SIAM Review
Vol. 58, No. 1, pp. 34–65

Approximating Spectral Densities of Large Matrices*

Lin Lin†
Yousef Saad‡
Chao Yang§

Abstract. In physics, it is sometimes desirable to compute the so-called *density of states* (DOS), also known as the *spectral density*, of a real symmetric matrix A. The spectral density can be viewed as a probability density distribution that measures the likelihood of finding eigenvalues near some point on the real line. The most straightforward way to obtain this density is to compute all eigenvalues of A, but this approach is generally costly and wasteful, especially for matrices of large dimension. There exist alternative methods that allow us to estimate the spectral density function at much lower cost. The major computational cost of these methods is in multiplying A with a number of vectors, which makes them appealing for large-scale problems where products of the matrix A with arbitrary vectors are relatively inexpensive. This article defines the problem of estimating the spectral density carefully and discusses how to measure the accuracy of an approximate spectral density. It then surveys a few known methods for estimating the spectral density and considers variations of existing methods. All methods are discussed from a numerical linear algebra point of view.

Key words. spectral density, density of states, large scale sparse matrix, approximation of distribution, quantum mechanics

AMS subject classifications. 15A18, 65F15

DOI. 10.1137/130934283

1. Introduction.

Given an $n \times n$ real symmetric and sparse matrix A, scientists in various disciplines often want to compute its *density of states* (DOS), or *spectral density*. Formally, the DOS is defined as

$$\phi(t) = \frac{1}{n} \sum_{j=1}^{n} \delta(t - \lambda_j), \tag{1.1}$$

*Received by the editors August 26, 2013; accepted for publication (in revised form) July 6, 2015; published electronically February 4, 2016. The work of the first and third authors was partially supported by the Laboratory Directed Research and Development Program of Lawrence Berkeley National Laboratory under U.S. Department of Energy contract DE-AC02-05CH11231. The work of the second and third authors was partially supported by the Scientific Discovery through the Advanced Computing (SciDAC) program funded by the U.S. Department of Energy, Office of Science, Advanced Scientific Computing Research and Basic Energy Sciences through grant DE-SC0008877.
http://www.siam.org/journals/sirev/58-1/93428.html

†Department of Mathematics, University of California, Berkeley, and Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720 (linlin@math.berkeley.edu).

‡Department of Computer Science and Engineering, University of Minnesota, Twin Cities, Minneapolis, MN 55455 (saad@cs.umn.edu).

§Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720 (cyang@lbl.gov).

where δ is the Dirac distribution commonly referred to as the Dirac δ-"function" [34, 4, 32] and the λ_j's are the eigenvalues of A, assumed here to be labeled nondecreasingly. Using the DOS, the number of eigenvalues in an interval $[a, b]$ can be formally expressed as

$$\nu_{[a,b]} = \int_a^b \sum_j \delta(t - \lambda_j)\ dt \equiv \int_a^b n\phi(t)dt. \tag{1.2}$$

Therefore, one can view $\phi(t)$ as a probability distribution "function," which gives the probability of finding eigenvalues of A in a given infinitesimal interval near t. If one had access to all the eigenvalues of A, the task of computing the DOS would become trivial. However, in many applications, the dimension of A is large and the computation of its entire spectrum is prohibitively expensive, which leads to the need to develop efficient alternative methods to estimate $\phi(t)$ without computing eigenvalues of A. Since $\phi(t)$ is not a proper function, we need to clarify what we mean by "estimating" $\phi(t)$, and this will be addressed in detail shortly. For now we can use our intuition to argue that $\phi(t)$ can be approximated by dividing the interval containing the spectrum of A into many subintervals and use a tool like Sylvester's law of inertia to count the number of eigenvalues within each of these subintervals. This approach yields a histogram of the eigenvalues. Expression (1.2) will then provide us with an average "value" of $\phi(t)$ in each small subinterval $[a, b]$. As the size of each subinterval decreases, the histogram approaches the spectral density of A. However, this is not a practical approach since performing such an inertia count requires us to compute the LDL^T factorization [15] of $A - t_i I$, where the t_i's are the end points of the subintervals. In general, this approach is prohibitively expensive because of the large number of intervals needed and the sheer cost of each factorization. Therefore, a procedure that relies entirely on multiplications of A with vectors is the only viable approach.

Because calculating the spectral density is such an important problem in quantum mechanics, there is an abundant literature devoted to this problem and research in this area was extremely active in the 1970s and 1980s, leading to clever and powerful methods developed by physicists and chemists [10, 39, 9, 43] for this purpose.

In this survey paper, we review two classes of methods for approximating the spectral density of a real symmetric matrix from a numerical linear algebra perspective. For simplicity, all methods are presented using real arithmetic operations, i.e., we assume the matrix is real symmetric. The generalization to Hermitian matrices is straightforward. The first class of methods contains the kernel polynomial method (KPM) [36, 41] and its variants. The KPM can be viewed as a formal polynomial expansion of the spectral density. It uses a moment matching method to derive the coefficients for the polynomials. The method, which is widely used in a variety of calculations that require the DOS [42], has continued to receive a tremendous amount of interest over the last few years [42, 6, 22, 35]. We show that a less well known, but rather original, method due to Lanczos and known as the "Lanczos spectroscopic" procedure, which samples the cosine transform of the spectral density, is closely related to the KPM. As another variant of the KPM, we present a spectral density probing method called the Delta-Gauss–Legendre method, which can be viewed as a polynomial expansion of a smoothed spectral density. The second class of methods we consider uses the classical Lanczos procedure to partially tridiagonalize A. The eigenvalues and eigenvectors of a tridiagonal matrix are used to construct approximations to the spectral density.

One of the key ingredients used in most of these methods is a well-established artifice for estimating the trace of a matrix. For example, the expansion coefficients

in the abovementioned KPM method can be obtained from the traces of the matrix polynomials $T_k(A)$, where T_k is the Chebyshev polynomial of degree k. Each of these is in turn estimated as the mean of $v^T T_k(A)v$ over a number of random vectors v. This procedure for estimating the trace has been discovered more or less independently by statisticians [18] and physicists and chemists [36, 41].

A natural question to ask is: among all the methods reviewed here, which is the best one to use? The answer to this question is not simple. Since the methods discussed in this paper are all based on matrix-vector product operations (MATVECs), the criterion for choosing the best method should be based on the quality of the approximation when approximately the same number of MATVECs is used. In order to determine the quality, we must first establish a way to measure the accuracy of the approximation. Because the spectral density is defined in terms of the Dirac δ-"functions", which are not proper functions but distributions [4, 32], the standard error metrics used for approximating smooth functions are not appropriate. Furthermore, the accuracy measure should depend on the desired resolution. In many applications, it is not necessary to obtain a high-resolution spectral density. If fact, such a high-resolution density would be highly discontinuous, considering the previously mentioned intuitive interpretation in terms of a histogram. For these reasons our proposed metric for measuring the accuracy of spectral density approximation, as defined in section 2, allows rigorous quantitative comparisons of the different spectral density approximation methods, instead of relying on a subjective visual measure as is often done in practice.

All the approximation methods we consider are presented in section 3. We give some numerical examples in section 4 to compare different numerical methods for approximating spectral densities. We illustrate the effectiveness of our error metric for evaluating the quality of the approximation, and describe some general observations on the behavior of different methods.

2. Assessing the Quality of Spectral Density Approximation. We will denote the approximate spectral density by $\widetilde{\phi}(t)$, which is a regular function. The types of approximate spectral densities we consider in this paper are all continuous functions. However, since $\phi(t)$ is defined in terms of a number of Dirac δ-functions that are not proper functions but distributions, we cannot use the standard L^p-norm with e.g., $p = 1, 2$, or ∞, to evaluate the approximation error defined in terms of $\phi(t) - \widetilde{\phi}(t)$; note that in this difference, $\widetilde{\phi}$ is interpreted as a distribution.

We discuss two approaches to get around this difficulty. In the first approach, we use the fact that $\delta(t)$ is a distribution, i.e., it is formally defined through applications to a test function g:

$$\langle \delta(\cdot - \lambda), g \rangle \equiv \int_{-\infty}^{\infty} \delta(t - \lambda) g(t) dt \equiv g(\lambda).$$

Here we use $\delta(\cdot - \lambda)$ to denote a Dirac δ centered at λ. The test function $g \in C^\infty(\mathbb{R})$, and for all $p, k \in \mathbb{N}$,

$$\sup_{t \in \mathbb{R}} |t^p g^{(k)}(t)| < \infty.$$

Here $g^{(k)}(t)$ is the kth derivative of $g(t)$. The test function g is chosen to be a member of the Schwartz space (or Schwartz class) [32], denoted by $\mathcal{S}$. In other words, the test function g should be smooth and decay sufficiently fast toward 0 when $|t|$ approaches

infinity. The error is then measured as

$$\epsilon_1 = \sup_{g\in\mathcal{S}} |\langle \phi, g\rangle - \langle \widetilde{\phi}, g\rangle|. \tag{2.1}$$

In practice, we restrict $\mathcal{S}$ to be a subspace of the Schwartz space that allows us to compute (2.1) at a finite resolution. We will elaborate on the choice of g and $\mathcal{S}$ in section 2.1.

In the second approach, we regularize δ-functions and replace them with continuous and smooth functions such as Gaussians with an appropriately chosen standard deviation σ. The resulting regularized spectral density, which we denote by $\phi_\sigma(t)$, is a well-defined function. Hence, it is meaningful to compute the approximation error

$$\epsilon_2 = \|\phi_\sigma(t) - \widetilde{\phi}(t)\|_p \tag{2.2}$$

for $p = 1, 2$, and ∞. There is a close connection between the first and second approaches, on which we will elaborate in the next section.

We should note that the notion of regularization, which is rarely discussed in the existing physics and chemistry literature, is important for assessing the accuracy of spectral density approximation. A fully accurate approximation amounts to computing all eigenvalues of A, but for most applications, one only needs to know the number of eigenvalues within any small subinterval contained in the spectrum of A. The size of the interval represents the "resolution" of the approximation. The accuracy of the approximation is only meaningful up to the desired resolution. When (2.2) is used to assess the quality of the approximation, the resolution is defined in terms of the regularization parameter σ. A smaller σ corresponds to higher resolution.

The notion of resolution can also be built into the error metric (2.1) if the trial function g belongs to a certain class of functions, which we will discuss in the next section.

2.1. Restricting the Test Function Space $\mathcal{S}$. The fact that the spectral density $\phi(t)$ is defined in terms of Dirac δ-functions suggests that no fixed smooth function can approximate the spectral density well in the limit of infinite resolution.

To see this, consider $\nu_{[a,b]}$ defined in (1.2) and the associated approximation obtained from a smooth approximation $\widetilde{\phi}(t)$ as

$$\widetilde{\nu}_{[a,b]} = \int_a^b n\widetilde{\phi}(t)dt.$$

For simplicity, let the spectral density $\phi(t) = \delta(t)$ be a single δ-function and the number of eigenvalues be $n = 1$. Infinite resolution means that $|\nu_{[a,b]} - \widetilde{\nu}_{[a,b]}|$ should be small for any choice of $[a, b]$. Now suppose $a = -\varepsilon, b = \varepsilon$. It is easy to verify that

$$\lim_{\varepsilon\to 0+} \nu_{[-\varepsilon,\varepsilon]} = 1, \quad \lim_{\varepsilon\to 0+} \widetilde{\nu}_{[-\varepsilon,\varepsilon]} = 0.$$

In this sense, *all* smooth approximations of the spectral density result in the same accuracy, i.e., there is no difference between a carefully designed approximation of the spectral density and a constant approximation. Hence, the distribution $\phi(t)$ behaves very much like a highly discontinuous function and cannot be approximated by smooth functions with infinite resolution.

In practice, physical quantities and observables can often be deduced from spectral density at finite resolution, i.e., the eigenvalue count only needs to be approximately

correct for an interval of a given finite size. For instance, in condensed matter physics, such information is sufficient to provide material properties such as the band gap or the Van Hove singularity [1] within a given target accuracy. The reduced resolution requirement suggests that we might not need to take the test space S in (2.1) to be the whole Schwartz space. Instead, we can choose functions that have "limited resolution" as test functions. For example, we may consider using Gaussian functions of the form

$$g_\sigma(t) = \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{t^2}{2\sigma^2}}, \tag{2.3}$$

and restrict $\mathcal{S}$ to the subspace

$$\mathcal{S}(\sigma; [\lambda_{lb}, \lambda_{ub}]) = \left\{ g \middle| g(t) \equiv g_\sigma(t-\lambda), \quad \lambda \in [\lambda_{lb}, \lambda_{ub}] \right\},$$

where λ_{lb} and λ_{ub} are lower and upper bounds of the eigenvalues of A, respectively, and the parameter σ defines the *target resolution* up to which we intend to measure. The use of Gaussian functions in the space $\mathcal{S}(\sigma; [\lambda_{lb}, \lambda_{ub}])$ can be understood as a smooth way of counting the number of eigenvalues in an interval whose size is proportional to σ.

