SIAM J. MATRIX ANAL. APPL.
Vol. 31, No. 5, pp. 2681–2699

SPECTRAL METHODS FOR PARAMETERIZED MATRIX EQUATIONS*

PAUL G. CONSTANTINE†, DAVID F. GLEICH‡, AND GIANLUCA IACCARINO§

Abstract. We apply polynomial approximation methods—known in the numerical PDEs context as *spectral methods*—to approximate the vector-valued function that satisfies a linear system of equations where the matrix and the right-hand side depend on a parameter. We derive both an interpolatory pseudospectral method and a residual-minimizing Galerkin method, and we show how each can be interpreted as solving a truncated infinite system of equations; the difference between the two methods lies in where the truncation occurs. Using classical theory, we derive asymptotic error estimates related to the region of analyticity of the solution, and we present a practical residual error estimate. We verify the results with two numerical examples.

Key words. parameterized systems, spectral methods

AMS subject classifications. 65D30, 65F99

DOI. 10.1137/090755965

1. Introduction.

We consider a system of linear equations where the elements of the matrix of coefficients and the right-hand side depend analytically on a parameter. Such systems often arise as an intermediate step within computational methods for engineering models which depend on one or more parameters. A large class of models employs such parameters to represent uncertainty in the input quantities; examples include PDEs with random inputs [3, 14, 32], image deblurring models [9], and noisy inverse problems [8]. Other examples of parameterized linear systems occur in electronic circuit design [23], applications of PageRank [6, 10], and dynamical systems [11]. Additionally, we note a class of interpolation schemes [30, 25] where each evaluation of the interpolant involves the solution of a linear system of equations that depends on the interpolation point. Parameterized linear operators have been analyzed in their own right in the context of perturbation theory; the standard reference for this work is Kato [22].

In our case, we are interested in approximating the vector-valued function that satisfies the parameterized matrix equation. We will analyze the use of polynomial approximation methods, which have evolved under the heading "spectral methods" in the context of numerical methods for PDEs [4, 7, 21]. In their most basic form, these methods are characterized by a global approximation of the function of interest by a finite series of orthogonal (algebraic or trigonometric) polynomials. For smooth functions, these methods converge geometrically, which is the primary reason

*Received by the editors April 14, 2009; accepted for publication (in revised form) by D. Boley July 23, 2010; published electronically September 23, 2010.
http://www.siam.org/journals/simax/31-5/75596.html

†Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA 94305. Current address: Sandia National Laboratories, Albuquerque, NM 87123 (pconsta@sandia.gov). This author was funded by the Department of Energy Predictive Science Academic Alliance Program.

‡Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA 94305. Current address: Sandia National Laboratories, Livermore, CA 94550 (dfgleic@sandia.gov). This author was supported by a Microsoft Live Labs Fellowship.

§Mechanical Engineering and Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA 94305 (jops@stanford.edu). This author was funded by the Department of Energy Predictive Science Academic Alliance Program.

TABLE 1.1
Notation.

Notation	Meaning
$A(s)$	a square matrix-valued function of a parameter s
$b(s)$	a vector-valued function of the parameter s
$\mathbf{A}$	a constant matrix
$\mathbf{b}$	a constant vector
$\langle\cdot\rangle$	the integral with respect to a given weight function
$\langle\cdot\rangle_n$	the integral $\langle\cdot\rangle$ approximated by an n-point Gaussian quadrature rule
$[\mathbf{M}]_{r\times r}$	the first $r \times r$ principal minor of a matrix $\mathbf{M}$

for their popularity. The use of spectral methods for parameterized equations is not unprecedented. In fact, we were motivated primarily by the so-called polynomial chaos methods [17, 32] and related work [3, 2, 31] in the burgeoning field of uncertainty quantification. There has been some work in the linear algebra community analyzing the fully discrete problems that arise in this context [13, 27, 12], but we know of no existing work addressing the more general problem of parameterized matrix equations.

There is an ongoing debate in spectral methods communities surrounding the relative advantages of Galerkin methods versus pseudospectral methods. In the case of parameterized matrix equations, the interpolatory pseudospectral methods require only the solution of the parameterized model evaluated at a discrete set of points, which makes parallel implementation straightforward. In contrast, the Galerkin method requires the solution of a coupled linear system whose dimension is many times larger than the original parameterized set of equations. We offer insight into this contest by establishing a formalism for rigorous comparison and deriving concrete relationships between the two methods using the Jacobi matrix of recurrence coefficients for the orthogonal polynomial basis. The specific relationships we derive have been established in numerical PDEs, but their consequence is arguably greater for parameterized matrix equations because of the way that the pseudospectral method *decouples* into a set of smaller subproblems while the Galerkin method does not, in general. In short, the primary contribution of this work is a unifying perspective.

