# Lanczos Approximation of Joint Spectral Quantities and Applications

**A DISSERTATION**
**SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL**
**OF THE UNIVERSITY OF MINNESOTA**
**BY**

**Tyson S. Loudon**

**IN PARTIAL FULFILLMENT OF THE REQUIREMENTS**
**FOR THE DEGREE OF**
**DOCTOR OF PHILOSOPHY**

**Advisor: Douglas N. Arnold**

**April, 2021**

# Acknowledgements

I would like to acknowledge the extensive help and guidance provided by my advisor Professor Arnold. His generosity with his time and ability to explain complex and (seemingly) incomprehensible ideas in a perfectly understandable manner has been a great aid. His guiding principle to make things as simple as possible, but not simpler, is one of the most valuable lessons I received during the course of my education.

# Dedication

I dedicate this thesis to Emily, the love of my life. Without whom I would be aimless and morbidly obese, and with whom I have purpose, support, and most importantly, family.

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

In the physics literature, spectral quantities abound. These are objects of the form $\sum_i w_i \delta(\lambda - \lambda_i)$, where the $\lambda_i$'s are the eigenvalues, or energies, of a Hermitian operator, and the coefficients, $w_i$, oftentimes depend on the corresponding eigenvectors. One frequently encountered spectral quantity, which much work has been devoted to approximating, is the density of states $\sum_i \delta(\lambda - \lambda_i)$, see, e.g., [33, 67, 51, 32] and references therein. In quantum physics, the density of states represents the distribution of energies at which quantum states are available for occupation. In numerical linear algebra, the density of states can be used to determine the number of eigenvalues of a matrix in a given interval, and an approximation of the density of states is useful in large-scale eigenvalue problems for this purpose. Another spectral quantity, which a large portion of this thesis is devoted to, is the spectral function $\sum_i |(x_i, v)|^2 \delta(\lambda - \lambda_i)$, where $x_i$ is the eigenvector corresponding to $\lambda_i$, $v$ is an arbitrary vector, and $(\,\cdot\,,\,\cdot\,)$ denotes an appropriate inner product which varies depending on the context. As we will see, the spectral function is central to the approximation of all other spectral quantities.

This thesis addresses the approximation of joint spectral quantities. Joint spectral quantities are a natural extension of the notion of spectral quantities to two distinct systems. These are quantities of the form $\sum_{i,j} w_{ij} \delta\big(\lambda - (\lambda_i + \lambda'_j)\big)$, where the eigenvalues $\lambda_i$ and $\lambda'_j$ are those of two distinct operators, and the coefficients, $w_{ij}$, may depend on the corresponding eigenvectors. It is often possible to consider joint spectral quantities as spectral quantities associated with a larger matrix. In some instances, this insight allows us to use known methods in a slightly new way.

Joint spectral quantities are of particular interest when modeling semiconductors, or other materials, where the fundamental material properties are determined by the distribution of electrons in the conductance band and holes in the valence band. While the electrons and holes are related, the Hamiltonian systems governing each are distinct, and the sum of eigenvalues of these two systems represents the energy required to excite an electron into the conductance band, and spawn the creation of an electron-hole pair. This process is fundamental in the operation of many electronic devices, such as, light-emitting diodes, laser diodes, and solar cells.

The most direct method to compute spectral and joint spectral quantities involves solving for all eigenpairs of one or more Hermitian operator(s). While many problems in physics and engineering require the largest or smallest eigenvalues in magnitude of an operator, spectral quantities are complicated in that they require all eigenvalues, and possibly all eigenvectors. Many methods exists for computing select eigenpairs of a linear operator, e.g., shift-and-invert type methods. However, computing all eigenpairs of an operator is a daunting task, and often represents a bottleneck in engineering applications. Thus, methods for approximating spectral and joint spectral quantities which avoid costly eigenvalue problems are necessary.

The two most prevalent methods for approximating spectral quantities in the literature are the Kernel Polynomial Method (KPM) and, as we refer to it in this thesis, the Lanczos process. The KPM was developed in the 1990's by physicists for use in approximating spectral functions and densities of states, and involves performing a formal Chebyshev polynomial expansion of a spectral function [61, 51, 53, 52]. The Chebyshev coefficients in the expansion are the so called "modified moments," which are quadratic forms involving Chebyshev polynomials of a matrix. To compute these modified moments, the three-term Chebyshev recursion may be used. The other method for approximating spectral functions, the Lanczos process, has classically been used to approximate bilinear forms $u^T f(A)v$ where $f$ is smooth, $A$ is Hermitian, and $u$ and $v$ are given vectors. A systematic introduction to using the Lanczos algorithm to approximate bilinear forms is given in [17]. These bilinear forms have many uses, including error estimation in iterative linear solvers [13], matrix function trace estimation [59, 7], and partial eigenvalue sums [6]. The Lanczos process can be viewed as a Gaussian quadrature approximation of an integral with respect to an unknown measure which depends on the spectrum of $A$ and the vector $v$.

Both the Lanczos process and the KPM have their benefits and drawbacks. The main benefit of the Lanczos process is the accuracy achieved. Because of the moment matching property of the Lanczos process, which follows from the relationship to Gauss quadrature, the Lanczos process is able to approximate spectral functions quite accurately at the cost of relatively few iterations of the Lanczos algorithm. The main drawback of the Lanczos process is the deterioration of mutual orthogonality between the basis vectors in the Lanczos algorithm due to finite precision arithmetic. Because of this loss of orthogonality, the beautifully simple three-term recurrence in the Lanczos algorithm cannot be used, and more costly orthogonalization techniques must be applied, e.g., full Gram–Schmidt orthogonalization. The main benefit of the KPM is its use of the three-term Chebyshev recurrence in order to compute expansion coefficients. However, the efficiency of the KPM comes at the cost of accuracy. Because the KPM relies on a Chebyshev expansion of Dirac measures, high degree polynomials are required in order to obtain accurate approximations. In contrast, the Lanczos process forms an approximation in terms of a linear combination of Dirac measures, and so is of the same form as the spectral function, albeit with fewer terms.

The main ingredient necessary for the Lanczos process is the Lanczos partial tridiagonalization of a matrix with respect to a given vector determined by the Lanczos algorithm. From the Lanczos partial tridiagonalization, we are able to compute quadrature nodes and weights which determine the Lanczos approximation to a spectral function. The Lanczos algorithm deviates drastically from theory once finite precision effects are taken into consideration. Necessarily, any discussion of the Lanczos process would be lacking without taking into consideration the effects of finite precision in the Lanczos algorithm. In Chapter 2 we discuss the Lanczos algorithm in exact and finite precision, taking special care to focus on Lanczos partial tridiagonalizations. Also discussed are methods developed to overcome issues encountered in finite precision.

In Chapter 3 we discuss the theory of the Lanczos process, and how the Lanczos process for approximating bilinear forms can be viewed as an approximation to the spectral function associated to a symmetric matrix and given vector. A priori error estimates for the Lanczos process currently in the literature are only available with respect to analytic functions. However, it is more appropriate to consider the error in a distributional sense, i.e., in negative Sobolev norms. This is also accomplished in Chapter 3 using a mainstay of approximation theory, Jackson's Theorem, as well as Sobolev imbedding theorems.

Next, we introduce spectral quantities, joint spectral quantities, and their Lanczos approximations in Chapter 4. Beginning with spectral quantities, we overview the existing literature on using Hutchinson's method, a Monte Carlo trace estimator, to form approximations of the density of states. This method is used extensively in the applications discussed in Chapter 5. Moving to joint spectral quantities, we define the joint density of states and joint spectral function. The joint density of states is a natural extension of the density of states to two distinct linear systems. We show how to approximate the joint density of states using methods pertaining to the density of states. Additionally, we develop another method for approximating the joint density of states which relies on the notion of convolution of measures. The final, and most difficult, joint spectral quantity is the joint spectral function. The joint spectral function, when computed exactly, requires full knowledge of all eigenvalues and eigenvectors of both systems under consideration. We show how using the spectrum of one operator or the other, but not both, we are able to accurately approximate the joint spectral function using the Lanczos process. Furthermore, if we only wish to approximate the joint spectral function in a small interval, we show that only a few select eigenpairs of one operator are necessary. In many instances, this makes the approximation of joint spectral functions tractable. For all cases of spectral and joint spectral quantities, we consider both standard eigenvalue problems and generalized eigenvalue problems.

In Chapter 5 we apply the theory developed in previous chapters to modeling random alloys. For this application we use the effective mass Schrödinger equation to model electrons and holes in an indium gallium nitride (InGaN) alloy. Using the Lanczos process, we analyze properties of InGaN alloys, and show how random alloys deviate from simpler homogeneous alloys. To the best of the authors knowledge, this is the first time a full numerical analysis of the effective mass Schrödinger equations in one, two, and three spatial dimensions has been performed for random alloys.

# Chapter 2

# Lanczos Partial Tridiagonalization

In this chapter we introduce the theory of partially tridiagonalizing a matrix with respect to a given starting vector using the Lanczos algorithm. We first discuss the Lanczos algorithm in infinite precision, and then take into account the effects of finite precision.

## 2.1 Krylov Subspaces

Let $A \in \mathbb{R}^{n \times n}$ be a nonsingular matrix and $v \in \mathbb{R}^n$ a nonzero vector. Denote the family of Krylov spaces generated by $A$ and $v$ as

$$\mathcal{K}_m(A, v) = \operatorname{span}\{v, Av, \ldots, A^{m-1}v\}, \quad m \in \mathbb{N}. \tag{2.1}$$

Krylov subspaces are the foundation of iterative methods for solving linear systems, e.g., conjugate gradient and GMRES, and for eigenvalue problems, e.g., Lanczos and Arnoldi iterations. We begin this section by understanding why Krylov spaces are natural to consider. Our presentation closely follows that of [25].

Suppose we are interested in determining the solution, $x \in \mathbb{R}^n$, to the linear system $Ax = b$. Recall the minimal polynomial of the matrix $A$ is the monic polynomial $p$, of minimal degree, for which $p(A) = 0$. The polynomial $p$ is easily constructed from the Jordan normal form of $A$. If $\lambda_1, \ldots, \lambda_d$ are the unique eigenvalues of $A$ and $m_i$ is the size of the largest Jordan block corresponding to $\lambda_i$, then the minimal polynomial of $A$ is given by $p(t) = \Pi_{i=1}^{d}(t - \lambda_i)^{m_i}$. For example, if the Jordan form of $A$ is

$$\begin{pmatrix} 1 & & & & \\ & 1 & 1 & & \\ & & 1 & & \\ & & & 2 & \\ & & & & 2 \end{pmatrix},$$

then, $\lambda_1 = 1$ has two Jordan blocks, the larger of which has size $m_1 = 2$, and $\lambda_2 = 2$ has two blocks of size $m_2 = 1$. For this example, the minimal polynomial is $p(t) = (t-1)^2(t-2)$.

Expanding out $p(t) = \prod_{i=1}^{d}(t - \lambda_i)^{m_i}$ in terms of the monomials, $t^i$, gives $p(t) = \sum_{i=0}^{m} c_i t^i$ where $m = m_1 + \ldots + m_d$. Note that $c_0 = \Pi_{i=1}^{d}(-\lambda_i)^{m_i} \neq 0$ since we assumed $A$ is invertible. Using $A^{-1}p(A) = 0$, it is easily seen that

$$A^{-1} = \sum_{i=1}^{m}\Big(-\frac{c_i}{c_0}\Big)A^{i-1} = q(A) \quad \text{with} \quad q(t) = \sum_{i=0}^{m-1}\Big(-\frac{c_{i+1}}{c_0}\Big)t^i.$$

Hence, the solution of the linear system $Ax = b$ satisfies $x = A^{-1}b = q(A)b$. We have just proved the following theorem.

**Theorem 1.** *If the minimal polynomial of a nonsingular matrix $A \in \mathbb{R}^{n\times n}$ has degree $m$, the solution of $Ax = b$ lies in $\mathcal{K}_m(A, b)$.*

If the degree of the minimal polynomial of $A$ is small, we can search for the solution to $Ax = b$ in the low-dimensional Krylov space associated with $A$ and $b$. In practice, it is typically not the case that the degree of the minimal polynomial of $A$ is small. For example, if all eigenvalues of $A$ are simple, then the degree of the minimal polynomial is $n$. The purpose of what has been presented so far, is to show that the action of the inverse of the matrix $A$ on the vector $b$, and the Krylov spaces associated with $A$ and $b$, are intimately related. The main utility of methods involving Krylov spaces resides in the fact that satisfactory approximate solutions to the system $Ax = b$ can typically be found in $\mathcal{K}_m(A, b)$ for $m \ll n$, even if the degree of the minimal polynomial is $n$.

Note that in Theorem 1 we did not use any information about the right-hand side vector in determining the dimension of the Krylov subspace. Rather, we showed that $A^{-1}v \in \mathcal{K}_m(A, v)$ for all $v \in \mathbb{R}^n$, so long as the degree of the minimal polynomial of $A$ is $m$. Next, we investigate how the relationship between $A$ and $v$ influences the dimension of the Krylov space $\mathcal{K}_m(A, v)$.

From the definition of Krylov subspaces (2.1), it is clear that $\mathcal{K}_j(A, v) \subseteq \mathcal{K}_{j+1}(A, v)$ for all natural numbers $j$, and that $\dim \mathcal{K}_j(A, v) \leq j$. The next theorem illustrates that there is a maximal dimension, $\overline{m}$, such that

$$\mathcal{K}_1(A, v) \subsetneq \mathcal{K}_2(A, v) \subsetneq \ldots \subsetneq \mathcal{K}_{\overline{m}}(A, v) = \mathcal{K}_{\overline{m}+1}(A, v) = \ldots \quad . \tag{2.2}$$

Recall, the minimal polynomial of $v$ with respect to $A$ is the monic polynomial of minimal degree for which $p(A)v = 0$. The grade of $v$, denoted grade$(v)$, is the degree of the minimal polynomial of $v$ with respect to $A$. Note that grade$(v)$ is always less than or equal to the degree of the minimal polynomial of $A$. The following proposition can be found in [48].

**Theorem 2.** *For $A \in \mathbb{R}^{n\times n}$ and $v \in \mathbb{R}^n$, $\dim(\mathcal{K}_m(A, v)) = \min(m, \mathrm{grade}(v))$.*

*Proof.* We first show that for $\overline{m} = \mathrm{grade}(v)$, $A^{\overline{m}+j}v \in \mathcal{K}_{\overline{m}}(A, v)$ for all $j \geq 0$. We proceed by induction. Denote the vector space of polynomials of degree less than or equal to $k$ as $\mathscr{P}_k$. Let $p \in \mathscr{P}_{\overline{m}}$ be the minimal polynomial of $v$ with respect to $A$, and write $p(t) = t^{\overline{m}} - q(t)$ for some $q \in \mathscr{P}_{\overline{m}-1}$. Using $p(A)v = 0$, we see that $A^{\overline{m}}v = q(A)v \in \mathcal{K}_{\overline{m}}(A, v)$. Next, assume that $A^{\overline{m}+j}v \in \mathcal{K}_{\overline{m}}(A, v)$ for $j = 0, \ldots, k$. Expressing $A^{\overline{m}+k}v$ in terms of the vectors $v, Av, \ldots, A^{\overline{m}-1}$, we see that

$$A^{\overline{m}+k+1}v = A\Big(A^{\overline{m}+k}v\Big) = A\sum_{i=0}^{\overline{m}-1} c_i A^i v = c_{\overline{m}-1}A^{\overline{m}}v + \sum_{i=1}^{\overline{m}-1} c_{i-1}A^i v,$$

for some constants $c_i$, $i = 0, \ldots, \overline{m} - 1$. Using again $A^{\overline{m}}v \in \mathcal{K}_{\overline{m}}(A, v)$, we see that $A^{\overline{m}+k+1}v \in \mathcal{K}_{\overline{m}}(A, v)$, as desired.

This demonstrates that $\dim \mathcal{K}_m(A, v) \leq \mathrm{grade}(v)$ for all $m$. Lastly, to complete the proof, we show that $\dim \mathcal{K}_m(A, v) = m$ if and only if $m \leq \mathrm{grade}(v)$. The vectors $v, Av, \ldots, A^{m-1}v$, are linearly independent if and only if for any collection of constants $c_i$, $i = 0, \ldots, m-1$, not all zero, the sum, $\sum_{i=0}^{m-1} c_i A^i v$, is nonzero. This is equivalent to saying there is no polynomial $q \in \mathscr{P}_{m-1}$ such that $q(A)v = 0$, i.e., $m \leq \mathrm{grade}(v)$. □

Theorem 2 shows that $\dim \mathcal{K}_m(A, v) = m$, so long as grade$(v) \geq m$. In what follows, we always assume $\dim \mathcal{K}_m(A, v) = m \ll n$ to simplify the analysis. For a symmetric matrix $A$ with simple eigenvalues, as long as $v$ has nonzero components in the direction of each eigenvector of $A$, $\dim \mathcal{K}_m(A, v) = m$ for all $m \leq n$, and so this assumption holds in most practical situations. To see this, note that when $v$ has nonzero components in the direction of each eigenvector of $A$ then the minimal polynomial of $A$, the characteristic polynomial of $A$, and the minimal polynomial of $v$ with respect to $A$ coincide, and are degree $n$. That the characteristic polynomial of $A$ and the minimal polynomial of $A$ are equal in this situation follows from the previous discussion of constructing the minimal polynomial from the Jordan normal form. To see that the minimal polynomial of $v$ with respect to $A$ equals the characteristic polynomial, assume the orthogonal eigenvectors of $A$ are $x_i$, with corresponding eigenvalues $\lambda_i$, for $i = 1, \ldots, n$. If the coefficients of $v$ in the eigenbasis are $\gamma_i$, then for any polynomial $p$,

$$p(A)v = \sum_{i=1}^{n} \gamma_i p(\lambda_i) x_i. \tag{2.3}$$

From (2.3) we see that if $p(A)v = 0$ and the $\gamma_i$'s are nonzero, then $p(\lambda_i) = 0$ for $i = 1, \ldots, n$. In other words, the minimal polynomial of $v$ with respect to $A$ is the same as the characteristic polynomial of $A$ when $v$ has nonzero components in the direction of each eigenvector of $A$.

## 2.2 Arnoldi Algorithm

Using the facts established about Krylov spaces in the previous section, we now turn to the Arnoldi algorithm for constructing an orthonormal basis of $\mathcal{K}_m(A, v)$. We show that this is the same basis determined by performing the Gram–Schmidt algorithm on the vectors $\{v, Av, \ldots, A^{m-1}v\}$. Recall that we always assume $\dim \mathcal{K}_m(A, v) = m$, meaning that the vectors $\{v, Av, \ldots, A^{m-1}v\}$ are linearly independent for $m \leq n$.

The basis for $\mathcal{K}_m(A, v)$ as given in (2.1), while useful for theory, is of little use in practice. This is because as the Krylov dimension increases, the vectors $A^j v$ become closely aligned with the dominant eigenvector, as in power iteration. While this is not an issue in perfect arithmetic, it does pose an issue on finite precision computers. For this reason, we need to determine a basis more suited for finite precision computations. This is precisely what the Arnoldi algorithm does [3].

Assume that $H \in \mathbb{R}^{n \times n}$ is an upper Hessenberg matrix orthogonally similar to $A$. That is,

$$AV = VH, \tag{2.4}$$

for an orthogonal matrix $V$. Equating the $j$th columns in (2.4), we have relation

$$Av_j = \sum_{i=1}^{j+1} h_{ij} v_i, \quad j = 1, \ldots, n, \tag{2.5}$$

where $v_j$ is the $j$th column of $V$, $(H)_{ij} = h_{ij}$, and $h_{n+1\,n} v_{n+1} = 0$. Rearranging (2.5), we can express the vector $v_{j+1}$ using a $j+1$ term recurrence involving the vectors $v_1, \ldots, v_j$,

$$h_{j+1\,j}\, v_{j+1} = Av_j - \sum_{i=1}^{j} h_{ij} v_i, \quad j = 1, \ldots, n. \tag{2.6}$$

From the orthonormality of the vectors $v_i$, the coefficients in (2.6) satisfy

$$h_{ij} = v_i^T A v_j, \qquad i = 1, \ldots, j, \qquad h_{j+1\,j} = \Big\| Av_j - \sum_{i=1}^{j} h_{ij} v_i \Big\|. \tag{2.7}$$

From recurrence (2.6) and the formulas for the coefficients in (2.7), given a nonzero starting vector $v \in \mathbb{R}^n$, we are able to construct the columns of an orthogonal matrix $V$, with first column $v_1 = v/\|v\|$, and upper Hessenberg matrix $H$ which satisfy (2.4). This is known as the Arnoldi algorithm, and the vectors $v_j$ are known as the Arnoldi vectors.

In practice, we are rarely interested in constructing $V$ and $H$ in full. Rather, we are mostly interested in utilizing (2.6) and (2.7) for $j = 1, \ldots, m$, with $m \ll n$. If we let $V_m$ be the first $m$ columns of $V$, and $H_m$ be the $m \times m$ principal submatrix of $H$, then we can write

$$AV_m = V_m H_m + h_{m+1\,m} v_{m+1} e_m^T, \tag{2.8}$$

where $e_m$ is the $m$th column of the $m \times m$ identity matrix. The $m$-step Arnoldi algorithm constructs all terms in (2.8), and is summarized in Algorithm 1.

**Algorithm 1** Arnoldi Algorithm

```
Initialize $v_1 = v/\|v\|$.
for $j = 1, \dots, m$ do
    $\tilde{v} = Av_j$
    for $i = 1, \dots, j$ do
        $h_{ij} = (\tilde{v}, v_i)$
        $\tilde{v} \leftarrow \tilde{v} - h_{ij} v_i$
    end for
    $h_{j+1\,j} = \|\tilde{v}\|$
    if $h_{j+1\,j} = 0$ then stop
    else
        $v_{j+1} = \frac{\tilde{v}}{h_{j+1\,j}}$
    end if
end for
```

Consider the stopping criteria, $h_{j+1\,j} = 0$, in Algorithm 1. When this occurs at step $j \leq m$, we have that $h_{j+1\,j} v_{j+1} = 0$, and so $AV_j = V_j H_j$. This means that the columns of $V_j$ span an invariant subspace of $A$. This is useful if we are interested in solving for eigenpairs of $A$. Indeed, if $\theta$ is an eigenvalue of $H_j$ with associated eigenvector $y$, then $\theta$ is an eigenvalue of $A$ with associated eigenvector $V_j y$. So, using an upper Hessenberg matrix of order $j$, we are able to determine spectral properties of $A$. This is a rare occurrence in practice. We show next that the Arnoldi vectors are a basis of the Krylov space $\mathcal{K}_m(A, v)$, and so our general assumption that the vectors $v, Av, \dots, A^j v$, are linearly independent for $j < n$ preclude the stopping criteria, $h_{j+1\,j} = 0$, from being achieved.

Carrying out the Arnoldi algorithm in full, with starting vector $v_1 = v/\|v\|$, results in $V, H \in \mathbb{R}^{n \times n}$, such that $AV = VH$, where $V$ is orthogonal and $H$ is upper Hessenberg with positive subdiagonal. Let $K = QR$ be the QR-factorization of the Krylov matrix $K = [v_1, Av_1, \dots, A^{n-1} v_1]$, where $Q$ is orthogonal and $R$ is upper triangular with positive diagonal. Since the QR-algorithm is simply the Gram–Schmidt algorithm applied to the columns of $K$, we know that the columns of $Q$ are an orthonormal basis for the space $\mathcal{K}_n(A, v)$, and the first $m$ columns of $Q$ are an orthonormal basis of $\mathcal{K}_m(A, v)$. Using $H$ and $V$, we can also write $K = VV^T K = V[e_1, He_1, H^2 e_1, \dots, H^{n-1} e_1]$. Since $H$ is upper Hessenberg with positive elements on the subdiagonal, $H^k$ has positive elements on the $k$th subdiagonal and is zero below. Hence, the matrix $[e_1, He_1, \dots, H^{n-1} e_1]$ is upper triangular with positive elements on the diagonal, and so $K = V[e_1, He_1, \ldots, H^{n-1}e_1]$ is a QR-factorization of $K$. By uniqueness of QR-factorizations we have $Q = V$, and so the columns of $V_m$ are an orthonormal basis for $\mathcal{K}_m(A, v)$.

## 2.3 Lanczos Algorithm

In this section we specialize to the case when $A$ is symmetric. We assume throughout this section that all operations are performed with exact arithmetic. The effects of finite precision will be taken into consideration in the next section.

Rewriting relationship (2.8) as $H_m = V_m^T A V_m$, we see that if $A$ is symmetric then the upper Hessenberg matrix $H_m$ is also symmetric. A symmetric upper Hessenberg matrix is a symmetric tridiagonal matrix. Throughout the rest of this chapter we replace $H_m$ with $T_m \in \mathbb{R}^{m \times m}$ to emphasize that we are dealing with a tridiagonal matrix. Writing $\alpha_j = h_{jj}$ and $\beta_j = h_{j+1\,j}$ for $j = 1, \ldots, m$, the Arnoldi recurrence now simplifies to

$$Av_j = \beta_{j-1}v_{j-1} + \alpha_j v_j + \beta_j v_{j+1}, \quad j = 1, \ldots, m, \tag{2.9}$$

where $\beta_0 v_0 = 0$, or, equivalently,

$$AV_m = V_m T_m + \beta_m v_{m+1} e_m^T, \tag{2.10}$$

where

$$T_m = V_m^T A V_m = \begin{pmatrix} \alpha_1 & \beta_1 & & \\ \beta_1 & \ddots & \ddots & \\ & \ddots & & \beta_{m-1} \\ & & \beta_{m-1} & \alpha_m \end{pmatrix}. \tag{2.11}$$

The vectors $v_j$ are now referred to as Lanczos vectors.

The Lanczos algorithm is widely used for many different applications [29]. Often, it is used to find a few extremal eigenvalues of sparse symmetric matrices. However, when we use the term "Lanczos algorithm" or "Lanczos iteration" in this thesis, we will be referring to algorithms for constructing $V_m$ and $T_m$ in (2.10). Oftentimes, we will use the phrase partial tridiagonalization when specifically referring to construction of the matrix $T_m$.

Because we will need them later, we introduce some terminology regarding the matrices $T_m$ and $V_m$. Denote the eigenvalues and eigenvectors of $T_m$ as $\theta_j$ and $y_j$ respectively, $j = 1, \ldots, m$ (we drop the dependence on $m$ for notational convenience). The values $\{\theta_j\}_{j=1}^m$ are called the Ritz values, and the vectors $\{V_m y_j\}_{j=1}^m$ are called the Ritz vectors. This terminology comes from the fact that the Lanczos algorithm can be viewed as a Rayleigh-Ritz method for approximating eigenpairs of the matrix $A$.

Before discussing the properties of the Lanczos algorithm, we present the equivalent of Algorithm 1 for the case of a symmetric matrix with the notation used in (2.9). This is given in Algorithm 2.

**Algorithm 2** Lanczos Algorithm

```
Initialize $v_1 = v/\|v\|$, $\beta_0 v_0 = 0$.
for $j = 1, \ldots, m$ do
    $\tilde{v} = Av_j - \beta_{j-1} v_{j-1}$
    $\alpha_j = (\tilde{v}, v_j)$
    $\tilde{v} \leftarrow \tilde{v} - \alpha_j v_j$
    $\beta_j = \|\tilde{v}\|$
    if $\beta_j = 0$ then stop
    else
        $v_{j+1} = \frac{\tilde{v}}{\beta_j}$
    end if
end for
```

We remark on several desirable features of the Lanczos algorithm:

1. Only one matrix vector multiplication is needed each iteration. Furthermore, we do not need to have the matrix $A$ stored in memory. Rather, it is sufficient to supply a routine which, given a vector $v$, returns $Av$. Thus, we are able to take full advantage of the case when $A$ is large and sparse.

2. For some applications we are only interested in creating a partial tridiagonalization of $A$. For example, this is the case when only eigenvalue, and not eigenvector, approximations of $A$ are desired. In this case we do not need to create the matrix $V_m$. Instead, at step $j$, only the coefficients $\alpha_j$ and $\beta_j$ need to be computed. For this, only the previous two Lanczos vectors, $v_{j-1}$ and $v_j$, need to be stored.

3. There are several ways to reduce a symmetric matrix to tridiagonal form, e.g.,

> Givens and Householder rotations. However, algorithms utilizing Givens or Householder rotations must be carried out in full before the matrix $A$ is reduced to tridiagonal form. At no intermediate step is the matrix tridiagonal. Contrast this with the Lanczos algorithm, which provides a partial tridiagonalization, $T_j = V_j^T AV_j$, *at each step* $j$.

The benefits of the Lanczos algorithm are manifold, which is why it is so widely used. However, as was known originally to Lanczos [29], the algorithm behaves differently in the presence of roundoff error. These differences are considered in the next section.

## 2.4 Lanczos Algorithm in Finite Precision

The desirable features of the Lanczos algorithm mentioned in the previous section, make it suitable for many applications, including eigenvalue problems, the solution of linear systems, and singular value decompositions. However, one of the main drawbacks, noticed in the original paper by Lanczos [29], is the loss of orthogonality of the Lanczos vectors due to rounding error. This widely known issue is discussed at length in all serious textbooks and articles on the Lanczos algorithm. In this section we discuss the loss of orthogonality in the context of the partial tridiagonalization of the matrix $A$. The original work on this issue is the PhD thesis of Paige [39], and his subsequent publications [40, 41].

In this section we will be dealing with computable quantities on finite precision computers. In order to avoid an abundance of tildes, or other methods to distinguished exact quantities and computed quantities, we use the same notation as in the previous section, with the knowledge that all quantities are computed in finite precision, unless stated otherwise. That is, the $v_j$'s, $\alpha_j$'s, and $\beta_j$'s represent quantities which have been computed with roundoff error, and are not identical to their counterparts in the previous section. For simplicity, we make the assumption that all Lanczos vectors have been normalized exactly, i.e., that $v_j^T v_j \equiv 1$. All equations will now include an error term, e.g., the recurrence relation (2.9) now becomes

$$Av_j = \beta_{j-1} v_{j-1} + \alpha_j v_j + \beta_j v_{j+1} + f_j, \quad j = 1, \dots, m, \tag{2.12}$$

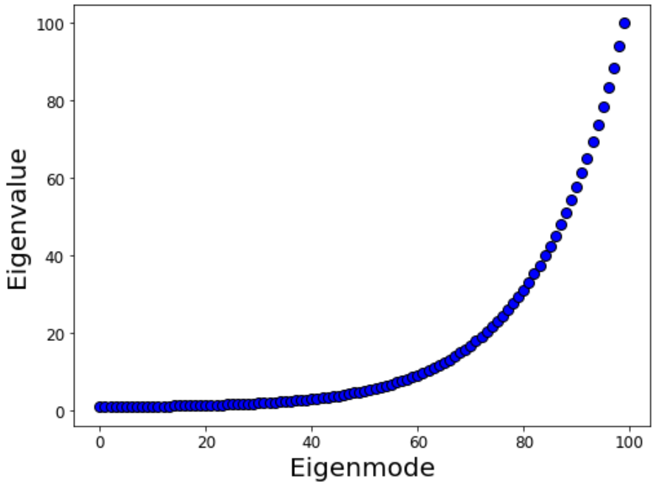


Figure 2.1: Eigenvalue distribution (2.14).

where $\beta_0 v_0 = 0$ and $f_j$ is a vector with entries accounting for the roundoff error at iteration $j$. Or, equivalently, in matrix form,

$$AV_m = V_m T_m + \beta_m v_{m+1} e_m^T + F_m, \tag{2.13}$$

where the $j$th column of matrix $F_m \in \mathbb{R}^{n\times m}$ accounts for the roundoff error at iteration $j$. Throughout this section $\epsilon$ represents the unit roundoff error, which for 64-bit computations is of order $10^{-16}$.

Before discussing the details surrounding roundoff errors in the Lanczos algorithm, we give a simple example illustrating the deviation of finite precision from exact arithmetic taken from [36]. We perform the Lanczos algorithm on a symmetric matrix with eigenvalues given by

$$\lambda_i = \underline{\lambda} + \frac{i-1}{n-1}(\overline{\lambda} - \underline{\lambda})\rho^{n-i} \quad i = 1, \ldots, n, \tag{2.14}$$

where $\rho$ is a parameter controlling the eigenvalue distribution, $\underline{\lambda}$ and $\overline{\lambda}$ are the beginning and end of the spectrum, and $n$ is the size of the matrix. For this example, we choose $\underline{\lambda} = 1$, $\overline{\lambda} = 100$, $\rho = 0.95$, $n = 100$, and we construct $V_m$ for $m = 50$ using two different methods. The eigenvalues for this example can be seen in Figure 2.1, which illustrates the clustering of eigenvalues near $\underline{\lambda}$ due to the parameter $\rho$ in (2.14). We

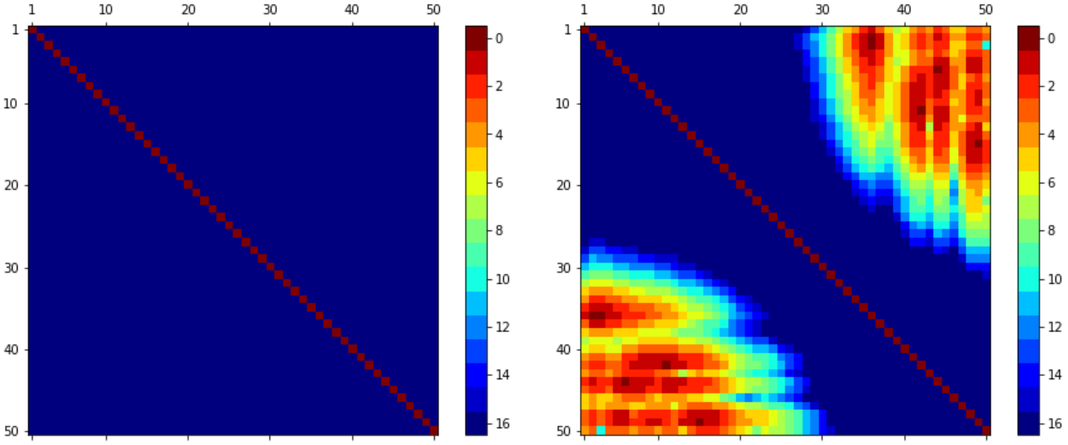


Figure 2.2: Plot of the matrix entries $-\log(|V_m^T V_m|)$ for $m = 50$ using Algorithm 1 (left) and Algorithm 2 (right) for a symmetric matrix.

use Algorithms 1 and 2 to construct the basis vectors, the difference being that Algorithm 2 constructs $v_{j+1}$ by orthogonalizing $Av_j$ against the two previously computed basis vectors $\{v_{j-1}, v_j\}$, whereas Algorithm 1 orthogonalizes $Av_j$ against all previously computed basis vectors $\{v_1, \ldots, v_j\}$, which is equivalent in exact arithmetic. The results, visualized by plotting the matrix elements $-\log|V_m^T V_m|$, can be seen in Figure 2.2. From Figure 2.2 we see that using the three term recurrence in the Lanczos algorithm results in a significant loss of orthogonality, while orthogonalizing each new Lanczos vector against all previously computed basis vectors retains mutual orthogonality to machine precision. Indeed, using the three term recurrence of Algorithm 2 results in $|v_i^T v_j| = \mathcal{O}(1)$ for some $i \neq j$, when it should ideally be $\mathcal{O}(\epsilon)$.

When taking into account roundoff error, we can expect that relationship (2.12) holds to within machine precision, i.e., the entries of $f_j$ are $\mathcal{O}(\epsilon)$. This is clearly seen in Figure 2.3, where the Lanczos vectors are computed using Algorithm 2 (same example as Figure 2.2 (right)), and the norm of $f_j = Av_j - \beta_{j-1}v_{j-1} - \alpha_j v_j - \beta_j v_{j+1}$ is plotted for each iteration. Even though the Lanczos vectors are far from orthogonal, as seen in Figure 2.2 (right), (2.12) holds for $f_j$ with entries of order $\epsilon$. In [39] it is shown that $\|f_j\| \leq C_n \epsilon \|A\|$, where $C_n$ depends on the matrix size $n$, and in [44] it is stated that no exception to the rule $\|f_j\| \leq \epsilon \|A\|$ has been found. In other words, even though the vector accounting for the roundoff error, $f_j$, remains small, we lose mutual orthogonality

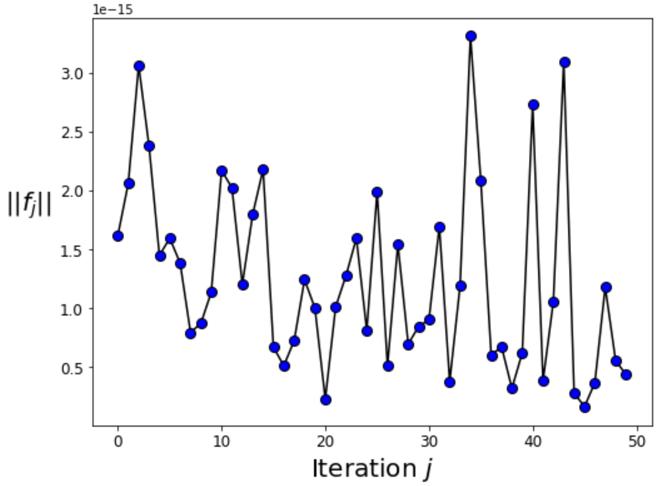


Figure 2.3: Norm of roundoff error $f_j$ computed using the Lanczos algorithm.

between the Lanczos vectors.

A major issue that arises as a consequence of the loss of orthogonality is $T_m \not\approx V_m^T A V_m$, i.e., we no longer have a partial tridiagonalization of the matrix $A$. This can be seen by premultiplying (2.13) by $V_m^T$,

$$V_m^T A V_m = (V_m^T V_m) T_m + V_m^T F_m. \tag{2.15}$$

While we expect the entries of $V_m^T F_m$ to be small since each column of $F_m$ has norm $\mathcal{O}(\epsilon \|A\|)$ and the columns of $V_m$ are unit length, $V_m^T V_m$ is far from the identity, as seen in Figure 2.2 (right). Thus, if we are interested in producing a partial tridiagonalization of $A$, we must take into account the loss of mutual orthogonality of the Lanczos vectors in finite precision.

Note that Figure 2.2 shows a clear structure to the loss of mutual orthogonality. We investigate this structure and the propagation of the loss of mutual orthogonality next. Let the matrix $K_m \in \mathbb{R}^{m \times m}$ have entries $k_{ij} = v_i^T v_j$, $i, j = 1, \ldots, m$, which sets $K_m = V_m^T V_m$. Ideally, $K_m$ would approximate the identity matrix to machine precision, however, the previous example shows that this does not hold when using Algorithm 2. We can characterize the propagation of diminishing orthogonality using difference equations for the entries $k_{ij}$ [54, 55].

**Theorem 3.** *Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix and $v_1 \in \mathbb{R}^n$ be a unit vector. Let $v_1, \ldots, v_{m+1}$ be the Lanczos vectors after $m$ iterations of the Lanczos algorithm (Algorithm 2) in finite precision and define $k_{ij} = k_{ji} = v_i^T v_j$ for $i, j = 1, \ldots, m+1$. Then, the terms $k_{ij}$ satisfy the following:*

$$\begin{aligned} k_{ii} &= 1 \qquad\qquad i = 1, \ldots, m+1, \\ k_{i\,i+1} &= v_i^T v_{i+1} \qquad i = 1, \ldots, m, \\ \beta_j k_{i\,j+1} &= \beta_i k_{i+1\,j} + (\alpha_i - \alpha_j) k_{ij} + \beta_{i-1} k_{i-1\,j} - \beta_{j-1} k_{i\,j-1} + v_j^T f_i - v_i^T f_j, \end{aligned}$$

*for $1 \le i < j \le m$ and $k_{0j} = 0$.*

*Proof.* The first equation is due to our assumption that the Lanczos vectors are normalized exactly while the second is a definition. Taking the inner product of (2.12) with $v_i$ gives

$$v_i^T A v_j = \beta_{j-1} k_{i\,j-1} + \alpha_j k_{ij} + \beta_j k_{i\,j+1} + v_i^T f_j, \tag{2.16}$$

for $i, j = 1, \ldots, m$. Subtract from (2.16) the same expression with $i$ and $j$ switched, and use $k_{ji} = k_{ij}$, to get

$$0 = \beta_{j-1} k_{i\,j-1} + \alpha_j k_{ij} + \beta_j k_{i\,j+1} + v_i^T f_j - \beta_{i-1} k_{i-1\,j} - \alpha_i k_{ij} - \beta_i k_{i+1\,j} - v_j^T f_i. \tag{2.17}$$

Rearranging terms in (2.17) gives the final result. Note that because $k_{ij} = k_{ji}$, and we have specified the diagonal as well as the super-diagonal, all that remains to be determined is $k_{i\,j+1}$ for $1 \le i < j \le m$. □

Using Theorem 3 we can determine how the loss of mutual orthogonality is propagated forward to newly created Lanczos vectors. The loss of orthogonality is initiated by local roundoff errors $f_j$, and then is propagated forward to newly created Lanczos vectors according the recurrence in Theorem 3. Because the $f_j$'s remain small, it is not due to an accumulation of roundoff errors that loss of orthogonality occurs. Rather, at step $j$, the deviation of $k_{i\,j+1} = v_i^T v_{j+1}$ from zero depends mainly on the level of orthogonality of the Lanczos vectors at the previous two iterations and the $\alpha$'s and $\beta$'s. The dependence of the level of orthogonality of the Lanczos vectors at step $j$ can be visualized using a finite difference stencil for $k_{i\,j+1}$. This is shown in Figure 2.4, which

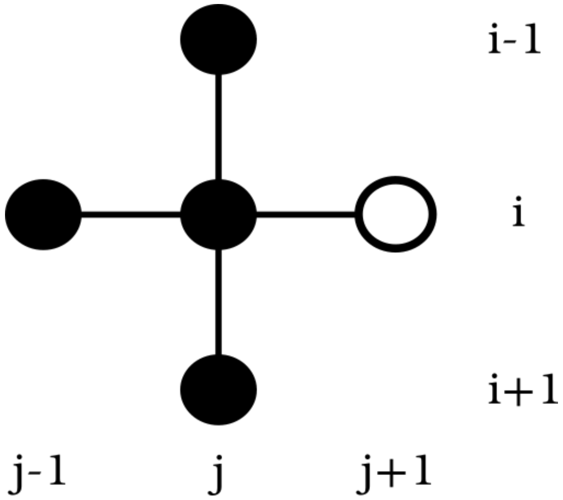


Figure 2.4: Finite difference stencil for $k_{i\,j+1} = v_i^T v_{j+1}$.

is a visual representation of the difference equation in Theorem 3, with the white circle representing $k_{i\,j+1}$, and the black circles representing the terms which $k_{i\,j+1}$ depends on.

In order to understand how far the Lanczos vectors deviate from orthogonal, we define

$$\kappa_m = \max_{1 \leq i,j \leq m} |k_{ij} - \delta_{ij}|, \tag{2.18}$$

where $\delta_{ij}$ is the Kronecker delta. In perfect arithmetic, $\kappa_m = 0$, however, as our example showed, this is not the case in the presence of roundoff error. Due to $\kappa_m$ being nonzero, or equivalently, the Lanczos vectors failing to be mutually orthogonal, the Lanczos algorithm fails to produce a partial tridiagonalization in many situations. Obviously, if we are interested in producing a partial tridiagonalization, we need to alter Algorithm 2. Work on this issue by Parlett and his students [43, 50, 54, 55, 56] has shown that by modifying the Lanczos algorithm to keep the Lanczos vectors “sufficiently orthogonal” (made specific shortly), ensures that we produce a partial tridiagonalization of $A$. Specific methods of ensuring the Lanczos vectors remain “sufficiently orthogonal” will be discussed in section 2.5, however, they all follow a similar format, given below in Algorithm 3.

Note that the only distinction between Algorithms 2 and 3 are lines 6 through 8 in

**Algorithm 3** Modified Lanczos Algorithm

```
1:  Initialize $v_1 = v/\|v\|$, $\beta_0 v_0 = 0$.
2:  for $j = 1, \dots, m$ do
3:      $\tilde{v} = Av_j - \beta_{j-1}v_{j-1}$
4:      $\alpha_j = (\tilde{v}, v_j)$
5:      $\tilde{v} \leftarrow \tilde{v} - \alpha_j v_j$
6:      if $\tilde{v}$ requires additional orthogonalization then
7:          modify $\tilde{v}$
8:      end if
9:      $\beta_j = \|\tilde{v}\|$
10:     if $\beta_j = 0$ then stop
11:     else
12:         $v_{j+1} = \frac{\tilde{v}}{\beta_j}$
13:     end if
14: end for
```

Algorithm 3. This is where we ensure the new created Lanczos vector, $v_{j+1}$, is "sufficiently orthogonal" to the previous Lanczos vectors $v_1, \dots, v_j$. In fact, Algorithm 3 is the same as Algorithm 2 if we always assume $\tilde{v}$ does not need additional orthogonalization. Most importantly, both algorithms produce all quantities in (2.13), with the exception of the roundoff matrix $F_m$. Next, we discuss what "sufficiently orthogonal" entails, in order to ensure the symmetric tridiagonal matrix produced in Algorithm 3 is a partial tridiagonalization of $A$.

In what follows, we need to perform a QR-factorization of $V_m = [v_1 \dots v_m]$, and so we need conditions which ensure the linear independence of the Lanczos vectors produced by Algorithm 3. For this we give a simple lemma from linear algebra.

**Lemma 1.** *Let $V \in \mathbb{R}^{n \times m}$ have columns of unit length, and define $K = V^T V$ and $\kappa = \max_{1 \le i,j \le m} |k_{ij} - \delta_{ij}|$. If $\kappa < (m-1)^{-1}$, then the columns of $V$ are linearly independent.*

*Proof.* By the definition of $K$, the columns of $V$ are linearly independent if and only if $K$ is invertible. Furthermore, $K$ is invertible if and only if it has strictly positive eigenvalues. Therefore, any condition which ensures the positivity of the eigenvalues of $K$, also ensures $V$ is full rank. To characterize the eigenvalues of $K$, we apply Gershgorin's circle theorem, which states that each eigenvalue, $\lambda$, of $K$ satisfies $|\lambda - k_{ii}| \le$

$\sum_{j \neq i} |k_{ij}|$ for some $1 \leq i \leq m$. Therefore, the eigenvalues of $K$ satisfy

$$1 - (m-1)\kappa \leq \lambda \leq 1 + (m-1)\kappa.$$

Note in the application of Gershgorin's circle theorem we used $|k_{ij}| \leq \kappa$ for $i \neq j$ and $k_{ii} = 1$ (follows from the assumption that the columns of $V$ have unit norm). It immediately follows that the eigenvalues of $K$ are strictly positive if $\kappa < (m-1)^{-1}$. □

Applying Lemma 1 with $V_m$ and $K_m = V_m^T V_m$, we see that the Lanczos vectors (the columns of $V_m$) are linearly independent if $\kappa_m < (m-1)^{-1}$, where $\kappa_m$ is given by (2.18). Note, this is a very weak condition. Indeed, for a standard problem, $m$ is of order $10^3$ or less, and so $\kappa_m$ can grow to the level of $10^{-3}$ with linear independence of the Lanczos vectors intact. This shows that while the Lanczos vectors may become exceedingly non-orthogonal, they remain linearly independent. In the following theorem we make the stronger assumption $\kappa_m \leq \sqrt{\epsilon/m}$, and assume this implies the linear independence of the Lanczos vectors (holds so long as $\epsilon < m/(m-1)^2 \sim 1/m$, which is true in all practical scenarios).

Next, we state the fundamental theorem, due to Simon [54], about partially tridiagonalizing a matrix using the Lanczos algorithm in finite precision.

**Theorem 4.** *Let $A \in \mathbb{R}^{n \times n}$ be a symmetric matrix, $v_1 \in \mathbb{R}^n$ be a unit vector and assume $T_m \in \mathbb{R}^{m \times m}$, $V_m \in \mathbb{R}^{n \times m}$, $\beta_m \in \mathbb{R}$, and $v_{m+1} \in \mathbb{R}^n$ are produced by Algorithm 3, i.e., they satisfy $AV_m = V_m T_m + \beta_m v_{m+1} e_m^T + F_m$, where the entries of $F_m$ are of order $\mathcal{O}(\epsilon \|A\|)$. Then, if $\kappa_m$, defined as in* (2.18)*, satisfies $\kappa_m \leq \sqrt{\epsilon/m}$, we have*

$$T_m = Q_m^T A Q_m + E_m,$$

*where $V_m = Q_m R_m$ is the exact QR factorization of $V_m$ and the entries of $E_m$ are of order $\mathcal{O}(\epsilon \|A\|)$.*

*Proof.* See [54, 55]. We remark that the $QR$ factorization in the statement of Theorem 4 is well defined due to our assumption that $\kappa_m \leq \sqrt{\epsilon/m}$ implies $\kappa_m < (m-1)^{-1}$ (which guarantees linear independence of the columns of $V_m$). □

Theorem 4 deserves considerable attention. The requirement that $\kappa_m = \mathcal{O}(\epsilon)$ is referred to as “orthogonal to working precision”, while $\kappa_m = \mathcal{O}(\sqrt{\epsilon})$ is referred to as “semi-orthogonality”. Theorem 4 tells us that as long as the Lanczos vectors are semi-orthogonal for steps $j = 1, \ldots, m$, the partial tridiagonalization of $A$, $T_m$, is accurate to order $\mathcal{O}(\epsilon)$. In fact, we do not gain any advantage when keeping the Lanczos vectors orthogonal to working precision. When using the Lanczos algorithm to approximate the spectrum of $A$ using the Rayleigh-Ritz method, the weaker condition of semi-orthogonality ensures that we do not obtain redundant copies of eigenvalues of $A$, which could otherwise be a serious issue.

Another important consideration in Theorem 4 is the starting vector $v_1$. Note that the partial tridiagonalization produced by the Lanczos algorithm depends on both the matrix $A$ and the starting vector $v_1$. Different starting vectors produce different partial tridiagonalizations. In many applications of the Lanczos algorithm, the starting vector $v_1$ is chosen randomly, and so is unimportant. However, in this thesis we will often be interested in producing partial tridiagonalizations of a matrix *with respect to specific starting vectors*. Theorem 4 tells us that the partial tridiagonalization produced by the Lanczos algorithm in finite precision is, up to roundoff error, the same as the partial tridiagonalization with starting vector $q_1$ (the first column of $Q_m$, the $Q$-factor of the exact $QR$ factorization of $V_m$) in the absence of roundoff error, as long as the Lanczos vectors are semi-orthogonal. Because we have assumed the Lanczos vectors are normalized exactly, the first column of $Q_m$ is $v_1$, i.e., $q_1 = v_1$. Therefore, so long as we ensure the Lanczos vectors remain semi-orthogonal, the partial tridiagonalization produced by Algorithm 3 is the one we desire.

Much work has been devoted to dealing with the loss of orthogonality of the Lanczos vectors. In the next section we give a brief overview of a few methods designed to handle this issue.

## 2.5 Orthogonalization Strategies

Several methods have been designed to overcome the loss of orthogonality in the Lanczos vectors. Determining which is best depends on the application at hand. Here we discuss a few methods designed to ensure the Lanczos vectors are at least semi-orthogonal, which ensures that the Lanczos algorithm produces a partial tridiagonalization of $A$

with respect to a supplied starting vector. All of the following methods fit in the framework of Algorithm 3, with each method using different criteria for determining if additional orthogonalization is necessary as well as different methods for producing semi-orthogonal Lanczos vectors.

### 2.5.1 Full Orthogonalization

Full orthogonalization takes the most conservative approach, and always performs additional orthogonalizations (with respect to Algorithm 3). Full orthogonalization entails explicit (Gram–Schmidt) orthogonalization against all previous Lanczos vectors for every iteration. It is one of the simplest ways to maintain a sufficient level of mutual orthogonality among the Lanczos vectors, however, it is also one of the most costly. Full orthogonalization was the method advocated by Lanczos [29] and Wilkinson [64].

Full orthogonalization is essentially applying the Arnoldi algorithm, Algorithm 1, to the symmetric matrix $A$. This method can be expected to retain mutual orthogonality to machine precision. This makes full orthogonalization useful for its simplicity and robustness. However, all computed Lanczos vectors need to be saved in order to form the inner products, which would not otherwise be necessary for some applications, e.g., if we are only interested in producing a partial tridiagonalization of $A$. Furthermore, each iteration involves more work since additional Lanczos vectors are present. The Lanczos method with full orthogonalization is presented in Algorithm 4 [44].

### 2.5.2 Selective Orthogonalization

Here we present a brief and simplified overview of the selective orthogonalization strategy of Parlett and Scott [43, 42]. Their orthogonalization strategy relies heavily on the insights of Paige about the behavior of the Lanczos algorithm in the presence of round-off error. Paige's thesis illustrated that orthogonality between Lanczos vectors is lost precisely when a Ritz value converges to an eigenvalue of $A$. Additionally, Paige showed in which direction orthogonality is lost. Selective orthogonalization takes advantage of this knowledge to retain semi-orthogonality.

The following two results form the basis for selective orthogonalization. Note that the first applies in the case of exact arithmetic while the second takes into account

**Algorithm 4** Lanczos Algorithm (full orthogonalization)

```
Initialize $v_1 = v/\|v\|$, $\beta_0 v_0 = 0$.
for $j = 1, \dots, m$ do
    $\tilde{v} = Av_j - \beta_{j-1} v_{j-1}$
    $\alpha_j = (\tilde{v}, v_j)$
    $\tilde{v} \leftarrow \tilde{v} - \alpha_j v_j$
    for $i = 1, \dots, j$ do
        $\tilde{v} \leftarrow \tilde{v} - (\tilde{v}, v_i) v_i$
    end for
    $\beta_j = \|\tilde{v}\|$
    if $\beta_j = 0$ then stop
    else
        $v_{j+1} = \frac{\tilde{v}}{\beta_j}$
    end if
end for
```

roundoff error.

**Theorem 5.** *Assume the Lanczos algorithm (Algorithm 2) for $j$ steps is conducted in exact arithmetic, i.e., we have computed the symmetric tridiagonal matrix $T_j \in \mathbb{R}^{j \times j}$ and the orthonormal Lanczos vectors $v_1, \dots, v_{j+1}$ such that*

$$AV_j = V_j T_j + \beta_j v_{j+1} e_j^T,$$

*where $V_j = [v_1 \dots v_j]$ and $\beta_j = \|(AV_j - V_j T_j) e_j\|$. Furthermore, assume the exact eigendecomposition of $T_j$ is $T_j = Y \Theta Y^T$, $y_k^T y_l = \delta_{kl}$, $k, l = 1, \dots, j$. Then, for each Ritz value $\theta_i$, there is a corresponding eigenvalue, $\lambda_{i'}$, of $A$, such that*

$$|\lambda_{i'} - \theta_i| \leq \beta_j |y_{ji}|, \quad i = 1, \dots, j,$$

*where $y_{ji} = e_j^T y_i$.*

**Theorem 6.** *Assume the Lanczos algorithm (Algorithm 2) for $j$ steps is conducted in finite precision, i.e., we have computed the symmetric tridiagonal matrix $T_j \in \mathbb{R}^{j \times j}$ and the Lanczos vectors $v_1, \dots, v_{j+1}$ such that*

$$AV_j = V_j T_j + \beta_j v_{j+1} e_j^T + F_j,$$

*where* $V_j = [v_1 \dots v_j]$, $\beta_j = \|(AV_j - V_jT_j)e_j\|$, *and* $F_j$ *is a roundoff matrix (columns of norm* $\mathcal{O}(\epsilon\|A\|)$*). Furthermore, assume the exact eigendecomposition of* $T_j$ *is* $T_j = Y\Theta Y^T$, $y_k^T y_l = \delta_{kl}$, $k, l = 1, \dots, j$, *and denote the Ritz vectors as* $z_i = V_j y_i$, $i = 1, \dots, j$. *Then,*

$$(z_i, v_{j+1}) = \frac{\gamma_i}{\beta_j |y_{ji}|}, \quad i = 1, \dots, j,$$

*where* $y_{ji} = e_j^T y_i$ *and the* $\gamma_i$*'s are of order* $\mathcal{O}(\epsilon\|A\|)$*.*

Theorem 5 [26, 43] shows that if the last entry of a normalized eigenvector of $T_j$ is small, then a Ritz value is close to an eigenvalue of $A$. Note that the quantity $\beta_j|y_{ji}|$ is equal to the norm of the residual of the Ritz pair as an eigenpair of $A$, i.e., $\|Az_i - \theta_i z_i\| = \beta_j|y_{ji}|$. While Theorem 5 is stated in exact arithmetic, a similar inequality holds with a slightly different constant when roundoff error is taken into consideration and the Lanczos vectors are no longer perfectly orthogonal [43]. The authors of [43] say the bounds in Theorem 5 “fail gracefully” when the Lanczos vectors are no longer orthonormal. Theorem 6, due to Paige [39], shows that when a Ritz value is near an eigenvalue, the Lanczos vector loses orthogonality precisely in the direction of the corresponding Ritz vector. Note that in exact arithmetic $(z_i, v_{j+1}) = 0$. We can use this knowledge to our advantage when orthogonalizing Lanczos vectors. When computing a new Lanczos vector, $v_{j+1}$, we can use the eigenvectors of $T_j$ to check if a Ritz value has converged to an eigenvalue of $A$ via Theorem 5. If so, according to Theorem 6, we should orthogonalize the Lanczos vector against the corresponding Ritz vector. This is summarized in Algorithm 5 where $\tau$, a user defined tolerance, should be $\mathcal{O}(\sqrt{\epsilon})$ in order to retain semi-orthogonality.

In Algorithm 5, the spectral decomposition is computed every step. Since $m \ll n$, and there are specialized algorithms for computing the eigenpairs of a symmetric tridiagonal matrix, this is not too cumbersome. However, there are several different ways to relax this requirement, see [43] for more details.