Using this choice of the test space, we can measure the quality of any approximation by the metric

$$E[\widetilde{\phi}; \mathcal{S}(\sigma; [\lambda_{lb}, \lambda_{ub}])] = \sup_{g \in \mathcal{S}(\sigma; [\lambda_{lb}, \lambda_{ub}])} |\langle \phi, g \rangle - \langle \widetilde{\phi}, g \rangle|. \tag{2.4}$$

We remark that the use of Gaussians is not the only way to restrict the test space. In some applications, the DOS is often used as a measure for integrating certain physical quantities of interest. If the quantity of interest can be expressed as

$$\langle \phi, g \rangle \equiv \int g(\lambda)\phi(\lambda) d\lambda \equiv \frac{1}{n} \sum_{j=1}^{n} g(\lambda_j)$$

for some smooth function g, then $\mathcal{S}$ can be chosen to contain only one function g, and the approximation error is naturally defined as

$$E[\widetilde{\phi}; g] = |\langle \phi(t), g \rangle - \langle \widetilde{\phi}(t), g \rangle| \tag{2.5}$$

for that particular function g. We give an example of this measure in section 4.3.

2.2. Regularizing the Spectral Density. The error metric in (2.4) can also be understood in the following sense. Let

$$\phi_\sigma(t) = \langle \phi(\cdot), g_\sigma(\cdot - t) \rangle = \sum_{j=1}^{n} g_\sigma(t - \lambda_j). \tag{2.6}$$

Then $\phi_\sigma(t)$ is nothing but a blurred or regularized spectral density, and the blurring is given by a Gaussian function with width σ. Similarly, $\langle \widetilde{\phi}(\cdot), g_\sigma(\cdot - t) \rangle$ can be understood as a blurred version of an approximate spectral density. Therefore, the error metric in (2.4) is equivalent to the L^∞ error between two well-defined functions $\langle \widetilde{\phi}(\cdot), g_\sigma(\cdot - t) \rangle$ and $\phi_\sigma(t)$.

This point of view leads to another way to construct and measure the approximation to the spectral density function. Instead of trying to approximate $\phi(t)$ directly, which might be difficult due to the presence of δ-functions in $\phi(t)$, we first construct a smooth representation of the δ-function. The representation we choose should be commensurate with the desired resolution of the spectral density. This regularization process allows us to expand smooth functions in terms of other smooth functions such as orthogonal polynomials, and the approximation error associated with such expansions can be evaluated directly and without introducing an additional regularization procedure.

The $\phi_\sigma(t)$ function defined in (2.6) is one way to construct a regularized spectral density. Again, the parameter σ controls the resolution. Larger values of σ will lead to smooth curves at the expense of accuracy. Smaller values of σ will lead to rough curves that have peaks at the eigenvalues and zeros elsewhere. This is illustrated in Figure 2.1, where σ takes four different values. We can see that as σ increases, ϕ_σ becomes smoother. When $\sigma = 0.96$, which corresponds to a very smooth spectral density, we can still see the global profile of the eigenvalue distribution, although local variation of the spectral density is mostly averaged out.

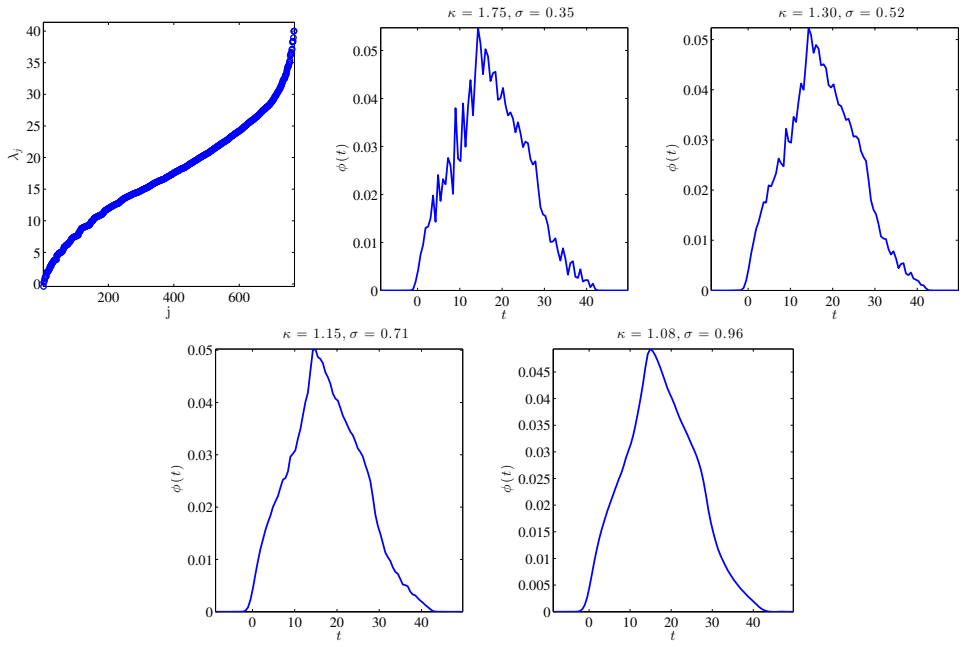

Fig. 2.1 *The eigenvalues as well as various regularized DOS ϕ_σ obtained by blurring the exact DOS (sum of δ-functions positioned at eigenvalues) of a matrix with Gaussians of the form* (2.3).

We remark that the optimal choice of σ, and therefore the smoothness of the approximate DOS, is application dependent. On the one hand, σ should be chosen to be as large as possible so that the regularized DOS ϕ_σ is easy to approximate numerically. On the other hand, increasing σ could cause an undesirable loss of detail and yield an erroneous result. It is up to the user to select a value of σ that balances accuracy and efficiency: σ should be chosen to be small enough to reach the target

accuracy, but not too small, which would require a large number of MATVECs to approximate the DOS.

We should also note that (2.6) is not the only way to regularize the DOS. Another choice is the Lorentzian function defined as

$$\frac{\eta}{(t-\lambda)^2+\eta^2} = -\mathrm{Im}\left(\frac{1}{t-\lambda+i\eta}\right), \tag{2.7}$$

where $\mathrm{Im}(z)$ denotes the imaginary part of a complex number z, and η is a small regularization constant that controls the width of the peak centered at λ. As η approaches zero, (2.7) approaches a Dirac δ-function centered at λ. This approach is used in Haydock's method, to be discussed in section 3.2.2. We also examine the differences between the regularization procedures in section 4 through numerical experiments.

2.3. Nonnegativity Condition. Since the DOS can be viewed as a probability distribution function, it satisfies the nonnegativity condition in the sense that

$$\langle \phi, g \rangle \geq 0 \tag{2.8}$$

for all nonnegative functions $g(t) \geq 0$ in the Schwartz space. Not all numerical methods described in what follows satisfy the nonnegativity condition by construction. We will see in section 4 that failure to preserve the nonnegativity condition can possibly lead to large numerical errors.

3. Numerical Methods for Estimating Spectral Density. In this section, we review two classes of methods for approximating the DOS of A. We begin with the KPM, which can be viewed as a polynomial approximation to the DOS. We show that two other approaches that are derived from different viewpoints are equivalent to KPM. We then describe a second class of methods based on the use of the familiar Lanczos partial tridiagonalization procedure [24]. These methods use blurring (or regularization) techniques to construct an approximate DOS from Ritz values. They differ in the types of blurring they utilize. One of them, which we will simply call the Lanczos method, uses Gaussian blurring, whereas the other method, which we call Haydock's method [17, 3, 28, 5], uses a Lorentzian blurring.

A common characteristic of these methods is that they all use a stochastic sampling and averaging technique to obtain an approximate DOS. The stochastic sampling and averaging technique is based on the following key result [18, 36, 2].

THEOREM 3.1. *Let A be a real symmetric matrix of dimension $n \times n$ with eigendecomposition $A = \sum_{j=1}^{n} \lambda_j u_j u_j^T$ and $u_i^T u_j = \delta_{ij}$, $i, j = 1, \ldots, n$. Here, δ_{ij} is the Kronecker δ symbol. Let v be a vector of dimension n, and suppose v can be represented as the following linear combination of $\{u_i\}_{i=1}^n$:*

$$v = \sum_{j=1}^{n} \beta_j u_j. \tag{3.1}$$

If each component of v is obtained independently from a normal distribution with zero mean and unit standard deviation, i.e.,

$$\mathbb{E}[v] = 0, \quad \mathbb{E}[vv^T] = I, \tag{3.2}$$

then

$$\mathbb{E}[\beta_i \beta_j] = \delta_{ij}, \quad i, j = 1, \ldots, n. \tag{3.3}$$

The proof of Theorem 3.1 is straightforward. The theorem suggests that the trace of a matrix function $f(A)$, which we need to compute in the KPM, for example, can be obtained by simply averaging $v^T f(A)v$ for a number of randomly generated vectors v that satisfy the conditions given in (3.2), because

$$\mathbb{E}[v^T f(A)v] = \mathbb{E}\left[\sum_{j=1}^{n} \beta_j^2 f(\lambda_j)\right] = \sum_{j=1}^{n} f(\lambda_j) = \text{Trace}[f(A)]. \tag{3.4}$$

3.1. The Kernel Polynomial Method. The KPM was proposed by Silver and Röder [36] and Wang [41] in the mid-1990s to calculate the DOS. See also [37, 38, 9, 30], among others, where similar approaches were also used.

3.1.1. Derivation of the original KPM. The KPM constructs an approximation to the exact DOS of a matrix A by formally expanding Dirac δ-functions in terms of Chebyshev polynomials $T_k(t) = \cos(k \arccos(t))$. For simplicity, we assume that the eigenvalues are in the interval $[-1, 1]$. As is the case for all methods which rely on Chebyshev expansions, a change of variables must first be performed to map an interval that contains $[\lambda_{\min}, \lambda_{\max}]$ to $[-1, 1]$ if this assumption does not hold. Following the Silver–Röder paper [36], we include, for convenience, the inverse of the weight function into the spectral density function,

$$\hat{\phi}(t) = \sqrt{1-t^2}\phi(t) = \sqrt{1-t^2} \times \frac{1}{n}\sum_{j=1}^{n} \delta(t - \lambda_j). \tag{3.5}$$

Then we expand the distribution $\hat{\phi}(t)$ as

$$\hat{\phi}(t) = \sum_{k=0}^{\infty} \mu_k T_k(t). \tag{3.6}$$

Equation (3.6) should be understood in the sense of distributions, i.e., for any test function $g \in \mathcal{S}$,

$$\int_{-1}^{1} \hat{\phi}(t)g(t)\ dt = \int_{-1}^{1} \sum_{k=0}^{\infty} \mu_k T_k(t)g(t)\ dt.$$

The same notation applies to the expansion of the DOS using other methods in the following discussion. By means of a formal moment matching procedure, the expansion coefficients μ_k are also defined by

$$\begin{aligned} \mu_k &= \frac{2-\delta_{k0}}{\pi} \int_{-1}^{1} \frac{1}{\sqrt{1-t^2}} T_k(t)\hat{\phi}(t)dt \\ &= \frac{2-\delta_{k0}}{\pi} \int_{-1}^{1} \frac{1}{\sqrt{1-t^2}} T_k(t)\sqrt{1-t^2}\phi(t)dt \\ &= \frac{2-\delta_{k0}}{n\pi} \sum_{j=1}^{n} T_k(\lambda_j). \end{aligned} \tag{3.7}$$

Here δ_{ij} is the Kronecker δ symbol so that $2 - \delta_{k0}$ is equal to 1 when $k = 0$ and to 2 otherwise.

Thus, apart from the scaling factor $(2-\delta_{k0})/(n\pi)$, μ_k is the trace of $T_k(A)$. It follows from Theorem 3.1 that

$$\zeta_k = \frac{1}{n_{\rm vec}} \sum_{l=1}^{n_{\rm vec}} \left(v_0^{(l)}\right)^T T_k(A) v_0^{(l)} \tag{3.8}$$

is a good estimation of the trace of $T_k(A)$, for a set of randomly generated vectors $v_0^{(1)}, v_0^{(2)}, \ldots, v_0^{(n_{\rm vec})}$ that satisfy the conditions given by (3.2). Here the subscript 0 is added to indicate that the vectors have not been multiplied by the matrix A. Then μ_k can be estimated by

$$\mu_k \approx \frac{2-\delta_{k0}}{n\pi}\zeta_k. \tag{3.9}$$

Now we consider the computation of each term $\left(v_0^{(l)}\right)^T T_k(A) v_0^{(l)}$. For simplicity, we drop the superscript l so that $v_0 \equiv v_0^{(l)}$. The three-term recurrence of the Chebyshev polynomial is exploited to compute $T_k(A)v_0$,

$$T_{k+1}(A)v_0 = 2AT_k(A)v_0 - T_{k-1}(A)v_0,$$

so if we let $v_k \equiv T_k(A)v_0$, we have

$$v_{k+1} = 2Av_k - v_{k-1}. \tag{3.10}$$

Once the scalars $\{\mu_k\}$ are determined, we should in theory obtain the expansion for $\phi(t) = \frac{1}{\sqrt{1-t^2}}\hat{\phi}(t)$. In practice, μ_k decays to 0 as $k \to \infty$, and the approximate DOS will be limited to Chebyshev polynomials of degree M. So ϕ is approximated by

$$\widetilde{\phi}_M(t) = \frac{1}{\sqrt{1-t^2}} \sum_{k=0}^{M} \mu_k T_k(t). \tag{3.11}$$

For a general matrix A whose eigenvalues are not necessarily in the interval $[-1,1]$, a linear transformation is first applied to A to bring its eigenvalues into the desired interval. Specifically, we will apply the method to the matrix

$$B = \frac{A-cI}{d},$$

where

$$c = \frac{\lambda_{lb}+\lambda_{ub}}{2}, \quad d = \frac{\lambda_{ub}-\lambda_{lb}}{2}, \tag{3.12}$$

and λ_{lb} and λ_{ub} are lower and upper bounds of the smallest and largest eigenvalues $\lambda_{\min}$ and $\lambda_{\max}$ of A, respectively.

