the performance of the planewave method may be improved.

We now take the computational cost of the offline stage into consideration. If we use the actual error $\epsilon^{\lambda}_{N,n_b}(\mathbf{k})$ given by (5.50) in our adaptive sampling procedure, we then need to first obtain $n_T \equiv dim \Xi_T$ "truth" solutions to (5.16). If our objective is to solve (5.16) for number of parameter points much greater than dim Ξ_T, then our reduced basis approximation will be competitive. However, with *a posteriori* error estimator, we can reduce the number of "truth" solutions required to N_s, where N_s is usually very small. This significantly reduces the overall offline computational cost as described in Section 2.3.6. We shall thus use the *a posteriori* error estimator for $\hat{\boldsymbol{\lambda}}_N(\mathbf{k})$, $\triangle^{\lambda}_{N,1}(\mathbf{k})$, to be outlined in Section 5.5, in constructing the reduced basis approximation space.

The total offline computational cost is also determined by the maximum N_s, $N_{s,\max}$, usually

chosen based on the highest accuracy we would like our approximation to be. However, since $\Delta^{\lambda}_{N,1}(\mathbf{k})$ is not sharp and diverges as N_s (and thus N) increases (as explained in Section 5.5), $N_{s,\max}$ can be rather large if the tolerance specified is too small. In this section, we choose $N_{s,\max}$ to be 14 so that it corresponds to the maximum value of N in Table 5.5 [3].

For $n_b = 4$, the offline stage requires a total computational time of 67s. Even with the *a posteriori* error estimation procedure, there must be a need to evaluate (5.16) at more than 700 **k**-points in order to justify the offline computational cost, assuming we only require $\varepsilon^{\lambda}_{N,n_b}$ to be of $O(10^{-4})$. This emphasizes the many query limit in which reduced basis method is most useful for. We shall provide in Section 5.4 some examples where we indeed need to determine band energies at many **k**-points.

5.3.5 Extension to *Ab Initio* Models

In a typical calculation based on pseudopotential Density Functional Theory model, V_{eff} is either not explicitly constructed, for example, due to the use of nonlocal pseudopotential operator, or not easily accessible to the user. The inaccessibility of V_{eff} does not allow the construction of the discrete reduced basis matrix $A^{N,1}_{i,j} = a_1(\zeta_j, \zeta_i; V_{\text{eff}}), 1 \le i,j \le N$ as outlined in Section 5.3.1 or Section 5.3.2. Here we shall demonstrate a trick by which we obtain $A^{N,1}$ based solely on the solutions $(\hat{\mathbf{u}}(\mathbf{k}), \hat{\boldsymbol{\lambda}}(\mathbf{k}))$, $\mathbf{k} \in S^A_N$, which are typical outputs of any electronic structure calculation. We illustrate the construction of $A^{N,1}$ for the augmented reduced basis approximation; similar procedure can be applied to the vectorial reduced basis space with little modification.

Suppose we are given a sample set $S^A_N = \{\mathbf{k}_1, \ldots, \mathbf{k}_{N_s}\}$ and associated solutions $(\hat{\mathbf{u}}(\mathbf{k}_n), \hat{\boldsymbol{\lambda}}(\mathbf{k}_n)) \in (Y^{n_b} \times \mathbb{R}^{n_b})$, $1 \le n \le N_s$. From (5.16), we can write

$$a_1(u_m(\mathbf{k}_n), u_{m'}(\mathbf{k}_{n'}); V_{\text{eff}}) = \lambda_{m'}(\mathbf{k}_{n'}) m(u_m(\mathbf{k}_n), u_{m'}(\mathbf{k}_{n'})) - \sum_{l=1}^{3} k_{l,n'} a_{2,l}(u_m(\mathbf{k}_n), u_{m'}(\mathbf{k}_{n'})), \quad (5.63)$$

for $1 \le m, m' \le n_b$ and $1 \le n, n' \le N_s$ since all $u_m(\mathbf{k}_n)$, $1 \le m \le n_b$, $1 \le n \le N_s$ reside in the

[3] In Section 5.4, we instead choose a (very) rough tolerance criteria of $\Delta^{\lambda}_{N,1} \le 1\,\mathrm{E}-2$ for terminating the "greedy" algorithm.

same space Y. In addition,

$$\zeta_i = \sum_{m=1}^{n_b} \sum_{n=1}^{N} \alpha_{m,n}^i u_m(k_n), \quad 1 \le i \le N, \tag{5.64}$$

where $N = N_s n_b$ and $\alpha_{m,n}^i$ are known from our orthogonalization procedure. The matrix $A^{N,1}$ is then simply given by

$$A_{i,j}^{N,1} = \sum_{m=1}^{n_b} \sum_{n=1}^{N} \sum_{m'=1}^{n_b} \sum_{n'=1}^{N} \alpha_{m,n}^i \alpha_{m',n'}^j a_1(u_m(\mathbf{k}_n), u_{m'}(\mathbf{k}_{n'}); V_{\text{eff}}), \quad 1 \le i,j \le N. \tag{5.65}$$

We have demonstrated the reduced basis solutions obtained through this procedure is identical for our numerical example.

The equation (5.63) is based on the formulation given by (5.16). However, there is an additional term $-\frac{1}{2}|\mathbf{k}|^2 m(u_m(\mathbf{k}_n), u_{m'}(\mathbf{k}_{n'}))$ on the right hand side in the more common formulation used in computational chemistry codes, such as ABINIT [43]. This is because these codes usually do not work with the parameterized form of (5.16); instead the following equations are solved: we find $u_i^{\mathcal{N}}(\mathbf{k}) \in Y^{\mathcal{N}}(\mathbf{k})$, $1 \le i \le n_b$, $E_i^{\mathcal{N}}(\mathbf{k}) \in \mathbb{R}$, $1 \le i \le n_b$ such that

$$\frac{1}{2}\int_\Omega \nabla u_i^{\mathcal{N}}(\mathbf{k}) \nabla v + \int_\Omega u_i^{\mathcal{N}}(\mathbf{k}) V_{\text{eff}} v = E_i^{\mathcal{N}}(\mathbf{k}) \int_\Omega u_i^{\mathcal{N}}(\mathbf{k}) v, \; v \in Y^{\mathcal{N}}(\mathbf{k}), \; 1 \le i \le n_b, \tag{5.66}$$

$$\int_\Omega u_i^{\mathcal{N}}(\mathbf{k}) \; u_j^{\mathcal{N}}(\mathbf{k}) = \delta_{i,j}, \quad 1 \le i \le j \le n_b; \tag{5.67}$$

where $Y^{\mathcal{N}}(\mathbf{k}) \equiv \text{span}\,\{e^{i(\mathbf{k}+\mathbf{G})\mathbf{x}}, \mathbf{x} \in \Omega, \frac{1}{2}|\mathbf{k}+\mathbf{G}|^2 \le E_{\text{cut}}\}$. The parameter $\mathbf{k}$ then enters the above equation through the basis set used, leading to the addition term $-\frac{1}{2}|\mathbf{k}|^2 m(u_m(\mathbf{k}_n), u_{m'}(\mathbf{k}_{n'}))$.

The formulation (5.66) also leads to a slight difficulty in applying this procedure as a post-processing tool to existing computational chemistry codes. The dependence of the approximation space on $\mathbf{k}$ means $\hat{\mathbf{u}}(\mathbf{k}_n)$, $1 \le n \le N_s$ do not all belong to the same approximation space. As a result, for $n \ne n'$, $\hat{\mathbf{u}}(\mathbf{k}_n)$ does not correspond to a test function for $\hat{\mathbf{u}}(\mathbf{k}_{n'})$, and (5.63) is not exactly satisfied. This discrepancy is sufficiently significant that the resulting $A^{N,1}$ is non-Hermitian, leading to complex eigenvalues with non-negligible imaginary component, especially as N_s increases. Hence only if we are able to use a consistent basis set for determining $(\hat{\mathbf{u}}(\mathbf{k}), \hat{\boldsymbol{\lambda}}(\mathbf{k}))$ for all $\mathbf{k} \in \mathcal{D}$, this procedure will allow us to construct a reduced basis approximation for our band structure

problem without explicit knowledge of V_{eff} from outputs of any electronic structure codes.

5.4 Applications in Determination of Spectral Properties of Solids

We use outputs $E_i(\mathbf{k})$, $1 \leq i \leq n_b$ of band structure calculations to determine spectral properties of solids [40]. We will concentrate on 3 properties, namely the integrated density of states, the joint density of states and the dielectric function. In this thesis, $E_i(\mathbf{k})$ are solutions to (5.14), a linear eigenvalue problem derived from an empirical pseudopotential model. Determination of spectral properties of solids can also be a post-processing step following a full *ab initio* calculation.

There are other applications in which the reduced basis method for linear eigenvalue problems can be useful within the computational chemistry context. For example, in *ab initio* calculations based on Density Functional Theory models, each fixed point iteration in the SCF scheme (Section 4.2.2) may require solutions to an linear eigenvalue problem at n_k k-points — these solutions are then used to accurately determine the electron density, and related functionals [76]. If n_k required is large, for example in DFT calculations for metals, a reduced basis approximation within each iteration can significantly speed up evaluations of the n_k eigensolutions, thus improving the overall efficiency of the SCF algorithm.

In the next three sections, we will use energy unit eV so that direct comparison with results in literature can be made easily. Our problems are formulated in atomic units (as emphasized in Section 1.5) — 1 a.u. of energy is equivalent to 27.21 eV.

5.4.1 Preliminaries

Here, we would like to first describe the notion of density of states for our problem (5.14).

Definition 5.1. *Let $E_n(\mathbf{k})$ be eigenvalues of (5.14) ordered by $E_1(\mathbf{k}) \leq E_2(\mathbf{k}) \leq \ldots$. The density of states measure ρ is the measure on $\mathbb{R}$ defined by*[4]

$$\rho(\infty, E] = \frac{1}{|\mathcal{D}|} \sum_n |\{k \in \mathcal{D}, E_n(\mathbf{k}) \leq E\}|, \tag{5.68}$$

where $|\mathcal{D}|$ is the Lebesgue measure of $\mathcal{D}$ (given by the volume of $\mathcal{D}$) and $|\{\ldots\}|$ is the Lebesgue measure of $\{\ldots\}$. [103]

[4]We have ignored the spin degree of freedom — thus we leave out a factor of 2.

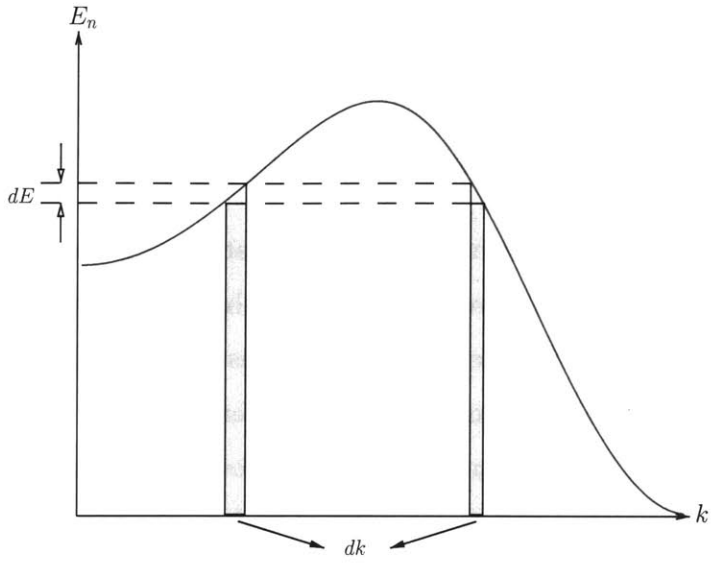

Figure 5-12: Pictorial representation of the Lebesgue sum in one dimension.

Remark 5.1. *Since $E_n(\mathbf{k}) \to \infty$ uniformly in $\mathbf{k}$ as $n \to \infty$, $\rho(-\infty, E]$ is finite. In addition ρ is absolutely continuous with respect to dE, Lebesgue measure on $\mathbb{R}$. The Radon-Nikodym derivative $d\rho/dE$ is called the density of states. [103]*

We denote the density of states, $d\rho/dE$, by $g(E)$. From (5.68) and Remark 5.1, we can express $g(E)$ as [60]

$$g(E) = \frac{1}{|\mathcal{D}|} \sum_n \int_{\mathcal{D}} \delta(E - E_n(\mathbf{k}))d\mathbf{k}, \tag{5.69}$$

where $\delta(\,\cdot\,)$ is the Kronecker delta function. Then, given a small change of E, dE, the change in the number of states, $|\mathcal{D}|d\rho = g(E)dE$ is given by the summation of the corresponding $d\mathbf{k}$ — the evaluation of this integral is illustrated for a single energy in one dimension in Figure 5-12.

The energy bands, $E_n(\mathbf{k})$ are also divided two categories, as shown in Figure 5-13. The first category, called valence bands, consists of the first n_v bands, i.e. $E_1(\mathbf{k}), \ldots, E_{n_v}(\mathbf{k})$, where n_v is the number of valence electrons — for silicon, $n_v = 4$. The second category, called conduction bands, consists of the rest of energy bands. In order to function as charge carriers, electrons must first be excited from the valence bands to the conduction bands. This division only exists in insulators; the gaps between valence bands and conduction bands determines how well an insulator can function as a semiconductor. No such division exists in conductors.

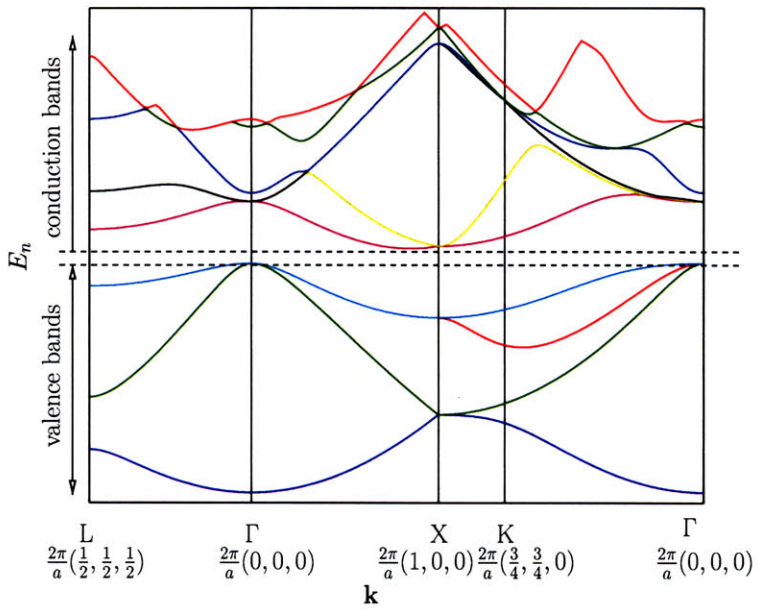

Figure 5-13: Division between valence bands and conduction bands for crystalline silicon.

5.4.2 Integrated Density of States

Problem Statement

Given an energy level E_0, the integrated density of states $I(E_0)$ is given by

$$I(E) = \int_{-\infty}^{E} g(\mathcal{E}) d\mathcal{E} \; ; \tag{5.70}$$

where $g(\,\cdot\,)$ is as defined in (5.69). For the current analysis, we are interested in determining $I_{\text{dif}}(6$ eV) where

$$I_{\text{dif}}(E) = I(E) - I(E_4(\mathbf{k} = (0,0,0))); \tag{5.71}$$

$E_4(\mathbf{k})$ at $\mathbf{k} = (0,0,0)$ corresponds to the highest energy in valence bands — this is simply to make the problem more interesting since we know $I(E_4(\mathbf{k}))$ for $\mathbf{k} = (0,0,0)$ is 4. Nevertheless, it is worthy to point out that we are typically interested in the inverse problem, where we would like to determine a E_0 such that $I(E_0)$ is equal to a certain value. For example, in the determination of the Fermi level, we would like $I(E_0)$ to be equal to the total valence charge, n_v.

Standard Techniques

Two standard techniques routinely used for the determination of the density of states and the integrated density of states are the Gaussian smearing technique and the tetrahedron technique. The Gaussian smearing technique involves convoluting the delta function with a Gaussian function [40]. The tetrahedron technique, on the other hand, partitions the Brillouin zone into several tetrahedra and assume a linear interpolation scheme within each tetrahedron [14, 58]. Although the Gaussian smearing technique is more widely used, the tetrahedron technique is usually more accurate [40] and will thus be used in the current analysis.

With the tetrahedron technique, we first partition the irreducible Brillouin zone into a set of tetrahedra $\mathcal{T}$, as shown in Figure 5-14. Then, for each tetrahedron $T \in \mathcal{T}$, linear interpolation of values at the vertices of T is assumed. An analytical expression of I can then be derived [14]: for a given band i, let $E_{i,1}^T$, $E_{i,2}^T$, $E_{i,3}^T$ and $E_{i,4}^T$ be the E_i at the vertices of the tetrahedron T such that $E_{i,1}^T \leq E_{i,2}^T \leq E_{i,3}^T \leq E_{i,4}^T$. We further denote $E_{i,mn}^T$ by $E_{i,m}^T - E_{i,n}^T$, V^T as the volume of the tetrahedron T, and $V^G = \sum_{T \in \mathcal{T}} V_T$. The integrated density of states $I(E)$ is then given by

$$I(E) = \sum_{i=1}^{n_b} \sum_{T \in \mathcal{T}} I_i^T(E), \tag{5.72}$$

where [14]

$$I_i^T(E) = \begin{cases} 0, & E < E_{i,1}^T \\ \frac{V^T}{V^G} \frac{(E - E_{i,1}^T)^3}{E_{i,21}^T E_{i,31}^T E_{i,41}^T}, & E_{i,1}^T < E < E_{i,2}^T \\ \frac{V^T}{V^G} \frac{1}{E_{i,31}^T E_{i,41}^T} ((E_{i,21}^T)^2 + 3E_{i,21}^T (E - E_{i,2}^T) + \ldots & \\ \qquad 3(E - E_{i,2}^T)^2 - \frac{E_{i,31}^T + E_{i,42}^T}{E_{i,32}^T E_{i,42}^T} (E - E_{i,2}^T)^3), & E_{i,2}^T < E < E_{i,3}^T \\ \frac{V^T}{V^G} (1 - \frac{(E_{i,4}^T - E)^3}{E_{i,41}^T E_{i,42}^T E_{i,43}^T}, & E_{i,3}^T < E < E_{i,4}^T \\ \frac{V^T}{V^G}, & E > E_{i,4}^T \end{cases} \tag{5.73}$$

We now introduce a Fourier basis set $Y^{\mathcal{N}}$ of size $\mathcal{N}$ and approximate E_n, $1 \leq n \leq n_b$ by $E_n^{\mathcal{N}}$, $1 \leq n \leq n_b$, the planewave approximation based on $Y^{\mathcal{N}}$. We define $I_{\text{dif},n_k,\mathcal{N}}(E)$ as

$$I_{\text{dif},n_k,\mathcal{N}}(E) = I_{n_k,\mathcal{N}}(E) - I(E_4(0)), \tag{5.74}$$

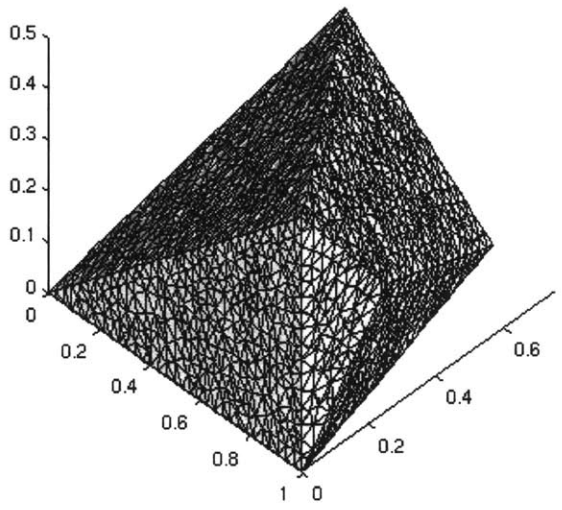

Figure 5-14: $\mathcal{T}$, the tetrahedra mesh of $\mathcal{D}$ with 4440 vertices.

where $I_{n_k,\mathcal{N}}(\cdot)$ is the integrated density of states computed on a tetrahedra mesh $\mathcal{T}$ with n_k vertices based on $E_n^{\mathcal{N}}(\mathbf{k})$ instead of $E_n(\mathbf{k})$. We further introduce a "truth" approximation, $I_{\text{dif},0}(\cdot)$ given by $I_{\text{dif},n_k,\mathcal{N}}(\cdot)$ evaluated at $n_k = n_{k,t} = 4440$ and $\mathcal{N} = \mathcal{N}_t = 1807$; $n_{k,t}$ and $\mathcal{N}_t$ are simply the dimensions of our "truth" approximation. Table 5.6 shows that for $|I_{\text{dif},n_k,\mathcal{N}} - I_{\text{dif},0}| \leq 0.01$, a combination of $\mathcal{N} = 137$ and $n_k = 572$ suffices; the computational cost is 62s.

Reduced Basis Approach

We now introduce a reduced basis approximation of dimension N and approximate $E_i(\mathbf{k})$, $1 \leq i \leq n_b$ by a reduced basis approximant, $E_{N,i}(\mathbf{k})$, $1 \leq i \leq n_b$. We define $I_{\text{dif},n_k,N}(E)$ as

$$I_{\text{dif},n_k,N}(E) = I_{n_k,N}(E) - I(E_4(0)), \tag{5.75}$$

where $I_{n_k,N}(\cdot)$ is the integrated density of states computed on a tetrahedra mesh $\mathcal{T}$ with n_k vertices based on $E_{N,n}(\mathbf{k})$ instead of $E_n(\mathbf{k})$. Note that we build a reduced basis approximation for eigensolutions up to $n_b = 9$ since the $\min_{\mathbf{k}\in\mathcal{D}} E_{N,9}(\mathbf{k}) - E_4(\mathbf{k} = (0,0,0))$ is greater than 6 eV.

Table 5.7 shows that we only require $N = 36$ and $n_k = 572$ to achieve the convergence criteria $|I_{\text{dif},n_k,N} - I_{\text{dif},0}| \leq 1\,\text{E}{-2}$. For this calculation, the total computational cost during the online stage is 3.8s. Comparing this to the 26s required by the planewave method, the reduced basis method is 7 times faster than the planewave method. However, we note that the offline computational cost

	$\lvert I_{\text{dif},n_k,\mathcal{N}} - I_{\text{dif},0} \rvert$			
n_k	$\mathcal{N} = 137$	$\mathcal{N} = 229$	$\mathcal{N} = 339$	$\mathcal{N} = 531$
6	1.9780E−1	8.0590E−1	1.9373E−1	1.9373E−1
8	1.5740E−1	1.5510E−1	1.5490E−1	1.5491E−1
15	1.1345E−1	1.1113E−1	1.1098E−1	1.1099E−1
23	1.5392E−1	1.5242E−1	1.5229E−1	1.5229E−1
38	3.7813E−2	3.5717E−2	3.5525E−2	3.5531E−2
58	2.2991E−2	2.0081E−2	1.9834E−2	1.9841E−2
80	4.8691E−2	4.6003E−2	4.5785E−2	4.5793E−2
114	3.2089E−2	2.9134E−2	2.8873E−2	2.8880E−2
149	2.7976E−2	2.5023E−2	2.4771E−2	2.4779E−2
202	1.8873E−2	1.5884E−2	1.5634E−2	1.5641E−2
249	2.2724E−2	1.9632E−2	1.9370E−2	1.9378E−2
324	1.5065E−2	1.1909E−2	1.1640E−2	1.1648E−2
389	1.4614E−2	1.1588E−2	1.1336E−2	1.1344E−2
490	1.1443E−2	8.3680E−3	8.1124E−3	8.1200E−3
572	6.1874E−3	3.3251E−3	3.0825E−3	3.0897E−3

Table 5.6: Variation of error in $I_{\text{dif},n_k,\mathcal{N}}$ with n_k and $\mathcal{N}$. $I_{\text{dif},0}$ is the "truth" approximation given by $I_{\text{dif},n_k,\mathcal{N}}$ computed at $n_k = 4440$ and $\mathcal{N} = 1807$.

(based on a "greedy" sampling algorithm that utilizes the *a posteriori* error estimation procedure and tolerance criteria $\Delta^{\lambda}_{N,1}(\mathbf{k}) \leq 10^{-2}$) is 102s. Reduced basis method is thus only competitive if we need to evaluate the integrated density of states at more than 6 different values of E — a requirement easily fulfilled when considering the inverse problem of determining a E_0 given $I(E_0)$, which we have described earlier.

5.4.3 Joint Density of States

Problem Statement

Given a photon energy E, we would like to determine the joint density of states defined as

$$J(E) = \sum_{i=1}^{n_v} \sum_{j=n_v+1}^{n_b} J_{i,j}(E), \tag{5.76}$$

where

$$J_{i,j}(E) = \frac{1}{|\mathcal{D}|} \int_{\mathcal{D}} \delta(E_j(\mathbf{k}) - E_i(\mathbf{k}) - E) d\mathbf{k}, \tag{5.77}$$

	$\lvert I_{\text{dif},n_k,N} - I_{\text{dif},0} \rvert$			
n_k	$N = 9$	$N = 18$	$N = 27$	$N = 36$
6	6.6524E−1	5.7391E−1	7.9422E−1	8.0044E−1
8	8.4275E−1	4.6119E−2	1.4645E−1	1.5495E−1
15	8.6394E−1	3.0609E−1	2.7881E−2	2.2031E−2
23	7.8922E−1	3.1237E−1	1.2205E−1	1.1781E−1
38	7.6262E−1	2.6901E−1	2.8294E−2	2.2474E−2
58	7.6250E−1	2.5123E−1	1.6408E−2	1.2404E−2
80	7.8762E−1	2.8170E−1	5.0466E−2	4.4836E−2
110	7.6507E−1	2.5271E−1	3.3282E−2	2.9116E−2
149	7.6276E−1	2.5477E−1	3.2175E−2	2.7561E−2
202	7.6703E−1	2.5138E−1	2.2609E−2	1.8471E−2
249	7.6386E−1	2.4893E−1	2.8111E−2	2.3542E−2
324	7.6605E−1	2.4583E−1	1.9893E−2	1.5752E−2
389	7.6163E−1	2.4455E−1	1.9957E−2	1.5857E−2
490	7.6826E−1	2.3807E−1	1.6392E−2	1.2511E−2
572	7.6307E−1	2.4099E−1	1.1567E−2	7.7170E−3

Table 5.7: Variation of error in $I_{\text{dif},n_k,N}$ with n_k and N. $I_{\text{dif},0}$ is is the "truth" approximation given by $I_{\text{dif},n_k,\mathcal{N}}$ computed at $n_k = 4440$ and $\mathcal{N} = 1807$.

where n_v is the number of valence bands; and $n_b - n_v$ is the number of conduction bands we need to examine, depending on the range of E we are interested in — larger E will require us to consider higher n_b. For silicon, the number of valence bands is 4, and we will look at 8 conduction bands; thus $n_b = 12$. The joint density of states then determines the density of pairs of states (one from valence bands and the other from conduction bands) that have energy difference E.

The joint density of states is particularly useful in molecular dynamics simulations in solid state physics. For example, it determines the transition rate of an electron from a ground state (one of the valence bands) to an excited state (one of the conduction bands), relevant to simulation of charge transport phenomena in semiconductor [38, 57]. It can also be used for the determination of the dielectric function, as described in Section 5.4.4.

Standard Techniques

We will again use the tetrahedron technique. For each tetrahedron $T \in \mathcal{T}$ and two indices i and j, let $\Delta^T_{i,j,m} = E^T_{j,m} - E^T_{i,m}$, $1 \le m \le 4$ and $\Delta^T_{i,j,1} < \Delta^T_{i,j,2} < \Delta^T_{i,j,3} < \Delta^T_{i,j,4}$. In addition,

$\Delta^T_{i,j,mn} = \Delta^T_{i,j,m} - \Delta^T_{i,j,n}$. Then, the joint density states is given by

$$J_{i,j}(E) = \sum_{T \in \mathcal{T}} J^T_{i,j}(E), \tag{5.78}$$

where [14, 58]

$$J^T_{i,j}(E) = \begin{cases} 0, & E < \Delta^T_{i,j,1} \\ \frac{V^T}{V^G} \frac{3(E-\Delta^T_{i,j,1})^2}{\Delta^T_{i,j,21}\Delta^T_{i,j,31}\Delta^T_{i,j,41}}, & \Delta^T_{i,j,1} < E < \Delta^T_{i,j,2} \\ \frac{V^T}{V^G} \frac{1}{\Delta^T_{i,j,31}\Delta^T_{i,j,41}}((3\Delta^T_{i,j,21}) + 6(E - \Delta^T_{i,j,2}) & \\ \qquad -3\frac{\Delta^T_{i,j,31}+\Delta^T_{i,j,42}}{\Delta^T_{i,j,32}\Delta^T_{i,j,42}}(E - \Delta^T_{i,j,2})^2), & \Delta^T_{i,j,2} < E < \Delta^T_{i,j,3} \\ \frac{V^T}{V^G} \frac{3(\Delta^T_{i,j,4}-E)^2}{\Delta^T_{i,j,41}\Delta^T_{i,j,42}\Delta^T_{i,j,43}}, & \Delta^T_{i,j,3} < E < \Delta^T_{i,j,4} \\ 0, & E > \Delta^T_{i,j,4} \end{cases} \tag{5.79}$$

It is computationally more intensive than the calculation of the integrated density of states due to the additional double summations.