### 2.5.3 Partial Orthogonalization

Partial orthogonalization, introduced by Simon [54, 56], approximates the solution to the difference equation in Theorem 3 to monitor mutual orthogonality of Lanczos vectors. If orthogonality drops below a certain threshold, then extra steps are taken to

**Algorithm 5** Lanczos Algorithm (selective orthogonalization)

```
Initialize tolerance $\tau > 0$, $v_1 = v/\|v\|$, $\beta_0 v_0 = 0$.
for $j = 1, \ldots, m$ do
    $\tilde{v} = Av_j - \beta_{j-1} v_{j-1}$
    $\alpha_j = (\tilde{v}, v_j)$
    $\tilde{v} \leftarrow \tilde{v} - \alpha_j v_j$
    $\tilde{\beta}_j = \|\tilde{v}\|$
    Diagonalize $T_j$, $T_j = Y \Theta Y^T$, $Y^T Y = I$
    Determine $\mathcal{I} = \{1 \le i \le j \mid \tilde{\beta}_j |y_{ji}| < \tau\}$
    for $i \in \mathcal{I}$ do
        $z_i = V_j y_i$
        $\tilde{v} \leftarrow \tilde{v} - \frac{(\tilde{v}, z_i)}{\|z_i\|^2} z_i$
    end for
    $\beta_j = \|\tilde{v}\|$
    if $\beta_j = 0$ then stop
    else
        $v_{j+1} = \frac{\tilde{v}}{\beta_j}$
    end if
end for
```

orthogonalize the current Lanczos vector against previous Lanczos vectors. By monitoring the level of orthogonality each iteration, we are able to keep the Lanczos vectors semi-orthogonal.

Let $k_{ij} = v_i^T v_j$, as in Theorem 3. Using the Lanczos vectors we can compute $k_{ij}$ up to roundoff error, however this requires many inner products and is the same work required in full orthogonalization. In order to be more efficient, we approximate the terms $k_{ij}$. The main issue in approximating $k_{ij}$ is the roundoff terms $v_j^T f_i - v_i^T f_j$. Since we cannot approximate the roundoff terms, but we know their order of magnitude is $\mathcal{O}(\epsilon\|A\|)$, we use random numbers in their stead. Denote the approximation to $k_{ij}$ as $\tilde{k}_{ij}$. We modify the recurrence relation in Theorem 3 so that $\tilde{k}_{ij}$ satisfies

$$\begin{aligned} \tilde{k}_{ii} &= 1 & i &= 1, \ldots, m+1, \\ \tilde{k}_{i\,i+1} &= \zeta_i & i &= 1, \ldots, m, \\ \beta_j \tilde{k}_{i\,j+1} &= \beta_i \tilde{k}_{i+1\,j} + (\alpha_i - \alpha_j)\tilde{k}_{ij} + \beta_{i-1}\tilde{k}_{i-1\,j} - \beta_{j-1}\tilde{k}_{i\,j-1} + \eta_{ij}, \end{aligned} \tag{2.19}$$

for $1 \le i < j \le m$, $\tilde{k}_{0j} = 0$, where $\zeta_i$ and $\eta_{ij}$ are random numbers chosen from appropriate distributions, e.g., $\zeta_i \in \mathcal{N}(0, \epsilon)$ and $\eta_{ij} \in \mathcal{N}(0, \epsilon\|A\|)$, where $\mathcal{N}(\mu, \sigma)$ is the normal distribution of mean $\mu$ and standard deviation $\sigma$.

In partial orthogonalization, at each iteration $1 \leq j \leq m$ we compute $\tilde{k}_{i\,j+1}$, $1 \leq i \leq j+1$, as an approximation to $k_{i\,j+1} = v_i^T v_{j+1}$, according to (2.19). Once $\tilde{k}_{i\,j+1}$ reaches a user specified tolerance, $\tau$, for some $1 \leq i \leq j$, we orthogonalize against all previous Lanczos vectors $v_1, \ldots, v_j$. The tolerance $\tau$ should be $\mathcal{O}(\sqrt{\epsilon})$ to maintain semi-orthogonality. Notice, however, it is insufficient to orthogonalize against all previous Lanczos vectors for just one iteration. Indeed, orthogonalizing against all previous Lanczos vectors at iteration $j$ implies $k_{i\,j+1} = v_i^T v_{j+1}$ is order $\mathcal{O}(\epsilon)$ for $i = 1, \ldots, j$. For the next step $j+1$, Theorem 3 tells us that

$$\beta_{j+1} k_{i\,j+2} = -\beta_j k_{i\,j} + \mathcal{O}(\epsilon\|A\|).$$

If $\tilde{k}_{i\,j+1} \approx k_{i\,j+1}$ reached the tolerance $\tau$, it is likely $\tilde{k}_{ij} \approx k_{ij}$ is also close to the tolerance. Accordingly, it is necessary to orthogonalize against all previous Lanczos vectors for two consecutive iterations. This brings the level of orthogonality at the next iteration down to machine precision, and we may proceed to use the standard Lanczos algorithm until orthogonality again deprecates to the tolerance $\tau$.

After the orthogonalizations are performed, the values $\tilde{k}_{i\,j+1}$ need to be updated. In perfect arithmetic they would be zero, however we need to take into consideration roundoff error. The author in [56] undertook a statistical study and found that replacing the $\tilde{k}_{i\,j+1}$ with values from the distribution $\mathcal{N}(0, 1.5\epsilon)$ performed well after reorthgonalization. The Lanczos algorithm with partial orthogonalization is given in Algorithm 6.

It is also possible to orthogonalize against select Lanczos vectors, instead of all previous vectors. However, we will not pursue the particulars here. For more details, see [56].

## 2.6 B-Lanczos Algorithm

In this section we discuss the generalization of the Lanczos algorithm to the case of partially tridiagonalizing a pair of matrices $A$ and $B$, where $A$ is symmetric and $B$ is symmetric positive definite. This extension is of primary interest when the matrices $A$ and $B$ define a generalized eigenvalue problem. Previously, the Lanczos algorithm

**Algorithm 6** Lanczos Algorithm (partial orthogonalization)

```
Initialize tolerance $\tau > 0$, $v_1 = v/\|v\|$, $\beta_0 v_0 = 0$.
for $j = 1, \ldots, m$ do
    $\tilde{v} = Av_j - \beta_{j-1} v_{j-1}$
    $\alpha_j = (\tilde{v}, v_j)$
    $\tilde{v} \leftarrow \tilde{v} - \alpha_j v_j$
    if orthogonalized previous iteration then                      Must
        for $i = 1, \ldots, j$ do                                  orthogonalize
            $\tilde{v} \leftarrow \tilde{v} - (\tilde{v}, v_i) v_i$                consecutive
            Update $\tilde{k}_{i\,j+1}$                                   iterations.
        end for
    else
        $\tilde{\beta}_j = \|\tilde{v}\|$
        Compute $\tilde{k}_{i\,j+1}$ according to (2.19) with $\beta_j = \tilde{\beta}_j$.
        if $\tilde{k}_{i\,j+1} > \tau$ for any $i = 1, \ldots, j$ then
            for $i = 1, \ldots, j$ do                              Approximate
                $\tilde{v} \leftarrow \tilde{v} - (\tilde{v}, v_i) v_i$            $v_i^T v_{j+1}$ and
                Update $\tilde{k}_{i\,j+1}$                               orthogonalize
            end for                                                if necessary.
        end if
    end if
    $\beta_j = \|\tilde{v}\|$
    if $\beta_j = 0$ then stop
    else
        $v_{j+1} = \frac{\tilde{v}}{\beta_j}$
    end if
end for
```

produced a partial tridiagonalization of a single matrix, from which spectral properties can be approximated with Ritz values and Ritz vectors. In this section, we extend the Lanczos algorithm so that the Ritz values and Ritz vectors now approximate the eigenpairs of a generalized eigensystem. The algorithms in this section can be found in [5, 49]. Throughout this section we assume all operations are done in exact arithmetic. The effects of roundoff error being similar to the case previously discussed.

Let $A, B \in \mathbb{R}^{n\times n}$ be symmetric with $B$ positive definite. We are interested in determining a symmetric tridiagonal matrix with spectrum related to that of the system $Ax = \lambda Bx$, $x^T Bx = 1$. The first step in realizing such a tridiagonal matrix is transforming the generalized system into a standard eigenvalue problem. The simplest way to do this is by working with the matrix $B^{-1}A$. The issue is that $B^{-1}A$ is no longer symmetric. However, it is self-adjoint with respect to the $B$-inner product $(x, y)_B \coloneqq x^T By$. This allows us to use the standard Lanczos algorithm with the matrix $B^{-1}A$, and the $B$-inner product and induced $B$-norm $\|\cdot\|_B$. Starting with nonzero $v \in \mathbb{R}^n$, the Lanczos recurrence after $m$-steps becomes

$$B^{-1}AV_m = V_m T_m + \beta_m v_{m+1} e_m^T, \tag{2.20}$$

or, equivalently,

$$AV_m = BV_m T_m + \beta_m Bv_{m+1} e_m^T, \tag{2.21}$$

where the columns of $V_m$ are a $B$-orthonormal basis of the Krylov space $\mathcal{K}_m(B^{-1}A, v)$.

First, consider the following naive implementation of the standard Lanczos algorithm with matrix $B^{-1}A$, vector $v$, and $B$-inner product and induced norm. With these modifications, Algorithm 2 becomes Algorithm 7 shown below. We refer to Algorithm 7 as the "naive" $B$-Lanczos algorithm due to the extra computational cost relative to other implementations. In Algorithm 7 there is one matrix vector multiplication with $A$, two matrix vector multiplications with $B$, and one linear solve with $B$. Next, we show how to eliminate the matrix vector multiplications with $B$ at the cost of storing additional vectors.

In order to reduce the costs associated with Algorithm 7 we work with the auxiliary vector $\tilde{w} = B\tilde{v}$, instead of $\tilde{v}$ from Algorithm 7. That is, instead of forming $\tilde{v} = B^{-1}Av_j - \beta_{j-1}v_{j-1}$ at the outset, we form $\tilde{w} = B\tilde{v} = Av_j - \beta_{j-1}Bv_{j-1}$. To illustrate the cost

**Algorithm 7** Naive B-Lanczos Algorithm

```
 1: Initialize $v_1 = v/\|v\|_B$, $\beta_0 v_0 = 0$.
 2: for $j = 1, \ldots, m$ do
 3:     $\tilde{v} = B^{-1}Av_j - \beta_{j-1}v_{j-1}$
 4:     $\alpha_j = (\tilde{v}, v_j)_B$
 5:     $\tilde{v} \leftarrow \tilde{v} - \alpha_j v_j$
 6:     $\beta_j = \|\tilde{v}\|_B$
 7:     if $\beta_j = 0$ then stop
 8:     else
 9:         $v_{j+1} = \frac{\tilde{v}}{\beta_j}$
10:     end if
11: end for
```

savings, we move through one iteration of the $B$-Lanczos algorithm. Let us assume the vectors $w_i = Bv_i$, $i = 1, \ldots, j$, have already been computed and note that $v_k^T w_l = \delta_{kl}$ for $k, l = 1, \ldots, j$. At the beginning of iteration $j$ we form $\tilde{w}$ as

$$\tilde{w} = Av_j - \beta_{j-1}w_{j-1}.$$

Notice that because we are working with $\tilde{w} = B\tilde{v}$, all $B$-inner products with $\tilde{v}$ translate to standard Euclidean inner products with $\tilde{w}$. Line 4 in Algorithm 7 becomes

$$\alpha_j = (\tilde{v}, v_j)_B = (\tilde{w}, v_j).$$

After $\alpha_j$ has been computed, $\tilde{w} = B\tilde{v}$ is updated, $\tilde{w} \leftarrow \tilde{w} - \alpha_j w_j$, giving

$$\tilde{w} = Av_j - \beta_{j-1}w_{j-1} - \alpha_j w_j.$$

Next, at line 6 we compute $\beta_j = \|\tilde{v}\|_B = \sqrt{(\tilde{v}, \tilde{w})}$, for which we need both $\tilde{v}$ and $\tilde{w}$. Hence, it is necessary to solve the linear system $B\tilde{v} = \tilde{w}$ for $\tilde{v}$. Finally, by defining $v_{j+1} = \tilde{v}/\beta_j$ we have the desired Lanczos vector. Note that we also define $w_{j+1} = \tilde{w}/\beta_j$, since we need it the following iteration. In this way, we form a bi-orthogonal system, instead of only the $B$-orthonormal basis as in Algorithm 7.

By working with the vector $\tilde{w} = B\tilde{v}$, rather than $\tilde{v}$ itself, we have eliminated two matrix vector multiplications with $B$. However, in order to utilize this method we also have to store the vectors $w_i = Bv_i$, $i = 1, \ldots, m+1$ for the $m$-step $B$-Lanczos algorithm. By defining $W_m = BV_m$, (2.21) becomes

$$AV_m = W_m T_m + \beta_m w_{m+1} e_m^T, \tag{2.22}$$

where $V_m$ and $W_m$ are bi-orthogonal, i.e., $V_m^T W_m = I$. Thus, we arrive at Algorithm 8.

**Algorithm 8** B-Lanczos Algorithm

```
Initialize $w = Bv$, $\beta_0 = \sqrt{v^T w}$, $v_1 = v/\beta_0$, $w_1 = w/\beta_0$, $w_0 = 0$.
for $j = 1, \ldots, m$ do
    $\tilde{w} = Av_j - \beta_{j-1} w_{j-1}$
    $\alpha_j = (\tilde{w}, v_j)$
    $\tilde{w} \leftarrow \tilde{w} - \alpha_j w_j$
    Solve $B\tilde{v} = \tilde{w}$ for $\tilde{v}$
    $\beta_j = \sqrt{(\tilde{v}, \tilde{w})}$
    if $\beta_j = 0$ then stop
    else
        $v_{j+1} = \frac{\tilde{v}}{\beta_j}$
        $w_{j+1} = \frac{\tilde{w}}{\beta_j}$
    end if
end for
```

Comparing the $B$-Lanczos algorithm (Algorithm 8) and the standard Lanczos algorithm (Algorithm 2), we see that there are two main distinctions. First, for the $B$-Lanczos algorithm we save two sets of vectors which are bi-orthogonal, instead of a single orthonormal basis. Secondly, each iteration of the $B$-Lanczos algorithm requires a linear solve with $B$, which is not present in the standard Lanczos algorithm. Both the standard Lanczos algorithm and the $B$-Lanczos algorithm require one matrix vector multiplication with $A$.

As in the standard Lanczos algorithm, we do not expect the bi-orthogonal basis to remain bi-orthogonal in the presence of roundoff error. At step $j$, we must now ensure the off diagonal entries of $W_j^T V_j$ remain $\mathcal{O}(\sqrt{\epsilon})$ for $T_m$ in (2.22) to be a partial tridiagonalization of the matrix pair $A$ and $B$. All methods mentioned in section 2.5 work for the $B$-Lanczos algorithm with minor modifications. We include the full orthogonalization version of the $B$-Lanczos algorithm, Algorithm 9, for completeness.

**Algorithm 9** $B$-Lanczos Algorithm (full orthogonalization)

```
Initialize $w = Bv$, $\beta_0 = \sqrt{v^T w}$, $v_1 = v/\beta_0$, $w_1 = w/\beta_0$, $w_0 = 0$.
for $j = 1, \ldots, m$ do
    $\tilde{w} = Av_j - \beta_{j-1} w_{j-1}$
    $\alpha_j = (\tilde{w}, v_j)$
    $\tilde{w} \leftarrow \tilde{w} - \alpha_j w_j$
    for $i = 1, \ldots, j$ do
        $\tilde{w} \leftarrow \tilde{w} - (\tilde{w}, v_i) w_i$
    end for
    Solve $B\tilde{v} = \tilde{w}$ for $\tilde{v}$
    $\beta_j = \sqrt{(\tilde{v}, \tilde{w})}$
    if $\beta_j = 0$ then stop
    else
        $v_{j+1} = \frac{\tilde{v}}{\beta_j}$
        $w_{j+1} = \frac{\tilde{w}}{\beta_j}$
    end if
end for
```

# Chapter 3

# The Lanczos Process

## 3.1 Quadratic Forms and Quadrature

We begin this chapter by motivating interest in quadratic forms, $v^T f(A)v$, where $v$ is a given vector, $A$ is a matrix, and $f$ is a function. We give further conditions on $A$, $v$, and $f$ later. For now, suppose we are interested in iteratively approximating the solution to the linear system $Ax = b$, where $A$ is nonsingular and symmetric. Let the approximate solution, after a number of iterations, be denoted by $\tilde{x}$. In order to know if we should accept our iterative solution as satisfactory, we would like to efficiently approximate the error $\|x - \tilde{x}\|$, where $\|\cdot\|$ is the standard Euclidean norm. Notice that we can write the error as $x - \tilde{x} = A^{-1}r$, where $r = b - A\tilde{x}$ is the residual vector. Using the relation between the error and the residual vector we can write

$$\|x - \tilde{x}\|^2 = (x - \tilde{x})^T(x - \tilde{x}) = (A^{-1}r)^T(A^{-1}r) = r^T f(A)r,$$

where $f(\lambda) = \lambda^{-2}$. By approximating the quadratic form $r^T f(A)r$, or by providing upper and lower bounds, we can determine when to stop the iterative linear solver, and consider our approximate solution converged. The methods given in this chapter for approximating quadratic forms were first proposed in [13], and our presentation closely follows that of [17].

To begin, let $A \in \mathbb{R}^{n\times n}$ and $v \in \mathbb{R}^n$ be a given symmetric matrix and unit vector respectively. Due to the symmetry of $A$, we know that all eigenvalues are real, and the eigenvectors form an orthonormal basis for $\mathbb{R}^n$. This allows us to form the orthogonal eigendecomposition $A = X\Lambda X^T$, where $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, $X = [x_1 \ldots x_n]$, and $X^T X = I$. We assume, without loss of generality, that the eigenvalues are arranged in ascending order, i.e., $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$.

For a smooth function, $f$, defined on the real line, the matrix $f(A)$ is defined as $f(A) = Xf(\Lambda)X^T$, with $f(\Lambda) = \operatorname{diag}(f(\lambda_1), \ldots, f(\lambda_n))$, see, e.g., [18]. With the spectral decomposition of $A$, and the definition of the matrix $f(A)$, the quadratic form $v^T f(A)v$ can be expressed as

$$\begin{aligned} v^T f(A)v &= v^T X f(\Lambda) X^T v, \\ &= (X^T v)^T f(\Lambda)(X^T v), \\ &= \sum_{i=1}^{n} |(x_i, v)|^2 f(\lambda_i), \end{aligned} \tag{3.1}$$

where $(\,\cdot\,,\,\cdot\,)$ is the standard Euclidean inner product. From (3.1) we see that the quadratic form $v^T f(A)v$ is completely determined by the magnitude of the components of $v$ in the direction of the eigenvectors of $A$, and $f$ evaluated at the eigenvalues.

While we have made the assumption that $v$ is a unit vector, obviously (3.1) holds for all $n$-vectors $v$. However, the formulas we develop to approximate $v^T f(A)v$ are simplest in the case that $v$ is a unit vector. If we are interested in approximating $u^T f(A)u$ for $\|u\| \neq 1$, we can easily rephrase the problem in terms of the unit vector $v = u/\|u\|$, using the relation $u^T f(A)u = \|u\|^2 v^T f(A)v$.

Next, we prepare to approximate $v^T f(A)v$ using Gaussian quadrature. To this end, we express (3.1) in integral form. Define a measure on the real line, $s(\lambda)$, depending on $v$ and the spectrum of $A$, as

$$s(\lambda) = \sum_{i=1}^{n} |(x_i, v)|^2 \delta(\lambda - \lambda_i), \tag{3.2}$$

where $\delta(\lambda)$ is the Dirac delta distribution concentrated at the value $\lambda$. Using the measure $s(\lambda)$, we can rewrite (3.1) as

$$v^T f(A)v = \int_a^b s(\lambda) f(\lambda) d\lambda, \tag{3.3}$$

where the limits of integration satisfy $a \leq \lambda_1$ and $b \geq \lambda_n$. The measure, or weight, $s(\lambda)$, will be referred to as the spectral function for its obvious relation to the spectrum of $A$.

A few comments are necessary at this point. First, the integral in (3.3) is well-defined, even though the weight is not defined in a pointwise sense. Indeed, the integral in (3.3) can be rewritten as a Riemann–Stieltjes integral

$$\int_a^b f d\mu, \qquad \mu(\lambda) = \begin{cases} 0, & \lambda < \lambda_1, \\ \sum_{i=1}^{k} |(x_i, v)|^2, & \lambda_k \leq \lambda < \lambda_{k+1}, \ k = 1, \ldots, n-1, \\ \sum_{i=1}^{n} |(x_i, v)|^2, & \lambda \geq \lambda_n, \end{cases}$$

where, for simplicity, we have assumed all eigenvalues of $A$ are simple, i.e., $\lambda_1 < \ldots < \lambda_n$. Since $\mu$ is non-decreasing and $f$ is continuous, the Riemann–Stieltjes integral exists [17, 63]. Second, while we are phrasing the problem as approximating the quadratic form $v^T f(A)v$, we are, in fact, approximating the spectral function $s(\lambda)$. This is obvious when looked at from the perspective of quadrature. Because the Gauss quadrature nodes and weights are independent of the integrand, we are approximating the action of $s(\lambda)$ on $f$, for arbitrary $f$, i.e., we are approximating $s(\lambda)$. We discuss in what sense we are approximating $s(\lambda)$ in greater detail in Section 3.3.

Using $s(\lambda)$, we define a (discrete) semi-inner product

$$\langle f, g \rangle_s := \int_a^b s(\lambda) f(\lambda) g(\lambda) d\lambda = (f(A)v, g(A)v), \tag{3.4}$$

and the corresponding induced semi-norm $\|f\|_s = \sqrt{\langle f, f \rangle_s}$, where $f$ and $g$ are smooth functions. A simple example illustrating why $\langle \cdot, \cdot \rangle_s$ is a semi-inner product, and not an inner product, is the minimal polynomial of $A$, i.e., if $p$ is the minimal polynomial of $A$, then $\langle p, p \rangle_s = \|p(A)v\|^2 = 0$ with $p \not\equiv 0$. We return to this point shortly. Most importantly, with our assumptions the semi-inner product $\langle \cdot, \cdot \rangle_s$ is well defined for all smooth functions.

Continuing our quest of approximating $v^T f(A)v$ with Gaussian quadrature, we turn towards determining the nodes and weights for the $m$-point quadrature rule corresponding to the weight $s(\lambda)$. For this, we follow the well known Golub–Welsch algorithm [19].

The Golub–Welsch algorithm determines the nodes and weights of a (weighted) Gauss quadrature rule using the eigenpairs of a certain symmetric tridiagonal matrix called the Jacobi matrix. The entries of the Jacobi matrix are the coefficients of the three term recurrence relation for the orthonormal polynomials with respect to the semi-inner product $\langle \cdot\,, \cdot \rangle_s$. We will see shortly that $\langle \cdot\,, \cdot \rangle_s$ is an inner product, and not just a semi-inner product, on the space of polynomials below a certain degree. To determine the entries of the Jacobi matrix, we construct the family of orthonormal polynomials with respect to $\langle \cdot\,, \cdot \rangle_s$. We begin by forming the family of monic orthogonal polynomials with respect to $\langle \cdot\,, \cdot \rangle_s$, and then specialize to orthonormal polynomials.

First, let $\mathscr{P}$ and $\mathscr{P}_k$ denote the vector space of all polynomials and the vector space of all polynomials of degree less than or equal to $k$ respectively. Similarly, let $\hat{\mathscr{P}}_k$ denote the space of monic polynomials of exact degree $k$, i.e., $\hat{p} \in \hat{\mathscr{P}}_k$ has the form $\hat{p}(\lambda) = \lambda^k + c_{k-1}\lambda^{k-1} + \ldots + c_0$ for some constants $c_i \in \mathbb{R}$, $i = 0, 1, \ldots, k-1$. As mentioned previously, $\langle \cdot\,, \cdot \rangle_s$ is not an inner product on the entire space of polynomials $\mathscr{P}$. This is important because in order to construct the family of monic orthogonal polynomials, the Gram–Schmidt orthogonalization process is utilized, which requires an inner product. To see where $\langle \cdot\,, \cdot \rangle_s$ fails to be an inner product (and on what space it is an inner product), we have the following lemma.

**Lemma 2.** *$\langle \cdot\,, \cdot \rangle_s$ is positive definite on $\mathscr{P}_{\overline{m}-1}$, where $\overline{m} = \text{grade}(v)$.*

*Proof.* From (3.4), for any $p \in \mathscr{P}$, we have $\langle p, p \rangle_s = \|p(A)v\|^2$, and so, $\langle p, p \rangle_s = 0$ if and only if $p(A)v = 0$. By definition, any polynomial $p$ for which $p(A)v = 0$, is divisible by the minimal polynomial of $v$ with respect of $A$, and the degree of the minimal polynomial of $v$ with respect to $A$ is $\overline{m} = \text{grade}(v)$. Hence, for any nonzero polynomial $p \in \mathscr{P}_{\overline{m}-1}$, $p(A)v \neq 0$, and so $\langle p, p \rangle_s > 0$. □

Because $\langle \cdot\,, \cdot \rangle_s$ is positive definite on $\mathscr{P}_k$ for $k < \overline{m} = \text{grade}(v)$, we know there exists a finite family of monic orthogonal polynomials, $\hat{p}_k \in \hat{\mathscr{P}}_k$, $k = 0, 1, \ldots, \overline{m} - 1$, with respect to the inner product $\langle \cdot\,, \cdot \rangle_s$ [16]. Additionally, these polynomials form a basis of $\mathscr{P}_{\overline{m}-1}$. For convenience, we define $\hat{p}_{-1} = 0$.

A useful feature of monic orthogonal polynomials is that they satisfy a three-term recurrence. To see that monic orthogonal polynomials satisfy a three-term recurrence, notice that $\hat{p}_{k+1}(\lambda) - \lambda\hat{p}_k(\lambda)$ has degree less than or equal to $k$, and so

$$\hat{p}_{k+1} - \lambda\hat{p}_k = -\hat{\alpha}_{k+1}\hat{p}_k - \hat{\beta}_k\hat{p}_{k-1} + \sum_{i=0}^{k-2} \hat{c}_i\hat{p}_i, \tag{3.5}$$

for some constants $\hat{\alpha}_{k+1}$, $\hat{\beta}_k$, and $\hat{c}_i$, $i = 0, \ldots, k-2$. Taking the inner product of (3.5) with $\hat{p}_k$ and $\hat{p}_{k-1}$, and exploiting orthogonality, gives

$$\hat{\alpha}_{k+1} = \frac{\langle \lambda\hat{p}_k, \hat{p}_k\rangle_s}{\langle \hat{p}_k, \hat{p}_k\rangle_s} \qquad k = 0, \ldots, \overline{m} - 2, \tag{3.6}$$

and,

$$\hat{\beta}_k = \frac{\langle \lambda\hat{p}_k, \hat{p}_{k-1}\rangle_s}{\langle \hat{p}_{k-1}, \hat{p}_{k-1}\rangle_s} = \frac{\langle \hat{p}_k, \lambda\hat{p}_{k-1}\rangle_s}{\langle \hat{p}_{k-1}, \hat{p}_{k-1}\rangle_s} = \frac{\langle \hat{p}_k, \hat{p}_k\rangle_s}{\langle \hat{p}_{k-1}, \hat{p}_{k-1}\rangle_s}, \tag{3.7}$$

for $k = 1, \ldots, \overline{m} - 2$. Lastly, to see that $\hat{c}_j = 0$, take the inner product of (3.5) with $\hat{p}_l$, for some $0 \leq l \leq k-2$, to get

$$\hat{c}_l = \frac{\langle \lambda\hat{p}_k, \hat{p}_l\rangle_s}{\langle \hat{p}_l, \hat{p}_l\rangle_s} = \frac{\langle \hat{p}_k, \lambda\hat{p}_l\rangle_s}{\langle \hat{p}_l, \hat{p}_l\rangle_s} = 0, \tag{3.8}$$

where the last equality in (3.8) holds because $\lambda\hat{p}_l(\lambda)$ is a polynomial of degree strictly less than $k$. Therefore, the family of monic orthogonal polynomials with respect to $\langle\,\cdot\,,\,\cdot\,\rangle_s$ satisfies

$$\hat{p}_{k+1} = (\lambda - \hat{\alpha}_{k+1})\hat{p}_k - \hat{\beta}_k\hat{p}_{k-1} \quad k = 0, 1, \ldots, \overline{m} - 2, \tag{3.9}$$

with $\hat{p}_{-1} = 0$, $\hat{p}_0 = 1$, and $\hat{\alpha}_{k+1}$ and $\hat{\beta}_k$ given by (3.6) and (3.7) respectively. Note that we do not need to define $\hat{\beta}_0$ since it multiplies $\hat{p}_{-1} = 0$.

Next, we obtain the family of orthonormal polynomials with respect to $\langle\,\cdot\,,\,\cdot\,\rangle_s$ by normalization, i.e., we define $p_k(\lambda) = \hat{p}_k(\lambda)/\|\hat{p}_k\|_s$. In order to determine the recurrence satisfied by the $p_k$'s, and therefore the elements of the Jacobi matrix, we manipulate (3.9). Writing $\hat{p}_j(\lambda) = p_j(\lambda)\|\hat{p}_j\|_s$ for $j = k-1$, $k$, $k+1$, in (3.9), and dividing through by $\|\hat{p}_k\|_s$, we obtain

$$\begin{aligned}\frac{\|\hat{p}_{k+1}\|_s}{\|\hat{p}_k\|_s}\,p_{k+1} &= (\lambda-\hat{\alpha}_{k+1})p_k - \hat{\beta}_k\frac{\|\hat{p}_{k-1}\|_s}{\|\hat{p}_k\|_s}\,p_{k-1},\\ &= (\lambda-\hat{\alpha}_{k+1})p_k - \frac{\|\hat{p}_k\|_s}{\|\hat{p}_{k-1}\|_s}\,p_{k-1},\end{aligned} \tag{3.10}$$

where in the last line of (3.10) we used the fact that $\hat{\beta}_k = \|\hat{p}_k\|_s^2/\|\hat{p}_{k-1}\|_s^2$ (see (3.7)). Writing $\beta_k = \|\hat{p}_k\|_s/\|\hat{p}_{k-1}\|_s = \sqrt{\hat{\beta}_k}$ and $\alpha_{k+1} = \hat{\alpha}_{k+1}$, we arrive at

$$\beta_{k+1}p_{k+1} = (\lambda-\alpha_{k+1})p_k - \beta_k p_{k-1} \quad k = 0,1,\ldots,\overline{m}-2. \tag{3.11}$$

Note that $p_0(\lambda) = \hat{p}_0(\lambda) = 1$ because we have assumed $v$ is a unit vector, and therefore $\langle 1,1\rangle_s = \|v\|^2 = 1$. Again, we define $p_{-1} = 0$ for convenience, and for this reason do not need to specify $\beta_0$.

With the (finite) family of orthonormal polynomials with respect to $\langle\,\cdot\,,\,\cdot\,\rangle_s$ defined, we are ready to define the Jacobi matrix corresponding to the weight $s(\lambda)$. Letting $P_m = P_m(\lambda) = [p_0(\lambda),\ldots,p_{m-1}(\lambda)]^T$ for $m < \overline{m}$, we can rewrite (3.11) in vector form as

$$\lambda P_m = J_m P_m + \beta_m p_m e_m, \quad J_m = \begin{pmatrix}\alpha_1 & \beta_1 & & \\ \beta_1 & \ddots & \ddots & \\ & \ddots & & \beta_{m-1}\\ & & \beta_{m-1} & \alpha_m\end{pmatrix}, \tag{3.12}$$

where $e_m$ is the last column of the $m\times m$ identity matrix and $J_m$ is the Jacobi matrix of order $m$ corresponding to the weight $s(\lambda)$. It is well known that the nodes for an $m$-point weighted Gaussian quadrature rule are the distinct roots of the degree $m$ orthonormal polynomial with respect to the weighted inner product. In our case, denote the roots of $p_m(\lambda)$ as $\theta_1 < \ldots < \theta_m$. From (3.12), it is immediate that the roots of $p_m$ are the eigenvalues of the Jacobi matrix, and the corresponding (unnormalized) eigenvectors of $J_m$ are $P_m(\theta_j)$, $j = 1,\ldots,m$. Next, as proved in [19], we show that the weights for the $m$-point Gaussian quadrature rule are the square of the first component of the normalized eigenvectors of the Jacobi matrix.

Given the quadrature nodes $\{\theta_j\}_{j=1}^m$, the weights, $\{w_j\}_{j=1}^m$, for any interpolatory quadrature rule (which includes Gaussian quadrature), are given by

$$w_j = \int_a^b s(\lambda) l_j(\lambda) d\lambda, \quad l_j(\lambda) = \prod_{\substack{i=1 \\ i \neq j}}^m \frac{\lambda - \theta_i}{\theta_j - \theta_i}, \quad j = 1, \ldots, m, \tag{3.13}$$

see, e.g., [2, 16]. The polynomials $l_j(\lambda)$ in (3.13) are the Lagrange polynomials defined by the nodes $\theta_1, \ldots, \theta_m$, and satisfy $l_j(\theta_i) = \delta_{ij}$. Before relating the quadrature weights to the eigenvectors of the Jacobi matrix, we first need a useful lemma which can be found, e.g., in [16, 17].

**Lemma 3** (Christoffel–Darboux formula)**.** *Let $p_k \in \mathscr{P}_k$, $k = 0, 1, \ldots,$ be a family of polynomials satisfying the recurrence*

$$\beta_{k+1} p_{k+1}(\lambda) = (\lambda - \alpha_{k+1}) p_k(\lambda) - \beta_k p_{k-1}(\lambda) \quad k = 0, 1, \ldots, \tag{3.14}$$

*for some $\alpha_k, \beta_k \in \mathbb{R}$, $k = 1, 2, \ldots,$ with $p_{-1} = 0$ (note that $\beta_0$ need not be defined since it multiplied $p_{-1}$). Then, for any $m \geq 0$,*

$$\sum_{k=0}^m p_k(\lambda) p_k(\theta) = \beta_{m+1} \frac{p_{m+1}(\lambda) p_m(\theta) - p_m(\lambda) p_{m+1}(\theta)}{\lambda - \theta}, \quad \lambda \neq \theta, \tag{3.15}$$

$$\sum_{k=0}^m p_k(\lambda)^2 = \beta_{m+1} \Big( p'_{m+1}(\lambda) p_m(\lambda) - p'_m(\lambda) p_{m+1}(\lambda) \Big), \tag{3.16}$$

*where*

$$p'(\lambda) = \frac{d}{d\lambda} p(\lambda).$$

*Proof.* Multiplying (3.14) by $p_k(\theta)$ gives

$$\beta_{k+1} p_{k+1}(\lambda) p_k(\theta) = (\lambda - \alpha_{k+1}) p_k(\lambda) p_k(\theta) - \beta_k p_{k-1}(\lambda) p_k(\theta). \tag{3.17}$$

Subtract from (3.17) the same expression with the roles of $\lambda$ and $\theta$ reversed, and rearrange to get

$$(\lambda - \theta)p_k(\lambda)p_k(\theta) = \beta_{k+1}\Big[p_{k+1}(\lambda)p_k(\theta) - p_k(\lambda)p_{k+1}(\theta)\Big] \\ - \beta_k\Big[p_k(\lambda)p_{k-1}(\theta) - p_{k-1}(\lambda)p_k(\theta)\Big]. \qquad (3.18)$$

Summing (3.18) from $k = 0$ to $k = m$, taking into consideration the telescoping structure and the definition $p_{-1} = 0$, establishes (3.15). Taking the limit of (3.15) as $\theta \to \lambda$ gives (3.16). □

The Christoffel–Darboux formulae are key in relating the Gaussian quadrature weights to the eigenvectors of the Jacobi matrix, as we show next.

**Theorem 7** (Golub–Welsch). *Let $J_m \in \mathbb{R}^{m\times m}$, $m \geq 1$, be the symmetric tridiagonal Jacobi matrix corresponding to the weight $s(\lambda)$ with diagonal entries $\alpha_j$, $j = 1, \ldots, m$, and positive super/sub-diagonal entries $\beta_j$, $j = 1, \ldots, m-1$. Let $\theta_j$, $j = 1, \ldots, m$, be the eigenvalues of $J_m$, and $y_j \in \mathbb{R}^m$, $j = 1, \ldots, m$, the corresponding normalized eigenvectors. The nodes and weights for the $m$-point Gaussian quadrature rule corresponding to the weight $s(\lambda)$ are given by $\theta_j$ and $w_j = |(y_j, e_1)|^2$, $j = 1, \ldots, m$, respectively.*

*Proof.* The order $m$ Jacobi matrix satisfies (3.12), from which it is immediate that the nodes for the quadrature rule are the distinct eigenvalues of $J_m$, with corresponding (unnormalized) eigenvectors $P_m(\theta_j) = [p_0(\theta_j), p_1(\theta_j), \ldots, p_{m-1}(\theta_j)]^T$, $j = 1, \ldots, m$. The nodes for $m$-point Gaussian quadrature are the roots of $p_m(\lambda)$, and therefore we can write $p_m(\lambda) = c\prod_{j=1}^{m}(\lambda - \theta_j)$, for some nonzero constant $c$. Notice that the Lagrange polynomials, $l_j(\lambda)$, corresponding to the roots of $p_m$ (defined in (3.13)), can be written as

$$l_j(\lambda) = \frac{p_m(\lambda)}{p_m'(\theta_j)(\lambda - \theta_j)}, \qquad j = 1, \ldots, m. \qquad (3.19)$$

Apply the first identity in Lemma 3 with $\theta = \theta_j$, noting that $p_m(\theta_j) = 0$, and solve for $p_m(\lambda)/(\lambda - \theta_j)$ to get

$$\frac{p_m(\lambda)}{(\lambda - \theta_j)} = \frac{-1}{\beta_{m+1}p_{m+1}(\theta_j)} \sum_{k=0}^{m-1} p_k(\theta_j)p_k(\lambda). \qquad (3.20)$$

Rewriting (3.19) using (3.20), the Lagrange polynomials become

$$l_j(\lambda) = \frac{-1}{\beta_{m+1} p'_m(\theta_j) p_{m+1}(\theta_j)} \sum_{k=0}^{m-1} p_k(\theta_j) p_k(\lambda), \qquad j = 1, \ldots, m. \tag{3.21}$$

With an explicit expression of the Lagrange polynomial in terms of the orthonormal polynomials, can solve for the quadrature weights $w_j$, $j = 1, \ldots, m$, using (3.13)

$$\begin{aligned} w_j = \int_a^b s(\lambda) l_j(\lambda) d\lambda &= \frac{-1}{\beta_{m+1} p'_m(\theta_j) p_{m+1}(\theta_j)} \sum_{k=0}^{m-1} p_k(\theta_j) \underbrace{\int_a^b s(\lambda) p_k(\lambda) d\lambda}_{\delta_{k0}}, \\ &= \frac{-1}{\beta_{m+1} p'_m(\theta_j) p_{m+1}(\theta_j)}, \end{aligned} \tag{3.22}$$

where in the last line we used $p_0(\lambda) = 1$. Next, using the second identity in Lemma 3 with $\lambda = \theta_j$, and again using $p_m(\theta_j) = 0$, we can relate the norm of the (unnormalized) eigenvector and quadrature weight as follows:

$$\begin{aligned} \|P_m(\theta_j)\|^2 &= P_m(\theta_j)^T P_m(\theta_j), \\ &= \sum_{k=0}^{m-1} p_k(\theta_j)^2, \\ &= \sum_{k=0}^{m-1} p_k(\theta_j)^2 + p_m(\theta_j)^2, \\ &= -\beta_{m+1} p'_m(\theta_j) p_{m+1}(\theta_j), \\ &= \frac{1}{w_j}, \end{aligned} \tag{3.23}$$

where the last equality follows from (3.22). Solving (3.23) for the quadrature weight gives $w_j = \|P_m(\theta_j)\|^{-2}$. By definition, $y_j = P_m(\theta_j)/\|P_m(\theta_j)\|$, and therefore

$$y_j = \sqrt{w_j} P_m(\theta_j). \tag{3.24}$$

Equation (3.24) shows that we can determine $w_j$ by equating any nonzero component of the vectors $y_j$ and $\sqrt{w_j} P_m(\theta_j)$, with the simplest being the first component since

$p_0(\lambda) = 1$ for all $\lambda$. Taking the inner product of (3.24) with $e_1$ and squaring gives $w_j = |(y_j, e_1)|^2$ for $j = 1, \ldots, m$. □

From Theorem 7, if the Jacobi matrix $J_m$ of order $m$ has eigenvalues $\theta_j$, and corresponding normalized eigenvectors $y_j$, $j = 1, \ldots, m$, the $m$-point Gaussian quadrature approximation to $v^T f(A)v$ is given by

$$v^T f(A)v = \sum_{i=1}^{n} |(x_i, v)|^2 f(\lambda_i) \approx \sum_{j=1}^{m} |(y_j, e_1)|^2 f(\theta_j) = e_1^T f(J_m) e_1. \tag{3.25}$$

According to (3.25), the quadratic form $v^T f(A)v$ is well-approximated by the entry in the first row and first column of the matrix $f(J_m)$.

As is well known, an $m$-point Gaussian quadrature rule is exact for polynomials of degree $2m - 1$. Accordingly, $v^T f(A)v = e_1^T f(J_m)e_1$ if $f \in \mathscr{P}_{2m-1}$. For $f \notin \mathscr{P}_{2m-1}$, standard error estimates for Gaussian quadrature apply. For example, assuming $f$ has $2m$ continuous derivatives, a standard error estimate is

$$v^T f(A)v - e_1^T f(J_m)e_1 = \frac{f^{(2m)}(\xi)}{(2m)!} \int_a^b s(\lambda) \left( \prod_{j=1}^{m} (\lambda - \theta_j)^2 \right) d\lambda, \tag{3.26}$$

for some $\xi \in (a, b)$, see [2, 17, 18]. Note that $\prod_{j=1}^{m} (\lambda - \theta_j)^2$ is proportional to $p_m(\lambda)^2$. In fact, it is not difficult to show that $\prod_{j=1}^{m} (\lambda - \theta_j) = \beta_1 \cdots \beta_m p_m(\lambda)$, and by the orthonormality of $p_m(\lambda)$, the integral in (3.26) is

$$\int_a^b s(\lambda) \left( \prod_{j=1}^{m} (\lambda - \theta_j)^2 \right) d\lambda = \prod_{j=1}^{m} \beta_j^2. \tag{3.27}$$

So, if the $\beta_j$'s are known along with bounds on $f^{(2m)}$, we can quantify the error accurately.

In summary, to approximate the quadratic form $v^T f(A)v$ using Gaussian quadrature, we first compute the Jacobi matrix for the inner product $\langle \cdot , \cdot \rangle_s$, and then the quadrature rule gives $v^T f(A)v \approx \sum_{j=1}^{m} |(y_j, e_1)|^2 f(\theta_j)$, where $\theta_j$ and $y_j$ are the eigenvalues and normalized eigenvectors of the Jacobi matrix. The main problem is that the weight, $s(\lambda)$, is defined in terms of the spectrum of $A$, which is unknown and, in general, difficult to compute. Next, we overcome this hurdle by relating the Jacobi matrix, $J_m$, to the Lanczos algorithm.

## 3.2 Lanczos Polynomials and Lanczos Vectors

As the last section showed, the most important element for constructing the Gaussian quadrature approximation of the quadratic form, $v^T f(A)v$, is the Jacobi matrix. The nodes and weights of the $m$-point quadrature rule are completely determined by the eigenpairs of the Jacobi matrix of order $m$ corresponding to the semi-inner product $\langle\,\cdot\,,\,\cdot\,\rangle_s$. However, one serious issue remains. Namely, the Jacobi matrix is determined by an unknown measure $s(\lambda)$. In order to overcome this hurdle, we relate the Jacobi matrix to the Lanczos partial tridiagonalization of $A$ with starting unit vector $v$. Throughout this section we assume $m$ is an integer satisfying $m < \overline{m} = \text{grade}(v)$.

Recall from last chapter, the order $m$ Lanczos partial tridiagonalization of the symmetric matrix $A$, with respect to the starting vector $v$, is the symmetric tridiagonal matrix, $T_m \in \mathbb{R}^{m\times m}$, given by $T_m = V_m^T A V_m$, where the columns of $V_m$ are an orthonormal basis for the Krylov space $\mathcal{K}_m(A, v) = \text{span}\{v, Av, \ldots, A^{m-1}v\}$. The entries of $T_m$ are the coefficients of the three term recurrence satisfied by the columns of $V_m$, which are referred to as the Lanczos vectors. When referring to the Lanczos algorithm in this chapter, we assume that all computations are done in infinite precision, the effects of finite precision arithmetic having already been considered in the previous chapter.

Before outlining how to relate the Jacobi matrix and the Lanczos partial tridiagonalization, we first want to point out a few hints relating polynomials, such as the orthonormal polynomials constructed in the last section, to the Lanczos vectors. The first connection is the definition of the Krylov space itself. It is easily seen that the Krylov space, $\mathcal{K}_m(A, v)$, can be expressed as $\mathcal{K}_m(A, v) = \{p(A)v \mid p \in \mathscr{P}_{m-1}\}$. It then follows that the Lanczos vectors can be expressed as $p(A)v$ for properly chosen polynomials $p$. The next connection comes from the definition of the semi-inner product $\langle\,\cdot\,,\,\cdot\,\rangle_s$. For any polynomials $p, q \in \mathscr{P}$, $\langle p, q\rangle_s = (p(A)v, q(A)v)$, and therefore, $\langle\,\cdot\,,\,\cdot\,\rangle_s$ directly relates the inner product of a polynomial $p$, to the vector $p(A)v$. This relation also highlights that if the polynomials $p$ and $q$ are orthogonal with respect to $\langle\,\cdot\,,\,\cdot\,\rangle_s$, the vectors $p(A)v$ and $q(A)v$ are orthogonal with respect to the Euclidean inner product. Lastly, and most importantly, is the three term recurrence satisfied by orthonormal polynomials and the Lanczos vectors. The three term recurrences will allow us to transition directly from the orthonormal polynomials with respect to $\langle\,\cdot\,,\,\cdot\,\rangle_s$ to the Lanczos vectors.

Let $\{p_k\}_{k=0}^{m}$ be the family of orthonormal polynomials with respect to the spectral function $s(\lambda) = \sum_{i=1}^{n} |(x_i, v)|^2 \delta(\lambda - \lambda_i)$, where $p_k \in \mathscr{P}_k$ is of exact degree $k$, with $p_0 = 1$. For convenience, we define $p_{-1} = 0$. As shown previously, the polynomials satisfy the three term recurrence relation

$$\beta_k p_k(\lambda) = (\lambda - \alpha_k) p_{k-1}(\lambda) - \beta_{k-1} p_{k-2}(\lambda) \quad k = 1, \ldots, m, \tag{3.28}$$

which is simply (3.11) (reindexed). The coefficients in the three term recurrence are given by

$$\alpha_k = \langle \lambda p_{k-1}, p_{k-1} \rangle_s, \quad \text{and} \quad \beta_k = \| (\lambda - \alpha_k) p_{k-1} - \beta_{k-1} p_{k-2} \|_s, \tag{3.29}$$

for $k = 1, \ldots, m$.

Now, we come to the truly amazing relation between the orthonormal polynomials with respect to $\langle\,\cdot\,,\,\cdot\,\rangle_s$, and the Lanczos vectors which result from the $m$-step Lanczos algorithm applied to the symmetric matrix $A$, with starting unit vector $v$. Defining vectors $v_k = p_{k-1}(A)v$, $k = 1, \ldots, m+1$, we now show that these are the Lanczos vectors. As mentioned previously, the orthonormality of the polynomials translates to the orthonormality of the vectors,

$$(v_i, v_j) = (p_{i-1}(A)v, p_{j-1}(A)v) = \langle p_{i-1}, p_{j-1} \rangle_s = \delta_{ij},$$

for $i, j = 1, \ldots, m+1$. By orthonormality and a dimension argument, we also see that the vectors, $v_k$, $k = 1, \ldots, m$, form an orthonormal basis of $\mathcal{K}_m(A, v)$ (note that $v_1 = v$). Additionally, the vectors satisfy a three term recurrence like the polynomials $p_k$. Evaluating (3.28) at $\lambda = A$, and multiplying by the vector $v$, we arrive at the standard three term recurrence of the Lanczos vectors

$$\beta_k v_{k+1} = (A - \alpha_k I)v_k - \beta_{k-1}v_{k-1}. \tag{3.30}$$

Finally, we show (3.30) is in fact the Lanczos recurrence by looking at the formulas for the coefficients. Using (3.29), we have

$$\begin{aligned}\alpha_k &= \langle \lambda p_{k-1}, p_{k-1}\rangle_s = (Ap_{k-1}(A)v, p_{k-1}(A)v) = (Av_k, v_k),\\ \beta_k &= \|(\lambda - \alpha_k)p_{k-1} - \beta_{k-1}p_{k-2}\|_s = \|(A - \alpha_k I)v_k - \beta_{k-1}v_{k-1}\|,\end{aligned}$$

which are the formulas for the coefficients in the Lanczos algorithm. Because the coefficients defining the recurrence relation for the polynomials $p_k$ are the same as the coefficients for the Lanczos recurrence, we have that the Jacobi matrix of order $m$ with respect to the (unknown) measure $s(\lambda)$ is the order $m$ Lanczos partial tridiagonalization of $A$ with respect to the starting vector $v$.

Utilizing this newfound relationship between the Jacobi matrix and the Lanczos partial tridiagonalization, we can approximate the quadratic form $v^T f(A)v$ using Gaussian quadrature by simply computing the Lanczos partial tridiagonalization of $A$ with starting vector $v$, as opposed to forming a family of orthonormal polynomials with respect to an unknown measure. Using the eigenpairs of the Lanczos partial tridiagonalization to determine the quadrature nodes and weights we are able to approximate $v^T f(A)v$. This makes approximation of the quadratic form straightforward since the Lanczos algorithm is well understood, and most numerical software packages have a routine for performing the Lanczos algorithm.

What we have shown is that it is possible to construct the Lanczos vectors from the orthonormal polynomials with respect to the inner product $\langle\,\cdot\,,\,\cdot\,\rangle_s$. We now show that the converse is also true. Given the partial tridiagonalization of $A$ with respect to the starting unit vector $v$, we can define a family of orthonormal polynomials with respect to $\langle\,\cdot\,,\,\cdot\,\rangle_s$. The $m$-step Lanczos algorithm applied to $A$ with starting unit vector $v$ is succinctly written as

$$AV_m = V_m T_m + \beta_m v_{m+1} e_m^T, \tag{3.31}$$

where the columns of $V_m \in \mathbb{R}^{n\times m}$ (the Lanczos vectors) are an orthonormal basis of

$\mathcal{K}_m(A, v) = \text{span}\{v, Av, \ldots, A^{m-1}v\}$, $T_m \in \mathbb{R}^{m\times m}$ is symmetric and tridiagonal, and $V_m^T v_{m+1} = 0$. The partial tridiagonalization of $A$ with respect to $v$ can be written as

$$T_m = V_m^T A V_m = \begin{pmatrix} \alpha_1 & \beta_1 & & \\ \beta_1 & \ddots & \ddots & \\ & \ddots & & \beta_{m-1} \\ & & \beta_{m-1} & \alpha_m \end{pmatrix}. \tag{3.32}$$

Note that at any iteration $1 \le k \le m$, $T_k = V_k^T A V_k \in \mathbb{R}^{k\times k}$ is the leading principal submatrix of $T_m$.

With the knowledge that the roots of the orthonormal polynomials with respect to $\langle\,\cdot\,,\,\cdot\,\rangle_s$ are the eigenvalues of the matrix $T_k$ (see (3.12)), define the polynomials

$$p_k(\lambda) = c_k \chi_k(\lambda), \quad \chi_k(\lambda) = \det(T_k - \lambda I), \quad c_k = \frac{(-1)^k}{\prod\limits_{i=1}^{k} \beta_i}, \tag{3.33}$$

for $k = 1, \ldots, m$, and $p_0(\lambda) = 1$. The polynomials, $p_k(\lambda)$, defined as in (3.33), are referred to as orthonormal Lanczos polynomials. In order to verify that the orthonormal Lanczos polynomials satisfy the three term recurrence (3.28), we use the following lemma.

**Lemma 4.** *Let $\eta_k$ and $\nu_k$ be real numbers for $k \in \mathbb{N}$ and define a family of symmetric tridiagonal matrices,*

$$S_k = \begin{pmatrix} \eta_1 & \nu_1 & & \\ \nu_1 & \ddots & \ddots & \\ & \ddots & & \nu_{k-1} \\ & & \nu_{k-1} & \eta_k \end{pmatrix}.$$

*Then, the determinants satisfy the recurrence*

$$\det S_k = \eta_k \det S_{k-1} - \nu_{k-1}^2 \det S_{k-2}, \quad k = 1, 2 \ldots,$$

*with initial conditions* $\det S_{-1} = 0$ *and* $\det S_0 = 1$ *(*$\nu_0$ *need not be defined since* $\det S_{-1} = 0$*).*

*Proof.* Expand the determinant of $S_k$ along the last row or column. □

We now show that the Lanczos polynomials satisfy the recurrence (3.28). Using Lemma 4, $\chi_k(\lambda) = \det(T_k - \lambda I)$, satisfies

$$\chi_k(\lambda) = (\alpha_k - \lambda)\chi_{k-1}(\lambda) - \beta_{k-1}^2 \chi_{k-2}(\lambda), \tag{3.34}$$

for $k = 1, 2, \ldots, m$ where we define $\chi_{-1} = 0$ and $\chi_0 = 1$. The normalization coefficients, $c_k$, defined in (3.33) satisfy $\beta_k c_k = -c_{k-1}$, and so multiplying the left hand side of (3.34) by $\beta_k c_k$, and the right hand side by $-c_{k-1}$, gives

$$\beta_k c_k \chi_k(\lambda) = (\lambda - \alpha_k) c_{k-1} \chi_{k-1}(\lambda) + \beta_{k-1}^2 c_{k-1} \chi_{k-2}(\lambda). \tag{3.35}$$

Using again $\beta_{k-1} c_{k-1} = -c_{k-2}$ on the rightmost term in (3.35), and the definition of the orthonormal Lanczos polynomials from (3.33), we have

$$\beta_k p_k(\lambda) = (\lambda - \alpha_k) p_{k-1}(\lambda) - \beta_{k-1} p_{k-2}(\lambda),$$

which is (3.28), as claimed.

The following theorem summarizes this section.

**Theorem 8.** *Let $A \in \mathbb{R}^{n \times n}$ be symmetric and $v$ be a unit $n$-vector. Denote the eigenpairs of $A$ as $Ax_i = \lambda_i x_i$, $x_i^T x_j = \delta_{ij}$, $i, j = 1, \ldots, n$, and define the measure $s(\lambda) = \sum_{i=1}^n |(x_i, v)|^2 \delta(\lambda - \lambda_i)$. Assuming $m <$ grade$(v)$, the Jacobi matrix of order $m$ corresponding to the measure $s(\lambda)$ is the order $m$ Lanczos partial tridiagonalization of $A$ with respect to the starting vector $v$.*

## 3.3 Approximating the Spectral Function

In this section we illustrate that using the Lanczos process for approximating the quadratic form $v^T f(A) v$ is equivalent to approximating the spectral function $s(\lambda) = \sum_{i=1}^n |(x_i, v)|^2 \delta(\lambda - \lambda_i)$. This has many applications in physics, which is the basis of this thesis. Examples include the density of states [33, 67], the joint density of states [57, 68], and the optical absorption curve [30]. Additionally, we state known error estimates in the Lanczos process for approximating $v^T f(A)v$ when $f$ is analytic, and give new error estimates, in appropriate Sobolev spaces, when $f$ has less regularity.

From the previous section, we discovered that by performing the $m$-step Lanczos algorithm on $A$, with starting unit vector $v$, the Ritz values, $\theta_j$, and corresponding normalized eigenvectors of the partial tridiagonalization, $y_j$, $j = 1, \ldots, m$, can be used to approximate $v^T f(A)v$ as

$$v^T f(A)v = \sum_{i=1}^{n} |(x_i, v)|^2 f(\lambda_i) \approx \sum_{j=1}^{m} |(y_j, e_1)|^2 f(\theta_j), \tag{3.36}$$

where the $\lambda_i$, $i = 1, \ldots, n$, are the eigenvalues of $A$ and the $x_i$ are the corresponding orthonormal eigenvectors. Using the spectral function $s(\lambda)$, we can represent $v^T f(A)v$ as $\int_a^b sf$ (see (3.3)). The approximation to $v^T f(A)v$ (right hand side of (3.36)) is then the integral of $f$ with measure $\tilde{s}$ defined as

$$\tilde{s}(\lambda) = \sum_{j=1}^{m} |(y_j, e_1)|^2 \delta(\lambda - \theta_j). \tag{3.37}$$

Since $\int_a^b sf$ is approximated by $\int_a^b \tilde{s}f$ for every test function $f$ (we have equality when $f$ is a polynomial of degree less than or equal to $2m - 1$), we can use the Lanczos process to construct $\tilde{s}$ as an approximation to the spectral function $s$.

The property that $\int_a^b sf = \int_a^b \tilde{s}f$ for all $f \in \mathscr{P}_{2m-1}$ is a well known fact about the Lanczos process, and is known as the moment matching property. The moment matching property states that

$$\sum_{i=1}^{n} |(x_i, v)|^2 \lambda_i^k = \sum_{j=1}^{m} |(y_j, e_1)|^2 \theta_j^k, \tag{3.38}$$

for $k = 0, 1, \ldots, 2m - 1$. This property, a consequence of the degree of precision of Gaussian quadrature, is very powerful. It allows us to approximate sums involving the unknown spectrum of $A$, using the $m$-step Lanczos partial tridiagonalization of $A$ with starting vector $v$. Using the eigenpairs of the partial tridiagonalization of $A$, we are able to determine "bulk" properties of the spectrum of $A$ with relatively few iterations of the Lanczos algorithm. This is in contrast to diagonalizing the matrix $A$, which gives moments of all orders, but is significantly more expensive.

### 3.3.1 Error Estimates for Analytic Functions

In order to understand how well the Lanczos process performs, we begin by formulating error estimates for the simplest case. Namely, we consider the error in the Lanczos approximation to $v^T f(A)v$ for analytic $f$. The results presented here for analytic functions are very similar to those given in [67]. We assume throughout this section that $f : [-1, 1] \to \mathbb{R}$ is analytic.

For the error estimate we use Chebyshev expansions, and known results on the decay rate of Chebyshev coefficients. The Chebyshev polynomials of the first kind are defined as

$$t_k(\lambda) = \cos(k \cos^{-1}(\lambda)), \quad \lambda \in [-1, 1], \quad k = 0, 1, \ldots. \tag{3.39}$$

Note that Chebyshev polynomials of the first kind are typically denoted by $T_k$, however we reserve this notation for Lanczos partial tridiagonalizations. To see that (3.39) in fact defines a family of polynomials one can use trigonometric identities to deduce that $t_0(\lambda) = 1$, $t_1(\lambda) = \lambda$, and $t_{k+1}(\lambda) = 2\lambda t_k(\lambda) - t_{k-1}(\lambda)$ for $k = 1, 2, \ldots$. Thus, $t_k(\lambda)$ is a polynomial of exact degree $k$ which satisfies $-1 \leq t_k(\lambda) \leq 1$ for all $\lambda \in [-1, 1]$.