It is important to ensure that the eigenvalues of B are within the interval $[-1,1]$. Otherwise, the magnitude of the Chebyshev polynomial, hence the product of $T_k(B)$ and v_0 computed through a three-term recurrence, will grow exponentially with k.

There are a number of ways [40, 46, 27] to obtain good lower and upper bounds λ_{lb} and λ_{ub} of the spectrum of A. For example, we can set λ_{ub} to $\theta_k + \|(A-\theta_k I)u_k)\|$ and λ_{lb} to $\theta_1 - \|(A-\theta_1 I)u_1)\|$, where θ_1 (resp., θ_k) is the algebraically smallest (resp.,

largest) Ritz value obtained from a k-step Lanczos iteration and u_1 (resp., u_k) is the associated normalized Ritz vector. Note that these residual norms are inexpensive to compute since $\|(A - \theta_j I)u_j\|$ can be easily expressed from the bottom entry of z_j, the unit norm eigenvector of the $k \times k$ tridiagonal matrix obtained from the Lanczos process. For details, see Parlett [31, sec. 13.2]. We should point out that λ_{lb} and λ_{ub} do not have to be very accurate approximations to $\lambda_{\min}$ and $\lambda_{\max}$. It is demonstrated in [27] that tight bounds can be obtained from a 20-step Lanczos iteration for matrices of dimension larger than 100,000.

ALGORITHM 1: The KPM.

Input: Real symmetric matrix A with eigenvalues between $[-1, 1]$. A set of points $\{t_i\}$ at which the DOS is to be evaluated, the degree M of the expansion polynomial.

Output: Approximate DOS $\{\widetilde{\phi}_M(t_i)\}$.

Set $\zeta_k = 0$ for $k = 0, \ldots, M$;
for $l = 1 : n_{\text{vec}}$ **do**
  Select a new random vector $v_0^{(l)}$;
  for $k = 0 : M$ **do**
    Compute $\zeta_k \leftarrow \zeta_k + \left(v_0^{(l)}\right)^T v_k^{(l)}$;
    Compute $v_{k+1}^{(l)}$ via the three-term recurrence $v_{k+1}^{(l)} = 2Av_k^{(l)} - v_{k-1}^{(l)}$ (for $k = 0$, $v_1^{(l)} = Av_0^{(l)}$);
  end for
end for
Set $\zeta_k \leftarrow \zeta_k / n_{\text{vec}}, \mu_k \leftarrow \frac{2-\delta_{k0}}{n\pi}\zeta_k$ for $k = 0, 1, \ldots, M$;
Evaluate $\{\widetilde{\phi}_M(t_i)\}$ using $\{\mu_k\}$ and (3.11);

Because the KPM can be viewed as a way to approximate a series of δ-functions which are highly discontinuous, Gibbs oscillation [21] can be observed near the peaks of the spectral density. Figure 3.1 shows an approximation to $\delta(t)$ by a Chebyshev polynomial expansion of the form (3.11) with $M = 40$. It can be seen that the approximation oscillates around 0 away from $t = 0$. As a result, it does not preserve the nonnegativity of $\delta(t)$. The Gibbs oscillation can be reduced by a technique called Jackson damping, which modifies the coefficients of the Chebyshev expansion. The details of Jackson damping are given in Appendix A, and Figure 3.1 shows that Jackson damping indeed reduces the amount of Gibbs oscillation significantly. However, Jackson damping tends to overregularize the approximate DOS and yields an approximation to $\delta(t)$ that has a wider spread. We will discuss this again with numerical results in section 4.

We should also point out that it is possible to replace Chebyshev polynomials in the KPM by other orthogonal polynomials such as the Legendre polynomials. We will denote by KPML the variant that uses Legendre polynomials to expand the spectral density.

We also note that the cost for constructing the KPM can be reduced by techniques presented in Appendix A. Because the KPM provides a finite polynomial expansion, the approximate DOS $\widetilde{\phi}$ can be evaluated at any arbitrary point t once the $\{\mu_k\}$'s have been determined.

3.1.2. The Spectroscopic View of the KPM. In his 1956 Prentice-Hall book *Applied Analysis* (reprinted by Dover as [25]), Lanczos described a method for computing spectra of real symmetric matrices, which he termed "spectroscopic." This

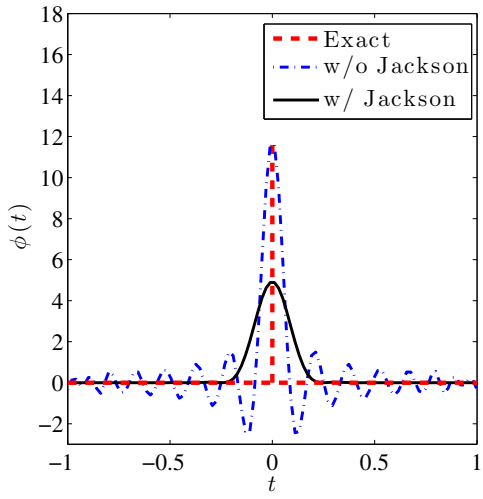

Fig. 3.1 *Chebyshev expansion with and without Jackson damping for Dirac δ-function $\delta(t)$. The Chebyshev polynomial degree is set to* 40.

approach, which also relies heavily on Chebyshev polynomials, is rather unusual in that it associates the spectrum of a matrix to a collection of frequencies, and the goal is to detect these frequencies by Fourier analysis. Because it is not competitive with modern methods for computing eigenvalues, this technique has lost its appeal. However, we will show in what follows that the spectroscopic approach is well suited for computing approximate spectral densities, and it is closely connected to the KPM.

Let us assume that the eigenvalues of the matrix A are in $[-1, 1]$. The spectroscopy approach takes samples of a function of the form

$$f(t) = \sum_{j=1}^{n} \beta_j^2 \cos(\theta_j t), \tag{3.13}$$

where the θ_j's are related to eigenvalues of A by $\theta_j = \cos\lambda_j$ at $t = 0, 1, 2, \ldots, M$. Then one can take the Fourier transform of $f(t)$ to reveal the spectral density of A. If a sufficient number of samples are taken, then the Fourier transform of the sampled function should have peaks near $\cos\lambda_j$, $j = 1, 2, \ldots, n$, and an approximate spectral density can be obtained.

Because λ_j's are not known, (3.13) cannot be evaluated directly. However, $M+1$ uniform samples of $f(t)$, i.e., $f(0)$, $f(1)$, $\ldots$, $f(M)$, can be obtained from the average of

$$v_0^T v_0,\ v_0^T T_1(A) v_0,\ \ldots,\ v_0^T T_M(A) v_0, \tag{3.14}$$

where $T_k(t)$ is the same kth degree Chebyshev polynomial of the first kind used in the previous section and v_0 is a random starting vector.

Taking a discrete cosine transform of (3.14) yields

$$F(p) = \frac{1}{2}\left(f(0) + (-1)^p f(M)\right) + \sum_{k=1}^{M-1} f(k) \cos\frac{kp\pi}{M}, \quad p = 0, \ldots, M. \tag{3.15}$$

Note that, as is customary, the end values are halved to account for the discontinuity

of the data at the interval boundaries. An approximation to the spectral density $\phi(t)$ can be obtained from $F(p)$, $p = 0, 1, \ldots, M$, via an interpolation procedure.

We now show that the spectroscopic approach is closely connected to the KPM. This connection can be seen by noticing that the coefficient ζ_k in (3.8) essentially gives an estimate of the following transform of the spectral density $\phi(t)$,

$$f(s) = \int_{-1}^{1} \cos(s \arccos t)\phi(t)dt, \tag{3.16}$$

evaluated at an integer k.

Since $t \in [-1, 1]$, we can rewrite (3.16) as the continuous cosine transform of a related function by introducing an auxiliary variable $\xi = \arccos t$ and defining

$$\psi(\xi) = \phi(\cos\xi)\sin\xi.$$

It is then easy to verify that (3.16) can be written as

$$f(s) = \int_{0}^{\infty} \cos(s\xi)\psi(\xi)\ d\xi.$$

Here we take $\psi(\xi)$ to be an even function and assume $\psi(\xi) = 0$ for $\xi > \pi$. Thus, the integration range can be taken from 0 to ∞, and $f(s)$ can indeed by obtained by performing a cosine transform of $\psi(\xi)$.

If $f(s)$ is given, we can obtain $\psi(\xi)$ via the inverse cosine transform

$$\psi(\xi) = \frac{2}{\pi}\int_{0}^{\infty} \cos(s\xi)f(s)\ ds. \tag{3.17}$$

Substituting $\xi = \arccos t$ back into (3.17) yields

$$\phi(t) = \frac{2}{\pi\sqrt{1-t^2}}\int_{0}^{\infty} \cos(s \arccos t)f(s)\ ds. \tag{3.18}$$

However, since we can only compute $f(s)$ for $s = k, k \in \mathbb{N}$ by estimating the trace of $T_k(A)$ using a stochastic averaging technique discussed in section 3.1.1, the integration in (3.18) can only be performed numerically using, for example, a composite trapezoidal rule:

$$\phi(t) \approx \frac{2}{\pi\sqrt{1-t^2}}\left[\frac{1}{2}f(0) + \sum_{k=1}^{M-1} f(k)T_k(t) + \frac{1}{2}f(M)T_M(t)\right]. \tag{3.19}$$

Comparing (3.19) with (3.6), we find that (3.19) is exactly the KPM expansion, aside from the fact that the coefficient for $T_M(t)$ is multiplied by a factor $\frac{1}{2}$. Therefore, the spectroscopic method and the KPM are essentially equivalent.

3.1.3. The Delta-Gauss–Legendre Expansion Approach. In some sense, the spectroscopy method discussed in the previous section samples the "reciprocal" space of the spectrum by computing $v_0^T T_j(A)v_0$, where $T_j(t)$ is the jth degree Chebyshev polynomial of the first kind, for a number of different randomly generated vectors v_0 and $j = 0, 1, \ldots, M$. It uses the discrete cosine transform to reveal the spectral density. In this section, we examine another way to directly sample or probe the spectrum of A at an arbitrary point $t_i \in [-1, 1]$ by computing $\{v_0^T p_{M_i}(A)v_0\}$, where

$p_{M_i}(t)$ is an M_ith degree polynomial of the form

$$p_{M_i}(t) \equiv \sum_{k_i=0}^{M_i} \mu_{k_i}(t_i) T_{k_i}(t). \tag{3.20}$$

The expansion coefficient $\mu_{k_i}(t_i)$ in the above expression is chosen, for each t_i, to be

$$\mu_{k_i}(t_i) = \frac{2-\delta_{k0}}{n\pi} \int_{-1}^{1} \frac{1}{\sqrt{1-t^2}} T_{k_i}(t)\delta(t-t_i)dt = \frac{2-\delta_{k0}}{n\pi} \frac{T_{k_i}(t_i)}{\sqrt{1-t_i^2}}. \tag{3.21}$$

The polynomial $p_{M_i}(t)$ defined in (3.20) can be viewed as a polynomial approximation to the δ-function $\delta(t-t_i)$, which can be regarded as a spectral probe placed at t_i.

The reason why $v_0^T p_{M_i}(A)v_0$ can be regarded as a sample of the spectral density at t_i can be explained as follows. The presence of an eigenvalue at t_i can be detected by integrating $\delta(t-t_i)$ over the entire spectrum of A with respect to a spectral point measure defined at eigenvalues only. The integral returns $+\infty$ if t_i is an eigenvalue of A and 0, otherwise. However, in practice, this integration cannot be performed without knowing the eigenvalues of A in advance.

A practical probing scheme can be devised by replacing $\delta(t-t_i)$ with a polynomial approximation such as the one given in (3.20) and integrating, which amounts to evaluating the trace of $p_{M_i}(A)$. This trace evaluation can be done using the same stochastic approach we introduced earlier for the KPM. We call this technique the *Delta-Chebyshev* method.

If v_0 is some random vector normalized to $\|v_0\| = 1$, whose expansion in the eigenbasis $\{u_j\}$ is given by

$$v_0 = \sum_{j=1}^{n} \beta_j u_j, \tag{3.22}$$

then $v_{M_i} \equiv p_{M_i}(A)v_0$ will have the expansion

$$v_{M_i} = \sum_{j=1}^{n} \beta_j p_{M_i}(\lambda_j) u_j. \tag{3.23}$$

Taking the inner product between v_0 and v_{M_i} yields

$$(v_{M_i})^T v_0 = \sum_{j=1}^{n} \beta_j^2 p_{M_i}(\lambda_j). \tag{3.24}$$

Since $\sum_{j=1}^{n} \beta_j^2 = 1$, (3.24) can be viewed as an integral of $p_{M_i}(t)$ associated with a point measure $\{\beta_j^2\}$ defined at eigenvalues of the matrix A. In the case when $\beta_j^2 = 1/n$ for all j, we can then simply rewrite the integral as $\mathrm{Trace}[p_{M_i}(A)]/n$. As we have already shown in previous sections, such a trace can be approximated by choosing multiple random vectors v_0 that satisfy the conditions (3.2) and averaging $v_0^T p_{M_i}(A)v_0$ for all these vectors. The averaged value yields $\widetilde{\phi}(t_i)$, which is the approximation to the spectral density $\phi(t)$ at an arbitrary sample point t_i.