As in previous section, we introduce a Fourier basis set $Y^{\mathcal{N}}$ of size $\mathcal{N}$ and approximate $E_n(\mathbf{k})$, $1 \le n \le n_b$ by $E^{\mathcal{N}}_n(\mathbf{k})$, $1 \le n \le n_b$, the planewave approximation based on $Y^{\mathcal{N}}$. We then define $J_{\mathcal{N}}(E)$ as

$$J_{\mathcal{N}}(E) = \sum_{i=1}^{n_v} \sum_{j=n_v+1}^{n_b} J_{\mathcal{N},i,j}(E), \tag{5.80}$$

where $J_{\mathcal{N},i,j}(E)$ is evaluated from (5.78) with $E_n(\mathbf{k})$, $1 \le n \le n_b$ replaced by $E^{\mathcal{N}}_n(\mathbf{k})$, $1 \le n \le n_b$. We evaluate $J_{\mathcal{N},i,j}(\cdot)$ on a tetrahedra mesh $\mathcal{T}$ with 4440 vertices, shown in Figure 5-14. Similarly, we introduce a "truth" approximation, $J_{\mathcal{N}_t}(\cdot)$ given by $J_{\mathcal{N}}(\cdot)$ evaluated at $\mathcal{N} = \mathcal{N}_t = 1807$. The size of n_k is dependent on the accuracy we seek for approximate $J(\cdot)$; our choice here is simply for the purpose of demonstrating the utility of reduced basis method in the large n_k limit.

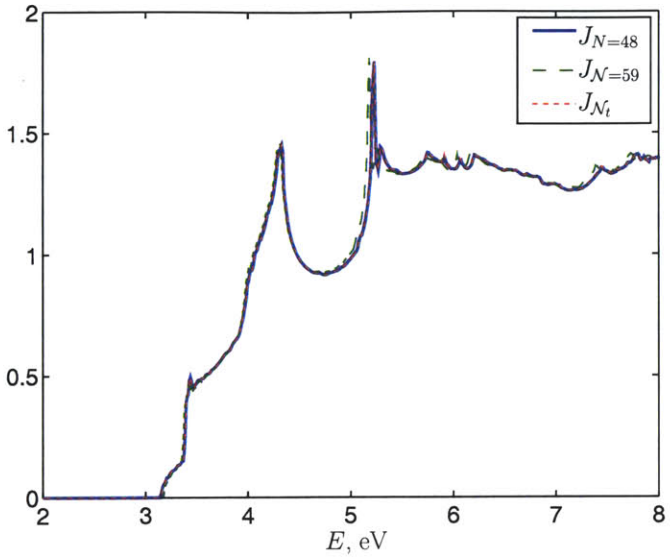

Figure 5-15: Different approximations to the joint density of states — $J_{N=48}$, $J_{\mathcal{N}=59}$ and $J_{\mathcal{N}=\mathcal{N}_t=1807}$ — versus energy E in eV.

Reduced Basis Approach

We introduce a reduced basis approximation of dimension N and approximate $E_i(\mathbf{k})$, $1 \leq i \leq n_b$ by a reduced basis approximant, $E_{N,i}(\mathbf{k})$, $1 \leq i \leq n_b$. We then define $J_N(E)$ as

$$J_N(E) = \sum_{i=1}^{n_v} \sum_{j=n_v+1}^{n_b} J_{N,i,j}(E), \tag{5.81}$$

where $J_{N,i,j}(E)$ is evaluated from (5.78) with $E_n(\mathbf{k})$, $1 \leq n \leq n_b$ replaced by $E_{N,n}(\mathbf{k})$, $1 \leq n \leq n_b$. Again, we evaluate $J_{N,i,j}(\,\cdot\,)$ on a tetrahedra mesh $\mathcal{T}$ with 4440 vertices, shown in Figure 5-14.

From Figure 5-15, we see that there is no discernible difference between the reduced basis approximation J_N for $N = 48$ and the "truth" approximation, $J_{\mathcal{N}_t}$. The online and offline computational costs of determining J_N for $N = 48$ are shown in Table 5.8. The offline computational cost is evaluated based on a "greedy" sampling algorithm that utilizes the *a posteriori* error estimation procedure and a tolerance criteria of $\triangle_{N,1}^{\lambda} \leq 10^{-2}$. This gives $N_{s,\max} = 7$ and an offline computational cost of 85s. The online computational cost is 88s, thus giving a total computational cost of 173s.

We now consider the planewave approximation, $J_{\mathcal{N}}$. In particular, we determine a $\mathcal{N}$ for which the evaluation of $J_{\mathcal{N}}$ will take approximately the same total amount of time (online + offline) as

	J_N	$J_{\mathcal{N}}$
Dimension	$N = 48$	$\mathcal{N} = 59$
Computational time	Online : 88s Offline : 85s Total : 173s	Total : 151s

Table 5.8: The computational cost of J_N and $J_{\mathcal{N}}$ for the results shown in Figure 5-15.

that for J_N. We see from Table 5.8 that this is given by $\mathcal{N} = 59$. From Figure 5-15, we note that $J_{\mathcal{N}=59}$ deviates from $J_{\mathcal{N}_t}$, and fails to give the correct location for the second peak of J. This slight deviation may not be significant physically in this particular case, but it does demonstrate the higher accuracy that reduced basis approximation can achieve given same amount of computing resources.

5.4.4 Dielectric Function

Problem Statement

Dielectric function describes the response of a crystalline solid to an external electric field. It is a measurable physical quantity, and as such acts as a valuable tool in validating quantum models through experimental results. For other practical applications, dielectric function is essential in the design of capacitor and dielectric waveguide, such as optical fibers [46, 60].

Since the real and imaginary part of the dielectric function are related by the Kramers-Kronig relations, we only need to determine the imaginary component of the dielectric function, $\epsilon_2(E)$, given by [60]

$$\epsilon_2(E) = \left(\frac{2\pi}{E}\right)^2 \sum_{i=1}^{n_v} \sum_{j=n_v+1}^{n_b} K_{i,j}(E), \tag{5.82}$$

where

$$K_{i,j}(E) = \frac{1}{|\mathcal{D}|} \int_{\mathcal{D}} \left(\sum_{\ell=1}^{3} |a_{2,\ell}(u_i(\mathbf{k}), u_j(\mathbf{k}))|^2 \right) \delta(E_j(\mathbf{k}) - E_i(\mathbf{k}) - E) d\mathbf{k}, \tag{5.83}$$

where $a_{2,\ell} = -\mathrm{i} \int_\Omega \frac{\partial w}{\partial x_\ell} v^*$. We choose $n_v = 4$ and $n_b = 12$ for the same reasons given in Section 5.4.3.

Standard Techniques

We will again use the tetrahedron technique. There are two ways in which we can treat the $a_{2,\ell}(u_i(\mathbf{k}), u_j(\mathbf{k}))$, $\ell = 1, \ldots, 3$ term. In [122], $a_{2,\ell}(u_i(\mathbf{k}), u_j(\mathbf{k}))$ is assumed to be weakly dependent on $\mathbf{k}$ and as such an average value $P_{i,j} = \langle \sum_{\ell=1}^{3} a_{2,\ell}(u_i(\,\cdot\,), u_j(\,\cdot\,)) \rangle$ can be used. Then, $K_{i,j}(E) = P_{i,j} J_{i,j}(E)$ — $J_{i,j}(E)$ can be computed based on the procedure outlined in Section 5.4.3. The other approach assumes a linear interpolation of $a_{2,\ell}(u_i(\mathbf{k}), u_j(\mathbf{k}))$ within each tetrahedron; the resulting formulation is more complicated and it is given in [68]; this is the formulation we will use.

As in previous two sections, we introduce a Fourier basis set $Y^{\mathcal{N}}$ of size $\mathcal{N}$ and approximate $(u_n(\mathbf{k}), E_n(\mathbf{k}))$, $1 \le n \le n_b$ by $(u_n^{\mathcal{N}}(\mathbf{k}), E_n^{\mathcal{N}}(\mathbf{k}))$, $1 \le n \le n_b$, the planewave approximation based on $Y^{\mathcal{N}}$. We then define $\epsilon_{2,\mathcal{N}}(E)$ as

$$\epsilon_{2,\mathcal{N}}(E) = \sum_{i=1}^{n_v} \sum_{j=n_v+1}^{n_b} K_{\mathcal{N},i,j}(E), \tag{5.84}$$

where $K_{\mathcal{N},i,j}(E)$ is evaluated from with $(u_n(\mathbf{k}), E_n(\mathbf{k}))$, $1 \le n \le n_b$ replaced by $(u_n^{\mathcal{N}}, E_n^{\mathcal{N}}(\mathbf{k}))$, $1 \le n \le n_b$. Here, in addition to the calculation of the energy difference at each $\mathbf{k}$, we also need to determine $a_{2,\ell}(u_i^{\mathcal{N}}(\mathbf{k}), u_j^{\mathcal{N}}(\mathbf{k}))$ which can be computationally expensive if $\mathcal{N}$ is large. Similarly, we introduce a "truth" approximation, $\epsilon_{2,\mathcal{N}_t}(\,\cdot\,)$ given by $\epsilon_{2,\mathcal{N}}(\,\cdot\,)$ evaluated at $\mathcal{N} = \mathcal{N}_t = 1807$. We evaluate $\epsilon_{2,\mathcal{N}}(\cdot)$ on a tetrahedron mesh with $n_k = 4440$ vertices — the size of n_k is dependent on the accuracy we seek to approximate $\epsilon_2(\cdot)$; our choice here is simply for the purpose of demonstrating the utility of reduced basis method in the large n_k limit. We note that in addition to accuracy in $E_{\mathcal{N},n}$, we must also evaluate $a_{2,\ell}(u_i^{\mathcal{N}}(\mathbf{k}), u_j^{\mathcal{N}}(\mathbf{k}))$ sufficient accurately. However, error in $a_{2,\ell}(u_i^{\mathcal{N}}(\mathbf{k}), u_j^{\mathcal{N}}(\mathbf{k}))$ is of the same order as error in $E_{\mathcal{N},n}$; as such, the accuracy requirement is not higher than previous two problems.

Reduced Basis Approach

We introduce an reduced basis approximation of dimension N and approximate $(u_i(\mathbf{k}), E_i(\mathbf{k}))$, $1 \le i \le n_b$ by a reduced basis approximant, $(u_{N,i}(\mathbf{k}), E_{N,i}(\mathbf{k}))$, $1 \le i \le n_b$. We then define $\epsilon_N(E)$ as

$$\epsilon_{2,N}(E) = \sum_{i=1}^{n_v} \sum_{j=n_v+1}^{n_b} K_{N,i,j}(E), \tag{5.85}$$

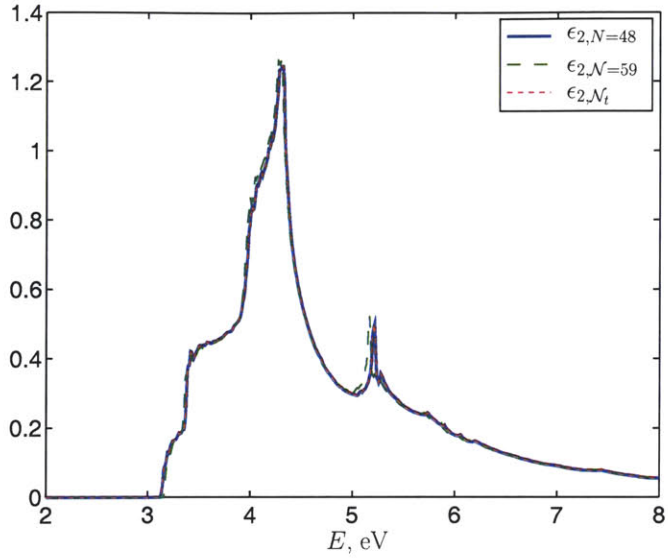

Figure 5-16: Different approximations to the dielectric function — $\epsilon_{2,N=48}$, $\epsilon_{2,\mathcal{N}=59}$ and $\epsilon_{2,\mathcal{N}=\mathcal{N}_t=1807}$ — versus energy E in eV.

	$\epsilon_{2,N}$	$\epsilon_{2,\mathcal{N}}$
Dimension	$N = 48$	$\mathcal{N} = 59$
Computational time	Online : 141s Offline : 88s Total : 229s	Total : 215s

Table 5.9: The computational cost of $\epsilon_{2,N}$ and $\epsilon_{2,\mathcal{N}}$ for the results shown in Figure 5-16.

where $K_{N,i,j}(E)$ is an approximation to $K_{i,j}(E)$ where $(u_n(\mathbf{k}), E_n(\mathbf{k}))$, $1 \leq n \leq n_b$ replaced by $(u_{N,i}(\mathbf{k}), E_{N,i}(\mathbf{k}))$, $1 \leq n \leq n_b$. Again, we evaluate $K_{N,i,j}(\,\cdot\,)$ on a tetrahedra mesh $\mathcal{T}$ with 4440 vertices, shown in Figure 5-14.

We now reuse the reduced basis approximation we have constructed in Section 5.4.3. From Figure 5-16, we see that there is no discernible difference between the reduced basis approximation $\epsilon_{2,N}$ for $N = 48$ and the "truth" approximation, $\epsilon_{2,\mathcal{N}_t}$. The computational costs of determining $\epsilon_{2,N}$ for $N = 48$ are shown in Table 5.9. Note that the offline computational cost is taken from Table 5.8. However, we can argue that since we are using the same reduced basis approximation from the previous section, the actual offline cost is 0.

We now consider the planewave approximation, $\epsilon_{2,\mathcal{N}}$. We use $\mathcal{N} = 59$ as in Section 5.4.3 which leads to a computational costs that is similar to the total computational time (online + offline) of a reduced basis approximation as shown in Table 5.9. From Figure 5-16, we note that $\epsilon_{2,\mathcal{N}=59}$

deviates from $\epsilon_{2,\mathcal{N}_t}$, and fails to give the correct location for the second peak of ϵ_2. In addition, if we do not consider the offline computational cost, we achieve a computational saving of 50% with reduced basis approximation. Again, the slight deviation may not be physically significant, but it does demonstrate the higher accuracy that a reduced basis approximation can achieve given same amount of computing resources.

5.5 *A Posteriori* Error Estimation

The derivation of the *a posteriori* error estimator for this problem follows Section 2.4 closely. The are two main differences: (*i*) the eigenvalues $\lambda_i(\mathbf{k})$ are not of multiplicity *one* and the multiplicity depends on both i and $\mathbf{k}$; and (*ii*) $a(v, v; V_{\text{eff}}; \mathbf{k})$ is not strictly positive for all $\mathbf{k} \in \mathcal{D}$.

5.5.1 Derivation

For $i = 1, \ldots, n_b$, we define the residual as

$$R_i(v; \mathbf{k}) = a(u_{N,i}(\mathbf{k}), v; V_{\text{eff}}; \mathbf{k}) - \lambda_{N,i}(\mathbf{k}) m(u_{N,i}(\mathbf{k}), v), \tag{5.86}$$

for $\forall v \in Y$. We also define a reconstructed error $\hat{e}_i$ in Y, such that

$$\hat{a}(\hat{e}_i, v) = R_i(v; \mathbf{k}), \quad \forall v \in Y, \tag{5.87}$$

where

$$\hat{a}(w, v) = a_1(w, v; V_{\text{eff}}) + \gamma m(w, v); \quad \gamma = 1 + |\lambda_1(0)|; \tag{5.88}$$

$$\|R_i(\,\cdot\,; \mathbf{k})\| \equiv \sup_{v \in Y} \frac{R_i(v; \mathbf{k})}{\hat{a}(v, v)^{1/2}} = \hat{a}(\hat{e}_i, \hat{e}_i)^{1/2}; \tag{5.89}$$

and $\|\cdot\| = \hat{a}(\,\cdot\,, \cdot\,)^{1/2}$.

We now define $a^+(w, v; V_{\text{eff}}; \mathbf{k}) = a(w, v; V_{\text{eff}}; \mathbf{k}) + \gamma m(w, v)$ and introduce the following eigenvalue problem: for $\mathbf{k} \in \mathcal{D}$, find $(\hat{\mathbf{u}}^+(\mathbf{k}), \hat{\boldsymbol{\lambda}}^+(\mathbf{k})) \in (Y^{n_b} \times \mathbb{R}^{n_b})$ such that

$$a^+(u_i^+(\mathbf{k}), v; V_{\text{eff}}; \mathbf{k}) = \lambda_i^+(\mathbf{k}) m(u_i^+(\mathbf{k}), v), \quad \forall v \in Y, \quad 1 \le i \le n_b, \tag{5.90}$$

$$m(u_i^+(\mathbf{k}), u_j^+(\mathbf{k})) = \delta_{ij}, \quad 1 \le i \le j \le n_b. \tag{5.91}$$

It is clear that $\hat{\mathbf{u}}^+(\mathbf{k}) = \hat{\mathbf{u}}(\mathbf{k})$ and $\lambda_i^+ = \lambda_i + \gamma$.

Proposition 5.1. *Given* $\hat{a}(w,v) = a_1(w,v;V_{\text{eff}}) + \gamma m(w,v)$ *and* $\gamma = 1 + |\lambda_1(0)|$, *we have*

$$a^+(v,v;\mathbf{k}) \equiv \hat{a}(v,v) \geq m(v,v) \geq 0. \tag{5.92}$$

Proof. First, we note that $a_{2,\ell}(v,v) = 0$, for $\ell = 1,\ldots,3$: let $v = v_1 + \mathrm{i}v_2$, and $v_1, v_2 \in \mathbb{R}$; then

$$\begin{aligned} a_{2,\ell}(v,v) &= -\mathrm{i}\int_\Omega \left(\frac{\partial v_1}{\partial x_\ell} + \mathrm{i}\frac{\partial v_2}{\partial x_\ell}\right)(v_1 - \mathrm{i}v_2) \\ &= -\mathrm{i}\int_\Omega \left(\frac{\partial v_1}{\partial x_\ell}v_1 + \frac{\partial v_2}{\partial x_\ell}v_2\right) - \int_\Omega \frac{\partial v_1}{\partial x_\ell}v_2 + \int_\Omega \frac{\partial v_2}{\partial x_\ell}v_1 \\ &= 0, \end{aligned} \tag{5.93}$$

since

$$\int_\Omega \frac{\partial v_1}{\partial x_\ell}v_2 = -\int_\Omega \frac{\partial v_2}{\partial x_\ell}v_1; \quad \int_\Omega \frac{\partial v_1}{\partial x_\ell}v_1 = 0; \quad \int_\Omega \frac{\partial v_2}{\partial x_\ell}v_2 = 0. \tag{5.94}$$

We first prove the left equality:

$$\begin{aligned} \hat{a}(v,v) &= a_1(v,v;V_{\text{eff}}) + \gamma m(v,v) \\ &= a(v,v;V_{\text{eff}};\mathbf{k}) + \gamma m(v,v) \\ &= a^+(v,v;V_{\text{eff}};\mathbf{k}). \end{aligned} \tag{5.95}$$

since for $\ell = 1,\ldots,3$, $a_{2,\ell}(v,v) = 0$.

To prove the right inequality, we note that

$$a_1(v,v;V_{\text{eff}}) \geq \lambda_1(0)m(v,v). \tag{5.96}$$

Then,

$$\begin{aligned} \hat{a}(v,v) &= a_1(v,v;V_{\text{eff}}) + (1 + |\lambda_1(0)|)m(v,v) \\ &\geq (1 + \lambda_1(0) + |\lambda_1(0)|)m(v,v) \\ &\geq m(v,v). \end{aligned} \tag{5.97}$$

This concludes the proof of Proposition 5.1. □

Hypothesis 5.1. *Assuming our reduced-basis approximation is convergent in the sense that*

$$\lambda_{N,i}(\mathbf{k}) \to \lambda_i(\mathbf{k}), \quad 1 \le i \le n_b, \quad as \quad N \to \infty. \tag{5.98}$$

Then, for sufficiently large N,

$$i = \arg \min_{1 \le j \le \mathcal{N}_t} \left| \frac{\lambda_j(\mathbf{k}) - \lambda_{N,i}(\mathbf{k})}{\lambda_j^+(\mathbf{k})} \right|. \tag{5.99}$$

Proposition 5.2. *Assume our reduced-basis approximation is convergent in the sense that*

$$\lambda_{N,i}(\mathbf{k}) \to \lambda_i(\mathbf{k}), \quad 1 \le i \le n_b, \quad as \quad N \to \infty. \tag{5.100}$$

Then, for large N and $i = 1, \ldots, n_b$,

$$\left| \frac{\lambda_i(\mathbf{k}) - \lambda_{N,i}(\mathbf{k})}{\lambda_i(\mathbf{k}) + \gamma} \right| \le \frac{\|R_i(\,\cdot\,;\mathbf{k})\|}{(\lambda_{N,i}(\mathbf{k}) + \gamma)^{1/2}} \tag{5.101}$$

In addition, for $\lambda_{N,i}(\mathbf{k})$ of multiplicity one and associated $u_{N,i}(\mathbf{k})$, we have

$$\|u_{N,i}(\mathbf{k}) - u_i(\mathbf{k})\| \le \frac{\|R_i(\,\cdot\,;\mathbf{k})\|}{d_i}, \tag{5.102}$$

and

$$|\lambda_{N,i}(\mathbf{k}) - \lambda_i(\mathbf{k})| \le \frac{\|R_i(\,\cdot\,;\mathbf{k})\|^2}{d_i^2}, \tag{5.103}$$

where $d_i = \min_{j \ne i} \left| \frac{\lambda_{N,j}(\mathbf{k}) - \lambda_{N,i}(\mathbf{k})}{\lambda_{N,j}(\mathbf{k}) + \gamma} \right|$.

Proof. For $i = 1, \ldots, n_b$, we define $\tilde{e} \in Y$ as

$$a^+(\tilde{e}_i, v; V_{\text{eff}}; \mathbf{k}) = R_i(v; \mathbf{k}), \quad \forall v \in Y; \tag{5.104}$$

$$|||R_i(\,\cdot\,;\mathbf{k})||| \equiv \sup_{v \in Y} \frac{R_i(v;\mathbf{k})}{a^+(v,v;\mathbf{k})^{1/2}} = a^+(\tilde{e}_i, \tilde{e}_i; \mathbf{k})^{1/2}; \tag{5.105}$$

and $|||\cdot||| = a^+(\cdot\,,\cdot\,;V_{\text{eff}};\mathbf{k})^{1/2}$. From (5.92), we then have

$$|||R_i(\cdot\,;\mathbf{k})||| \leq \|R_i(\cdot\,;\mathbf{k})\|. \tag{5.106}$$

From here onwards, the proof follows closely that of Proposition 2.2. Let $u_{N,i}(\mathbf{k}) = \sum_{j=1}^{\mathcal{N}_t} \alpha_j u_j(\mathbf{k})$ and $\tilde{e}_i = \sum_{j=1}^{\mathcal{N}_t} \beta_j u_j(\mathbf{k})$; $\mathcal{N}_t$ is the dimension of our "truth" approximation. From (5.104),

$$\begin{aligned}
a^+\left(\sum_{j'=1}^{\mathcal{N}_t} \beta_{j'} u_{j'}(\mathbf{k}), v; V_{\text{eff}};\mathbf{k}\right) &= a\left(\sum_{j=1}^{\mathcal{N}_t} \alpha_j u_j(\mathbf{k}), v; V_{\text{eff}};\mathbf{k}\right) - \lambda_{N,i}(\mathbf{k}) m\left(\sum_{j=1}^{\mathcal{N}_t} \alpha_j u_j(\mathbf{k}), v\right) \\
\sum_{j'=1}^{\mathcal{N}_t} \beta_{j'} \lambda_{j'}^+(\mathbf{k}) m(u_{j'}(\mathbf{k}), v;\mathbf{k}) &= \sum_{j=1}^{\mathcal{N}_t} \alpha_j (\lambda_j(\mathbf{k}) - \lambda_{N,i}(\mathbf{k})) m(u_j(\mathbf{k}), v;\mathbf{k}) \\
\beta_j &= \alpha_j \left(\frac{\lambda_j(\mathbf{k}) - \lambda_{N,i}(\mathbf{k})}{\lambda_j^+(\mathbf{k})}\right).
\end{aligned} \tag{5.107}$$

Then,

$$\begin{aligned}
|||R_i(\cdot\,;\mathbf{k})|||^2 &= a^+(\tilde{e}_i, \tilde{e}_i; V_{\text{eff}};\mathbf{k}) \\
&= \sum_{j=1}^{\mathcal{N}_t} \beta_j^2 \lambda_j^+(\mathbf{k}) m(u_j(\mathbf{k}), u_j(\mathbf{k})) \\
&= \sum_{j=1}^{\mathcal{N}_t} \alpha_j^2 \left(\frac{\lambda_j(\mathbf{k}) - \lambda_{N,i}(\mathbf{k})}{\lambda_j^+(\mathbf{k})}\right)^2 \lambda_j^+(\mathbf{k}).
\end{aligned} \tag{5.108}$$

Dividing by $\lambda_{N,i}(\mathbf{k}) + \gamma$, we have

$$\begin{aligned}
\frac{|||R_i(\cdot\,;\mathbf{k})|||^2}{\lambda_{N,i}(\mathbf{k}) + \gamma} &= \frac{1}{a^+(u_{N,i}(\mathbf{k}), u_{N,i}(\mathbf{k}); V_{\text{eff}};\mathbf{k})} \sum_{j=1}^{\mathcal{N}_t} \alpha_j^2 \left(\frac{\lambda_j(\mathbf{k}) - \lambda_{N,i}(\mathbf{k})}{\lambda_j^+(\mathbf{k})}\right)^2 \lambda_j^+(\mathbf{k}) \\
&\geq \min_{1\leq j\leq \mathcal{N}_t} \left(\frac{\lambda_j(\mathbf{k}) - \lambda_{N,i}(\mathbf{k})}{\lambda_j^+(\mathbf{k})}\right)^2 \frac{\sum_{j=1}^{\mathcal{N}_t} \alpha_j^2 \lambda_j^+(\mathbf{k})}{\sum_{j'=1}^{\mathcal{N}_t} \alpha_{j'}^2 \lambda_{j'}^+(\mathbf{k})} \\
&= \min_{1\leq j\leq \mathcal{N}_t} \left(\frac{\lambda_j(\mathbf{k}) - \lambda_{N,i}(\mathbf{k})}{\lambda_j^+(\mathbf{k})}\right)^2.
\end{aligned} \tag{5.109}$$

Therefore, based on Hypothesis 5.1,

$$\begin{aligned}\left|\frac{\lambda_i(\mathbf{k})-\lambda_{N,i}(\mathbf{k})}{\lambda_i(\mathbf{k})+\gamma}\right| &\leq \frac{|||R_i(\,\cdot\,;\mathbf{k})|||}{(\lambda_{N,i}(\mathbf{k})+\gamma)^{1/2}} \\ &\leq \frac{\|R_i(\,\cdot\,;\mathbf{k})\|}{(\lambda_{N,i}(\mathbf{k})+\gamma)^{1/2}}, \end{aligned} \tag{5.110}$$

from (5.106). This proves (5.101).

To prove (5.102), we first note that

$$u_{N,i}(\mathbf{k})-u_i(\mathbf{k})=\sum_{j\neq i}\alpha_j u_j(\mathbf{k})+(\alpha_i-1)u_i(\mathbf{k}), \tag{5.111}$$

which leads to

$$\begin{aligned}|||u_{N,i}(\mathbf{k})-u_i(\mathbf{k})|||^2 &= a^+\left(\sum_{j\neq i}\alpha_j u_j(\mathbf{k})+(\alpha_i-1)u_i(\mathbf{k}),\sum_{j\neq i}\alpha_j u_j(\mathbf{k})+(\alpha_i-1)u_i(\mathbf{k});V_{\text{eff}};\mathbf{k}\right) \\ &= a^+\left(\sum_{j\neq i}\alpha_j u_j(\mathbf{k}),\sum_{j\neq i}\alpha_j u_j(\mathbf{k});V_{\text{eff}};\mathbf{k}\right) \\ &\quad + a^+((\alpha_i-1)u_i(\mathbf{k}),(\alpha_i-1)u_i(\mathbf{k});V_{\text{eff}};\mathbf{k}) \\ &\quad + a^+\left(\sum_{j\neq i}\alpha_j u_j(\mathbf{k}),(\alpha_i-1)u_i(\mathbf{k});V_{\text{eff}};\mathbf{k}\right) \\ &\quad + a^+\left((\alpha_i-1)u_i(\mathbf{k}),\sum_{j\neq i}\alpha_j u_j(\mathbf{k});V_{\text{eff}};\mathbf{k}\right) \\ &= \sum_{j\neq i}\alpha_j^2\lambda_j^+(\mathbf{k})+(\alpha_i-1)^2\lambda_i^+(\mathbf{k}). \end{aligned} \tag{5.112}$$

From (5.108),

$$\begin{aligned}|||R_i(\,\cdot\,;\mathbf{k})|||^2 &\geq \sum_{j\neq i}\alpha_j^2\left(\frac{\lambda_j(\mathbf{k})-\lambda_{N,i}(\mathbf{k})}{\lambda_j^+(\mathbf{k})}\right)^2\lambda_j^+(\mathbf{k}) \\ &\geq \min_{j\neq i}\left(\frac{\lambda_j(\mathbf{k})-\lambda_{N,i}(\mathbf{k})}{\lambda_j^+(\mathbf{k})}\right)^2\sum_{j\neq i}\alpha_j^2\lambda_j^+(\mathbf{k}), \end{aligned} \tag{5.113}$$

and from (5.92),

$$\begin{aligned} |||u_{N,i}(\mathbf{k}) - u_i(\mathbf{k})|||^2 &\geq m(u_{N,i}(\mathbf{k}) - u_i(\mathbf{k}), u_{N,i}(\mathbf{k}) - u_i(\mathbf{k})) \\ &= 2(1-\alpha_i). \end{aligned} \tag{5.114}$$

Let $\tilde{d}_i = \min_{j\neq i} \left| \frac{\lambda_j(\mathbf{k}) - \lambda_{N,i}(\mathbf{k})}{\lambda_j^+(\mathbf{k})} \right|$. Then

$$|||u_{N,i}(\mathbf{k}) - u_i(\mathbf{k})|||^2 - \frac{\lambda_i^+(\mathbf{k})}{4} |||u_{N,i}(\mathbf{k}) - u_i(\mathbf{k})|||^4 \leq \frac{|||R_i(\,\cdot\,;\mathbf{k})|||^2}{d_i^2}; \tag{5.115}$$

by solving for $|||u_{N,i}(\mathbf{k}) - u_i(\mathbf{k})|||^2$ and expanding the square root term, we obtain

$$\begin{aligned} |||u_{N,i}(\mathbf{k}) - u_i(\mathbf{k})|||^2 &\leq \frac{|||R_i(\,\cdot\,;\mathbf{k})|||^2}{d_i^2} \\ &\leq \frac{\|R_i(\,\cdot\,;\mathbf{k})\|^2}{d_i^2}, \end{aligned} \tag{5.116}$$

based on (5.106), after ignoring the higher-order term involving $|||\cdot|||^4$. Finally, in the asymptotic limit of (5.100), we have

$$\tilde{d}_i \approx \min_{j\neq i} \left| \frac{\lambda_j(\mathbf{k}) - \lambda_{N,i}(\mathbf{k})}{\lambda_{N,j}^+(\mathbf{k})} \right| \equiv d_i. \tag{5.117}$$

This proves (5.102).