In this paper, we will first describe the parameterized matrix equation and characterize its solution in section 2. We then derive a spectral Galerkin method and a pseudospectral method for approximating the solution to the parameterized matrix equation in section 3. In section 4, we analyze the relationship between these methods using the symmetric, tridiagonal Jacobi matrices—techniques which are reminiscent of the analysis of Gaussian quadrature by Golub and Meurant [18] and Gautschi [15]. We derive error estimates for the methods that relate the geometric rate of convergence to the size of the region of analyticity of the solution in section 5, and we conclude with simple numerical examples in section 6. See Table 1.1 for a list of notational conventions, and note that *all index sets begin at 0 to remain consistent with the ordering of a set of polynomials by their largest degree.*

2. Parameterized matrix equations. In this section, we define the specific problem we will study and characterize its solution. We consider problems that depend on a single parameter s that takes values in the finite interval $[-1, 1]$. Assume that the interval $[-1, 1]$ is equipped with a positive scalar weight function $w(s)$ such that all moments exist, i.e.,

$$\langle s^k \rangle \equiv \int_{-1}^{1} s^k w(s)\, ds < \infty, \qquad k = 1, 2, \ldots, \tag{2.1}$$

and the integral of $w(s)$ is equal to 1. We will use the bracket notation to denote an integral against the given weight function. In a stochastic context, one may interpret this as an expectation operator, where $w(s)$ is the density function of the random variable s.

Let the $\mathbb{R}^N$-valued function $x(s)$ satisfy the linear system of equations

$$A(s)x(s) = b(s), \qquad s \in [-1, 1], \tag{2.2}$$

for a given $\mathbb{R}^{N\times N}$-valued function $A(s)$ and $\mathbb{R}^N$-valued function $b(s)$. We assume that the elements of both $A(s)$ and $b(s)$ are analytic in a region containing $[-1, 1]$. Additionally, we assume that $A(s)$ is bounded away from singularity for all $s \in [-1, 1]$. This implies that we can write $x(s) = A^{-1}(s)b(s)$.

The elements of the solution $x(s)$ can also be written using Cramer's rule [24, Chapter 6] as a ratio of determinants,

$$x_i(s) = \frac{\det(A_i(s))}{\det(A(s))}, \qquad i = 0, \ldots, N-1, \tag{2.3}$$

where $A_i(s)$ is the parameterized matrix formed by replacing the ith column of $A(s)$ by $b(s)$. From (2.3) and the invertibility of $A(s)$, we can conclude that $x(s)$ is analytic in a region containing $[-1, 1]$.

(2.3) reveals the underlying structure of the solution as a function of s. If $A(s)$ and $b(s)$ depend polynomially on s, then (2.3) tells us that $x(s)$ contains rational functions. Note also that this structure is independent of the particular weight function $w(s)$.

Parameterized matrix equations with similar constraints appear in a variety of applications. Many have studied elliptic PDEs with random coefficients in the elliptic operator [14, 1]. In this context, one may choose to model the random coefficients with a mean-plus-fluctuation type decomposition where a parameter may control the fluctuation. To ensure that the coefficients remain positive, certain analytic functions, such as the exponential function, may be applied to the parameter. Another interesting example occurs in the PageRank model for ranking nodes in a graph [26]. To compute the ranking vector, one must solve a large matrix equation, where the matrix contains a damping parameter that linearly affects the matrix elements. Finally, we note that, in many electronic circuit design applications, one must solve a matrix equation, that depends on the square of a given parameter which represents frequency [23]. The output of this model is the frequency response function. As an aside, we note that many applications may include matrices and right-hand sides that depend on multiple parameters. The techniques, analyses, and formalism we present extend to the multivariate case using standard tensor product constructions. However, the cost of such an approximation increases, exponentially as the number of parameters increases, rendering the spectral methods infeasible for more than a handful of parameters. Therefore, we focus on the univariate problems.

3. Spectral methods. In this section, we derive the spectral methods we use to approximate the solution $x(s)$. We begin with a brief review of the relevant theory of orthogonal polynomials, Gaussian quadrature, and Fourier series. We include this section primarily for the sake of notation, and we refer the reader to a standard text on orthogonal polynomials [29] for further theoretical details and [16] for a modern perspective on computation.

3.1. Orthogonal polynomials and Gaussian quadrature. Let $\mathbb{P}$ be the space of real polynomials defined on $[-1, 1]$, and let $\mathbb{P}_n \subset \mathbb{P}$ be the space of polynomials

of degree at most n. For any p, q in $\mathbb{P}$, we define the inner product as

$$\langle pq \rangle \equiv \int_{-1}^{1} p(s)q(s)w(s)\,ds. \tag{3.1}$$

We define a norm on $\mathbb{P}$ as $\|p\|_{L^2} = \sqrt{\langle p^2 \rangle}$, which is the standard L^2 norm for the given weight $w(s)$. Let $\{\pi_k(s)\}$ be the set of polynomials that are orthonormal with respect to $w(s)$, i.e., $\langle \pi_i \pi_j \rangle = \delta_{ij}$. It is known that $\{\pi_k(s)\}$ satisfy a three-term recurrence relation

$$\beta_{k+1}\pi_{k+1}(s) = (s - \alpha_k)\pi_k(s) - \beta_k \pi_{k-1}(s), \qquad k = 0, 1, 2, \ldots, \tag{3.2}$$

with $\pi_{-1}(s) = 0$ and $\pi_0(s) = 1$. If we consider only the first n equations, then we can rewrite (3.2) as

$$s\pi_k(s) = \beta_k \pi_{k-1}(s) + \alpha_k \pi_k(s) + \beta_{k+1}\pi_{k+1}(s), \qquad k = 0, 1, \ldots, n-1. \tag{3.3}$$