As we indicated in section 2, because $\delta(t-t_i)$ is not a proper function, directly constructing a good polynomial approximation may be difficult. A more plausible approach is to "regularize" $\delta(t-t_i)$ first by replacing it with a smooth function that has a peak at $t=t_i$, and then constructing a polynomial approximation to this smooth function.

We choose the regularized δ-function to be the Gaussian $g_\sigma(t - t_i)$, where g_σ is defined in (2.3) and the standard deviation σ controls the smoothness or the amount of regularization of the function.

It is possible to expand $g_\sigma(t - t_i)$ in terms of Chebyshev polynomials. However, we have found that it is easier to derive an expansion with closed form in terms of Legendre polynomials. It can been shown (see Appendix A) that

$$g_\sigma(t - t_i) = \frac{1}{n(2\pi\sigma^2)^{1/2}} \sum_{k=0}^{\infty} \left(k + \frac{1}{2}\right) \gamma_k(t_i) L_k(t), \tag{3.25}$$

where $L_k(t)$ is the Legendre polynomial of degree k, and the expansion coefficient $\gamma_k(t_i)$ is defined by

$$\gamma_k(t_i) = \int_{-1}^{1} L_k(s) e^{-\frac{1}{2}((s-t_i)/\sigma)^2} ds. \tag{3.26}$$

It can be also shown (see Appendix A) that $\gamma_k(t_i)$ can be determined by a recursive procedure that does not require us to explicitly compute the integral in (3.26).

If we take an approximation to $g_\sigma(t - t_i)$ to be the first $M_i + 1$ terms in the expansion (3.25), i.e.,

$$\widetilde{\phi}_{M_i}(t) = \frac{1}{(2\pi\sigma^2)^{1/2}} \sum_{k=0}^{M_i} \left(k + \frac{1}{2}\right) \gamma_k(t_i) L_k(t), \tag{3.27}$$

then a practical scheme for sampling the spectral density of A can be devised by computing $v_0^T \widetilde{\phi}_{M_i}(A) v_0$ for randomly generated and normalized v_0's and averaging these quantities. Because this scheme is based on regularizing the δ function with a Gaussian and expanding the Gaussian in Legendre polynomials, we call this scheme a Delta-Gauss–Legendre (DGL) method and summarize it in Algorithm 2.

ALGORITHM 2: Multipoint DGL expansion.

Input: Real symmetric matrix A with eigenvalues between $[-1, 1]$. A set of points $\{t_i\}$ at which the DOS is to be evaluated, with $M_{\max}$ the maximum degree employed for all the points.

Output: Approximate DOS $\{\widetilde{\phi}_M(t_i)\}$.

for each t_i **do**
  Compute and store the expansion coefficients $\{\gamma_k(t_i)\}_{k=0}^{M_i}$ using (3.26);
end for
Set $\zeta_k = 0$ for $k = 0, \ldots, M_{\max}$;
for $l = 1 : n_{\text{vec}}$ **do**
  Select a new random vector $v_0^{(l)}$;
  for $k = 0 : M_{\max}$ **do**
    Compute $\zeta_k \leftarrow \zeta_k + \left(v_0^{(l)}\right)^T v_k^{(l)}$;
    Compute $v_{k+1}^{(l)}$ via the three-term recurrence $v_{k+1}^{(l)} = \frac{2k+1}{k+1} A v_k^{(l)} - \frac{k}{k+1} v_{k-1}^{(l)}$ (for $k = 0$, $v_1^{(l)} = A v_0^{(l)}$);
  end for
end for
Set $\zeta_k \leftarrow \zeta_k / n_{\text{vec}}$ for all $k = 0, 1, \ldots, M_{\max}$;
Evaluate $\widetilde{\phi}_{M_i}(t_i)$ using (3.27) with $\{\zeta_k\}$ and the stored $\{\gamma_k(t_i)\}$;

Note that both the Delta-Chebyshev and the DGL methods compute quantities such as $v_0^T p_{M_i} v_0$ at sampled point t_i within the spectrum of A. This would be an unacceptably expensive procedure were it not for the fact that the same vector sequences $\{T_k(A)v_0\}$ and $\{L_k(A)v_0\}$ for $k = 0, 1, \ldots$ can be used for all points t_i at the same time. They only need to be generated once.

Although the Delta-Chebyshev method and the KPM are derived from somewhat different principles, there is a close connection between the two which may not be entirely obvious. The key to recognizing this connection is to notice that the average value of $v_0^T p_{M_i}(A) v_0$ can be viewed as an approximation to $\mathrm{Trace}(p_{M_i}(A))$, which can be written as

$$\begin{aligned}\mathrm{Trace}(p_{M_i}(A)) &= \frac{1}{n}\sum_{k_i=0}^{M_i} \mu_{k_i}(t_i) \sum_{j=1}^{n} T_{k_i}(\lambda_j) \\ &= \frac{1}{n}\sum_{k_i=0}^{M_i} \frac{2-\delta_{k_i,0}}{\pi} \frac{T_{k_i}(t_i)}{\sqrt{1-t_i^2}} \mathrm{Trace}(T_{k_i}(A)) \\ &= \sum_{k_i=0}^{M_i} \left[\frac{2-\delta_{k_i,0}}{n\pi}\mathrm{Trace}(T_{k_i}(A))\right] \frac{T_{k_i}(t_i)}{\sqrt{1-t_i^2}}. \end{aligned} \tag{3.28}$$

Note that the coefficients within the square bracket in (3.28) are exactly the same coefficients as those in the KPM that appear in the expansion (3.6) of the function $\hat{\phi}(t) = \sqrt{1-t^2}\phi(t)$. Therefore, when $M_i = M$ for all i, the Delta-Chebyshev expansion method is identical to the KPM. Hence, the cost of this approach is the same as that of the KPM if polynomials of the same degree and the same number of sampling vectors are used at each t_i.

When M_i is allowed to vary with respect to i, there is a slight advantage to using the Delta-Chebyshev method in terms of flexibility. We can use polynomials of different degrees in different parts of the spectrum to obtain a more accurate approximation. Note that, in this situation, if $M_{\max}$ is the maximum degree employed for all the points, the number of MATVECs employed remains the same and equal to $M_{\max}$, since we will need to compute, for each random vector v_0, the vectors $T_k(A)v_0$ for $k = 0, \ldots, M_{\max}$, which are needed by the points requiring the highest degree. However, some of the other calculations (inner products) required to obtain the spectral density can be avoided, though in most cases applying $T_k(A)$ to v_0 dominates the computational cost in the DOS calculation. The computational cost of DGL is similar to that of the Delta-Chebyshev method. Similarly, one can also show that DGL is closely related to KPML, i.e., it is an expansion of a regularized spectral density in terms of Legendre polynomials. However, we will omit the alternative derivation here.

The close connection between the Delta-Chebyshev method and the KPM also suggests that Gibbs oscillation can be observed in the approximate DOS produced by Delta-Chebyshev and DGL, especially when σ is small. There is no guarantee that the nonnegativity of $\phi(t)$ can be preserved by DGL.

3.2. The Lanczos Algorithm. Because finding a highly accurate DOS essentially amounts to computing all eigenvalues of A, any method that can provide approximations to the spectrum of A can be used to construct an approximate DOS as well. Since the Lanczos algorithm yields good approximations to extreme eigenvalues, it is a good candidate for computing localized spectral densities at least at both ends of the spectrum. In this section, we show that it is also possible to combine the

Lanczos algorithm with multiple randomly generated starting vectors to construct a good approximation to the complete DOS.

It should be noted that the spectral density ϕ as a probability distribution is nonnegative, i.e., $\langle \phi, g \rangle \geq 0$ if $g \geq 0$ everywhere. This is an important property, but the KPM and its variants as introduced in previous sections do not preserve the nonnegativity of the spectral density. In contrast, the methods introduced in this section, including the Lanczos method and the Haydock method, preserve the nonnegativity by construction. This will become a clear advantage for certain spectral densities, as is illustrated in section 4 using numerical experiments.

3.2.1. Constructing Spectral Density Approximations from Ritz Values and Gaussian Blurring. For a given starting vector v_0, an M-step Lanczos procedure for a real symmetric matrix A can be succinctly described by

$$AV_M = V_M T_M + f_M e_M^T, \quad V_M^T V_M = I, \quad V_M^T f_M = 0. \tag{3.29}$$

Here, T_M is an $M \times M$ tridiagonal matrix, V_M is an $n \times M$ matrix, and I_M is an $M \times M$ identity matrix. It is well known [15] that the kth column of V_M can be expressed as

$$V_M e_k = p_{k-1}(A) v_0, \quad k = 1, \ldots, M,$$

where $\{p_k(t)\}$, $k = 0, 1, 2, \ldots, M-1$, is a set of polynomials orthogonal with respect to the weighted spectral distribution $\phi_{v_0}(t)$ that takes the form

$$\phi_{v_0}(t) = \sum_{j=1}^{n} \beta_j^2 \delta(t - \lambda_j), \tag{3.30}$$

where the β_j's are the expansion coefficients obtained from expanding v_0 in the eigenvector basis of A as in (3.22).

It is also well known that these orthogonal polynomials can be generated by a three-term recurrence whose coefficients are defined by the matrix elements of T_M [12]. If (θ_k, y_k), $k = 0, 1, 2, \ldots, M$, are eigenpairs of the tridiagonal matrix T_M, and τ_k is the first entry of y_k, then the distribution function defined by

$$\sum_{k=0}^{M} \tau_k^2 \delta(t - \theta_k) \tag{3.31}$$

serves as an approximation to the weighted spectral density function $\phi_{v_0}(t)$, in the sense that

$$\sum_{j=1}^{n} \beta_j^2 p_q(\lambda_j) = \sum_{k=0}^{M} \tau_k^2 p_q(\theta_k) \tag{3.32}$$

for all polynomials of degree $0 \leq q \leq 2M + 1$. The moment matching property described by (3.32) is well known [14] and is related to the Gaussian quadrature rules [13, 16, 15].

Since, in most cases, we are interested in the standard spectral density defined by (1.1), we would like to choose a starting vector v_0 such that β_j^2 is uniform. However, this is generally not possible without knowing the eigenvectors $\{u_j\}$ of A in advance. To address this issue, we resort to the same stochastic approach we used in previous sections.

We repeat the Lanczos process with multiple randomly generated starting vectors $v_0^{(l)}$, $l = 1, 2, \ldots, n_{\mathrm{vec}}$, that satisfy the conditions given by (3.2). It follows from (3.4) that

$$\frac{1}{n_{\mathrm{vec}} n} \sum_{l=1}^{n_{\mathrm{vec}}} \left(v_0^{(l)}\right)^T \delta(tI - A) v_0^{(l)} = \frac{1}{n} \sum_{j=1}^{n} \left(\frac{1}{n_{\mathrm{vec}}} \sum_{l=1}^{n_{\mathrm{vec}}} \left(\beta_j^{(l)}\right)^2 \right) \delta(t - \lambda_j) \tag{3.33}$$

is a good approximation to the standard spectral density $\phi(t)$ in (1.1). Since each distribution (3.31) generated by the Lanczos procedure is a good approximation to (3.30), the average of (3.31) over l, i.e.,

$$\widetilde{\phi}(t) = \frac{1}{n_{\mathrm{vec}}} \sum_{l=1}^{n_{\mathrm{vec}}} \left(\frac{1}{n} \sum_{k=0}^{M} \left(\tau_k^{(l)}\right)^2 \delta(t - \theta_k^{(l)}) \right), \tag{3.34}$$

should yield a good approximation to the standard spectral density (1.1).

Since (3.34) has far fewer peaks than $\phi(t)$ when M is small, a direct comparison of (3.34) with $\phi(t)$ is not very meaningful. However, when $\phi(t)$ is regularized by replacing $\delta(t - \lambda_i)$ with $g_\sigma(t - \lambda_i)$, we can replace $\delta(t - \theta_k^{(l)})$ in (3.34) with a Gaussian centered at the $\theta_k^{(l)}$ to yield a regularized DOS approximation, i.e., we define the approximate DOS as

$$\widetilde{\phi}_\sigma(t) = \frac{1}{n_{\mathrm{vec}}} \sum_{l=1}^{n_{\mathrm{vec}}} \left(\frac{1}{n} \sum_{k=0}^{M} \left(\tau_k^{(l)}\right)^2 g_\sigma(t - \theta_k^{(l)}) \right). \tag{3.35}$$

This regularization is well justified because in the limit of $M = n$, all Ritz values are the eigenvalues and $\widetilde{\phi}_\sigma(t)$ is exactly the same as the regularized DOS $\phi_\sigma(t)$ for the same σ. We will refer to the method that constructs the DOS approximation from Ritz values obtained from an M-step Lanczos iteration as the Lanczos method in what follows.

Because $g_\sigma(t) \geq 0$, the approximate DOS produced by the Lanczos method is nonnegative. This is a desirable property not shared by the KPM, DGL, or the spectroscopic method.