Because Chebyshev polynomials are defined in the interval $[-1, 1]$, we need to perform a spectral transformation to put the eigenvalues of $A$ in the interval $[-1, 1]$. Recall that the integration bounds in the definition of $\langle \cdot , \cdot \rangle_s$ (see (3.4)) satisfy $a \leq \lambda_1$ and $\lambda_n \leq b$. Using these bounds, define $c = (b + a)/2$ and $d = (b - a)/2$. Then, the matrix

$$\hat{A} = \frac{1}{d}(A - cI), \tag{3.40}$$

has its spectrum inside $[-1, 1]$. We assume this has already been performed, and drop the circumflex.

Expand $f$ in terms of Chebyshev polynomials of the first kind as

$$f(\lambda) = \sum_{k=0}^{\infty} \mu_k t_k(\lambda), \tag{3.41}$$

where the coefficients $\mu_k$ are given by

$$\mu_k = \frac{2-\delta_{k0}}{\pi} \int_{-1}^{1} (1-\lambda^2)^{-1/2} t_k(\lambda) f(\lambda) d\lambda. \tag{3.42}$$

Note that the constant before the integral in (3.42) comes from the fact that the Chebyshev polynomials satisfy

$$\int_{-1}^{1} (1-\lambda^2)^{-1/2} t_k(\lambda) t_l(\lambda) d\lambda = \begin{cases} 0 & \text{if } k \neq l, \\ \pi & \text{if } k = l = 0, \\ \frac{\pi}{2} & \text{if } k = l \neq 0. \end{cases}$$

By assuming $f$ is analytic in $[-1,1]$, we in fact get the stronger result that $f$ is analytic in a region of the complex plane containing the closed interval $[-1,1]$ in its interior. The larger the region in the complex plane in which $f$ is analytic, the faster the Chebyshev coefficients decay. The region of analyticity which is important in this regard is the interior of a certain ellipse known as a Bernstein ellipse [58]. The Bernstein ellipses are the image of circles of radius $\rho > 1$ centered at the origin under the Joukowsky map given by $\zeta(z) = 1/2(z + z^{-1})$. Putting this all together, the Bernstein ellipse, $E_\rho$, for a parameter $\rho > 1$, is given by

$$E_\rho = \{1/2(z + z^{-1}) \mid z = \rho e^{i\theta} \text{ for all } \theta \in [0, 2\pi)\}. \tag{3.43}$$

The Bernstein ellipse $E_\rho$ has foci at ±1 and semi-major axis $1/2(\rho + \rho^{-1})$ and semi-minor axis $1/2(\rho - \rho^{-1})$. Bernstein ellipses for several values of $\rho$ are shown in Figure 3.1.

The following lemma on the decay rate of Chebyshev coefficients for analytic functions is taken from [58].

**Lemma 5.** *Let $f(\lambda)$ be analytic on $[-1,1]$ with Chebyshev expansion $f(\lambda) = \sum_{k=0}^{\infty} \mu_k t_k(\lambda)$, and, for $\rho > 1$, analytically continuable to the interior of $E_\rho$. Then, the coefficients of the Chebyshev expansion satisfy*

$$|\mu_0| \leq M, \qquad |\mu_k| \leq 2M\rho^{-k} \quad \textit{for} \quad k \geq 1,$$

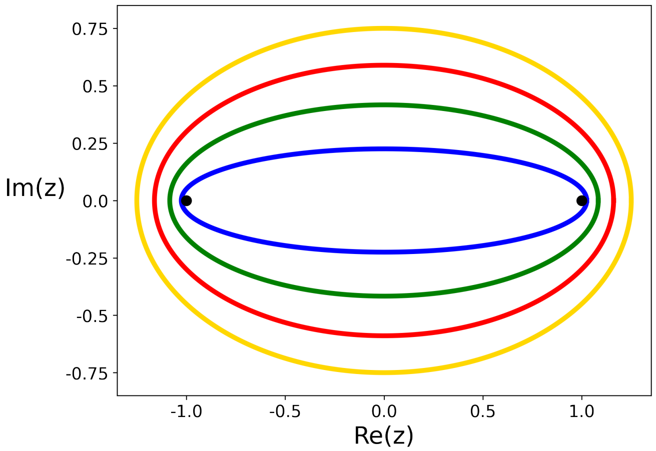


Figure 3.1: Bernstein ellipses $E_\rho$ for $\rho = 1.25$ (blue), $\rho = 1.5$ (green), $\rho = 1.75$ (red), and $\rho = 2.0$ (yellow).

*where* $|f| \leq M$ *inside* $E_\rho$*.*

Finally, we are ready to state error results for the approximation of $v^T f(A)v$ via the Lanczos process for analytic $f$.

**Theorem 9.** *Let* $A \in \mathbb{R}^{n\times n}$ *be symmetric with eigenvalues in the interval* $[-1, 1]$ *and* $v \in \mathbb{R}^n$ *be a unit vector, let* $f$ *be analytic in* $[-1, 1]$ *and analytically continuable inside* $E_\rho$ *for* $\rho > 1$*, and let* $s$ *and* $\tilde{s}$ *be distributions as defined in* (3.2) *and* (3.37) *respectively (*$\tilde{s}$ *being determined by the* $m$*-step Lanczos process with* $A$ *and* $v$*). Then, the error in the* $m$*-step Lanczos process approximation to the quadratic form* $v^T f(A)v$ *is*

$$\left|\int_{-1}^{1} \big(s(\lambda) - \tilde{s}(\lambda)\big) f(\lambda) d\lambda\right| \leq \frac{4M\rho}{\rho^{2m}(\rho-1)},$$

*where* $|f| \leq M$ *inside* $E_\rho$*.*

*Proof.* Let $\tilde{f}(\lambda)$ be the $2m-1$ degree Chebyshev expansion of $f$, i.e.,

$$\tilde{f}(\lambda) = \sum_{k=0}^{2m-1} \mu_k t_k(\lambda) \approx f(\lambda) = \sum_{k=0}^{\infty} \mu_k t_k(\lambda). \tag{3.44}$$

Since the quadrature formula is exact for polynomials of degree up to $2m-1$ we have

$\int_{-1}^{1} s\tilde{f} = \int_{-1}^{1} \tilde{s}\tilde{f}$, and therefore

$$\begin{aligned}\left|\int_{-1}^{1} \big(s(\lambda) - \tilde{s}(\lambda)\big) f(\lambda) d\lambda\right| &= \left|\int_{-1}^{1} \big(s(\lambda) - \tilde{s}(\lambda)\big)\big(f(\lambda) - \tilde{f}(\lambda)\big) d\lambda\right| \\ &\leq \int_{-1}^{1} s(\lambda)\big|f(\lambda) - \tilde{f}(\lambda)\big| d\lambda + \int_{-1}^{1} \tilde{s}(\lambda)\big|f(\lambda) - \tilde{f}(\lambda)\big| d\lambda.\end{aligned} \tag{3.45}$$

Next, we bound the term $\int_{-1}^{1} s|f - \tilde{f}|$ (the other term being nearly identical). Expanding $f - \tilde{f}$ in a Chebyshev series, see (3.44), gives

$$\int_{-1}^{1} s(\lambda)\big|f(\lambda) - \tilde{f}\big| d\lambda = \int_{-1}^{1} s(\lambda)\left|\sum_{k=2m}^{\infty} \mu_k t_k(\lambda)\right| d\lambda \leq \sum_{k=2m}^{\infty} |\mu_k| \int_{-1}^{1} s(\lambda)\big|t_k(\lambda)\big| d\lambda. \tag{3.46}$$

Using $|t_k(\lambda)| \leq 1$ for all $\lambda \in [-1, 1]$ and $\int_{-1}^{1} s = 1$ (follows from $v$ being a unit vector) we have $\int_{-1}^{1} s|t_k| \leq 1$ for all $k$, and therefore from (3.46) we conclude

$$\int_{-1}^{1} s(\lambda)\big|f(\lambda) - \tilde{f}\big| d\lambda \leq \sum_{k=2m}^{\infty} |\mu_k|. \tag{3.47}$$

We remark that $\int_{-1}^{1} \tilde{s} = 1$, and so by the same argument $\int_{-1}^{1} \tilde{s}|f - \tilde{f}| \leq \sum_{k=2m}^{\infty} |\mu_k|$. Using the bounds on $\mu_k$ from Lemma 5 we have

$$\sum_{k=2m}^{\infty} |\mu_k| \leq 2M \sum_{k=2m}^{\infty} \rho^{-k} = \frac{2M\rho}{\rho^{2m}(\rho - 1)}, \tag{3.48}$$

where $M$ is the bound of $f$ in $E_\rho$. Putting together (3.47) and (3.48) gives

$$\int_{-1}^{1} s(\lambda)\big|f(\lambda) - \tilde{f}(\lambda)\big| d\lambda \leq \frac{2M\rho}{\rho^{2m}(\rho - 1)}.$$

As noted previously, the same bound holds for $\int_{-1}^{1} \tilde{s}|f - \tilde{f}|$. Combining the bounds on $\int_{-1}^{1} s|f - \tilde{f}|$ and $\int_{-1}^{1} \tilde{s}|f - \tilde{f}|$ with (3.45) gives the desired result. □

The above theorem shows that for analytic $f$, the action of the distribution $s(\lambda)$ on $f$ matches that of the approximation $\tilde{s}(\lambda)$, determined by the $m$-step Lanczos process, to within $\mathcal{O}(\rho^{-2m})$ for some value $\rho > 1$. This is impressive, as other known methods for approximating the spectral function (in terms of the action on analytic functions) are order $\mathcal{O}(\rho^{-m})$ [33]. This includes the well known Kernel Polynomial Method conceived in the 1990's for approximating the density of states (and which can also be used to approximate $s$) [61, 53, 52, 51]. In other words, the Lanczos process is twice as accurate as other known methods! While the error estimates of Theorem 9 are interesting, and useful for comparison, they are a best case scenario since it involves analytic functions. Next, we investigate error results in Sobolev spaces.

### 3.3.2 Error Estimates in Sobolev Spaces

In the last section we gave error estimates in the Lanczos process for approximating $v^T f(A)v$ for analytic functions $f$. However, as mentioned previously, we are more interested in estimates for the error in the Lanczos approximation to the spectral function, $s(\lambda) - \tilde{s}(\lambda)$. To accomplish this, we consider the norm of $s - \tilde{s}$ in the dual space of appropriate Sobolev spaces, and use Jackson type estimates to get an a priori rate of convergence. We begin construction of these error estimates by using general estimates in the dual space of continuous functions on a closed interval. Then, using Sobolev imbeddings in the space of Hölder continuous functions we are able to establish the desired results.

First, we have a need to introduce some notation and terminology from the theory of Sobolev spaces and Hölder spaces in one dimension. For this background material we closely follow [1, 15]. The domain on which we define all of the following spaces is the open interval $\Omega = (-1, 1)$. Let $L^p(\Omega)$, $1 \le p \le \infty$, denote the standard Lebesgue space of measurable functions $f : \Omega \to \mathbb{R}$ with finite norm

$$\|f\|_{L^p} := \begin{cases} \Big( \int\limits_{-1}^{1} |f|^p d\lambda \Big)^{1/p}, & 1 \le p < \infty, \\ \operatorname*{ess\,sup}\limits_{\Omega} |f|, & p = \infty. \end{cases}$$

Note that for notational convenience we suppress dependence on the domain $\Omega$ when denoting the norm $\|\cdot\|_{L^p}$. This should not cause confusion since the domain does not change in this section. For a nonnegative integer $k$, the Sobolev space, $W^{k,p}(\Omega)$, is defined as

$$W^{k,p}(\Omega) = \left\{ f \in L^p(\Omega) \;\middle|\; f^{(\ell)} \in L^p(\Omega) \text{ for } 0 \le \ell \le k \right\},$$

where, $f^{(\ell)} = d^\ell f/d\lambda^\ell$, with derivatives understood in a weak (distributional) sense, and $W^{0,p}(\Omega) = L^p(\Omega)$. When equipped with the norm

$$\|f\|_{W^{k,p}} := \left( \sum_{\ell=0}^{k} \|f^{(\ell)}\|_{L^p}^p \right)^{1/p},$$

$W^{k,p}(\Omega)$ is a Banach space. Also important are the spaces dual to $W^{k,p}(\Omega)$, denoted $\big(W^{k,p}(\Omega)\big)'$, i.e., the space of bounded linear functions on $W^{k,p}(\Omega)$. The space $\big(W^{k,p}(\Omega)\big)'$ is a Banach space with the standard operator norm

$$\|L\|_{(W^{k,p})'} = \sup_{\substack{u \in W^{k,p}(\Omega) \\ \|u\|_{W^{k,p}} \le 1}} |L(u)|.$$

Next, we define the Hölder spaces. Let $C(\overline{\Omega})$ denote the Banach space of continuous functions $f : \overline{\Omega} \to \mathbb{R}$ with uniform norm $\|f\|_\infty = \sup_{-1 \le x \le 1} |f(x)|$. The $\gamma^{\text{th}}$-Hölder semi-norm is defined as

$$|f|_{C^{0,\gamma}} := \sup_{\substack{x,y \in [-1,1] \\ x \ne y}} \frac{|f(x) - f(y)|}{|x - y|^\gamma},$$

and the $\gamma^{\text{th}}$-Hölder norm is

$$\|f\|_{C^{0,\gamma}} = \|f\|_\infty + |f|_{C^{0,\gamma}}.$$

The Hölder space, $C^{k,\gamma}(\Omega)$, is then defined as the space of functions for which the norm

$$\|f\|_{C^{k,\gamma}} := \sum_{\ell=0}^{k} \|f^{(\ell)}\|_\infty + |f^{(k)}|_{C^{0,\gamma}},$$

is finite.

To begin, let $w(\lambda)$ be a general weight (we use $w$ so as to not confuse with the spectral function $s$), and suppose we are interested in using quadrature to approximate the integral $\int_{-1}^{1} w(\lambda) f(\lambda) d\lambda$, for $f \in C(\overline{\Omega})$. Choosing distinct nodes, $\theta_j^{(m)} \in \overline{\Omega}$, and weights, $\tau_j^{(m)} \in \mathbb{R}$, $j = 1, \ldots, m$, we approximate the integral as

$$\int\limits_{-1}^{1} w(\lambda) f(\lambda) d\lambda \approx \sum_{j=1}^{m} \tau_j^{(m)} f\big(\theta_j^{(m)}\big). \tag{3.49}$$

The $m$-point quadrature rule can be written using a discrete weight, $\tilde{w}$, defined as

$$\tilde{w}(\lambda) = \sum_{j=1}^{m} \tau_j^{(m)} \delta\big(\lambda - \theta_j^{(m)}\big), \tag{3.50}$$

and (3.49) can be written as $\int_{-1}^{1} wf \approx \int_{-1}^{1} \tilde{w} f$. Throughout this section we assume the quadrature rule has degree of precision $d = d(m)$, so that

$$\int\limits_{-1}^{1} w(\lambda) f(\lambda) d\lambda = \int\limits_{-1}^{1} \tilde{w}(\lambda) f(\lambda) d\lambda = \sum_{j=1}^{m} \tau_j^{(m)} f(\theta_j^{(m)}) \quad \text{for all} \quad f \in \mathscr{P}_d.$$

For most interpolatory quadrature rules, $d(m) = m - 1$, and for Gaussian quadrature, $d(m) = 2m - 1$.

In order for (3.49) to be meaningful, we obviously need to put conditions on the weight $w$. Since we are most interested in measuring the error in linear combinations of Dirac distributions, and the Dirac distribution can most generally be seen as an element of the dual space of continuous functions, this is where we begin the analysis. Therefore, we consider the quadrature approximation (3.49) for $w \in (C(\overline{\Omega}))'$, and write $w(f) = \int_{-1}^{1} w(\lambda) f(\lambda) d\lambda$ for $f \in C(\overline{\Omega})$. Every element of $(C(\overline{\Omega}))'$ is representable as a Riemann–Stieltjes integral, i.e., for every $w \in (C(\overline{\Omega}))'$, there exists a function $\mu$, of bounded variation, such that $\int_{-1}^{1} wf = \int_{-1}^{1} f d\mu$ for every $f \in C(\overline{\Omega})$, where $\int_{-1}^{1} f d\mu$ is a Riemann–Stieltjes integral [28]. The correspondence between $w \in (C(\overline{\Omega}))'$ and $\mu$ is unique if we impose the normalization conditions $\mu(-1) = 0$ and that $\mu$ be right continuous. Furthermore, $\|w\|_{(C(\bar{\Omega}))'} = \mathrm{Var}(\mu)$ where

$$\mathrm{Var}(\mu) = \sup \sum_{j=1}^{\ell} |\mu(x_j) - \mu(x_{j-1})|,$$

is the total variation in $\mu$, with the supremum taken over all partitions, $-1 = x_0 < x_1 < \ldots < x_\ell = 1$, $\ell$ being arbitrary. In keeping with the terminology of this chapter, we refer to elements of $(C(\bar{\Omega}))'$ as weights or measures interchangeably.

To estimate the quadrature error in $(C(\bar{\Omega}))'$, we need to bound $|\int_{-1}^{1} wf - \int_{-1}^{1} \tilde{w}f|$ for arbitrary $f \in C(\bar{\Omega})$, $\|f\|_\infty \leq 1$. In order to accomplish this, we use Jackson type theorems, as opposed to the decay rate of Chebyshev coefficients which was used for the case of analytic $f$. Toward this goal, we introduce the modulus of continuity for a function $f : [-1, 1] \to \mathbb{R}$,

$$\omega_f(\delta) \coloneqq \sup\{|f(x) - f(y)| \mid |x - y| \leq \delta, \quad x, y \in [-1, 1]\}. \tag{3.51}$$

A function $f$ is continuous if $\omega_f(\delta) \to 0$ as $\delta \to 0$ and continuously differentiable if $\omega_f(\delta) = \mathcal{O}(\delta)$. The modulus of continuity therefore allows us to measure levels of continuity which lie somewhere between these extremes, in a similar manner to the Hölder spaces.

The last ingredient for the Jackson theorem is the best uniform approximation. For a continuous function $f : [-1, 1] \to \mathbb{R}$, define the best uniform approximation of degree $k$, denoted $\tilde{f}$, as the unique degree $k$ polynomial which satisfies $\|f - \tilde{f}\|_\infty \leq \|f - p\|_\infty$ for all $p \in \mathscr{P}_k$. For a proof of existence and uniqueness of best uniform approximates see [58]. Also, define the error in the best uniform approximation as

$$E_k(f) \coloneqq \|f - \tilde{f}\|_\infty = \inf_{p \in \mathscr{P}_k} \|f - p\|_\infty. \tag{3.52}$$

With the modulus of continuity and best uniform approximation defined, we are now ready to state Jackson's Theorem, which gives uniform error bounds for polynomial approximation. The following is taken from [10].

**Theorem 10** (Jackson)**.** *Let $f : [-1, 1] \to \mathbb{R}$ be continuous. Then, $E_k(f) \leq \omega_f(\pi/(k + 1))$. Furthermore, if $f$ is Lipschitz continuous, i.e., if there exists a constant $L > 0$ such* *that* $|f(x) - f(y)| \leq L|x - y|$ *for all* $x, y \in [-1, 1]$*, then* $E_k(f) \leq L\pi/(2k+2)$*.*

Next, using Jackson's Theorem, we state another useful lemma pertaining to bounded linear functionals on the space of continuous functions which vanish on a subspace of polynomials. This is one characteristic of the error in quadrature routines, which is zero on all polynomials of degree less than or equal to the degree of precision.

**Lemma 6.** *Let* $e \in (C(\overline{\Omega}))'$ *vanish on* $\mathscr{P}_d$*. Then, for all* $f \in C(\overline{\Omega})$

$$\left|\int_{-1}^{1} e(\lambda) f(\lambda) d\lambda\right| \leq \|e\|_{(C(\overline{\Omega}))'}\, \omega_f\Big(\frac{\pi}{d+1}\Big).$$

*Proof.* Assume $e \in (C(\overline{\Omega}))'$ vanishes on $\mathscr{P}_d$, $f \in C(\overline{\Omega})$, and $\tilde{f}$ is the best uniform approximation to $f$ in $\mathscr{P}_d$. Using the linearity of $e$ and the fact that $\int_{-1}^{1} e\tilde{f} = 0$, we have

$$\left|\int_{-1}^{1} e(\lambda) f(\lambda) d\lambda\right| = \left|\int_{-1}^{1} e(\lambda)\big(f(\lambda) - \tilde{f}(\lambda)\big) d\lambda\right| \leq \|e\|_{(C(\overline{\Omega}))'} \|f - \tilde{f}\|_\infty.$$

By Jackson's Theorem, $\|f - \tilde{f}\|_\infty \leq \omega_f\big(\pi/(d+1)\big)$, giving the desired result. □

In most practical cases, we are not interested in using quadrature to approximate all elements of $(C(\overline{\Omega}))'$, but rather those with a special property. Namely, that of positivity. We call $w \in (C(\overline{\Omega}))'$ positive if $\int_{-1}^{1} wf \geq 0$ for all nonnegative $f \in C(\overline{\Omega})$. The following lemma applies the results of Lemma 6 to the case of a quadrature rule with degree of precision $d$, approximating a positive measure. A similar result, in a less general context, can be found in [14].

**Lemma 7.** *Let* $w \in (C(\overline{\Omega}))'$ *be positive, and for* $\tau_j^{(m)} \in \mathbb{R}$ *and distinct* $\theta_j^{(m)} \in \overline{\Omega}$*,* $j = 1, \ldots, m$*, define* $\tilde{w} \in (C(\overline{\Omega}))'$ *as* $\tilde{w}(\lambda) = \sum_{j=1}^{m} \tau_j^{(m)} \delta\big(\lambda - \theta_j^{(m)}\big)$*. If* $w - \tilde{w}$ *vanishes on* $\mathscr{P}_d$ *for* $d = d(m) \geq 0$*, then for all* $f \in C(\overline{\Omega})$

$$\left|\int_{-1}^{1} \big(w(\lambda) - \tilde{w}(\lambda)\big) f(\lambda) d\lambda\right| \leq \bigg(\sum_{j=1}^{m} \Big(\tau_j^{(m)} + \big|\tau_j^{(m)}\big|\Big)\bigg) \omega_f\Big(\frac{\pi}{d+1}\Big).$$

*Proof.* For $f \in C(\overline{\Omega})$ we apply Lemma 6 to $w - \tilde{w}$ to get

$$\left| \int_{-1}^{1} \big(w(\lambda) - \tilde{w}(\lambda)\big) f(\lambda) d\lambda \right| \leq \|w - \tilde{w}\|_{(C(\overline{\Omega}))'} \omega_f\left(\frac{\pi}{d+1}\right). \tag{3.53}$$

By the triangle inequality, $\|w-\tilde{w}\|_{(C(\overline{\Omega}))'} \leq \|w\|_{(C(\overline{\Omega}))'} + \|\tilde{w}\|_{(C(\overline{\Omega}))'}$. Next, we show that $\|w\|_{(C(\overline{\Omega}))'} = \sum_{j=1}^{m} \tau_j^{(m)}$ and $\|\tilde{w}\|_{(C(\overline{\Omega}))'} = \sum_{j=1}^{m} |\tau_j^{(m)}|$, which establishes the result. Let $\mu$ be the function of bounded variation for which $\int_{-1}^{1} wf = \int_{-1}^{1} f d\mu$ for all $f \in C(\overline{\Omega})$. The positivity of $w$ implies $\mu$ is monotonically increasing. Because of this, $\|w\|_{(C(\overline{\Omega}))'} = \mathrm{Var}(\mu) = \int_{-1}^{1} d\mu = \int_{-1}^{1} w(\lambda) d\lambda$, i.e., the norm of $w$ is the zeroth moment. Because $w - \tilde{w}$ vanishes on the constants, the zeroth moment is given by $\sum_{j=1}^{m} \tau_j^{(m)}$. The function of bounded variation, $\tilde{\mu}$, for which $\int_{-1}^{1} \tilde{w} f = \int_{-1}^{1} f d\tilde{\mu}$ for all $f \in C(\overline{\Omega})$, is given by

$$\tilde{\mu}(\lambda) = \begin{cases} 0, & \lambda < \theta_1, \\ \sum\limits_{j=1}^{k} \tau_j^{(m)}, & \theta_k \leq \lambda < \theta_{k+1},\ k = 1, \ldots, m-1, \\ \sum\limits_{j=1}^{m} \tau_j^{(m)}, & \lambda \geq \theta_m, \end{cases}$$

and has total variation $\|\tilde{w}\|_{(C(\overline{\Omega}))'} = \mathrm{Var}(\tilde{\mu}) = \sum_{j=1}^{m} |\tau_j^{(m)}|$. □

For a given weight, $w$, let $\theta_j^{(m)} \in [-1, 1]$ and $\tau_j^{(m)} \in \mathbb{R}$, $j = 1, \ldots, m$, be the nodes and weights respectively of a family of $m$-point quadrature rules for $m = 1, 2, \ldots$. A family of quadrature rules is called convergent of class $X$ if $\sum_{j=1}^{m} \tau_j^{(m)} f(\theta_j^{(m)}) \to \int_{-1}^{1} wf$ as $m \to \infty$ for all $f \in X$. This definition aligns with intuition; as we include more quadrature nodes and weights, the quadrature rule should approximate the true value of the integral more closely. Commonly used function spaces include $C(\overline{\Omega})$ and Riemann integrable functions. Lemma 7 shows that if the degree of precision tends to infinity as the number of points $m$ tends to infinity (true for all interpolatory quadrature rules), a sufficient condition for the quadrature scheme to be convergent of class $C(\overline{\Omega})$ is $\sup_{m \in \mathbb{N}} \sum_{j=1}^{m} |\tau_j^{(m)}| < \infty$. This follows from $\omega_f(\delta) \to 0$ as $\delta \to 0$ for $f \in C(\overline{\Omega})$, and $\sum_{j=1}^{m} (\tau_j^{(m)} + |\tau_j^{(m)}|) \leq 2 \sum_{j=1}^{m} |\tau_j^{(m)}|$. In fact, in 1933 Pólya showed that these two conditions are necessary and sufficient for a quadrature rule to be convergent of class $C(\overline{\Omega})$ [45]. The condition, $\sup_{m \in \mathbb{N}} \sum_{j=1}^{m} |\tau_j^{(m)}| < \infty$ is sometimes referred to as the

Pólya condition.

Because the sum of the quadrature weights is always the (finite) zeroth moment, it is desirable to have positive weights since in this case the Pólya condition will automatically be satisfied. As a corollary to Lemma 7, we can say that any interpolatory quadrature rule with positive weights is convergent of class $C(\overline{\Omega})$. This is sometimes referred to as Steklov's Theorem [28]. Since the Lanczos process corresponds to Gaussian quadrature, which always has positive weights, we know the Lanczos process for approximating $v^T f(A)v$ is convergent for all $f \in C(\overline{\Omega})$. This should come as no surprise, since for $m = n$, the $m$-step Lanczos algorithm creates a symmetric tridiagonal matrix orthogonally similar to $A$, i.e., $A = V_m T_m V_m^T$, and so $v^T f(A)v = e_1^T f(T_m)e_1$. What we would like to focus on now, is the rate at which the Lanczos process converges.

We remark that this is as far as we may proceed while considering the action of $w \in (C(\overline{\Omega}))'$ on arbitrary $f \in C(\overline{\Omega})$, in the sense that we may only arrive at bounds which tend to zero, and not on convergence rates. This follows because the modulus of continuity of a continuous function can decay arbitrarily slow. In [34], it was demonstrated that for an arbitrary family of quadrature rules which is convergent of class $C(\overline{\Omega})$, a continuous function can be constructed such that quadrature error tends to zero as slowly as desired. Specifically, for any family of quadrature rules convergent of class $C(\overline{\Omega})$, and for any sequence of positive numbers tending to zero, $\{\epsilon_k\}_{k=1}^{\infty}$, one can construct a function $f \in C(\overline{\Omega})$, and an increasing sequence $\{n_k\}_{k=1}^{\infty}$, such that the error in the $n_k$-point quadrature approximation to the integral of $f$ is $\epsilon_k$. Therefore, in order to gain information on the convergence rates of the Lanczos process, we specialize to the space of Hölder continuous functions.

From Jackson's Theorem, we know we can bound the error in the best approximation by the modulus of continuity for a continuous function. The modulus of continuity is particularly easy to characterize for Hölder continuous functions. This will be useful when considering Sobolev spaces, since standard imbedding theorems give conditions for Sobolev spaces to be contained in Hölder spaces. Bounding the modulus of continuity for Hölder continuous functions is done in the following lemma.

**Lemma 8.** *For $f \in C^{0,\gamma}(\Omega)$, $\omega_f(\delta) \leq |f|_{C^{0,\gamma}} \delta^\gamma$.*

*Proof.* The inequality obviously holds for $\delta = 0$. For $\delta > 0$, we have

$$|f|_{C^{0,\gamma}} = \sup_{\substack{x,y\in[-1,1]\\ x\neq y}} \frac{|f(x)-f(y)|}{|x-y|^\gamma} \geq \sup_{\substack{x,y\in[-1,1]\\ |x-y|\leq\delta\\ x\neq y}} \frac{|f(x)-f(y)|}{|x-y|^\gamma}$$

$$\geq \sup_{\substack{x,y\in[-1,1]\\ |x-y|\leq\delta}} \frac{|f(x)-f(y)|}{\delta^\gamma} = \frac{\omega_f(\delta)}{\delta^\gamma}.$$

□

Using Lemma 7 and 8 we are now ready to derive error estimates in Sobolev spaces. From standard Sobolev imbeddings we know that elements of $W^{1,p}(\Omega)$, $p > 1$, are Hölder continuous with exponent $\gamma = 1 - 1/p$ [1, 15]. Therefore, we can apply the results of Lemma 8 using the exponent $1 - 1/p$ for arbitrary $f \in W^{1,p}(\Omega)$.

**Theorem 11.** *Let $1 < p \leq \infty$, $w \in (C(\bar{\Omega}))'$ be positive, and for $\tau_j^{(m)} \in \mathbb{R}$ and distinct $\theta_j^{(m)} \in \bar{\Omega}$, $j = 1, \ldots, m$, define $\tilde{w} \in (C(\bar{\Omega}))'$ as $\tilde{w}(\lambda) = \sum_{j=1}^m \tau_j^{(m)} \delta(\lambda - \theta_j)$. If $w - \tilde{w}$ vanishes on $\mathscr{P}_d$ for $d = d(m) \geq 0$, and $\tau = \sup_{m\in\mathbb{N}} \sum_{j=1}^m |\tau_j^{(m)}| < \infty$, then there exists a constant, $C > 0$, independent of $m$, such that*

$$\|w - \tilde{w}\|_{(W^{1,p})'} \leq \frac{C}{(d+1)^{1-1/p}}.$$

*Proof.* Let $f \in W^{1,p}(\Omega)$ for $p > 1$. Then, there exists a positive constant, independent of $f$, such that $\|f\|_{C^{0,\gamma}} \leq C\|f\|_{W^{1,p}}$ for $\gamma = 1 - 1/p$, see, e.g., [1, 15]. From Lemma 7 we have

$$\left|\int_{-1}^{1} \big(w(\lambda) - \tilde{w}(\lambda)\big) f(\lambda) d\lambda\right| \leq \left(\sum_{j=1}^m \left(\tau_j^{(m)} + \left|\tau_j^{(m)}\right|\right)\right) \omega_f\left(\frac{\pi}{d+1}\right).$$

Using the assumption that the Pólya condition is satisfied, i.e., $\tau = \sup_{m\in\mathbb{N}} \sum_{j=1}^m |\tau_j^{(m)}| < \infty$, we have $\sum_{j=1}^m (\tau_j^{(m)} + |\tau_j^{(m)}|) \leq 2\tau$, with $\tau$ independent of $m$. Additionally, from

Lemma 8 we have that $\omega_f(\pi/(d+1)) \leq |f|_{C^{0,\gamma}}(\pi/(d+1))^\gamma$. Combining these results

$$\begin{aligned}\left|\int_{-1}^{1}\big(w(\lambda)-\tilde{w}(\lambda)\big)f(\lambda)d\lambda\right| &\leq \frac{2\tau\pi^\gamma}{(d+1)^\gamma}|f|_{C^{0,\gamma}},\\ &\leq \frac{2\tau\pi^\gamma}{(d+1)^\gamma}\|f\|_{C^{0,\gamma}},\\ &\leq \frac{C}{(d+1)^\gamma}\|f\|_{W^{1,p}},\end{aligned}$$

which immediately gives the desired result. □

Next, we extend the results of the previous theorem to gain error results in $\big(W^{k,p}(\Omega)\big)'$. For this, we follow [10] and prove $E_k(f) \leq \pi/(2(k+1))E_{k-1}(f')$ in the case that $f$ and $f'$ are continuous.

**Lemma 9.** *Let $f : [-1,1] \to \mathbb{R}$ be continuous with continuous derivative. Then,*

$$E_k(f) \leq \frac{\pi}{2(k+1)}E_{k-1}(f').$$

*Proof.* Let $p_{k-1} \in \mathscr{P}_{k-1}$ be the best uniform approximation of $f'$ and define $p_k(\lambda) = \int_{-1}^{\lambda} p_{k-1}(\theta)d\theta$. Using $p_k'(\lambda) = p_{k-1}(\lambda)$, we see that $E_{k-1}(f') = \|(f-p_k)'\|_\infty$. Furthermore, $f - p_k$ is Lipschitz with Lipschitz constant $L = E_{k-1}(f')$ (follows from the Mean Value Theorem). Applying Jackson's Theorem for the case of a Lipschitz function gives

$$E_k(f) = E_k(f-p_k) \leq \frac{\pi}{2(k+1)}E_{k-1}(f').$$

□

Using Lemma 9 we are now ready to extend Theorem 11 to estimates in $\big(W^{k,p}(\Omega)\big)'$.

**Theorem 12.** *Let $k \geq 1$ be an integer, $1 \leq p \leq \infty$ (with $p > 1$ in the case $k = 1$), let $w \in (C(\overline{\Omega}))'$ be positive, and for $\tau_j^{(m)} \in \mathbb{R}$ and distinct $\theta_j^{(m)} \in \overline{\Omega}$, $j = 1, \ldots, m$, define $\tilde{w}(\lambda) \in (C(\overline{\Omega}))'$ as $\tilde{w} = \sum_{j=1}^m \tau_j^{(m)}\delta(\lambda-\theta_j)$. If $w - \tilde{w}$ vanishes on $\mathscr{P}_d$ for $d = d(m) \geq k-1$ and $\tau = \sup_{m\in\mathbb{N}}\sum_{j=1}^m|\tau_j^{(m)}| < \infty$, then there exists a constant,*

*$C > 0$, independent of $m$, such that*

$$\|w-\tilde{w}\|_{(W^{k,p})'} \leq \frac{C}{(d-k+2)^{k-1/p}}, \qquad k \geq 1,\ p \neq 1,$$

*and, for any $\epsilon \in (0,1)$,*

$$\|w-\tilde{w}\|_{(W^{k,1})'} \leq \frac{C}{(d-k+3)^{k-1-\epsilon}}, \qquad k \geq 2,\ p = 1.$$

*Proof.* We first consider the case $p \neq 1$. With the assumptions on $k$ and $p$, $f \in W^{k,p}(\Omega)$ is an element of $C^{k-1,\gamma}(\Omega)$ for $\gamma = 1 - 1/p$ and satisfies $\|f\|_{C^{k-1,\gamma}} \leq C\|f\|_{W^{k,p}}$ for some constant, $C > 0$, independent of $f$. As is the proof of Lemma 6, we have

$$\Big|\int\limits_{-1}^{1} \big(w(\lambda)-\tilde{w}(\lambda)\big)f(\lambda)d\lambda\Big| \leq \|w-\tilde{w}\|_{(C(\bar{\Omega}))'}\, E_d(f), \tag{3.54}$$

where $E_d(f)$ (see (3.52)) is the error in the best uniform approximation of $f$ in $\mathscr{P}_d$. Using the positivity of $w$, we showed in Lemma 7 that $\|w-\tilde{w}\|_{(C(\bar{\Omega}))'} \leq \sum_{j=1}^{m}(\tau_j^{(m)} + |\tau_j^{(m)}|)$, and so by using our assumption that the Pólya condition is satisfied, $\|w-\tilde{w}\|_{(C(\bar{\Omega}))'} \leq 2\tau$ where $\tau = \sup_{m\in\mathbb{N}}\sum_{j=1}^{m}|\tau_j^{(m)}|$ is independent of $m$. Updating the error (3.54) with the bound on $\|w-\tilde{w}\|_{C(\bar{\Omega})'}$ gives

$$\Big|\int\limits_{-1}^{1} \big(w(\lambda)-\tilde{w}(\lambda)\big)f(\lambda)d\lambda\Big| \leq 2\tau E_d(f). \tag{3.55}$$

In order to bound $E_d(f)$, we apply Lemma 9 $k-1$ times,

$$\begin{aligned} E_d(f) &\leq \left(\frac{\pi}{2}\right)\frac{1}{(d+1)}E_{d-1}(f'), \\ &\leq \left(\frac{\pi}{2}\right)^2\frac{1}{(d+1)(d)}E_{d-2}(f''), \\ &\vdots \\ &\leq \left(\frac{\pi}{2}\right)^{k-1}\frac{1}{(d+1)(d)\cdots(d-k+3)}E_{d-k+1}\big(f^{(k-1)}\big). \end{aligned} \tag{3.56}$$

Using,

$$\frac{1}{(d+1)(d)\cdots(d-k+3)} \leq \frac{1}{(d-k+2)^{k-1}},$$

and,

$$\begin{aligned} E_{d-k+1}(f^{(k-1)}) &\leq \omega_{f^{(k-1)}}\left(\frac{\pi}{(d-k+2)}\right), \\ &\leq \left(\frac{\pi}{(d-k+2)}\right)^{\gamma} \left|f^{(k-1)}\right|_{C^{0,\gamma}}, \\ &\leq \left(\frac{\pi}{(d-k+2)}\right)^{\gamma} \|f\|_{C^{k-1,\gamma}}, \\ &\leq \frac{C}{(d-k+2)^{\gamma}}\|f\|_{W^{k,p}}, \end{aligned}$$

(3.56) becomes

$$E_d(f) \leq \frac{C}{(d-k+2)^{k-1+\gamma}}\|f\|_{W^{k,p}} = \frac{C}{(d-k+2)^{k-1/p}}\|f\|_{W^{k,p}}. \tag{3.57}$$

Combining (3.55) and (3.57) completes the claim for $p \neq 1$. For $p = 1$ and $k \geq 2$, Sobolev imbeddings tell us $f \in W^{k,1}(\Omega)$ is an element of $C^{k-2,1-\epsilon}(\overline{\Omega})$ for any $\epsilon \in (0,1)$. Hence, we perform the same analysis as before, but apply the results of Lemma 9 $k-2$ times, as opposed to the $k-1$ times done previously. In this case the bound on the error in the best uniform approximation to $f$ in $\mathscr{P}_d$ satisfies

$$E_d(f) \leq \frac{C}{(d-k+3)^{k-1-\epsilon}}\|f\|_{W^{k,1}}.$$

□

We may now apply the results of Theorem 12 to the Lanczos process approximation of the spectral function. The next corollary follows from Theorem 12 by using $d(m) = 2m-1$ and noting that because the Gaussian quadrature weights are positive, the Pólya condition is always satisfied.

**Corollary 1.** *Let $A \in \mathbb{R}^{n\times n}$ be symmetric with eigenvalues in the interval $[-1,1]$ and $v \in \mathbb{R}^n$ be a unit vector, let $k \geq 1$ be an integer, $1 \leq p \leq \infty$ (with $p > 1$ in the case $k = 1$), let $s(\lambda)$ be the spectral function corresponding to $A$ and $v$, and $\tilde{s}(\lambda)$ be the approximation determined by the $m$-step Lanczos process with $A$ and $v$. Then, there exists a constant, $C > 0$, independent of $m$, such that the error in the $m$-step Lanczos* *approximation to the spectral function is*

$$\|s-\tilde{s}\|_{(W^{k,p})'} \leq \frac{C}{(2m-k+1)^{k-1/p}}, \qquad k \geq 1,\ p \neq 1,$$

*and, for any* $\epsilon \in (0,1)$,

$$\|s-\tilde{s}\|_{(W^{k,1})'} \leq \frac{C}{(2m-k+2)^{k-1-\epsilon}}, \qquad k \geq 2,\ p = 1.$$

In this section we showed that performing the Lanczos process to approximate the quadratic form $v^T f(A)v$ is equivalent to approximating the spectral function $s(\lambda)$. We also stated a known error estimate for the Lanczos approximation to $v^T f(A)v$ when $f$ is analytic. Finally, we gave a new bound for the error, $s-\tilde{s}$, in Sobolev spaces which has not previously appeared in the literature.

## 3.4 Quadratic Forms for Generalized Systems

Thus far, we have looked at approximating the quadratic form $v^T f(A)v$, which is determined by the inner products of $v$ with the eigenvectors of $A$, and $f$ evaluated at the eigenvalues. In this section we look at the equivalent situation where the eigenvalues and eigenvectors stem from a generalized eigenvalue problem. We are most interested in finite element discretizations of eigenvalue problems for elliptic operators, which results in a generalized algebraic eigenvalue problem.

Let $A, B \in \mathbb{R}^{n\times n}$ be symmetric with $B$ positive definite. Because $B$ is symmetric positive definite, it can be used to define an inner product $(x,y)_B = (x,By) = x^T By$, for all $x, y \in \mathbb{R}^n$, with corresponding induced norm $\|x\|_B = \sqrt{(x,x)_B}$. We are interested in the analogs of (3.1) and (3.2) using the eigenpairs of the generalized eigenvalue problem

$$Ax_i = \lambda_i Bx_i, \quad x_i^T Bx_j = \delta_{ij}, \tag{3.58}$$

for $i, j = 1, \ldots, n$. Letting $X = [x_1 \ldots x_n]$ and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, we can rewrite (3.58) as $AX = BX\Lambda$ and $X^T BX = I$.

We can rewrite the generalized eigensystem (3.58) as a standard eigenvalue problem with the matrix $B^{-1}A$, however, the matrix $B^{-1}A$ is not symmetric with respect to the standard Euclidean inner product. Fortunately, it is symmetric with respect to the $B$-inner product. So, given a smooth function $f$ and a vector $v \in \mathbb{R}^n$ satisfying $\|v\|_B = 1$, a natural extension of the quadratic form (3.1) for the case of a generalized eigensystem is

$$v^T f_{A,B} v = \sum_{i=1}^{n} |(x_i, v)_B|^2 f(\lambda_i), \tag{3.59}$$

where $f_{A,B}$, a matrix to be determined, depends on $f$, $A$, and $B$. By developing the right-hand side of (3.59) it is easily checked that $f_{A,B} = BXf(\Lambda)X^T B$. The "generalized" spectral function in this case is now $s(\lambda) = \sum_{i=1}^{n} |(x_i, v)_B|^2 \delta(\lambda - \lambda_i)$.

In order to investigate approximation of the spectral function corresponding to $A$, $B$, and $v$, we first transform the generalized eigenvalue problem to a standard eigenvalue problem, and then relate (3.59) to results of the previous section. Let $B = LL^T$ be the Cholesky factorization of $B$, where $L$, the Cholesky factor, is lower triangular with positive entries on the diagonal. Using the Cholesky factor $L$, we rewrite (3.58) as

$$\left(L^{-1} A L^{-T}\right)\left(L^T x_i\right) = \lambda_i \left(L^T x_i\right), \quad \left(L^T x_i\right)^T \left(L^T x_j\right) = \delta_{ij}, \tag{3.60}$$

for $i, j = 1, \ldots, n$. Therefore, defining $C = L^{-1} A L^{-T}$ and $z_i = L^T x_i$, $i = 1, \ldots, n$, we have a standard eigenvalue problem for the symmetric matrix $C$. The eigenpairs of the matrix $C$ completely determine the eigenpairs of the generalized system, and vice versa.

With $Z = [z_1 \ldots z_n] = L^T X$ and $u = L^T v$ (note $\|u\| = 1$ since we have assumed $\|v\|_B = 1$), we have $(x_i, v)_B = (z_i, u)$ for $i = 1, \ldots, n$. Therefore,

$$\begin{aligned} v^T f_{A,B} v &= \sum_{i=1}^{n} |(x_i, v)_B|^2 f(\lambda_i) = \sum_{i=1}^{n} |(z_i, u)|^2 f(\lambda_i), \\ &= (Z^T u)^T f(\Lambda)(Z^T u) = u^T f(C) u. \end{aligned} \tag{3.61}$$

In other words, by performing the Lanczos process with the matrix $C = L^{-1} A L^{-T}$ and vector $u = L^T v$, we can approximate the quadratic form $v^T f_{A,B} v$, and hence the spectral function $s$. Next, we show that the Lanczos partial tridiagonalization of $C = L^{-1} A L^{-T}$ with starting vector $u = L^T v$, is the same as the partial tridiagonalization resulting from the $B$-Lanczos algorithm with $A$, $B$, and $v$.

The $m$-step Lanczos algorithm applied to $C$ with starting vector $u = L^T v$ gives

$$CU_m = U_m T_m + \beta_m u_{m+1} e_m^T, \tag{3.62}$$

where the columns of $U_m$ are an orthonormal basis of the Krylov space $\mathcal{K}_m(C, u)$, $T_m \in \mathbb{R}^{m \times m}$ is a symmetric tridiagonal matrix, and $\beta_m u_{m+1} e_m^T$ is the rank-one remainder term. Notice that by premultiplying (3.62) by $L$ and defining $V_m = L^{-T} U_m$ and $v_{m+1} = L^{-T} u_{m+1}$, we find

$$AV_m = BV_m T_m + \beta_m B v_{m+1} e_m^T, \tag{3.63}$$

which is the $B$-Lanczos algorithm with starting vector $v$. Notice that, as expected, the columns of $V_m$ are $B$-orthonormal since $V_m^T B V_m = U_m^T U_m = I$.

As in Section 3.1, the Lanczos partial tridiagonalization is the Jacobi matrix, and therefore the nodes and weights of the $m$-point Gaussian quadrature rule are determined by the eigenpairs of $T_m$. Letting $\theta_j$ and $y_j$, $j = 1, \ldots, m$, denote the eigenvalues and orthonormal eigenvectors of $T_m$, respectively, the Lanczos process approximation of the quadratic form (3.59) is

$$v^T f_{A,B} v = \sum_{i=1}^{n} |(x_i, v)_B|^2 f(\lambda_i) \approx \sum_{j=1}^{m} |(y_j, e_1)|^2 f(\theta_j) = e_1 f(T_m) e_1.$$

Note that the moment matching property for the Lanczos process with matrix $C = L^{-1} A L^{-T}$ and vector $u = L^T v$ states that,

$$\sum_{i=1}^{n} |(z_i, u)|^2 \lambda_i^k = \sum_{j=1}^{m} |(y_j, e_1)|^2 \theta_j^k, \qquad k = 0, 1, \ldots, 2m - 1.$$

Using the fact that $(z_i, u) = (x_i, v)_B$, we derive a moment matching property for the $B$-Lanczos process applied to the matrix pair $A$, $B$, and $B$-unit vector $v$

$$\sum_{i=1}^{n} |(x_i, v)_B|^2 \lambda_i^k = \sum_{j=1}^{m} |(y_j, e_1)|^2 \theta_j^k, \qquad k = 0, 1, \ldots, 2m - 1. \tag{3.64}$$

We have discussed two options for using the Lanczos process to approximate the quadratic form $v^T f_{A,B} v$:

1. Perform the Lanczos algorithm on $C = L^{-1}AL^{-T}$ with starting vector $u = L^T v$. Each iteration requires two linear solves (one with $L^T$ and another with $L$) and one matrix vector multiplication with $A$.

2. Perform the $B$-Lancozs algorithm with $A$, $B$, and starting vector $v$. Each iteration requires one linear solve with $B$ and one matrix vector multiplication with $A$.

Clearly (2) is the more cost-effective option without even taking into consideration the cost of computing the Cholesky factor of $B$.

To summarize, the Lanczos process for approximating the spectral function $s(\lambda) = \sum_{i=1}^{n} |(x_i, v)_B|^2 f(\lambda_i)$ using an $m$-point quadrature rule is:

1. Perform the $m$-step $B$-Lanczos algorithm with $A$, $B$, and starting vector $v$, $\|v\|_B = 1$, to get the symmetric tridiagonal matrix $T_m$.

2. Compute the eigenpairs of $T_m$, $T_m y_j = \theta_j y_j$, $\|y_j\| = 1$, for $j = 1, \ldots m$.

3. $v^T f_{A,B} v \approx \sum\limits_{j=1}^{m} |(y_j, e_1)|^2 f(\theta_j)$.

# Chapter 4

# Lanczos Approximation of Joint Spectral Quantities

## 4.1 Spectral Quantities

In this chapter we discuss approximating linear combinations of Dirac measures of the form $q(\lambda) = \sum_i w_i \delta(\lambda - \lambda_i)$, where the $\lambda_i$'s are eigenvalues of a symmetric matrix, and the coefficients, $w_i$, may or may not depend on the eigenvector corresponding to $\lambda_i$. We refer to $q(\lambda)$ as a "spectral quantity." A simple example of a spectral quantity which we have already encountered is the spectral function corresponding to a symmetric matrix and given vector. In order to approximate spectral quantities, we utilize the Lanczos process discussed in the previous chapter. We begin by discussing the Lanczos approximation to the density of states for a matrix, the results of which are known. Afterward, we advance to joint spectral quantities, which are spectral quantities involving the eigenpairs of two separate systems. These are of the form $\sum_{i,j} w_{ij} \delta\big(\lambda - (\lambda_i + \lambda'_j)\big)$, where the $\lambda_i$ and $\lambda'_j$ are the eigenvalues of two different symmetric matrices, and the coefficients, $w_{ij} \in \mathbb{R}$, may depend on the corresponding eigenvectors. The two joint spectral quantities discussed in this chapter are the joint density of states and the joint spectral function. The joint density of states is a natural extension of the density of states for two separate eigenvalue problems, and the joint spectral function is of great utility when determining optical properties of semiconductors. In all cases of spectral and joint spectral quantities, we first discuss the case of a standard eigenvalue problem, and then the extension to generalized eigenvalue problems. To the best of the authors knowledge, the methods presented here for the joint density of states and the joint spectral function are new. All methods are discussed from a numerical linear algebra perspective, with applications considered in the next chapter.

Throughout this chapter we use $A, A' \in \mathbb{R}^{n\times n}$ to denote symmetric matrices, and $B \in \mathbb{R}^{n\times n}$ for a symmetric positive definite matrix. For standard eigenvalue problems, we use the following notation for eigenpairs

$$\begin{aligned} Ax_i &= \lambda_i x_i, & x_i^T x_j &= \delta_{ij}, \\ A'x_i' &= \lambda_i' x_i', & x_i'^T x_j' &= \delta_{ij}, \end{aligned} \tag{4.1}$$

for $i, j = 1, \ldots, n$, and for the case of generalized eigenvalue problems we use

$$\begin{aligned} Ax_i &= \lambda_i B x_i, & x_i^T B x_j &= \delta_{ij}, \\ A'x_i' &= \lambda_i' B x_i', & x_i'^T B x_j' &= \delta_{ij}, \end{aligned} \tag{4.2}$$

for $i, j = 1, \ldots, n$. In both cases we assume the eigenvalues are in ascending order, $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$, and similarly for the $\lambda_i'$'s. Note in the case of generalized eigensystems, the right hand side matrix is $B$ for both systems (as opposed to using $B$ for one and $B'$ for the other). This is due to the fact that the generalized systems (4.2) we are interested in are finite element discretizations of eigenvalue problems for elliptic operators. In this case, using the same $B$ (Galerkin mass) matrix for the primed and unprimed systems represents using the same finite dimensional subspace to approximate infinite dimensional eigenfunctions.

As we saw in the previous chapter, given a vector $v \in \mathbb{R}^n$, the Lanczos process can be used to construct approximations to a measure on the real line which is dependent on the spectrum of $A$ and the vector $v$. In this chapter we exclusively refer to this measure as the "spectral function," given by

$$s(\lambda; A, v) = \sum_{i=1}^{n} |(x_i, v)|^2 \delta(\lambda - \lambda_i), \tag{4.3}$$

or, in the case of a generalized eigensystem,

$$s(\lambda; A, B, v) = \sum_{i=1}^{n} |(x_i, v)_B|^2 \delta(\lambda - \lambda_i), \tag{4.4}$$

where $\delta$ is the Dirac distribution, $(\,\cdot\,,\,\cdot\,)$ and $(\,\cdot\,,\,\cdot\,)_B$ are the Euclidean and $B$-inner products, and the eigenpairs are as in (4.1) or (4.2) respectively. In the last chapter we showed how to construct an approximation to the spectral function using the Lanczos process, in addition to giving error estimates. Specifically, we use the $m$-step Lanczos algorithm to partially tridiagonalize the matrix $A$, with starting vector $v$, to obtain a symmetric tridiagonal matrix $T_m \in \mathbb{R}^{m \times m}$, from which we are able to approximate the spectral function. Denoting the eigenpairs of the partial tridiagonalization as $T_m y_j = \theta_j y_j$, $j = 1, \ldots, m$, (note that we suppress the dependence of the eigenpairs on $m$ for notational convenience), the Lanczos approximation to the spectral function (4.3) is

$$\tilde{s}(\lambda) = \|v\|^2 \sum_{j=1}^{m} |(y_j, e_1)|^2 \delta(\lambda - \theta_j), \tag{4.5}$$

where $e_1$ is the first column of the $m \times m$ identity matrix. Previously, when discussing the Lanczos process we assumed $v$ was a unit vector, and so the prefactor, $\|v\|^2$, was absent. However, in this chapter we will mostly be dealing with non-unit vectors, and so the extra term is necessary to incorporate. Note that the Lanczos approximation to the spectral function (4.4) is the same as above, the only difference being that the matrix $T_m$ is constructed using the $B$-Lanczos algorithm and the prefactor becomes $\|v\|_B^2$. Due to the relationship between the Lanczos process and Gauss quadrature, the Lanczos approximation to the spectral function matches the first $2m - 1$ moments of the spectral function. That is, the Lanczos process approximation satisfies,

$$\sum_{i=1}^{n} |(x_i, v)|^2 \lambda_i^k = \sum_{j=1}^{m} |(y_j, e_1)|^2 \theta_j^k, \qquad k = 0, 1, \ldots, 2m - 1, \tag{4.6}$$

or, for the case of a generalized system,

$$\sum_{i=1}^{n} |(x_i, v)_B|^2 \lambda_i^k = \sum_{j=1}^{m} |(y_j, e_1)|^2 \theta_j^k, \qquad k = 0, 1, \ldots, 2m - 1. \tag{4.7}$$

For notational convenience, in this chapter we use $\langle\,\cdot\,,\,\cdot\,\rangle$ to denote the dual pairing of $(C(\overline{\Omega}))' \times C(\overline{\Omega})$, where $\Omega$ is an open interval, and its closure, $\overline{\Omega}$, contains the eigenvalues associated to the spectral quantity. That is, for a spectral quantity $q(\lambda) = \sum_i w_i \delta(\lambda - \lambda_i)$, with $\lambda_i \in \overline{\Omega}$ and $w_i \in \mathbb{R}$ for all $1 \leq i \leq n$, the dual pairing is $\langle q, f \rangle = \sum_i w_i f(\lambda_i)$ for all $f \in C(\overline{\Omega})$. Throughout this chapter we refer to elements of $C(\overline{\Omega})$ as test functions.

For all spectral and joint spectral quantities discussed in this chapter, the template for producing Lanczos approximations is the same. First, we relate the spectral quantity to a spectral function (4.3) or (4.4), depending on whether a standard or generalized eigenvalue problem is under consideration. We then use the Lanczos process to construct an approximation to the spectral function. By following this template, we construct accurate approximations to the spectral quantity of interest. We begin with the Lanczos approximation to the density of states, which is the simplest example of a spectral quantity.

## 4.2 Density of States

The first spectral quantity we approximate using the Lanczos process is the density of states [33]. Formally, the density of states of the symmetric matrix $A$ is

$$\phi(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \delta(\lambda - \lambda_i), \tag{4.8}$$

where the eigenvalues, $\lambda_i$, are as in (4.1). This is an example of a spectral quantity where the coefficients are uniformly equal to $1/n$. The density of states is of great interest in electronic structure calculations as well as in large scale parallel eigenvalue computations.

We briefly explain one practical use of the density of states. Given a symmetric matrix $A$, and real numbers $\mu < \nu$, distinct from any eigenvalues of $A$, suppose we wish to know how many eigenvalues are inside the interval $(\mu, \nu)$. Denote this quantity as $N(\mu, \nu)$. Classically, one computes $N(\mu, \nu)$ using the inertia of spectral transformations of $A$. As a consequence of Sylvester's law of inertia, the number of eigenvalues in the interval $(\mu, \nu)$ is the difference in the number of positive entries on the diagonal of $D_\mu$ and $D_\nu$, where $D_\mu$ and $D_\nu$ are the diagonal matrices in the $LDL^T$ factorizations of

$A - \mu I$ and $A - \nu I$ respectively (assuming they exist) [44]. This generalizes to computing the $LDL^T$ factorization of $A - \mu B$ and $A - \nu B$ in the case of a generalized eigenvalue problem. However, computing $LDL^T$ factorizations is prohibitively expensive if the matrix is large, or if we wish to compute $N(\mu, \nu)$ for several different values of $\mu$ and $\nu$. On the other hand, using the density of states we can compute $N(\mu, \nu)$ as

$$N(\mu, \nu) = \int_{\mu}^{\nu} n\phi(\lambda) d\lambda. \tag{4.9}$$

Note that the above shows how the density of states is, indeed, a density. Using the density of states, or an approximation thereof, computing $N(\mu, \nu)$ for several different values of $\mu$ and $\nu$ is simple and inexpensive using (4.9). An approximation to the density of states can be used, for example, to partition the spectrum of $A$ into subintervals with a roughly equal number of eigenvalues. That is, by determining values $t_0 < t_1 < \ldots < t_{n_s}$, such that $N(t_i, t_{i+1}) \approx n/n_s$ for $i = 0, \ldots, n_s - 1$, we can divide the spectrum into $n_s$ subintervals, each of which contain a similar number of eigenvalues. This is used in the parallel spectrum slicing software Eigenvalue Slicing Library [31].