Setting $\boldsymbol{\pi}_n(s) = [\pi_0(s), \pi_1(s), \ldots, \pi_{n-1}(s)]^T$, we can write this conveniently in matrix form as

$$s\boldsymbol{\pi}_n(s) = \mathbf{J}_n \boldsymbol{\pi}_n(s) + \beta_n \pi_n(s)\mathbf{e}_n, \tag{3.4}$$

where $\mathbf{e}_n$ is a vector of zeros with a one in the last entry and $\mathbf{J}_n$ (known as the *Jacobi matrix*) is a symmetric, tridiagonal matrix defined as

$$\mathbf{J}_n = \begin{bmatrix} \alpha_0 & \beta_1 & & & \\ \beta_1 & \alpha_1 & \beta_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \beta_{n-2} & \alpha_{n-2} & \beta_{n-1} \\ & & & \beta_{n-1} & \alpha_{n-1} \end{bmatrix}. \tag{3.5}$$

The zeros $\{\lambda_i\}$ of $\pi_n(s)$ are the eigenvalues of $\mathbf{J}_n$, and $\boldsymbol{\pi}_n(\lambda_i)$ are the corresponding eigenvectors; this follows directly from (3.4). Let $\mathbf{Q}_n$ be the orthogonal matrix of eigenvectors of $\mathbf{J}_n$, i.e., let $\mathbf{q}_i^n$ be the ith column of $\mathbf{Q}_n$, where

$$\mathbf{q}_i^n = \frac{1}{\|\boldsymbol{\pi}_n(\lambda_i)\|_2} \boldsymbol{\pi}_n(\lambda_i), \tag{3.6}$$

where $\|\cdot\|_2$ is the standard 2-norm on $\mathbb{R}^n$. Then we write the eigenvalue decomposition of $\mathbf{J}_n$ as

$$\mathbf{J}_n = \mathbf{Q}_n \boldsymbol{\Lambda}_n \mathbf{Q}_n^T. \tag{3.7}$$

It is known (cf. [16]) that the eigenvalues $\{\lambda_i\}$ are the familiar Gaussian quadrature points associated with the weight function $w(s)$. The quadrature weight ν_i corresponding to λ_i is equal to the square of the first component of the eigenvector associated with λ_i, i.e.,

$$\nu_i \;=\; \mathbf{Q}(0, i)^2 \;=\; \frac{1}{\|\boldsymbol{\pi}_n(\lambda_i)\|_2^2}. \tag{3.8}$$

The weights $\{\nu_i\}$ are known to be strictly positive. We will use these facts repeatedly in the sequel. For an integrable scalar function $f(s)$, we can approximate its integral by an n-point Gaussian quadrature rule, which is a weighted sum of function evaluations,

$$\int_{-1}^{1} f(s)w(s)\,ds = \sum_{i=0}^{n-1} f(\lambda_i)\nu_i + R_n(f). \tag{3.9}$$

If $f \in \mathbb{P}_{2n-1}$, then $R_n(f) = 0$, that is to say, the *degree of exactness* of the Gaussian quadrature rule is $2n-1$. We use the notation

$$\langle f \rangle_n \equiv \sum_{i=0}^{n-1} f(\lambda_i)\nu_i \tag{3.10}$$

to denote the Gaussian quadrature rule. This is a discrete approximation to the true integral.

3.2. Fourier series. The polynomials $\{\pi_k(s)\}$ form an orthonormal basis for the Hilbert space

$$L^2 \equiv L^2_w([-1,1]) = \{f : [-1,1] \to \mathbb{R} \mid \|f\|_{L^2} < \infty\}. \tag{3.11}$$

Therefore, any $f \in L^2$ admits a convergent *Fourier series*

$$f(s) = \sum_{k=0}^{\infty} \langle f\pi_k \rangle \, \pi_k(s). \tag{3.12}$$

The coefficients $\langle f\pi_k \rangle$ are called the *Fourier coefficients*. If we truncate the series (3.12) after n terms, we are left with a polynomial of degree $n-1$ that is the best approximation polynomial in the L^2 norm. In other words, if we denote

$$P_n f(s) = \sum_{k=0}^{n-1} \langle f\pi_k \rangle \, \pi_k(s), \tag{3.13}$$

then

$$\|f - P_n f\|_{L^2} = \inf_{p \in \mathbb{P}_{n-1}} \|f - p\|_{L^2}. \tag{3.14}$$

In fact, the error made by truncating the series is equal to the sum of squares of the neglected coefficients,

$$\|f - P_n f\|^2_{L^2} = \sum_{k=n}^{\infty} \langle f\pi_k \rangle^2. \tag{3.15}$$

These properties of the Fourier series motivate the theory and practice of spectral methods.

We have shown that each element of the solution $x(s)$ of the parameterized matrix equation is analytic in a region containing the closed interval $[-1,1]$. Therefore, it is continuous and bounded on $[-1,1]$, which implies that $x_i(s) \in L^2$ for $i = 0, \ldots, N-1$. We can thus write the convergent Fourier expansion for each element using vector notation as

$$x(s) = \sum_{k=0}^{\infty} \langle x\pi_k \rangle \, \pi_k(s), \tag{3.16}$$

where the equality is in the L^2 sense. Note that we are abusing the bracket notation here, but this will make further manipulations very convenient. The computational strategy is to choose a truncation level $n-1$ and estimate the coefficients of the truncated expansion.