To prove (5.103), we note that

$$\begin{aligned} \lambda_{N,i}(\mathbf{k}) - \lambda_i(\mathbf{k}) &= a^+(u_{N,i}(\mathbf{k}), u_{N,i}(\mathbf{k}); V_{\text{eff}}; \mathbf{k}) - a^+(u_i(\mathbf{k}), u_i(\mathbf{k}); V_{\text{eff}}; \mathbf{k}) \\ &= \sum_{j=1}^{\mathcal{N}_t} \alpha_j^2 \lambda_j^+(\mathbf{k}) - \lambda_i^+(\mathbf{k}) \\ &= \sum_{j\neq i} \alpha_j^2 \lambda_j^+(\mathbf{k}) - (1-\alpha_i^2)\lambda_i^+(\mathbf{k}). \end{aligned} \tag{5.118}$$

Substituting (5.112) and (5.116) into (5.118), we get

$$\begin{aligned} \lambda_{N,i}(\mathbf{k}) - \lambda_i(\mathbf{k}) &= |||u_{N,i}(\mathbf{k}) - u_i(\mathbf{k})|||^2 - (\alpha_i - 1)^2 \lambda_i^+(\mathbf{k}) - (1-\alpha_i^2)\lambda_i^+(\mathbf{k}) \\ &\leq |||u_{N,i}(\mathbf{k}) - u_i(\mathbf{k})|||^2 \\ &\leq \frac{\|R_i(\,\cdot\,;\mathbf{k})\|^2}{\tilde{d}_i^2}; \end{aligned} \tag{5.119}$$

since $1 - \alpha_i^2 = \sum_{j \neq i} \alpha_j^2 \geq 0$ and $\lambda_i^+ > 0$. In the asymptotic limit of (5.100) where $\tilde{d}_i \approx d_i$, this proves (5.103). □

5.5.2 Numerical Results

The offline-online decomposition of the error estimator calculation is similar to Section 2.4 and thus will not be repeated here. In addition, since we are dealing with eigenvalues of multiplicity greater than one, we will only consider (5.101). We define our error estimator $\Delta^{\lambda}_{N,n_b}(\mathbf{k})$ as

$$\Delta^{\lambda}_{N,n_b}(\mathbf{k}) = \max_{1 \leq i \leq n_b} \frac{\|R_i(\,\cdot\,;\mathbf{k})\|}{(\lambda_{N,i}(\mathbf{k}))^{1/2}}, \tag{5.120}$$

and the effectivity measure:

$$\eta^{\lambda}_{N,n_b}(\mathbf{k}) = \frac{\Delta^{\lambda}_{N,n_b}(\mathbf{k})}{\epsilon^{\lambda}_{N,n_b}(\mathbf{k})}, \tag{5.121}$$

We present in Figure 5-17 $\bar{\eta}^{\lambda}_{N,n_b}$, the mean of $\eta^{\lambda}_{N,n_b}(\mathbf{k})$ for $\mathbf{k} \in \Xi_{\mathbf{k}}$, for our results obtained for the augmented reduced basis approximation. We obtain error estimator with effectivity closer to 1 at smaller N. However, this effectivity diverges as N increases. In Section 2.4.3, we have explained why (5.121) will diverge as N increases. As a result, the use of (5.120) as an error measure in the "greedy" sampling procedure may lead to unnecessarily large N. Thus, we would like to emphasize that for the current problem, the error estimation procedure is only used to determine a good set of sample points given $N_{\max}$; it is not used to determine the size of $N_{\max}$. Certainly if we are able to extend the bounds in (5.102) and (5.103) to eigensolutions that degenerate, we will have better error estimators.

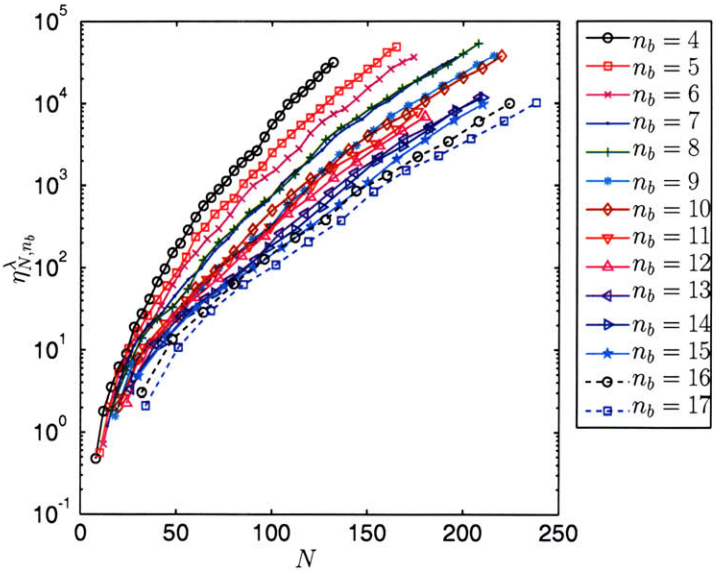

Figure 5-17: Variation of the average effectivity of Δ^{λ}_{N,n_b}, $\bar{\eta}^{\lambda}_{N,n_b}(\mathbf{k})$, with N for $4 \leq n_b \leq 17$ and $\hat{\mathbf{u}} \in (W_N^A)^{n_b}$.

Chapter 6

One Dimensional Kohn Sham Equations

6.1 Introduction

In this chapter, we develop the reduced basis method for one dimensional Kohn Sham equations. Our goal is twofold. First, it offers the opportunity to assimilate methodologies developed in previous chapters. We generalize the augmented reduced basis approximation and the vectorial reduced basis approximation, first introduced in Chapter 2, to nonlinear eigenvalue problems by following closely the procedure outlined in Chapter 3 and 4. Second, it functions as a feasibility study for the three dimensional Kohn Sham equations, the workhorse of density functional theory. Nevertheless, one dimensional models are important in their own right; they are rudimentary in the understanding of crystalline solids and has been studied in [102], and more recently [13]. We also introduce two new ingredients — the geometric parameterization of the Kohn Sham equations and the handling of coupled equations within the reduced basis framework. Results in this chapter were first reported in [22].

6.2 Abstract Formulation

6.2.1 Problem Statement

We consider a one dimensional periodic system with lattice parameter $\mu \in \mathcal{D} \subset \mathbb{R}_+$, and hence unit cell $\tilde{\Omega}(\mu) \equiv]-\frac{\mu}{2}, \frac{\mu}{2}]$. In addition, a single nucleus of charge Z lies at the center of the cell and the number of electrons per nucleus is n_e, with $n_e = Z$ for charge neutrality.

Our output of interest is the ground state energy of the system, $\tilde{\mathcal{E}}$, which we shall determine based on the spinless Density Functional Theory [18, 32, 76, 90]. Our input parameter is the lattice length μ. For simplicity, we shall not include Z in our parameter space; as such, each new Z constitutes a new problem in our reduced-basis approximation. This input-output relation then provides an abstraction for studying how ground state energy changes with lattice parameter. It also provides a convenient framework for the determination of forces exerted on the nuclei when the structure is deformed, or the characterization of the nonlinear behavior of elasticity constant.

6.2.2 Energy Statement

Based on the Density Functional Theory, the equilibrium ground state energy is obtained by solving a minimization problem for $\tilde{\mathbf{u}}([Z,\mu^*]) \equiv (\tilde{u}_1([Z,\mu^*]), \ldots, \tilde{u}_{n_e}([Z,\mu^*]))$, where [11, 25, 64, 66]

$$\begin{aligned} \tilde{\mathbf{u}}([Z,\mu]) &= \arg\inf_{\tilde{\mathbf{w}}} \Big\{ \tilde{E}(\tilde{\mathbf{w}} \equiv (\tilde{w}_1, \cdots, \tilde{w}_{n_e}); [Z,\mu]), \tilde{w}_i \in \tilde{Y}, \\ &\qquad \int_{\tilde{\Omega}(\mu)} \tilde{w}_i \tilde{w}_j = \delta_{ij}, 1 \le i \le j \le n_e \Big\}, \end{aligned} \tag{6.1}$$

$$\mu^*(Z) = \arg\inf_{\mu} \Big\{ \tilde{\mathcal{E}}(\tilde{\mathbf{u}}([Z,\mu]); [Z,\mu]); \mu > 0 \Big\}; \tag{6.2}$$

here $\tilde{Y} \equiv H^1_{\text{per}}(\tilde{\Omega}(\mu))$ is the space of μ-periodic functions in $H^1(\mathbb{R})$; $\delta_{ij} = \{1 \text{ if } i = j,\ 0 \text{ otherwise}\}$; and $\tilde{u}_i$ is the Kohn-Sham orbital associated with the ith electron. We denote $\tilde{x}$ as a point in $\tilde{\Omega}(\mu)$.

The electronic energy $\tilde{E}(\tilde{\hat{\mathbf{w}}}; [Z, \mu])$ is defined as

$$\begin{aligned}\tilde{E}(\tilde{\hat{\mathbf{w}}}; [Z, \mu]) \;=\; & C_w \sum_{i=1}^{n_e} \int_{\tilde{\Omega}(\mu)} (\nabla \tilde{w}_i)^2 - Z \sum_{i=1}^{n_e} \int_{\tilde{\Omega}(\mu)} \tilde{G}\, \tilde{w}_i{}^2 \\ & + \frac{1}{2} C_c \int_{\tilde{\Omega}(\mu)} \int_{\tilde{\Omega}(\mu)} \left(\sum_{i=1}^{n_e} \tilde{w}_i^2(\tilde{x}) \right) \tilde{G}(\tilde{x} - \tilde{y}) \left(\sum_{j=1}^{n_e} \tilde{w}_j^2(\tilde{y}) \right) d\tilde{x}\, d\tilde{y} \\ & - C_x \int_{\tilde{\Omega}(\mu)} \left(\sum_{j=1}^{n_e} \tilde{w}_i^2 \right)^{4/3}, \end{aligned} \tag{6.3}$$

where we have used the X-α approximation to approximate the exchange term and neglected the correlation term; and C_w, C_c, and C_x are model constants (for which we have used $C_w = 0.5$, $C_c = 1$ and $C_x = 0.7386$). The exchange term $-C_x \int_{\tilde{\Omega}(\mu)} (\sum_{j=1}^{n_e} \tilde{w}_i^2)^{4/3}$ in (6.3) has the form that is appropriate for the three dimension case — it does not have the correct homogeneity in this one dimensional problem. For one dimensional problem, we can derive formally based on [15, 62, 108] from the free electron ansatz that the correct expression for the exchange term in one dimension is of the form $-C_{x,\mathrm{1d}} \int_{\tilde{\Omega}(\mu)} (\sum_{j=1}^{n_e} \tilde{w}_i^2)^2$, where $C_{x,\mathrm{1d}} \neq C_x$. We have purposely use the form given in (6.3) to demonstrate the reduced basis approximation of problems with nonpolynomial nonlinearities and thus allow a direct extension to the three dimensional Kohn Sham equations in Chapter 7.

The periodic Green's function $\tilde{G}(\cdot\,; \mu)\colon\ \tilde{\Omega}(\mu) \to \mathbb{R}$ satisfies

$$-\Delta \tilde{G} = \left\{ \delta(\tilde{x}) - \frac{1}{|\tilde{\Omega}(\mu)|} \right\}, \qquad \int_{\tilde{\Omega}(\mu)} \tilde{G} = 0, \tag{6.4}$$

where Δ is the Laplacian operator, $\delta(\tilde{x})$ is the Dirac delta distribution, and $|\tilde{\Omega}(\mu)| = \mu$ is the length of $\tilde{\Omega}(\mu)$. The function $\tilde{G}$ is simply the one-dimensional periodic Coulomb potential. The term $\sum_{j=1}^{n_e} \tilde{u}_i^2(\mu)$ is the electron density of the system. It is usually denoted by $\tilde{\rho}(\mu)$ to underscore the dependence of functionals on electron density in Density Functional Theory. Nevertheless, due to the kinetic energy term in (6.3), solving (6.1) still requires the determination of the wavefunctions $\hat{\mathbf{u}}$.

The total energy $\tilde{\mathcal{E}}(\tilde{\hat{\mathbf{w}}}; [Z, \mu])$ — our output of interest — is then given by

$$\tilde{\mathcal{E}}(\tilde{\hat{\mathbf{w}}}; [Z, \mu]) = \tilde{E}(\tilde{\hat{\mathbf{w}}}; [Z, \mu]) + \frac{Z^2}{2} \eta(\mu), \tag{6.5}$$

where $\eta(\mu)$ is the nuclear - nuclear correction term given by

$$\eta(\mu) = \lim_{\tilde{x}\to 0}\left\{\tilde{G}(\tilde{x};\mu) - \frac{|\tilde{x}|}{2}\right\} \tag{6.6}$$

By Fourier considerations, $\tilde{G}(\,\cdot\,;\mu)$ can be expressed as

$$\tilde{G}(\tilde{x};\mu) = \frac{\mu}{2\pi^2}\sum_{k=1}^{\infty}\frac{\cos(2\pi k\tilde{x}/\mu)}{k^2}. \tag{6.7}$$

Since $\tilde{G}(0;\mu) = \frac{\mu}{2\pi^2}\sum_{k=1}^{\infty}\frac{1}{k^2}$ and $\sum_{k=1}^{\infty}\frac{1}{k^2} = \frac{\pi^2}{6}$, we obtain

$$\begin{aligned}\eta(\mu) &= \tilde{G}(0;\mu)\\ &= \frac{\mu}{12}.\end{aligned} \tag{6.8}$$

6.2.3 Euler-Lagrange Equations

We now derive the equivalent Euler-Lagrange equations for the constrained minimization problem (6.2). We first introduce a set of Lagrange multipliers $\hat{\boldsymbol{\lambda}} \equiv (\lambda_{ij}, 1 \le i \le j \le n_e)$ for the constraints $\int_{\tilde{\Omega}(\mu)} \tilde{w}_i\tilde{w}_j = \delta_{ij}$, $1 \le i \le j \le n_e$. Then, the Lagrange equation can be defined as

$$\mathcal{L}(\tilde{\hat{\mathbf{w}}}, \hat{\boldsymbol{\lambda}}; [Z,\mu]) = \tilde{E}(\tilde{\hat{\mathbf{w}}}; [Z,\mu]) - \sum_{i=1}^{n_e}\sum_{j=i}^{n_e}\lambda_{ij}\left\{\int_{\tilde{\Omega}(\mu)}\tilde{w}_i\tilde{w}_j = \delta_{ij}\right\}. \tag{6.9}$$

Now, let $\tilde{\hat{\mathbf{w}}} = \tilde{\hat{\mathbf{u}}} + \tilde{\hat{\mathbf{v}}}$ where $\tilde{\hat{\mathbf{v}}} \equiv (\tilde{v}_1, \cdots, \tilde{v}_{n_e})$. Then

$$\begin{aligned}\mathcal{L}(\tilde{\hat{\mathbf{u}}} + \tilde{\hat{\mathbf{v}}}, \hat{\boldsymbol{\lambda}}; [Z,\mu]) &= \mathcal{L}(\tilde{\hat{\mathbf{u}}}, \hat{\boldsymbol{\lambda}}; [Z,\mu])\\ &\quad + 2C_w\sum_{i=1}^{n_e}\int_{\tilde{\Omega}(\mu)}\nabla\tilde{u}_i\nabla\tilde{v}_i - 2Z\sum_{i=1}^{n_e}\int_{\tilde{\Omega}(\mu)}\tilde{G}\tilde{u}_i\tilde{v}_i\\ &\quad + 2C_c\sum_{i=1}^{n_e}\int_{\tilde{\Omega}(\mu)}\int_{\tilde{\Omega}(\mu)}\tilde{u}_i(\tilde{x})\,\tilde{G}(\tilde{x}-\tilde{y})\left(\sum_{j=1}^{n_e}\tilde{u}_j^2(\tilde{y})\right)\tilde{v}_i(\tilde{x})\,d\tilde{x}\,d\tilde{y}\\ &\quad - C_x\frac{8}{3}\sum_{i=1}^{n_e}\int_{\tilde{\Omega}(\mu)}\tilde{u}_i\left(\sum_{j=1}^{n_e}\tilde{u}_j^2(\tilde{y})\right)^{1/3}\tilde{v}_i\\ &\quad - \sum_{i=1}^{n_e}\sum_{j=i}^{n_e}\lambda_{ij}\left\{\int_{\tilde{\Omega}(\mu)}\tilde{u}_i\tilde{v}_j + \tilde{u}_j\tilde{v}_i\right\} + \mathcal{O}(\tilde{v}_1^2,\ldots,\tilde{v}_{n_e}^2).\end{aligned} \tag{6.10}$$

Since the first variation must vanish when $\tilde{\mathbf{u}}$ is the unique minimizer,

$$
\begin{aligned}
& C_w \sum_{i=1}^{n_e} \int_{\tilde{\Omega}(\mu)} \nabla \tilde{u}_i \nabla \tilde{v}_i - Z \sum_{i=1}^{n_e} \int_{\tilde{\Omega}(\mu)} \tilde{G} \tilde{u}_i \tilde{v}_i \\
& + C_c \sum_{i=1}^{n_e} \int_{\tilde{\Omega}(\mu)} \int_{\tilde{\Omega}(\mu)} \tilde{u}_i(\tilde{x})\, \tilde{G}(\tilde{x}-\tilde{y}) \left(\sum_{j=1}^{n_e} \tilde{u}_j^2(\tilde{y}) \right) \tilde{v}_i(\tilde{x})\, d\tilde{x}\, d\tilde{y} \\
& - C_x \frac{4}{3} \sum_{i=1}^{n_e} \int_{\tilde{\Omega}(\mu)} \tilde{u}_i \left(\sum_{j=1}^{n_e} \tilde{u}_j^2(\tilde{y}) \right)^{1/3} \tilde{v}_i - \frac{1}{2} \sum_{i=1}^{n_e} \sum_{j=i}^{n_e} \lambda_{ij} \left\{ \int_{\tilde{\Omega}(\mu)} \tilde{u}_i \tilde{v}_j + \tilde{u}_j \tilde{v}_i \right\} \;=\; 0 \qquad (6.11)
\end{aligned}
$$

The above holds separately for $\forall \tilde{v}_i \in \tilde{Y}$; therefore

$$
\begin{aligned}
& -C_w \triangle \tilde{u}_i - Z\, \tilde{G}\, \tilde{u}_i + C_c\, \tilde{u}_i \int_{\tilde{\Omega}(\mu)} \left(\sum_{j=1}^{n_e} \tilde{u}_j^2(\tilde{y}) \right) \tilde{G}(\tilde{x}-\tilde{y})\, d\tilde{y} \\
& \qquad - C_x \frac{4}{3} \left(\sum_{j=1}^{n_e} \tilde{u}_j^2 \right)^{1/3} \tilde{u}_i - \frac{1}{2} \lambda_{ii} \tilde{u}_i - \frac{1}{2} \sum_{j=1}^{n_e} \lambda_{ij} \tilde{u}_j \;=\; 0, \quad 1 \le i \le n_e. \qquad (6.12)
\end{aligned}
$$

Now, let $-\tilde{\phi} = C_c \int_{\tilde{\Omega}(\mu)} \left(\sum_{j=1}^{n_e} \tilde{u}_j^2(\tilde{y}) \right) \tilde{G}(\tilde{x}-\tilde{y})\, d\tilde{y}$. Then

$$
\begin{aligned}
-\triangle \tilde{\phi} &= C_c \int_{\tilde{\Omega}(\mu)} \left(\sum_{j=1}^{n_e} \tilde{u}_j^2(\tilde{y}) \right) \triangle \tilde{G}(\tilde{x}-\tilde{y})\, d\tilde{y} \\
&= -C_c \int_{\tilde{\Omega}(\mu)} \left(\sum_{j=1}^{n_e} \tilde{u}_j^2(\tilde{y}) \right) \left\{ \delta(\tilde{x}-\tilde{y})\, d\tilde{y} - \frac{1}{|\tilde{\Omega}(\mu)|} \right\} \\
&= C_c \left\{ \frac{1}{|\tilde{\Omega}(\mu)|} \sum_{j=1}^{n_e} \int_{\tilde{\Omega}(\mu)} \tilde{u}_j^2(\tilde{y})\, d\tilde{y} - \sum_{j=1}^{n_e} \int_{\tilde{\Omega}(\mu)} \tilde{u}_j^2(\tilde{y})\, \delta(\tilde{x}-\tilde{y})\, d\tilde{y} \right\} \\
&= C_c \left\{ \frac{Z}{|\tilde{\Omega}(\mu)|} - \sum_{j=1}^{n_e} \tilde{u}_j^2 \right\}, \qquad (6.13)
\end{aligned}
$$

since $\int_{\tilde{\Omega}(\mu)} \tilde{u}_j^2(\tilde{y})\, d\tilde{y} = 1$ and $Z = n_e$ for charge neutrality.

The Euler-Lagrange equations are then given by:

$$C_w \Delta \tilde{u}_i - \tilde{G}\tilde{u}_i - \tilde{\phi}\tilde{u}_i + C_x \frac{4}{3}\left(\sum_{j=1}^{n_e} \tilde{u}_j^2\right)^{1/3} \tilde{u}_i = \lambda_{ii}\tilde{u}_i, \qquad 1 \le i \le n_e, \tag{6.14}$$

$$-\Delta\tilde{\phi} - C_c\left[\frac{Z}{|\tilde{\Omega}(\mu)|} - \sum_{j=1}^{n_e} \tilde{u}_j^2\right] = 0, \tag{6.15}$$

$$\int_{\tilde{\Omega}(\mu)} \tilde{u}_i \tilde{u}_j = \delta_{ij}, \qquad 1 \le i \le j \le n_e \tag{6.16}$$

$$\int_{\tilde{\Omega}(\mu)} \tilde{\phi} = 0; \tag{6.17}$$

$\tilde{\phi}$ is simply the Hartree potential [25] with a normalization of $\int_{\tilde{\Omega}(\mu)} \tilde{\phi} = 0$. Equations (6.14)-(6.17) are the Kohn Sham equations, albeit in its one dimensional form. We note that the solutions $\tilde{\mathbf{u}}$, $\hat{\boldsymbol{\lambda}}$ and $\tilde{\phi}$ are coupled. In addition, the nonlinear term $(\sum_{j=1}^{n_e} \tilde{u}_j^2)^{1/3}\tilde{u}_i$ in (6.14) is similar to the nonlinear term examined in Chapter 4. But we now have an nonlinear eigenvalue problem for which we need to determine n_e number of eigenvectors and eigenvalues. To facilitate our reduced basis formulation, we now rewrite (6.14)-(6.17) in the weak form.

6.2.4 Abstract Formulation

The weak form of the Euler-Lagrange equations (6.14)-(6.17) is: find $\tilde{\mathbf{u}}([Z,\mu]) \equiv (\tilde{\tilde{\mathbf{u}}}([Z,\mu]), \tilde{\phi}([Z,\mu]), \tilde{G}([Z,\mu]), \hat{\boldsymbol{\lambda}}([Z,\mu]), \tau([Z,\mu])) \in \tilde{\mathcal{Y}} \equiv (\tilde{Y}^{n_e} \times \tilde{Y} \times \tilde{Y}_0 \times \mathbb{R}^{n_e(n_e+1)/2} \times \mathbb{R})$ such that

$$\begin{aligned} & C_w \int_{\tilde{\Omega}(\mu)} \nabla\tilde{u}_i \nabla\tilde{v} - Z\int_{\tilde{\Omega}(\mu)} \tilde{G}\tilde{u}_i\tilde{v} - \int_{\tilde{\Omega}(\mu)} \tilde{\phi}\tilde{u}_i\tilde{v} \\ & - C_x\frac{4}{3}\int_{\tilde{\Omega}(\mu)} \left(\sum_{j=1}^{n_e} \tilde{u}_j^2\right)^{1/3} \tilde{u}_i\tilde{v} - \frac{1}{2}\lambda_{ii}\int_{\tilde{\Omega}(\mu)} \tilde{u}_i\tilde{v} \\ & \qquad - \frac{1}{2}\int_{\tilde{\Omega}(\mu)} \sum_{j=1}^{n_e} \lambda_{ij}\tilde{u}_j\tilde{v} = 0, \quad \forall \tilde{v} \in \tilde{Y}, \quad 1 \le i \le n_e, \end{aligned} \tag{6.18}$$

$$\int_{\tilde{\Omega}(\mu)} \tilde{u}_i\tilde{u}_j - \delta_{ij} = 0, \quad 1 \le i \le j \le n_e, \tag{6.19}$$

$$\int_{\tilde{\Omega}(\mu)} \nabla\tilde{\phi}\nabla\tilde{v} - C_c\left[\frac{Z}{|\tilde{\Omega}(\mu)|}\int_{\tilde{\Omega}(\mu)}\tilde{v} - \sum_{j=1}^{n_e}\int_{\tilde{\Omega}(\mu)}\tilde{u}_j^2\tilde{v}\right] + \tau\int_{\tilde{\Omega}(\mu)}\tilde{v} = 0, \quad \forall\tilde{v}\in\tilde{Y}, \tag{6.20}$$

$$\int_{\tilde{\Omega}(\mu)}\tilde{\phi} = 0, \tag{6.21}$$

$$\int_{\tilde{\Omega}(\mu)} \nabla\tilde{G}\nabla\tilde{v} - \left[\tilde{v}(0) - \frac{1}{|\tilde{\Omega}(\mu)|}\int_{\tilde{\Omega}(\mu)}\tilde{v}\right] = 0, \quad \forall\tilde{v}\in\tilde{Y}_0, \tag{6.22}$$

where $\hat{\boldsymbol{\lambda}}([Z,\mu]) \equiv (\lambda_{ij}([Z,\mu]), 1\le i\le j\le n_e)$ and τ is the Lagrange multipliers associated with the constraint $\int_{\tilde{\Omega}(\mu)}\tilde{\phi} = 0$; and $\tilde{Y}_0 = \{\tilde{v}\in\tilde{Y} | \int_{\tilde{\Omega}(\mu)}\tilde{v} = 0\}$.

We note that τ is a computational convenience; in fact $\tau = 0$. Let $\tilde{v} = 1$. From (6.20),

$$\int_{\tilde{\Omega}(\mu)} \nabla\tilde{\phi} - C_c\left[Z - \sum_{j=1}^{n_e}\int_{\tilde{\Omega}(\mu)}\tilde{u}_j^2\right] + \tau|\Omega(\mu)| = 0. \tag{6.23}$$

Since $\int_{\tilde{\Omega}(\mu)}\tilde{u}_j^2 = 1$, $1\le j\le n_e$, $Z = n_e$, and $\int_{\tilde{\Omega}(\mu)}\nabla\tilde{\phi} = 0$ due to the periodic boundary condition, (6.23) gives $\tau = 0$.

To facilitate the variational description of the problem, we define the following functional forms

$$\tilde{a}_0(\tilde{w},\tilde{v};\mu) \equiv \int_{\tilde{\Omega}(\mu)}\nabla\tilde{w}\,\nabla\tilde{v}, \tag{6.24}$$

$$\tilde{a}_1(\tilde{w},\tilde{v};\mu) \equiv \int_{\tilde{\Omega}(\mu)}\tilde{w}\,\tilde{v}, \tag{6.25}$$

$$\tilde{a}_2(\tilde{w},\tilde{s},\tilde{v};\mu) \equiv \int_{\tilde{\Omega}(\mu)}\tilde{w}\,\tilde{s}\,\tilde{v}, \tag{6.26}$$

$$\tilde{a}^{\rm nl}(\tilde{w},\tilde{t},\tilde{v};\mu) \equiv \int_{\tilde{\Omega}(\mu)}\tilde{w}\,\tilde{t}^{1/3}\tilde{v}, \tag{6.27}$$

$$\tilde{l}(\tilde{w};\mu) \equiv \int_{\tilde{\Omega}(\mu)}\tilde{w}, \tag{6.28}$$

for any $\mu\in\mathcal{D}$, $\tilde{w}\in\tilde{Y}$, $\tilde{v}\in\tilde{Y}$, $\tilde{s}\in\tilde{Y}$, and non-negative $\tilde{t}\in\tilde{Y}$.