Next, we focus on using the Lanczos process to approximate (4.8). If we are able to construct a unit vector $v \in \mathbb{R}^n$ equally weighted in the direction of each eigenvector, i.e., a vector which satisfies $(x_i, v) = \pm 1/\sqrt{n}$ for $i = 1, \ldots, n$, then the spectral function corresponding to $A$ and $v$ would be exactly the density of states. Hence, the Lanczos process with matrix $A$ and starting vector $v$ would give an accurate approximation of the density of states $\phi(\lambda)$. However, without complete knowledge of the spectrum of $A$ a priori, this is not possible. Instead, methods to approximate the density of states using the Lanczos process rely on a stochastic technique to remove the influence of the coefficients $|(x_i, v)|^2$ (see (4.3)) in the spectral function corresponding to $A$ and $v$. This concept is closely related to Monte Carlo trace estimators.

The method of approximating the density of states we discuss utilizes a result of Hutchinson [24]. Following [32], we refer to a stochastic method which utilizes the following lemma as Hutchinson's method. We use $\mathcal{N}(0, 1)$ to denote the standard normal distribution, and for a random variable $w$ taking values in $\mathbb{R}^n$, we write $w \sim \mathcal{N}(0, 1)$

when the entries of $w$ are drawn from $\mathcal{N}(0,1)$ randomly with independently and identically distributed (i.i.d.) values. We refer to such vectors as standard normal random $n$-vectors.

**Lemma 10** (Hutchinson)**.** *Let $C \in \mathbb{R}^{n\times n}$ and $w \sim \mathcal{N}(0,1)$. Then,*

$$\mathbb{E}[w^T C w] = \mathrm{tr}(C).$$

*Proof.* Because $w$ is a standard normal random $n$-vector, the entries satisfy

$$\mathbb{E}[w_i w_j] = \mathbb{E}[w_i]\mathbb{E}[w_j] = 0, \quad i \neq j, \qquad \text{and} \qquad \mathbb{E}[w_i^2] = 1.$$

In other words, $\mathbb{E}[ww^T] = I$, with expectation understood componentwise. Using the linearity of expectation we find

$$\mathbb{E}[w^T C w] = \mathbb{E}\Bigg[\sum_{i,j=1}^{n} c_{ij} w_i w_j\Bigg] = \sum_{i,j=1}^{n} c_{ij}\,\underbrace{\mathbb{E}[w_i w_j]}_{\delta_{ij}} = \mathrm{tr}(C).$$

□

We remark that for any $n \times n$ matrix $C$, and random variable $w$ taking values in $\mathbb{R}^n$ with i.i.d. entries drawn from a probability distribution, the only requirement necessary for $\mathbb{E}[w^T C w] = \mathrm{tr}(C)$ is $\mathbb{E}[ww^T] = I$. The standard normal distribution is not the only distribution satisfying this property. In fact, it is not difficult to show that if $w$ has i.i.d. entries drawn from a distribution with mean $\mu$ and variance $\sigma^2$, then $\mathbb{E}[w^T C w] = \sigma^2 \mathrm{tr}(C) + \mu^2 \sum_{i,j=1}^{n} c_{ij}$ (follows directly from $\mathbb{E}[w_i w_j] = \mu^2 + \sigma^2 \delta_{ij}$). Therefore, if the entries of $w$ are drawn from any distribution with zero mean and unit variance, then $\mathbb{E}[w^T C w] = \mathrm{tr}(C)$. Another distribution satisfying this property is the Rademacher distribution, in which random variables take values ±1 with equal probability. Other distributions which work, along with error bounds for stochastic trace estimation, can be found in [4].

Next, we apply Lemma 10 to a matrix given by an outer product. This will illustrate how we intend to use stochastic processes to remove the influence of the coefficients $|(x_i, v)|^2$ in the spectral function.

**Corollary 2.** *For $x, y \in \mathbb{R}^n$ and $w \sim \mathcal{N}(0,1)$, $\mathbb{E}\big[(x,w)(y,w)\big] = (x,y)$. In particular, if $x = y$, $\mathbb{E}\big[|(x,w)|^2\big] = \|x\|^2$*

*Proof.* Apply Lemma 10 with $C = xy^T$, noting that $\mathrm{tr}(C) = (x,y)$. □

Corollary 2 gives the main intuition on how we plan on approximating the density of states using spectral functions. For a standard normal random $n$-vector $w$, Corollary 2 tells us $\mathbb{E}\big[|(x_i,w)|^2\big] = 1$, where the $x_i \in \mathbb{R}^n$, $i = 1, \dots, n$, are the orthonormal eigenvectors of the matrix $A$ given by (4.1). Therefore, the spectral function corresponding to $A$ and $w$, $s(\lambda; A, w)$, is a linear combination of Dirac deltas concentrated at the eigenvalues of $A$, with coefficients $|(x_i,w)|^2$. Each of these coefficients has an expectation of unity, and thus matches the coefficients of the density of states on average (disregarding the prefactor $1/n$). By averaging spectral functions corresponding to several standard normal random $n$-vectors, we stochastically approximate the density of states by removing the influence of the coefficients, $|(x_i,w)|^2$, from each individual spectral function. Next, we proceed more formally in outlining the Lanczos approximation to the density of states.

Notice that the action of the density of states on any test function $f$ satisfies

$$\langle \phi, f \rangle = \frac{1}{n}\mathrm{tr}\big(f(A)\big),$$

where we used the property that the trace of a matrix is the sum of its eigenvalues. Applying Lemma 10 with the matrix $C = f(A)$, we see that $\langle \phi, f \rangle = 1/n\mathbb{E}[w^T f(A) w]$ for any test function $f$, where $w$ is a standard normal random $n$-vector. The spectral function corresponding to $A$ and $w$, and the quadratic form $w^T f(A) w$, are related by

$$w^T f(A) w = \langle s(\lambda; A, w), f \rangle,$$

for any test function $f$. Putting everything together, by choosing $w \sim \mathcal{N}(0,1)$, the spectral function corresponding to $A$ and $w$ is related to the density of states by

$$\langle \phi, f \rangle = \frac{1}{n}\mathrm{tr}(f(A)) = \frac{1}{n}\mathbb{E}[w^T f(A) w] = \frac{1}{n}\mathbb{E}\Big[\langle s(\lambda; A, w), f \rangle\Big], \tag{4.10}$$

for all test functions $f$. Written another way,

$$\phi(\lambda) = \frac{1}{n}\mathbb{E}\Big[s(\lambda; A, w)\Big]. \tag{4.11}$$

What (4.11) shows is that we can approximate the density of states by averaging spectral functions corresponding to $A$, and vectors with i.i.d. entries drawn from $\mathcal{N}(0, 1)$. By averaging spectral functions, $s(\lambda; A, w)$, for several standard normal random $n$-vectors $w$, we create something near the expected value, which, according to (4.11), is the density of states. We refer to the standard normal random $n$-vectors used in the averaging process as trial vectors. Choosing $n_v$ trial vectors, $w^{(k)} \sim \mathcal{N}(0, 1)$, $k = 1, \dots, n_v$, the stochastic approximation to the density of states using spectral functions is

$$\phi(\lambda) \approx \frac{1}{n_v n}\sum_{k=1}^{n_v} s(\lambda; A, w^{(k)}) = \frac{1}{n_v n}\sum_{k=1}^{n_v}\sum_{i=1}^{n} |(x_i, w^{(k)})|^2 \delta(\lambda - \lambda_i), \tag{4.12}$$

where, as in (4.11), the approximation is an equality with respect to expected value. Due to the law of large numbers, the more trial vectors we choose, the closer we expect the right hand side of (4.12) to match the density of states.

Now, with the density of states related to spectral functions through (4.12), we are primed to use the Lanczos process to approximate the density of states. Since the Lanczos process produces an approximation to a spectral function, we simply replace all spectral functions in (4.12) with the corresponding Lanczos approximation. For any one trial vector $w$, we perform the $m$-step Lanczos algorithm on $A$ with $w$ as starting vector, obtaining the partial tridiagonalization $T_m \in \mathbb{R}^{m \times m}$. The eigenvalues, also known as Ritz values, of $T_m$, $\theta_j$, and corresponding orthonormal eigenvectors $y_j$, $j = 1, \dots, m$, determine the nodes and weights for the Lanczos approximation to $s(\lambda; A, w)$ by

$$s(\lambda; A, w) \approx \|w\|^2 \sum_{j=1}^{m} |(y_j, e_1)|^2 \delta(\lambda - \theta_j). \tag{4.13}$$

By replacing each spectral function in (4.12) with the approximation from the Lanczos process, as in (4.13), we stochastically approximate the density of states.

To this end, let $w^{(k)} \sim \mathcal{N}(0, 1)$, $k = 1, \dots, n_v$, denote trial vectors. For each trial vector we partially tridiagonalize $A$ by performing the $m$-step Lanczos algorithm on

$A$ with starting vector $w^{(k)}$, obtaining $T_m^{(k)} \in \mathbb{R}^{m\times m}$. Computing the eigenpairs of each partial tridiagonalization, $T_m^{(k)} y_j^{(k)} = \theta_j^{(k)} y_j^{(k)}$, $(y_i^{(k)}, y_j^{(k)}) = \delta_{ij}$, $i, j = 1, \ldots, m$, the Lanczos approximation to the density of states is

$$\tilde{\phi}(\lambda) = \frac{1}{n_v n} \sum_{k=1}^{n_v} \sum_{j=1}^{m} \|w^{(k)}\|^2 \big|(y_j^{(k)}, e_1)\big|^2 \delta\big(\lambda - \theta_j^{(k)}\big). \tag{4.14}$$

Approximating the density of states using the Lanczos process is summarized in Algorithm 10. Note that each of the $n_v$ Lanczos processes needed to compute (4.14) are completely independent of the others, making the computation of $\tilde{\phi}$ embarrassingly parallel. For this reason we dispense with the superscripts on all vectors in Algorithm 10.

**Algorithm 10** Lanczos Approximation of the Density of States

```
Initialize $n_v$, $m$, and set $k = 0$ and $\tilde{\phi}(\lambda) = 0$.
while $k < n_v$ do
    Draw trial vector $w \sim \mathcal{N}(0, 1)$.
    Partially tridiagonalize $A$ with starting vector $w$ to get $T_m \in \mathbb{R}^{m\times m}$.
    Compute eigenpairs $T_m y_j = \theta_j y_j$, $y_i^T y_j = \delta_{ij}$, $i, j = 1, \ldots, m$.
    $\tilde{\phi}(\lambda) \leftarrow \tilde{\phi}(\lambda) + \frac{\|w\|^2}{n_v n} \sum_{j=1}^{m} |(y_j, e_1)|^2 \delta(\lambda - \theta_j)$.
    $k \leftarrow k + 1$.
end while
```

## 4.3 Density of States for Generalized Eigenvalue Problems

In this section we are interested in computing the density of states, $\phi(\lambda) = 1/n \sum_{i=1}^{n} \delta(\lambda - \lambda_i)$, associated to the generalized eigenvalue system (4.2). The Lanczos approximation to the density of states for the generalized eigenvalue system is the same as in the standard eigenvalue problem, with one major exception. Namely, special care must be taken when determining the starting vector for the Lanczos process. The definitive source for approximating the density of states for a generalized eigensystem is [67], which our presentation closely follows.

Assuming the matrices in the generalized eigensystem (4.2) are the result of a finite element discretization, we first modify the stiffness and mass matrices without altering the eigenvalues of the system. We do this for two main reasons. First, as with all

Lanczos methods for generalized eigensystems, we expect to use the $B$-Lanczos algorithm, which requires a linear solve with $B$ each iteration. Iterative linear solves can obviously be expected to converge faster if the coefficient matrix is well conditioned. Second, the methods presented in this section require us to factor the matrix $B$. Matrix factorization can be prohibitively expensive for large matrices, and so we overview an economical method to approximate the factorization using Chebyshev polynomials. When the matrix $B$ is well conditioned, we show that the approximate factorization is more efficient.

In order to ensure the mass matrix is well conditioned, we utilize a well known fact about the finite element method. Let $D = \text{diag}(B)$ be the diagonal matrix with entries on the diagonal equal to those of the diagonal of the mass matrix. Because we have assumed the mass matrix is positive definite, all elements on the diagonal of $B$ are positive, and therefore all elements of $D$ are nonnegative. Using $D$, we replace the stiffness and mass matrix with $A \leftarrow D^{-1/2}AD^{-1/2}$ and $B \leftarrow D^{-1/2}BD^{-1/2}$, transforming the generalized eigensystem to

$$\Big(D^{-1/2}AD^{-1/2}\Big)\big(D^{1/2}x\big) = \lambda\Big(D^{-1/2}BD^{-1/2}\Big)\big(D^{1/2}x\big), \tag{4.15}$$

where $D^{1/2}$ is the diagonal matrix with entries equal to the square root of the entries of $D$ and $D^{-1/2} = (D^{1/2})^{-1}$. The new stiffness and mass matrices, $D^{-1/2}AD^{-1/2}$ and $D^{-1/2}BD^{-1/2}$ respectively, are inexpensive to compute at the outset, and the new mass matrix is well-conditioned. In [62] it was shown that for any conforming mesh of linear triangles (two-dimensions), the scaled mass matrix has spectral condition number bounded by four. In three dimensions, for conforming linear tetrahedral elements, the condition number of the scaled mass matrix is bounded by five. In what follows we assume the scaling (4.15) has already been performed, and denote the scaled stiffness and mass matrices by $A$ and $B$ respectively.

Next, we transform the generalized eigenvalue problem into a standard eigenvalue problem in order to show how the stochastic methods from the previous section translate to the generalized eigenvalue problem. Toward this end, let $B = LL^T$ denote a factorization of $B$, e.g., the Cholesky factorization or the square root factorization. As seen previously, using the factorization of the mass matrix, the generalized eigenvalue problem becomes the standard eigenvalue problem

$$Cz_i = \lambda_i z_i, \qquad z_i^T z_j = \delta_{ij}, \qquad i,j = 1,\ldots,n, \tag{4.16}$$

where $C = L^{-1}AL^{-T}$ and $z_i = L^T x_i$.

Because the matrix $C$ has the same eigenvalues as the generalized eigenvalue problem with $A$ and $B$, we can apply the methods of the previous section to approximate the density of states. Choosing a trial vector $w \sim \mathcal{N}(0,1)$, the spectral function corresponding to $C$ and $w$ equals the density of states in expected value,

$$\phi(\lambda) = \frac{1}{n}\mathbb{E}\Big[s(\lambda; C, w)\Big], \tag{4.17}$$

as in (4.11). We next relate the spectral function $s(\lambda; C, w)$, to that of $s(\lambda; A, B, v)$, for a properly chosen vector $v$.

For the eigenvectors, $z_i = L^T x_i$, of $C = L^{-1}AL^{-T}$ (as in (4.16)), notice that for any vector $u \in \mathbb{R}^n$ we have

$$(z_i, u) = z_i^T u = x_i^T L u = x_i^T \underbrace{(LL^T)}_{B}(L^{-T}u) = (x_i, v)_B, \tag{4.18}$$

where $v = L^{-T}u$. From (4.18) we are able to relate the spectral function corresponding to $C$ and trial vector $w$, and the spectral function corresponding to the generalized system with $A$ and $B$, and starting vector $v = L^{-T}w$. Indeed, for $w \sim \mathcal{N}(0,1)$, and $v = L^{-T}w$, by (4.18) we have

$$s(\lambda; C, w) = \sum_{i=1}^{n} |(z_i, w)|^2 \delta(\lambda - \lambda_i) = \sum_{i=1}^{n} |(x_i, v)_B|^2 \delta(\lambda - \lambda_i) = s(\lambda; A, B, v).$$

This illustrates the main distinction between the Lanczos approximation to the density of states for a standard eigenvalue problem and a generalized eigenvalue problem. For the Lanczos approximation to the density of states of a matrix, the trial vector $w \sim \mathcal{N}(0,1)$ is the same as the starting vector for the Lanczos algorithm. On the other hand, when approximating the density of states for a generalized eigenvalue problem, the trial vector and starting vector are different. In summary, when applying Hutchinson's method to a generalized eigenvalue problem we use $L^{-T}w$ as starting vector for the $B$-Lanczos algorithm, where $w$ is a standard normal random $n$-vector and $B = LL^T$, as opposed to using $w$ as the starting vector in the Lanczos algorithm for the standard eigenvalue problem. Note that by defining $v = L^{-T}w$, we have $\|v\|_B^2 = \|w\|^2$, and so when using the Lanczos process to approximate the density of states for a generalized eigensystem, each trial will have a prefactor of $\|w\|^2$, the same as the Lanczos approximation to the density of states for a matrix.

As noted previously, we produce an approximation to the density of states by stochastically averaging spectral functions over many different trial vectors $w$. By replacing the spectral functions with their corresponding Lanczos approximation, we determine a computable approximation to the density of states for a generalized system. The recipe for using Hutchinson's method in conjunction with the Lanczos process to approximate the density of states for a generalized system is given in Algorithm 11.

**Algorithm 11** $B$-Lanczos Approximation of the Density of States

```
Initialize $n_v$, $m$, and set $k = 0$ and $\tilde{\phi}(\lambda) = 0$.
Factor $B = LL^T$.
**while** $k < n_v$ **do**
    Draw trial vector $w \sim \mathcal{N}(0, 1)$.
    Form starting vector $v = L^{-T}w$.
    Perform $B$-Lanczos with $A$, $B$, and vector $v$, to get $T_m \in \mathbb{R}^{m \times m}$.
    Compute eigenpairs $T_m y_j = \theta_j y_j$, $y_i^T y_j = \delta_{ij}$, $i, j = 1, \ldots, m$.
    $\tilde{\phi}(\lambda) \leftarrow \tilde{\phi}(\lambda) + \frac{\|w\|^2}{n_v n} \sum_{j=1}^{m} |(y_j, e_1)|^2 \delta(\lambda - \theta_j)$.
    $k \leftarrow k + 1$.
**end while**
```

The method for approximating the density of states for a generalized eigensystem, as presented is Algorithm 11, poses a major issue in that we have to perform a Cholesky factorization or square root factorization of $B$. For two and three dimensional problems this is a significant bottleneck. In order to overcome this issue, we follow [67] and use a polynomial approximation of the operator $L^{-T}$.

Let $S$ be the unique symmetric positive definite square root factorization of $B$, i.e., $B = S^2$ with $S$ symmetric positive definite [23]. Note that with the previous notation

$S = L = L^T$. In order to apply Algorithm 11, we approximate $S^{-1}w$ using Chebyshev polynomials. In essence, this removes the costly factorization in line 2 of Algorithm 11, and replaces line 5 with an approximation to $S^{-1}w$ where $w$ is a standard normal random $n$-vector. Before diving into the technical details, we note that this amounts to approximating $\bar{g}(B)w$ where $\bar{g}(\nu) = \nu^{-1/2}$.

As with all things involving Chebyshev polynomials, the first step is a linear scaling to the interval $[-1, 1]$. Let $a$ and $b$ be the minimum and maximum eigenvalues of $B$ respectively, i.e., $a = \lambda_{\min}(B) > 0$, $b = \lambda_{\max}(B)$, and $a < b$ (strict inequality is required since if $a = b$, then $B$ is a multiple of the identity matrix, in which case this is a standard eigenvalue problem). By definition, the eigenvalues of $B$ lie in the interval $[a, b]$. Letting $c = 1/2(b + a)$ and $d = 1/2(b - a)$, we define the linear scaling $\lambda(\nu) = d^{-1}(\nu - c)$ between the intervals $\nu \in [a, b]$ and $\lambda \in [-1, 1]$. After scaling to the interval $[-1, 1]$, we approximate $\bar{g}(\nu) = g(\lambda) = (c + d\lambda)^{-1/2}$ using Chebyshev polynomials.

The degree $k$ Chebyshev expansion of $g(\lambda)$ is given by

$$g_k(\lambda) = \sum_{j=0}^{k} \mu_j t_j(\lambda), \tag{4.19}$$

where $t_j(\lambda)$ denotes the degree $j$ Chebyshev polynomial of the first kind, and the coefficients, $\mu_j$, are given by

$$\mu_j = \frac{2 - \delta_{j0}}{\pi} \int_{-1}^{1} \frac{t_j(\lambda) g(\lambda)}{\sqrt{1 - \lambda^2}} d\lambda, \qquad j = 0, 1, \ldots, k. \tag{4.20}$$

Note that the standard notation for Chebyshev polynomials of the first kind is $T_j$, however we reserve this notation for Lanczos partial tridiagonalizations. An accurate and economical method to approximate the Chebyshev coefficients is Gauss-Chebyshev quadrature. The authors in [67] recommend taking a conservative approach and using a $4k$ point Gauss-Chebyshev quadrature rule, noting that the quadrature is performed only once at the outset, and the cost is negligible relative to the total cost of approximating the density of states. The last free parameter to choose is the polynomial degree $k$, which is discussed in Section 4.3.1.

Define the matrix $\hat{B} = d^{-1}(B - cI)$, and note that the eigenvalues of $\hat{B}$ lie in the interval $[-1, 1]$. Once the Chebyshev coefficients, $\mu_j$, $j = 0, 1, \ldots, k$, have been computed, the approximation to $S^{-1}$ is given by

$$S^{-1} \approx g_k(\hat{B}) = \sum_{j=0}^{k} \mu_j t_j(\hat{B}). \tag{4.21}$$

Using (4.21), the factorization of the mass matrix $B$ in Algorithm 11 is skipped, and in line 5 we use as starting vector for the $B$-Lanczos algorithm $v = \sum_{j=0}^{k} \mu_j t_j(\hat{B})w$, where $w$ is the trial vector. The rest of the algorithm proceeds as given. Note that a matrix approximating $S^{-1}$ is never constructed as implied by (4.21). Instead, we use the Chebyshev recurrence relation to construct the approximation to $S^{-1}w$. Defining $w_j = t_j(\hat{B})w$, $j = 0, 1, \ldots, k$, the Chebyshev recurrence gives

$$w_{j+1} = 2\hat{B}w_j - w_{j-1}, \quad j = 1, 2, \ldots, k-1, \tag{4.22}$$

where $w_0 = w$ and $w_1 = \hat{B}w$. Therefore, given a trial vector $w_0 = w$, we construct $w_j$, $j = 1, \ldots, k$, using the Chebyshev recurrence (4.22), and the starting vector to be used in Algorithm 11 is

$$v = \sum_{j=0}^{k} \mu_j w_j. \tag{4.23}$$

In fact, we can form the approximation of $S^{-1}w$ more economically by only storing three vectors $w_j = t_j(\hat{B})w$ at a time. For $v$ defined as in (4.23), $v$ can be recursively updated, $v \leftarrow v + \mu_{j+1} w_{j+1}$, where $w_{j+1}$ is computed using the previous two vectors $w_{j-1}$ and $w_j$ according to (4.22). This is illustrated in Algorithm 12. Next, we give details on the choice of the polynomial degree $k$ in the Chebyshev expansion of $g(\lambda) = (c + d\lambda)^{-1/2}$.

### 4.3.1 Choosing Degree of Chebyshev Expansion

Now, we address the important question of what polynomial degree is necessary in the Chebyshev approximation of the inverse square root of $B$. While we expect a higher degree polynomial to correspond to a more accurate approximation of the matrix-vector product $S^{-1}w$, a higher degree polynomial also means more matrix vector multiplications in the formation of each starting vector for the B-Lanczos algorithm. Hence, we

**Algorithm 12** Approximation of $S^{-1}w$ using the Chebyshev Recurrence (see (4.23))

```
Initialize symmetric $\hat{B} \in \mathbb{R}^{n \times n}$ with eigenvalues in $[-1, 1]$, coefficients $\mu_j$, $j = 0, \ldots, k$, and $w \in \mathbb{R}^n$.
Set $w_0 = w$, $w_{-1} = w_1 = 0$, $v = \mu_0 w_0$, and $j = 0$.
while $j < k$ do
    $w_1 \leftarrow \hat{B} w_0$.
    if $j \neq 0$ then
        $w_1 \leftarrow 2w_1 - w_{-1}$
    end if
    $v \leftarrow v + \mu_{j+1} w_1$
    $w_{-1} \leftarrow w_0$
    $w_0 \leftarrow w_1$
    $j \leftarrow j + 1$
end while
```

do not want to use a needlessly high degree polynomial. In order to understand the degree polynomial necessary, we look at how closely the degree $k$ Chebyshev expansion, $g_k(\lambda)$, approximates $g(\lambda) = (c + d\lambda)^{-1/2}$. For this, we again turn to the theory of analytic functions in Bernstein ellipses, as we did in understanding the error in the Lanczos process for analytic functions.

We begin by looking at where the function $g(\lambda) = (c + d\lambda)^{-1/2}$ is analytic. Because we have assumed the $B$-matrix is symmetric positive definite, it has positive eigenvalues, i.e., $a = \lambda_{\min}(B) > 0$, and so the scaling constants $c = 1/2(b + a)$ and $d = 1/2(b - a)$ satisfy $c > d > 0$. Therefore, $g$ is analytic in $[-1, 1]$ and has a singularity at $\lambda = -c/d < -1$. Recall, the Bernstein ellipse corresponding to a parameter $\rho > 1$ is defined as

$$E_\rho = \{1/2(z + z^{-1}) \mid z = \rho e^{i\theta} \text{ for } \theta \in [0, 2\pi)\}. \tag{4.24}$$

The Bernstein ellipse $E_\rho$ is the ellipse in the complex plane with foci at ±1 and semi-major axis $1/2(\rho + \rho^{-1})$ and semi-minor axis $1/2(\rho - \rho^{-1})$. The following theorem, taken from [58], gives the rate of convergence for truncated Chebyshev expansions of functions which are analytic in the interval $[-1, 1]$, and analytically continuable to the interior of a Bernstein ellipse $E_\rho$.

**Theorem 13.** *Let a function $g$ analytic in $[-1, 1]$ be analytically continuable to the interior of the Bernstein ellipse $E_\rho$ for $\rho > 1$. Then, the degree $k$ Chebyshev expansion* *satisfies*

$$\|g - g_k\|_\infty \leq \frac{2M(\rho)\rho^{-k}}{\rho - 1},$$

*where* $|g| \leq M(\rho)$ *inside* $E_\rho$.

In order to apply Theorem 13 to the expansion (4.19), we need to determine the value of the ellipse parameter $\rho > 1$, and a bound, $M = M(\rho)$, for the function $g(\lambda) = (c + d\lambda)^{-1/2}$ in the interior of $E_\rho$. With knowledge of the singularity at $\lambda = -c/d$, we choose a Bernstein ellipse with semi-major axis, $1/2(\rho + \rho^{-1})$, satisfying

$$\frac{\rho + \rho^{-1}}{2} < \frac{c}{d}. \tag{4.25}$$

Solving the quadratic equation resulting from (4.25), we take any $\rho$ satisfying

$$1 < \rho < \bar{\rho}, \qquad \bar{\rho} = \frac{c}{d} + \sqrt{\left(\frac{c}{d}\right)^2 - 1}. \tag{4.26}$$

Notice that we can write $c/d = (\kappa + 1)/(\kappa - 1)$, where $\kappa = b/a$ is the spectral condition number of $B$. Rewriting the upper bound $\bar{\rho}$ in terms of the spectral condition number gives

$$\bar{\rho} = \frac{\kappa + 1}{\kappa - 1} + \sqrt{\left(\frac{\kappa + 1}{\kappa - 1}\right)^2 - 1}. \tag{4.27}$$

The dependence of $\bar{\rho}$ on $\kappa$, as well as plots of several Bernstein ellipses corresponding to $\bar{\rho} = \bar{\rho}(\kappa)$ for different values of $\kappa$ are shown in Figure 4.1. We see that the better conditioned the matrix $B$, i.e., the closer $\kappa$ is to unity, the larger the Bernstein ellipse in which $g(\lambda)$ is analytic, and hence the faster the Chebyshev expansion of $g$ converges. This is the main reason the diagonal scaling (4.15) is important.

Next, we determine the maximum value of $g(\lambda) = (c + d\lambda)^{-1/2}$ inside the Bernstein ellipse $E_\rho$. The following lemma is taken from [67].

**Lemma 11.** *If* $c > d > 0$ *and* $\rho$ *satisfies* (4.26)*, the maximum modulus of* $g(\lambda) = (c + d\lambda)^{-1/2}$ *inside the Bernstein ellipse* $E_\rho$ *is*

$$M(\rho) = \left(c - d^r\right)^{-1/2},$$

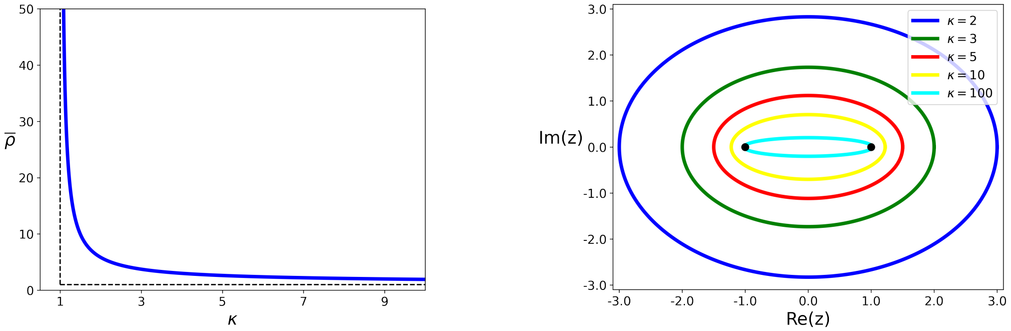


Figure 4.1: Bernstein ellipse parameter $\bar{\rho}$ as a function of the spectral condition number $\kappa$ (left). Bernstein ellipses, $E_{\bar{\rho}}$, for various spectral condition numbers (right).

*where, $r = 1/2(\rho + \rho^{-1})$, is the semi-major axis.*

Combining Lemma 11 and (4.26) with Theorem 13, we bound the uniform error in the Chebyshev expansion of $g$ in the following theorem.

**Theorem 14.** *Let $g(\lambda) = (c + d\lambda)^{-1/2}$ for $\lambda \in [-1, 1]$ with $c > d > 0$. The degree $k$ Chebyshev expansion of $g$ satisfies*

$$\|g - g_k\|_\infty \leq \frac{2\rho^{-k}}{(\rho - 1)\sqrt{c - d^r}},$$

*where $r = 1/2(\rho + \rho^{-1})$, and $\rho$ is any real number satisfying*

$$1 < \rho < \frac{c}{d} + \sqrt{\left(\frac{c}{d}\right)^2 - 1}.$$

With the uniform error in the Chebyshev expansion bounded by computable quantities in Theorem 14, we now outline our procedure for determining what degree Chebyshev expansion to use. For the first step, we compute, to a high degree of accuracy, the minimum and maximum eigenvalues of $B$, $a = \lambda_{\min}(B)$ and $b = \lambda_{\min}(B)$, in order to scale the matrix to have eigenvalues in the interval $[-1, 1]$. Note that scaling to the interval $[-1, 1]$ only requires lower and upper bounds on $\lambda_{\min}(B)$ and $\lambda_{\max}(B)$ respectively. However, since the smallest and largest eigenvalues of a matrix are typically the easiest to acquire, and these values allow us to compute the rate of decay of the error from Theorem 14, we find it advantageous to perform this computation. Furthermore, the cost of computing the largest and smallest eigenvalues of $B$ is negligible in comparison to the overall cost of approximating the density of states.

Next, from $a$ and $b$ we determine the scaling constants $c = 1/2(b + a)$ and $d = 1/2(b - a)$, which allow us to scale from the interval $[a, b]$ to $[-1, 1]$, and set

$$\bar{\rho} = \frac{c}{d} + \sqrt{\left(\frac{c}{d}\right)^2 - 1}. \tag{4.28}$$

From (4.26) we know that $g$ is analytic in the Bernstein ellipse $E_\rho$ for any $1 < \rho < \bar{\rho}$. Accordingly, we choose $k$ to be the smallest integer satisfying

$$\frac{2\bar{\rho}^{-k}}{(\bar{\rho} - 1)\sqrt{c - d\bar{r}}} < \tau, \tag{4.29}$$

where $\bar{r} = 1/2(\bar{\rho} + \bar{\rho}^{-1})$, and $\tau$ is a chosen tolerance. The proper value of $k$ can be found using any root-finding method, e.g., Newton's method or bisection, to determine the real number which makes (4.29) an equality, and then take $k$ to be the smallest integer larger than this root. To give an idea of standard values of $\bar{\rho}$ and $k$, if $a = 0.50$ and $b = 1.48$ (these are actual values from a one dimensional finite element mass matrix corresponding to cubic Lagrange polynomials on a uniform mesh after the diagonal scaling (4.15) has been performed), then $\bar{\rho} = 3.78$. In this case, choosing the tolerance $\tau$ as small as $10^{-16}$ results in $k = 28$, a very manageable request.

## 4.4 Joint Density of States

Let $A, A' \in \mathbb{R}^{n \times n}$ be the symmetric matrices with eigenpairs as in (4.1). In this section we discuss approximating the joint density of states, $J(\lambda)$, defined as

$$J(\lambda) = \frac{1}{n^2} \sum_{i,j=1}^{n} \delta\big(\lambda - (\lambda_i + \lambda'_j)\big). \tag{4.30}$$

The joint density of states is a joint spectral quantity with uniform coefficients $1/n^2$, and is the density of states corresponding to a matrix which has $n^2$ eigenvalues $\lambda_i + \lambda'_j$, $i, j = 1, \ldots, n$. We describe this matrix momentarily. The joint density of states has many uses in solid state physics and semiconductor design [57, 68, 37, 38].

As a first step towards approximating (4.30), we relate the joint density of states to the density of states of an $n^2 \times n^2$ matrix. In order to accomplish this, we recall the definition of the Kronecker product.

**Definition 1.** *For $C \in \mathbb{R}^{m\times n}$ and $D \in \mathbb{R}^{p\times q}$, the Kronecker product, $C \otimes D$, is the $mp \times nq$ block matrix*

$$C \otimes D = \begin{pmatrix} c_{11}D & \cdots & c_{1n}D \\ \vdots & \ddots & \vdots \\ c_{m1}D & \cdots & c_{mn}D \end{pmatrix}.$$

In this chapter we use many properties of Kronecker products, which we record now. We do not state these results in their full generality, but rather in the manner in which we intend to use them. These properties can be found in linear algebra textbooks, see, for example, [18].

**Lemma 12** (Properties of Kronecker Products)**.** *Let $C, D, \in \mathbb{R}^{n\times n}$ and $U, V \in \mathbb{R}^{n\times m}$. Then,*

1. $(C \otimes D)^T = C^T \otimes D^T$,
2. $(C \otimes D)(U \otimes V) = (CU) \otimes (DV)$.

It is straightforward to show that the eigenvalues of the Kronecker product of two $n \times n$ matrices are the $n^2$ products of the eigenvalues of the two matrices. For the joint density of states, we want a similar result involving sums of eigenvalues, rather than products. The following linear operator accomplishes exactly this.

**Definition 2.** *For $C \in \mathbb{R}^{m\times m}$ and $D \in \mathbb{R}^{n\times n}$, the Kronecker sum, $C \oplus D$, is the $mn \times mn$ block matrix*

$$C \oplus D = C \otimes I_n + I_m \otimes D, \tag{4.31}$$

*where $I_k$ is the $k \times k$ identity matrix. When $m = n$, we drop the subscript on the identity matrix and write $C \oplus D = C \otimes I + I \otimes D$, the dimension of the identity matrix being clear from context.*

In some references, the Kronecker sum is defined in the opposite manner as $I_n \otimes C + D \otimes I_m$, e.g., [22]. This is not equivalent to our definition since, in general, $C \oplus D \neq D \oplus C$.

The alternate definition is due to the structure of the Sylvester equation, a standard application of the Kronecker sum. However, the properties of the Kronecker sum we wish to exploit do not change with the definition (4.31), and, for our purposes, (4.31) is more satisfactory.

We now state a useful result about the eigenpairs of a Kronecker sum of two matrices, which can be found in [22].

**Theorem 15.** *For $A$ and $A'$ be as in* (4.1)*, the eigenvalues of $A \oplus A'$ are $\lambda_i + \lambda'_j$, with corresponding eigenvectors $x_i \otimes x'_j$, $i, j = 1, \ldots, n$.*

*Proof.* Let $Ax = \lambda x$ and $A'x' = \lambda' x'$ for $\lambda, \lambda' \in \mathbb{R}$ and nonzero $x, x' \in \mathbb{R}^n$. Then,

$$\begin{aligned}(A \oplus A')(x \otimes x') &= (A \otimes I + I \otimes A')(x \otimes x'),\\ &= (Ax \otimes x' + x \otimes A'x'),\\ &= (\lambda x \otimes x' + x \otimes \lambda' x'),\\ &= (\lambda + \lambda')x \otimes x',\end{aligned}$$

where in the second equality we used property (2) in Lemma 12. □

Theorem 15 tells us that the eigenvalues of $A \oplus A'$ are the sum of all combinations of eigenvalues of $A$ and $A'$. Hence, the joint density of states (4.30) is equivalent to the density of states of $A \oplus A'$. Therefore, we can think about the joint density of states in two ways. On one hand, it is the joint spectral quantity corresponding to the matrices $A$ and $A'$ with uniform coefficients $1/n^2$. On the other hand, it is the spectral quantity corresponding to the matrix $A \oplus A'$ with uniform coefficients $1/n^2$.

Next, we look at applying the techniques from Section 4.2 to approximate the density of states for the matrix $A \oplus A'$. We know that by choosing a standard normal random $n^2$-vector $w$, the spectral function $1/n^2 s(\lambda; A \oplus A', w)$ is equal to the joint density of states in expectation, as in (4.11). By stochastically averaging the results of the Lanczos process applied to $A \oplus A'$, we can approximate the joint density of states using Algorithm 10. The main issue with this method is for $n \gg 1$, storing $A \oplus A'$, or performing matrix operations with $A \oplus A'$ is not feasible. Instead, we wish to approximate the joint density of by performing matrix operations with $A$ and $A'$ individually. Next, we discuss two different methods for approximating the joint density of states, both of which perform the Lanczos process on $A$ and $A'$ individually, rather than on the matrix $A \oplus A'$.

### 4.4.1 Method I

For the method I Lanczos approximation to the joint density of states, we implicitly perform the Lanczos algorithm on $A \oplus A'$ by performing the Lanczos algorithm on the individual matrices.

At first glance, one might hope that by performing the Lanczos algorithm on $A$ and $A'$ separately, we may then use the Kronecker sum to realize a Lanczos partial tridiagonalization of $A \oplus A'$. To see why this fails, let $v, v' \in \mathbb{R}^n$ be the starting vectors for an $m$-step Lanczos algorithm applied to $A$ and $A'$ respectively, i.e., we have

$$\begin{aligned} AV_m &= V_m T_m + \beta_m v_{m+1} e_m^T, \\ A'V'_m &= V'_m T'_m + \beta'_m v'_{m+1} e_m^T, \end{aligned} \tag{4.32}$$

where $T_m, T'_m \in \mathbb{R}^{m \times m}$ are symmetric and tridiagonal, the columns of $V_m, V'_m \in \mathbb{R}^{n \times m}$ are an orthonormal basis for $\mathcal{K}_m(A, v)$ and $\mathcal{K}_m(A', v')$ respectively, and $V_m^T v_{m+1} = 0 = {V'_m}^T v'_{m+1}$. Let the entries of $T_m$ and $T'_m$ be given by

$$\begin{aligned} T_m = V_m^T A V_m &= \begin{pmatrix} \alpha_1 & \beta_1 & & & \\ \beta_1 & \alpha_2 & & & \\ & \ddots & \ddots & \ddots & \\ & & & \alpha_{m-1} & \beta_{m-1} \\ & & & \beta_{m-1} & \alpha_m \end{pmatrix}, \\ T'_m = {V'_m}^T A' V'_m &= \begin{pmatrix} \alpha'_1 & \beta'_1 & & & \\ \beta'_1 & \alpha'_2 & & & \\ & \ddots & \ddots & \ddots & \\ & & & \alpha'_{m-1} & \beta'_{m-1} \\ & & & \beta'_{m-1} & \alpha'_m \end{pmatrix}. \end{aligned} \tag{4.33}$$

Using the definition of the Kronecker sum and the Lanczos relations (4.32), we see that

$$\begin{aligned}(A \oplus A')(V_m \otimes V'_m) &= AV_m \otimes V'_m + V_m \otimes A'V'_m, \\ &= (V_m T_m + \beta_m v_{m+1} e_m^T) \otimes V'_m + \\ &\qquad\qquad V_m \otimes (V'_m T'_m + \beta'_m v'_{m+1} e_m^T), \\ &= (V_m T_m \otimes V'_m + V_m \otimes V'_m T'_m) + R, \\ &= (V_m \otimes V'_m)(T_m \oplus T'_m) + R,\end{aligned} \tag{4.34}$$

where $R = (V_m \otimes \beta'_m v'_{m+1} e_m^T + \beta_m v_{m+1} e_m^T \otimes V'_m)$. The main takeaway is that (4.34) does NOT constitute a Lanczos algorithm applied to $A \oplus A'$ with starting vector $v \otimes v'$. This follows because the matrix $T_m \oplus T'_m$ is not tridiagonal, but block tridiagonal. Therefore, in order to perform the Lanczos process on $A \oplus A'$, we must look elsewhere.

While the naive first attempt to perform the Lanczos algorithm on $A \oplus A'$ failed, the idea of performing the Lanczos algorithm on the operators $A$ and $A'$ independently in order to partially tridiagonalize $A \oplus A'$ is not without merit. Next, we show that the spectral function corresponding to the matrix $A \oplus A'$ and a rank one starting vector equals the joint density of states in expected value. Consider the following lemma.

**Lemma 13.** *If $w$ and $w'$ are independent standard normal random $n$-vectors, then $\mathbb{E}[(w \otimes w')(w \otimes w')^T] = I$.*

*Proof.* Using property (1) and (2) in Lemma 12 we have

$$\begin{aligned}\mathbb{E}\big[(w \otimes w')(w \otimes w')^T\big] &= \mathbb{E}\big[(w \otimes w')(w^T \otimes w'^T)\big] = \mathbb{E}\big[(ww^T) \otimes (w'w'^T)\big] \\ &= \mathbb{E}[ww^T] \otimes \mathbb{E}[w'w'^T] = I \otimes I = I.\end{aligned}$$

□

As mentioned in Section 4.2, we choose a trial vector $w \sim \mathcal{N}(0,1)$ because $\mathbb{E}[w_i w_j] = \delta_{ij}$, or equivalently $\mathbb{E}[ww^T] = I$. This allows us to use Corollary 2 to equate the density of states and the expectation of a spectral function. Lemma 13 tells us that the same property, $\mathbb{E}[(w \otimes w')(w \otimes w')^T] = I$, is satisfied when both random vectors, $w$ and $w'$, have i.i.d. entries in $\mathcal{N}(0,1)$. Meaning, the joint density of states is equal in expectation to the spectral function corresponding to the matrix $A \oplus A'$ and vector $w \otimes w'$ (with a prefactor of $1/n^2$). That is,

$$J(\lambda) = \frac{1}{n^2}\mathbb{E}\Big[s(\lambda; A \oplus A', w \otimes w')\Big], \tag{4.35}$$

where $w, w' \sim \mathcal{N}(0,1)$. Next, we show that for the special case of a rank one starting vector, the Lanczos algorithm applied to $A \oplus A'$, can be deduced from the Lanczos algorithm applied to $A$ and $A'$ individually.

The first crucial piece of the puzzle is to notice that the Krylov space $\mathcal{K}_m(A \oplus A', v \otimes v')$ is spanned by the columns of $V_m \otimes V_m'$, where $V_m$ and $V_m'$ are as in (4.32). This is detailed in the following theorem.

**Theorem 16** (Remark 3.3 in [27] for $d = 2$)**.** *Let $A, A' \in \mathbb{R}^{n\times n}$ and $v, v' \in \mathbb{R}^{n\times n}$. Then,* $\mathcal{K}_m(A \oplus A', v \otimes v') \subset \mathcal{K}_m^{\otimes}(A, v; A', v') := \operatorname{span}\{u \otimes u' \mid u \in \mathcal{K}_m(A, v) \text{ and } u' \in \mathcal{K}_m(A', v')\}$.

*Proof.* We proceed with induction. For $m = 1$ we have equality. Next, assume $\mathcal{K}_m(A \oplus A', v \otimes v') \subset \mathcal{K}_m^{\otimes}(A, v; A', v')$. Any $u \in \mathcal{K}_{m+1}(A \oplus A', v \otimes v')$ can be expressed as

$$u = c\big(A \oplus A'\big)^m (v \otimes v') + u_m,$$

for some constant $c$ and $u_m \in \mathcal{K}_m(A \oplus A', v \otimes v')$. By the inductive hypothesis we only need to focus on the $(A \oplus A')^m(v \otimes v')$ term. Notice that by the binomial theorem and property (2) in Lemma 12

$$(A \oplus A')^m = \sum_{k=0}^{m} \binom{m}{k} A^{m-k} \otimes A'^k,$$

and therefore

$$(A \oplus A')^m(v \otimes v') = \sum_{k=0}^{m} \binom{m}{k} A^{m-k}v \otimes A'^k v',$$

which is plainly an element of $\mathcal{K}_{m+1}^{\otimes}(A, v; A', v')$. □

Next, we show how Theorem 16 allows us to apply the Lanczos algorithm to $A \oplus A'$ with starting vector $v \otimes v'$ implicitly. Suppose we perform an $m$-step Lanczos iteration on $A \oplus A'$ with starting vector $u = v \otimes v'$. Assuming $m$ is less than the grade of $v \otimes v'$

with respect to $A \oplus A'$, we have

$$(A \oplus A')U_m = U_m T_m^{\oplus} + \beta_m^{\oplus} u_{m+1} e_m^T, \tag{4.36}$$

where the columns of $U_m = [u_1, \ldots, u_m] \in \mathbb{R}^{n^2 \times m}$ form an orthonormal basis for $\mathcal{K}_m(A \oplus A', v \otimes v')$, and the entries of the symmetric tridiagonal matrix $T_m^{\oplus} \in \mathbb{R}^{m \times m}$ are

$$T_m^{\oplus} = U_m^T (A \oplus A') U_m = \begin{pmatrix} \alpha_1^{\oplus} & \beta_1^{\oplus} & & & \\ \beta_1^{\oplus} & \alpha_2^{\oplus} & & & \\ & \ddots & \ddots & \ddots & \\ & & & \alpha_{m-1}^{\oplus} & \beta_{m-1}^{\oplus} \\ & & & \beta_{m-1}^{\oplus} & \alpha_m^{\oplus} \end{pmatrix}. \tag{4.37}$$

We now use Theorem 16 to relate the Lanczos vectors of $A \oplus A'$, $U_m$, to the Kronecker product $V_m \otimes V_m'$, where $V_m$ and $V_m'$ are as in (4.32). In so doing, we determine a method to compute $T_m^{\oplus}$, which is the main ingredient for the Lanczos process. We then are able to use the Lanczos process to approximate $1/n^2 s(\lambda; A \oplus A', w \otimes w')$ (which equals the joint density of states in expected value, see (4.35)), where $w$ and $w'$ are standard normal random $n$-vectors. All this without ever needing to form the matrix $A \oplus A'$!

Theorem 16 tells us the columns of $V_m \otimes V_m'$ span $\mathcal{K}_m(A \oplus A, v \otimes v')$. Hence, the columns of $U_m$, an orthonormal basis of $\mathcal{K}_m(A \oplus A', v \otimes v')$, can be expressed as

$$u_k = \sum_{i,j=1}^{k} \gamma_{ij}^k v_i \otimes v_j' \quad \text{for} \quad k = 1, \ldots, m, \tag{4.38}$$

for some coefficient matrix $\gamma^k \in \mathbb{R}^{k \times k}$, where $v_j$ and $v_j'$ are the $j$-th columns of $V_m$ and $V_m'$ respectively. We refer to the basis, $\{v_i \otimes v_j'\}$, $i, j = 1, \ldots, m$, of $\mathcal{K}_m^{\otimes}(A, v; A', v')$ as the tensorial basis (here we assume the grades of the vectors $v$ and $v'$ are such that the term basis is justified). We assume the tensorial basis vectors are orthonormal, which follows automatically if the columns of $V_m$ and $V_m'$ are orthonormal, i.e., $V_m^T V_m = I = {V_m'}^T V_m'$. Given that we know $u_1 = v_1 \otimes v_1'$, i.e., $\gamma_{11}^1 = 1$, we now proceed to use (4.36) and (4.38) to determine the elements of the tridiagonal matrix $T_m^{\oplus}$ from $T_m$ and $T_m'$.

The first step in the Lanczos algorithm involves forming the matrix vector product $(A \oplus A')u_k$. Meaning, given $u_k$ in the tensorial basis as in (4.38), i.e., given the coefficient matrix $\gamma^k \in \mathbb{R}^{k\times k}$, we must determine $(A \oplus A')u_k$ in the tensorial basis. This can easily be done using the three-term recurrence formulas for the Lanczos algorithms on $A$ and $A'$. Define $\eta^{k+1} \in \mathbb{R}^{(k+1)\times(k+1)}$ to be the coefficients of $(A \oplus A')u_k$ in the tensorial basis, i.e.,

$$(A \oplus A')u_k = \sum_{i,j=1}^{k+1} \eta_{ij}^{k+1} v_i \otimes v'_j. \tag{4.39}$$

Inserting the expansion (4.38) into (4.39), and using the three-term Lanczos recurrence gives

$$\begin{aligned} \sum_{i,j=1}^{k+1} \eta_{ij}^{k+1} v_i \otimes v'_j &= \sum_{i,j=1}^{k} \gamma_{ij}^k (A \oplus A')(v_i \otimes v'_j), \\ &= \sum_{i,j=1}^{k} \gamma_{ij}^k (Av_i \otimes v'_j + v_i \otimes A'v'_j), \\ &= \sum_{i,j=1}^{k} \gamma_{ij}^k \Big[ (\beta_{i-1}v_{i-1} + \alpha_i v_i + \beta_i v_{i+1}) \otimes v'_j, \\ &\qquad\qquad + v_i \otimes (\beta'_{j-1}v'_{j-1} + \alpha'_j v'_j + \beta'_j v'_{j+1}) \Big]. \end{aligned} \tag{4.40}$$

Re-indexing individual terms in the final summation of (4.40) and adopting the convention that $\beta_0 = \beta'_0 = 0$ and $\gamma_{ij}^k = 0$ for $i, j \notin \{1, \ldots, k\}$, we find that the entries of the matrix $\eta^{k+1}$ are

$$\eta_{ij}^{k+1} = \gamma_{ij}^k(\alpha_i + \alpha'_j) + \gamma_{i+1\,j}^k \beta_i + \gamma_{i\,j+1}^k \beta'_j + \gamma_{i-1\,j}^k \beta_{i-1} + \gamma_{i\,j-1}^k \beta'_{j-1}, \tag{4.41}$$

for $i, j = 1, \ldots, k+1$.

The next step in the Lanczos algorithm creates the coefficient $\alpha_k^\oplus = (u_k, (A \oplus A')u_k)$. Due to the orthonormality of the tensorial basis vectors $v_i \otimes v'_j$, we can now easily compute $\alpha_k^\oplus$ as

$$\alpha_k^\oplus = \sum_{i,j=1}^{k} \gamma_{ij}^k \eta_{ij}^{k+1} = (\gamma^k, \eta^{k+1})_F, \tag{4.42}$$

where $(\,\cdot\,,\,\cdot\,)_F$ is the Frobenius inner product, and again, we have made use of the convention that $\gamma^k_{ij} = 0$ for $i, j \notin \{1, \dots, k\}$.

Define $\tilde{u}_{k+1}$ as

$$\tilde{u}_{k+1} = (A \oplus A')u_k - \alpha^{\oplus}_k u_k - \beta^{\oplus}_{k-1} u_{k-1}, \tag{4.43}$$

and let the coefficients of $\tilde{u}_{k+1}$ in the tensorial basis be $\tilde{\gamma}^{k+1} \in \mathbb{R}^{(k+1)\times(k+1)}$. From (4.43), the entries of $\tilde{\gamma}^{k+1}$ satisfy

$$\tilde{\gamma}^{k+1}_{ij} = \eta^{k+1}_{ij} - \alpha^{\oplus}_k \gamma^k_{ij} - \beta^{\oplus}_{k-1} \gamma^{k-1}_{ij} \qquad i, j = 1, \dots, k+1. \tag{4.44}$$

Using again the convention that $\gamma^k_{ij} = 0$ for $i, j \notin \{1, \dots, k\}$, we can rewrite (4.44) in matrix form as

$$\begin{aligned}\tilde{\gamma}^{k+1} &= \eta^{k+1} - \alpha^{\oplus}_k \gamma^k - \beta^{\oplus}_{k-1} \gamma^{k-1} \\ &= \eta^{k+1} - \alpha^{\oplus}_k \begin{pmatrix} & & 0 \\ & \gamma^k & \vdots \\ 0 & \dots & 0 \end{pmatrix} - \beta^{\oplus}_{k-1} \begin{pmatrix} & & 0 & 0 \\ & \gamma^{k-1} & \vdots & \vdots \\ 0 & \dots & 0 & 0 \\ 0 & \dots & 0 & 0 \end{pmatrix}.\end{aligned} \tag{4.45}$$

Finally,

$$\beta^{\oplus}_k = \sqrt{(\tilde{u}_{k+1}, \tilde{u}_{k+1})} = \|\tilde{\gamma}^{k+1}\|_F \quad \text{and} \quad \gamma^{k+1} = \frac{1}{\beta^{\oplus}_k}\tilde{\gamma}^{k+1}, \tag{4.46}$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

A concise overview of the partial tridiagonalization of $(A \oplus A')$, with rank one starting vector $v \otimes v'$ is given in Algorithm 13. Full orthogonalization is used in the Lanczos algorithm on $A$ and $A'$ to ensure the tensorial basis is orthonormal to working precision. Using Algorithm 13, we are now able to perform the partial tridiagonalization of $A \oplus A'$, without explicitly forming the matrix $A \oplus A' \in \mathbb{R}^{n^2 \times n^2}$. This, along with Lemma 13, allows us to use the methods in Section 4.2 to approximate the density of states for $A \oplus A'$, which is equivalent to the joint density of states.

Before utilizing Algorithm 13 to approximate the joint density of states, we comment on the algorithmic complexity. For this discussion we assume the matrices $A$ and $A'$ are sparse, and matrix vector products require $\mathcal{O}(n)$ floating point operations (flops). More generally, for a matrix with at most $n_z$ nonzeros per row, matrix vector multiplication

**Algorithm 13** Lanczos Algorithm of Kronecker Sum (full orthogonalization)

```
Initialize $v_1 = v/\|v\|$, $v'_1 = v'/\|v'\|$, $\beta_0 = \beta'_0 = 0$, $v_0 = v'_0 = 0$, $\gamma^1_{11} = 1$.
for $k = 1, \ldots, m$ do
    $\tilde{v} = Av_k - \beta_{k-1}v_{k-1}$
    $\tilde{v}' = A'v'_k - \beta'_{k-1}v'_{k-1}$
    $\alpha_k = (\tilde{v}, v_k)$
    $\alpha'_k = (\tilde{v}', v'_k)$
    for $i = 1, \ldots, k$ do
        $\tilde{v} \leftarrow \tilde{v} - (\tilde{v}, v_i)v_i$
        $\tilde{v}' \leftarrow \tilde{v}' - (\tilde{v}', v'_i)v'_i$
    end for
    $\beta_k = \|\tilde{v}\|$
    $\beta'_k = \|\tilde{v}'\|$
    if $\beta_k = 0$ or $\beta'_k = 0$ then stop
    else
        $v_{k+1} = \frac{\tilde{v}}{\beta_k}$
        $v'_{k+1} = \frac{\tilde{v}'}{\beta'_k}$
    end if
    for $i, j = 1, \ldots, k+1$ do
        $\eta^{k+1}_{ij} = \gamma^k_{ij}(\alpha_i + \alpha'_j) + \gamma^k_{i+1\,j}\beta_i + \gamma^k_{i\,j+1}\beta'_j + \gamma^k_{i-1\,j}\beta_{i-1} + \gamma^k_{i\,j-1}\beta'_{j-1}$
    end for
    $\alpha^{\oplus}_k = (\gamma^k, \eta^{k+1})_F$
    $\tilde{\gamma}^{k+1} = \eta^{k+1} - \alpha^{\oplus}_k\gamma^k - \beta^{\oplus}_{k-1}\gamma^{k-1}$
    $\beta^{\oplus}_k = \|\tilde{\gamma}^{k+1}\|_F$
    $\gamma^{k+1} = \frac{1}{\beta^{\oplus}_k}\tilde{\gamma}^{k+1}$
end for
```

is $\mathcal{O}(n_z n)$. The standard $m$-step Lanczos algorithm in exact arithmetic for the matrix $A$ and vector $v$ has $\mathcal{O}(mn)$ complexity. This follows from all steps, in each of the $m$ iterations, requiring $\mathcal{O}(n)$ flops. Therefore, due to the Kronecker sum of $A$ and $A'$ being an order $n^2$ matrix, applying the standard $m$-step Lanczos algorithm to $A \oplus A'$ with starting vector $v \otimes v'$ is $\mathcal{O}(mn^2)$ complexity. On the other hand, Algorithm 13 performs the Lanczos algorithm on $A \oplus A'$ with rank one starting vector in $\mathcal{O}(mn + m^2)$ operations. Assuming $m \ll n$, Algorithm 13 has complexity $\mathcal{O}(mn)$ which is no more than the standard $m$-step Lanczos algorithm for an $n \times n$ matrix!

Before approximating the joint density of states with several trial vectors, we review Hutchinson's method for the density of states of $A \oplus A'$ with one trial vector. Specifically, letting $w, w' \sim \mathcal{N}(0, 1)$, Lemma 13 tells us that $\mathbb{E}[(w \otimes w')(w \otimes w')^T] = I$, and so we may apply Lemma 10 with trial vector $w \otimes w'$. The spectral function corresponding to $A \oplus A'$ and vector $w \otimes w'$, $s(\lambda; A \oplus A', w \otimes w')$, is proportional to the joint density of states in expectation, i.e., $\langle J, f \rangle = 1/n^2 \mathbb{E}[\langle s(\lambda; A \oplus A', w \otimes w'), f \rangle]$ for all test functions $f$ (see (4.35)). With the aid of Algorithm 13, we can approximate $s(\lambda; A \oplus A', w \otimes w')$ using the Lanczos process. By forming the Lanczos partial tridiagonalization $T_m^{\oplus} \in \mathbb{R}^{m \times m}$ from (4.37), and denoting the eigenvalues of $T_m^{\oplus}$ as $\theta_j$, with corresponding normalized eigenvector $y_j$, $j = 1, \ldots, m$, an approximation to the spectral function for $A \oplus A'$ and $w \otimes w'$ is given by

$$s(\lambda; A \oplus A', w \otimes w') \approx \|w\|^2 \|w'\|^2 \sum_{j=1}^{m} |(y_j, e_1)|^2 \delta(\lambda - \theta_j). \tag{4.47}$$

The right hand side of (4.47) is an approximation of the spectral function in the sense that the first $2m - 1$ moments of $\|w\|^2 \|w'\|^2 \sum_{j=1}^{m} |(y_j, e_1)|^2 \delta(\lambda - \theta_j)$ match those of $s(\lambda; A \oplus A', w \otimes w')$.