3.3. Spectral collocation. The term *spectral collocation* typically refers to the technique of constructing a Lagrange interpolating polynomial through the exact solution evaluated at the Gaussian quadrature points. Suppose that λ_i, $i = 0, \ldots, n-1$ are the Gaussian quadrature points for the weight function $w(s)$. We can construct an $n-1$ degree polynomial interpolant of the solution through these points as

$$x_{c,n}(s) \;=\; \sum_{i=0}^{n-1} x(\lambda_i)\ell_i(s) \;\equiv\; \mathbf{X}_c \mathbf{l}_n(s). \tag{3.17}$$

The vector $x(\lambda_i)$ is the solution to the equation $A(\lambda_i)x(\lambda_i) = b(\lambda_i)$. The $n-1$ degree polynomial $\ell_i(s)$ is the standard Lagrange basis polynomial defined as

$$\ell_i(s) = \prod_{j=0,\; j\neq i}^{n-1} \frac{s-\lambda_j}{\lambda_i - \lambda_j}. \tag{3.18}$$

The $N \times n$ constant matrix $\mathbf{X}_c$ (the subscript c is for *collocation*) has one column for each $x(\lambda_i)$, and $\mathbf{l}_n(s)$ is a vector of the Lagrange basis polynomials.

By construction, the collocation polynomial $x_{c,n}$ interpolates the true solution $x(s)$ at the Gaussian quadrature points. We will use this construction to show the connection between the pseudospectral method and the Galerkin method.

3.4. Pseudospectral methods. Notice that computing the true coefficients of the Fourier expansion of $x(s)$ requires the exact solution. The essential idea of the pseudospectral method is to approximate the Fourier coefficients of $x(s)$ by a Gaussian quadrature rule. In other words,

$$x_{p,n}(s) \;=\; \sum_{i=0}^{n-1} \langle x\pi_k \rangle_n \, \pi_k(s) \;\equiv\; \mathbf{X}_p \boldsymbol{\pi}_n(s), \tag{3.19}$$

where $\mathbf{X}_p$ is an $N \times n$ constant matrix of the approximated Fourier coefficients and the subscript p is for *pseudospectral*. For clarity, we recall

$$\langle x\pi_k \rangle_n = \sum_{i=0}^{n-1} x(\lambda_i)\pi_k(\lambda_i)\nu_i, \tag{3.20}$$

where $x(\lambda_i)$ solves $A(\lambda_i)x(\lambda_i) = b(\lambda_i)$. In general, the number of points in the quadrature rule need not have any relationship to the order of truncation. However, when the number of terms in the truncated series is equal to the number of points in the quadrature rule, the pseudospectral approximation is equivalent to the collocation approximation. This relationship is well known in the context of numerical PDEs [4, 21]; nevertheless, we include the following lemma and theorem for use in later proofs.

LEMMA 3.1. *Let* $\mathbf{q}_0$ *be the first row of* $\mathbf{Q}_n$*, and define* $\mathbf{D}_{\mathbf{q}_0} = \operatorname{diag}(\mathbf{q}_0)$*. The matrices* $\mathbf{X}_p$ *and* $\mathbf{X}_c$ *are related by* $\mathbf{X}_p = \mathbf{X}_c \mathbf{D}_{\mathbf{q}_0} \mathbf{Q}_n^T$.

Proof. Using (3.8), write

$$\begin{aligned}\mathbf{X}_p(:,k) &= \langle x\pi_k\rangle_n \\ &= \sum_{j=0}^{n-1} x(\lambda_j)\pi_k(\lambda_j)\nu_j \\ &= \sum_{j=0}^{n-1} \mathbf{X}_c(:,j)\frac{1}{\|\boldsymbol{\pi}_n(\lambda_j)\|_2}\frac{\pi_k(\lambda_j)}{\|\boldsymbol{\pi}_n(\lambda_j)\|_2} \\ &= \mathbf{X}_c\mathbf{D}_{\mathbf{q}_0}\mathbf{Q}_n^T(:,k)\end{aligned}$$

which implies $\mathbf{X}_p = \mathbf{X}_c\mathbf{D}_{\mathbf{q}_0}\mathbf{Q}_n^T$ as required. ∎

Principally one can interpret the matrix $\mathbf{Q}_n\mathbf{D}_{\mathbf{q}_0}^{-1}$ as a *discrete Fourier transform* between the parameter space and the Fourier space. However, we do not adopt this terminology to avoid confusing the reader with the more standard usage involving trigonometric polynomials.

THEOREM 3.2. *The $n-1$ degree collocation approximation is equal to the $n-1$ degree pseudospectral approximation using an n-point Gaussian quadrature rule, i.e.,*

$$x_{c,n}(s) = x_{p,n}(s) \tag{3.21}$$

for all s.

Proof. Note that the elements of $\mathbf{q}_0$ are all nonzero, so $\mathbf{D}_{\mathbf{q}_0}^{-1}$ exists. Then lemma 3.1 implies $\mathbf{X}_c = \mathbf{X}_p\mathbf{Q}_n\mathbf{D}_{\mathbf{q}_0}^{-1}$. Using this change of variables, we can write

$$x_{c,n}(s) \;=\; \mathbf{X}_c\mathbf{l}_n(s) \;=\; \mathbf{X}_p\mathbf{Q}_n\mathbf{D}_{\mathbf{q}_0}^{-1}\mathbf{l}_n(s). \tag{3.22}$$

Thus it is sufficient to show that $\boldsymbol{\pi}_n(s) = \mathbf{Q}_n\mathbf{D}_{\mathbf{q}_0}^{-1}\mathbf{l}_n(s)$. Since this is just a vector of polynomials with degree at most $n-1$, we can do this by multiplying each element by each orthonormal basis polynomial up to order $n-1$ and integrating. Toward this end, we define $\boldsymbol{\Theta} \equiv \langle \mathbf{l}_n\boldsymbol{\pi}_n^T\rangle$.