An alternative way to refine the Lanczos-based DOS approximation from an M-step Lanczos run is to first construct an approximate cumulative spectral density or cumulative density of states (CDOS), which is a monotonically increasing function, and then take the derivative of the CDOS through a finite difference procedure or other means. This technique is discussed in Appendix C.

3.2.2. Haydock's Method. As indicated earlier, the use of Gaussians is not the only way to regularize the spectral density. Another possibility is to replace $\delta(t - \lambda_i)$ in (1.2) with a Lorentzian of the form (2.7) and centered at λ_i. The regularized DOS can be written as

$$\phi_\eta(t) = \frac{1}{n\pi} \sum_{j=1}^{n} \frac{\eta}{(t - \lambda_j)^2 + \eta^2}.$$

Consequently, an alternative approximation to the spectral density can be obtained by simply replacing $\delta(t - \theta_k^{(k)})$ in (3.34) with a Lorentzian centered at $\theta_k^{(k)}$, i.e.,

$$\widetilde{\phi}_\eta(t) = \frac{1}{n_{\mathrm{vec}}} \sum_{l=1}^{n_{\mathrm{vec}}} \left[\frac{1}{n\pi} \sum_{k=0}^{M} \left(\tau_k^{(l)}\right)^2 \frac{\eta}{(t - \theta_k^{(l)})^2 + \eta^2} \right],$$

where $\theta_k^{(l)}$ and $\tau_k^{(l)}$ are the same Ritz values and weighting factors that appear in (3.35) and η is an appropriately chosen constant that corresponds to the resolution of the spectral density to be approximated. This approximation was first suggested by Haydock, Heine, and Kelly [17]. We will refer to this approach as Haydock's method.

Haydock's original method does not require computing Ritz values, even though computing the eigenvalues of a small tridiagonal matrix is by no means costly nowadays. The method makes use of the fact that

$$\phi_\eta(t) = -\frac{1}{n\pi}\mathrm{Im}\ \sum_{j=1}^{n}\frac{1}{t-\lambda_j+i\eta} = -\frac{1}{n\pi}\mathrm{Im}\ \mathrm{Trace}\left[(tI-A+i\eta I)^{-1}\right].$$

Hence, once again, the task of approximating $\phi_\eta(t)$ reduces to that of approximating the trace of $(tI-A+i\eta I)^{-1}$, which can be obtained by

$$\frac{1}{n_{\mathrm{vec}}}\sum_{l=1}^{n_{\mathrm{vec}}}\left(v_0^{(l)}\right)^T(tI-A+i\eta I)^{-1}v_0^{(l)} \tag{3.36}$$

for n_{vec} randomly generated vectors $v_0^{(l)}$ that satisfy the conditions (3.2).

Note that a direct calculation of (3.36) requires solving linear systems of the form $[A-(t+i\eta)I]z = v_0$ repeatedly for any point t at which the spectral density is to be evaluated. This approach can be prohibitively expensive. Haydock's approach approximates $v_0^T(tI-A+i\eta I)^{-1}v_0$ for multiple t's at the cost of performing a single Lanczos factorization and some additional calculations that are much lower in complexity.

If v_0 is used as the starting vector of the Lanczos procedure, then it follows from the shift-invariant property of the Lanczos algorithm that

$$[A-(t+i\eta)I]V_M = V_M[T_M-(t+i\eta)I] + fe_{M+1}^T, \tag{3.37}$$

where V_M and T_M are the same orthonormal and tridiagonal matrices, respectively, that appear in (3.29). After multiplying (3.37) from the left by $[A-(t+i\eta)I]^{-1}$, from the right by $[T_M-(t+i\eta)I]^{-1}$, and rearranging terms, we obtain

$$[A-(t+i\eta)I]^{-1}V_M = V_M[T_M-(t+i\eta)I]^{-1} - [A-(t+i\eta)I]^{-1}fe_{M+1}^T[T_M-(t+i\eta)I]^{-1}.$$

It follows that

$$\begin{aligned} v_0^T\left[A-(t+i\eta)I\right])^{-1}v_0 &= e_1^TV_M^T[A-(t+i\eta)I]^{-1}V_Me_1 \\ &= e_1^T[T_M-(t+i\eta)I]^{-1}e_1 + \xi, \end{aligned}$$

where $\xi = -\left(v_0^T[A-(t+i\eta)I]^{-1}f\right)\left(e_{M+1}^T[T_M-(t+i\eta)I]^{-1}e_1\right)$. If ξ is sufficiently small, computing $v_0^T(tI-A+i\eta I)^{-1}v_0$ reduces to computing the $(1,1)$-th entry of the inverse of $T_M-(t+i\eta)I$. It is not difficult to show that this entry is exactly the same as the expression given in (3.36) up to a constant scaling factor.

Because T_M is tridiagonal with $\alpha_1,\alpha_2,\ldots,\alpha_M$ on the diagonal and $\beta_2,\beta_3,\ldots,\beta_M$ on the subdiagonals and superdiagonals, $e_1^T(zI-T_M)^{-1}e_1$ can be computed in a recursive fashion using the continued fraction formula

$$e_1^T(zI-T_M)^{-1}e_1 = \cfrac{1}{z-\alpha_1+\cfrac{\beta_2^2}{z-\alpha_2+\cdots}}. \tag{3.38}$$

This formula can be verified from the identity following Cramer's rule

$$e_1^T (zI - T_M)^{-1} e_1 \equiv \frac{\det(zI - \hat{T}_M)}{\det(zI - T_M)},$$

where $\hat{T}_M$ is the trailing submatrix starting from the $(2,2)$ entry of T_M (3.38) and assuming the tridiagonal structure of both T_M and $\hat{T}_M$ matrices. It is also related to the generation of Sturm sequences used in bisection methods for computing eigenvalues of tridiagonal matrices [31]. Although this is an elegant way to compute $e_1^T(T_M - zI)^{-1}e_1$, its cost is not much lower than that of solving the linear system $(T_M - zI)w = e_1$ and taking its first entry. For most problems, the cost for this procedure is small compared to that required to perform the Lanczos procedure to obtain T_M.

We should point out that the nonnegativity of $\phi(t)$ is preserved by the Haydock method. However, the Lorentzian function defined by (2.7) decreases to 0 at a much slower rate than a Gaussian as t moves away from λ. Hence, when a high-resolution approximation is required, we may need to choose a very small η in order to produce an accurate approximation.

4. Numerical Results. In this section, we compare all methods discussed above for approximating the spectral density of A through numerical examples. For a given test problem, we first define the target resolution of the DOS to be approximated by setting the parameter σ in either (2.2) or (2.4) or the parameter η in (2.7). We use the metric defined in (2.4) to measure the approximation errors associated with the KPM and its variants, with the exception of DGL. For DGL and the Lanczos and Haydock methods, we simply use the error metric (2.2) with $p = \infty$. Since the spectroscopic method is equivalent to the KPM, we do not show any numerical results for the spectroscopic method.

4.1. Modified Laplacian Matrix. The first example is a modified two-dimensional Laplacian operator with zero Dirichlet boundary condition defined on the domain $[0, 30] \times [0, 30]$. The operator is discretized using a five-point finite difference stencil with $\Delta h = 1$. The modification involves adding a diagonal matrix, which can be regarded as a discretized potential function. The diagonal matrix is generated by adding two Gaussians, one centered at the point $(4, 5)$ of the domain and the other at the point $(25, 15)$. The dimension of the matrix is 750, which is relatively small. We set the parameters σ and η to 0.35. For all calculations shown in this section, we use $n_{\text{vec}} = 100$ random vectors whenever stochastic averaging is needed. Each calculation is repeated 10 times. Each plotted value is the mean value of the computed quantities produced from the 10 runs, with the error bar indicating the standard deviation of the 10 runs.

In Figure 4.1, we compare all methods presented in the previous section. We observe that the Lanczos method seems to outperform all other methods, especially when M is relatively small. The use of Jackson damping in the KPM does not appear to improve the accuracy of the approximation. To some extent, this is not surprising because the true DOS has many sharp peaks (see Figure 4.2), even after it is regularized. Hence, Jackson damping, which tends to overregularize the KPM approximation, may not be able to capture these sharp peaks. The DGL method, the KPM, and KPML behave similarly, as is expected.

In Figure 4.2, we compare $\phi_\sigma(t)$ and $\widetilde{\phi}(t)$ directly for the Lanczos and Haydock methods and the KPM with and without Jackson damping. To see the accuracy of

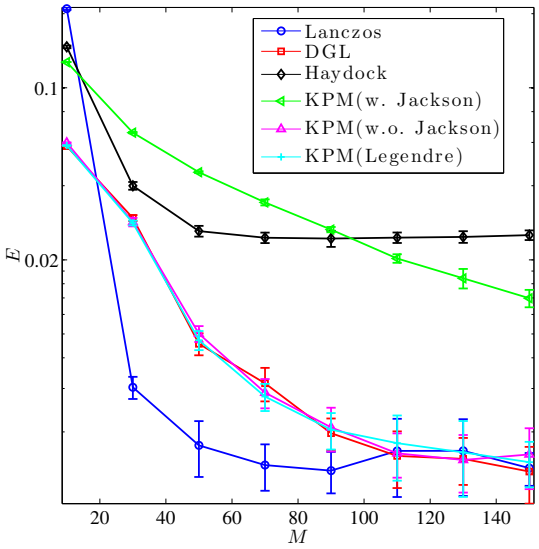

Fig. 4.1 *A comparison of approximation errors of all methods applied to the modified Laplacian matrix for different M values.*

the different methods more clearly, we choose a higher resolution by setting σ and η to 0.05. We note that the meaning of $\widetilde{\phi}(t)$ is different for different methods. For Lanczos, $\widetilde{\phi}(t)$ is the approximate DOS obtained using Gaussian blurring. For Haydock, $\widetilde{\phi}(t)$ is the approximate DOS obtained using Lorentzian Gaussian blurring. For the KPM (with and without Jackson damping), we first evaluate $\widetilde{\phi}(t)$ as in section 3.1 and then plot instead the quantity $\langle \widetilde{\phi}(\cdot), g_\sigma(\cdot - t)\rangle$. In this sense, the exact and approximate DOS are regularized on the same footing. The same procedure is adopted for other numerical examples in this section as well.

We use $M = 100, n_{\text{vec}} = 100$ for all methods. In this case, a visual inspection of the approximate DOS plots in Figure 4.2 yields the same conclusion that we reached earlier based on the measured errors shown in Figure 4.1. Lanczos appears to be the most accurate among all the methods. The DOS curves generated from both the Lanczos and the Haydock methods are above zero. The peaks in the DOS curve produced by the Haydock method are not as sharp as those produced by the Lanczos method, because Haydock uses a Lorentzian to regularize the Dirac δ-function, whereas the Lanczos method uses a Gaussian function to blur the Dirac δ-function centered at Ritz values. The KPM method without Jackson damping does not preserve the nonnegativity of the approximate DOS, and Gibbs oscillation is clearly observed in Figure 4.2 (d). Finally, the KPM with Jackson damping preserves the nonnegativity of the approximate DOS. However, the use of Jackson damping leads to missing several peaks in the DOS, as is illustrated in Figure 4.1. The behaviors of DGL and KPML are similar to that of the KPM without Jackson damping.

4.2. Other Test Matrices. In this section, we compare different DOS approximation methods for two other matrices taken from the University of Florida Sparse Matrix collection [8]. The *pe3k* matrix originates from the vibrational mode calculation of a polyethylene molecule with 3,000 atoms [45]. The *shwater* matrix originates from a computational fluid dynamics simulation. The size of the *pe3k* matrix is 9,000,

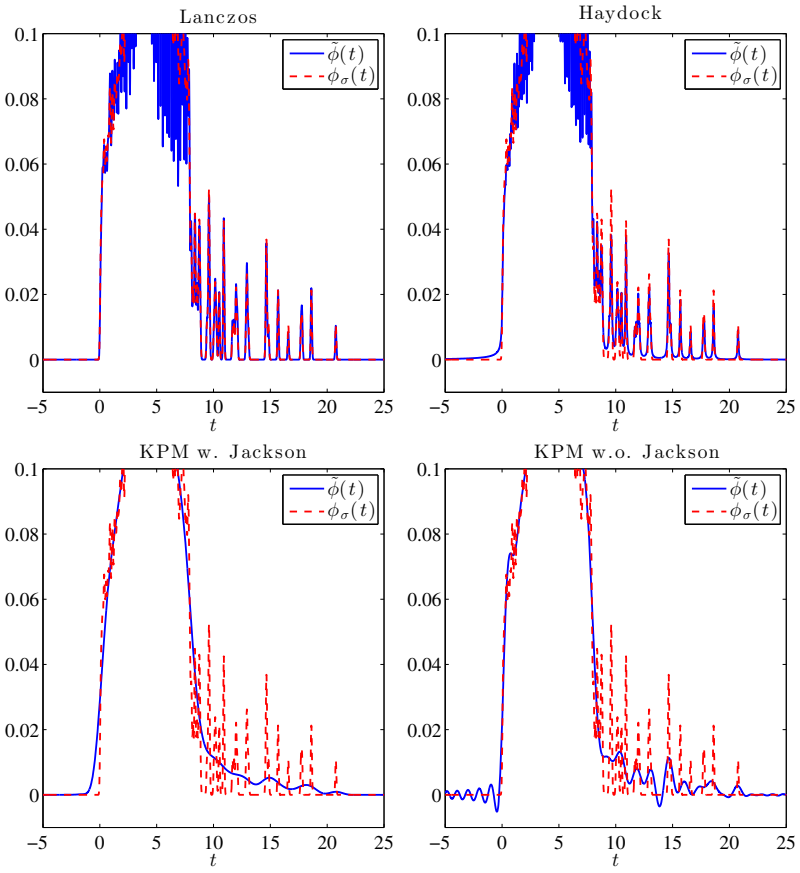

Fig. 4.2 *Comparing the regularized DOS (with $\sigma = 0.05$) with the approximate DOS produced by* (a) *the Lanczos method,* (b) *the Haydock method,* (c) *the KPM with Jackson damping, and* (d) *the KPM without Jackson damping for $M = 100$.*

and the size of the shwater matrix is 81,920. These two test matrices have quite different characteristics in their DOS. The spectrum of the *pe3k* matrix contains a large gap as well as many peaks. The DOS of the *shwater* matrix is relatively smooth, as we will see in what follows.