As in the case of the density of states, we approximate the joint density of states by averaging the results of many Lanczos processes with independent starting vectors. Using Algorithm 13 with trial vectors, $w^{(k)} \otimes w'^{(k)}$, $k = 1, \ldots, n_v$, results in $n_v$ order $m$ Lanczos partial tridiagonalizations of $A \oplus A'$. Denoting, respectively, the eigenvalues and corresponding normalized eigenvectors of the partial tridiagonalizations as $\theta_j^{(k)}$ and $y_j^{(k)}$, $j = 1, \ldots, m$, and $k = 1, \ldots, n_v$, the method I approximation to the joint density of states is

$$\tilde{J}(\lambda) = \frac{1}{n_v n^2} \sum_{k=1}^{n_v} \sum_{j=1}^{m} \left\| w^{(k)} \right\|^2 \left\| w'^{(k)} \right\|^2 |(y_j^{(k)}, e_1)|^2 \delta(\lambda - \theta_j^{(k)}). \tag{4.48}$$

The full algorithm for the method I Lanczos approximation to the joint density of states is given in Algorithm 14.

We remark that the procedure given in this section for performing the Lanczos algorithm on a Kronecker sum has not, to the best of the authors knowledge, appeared in the literature. In this thesis, we use the Lanczos algorithm for its relation to the spectral function and Gauss quadrature. However, the Lanczos algorithm has many uses, and this new method of performing the Lanczos algorithm on a Kronecker sum with a rank one starting vector, may be useful in other areas. For example, when approximating the solution of the Poisson equation in two dimensions with the finite element or finite difference methods, if the source term is separable, then in many cases the linear system is of the form $(A_1 \oplus A_2)x = b_1 \otimes b_2$. When using Krylov methods, such as the Lanczos algorithm, if $x_0$ is the initial guess, the residual vector $r = b_1 \otimes b_2 - (A_1 \oplus A_2)x_0$ is typically chosen as the starting vector. Choosing a zero initial guess gives $b_1 \otimes b_2$ as the starting vector, and hence Algorithm 13 can be used to perform the Lanczos algorithm on $A_1 \oplus A_2$ without needing to explicitly form the Kronecker sum.

**Algorithm 14** Lanczos Approximation of the Joint Density of States (method I)

```
Initialize $m$, $n_v$, and set $k = 0$ and $\tilde{J}(\lambda) = 0$.
while $k < n_v$ do
    Draw trial vectors $w, w' \sim \mathcal{N}(0, 1)$.
    Partially tridiagonalize $A \oplus A'$ with starting vector $w \otimes w'$ (Algorithm 13) to get
    $T_m^{\oplus} \in \mathbb{R}^{m \times m}$.
    Compute eigenpairs $T_m^{\oplus} y_j = \theta_j y_j$, $y_i^T y_j = \delta_{ij}$, $i, j = 1, \ldots, m$.
    $\tilde{J}(\lambda) \leftarrow \tilde{J}(\lambda) + \frac{\|w\|^2 \|w'\|^2}{n_v n^2} \sum_{j=1}^{m} |(y_j, e_1)|^2 \delta(\lambda - \theta_j)$.
    $k \leftarrow k + 1$.
end while
```

### 4.4.2 Method II

In this section we give a second method to approximate the joint density of states. If $T_m$ and $T'_m$ are the Lanczos partial tridiagonalizations of $A$ and $A'$ respectively, we showed in the previous section that the Kronecker sum, $T_m \oplus T'_m$, is not a partial tridiagonalization of $A \oplus A'$. This follows because the Kronecker sum of two tridiagonal matrices is not tridiagonal, but block tridiagonal. Therefore, $T_m \oplus T'_m$ can not be used for the Lanczos process, with its desired moment matching property. However, we show in this section that we can indeed use the eigenpairs of $T_m \oplus T'_m$ to determine the nodes and weights for a quadrature rule approximating a spectral function corresponding to the matrix $A \oplus A'$, albeit not Gaussian quadrature.

Before introducing method II, we define the convolution of measures on the real line. Let $\mu$ and $\mu'$ be measures with compact support (meaning $d\mu$ and $d\mu'$ are zero outside of some finite interval). Then, the measure $\mu * \mu'$ is defined as

$$\int_{-\infty}^{+\infty} f(\lambda) d(\mu * \mu')(\lambda) := \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(\lambda + \lambda') d\mu(\lambda) d\mu'(\lambda'), \tag{4.49}$$

for any test function $f$, where the integrals in (4.49) are Riemann–Stieltjes integrals [47]. We call $\mu * \mu'$ the convolution of the measures $\mu$ and $\mu'$. Recall, to each measure there corresponds an element of the dual space of continuous functions, i.e., to the measure $\mu$ there exists $\eta$ such that $\langle \eta, f \rangle = \int f d\mu$ for all test function $f$. This correspondence is unique assuming some normalization conventions are satisfied. For the action of $\eta$ on a test function $f$, we write $\langle \eta, f \rangle = \int \eta(\lambda) f(\lambda) d\lambda$. Let $\eta$ and $\eta'$ be the elements of the dual space of continuous functions corresponding to $\mu$ and $\mu'$ respectively. Then, $\mu * \mu'$ corresponds to an element of the dual space of continuous functions, denoted $\eta * \eta'$, which satisfies

$$\int_{-\infty}^{+\infty} \big(\eta * \eta'\big)(\lambda) f(\lambda) d\lambda = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(\lambda + \lambda') d\mu(\lambda) d\mu'(\lambda'). \tag{4.50}$$

It is straightforward to show that condition (4.50) is satisfied by

$$(\eta * \eta')(\lambda) = \int_{-\infty}^{+\infty} \eta(\theta) \eta'(\lambda - \theta) d\theta. \tag{4.51}$$

With the convolution of measures defined, we now represent the joint density of states as a convolution. If we write $\phi(\lambda)$ as the density of state of $A$, and $\phi'(\lambda)$ as the density of states of $A'$, the joint density of states is the convolution of $\phi$ and $\phi'$, i.e.,

$$J(\lambda) = (\phi * \phi')(\lambda) = \int\limits_{-\infty}^{+\infty} \phi(\theta)\phi'(\lambda - \theta)d\theta, \tag{4.52}$$

see, e.g., [37, 38]. We can write (4.52) as $J(\lambda) = 1/n \sum_{i=1}^{n} \phi'(\lambda - \lambda_i)$. Similarly, because convolution is commutative, it holds that $J(\lambda) = 1/n \sum_{i=1}^{n} \phi(\lambda - \lambda_i')$.

The relation (4.52) gives a direct method to approximate the joint density of states. By replacing both densities of states in (4.52) with approximations from the Lanczos process, we derive a new Lanczos approximation to the joint density of states, distinct from that created by Method I. As outlined in Section 4.2, the densities of states $\phi$ and $\phi'$ are constructed using several trial vectors with entries in $\mathcal{N}(0,1)$. For simplicity, we begin by constructing an approximation to the joint density of states when the densities of states of $A$ and $A'$ have been approximated using one trial vector.

Let $w, w' \sim \mathcal{N}(0,1)$ be trial vectors. Then, as in (4.11), we know that $1/ns(\lambda; A, w)$ and $1/ns(\lambda; A', w')$ are equal in expectation to the densities of states $\phi$ and $\phi'$ respectively. Let $T_m \in \mathbb{R}^{m\times m}$ be the Lanczos partial tridiagonalization of $A$ with starting vector $w$, and $T_m' \in \mathbb{R}^{m\times m}$ the Lanczos partial tridiagonalization of $A'$ with starting vector $w'$. Replacing the spectral functions by their corresponding Lanczos approximations gives the following (single trail vector) approximations to the densities of states

$$\tilde{\phi}(\lambda) = \frac{\|w\|^2}{n}\sum_{j=1}^{m}|(y_j, e_1)|^2\delta(\lambda - \theta_j), \quad \text{and} \quad \tilde{\phi}'(\lambda) = \frac{\|w'\|^2}{n}\sum_{j=1}^{m}|(y_j', e_1)|^2\delta(\lambda - \theta_j'),$$

where $\theta_j$ and $y_j$, $j = 1, \ldots, m$, are the eigenvalues and normalized eigenvectors of $T_m$ respectively, and $\theta_j'$ and $y_j'$, $j = 1, \ldots, m$, are the eigenvalues and normalized eigenvectors of $T_m'$ respectively. The new Lanczos approximation to the joint density of states from (4.52) is,

$$\begin{aligned}\tilde{J}(\lambda) &= \frac{\|w\|^2}{n}\sum_{i=1}^{m}|(y_i,e_1)|^2\tilde{\phi}'(\lambda-\theta_i),\\ &= \frac{\|w\|^2\|w'\|^2}{n^2}\sum_{i,j=1}^{m}|(y_i,e_1)|^2|(y_j',e_1)|^2\delta\big(\lambda-(\theta_i+\theta_j')\big).\end{aligned} \tag{4.53}$$

Notice that the Dirac distributions in (4.53) are concentrated at the $m^2$ eigenvalues of the matrix $T_m \oplus T_m'$, and the coefficients, $|(y_i,e_1)|^2|(y_j',e_1)|^2$, are the square of the first component of the normalized eigenvectors, $y_i \otimes y_j'$. So, while we rejected $T_m \oplus T_m'$ in the previous section because it is not a Lanczos partial tridiagonalization of $A \oplus A'$, what (4.53) shows is that the eigenpairs of $T_m \oplus T_m'$ may be used to form an approximation to the joint density of states. Next, we show in what sense $\tilde{J}$ defined in (4.53) approximates the joint density of states. First, we state a simple lemma.

**Lemma 14.** *Let $\lambda_i$, $\lambda_i'$, $w_i$, and $w_i'$, $i=1,\dots,n$, be given real numbers, and for an integer, $d \geq 0$, suppose $\theta_j$, $\theta_j'$, $\tau_j$, and $\tau_j'$, $j=1,\dots,m$, are real numbers which satisfy the moment matching criterion*

$$\sum_{i=1}^{n} w_i\lambda_i^{\ell} = \sum_{j=1}^{m}\tau_j\theta_j^{\ell} \quad \textit{and} \quad \sum_{i=1}^{n} w_i'{\lambda_i'}^{\ell} = \sum_{j=1}^{m}\tau_j'{\theta_j'}^{\ell},$$

*for $\ell = 0,1,\dots,d$. Then,*

$$\sum_{i,j=1}^{n} w_i w_j'(\lambda_i+\lambda_j')^{\ell} = \sum_{i,j=1}^{m}\tau_i\tau_j'(\theta_i+\theta_j')^{\ell},$$

*for $\ell = 0,1,\dots,d$.*

*Proof.* By the Binomial Theorem $(\lambda_i+\lambda_j')^{\ell} = \sum_{k=0}^{\ell}\binom{\ell}{k}\lambda_i^{\ell-k}{\lambda_j'}^{k}$, and so for any integer

$0 \le \ell \le d$,

$$\begin{aligned}\sum_{i,j=1}^{n} w_i w_j'(\lambda_i + \lambda_j')^\ell &= \sum_{i,j=1}^{n}\sum_{k=0}^{\ell}\binom{\ell}{k} w_i w_j' \lambda_i^{\ell-k}{\lambda_j'}^k,\\ &= \sum_{k=0}^{\ell}\binom{\ell}{k}\Big(\sum_{i=1}^{n} w_i \lambda_i^{\ell-k}\Big)\Big(\sum_{j=1}^{n} w_j' {\lambda_j'}^k\Big),\\ &= \sum_{k=0}^{\ell}\binom{\ell}{k}\Big(\sum_{i=1}^{m} \tau_i \theta_i^{\ell-k}\Big)\Big(\sum_{j=1}^{m} \tau_j' {\theta_j'}^k\Big),\\ &= \sum_{i,j=1}^{n}\sum_{k=0}^{\ell}\binom{\ell}{k} \tau_i \tau_j' \theta_i^{\ell-k}{\theta_j'}^k,\\ &= \sum_{i,j=1}^{m} \tau_i \tau_j'(\theta_i + \theta_j')^\ell.\end{aligned}$$

□

Next, we show how Lemma 14, a simple consequence of the binomial theorem, explains the utility of the joint density of states approximation (4.53).

**Theorem 17.** *Let $A, A' \in \mathbb{R}^{n\times n}$ be symmetric and $v, v' \in \mathbb{R}^n$. Additionally, let $T_m, T_m' \in \mathbb{R}^{m\times m}$, with normalized eigenpairs $T_m y_j = \theta_j y_j$ and $T_m' y_j' = \theta_j y_j'$, $j = 1, \dots, m$, be the order $m$ Lanczos partial tridiagonalizations of $A$ with starting vector $v$ and $A'$ with starting vector $v'$ respectively. Then, the first $2m-1$ moments of $\|v\|^2\|v'\|^2 \sum_{i,j=1}^{m} |(y_i, e_1)|^2 |(y_j', e_1)|^2 \delta\big(\lambda - (\theta_i + \theta_j')\big)$ match those of $s(\lambda; A \oplus A', v \otimes v')$.*

*Proof.* Let $A$ and $A'$ have eigenpairs as in (4.1). The measure $\|v\|^2 \sum_{j=1}^{m} |(y_j, e_1)|^2 \delta(\lambda - \theta_j)$ is the result of the Lanczos process with the matrix $A$ and vector $v$, and hence an approximation to the spectral function, $s(\lambda; A, v) = \sum_{i=1}^{n} |(x_i, v)|^2 \delta(\lambda - \lambda_i)$, in the sense that the first $2m-1$ moments match (see (4.6)), i.e.,

$$\sum_{i=1}^{n} |(x_i, v)|^2 \lambda_i^\ell = \|v\|^2 \sum_{j=1}^{m} |(y_j, e_1)|^2 \theta_j^\ell, \qquad \ell = 0, 1, \dots, 2m-1.$$

The same holds true for all primed quantities,

$$\sum_{i=1}^{n} |(x_i', v')|^2 {\lambda_i'}^{\ell} = \|v'\|^2 \sum_{j=1}^{m} |(y_j', e_1)|^2 {\theta_j'}^{\ell}, \qquad \ell = 0, 1, \ldots, 2m-1.$$

A straightforward application of Lemma 14 with $d = 2m-1$, $w_i = |(x_i, v)|^2$, $w_i' = |(x_i', v')|^2$, $\tau_j = \|v\|^2 |(y_j, e_1)|^2$, and $\tau_j' = \|v'\|^2 |(y_j', e_1)|^2$ shows

$$\sum_{i,j=1}^{n} |(x_i, v)|^2 |(x_j', v')|^2 (\lambda_i + \lambda_j')^{\ell} = \|v\|^2 \|v'\|^2 \sum_{i,j=1}^{m} |(y_i, e_1)|^2 |(y_j', e_1)|^2 (\theta_i + \theta_j')^{\ell}, \quad (4.54)$$

for $\ell = 0, 1, \ldots, 2m-1$. The result follows by noticing the left hand side of (4.54) is the $\ell$th moment of $s(\lambda; A \oplus A', v \otimes v')$ and the right hand side is the $\ell$th moment of $\|v\|^2 \|v'\|^2 \sum_{i,j=1}^{m} |(y_i, e_1)|^2 |(y_j', e_1)|^2 \delta\big(\lambda - (\theta_i + \theta_j')\big)$. □

Theorem 17 shows us that $\tilde{J}$ from (4.53) matches the first $2m-1$ moments of the spectral function $1/n^2 s(\lambda; A \oplus A, w \otimes w')$. In turn, the spectral function $1/n^2 s(\lambda; A \oplus A, w \otimes w')$ is equal in expectation to the joint density of states when $w, w' \sim \mathcal{N}(0,1)$. In other words, by using the Lanczos process to approximate the densities of states for $A$ and $A'$, we can form an accurate approximation to the density of states of $A \oplus A'$.

We remark that while Theorem 17 is stated for two order $m$ Lanczos approximations to the spectral functions $s(\lambda; A, v)$ and $s(\lambda; A', v')$, the user may also choose different order approximations for each. In this case, if one produces an order $m$ Lanczos approximation to $s(\lambda; A, v)$, an order $m'$ Lanczos approximation to $s(\lambda; A', v')$, and convolve them (as in (4.53)), the result matches the first $2\min\{m, m'\}-1$ moments of the spectral function $s(\lambda; A \oplus A', v \otimes v')$.

As in the case of the density of states, for the method II Lanczos approximation to the joint density of states we stochastically average results over many trial vectors. To this end, let $w^{(k)}, {w'}^{(k)} \sim \mathcal{N}(0,1)$, $k = 1, \ldots, n_v$, denote trial vectors. For each trial vector we partially tridiagonalize $A$ and $A'$ by performing the $m$-step Lanczos algorithm with starting vectors $w^{(k)}$ and ${w'}^{(k)}$ respectively, obtaining $T_m^{(k)}, {T_m'}^{(k)} \in \mathbb{R}^{m \times m}$. Denote the eigenpairs of each partial tridiagonalization as, $T_m^{(k)} y_i^{(k)} = \theta_i^{(k)} y_i^{(k)}$, $\|y_i^{(k)}\| = 1$, and ${T_m'}^{(k)} {y_j'}^{(k)} = {\theta_j'}^{(k)} {y_j'}^{(k)}$, $\|{y_j'}^{(k)}\| = 1$, for $i, j = 1, \ldots, m$. Then, the

Lanczos approximation to the joint density of states is

$$\tilde{J}(\lambda) = \frac{1}{n_v^2 n^2} \sum_{k,\ell=1}^{n_v} \sum_{i,j=1}^{m} \|w^{(k)}\|^2 \|w'^{(\ell)}\|^2 \big|\big(y_i^{(k)}, e_1\big)\big|^2 \big|\big(y_j'^{(\ell)}, e_1\big)\big|^2 \delta\big(\lambda - (\theta_i^{(k)} + \theta_j'^{(\ell)})\big). \tag{4.55}$$

The method II approximation to the joint density of states is summarized in Algorithm 15.

**Algorithm 15** Lanczos Approximation of the Joint Density of States (method II)

```
Initialize $m$, $n_v$.
for $k = 1, \dots, n_v$ do
    Draw trial vectors $w^{(k)}$, $w'^{(k)} \sim \mathcal{N}(0,1)$.
    Partially tridiagonalize $A$ and $A'$ with starting vectors $w^{(k)}$ and $w'^{(k)}$ respectively to get $T_m^{(k)}$, $T'^{(k)}_m \in \mathbb{R}^{m \times m}$.
    Compute eigenpairs $T_m^{(k)} y_i^{(k)} = \theta_i^{(k)} y_i^{(k)}$, $\|y_i^{(k)}\| = 1$, and $T'^{(k)}_m y'^{(k)}_i = \theta'^{(k)}_i y'^{(k)}_i$, $\|y'^{(k)}_i\| = 1$, $i = 1, \dots, m$.
end for
$\tilde{J}(\lambda) = \frac{1}{n_v^2 n^2} \sum_{k,\ell=1}^{n_v} \sum_{i,j=1}^{m} \|w^{(k)}\|^2 \|w'^{(\ell)}\|^2 |(y_i^{(k)}, e_1)|^2 |(y'^{(\ell)}_j, e_1)|^2 \delta\big(\lambda - (\theta_i^{(k)} + \theta'^{(\ell)}_j)\big)$.
```

Finally, we make a few comments on the differences in method I and method II for producing a Lanczos approximation to the joint density of states. First, we notice that method I, with one trial vector $w \otimes w'$, is an $m$-point Gauss quadrature approximation to the spectral function $s(\lambda; A \oplus A', w \otimes w')$, while method II is an $m^2$-point quadrature approximation to the same spectral function. Both methods match the first $2m - 1$ moments of $s(\lambda; A \oplus A', w \otimes w')$, and both methods have positive weights. Hence, both quadrature rules are convergent of class $C(\overline{\Omega})$, as Steklov's Theorem (mentioned in the previous chapter) stipulates. Both methods post-process Lanczos partial tridiagonalizations of $A$ and $A'$ to create an approximation to the joint density of states. In short, for both methods we get three spectral quantities for the price of two (the eigenpairs of Lanczos partial tridiagonalizations giving approximations to the density of states of $A$ and $A'$ by Algorithm 10, and by post-processing we get an approximation to the joint density of states). Method I is optimal in the sense that it recreates the maximal number of moments for a given number of quadrature nodes (due to its relation to Gauss quadrature). Method II is ideal because for $n_v$ partial tridiagonalizations of $A$ and $A'$, we average $n_v^2$ approximations to the joint density of states. This is in contrast to method I, where $n_v$ partial tridiagonalizations of $A$ and $A'$ results in averaging $n_v$ approximations to the joint density of states.

## 4.5 Joint Density of States for Generalized Eigenvalue Problems

In this section we approximate the joint density of states, $J(\lambda) = 1/n^2 \sum_{i,j=1}^{n} \delta(\lambda - (\lambda_i + \lambda_j'))$, where the eigenvalues stem from the generalized eigenvalue problems (4.2). As for the density of states for a generalized eigensystem, the main difference is the choice of starting vector for the Lanczos algorithm. Instead of using a random vector with entries in $\mathcal{N}(0,1)$, we need to perform a linear operation on this vector first. Most of the heavy lifting for this section has already been performed in Section 4.3 and 4.4, and so we give a simple overview of approximating the joint density of states by method I and II. As we saw in Section 4.3, by scaling using the diagonal of the mass matrix (see (4.15)), the new mass matrix $D^{-1/2}BD^{-1/2}$ is better conditioned. Moving forward, we assume this scaling has already been performed.

As seen previously, the first step is to factor the mass matrix in order to reformulate the generalized eigenvalue problem as a standard eigenvalue problem. Letting $B = LL^T$ be the Cholesky or square root factorization of the mass matrix, the generalized eigensystems in (4.2) become

$$\begin{aligned} Cz_i &= \lambda_i z_i, & z_i^T z_j &= \delta_{ij} \\ C'z_i' &= \lambda_i' z_i', & z_i'^T z_j' &= \delta_{ij}, \end{aligned} \tag{4.56}$$

for $i, j = 1, \ldots, n$, where $C = L^{-1}AL^{-T}$, $C' = L^{-1}A'L^{-T}$, $z_i = L^T x_i$, and $z_i' = L^T x_i'$. Note that when transitioning from the generalized eigenvalue problem (4.2), to the standard eigenvalue problem (4.56), the eigenvalues remain unaltered (in contrast with the eigenvectors). This tells us the joint density of states corresponding to the pair of generalized eigenvalue problems (4.2) is the joint density of states for the standard eigenvalue problems (4.56).

Using the definition of the Kronecker sum, Theorem 15 tells us the joint density of states for the systems (4.56) is the density of states for $C \oplus C'$. We are now in position to use the results of Section 4.4 to approximate the density of states for $C \oplus C'$. Choosing trial vectors $w, w' \sim \mathcal{N}(0,1)$, we know that the spectral function corresponding to the matrix $C \oplus C'$ and vector $w \otimes w'$ satisfies

$$J = \frac{1}{n^2}\mathbb{E}\Big[s(\lambda; C \oplus C', w \otimes w')\Big]. \tag{4.57}$$

Therefore, in order to approximate the joint density of states for the generalized eigenvalue problems (4.2), we use the Lanczos process to approximate the spectral function $s(\lambda; C \oplus C', w \otimes w')$.

In order to relate $s(\lambda; C \oplus C', w \otimes w')$ to the $B$-Lanczos method we use the same device deployed in (4.18). Namely, for any vectors $u, u' \in \mathbb{R}^n$, the eigenvectors of $C$ and $C'$ defined in (4.56) satisfy

$$(z_i, u) = (x_i, v)_B \quad \text{and} \quad (z_i', u') = (x_i', v')_B, \tag{4.58}$$

where $v = L^{-T}u$ and $v' = L^{-T}u'$. This shows that the spectral function $s(\lambda; C, u)$ is equivalent to the spectral function $s(\lambda; A, B, v)$ with $v = L^{-T}u$. Similarly, $s(\lambda; C', u') = s(\lambda; A', B, v')$ with $v' = L^{-T}u'$. We next show how to use $B$-Lanczos algorithms to approximate the joint density of states using method I and II of the previous section.

### 4.5.1 Method I

Method I relies on implicitly forming the Lanczos partial tridiagonalization of $C \oplus C'$. This is accomplished (see Algorithm 13) by performing the partial tridiagonalization of $C$ and $C'$ individually, and then combining the results to get the partial tridiagonalization of $C \oplus C'$. Recall from Chapter 2 that the Lanczos partial tridiagonalization of $C$, with starting vector $u$, is the same as the symmetric tridiagonal matrix resulting from the $B$-Lanczos algorithm with the matrices $A$ and $B$, and starting vector $v = L^{-T}u$. Similarly, the Lanczos partial tridiagonalization of $C'$, with starting vector $u'$, is the symmetric tridiagonal matrix resulting from the $B$-Lanczos algorithm with matrices $A'$ and $B$, and starting vector $v' = L^{-T}u'$. Therefore, we can alter Algorithm 13 to use the $B$-Lanczos algorithm, with the proper starting vector, and in this way create the Lanczos partial tridiagonalization of $C \oplus C'$ with starting vector $u \otimes u'$ for arbitrary vectors $u, u' \in \mathbb{R}^n$. This is shown in Algorithm 16. Note that while Algorithm 16 involves the matrix vector products $L^{-T}u$ and $L^{-T}u'$, these are to be approximated using the Chebyshev expansion of the inverse square root of $B$, discussed at length in Section 4.3. Also, as in Section 4.4, we use the full orthogonalization variant of the B-Lanczos algorithm for robustness.

**Algorithm 16** *B*-Lanczos Algorithm on Kronecker Sum (full orthogonalization)

```
Initialize $m$, $v_0 = L^{-T}u$, $v'_0 = L^{-T}u'$, $w_0 = Bv_0$, $w'_0 = Bv'_0$, $v_1 = v_0/\sqrt{v_0^T w_0}$, $w_1 = w_0/\sqrt{v_0^T w_0}$, $v'_1 = v'_0/\sqrt{v_0'^T w'_0}$, $w'_1 = w'_0/\sqrt{v_0'^T w'_0}$, $\beta_0 = \beta'_0 = 0$, $\gamma^1_{11} = 1$.
for $k = 1, \dots, m$ do
    $\tilde{w} = Av_k - \beta_{k-1} w_{k-1}$
    $\tilde{w}' = A'v'_k - \beta'_{k-1} w'_{k-1}$
    $\alpha_k = (\tilde{w}, v_k)$
    $\alpha'_k = (\tilde{w}', v'_k)$
    for $i = 1, \dots, k$ do
        $\tilde{w} \leftarrow \tilde{w} - (\tilde{w}, v_i) w_i$
        $\tilde{w}' \leftarrow \tilde{w}' - (\tilde{w}', v'_i) w'_i$
    end for
    Solve $B\tilde{v} = \tilde{w}$ for $\tilde{v}$
    Solve $B\tilde{v}' = \tilde{w}'$ for $\tilde{v}'$
    $\beta_k = (\tilde{v}, \tilde{w})$
    $\beta'_k = (\tilde{v}', \tilde{w}')$
    if $\beta_k = 0$ or $\beta'_k = 0$ then stop
    else
        $v_{k+1} = \frac{\tilde{v}}{\beta_k}$ and $w_{k+1} = \frac{\tilde{w}}{\beta_k}$
        $v'_{k+1} = \frac{\tilde{v}'}{\beta'_k}$ and $w'_{k+1} = \frac{\tilde{w}'}{\beta'_k}$
    end if
    for $i, j = 1, \dots, k+1$ do
        $\eta^{k+1}_{ij} = \gamma^k_{ij}(\alpha_i + \alpha'_j) + \gamma^k_{i+1\,j}\beta_i + \gamma^k_{i\,j+1}\beta'_j + \gamma^k_{i-1\,j}\beta_{i-1} + \gamma^k_{i\,j-1}\beta'_{j-1}$
    end for
    $\alpha^{\oplus}_k = (\gamma^k, \eta^{k+1})_F$
    $\tilde{\gamma}^{k+1} = \eta^{k+1} - \alpha^{\oplus}_k \gamma^k - \beta^{\oplus}_{k-1}\gamma^{k-1}$
    $\beta^{\oplus}_k = \|\tilde{\gamma}^{k+1}\|_F$
    $\gamma^{k+1} = \frac{1}{\beta^{\oplus}_k}\tilde{\gamma}^{k+1}$
end for
```

Using Algorithm 16, we can produce the partial tridiagonalization of $C \oplus C'$ with starting vector $w \otimes w'$, where $w, w' \sim \mathcal{N}(0, 1)$ in order to approximate the joint density of states. Using the partial tridiagonalization, say $T_m^{\oplus} \in \mathbb{R}^{m \times m}$, we are able to use the Lanczos process to approximate $s(\lambda; C \oplus C', w \otimes w')$, with the spectral function $s(\lambda; C \oplus C', w \otimes w')$ being equal in expectation to the joint density of states (see (4.57)). Let the eigenvalues and normalized eigenvectors of $T_m^{\oplus}$ be $\theta_j$ and $y_j$, $j = 1, \ldots, m$, respectively. The method I approximation to the joint density of states (for one trial vector) is

$$\tilde{J}(\lambda) = \frac{\|w\|^2 \|w'\|^2}{n^2} \sum_{j=1}^{m} |(y_j, e_1)|^2 \delta(\lambda - \theta_j). \tag{4.59}$$

Obviously, by performing the Lanczos process for several trial vectors, and averaging the result, we produce a more accurate approximation to the joint density of states. This is included in Algorithm 17 for $n_v$ trial vectors.

**Algorithm 17** B-Lanczos Approximation of the Joint Density of States (method I)

```
Initialize $m$, $n_v$, and set $k = 0$ and $\tilde{J}(\lambda) = 0$.
while $k < n_v$ do
    Draw trial vectors $w, w' \sim \mathcal{N}(0, 1)$.
    Partially tridiagonalize $C \oplus C'$ with starting vector $w \otimes w'$ (Algorithm 16) to get $T_m^{\oplus} \in \mathbb{R}^{m \times m}$.
    Compute eigenpairs $T_m^{\oplus} y_j = \theta_j y_j$, $y_i^T y_j = \delta_{ij}$, $i, j = 1, \ldots, m$.
    $\tilde{J}(\lambda) \leftarrow \tilde{J}(\lambda) + \frac{\|w\|^2 \|w'\|^2}{n_v n^2} \sum_{j=1}^{m} |(y_j, e_1)|^2 \delta(\lambda - \theta_j)$.
    $k \leftarrow k + 1$.
end while
```

### 4.5.2 Method II

For the method II approximation to the joint density of states for generalized eigensystems, we convolve the Lanczos approximations to the density of states for the individual generalized eigensystems. To make concepts concrete, we start by constructing the density of states approximations for each eigensystem using one trial vector, before generalizing to many trial vectors. Let $B = LL^T$ be a factorization of the symmetric positive definite mass matrix, and let $w, w' \sim \mathcal{N}(0, 1)$ be trial vectors. By performing the $B$-Lanczos algorithm on $A$ and $B$ with starting vector $v = L^{-T} w$, we obtain the order $m$ partial tridiagonalization $T_m \in \mathbb{R}^{m \times m}$. Similarly, using the $B$-Lanczos algorithm with starting vector $v' = L^{-T} w'$, we obtain $T_m' \in \mathbb{R}^{m \times m}$, the order $m$ partial tridiagonalization of $A'$ and $B$. Let the eigenvalues and normalized eigenvectors of the partial tridiagonalizations be given by $T_m y_j = \theta_j y_j$, $T'_m y'_j = \theta'_j y'_j$, for $j = 1, \ldots, m$. The Lanczos approximation to the density of states (with one trial vector) for the matrix pair $A$ and $B$ is

$$\tilde{\phi}(\lambda) = \frac{\|w\|^2}{n} \sum_{j=1}^{m} |(y_j, e_1)|^2 \delta(\lambda - \theta_j),$$

and for the matrix pair $A'$ and $B$,

$$\tilde{\phi}'(\lambda) = \frac{\|w'\|^2}{n} \sum_{j=1}^{m} |(y_j, e_1)|^2 \delta(\lambda - \theta_j).$$

With approximations to the densities of states, we now convolve $\phi$ and $\phi'$ to construct an approximation to the joint density of states given by

$$\begin{aligned} \tilde{J}(\lambda) &= \frac{\|w\|^2}{n} \sum_{i=1}^{m} |(y_i, e_1)|^2 \tilde{\phi}'(\lambda - \theta_i), \\ &= \frac{\|w\|^2 \|w'\|^2}{n^2} \sum_{i,j=1}^{m} |(y_i, e_1)|^2 |(y'_j, e_1)|^2 \delta\big(\lambda - (\theta_i + \theta'_j)\big). \end{aligned} \tag{4.60}$$

As in method II for the standard eigenvalue problems, the above approximation simply amounts to multiplying all coefficients from the approximate densities of states, and adding all combinations of Ritz values. Because the densities of states, $\phi$ and $\phi'$, match the first $2m - 1$ moments of $s(\lambda; C, w) = s(\lambda; A, B, v)$ and $s(\lambda; C', w') = s(\lambda; A', B, v')$ respectively, the approximate joint density of states (4.60) will match the first $2m - 1$ moments of $s(\lambda; C \oplus C', w \otimes w')$ as shown in Theorem 17. In turn, as illustrated in (4.57), $1/n^2 s(\lambda; C \oplus C', w \otimes w')$ equals the joint spectral function in expectation. Therefore, by approximating the densities of states for both systems, we gain an approximation to the joint density of states at no (significant) additional expense.

By performing the same approximation using additional trial vectors, we gain a better approximation to the joint density of states. This is summarized in Algorithm 18. Note that while the Cholesky factor $L$ is used ($B = L^T L$) in Algorithm 18, in practice we utilize a Chebyshev approximation $L^{-T}$ in order to form the starting vectors, as outlined in Section 4.3.

**Algorithm 18** $B$-Lanczos Approximation of the Joint Density of States (method II)

```
Initialize $m$, $n_v$.
for $k = 1, \dots, n_v$ do
    Draw trial vectors $w^{(k)}, w'^{(k)} \sim \mathcal{N}(0,1)$.
    Form starting vectors $v^{(k)} = L^{-T}u^{(k)}$ and $v'^{(k)} = L^{-T}u'^{(k)}$.
    Perform $B$-Lanczos with $A$ and $A'$ and starting vectors $w^{(k)}$ and $w'^{(k)}$ to get $T_m^{(k)}, T'^{(k)}_m \in \mathbb{R}^{m \times m}$.
    Compute eigenpairs $T_m^{(k)} y_i^{(k)} = \theta_i^{(k)} y_i^{(k)}$, $\|y_i^{(k)}\| = 1$, and $T'^{(k)}_m y'^{(k)}_i = \theta'^{(k)}_i y'^{(k)}_i$, $\|y'^{(k)}_i\|$, $i = 1, \dots, m$.
end for
$\tilde{J}(\lambda) = \frac{1}{n_v^2 n^2} \sum_{k,\ell=1}^{n_v} \sum_{i,j=1}^{m} \|w^{(k)}\|^2 \|w'^{(\ell)}\|^2 |(y_i^{(k)}, e_1)|^2 |(y'^{(\ell)}_j, e_1)|^2 \delta\big(\lambda - (\theta_i^{(k)} + \theta'^{(\ell)}_j)\big)$.
```

## 4.6 Joint Spectral Function

In this section we introduce one final joint spectral quantity. Namely, that of a joint spectral function for a pair of eigenvalue problems. Let $A, A' \in \mathbb{R}^{n \times n}$ be symmetric matrices with eigenpairs as in (4.1). We define the joint spectral function as

$$\alpha(\lambda) = \sum_{i,j=1}^{n} |(x_i, x'_j)|^2 \delta\big(\lambda - (\lambda_i + \lambda'_j)\big). \tag{4.61}$$

To produce the joint spectral function exactly, a complete accounting of all eigenpairs of $A$ and $A'$ is required. This is in contrast to the joint density of states where only the eigenvalues are present. We will show that the joint spectral function is the spectral function for the matrix $A \oplus A'$ and vector $\sum_{i=1}^{n} e_i \otimes e_i$. It is used in semiconductor physics, and is important in the determination of optical properties of light emitting diodes. We will see specific applications of the joint spectral function in the next chapter.

Recall the outline for approximating spectral quantities. We first relate the spectral quantity to a spectral function, and then use the Lanczos process to approximate the spectral function. Following this roadmap, we look at spectral functions for the matrix $A \oplus A'$. Using results of Theorem 15, we see that the spectral function of $A \oplus A'$ with starting vector $u \in \mathbb{R}^{n^2}$ is given by

$$s(\lambda; A \oplus A', u) = \sum_{i,j=1}^{n} \big|(x_i \otimes x'_j, u)\big|^2 \delta\big(\lambda - (\lambda_i + \lambda'_j)\big). \tag{4.62}$$

Equation (4.62) shows that if we choose $u$ such that $(x \otimes x', u) = (x, x')$ for arbitrary $x, x' \in \mathbb{R}^n$, then the spectral function corresponding to $A \oplus A'$ and $u$ is equal to the joint spectral function. This would allow us to perform the Lanczos process on $A \oplus A'$ (potentially utilizing Algorithm 13) in order to approximate the joint spectral function. It is easily verified that the desired $u$ is given by

$$u = \sum_{i=1}^{n} e_i \otimes e_i, \tag{4.63}$$

where $e_i$ is the $i$th column of the identity matrix of order $n$.

The starting vector (4.63) poses several issues. First, we see that $u$ is a sum of rank one vectors. This is problematic because the spectral function is not linear in the starting vector, i.e., for vectors $v, w \in \mathbb{R}^n$,

$$s(\lambda; A, v + w) \neq s(\lambda; A, v) + s(\lambda; A, w). \tag{4.64}$$

Another way of saying (4.64) is, $(v+w)^T f(A)(v+w) \neq v^T f(A)v + w^T f(A)w$ for smooth $f$ in general. Hence, we are unable to perform $n$ Lanczos processes on the matrix $A \oplus A'$ with rank one vectors, $e_i \otimes e_i$, $i = 1, \ldots, n$, in order to approximate the spectral function $s(\lambda; A \oplus A', u)$. In fact, we can show that for any test function $f$ and $u$ given by (4.63) we have

$$\langle s(\lambda; A \oplus A', u), f \rangle = \sum_{i,j=1}^{n} (e_i \otimes e_i)^T f(A \oplus A')(e_j \otimes e_j). \tag{4.65}$$

While it is possible to approximate bilinear forms $v^T f(A)w$, $v \neq w$, the standard method is to use the identity

$$v^T f(A)w = 1/4\big((v + w)^T f(A)(v + w) - (v - w)^T f(A)(v - w)\big),$$

see, e.g., [17]. This is again problematic because $e_i \otimes e_i + e_j \otimes e_j$ and $e_i \otimes e_i - e_j \otimes e_j$ are rank two for $i \neq j$, and so do not fit within the framework of Algorithm 13 which requires a rank one starting vector. Second, even if (4.64) were true, because $n \gg 1$ for problems of interest, it would be too costly to approximate $s(\lambda; A \oplus A', u)$ using $n$ Lanczos processes on the spectral functions $s(\lambda; A \oplus A', e_i \otimes e_i)$, $i = 1, \ldots, n$. Thus, we must devise a new strategy to approximate the joint spectral function, rather than rely on old tools.

Next, consider the following scenario. Suppose we know the eigenpairs of $A'$, i.e., we have exactly computed eigenvalues $\lambda'_j \in \mathbb{R}$ and corresponding orthonormal eigenvectors $x'_j \in \mathbb{R}^n$ for $j = 1, \ldots, n$. The spectral function corresponding to $A$ and $x'_j$ is given by

$$s_j(\lambda) \coloneqq s(\lambda; A, x'_j) = \sum_{i=1}^{n} |(x_i, x'_j)|^2 \delta(\lambda - \lambda_i). \tag{4.66}$$

Other than shifting by a factor $\lambda'_j$, (4.66) is the marginal obtained by fixing the index $j$ in (4.61). In other words,

$$\alpha(\lambda) = \sum_{j=1}^{n} s_j(\lambda - \lambda'_j). \tag{4.67}$$

Similarly, if we know the eigenvalues $\lambda_i \in \mathbb{R}$ and corresponding orthonormal eigenvectors $x_i \in \mathbb{R}^n$ of $A$, $i = 1, \ldots, n$, then the spectral function corresponding to $A'$ and $x_i$ is

$$s'_i(\lambda) \coloneqq s(\lambda; A', x_i) = \sum_{j=1}^{n} |(x_i, x'_j)|^2 \delta(\lambda - \lambda'_j),$$

and the joint spectral function is given by

$$\alpha(\lambda) = \sum_{i=1}^{n} s'_i(\lambda - \lambda_i).$$

Equation (4.67) tells us that if we know the eigenpairs of $A'$, then we can obtain the joint spectral function using the spectral functions corresponding to $A$ and the eigenvectors of $A'$. So, by approximating the eigenpairs of $A'$, and performing $n$ Lanczos processes on $A$ with the approximate eigenvectors of $A'$, we can approximate the joint spectral function. Continuing with our assumption that we know the eigenpairs of $A'$, let $T_m^{(j)} \in \mathbb{R}^{m \times m}$ be the partial tridiagonalization of $A$ with starting vector $x'_j$, $j = 1, \ldots, n$, and denote the eigenpairs of the partial tridiagonalization as $T_m^{(j)} y_k^{(j)} = \theta_k^{(j)} y_k^{(j)}$, $k = 1, \ldots, m$. Then, the Lanczos approximation to the joint spectral function is given by

$$\tilde{\alpha}(\lambda) = \sum_{j=1}^{n} \tilde{s}_j(\lambda - \lambda'_j), \tag{4.68}$$

where,

$$\tilde{s}_j(\lambda) = \sum_{k=1}^{m} |(y_k^{(j)}, e_1)|^2 \delta(\lambda - \theta_k^{(j)}). \tag{4.69}$$

While this may seem like an exercise in futility, given that we need to fully diagonalize a matrix in order to approximate the joint spectral function, our examples in the next chapter prove otherwise. The reasoning is simple. First, note that we have reduced the workload by half. Instead of approximating eigenpairs of both matrices $A$ and $A'$, we now only need to approximate the eigenpairs of one or the other to estimate the joint spectral function. Second, in certain situations, we are only interested in approximating the joint spectral function in an interval $[\underline{\lambda}, \overline{\lambda}]$. This means that we are only required to approximate the eigenpairs of $A'$ (or of $A$) for eigenvalues in a certain range. We discuss specifically which eigenpairs of $A'$ are required next.

Assume we want to approximate the joint spectral function in the interval $[\underline{\lambda}, \overline{\lambda}]$ where $\underline{\lambda} \leq \lambda_1 + \lambda'_1$, i.e., we are interested in the "bottom" portion of the joint spectral function (recall the eigenvalues of both $A$ and $A'$ are in ascending order). Next, we show that we only need to compute a portion of the spectrum of $A'$ for the Lanczos process, dependent on the magnitude of $\overline{\lambda}$. This greatly reduces the complexity of the problem, and makes the method suitable for two and three dimensional computations.

We begin with a simple observation. Assume for the moment we know the first $\overline{i}$ eigenpairs of $A$ and $\overline{j}$ eigenpairs of $A'$, i.e., we have eigenvalues $\lambda_i$ and corresponding eigenvectors $x_i$, for $i = 1, \ldots, \overline{i}$ and $\lambda'_j$ and $x'_j$ for $j = 1, \ldots, \overline{j}$, with $\overline{i}, \overline{j} \leq n$. Using only these eigenpairs we may approximate the joint spectral function as

$$\alpha(\lambda) \approx \sum_{i=1}^{\overline{i}} \sum_{j=1}^{\overline{j}} |(x_i, x'_j)|^2 \delta\big(\lambda - (\lambda_i + \lambda'_j)\big). \tag{4.70}$$

Notice, however, that we can only rely on approximation (4.70) for $\lambda$ between

$$\underline{\lambda} \leq \lambda \leq \min(\lambda_1 + \lambda'_{\overline{j}}, \lambda_{\overline{i}} + \lambda'_1). \tag{4.71}$$

The maximum value for which we trust the approximation (4.70) is determined by (4.71) by the following reasoning: if more eigenvalues of $A$ and $A'$ are computed, then the new terms added to the right hand side of (4.70) involve Dirac masses concentrated at

$\lambda_{\bar{i}+i} + \lambda'_j$ and $\lambda_i + \lambda'_{\bar{j}+j}$ for $i, j \geq 1$, which are all greater than $\min(\lambda_1 + \lambda'_{\bar{j}}, \lambda_{\bar{i}} + \lambda'_1)$. Hence, we know that for all values $\lambda$ satisfying (4.71), the joint spectral function is unchanged with the addition of these new terms. The conclusion being, if we want to use (4.70) to approximate the joint spectral function for $\lambda \in [\underline{\lambda}, \overline{\lambda}]$, we need to compute enough eigenpairs of $A$ and $A'$ such that $\overline{\lambda} \leq \min(\lambda_1 + \lambda'_{\bar{j}}, \lambda_{\bar{i}} + \lambda'_1)$, where $\lambda_{\bar{i}}$ and $\lambda'_{\bar{j}}$ are the largest eigenvalues computed of $A$ and $A'$ respectively.

Now, we translate the above reasoning to the approximation of the joint spectral function using the Lanczos process on $A$ with starting vectors equal to the eigenvectors of $A'$, as in (4.67). The above reasoning suggests that if we are interested in approximating the joint spectral function for values of $\lambda \leq \overline{\lambda}$, we need to compute all eigenpairs of $A'$ with eigenvalues less than or equal to $\overline{\lambda} - \lambda_1$. Similarly, if we approximate the joint spectral function using the Lanczos process on $A'$ with starting vectors equal to the eigenvectors of $A$, then we need to compute all eigenpairs of $A$ with eigenvalues less than or equal to $\overline{\lambda} - \lambda'_1$. Note that, assuming eigenvalues of $A$ and $A'$ are positive, $\lambda_1$ and $\lambda'_1$ are easy to approximate using a few iterations of the inverse power method, or other more sophisticated methods. Note also that, in practice, all that is needed are lower bounds for $\lambda'_1$ or $\lambda_1$. However, a poor lower bound will require the computation of additional eigenpairs to ensure (4.71) is satisfied. Hence, it may be worthwhile to compute $\lambda_1$ to reasonable accuracy.

Using the above reasoning, the Lanczos approximation to the joint spectral function in the interval $[\underline{\lambda}, \overline{\lambda}]$, with $\underline{\lambda} \leq \lambda_1 + \lambda'_1$, is summarized in Algorithm 19.

**Algorithm 19** Lanczos Approximation to the Joint Spectral Function

```
Initialize $m$, $\overline{\lambda}$, and set $\tilde{\alpha}(\lambda) = 0$.
Compute lower bound $\tilde{\lambda}_1$ such that $\tilde{\lambda}_1 \leq \lambda_1 = \lambda_{\min}(A)$.
Compute eigenpairs of $A'$ with eigenvalues less than $\overline{\lambda} - \tilde{\lambda}_1$, $A'x'_j = \lambda'_j x'_j$, $\|x'_j\| = 1$,
  $j = 1, \ldots, \bar{j}$.
for $j = 1, \ldots, \bar{j}$ do
    Partially tridiagonalize $A$ with starting vector $x'_j$, obtaining $T_m \in \mathbb{R}^{m \times m}$.
    Compute eigenpairs $T_m y_k = \theta_k y_k$, $y_k^T y_\ell = \delta_{k\ell}$, for $k, \ell = 1, \ldots, m$.
    $\tilde{\alpha}(\lambda) \leftarrow \tilde{\alpha}(\lambda) + \sum_{k=1}^{m} |(y_k, e_1)|^2 \delta\big(\lambda - (\theta_k + \lambda'_j)\big)$
end for
```

## 4.7 Joint Spectral Function for Generalized Eigenvalue Problems

In this section we present the Lanczos approximation to the joint spectral function corresponding to a pair of generalized eigenvalue problems. We use the same notation as in last section, with the only difference being the use of the $B$-inner product, e.g., the joint spectral function becomes

$$\alpha(\lambda) = \sum_{i,j=1}^{n} |(x_i, x'_j)_B|^2 \delta\big(\lambda - (\lambda_i + \lambda'_j)\big), \tag{4.72}$$

where the eigenpairs are as in (4.2). Similarly, the marginal (4.66) becomes

$$s_j(\lambda) := s(\lambda; A, B, x'_j) = \sum_{i=1}^{n} |(x_i, x'_j)_B|^2 \delta(\lambda - \lambda_i). \tag{4.73}$$

The joint spectral function (4.72) and marginal (4.73) are related by

$$\alpha(\lambda) = \sum_{j=1}^{n} s_j(\lambda - \lambda'_j). \tag{4.74}$$

We can use the $B$-Lanczos process to approximate $s_j(\lambda)$ in (4.73), and by replacing $s_j(\lambda - \lambda'_j)$ in (4.74) with the corresponding Lanczos approximation, we create an approximation to the joint spectral function. If we write $\tilde{s}_j(\lambda)$ as the Lanczos approximation to the spectral function $s(\lambda; A, B, x'_j)$, then the approximation to the joint spectral function is given by

$$\tilde{\alpha}(\lambda) = \sum_{j=1}^{n} \tilde{s}_j(\lambda - \lambda'_j). \tag{4.75}$$

Note that, as in the previous section, if we are only interested in approximating the joint spectral function for values $\lambda \in [\underline{\lambda}, \overline{\lambda}]$, we only need to approximate some of the eigenpairs of the matrix pencil involving $A'$ and $B$. The specific number determined by the value $\overline{\lambda} - \lambda_1$ (here we again assume $\underline{\lambda} \leq \lambda_1 + \lambda'_1$).

In the case of the density of states and joint density of states, it was necessary to change the starting vector when transitioning from the standard eigenvalue problem to the generalized eigenvalue problem, in addition to using the $B$-Lanczos method. However, the situation is simpler in the case of the joint spectral function. The only modification needed is the use of the $B$-Lanczos method instead of the standard Lanczos algorithm. The Lanczos approximation to the joint spectral function corresponding to a pair of generalized eigenvalue systems is given in Algorithm 20.

**Algorithm 20** $B$-Lanczos Approximation of the Joint Spectral Function

```
Initialize $m$, $\overline{\lambda}$, and set $\tilde{\alpha}(\lambda) = 0$.
Compute lower bound $\tilde{\lambda}_1$ such that $\tilde{\lambda}_1 \leq \lambda_1$.
Compute eigenpairs of matrix pair $A'$ and $B$ with eigenvalues less than $\overline{\lambda} - \tilde{\lambda}_1$,
    $A'x'_j = \lambda'_j B x'_j$, $\|x'_j\|_B = 1$, $j = 1, \ldots, \overline{j}$.
for $j = 1, \ldots, \overline{j}$ do
    Perform $m$-steps of $B$-Lanczos with $A$, $B$, and $x'_j$, to get $T_m \in \mathbb{R}^{m \times m}$.
    Compute eigenpairs $T_m y_k = \theta_k y_k$, $y_k^T y_\ell = \delta_{k\ell}$, for $k, \ell = 1, \ldots, m$.
    $\tilde{\alpha}(\lambda) \leftarrow \tilde{\alpha}(\lambda) + \sum_{k=1}^{m} |(y_k, e_1)|^2 \delta\big(\lambda - (\theta_k + \lambda'_j)\big)$.
end for
```

# Chapter 5

# Joint Spectral Quantities and Semiconductor Applications

## 5.1 Modeling Random Alloys

The opto-electronic properties of semiconductors are governed by electric charge carrier distributions and their energy levels. Charge carriers in a semiconductor include electrons in the conductance band and holes in the valence band. In order to understand the quantum effects governing semiconductor behavior, we model carriers using the time-independent Schrödinger equation. Solving the Schrödinger eigenvalue problem for the electron and hole systems is a computationally intense exercise, and for many practical problems is outside of the capability of even the largest supercomputing clusters. Therefore, numerical devices which obviate the need for a full diagonalization of the electron and hole Hamiltonian are necessary. In this chapter we use the Lanczos process as just such a device.

In this chapter we focus on applying the Lanczos process to approximate joint spectral quantities corresponding to the ternary alloy indium gallium nitride (InGaN or $\mathrm{In}_X\mathrm{Ga}_{1-X}\mathrm{N}$ when specifying the indium fraction $X$). InGaN is a promising semiconductor material with many beneficial properties. Most important is the ability to tailor the bandgap to a wide range of energies based on the indium composition. InGaN alloys are used in many industrial applications, including green and blue light emitting diodes and lasers. When modeling InGaN, the random indium content of the material is critical [66], and accounts for a phenomena called localization wherein the charge carriers become concentrated in small regions of the domain. We follow the modeling paradigms of [30], which are outlined in the upcoming sections.

The fundamental equation we work with when modeling quantum mechanical effects of random alloys is the Schrödinger equation

$$-\frac{\hbar^2}{2}\nabla\cdot\left(\frac{1}{m}\nabla\psi\right)+V\psi=E\psi, \tag{5.1}$$

where $\psi$ is the wavefunction, $E$ the discrete energy level, $V$ the potential, $m$ the particle mass, and $\hbar$ the reduced Planck constant. The elliptic operator, $-\hbar^2/2\nabla\cdot\big(1/m\nabla\big)+V$, is referred to as the Hamiltonian, and the Schrödinger equation is an eigenvalue problem for the Hamiltonian. The eigenfunctions are referred to as wavefunctions, and the eigenvalues of the Hamiltonian represent discrete energy levels of the quantum system. Accordingly, in this chapter we use the term eigenvalue and energy interchangeably and similarly for the terms eigenfunction and wavefunction. Oftentimes, when a single particle is under consideration, the Hamiltonian is simplified to $-\hbar^2/(2m)\Delta+V$. However, in this Chapter we consider a generalization of the Schrödinger equation, referred to as the effective mass Schrödinger equation, in which the mass term is spatially varying. Therefore, keeping the reciprocal of the mass inside the divergence term is necessary.

### 5.1.1 Indium Fraction

When modeling InGaN alloys, we use a periodic cubic lattice in $d$ dimensions ($d=1,2$, or 3) with lattice spacing $a=2.833$ Å. At each lattice point, an InN or GaN cation is randomly placed using a random Bernoulli trial with probability of success (success meaning an InN cation is located at the lattice position) equal to $\overline{X}$. The value $0\leq\overline{X}\leq 1$ is the "bulk" indium content of the random alloy. Pure GaN corresponds to $\overline{X}=0$ and pure InN corresponds to $\overline{X}=1$. Once the InN and GaN are randomly distributed in the lattice, the spatially varying indium fraction, $X(x)$, is determined by a Gaussian averaging process. If we let the values $\{r_i\}_{i=1}^{n_\ell}$ represent the $n_\ell$ lattice coordinates and $\chi_i$, $i=1,\dots,n_\ell$, boolean indicators of an InN cation in lattice position $r_i$ ($\chi_i=1$ if InN is located at lattice position $r_i$ and 0 otherwise), the indium fraction at any point

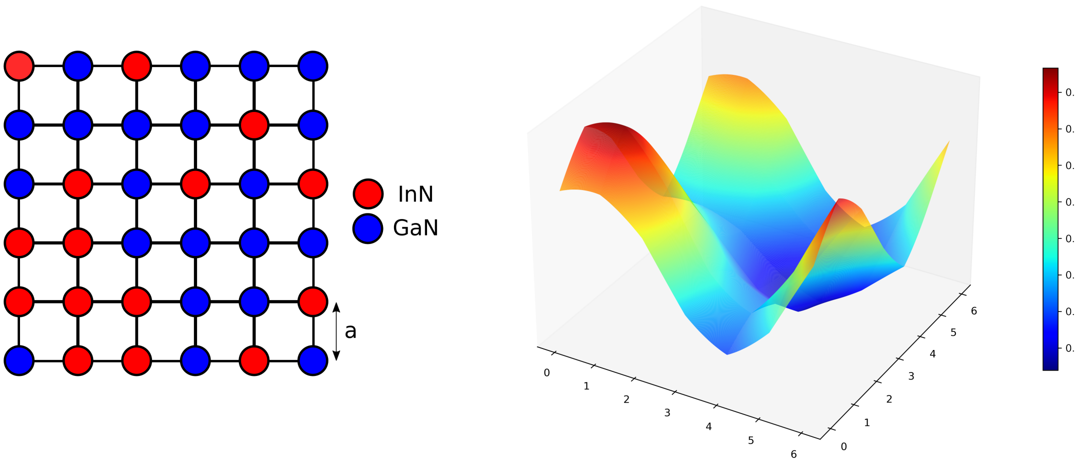


Figure 5.1: Depiction of two dimensional InGaN lattice with lattice spacing $a = 2.833$ Å (left) and corresponding periodic Gaussian averaged indium fraction $X(x)$ (right).

$x$ in the domain is defined as

$$X(x) = \frac{\sum_{i=1}^{n_\ell} \chi_i w(x, r_i)}{\sum_{i=1}^{n_\ell} w(x, r_i)}, \tag{5.2}$$

where $w(x, y) = \exp(-|x-y|^2/2(2a)^2)$ and $|x-y|$ denote the Euclidean distance between points $x$ and $y$. Note that periodicity must be taken into account when computing distances $|x - y|$. The value of twice the lattice spacing, i.e., $2a$, as the standard deviation in the Gaussian averaging is a modeling choice taken from [30].