Using the polynomial exactness of the Gaussian quadrature rule, we compute the i,j element of $\boldsymbol{\Theta}$:

$$\begin{aligned}\boldsymbol{\Theta}(i,j) &= \langle l_i\pi_j\rangle \\ &= \sum_{k=0}^{n-1} \ell_i(\lambda_k)\pi_j(\lambda_k)\nu_k \\ &= \frac{1}{\|\boldsymbol{\pi}_n(\lambda_i)\|_2}\frac{\pi_j(\lambda_i)}{\|\boldsymbol{\pi}_n(\lambda_i)\|_2} \\ &= \mathbf{Q}_n(0,i)\mathbf{Q}_n(j,i),\end{aligned}$$

which implies that $\boldsymbol{\Theta} = \mathbf{D}_{\mathbf{q}_0}\mathbf{Q}_n^T$. Therefore,

$$\begin{aligned}\left\langle \mathbf{Q}_n\mathbf{D}_{\mathbf{q}_0}^{-1}\mathbf{l}_n\boldsymbol{\pi}_n^T\right\rangle &= \mathbf{Q}_n\mathbf{D}_{\mathbf{q}_0}^{-1}\left\langle \mathbf{l}_n\boldsymbol{\pi}_n^T\right\rangle \\ &= \mathbf{Q}_n\mathbf{D}_{\mathbf{q}_0}^{-1}\boldsymbol{\Theta} \\ &= \mathbf{Q}_n\mathbf{D}_{\mathbf{q}_0}^{-1}\mathbf{D}_{\mathbf{q}_0}\mathbf{Q}_n^T \\ &= \mathbf{I}_n,\end{aligned}$$

which completes the proof. ∎

Some refer to the pseudospectral method explicitly as an interpolation method [4]. See [21] for an insightful interpretation in terms of a discrete projection. Because of this property, we will freely interchange the collocation and pseudospectral approximations when convenient in the ensuing analysis.

The work required to compute the pseudospectral approximation is highly dependent on the parameterized system. In general, we assume that the computation of $x(\lambda_i)$ dominates the work; in other words, the cost of computing Gaussian quadrature formulas is negligible compared to computing the solution to each linear system. Then if each $x(\lambda_i)$ costs $\mathcal{O}(N^3)$, the pseudospectral approximation with n terms costs $\mathcal{O}(nN^3)$.

3.5. Spectral Galerkin method. The spectral Galerkin method computes a finite dimensional approximation to $x(s)$ such that each element of the equation residual is orthogonal to the approximation space. Define

$$r(y,s) = A(s)y(s) - b(s). \tag{3.23}$$

The finite dimensional approximation space for each component $x_i(s)$ will be the space of polynomials of degree at most $n-1$. This space is spanned by the first n orthonormal polynomials, i.e., $\text{span}(\pi_0(s),\ldots,\pi_{n-1}(s)) = \mathbb{P}_{n-1}$. We seek an $\mathbb{R}^N$-valued polynomial $x_{g,n}(s)$ of maximum degree $n-1$ such that

$$\langle r_i(x_{g,n})\pi_k\rangle = 0, \qquad i = 0,\ldots,N-1, \qquad k = 0,\ldots,n-1, \tag{3.24}$$

where $r_i(x_{g,n})$ is the ith component of the residual. We can write (3.24) in matrix notation as

$$\left\langle r(x_{g,n})\boldsymbol{\pi}_n^T\right\rangle = \mathbf{0} \tag{3.25}$$

or equivalently

$$\left\langle Ax_{g,n}\boldsymbol{\pi}_n^T\right\rangle = \left\langle b\boldsymbol{\pi}_n^T\right\rangle. \tag{3.26}$$

Since each component of $x_{g,n}(s)$ is a polynomial of degree at most $n-1$, we can write its expansion in $\{\pi_k(s)\}$ as

$$x_{g,n}(s) \;=\; \sum_{k=0}^{n-1} \mathbf{x}_{g,k}\pi_k(s) \;\equiv\; \mathbf{X}_g\boldsymbol{\pi}_n(s), \tag{3.27}$$

where $\mathbf{X}_g$ is a constant matrix of size $N\times n$ and the subscript g is for *Galerkin*. Then (3.26) becomes

$$\left\langle A\mathbf{X}_g\boldsymbol{\pi}_n\boldsymbol{\pi}_n^T\right\rangle = \left\langle b\boldsymbol{\pi}_n^T\right\rangle. \tag{3.28}$$