Then, $\tilde{\mathbf{u}}([Z,\mu])\in\tilde{\mathcal{Y}}$ satisfies

$$\tilde{\mathcal{A}}(\tilde{\mathbf{u}}([Z,\mu]),\tilde{\mathbf{v}};[Z,\mu]) = 0, \quad \forall\tilde{\mathbf{v}}\in\tilde{\mathcal{Y}}, \tag{6.29}$$

where

$$
\begin{aligned}
\tilde{\mathcal{A}}(\tilde{\mathbf{w}} \equiv (\tilde{\tilde{\mathbf{w}}}, \tilde{s}, \tilde{t}, \hat{\boldsymbol{\sigma}}, \kappa), \tilde{\mathbf{v}} \equiv (\tilde{\hat{\mathbf{v}}}, \tilde{\varsigma}, \tilde{\varrho}, \hat{\boldsymbol{\varphi}}, \varpi); [Z, \mu]) \equiv \\
&\sum_{i=1}^{n_e} \left[C_w \; \tilde{a}_0(\tilde{w}_i, \tilde{v}_i; \mu) - \; Z \; \tilde{a}_2(\tilde{w}_i, \tilde{t}, \tilde{v}_i; \mu) - \tilde{a}_2(\tilde{w}_i, \tilde{s}, \tilde{v}_i; \mu) - \frac{4}{3} C_x \tilde{a}^{\mathrm{nl}}(\tilde{w}_i, \sum_{j=1}^{n_e} \tilde{w}_j^2, \tilde{v}_i) \right. \\
&\left. - \; \frac{1}{2} \sigma_{ii} a_1(\tilde{u}_i, \tilde{v}_i; \mu) - \frac{1}{2} \sum_{j=1}^{n_e} \sigma_{ij} \; \tilde{a}_1(\tilde{w}_j, \tilde{v}_i; \mu) \right] \\
&+ \sum_{i=1}^{n_e} \sum_{j=i}^{n_e} \varphi_{ij} \left\{ \tilde{a}_1(\tilde{w}_i, \tilde{w}_j; \mu) - \delta_{ij} \right\} \\
&+ \left[\tilde{a}_0(\tilde{s}, \tilde{\varsigma}; \mu) + \; C_c \sum_{j=1}^{n_e} \tilde{a}_2(\tilde{w}_j, \tilde{w}_j, \tilde{\varsigma}; \mu) - \; \frac{Z C_c}{\mu} \tilde{l}(\tilde{\varsigma}; \mu) + \kappa \; \tilde{l}(\tilde{\varsigma}; \mu) \right] \\
&+ \varpi \; \tilde{l}(\tilde{s}; \mu) \\
&+ \left[\tilde{a}_0(\tilde{t}, \tilde{\varrho}; \mu) + \frac{1}{\mu} \tilde{l}(\tilde{\varrho}; \mu) - \tilde{\varrho}(0) \right]. \qquad (6.30)
\end{aligned}
$$

6.2.5 Parameterized Abstract Formulation

To obtain the parameterized weak form, we first define an affine geometric mapping, $\mathcal{G}(\mu)$, from $\tilde{\Omega}(\mu)$ to $\Omega \equiv]-\frac{1}{2}, \frac{1}{2}]$. This can be expressed as

$$x = \mathcal{G}(\tilde{x}; \mu) \equiv \frac{1}{\mu} \tilde{x}. \qquad (6.31)$$

We further define $Y \equiv H^1_{\mathrm{per}}(\Omega)$, the space of 1-periodic functions in $H^1(\mathbb{R})$ with the associated inner product $(w, v)_Y \equiv \int_\Omega \nabla w \cdot \nabla v + \int_\Omega wv$ and norm $\| \cdot \| = (\,\cdot\,, \cdot\,)_Y^{1/2}$; and for any $w \in Y$, $v \in Y$, $s \in Y$, and non-negative $t \in Y$, the parameter-independent functional forms

$$
\begin{aligned}
a_0(w, v) &\equiv \int_\Omega \nabla u \nabla v, \qquad (6.32) \\
a_1(w, v) &\equiv \int_\Omega u \, v, \qquad (6.33) \\
a_2(w, s, v) &\equiv \int_\Omega u \, s \, v, \qquad (6.34)
\end{aligned}
$$

$$a^{\rm nl}(w,t,v) \equiv \int_\Omega w\, t^{1/3} v, \tag{6.35}$$

$$l(w) \equiv \int_\Omega w. \tag{6.36}$$

It is then a simple matter to demonstrate that, for any $\tilde{w} \in \tilde{Y}$ and $w = (\tilde{w} \circ \mathcal{G}^{-1}(\,\cdot\,;\mu)) \in Y$, $\tilde{v} \in \tilde{Y}$ and $v = (\tilde{v} \circ \mathcal{G}^{-1}(\,\cdot\,;\mu)) \in Y$, $\tilde{s} \in \tilde{Y}$ and $s = (\frac{1}{\mu}\, \tilde{s} \circ \mathcal{G}^{-1}(\,\cdot\,;\mu)) \in Y$, and non-negative $\tilde{t} \in \tilde{Y}$ and $t = (\tilde{t} \circ \mathcal{G}^{-1}(\,\cdot\,;\mu)) \in Y$,

$$\tilde{a}_0(\tilde{w},\tilde{v};\mu) = \frac{1}{\mu}\, a_0(w,v), \tag{6.37}$$

$$\tilde{a}_1(\tilde{w},\tilde{v};\mu) = \mu\, a_1(w,v), \tag{6.38}$$

$$\tilde{a}_2(\tilde{w},\tilde{s},\tilde{v};\mu) = \mu^2\, a_2(w,s,v), \tag{6.39}$$

$$\tilde{a}_2(\tilde{w},\tilde{w},\tilde{v};\mu) = \mu\, a_2(w,w,v), \tag{6.40}$$

$$\tilde{a}^{\rm nl}(\tilde{w},\tilde{t},\tilde{v};\mu) = \mu\, a^{\rm nl}(w,t,v), \tag{6.41}$$

$$\tilde{l}(\tilde{w};\mu) = \mu\, l(w). \tag{6.42}$$

We shall exploit (6.37)–(6.42) to transform our problem to the fixed reference domain Ω.

We first note that the weak form for $\tilde{G}$ (corresponding to nonzero $\tilde{\varrho}$ in (6.30) transforms to

$$a_0(G,\varrho) + \{l(\varrho) - \varrho(0)\} = 0, \quad \forall \varrho \in Y_0, \tag{6.43}$$

for

$$G(\,\cdot\,) = \frac{1}{\mu}\, \tilde{G} \circ \mathcal{G}^{-1}(\,\cdot\,;\mu), \tag{6.44}$$

and $Y_0 = \{v \in Y | \int_\Omega v = 0\}$. We observe that G is the Coulomb potential $\tilde{G}$ scaled by μ, and it is a universal function — independent of μ and Z.

We now define $u_i([Z,\mu]) = \tilde{u}_i \circ \mathcal{G}^{-1}(\,\cdot\,;\mu)$ and $\phi([Z,\mu]) = \frac{1}{\mu}\tilde{\phi} \circ \mathcal{G}^{-1}(\,\cdot\,;\mu)$. Then, $\mathbf{u}([Z,\mu]) \equiv (\hat{\mathbf{u}}([Z,\mu]), \phi([Z,\mu]), \hat{\boldsymbol{\lambda}}([Z,\mu]), \tau([Z,\mu])) \in \mathcal{Y} \equiv (Y^{n_e} \times Y \times \mathbb{R}^{n_e(n_e+1)/2} \times \mathbb{R})$ satisfies

$$\mathcal{A}(\mathbf{u}([Z,\mu]), \mathbf{v}; G; [Z,\mu]) = 0, \quad \forall \mathbf{v} \in \mathcal{Y}, \tag{6.45}$$

where $\hat{\mathbf{u}}([Z,\mu]) \equiv (u_i([Z,\mu]), 1 \le i \le n_e)$; $\hat{\boldsymbol{\lambda}}([Z,\mu]) \equiv (\lambda_{ij}([Z,\mu]), 1 \le i \le j \le n_e)$; and $\mathcal{A}$ is defined

as

$$\begin{aligned}
&\mathcal{A}(\mathbf{w} \equiv (\hat{\mathbf{w}}, s, \hat{\boldsymbol{\sigma}}, \kappa), \mathbf{v} \equiv (\hat{\mathbf{v}}, \varsigma, \hat{\boldsymbol{\varphi}}, \varpi); t; [Z, \mu]) \equiv \\
&\quad \sum_{i=1}^{n_e} \Big[\theta_1([Z,\mu])\ a_0(w_i, v_i) + \ \theta_2([Z,\mu])\ a_2(w_i, t, v_i) + \theta_3([Z,\mu])\ a_2(w_i, s, v_i) \\
&\quad + \theta_5([Z,\mu])\ a^{\mathrm{nl}}(w_i, \sum_{j=1}^{n_e} w_j^2, v_i) + \theta_4([Z,\mu])\ \sigma_{ii} a_1(w_i, v_i) + \theta_4([Z,\mu])\ \sum_{j=1}^{n_e} \sigma_{ij}\ a_1(w_j, v_i) \Big] \\
&\quad + \sum_{i=1}^{n_e} \sum_{j=i}^{n_e} \varphi_{ij} \left\{ \beta_1([Z,\mu]) a_1(w_i, w_j) + \beta_2([Z,\mu]) \delta_{ij} \right\} \\
&\quad + \left[\alpha_1([Z,\mu]) a_0(s, \varsigma) + \alpha_2([Z,\mu]) \sum_{j=1}^{n_e} a_2(w_j, w_j, \varsigma) + \alpha_3([Z,\mu]) l(\varsigma) + \kappa \alpha_4([Z,\mu])\ l(\varsigma) \right] \\
&\quad + \varpi\ l(s). \qquad (6.46)
\end{aligned}$$

For prescribed C_w, C_x, and C_c,

$$\theta([Z,\mu]) = \left\{ \frac{C_w}{\mu}, -Z\mu^2, -\mu^2, -\frac{\mu}{2}, -C_x \frac{4}{3}\mu \right\}, \qquad (6.47)$$

$$\alpha([Z,\mu]) = \{1, C_c\mu, -C_c Z, \mu\}, \quad \text{and} \qquad (6.48)$$

$$\beta([Z,\mu]) = \{\mu, -1\}. \qquad (6.49)$$

The total energy $\mathcal{E}(\mathbf{u}([Z,\mu]); G; [Z,\mu])$ is then given by

$$\begin{aligned}
\mathcal{E}(\mathbf{w} \equiv (\hat{\mathbf{w}}, s, \hat{\boldsymbol{\sigma}}, \kappa); G; [Z,\mu]) &= \sum_{i=1}^{n_e} \Big[\theta_1([Z,\mu]) a_0(w_i, w_i) + \theta_2([Z,\mu])\ a_2(w_i, G, w_i) \\
&\quad + \frac{1}{2}\theta_3([Z,\mu])\ a_2(w_i, s, w_i) + \frac{3}{4}\theta_5([Z,\mu])\ a^{\mathrm{nl}}(w_i, \sum_{j=1}^{n_e} w_j^2, w_i) \Big] \\
&\quad + \frac{Z^2}{2}\eta(\mu). \qquad (6.50)
\end{aligned}$$

6.2.6 "Truth" Finite Element Approximation

As in previous chapters, we now introduce our finite element "truth" approximation to (6.45). We first define our finite element space $Y_h \subset Y$ of dimension $\mathcal{N}$ as

$$Y_h \equiv \{v \in Y \mid v|_{\mathbf{T}_h} \in \mathbb{P}_1(\mathbf{T}_h),\ \forall \mathbf{T}_h \in \mathcal{T}_h\}, \tag{6.51}$$

$$\mathbb{P}_1(\mathbf{T}_h) \equiv \text{span}\{1,\ x\}, \tag{6.52}$$

where $\mathcal{T}_h$ is a (regular) uniform "triangulation" of the domain Ω comprising linear elements $\mathbf{T}_h$ of length h. Our finite element approximation to (6.45) is then given by: find $\mathbf{u}_h([Z,\mu]) \equiv (\hat{\mathbf{u}}_h([Z,\mu]), \phi_h([Z,\mu]), \hat{\boldsymbol{\lambda}}_h([Z,\mu]), \tau_h([Z,\mu])) \in \mathcal{Y}_h \equiv (Y_h^{n_e} \times Y_h \times \mathbb{R}^{n_e(n_e+1)/2} \times \mathbb{R})$ such that

$$\mathcal{A}_h(\mathbf{u}_h([Z,\mu]), \mathbf{v}; G_h; [Z,\mu]) = 0, \quad \forall \mathbf{v} \in \mathcal{Y}_h. \tag{6.53}$$

Here G_h is the finite element approximation to G. The one dimensional Green's function can be easily solved numerically; since the Dirac delta function is a bounded function in one dimension, it can be treated satisfactorily with finite element method. In addition $\mathcal{A}_h$ is an approximation to $\mathcal{A}$ in which the terms $a^{\text{nl}}(w,t,v)$, and $a_2(w,s,v)$ are replaced by quadrature sums: we approximate $a^{\text{nl}}(w,t,v)$ by

$$a^{\text{nl}}(w,t,v) = \sum_{\text{quad}} w(\,\cdot\,)\, t(\,\cdot\,)\, v(\,\cdot\,). \tag{6.54}$$

The term $a_2(\,\cdot\,,\,\cdot\,,\,\cdot\,)$ is treated analogously.

The finite element approximation to the total energy $\mathcal{E}(w; G; [Z,\mu])$, $\mathcal{E}_h(\mathbf{w} \equiv (\hat{\mathbf{u}}, s, \hat{\boldsymbol{\sigma}}, \kappa);\ G_h;\ [Z,\mu])$, is then given by

$$\begin{aligned}
\mathcal{E}_h(\mathbf{w}; G_h; [Z,\mu]) \;=\; & \sum_{i=1}^{n_e} \Big[\theta_1([Z,\mu])\ a_0(w_i, w_i) \\
& + \theta_2([Z,\mu])\ a_2(w_i, G_h, w_i) + \frac{1}{2}\theta_3([Z,\mu])\ a_2(w_i, s, w_i) \\
& + \frac{3}{4}\theta_5([Z,\mu])\ a^{\text{nl}}(w_i, \sum_{j=1}^{n_e} w_j^2, w_i) \Big] + \frac{Z^2}{2}\eta(\mu);
\end{aligned} \tag{6.55}$$

again, quadrature sums are applied to $a_2(\,\cdot\,,\,\cdot\,,\,\cdot\,)$ and $a^{\text{nl}}(\,\cdot\,,\,\cdot\,,\,\cdot\,)$.

The resulting discrete coupled nonlinear equations will now be solved. We shall solve these

equations based on the fixed-point method outlined in Section 4.2.2. The algorithm can be stated as follows: if the density $\rho_h([Z,\mu])$ defined as

$$\rho_h([Z,\mu]) \equiv \sum_{j=1}^{n_e} u_{h,j}^2([Z,\mu]) \tag{6.56}$$

is known, we may compute $\phi_h([Z,\mu])$ explicitly. The weak form for ϕ corresponds to nonzero ς and ϖ of (6.45). Then, $(\hat{\mathbf{u}}_h([Z,\mu]), \hat{\boldsymbol{\lambda}}_h([Z,\mu]))$ satisfy the following weak forms (corresponding to nonzero $\hat{\mathbf{v}}$ and $\hat{\boldsymbol{\varphi}}$ of (6.45)):

$$\begin{aligned} &\theta_1([Z,\mu])\ a_0(u_{h,i},v) + \ \theta_2([Z,\mu])\ a_2(u_{h,i},G_h,v) \\ &+ \theta_3([Z,\mu])\ a_2(u_{h,i},\phi_h,v) + \ \theta_5([Z,\mu])\ a^{\mathrm{nl}}(u_{h,i},\rho_h,v) \\ &\qquad = \ 2\theta_4([Z,\mu])\ \lambda_{h,ii} a_1(u_{h,i},v), \quad \forall v \in Y_h, \quad 1 \le i \le n_e, \end{aligned} \tag{6.57}$$

$$\beta_1([Z,\mu]) a_1(u_{h,i},u_{h,j}) = \ \beta_2([Z,\mu])\delta_{ij}, \quad 1 \le i \le j \le n_e. \tag{6.58}$$

This corresponds to a symmetric eigenvalue problem for which the solutions $\hat{\mathbf{u}}_h([Z,\mu])$ and $\hat{\boldsymbol{\lambda}}_h([Z,\mu])$, are real.

However, we usually do not know ρ_h. We therefore proceed iteratively: we start with an initial guess ρ_h^0; then for $k > 0$, we compute ϕ_h^k from ρ_h^{k-1}, solve (6.58) to obtain $\hat{\mathbf{u}}_h^k$, and compute a new density given by $\rho_h^k = (\rho_h^{k-1} - \sum_{j=1}^{n_e}(u_{h,j}^k)^2)/2$. We repeat this procedure until $\|\rho_h^{k-1} - \sum_{j=1}^{n_e}(u_{h,j}^k)^2\|_2 < \varepsilon_{\mathrm{tol}}$ where $\|\cdot\|_2$ is the vector norm and $\varepsilon_{\mathrm{tol}}$ is an user defined tolerance.

From next section onward, we will drop the subscript h, and assume the finite element solution is our "truth" solution, i.e. Y, $\hat{\mathbf{u}}$, $\hat{\boldsymbol{\lambda}}$, ϕ and G refer to Y_h, $\hat{\mathbf{u}}_h$, $\hat{\boldsymbol{\lambda}}_h$, ϕ_h and G_h. Our truth approximations are obtained for $\mathcal{N} = 400$. As shown in Figure 6-1, the solutions in $\hat{\mathbf{u}}$ exhibit considerable variation with respect to μ. In particular, we note that the solutions are equivalent up to a sign. In addition, the mode shapes of $u_i([Z,\mu])$ can change as μ varies.

6.3 Reduced-Basis Formulation

In chapter 2, we show that for well separated eigenvalues, we can usually recover a sufficiently smooth variation of the eigenvectors with the parameter. Then, the vectorial reduced basis approximation can be very efficient. On the other hand, if the pre-processing steps are unable to

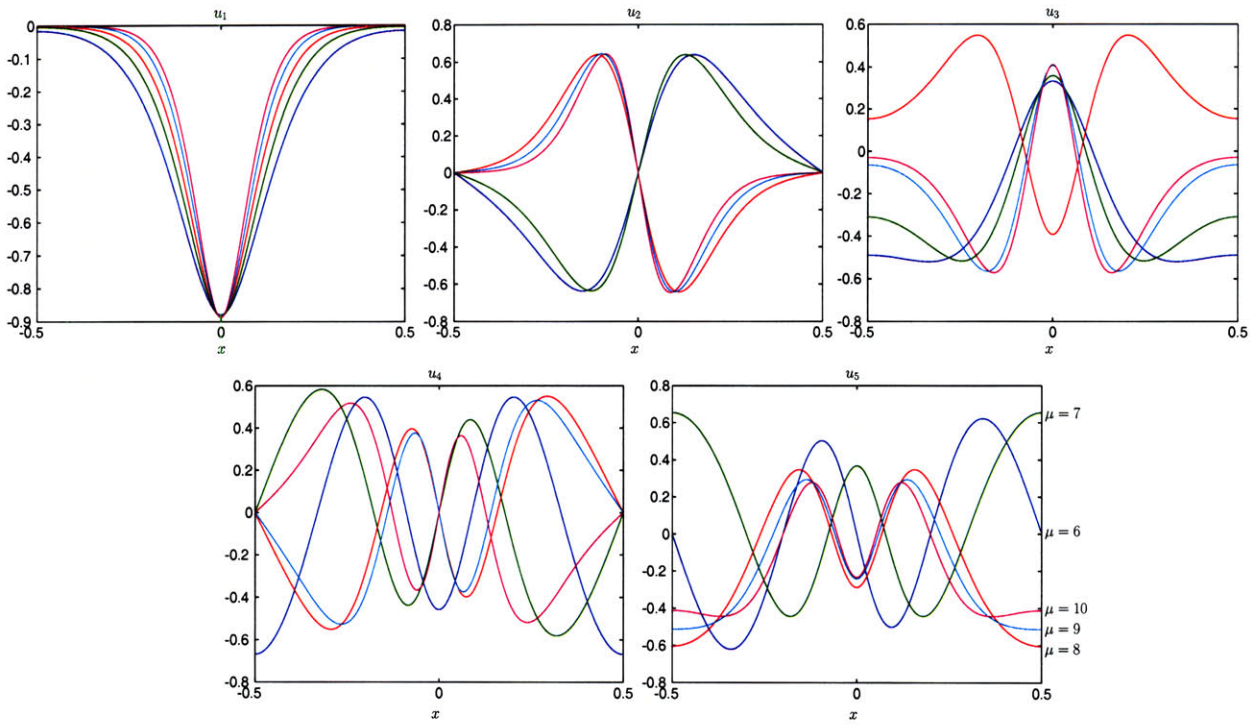

Figure 6-1: Solutions of $\hat{\mathbf{u}}([Z, \mu])$ for $n_e = 5$ at $\mu = 6$, 7, 8, 9, and 10.

approximate this smooth solution manifold adequately, as in Chapter 5, then the augmented reduced basis space may be more appropriate. From Figure 6-1, visual inspection suggests that we can recover the required smooth properties with the pre-processing steps described in Section 5.3.2. However, for comparison, we will examine the performance of both the augmented reduced basis approximation and the vectorial reduced basis approximation for the one dimensional Kohn Sham equations. Most of the required ingredients have already been described in details in previous chapters; thus we shall be brief with materials already covered.

We first note that the Kohn Sham equation is a set of coupled equations with two field variables. The reduced basis approximation of coupled equations is, nevertheless, straightforward. Instead of one reduced basis approximation space, we construct two independent reduced basis approximation spaces for $\hat{\mathbf{u}}$ and ϕ. However, the construction of these two reduced basis approximation spaces are coupled; we explore this issue in Section 6.3.5. Consider, for now, the case where we have constructed for ϕ the nested sample sets $S^{\phi}_{N^{\phi}} = \{\mu^{\phi}_1, \ldots, \mu^{\phi}_{N^{\phi}}\}$, $1 \leq N^{\phi} \leq N^{\phi}_{\max}$. Since the field variable ϕ is a scalar quantity, the construction of the reduced basis space is then standard and is

given by

$$\begin{aligned} W^{\phi}_{N^{\phi}} &= \text{span } \{\phi([Z,\mu^{\phi}_n]), 1 \le n \le N^{\phi}\}, \quad 1 \le N^{\phi} \le N^{\phi}_{\max}, \\ &= \text{span } \{\chi_n, 1 \le n \le N^{\phi}\}, \quad 1 \le N^{\phi} \le N^{\phi}_{\max}. \end{aligned} \tag{6.59}$$

The χ_n, $1 \le n \le N^{\phi}$ are obtained by orthonormalizing $\phi([Z,\mu^{\phi}_n])$, $1 \le n \le N^{\phi}$ relative to the $(\cdot\,;\cdot)_Y$ inner product.

Analogously, we shall assume we have constructed for $\hat{\mathbf{u}}$, the nested sample sets $S^{A,u}_{N^u} = (\mu^u_1, \ldots, \mu^u_{N^u_s})$, $1 \le N^u_s \le N^u_{s,\max}$. However, the construction of the reduced basis approximation spaces of $\hat{\mathbf{u}}$ warrants a longer explanation. We shall precede this with a description of issues related to the pre-processing of the eigenvectors, specialized for this problem.

6.3.1 Pre-processing

Unlike chapter 2, pre-processing the eigenvectors is essential to the construction of both reduced basis approximations. Its role will be further elaborated when the two reduced basis approximations are discussed. Here, we shall outline the procedure by which eigenvectors are sorted and aligned.

For this one-dimensional problem, the eigenvalues are non-degenerate; the sorting and alignment procedure described in Section 5.3.2 can be significantly simplified. In Figure 6-1, we can identify two types of discontinuities with respect to μ: (*i*) sign switching (as demonstrated by u_2, u_3, u_4, and u_5); and (*ii*) mode crossing, where u_4 at $\mu = 6$ should be a smooth transition of u_5 at $\mu = 7$, 8, 9 and 10, and vice versa. There is a third type of discontinuity not exhibited by the solutions in Figure 6-1: mode entering, i.e., there are more than n_e forms of mode shapes after taking our eigenvectors $\hat{\mathbf{u}}(\mu)$ for all $\mu \in \mathcal{D}$ into considerations. These discontinuities do not however represent the actual smoothness of our solution manifold. In fact, they are artifacts of the eigenvalue solvers: the components of $\hat{\mathbf{u}}([Z,\mu])$ are not arranged according to any particular structure identified by the mode shapes of the eigenvectors $u_i([Z,\mu])$, $1 \le i \le n_e$, and they are only equivalent up to a sign. With the following sorting and alignment procedure, we can usually recover the smoothness of the solution manifold.

The simplified sorting and alignment procedure can be described as follows: given a sample set $S_{N^u} \equiv \{\mu^u_n, 1 \le n \le N^u\}$ and the associated pre-sorted set $U_{N^u} \equiv \{\hat{\zeta}^s_n, 1 \le n \le N^u\}$ where $\hat{\zeta}^s_n$, $1 \le n \le N^u$ are the sorted basis functions of $\hat{\mathbf{u}}_n$, $1 \le n \le N^u$, we wish to add $\hat{\mathbf{u}}([Z,\mu^u_{N^u+1}])$

```
given μ^u_1, ... μ^u_{N^u_max};
let U_1 = {ζ̂^s_1 ≡ û([Z, μ^u_1])};
for N^u = 2 : N^u_max
        n* = arg min_{1≤n≤N^u−1} |μ^u_n − μ^u_{N^u}|;
        for i = 1 : n_e
                e^+_j = ||ζ^s_{n*,i} + u_j([Z, μ^u_{N^u}])||_Y,   1 ≤ j ≤ n_e;
                e^−_j = ||ζ^s_{n*,i} − u_j([Z, μ^u_{N^u}])||_Y,   1 ≤ j ≤ n_e;
                j* = arg min_{1≤j≤n_e} {e^+_j, e^−_j};
                if e^−_{j*} > e^+_{j*}
                        ζ^s_{N^u,i} = −u_i([Z, μ^u_{N^u}]);
                else
                        ζ^s_{N^u,i} = u_i([Z, μ^u_{N^u}]);
                end
        end
        U_{N^u} = U_{N^u−1} ∪ ζ̂^s_{N^u};
end.
```

Figure 6-2: The sorting and aligning algorithm for eigenvectors of the one-dimensional Kohn Sham equations.

to U_{N^u} to form U_{N^u+1}. We first select a $\hat{\zeta}^s_n \in U_{N^u}$ such that $\mu^u_n \in S_{N^u}$ is closest to $\mu^u_{N^u+1}$. We compute $e^+_j = ||\zeta^s_{n,1} + u_j(\mu^u_{N^u+1})||_Y$ and $e^-_j = ||\zeta^s_{n,1} - u_j(\mu^u_{N^u+1})||_Y$ for $1 \le j \le n_e$; we then determine $j^* \equiv \arg\min_{1\le j\le n_e}\{e^+_j, e^-_j\}$. If $e^-_{j^*} > e^+_{j^*}$, then $\zeta^s_{N^u+1,1} = -u_{j^*}(\mu^u_{N^u+1})$; otherwise $\zeta^s_{N^u+1,1} = u_{j^*}(\mu^u_{N^u+1})$. This is then repeated for $\zeta^s_{N^u+1,i}$, $i = 2, \ldots, n_e$. First part of the algorithm associates $u_j(\mu)$ to the correct mode shape and the second part of the algorithm remove the sign variation in ζ^s_i. Figure 6-2 summarizes the procedure.

Figure 6-3 shows the same set of $\hat{\mathbf{u}}([Z,\mu]) \equiv (u_1([Z,\mu]), \ldots, u_{n_e}([Z,\mu]))$ of Figure 6-1 after the sorting and aligning algorithm. The sorted and aligned solutions are denoted by $\hat{\zeta}^s_n \equiv (\zeta^s_{n,1}, \ldots, \zeta^s_{n,n_e})$, where n denotes the nth sample points in S_{N^u}. Comparing Figure 6-3 and Figure 6-1, we note sign changes have been effected in $\zeta^s_{n,2}$, $\zeta^s_{n,3}$, $\zeta^s_{n,4}$ and $\zeta^s_{n,5}$, and mode switching between $u_4([Z,\mu_n])$ and $u_5([Z,\mu_n])$) has been effected for $\mu_n = 6$. This leads to a smooth variation of $\hat{\zeta}^s_n$ with respect to μ_n.