We set σ to 0.3 in tests presented in this section. We observe that the KPM with Jackson damping only becomes accurate when the degree of the expanding polynomials (M) is high enough, and the convergence with respect to M is rather slow. The DGL method, the KPM without Jackson damping, and KPML behave similarly.

For the *shwater* matrix, which has a relatively smooth spectral density, the Lanczos method is still the most accurate, as we can see from Figure 4.3. It only takes $M = 50$ Lanczos steps to reach 10^{-3} accuracy. The KPM (without Jackson damping) and KPML, as well as the DGL method, all require $M > 110$ terms to reach the same level of accuracy.

Figure 4.4 shows that when we set $M = 100$, $n_{\mathrm{vec}} = 100$, the KPM without Jackson damping yields an accurate approximation to the DOS, whereas Jackson damping introduces slightly larger errors near the locations of the peaks and valleys

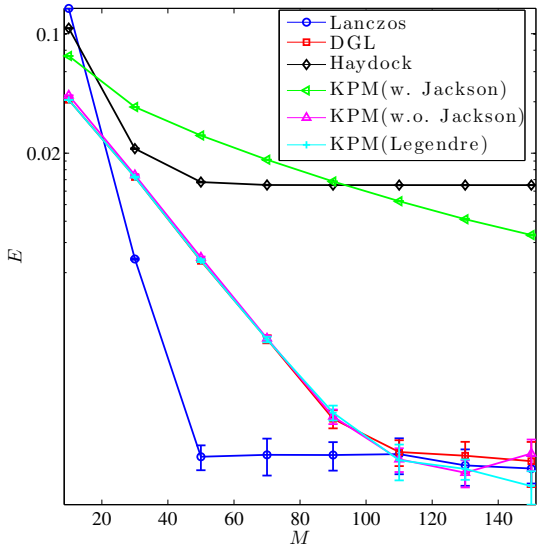

Fig. 4.3 *A comparison of approximation errors of all DOS approximation methods applied to the* shwater *matrix for different* M *values. The regularization parameter* σ *is set to* 0.3.

of the DOS curve. This error is due to the use of extra smoothing. The DOS generated by the Haydock method also has larger errors near the peaks and valleys of the DOS curves, due to the use of Lorentzian regularization.

In Figure 4.5, we zoom into the tail of the DOS curves produced by the Lanczos method and the KPM. It can be seen that Lanczos preserves the nonnegativity of the DOS, whereas the KPM does not. However, since the DOS is smooth, the Gibbs oscillation is very small and can only be seen clearly at the tail of the DOS curve.

For the *pe3k* matrix, Figure 4.6 shows that Lanczos method is significantly more accurate than other methods, followed by the Haydock method. This difference in accuracy can be further observed in Figure 4.7, which compares the regularized DOS with the approximate DOS for the Lanczos method, the Haydock method, and the KPM with and without Jackson damping. We use $M = 100$ and $n_{\text{vec}} = 100$. The *pe3k* matrix has a large gap between the low and high ends of the spectrum. Without the use of Jackson damping, the KPM produces large oscillations over the entire spectrum. We observed similar behavior for DGL and KPML. Adding Jackson damping reduces oscillations in the approximate DOS. However, it leads to an overregularized DOS approximation and is not accurate.

4.3. Application: Heat Capacity Calculation. At the end of section 2.1 we described how there are different ways to regularize the DOS depending on the applications. Here we give an example of a heat capacity calculation for a molecule. The heat capacity is a thermodynamic property and is defined as [29, 44]

$$C_v = \int_0^\infty k_B \frac{(\hbar\omega c/k_B T)^2 e^{-\hbar\omega c/k_B T}}{(1 - e^{-\hbar\omega c/k_B T})^2} \phi(\omega) d\omega, \tag{4.1}$$

where k_B is the Boltzmann constant, c is the speed of light, $\hbar$ is Planck's constant, T is the temperature, and $\omega = \sqrt{\lambda}$ is the vibration frequency.

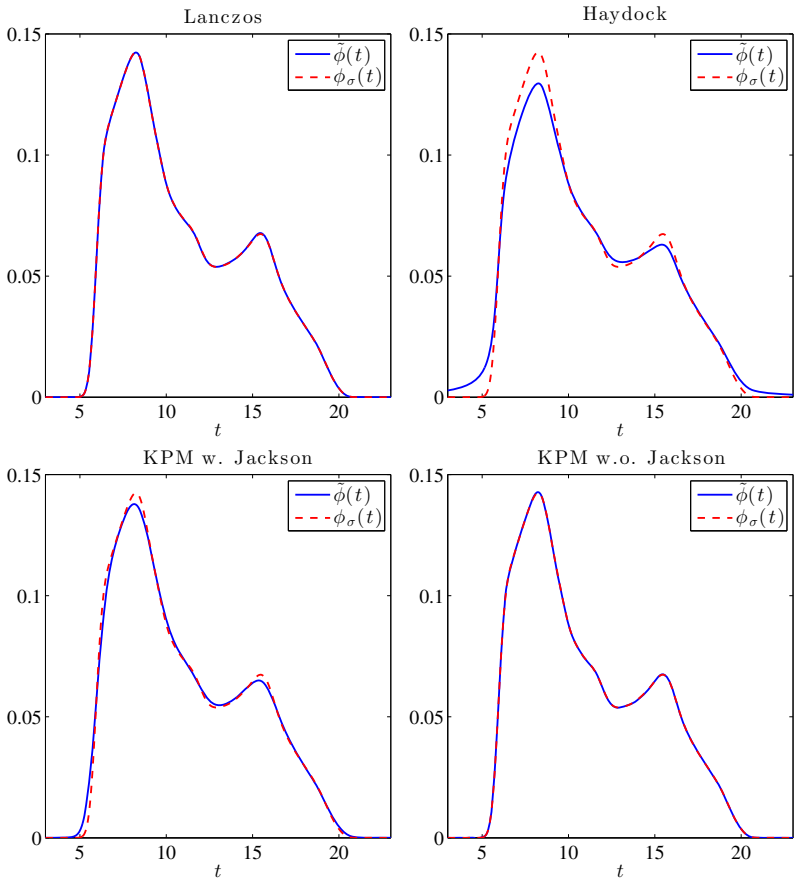

Fig. 4.4 *Comparing the regularized DOS (with $\sigma = 0.3$) with the approximate DOS produced by* (a) *the Lanczos method,* (b) *the Haydock method,* (c) *the KPM with Jackson damping, and* (d) *the KPM without Jackson damping for the* shwater *matrix.*

Here, if we define

$$g(\omega) = k_B \frac{(\hbar\omega c/k_B T)^2 e^{-\hbar\omega c/k_B T}}{(1 - e^{-\hbar\omega c/k_B T})^2} \tag{4.2}$$

and define the DOS $\phi(\omega)$ using the square root of the eigenvalues of the Hessian associated with a molecular potential function with respect to atomic coordinates of the molecule, we have

$$C_v = \langle \phi, g \rangle.$$

Therefore, the error can be measured directly using (2.5).

In what follows, we take the Hessian to be the modified Laplacian matrix and the *pe*3*k* matrix and compute the corresponding heat capacity $C_v(T)$ for different temperature values T. We note that here the computed values of $C_v(T)$ do not carry any physical meaning, but merely serve as a proof of principle for assessing the accuracy of the estimated DOS. We compare the KPM and the Lanczos method. All

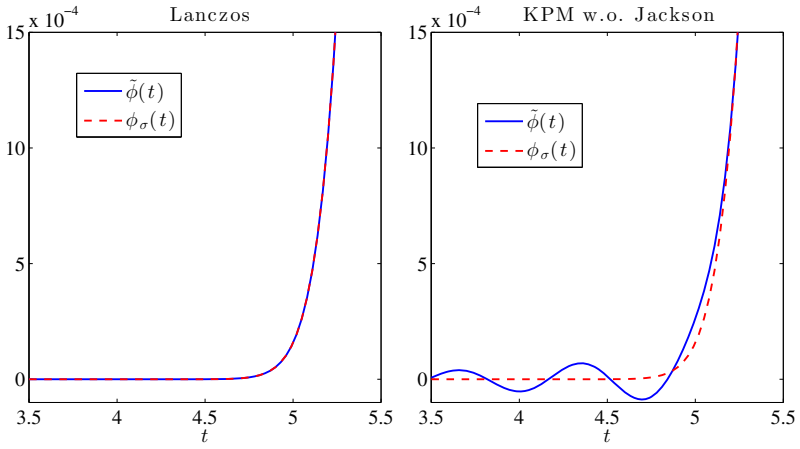

Fig. 4.5 *A comparison of approximation errors of the Lanczos method (left) with that of the KPM without Jackson damping (right) at the higher end of the spectrum of the* shwater *matrix.*

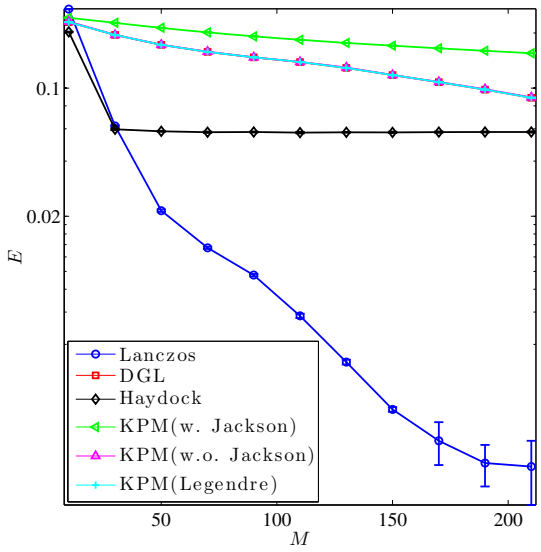

Fig. 4.6 *A comparison of approximation errors of all DOS approximation methods applied to the* pe3k *matrix for different M values. The regularization parameter σ is set to* 0.3.

computations use $M = 40$ MATVECs. Each computed C_v is an averaged value over 100 runs. To facilitate the comparison, we normalize C_v so that its maximum value is 1. The Lanczos method is also fully flexible when the error metric is changed. To this end, we regularize the distribution obtained from Ritz values not by Gaussians, but by the function g in this application. In other words, in (3.35) we replace g_σ by the function g in (4.2).

Figure 4.8 shows that both the KPM and the Lanczos method correctly reproduce the normalized $C_v(T)$ for the modified Laplacian matrix. We also plot the error generated in both the KPM and the Lanczos method. We observe that the error

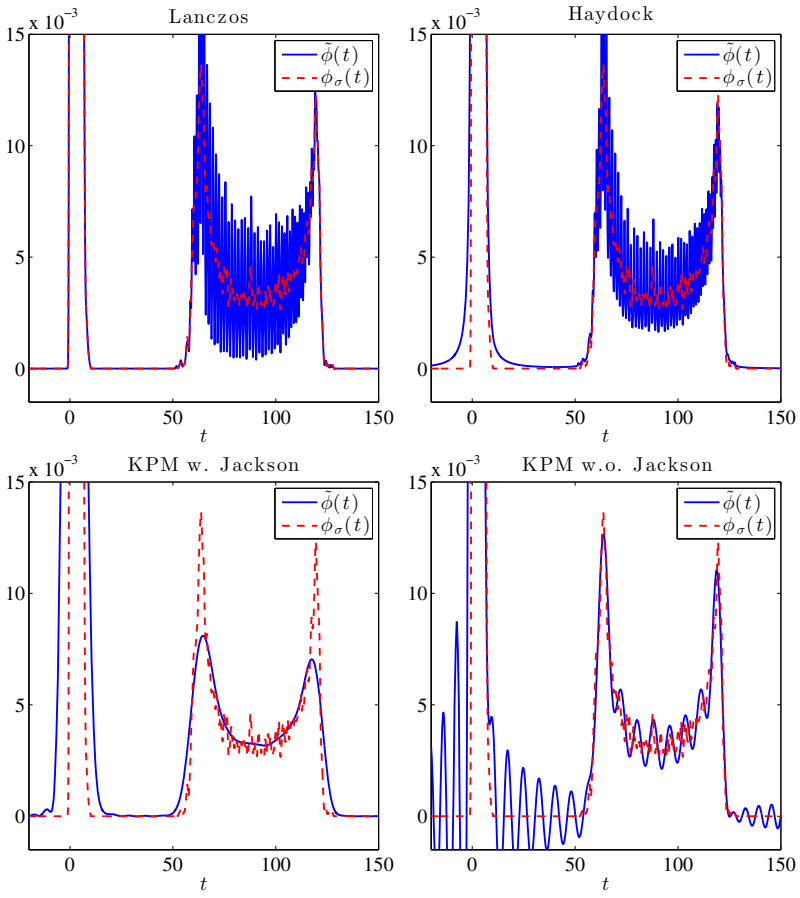

Fig. 4.7 *Comparing the regularized DOS (with $\sigma = 0.3$) with the approximate DOS produced by* (a) *the Lanczos method,* (b) *the Haydock method,* (c) *the KPM with Jackson damping, and* (d) *the KPM without Jackson damping for the* pe3k *matrix.*

associated with the Lanczos method is slightly smaller. This observation agrees with previous results that demonstrate the effectiveness and accuracy of both the KPM and the Lanczos method for computing a relatively smooth DOS.