Using the vec notation [19, section 4.5], we can rewrite (3.28) as

$$\left\langle \boldsymbol{\pi}_n\boldsymbol{\pi}_n^T \otimes A\right\rangle \text{vec}(\mathbf{X}_g) = \left\langle \boldsymbol{\pi}_n \otimes b\right\rangle, \tag{3.29}$$

where $\text{vec}(\mathbf{X}_g)$ is an $Nn\times 1$ constant vector equal to the columns of $\mathbf{X}_g$ stacked on top of each other. The constant matrix $\left\langle \boldsymbol{\pi}_n\boldsymbol{\pi}_n^T \otimes A\right\rangle$ has size $Nn\times Nn$ and a distinct block structure; the i,j block of size $N\times N$ is equal to $\langle\pi_i\pi_j A\rangle$. More explicitly,

$$\left\langle \boldsymbol{\pi}_n\boldsymbol{\pi}_n^T \otimes A\right\rangle = \begin{bmatrix} \langle\pi_0\pi_0 A\rangle & \cdots & \langle\pi_0\pi_{n-1}A\rangle \\ \vdots & \ddots & \vdots \\ \langle\pi_{n-1}\pi_0 A\rangle & \cdots & \langle\pi_{n-1}\pi_{n-1}A\rangle \end{bmatrix}. \tag{3.30}$$

Similarly, the ith block of the $Nn \times 1$ vector $\langle \boldsymbol{\pi}_n \otimes b \rangle$ is equal to $\langle b\pi_i \rangle$, which is exactly the ith Fourier coefficient of $b(s)$. In the language of signal processing, (3.29) can be interpreted as a deconvolution. But we will not adopt this terminology.

Since $A(s)$ is bounded and nonsingular for all $s \in [-1, 1]$, it is straightforward to show that $x_{g,n}(s)$ exists and is unique using the classical Galerkin theorems presented and summarized in Brenner and Scott [5, Chapter 2]. This implies that $\mathbf{X}_g$ is unique, and since $b(s)$ is arbitrary, we conclude that the matrix $\left\langle \boldsymbol{\pi}_n \boldsymbol{\pi}_n^T \otimes A \right\rangle$ is nonsingular for all finite truncations n.

The work required to compute the Galerkin approximation depends on how one computes the integrals in (3.29). If we assume that the cost of forming the system is negligible, then the costly part of the computation is solving the system (3.29). The size of the matrix $\left\langle \boldsymbol{\pi}_n \boldsymbol{\pi}_n^T \otimes A \right\rangle$ is $Nn \times Nn$, so we expect an operation count of $\mathcal{O}(N^3 n^3)$, in general. However, many applications beget systems with sparsity or exploitable structure that can considerably reduce the required work. In particular, if $A(s)$ is sparse (e.g., the discretization of a parameterized PDE), then the blocks $\langle A\pi_i \pi_j \rangle$ will inherit the same sparsity pattern.

Additionally, the block sparsity pattern of the matrix (3.30) will depend on the specific type of parameteric dependence in $A(s)$. If the parameterized matrix $A(s)$ depends polynomially on s, then the matrix (3.30) will have a fixed block bandwidth (for sufficiently large n) that depends on the degree of the polynomial. We make this precise in the following lemma.

LEMMA 3.3. *Suppose that $A(s)$ contains polynomials in s of degree at most m_a. Then for $n > m_a$, the matrix $\langle \boldsymbol{\pi}_n \boldsymbol{\pi}_n \otimes A \rangle$ will have a block bandwidth $2m_a + 1$.*

Proof. Consider the block $\langle A\pi_i \pi_j \rangle$, where we assume that $j > i$. For a polynomial $p_k(s)$ of degree k, we know that $\langle p_k \pi_j \rangle = 0$ for all $j > k$. Therefore,

$$\langle A\pi_i \pi_j \rangle = 0, \qquad j > m_a + i, \tag{3.31}$$

which completes the proof. ∎

3.6. Summary. We have discussed two classes of spectral methods: (i) the interpolatory pseudospectral method which approximates the truncated Fourier series of $x(s)$ by using a Gaussian quadrature rule to approximate each Fourier coefficient, and (ii) the Galerkin projection method which finds an approximation in a finite dimensional subspace of polynomials such that the residual $A(s)x_{g,n}(s) - b(s)$ is orthogonal to the approximation space. In general, the n-term pseudospectral approximation requires n solutions of the original parameterized matrix equation (2.2) evaluated at the Gaussian quadrature points, while the Galerkin method requires the solution of the coupled linear system of equations (3.29) that is n times as large as the original parameterized matrix equation.

Before discussing asymptotic error estimates, we first derive some interesting and useful connections between these two classes of methods. In particular, we can interpret each method as a set of functions acting on the infinite Jacobi matrix for the weight function $w(s)$; the difference between the methods lies in where each truncates the infinite system of equations.

4. Connections between pseudospectral and Galerkin methods. We begin with a useful lemma for representing a matrix of Gaussian quadrature integrals in terms of functions of the Jacobi matrix.