The 3rd type of discontinuities — mode entering — has to be handled differently. In the sim-

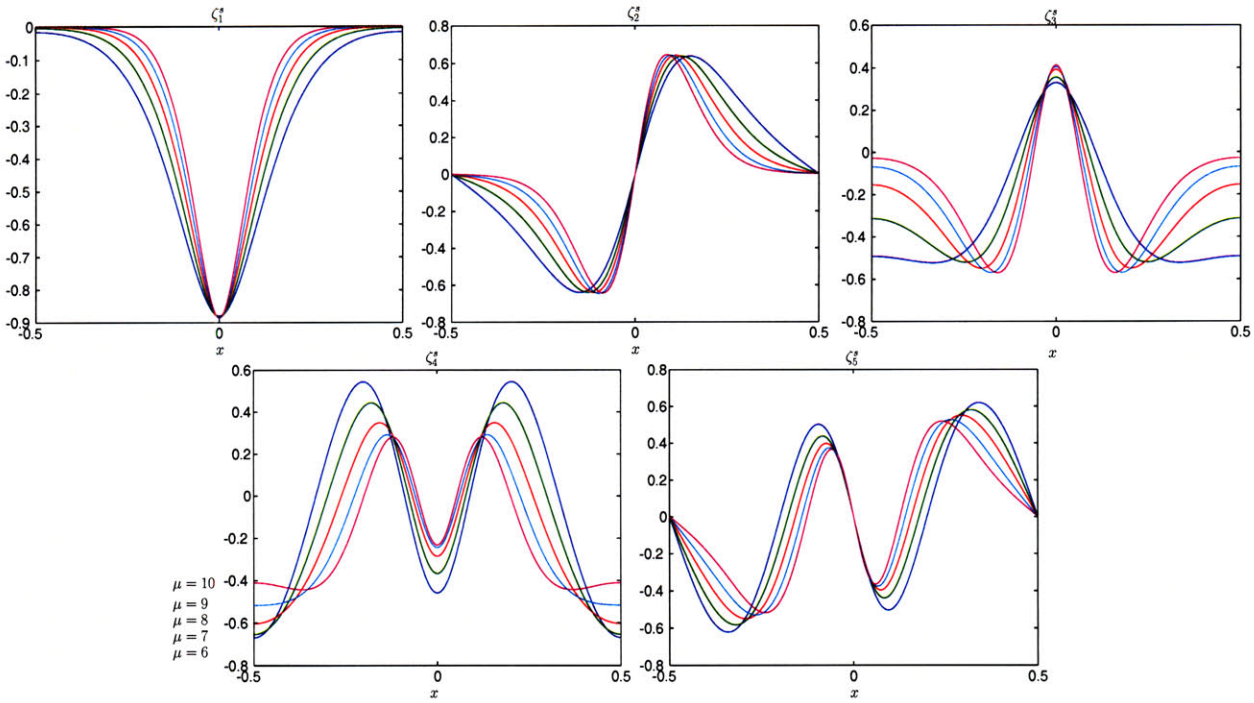

Figure 6-3: Pre-processed $\hat{\mathbf{u}}([Z, \mu])$, given by $\hat{\boldsymbol{\zeta}}^s$, for $n_e = 5$ at $\mu = 6$, 7, 8, 9, and 10.

plest scenario, this occurs when the highest eigenvector considered, $u_{n_e}([Z, \mu])$ changes its mode shape as μ varies, with no corresponding change in $u_{n_e-1}([Z, \mu])$ — it is a mode crossing involving $u_{n_e+1}([Z, \mu])$, not $u_{n_e-1}([Z, \mu])$. As the algorithm in Figure 6-2 cannot remove this type of discontinuities, the solution manifold is then discontinuous at the μ-point at which this mode entering phenomenon occurs. The $u_{n_e}([Z, \mu])$ and thus $\hat{\mathbf{u}}([Z, \mu])$ are then piecewise continuous with respect to μ. This can lead to a sub-optimal approximation — it is equivalent to solving two separate solution manifolds (assuming there is only one case of mode entering within $\mathcal{D}$); sufficient sample points from both manifolds must be selected in order to obtain a good approximation.

With the augmented reduced basis approximation, this discontinuity can be more readily accommodated. By taking the span of all eigenvectors at all sample points, the approximation space is very rich; the approximation relies more on the Galerkin procedure to find the optimal combination of the basis functions, and less on approximating the solution manifold. For the vectorial reduced basis approximation however, the convergence rate may be poor since $u_i([Z, \mu])$, $1 \leq i \leq n_e$ do not have the common smoothness property, as explained in Section 2.3.7.

An alternative approach to handling mode entering is to construct a reduced-basis space for n_e+1 eigenvectors so that the resulting discontinuity can be treated by the algorithm of Figure 6-2 as a

case of mode switching. We then recover the case where we are only approximating a single solution manifold. Then, vectorial reduced basis approximation will be more rapidly convergent since the solution manifold is no longer discontinuous. With the augmented reduced basis approximation, this approach may lead to a larger N since N scales as $N_s n_b$, depending on N_s required in the two approaches.

The discontinuity can also affect the convergence rate of the approximation of nonlinear functions based on the empirical interpolation method; this will be further elaborated in Section 6.3.2. From here onward, we will assume $\hat{\mathbf{u}}([Z,\mu_n])$ is the sorted set of eigenvectors $\hat{\boldsymbol{\zeta}}_n^s$ and the $\hat{\boldsymbol{\lambda}}([Z,\mu_n])$ has been reordered accordingly.

6.3.2 Augmented Reduced Basis Space

The space

We first introduce nested sample sets $S_{N^u}^{A,u} = (\mu_1^u, \ldots, \mu_{N_s^u}^u)$, $1 \le N_s^u \le N_{s,\max}^u$ and define the associated nested reduced-basis spaces as

$$W_{N^u}^{A,u} = \text{span}\ \{u_i([Z,\mu_j]), 1 \le i \le n_e, 1 \le j \le N_s^u\}, \quad 1 \le N_s^u \le N_{s,\max}^u, \tag{6.60}$$

$$= \text{span}\ \{\zeta_n, 1 \le n \le N^u \equiv N_s^u n_e\}, \quad 1 \le N_s^u \le N_{s,\max}^u; \tag{6.61}$$

where $u_1([Z,\mu_j]), \ldots, u_{n_e}([Z,\mu_j])$ are the solutions of (6.45) at $\mu = \mu_j$; and ζ_n are basis functions obtained after $u_i(\mu_j)$, $1 \le i \le n_e$, $1 \le j \le N_s$ are orthonormalized. Then, an approximation of $u_i(\mu)$ in $W_{N^u}^{A,u}$ is represented by $u_{N,i}(\mu) = \sum_{n=1}^{N^u} \alpha_n(\mu)\zeta_n$.

The approximation

Our reduced basis approximation to (6.45) is given by: find $\mathbf{u}_{N,M}([Z,\mu]) \equiv (\hat{\mathbf{u}}_{N,M}([Z,\mu]), \phi_{N,M}([Z,\mu]), \hat{\boldsymbol{\lambda}}_{N,M}([Z,\mu]), \tau_{N,M}([Z,\mu])) \in \mathcal{Y}_N \equiv ((W_{N^u}^{A,u})^{n_e} \times W_N^{\phi} \times \mathbb{R}^{n_e} \times \mathbb{R})$ such that

$$\mathcal{A}_{M,A}(\mathbf{u}_{N,M}([Z,\mu]), \mathbf{v}; G; [Z,\mu]) = 0, \quad \forall \mathbf{v} \in \mathcal{Y}_N, \tag{6.62}$$

where $\mathcal{A}_{M,A}$ is defined as

$$\begin{aligned}
\mathcal{A}_{M,A}(\mathbf{w} \equiv (\hat{\mathbf{w}}, s, \hat{\boldsymbol{\sigma}}, \kappa), \mathbf{v} \equiv (\hat{\mathbf{v}}, \varsigma, \hat{\boldsymbol{\varphi}}, \varpi); t; [Z, \mu]) \equiv \\
\sum_{i=1}^{n_e} \Big[\theta_1([Z,\mu])\ a_0(w_i, v_i) + \ \theta_2([Z,\mu])\ a_2(w_i, t, v_i) + \theta_3([Z,\mu])\ a_2(w_i, s, v_i) \\
+ \theta_5([Z,\mu])\ a^{\mathrm{nl,M}}(w_i, \sum_{j=1}^{n_e} w_j^2, v_i) + \ \theta_4([Z,\mu])\ \sigma_{ii} a_1(w_i, v_i) + \theta_4([Z,\mu])\ \sum_{j=1}^{n_e} \sigma_{ij}\ a_1(w_j, v_i) \Big] \\
+ \sum_{i=1}^{n_e} \sum_{j=i}^{n_e} \varphi_{ij} \left\{ \beta_1([Z,\mu]) a_1(w_i, w_j) + \beta_2([Z,\mu]) \delta_{ij} \right\} \\
+ \left[\alpha_1([Z,\mu]) a_0(s, \varsigma) + \alpha_2([Z,\mu]) \sum_{j=1}^{n_e} a_2(w_j, w_j, \varsigma) + \alpha_3([Z,\mu]) l(\varsigma) + \kappa \alpha_4([Z,\mu])\ l(\varsigma) \right] \\
+ \varpi\ l(s). \qquad (6.63)
\end{aligned}$$

When compared to (6.46), we have approximated $a^{\mathrm{nl}}(w_i, \sum_{j=1}^{n_e} w_j^2, v_i)$ by

$$a^{\mathrm{nl,M}}(w_i, \sum_{j=1}^{n_e} w_j^2, v_i) \equiv \int_\Omega g^w_{M^{g_i},i}\, v, \qquad (6.64)$$

where $g^u_{M^{g_i},i}$ is an empirical interpolation approximation to $g_i(\hat{\mathbf{u}}) \equiv u_i(\sum_{j=1}^{n_e} u_j^2)^{1/3}$. Since there are n_e components to $\hat{\mathbf{u}}$, we need to construct n_e empirical interpolation approximations — we have described the construction procedure in Section 3.2.3. In particular, for each g_i^u (and thus $g^u_{M^{g_i},i}$), we construct the sample set $S^{g_i}_{M^{g_i}}$, the approximation space $W^{g_i}_{M^{g_i}}$ and the set of interpolation points $T^{g_i}_{M^{g_i}}$. Here, M^{g_i} denotes the size of ${}^{g_i}_{M^{g_i}}$, which is equivalent to the dimension of $W^{g_i}_{M^{g_i}}$. We further denote M as $\max_{1\le i\le n_e} M^{g_i}$.

Note that if u_i where not pre-processed, $g_i(\hat{\mathbf{u}})$ would not be varying smoothly with μ — there will be discontinuities. In order to have an efficient approximation based on empirical interpolation method, it is thus necessary to preprocess the eigenvectors according to Section 6.3.1. However, it is important to point out that the term $\sum_{j=1}^{n_e} u_j^2$ will always lead to a discontinuity in g_i when the third type of discontinuities — mode entering — occurs. The remedies suggested for reduced basis approximation of $\hat{\mathbf{u}}$ in Section 6.3.1 do not remove the discontinuity in g_i since u_1, ..., u_{n_e} in the summation term are those associated with the lowest n_e eigenvalues. However, g_i is piecewise continuous in $\mathcal{D}$ and thus can be approximated using the empirical interpolation method, as explained in Section 3.3. Nevertheless, as noted in Section 3.3 as well, the convergence rate of

the approximation will be degraded — we expect M required to achieve a certain tolerance to be larger than that required in cases where this type of discontinuities is absent.

The approximation to the total energy, $\mathcal{E}_{N,M}(\mathbf{u}_{N,M}([Z,\mu]);G;[Z,\mu])$, is finally given by

$$\begin{aligned}\mathcal{E}_{N,M}(\mathbf{w}\equiv(\hat{\mathbf{w}},s,\hat{\sigma},\kappa);G;[Z,\mu]) &= \sum_{i=1}^{n_e}\Big[\theta_1([Z,\mu])a_0(w_i,w_i)+\theta_2([Z,\mu])\ a_2(w_i,G,w_i)\\ &\quad+\frac{1}{2}\theta_3([Z,\mu])\ a_2(w_i,s,w_i)\\ &\quad+\frac{3}{4}\theta_5([Z,\mu])\ a^{\mathrm{nl,M}}(w_i,\sum_{j=1}^{n_e}w_j^2,w_i)\Big]+\frac{Z^2}{2}\eta(\mu). \quad (6.65)\end{aligned}$$

We note that the above is consistent with our reduced basis approximation (6.62) — $a^{\mathrm{nl}}(w_i, \sum_{j=1}^{n_e} w_j^2, v_i)$ has been replaced by $a^{\mathrm{nl,M}}(w_i, \sum_{j=1}^{n_e} w_j^2, v_i)$.

Offline-online decomposition

Equation (6.63) has the affine parameter dependence form we desired and thus (6.62) readily submits to the offline-online treatment described in Section 4.3.3. During the online stage, the SCF algorithm outlined in Section 6.2.6 can be used to solve the resulting discrete equation. For each SCF iteration and given $\hat{\mathbf{u}}_{N,M}^k$ and $\phi_{N,M}^k$, we solve $\phi_{N,M}^{k+1}$ at a cost of $O((N^\phi)^3 + n_e N^u (N^\phi)^2)$ and $\hat{\mathbf{u}}_{N,M}^{k+1}$ at a cost of $O((N^u)^3 + N^\phi (N^u)^2 + n_e M (N^u)^2)$. With the convergence criteria $\max_{1\le i\le n_e} |\lambda_{N,M,i}^{k+1} - \lambda_{N,M,i}^k| < 1\mathrm{E}{-}10$, it typically takes less than 10 iterations to converge for the current problem. The total cost of the online computation is then independent of $\mathcal{N}$.

6.3.3 Vectorial Reduced Basis Space

The space

Here, we introduce nested sample sets $S_{N^u}^{V,u} = \{\mu_1^u, \ldots, \mu_{N^u}^u\}, 1 \le N^u \le N_{\max}^u$ and define the associated nested reduced-basis spaces as

$$\begin{aligned} W_{N^u}^{V,u} &= \text{span}\ \{\hat{\mathbf{u}}([Z,\mu_n^u]), 1\le n\le N^u\}, \quad 1\le N^u\le N_{\max}^u, && (6.66)\\ &= \text{span}\ \{\hat{\boldsymbol{\zeta}}_n, 1\le n\le N^u\}, \quad 1\le N^u\le N_{\max}^u; && (6.67)\end{aligned}$$

where $\hat{\mathbf{u}}([Z,\mu_n^u]) \equiv (u_1([Z,\mu_n^u]),\ldots,u_{n_e}([Z,\mu_n^u]))$ are the solutions of (6.46) at $\mu=\mu_n^u$ for a given Z; and $\hat{\boldsymbol{\zeta}}_n \equiv (\zeta_{n,1},\ldots,\zeta_{n,n_e})$ are basis functions obtained after $\hat{\mathbf{u}}([Z,\mu_n^u]), 1\le n\le N^u$ are sorted and aligned (described in Section 6.3.1), and pseudo-orthogonalized (described in Section 2.3.5). This allows us to obtain a smaller N^u and a better conditioned discrete system. Then, an approximation of $\hat{\mathbf{u}}$ in $W_{N^u}^{V,u}$ is given by $\hat{\mathbf{u}}_{N,M}([Z,\mu]) = \sum_{n=1}^{N^u}\psi_n([Z,\mu])\hat{\boldsymbol{\zeta}}_n$ — the ith component of $\hat{\mathbf{u}}_{N,M}([Z,\mu])$ is given by $u_{N,M,i}([Z,\mu]) = \sum_{n=1}^{N^u}\psi_n([Z,\mu])\zeta_{n,i}$, $1\le i\le n_e$.

The approximation

For the approximation based on the vectorial reduced basis space, a better starting point will be the energy statement: the equilibrium ground state of the resulting neutral structure for a particular Z $(=n_e)$ is given by $\hat{\mathbf{u}}_{N,M}([Z,\mu]) \equiv (u_{N,M,1}([Z,\mu]),\ \ldots,\ u_{N,M,n_e}([Z,\mu]))$, where

$$\begin{aligned}\hat{\mathbf{u}}_{N,M}([Z,\mu]) &= \arg\inf_{\hat{\mathbf{w}}}\Big\{E_{N,M}(\hat{\mathbf{w}}\equiv(w_1,\ldots,w_{n_e});[Z,\mu]), w_i\in W_{N^u}^{V,u}, \\ &\qquad \mu\int_\Omega w_i^2 = 1, 1\le i\le n_e\Big\}, \qquad (6.68)\\ \mu^*(Z) &= \arg\inf_{\mu}\{\mathcal{E}_{N,M}(\hat{\mathbf{u}}_{N,M}([Z,\mu]);[Z,\mu]);\mu>0\}. \qquad (6.69)\end{aligned}$$

This clearly shows that (6.68) is different from (6.1) — we only impose the constraints $\mu\int_\Omega u_{N,M,i}^2 = 1$, $1\le i\le n_e$; the orthogonality constraints of (6.1) are not present in (6.68). We hypothesized in Hypothesis 2.1 that the orthogonality property inherent in the basis functions $\hat{\boldsymbol{\zeta}}_n$ will lead an approximate solution $\hat{\mathbf{u}}_{N,M}$ that approximately obeys the orthogonality constraints, without explicit imposition of the orthogonality constraints — we demonstrate this empirically in Section 6.4.1.

The solution $\hat{\mathbf{u}}_{N,M}([Z,\mu])$ is obtained by solving the following Euler Lagrange equations: find $\mathbf{u}_{N,M}([Z,\mu]) \equiv (\hat{\mathbf{u}}_{N,M}([Z,\mu]),\ \phi_{N,M}([Z,\mu]),\ \hat{\boldsymbol{\lambda}}_{N,M}([Z,\mu]),\ \tau_{N,M}([Z,\mu])) \in \mathcal{Y}_N \equiv (W_{N^u}^{V,u}\times W_{N^\phi}^{\phi}\times \mathbb{R}^{n_e}\times\mathbb{R})$ such that

$$\mathcal{A}_{M,V}(\mathbf{u}_{N,M}([Z,\mu]),\mathbf{v};G;[Z,\mu]) = 0, \quad \forall \mathbf{v}\equiv(\hat{\mathbf{v}},\varsigma,\hat{\boldsymbol{\varphi}},\varpi)\in\mathcal{Y}_N, \qquad (6.70)$$

where $\hat{\lambda}_{N,M}([Z,\mu]) \equiv ((\lambda_{N,M})_{ii}, 1 \le i \le n_e)$ and

$$
\begin{aligned}
\mathcal{A}_{M,V}(\mathbf{w} \equiv (\hat{\mathbf{w}}, s, \hat{\boldsymbol{\sigma}}, \kappa), \mathbf{v} \equiv (\hat{\mathbf{v}}, \varsigma, \hat{\boldsymbol{\varphi}}, \varpi); t; [Z,\mu]) \equiv \\
\sum_{i=1}^{n_e} \Big[\theta_1([Z,\mu])\ a_0(w_i, v_i) + \ \theta_2([Z,\mu])\ a_2(w_i, t, v_i) + \theta_3([Z,\mu])\ a_2(w_i, s, v_i) \\
+ \theta_5([Z,\mu])\ a^{\mathrm{nl,M}}(w_i, \sum_{j=1}^{n_e} w_j^2, v_i) + 2\theta_4([Z,\mu])\ \sigma_{ii} a_1(w_i, v_i) \Big] \\
+ \sum_{i=1}^{n_e} \varphi_{ii} \{\beta_1([Z,\mu]) a_1(w_i, w_i) + \beta_2([Z,\mu])\delta_{ii}\} \\
+ \Big[\alpha_1([Z,\mu]) a_0(s,\varsigma) + \alpha_2([Z,\mu]) \sum_{j=1}^{n_e} a_2(w_j, w_j, \varsigma) + \alpha_3([Z,\mu]) l(\varsigma) + \kappa\alpha_4([Z,\mu])\ l(\varsigma) \Big] \\
+ \varpi\ l(s). \qquad (6.71)
\end{aligned}
$$

Finally the reduced-basis approximation for the electronic energy, $E_{N,M}(\mathbf{u}_{N,M}([Z,\mu]); G; [Z,\mu])$, is given by

$$
\begin{aligned}
E_{N,M}(\mathbf{w} \equiv (\hat{\mathbf{w}}, s, \hat{\boldsymbol{\sigma}}, \kappa); G; [Z,\mu]) \ = \ & \sum_{i=1}^{n_e} \Big[\theta_1([Z,\mu]) a_0(w_i, w_i) + \ \theta_2([Z,\mu])\ a_2(w_i, G, w_i) \\
& + \frac{1}{2}\theta_3([Z,\mu])\ a_2(w_i, s, w_i) \\
& + \frac{3}{4}\theta_5([Z,\mu])\ a^{\mathrm{nl,M}}(w_i, \sum_{j=1}^{n_e} w_j^2, w_i) \Big]. \qquad (6.72)
\end{aligned}
$$

and $\mathcal{E}_{N,M}(\mathbf{u}_{N,M}([Z,\mu]); G; [Z,\mu])$ is given by $E_{N,M}(\mathbf{u}_{N,M}([Z,\mu]); G; [Z,\mu]) + \frac{Z^2}{2}\eta(\mu)$.

When compared to (6.1) and (6.46), we have made two approximations. First, similar to the augmented reduced basis approximation, we approximate $a^{\mathrm{nl}}(w_i, \sum_{j=1}^{n_e} w_j^2, v_i)$

$$
a^{\mathrm{nl,M}}(w_i, \sum_{j=1}^{n_e} w_j^2, v_i) \equiv \int_\Omega g^w_{M^{g_i},i}\, v, \qquad (6.73)
$$

where $g^u_{M^{g_i},i}$ is an empirical interpolation approximation to $g_i(\hat{\mathbf{u}}) \equiv u_i(\sum_{j=1}^{n_e} u_j^2)^{1/3}$. Second, we only impose the constraints $\mu \int_\Omega u_{N,M,i}^2 = 1, 1 \le i \le n_e$. Finally, since $\int_\Omega \chi_n = 0$, $1 \le n \le N^u$, $\int_\Omega \phi_{N,M}$ is perforce zero; our discrete (nonlinear) algebraic system will thus have an actual dimension of $N^u + N^\phi + n_e$.

Offline-online decomposition

Equation (6.71) again has the affine parameter dependence form we desired and thus (6.70) readily submits to the offline-online computational decomposition. The fixed point method used in Section 6.3.2, however, cannot be used to solve the discrete equations at the online stage. We thus resort to Newton's method as described in Section 4.3.3. To assist convergence, we exploit a homotopy procedure in $\epsilon \in [0,1]$ and $\mu_k = \mu_i + \epsilon(\mu - \mu_i)$, where μ_k is the μ at k intermediate homotopy step and μ_i is the initial $\mu_{k=0}$ at start of the homotopy procedure — usually chosen to be the closest $\mu^u \in S^u_{N^u}$ to μ. The online complexity of the method is $O((N^\phi)^3 + n_e N^\phi (N^u)^2 + (N^u)^3 + n_e N^u M)$ per Newton iteration.

6.3.4 Error Measures

For a parameter test sample Ξ_T, we define the following error measures:

$$\varepsilon^u_{N,M} = \max_{\mu \in \Xi_T} \epsilon^u_{N,M}([Z,\mu]), \tag{6.74}$$

$$\varepsilon^\phi_{N,M} = \max_{\mu \in \Xi_T} \epsilon^\phi_{N,M}([Z,\mu]), \tag{6.75}$$

$$\varepsilon^{\mathcal{E}}_{N,M} = \max_{\mu \in \Xi_T} \epsilon^{\mathcal{E}}_{N,M}([Z,\mu]), \tag{6.76}$$

$$\varepsilon^{\text{ortho}}_{N,M} = \max_{\mu \in \Xi_T} \epsilon^{\text{ortho}}_{N,M}([Z,\mu]); \tag{6.77}$$

where

$$\epsilon^u_{N,M}([Z,\mu]) = \frac{(\sum_{i=1}^{n_e} \|u_i([Z,\mu]) - u_{N,M,i}([Z,\mu])\|_Y^2)^{1/2}}{(\sum_{i=1}^{n_e} \|u_i([Z,\mu])\|_Y^2)^{1/2}}, \tag{6.78}$$

$$\epsilon^\phi_{N,M}([Z,\mu]) = \frac{\|\phi([Z,\mu]) - \phi_{N,M}([Z,\mu])\|_Y}{\|\phi([Z,\mu])\|_Y}, \tag{6.79}$$

$$\epsilon^{\mathcal{E}}_{N,M}([Z,\mu]) = \frac{|\mathcal{E}_{N,M}([Z,\mu]) - \mathcal{E}([Z,\mu])|}{|\mathcal{E}([Z,\mu])|}, \tag{6.80}$$

$$\epsilon^{\text{ortho}}_{N,M}([Z,\mu]) = \max_{1 \le i < j \le n_e} \int_\Omega u_{N,M,i}([Z,\mu]) u_{N,M,j}([Z,\mu]). \tag{6.81}$$

Then, $\varepsilon^u_{N,M}$, $\varepsilon^\phi_{N,M}$ and $\varepsilon^{\mathcal{E}}_{N,M}$ are respectively the maximum error in the reduced-basis approximation of $\hat{\mathbf{u}}$, ϕ and $\mathcal{E}$ within a given sample Ξ_T; and $\varepsilon^{\text{ortho}}_{N,M}$ is a measure of non-compliance in the orthogonality constraints. The subscript N, M emphasizes the dependence of the approximation errors defined by (6.74) –(6.77) on N^u, N^ϕ and M^{g_i}, $1 \le i \le n_b$.

While error measures (6.74) –(6.77) are all relevant in the vectorial reduced basis approximation, only $\varepsilon^{\phi}_{N,M}$ and $\varepsilon^{\mathcal{E}}_{N,M}$ are relevant in the augmented reduced basis approximation[1]. With augmented reduced basis approximation, the components of $\hat{\mathbf{u}}_{N,M} \in W^{A,u}_{N^u}$ are not sorted or aligned, rendering the measure $\varepsilon^{u}_{N,M}$ meaningless. On the hand, the error measure $\varepsilon^{\text{ortho}}_{N,M}$ is identically zero as the eigenvalue solver within each SCF iteration ensures the orthogonality in $\hat{\mathbf{u}}_{N,M}$.

We will now define the projection errors for $\hat{\mathbf{u}}$ and ϕ. For a given $W^{\bullet,u}_{N^u}$ (referring to either $W^{A,u}_{N^u}$ or $W^{V,u}_{N^u}$) and $W^{\phi}_{N^\phi}$, we define

$$\varepsilon^{u}_{N,p} = \max_{\mu\in\Xi_T} \epsilon^{u}_{N,p}([Z,\mu]), \tag{6.82}$$

$$\varepsilon^{\phi}_{N,p} = \max_{\mu\in\Xi_T} \epsilon^{\phi}_{N,p}([Z,\mu]); \tag{6.83}$$

where

$$\epsilon^{u}_{N,p}([Z,\mu]) = \left(\frac{\sum_{i=1}^{n_e} \|u_{p,i}([Z,\mu]) - u_i([Z,\mu])\|^2_Y}{\sum_{i=1}^{n_e} \|u_i(\mu)\|^2_Y}\right)^{1/2}, \tag{6.84}$$

$$\epsilon^{\phi}_{N,p}([Z,\mu]) = \min_{\varphi\in W^{\phi}_{N^\phi}} \frac{\|\varphi - \phi([Z,\mu])\|_Y}{\|\phi([Z,\mu])\|_Y}. \tag{6.85}$$

Here $\hat{\mathbf{u}}_p([Z,\mu]) = (u_{p,1}([Z,\mu]),\ldots,u_{p,n_e}([Z,\mu])$ is the best projection of $\hat{\mathbf{u}}$ onto $W^{\bullet,u}_{N^u}$. For $W^{A,u}_{N^u}$, this is given by

$$u_{p,i}([Z,\mu]) = \arg\min_{w\in W^{A,u}_{N^u}} \|w - u_i([Z,\mu])\|_Y, \quad 1\le i\le n_e, \tag{6.86}$$

while for $W^{V,u}_{N^u}$, $\hat{\mathbf{u}}_p([Z,\mu]) = \sum_{j=1}^{N^u} \alpha^*_j([Z,\mu])\hat{\boldsymbol{\zeta}}_j$ where

$$\alpha^*([Z,\mu]) = \arg\min_{\alpha\in\mathbb{R}^{N^u}} \sum_{i=1}^{n_e} \left\| \sum_{j=1}^{N^u} \alpha_j \zeta_{j,i} - u_i([Z,\mu]) \right\|^2_Y. \tag{6.87}$$

Finally, we note that unlike Chapter 2 and 5, we do not have an error estimator for our quantities of interest. Thus, computation of the errors in the field variables will always be of $O(\mathcal{N})$.

[1]It may be possible to compute $\varepsilon^{u}_{N,M}$ with the augmented reduced basis approximation, but it requires additional efforts; we could use the alignment procedure to first align $\hat{\mathbf{u}}_{N,M}([Z,\mu])$ with $\hat{\mathbf{u}}([Z,\mu])$ before computing $\varepsilon^{u}_{N,M}$.

Given S_1^ϕ, W_1^ϕ;

Repeat $N^\phi = 2, \ldots$

$$\mu_{N^\phi}^* = \arg\max_{\mu \in \Xi_T} \epsilon_{N^\phi - 1, p}^\phi([Z, \mu]);$$
$$\varepsilon_{\max}^\phi = \epsilon_{N^\phi - 1, p}^\phi([Z, \mu_{N^\phi}^*]);$$
$$S_{N^\phi}^\phi = S_{N^\phi - 1}^\phi \cup \mu_{N^\phi}^*;$$
$$W_{N^\phi}^\phi = W_{N^\phi - 1}^\phi + \text{span } \{\phi([Z, \mu_{N^\phi}^*])\};$$

until $\varepsilon_{\max}^\phi \leq 1\text{E}{-5}$.

Given $S_1^{\bullet,u}$, $W_1^{\bullet,u}$;

Repeat $N^u = 2, \ldots$

$$\mu_{N^u}^* = \arg\max_{\mu \in \Xi_T} \epsilon_{N^u - 1}^{\mathcal{E}}([Z, \mu]);$$
$$\varepsilon_{\max}^{\mathcal{E}} = \epsilon_{N^u - 1}^{\mathcal{E}}([Z, \mu_{N^u}^*]);$$
$$S_{N^u}^{\bullet,u} = S_{N^u - 1}^{\bullet,u} \cup \mu_{N^u}^*;$$
$$W_{N^u}^{\bullet,u} = W_{N^u - 1}^{\bullet,u} + \text{span } \{u([Z, \mu_{N^u}^*])\};$$

until $\varepsilon_{\max}^{\mathcal{E}} \leq 1\text{E}{-10}$.

Figure 6-4: The two-pass sampling procedure to construct $S_{N^\phi}^\phi$ and $S_{N^u}^{\bullet,u}$.