Figure 4.9 shows that, for the *pe3k* matrix, the KPM approximation of $C_v(T)$ exhibits much larger error than that produced by the Lanczos method. This observation agrees with the results shown in Figure 4.7, which suggests that the Lanczos method yields a much more accurate DOS estimate, especially when M is relatively small.

5. Conclusion. We have surveyed numerical algorithms for estimating the spectral density of a real symmetric matrix A from a numerical linear algebra perspective. The algorithms can be categorized into two classes. The first class contains the KPM method and its variants. The KPM is based on constructing polynomial approximations to Dirac δ-"functions" or regularized δ-"functions." We showed that the Lanczos spectroscopic method is equivalent to the KPM even though it is derived from a different viewpoint. The DGL method is slightly different, but can be viewed as a

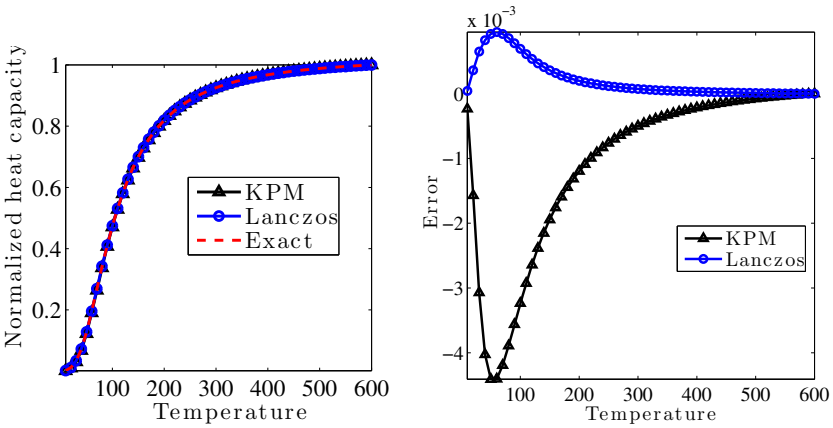

Fig. 4.8 *A comparison of the approximate heat capacity produced by the Lanczos method and the KPM with the "exact" heat capacity at different temperatures (left), and the approximation errors produced by these methods at different temperature values (right), for the modified Laplacian.*

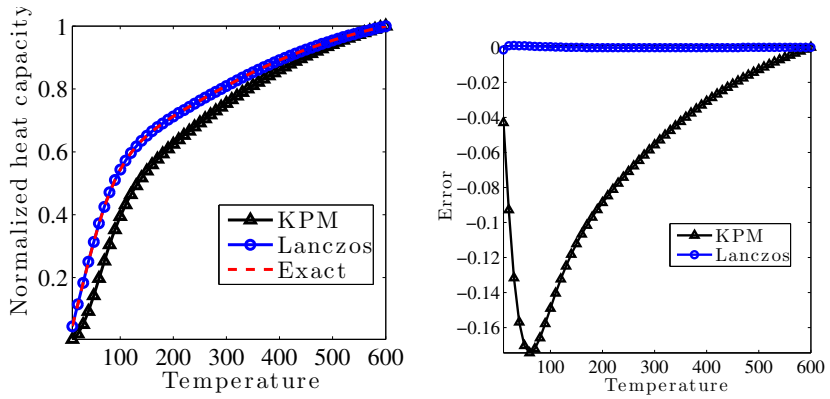

Fig. 4.9 *A comparison of the approximate heat capacity produced by the Lanczos method and the KPM with the "exact" heat capacity at different temperatures (left), and the approximation errors produced by these methods at different temperature values (right), for the* pe3k *matrix.*

polynomial expansion of a regularized spectral density. It is more flexible because it allows polynomials of different degrees to be used at different spectral locations.

The second class of methods is based on the classical Lanczos procedure for partially tridiagonalizing A. Both the Lanczos and the Haydock methods make use of eigenvalues and eigenvectors of the tridiagonal matrix to construct approximations to the DOS. They differ only in the type of regularization they use to interpolate spectral density from Ritz values to other locations in the spectrum. The Lanczos method uses a Gaussian blurring function, whereas the Haydock method uses a Lorentzian. Because a Lorentzian decreases to zero at a much slower rate than a Gaussian away from its peak, it can be less effective when a high-resolution spectral density is needed.

Regularization through the use of Gaussian blurring of δ-"functions" not only allows us to specify the desired resolution of the approximation, but also allows us to

properly define an error metric for measuring the accuracy of the approximation in a rigorous and quantitative manner.

The KPM and its variants require estimating the trace of A or $p(A)$, where $p(t)$ is a polynomial. An important technique for obtaining such an estimate is the stochastic sampling and averaging of the Rayleigh quotient $v_0^T p(A) v_0 / v_0^T v_0$. Averaging the tridiagonal matrices produced by the Lanczos procedure started from randomly generated starting vectors ensures that the approximation contains equal contributions from all spectral components of A. This is an important requirement of the Lanczos and Haydock algorithms.

Our numerical tests show that the Lanczos method consistently outperforms the other methods in terms of the accuracy of the approximation, especially when a few MATVECs are used in the computation. Furthermore, both the Lanczos and Haydock algorithms guarantee that the approximate DOS is nonnegative. This is a desirable feature of any DOS approximation. Another nice property of the Lanczos and Haydock algorithms is that in the limit of $M = n$, they fully recover a regularized DOS.

The KPM and its variants appear to work well when the DOS to be approximated is relatively smooth. They are less effective when the DOS contains many peaks or when the spectrum of A contains large gaps. We found the use of Jackson damping can remove the Gibbs oscillation of the KPM. However, it tends to overregularize the approximate DOS and misses important features (peaks) of the DOS.

Appendix A. Further Discussion on the KPM. For the KPM, a common approach used to damp the Gibbs oscillations is to use the Chebyshev–Jackson approximation [19, 33, 20], which modulates the coefficients μ_k with a damping factor g_k^M defined by

$$g_k^M = \frac{\left(1 - \frac{k}{M+2}\right)\sin(\alpha_M)\cos(k\alpha_M) + \frac{1}{M+2}\cos(\alpha_M)\sin(k\alpha_M)}{\sin(\alpha_M)}, \tag{A.1}$$

where $\alpha_M = \frac{\pi}{M+2}$. Consequently, the damped Chebyshev expansion has the form

$$\widetilde{\phi}_M(t) = \sum_{k=0}^{M} \mu_k g_k^M T_k(t).$$

The approximation of Jackson damping is demonstrated in Figure 3.1.

Another variant can be derived assuming that Chebyshev polynomials are not the only type of orthogonal polynomials that can be used in the expansion. The only practical requirement for an orthogonal polynomial is that we explicitly know its three-term recurrence. For example, we can use the Legendre polynomials $L_k(t)$, which obey the three-term recursion

$$L_0(t) = 1, \quad L_1(t) = t, \quad (k+1)L_{k+1}(t) = (2k+1)tL_k(t) - kL_{k-1}(t).$$

See, for example, [7], for three-term recurrences for a wide class of such polynomials, e.g., all those belonging to the Jacobi class, which include Legendre and Chebyshev polynomials as particular cases.

From a computational point of view, some savings in time can be achieved if we are willing to store more vectors. This is due to the formula

$$T_p(t)T_q(t) = \frac{1}{2}\left[T_{p+q}(t) - T_{|p-q|}(t)\right],$$

from which we obtain

$$T_{p+q}(t) = 2\ T_p(t)T_q(t) + T_{|p-q|}(t).$$

For a given k we can use the above formula with $p = \lceil k/2 \rceil$ and $q = k - p$. This requires that we compute and store $v_r = T_r(A)v_0$ for $r \leq p$. Then, the moments $v_0^T T_r(A)v_0$ for $r \leq p$ can be computed in the usual way, and for $r = p + q > p$ we can use the formula

$$v_0^T T_{p+q}(A)v_0 = 2\ v_p^T v_q + v_0^T v_{|p-q|}.$$

This saves half of the matrix-vector products at the expense of storing all the previous $\{v_r\}$, and therefore it is not practical for high degree polynomials.

Appendix B. Details on the Derivation of the DGL Method. We now calculate the γ_k's starting with γ_0. Since $L_0(\lambda) = 1$, a change of variable $t \leftarrow (s-t)/\sqrt{2\sigma^2}$ yields

(B.1)
$$\gamma_0 = \sigma\sqrt{\frac{\pi}{2}}\left[\operatorname{erf}\left(\frac{1-t}{\sqrt{2}\sigma}\right) - \operatorname{erf}\left(\frac{-1-t}{\sqrt{2}\sigma}\right)\right] = \sigma\sqrt{\frac{\pi}{2}}\left[\operatorname{erf}\left(\frac{1-t}{\sqrt{2}\sigma}\right) + \operatorname{erf}\left(\frac{1+t}{\sqrt{2}\sigma}\right)\right],$$

where we have used the standard error function

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

Now consider a general coefficient γ_{k+1} with $k \geq 0$. There does not seem to exist a closed form formula for γ_k for a general k. However, these coefficients can be obtained by a recurrence relation. To this end, we need to determine concurrently the sequence

$$\psi_k = \int_{-1}^{1} L_k'(s)e^{-\frac{1}{2}((s-t)/\sigma)^2}ds. \tag{B.2}$$

From the three-term recurrence of the Legendre polynomials,

$$(k+1)L_{k+1}(\lambda) = (2k+1)\lambda L_k(\lambda) - kL_{k-1}(\lambda), \tag{B.3}$$

we find by integration that

$$(k+1)\gamma_{k+1} = (2k+1)\int_{-1}^{1} sL_k(s)e^{-\frac{1}{2}((s-t)/\sigma)^2}ds - k\gamma_{k-1}. \tag{B.4}$$

A useful observation is that the above formula is valid for $k = 0$ if we set $\gamma_{-1} \equiv 0$. This comes from (B.3), which is valid for $k = 0$ on setting $L_{-1}(\lambda) \equiv 0$. Next we expand the integral term in the above equality,

$$\int_{-1}^{1} se^{-\frac{1}{2}((s-t)/\sigma)^2}L_k(s)ds = \sigma^2\ \int_{-1}^{1}\frac{s-t}{\sigma^2}e^{-\frac{1}{2}((s-t)/\sigma)^2}L_k(s)ds + t\gamma_k \tag{B.5}$$

$$= \sigma^2\ \int_{-1}^{1}\frac{d}{ds}[-e^{-\frac{1}{2}((s-t)/\sigma)^2}]L_k(s)ds + t\gamma_k. \tag{B.6}$$

The next step is to proceed with integration by parts for the integral in the above expression:

$$\int_{-1}^{1} \frac{d}{ds}[-e^{-\frac{1}{2}((s-t)/\sigma)^2}]L_k(s)ds = -L_k(s)e^{-\frac{1}{2}((s-t)/\sigma)^2}\Big|_{-1}^{1} + \int_{-1}^{1} e^{-\frac{1}{2}((s-t)/\sigma)^2} L_k'(s)ds. \tag{B.7}$$

Noting that $L_k(1) = 1$ and $L_k(-1) = (-1)^k$ for all k, we find

$$\int_{-1}^{1} \frac{d}{ds}[-e^{-\frac{1}{2}((s-t)/\sigma)^2}]L_k(s)ds = -e^{-\frac{1}{2}((1-t)/\sigma)^2} + (-1)^k e^{-\frac{1}{2}((1+t)/\sigma)^2} + \psi_k \tag{B.8}$$

$$= -e^{-\frac{1}{2}(1+t^2)/\sigma^2}\left[e^{t/\sigma^2} - (-1)^k e^{-t/\sigma^2}\right] + \psi_k \tag{B.9}$$

$$\equiv \psi_k - \zeta_k, \tag{B.10}$$

where we have defined

$$\begin{aligned} \psi_k &= \int_{-1}^{1} e^{-\frac{1}{2}((s-t)/\sigma)^2} L_k'(s)ds, \\ \zeta_k &= e^{-\frac{1}{2}((1-t)/\sigma)^2} - (-1)^k e^{-\frac{1}{2}((1+t)/\sigma)^2}. \end{aligned} \tag{B.11}$$

We note in passing that according to (B.9), ζ_k can be written as

$$\zeta_k = \begin{cases} 2e^{-\frac{1}{2}(1+t^2)/\sigma^2}\mathrm{sh}(t/\sigma^2) & \text{for } k \text{ even}, \\ 2e^{-\frac{1}{2}(1+t^2)/\sigma^2}\mathrm{ch}(t/\sigma^2) & \text{for } k \text{ odd}. \end{cases}$$

Substituting (B.8) into (B.6) and the result into (B.4) yields

$$(k+1)\gamma_{k+1} = (2k+1)\left[\sigma^2(\psi_k - \zeta_k) + t\gamma_k\right] - k\gamma_{k-1}. \tag{B.12}$$

The only thing left to do is to find a recurrence for the ψ_k's. Here we use the elegant formula found in, e.g., [26, p. 47],

$$L_{k+1}'(\lambda) = (2k+1)L_k(\lambda) + L_{k-1}'(\lambda). \tag{B.13}$$

Integrating over $[-1, 1]$ yields the relation

$$\psi_{k+1} = (2k+1)\gamma_k + \psi_{k-1}. \tag{B.14}$$

Note that the initial values of ψ_k are $\psi_0 = 0$, $\psi_1 = \gamma_0$. Ultimately, we obtain the following recurrence relations:

$$\begin{cases} \gamma_{k+1} = \frac{2k+1}{k+1}\left[\sigma^2(\psi_k - \zeta_k) + t\gamma_k\right] - \frac{k}{k+1}\gamma_{k-1}, \\ \psi_{k+1} = (2k+1)\gamma_k + \psi_{k-1}. \end{cases} \tag{B.15}$$

It can be noted that the above formulas work for $k = 0$ by setting $\gamma_{-1} = \psi_{-1} = 0$. The recurrence starts with $k = 0$, using the initial values γ_0 given by (B.1), $\psi_1 = \gamma_0$, and $\psi_0 = 0$.