An example of a small two dimensional InGaN lattice is shown in Figure 5.1. InN, shown in red, is useful for producing the infrared portion of the spectrum and GaN, shown in blue, is commonly used in blue light emitting diodes. When combined, InGaN alloys are capable of producing an array of colors, depending on the indium concentration. In Figure 5.1 a $6 \times 6$ lattice is shown with the bottom left position being $(0, 0)$ and the top right position as $(5a, 5a)$. The lattice then repeats periodically. In other words, the lattice positions $(ka, 0)$ are identical to $(ka, 6a)$ for $k = 0, \ldots, 5$. Similarly, the lattice positions $(0, ka)$ are identically $(6a, ka)$ for $k = 0, \ldots, 5$. The periodic nature

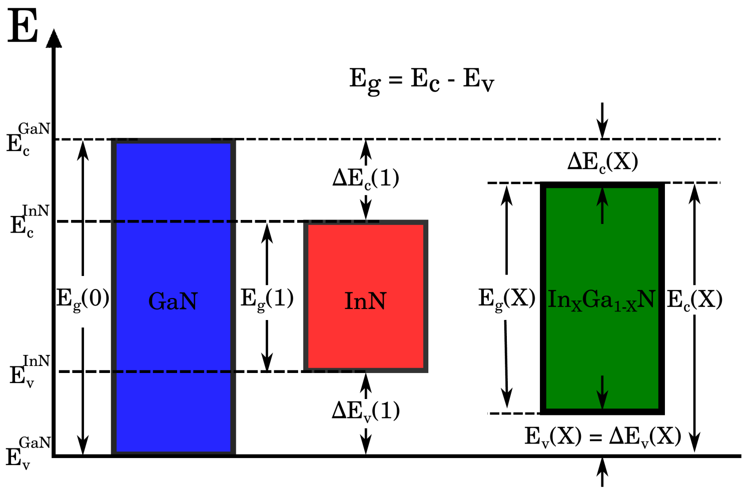


Figure 5.2: Bandgap of InGaN alloy.

of the lattice is apparent in the spatially varying indium fraction shown in Figure 5.1, and is a consequence of taking periodicity into account when computing the indium fraction using (5.2).

### 5.1.2 Bandgap, Conductance Band, and Valence Band

The distinctive feature of semiconductors is the bandgap energy, $E_g$, which is the difference in the discrete energy level between the conduction band, $E_c$, and the valence band, $E_v$. The valence band is the highest energy fully occupied orbital and the conductance band the lowest energy partially filled orbital. By alloying InN and GaN, engineers are able to modulate the bandgap for a desired purpose. Figure 5.2 illustrates the bandgap of $\mathrm{In}_X\mathrm{Ga}_{1-X}\mathrm{N}$ in reference to that of InN and GaN.

A simple method to estimate the bandgap of $\mathrm{In}_X\mathrm{Ga}_{1-X}\mathrm{N}$ alloys is to use Vegard's law which takes the convex combination of the bandgap of InN, $E_g^{\mathrm{InN}} = 0.61$ eV, and that of GaN, $E_g^{\mathrm{GaN}} = 3.437$ eV, i.e., to use $XE_g^{\mathrm{InN}} + (1-X)E_g^{\mathrm{GaN}}$. A better method is to use a quadratic correction to Vegard's law

$$E_g(x) = X(x)E_g^{\mathrm{InN}} + (1 - X(x))E_g^{\mathrm{GaN}} - \gamma X(x)(1 - X(x)), \tag{5.3}$$

where $\gamma = 1.4$ eV, is referred to as the bowing parameter [60].

Using the definition of the bandgap as the difference in energy level between the conductance band, $E_c$, and valence band, $E_v$, we write $E_g(X(x)) = E_c(X(x)) - E_v(X(x))$. In order to determine the potentials influencing the electrons in the conductance band and holes in the valence band, we must determine expressions for $E_c$ and $E_v$. Let $\Delta E_c(1)$ and $\Delta E_v(1)$ denote the change in energy levels from the conductance and valence band, respectively, of GaN and InN. Experimentally, one may determine that $\Delta E_c(1) \approx 2/3(E_g^{\text{GaN}} - E_g^{\text{InN}})$ and $\Delta E_v(1) \approx 1/3(E_g^{\text{GaN}} - E_g^{\text{InN}})$. Interpolating linearly to values of the indium fraction between zero and one, we use $\Delta E_c(X) = 2/3(E_g^{\text{GaN}} - E_g(X))$ and $\Delta E_v(X) = 1/3(E_g^{\text{GaN}} - E_v(X))$, which gives the conductance and valence band potentials for any value $x$ in the domain

$$E_c(x) = E_g^{\text{GaN}} - \frac{2}{3}\Big(E_g^{\text{GaN}} - E_g(x)\Big) \quad \text{and} \quad E_v(x) = \frac{1}{3}\Big(E_g^{\text{GaN}} - E_g(x)\Big). \tag{5.4}$$

### 5.1.3 Effective Mass

The last ingredient necessary to write the Schrödinger equations is the mass term. Because we are dealing with an alloy, we are unable to use the mass of InN carriers or GaN carriers. Intuitively, it makes sense to use the carrier mass of InN in regions of dense InN, and similarly for regions of dense GaN. The question is how to "interpolate" between the two masses in regions mixed with InN and GaN. In order to accomplish this, we investigate a model one dimensional problem.

For the sake of simplicity, suppose we are interested in determining the energy of a particle with constant mass $m^{(0)}$ under the influence of a constant potential $V_0$ in one dimension on the domain $(0, \pi)$. Solving the time-independent Schrödinger equation with zero Dirichlet boundary conditions gives energy levels

$$E_k^{(0)} = V_0 + \frac{\hbar^2}{2m^{(0)}}k^2, \quad k \in \mathbb{N}. \tag{5.5}$$

Note that the energies in (5.5) are just the eigenvalues of the Laplacian (scaled by $\hbar^2/(2m^{(0)})$) plus the value of the constant potential. Similarly, considering the energy level of a different particle of mass $m^{(1)}$, under the influence of the same constant potential energy, results in energy levels

$$E_k^{(1)} = V_0 + \frac{\hbar^2}{2m^{(1)}}k^2, \quad k \in \mathbb{N}. \tag{5.6}$$

Taking the convex combination of the energies (5.5) and (5.6) we find

$$E_k^{(\theta)} := (1-\theta)E_k^{(0)} + \theta E_k^{(1)} = V_0 + \frac{\hbar^2}{2m^{(\theta)}}k^2, \quad k \in \mathbb{N}, \tag{5.7}$$

where,

$$m^{(\theta)} := \left(\frac{1-\theta}{m^{(0)}} + \frac{\theta}{m^{(1)}}\right)^{-1}, \tag{5.8}$$

for some $0 \leq \theta \leq 1$. Results (5.7) and (5.8) indicate that when determining energy levels by interpolating between two species, the resulting energy is determined by a particle which has mass equal to the harmonic mean of the two masses $m^{(0)}$ and $m^{(1)}$.

Applying the above reasoning to the case of an InGaN alloy, we use as masses for the electrons and holes

$$m_e(x) = \left(\frac{1-X(x)}{m_e^{\mathrm{GaN}}} + \frac{X(x)}{m_e^{\mathrm{InN}}}\right)^{-1} \quad \text{and} \quad m_h(x) = \left(\frac{1-X(x)}{m_h^{\mathrm{GaN}}} + \frac{X(x)}{m_h^{\mathrm{InN}}}\right)^{-1}, \tag{5.9}$$

where $m_e^{\mathrm{GaN}}$ and $m_e^{\mathrm{InN}}$ are the electron masses for GaN and InN respectively and $m_h^{\mathrm{GaN}}$ and $m_h^{\mathrm{InN}}$ are the hole masses of GaN and InN respectively. We refer to $m_e$ and $m_h$ as the effective electron and hole mass respectively. The carrier masses for GaN and InN are all expressible in terms of the electron rest mass, $m_e^0$,

$$m_e^{\mathrm{GaN}} = 0.21m_e^0, \;\; m_e^{\mathrm{InN}} = 0.07m_e^0, \;\; m_h^{\mathrm{GaN}} = 1.87m_e^0, \;\; m_h^{\mathrm{InN}} = 1.61m_e^0, \tag{5.10}$$

where the electron rest mass is approximately 511 keV/$c^2$ with $c$ being the speed of light.

Finally, we remark that (5.9) should be considered as a first order approximation to the carrier masses in an InGaN alloy. Higher order approximations to the effective mass are discussed in [20].

### 5.1.4 Effective Mass Schrödinger Equation

With the potentials defined in (5.4) and the effective masses defined in (5.9), we have all terms necessary to write the Schrödinger equation satisfied by the electrons and holes in a random alloy:

$$\begin{aligned} -\frac{\hbar^2}{2}\nabla\cdot\left(\frac{1}{m_e}\nabla\psi^e\right)+E_c\psi^e=E^e\psi^e,\\ \frac{\hbar^2}{2}\nabla\cdot\left(\frac{1}{m_h}\nabla\psi^h\right)+E_v\psi^h=E^h\psi^h, \end{aligned} \tag{5.11}$$

where the electron energies and wavefunctions are $E^e$ and $\psi^e$ respectively, and the hole energies and wavefunctions are $E^h$ and $\psi^h$ respectively. Note the change in sign in front of the second order term in the equation satisfied by the holes. The second order term represents the kinetic energy of the carriers and the zeroth order term represents the potential energy. With the potentials and sign conventions of (5.11), the electron energies are positive (the second order term is positive definite and the conductance potential is positive) and larger eigenvalues correspond to higher energy quantum states. On the other hand, the hole energies can be positive and negative. More energetic states for the holes correspond to negative eigenvalues of increasing magnitude.

Note that the electron and hole energies are measured with respect to the valence band energy of GaN. As shown in Figure 5.2, we made an arbitrary modeling choice and defined the reference energy, or the zero on the energy scale, to be the valence band energy of GaN. We could just as easily defined the valence band energy of InN to be zero. Independent of the arbitrary reference energy, of fundamental importance are the energy differences $E_i^e - E_j^h$, $i, j = 1, 2, \ldots$, which represents the energy required to excite an electron to the conductance band and create an electron-hole pair. Next, we discuss a different normalization convention which is more in line with the theory of joint spectral functions outlined in the previous Chapter.

### 5.1.5 Normalization Convention

As mentioned in Section 5.1.4, the fundamental quantity of interest is the energy difference of the electrons in the conductance band and the holes in the valence band. In this section we reformulate the effective mass Schrödinger equations (5.11) in a way that leaves the wavefunctions unaltered, but modifies the energies. As we will see, this change still allows us to compute the energy differences of interest.

Define a new conductance and valence band potentials as

$$V_c(x) = \frac{2}{3}E_g(x) \quad \text{and} \quad V_v(x) = \frac{1}{3}E_g(x). \tag{5.12}$$

Note that relative to the potentials $E_c$ and $E_v$ defined in (5.4), the new potentials satisfy

$$V_c = E_c - \frac{1}{3}E_g^{\text{GaN}} \quad \text{and} \quad V_v = \frac{1}{3}E_g^{\text{GaN}} - E_v.$$

In other words, $V_c$ is simply a shift down of the potential $E_c$ by $1/3E_g^{\text{GaN}}$, while $V_v$ is a sign reversal of $E_v$, followed by a shift up by $1/3E_g^{\text{GaN}}$. Using the potentials $V_c$ and $V_v$, we are interested in solutions to the Schrödinger equation

$$\begin{aligned} -\frac{\hbar^2}{2}\nabla\cdot\left(\frac{1}{m_e}\nabla\psi^e\right) + V_c\psi^e &= \lambda^e\psi^e, \\ -\frac{\hbar^2}{2}\nabla\cdot\left(\frac{1}{m_h}\nabla\psi^h\right) + V_v\psi^h &= \lambda^h\psi^h. \end{aligned} \tag{5.13}$$

Note that our notational use of $\psi^e$ and $\psi^h$ in both (5.11) and (5.13) is justified, as they are equivalent. Both second order terms in (5.13) are positive definite, as opposed to (5.11), where the second order term in the hole equation was negative definite. Also, because both potentials $V_c$ and $V_v$ are positive, the energies $\lambda^e$ and $\lambda^h$ are positive. Furthermore, it is straightforward to see that the energies of systems (5.11) and (5.13) are related by

$$E_i^e = \frac{1}{3}E_g^{\text{GaN}} + \lambda_i^e \quad \text{and} \quad E_j^h = \frac{1}{3}E_g^{\text{GaN}} - \lambda_j^h, \quad i,j \in \mathbb{N}. \tag{5.14}$$

Therefore, the fundamental quantity of interest, namely the energy required to excite an electron to the conductance band, is given by

$$E_i^e - E_j^h = \lambda_i^e + \lambda_j^h, \quad i,j \in \mathbb{N}. \tag{5.15}$$

Moving forward, we will work with numerical discretizations of (5.13). Obviously, if one is interested in the energies corresponding to the system (5.11), a simple application of (5.14) transitions from one convention to the other.

### 5.1.6 Nondimensionalization

When discretizing the system (5.13), it is convenient to first nondimensionalize the system. Nondimensionalization involves choosing a standard length and mass to measure all others in relation to. To illustrate, we nondimensionalize the system (5.1), with application to the systems (5.13) straightforward.

First, we choose the InGaN lattice spacing constant, $a = 2.833\text{Å}$, as a characteristic length and define a new dimensionless length scale as $\hat{x} = x/a$. The InGaN lattice, $a\mathbb{Z}^d$, becomes $\mathbb{Z}^d$, and the new eigenfunctions we are interested in computing are $\hat{\psi}(\hat{x}) := \psi(x)$. This change of variables modifies the spatial derivatives according to $\partial/\partial\hat{x}_i = a\partial/\partial x_i$, $i = 1, \ldots, d$, and so the gradient becomes $\nabla = a^{-1}\hat{\nabla}$.

Next, we choose the electron rest mass, $m_e^0 \approx 511$ keV/c$^2$, as the characteristic mass. This is a natural choice since all relevant masses (5.10) are already expressed in terms of the electron rest mass.

With the characteristic length and mass decided, we choose the characteristic (reference) energy as

$$E_r = \frac{\hbar^2}{2m_e^0 a^2} \approx 0.4747 \text{ eV}. \tag{5.16}$$

Using the characteristic length, mass, and energy we can nondimensionalize the remaining variables in (5.1) as

$$\hat{V}(\hat{x}) = \frac{V(x)}{E_r}, \qquad \hat{m}(\hat{x}) = \frac{m(x)}{m_e^0}, \qquad \hat{E} = \frac{E}{E_r}. \tag{5.17}$$

With all terms nondimensionalized, we can rewrite (5.1) in terms of dimensionless quantities as

$$-\hat{\nabla} \cdot \left(\frac{1}{\hat{m}}\hat{\nabla}\hat{\psi}\right) + \hat{V}\hat{\psi} = \hat{E}\hat{\psi}. \tag{5.18}$$

When performing computations, the discretization of (5.18) is used. Then, when reporting the results, the energies are converted back to physical units using $E = \hat{E}E_r$, with the reference energy, $E_r$, given by (5.16). On the other hand, when referencing spatial variables, e.g., plotting potentials or wavefunctions, we report results with respect to the transformed variable, $\hat{x}$, rather than transitioning to physical units. For example, when performing one dimensional computations on a lattice with 5001 cation sites, we report results using the domain $\Omega = [0, 5000]$, rather than transitioning to physical units on the domain $[0, L]$, where $L = 5000a \approx 1.4$ $\mu$m. Obviously, we can easily transition between one convention and the other using the lattice spacing $a$.

### 5.1.7 Spectral and Joint Spectral Quantities

The fundamental optoelectronic properties of InGaN semiconductors are determined by spectral quantities defined in terms of the eigenpairs of the effective mass Schrödinger equations (5.13). These are the densities of states,

$$\phi_e(\lambda) = \sum_{i=1}^{\infty} \delta(\lambda - \lambda_i^e) \quad \text{and} \quad \phi_h(\lambda) = \sum_{j=1}^{\infty} \delta(\lambda - \lambda_j^h), \tag{5.19}$$

in addition to the joint densities of states and absorption curve (sometimes called the absorption coefficient),

$$J(\lambda) = \sum_{i,j=1}^{\infty} \delta\big(\lambda - (\lambda_i^e + \lambda_j^h)\big) \quad \text{and} \quad \alpha(\lambda) = \sum_{i,j=1}^{\infty} |\langle \psi_i^e, \psi_j^h \rangle|^2 \delta\big(\lambda - (\lambda_i^e + \lambda_j^h)\big), \tag{5.20}$$

where $\langle \cdot\,, \cdot \rangle$ represents the $L^2(\Omega)$ inner product. Note the resemblance of the absorption curve and the joint spectral function defined in the previous chapter.

The quantities in (5.19) and (5.20) involve the solution of infinite dimensional problems, and so are out of reach except in the simplest situations (we describe such a contrivance in Section 5.3). Hence, in order to understand the properties of an InGaN alloy, we must first use a robust discretization in order to apply the Lanczos process. We use the standard finite element method, and discuss the discretization in Section 5.2.

### 5.1.8 Regularizing the Dirac Delta

In many instances, we would like to visualize spectral quantities (if known), and the Lanczos approximation of the spectral quantities, both of which involve linear combinations of Dirac distributions. Oftentimes, the Dirac distribution, concentrated at a value $E_0$, is thought of as a function with unit integral which is zero everywhere, except at the value $E_0$, where it is infinite. This is illustrated in Figure 5.3 (left). Although this is not technically correct, since the Dirac distribution is not defined in a pointwise

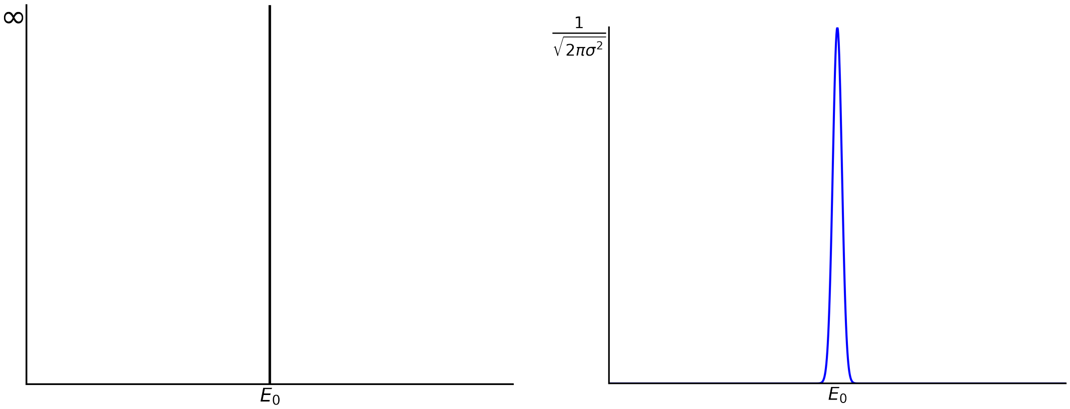


Figure 5.3: Visualization of the Dirac distribution concentrated at the value $E_0$ (left) and a Gaussian with standard deviation $\sigma$ and mean $E_0$ (right).

sense, we can use a smooth approximation which, in an appropriate limit, also has these properties. Let $\sigma$ be a small positive parameter. We “regularize” the Dirac distribution by replacing it with a Gaussian of standard deviation $\sigma$

$$\delta_\sigma(E) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{E^2}{2\sigma^2}}. \tag{5.21}$$

The regularized Dirac distribution, $\delta_\sigma(E - E_0)$, seen in Figure 5.3 (right), is an approximation to $\delta(E - E_0)$ in the sense that the following hold:

(i) $\int\limits_{-\infty}^{+\infty} \delta_\sigma(E)dE = 1$ for all $\sigma > 0$.

(ii) $\lim\limits_{\sigma\to 0} \delta_\sigma(E) = \begin{cases} 0, & E \neq 0, \\ \infty, & E = 0. \end{cases}$

(iii) $\lim\limits_{\sigma\to 0} \int\limits_{-\infty}^{+\infty} \delta_\sigma(E)f(E)dE = \int\limits_{-\infty}^{+\infty} \delta(E)f(E)dE$ for all $f \in C_0^\infty(\mathbb{R})$.

The first fact is easily established using polar coordinates. For the second, the case $E = 0$ follows directly from the definition (5.21). For the case $E \neq 0$, applying the Squeeze Theorem to $0 \leq \delta_\sigma(E) \leq \sqrt{2\pi\sigma^2}/(2\pi\sigma^2 + \pi E^2)$ (follows from $e^x \geq 1 + x$) as $\sigma \to 0$ gives the desired result. For the third fact, using the change of variables,

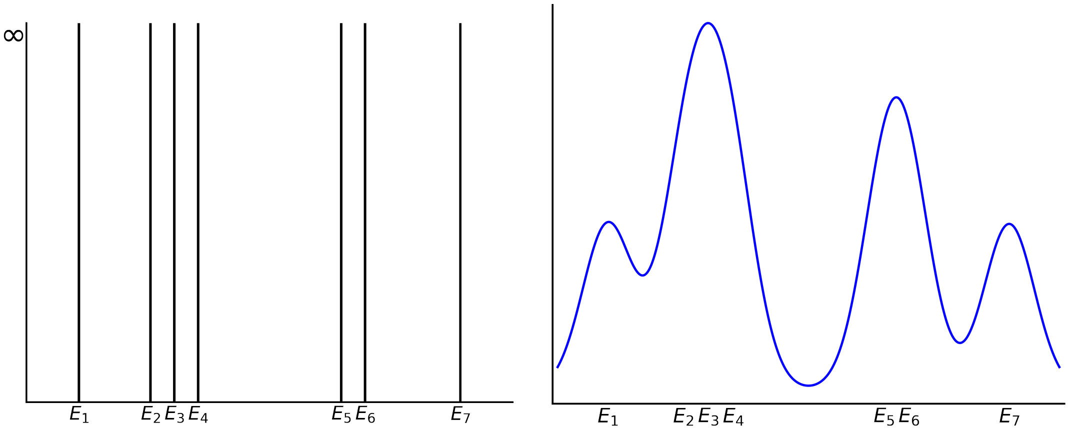


Figure 5.4: Visualization of the density of states, $\sum_{i=1}^{7} \delta(E - E_i)$, using "exact" Dirac mass (left) and the regularization using Gaussians, $\sum_{i=1}^{7} \delta_\sigma(E - E_i)$, for some regularization parameter $\sigma > 0$ (right).

$y = E/\sigma$, we have

$$\int\limits_{-\infty}^{+\infty} \delta_\sigma(E) f(E) dE = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{+\infty} e^{-y^2/2} f(\sigma y) dy, \tag{5.22}$$

where $f \in C_0^\infty(\mathbb{R})$ is arbitrary. Noticing that the integrand in the right-hand side of (5.22) is dominated by $e^{-y^2/2}\|f\|_\infty$ and $\int_{-\infty}^{+\infty} e^{-y^2/2}\|f\|_\infty dy < \infty$, we can use the Lebesgue Dominated Convergence Theorem to get

$$\begin{aligned} \lim_{\sigma\to 0} \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{+\infty} e^{-y^2/2} f(\sigma y) dy &= \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{+\infty} \lim_{\sigma\to 0} e^{-y^2/2} f(\sigma y) dy, \\ &= \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{+\infty} e^{-y^2/2} f(0) dy, \\ &= f(0), \\ &= \int\limits_{-\infty}^{+\infty} \delta(E) f(E) dE. \end{aligned} \tag{5.23}$$

Together, (5.22) and (5.23) give (iii).

In order to understand how regularization through the use of Gaussians influences spectral quantities, we use the density of states as an example. Let $E_i$, for $i = 1, \ldots, 7$, denote seven positive energy levels, and suppose we want to visualize $\sum_{i=1}^{7} \delta(E - E_i)$. Without regularization, the density of states can be visualized as seven spikes, each of which has unit area, infinite height, and infinitesimally small width. This is shown in Figure 5.4 (left). Replacing the Dirac distributions with the Gaussian (5.21) of some finite width results in Figure 5.4 (right). We see that the regularized density of states is largest where there is a cluster of eigenvalues (the region around $E_2$, $E_3$, and $E_4$), and smallest where there is a large gap between the eigenvalues (between $E_4$ and $E_5$).

An important property of the Dirac mass is its unit area

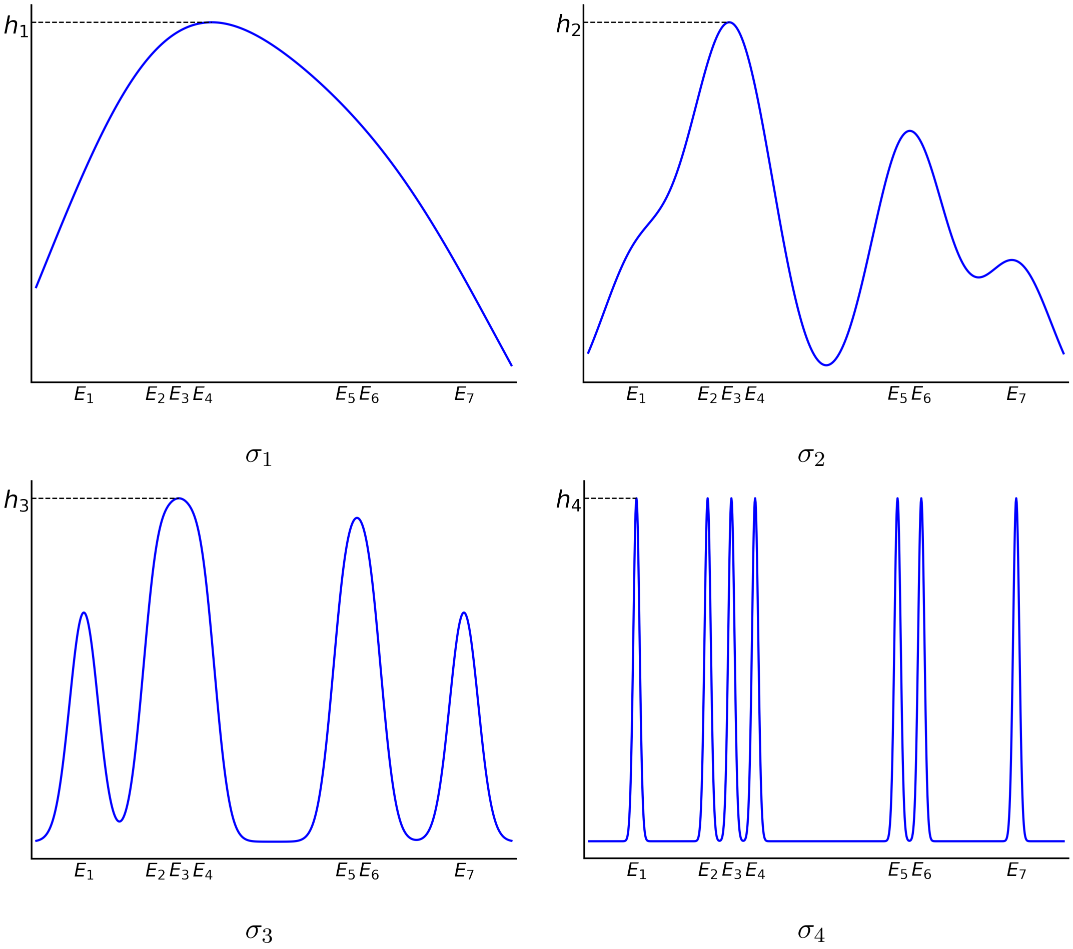
$$\int\limits_{E_0-\epsilon}^{E_0+\epsilon} \delta(E - E_0) dE = 1, \tag{5.24}$$

for any $\epsilon > 0$. By replacing the Dirac distribution with a Gaussian, $\delta_\sigma(E - E_0)$, the equality becomes and approximation. How well the approximation holds obviously depends on the ratio $\epsilon/\sigma$. If $\epsilon/\sigma \ll 1$, then we expect (5.24) to fail catastrophically, while if $\epsilon/\sigma \gg 1$, then (5.24) should hold closely. Depending on the precision with which we require (5.24) to hold will dictate the choice of $\sigma$. Figure 5.5 shows the density of states for four different values of $\sigma$. We see that as $\sigma$ increases, the Gaussian distributions blur together, while for smaller values of $\sigma$ we can see each individual Gaussian. In other words, larger $\sigma$ corresponds to less detail, while smaller $\sigma$ corresponds to more specificity. The correct choice of $\sigma$ depends on the level of precision with which we want to emulate the Dirac distribution.

There are many ways to approximate the Dirac distribution, and we have made an arbitrary choice in using (5.21). Other commonly used choices include

$$\tilde{\delta}_\sigma(E) = \begin{cases} \frac{1}{2\sigma}, & |E| < \sigma \\ 0, & \text{otherwise} \end{cases}, \quad \text{or} \quad \tilde{\delta}_\sigma(E) = \begin{cases} \frac{C}{\sigma} \exp\left(\frac{\sigma^2}{|E|^2 - \sigma^2}\right), & |E| < \sigma \\ 0, & \text{otherwise} \end{cases},$$

where the constant $C$ is a normalization factor. Note that both approximations, $\tilde{\delta}_\sigma$, have finite support, as opposed to using a Gaussian, $\delta_\sigma$, with infinite support. More generally,



$\sigma_1$ $\sigma_2$

$\sigma_3$ $\sigma_4$

Figure 5.5: Regularized density of states corresponding to different levels of regularization $\sigma_1 > \sigma_2 > \sigma_3 > \sigma_4$. The maximum height of the graphs satisfies $h_1 < h_2 < h_3 < h_4$.

finite width approximations of the Dirac distribution are known as approximations of unity (the Dirac distribution being the identity with respect to convolution) [63]. The specific choice of approximation will alter how the spectral quantities appear graphically.

Note that thus far, we have been discussing visualization of the *exact* density of states, and not an approximation. When discussing the approximation, there are other factors to consider. Namely, smaller values of $\sigma$ will require us to perform more Lanczos iterations since we are, in a sense, trying to recreate the exact position of each energy. On the other hand, larger values of $\sigma$ allow for easier approximation, since we are only trying to emulate the bulk properties of the spectral quantity.

Lastly we remark that if $E$ has units of energy, then $\delta(E)$ has units of reciprocal energy. This follows because multiplying $\delta(E)$ by an infinitesimal energy $dE$ and summing (integrating), gives a dimensionless constant. Also, the parameter $\sigma$ will also have units of energy since it measures a standard deviation, i.e., width, in energy space. The units of energy we use in this chapter are electron volts, or $e$V, and so a value $\sigma = 0.01$ corresponds to 10 m$e$V.

### 5.1.9 On the Choice of $\sigma$ and $m$

Previously, we discussed the choice of regularization parameter, $\sigma$, with respect to the exact spectral quantity. Here we discuss the impact of $\sigma$ with respect to the Lanczos approximation and the choice of the Krylov parameter $m$, which represents the number of Gram–Schmidt orthogonalization steps. This is an important and nuanced issue. In order to understand the relationship between $\sigma$ and $m$ we examine a simple example illustrating the finer points.

We start with a matrix related to the discretization of the one dimensional Laplace operator using central differences (or finite elements with mass lumping)

$$A = \begin{pmatrix} 2 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{pmatrix} \in \mathbb{R}^{n\times n}. \tag{5.25}$$

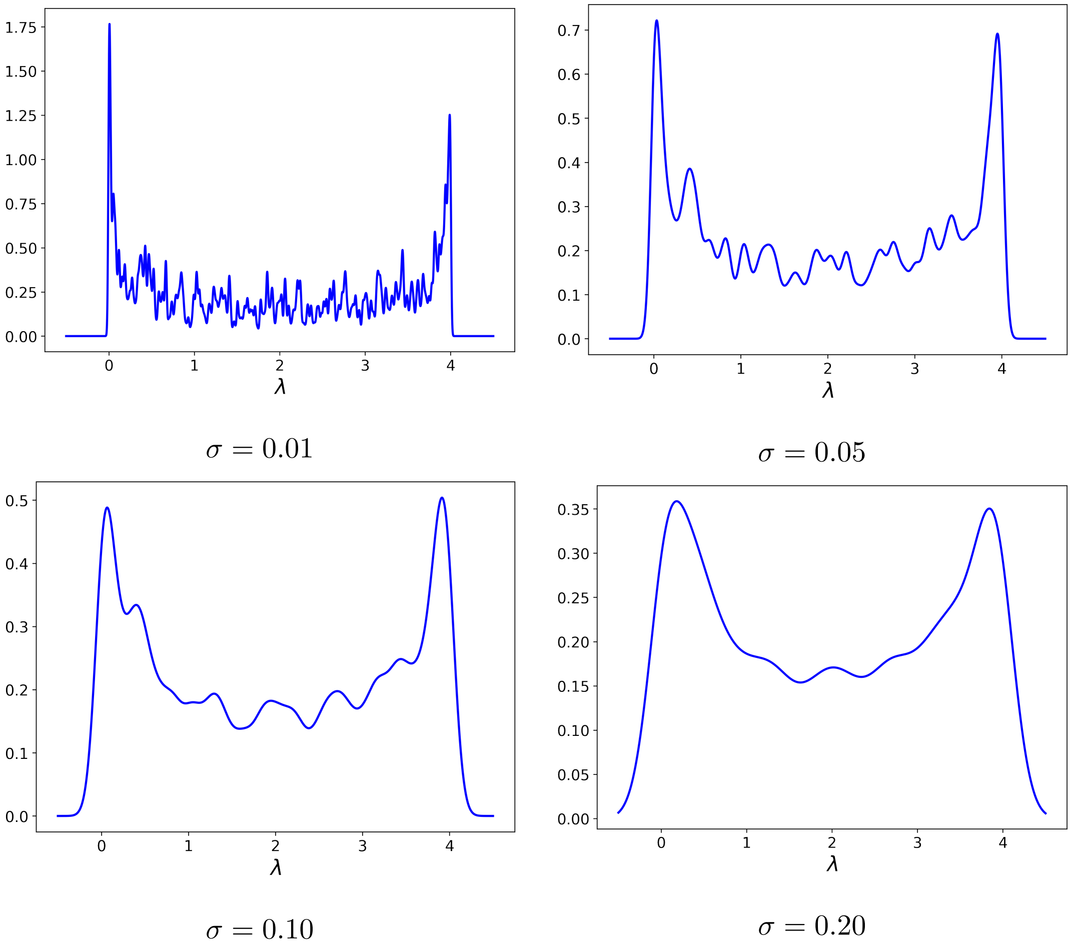
The matrix $A$ has eigenvalues

$$\lambda_i = 4\sin^2\left(\frac{i\pi}{2(n+1)}\right), \quad i = 1, \ldots, n, \tag{5.26}$$

and (unnormalized) eigenvectors which are the columns of the matrix with entries

$$\sin\left(\frac{ij\pi}{n+1}\right), \quad i, j = 1, \ldots, n.$$

Denote the normalized eigenvectors of $A$ as $x_i$, $i = 1, \ldots, n$. Choosing a vector $v$ with entries drawn from the standard normal distribution which has been normalized, the



$\sigma = 0.01$

$\sigma = 0.05$

$\sigma = 0.10$

$\sigma = 0.20$

Figure 5.6: Spectral function for the matrix defined in (5.25) and a vector with entries drawn from $\mathcal{N}(0,1)$ which has been normalized.

spectral function

$$s_\sigma(\lambda) = \sum_{i=1}^{n} |(x_i, v)|^2 \delta_\sigma(\lambda - \lambda_i),$$

is shown in Figure 5.6 for values of $\sigma$ ranging from 0.01 to 0.20 and $n = 2000$.

We focus on the regularization parameter $\sigma = 0.05$, and investigate the Lanczos approximation of $s_\sigma$ for different values of the Krylov space parameter $m$. Denote the regularized approximation to $s_\sigma$ from the Lanczos process by

$$\tilde{s}_\sigma(\lambda) = \sum_{j=1}^{m} |(y_j, e_1)|^2 \delta_\sigma(\lambda - \theta_j),$$

where the $\theta_j \in \mathbb{R}$ (Ritz values) and $y_j \in \mathbb{R}^m$ ($\|y_j\| = 1$) are the eigenpairs of the Lanczos partial tridiagonalization of $A$ with starting vector $v$. First, we purposely choose a value of $m$ which is insufficient, i.e., too small. Figure 5.7 (left) shows the Lanczos approximation of the spectral function depicted in Figure 5.6 for Krylov dimension $m = 25$ and $\sigma = 0.05$. The Ritz values, $\theta_j$, $j = 1, \ldots, m$ are shown on the $x$-axis with dotted vertical lines up to the Lanczos approximation $\tilde{s}_\sigma$. Also shown in Figure 5.7 (right) is the error in the Lanczos approximation, $\tilde{s}_\sigma - s_\sigma$, along with the Ritz values. Two important details about the Lanczos process are illustrated in Figure 5.7. First, the spectral function is approximated well at the extremities, and poorly in the interior. This is related to the fact that Gaussian quadrature nodes cluster at the endpoints of the interval of integration. Because the quadrature nodes are more dense at the endpoints of the interval, the Lanczos process matches the spectral function closely there. The second detail to be noted is the Lanczos approximation oscillates around the exact spectral function in the interior where the gap between Ritz values is largest. This is easily explained by the moment matching property, which requires

$$\sum_{i=1}^{n} |(x_i, v)|^2 \lambda_i^\ell = \sum_{j=1}^{m} |(y_j, e_1)|^2 \theta_j^\ell, \qquad \ell = 0, 1, \ldots, 2m-1. \tag{5.27}$$

When $m/n \ll 1$, the weights, $|(y_j, e_1)|^2$, must be relatively large in order for (5.27) to be satisfied. Due to this overcompensation, the Lanczos approximation is larger than the spectral function at the Ritz values. This is clearly be seen in Figure 5.7 (right), where the oscillations in the error have large positive amplitudes at the Ritz values, and large negative amplitudes midway between Ritz values.

With these two details, one, that the Lanczos approximation oscillates around the exact solution, and two, that the Lanczos approximation converges from the extremities, it is straightforward to devise a strategy to approximate any one spectral function accurately for a given value of $\sigma$. Namely, by continuing with the Lanczos process

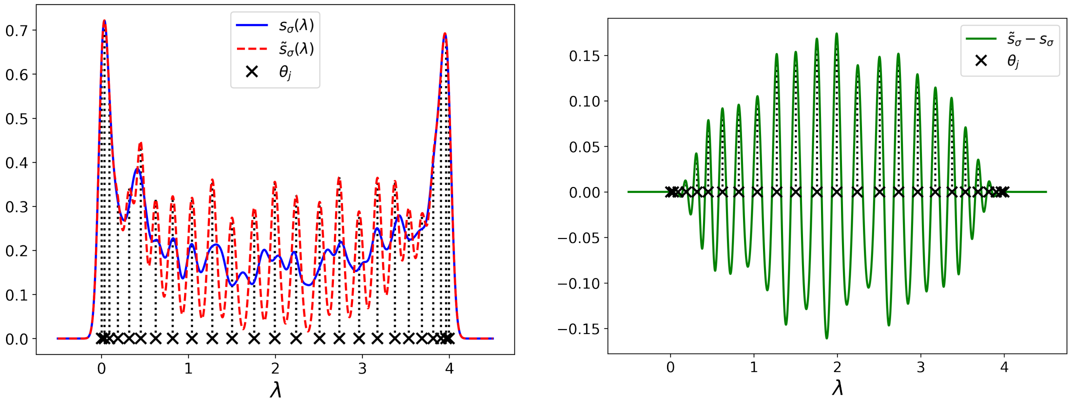


Figure 5.7: Lanczos approximation to a spectral function for Krylov dimension $m = 25$ and regularization parameter $\sigma = 0.05$.

until the difference in Ritz values is small enough, to be made precise momentarily, we can determine how accurately we have approximated a given spectral function. How small the gap in Ritz values needs to be depends on the value $\sigma$. For smaller values of $\sigma$, we need Ritz values (quadrature nodes) to be closer together in order to avoid the oscillations seen in Figure 5.7. Therefore, it makes sense to choose $\tau \in \mathbb{R}$, a multiple of $\sigma$, and continue with the Lanczos process until the gap between Ritz values is smaller than $\tau$.

Using this methodology, Table 5.1 shows the results for $\tau = 2.5\sigma = 0.125$. The value $\overline{m}$ is chosen to be the largest integer for which $\max_{1\leq i\leq \overline{m}-1}|\theta_i - \theta_{i+1}| < \tau$, and then $\bar{\theta}$ is set equal to $\theta_{\overline{m}}$. In other words, because the Ritz values $\theta_1 \leq \ldots \leq \theta_{\overline{m}} = \bar{\theta}$ are clustered sufficiently close together, we can trust the Lanczos approximation to the spectral function for $\lambda \leq \bar{\theta}$. The error, $\|s_\sigma - \tilde{s}_\sigma\|_{L^\infty(0,\bar{\theta})} = \sup_{\lambda\in(0,\bar{\theta})}|s_\sigma(\lambda) - \tilde{s}_\sigma(\lambda)|$, shown in the final column of Table 5.1, is seen to be stable for this fixed value of $\tau$.

| $m$ | $\overline{m}$ | $\bar{\theta}$ | $\|s_\sigma - \tilde{s}_\sigma\|_{L^\infty(0,\bar{\theta})}$ |
|---|---|---|---|
| 20 | 3 | 0.15 | $3.23 \times 10^{-2}$ |
| 30 | 7 | 0.45 | $3.55 \times 10^{-2}$ |
| 40 | 11 | 0.64 | $1.01 \times 10^{-2}$ |
| 50 | 16 | 0.86 | $2.91 \times 10^{-3}$ |
| 60 | 60 | 4.00 | $3.70 \times 10^{-3}$ |

Table 5.1: Uniform norm of error in Lanczos approximation to spectral function for $\tau = 2.5\sigma$ and $\sigma = 0.05$.

Similar results occur for smaller values of $\tau$, and can bee seen in Table 5.2 for $\tau = 1.5\sigma = 0.075$. We see that the error, $\|s_\sigma - \tilde{s}_\sigma\|_{L^\infty(0,\bar{\theta})}$, for tolerance $\tau = 1.5\sigma$ is a few orders of magnitude smaller than for $\tau = 2.5\sigma$. Again, the uniform error in the interval $(0, \bar{\theta})$ is quite stable. Clearly, larger values $\tau$ require few iterations of the Lanczos process (smaller values of Krylov parameter $m$), while smaller $\tau$ (tighter tolerances) requires larger values of $m$, and hence is more computationally intensive.

| $m$ | $\overline{m}$ | $\bar{\theta}$ | $\|s_\sigma - \tilde{s}_\sigma\|_{L^\infty(0,\bar{\theta})}$ |
|---|---|---|---|
| 60 | 15 | 0.54 | $1.76 \times 10^{-4}$ |
| 70 | 22 | 0.85 | $1.97 \times 10^{-4}$ |
| 80 | 26 | 0.92 | $3.28 \times 10^{-5}$ |
| 90 | 37 | 1.41 | $2.68 \times 10^{-5}$ |
| 100 | 100 | 4.00 | $7.09 \times 10^{-6}$ |

Table 5.2: Uniform norm of error in Lanczos approximation to spectral function for $\tau = 1.5\sigma$ and $\sigma = 0.05$.

By considering the gap between Ritz values in relation to the regularization parameter $\sigma$, we are able able to accurately determine when to stop the Lanczos process. When there are large gaps, relative to $\sigma$, the Lanczos approximation will oscillate about the exact spectral function as in Figure 5.7. Because solving for the eigenvalues of small symmetric tridiagonal matrices is efficient, we can inexpensively ensure the Lanczos process produces an accurate approximation to the spectral function by terminating the Lanczos algorithm when the gap between Ritz values is smaller than $\tau$ for a properly chosen $\tau \in \mathbb{R}$. Smaller values of $\tau$ correspond to tighter tolerances, more Lanczos iterations, and a more accurate approximation. Conversely, larger values of $\tau$ require fewer iterations and produce a less accurate approximation.

## 5.2 Finite Element Discretization

In this section we use the standard $H^1$-conforming Lagrange finite elements to discretize the effective mass Schrödinger equation [9, 12, 11]. We first write Equation (5.18) in weak form, which is the natural place to begin the finite element method. Multiplying (5.1) by a smooth function $\varphi$, and integrating by parts over a domain $\Omega \subset \mathbb{R}^d$, we have

$$\int\limits_{\Omega} \left( \frac{1}{m} \nabla\psi \cdot \nabla\varphi + V\psi\varphi \right) - \int\limits_{\partial\Omega} \frac{1}{m} \frac{\partial\psi}{\partial\nu} \varphi = \lambda \int\limits_{\Omega} \psi\varphi, \tag{5.28}$$

where $\nu$ is the outward unit normal vector on $\partial\Omega$ and $\partial\psi/\partial\nu = \nabla\psi \cdot \nu$. Assuming periodic boundary conditions, the boundary term vanishes. Let $H^1_{\text{per}}(\Omega)$ denote the subset of $H^1(\Omega) = \{u \in L^2(\Omega) \mid \partial u/\partial x_i \in L^2(\Omega) \text{ for } i = 1, \dots d\}$ satisfying periodic boundary conditions. Writing (5.28) in terms of bilinear operators, we are solving for the energies $\lambda \in \mathbb{R}$ and wavefunctions $\psi \in H^1_{\text{per}}(\Omega)$, such that

$$a(\psi, \varphi) = \lambda \langle \psi, \varphi \rangle \quad \text{for all} \quad \varphi \in H^1_{\text{per}}(\Omega), \tag{5.29}$$

where $a : H^1(\Omega) \times H^1(\Omega) \to \mathbb{R}$ is given by

$$a(\psi, \varphi) = \int\limits_{\Omega} \left( \frac{1}{m} \nabla\psi \cdot \nabla\varphi + V\psi\varphi \right). \tag{5.30}$$

Let $\mathcal{T}$ denote a shape regular conforming triangulation of the domain $\Omega$, and for a fixed natural number $p$, define the finite element space

$$X_p = \{f \in C^0(\overline{\Omega}) \mid f|_T \in \mathscr{P}_p(T) \text{ for all } T \in \mathcal{T}\} \cap H^1_{\text{per}}(\Omega). \tag{5.31}$$

Posing the infinite dimensional problem (5.29) over the finite dimensional space $X_p$ using the standard Galerkin method, the problem becomes: find $\tilde{\psi} \in X_h$ and $\tilde{\lambda} \in \mathbb{R}$ such that

$$a(\tilde{\psi}, \varphi) = \tilde{\lambda} \langle \tilde{\psi}, \varphi \rangle \quad \text{for all} \quad \varphi \in X_h. \tag{5.32}$$

Choosing a basis for the function space $X_p$, $\{\varphi_k\}_{k=1}^n$ say, the finite dimensional problem (5.32) becomes

$$Ax = \lambda Bx, \tag{5.33}$$

where $A_{ij} = a(\varphi_j, \varphi_i)$ is the stiffness matrix and $B_{ij} = \langle \varphi_j, \varphi_i \rangle$ is the mass matrix. The mass matrix is symmetric positive definite, and assuming the potential is positive, the stiffness matrix is as well. Therefore, the eigenvalues, or energies, of the system (5.33) are positive. The eigenvectors of the system (5.33) are the coefficients of the eigenfunctions $\tilde{\psi}$ in terms of the basis $\{\varphi_k\}_{k=1}^n$.

When performing the finite element discretization for the systems (5.13), we arrive at two generalized eigensystems, one for the electrons and another for the holes,

$$A^e x_i^e = \lambda_i^e B x_i^e \quad \text{and} \quad A^h x_j^h = \lambda_j^h B x_j^h. \tag{5.34}$$

Note that we use the same notation in (5.34) for the energies of the discretized system, $\lambda^e$ and $\lambda^h$, as we did for the exact energies in (5.13). This is done to avoid a deluge of extra tildes throughout this chapter. Both electron and hole eigenvectors are assumed to be $B$-orthonormalized, i.e., $(x_i^e, x_j^e)_B = \delta_{ij}$ and $(x_i^h, x_j^h)_B = \delta_{ij}$ where $(\,\cdot\,,\,\cdot\,)_B$ is the $B$ inner product.

Next, we overview the spectral and joint spectral quantities (5.19) and (5.20) in light of the finite element discretization (5.34). The density of states for the systems (5.34) are

$$\phi_e(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \delta(\lambda - \lambda_i^e) \quad \text{and} \quad \phi_h(\lambda) = \frac{1}{n}\sum_{j=1}^{n} \delta(\lambda - \lambda_j^h). \tag{5.35}$$

Similarly, the joint density of states is given by

$$J(\lambda) = \frac{1}{n^2}\sum_{i,j=1}^{n} \delta\big(\lambda - (\lambda_i^e + \lambda_j^h)\big). \tag{5.36}$$

For the absorption curve, notice that the approximations to the wavefunctions of the systems (5.13) are given by

$$\psi_i^e \approx \tilde{\psi}_i^e = \sum_{k=1}^{n} (x_i^e)_k \varphi_k \quad \text{and} \quad \psi_j^h \approx \tilde{\psi}_j^h = \sum_{k=1}^{n} (x_j^h)_k \varphi_k. \tag{5.37}$$

Therefore, using the relation $\langle \tilde{\psi}_i^e, \tilde{\psi}_j^h \rangle = (x_i^e, x_j^h)_B$, the approximate absorptions curve is

$$\alpha(\lambda) = \sum_{i,j=1}^{n} |(x_i^e, x_j^h)_B|^2 \delta\big(\lambda - (\lambda_i^e + \lambda_j^h)\big), \tag{5.38}$$

which is exactly the joint spectral function for the generalized eigensystems in (5.34). In the rest of this chapter we approximate the spectral quantities (5.35) and the joint spectral quantities (5.36) and (5.38) using the Lanczos process.

For all of the following problems we use the FEniCS finite element software [35] to assemble the stiffness and mass matrices and use the PETSc and SLEPc libraries [8, 21] for the solution of linear systems and eigenvalue problems respectively. When computing the absorption curve, we first need to approximate several eigenpairs of the electron (or hole) eigensystem. In SLEPc, a Krylov-Schur method is employed for the solution of all eigensystems, which for a symmetric matrix, is the thick-restart Lanczos algorithm [65]. The generalized eigenvalue problem uses the same method, with the only difference in the use of operator and inner product.

## 5.3 Homogeneous Alloys

Before modeling random alloys, we first look at the simple case of homogeneous alloys. For homogeneous alloys, we assume the indium fraction is constant throughout the domain. That is, we replace the spatially varying indium fraction, $X(x)$, with the bulk indium fraction $\overline{X}$. This removes the spatial dependence of the conductance and valence band potentials, making them constant. The effective masses for the electrons and holes are similarly constant. Because the potentials and effective masses are constant, the effective mass Schrödinger equation becomes the Laplace eigenvalue problem, which we can solve analytically. By understanding homogeneous alloys, which model so called "bulk" properties, we are able to understand and explain different phenomena that occur with random alloys.

For the homogeneous alloy, we choose the domain to be $\Omega = [0, L]^d$, in dimension $d$, where the domain length $L$ depends on the number of lattice sites and the lattice spacing. Specifically, for $N$ lattice sites with lattice spacing $a$, the domain length is given by $L = (N-1)a$. We choose to impose zero Dirichlet boundary conditions for simplicity.

In what follows we fix the indium fraction to be $\overline{X} = 0.2$. Vergard's Law with bowing parameter (5.3) gives a bandgap energy of $E_g = 2.65$ $e$V, from which we can determine the conductance and valence potential, $V_c = 2/3E_g$ and $V_v = 1/3E_g$. The effective masses are

$$
\begin{aligned}
m_e &= \left(\frac{1}{5m_e^{\text{InN}}} + \frac{4}{5m_e^{\text{GaN}}}\right)^{-1} \approx 0.15 m_e^0,\\
m_h &= \left(\frac{1}{5m_h^{\text{InN}}} + \frac{4}{5m_h^{\text{GaN}}}\right)^{-1} \approx 1.81 m_e^0,
\end{aligned}
$$

where $m_e^0$ is the electron rest mass $5.11 \times 10^5$ eV/$c^2$.

With all values in the Schrödinger equation specified, we are prepared to write the electron and hole wavefunctions , and corresponding energies. Using multi-index notation, the $L^2(\Omega)$ normalized solutions at $x = (x_1, \ldots, x_d) \in \Omega$ are given by

$$
\begin{aligned}
\psi_\mu^e(x) &= \left(\frac{2}{L}\right)^{d/2} \prod_{i=1}^{d} \sin\left(\frac{\mu_i \pi x_i}{L}\right), \quad & E_\mu^e &= V_c + \frac{\hbar^2\pi^2}{2m_e L^2}|\mu|^2,\\
\psi_\nu^h(x) &= \left(\frac{2}{L}\right)^{d/2} \prod_{i=1}^{d} \sin\left(\frac{\nu_i \pi x_i}{L}\right), \quad & E_\nu^h &= V_v + \frac{\hbar^2\pi^2}{2m_h L^2}|\nu|^2,
\end{aligned}
\tag{5.39}
$$

for $\mu, \nu \in \mathbb{N}^d$. In (5.39) we use the notation $|\mu|^2 = \sum_{i=1}^{d} \mu_i^2$ for $\mu \in \mathbb{N}^d$.

By the orthogonality of sinusoids,

$$
(\psi_\mu^e, \psi_\nu^h) = \delta_{\mu\nu} = \begin{cases} 1, & \mu = \nu,\\ 0, & \text{otherwise,} \end{cases}
$$

which means that most terms in the absorption curve are zero. Only those terms where the electron and hole multi-index are identical survive. Defining $E_{\mu\nu}$ as the sum of the electron and hole energies, from (5.39) we have

$$
E_{\mu\nu} := E_\mu^e + E_\nu^h = E_g + \frac{\hbar^2\pi^2}{2L^2}\left(\frac{|\mu|^2}{m_e} + \frac{|\nu|^2}{m_h}\right). \tag{5.40}
$$

The densities of state, joint density of state, and absorption curve for a homogeneous alloy are then given by

$$
\begin{gathered}
\phi^e(E) = \sum_{\mu} \delta\big(E - E^e_{\mu}\big), \qquad \phi^h(E) = \sum_{\nu} \delta\big(E - E^h_{\nu}\big), \\
J(E) = \sum_{\mu,\nu} \delta\big(E - E_{\mu\nu}\big), \\
\alpha(E) = \sum_{\mu} \delta\big(E - E_{\mu\mu}\big),
\end{gathered}
\tag{5.41}
$$

where all energies are given as in (5.39) or (5.40), and the summations are taken over all multi-indexes in $\mathbb{N}^d$.

When defining the density of states for a matrix, or joint density of states for a pair of matrices, we add a normalization factor, $1/n$ for the density of states or $1/n^2$ for the joint density of states, in front of the summation, where $n$ is the matrix size. In the case of infinite dimensional solutions like (5.39), this normalization is nonsensical. Instead, we plot the spectral densities for energy values less than or equal to 4 $e$V, and normalize each by the number of summands with energy values less than 4.5 $e$V (we extend beyond 4 $e$V because when using the Gaussian in place of the Dirac mass, terms with energy beyond 4 $e$V contribute to the value of the spectral quantity for energy values less than or equal to 4 $e$V). We denote the number of summands in the electron density of states, hole density of states, joint density of states, and absorption curve as $n_e$, $n_h$, $n_{eh}$, and $n_\alpha$ respectively. These values, along with the domain length used for each dimension, are given below in Table 5.3. The spectral and joint spectral quantities for homogeneous alloys in one, two, and three dimensions can be seen in Figures 5.8, 5.9, and 5.10 respectively. All plots are shown with regularization parameter $\sigma = 100$ m$e$V.

| $d$ | $N$ | $n_e$ | $n_h$ | $n_{eh}$ | $n_\alpha$ |
|---|---|---|---|---|---|
| 1 | 5001 | 1479 | 5913 | 4,043,948 | 1170 |
| 2 | 201 | 2693 | 43,699 | 20,032,672 | 1671 |
| 3 | 51 | 1445 | 104,176 | 7,870,992 | 681 |

Table 5.3: Homogeneous InGaN spectral quantity data.

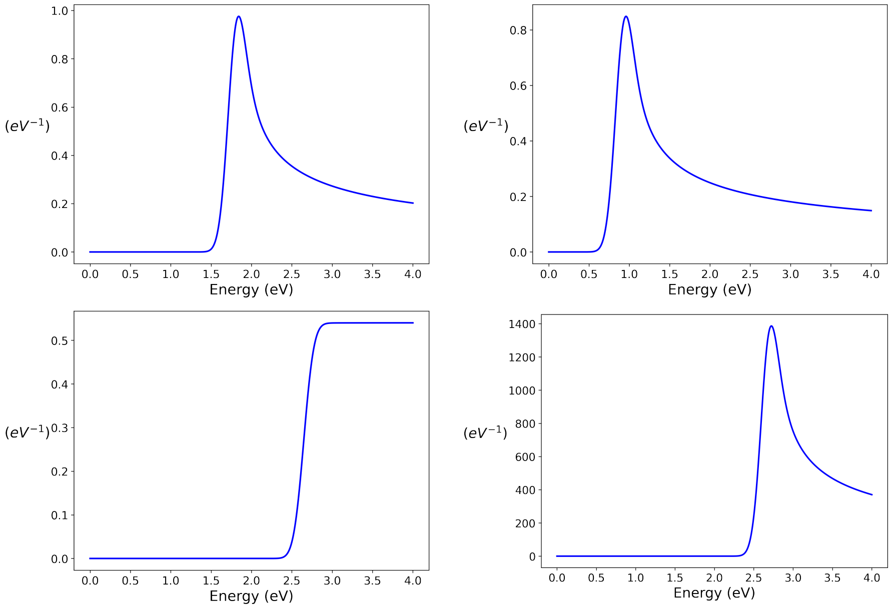


Figure 5.8: Electron DOS (top left), hole DOS(top right), JDOS (bottom left), and absorption curve (bottom right) for a one dimensional uniform InGaN alloy with twenty percent indium. Plotted using $\sigma = 100$ m$e$V.

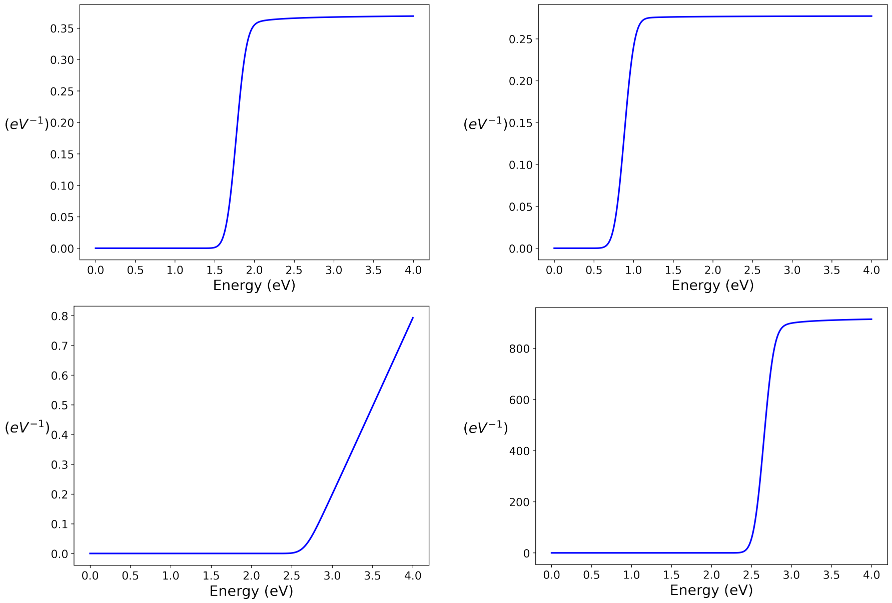


Figure 5.9: Electron DOS (top left), hole DOS(top right), JDOS (bottom left), and absorption curve (bottom right) for a two dimensional uniform InGaN alloy with twenty percent indium. Plotted using $\sigma = 100$ m$e$V.