6.3.5 Construction of Samples

We need to construct $n_e + 2$ approximation spaces: for each of $g_{M^{g_i},i}^u$, $1 \leq i \leq n_e$, we construct an independent $W_{M^{g_i}}^{g_i}$; for $\hat{\mathbf{u}}_{N,M}$, $W_{N^u}^{\bullet,u}$; and for $\phi_{N,M}$, $W_{N^\phi}^\phi$. As in Chapter 4, we first construct $W_{M^{g_i}}^{g_i}$, $1 \leq i \leq n_e$, followed by $W_{N^u}^{\bullet,u}$ and $W_{N^\phi}^\phi$. The spaces $W_{M^{g_i}}^{g_i}$ are constructed based on the empirical interpolation method described in Chapter 3. There are, however, several ways by which nested reduced basis sample sets $S_{N^u}^{\bullet,u}$ (referring to either $S_{N^u}^{A,u}$ or $S_{N^u}^{V,u}$) and $S_{N^\phi}^\phi$ can be constructed within the framework of the adaptive sampling procedure outlined in Chapter 4, each with a different effect on the final optimality of the spaces.

We choose to perform a two-pass adaptive sampling procedure to construct the spaces $W_{N^u}^{\bullet,u}$ and $W_{N^\phi}^\phi$. In the first pass, we construct $S_{N^\phi}^\phi$ (and correspondingly $W_{N^\phi}^\phi$) based on $\epsilon_{N,p}^\phi([Z, \mu])$ with the convergence criterion $\varepsilon_{N,p}^\phi < 1\text{E}{-5}$; note we have used the projection error to construct S_N^ϕ. In the second past, armed with a good approximation of ϕ, we construct $W_{N^u}^{\bullet,u}$ based on the $\epsilon_{N,M}^{\mathcal{E}}([Z, \mu])$ with the convergence criterion $\varepsilon_{N,M}^{\mathcal{E}} < 1\text{E}{-10}$. Figure 6-4 summarizes the procedure.

The use of projection error $\epsilon^{\phi}_{N,p}([Z,\mu])$ as an error measure allows us to construct a reduced basis space for ϕ independent of $\hat{\mathbf{u}}$, thus isolating the behavior of the space $W^{\phi}_{N^{\phi}}$ from $W^{\bullet,u}_{N^u}$. We can use the projection error $\epsilon^{\phi}_{N,p}([Z,\mu])$ as an error measure because we already have the solutions at all $\mu \in \Xi_T$ in order to construct the collateral spaces $W^{g_i}_{M^{g_i},i}$ for $g^u_{M^{g_i},i}$. While the use of projection error and the two-pass procedure are expensive, they constitute only a small portion of the overall costs since computing the solutions at all $\mu \in \Xi_T$ will constitute the bulk of the offline computational costs. Certainly, if error estimators are available, the sampling procedure will be more efficient. This can be addressed in future work.

There are of course other possible ways we can construct $S^{\phi}_{N^{\phi}}$ and $S^{\bullet,u}_{N^u}$. The simplest choice is to choose a sample set such that $S_N \equiv S^{\bullet,u}_{N^u} \equiv S^{\phi}_{N^{\phi}}$ by doing a single pass of the adaptive sampling procedure based on the error measure $\epsilon^{\mathcal{E}}_{N,M}([Z,\mu])$. This is, however, suboptimal since variation of $\hat{\mathbf{u}}([Z,\mu])$ with μ can differ from that of $\phi([Z,\mu])$. The choice of sample points and the size of the sample spaces can thus differ significantly from one another.

Another choice is to construct $S^{\bullet,u}_{N^u}$ and $S^{\phi}_{N^{\phi}}$ based on different error measures within a single pass of the adaptive sampling procedure. We construct, say, $S^{\bullet,u}_{N^u}$ based on $\epsilon^{\mathcal{E}}_{N,M}([Z,\mu])$ and $S^{\phi}_{N^{\phi}}$, $\epsilon^{\phi}_{N,M}([Z,\mu])$. At first glance, this may suggest $S^{\bullet,u}_{N^u}$ and $S^{\phi}_{N^{\phi}}$ have been constructed independently from one another. However, if approximation error resulting from, say $S^{\phi}_{N^{\phi}}$, significantly affects $\epsilon^{\mathcal{E}}_{N,M}([Z,\mu])$, especially during the intermediate steps of the adaptive sampling procedure, we cannot then satisfactorily conclude that $S^{\bullet,u}_{N^u}$ is indeed optimal.

6.4 Numerical Results

6.4.1 Convergence

We consider $\mu \in \mathcal{D} \equiv [7,12]$ and $n_e = 3$, 5, 7 and 9. These parameters are selected such that $\mathcal{E}_{N,M}([Z,\mu])$ attains its minimum within the range $\mathcal{D}$. We introduce a parameter test sample set $\Xi_T \subset \mathcal{D}$ consisting of 200 sample points distributed uniformly in $\mathcal{D}$. The same sample set is used as our training sample set for constructing our reduced basis spaces. For each $g^u_{M^{g_i},i}$, $1 \le i \le n_e$, we choose an M^{g_i} such that empirical interpolation errors in $g^u_{M^{g_i},i}$, $1 \le i \le n_e$ are less than 10^{-12}. In Table 6.1, we present the approximation errors $\varepsilon^{\mathcal{E}}_{N,M}$ and $\varepsilon^{\phi}_{N,M}$ for the augmented reduced basis space for the case $n_e = 5$. Here, $N^{\phi} = 5$ and $M = 13$. We observe a monotonic decrease in the

N^u	N_s^u	$\varepsilon_{N,M}^{\mathcal{E}}$	$\varepsilon_{N,M}^{\phi}$
5	1	2.6538 E−1	6.8988 E−1
10	2	2.3079 E−3	3.8350 E−2
15	3	5.3975 E−7	9.7910 E−5
20	4	2.2108 E−8	8.0856 E−6
25	5	6.8321 E−9	5.9510 E−6
30	6	4.2363 E−10	5.9509 E−6

Table 6.1: Variations of reduced-basis errors $\varepsilon_{N,M}^{\mathcal{E}}$, and $\varepsilon_{N,M}^{\phi}$ with N^u for $n_e = 5$ and $\hat{\mathbf{u}}_{N,M}(\mu) \in (W_{N^u}^{A,u})^{n_e}$. The corresponding N_s^u are also listed.

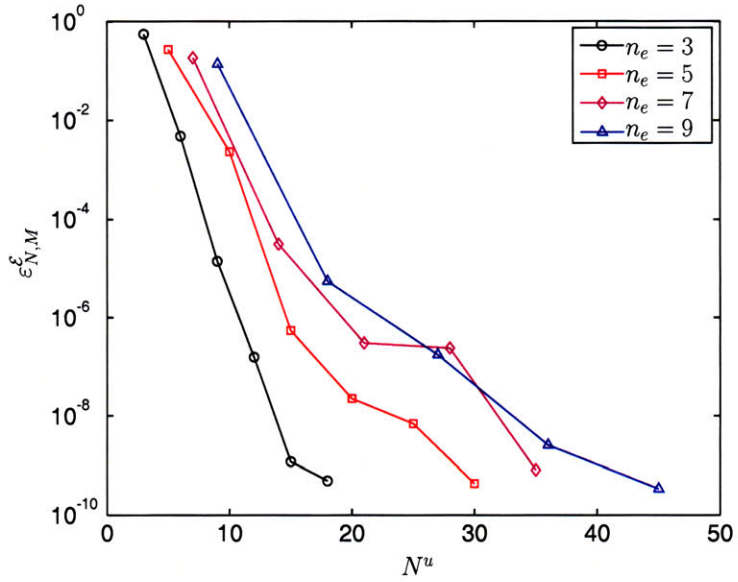

Figure 6-5: Convergence of the reduced basis error $\varepsilon_{N,M}^{\mathcal{E}}$ for $\hat{\mathbf{u}}_{N,M}(\mu) \in (W_{N^u}^{A,u})^{n_e}$.

approximation errors. We only require 20 basis functions to accurately approximate $\mathcal{E}$ to a relative error of 10^{-8}. In addition, from Figure 6-5, we see similar behavior for all the cases examined.

For vectorial reduced basis space, we also observe a monotonic decrease in the approximation errors $\varepsilon_{N,M}^{u}$, $\varepsilon_{N,M}^{\phi}$, $\varepsilon_{N,M}^{\mathcal{E}}$, and $\varepsilon_{N,M}^{\text{ortho}}$ as N^u increases for the case $n_e = 5$, as shown in Table 6.2. The N^{ϕ} and M are the same as that for the augmented reduced basis approximation. In this case, we only require $N^u = 9$ to reduce $\varepsilon_{N,M}^{\mathcal{E}}$ to a relative error of 10^{-9}. We also note that $\varepsilon_{N,M}^{\mathcal{E}}$ is approximately the square of $\varepsilon_{N,M}^{u}$ (and $\varepsilon_{N,M}^{\phi}$), indicating that $\mathcal{E}$ can in fact be approximated very accurately with very few basis functions by $\mathcal{E}_{N,M}$. At $N^u = n_e$, the apparent "good orthogonality" is deceiving. When the size of $N^u = n_e$, the number of constraints are equivalent to dimension of $W_{N^u}^{V,u}$. Our optimization problem then has zero degree of freedom and our solution is fully determined by

N^u	$\varepsilon^u_{N,M}$	$\varepsilon^{\mathcal{E}}_{N,M}$	$\varepsilon^{\phi}_{N,M}$	$\varepsilon^{\text{ortho}}_{N,M}$
5	1.7636 E−1	8.3993 E−3	1.8600 E−1	3.9327 E−14
6	7.3748 E−2	1.5655 E−3	8.2272 E−2	1.0626 E−3
7	5.4825 E−3	8.7358 E−6	2.3707 E−3	1.6400 E−4
8	3.5230 E−3	6.3719 E−7	9.5122 E−4	5.5515 E−5
9	1.2251 E−4	1.3157 E−9	5.0798 E−5	3.2761 E−6
10	1.0978 E−5	3.7331 E−11	5.9562 E−6	1.9997 E−7

Table 6.2: Variations of reduced-basis errors $\varepsilon^u_{N,M}$, $\varepsilon^{\mathcal{E}}_{N,M}$, $\varepsilon^{\phi}_{N,M}$ and $\varepsilon^{\text{ortho}}_{N,M}$ with N^u for $n_e = 5$ and $\hat{\mathbf{u}}_{N,M}([Z,\mu]) \in W^{V,u}_{N^u}$.

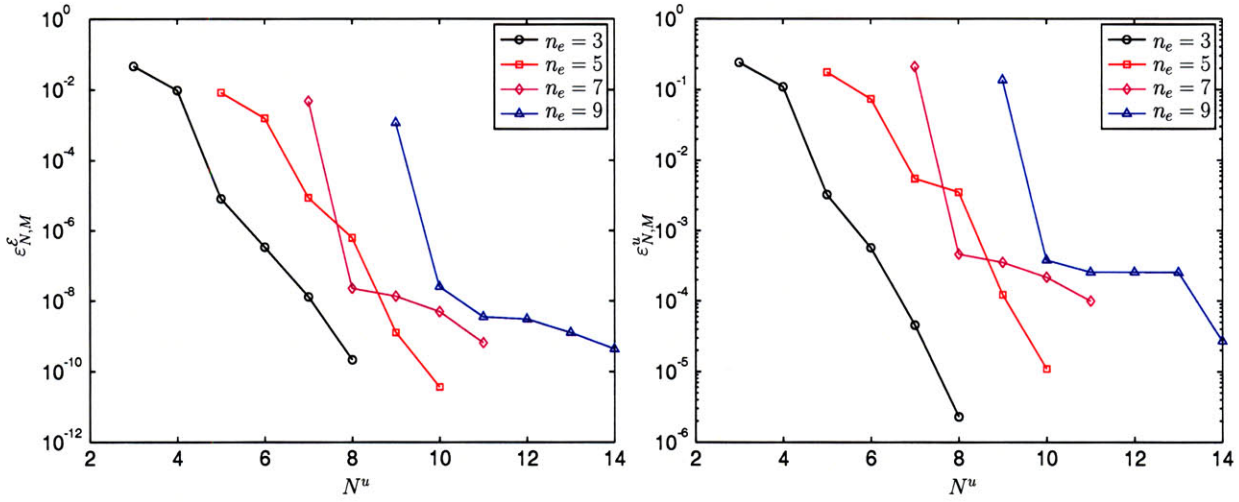

Figure 6-6: Convergence of the reduced basis error $\varepsilon^{\mathcal{E}}_{N,M}$ and $\varepsilon^u_{N,M}$ for $\hat{\mathbf{u}}_{N,M}(\mu) \in W^{V,u}_{N^u}$.

the constraints. For our case, the solution will always be one of the scaled solutions at $\mu \in S^{V,u}_{N^u}$, for which the orthogonality is obviously satisfied. In Figure 6-6, we show the convergence of $\varepsilon^u_{N,M}$ and $\varepsilon^{\mathcal{E}}_{N,M}$ for $n_e = 3$, 5, 7 and 9. We again observe that they decrease monotonically with increasing N^u in all these cases.

6.4.2 Comparison

From Table 6.3, we observe that the size of $W^{V,u}_{N^u}$ is smaller than $W^{A,u}_{N^u}$ for all n_e examined. In addition, for $W^{V,u}_{N^u}$, the number of basis functions required for each case is only slightly higher than n_e; N^u scales approximately as $n_e + C_V$, where C_V is a small integer. Here, we deduce $C_V \approx 5$. For $W^{A,u}_{N^u}$, we have N^u that scales approximately as $N_s n_e$, where N_s is relatively insensitive to n_e. For our current example, N_s is between 5 and 6.

We wish to contrast these results to that of Chapter 5. First, for the current problem, N_s

	N	
n_b	$W_{N^u}^{A,u}$	$W_{N^u}^{V,u}$
3	18	8
5	30	10
7	35	11
9	45	14

Table 6.3: Comparison between the augmented reduced basis approximation and the vectorial reduced basis approximation based on N required to reduce $\varepsilon_{N,M}^{\mathcal{E}}$ to below 1E−9 for $n_e = 3$, 5, 7 and 9.

does not decrease appreciably with n_e even when the solutions are approximated very accurately. Second we are also able to obtain an efficient vectorial reduced basis approximation. This is because the preprocessing algorithm is able to obtain a smooth variation of $u_i(\mu)$, $1 \leq i \leq n_e$ with respect to $\mu \in \mathcal{D}$ as shown in Figure 6-3. The better performance of the algorithm for this problem when compared to Section 5.3.2 is perhaps due to the smaller parameter domain and thus more limited variation in $\hat{\mathbf{u}}([Z, \mu])$ with respect to μ. The pre-processed basis functions then allows the approximation procedure to efficiently exploit the inherent orthogonality property in the vectorial reduced basis space $W_{N^u}^{V,u}$. In addition, as explained in Section 2.3.2 and 2.3.7, the common smoothness of $u_i([Z, \mu])$, $1 \leq i \leq n_e$ might have played a role, an important aspect of the vectorial reduced basis approximation which should be explored further in future work.

The above comparison between $W_{N^u}^{A^u}$ and $W_{N^u}^{V,u}$ is based solely on the dimension of the reduced basis spaces. However, the online computational cost is also strongly dependent on the solution method used. As described in Section 6.3.2 and 6.3.3, the solution method for $W_{N^u}^{A,u}$ is the SCF algorithm while for $W_{N^u}^{V,u}$, the Newton's method. For $n_e = 5$, online computation cost per evaluation of μ for $W_{N^u}^{A,u}$ is 0.015s while for $W_{N^u}^{V,u}$, 0.13s. Even though the dimension of $W_{N^u}^{V,u}$ is smaller than $W_{N^u}^{A,u}$, the online computation for $W_{N^u}^{V,u}$ is nearly 10 times slower than $W_{N^u}^{A,u}$, reflecting the greater overall efficiency of the augmented reduced basis approximation. Certainly, there is room for improvement in the online solution method used in our vectorial reduced basis approximation.

6.5 Application

As an illustrative application, we shall determine μ^*, μ at which the total energy is minimum, for $Z = n_e = 5$. We require an absolute accuracy of 10^{-5} in μ^*. We use the simple bisection method for

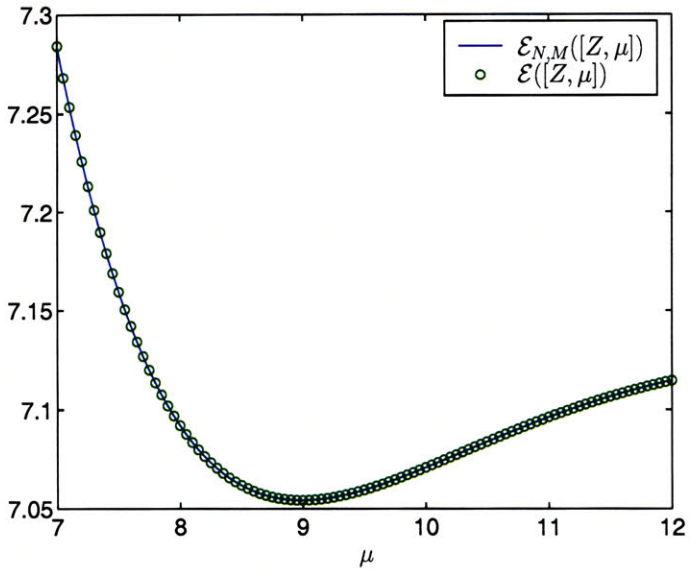

Figure 6-7: Comparison between the $\mathcal{E}([Z,\mu])$ and $\mathcal{E}_{N,M}([Z,\mu])$ for $Z = 5$ and $7 \leq \mu \leq 12$.

this purpose; it will thus involve repetitive evaluation of $\mathcal{E}([Z,\mu])$. The approach is straightforward. We approximate $\mathcal{E}([Z,\mu])$ by $\mathcal{E}_{N,M}([Z,\mu])$ based on our vectorial reduced basis approximation. We have use $N^u = 9$, $N^\phi = 5$ and maximum M is 13. The variation of $\mathcal{E}([Z,\mu])$ and $\mathcal{E}_{N,M}([Z,\mu])$ with μ is shown in Figure 6-7.

A total of 21 evaluations of $\mathcal{E}_{N,M}([Z,\mu])$ is required and μ^* is found to be 8.875. The total online computational cost is only 4s. On the contrary, with finite element approximation of $\mathcal{N} = 400$, the total computational cost is 700s. If we compare only the online cost with the total computational cost based on "truth" approximation, we achieve a computational saving of order $O(10)$.

However, taking the offline computational cost into consideration, the reduced basis method is not competitive for this particular problem. Due to absence of efficient *a posteriori* error estimation procedure, we must first compute solutions at all $\mu \in \Xi_T$ — for our example, there are 200 sample points in Ξ_T, leading to a total offline computational cost of 6500s. Clearly, if this reduced basis model is used only for this one application, it would have made no sense. However, if we reuse the model in other applications, we may be able to justify the initial computational overhead during the offline stage. For example, the current reduced basis model, which evaluates only the total energy, can be directly used to determine static structural properties such as elasticity constant and bulk modulus [120]. With some extensions to the current model, we can further develop efficient reduced basis strategies for lattice dynamics simulations. For study of harmonic vibrations

and phonon modes, we can develop a reduced basis model based on Density Functional Perturbation Theory [7, 42]. For study of both harmonic and anharmonic vibrations, we can employ the frozen phonon approach which requires the use of supercell — although this leads to a larger simulation cell and more complicated parameterizations due to large number of nuclei, we only need to perform total energy calculations and employ finite difference formulae to determine all the quantities of interests [28, 76]. All the above examples serve to emphasize the many query limit in which the reduced basis method is most useful for.

Chapter 7

Three Dimensional Kohn Sham Equations

7.1 Introduction

We now extend the methodology developed in Chapter 6 to three dimensional Kohn Sham equations. The overall methodology is unchanged, but the higher dimension leads to greater numerical complexity. In particular, the "truth" approximation is significantly more complicated — a complete development of the numerical codes would exceed the duration of this thesis. We thus limit the problems examined to those that can be successfully handled by our admittedly sub-optimal finite element codes. This restriction somewhat limits the complexity of the problems addressed in this chapter. As such, we examine only one numerical example: the electronic ground state energy calculation for a simple cubic structure of Beryllium.

7.2 Abstract Formulation

7.2.1 Problem Statement

We consider a simple cubic structure with lattice parameter μ, and hence unit cell $\tilde{\Omega}(\mu) \equiv]-\frac{\mu}{2}, \frac{\mu}{2}]^3$. A single nucleus of charge Z lies at the center of the cell. To each nucleus, we associate n_e orbitals[1];

[1]Note that in Chapter 6, n_e refers to the number of electrons while in this chapter n_e refers to the number of orbitals.

as we will be using models based on spinless Density Functional Theory [18, 32, 76, 90], 2 electrons are assigned to each orbital. For charge neutrality, we then have $n_e = Z/2$.

Our output of interest is again the ground state energy of the system, $\tilde{\mathcal{E}}$, which we shall determine based on the spinless Density Functional Theory [18, 32, 76, 90]. Our input parameter is the lattice length μ. For simplicity, we shall not include Z in our parameter space; as such, each new Z constitutes a new problem in our reduced-basis approximation.

The energy statement, the derivation of the Euler-Lagrange equations and their weak forms, and the parameterization procedure are little changed when compared to Chapter 6. As such, we will give a brief description of the energy statement followed directly by the parameterized abstract formulation; we will skip the derivation of the Euler-Lagrange equations and the abstract formulation in the original domain.

7.2.2 Energy Statement

The equilibrium ground state energy is obtained by solving a minimization problem for $\tilde{\mathbf{u}}([Z,\mu^*]) \equiv (\tilde{u}_1([Z,\mu^*]),\ldots,\tilde{u}_{n_e}([Z,\mu^*]))$, where [11, 25, 64, 66]

$$\begin{aligned}\tilde{\mathbf{u}}([Z,\mu]) &= \arg\inf_{\tilde{\mathbf{w}}}\Big\{\tilde{E}(\tilde{\mathbf{w}}\equiv(\tilde{w}_1,\cdots,\tilde{w}_{n_e});[Z,\mu]),\tilde{w}_i\in\tilde{Y}, \\ &\qquad \int_{\tilde{\Omega}(\mu)}\tilde{w}_i\tilde{w}_j=\delta_{ij},1\le i,j\le n_e\Big\}, \qquad (7.1)\\ \mu^*(Z) &= \arg\inf_{\mu}\Big\{\tilde{\mathcal{E}}(\tilde{\mathbf{u}}([Z,\mu]);[Z,\mu]);\mu>0\Big\}; \qquad (7.2)\end{aligned}$$

here $\tilde{Y}\equiv H^1_{\text{per}}(\tilde{\Omega}(\mu))$ is the space of μ-periodic functions in $H^1(\mathbb{R}^3)$; $\delta_{ij}=\{1\text{ if }i=j,\ 0\text{ otherwise}\}$; and $\tilde{u}_i$ is the ith Kohn-Sham orbital. The electronic energy $\tilde{E}(\tilde{\mathbf{w}};[Z,\mu])$ is defined as

$$\begin{aligned}\tilde{E}(\tilde{\mathbf{w}};[Z,\mu]) &= C_w\sum_{i=1}^{n_e}\int_{\tilde{\Omega}(\mu)}(\nabla\tilde{w}_i)^2 - Z\sum_{i=1}^{n_e}\int_{\tilde{\Omega}(\mu)}\tilde{G}\,\tilde{w}_i{}^2 \\ &\quad + \frac{1}{2}C_c\int_{\tilde{\Omega}(\mu)}\int_{\tilde{\Omega}(\mu)}\left(2\sum_{i=1}^{n_e}\tilde{w}_i^2(\tilde{x})\right)\tilde{G}(\tilde{x}-\tilde{y})\left(2\sum_{j=1}^{n_e}\tilde{w}_j^2(\tilde{y})\right)d\tilde{x}\,d\tilde{y} \\ &\quad - C_x\int_{\tilde{\Omega}(\mu)}\left(2\sum_{j=1}^{n_e}\tilde{w}_i^2\right)^{4/3}, \qquad (7.3)\end{aligned}$$

where $\tilde{x} \in \mathbb{R}^3$ denotes a point in $\tilde{\Omega}(\mu)$; and C_w, C_c, and C_x are model constants — we use $C_w = 0.5$, $C_c = 1$ and $C_x = 0.7386$. The periodic Green's function $\tilde{G}(\cdot\,;\mu)\colon\ \tilde{\Omega}(\mu) \rightarrow \mathbb{R}^3$ satisfies

$$-\triangle \tilde{G} = 4\pi \left\{ \delta(\tilde{x}) - \frac{1}{|\tilde{\Omega}(\mu)|} \right\}, \qquad \int_{\tilde{\Omega}(\mu)} \tilde{G} = 0, \tag{7.4}$$

where $\triangle$ is the Laplacian operator, $\delta(\tilde{x})$ is the Dirac delta distribution, and $|\tilde{\Omega}(\mu)| = \mu^3$ is the volume of $\tilde{\Omega}(\mu)$.

The total energy $\tilde{\mathcal{E}}(\tilde{\mathbf{w}}; [Z, \mu])$ — our output of interest — is then given by

$$\tilde{\mathcal{E}}(\tilde{\mathbf{w}}; [Z, \mu]) = \tilde{E}(\tilde{\mathbf{w}}; [Z, \mu]) + \frac{Z^2}{2}\eta(\mu), \tag{7.5}$$

where $\eta(\mu)$ is the nuclear - nuclear correction term given by

$$\eta(\mu) = \lim_{\tilde{x} \rightarrow 0} \left\{ \tilde{G}(\tilde{x}; \mu) - \frac{1}{|\tilde{x}|} \right\}. \tag{7.6}$$

Unlike the one dimensional Kohn Sham equations, $\eta(\mu)$ does not have a close form solution — we evaluate $\eta(\mu)$ numerically in Section 7.2.4.

7.2.3 Parameterized Abstract Formulation

From the weak form of the Euler Lagrange equations for the constrained minimization problem (7.1), we derive the equivalent parameterized abstract formulation. We first define an affine geometric mapping, $\mathcal{G}(\mu)$, from $\tilde{\Omega}(\mu)$ to $\Omega \equiv]-\frac{1}{2}, \frac{1}{2}[^3$ which for our simple cubic structure, can be expressed as

$$x = \mathcal{G}(\tilde{x}; \mu) \equiv \frac{1}{\mu}\tilde{x}. \tag{7.7}$$

We also define $u_i([Z, \mu]) = \tilde{u}_i \circ \mathcal{G}^{-1}(\cdot\,;\mu)$, $\phi([Z, \mu]) = \mu\tilde{\phi} \circ \mathcal{G}^{-1}(\cdot\,;\mu)$ and $G = \mu\tilde{G} \circ \mathcal{G}^{-1}(\cdot\,;\mu)$.[2] Then $\mathbf{u}([Z, \mu]) \equiv (\hat{\mathbf{u}}([Z, \mu]), \phi([Z, \mu]), \hat{\boldsymbol{\lambda}}([Z, \mu]), \tau([Z, \mu])) \in \mathcal{Y} \equiv (Y^{n_e} \times Y \times \mathbb{R}^{n_e(n_e+1)/2} \times \mathbb{R})$ satisfies

$$\mathcal{A}(\mathbf{u}([Z, \mu]), \mathbf{v}; G; [Z, \mu]) = 0, \qquad \forall \mathbf{v} \in \mathcal{Y}, \tag{7.8}$$

[2]Note that the dimensionality of the problem changes the scaling factor of ϕ and G; in Chapter 6 where we look at one dimensional problems, $\phi = \frac{1}{\mu}\tilde{\phi} \circ \mathcal{G}^{-1}(\cdot\,;\mu)$ and $G = \frac{1}{\mu}\tilde{G} \circ \mathcal{G}^{-1}(\cdot\,;\mu)$

where $Y \equiv H^1_{\rm per}(\Omega)$ is the space of periodic function in $H^1(\mathbb{R}^3)$; $\hat{\mathbf{u}}([Z,\mu]) \equiv (u_i([Z,\mu]), 1 \le i \le n_e)$; $\hat{\boldsymbol{\lambda}}([Z,\mu]) \equiv (\lambda_{ij}([Z,\mu]), 1 \le i \le j \le n_e)$; G satisfies

$$a_0(G,\varrho) + 4\pi\{l(\varrho) - \varrho(0)\} = 0, \quad \forall \varrho \in Y; \quad l(G) = 0; \tag{7.9}$$

and $\mathcal{A}$ is defined as

$$\begin{aligned}
&\mathcal{A}(\mathbf{w} \equiv (\hat{\mathbf{w}}, s, \hat{\boldsymbol{\sigma}}, \kappa), \mathbf{v} \equiv (\hat{\mathbf{v}}, \varsigma, \hat{\boldsymbol{\varphi}}, \varpi); t; \mu) \equiv \\
&\quad \sum_{i=1}^{n_e} \Big[\theta_1([Z,\mu])\ a_0(w_i, v_i) + \ \theta_2([Z,\mu])\ a_2(w_i, t, v_i) + \theta_3([Z,\mu])\ a_2(w_i, s, v_i) \\
&\quad + \theta_5([Z,\mu])\ a^{\rm nl}(w_i, \sum_{j=1}^{n_e} w_j^2, v_i) + \theta_4([Z,\mu])\ \sigma_{ii} a_1(u_i, v_i) + \theta_4([Z,\mu])\ \sum_{j=1}^{n_e} \sigma_{ij}\ a_1(w_j, v_i)\Big] \\
&\quad + \sum_{i=1}^{n_e}\sum_{j=i}^{n_e} \varphi_{ij}\left\{\beta_1([Z,\mu]) a_1(w_i, w_j) + \beta_2([Z,\mu])\delta_{ij}\right\} \\
&\quad + \Big[\alpha_1([Z,\mu]) a_0(s,\varsigma) + \alpha_2([Z,\mu]) \sum_{j=1}^{n_e} a_2(w_j, w_j, \varsigma) + \alpha_3([Z,\mu]) l(\varsigma) + \kappa\alpha_4([Z,\mu])\ l(\varsigma)\Big] \\
&\quad + \varpi\ l(s).
\end{aligned} \tag{7.10}$$

Here, $a_0(w,v) \equiv \int_\Omega \nabla w \nabla v$, $a_1(w,v) \equiv \int_\Omega w\, v$, $a_2(w,s,v) \equiv \int_\Omega w\, s\, v$, $a^{\rm nl}(w,t,v) \equiv \int_\Omega w\, t^{1/3} v$, and $l(w) \equiv \int_\Omega w$ for any $w \in Y$, $v \in Y$, $s \in Y$, and non-negative $t \in Y$. For prescribed C_w, C_x, and C_c, θ, α and β are given by

$$\theta([Z,\mu]) = \left[C_w\mu, Z\,\mu^2, -\mu^2, -\frac{\mu^3}{2}, -C_x\frac{4}{3}2^{1/3}\mu^3\right], \tag{7.11}$$

$$\alpha([Z,\mu]) = \left[1, C_c 2\mu^3, -C_c Z, \mu^3\right], \tag{7.12}$$

$$\beta([Z,\mu]) = \left[\mu^3, -1\right]. \tag{7.13}$$

The total energy $\mathcal{E}(\mathbf{u}(\mu); G; [Z,\mu])$ is then given by

$$\begin{aligned}
\mathcal{E}(\mathbf{w} \equiv (\hat{\mathbf{w}}, s, \hat{\boldsymbol{\sigma}}, \kappa); G; \mu) \ = \ &\sum_{i=1}^{n_e} \Big[\theta_1([Z,\mu]) a_0(w_i, w_i) + \theta_2([Z,\mu])\ a_2(w_i, G, w_i) \\
&+ \frac{1}{2}\theta_3([Z,\mu])\ a_2(w_i, s, w_i) + \ \frac{3}{4}\theta_5([Z,\mu])\ a_1^{\rm nl}(w_i, \sum_{j=1}^{n_e} w_j^2, w_i)\Big] \\
&+ \frac{Z^2}{2}\eta(\mu),
\end{aligned} \tag{7.14}$$

where

$$\eta(\mu) = \lim_{x \to 0} \left\{ G(x) - \frac{1}{|x|} \right\}. \tag{7.15}$$

7.2.4 "Truth" Approximation

Again, we will use finite element approximation as our "truth" approximation. We first define our finite element space $Y_h \subset Y$ of dimension $\mathcal{N}$ as

$$Y_h \equiv \{v \in Y \mid v|_{\mathbf{T}_h} \in \mathbf{Q}_3(\mathbf{T}_h),\ \forall \mathbf{T}_h \in \mathcal{T}_h\}, \tag{7.16}$$

$$\mathbf{Q}_3(\mathbf{T}_h) \equiv \text{span}\{1,\ x_1,\ x_2,\ x_3,\ x_1x_2,\ x_2x_3,\ x_1x_3,\ x_1x_2x_3\}, \tag{7.17}$$

where $\mathcal{T}_h$ is a (regular) uniform "triangulation" of the domain Ω comprising of cubical elements T_h of edge-length h. As we shall see, our low-order uniform mesh is far from optimal, in particular for higher Z.