LEMMA 4.1. *Let $f(s)$ be a scalar function analytic in a region containing $[-1, 1]$. Then $\left\langle f \boldsymbol{\pi}_n \boldsymbol{\pi}_n^T \right\rangle_n = f(\mathbf{J}_n)$.*

Proof. We examine the i, j element of the $n \times n$ matrix $f(\mathbf{J}_n)$.

$$\begin{aligned}
\mathbf{e}_i^T f(\mathbf{J}_n)\mathbf{e}_j &= \mathbf{e}_i^T \mathbf{Q}_n f(\mathbf{\Lambda}_n)\mathbf{Q}_n^T \mathbf{e}_j \\
&= (\mathbf{q}_i^n)^T f(\mathbf{\Lambda}_n)\mathbf{q}_j^n \\
&= \sum_{k=0}^{n-1} f(\lambda_k) \frac{\pi_i(\lambda_k)}{\|\boldsymbol{\pi}(\lambda_k)\|_2} \frac{\pi_j(\lambda_k)}{\|\boldsymbol{\pi}(\lambda_k)\|_2} \\
&= \sum_{k=0}^{n-1} f(\lambda_k)\pi_i(\lambda_k)\pi_j(\lambda_k)\nu_k^n \\
&= \langle f\pi_i\pi_j \rangle_n ,
\end{aligned}$$

which completes the proof. ∎

Note that Lemma 4.1 generalizes Theorem 3.4 in [18]. With this in the arsenal, we can prove the following theorem relating the pseudospectral approximation to the Galerkin approximation.

THEOREM 4.2. *The pseudospectral solution is equal to an approximation of the Galerkin solution, where each integral in* (3.29) *is approximated by an* n*-point Gaussian quadrature formula. In other words,* $\mathbf{X}_p$ *solves*

$$\left\langle \boldsymbol{\pi}_n \boldsymbol{\pi}_n^T \otimes A \right\rangle_n \mathrm{vec}(\mathbf{X}_p) = \langle \boldsymbol{\pi}_n \otimes \mathbf{b} \rangle_n . \tag{4.1}$$

Proof. Define the $N \times n$ matrix $\mathbf{B}_c = [b(\lambda_0) \cdots b(\lambda_{n-1})]$, and let $\mathbf{I}$ be the $N \times N$ identity matrix. We proceed to verify the equality as follows. First expand the quadrature rule, and employ (3.6),

$$\begin{aligned}
\left\langle \boldsymbol{\pi}_n \boldsymbol{\pi}_n^T \otimes A \right\rangle_n &= \sum_{k=0}^{n-1} \left(\boldsymbol{\pi}_n(\lambda_k)\boldsymbol{\pi}_n(\lambda_k)^T \otimes A(\lambda_k) \right) \nu_k \\
&= \sum_{k=0}^{n-1} \left(\boldsymbol{\pi}_n(\lambda_k)\boldsymbol{\pi}_n(\lambda_k)^T \otimes A(\lambda_k) \right) \frac{1}{\|\boldsymbol{\pi}_n(\lambda_k)\|_2^2} \\
&= \sum_{k=0}^{n-1} \mathbf{q}_k^n (\mathbf{q}_k^n)^T \otimes A(\lambda_k) \\
&= (\mathbf{Q}_n \otimes \mathbf{I}) A(\mathbf{\Lambda}_n) (\mathbf{Q}_n \otimes \mathbf{I})^T,
\end{aligned}$$

where

$$A(\mathbf{\Lambda}_n) = \begin{bmatrix} A(\lambda_0) & & \\ & \ddots & \\ & & A(\lambda_{n-1}) \end{bmatrix} \tag{4.2}$$

is a block diagonal matrix. Using Lemma 3.1, we have

$$\begin{aligned}
\left\langle \boldsymbol{\pi}_n \boldsymbol{\pi}_n^T \otimes A \right\rangle_n \mathrm{vec}(\mathbf{X}_p) &= (\mathbf{Q}_n \otimes \mathbf{I}) A(\mathbf{\Lambda}_n) (\mathbf{Q}_n \otimes \mathbf{I})^T (\mathbf{Q}_n \otimes \mathbf{I})(\mathbf{D}_{\mathbf{q}_0} \otimes \mathbf{I}) \mathrm{vec}(\mathbf{X}_c) \\
&= (\mathbf{Q}_n \otimes \mathbf{I})(\mathbf{D}_{\mathbf{q}_0} \otimes \mathbf{I}) A(\mathbf{\Lambda}_n) \mathrm{vec}(\mathbf{X}_c) \\
&= (\mathbf{Q}_n \otimes \mathbf{I})(\mathbf{D}_{\mathbf{q}_0} \otimes \mathbf{I}) \mathrm{vec}(\mathbf{B}_c).
\end{aligned}$$

By an argument identical to the proof of Lemma 3.1, we have

$$(\mathbf{Q}_n \otimes \mathbf{I})(\mathbf{D}_{\mathbf{q}_0} \otimes \mathbf{I}) \mathrm{vec}(\mathbf{B}_c) = \langle \boldsymbol{\pi}_n \otimes b \rangle_n \tag{4.3}$$

as required. ∎

Theorem 4.2 is well known in the numerical PDEs context; see Theorem 16 in [4]. This begets a corollary giving conditions for equivalence between Galerkin and pseudospectral approximations.