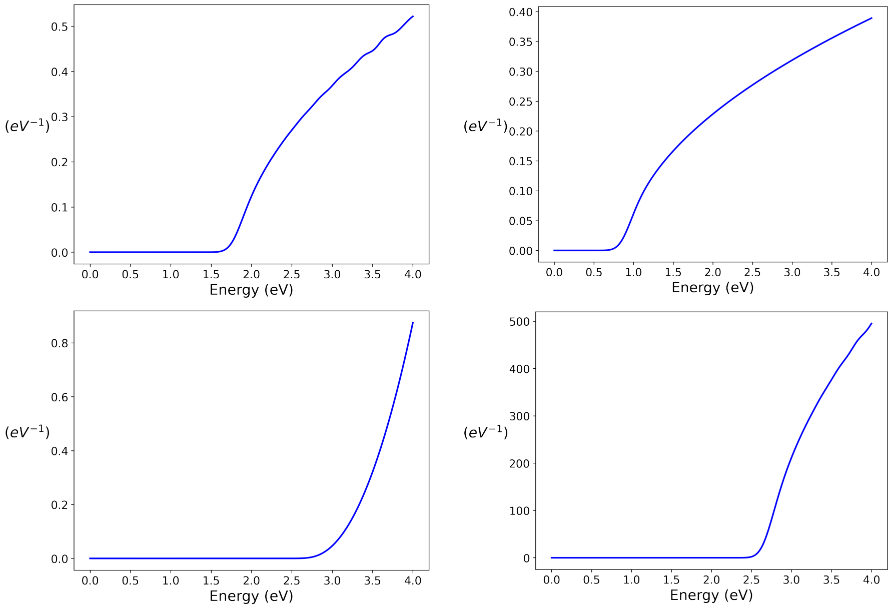


Figure 5.10: Electron DOS (top left), hole DOS(top right), JDOS (bottom left), and absorption curve (bottom right) for a three dimensional uniform InGaN alloy with twenty percent indium. Plotted using $\sigma = 100$ m$e$V.

## 5.4 Joint Spectral Approximation Workflow

In this section we describe the workflow used to approximate the absorption curve for a random InGaN alloy in $d$-dimensions. The outline is listed below:

1. Create random InGaN lattice, compute fundamental energy of electron and hole eigensystems, and minimal and maximal eigenvalue of scaled stiffness matrix.

2. Use the Lanczos process to approximate the electron and hole density of states.

3. Using the densities of states, determine how many and which eigenpairs need to be computed.

4. Compute eigenpairs.

5. Approximate absorption curve by performing the Lanczos process with each computed eigenvector as starting vector.

The first step, creating the random lattice, is where we compute the spatially dependent indium fraction $X(x)$. With the indium fraction, we are able to compute the conductance and valence potentials $V_c$ and $V_v$, as well as the effective masses $m_e$ and $m_h$. These terms give us all the ingredients to form the stiffness matrices for the electrons and holes $A^e$ and $A^h$. Note that the mass matrix is independent of the specific random realization for the InGaN lattice, and only depends on the choice of tessellation and piecewise polynomial basis. Once the matrices have been assembled, we begin approximating spectral quantities.

Before beginning with any Lanczos type methods, we first compute several preliminary eigenvalues. The first are the fundamental energies of the electron and hole eigensystems. These allow us to use the density of states (discussed shortly) to determine how many eigenpairs to compute. Recall, if we were computing the absorption curve exactly, need to compute $n_e$ electron and $n_h$ hole eigenpairs so that

$$\overline{E} \leq \min(\lambda_1^e + \lambda_{n_h}^h, \lambda_{n_e}^e + \lambda_1^h), \tag{5.42}$$

is satisfied, where $\overline{E}$ is the maximal energy we are interested in viewing the absorption curves. This way, we know that computing more eigenpairs does not change the absorption curve for energies less than $\overline{E}$. We choose $\overline{E} = 4\ e\text{V}$ throughout this chapter because this is slightly above the bandgap of GaN, and, as we will see, all of the interesting phenomena occur between the bandgap of InN and GaN. Condition (5.42) similarly applies to using the Lanczos process, except that we only need to compute one set of eigenpairs or the other. By knowing the fundamental electron and hole eigenvalues to very high accuracy, we can then determine the number of electron and hole eigenpairs necessary using

$$n_e = \int_0^{\overline{E}-\lambda_1^h} n\phi^e(\lambda)d\lambda \quad \text{and} \quad n_h = \int_0^{\overline{E}-\lambda_1^e} n\phi^h(\lambda)d\lambda. \tag{5.43}$$

In practice, we replace the spectral densities in (5.43) with the regularized approximations (replace Dirac delta with a Gaussian of variance $\sigma$ as in (5.21)) from the Lanczos process, and use a simple composite trapezoidal rule to approximate the integrals. More details on the choice of $\sigma$ are given in Section 5.5.

In addition to the fundamental electron and hole energies, we compute the largest and smallest eigenvalues of the mass matrix. The mass matrix does not change from one random realization to the next, and so the extremal eigenvalues can be stored and reused if many computations are performed using the same mesh and polynomial basis. We compute these values for use in the density of states computation. Using the minimal and maximal eigenvalues of the stiffness matrix, we are able to determine the degree Chebyshev expansion with which to approximate $S^{-1}$ where $S$ is the square root of the stiffness matrix $B$. This was discussed at length in the previous chapter, and so here we highlight the computations performed without explanation. If $a$ and $b$ are the smallest and largest eigenvalues of $B$, then we define $c = 1/2(b+a)$, $d = 1/2(b-a)$, and set

$$\rho = \frac{c}{d} + \sqrt{\left(\frac{c}{d}\right)^2 - 1}.$$

We choose a Chebyshev expansion of degree $k$, where $k$ is the smallest integer satisfying

$$\frac{2\rho^{-k}}{(\rho-1)\sqrt{c-d^r}} < 10^{-16},$$

where $r = 1/2(\rho + \rho^{-1})$. The value $k$ is determined using a bisection algorithm. The values of $k$, $a$, and $b$, along with the domain, $\Omega$, and degree finite elements $p$, used in this chapter for each dimension are shown in Table 5.4.

| $d$ | $\Omega$ | $p$ | $a$ | $b$ | $k$ |
|---|---|---|---|---|---|
| 1 | $[0, 5000]$ | 3 | 0.50 | 1.48 | 28 |
| 2 | $[0, 200]^2$ | 3 | 0.29 | 2.01 | 48 |
| 3 | $[0, 50]^3$ | 2 | 0.25 | 4.35 | 81 |

Table 5.4: Degree Chebyshev expansion to use for approximating inverse square root of $B$.

Next, we perform Step (2), which is to use the Lanczos process to approximate the densities of states for the electron and hole systems. For each system, we perform $n_v$

trials with starting vectors with entries drawn independently from the standard normal distribution as in Algorithm 2. Using the approximations to the densities of states $\tilde{\phi}_e$ and $\tilde{\phi}_h$ for the electrons and holes respectively, we can approximate the required number of electron and hole eigenpairs necessary according to (5.43). Call these approximations $\tilde{n}_e$ and $\tilde{n}_h$.

Assuming $\tilde{n}_e < \tilde{n}_h$, we compute $\tilde{n}_e$ electron eigenpairs in Step (3). Note that for InGaN alloys, it is always the case that $\tilde{n}_e < \tilde{n}_h$, by a considerable factor. This is due to the higher mass of the holes in comparison to that of the electrons. The density of hole eigenpairs at lower energies can be seen explicitly in the densities of states for the homogeneous alloys in Figures 5.8, 5.9, 5.10. We include the computation of $\tilde{n}_h$ in this thesis for completeness. Also, we need the hole density of states approximation in order to form the joint density of states approximation using method II described in the previous chapter. So, the only "unnecessary" work performed is the computation of the fundamental electron eigenvalue, and the approximation of $n_h$ by (5.43). The timing of these two steps is negligible when compared to the total timing of computing the approximate absorption curve.

Once the $\tilde{n}_e$ electron energies, $\lambda_i^e$, and eigenvectors, $x_i^e$, $i = 1, \ldots, \tilde{n}_e$, are computed, we then perform the Lanczos process with the matrices $A^h$, $B$, and starting vector $x_i^e$, to approximate the marginals

$$s_i^h(\lambda) := s(\lambda; A^h, B, x_i^e) = \sum_{j=1}^{n} |(x_i^e, x_j^h)|^2 \delta(\lambda - \lambda_j^h) \quad \text{for} \quad i = 1, \ldots, \tilde{n}_e. \tag{5.44}$$

If the Lanczos approximation to the marginal $s_i^h(\lambda)$ is denoted $\tilde{s}_i^h(\lambda)$, then the approximation to the absorption curve is

$$\tilde{\alpha}(\lambda) = \sum_{i=1}^{\tilde{n}_e} \tilde{s}_i^h(\lambda - \lambda_i^e). \tag{5.45}$$

Similarly, if it were the case that $\tilde{n}_h < \tilde{n}_e$, we would compute $\tilde{n}_h$ hole eigenvalues $\lambda_j^h$ and corresponding eigenvectors $x_j^h$ for $j = 1, \ldots, \tilde{n}_h$. Then, we use the eigenvectors $x_j^h$

as the starting vectors for the Lanczos process to approximate the marginals

$$s_j^e(\lambda) := s(\lambda; A^e, B, x_j^h) = \sum_{i=1}^{n} |(x_i^e, x_j^h)|^2 \delta(\lambda - \lambda_i^e). \tag{5.46}$$

Denoting by, $\tilde{s}_j^e$, the Lanczos approximation to (5.46), the Lanczos approximation to the absorption curve is then

$$\tilde{\alpha}(\lambda) = \sum_{j=1}^{\tilde{n}_h} \tilde{s}_j^e(\lambda - \lambda_j^h). \tag{5.47}$$

## 5.5 1D Random Alloys

We begin with the simple one dimensional case. For all computations we use a lattice with 5001 lattice spaces (the first lattice space being equal to the last due to the choice of periodic boundary conditions), meaning our computational domain is $\Omega = [0, 5000]$. In terms of physical units this is a domain of length $5000 \times 2.833$ Å $\approx 1.4\ \mu m$. For all computations we use unit length intervals to discretize the domain, and use degree three polynomials on each interval for the finite element method. The resulting matrices have dimension $15,000 \times 15,000$, which is small enough so that we can compute the exact absorption curve.

The absorption curves for varying levels of indium concentration can be seen in Figure 5.11. These are computed using the number of electron and hole eigenpairs displayed in Table 5.5. For each of the four indium concentrations, $n_e \times n_h$ overlap integrals are computed, and the absorption curve with as many summands is evaluated at an array of points.

| $\overline{X}$ | $n_e$ | $\lambda_1^e$ (eV) | $n_h$ | $\lambda_1^h$ (eV) |
|---|---|---|---|---|
| 0.05 | 1133 | 1.41 | 3867 | 0.58 |
| 0.10 | 1164 | 1.27 | 4096 | 0.52 |
| 0.15 | 1202 | 1.10 | 4353 | 0.42 |
| 0.20 | 1233 | 0.93 | 4582 | 0.32 |

Table 5.5: One dimensional absorption curve data.

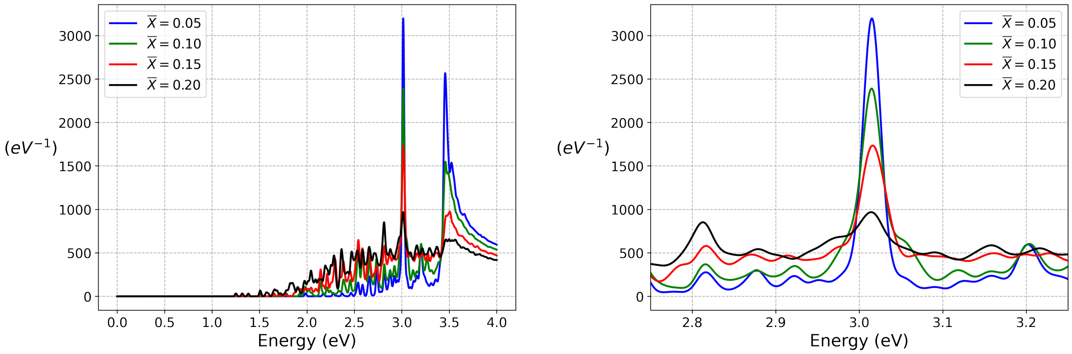


Figure 5.11: 1D absorption curves using $\sigma = 10$ m$e$V for indium concentrations between 5% and 20% (left). Zoom into energy levels around 3 eV (right).

We begin by explaining the discrepancy in the absorption curve for a random alloy (shown in Figure 5.11) and that of a homogeneous alloy (shown in Figure 5.8) in one dimension.

### 5.5.1 Spike near 3 eV

Here we discuss numerically the spike in the 1D absorption curve, which does not appear in the case of homogeneous alloys. In order to verify that the spike is indeed physical, and not a numerical artifact, we give results pertaining to quantum wells in 1D which justify the presence of the spike. All computations done in this section are performed by discretizing the effective mass Schrödinger equation, and solving the resulting generalized algebraic eigenvalue problem directly. No Lanczos approximations are used in this section, and the numerical discretization is done with enough precision so that we can consider the computed eigenpairs to be exact.

First we look at the densities of states for the four random realizations which give the results displayed in Figure 5.11. These densities of states can be seen in Figure 5.12. In both the electron spectral density (top figures) and hole spectral densities (bottom figures) we see an additional spike which is not present in the densities of state for homogeneous alloys. The electron densities for the four bulk indium concentrations plotted, exhibit a concentration of energies near 2.05 $e$V, while the hole densities exhibit one near 0.96 $e$V. Adding these energies together gives 3.01 $e$V, which corresponds to the location of the spike in the absorption curve.

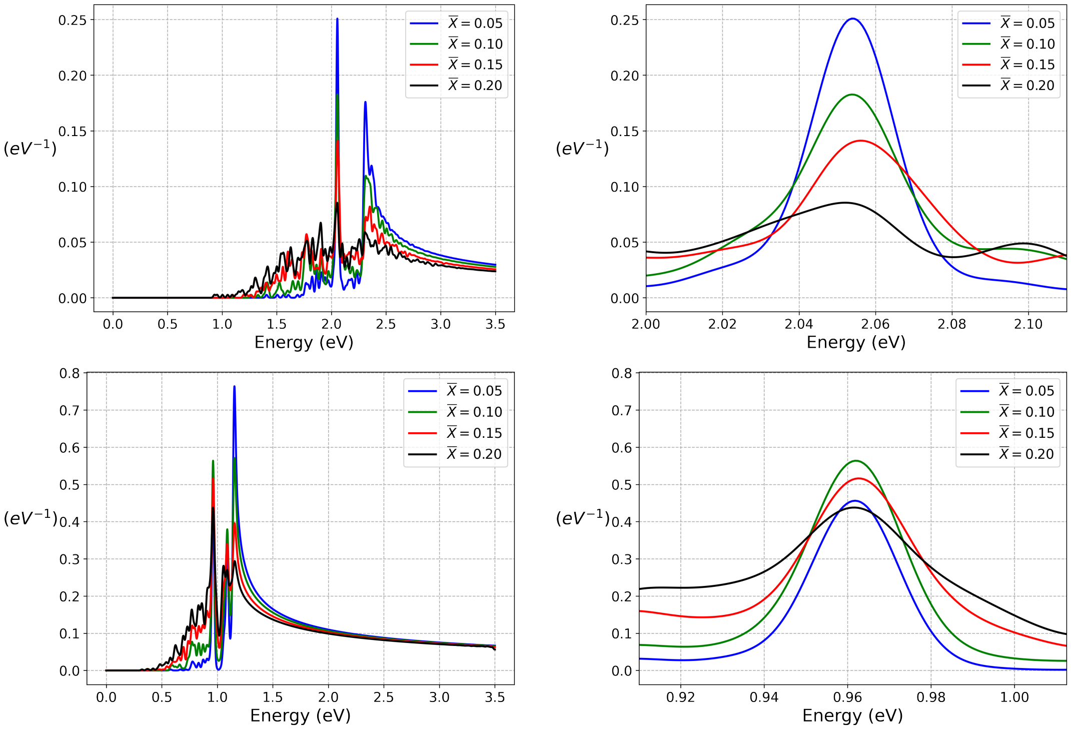


Figure 5.12: Electron and hole densities of states for varying bulk indium concentration $\overline{X}$. The electron density of states (top left) and zoom into the region around 2.05 eV (top right). The hole density of states (bottom left) and zoom into the region around 0.96 eV (bottom right).

To investigate what is special about the electron and hole energies 2.06 and 0.95, we perform an experiment. We simplify the problem to have 101 lattice points where the arrangement is 50 GaN cations, 1 InN cation, and another 50 GaN cations. The indium fraction for this specific arrangement is shown in Figure 5.13 (left). The black dashed lines indicate three standard deviations (six lattice spaces) from the InN position at $x = 50$. Here the standard deviation is in reference to the Gaussian averaged indium fraction. From basic properties of the normal distribution, 99.73% of the area under the indium fraction is accounted for within the region enclosed by the black dashed lines.

With the indium fraction determined for this special arrangement of GaN and InN, we next determine the bandgap $E_g(x)$ according to (5.3). This is the blue line shown in Figure 5.13 (right). The dashed black lines again show six lattice positions to the left and right of the InN cation at position $x = 50$. Outside of region enclosed by the

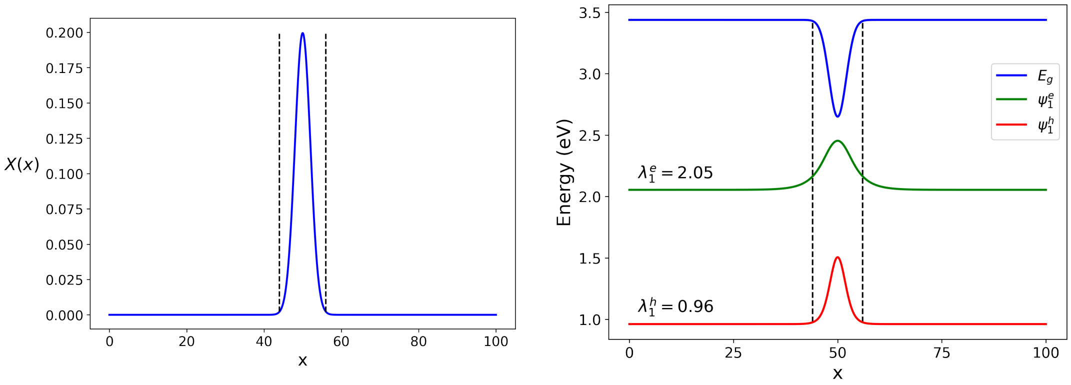


Figure 5.13: Indium fraction for a one dimensional lattice with fifty GaN cations, one InN cation, and another fifty GaN cations (left). Black lines indicating three standard deviations, or six lattice spaces, for the Gaussian averaged indium fraction to the left and right of the InN cation. Spatially dependent bandgap (see (5.3)) and fundamental electron and hole eigenfunction superimposed.

dashed black lines, the bandgap is essentially that of GaN, $E_g^{\text{GaN}} = 3.437\ eV$. Close to the InN cation, the bandgap drops to a minimum of 2.65 $e$V, which is the bandgap for a twenty percent indium fraction. With the conductance and valence potentials being proportional to the bandgap, both exhibit a minima at $x = 50$. This situation is similar to the classical finite square well potential studied in most quantum physics texts. The main difference being the conductance and valence energies are smooth, as opposed to the discontinuous finite square well. The $L^2$ normalized fundamental electron and hole eigenfunctions are plotted along with the bandgap in Figure 5.13 (right). The height of the electron and hole wavefunctions is their respective energy level.

We see that, indeed, both are localized in the region of low potential induced by the presence of the InN cation. We also see that the fundamental electron wavefunction “leaks” out of the region of low potential. Physically, this represents a probability of the electron being outside the region of low energy. This phenomena is referred to as quantum tunneling, see, e.g., [46]. Most importantly, the fundamental electron and hole energies are approximately 2.05 eV and 0.96 eV respectively, which matches the peaks in the electron and hole spectral densities.

When moving to the more complicated situation of 5001 lattice positions, the same

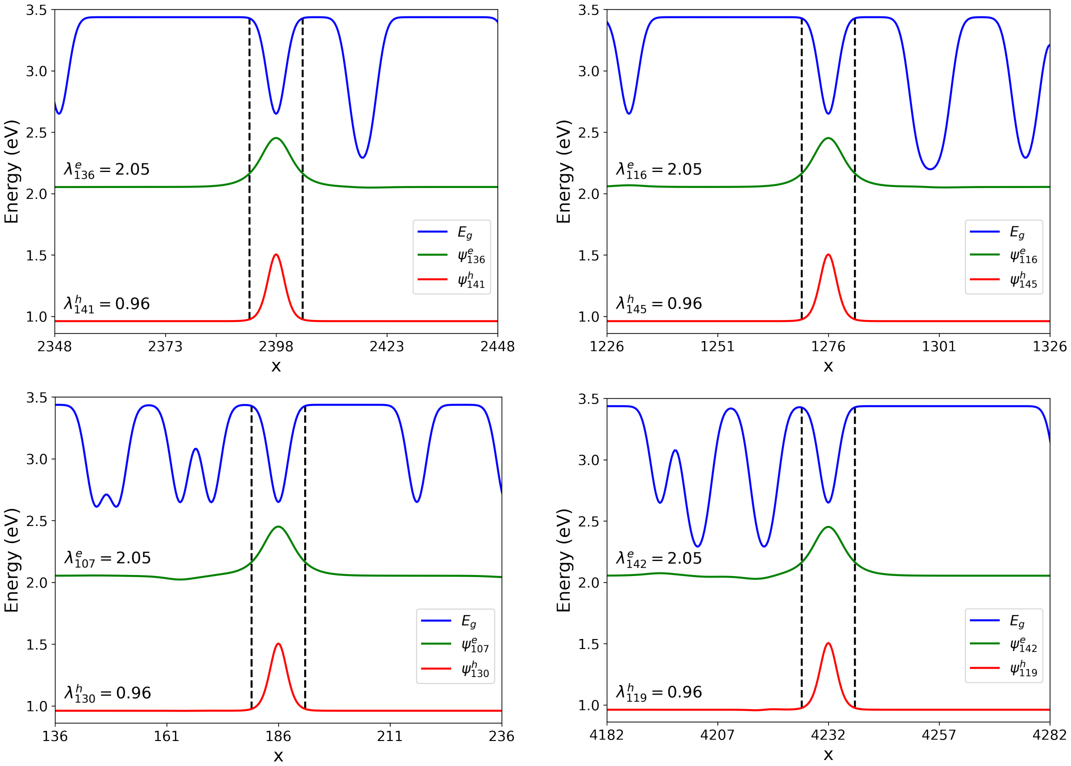


Figure 5.14: Four instances of electron and hole pairs with large overlap where the electron eigenfunctions have energy approximately 2.05 $e$V and the hole eigenfunctions have energy approximately 0.96 $e$V.

situation of one InN cation surrounded by many GaN cations occurs many times, resulting in a spike in the absorption curve around 3.01 eV. We look at the case of $\overline{X} = 0.05$, and show four occurrences in Figure 5.14 which are similar to the case of one InN cation surrounded by fifty GaN cations on either side. In all four instances, while the bandgap is more complicated, the end result is the same. The electron and hole wavefunctions localize in a local minima of the bandgap, and have energies of 2.05 $e$V and 0.96 $e$V respectively.

Additionally, due to the numerous random configurations of InN and GaN, there are additional eigenmodes contributing to the spike in the absorption curve. A few instances of these are shown in Figure 5.15.

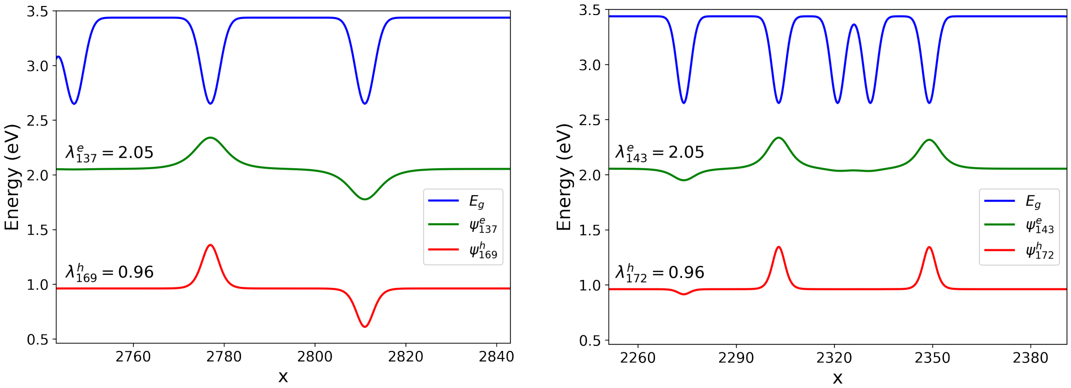


Figure 5.15: Additional examples of electron and hole eigenmodes adding to the spike in absorption curve at 3.01 $e$V.

### 5.5.2 One Dimensional Walk-through

Now that we have explained the discrepancy in the one dimensional absorption curve for homogeneous alloys and random alloys, we step through through the workflow outlined in Section 5.4 for the Lanczos approximation of the absorption curve and joint density of states. For these problems, the matrix sizes are sufficiently small so that the exact solution is computable. Therefore, we are able to compute errors in the Lanczos approximation to the absorption curve and joint density of states. Using techniques and parameter values given in this section, we will be able to approximate absorption curves in two and three dimensions, where the exact solutions are too computationally expensive to compute.

The example we use is an $\text{In}_X\text{Ga}_{1-X}\text{N}$ alloy with twenty percent indium concentration. We again use a lattice containing 5001 sites (the first being equal to the last) and the standard finite element method with cubic polynomials, which results in stiffness and mass matrices of order $n = 15,000$. For the random lattice realization in this example, the spatially varying indium fraction is displayed in Figure 5.16 (top left). A zoom in of the region $[10, 30]$ is also shown in Figure 5.16 (top right). By superimposing the lattice, we are able to see how the Gaussian averaging process is effected by individual InN cations in the lattice. With the indium fraction determined, we are able to construct the spatially varying bandgap, and hence the potentials and effective masses required for construction of the stiffness and mass matrices. The bandgap, shown in

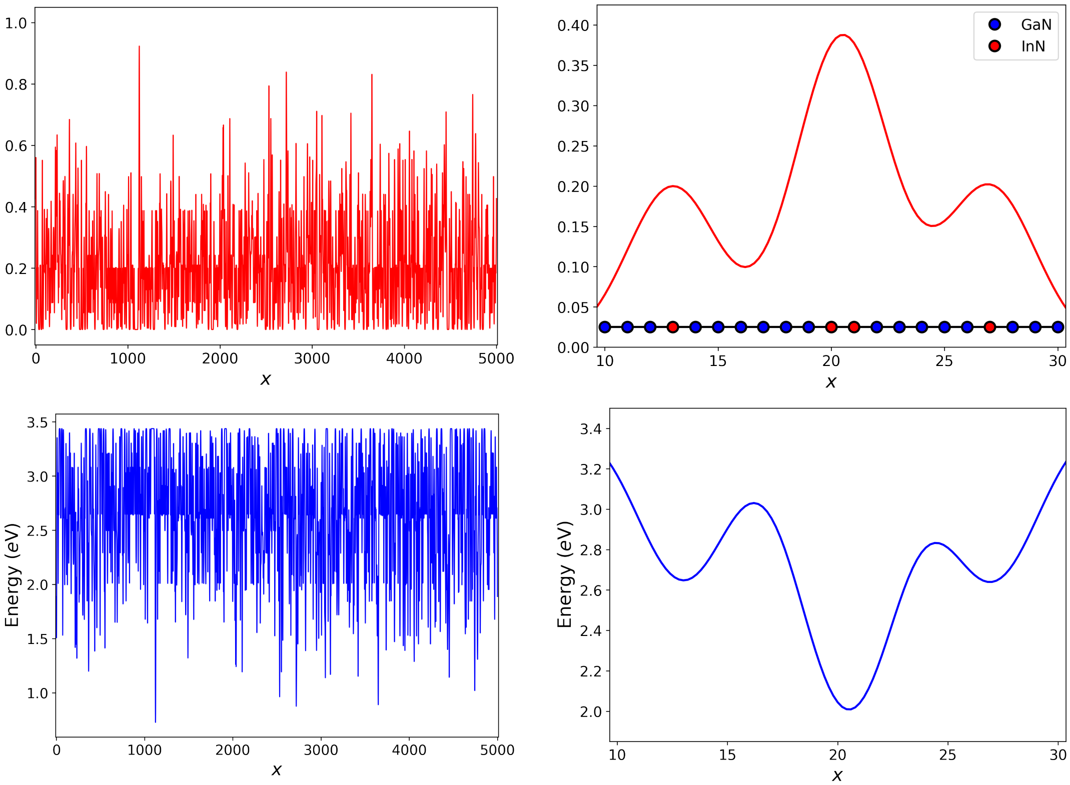


Figure 5.16: Indium fraction for example with 5001 lattice positions and $\overline{X} = 0.20$ (top left). Zoom into the region $[10, 30]$ with arrangement of InN and GaN cations on lattice (top right). Bandgap (bottom left) and zoom into the region $[10, 30]$ (bottom right).

Figure 5.16 (bottom left), oscillates between the bandgap of GaN and InN, varying with the indium composition. A zoom of the bandgap in the region $[10, 30]$ is shown in Figure 5.16 (bottom right).

The next step in the workflow is the approximation of the electron and hole densities of states. These are shown in Figure 5.17 for two different values of $\sigma$. Ten trials were performed and a Krylov parameter of $m = 150$ ($m = n/100$) was used. Note that we used a constant Krylov parameter, rather than increasing $m$ until the gap between Ritz values is small enough, as discussed at the end of Section 5.1.9. We do this because a coarse approximation to the density of states is more than sufficient for our purposes. We next show that for this value of Krylov parameter, while the Lanczos approximation to the density of states (especially in the electron case) is not accurate, the integral of

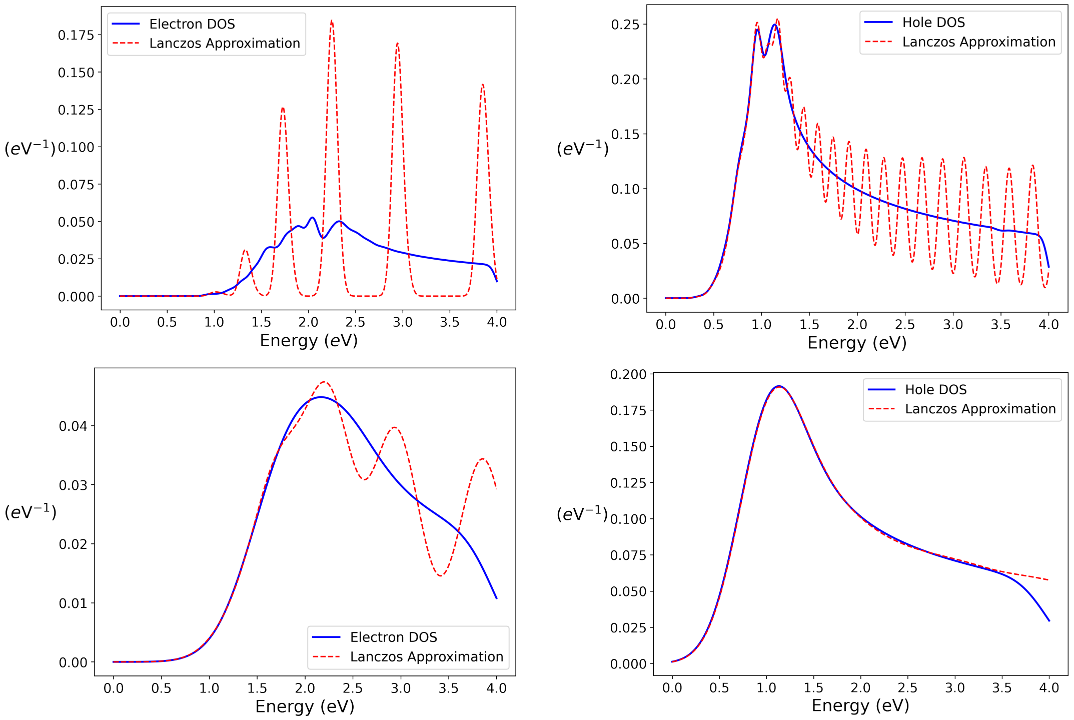


Figure 5.17: Lanczos approximation to the electron and hole densities of state using 10 trials and $m = 150$.

the density of states, i.e., the counting function, is accurate enough for our purposes.

The next step is to approximate the number of electron eigenpairs needed to approximate the absorption curve using the Lanczos process. The number of electron and hole eigenpairs necessary, $n_e$ and $n_h$ respectively, is defined by (5.43). For this example, the fundamental electron and hole energies are 0.93 eV and 0.32 eV respectively. Using the fundamental energies and criterion (5.42), we need to compute all electron eigenpairs up to energy level $4 - 0.32 = 3.68$ eV, or all hole eigenpairs up to energy level $4 - 0.93 = 3.07$ eV. The exact values of electron and hole eigenpairs necessary are $n_e = 1233$ and $n_h = 4582$ respectively. The approximation of these numbers is computed by replacing the exact density of states in (5.43) with the Lanczos approximation using several values of regularization parameter $\sigma$. The integrals in (5.43) are approximated using a composite trapezoidal rule. Once the electron eigenpairs are computed, we can easily check that we have computed enough using (5.42).

| $\sigma$ (meV) | $n_e$ | $\lceil \tilde{n}_e \rceil$ | $n_h$ | $\lceil \tilde{n}_h \rceil$ |
|---|---|---|---|---|
| 10 | | 1117 | | 4485 |
| 130 | | 1142 | | 4557 |
| 260 | 1233 | 1189 | 4582 | 4549 |
| 380 | | 1205 | | 4522 |
| 500 | | 1202 | | 4463 |

Table 5.6: Approximation of the values $n_e$ and $n_h$ defined in (5.43) using the Lanczos approximation to the densities of states for several values of $\sigma$.

The approximations to $n_e$ and $n_h$, $\tilde{n}_e$ and $\tilde{n}_h$ respectively, are shown in Table 5.6. From Table 5.6 we see that using a larger value of $\sigma$ allows us to approximate the number of electron eigenpairs necessary. For visualization of the density of states, a larger value of $\sigma$ blurs out details. Yet, for approximating the value $n_e$, larger values of $\sigma$, produce more accurate approximations. In order to ensure enough electron eigenpairs are computed we implement a five percent fudge factor, and request $\lceil 1.05\tilde{n}_e \rceil = 1263$ electron eigenpairs from SLEPc. A five percent fudge factor and a value of $\sigma = 500$ $m$eV is used in all subsequent computations in the approximation of $n_e$.

Because the number of electron eigenpairs necessary to compute the absorption curve is much smaller than the number of hole eigenpairs, we use the electron eigenpairs for the Lanczos process. Once the electron eigenpairs have been computed using SLEPc, we are ready to approximate the spectral functions, $s_i^h(E)$, defined in (5.44) for each of the computed eigenvectors, with the absorption curve a sum of such approximations (as in (5.45)). Before approximating these spectral functions, we view the exact first term in the absorption curve expansion, $s_1^h(E - \lambda_1^e)$, seen in Figure 5.18 in linear and log scale. We see an interesting discrepancy between the spectral function on a linear scale, and on a log scale. On the linear scale we see three spikes of decreasing amplitude, while on the log scale, several more spikes are present which are indiscernible on the linear scale. Also to be noted, is the trailing edge in the spectral function in the log scale after approximately 2.1 $e$V. Next, we investigate the qualitative aspects of the spectral function corresponding to the fundamental electron wavefunction, and its approximation using the Lanczos process.

We approximate the spectral functions, $s_1^h(E - \lambda_1^e)$, using the Lanczos process in

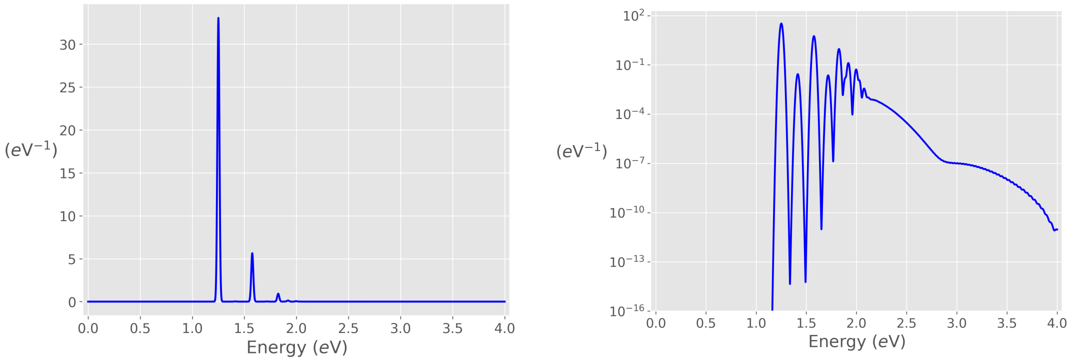


Figure 5.18: Exact spectral function corresponding to the fundamental electron eigenfunction and hole eigensystem, $s_1^h(E-\lambda_1^e)$, on linear scale (left) and log scale (right) for $\sigma = 10$ m$e$V.

Figure 5.19 for $\sigma = 10$ m$e$V and Krylov parameters $m = 25$ and $m = 200$. There are two items to note in Figure 5.19 with regard to the Lanczos approximation. First, that the small value of $m = 25$ does an exceptional job of capturing the details of the exact spectral function. Second, there does not appear to be much of a discrepancy between using $m = 25$ and $m = 200$. We investigate the difference in using $m = 25$ and $m = 200$ in more detail momentarily. We describe the "spikey" nature of the spectral function next.

As seen in Figure 5.16 for the simple case of fifty GaN cations, one InN cation, and another fifty GaN cations, the fundamental electron and hole eigenfunctions "localize" in the region of the domain where the bandgap, $E_g(x)$, is minimal. In the case of a random alloy, these finite wells occur many times and eigenfunctions localize at the local minima of the bandgap (local maxima of the indium fraction). The lowest energy eigenfunctions occurring where a large cluster of InN cations occur in the lattice. The $L^2$ normalized eigenfunctions responsible for the qualitative structure of the spectral function, $s_1^h(E-\lambda_1^e)$, are plotted in Figure 5.20. Because lower energy hole eigenfunctions each localize to their own respective well, the overlaps with the fundamental electron eigenfunction, $|(\psi_1^e, \psi_j^h)|^2$, are essentially zero except for those hole wavefunctions localizing in the same well as $\psi_1^e$. The overlaps responsible for the first two spikes in the spectral function are $|(\psi_1^e, \psi_1^h)|^2 = .83$ and $|(\psi_1^e, \psi_{88}^h)|^2 = .14$ and occur at energies $\lambda_1^e + \lambda_1^h = 1.25\ eV$ and $\lambda_1^e + \lambda_{88}^h = 1.58\ eV$ respectively. Interestingly, we can see that $\psi_1^h$ and $\psi_{88}^h$ are the first

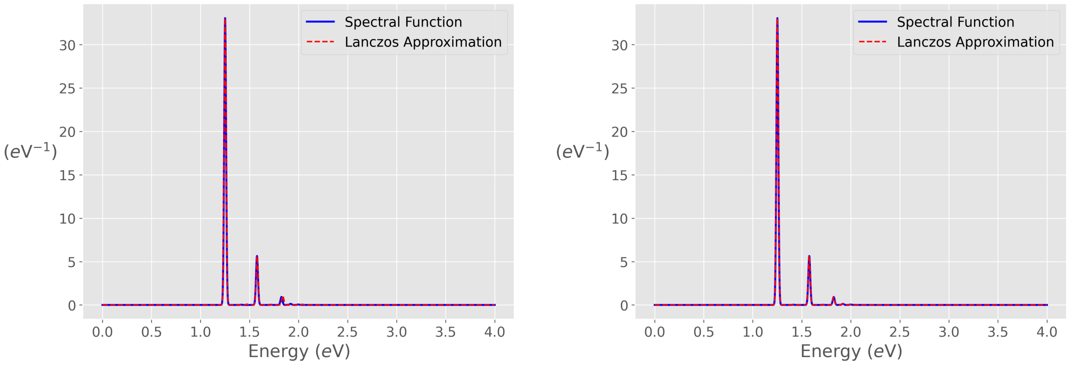


Figure 5.19: Lanczos approximation to spectral function corresponding to the fundamental electron eigenfunction and hole eigensystem for $m = 25$ (left) and $m = 200$ (right) on linear scale ($\sigma = 10$ meV).

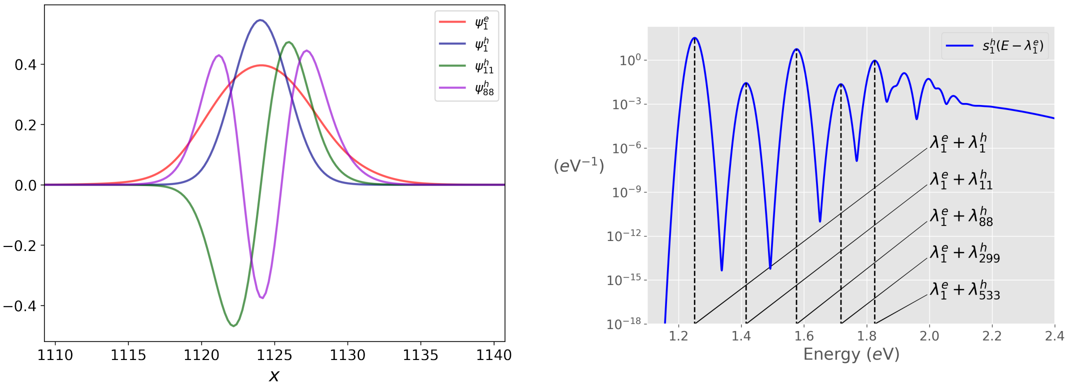


Figure 5.20: Eigenfunctions responsible for the qualitative aspects of the spectral function $s_1^h(E - \lambda_1^e)$ (left) and the exact spectral function on a log scale (right).

and third mode corresponding to the well where $\psi_1^e$ localizes. Therefore, the overlap corresponding to the second mode is relatively insignificant, and cannot even be seen in Figure 5.19. We can check that the second mode for this well, $\psi_{11}^h$, produces an overlap of $|(\psi_1^e, \psi_{11}^h)|^2 = 6.85 \times 10^{-4}$ at energy $\lambda_1^e + \lambda_{11}^h = 1.42\ eV$.

Next, we investigate the trailing edge in the spectral function corresponding to the fundamental electron eigenfunction. As seen in Figure 5.20, the overlaps determining the first five or so spikes are caused by hole eigenfunctions localized in the same region as the electron eigenfunction. Once we move to higher energies, we see a trailing edge in the log plot of $s_1^h(E-\lambda_1^e)$ beginning around 2.1 $e$V, as can be seen in Figures 5.18 (right)

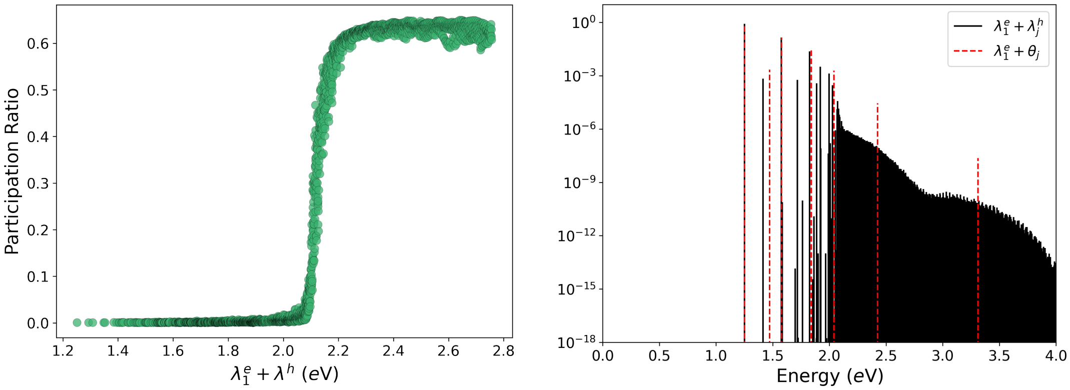


Figure 5.21: Participation ratio of for hole eigenfunctions, $\psi_j^h$ (ordinate) corresponding to energies $\lambda_1^e + \lambda_j^h$ (abscissa) showing contribution of delocalized eigenfunctions to trailing edge (left). Barcode plot of spectral function $s_1^h(E - \lambda_1^e)$ and Lanczos approximation for $m = 25$ (right).

and 5.20 (right). We show this is due to higher energy delocalized hole eigenfunctions.

One measure of localization of a function $\phi : \Omega \to \mathbb{R}$ is the participation ratio given by

$$\frac{1}{|\Omega|} \frac{\left( \int_\Omega \phi^2 \right)^2}{\int_\Omega \phi^4}, \tag{5.48}$$

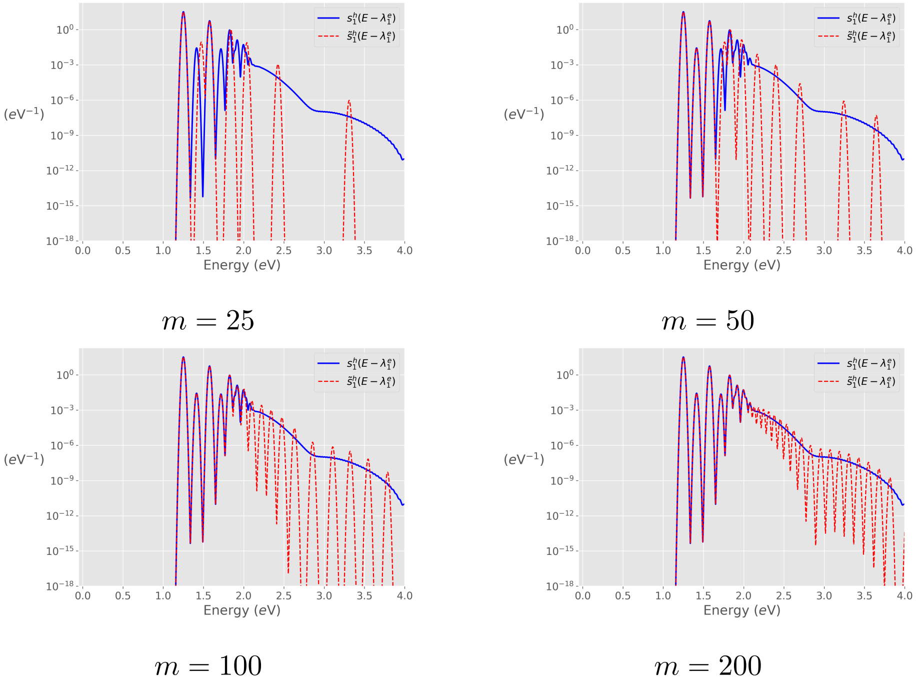
where $|\Omega|$ is the volume of the domain. This may also be considered a relative participation ratio due to the factor of $|\Omega|^{-1}$. Notice that the participation ratio of a constant is unity, while the participation ratio of the characteristic function of a subdomain $\Omega_0 \subset \Omega$ is $|\Omega_0|/|\Omega|$. In other words, the smaller the support of $\phi$, the closer to zero the participation ratio. In Figure 5.21 (left), the participation ratio of the hole eigenfunctions,$\psi_j^h$, is plotted on the y-axis, with energies $\lambda_1^e + \lambda_j^h$ shown on the x-axis. We can see that the hole eigenfunctions delocalize, i.e., begin to be supported on the entire domain, around 2.1 $e$V, which is exactly where the trailing edge begins in the spectral function $s_1^h(E - \lambda_1^e)$. We remark that the participation ratio of sinusoids in one dimension is 2/3, which is close to where the participation ratio of the hole eigenfunctions asymptotes.

Another way to visualize the spectral function (and Lanczos approximation) is shown in Figure 5.21 (right). At each energy $\lambda_1^e + \lambda_j^h$, $j = 1, \ldots, n_h$, a thin black line is shown with height equal to the overlap $|(\psi_1^e, \psi_j^h)|^2$. Also shown in red is the Lanczos approximation to the spectral function for $m = 25$. The red stripes are located at energies $\lambda_1^e + \theta_j$, for $j = 1, \ldots, m$, where the $\theta_j$'s are the Ritz values for the Lanczos process, and have heights equal to $|(y_j, e_1)|^2$, where $y_j$ is the eigenvector corresponding to $\theta_j$. With this "barcode" plot, we are able to again see the trailing edge near 2.1 $e$V. Furthermore, we see that the Lanczos approximation is able to match the first and third spike (almost) exactly, while the weights, $|(y_j, e_1)|^2$, for energies higher than 2 $e$V are larger than the exact overlaps in order to satisfy the moment matching criterion.

In order to further investigate the effectiveness of the Lanczos process in approximating the absorption curve, we look at the Lanczos approximation to the first spectral function, $s_1^h(E - \lambda_1^e)$, on a log scale for various values of the Krylov parameter $m$. This is shown in Figure 5.22 and 5.23 for regularization values $\sigma = 10$ m$e$V and $\sigma = 50$ m$e$V respectively. Figure 5.22 shows how truly remarkable the Lanczos process is. Even for the small value $m = 25$, the Lanczos approximation captures the first overlap to high accuracy. Then, as $m$ is increased, more and more of the character of the spectral function is captured. Figures 5.22 and 5.23 also show how the Lanczos approximation is influenced by the magnitude of the regularization parameter $\sigma$. The more blurring present, i.e., the larger $\sigma$ is, the easier the absorption curve is to approximate.

Finally, we are ready to approximate the absorption curve for a one dimensional InGaN alloy. The Lanczos approximation to the absorption curve works by approximating several spectral functions, each corresponding to one of the computed electron eigenfunctions (see (5.45)). We have gone into great detail of the spectral function corresponding to the fundamental electron eigenfunction, the others being similar. The Lanczos approximations to the absorption curve are shown in Figure 5.24 for several values of Krylov dimension $m$. The figures on the left are the absorption curve on a standard (linear) scale, while the figures on the right correspond to a log scale. As in the case of the Lanczos approximation to the spectral function corresponding to the fundamental electron eigenmode (Figure 5.22), the Lanczos process captures the initial take off from zero very well, even for small values of $m$. That is, the convergence is from low energy to higher energy. The more precisely we want to capture higher energy phenomena, the larger we need to take the Krylov dimension $m$.

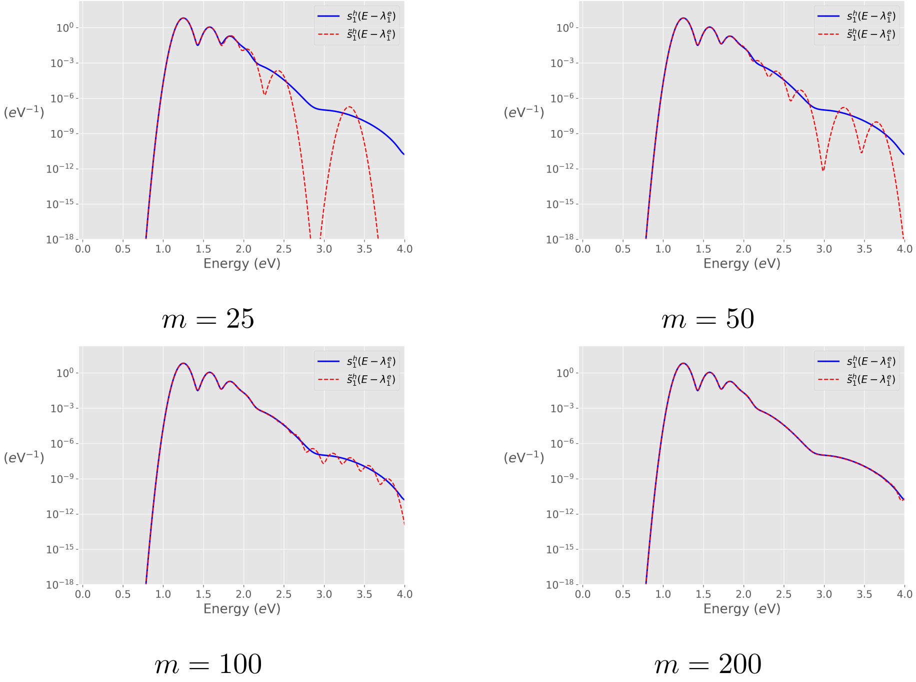
The last part of the computation is that of the joint density of states. We review the two methods for approximating the joint density of states next.



$m = 25$ $m = 50$

$m = 100$ $m = 200$

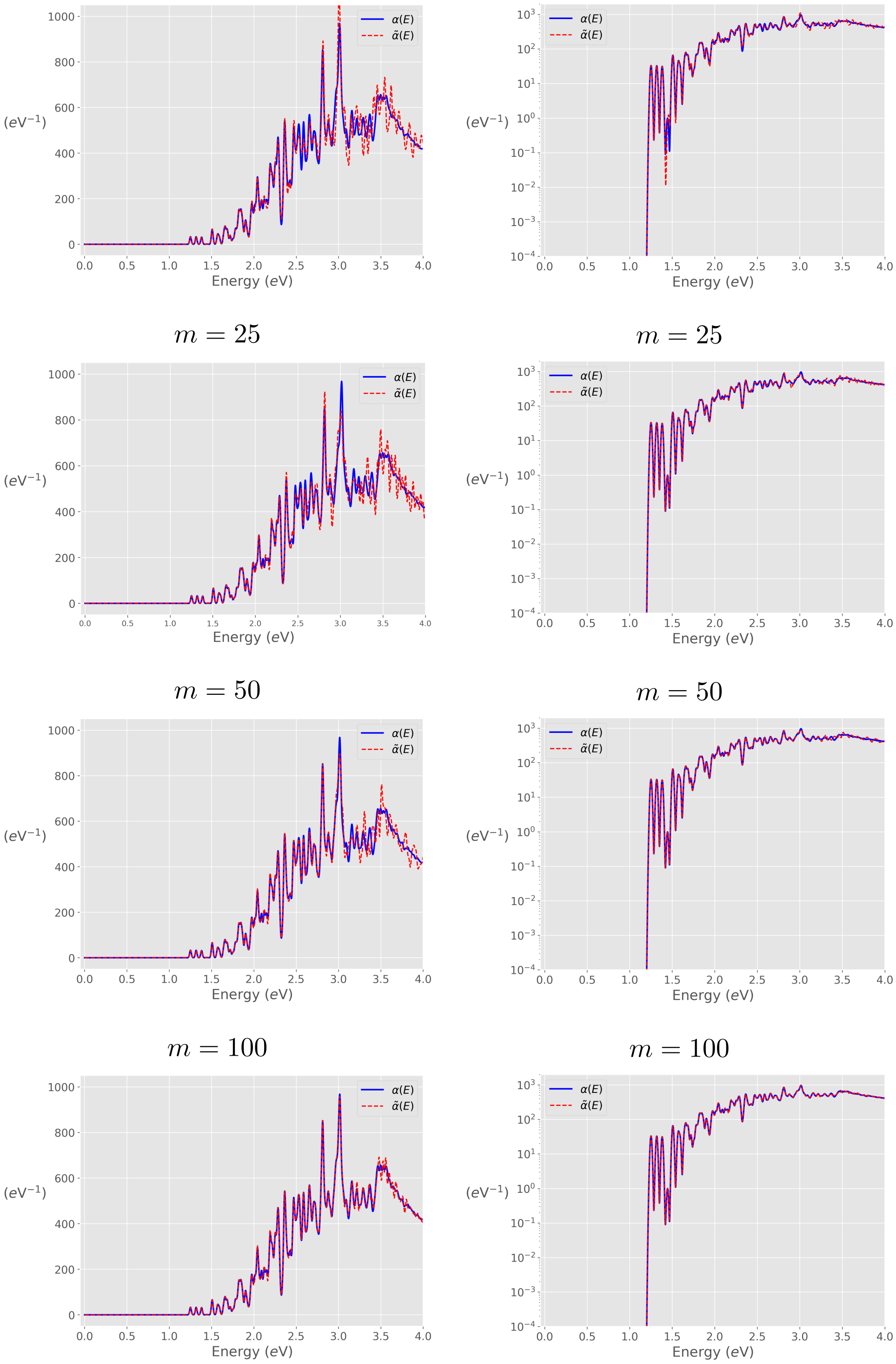
Figure 5.22: Lanczos approximation to the spectral function corresponding to the fundamental electron eigenfunction on a log scale for $\sigma = 10$ m$e$V and various values of Krylov dimension $m$.



$m = 25$ $m = 50$

$m = 100$ $m = 200$

Figure 5.23: Lanczos approximation to the spectral function corresponding to the fundamental electron eigenfunction on a log scale for $\sigma = 50$ m$e$V and various values of Krylov dimension $m$.