Our finite element approximation to (7.8) is then given by: find $\mathbf{u}_h([Z,\mu]) \equiv (\hat{\mathbf{u}}_h([Z,\mu]), \phi_h([Z,\mu]), \hat{\boldsymbol{\lambda}}_h([Z,\mu]), \tau_h([Z,\mu])) \in \mathcal{Y}_h \equiv ((Y_h)^{n_e} \times Y_h \times \mathbb{R}^{n_e(n_e+1)/2} \times \mathbb{R})$ such that

$$\mathcal{A}_h(\mathbf{u}_h([Z,\mu]), \mathbf{v}; G^{\text{ES}}; [Z,\mu]) = 0, \quad \forall \mathbf{v} \in \mathcal{Y}_h. \tag{7.18}$$

here G^{ES} is the Ewald Sum approximation to G, the periodic Green function; and $\mathcal{A}_h$ is an approximation to $\mathcal{A}$ in which the terms $a^{\text{nl}}(w,t,v)$ and $a_2(w,s,v)$ are replaced by the quadrature sums as described in Section 4.2.2. The finite element approximation to the total energy, $\mathcal{E}_h(\mathbf{u}_h([Z,\mu]); G^{\text{ES}}; [Z,\mu])$, is then given by

$$\begin{aligned}
\mathcal{E}_h(\mathbf{u}_h([Z,\mu]); G^{\text{ES}}; [Z,\mu]) = \sum_{i=1}^{n_e} \Big[& \theta_1([Z,\mu])\ a_0(u_{h,i}, u_{h,i}) + \theta_2([Z,\mu])\ a_2(u_{h,i}, G^{\text{ES}}, u_{h,i}) \\
& + \frac{1}{2}\theta_3([Z,\mu])\ a_2(u_{h,i}, \phi_i, u_{h,i}) + \frac{3}{4}\theta_5([Z,\mu])\ a^{\text{nl}}(u_{h,i}, \sum_{j=1}^{n_e} u_{h,j}^2, u_{h,i}) \Big] \\
& + \frac{Z^2}{2}\eta(\mu);
\end{aligned} \tag{7.19}$$

Finally (7.18) can be solved using the fixed point method described in Section 6.2.6.

Ewald summation

As derived earlier, G is independent of μ and Z (and of course $\hat{\mathbf{u}}$ and ϕ), and hence (7.9) need only be addressed once. However, unlike the one-dimensional case, the Dirac delta function is not a bounded functional with respect to H^1 in $\mathbb{R}^3$. Finite element method cannot satisfactorily handle the singularity in the resulting equation even with regularization; the convergence is poor and very high resolution in the vicinity of the singularity is required. We thus apply the standard Ewald Sum method [11]. To begin, we write [11]

$$G(x) = \hat{G}(x) + \frac{1}{|x|}, \tag{7.20}$$

where

$$\hat{G}(x) = -\frac{\pi}{\gamma} - \frac{\operatorname{erf}(\sqrt{\gamma}|x|)}{|x|} + \sum_{j \in \mathbb{Z}^3_\infty \setminus \{0\}} \frac{\operatorname{erfc}(\sqrt{\gamma}|x-j|)}{|x-j|} + \sum_{k \in \mathbb{Z}^3_\infty \setminus \{0\}} \frac{e^{\frac{-\pi^2|k|^2}{\gamma} + 2\pi i k \cdot x}}{\pi |k|^2}; \tag{7.21}$$

here erf and erfc are the "error function" and "complementary error function," $\mathbb{Z}_\infty \equiv \{-\infty, \cdots, -2, -1, 0, 1, 2, \cdots, \infty\}$, and γ is a convergence tuning parameter (which we set to unity). By truncating the the sums in (7.21), we obtain the Ewald Sum approximation to $G(\cdot)$:

$$G^{\mathrm{ES}}(x) \equiv \hat{G}^{\mathrm{ES}}(x) + \frac{1}{|x|}, \tag{7.22}$$

where

$$\hat{G}^{\mathrm{ES}}(x) \equiv -\frac{\pi}{\gamma} - \frac{\operatorname{erf}(\sqrt{\gamma}|x|)}{|x|} + \sum_{j \in \mathbb{Z}^3_{j_o} \setminus \{0\}} \frac{\operatorname{erfc}(\sqrt{\gamma}|x-j|)}{|x-j|} + \sum_{k \in \mathbb{Z}^3_{k_o} \setminus \{0\}} \frac{e^{\frac{-\pi^2|k|^2}{\gamma} + 2\pi i k \cdot x}}{\pi |k|^2}; \tag{7.23}$$

here $\mathbb{Z}_n \equiv \{-n, \cdots, -2, -1, 0, 1, 2, \cdots, n\}$; and the positive integers j_o and k_o are the physical-space and Fourier-space cut-offs, respectively. Note that the term $\lim_{x \to 0}\{G^{\mathrm{ES}}(x) - \frac{1}{|x|}\}$ term in (7.15) is now simply given by $\hat{G}^{\mathrm{ES}}(0)$.

We demonstrate in Table 7.1 the rapid convergence of the Ewald Sum approximation.[3] In

[3]The error is not measured exactly in the Y–norm: we calculate $\hat{G}^{\mathrm{ES}}_{j_o,k_o}$ at the nodes of the mesh and then evaluate the Y norm of the resulting interpolant.

p	$\|\hat{G}^{\rm ES}_{j_o=p,k_o=8} - \hat{G}^{\rm ES}_{j_o=8,k_o=8}\|_Y$	$\|\hat{G}^{\rm ES}_{j_o=8,k_o=p} - \hat{G}^{\rm ES}_{j_o=8,k_o=8}\|_Y$
1	2.3695	2.5631 E−4
2	3.6992 E−6	0
4	4.9873 E−19	0
6	0	0

Table 7.1: Convergence of the Ewald Sum $\hat{G}^{\rm ES}_{j_o,k_o}$.

particular, we take $G^{\rm ES}_{j_o=8,k_o=8}$ as the "truth," and present convergence in physical space, $G^{\rm ES}_{j_o=p,k_o=8}$, $1 \le p \le 8$, and in Fourier space $G^{\rm ES}_{j_o=8,k_o=p}$, $1 \le p \le 8$. Clearly, very small j_o, k_o suffices; and equally obvious, it is possible to choose a better tuning parameter γ.

Treatment of singular term

We now address the numerical quadrature issues related to the evaluation of functional involving the singular term — $a_2(w, G^{\rm ES}, v)$. We approximate $a_2(w, G^{\rm ES}, v)$ by

$$\begin{aligned} a_2(w, G^{\rm ES}, v) &= \int_\Omega w\, G^{\rm ES}\, v, \\ &= \int_\Omega w\, \hat{G}^{\rm ES}\, v + \int_\Omega \frac{w\, v}{|x|}, \\ &\approx \sum_{\rm quad} w(\cdot)\, \hat{G}^{\rm ES}(\cdot)\, v(\cdot) + \sum_{\rm quad} \frac{1}{|\cdot|} w(\cdot)\, v(\cdot), \end{aligned} \tag{7.24}$$

where $\sum_{\rm quad}$ denotes $Q \times Q \times Q$ tensorized Gauss-Legendre quadrature over each element $T_h \in \mathcal{T}_h$. Numerical tests confirm that the first term — very smooth — converges exponentially as Q increases; in contrast, the second term — singular — converges quite slowly. Though our choice $Q = 6$ probably suffices for our current study, clearly more "special purpose" quadratures must be developed for the singular term — at least in elements close to the nucleus (non-uniform meshes will also help in this regard).

From next section onward, we will drop the subscript h and superscript ES, and assume the finite element solution is our "truth" solution, i.e. Y, $\hat{\mathbf{u}}$, $\hat{\boldsymbol{\lambda}}$, ϕ and G refer to Y_h, $\hat{\mathbf{u}}_h$, $\hat{\boldsymbol{\lambda}}_h$, ϕ_h and G_h. We have used $\mathcal{N} = 20 \times 20 \times 20$.

7.3 Reduced-Basis Formulation

In Chapter 6, we showed that both reduced basis approximation — augmented and vectorized — can be applied successfully to Kohn Sham equations. Here, we consider only the vectorial reduced basis space, as we anticipate the resulting approximation space will be efficient. The augmented reduced basis approximation considered in Chapter 6 can be easily extended to the 3-dimensional Kohn Sham equations considered here.

7.3.1 The Approximation Space

We introduce nested sample sets $S^u_{N^u} = \{\mu^u_1, \ldots, \mu^u_{N^u}\}, 1 \leq N^u \leq N^u_{\max}$ and define the associated nested vectorial reduced-basis spaces as

$$W^u_{N^u} = \text{span } \{\hat{\mathbf{u}}([Z, \mu^u_n]), 1 \leq n \leq N^u\}, \quad 1 \leq N^u \leq N^u_{\max}, \tag{7.25}$$

$$= \text{span } \{\hat{\zeta}_n, 1 \leq n \leq N^u\}, \quad 1 \leq N^u \leq N^u_{\max}; \tag{7.26}$$

where $\hat{\mathbf{u}}([Z, \mu^u_n]) \equiv (u_1([Z, \mu^u_n]), \ldots, u_{n_e}([Z, \mu^u_n]))$ are the solutions of (7.10) at $\mu = \mu^u_n$ for a given Z; and $\hat{\zeta} \equiv (\zeta_1, \ldots, \zeta_{n_e})$ are basis functions obtained after $\hat{\mathbf{u}}([Z, \mu^u_n]), 1 \leq n \leq N^u$ are preprocessed — sorted, aligned and pseudo-orthogonalized for smaller N^u and better stability in the resulting discrete system. Then, an approximation of $\hat{\mathbf{u}}$ in W^u_N is given by $\hat{\mathbf{u}}_{N,M}([Z, \mu]) = \sum_{n=1}^{N^u} \psi_n([Z, \mu])\hat{\zeta}_n$ — the ith component of $\hat{\mathbf{u}}_{N,M}([Z, \mu])$ is given by $u_{N,M,i}([Z, \mu]) = \sum_{n=1}^{N^u} \psi_n([Z, \mu])\zeta_{n,i}$, $1 \leq i \leq n_e$. The preprocessing procedure outlined in Section 5.3.2 can be used. However, as $n_e \leq 2$, the eigenfunctions are well-behaved: there is only sign variation and the eigenvectors are well-separated for $\mathcal{D} \equiv [2, 4]$.

We may similarly define for ϕ the nested sample sets $S^\phi_{N^\phi} = \{\mu^\phi_1, \ldots, \mu^\phi_{N^\phi}\}$ and the associated reduced-basis space

$$W^\phi_{N^\phi} = \text{span } \{\phi([Z, \mu^\phi_n]), 1 \leq n \leq N^\phi\}, \quad 1 \leq N^\phi \leq N^\phi_{\max} \tag{7.27}$$

$$= \text{span } \{\chi_n, 1 \leq n \leq N^\phi\}, \quad 1 \leq N^\phi \leq N^\phi_{\max}.$$

The χ_n, $1 \leq n \leq N^\phi$ are obtained by orthonormalizing $\phi([Z, \mu^\phi_n])$, $1 \leq n \leq N^\phi$ relative to the $(\cdot; \cdot)_Y$ inner product.

Finally (although in actual practice, initially), we construct the $S^u_{N^u}$ and $S^\phi_{N^\phi}$ based on the greedy selection process described in Section 6.3.5.

7.3.2 The Approximation

This section follows closely Section 6.3.3. With vectorial reduced-basis approximation, the equilibrium ground state of the resulting neutral structure for a particular Z is given by $\hat{\mathbf{u}}_{N,M}([Z,\mu]) \equiv (u_{N,M,i}([Z,\mu]),\ 1 \le i \le n_e)$, where

$$\hat{\mathbf{u}}_{N,M}([Z,\mu]) = \arg\inf_{\hat{\mathbf{w}}} \Big\{ E_{N,M}(\hat{\mathbf{w}} \equiv (w_1,\ldots,w_{n_e});[Z,\mu]), w_i \in W^u_N, \quad \mu\int_\Omega w_i^2 = 1, 1 \le i \le n_e \Big\}, \tag{7.28}$$

$$\mu^*(Z) = \arg\inf_{\mu} \{ \mathcal{E}_{N,M}(\hat{\mathbf{u}}_{N,M}([Z,\mu]);[Z,\mu]);\mu > 0 \}. \tag{7.29}$$

Here, $\hat{\mathbf{u}}_{N,M}([Z,\mu])$ is obtained by solving the following Euler Lagrange equations: find $\mathbf{u}_{N,M}([Z,\mu]) \equiv (\hat{\mathbf{u}}_{N,M}([Z,\mu]),\ \phi_{N,M}([Z,\mu]),\ \hat{\boldsymbol{\lambda}}_{N,M}([Z,\mu]),\ \tau_{N,M}([Z,\mu])) \in \mathcal{Y}_N \equiv (W^u_{N^u} \times W^\phi_N \times \mathbb{R}^{n_e} \times \mathbb{R})$ such that

$$\mathcal{A}_M(\mathbf{u}_{N,M}([Z,\mu]),\mathbf{v};G;[Z,\mu]) = 0, \quad \forall \mathbf{v} \equiv (\hat{\mathbf{v}},\varsigma,\hat{\varphi},\varpi) \in \mathcal{Y}_N, \tag{7.30}$$

where $\hat{\boldsymbol{\lambda}}_{N,M}([Z,\mu]) \equiv (\lambda_{N,M,i}([Z,\mu]), 1 \le i \le n_e)$ and

$$\begin{aligned}
\mathcal{A}_M(\mathbf{w} \equiv (\hat{\mathbf{w}},s,\hat{\sigma},\kappa),\mathbf{v} \equiv (\hat{\mathbf{v}},\varsigma,\hat{\varphi},\varpi);t;[Z,\mu]) \equiv \\
\sum_{i=1}^{n_e} \Big[\theta_1([Z,\mu])\ a_0(w_i,v_i) + \ \theta_2([Z,\mu])\ a_2(w_i,t,v_i) + \theta_3([Z,\mu])\ a_2(w_i,s,v_i) \\
+ \theta_5([Z,\mu])\ a^{\mathrm{nl,M}}(w_i,\sum_{j=1}^{n_e} w_j^2,v_i) + 2\theta_4([Z,\mu])\ \sigma_i a_1(w_i,v_i) \Big] \\
+ \sum_{i=1}^{n_e} \varphi_i \{\beta_1([Z,\mu]) a_1(w_i,w_i) + \beta_2([Z,\mu])\} \\
+ \Big[\alpha_1([Z,\mu]) a_0(s,\varsigma) + \alpha_2([Z,\mu]) \sum_{j=1}^{n_e} a_2(w_j,w_j,\varsigma) + \alpha_3([Z,\mu]) l(\varsigma) + \kappa\alpha_4([Z,\mu])\ l(\varsigma) \Big] \\
+ \varpi\ l(s).
\end{aligned} \tag{7.31}$$

Then, $E_{N,M}(\mathbf{u}_{N,M}([Z,\mu]);G;[Z,\mu])$, the reduced-basis approximation for the electronic energy $E(\mathbf{u}([Z,\mu]);\,G;[Z,\mu])$, is given by

$$\begin{aligned} E_{N,M}(\mathbf{w}\equiv(\hat{\mathbf{w}},s,\hat{\sigma},\kappa);G;[Z,\mu]) \;=\; & \sum_{i=1}^{n_e}\Big[\theta_1([Z,\mu])a_0(w_i,w_i)+\theta_2([Z,\mu])\;a_2(w_i,G,w_i) \\ & +\frac{1}{2}\theta_3([Z,\mu])\;a_2(w_i,s,w_i) \\ & +\frac{3}{4}\theta_5([Z,\mu])\;a^{\mathrm{nl,M}}(w_i,\sum_{j=1}^{n_e}w_j^2,w_i)\Big]. \end{aligned} \tag{7.32}$$

and $\mathcal{E}_{N,M}(\mathbf{u}_{N,M}([Z,\mu]);G;[Z,\mu])$, the reduced-basis approximation to the total energy $\mathcal{E}(\mathbf{u}([Z,\mu]);$ $G;[Z,\mu])$, is given by $E_{N,M}(\mathbf{u}_{N,M}([Z,\mu]);G;[Z,\mu])+\frac{Z^2}{2}\eta(\mu)$.

Compared to (7.1), we have made two approximations in (7.31). First, we approximate $a^{\mathrm{nl}}(w_i,$ $\sum_{j=1}^{n_e}w_j^2,v_i)$ by $a^{\mathrm{nl,M}}(w_i,\sum_{j=1}^{n_e}w_j^2,v_i)\equiv\int_\Omega g^w_{M^{g_i},i}\,v$, where $g^u_{M^{g_i},i}$ is an empirical interpolation approximation to $g_i(\hat{\mathbf{u}})\equiv u_i(\sum_{j=1}^{n_e}u_j^2)^{1/3}$. Thus, we require n_e empirical interpolation approximations and M^{g_i} is the dimension of the approximation space for $W^{g_i}_{M^{g_i}}$. We also define M as $\max_{1\le i\le n_e}M^{g_i}$. Second, we only impose the constraints $\mu\int_\Omega u^2_{N,M,i}=1$, $1\le i\le n_e$. Finally, since $\int_\Omega\chi_n=0$, $1\le n\le N$, $\int_\Omega\phi_{N,M}$ is perforce zero; our discrete (nonlinear) algebraic system will thus have an actual dimension of $N^u+N^\phi+n_e$. This can then be solved based on the Newton iterative procedure outlined in Chapter 4 with an online complexity of $O((N^\phi)^3+n_eN^\phi(N^u)^2+(N^u)^2+n_eN^uM)$ per Newton iteration. Lastly, the online-offline computational procedure can be readily applied.

7.4 Numerical Results

Here, the largest problem we could solve (using MATLAB) is for $Z=4$; this corresponds to a simple cubic structure for Beryllium. Note that, as we have used a restricted model, two electrons will occupy a single energy level; consequently we only solve for two eigenfunctions — $n_e=2$.

We consider $\mu\in\mathcal{D}\equiv[2,4]$ and a parameter test sample Ξ_T of size 80 with sample points distributed uniformly in $\mathcal{D}$. We choose M^{g_i}, such that the approximation error in $g^u_{M^{g_i}}$ is less than

N^u	$\varepsilon^u_{N,M}$	$\varepsilon^{\mathcal{E}}_{N,M}$	$\varepsilon^{\phi}_{N,M}$	$\varepsilon^{\text{ortho}}_{N,M}$
2	4.6855E-1	8.0765E-2	3.0058E-1	8.1298E-8
3	1.1070E-1	1.0543E-2	5.5140E-2	2.6072E-2
4	2.7217E-2	1.1797E-3	1.6030E-2	2.4393E-2
5	3.5961E-4	1.1086E-6	7.1304E-5	2.0356E-3
6	1.6391E-4	2.1172E-7	2.8961E-5	1.1883E-3
7	1.2049E-4	1.5489E-7	2.0473E-5	9.3882E-4
8	2.6687E-5	4.4191E-8	4.1551E-6	3.8011E-4
9	1.1542E-6	1.5024E-8	2.7681E-7	1.0560E-4

Table 7.2: Convergence of the reduced basis errors — $\varepsilon^u_{N,M}$, $\varepsilon^{\mathcal{E}}_{N,M}$, $\varepsilon^{\phi}_{N,M}$, and $\varepsilon^{\text{ortho}}_{N,M}$ — with N^u for $n_e = 2$ and $2 \le \mu \le 4$.

10^{-10} for all i. We then define the following

$$\varepsilon^u_{N,M} = \max_{\mu\in\Xi_T} \frac{(\sum_{i=1}^{n_e} \|u_i([Z,\mu]) - u_{N,M,i}([Z,\mu])\|_Y^2)^{1/2}}{(\sum_{i=1}^{n_e} \|u_i([Z,\mu])\|_Y^2)^{1/2}}, \tag{7.33}$$

$$\varepsilon^{\phi}_{N,M} = \max_{\mu\in\Xi_T} \frac{\|\phi([Z,\mu]) - \phi_{N,M}([Z,\mu])\|_Y}{\|\phi([Z,\mu])\|_Y}, \tag{7.34}$$

$$\varepsilon^{\mathcal{E}}_{N,M} = \max_{\mu\in\Xi_T} \frac{|\mathcal{E}_N([Z,\mu]) - \mathcal{E}([Z,\mu])|}{|\mathcal{E}([Z,\mu])|} \tag{7.35}$$

$$\varepsilon^{\text{ortho}}_{N,M} = \max_{\mu\in\Xi_T} \max_{1\le i<j\le n_e} \int_\Omega u_{N,M,i}([Z,\mu])\, u_{N,M,j}([Z,\mu]). \tag{7.36}$$

where $\varepsilon^u_{N,M}$, $\varepsilon^{\phi}_{N,M}$ and $\varepsilon^{\mathcal{E}}_{N,M}$ are respectively the maximum error in the reduced-basis approximation of $\hat{\mathbf{u}}$, ϕ and $\mathcal{E}$ within a given sample Ξ_T; and $\varepsilon^{\text{ortho}}_{N,M}$ is a measure of non-compliance in the orthogonality constraints.

We again achieve a convergence results not unlike that obtained for the 1-dimensional model of Chapter 6: as shown in Table 7.2 and for $N^\phi = 7$ and $M = 13$, the errors in $\hat{\mathbf{u}}$ again decrease monotonically and the orthogonality of the solution is increasingly satisfied as N increases. However, we note that the computational savings achieved at the *online* stage is significantly higher for a 3-dimensional problem. For our finite element approximation with $\mathcal{N} = 8000$, the computational cost is 6576.45s. On the other hand, the computational cost during the online stage for $|\mathcal{E}_N([Z,\mu]) - \mathcal{E}([Z,\mu])|/|\mathcal{E}([Z,\mu])| \le 1\text{E}{-6}$ is 8s. We do however admit that finite element method may not be the most appropriate approximation method to use for this particular problem — planewave method will most likely converge faster and admit more efficient solution procedures.

7.5 Application

As an illustrative application, we shall again determine μ^*, μ at which the total energy is minimum, for $Z = 2n_e = 4$. We require an absolute accuracy of 10^{-5} in μ^*. We use the simple bisection method used in Section 6.5; it will thus involve repetitive evaluation of $\mathcal{E}([Z,\mu])$. The approach is straightforward. We approximate $\mathcal{E}([Z,\mu])$ by $\mathcal{E}_{N,M}([Z,\mu])$ based on our vectorial reduced basis approximation. We have use $N^u = 9$, $N^\phi = 7$ and $M = 13$. The variation of $\mathcal{E}([Z,\mu])$ and $\mathcal{E}_{N,M}([Z,\mu])$ with μ is shown in Figure 7-1.

A total of 20 evaluations of $\mathcal{E}_{N,M}([Z,\mu])$ is required and μ^* is found to be 2.9375. The total online computational cost is only 18.15s. On the contrary, with finite element approximation of $\mathcal{N} = 8000$, the total computational cost would have taken more than 36 hours[4]. If we compare only the online cost with the total computational cost based on "truth" approximation, we clearly achieve online computational savings of order $O(10^3)$.

However, taking the offline computational cost into consideration, the reduced basis method is clearly not competitive for this particular problem. Due to the lack of an efficient *a posteriori* error estimation procedure, we must first compute solutions at all $\mu \in \Xi_T$ — for our example, there are 80 sample points in Ξ_T, leading to a total offline computational cost of approximately 146 hours. Clearly, if the reduced basis model that we have developed is only used for this particular problem, it would have made no sense. Of course, if we could construct a sample set without the use of the "greedy" sampling procedure, the offline computational cost is then manageable — it would also mean we may have a suboptimal approximation with a dimension that is larger than is needed, or worse, we do not know how accurate the approximation is.

However, the low order model obtained through the reduced basis approximation can be reused to determine other static and dynamic properties of the crystal, as suggested in Section 6.5. For example, we can use the current approximation can be used to determine elasticity constants and the bulk modulus [120]. With some extensions to the current model, we can further perform lattice dynamics simulations based on the Density Functional Perturbation Theory [7, 42] or the frozen phonon approach [28, 76]. Under such circumstances, we may be able to justify the large initial computational overhead incurred during the offline stage.

[4]As mentioned in Section 7.4, the use of finite element method for this problem is clearly not optimal. We could perhaps obtain a much lower computational cost by, for example, the planewave method. The online computational saving achieved by the reduced basis method is then much lower.

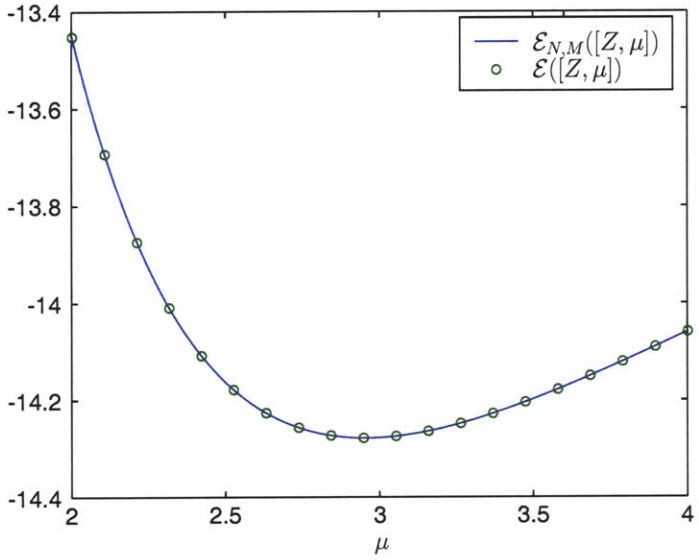

Figure 7-1: Comparison between the $\mathcal{E}([Z,\mu])$ and $\mathcal{E}_{N,M}([Z,\mu])$ for $Z=4$, $n_e=2$, and $2\leq\mu\leq 4$.

Chapter 8

Concluding Remarks

8.1 Summary

The main goal of this thesis is to develop reduced basis methods for eigenvalue problems encountered in computational chemistry. The essential components of the reduced basis method are (*i*) rapidly convergent global reduced basis approximation spaces generated by Galerkin projection onto a space spanned by solutions of the governing partial differential equations at N judiciously chosen samples in parameter space such that N, the dimension of the resulting reduced order model, is much smaller than $\mathcal{N}$, the dimension of the underlying discretization of the "truth" approximation; (*ii*) *a posteriori* error estimators to provide inexpensive estimation for the errors in the outputs of interest; (*iii*) an offline-online computational procedure to decouple the generation and projection stages of the approximation process; and (*iv*) optimal sampling strategies to pick parameter samples optimally.