An important remark here is that one has to be careful about the application of the recurrence (B.15). A perceptive reader may notice that such a recurrence runs the risk of being unstable. In fact, we observe the following behavior. For large values

of σ the Gaussian function can be very smooth and as a result a very small degree of polynomials may be needed, i.e., the value of γ_k drops to small values quite rapidly as k increases. If we ask for a high degree polynomial and continue the recurrence (B.15) beyond the point where the expansion has converged (indicated by small values of γ_k) we will essentially iterate with noise. As it turns out, this noise is amplified by the recurrence. This is because the coefficient $\psi_k - \zeta_k$ becomes just noise, which causes the recurrence to diverge. An easy remedy is to just stop iterating (B.15) as soon as two consecutive γ_k's are small. This takes care of two issues at the same time. First, it determines a sort of optimal degree to be used. Second, it avoids the unstable behavior observed by continuing the recurrence. Specifically, a test such as the following is performed:

$$|\gamma_{k-1}| + |\gamma_k| \leq k \cdot \text{tol}, \tag{B.16}$$

where tol is a small tolerance which can be set to 10^{-6}, for example.

With this we can now easily develop the DGL expansion algorithm, in which we will refer to formula (3.24). However, now p_M is the M-degree polynomial

$$p_M(t) = \frac{1}{(2\pi\sigma^2)^{1/2}} \sum_{k=0}^{M} \left(k + \frac{1}{2}\right) \gamma_k L_k(t) \tag{B.17}$$

obtained by truncating the sum (3.25) to $M + 1$ terms.

Appendix C. Cumulative Density of States from the Lanczos Method. An alternative way to refine the Lanczos-based DOS approximation from an M-step Lanczos run is to first construct an approximate cumulative spectral density or cumulative density of states (CDOS), defined as

$$\psi(t) = \int_{-\infty}^{t} \phi(s) ds.$$

Without applying regularization, the approximate CDOS can be computed from the Lanczos procedure as

$$\widetilde{\psi}(t) = \sum_{k=0}^{M} \eta_k^2 \delta(t - \theta_k), \tag{C.1}$$

where $\eta_k^2 = \sum_{i=1}^{k} \tau_i^2$ and θ_k and τ_k are eigenvalues and the first components of the eigenvectors of the tridiagonal matrix T_M defined in (3.34). This approximation is plotted as a staircase function in Figure C.1 for the modified two-dimensional Laplacian. Note that both $\psi(t)$ and $\widetilde{\psi}(t)$ are monotonically nondecreasing functions. Furthermore, it can been shown [23, 11] that $\psi(t) - \widetilde{\psi}(t)$ has precisely $2M - 1$ sign changes within the spectrum of A. A sign change occurs when $\psi(t)$ crosses either a vertical or a horizontal step of $\widetilde{\psi}(t)$. These properties allow us to construct an "interpolated" CDOS that matches $\psi(t)$ and $\widetilde{\psi}(t)$ at the points where $\psi(t)$ crosses $\widetilde{\psi}(t)$.

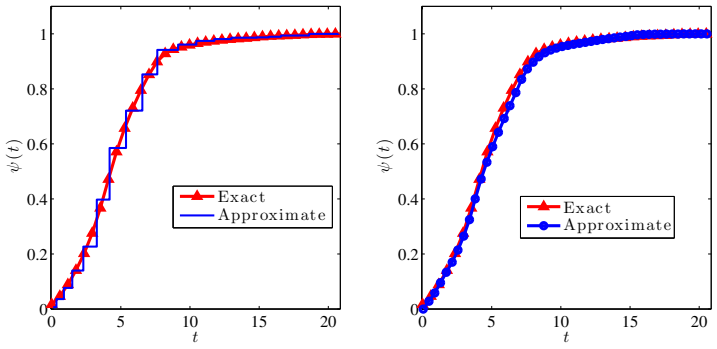

Fig. C.1 *The approximate cumulative spectral density associated with the modified two-dimensional Laplacian constructed directly from a 20-step Lanczos run (left) and its spline-interpolated and smooth version (right).*

REFERENCES

[1] N. Ashcroft and N. Mermin, *Solid State Physics*, Thomson Learning, Toronto, 1976.

[2] H. Avron and S. Toledo, *Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix*, J. ACM, 58 (2011), 8.

[3] S. R. Broderick and K. Rajan, *Eigenvalue decomposition of spectral features in density of states curves*, Europhys. Lett., 95 (2011), 57005.

[4] F. W. Byron and R. W. Fuller, *Mathematics of Classical and Quantum Physics*, Dover, New York, 1992.

[5] R. K. Chouhan, A. Alam, S. Ghosh, and A. Mookerjee, *Ab initio study of phonon spectrum, entropy and lattice heat capacity of disordered Re-W alloys*, J. Phys. Condens. Matter, 24 (2012), 375401.

[6] L. Covaci, F. M. Peeters, and M. Berciu, *Efficient numerical approach to inhomogeneous superconductivity: The Chebyshev-Bogoliubov-de Gennes method*, Phys. Rev. Lett., 105 (2010), 167006.

[7] P. J. Davis, *Interpolation and Approximation*, Blaisdell, Waltham, MA, 1963.

[8] T. A. Davis and Y. Hu, *The University of Florida sparse matrix collection*, ACM Trans. Math. Software, 38 (2011), 1.

[9] D. A. Drabold and O. F. Sankey, *Maximum entropy approach for linear scaling in the electronic structure problem*, Phys. Rev. Lett., 70 (1993), pp. 3631–3634.

[10] F. Ducastelle and F. Cyrot-Lackmann, *Moments developments and their application to the electronic charge distribution of d bands*, J. Phys. Chem. Solids, 31 (1970), pp. 1295–1306.

[11] B. Fischer and R. W. Freund, *An inner product-free conjugate gradient-like algorithm for Hermitian positive definite systems*, in Proceedings of the Cornelius Lanczos 1993 International Centenary Conference, 1994, SIAM, Philadelphia, pp. 288–290.

[12] W. Gautschi, *Computational aspects of three-term recurrence relations*, SIAM Rev., 9 (1967), pp. 24–82.

[13] W. Gautschi, *Construction of Gauss-Christoffel quadrature formulas*, Math. Comp., 22 (1968), pp. 251–270.

[14] W. Gautschi, *A survey of Gauss-Christoffel quadrature formulae*, in E. B. Christoffel: The Influence of His Work in Mathematics and the Physical Sciences, P. Butzer and F. Feher, eds., 1981, Birkhäuser, Basel, pp. 72–147.

[15] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 4th ed., Johns Hopkins University Press, Baltimore, MD, 2013.

[16] G. H. Golub and J. H. Welsch, *Calculation of Gauss quadrature rule*, Math. Comp., 23 (1969), pp. 221–230.

[17] R. Haydock, V. Heine, and M. J. Kelly, *Electronic structure based on the local atomic environment for tight-binding bands*, J. Phys. C Solid State Phys., 5 (1972), pp. 2845–2858.

[18] M. F. Hutchinson, *A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines*, Commun. Stat. Simul. Comput., 18 (1989), pp. 1059–1076.

[19] D. Jackson, *The Theory of Approximation*, AMS, Providence, RI, 1930.
[20] L. O. Jay, H. Kim, Y. Saad, and J. R. Chelikowsky, *Electronic structure calculations using plane wave codes without diagonalization*, Comput. Phys. Commun., 118 (1999), pp. 21–30.
[21] H. Jeffreys and B. Jeffreys, *Methods of Mathematical Physics*, 3rd ed., Cambridge University Press, Cambridge, UK, 1999.
[22] D. Jung, G. Czycholl, and S. Kettemann, *Finite size scaling of the typical density of states of disordered systems within the kernel polynomial method*, Internat. J. Modern Phys. Conf. Ser., 11 (2012), pp. 108–113.
[23] S. Karlin and L. S. Shapley, *Geometry of Moment Spaces*, Mem. Amer. Math. Soc., AMS, Providence, RI, 1953.
[24] C. Lanczos, *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*, J. Res. Nat. Bur. Stand., 45 (1950), pp. 255–282.
[25] C. Lanczos, *Applied Analysis*, Dover, New York, 1988.
[26] N. N. Lebedev, *Special Functions and Their Applications*, Dover, New York, 1972.
[27] R. Li and Y. Zhou, *Bounding the spectrum of large Hermitian matrices*, Linear Algebra Appl., 435 (2011), pp. 480–493.
[28] W. Li, H. Sevincili, S. Roche, and G. Cuniberti, *Efficient linear scaling method for computing the thermal conductivity of disordered materials*, Phys. Rev. B, 83 (2011), 155416.
[29] D. A. McQuarrie, *Statistical Mechanics*, Harper & Row, New York, 1976.
[30] G. A. Parker, W. Zhu, Y. Huang, D. Hoffman, and D. J. Kouri, *Matrix pseudo-spectroscopy: Iterative calculation of matrix eigenvalues and eigenvectors of large matrices using a polynomial expansion of the Dirac delta function*, Comput. Phys. Commun., 96 (1996), pp. 27–35.
[31] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Classics Appl. Math. 20, SIAM, Philadelphia, 1998.
[32] R. D. Richtmyer, *Principles of Advanced Mathematical Physics*, Vol. 1, Springer-Verlag, New York, 1981.
[33] T. J. Rivlin, *An Introduction to the Approximation of Functions*, Dover, New York, 2003.
[34] L. Schwartz, *Mathematics for the Physical Sciences*, Dover, New York, 1966.
[35] B. Seiser, D. G. Pettifor, and R. Drautz, *Analytic bond-order potential expansion of recursion-based methods*, Phys. Rev. B, 87 (2013), 094105.
[36] R. N. Silver and H. Röder, *Densities of states of mega-dimensional Hamiltonian matrices*, Internat. J. Modern Phys. C, 5 (1994), pp. 735–753.
[37] R. N. Silver and H. Röder, *Calculation of densities of states and spectral functions by Chebyshev recursion and maximum entropy*, Phys. Rev. E, 56 (1997), pp. 4822–4829.
[38] R. N. Silver, H. Röder, A. F. Voter, and J. D. Kress, *Kernel polynomial approximations for densities of states and spectral functions*, J. Comput. Phys., 124 (1996), pp. 115–130.
[39] I. Turek, *A maximum-entropy approach to the density of states within the recursion method*, J. Phys. C, 21 (1988), pp. 3251–3260.
[40] J. L. M. Van Dorsselaer, M. E. Hoschstenbach, and H. A. Van der Vorst, *Computing probabilistic bounds for extreme eigenvalues of symmetric matrices with the Lanczos method*, SIAM J. Matrix Anal. Appl., 22 (2000), pp. 837–852.
[41] L.-W. Wang, *Calculating the density of states and optical-absorption spectra of large quantum systems by the plane-wave moments method*, Phys. Rev. B, 49 (1994), ppp. 10154–10158.
[42] A. Weisse, G. Wellein, A. Alvermann, and H. Fehske, *The kernel polynomial method*, Rev. Modern Phys., 78 (2006), pp. 275–306.
[43] J. C. Wheeler and C. Blumstein, *Modified moments for harmonic solids*, Phys. Rev. B, 6 (1972), pp. 4380–4382.
[44] C. Yang, D. W. Noid, B. G. Sumpter, D. C. Sorensen, and R. E. Tuzun, *An efficient algorithm for calculating the heat capacity of a large-scale molecular system*, Macromol. Theory Simul., 10 (2001), pp. 756–761.
[45] C. Yang, B. W. Peyton, D. W. Noid, B. G. Sumpter, and R. E. Tuzun, *Large-scale normal coordinate analysis for molecular structures*, SIAM J. Sci. Comput., 23 (2001), pp. 563–582.
[46] Y. Zhou, Y. Saad, M. L. Tiago, and J. R. Chelikowsky, *Parallel self-consistent-field calculations via Chebyshev-filtered subspace acceleration*, Phy. Rev. E, 74 (2006), 066704.