$m = 25$ $m = 25$

$m = 50$ $m = 50$

$m = 100$ $m = 100$

$m = 200$ $m = 200$

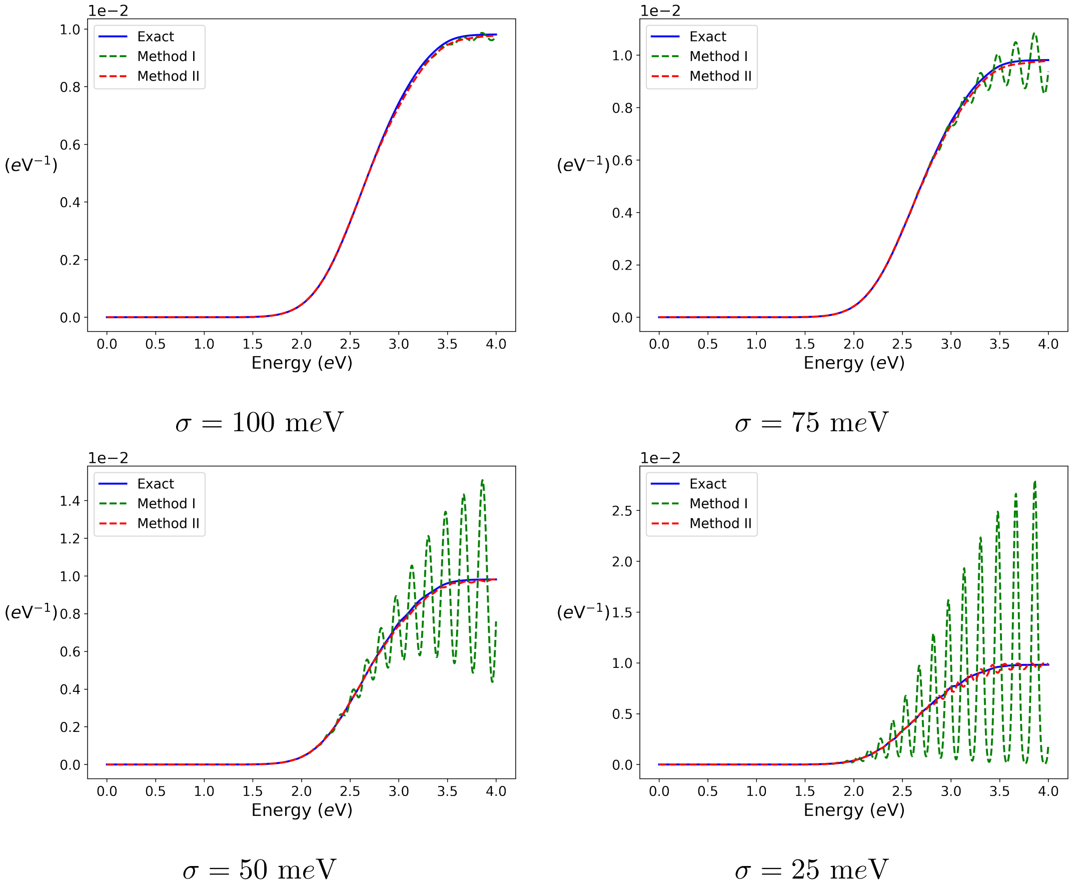
Figure 5.24: Lanczos approximation to the absorption curve for a one dimensional InGaN alloy with twenty percent indium for various Krylov dimensions $m$ and $\sigma = 10$ meV.

### 5.5.3 Joint Density of States Comparison

In this section we compare and contrast method I and method II for approximating the joint density of states. Recall, method I is based on Gaussian quadrature, and so recreates a maximal number of moments for a given Krylov parameter $m$. Method II is based on the convolution of the Lanczos approximations to the densities of state for the electrons and holes, and matches the same number of moments as method I. As discussed in the previous section (and visualized in Figure 5.17) we need approximations to the electron density of states in order to know how many electron eigenpairs to compute in order to approximate the absorption curve, i.e., we need to approximate $n_e$ defined in (5.43). Therefore, we are able to reuse the electron density of states computation for the method II approximation to the joint density of states.

For the first test, we again consider the same twenty percent indium content alloy on domain $\Omega = [0, 5000]$ used in the previous section. For the Monte Carlo method, we use ten trial vectors, and compute approximations to the joint density of states using Krylov parameter $m = 800$. Note that this value is significantly higher than that used in the approximation of the electron density of states for use in approximating the number of electron eigenpairs less than or equal to a certain energy ($m = 800$ versus $m = 150$). This is due to the fact that the exact joint density of states has $15,000^2$ energies, as opposed to the exact density of states, which has $15,000$ energies. The method I and method II Lanczos approximations to the joint density of states can be seen in Figure 5.25, along with the exact joint density of states.

From Figure 5.25 we see that both methods approximate the exact joint density of states for large values of $\sigma$, e.g., $\sigma = 100$ m$e$V. However, when we decrease $\sigma$, in order to gain more resolution, we see the standard Lanczos phenomena occur in the method I approximation. Namely, that of oscillating about the exact solution due to an insufficiently small Krylov parameter. This is easy to explain when we consider how method I is approximating the joint density of states. Because method I relies on performing the Lanczos process on a matrix which is a Kronecker sum of two matrices of size $15,000 \times 15,000$, we are performing the Lanczos process on a matrix of order $15,000^2 = 225,000,000$. This is a large matrix indeed! Therefore, taking a Krylov parameter of $m = 800$, or approximately $3.56 \times 10^{-4}\%$ of $225,000,000$, is woefully inadequate for a high resolution approximation to the joint density of states. Because



$\sigma = 100$ meV

$\sigma = 75$ meV

$\sigma = 50$ meV

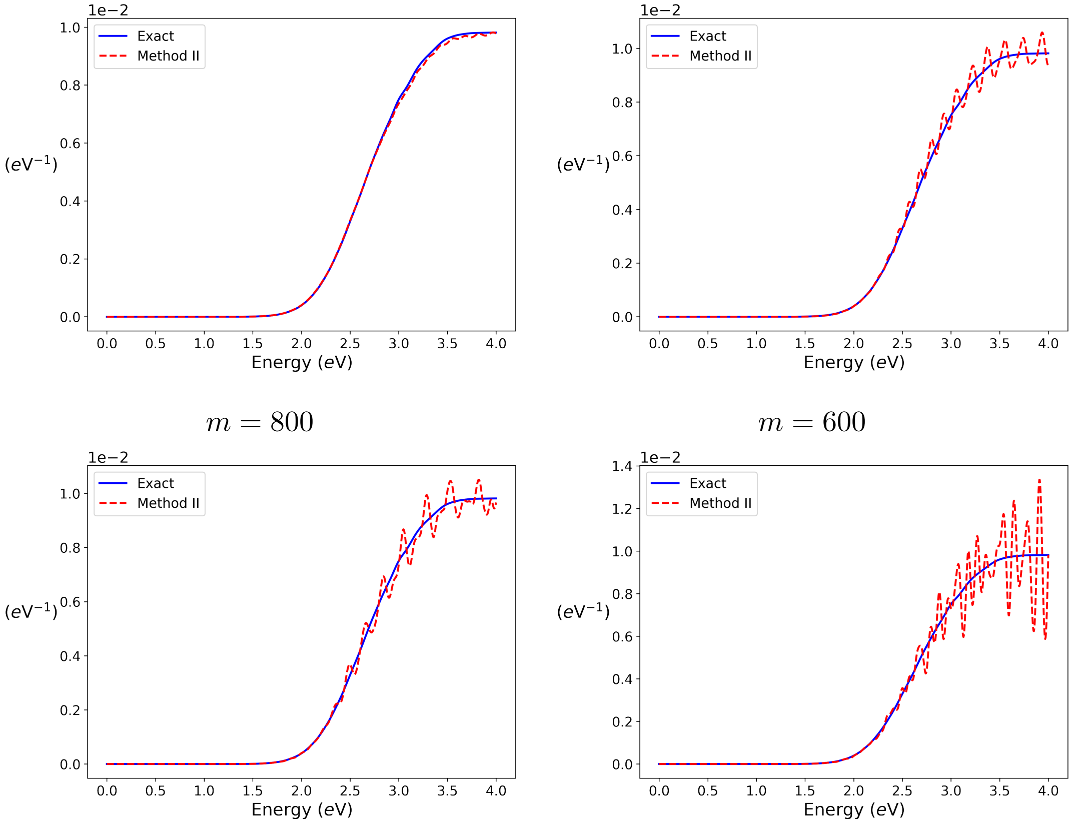
$\sigma = 25$ meV

Figure 5.25: Comparison of method I and method II for approximating the joint density of states with 10 trials and $m = 800$.

of this weakness in the method I approximation, we choose to use method II in the rest of this Chapter when approximating joint densities of state.

Next, we consider how small we can take the Krlov parameter $m$, and still obtain an acceptable approximation to the joint density of states. Figure 5.26 shows the approximations of the joint density of states by method II for several values of $m$ with the regularization parameter fixed at $\sigma = 50$ meV. From Figure 5.26, we see that decreasing the value of $m$ quickly compromises the method II joint density of states approximation. Note that this is not unique to the Lanczos approximation to the joint density of states, and also occurs in the Lanczos approximation to the density of states [33, 67].

Note that, while the approximation to the joint density of states is (visually) of



$m = 800$ $m = 600$

$m = 400$ $m = 200$

Figure 5.26: Method II of approximating the joint density of states for fixed regularization parameter $\sigma = 50$ m$e$V, 10 random trials, and various values of the dimension of the Krylov space $m$.

poor quality, the integrated joint density of states is still quite accurate. This is similar to approximating the number of electron eigenpairs needed for the Lanczos process. Figure 5.17 shows that the while the Lanczos approximation to the density of states may be inaccurate, Table 5.6 shows the integrated density of states can be quite accurate. We look at the similar case for the joint density of states. We denote the integrated joint density of states as

$$N_J(E) = \int_0^E n^2 J(\lambda) d\lambda = \sum_{i,j=1}^{n} U\big(E - (\lambda_i^e + \lambda_j^h)\big), \tag{5.49}$$

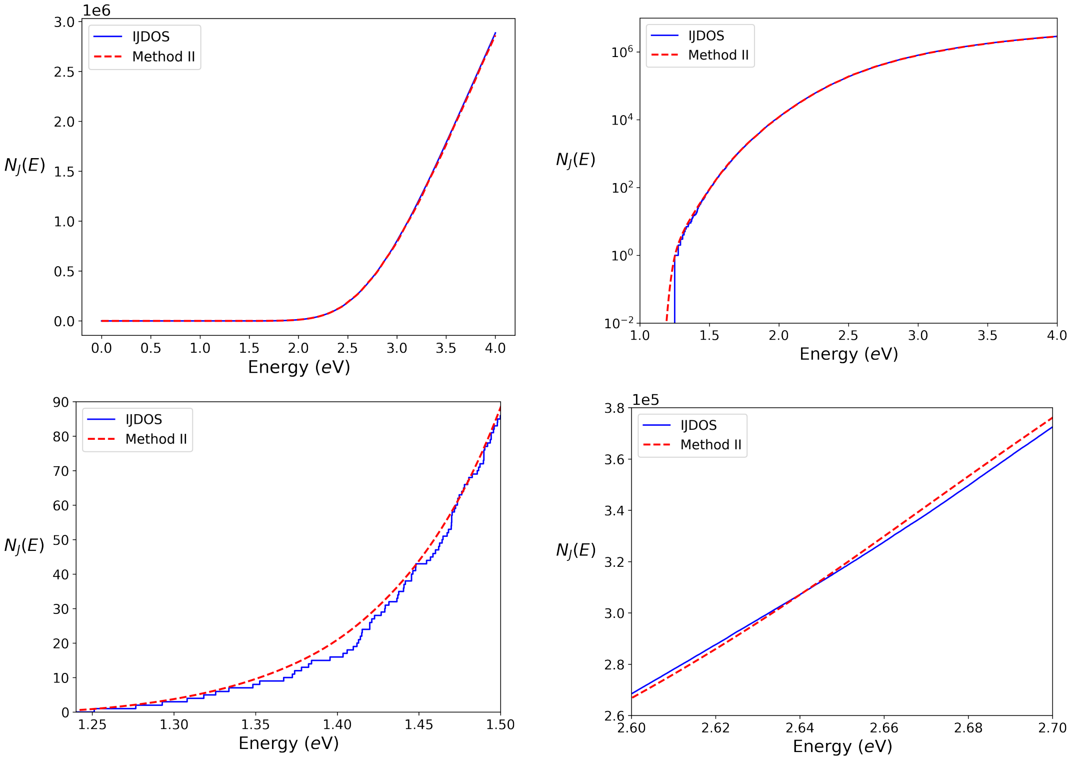


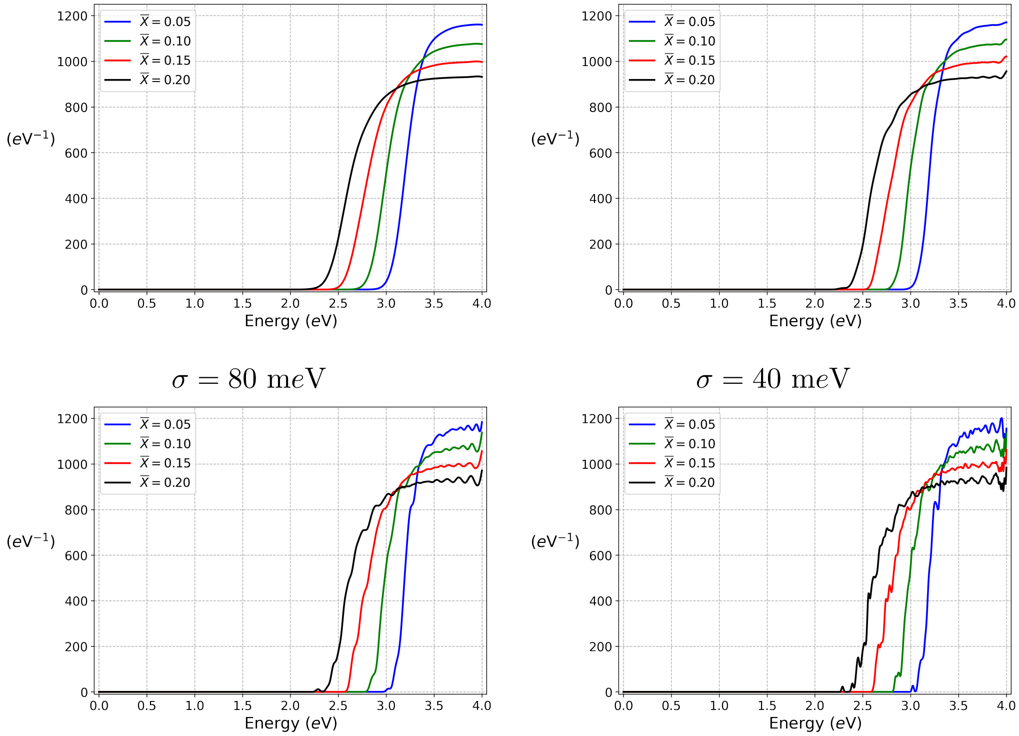
Figure 5.27: Integrated joint density of states and Lanczos approximation using $\sigma = 50$ m$e$V with $m = 400$ and ten trials on linear scale (top left) and log scale (top right). Zoom into the region of initial take off (bottom left) and middle region (bottom right).

where $U(E)$ is the Heaviside step function. For a given energy $E$, $N_J(E)$ tells us how many terms in the absorption curve have energy less than $E$. In Figure 5.27 the exact integrated joint density of states is shown in blue. Also, shown in Figure 5.27 is Lanczos approximation to $N_J(E)$, computed by replacing the exact joint density of states in the integral (5.49) with the method II Lanczos approximation using $m = 400$, $\sigma = 50$ m$e$V, and ten trials (same as Figure 5.26 (bottom left)). The integral in (5.49) is approximated using a composite trapezoidal rule. We see from Figures 5.26 and 5.27 that while the joint density of states may be inaccurate, in that it oscillates about the exact joint density of states, the integrated joint density of states approximation can be quite accurate.

## 5.6 2D Random Alloys

Now that we have given an overview of the Lanczos approximation of absorption curves and joint densities of states in one dimension, where we are able to compute the spectral quantities exactly, we move to the two dimensional case, where the exact solution is too costly to compute. Two dimensional random alloys are of practical interest, and are often used as one part of a larger three dimensional computation, or to simulate layered materials.

For each of the following computations, we use a 201 × 201 lattice, making the computational domain $\Omega = [0, 200]^2$. The mesh is obtained by uniformly discretizing the domain into unit squares, with each unit square further subdivided into two triangles. On each triangle, we use cubic Lagrange finite elements. Hence, the stiffness and mass matrices are of order $360,000$ ($360,000$ being equal to $(3 \times 200)^2$). For matrices of this size, the first step is to determine the number of electron eigenpairs needed for accurate representation of the absorption curve. For the Lanczos approximation of the densities of states, sixty-four trial vectors and a Krylov dimension of $m = 400$ are used. Again, we use a large regularization parameter $\sigma = 500$ m$e$V when replacing the exact densities of states in (5.43) with the corresponding Lanczos approximation. Using the Lanczos approximations to the densities of state, Table 5.7 shows the approximate number of electron eigenpairs necessary, as well as the exact number required. We again use a five percent fudge factor, and request $\lceil 1.05\tilde{n}_e \rceil$ electron eigenpairs from SLEPc. This ensures we compute more than enough electron eigenpairs, and allows us to report the exact number of electron eigenpairs, $n_e$, necessary for the satisfaction of criterion (5.42). Table 5.7 shows the value of the fundamental electron and hole energies, the exact number of electron eigenpairs needed, and the approximate number of electron and hole eigenpairs needed using the Lanczos approximation to the density of states (no five percent fudge factor present). Again, we see a pronounced difference in the number of electron and hole eigenpairs necessary for absorption computations. Indeed, for the four cases displayed in Table 5.7, on average, we need 10.3 times as many hole eigenpairs as electron eigenpairs, i.e., $\tilde{n}_h/\tilde{n}_e \approx 10.3$. Hence, there is significant advantage in using the Lanczos process to approximate the absorption curve, which requires either the electron or hole eigenpairs, but not both.



$\sigma$ = 80 meV $\qquad$ $\sigma$ = 40 meV

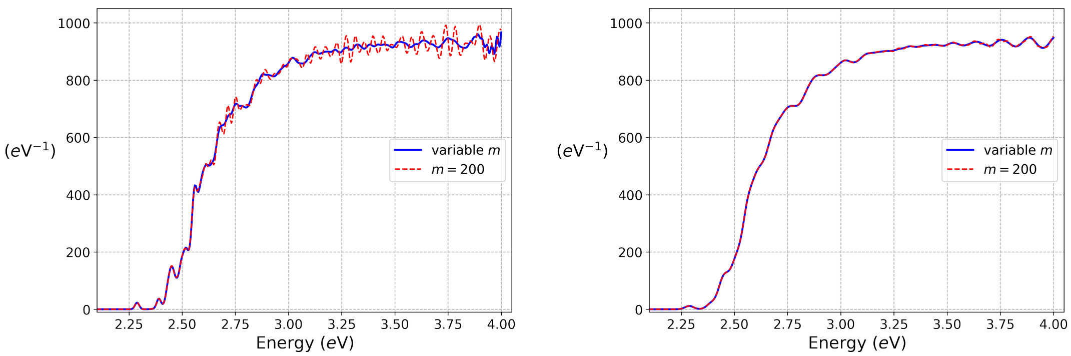
$\sigma$ = 20 meV $\qquad$ $\sigma$ = 10 meV

Figure 5.28: Two dimensional absorption curves for $\text{In}_X\text{Ga}_{1-X}\text{N}$ on a lattice of size 201 × 201 with adaptively chosen Krylov dimension using tolerance $\tau = 25$ meV.

| $\overline{X}$ | $\lambda_1^e$ (eV) | $\lambda_1^h$ (eV) | $n_e$ | $\lceil \tilde{n}_e \rceil$ | $\lceil \tilde{n}_h \rceil$ |
|---|---|---|---|---|---|
| 0.05 | 2.07 | 0.94 | 1162 | 1176 | 10706 |
| 0.10 | 1.93 | 0.87 | 1306 | 1312 | 12980 |
| 0.15 | 1.78 | 0.79 | 1434 | 1428 | 15425 |
| 0.20 | 1.60 | 0.66 | 1581 | 1577 | 18258 |

Table 5.7: Two dimensional $\text{In}_X\text{Ga}_{1-X}\text{N}$ absorption curve computation data.

Once the electron eigenpairs are determined, we are prepared to use the Lanczos process to approximate the absorption curves. This is done for bulk indium concentrations of five, ten, fifteen, and twenty percent. The results are shown in Figure 5.28. For the two dimensional case, we adaptively determine the correct Krylov dimension $m$ using



$\sigma$ = 10 m$e$V

$\sigma$ = 20 m$e$V

Figure 5.29: Lanczos approximation of two dimensional absorption curve for InGaN alloy ($\overline{X}$ = 0.20) with an adaptively chosen Krylov dimension (using tolerance $\tau$ = 25 m$e$V) and fixed Krylov dimension $m$ = 200.

the tolerance $\tau = 2.5\sigma$ with $\sigma$ = 10 m$e$V, as discussed in Section 5.1.9. This is accomplished by starting the Lanczos algorithm with an electron eigenvector, $x^e$, and every ten iterations we compute the Ritz values, and check if the gap between Ritz values less than $4 - \lambda^e$ is smaller than $\tau$ = 25 m$e$V where $\lambda^e$ is the eigenvalue corresponding to $x^e$. If so, the computation is terminated, if not, the computation is continued for another ten iterations before another check is performed. Because of our choice of $\tau$ = 25 m$e$V, we have confidence in our absorption curves for regularization parameter as small as $\sigma$ = 10 m$e$V.

A natural question one might raise: is it necessary to use a variable Krylov dimension? The answer depends on the level of specificity in which we wish to approximate the absorption curve. Figure 5.29 shows two absorption curves for an example with a twenty percent indium fraction (same as that seen in Figure 5.28 for $\overline{X}$ = 0.20 and $\sigma$ = 10 m$e$V). The absorption curve in blue adaptively chooses the Krylov dimension using $\tau$ = 25 m$e$V, and the dashed curved in red uses a small fixed value $m$ = 200. For the variable case, the maximum Krylov dimension is $m$ = 1510 (for the fifth electron eigenfunction) and decreases down to $m$ = 130. The exact value of the Krylov dimension for each energy level can be seen in Figure 5.30 (right). From Figure 5.29, we see that the absorption curve computed using $m$ = 200 matches the absorption curve computed using variable $m$ at the take off of the curve, and oscillates around it for the remainder

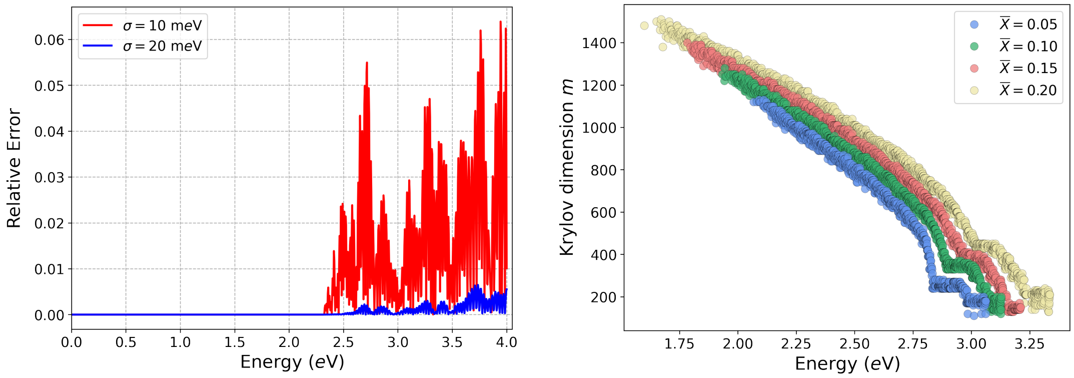
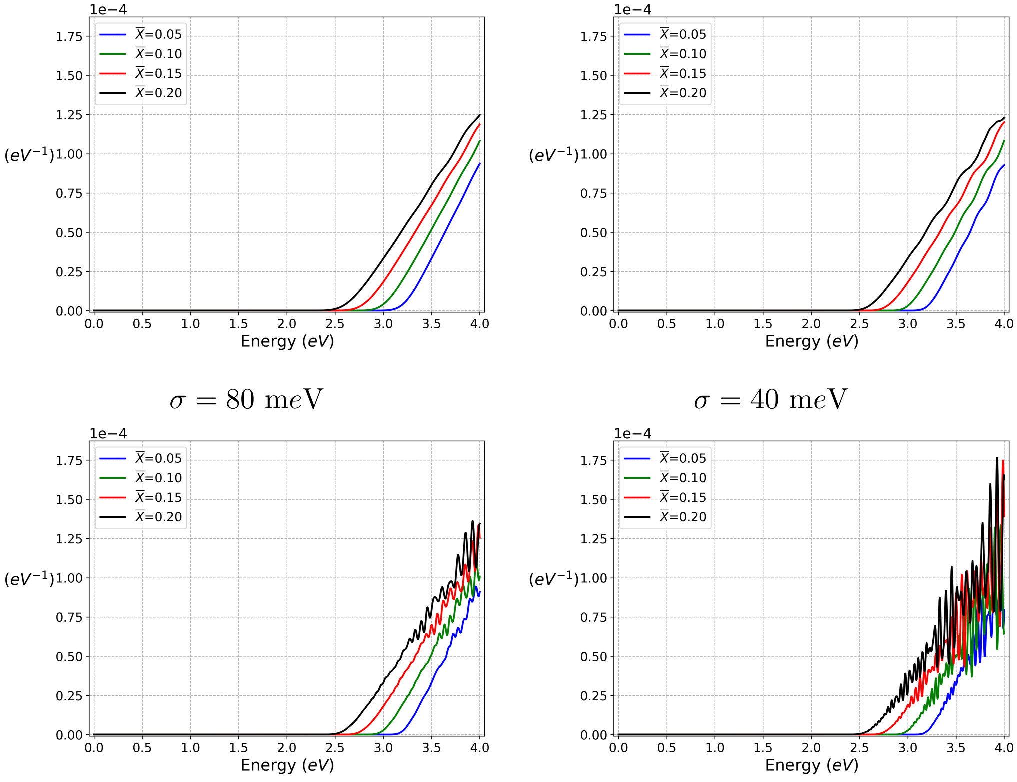


Figure 5.30: Relative error of two curves seen in Figure 5.29 (left). Krylov dimension for two dimensional absorption curves (see Figure 5.28) using tolerance $\tau = 25$ meV (right).

of the curve. But, when using a slightly larger regularization factor $\sigma$, these oscillations are eliminated, and we see how well using a small value of $m$ performs, for much less work. The relative error in the absorption curve using a variable Krylov dimension and a small fixed value is seen in Figure 5.30 (left). We see that even for $\sigma = 10$ meV, the relative error in the two curves stays below seven percent. When considering the relative error when using regularization factor of $\sigma = 20$ meV, the error stays below one percent. Based on this example, it seems unnecessary to perform the extra work of using a variable Krylov dimension. Oftentimes, one is interested in averaging absorption curves over many different random realizations of the InGaN lattice. If this is the purpose, rather than computing one realization with high fidelity, then using a small fixed value of the Krylov parameter may be a better use of resources.

Lastly, we investigate the joint density of states computation using method II in two spatial dimensions. The joint density of states for several values of regularization parameter $\sigma$ are shown in Figure 5.31. These are computed using 64 trial vectors and a Krylov dimension of $m = 400$. One thing to note is the small values on the y-axis in Figure 5.31. This is due to the prefactor of $1/n^2$ in the definition of $J(E)$. For these two dimensional problems $n = 360,000$, and so the factor of $1/n^2$ is of order $10^{-12}$. In Figure 5.31 we see that for small values of $\sigma$, e.g., $\sigma = 10$ meV or $\sigma = 20$ meV, the Lanczos approximation to the joint density of states is highly oscillatory. This is especially prevalent for the regularization parameter $\sigma = 10$ meV. This is due to the



$\sigma = 80$ meV

$\sigma = 40$ meV

$\sigma = 20$ meV

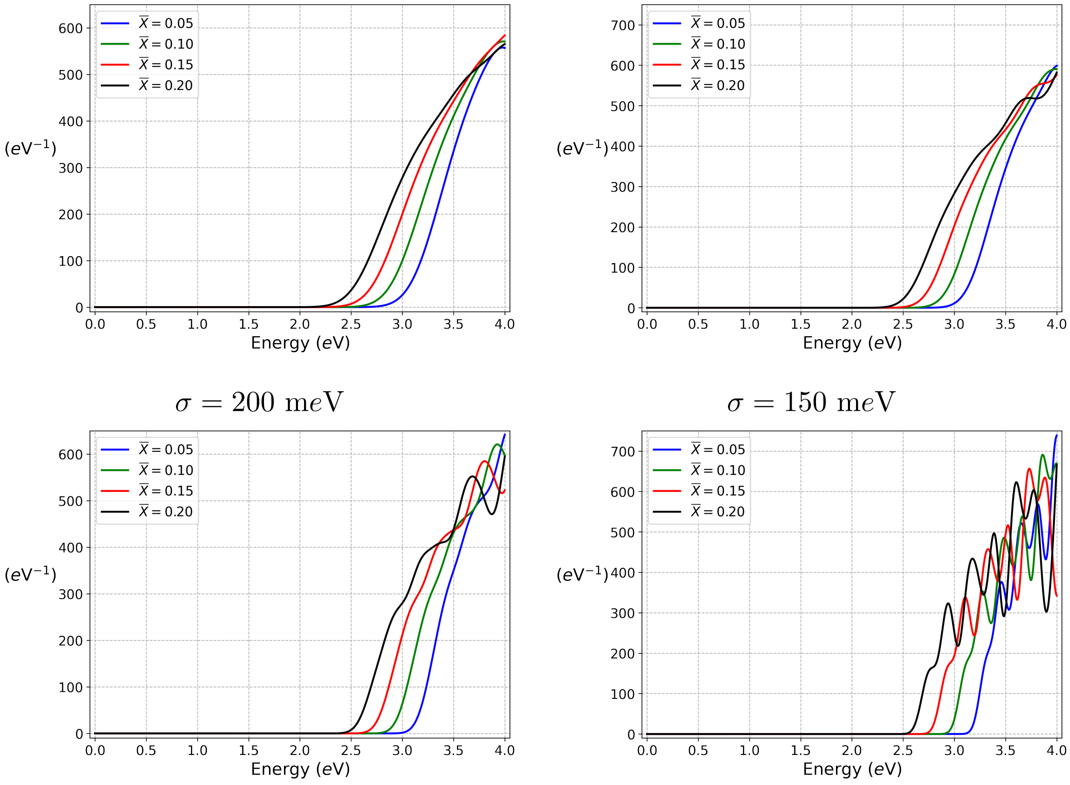
$\sigma = 10$ meV

Figure 5.31: Two dimensional Lanczos method II approximation to the joint density of states for $\mathrm{In}_X\mathrm{Ga}_{1-X}\mathrm{N}$ lattice of size $201 \times 201$ using Krylov dimension $m = 400$ and 64 trial vectors.

smaller value of the Krylov parameter $m = 400$. Recall for the one dimensional case, we needed a Krylov dimension of $m = 800$ to match the exact joint density of states for the regularization parameter $\sigma = 50$ meV (see Figure 5.26). Here the matrix sizes are much larger, and so if we require a high accuracy approximation to the joint density of states, then we need to compensate for this fact with a larger value of $m$.

## 5.7 3D Random Alloys

Finally we are ready for a full three dimensional realization of $\mathrm{In}_X\mathrm{Ga}_{1-X}\mathrm{N}$ lattices. In this section we use a $51 \times 51 \times 51$ lattice for bulk indium fractions between five



$\sigma$ = 200 meV

$\sigma$ = 150 meV

$\sigma$ = 100 meV

$\sigma$ = 50 meV

Figure 5.32: Three dimensional absorption curves for $\mathrm{In}_X\mathrm{Ga}_{1-X}\mathrm{N}$ on a lattice of size $51 \times 51 \times 51$ using Krylov dimension $m = 2000$.

and twenty percent. For the spatial discretization, each unit cube is subdivided into six tetrahedron, and quadratic Lagrange finite elements are used on each tetrahedron. This results in matrices of order $n = 1,000,000 = (2 \times 50)^3$. The Lanczos approximation to the absorption curve for Krylov parameter $m = 2000$ are shown in Figure 5.32. The statistics for the four computations are shown in Table 5.8. For the four cases listed in Table 5.8, on average, $\tilde{n}_h/\tilde{n}_e = 29.7$. In other words, we need thirty times as many hole eigenpairs as we do electron eigenpairs to satisfy the criterion (5.42). Therefore, the Lanczos approximation becomes more economical as the spatial dimension increases due to the increasing number of hole eigenpairs required.

| $\overline{X}$ | $\lambda_1^e$ ($e$V) | $\lambda_1^h$ ($e$V) | $n_e$ | $\lceil \tilde{n}_e \rceil$ | $\lceil \tilde{n}_h \rceil$ |
|---|---|---|---|---|---|
| 0.05 | 2.15 | 0.94 | 462 | 518 | 12453 |
| 0.10 | 2.01 | 0.90 | 516 | 572 | 16282 |
| 0.15 | 1.89 | 0.83 | 620 | 648 | 20155 |
| 0.20 | 1.76 | 0.79 | 624 | 687 | 24212 |

Table 5.8: Three dimensional $\mathrm{In}_X\mathrm{Ga}_{1-X}\mathrm{N}$ absorption curve data.

One thing to note from the absorption curves in Figure 5.8 is the larger regularization parameter used compared to one and two dimensional computations. The reason for using a larger regularization parameter is simple. If we consider the twenty percent bulk indium fraction case, for the first $n_e = 624$ electron eigenvalues computed, the largest spectral gap is given by

$$\max_i |\lambda_{i+1}^e - \lambda_i^e| = 0.088\ e\text{V}. \tag{5.50}$$

We take a closer look at the eigenvalues responsible for this spectral gap in Table 5.9. We see that there is a cluster of energies near 2.87 $e$V and another near 2.96 $e$V, but a relatively large gap between the two. Individual clusters of energies can be seen in the absorption curve in Figure 5.32 for $\sigma = 50$ m$e$V. These clusters of energies with gaps in between are due to the small size of the lattice, i.e., the gaps are due to the small number of possible random configurations of InN and GaN. With a larger lattice more configurations are possible, e.g., regions dense with InN or GaN, and as a consequence of additional configurations, there will be no gaps in the energies of the system. Therefore, while we are attempting to model a physical system using the effective mass Schrödinger equation, due to the small lattice size, we are seeing results which are nonphysical.

| | | | |
|---|---|---|---|
| $\lambda_{457}^e$ | 2.8707 | $\lambda_{462}^e$ | 2.9611 |
| $\lambda_{458}^e$ | 2.8711 | $\lambda_{463}^e$ | 2.9616 |
| $\lambda_{459}^e$ | 2.8716 | $\lambda_{464}^e$ | 2.9624 |
| $\lambda_{460}^e$ | 2.8718 | $\lambda_{465}^e$ | 2.9628 |
| $\lambda_{461}^e$ | 2.8726 | $\lambda_{466}^e$ | 2.9632 |

Table 5.9: Gap in electron energies.

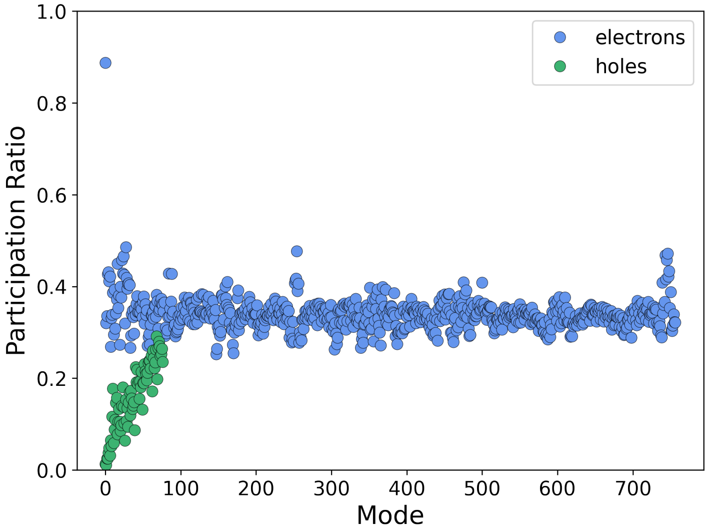
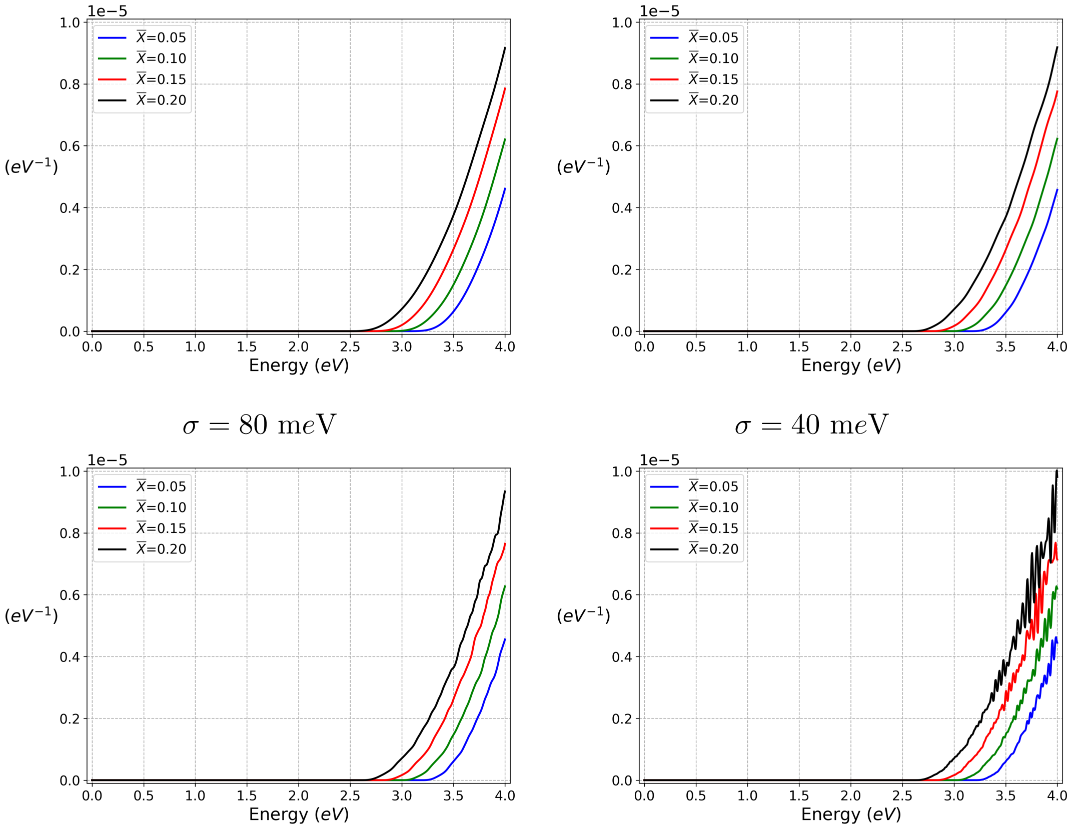


Figure 5.33: Participation ratio of first 757 electron wavefunctions and 77 hole wavefunctions for $\overline{X} = 0.20$ on a $51 \times 51 \times 51$ lattice.

The participation ratio, see (5.48), for the first 757 electron wavefunctions and 77 hole wavefunctions for the $\overline{X} = 0.20$ case is shown in Figure 5.33. The first thing to notice in Figure 5.33 is the magnitude of the participation ratio for the fundamental electron wavefunction. Recall the participation ratio of a constant is unity. Figure 5.33 tells us the fundamental electron wavefunction is nearly constant, or rather, a small perturbation of a constant. The remaining electron eigenfunctions, have an average participation ratio of 0.34, which is close to, $(2/3)^3 \approx 0.30$, the participation ratio of sinusoids in three dimensions. Essentially, the electron wavefunctions are the solutions of the Laplace eigenvalue problem on a cube with periodic boundary conditions. We also see that the hole wavefunctions, while initially localized, begin to delocalize immediately. However, on a larger lattice, there would be more local minima in the valence band energy, and hence more localized hole wavefunctions.

The behavior of the electron wavefunctions is a consequence of two factors. The first being the small lattice size, the second being the lack of variation in the indium fraction. As mentioned regarding the gap in electron energies, the small lattice allows relatively few InN and GaN configurations, and in particular, there is no region dense with InN in the lattice, which would create a region of low potential surrounded by high barriers for the electron wavefunctions to localize inside. Secondly, due to the chosen modeling paradigms, the indium fraction is averaged over cations within two lattice spaces. As the spatial dimension increases, the averaging is performed over more lattice sites. Recall the one dimensional example of one InN cation surrounded by fifty GaN cations on both sides. The indium fraction for this example, seen in Figure 5.13, reached twenty percent. For a similar three dimensional example on an $11 \times 11 \times 11$ lattice with one InN cation in the center, and all remaining lattice sites occupied by GaN, the indium fraction reaches a maximum of approximately 0.83%. Because there are more nearest neighbors in a three dimensional lattice, as opposed to a one dimensional lattice, there will be decreased fluctuation in indium fraction. This will in turn cause less fluctuation in the conductance and valence band energies, and less localization will occur.

Lastly, the joint density of states for the four bulk indium fractions can be seen in Figure 5.34. For these computations, a Krylov dimension of $m = 500$ was used for twenty trial vectors. Notice the difference in the two dimensional joint density of states approximations seen in Figure 5.31 and the three dimensional ones seen in Figure 5.34. For larger values of the regularization parameter $\sigma$, the two and three dimensional curves look qualitatively similar. On the other hand, for smaller values of $\sigma$, e.g., $\sigma = 10$ m$e$V, there is significantly less oscillations. This is due to a smaller gap in the nodes for the Lanczos approximation to the joint density of states in three dimensions. To investigate, we consider one of the electron density of states approximations and one of the hole density of states approximations. Recall, to create the joint density of states approximation by method II, we add all possible combinations of the nodes and multiply all possible combinations of the weights. This results in $25,000 = 500^2$ nodes and weights. For one specific realization, of the $25,000$ nodes, 483 of them are less than $\overline{E} = 4$ $e$V, and the largest gap between these 483 nodes is approximately 14 m$e$V. Assuming the other trials have similar results, this is why we see some small oscillation in Figure 5.34 for $\sigma = 10$ m$e$V, and none for $\sigma = 20$ m$e$V and larger.



$\sigma$ = 80 meV

$\sigma$ = 40 meV

$\sigma$ = 20 meV

$\sigma$ = 10 meV

Figure 5.34: Three dimensional joint densities of state for $\text{In}_X\text{Ga}_{1-X}\text{N}$ on a lattice of size $51 \times 51 \times 51$ using twenty trials and Krylov dimension $m = 500$.

# Chapter 6

# Conclusion

In this thesis we analyzed the Lanczos process for approximating spectral functions, and proposed methods for extending the Lanczos process for the computation of joint spectral quantities. The joint spectral quantities examined were the joint density of states and the joint spectral function, both of which have practical applications in semiconductor modeling. Two methods for approximating the joint density of states were proposed and applied to a random alloy modeled using the effective mass Schrödinger equation. The first method relies on realizing the joint density of states as the density of states for a larger matrix, while the second relies on the notion of convolution of measures. The other joint spectral quantity considered is the joint spectral function, which, if computed exactly, requires complete knowledge of the spectrum of two operators. The Lanczos approximation of the joint spectral function, is realized by rewriting the joint spectral function as a sum of spectral functions, each of which can be approximated by the Lanczos process. At the heart of both methods is a deep connection between the Lanczos algorithm for partially tridiagonalizing a matrix and Gauss quadrature.

The Lanczos type methods devised were seen to be accurate and efficient for approximating joint spectral quantities pertaining to a random InGaN alloy. This was determined by comparing the Lanczos methods with the exact solution in one (spatial) dimension. Using the knowledge gained from one dimension, we were able to approximate joint spectral quantities in two and three dimensions. For these cases, little work has been done due to the high cost of diagonalizing Schrödinger operators.

With the advantages of the Lanczos process, there remain a few drawbacks which must be mentioned. First and foremost, is the loss of orthogonality in the Lanczos vectors when performing Lanczos partial tridiagonalizations. The beauty of the simple three-term Lanczos recurrence is lost when moving from theory to finite precision computations. In order to avoid the loss of orthogonality, full Gram–Schmidt orthogonalization was used in this thesis. While not the most economical, essentially dismissing the advantages of symmetry (using the Arnoldi algorithm on a symmetric operator), it is the most robust. However, this use of full orthogonalization requires storage of all Lanczos vectors, and each iteration more orthogonalization steps are necessary.

One feature lacking in the Lanczos process is a posteriori error estimates. When working with semiconductor applications to spectral and joint spectral quantities, experience was necessary to determine the correct Krylov dimension and number of trial vectors to use. While we developed a heuristic for determining when the Lanczos approximation to a spectral function is adequate, namely that of continuing with the Lanczos process until the gap between Ritz values fell below a certain tolerance, this heuristic requires a priori knowledge of the operator spectrum. More beneficial would be computable error bounds determined by a Lanczos partial tridiagonalization of some order and a regularization parameter determining how closely we wish to approximate the Dirac measure.

For the computation of joint spectral function, one set of eigenpairs or the other is required. This essentially halves the work, with the joint spectral function requiring eigenvalues and eigenvectors of two distinct operators. However, for large problems, solving for the eigenpairs of any operator is a challenging task. It would be nice to determine a Monte Carlo type method for approximating the joint spectral function, similar to how densities of state and joint densities of state are approximated.

There are many avenues for continuing work described in this thesis in the areas of numerical analysis and engineering applications. One such application is in the computation of local densities of states. For a Hamiltonian with energies, $E_i$, and corresponding wavefunctions, $\psi_i$, $i = 1, 2, \ldots$, the local density of states is given by $\mathrm{LDOS}(x, E) = \sum_i |\psi_i(x)|^2 \delta(E - E_i)$. With the wavefunctions $L^2$ normalized, it is easy to see that the density of states is the integral of the local density of states over the domain. The local density of states is a quantity of great interest to physicists and engineers, and a Lanczos process type method seems natural for approximation.

Another area of work includes comparing the Lanczos type methods described in this thesis with the Kernel Polynomial Method (KPM). This thesis focused exclusively on using the Lanczos process to approximate joint spectral quantities. However, several methods devised in this thesis naturally lend themselves, without modification, to approximation by the KPM. Naturally, it makes sense to compare these two methods in terms of computational timing and accuracy of approximation. The KPM method may be advantageous in that it does not require costly Gram–Schmidt orthogonalization, the main weakness of the Lanczos process.

In conclusion, the Lanczos type methods for approximating joint spectral quantities are reliable and economical. The main purpose of the Lanczos process is to avoid costly eigenvalue solves, which the approximation methods derived in this thesis accomplish for the joint density of states. For the joint spectral function, we reduce the problem in half, and only require the spectrum of one operator. When only interested in the joint spectral function for a small range of values, only a portion of the spectrum of one operator is required. This was seen to be extremely beneficial in random InGaN alloy applications, where far fewer eigenpairs for the electron Hamiltonian were required than for the hole Hamiltonian.

# References

[1] Robert A. Adams and John J. F. Fournier. *Sobolev Spaces*, volume 140 of *Pure and Applied Mathematics (Amsterdam)*. Elsevier/Academic Press, Amsterdam, second edition, 2003.

[2] D. N. Arnold. *A Concise Introduction to Numerical Analysis.* online, 2001.

[3] W. E. Arnoldi. The principle of minimized iteration in the solution of the matrix eigenvalue problem. *Quart. Appl. Math.*, 9:17–29, 1951.

[4] Haim Avron and Sivan Toledo. Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix. *J. ACM*, 58(2):Art. 8, 17, 2011.

[5] Zhaojun Bai, James Demmel, Jack Dongarra, Axel Ruhe, and Henk van der Vorst, editors. *Templates for the Solution of Algebraic Eigenvalue Problems*, volume 11 of *Software, Environments, and Tools.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2000. A practical guide.

[6] Zhaojun Bai, Mark Fahey, Gene H Golub, M Menon, and E Richter. Computing partial eigenvalue sum in electronic structure calculations. Technical report, Tech. Report SCCM-98-03, Stanford University, 1998.

[7] Zhaojun Bai and Gene H. Golub. Bounds for the trace of the inverse and the determinant of symmetric positive definite matrices. volume 4, pages 29–38. 1997. The heritage of P. L. Chebyshev: a Festschrift in honor of the 70th birthday of T. J. Rivlin.

[8] Satish Balay, Shrirang Abhyankar, Mark F. Adams, Jed Brown, Peter Brune, Kris Buschelman, Lisandro Dalcin, Alp Dener, Victor Eijkhout, William D. Gropp, Dmitry Karpeyev, Dinesh Kaushik, Matthew G. Knepley, Dave A. May, Lois Curfman McInnes, Richard Tran Mills, Todd Munson, Karl Rupp, Patrick Sanan, Barry F. Smith, Stefano Zampini, Hong Zhang, and Hong Zhang. PETSc users manual. Technical Report ANL-95/11 - Revision 3.14, Argonne National Laboratory, 2020.

[9] Daniele Boffi. Finite element approximation of eigenvalue problems. *Acta Numer.*, 19:1–120, 2010.

[10] E. W. Cheney. *Introduction to approximation theory.* AMS Chelsea Publishing, Providence, RI, 1998. Reprint of the second (1982) edition.

[11] P. G. Ciarlet and J.-L. Lions, editors. *Handbook of numerical analysis. Vol. II.* Handbook of Numerical Analysis, II. North-Holland, Amsterdam, 1991. Finite element methods. Part 1.

[12] Philippe G. Ciarlet. *The finite element method for elliptic problems*, volume 40 of *Classics in Applied Mathematics.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2002. Reprint of the 1978 original [North-Holland, Amsterdam; MR0520174 (58 #25001)].

[13] Germund Dahlquist, Stanley C. Eisenstat, and Gene H. Golub. Bounds for the error of linear systems of equations using the theory of moments. *J. Math. Anal. Appl.*, 37:151–166, 1972.

[14] Philip J. Davis and Philip Rabinowitz. *Methods of numerical integration.* Dover Publications, Inc., Mineola, NY, 2007. Corrected reprint of the second (1984) edition.

[15] Lawrence C. Evans. *Partial differential equations*, volume 19 of *Graduate Studies in Mathematics.* American Mathematical Society, Providence, RI, second edition, 2010.

[16] Walter Gautschi. *Orthogonal Polynomials: Computation and Approximation.* Numerical Mathematics and Scientific Computation. Oxford University Press, New York, 2004. Oxford Science Publications.

[17] Gene H. Golub and Gérard Meurant. *Matrices, Moments and Quadrature with Applications.* Princeton Series in Applied Mathematics. Princeton University Press, Princeton, NJ, 2010.

[18] Gene H. Golub and Charles F. Van Loan. *Matrix Computations.* Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, fourth edition, 2013.

[19] Gene H. Golub and John H. Welsch. Calculation of Gauss quadrature rules. *Math. Comp. 23 (1969), 221-230; addendum, ibid.*, 23(106, loose microfiche suppl):A1–A10, 1969.

[20] Claudine Hermann and Claude Weisbuch. $\vec{\mathrm{k}} \cdot \vec{\mathrm{p}}$ perturbation theory in iii-v compounds and alloys: a reexamination. *Phys. Rev. B*, 15:823–833, Jan 1977.

[21] Vicente Hernandez, Jose E. Roman, and Vicente Vidal. SLEPc: a scalable and flexible toolkit for the solution of eigenvalue problems. *ACM Trans. Math. Software*, 31(3):351–362, 2005.

[22] Roger A. Horn and Charles R. Johnson. *Topics in matrix analysis.* Cambridge University Press, Cambridge, 1994. Corrected reprint of the 1991 original.

[23] Roger A. Horn and Charles R. Johnson. *Matrix analysis.* Cambridge University Press, Cambridge, second edition, 2013.

[24] M. F. Hutchinson. A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. *Comm. Statist. Simulation Comput.*, 19(2):433–450, 1990.

[25] Ilse C. F. Ipsen and Carl D. Meyer. The idea behind Krylov methods. *Amer. Math. Monthly*, 105(10):889–899, 1998.

[26] W. Kahan and B. Parlett. *Sparse Matrix Computation*, chapter How Far Should You Go With the Lanczos Algorithm? Academic Press, 1976.

[27] Daniel Kressner and Christine Tobler. Krylov subspace methods for linear systems with tensor product structure. *SIAM J. Matrix Anal. Appl.*, 31(4):1688–1714, 2009/10.

[28] Erwin Kreyszig. *Introductory functional analysis with applications*. Wiley Classics Library. John Wiley & Sons, Inc., New York, 1989.

[29] Cornelius Lanczos. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *J. Research Nat. Bur. Standards*, 45:255–282, 1950.

[30] Chi-Kang Li, Marco Piccardo, Li-Shuo Lu, Svitlana Mayboroda, Lucio Martinelli, Jacques Peretti, James S Speck, Claude Weisbuch, Marcel Filoche, and Yuh-Renn Wu. Localization landscape theory of disorder in semiconductors. iii. application to carrier transport and recombination in light emitting diodes. *Physical Review B*, 95(14):144206, 2017.

[31] Ruipeng Li, Yuanzhe Xi, Lucas Erlandson, and Yousef Saad. The eigenvalues slicing library (EVSL): algorithms, implementation, and software. *SIAM J. Sci. Comput.*, 41(4):C393–C415, 2019.

[32] Lin Lin. Randomized estimation of spectral densities of large matrices made accurate. *Numer. Math.*, 136(1):183–213, 2017.

[33] Lin Lin, Yousef Saad, and Chao Yang. Approximating spectral densities of large matrices. *SIAM Rev.*, 58(1):34–65, 2016.

[34] Peter R. Lipow and Frank Stenger. How slowly can quadrature formulas converge? *Math. Comp.*, 26:917–922, 1972.

[35] Anders Logg, Kent-Andre Mardal, and Garth N. Wells, editors. *Automated solution of differential equations by the finite element method*, volume 84 of *Lecture Notes in Computational Science and Engineering*. Springer, Heidelberg, 2012. The FEniCS book.

[36] Gérard Meurant and Zdeněk Strakoš. The Lanczos and conjugate gradient algorithms in finite precision arithmetic. *Acta Numer.*, 15:471–542, 2006.

[37] Stephen Karrer O'leary. An empirical density of states and joint density of states analysis of hydrogenated amorphous silicon: a review. *Journal of Materials Science: Materials in Electronics*, 15(7):401–410, 2004.

[38] Stephen Karrer O'Leary. An analytical density of states and joint density of states analysis of amorphous semiconductors. *Journal of applied physics*, 96(7):3680–3686, 2004.

[39] C. C. Paige. *The Computation of Eigenvalues and Eigenvectors of Very Large Sparse Matrices*. PhD thesis, University of London, 1971.

[40] C. C. Paige. Computational variants of the Lanczos method for the eigenproblem. *J. Inst. Math. Appl.*, 10:373–381, 1972.

[41] C. C. Paige. Error analysis of the Lanczos algorithm for tridiagonalizing a symmetric matrix. *J. Inst. Math. Appl.*, 18(3):341–349, 1976.

[42] B. N. Parlett. A new look at the Lanczos algorithm for solving symmetric systems of linear equations. *Linear Algebra Appl.*, 29:323–346, 1980.

[43] B. N. Parlett and D. S. Scott. The Lanczos algorithm with selective orthogonalization. *Math. Comp.*, 33(145):217–238, 1979.

[44] Beresford N. Parlett. *The Symmetric Eigenvalue Problem*, volume 20 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1998. Corrected reprint of the 1980 original.

[45] G. Pólya. Über die Konvergenz von Quadraturverfahren. *Math. Z.*, 37(1):264–286, 1933.

[46] J.W. Rohlf. *Modern Physics from $\alpha$ to $Z_0$*. Wiley, 1994.

[47] Walter Rudin. *Fourier Analysis on Groups*. Interscience Tracts in Pure and Applied Mathematics, No. 12. Interscience Publishers (a division of John Wiley and Sons), New York-London, 1962.

[48] Yousef Saad. *Iterative Methods for Sparse Linear Systems*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, second edition, 2003.

[49] Yousef Saad. *Numerical Methods for Large Eigenvalue Problems*, volume 66 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2011. Revised edition of the 1992 original [ 1177405].

[50] D. S. Scott. *Analysis of the Symmetric Lanczos Algorithm.* PhD thesis, University of California, Berkeley, 1978.

[51] RN Silver and H Röder. Densities of states of mega-dimensional hamiltonian matrices. *International Journal of Modern Physics C*, 5(04):735–753, 1994.

[52] RN Silver and H Röder. Calculation of densities of states and spectral functions by chebyshev recursion and maximum entropy. *Physical Review E*, 56(4):4822, 1997.

[53] RN Silver, H Roeder, AF Voter, and JD Kress. Kernel polynomial approximations for densities of states and spectral functions. *Journal of Computational Physics*, 124(1):115–130, 1996.

[54] H. D. Simon. *The Lanczos Algorithm for Solving Symmetric Linear Systems.* PhD thesis, University of California, Berkeley, 1982.

[55] Horst D. Simon. Analysis of the symmetric Lanczos algorithm with reorthogonalization methods. *Linear Algebra Appl.*, 61:101–131, 1984.

[56] Horst D. Simon. The Lanczos algorithm with partial reorthogonalization. *Math. Comp.*, 42(165):115–142, 1984.

[57] J. Singh. *Electronic and Optoelectronic Properties of Semiconductor Structures.* Cambridge University Press, 2007.

[58] Lloyd N. Trefethen. *Approximation Theory and Approximation Practice.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2013.

[59] Shashanka Ubaru, Jie Chen, and Yousef Saad. Fast estimation of $\mathtt{tr}(f(A))$ via stochastic Lanczos quadrature. *SIAM J. Matrix Anal. Appl.*, 38(4):1075–1099, 2017.

[60] W Walukiewicz, SX Li, J Wu, KM Yu, JW Ager III, EE Haller, Hai Lu, and William J Schaff. Optical properties and electronic structure of inn and in-rich group iii-nitride alloys. *Journal of Crystal Growth*, 269(1):119–127, 2004.

[61] Lin-Wang Wang. Calculating the density of states and optical-absorption spectra of large quantum systems by the plane-wave moments method. *Physical Review B*, 49(15):10154, 1994.

[62] A. J. Wathen. Realistic eigenvalue bounds for the Galerkin mass matrix. *IMA J. Numer. Anal.*, 7(4):449–457, 1987.

[63] Richard L. Wheeden and Antoni Zygmund. *Measure and Integral*. Pure and Applied Mathematics (Boca Raton). CRC Press, Boca Raton, FL, second edition, 2015. An introduction to real analysis.

[64] J. H. Wilkinson. *The Algebraic Eigenvalue Problem*. Monographs on Numerical Analysis. The Clarendon Press, Oxford University Press, New York, 1988. Oxford Science Publications.

[65] Kesheng Wu and Horst Simon. Thick-restart Lanczos method for large symmetric eigenvalue problems. *SIAM J. Matrix Anal. Appl.*, 22(2):602–616, 2000.

[66] Yuh-Renn Wu, Ravi Shivaraman, Kuang-Chung Wang, and James S Speck. Analyzing the physical properties of ingan multiple quantum well light emitting diodes from nano scale structure. *Applied Physics Letters*, 101(8):083505, 2012.

[67] Yuanzhe Xi, Ruipeng Li, and Yousef Saad. Fast computation of spectral densities for generalized eigenvalue problems. *SIAM J. Sci. Comput.*, 40(4):A2749–A2773, 2018.

[68] Chenggang Zhou, Thomas C Schulthess, Stefan Torbrügge, and DP Landau. Wang-landau algorithm for continuous models and joint density of states. *Physical review letters*, 96(12):120201, 2006.