The concept of constructing efficient approximation spaces that are problem-specific is not new in computational chemistry — chemists have developed numerous highly efficient basis sets for electronic structure calculations of a wide variety of molecular systems [35, 121]. However, these basis sets were obtained at a price — considerable effort was expended to optimize a particular basis set for a particular system with respect to some experimental data [31, 118, 119]. In addition, no such efficient basis set is available for extended systems. As such, we believe the reduced basis method provides a systematic approach by which efficient basis sets can be obtained for any system with the added advantage of good convergence properties.

We extend earlier work on reduced basis methods for linear eigenvalue problems [71]. We consider eigensolutions of vectorial nature, i.e. solution with multiple eigenvectors and eigenvalues. In particular, we introduce two approximation approaches — the augmented reduced basis approximation and the vectorial reduced basis approximation. The augmented reduced basis approximation relies on optimality of the Galerkin method to find the best linear combination of the basis functions in W_N^A, the augmented reduced basis space. Since W_N^A consists of all eigenvectors at all the sample points in the associated sample set, the resulting approximation space is in fact very rich; its dimension is equally large. On the other hand, the vectorial reduced basis approximation reduces the degree of freedoms available to the Galerkin procedure by first preprocessing the basis functions in the vectorial reduced basis space, W_N^V. The efficiency of W_N^V is contingent on the ability of the preprocessing step to obtain a smooth variation of the eigenvectors with respect to our input parameter. In our first example based on a harmonic oscillator problem, we show that the vectorial reduced basis approximation can indeed be very efficient compared to the augmented reduced basis approximation. We further equip our reduced basis approximations with efficient asymptotic *a posteriori* error estimation procedures. This enables us to control the accuracy of the approximation, and provide an inexpensive guide to efficiently construct an optimal reduced basis space based on the "greedy" adaptive sampling procedure.

In Chapter 5, we apply the reduced basis method developed for linear eigenvalue problem to band structure calculation — the rapid determination of band energies $E_i(\mathbf{k})$, $1 \leq i \leq n_b$, given any $\mathbf{k}$ in the first Brillouin zone for a periodic Hamiltonian operator with fixed background periodic potential. This allows the rapid determination of band structure properties required in the study of transport phenomena and the determination of macroscopic properties. Due to the rich variation of the solutions $\hat{\mathbf{u}}(\mathbf{k})$ with respect to the parameter $\mathbf{k}$, the augmented reduced basis approximation performs much better than the vectorial reduced basis approximation as it provides more degree of freedom to achieve the Galerkin optimality. The augmented reduced basis space proves to be more efficient than the vectorial reduced basis space. We demonstrate the utility of the reduced basis approach in the determination of spectral properties of crystalline silicon based on the empirical pseudopotential model.

In computational chemistry problems, nonlinear eigenvalue problems are commonly encountered [18, 32, 90, 111]. For reduced basis approximations of nonlinear eigenvalue problems, we first

describe the empirical interpolation method — a rapidly convergent interpolation procedure for parametric fields — in Chapter 3. We then show in Chapter 4 how the incorporation of the empirical interpolation method within our reduced basis framework can lead to an online computation where the complexity is independent of $\mathcal{N}$.

In Chapter 6 and 7, we consider the rapid determination of the ground state energy, $\mathcal{E}(\mu)$ of a crystal structure based on the Density Functional Theory. A significant improvement in the efficiency of such calculations has important implications in *ab-initio* geometry optimization of molecular systems and multiscale, multiphysics simulations. To achieve this, we seek an efficient reduced basis approximation of the associated Kohn-Sham equations. The equations are first parameterized by mapping the solutions onto a fixed reference domain with positions of nuclei mapped onto unique locations in the reference domain; we consider only cases where positions of nuclei does not vary. We render the nonlinear functions affine by using the empirical interpolation method. We further show that for the cases examined, the vectorial reduced basis approximation can be particularly efficient — it exploits the inherent orthogonality properties between the solutions $u_i(\mu)$, $1 \leq i \leq n_e$, and their common smoothness to achieve a significant reduction in the dimension of the resulting algebraic equations. The results based on a one dimensional periodic problem indicate that the reduced-basis space is rapidly convergent with N and depend weakly on n_e. The energy $\mathcal{E}(\mu)$ can also be easily approximated to a relative accuracy of 10^{-8}. For three dimensional problems, the result is more limited due to limitation in the solver used for obtaining the truth approximation. We show some results for a simple cubic structure of Beryllium.

8.2 Future Work

We conclude by proposing possible future work. We divide the discussion into two: the first section look at the possible numerical improvements while the second suggests possible avenues by which the reduced basis method can be more widely applied in "real" computational chemistry calculations.

8.2.1 Numerical Improvement

An important ingredient of the reduced basis framework is the efficient *a posteriori* error estimation procedure. We currently do not have an error estimation procedure for nonlinear eigenvalue prob-

lem. This leads to a very expensive offline procedure since we need to compute solutions for all the parameters in our test sample. For the linear eigenvalue problem, the error estimators developed in this thesis are also not rigorous bounds for our quantities of interest — they cannot be used as certificates of fidelity. Development of efficient, sharp and rigorous error bounds for reduced basis approximations of linear and nonlinear eigenvalue problems will greatly encourage wider adoption of the reduced basis method.

For our vectorial reduced basis approximation, the solution method at the online stage — a Newton iterative scheme — is currently not robust, especially if N is small. Its efficiency also pales against existing eigenvalue solver, even when compared to system of larger size. A more robust and efficient solution method would greatly increase the appeal of vectorial reduced basis approximation.

For many computational chemistry problems, the quantities of interest can usually be expressed as functions of the projector onto the eigenvectors associated with, say, the n_e lowest eigenvalues. Instead of constructing a reduced basis approximation for the n_e eigenpairs, we can construct a reduced basis approximation for the projector. The latter is a single object that is more well-defined. Therefore, efforts in this direction may yet yield an efficient approach in solving computational chemistry problems. At present however, it is not clear how we can construct *a posteriori* estimation procedure for reduced basis formulation based on projectors — current formulation of the error estimation procedure requires knowledge of the eigenpairs.

8.2.2 Applications

Computational chemistry problems are many and varied. We limit the discussion to possible extensions to the two applications considered in this thesis.

Extensions to band structure calculations

First, a numerical comparison between the Slater-Koster interpolation scheme based on Wannier functions [79, 109, 110] and the current reduced basis approach will allow us to better identify future applications for the reduced basis method. In addition, the Slater-Koster interpolation scheme is well-established in the solid states community and as such benchmarking our method with respect to this scheme may encourage wider adoption of the reduced basis approach.

In *ab initio* calculations based on Density Functional Theory, the density of electrons are usually determined from the wavefunctions evaluated at several **k**-points in the first Brillouin zone — for some elements, such as metals, the number of **k**-points required can be large. Coupling this with the need to compute a new density for each iteration of the SCF algorithm, the computational cost can be large — $O(n_k \mathcal{N}^\bullet)$ where n_k is the number of **k**-points needed to accurately evaluate the density. We consider possible computational savings based on reduced basis approach. In each SCF iteration, we construct a reduced basis approximation of dimension $N \ll \mathcal{N}$. Then, the reduced basis approximant to the wavefunctions at all n_k are used to construct the density of electrons — the computational cost is then $O(N\mathcal{N}^\bullet + n_k N^3)$. For large n_k, this can potentially lead to large computational savings.

Extensions to ground state energy calculations

For our three dimensional problems, we are limited by the size of Z that we can treat. For large nuclear charge, the cusps at the nuclear positions become very pronounced and thus very fine resolution is required, leading to exorbitant computational costs, not to mention the complexity involved in writing an efficient code for determining the "truth" approximations. One alternative is to couple the reduced basis approximation with existing electronic structure codes, which have been optimized by many researchers. There are however several challenges associated with this. First, as with any large-scale code, knowledge on the implementation details of these codes is crucial in developing a reduced basis approximation that is consistent with the "truth" approximation computed from these codes. Second, most of these codes utilize pseudopotential to reduce the effects of the cusps mentioned earlier. However, the current formulation in these codes are not amenable to the reduced basis treatment. Lastly, due to the symmetric properties of crystalline solids, common practice of symmetrizing the density with respect to some symmetry point groups also leads to some difficulties in developing efficient empirical interpolants for the nonlinear terms.

Implementation challenges aside, we wish to examine the formulation issues related to parameterization of problems with moving nuclei. One possible solution is through nonlinear geometric mapping of (moving) locations of nuclei in the original domain onto fixed locations in the reference domain. This will lead to nonaffine functionals, which can then be rendered affine using empirical interpolation method. Note, however, this significantly increases the complexity of the approxi-

mation process since, due to the nonlinear mapping, all functionals are now at least nonaffine. In addition, it is also limited to (*i*) small displacements as large displacement will lead to conditioning issues, and (*ii*) small number of nuclei as the nonlinearity of the mapping functions grows with the number of nuclei. Extension of the reduced basis approach to cases with moving nuclei will be of particular interest in computational chemistry.

Bibliography

[1] M. Ainsworth and J. T. Oden. *A posteriori error estimation in finite analysis.* Wiley, New York, 2000.

[2] B. O. Almroth, P. Stern, and F. A. Brogan. Automatic choice of global shape functions in structural analysis. *AIAA Journal*, 16:525–528, 1978.

[3] N. W. Ashcroft and N. D. Mermin. *Solid State Physics.* Saunders College, Philadelphia, 1976.

[4] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst. *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide.* SIAM, Philadelphia, 2000.

[5] A. Baldereschi. Mean-value point in the brillouin zone. *Phys. Rev. B*, 7(12):5212–5215, Jun 1973.

[6] S. Baroni, A. D. Corso, S. de Gironcoli, P. Giannozzi, C. Cavazzoni, G. Ballabio, S. Scandolo, G. Chiarotti, P. Focher, A. Pasquarello, K. Laasonen, A. Trave, R. Car, N. Marzari, and A. Kokalj. http://www.pwscf.org.

[7] S. Baroni, S. de Gironcoli, A. Dal Corso, and P. Giannozzi. Phonons and related crystal properties from density-functional perturbation theory. *Rev. Mod. Phys.*, 73(2):515–562, Jul 2001.

[8] 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.*, 339:667–672, 2004.

[9] A. Barrett and G. Reddien. On the reduced basis method. *Z. Angew. Math. Mech.*, 75(7):543–549, 1995.

[10] P. Bettess and R. A. Abram. Finite and infinite elements for a simple problem in quantum mechanics. *Communications in Numerical Methods in Engineering*, 18(5):325–334, 2002.

[11] X. Blanc. A mathematical insight into ab initio simulations of the solid phase. In *Mathematical models and methods for ab initio quantum chemistry*, volume 74 of *Lecture Notes in Chem.*, pages 133–158. Springer, Berlin, 2000.

[12] X. Blanc. Geometry optimization for crystals in Thomas-Fermi type theories of solids. *Comm. Partial Differential Equations*, 26(3-4):651–696, 2001.

[13] X. Blanc and C. Le Bris. Periodicity of the infinite-volume ground state of a one-dimensional quantum model. *Nonlinear Anal.*, 48(6):791–803, 2002.

[14] P. E. Blöchl, O. Jepsen, and O. K. Andersen. Improved tetrahedron method for brillouin-zone integrations. *Phys. Rev. B*, 49(23):16223–16233, Jun 1994.

[15] O. Bokanowski and N. J. Mauser. Local approximation for the Hartree-Fock exchange potential: a deformation approach. *Math. Models Methods Appl. Sci.*, 9(6):941–961, 1999.

[16] A. Buffa, Y. Maday, A. T. Patera, C. Prud'homme, and G. Turinici. *A Priori* convergence of multi-dimensional parametrized reduced–basis approximations. 2007. In progress.

[17] E. Cancès. Self-consistent field algorithms for kohn–sham models with fractional occupation numbers. *The Journal of Chemical Physics*, 114(24):10616–10622, 2001.

[18] E. Cancès, M. Defranceschi, W. Kutzelnigg, C. Le Bris, and Y. Maday. Computational quantum chemistry: a primer. In C. Le Bris, editor, *Handbook of numerical analysis, Vol. X*, Special Volume: Computational Chemistry, pages 3–270. North-Holland, Amsterdam, 2003.

[19] E. Cancès and C. Le Bris. Can we outperform the DIIS approach for electronic structure calculations? *International Journal of Quantum Chemistry*, 79(2):82–90, 2000.

[20] E. Cancès and C. Le Bris. On the convergence of SCF algorithms for the Hartree-Fock equations. *M2AN Math. Model. Numer. Anal.*, 34(4):749–774, 2000.

[21] E. Cancès, C. Le Bris, Y. Maday, and G. Turinici. Towards reduced basis approaches in ab initio electronic structure computations. In *Proceedings of the Fifth International Conference on Spectral and High Order Methods (ICOSAHOM-01) (Uppsala)*, volume 17, pages 461–469, 2002.

[22] E. Cancès, C. Le Bris, N. C. Nguyen, Y. Maday, A. T. Patera, and G. S. H. Pau. Feasibility and competitiveness of a reduced basis approach for rapid electronic structure calculations in quantum chemistry. In *Proceedings of the Workshop for High-dimensional Partial Differential Equations in Science and Engineering (Montreal)*, 2007. To be published.

[23] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang. *Spectral Methods in Fluid Dynamics*. Springer, New York, 1987.

[24] K. M. Carling and E. A. Carter. Orbital-free density functional theory calculations of the properties of Al, Mg and Al–Mg crystalline phases. *Modelling and Simulation in Materials Science and Engineering*, 11(3):339–348, 2003.

[25] I. Catto, C. Le Bris, and P.-L. Lions. Recent mathematical results on the quantum modeling of crystals. In *Mathematical models and methods for ab initio quantum chemistry*, volume 74 of *Lecture Notes in Chem.*, pages 95–119. Springer, Berlin, 2000.

[26] D. J. Chadi and M. L. Cohen. Electronic structure of $Hg_{1-x}Cd_xTe$ alloys and charge-density calculations using representative k points. *Phys. Rev. B*, 7(2):692–699, Jan 1973.

[27] D. J. Chadi and M. L. Cohen. Special points in the brillouin zone. *Phys. Rev. B*, 8(12):5747–5753, Dec 1973.

[28] D. J. Chadi and R. M. Martin. Calculation of lattice dynamical properties from electronic energies: Application to C, Si and Ge. *Solid State Communications*, 19(7):643–646, July 1976.

[29] M. L. Cohen and T. K. Bergstresser. Band structures and pseudopotential form factors for fourteen semiconductors of the diamond and zinc-blende structures. *Phys. Rev.*, 141(2):789–796, Jan 1966.

[30] M. Côté, P. D. Haynes, and C. Molteni. Material design from first principles: the case of boron nitride polymers. *Journal of Physics: Condensed Matter*, 14(42):9997–10009, 2002.

[31] E. R. Davidson and D. Feller. Basis set selection for molecular calculations. *Chem. Rev.*, 86:681 – 696, 1986.

[32] R. M. Dreizler and E. K. U. Gross. *Density Functional Theory: An Approach to the Quantum Many-Body Problem.* Springer, New York, 1991.

[33] J.-L. Fattebert and M. Buongiorno Nardelli. Finite difference methods for ab initio electronic structure and quantum transport calculations of nanostructures. In C. Le Bris, editor, *Handbook of numerical analysis, Vol. X*, Special Volume: Computational Chemistry, pages 571–612. North-Holland, Amsterdam, 2003.

[34] J. L. Fattebert, R. D. Hornung, and A. M. Wissink. Finite element approach for density functional theory calculations on locally-refined meshes. *J. Comput. Phys.*, 223(2):759–773, 2007.

[35] D. Feller and E. R. Davidson. Basis sets for ab initio molecular orbital calculations and intermolecular interactions. In *Reviews in Computational Chemistry*, pages 1–43. VCH, New York, 1990.

[36] 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(1):21–28, 1983.

[37] F. Finocchi, J. Goniakowski, X. Gonze, and C. Pisani. An introduction to first-principles simulations of extended systems. In C. Le Bris, editor, *Handbook of numerical analysis, Vol. X*, Special Volume: Computational Chemistry, pages 377–451. North-Holland, Amsterdam, 2003.

[38] M. V. Fischetti and S. E. Laux. Monte carlo analysis of electron transport in small semiconductor devices including band-structure and space-charge effects. *Phys. Rev. B*, 38(14):9721–9745, Nov 1988.

[39] C. J. García-Cervera. An efficient real space method for orbital-free density-functional theory. *Commun. Comput. Phys.*, 2(2):334–357, April 2007.

[40] G. Gilat. Analysis of methods for calculating spectral properties in solids. *Journal of Computational Physics*, 10(3):432–465, Dec 1972.

[41] G. Gilat and L. J. Raubenheimer. Accurate numerical method for calculating frequency-distribution functions in solids. *Phys. Rev.*, 144(2):390–395, Apr 1966.

[42] X. Gonze. Adiabatic density-functional perturbation theory. *Phys. Rev. A*, 52(2):1096–1114, Aug 1995.

[43] X. Gonze, J. M. Beuken, R. Caracas, F. Detraux, M. Fuchs, G. M. Rignanese, L. Sindic, M. Verstraete, G. Zerah, and F. Jollet. First-principles computation of material properties: the ABINIT software project. *Computational Materials Science*, 25(3):478–492, 2002.

[44] M. A. Grepl, Y. Maday, N. C. Nguyen, and A. T. Patera. Efficient reduced-basis treatment of nonaffine and nonlinear partial differential equations. *M2AN (Math. Model. Numer. Anal.)*, 2007.

[45] M. A. Grepl, N. C. Nguyen, K. Veroy, A. T. Patera, and G. R. Liu. Certified rapid solution of partial differential equations for real-time parameter estimation and optimization. In *Proceedings of the 2nd Sandia Workshop of PDE-Constrained Optimization: Towards Real-Time PDE-Constrained Optimization*, SIAM Computational Science and Engineering Book Series, 2007. In press.

[46] G. Grosso and G. P. Parravicini. *Solid State Physics.* Academic Press, Amsterdam, 2000.

[47] M. D. Gunzburger. *Finite Element Methods for Viscous Incompressible Flows.* Academic Press, 1989.

[48] F. Gygi. Large-scale first-principles molecular dynamics: moving from terascale to petascale computing. *Journal of Physics: Conference Series*, 46:268–277, 2006.

[49] J. Hama and M. Watanabe. General formulae for the special points and their weighting factors in k-space integration. *Journal of Physics: Condensed Matter*, 4(19):4583–4594, 1992.

[50] S. Hamel, A. Williamson, H. Wilson, F. Gigy, G. Galli, E. Ratner, and D. Wack. First Principles Computation of Optical Response in Silicon Nanostructures. *APS Meeting Abstracts*, page 43007, Mar. 2007.

[51] E. Isaacson and H. B. Keller. *Computation of Eigenvalues and Eigenvectors, Analysis of Numerical Methods.* Dover, New York, 1994.

[52] K. Ito and S. S. Ravindran. A reduced basis method for control problems governed by PDEs. In W. Desch, F. Kappel, and K. Kunisch, editors, *Control and Estimation of Distributed Parameter Systems*, pages 153–168. Birkhäuser, 1998.

[53] 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, 1998.

[54] 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.

[55] K. Ito and J. D. Schroeter. Reduced order feedback synthesis for viscous incompressible flows. *Mathematical And Computer Modelling*, 33(1-3):173–192, 2001.

[56] A. D. Izaak. Kolmogorov widths in finite-dimensional spaces with mixed norms. *Mathematical Notes*, 55(1):43–52, 1994.

[57] C. Jacoboni and P. Lugli. *The Monte Carlo Method for Semiconductor Device Simulation.* Springer, Wein, 1989.

[58] O. Jepson and O. K. Anderson. The electronic structure of h.c.p. Ytterbium. *Solid State Communications*, 9(20):1763–1767, Oct 1971.

[59] J. D. Joannopoulos, R. D. Meade, and J. N. Winn. *Photonic Crystals.* Princeton University Press, 1995.

[60] E. Kaxiras. *Atomic and Electronic Structure of Solids.* Cambridge University Press, Cambridge, 2003.

[61] C. Kittel. *Introduction to Solid State Physics.* Wiley, 1995.

[62] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Phys. Rev.*, 140(4A):A1133–A1138, Nov 1965.

[63] S. N. Kudryavtsev. Widths of classes of finitively smooth functions in Sobolev spaces. *Mathematical Notes*, 77(4):535–539, 2005.

[64] C. Le Bris. Computational chemistry from the perspective of numerical analysis. *Acta Numer.*, 14:363–444, 2005.

[65] C. Le Bris. *Private Communication.* Montreal, 2005.

[66] C. Le Bris and P.-L. Lions. From atoms to crystals: a mathematical journey. *Bull. Amer. Math. Soc. (N.S.)*, 42(3):291–363, 2005.

[67] M. Y. L. Lee. Estimation of the error in the reduced-basis method solution of differential algebraic equations. *SIAM Journal of Numerical Analysis*, 28:512–528, 1991.

[68] G. Lehmann and M. Taut. Numerical calculation of the density of states and related properties. *Physica Status Solidi (b)*, 54(2):469–477, Dec 1972.

[69] R. B. Lehoucq, D. C. Sorensen, and C. Yang. *ARPACK Users' Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods.* SIAM, Philadelphia, 1998.

[70] E. H. Lieb and B. Simon. The Thomas-Fermi theory of atoms, molecules and solids. *Advances in Math.*, 23(1):22–116, 1977.

[71] L. Machiels, Y. Maday, I. B. Oliveira, A. Patera, and D. Rovas. Output bounds for reduced-basis approximations of symmetric positive definite eigenvalue problems. *C. R. Acad. Sci. Paris, Série I*, 331(2):153–158, 2000.

[72] Y. Maday. Reduced–basis method for the rapid and reliable solution of partial differential equations. In *Proceedings of International Conference of Mathematicians, Madrid.* European Mathematical Society Eds., 2006.

[73] Y. Maday, N. C. Nguyen, A. T. Patera, and G. S. H. Pau. A general, multipurpose interpolation procedure: the magic points. in progress. 2007.

[74] Y. Maday, A. 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, 2002.

[75] 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.

[76] R. M. Martin. *Electronic Structure: basic theory and practical methods.* Cambridge University Press, 2004.

[77] D. Marx and J. Hutter. Ab initio molecular dynamics: Theory and implementation. In *Modern Methods and Algorithms of Quantum Chemistry*, pages 301–449. NIC, FZ Jülich, 2000.

[78] N. Marzari. *Private Communication.* MIT, 2007.

[79] N. Marzari and D. Vanderbilt. Maximally localized generalized wannier functions for composite energy bands. *Phys. Rev. B*, 56(20):12847–12865, Nov 1997.

[80] R. E. Miller and E. B. Tadmor. The quasicontinuum method: Overview, applications and current directions. *Journal of Computer-Aided Materials Design*, 9:203–239, 2002.

[81] H. J. Monkhorst and J. D. Pack. Special points for Brillouin zone integrations. *Phys. Rev. B*, 13(12):5188–5192, Jun 1976.

[82] N. C. Nguyen, K. Veroy, and A. T. Patera. Certified real-time solution of parametrized partial differential equations. In S. Yip, editor, *Handbook of Materials Modeling*, pages 1523–1558. Springer, 2005.

[83] A. K. Noor and I. Babuška. Quality assessment and control of finite element solutions. *Finite Elements in Analysis and Design*, 3(1):1–26, April 1987.

[84] 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.

[85] A. K. Noor and J. M. Peters. Reduced basis technique for nonlinear analysis of structures. *AIAA Journal*, 18(4):455–462, 1980.

[86] A. K. Noor, J. M. Peters, and C. M. Andersen. Mixed models and reduction techniques for large-rotation nonlinear problems. *Comp. Meth. Appl. Mech. Engrg.*, 44:67–89, 1984.

[87] J. T. Oden and L. F. Demkowicz. *Functional Analysis.* CRC Press, Boca Raton, 1996.

[88] G. B. Olson. Designing a New Material World . *Science*, 288(5468):993–998, 2000.

[89] B. N. Parlett. *The Symmetric Eigenvalue Problem.* Society for Industrial and Applied Mathematics, Philadelphia, 1998.

[90] R. G. Parr and W. Yang. *Density-Functional Theory of Atoms and Molecules (International Series of Monographs on Chemistry).* Oxford University Press, 1994.

[91] J. E. Pask and P. A. Sterne. TOPICAL REVIEW: Finite element methods in ab initio electronic structure calculations. *Modelling Simul. Mater. Sci. Eng.*, 13:71–96, Apr. 2005.

[92] A. Patera and G. Rozza. *Reduced Basis Approximation and A Posteriori Error Estimation for Parametrized Partial Differential Equations, Version 1.0 to appear in (tentative rubric) MIT Pappalardo Graduate Monographs in Mechanical Engineering.* MIT, 2007.

[93] M. C. Payne, M. P. Teter, D. C. Allan, T. A. Arias, and J. D. Joannopoulos. Iterative minimization techniques for ab initio total-energy calculations: molecular dynamics and conjugate gradients. *Rev. Mod. Phys.*, 64(4):1045–1097, Oct 1992.

[94] J. S. Peterson. The reduced basis method for incompressible viscous flow calculations. *SIAM J. Sci. Stat. Comput.*, 10(4):777–786, 1989.

[95] G. A. Petersson, S. Zhong, J. John A. Montgomery, and M. J. Frisch. On the optimization of gaussian basis sets. *The Journal of Chemical Physics*, 118(3):1101–1109, 2003.

[96] A. Pinkus. *n-Widths in Approximation Theory.* Springer, 1985.

[97] T. A. Porsching. Estimation of the error in the reduced basis method solution of nonlinear equations. *Mathematics of Computation*, 45(172):487–496, 1985.

[98] T. A. Porsching and M. Y. L. Lee. The reduced-basis method for initial value problems. *SIAM Journal of Numerical Analysis*, 24:1277–1287, 1987.

[99] C. Prud'homme, D. Rovas, K. Veroy, Y. Maday, A. Patera, and G. Turinici. Reliable real-time solution of parametrized partial differential equations: Reduced-basis output bounds methods. *Journal of Fluids Engineering*, 124(1):70–80, 2002.

[100] P. Pulay. Improved SCF convergence acceleration. *Journal of Computational Chemistry*, 3(4):556–560, 1982.

[101] A. Quarteroni, R. Sacco, and F. Saleri. *Numerical Mathematics*, volume 37 of *Texts in Applied Mathematics*. Springer, New York, 2000.

[102] C. Radin. Periodicity of classical ground states. *Phys. Rev. Lett.*, 51(8):621–622, Aug 1983.

[103] M. Reed and B. Simon. *Methods of modern mathematical physics. IV. Analysis of operators.* Academic Press [Harcourt Brace Jovanovich Publishers], New York, 1978.

[104] W. C. Rheinboldt. Numerical analysis of continuation methods for nonlinear structural problems. *Computers and Structures*, 13(1-3):103–113, 1981.

[105] W. C. Rheinboldt. On the theory and error estimation of the reduced basis method for multi-parameter problems. *Nonlinear Analysis, Theory, Methods and Applications*, 21(11):849–858, 1993.

[106] C. C. J. Roothaan. New developments in molecular orbital theory. *Rev. Mod. Phys.*, 23(2):69–89, Apr 1951.

[107] V. R. Saunders and I. H. Hillier. A level-shifting method for converging closed shell hartree-fock wave functions. *International Journal of Quantum Chemistry*, 7(4):699–705, 1973.

[108] J. C. Slater. A simplification of the hartree-fock method. *Phys. Rev.*, 81(3):385–390, Feb 1951.

[109] J. C. Slater and G. F. Koster. Simplified lcao method for the periodic potential problem. *Phys. Rev.*, 94(6):1498–1524, Jun 1954.

[110] I. Souza, N. Marzari, and D. Vanderbilt. Maximally localized wannier functions for entangled energy bands. *Phys. Rev. B*, 65(3):035109, Dec 2001.

[111] A. Szabo and N. S. Ostlund. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory.* Dover Publications, 1996.

[112] J. Torras, E. Deumens, and S. B. Trickey. Software integration in multi-scale simulations: the pupil system. *Journal of Computer-Aided Materials Design*, 13(1-3):201–212, Oct 2006.

[113] L. Van Hove. The occurrence of singularities in the elastic frequency distribution of a crystal. *Phys. Rev.*, 89(6):1189–1193, Mar 1953.

[114] 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. *International Journal for Numerical Methods in Fluids*, 47:773–788, 2005.

[115] 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. In *Proceedings of the 16th AIAA Computational Fluid Dynamics Conference*, 2003. Paper 2003-3847.

[116] X. Wang, J. R. Yates, I. Souza, and D. Vanderbilt. Ab initio calculation of the anomalous hall conductivity by wannier interpolation. *Physical Review B (Condensed Matter and Materials Physics)*, 74(19):195118, 2006.

[117] H. Wilson, G. Galli, F. Gygi, S. Hamel, A. Williamson, E. Ratner, and D. Wack. Efficient calculations of the dielectric response in semiconductor nanostructures for optical metrology. *APS Meeting Abstracts*, page 41015, Mar. 2007.

[118] S. Wilson. Basis sets. In *Methods in Computational Molecular Physics*, volume 113 of *NATO ASI Series*, pages 71–93. Reidel, Dordrecht, 1983.

[119] S. Wilson. Finite basis sets and the algebraic approximation. In *Handbook of Molecular Physics and Quantum Chemistry*, pages 585–640. Wiley, England, 1990.

[120] M. T. Yin and M. L. Cohen. Theory of static structural properties, crystal stability, and phase transformations: Application to Si and Ge. *Phys. Rev. B*, 26(10):5668–5687, Nov 1982.

[121] D. C. Young. Using existing basis sets. In *Computational Chemistry*, pages 78–91. Wiley, 2002.

[122] P. Y. Yu and M. Cardona. *Fundamentals of Semiconductors: Physics and Materials Properties.* Springer, New York, 2005.