COROLLARY 4.3. *If $b(s)$ contains only polynomials of maximum degree m_b and $A(s)$ contains only polynomials of maximum degree 1 (i.e., linear functions of s), then $x_{g,n}(s) = x_{p,n}(s)$ for $n \geq m_b$ for all $s \in [-1, 1]$.*

Proof. The parameterized matrix $\boldsymbol{\pi}_n(s)\boldsymbol{\pi}_n(s)^T \otimes A(s)$ and parameterized vector $\boldsymbol{\pi}_n(s) \otimes b(s)$ have polynomials of degree at most $2n-1$. Thus, by the polynomial exactness of the Gaussian quadrature formulas,

$$\left\langle \boldsymbol{\pi}_n \boldsymbol{\pi}_n^T \otimes A \right\rangle_n = \left\langle \boldsymbol{\pi}_n \boldsymbol{\pi}_n^T \otimes A \right\rangle, \qquad \left\langle \boldsymbol{\pi}_n \otimes b \right\rangle_n = \left\langle \boldsymbol{\pi}_n \otimes b \right\rangle. \tag{4.4}$$

Therefore, $\mathbf{X}_g = \mathbf{X}_p$, and consequently

$$x_{g,n}(s) \;=\; \mathbf{X}_g \boldsymbol{\pi}_n(s) \;=\; \mathbf{X}_p \boldsymbol{\pi}_n(s) \;=\; x_{p,n}(s) \tag{4.5}$$

as required. ∎

Using the vec properties and Lemma 4.1, we can squeeze another corollary out of Theorem 4.2.

COROLLARY 4.4. *First define $A(\mathbf{J}_n)$ to be the $Nn \times Nn$ constant matrix with the i, j block of size $n \times n$ equal to $A(i, j)(\mathbf{J}_n)$. Next define $b(\mathbf{J}_n)$ to be the $Nn \times n$ constant matrix with the ith $n \times n$ block equal to $b_i(\mathbf{J}_n)$. Then the pseudospectral coefficients $\mathbf{X}_p$ satisfy*

$$A(\mathbf{J}_n)\mathrm{vec}(\mathbf{X}_p^T) = b(\mathbf{J}_n)\mathbf{e}_0, \tag{4.6}$$

where $\mathbf{e}_0 = [1, 0, \ldots, 0]^T$ is an n-vector.

Proof. Using the vec property applied to (4.1),

$$\left\langle A\mathbf{X}_p \boldsymbol{\pi}_n \boldsymbol{\pi}_n^T \right\rangle_n = \left\langle b \boldsymbol{\pi}_n^T \right\rangle_n. \tag{4.7}$$

Taking the transpose, we get

$$\left\langle \boldsymbol{\pi}_n \boldsymbol{\pi}_n^T \mathbf{X}_p^T A^T \right\rangle_n = \left\langle \boldsymbol{\pi}_n b^T \right\rangle_n. \tag{4.8}$$

Using the vec property again, we get

$$\left\langle A \otimes \boldsymbol{\pi}_n \boldsymbol{\pi}_n^T \right\rangle_n \mathrm{vec}(\mathbf{X}_p^T) = \left\langle b \otimes \boldsymbol{\pi}_n \right\rangle_n. \tag{4.9}$$

Since $\boldsymbol{\pi}_n(s)^T \mathbf{e}_0 = 1$, we can use Lemma 4.1 to get

$$\begin{aligned} \left\langle A \otimes \boldsymbol{\pi}_n \boldsymbol{\pi}_n^T \right\rangle_n &= A(\mathbf{J}_n) \\ \left\langle b \otimes \boldsymbol{\pi}_n \right\rangle_n &= \left\langle b \otimes \boldsymbol{\pi}_n \boldsymbol{\pi}_n^T \mathbf{e}_0 \right\rangle_n \\ &= \left\langle b \otimes \boldsymbol{\pi}_n \boldsymbol{\pi}_n^T \right\rangle_n \mathbf{e}_0 \\ &= b(\mathbf{J}_n)\mathbf{e}_0, \end{aligned}$$

which completes the proof. ∎

Theorem 4.2 leads to a fascinating connection between the matrix operators in the Galerkin and pseudospectral methods, namely, that the matrix in the Galerkin system is equal to a submatrix of the matrix from a sufficiently larger pseudospectral computation. This is the key to understanding the relationship between the Galerkin

and pseudospectral approximations. In the following lemma, we denote the first $r \times r$ principal minor of a matrix $\mathbf{M}$ by $[\mathbf{M}]_{r\times r}$.

LEMMA 4.5. *Let $A(s)$ contain only polynomials of degree at most m_a, and let $b(s)$ contain only polynomials of degree at most m_b. Define*

$$m \equiv m(n) \geq \max\left(\left\lceil \frac{m_a + 2n - 1}{2} \right\rceil \, , \, \left\lceil \frac{m_b + n}{2} \right\rceil\right). \tag{4.10}$$

Then

$$\left\langle \boldsymbol{\pi}_n \boldsymbol{\pi}_n^T \otimes A \right\rangle = \left[\left\langle \boldsymbol{\pi}_m \boldsymbol{\pi}_m^T \otimes A \right\rangle_m\right]_{Nn\times Nn},$$
$$\left\langle \boldsymbol{\pi}_n \otimes b \right\rangle = \left[\left\langle \boldsymbol{\pi}_m \otimes b \right\rangle_m\right]_{Nn\times 1}.$$

Proof. The integrands of the matrix $\left\langle \boldsymbol{\pi}_n \boldsymbol{\pi}_n^T \otimes A \right\rangle$ are polynomials of degree at most $2n+m_a-2$. Therefore, they can be integrated exactly with a Gaussian quadrature rule of order m. A similar argument holds for $\langle \boldsymbol{\pi}_n \otimes b \rangle$. ∎

Combining Lemma 4.5 with Corollary 4.4, we get the following proposition relating the Galerkin coefficients to the Jacobi matrices for $A(s)$ and $b(s)$ that depend polynomially on s.

PROPOSITION 4.6. *Let m, m_a, and m_b be defined as in Lemma 4.5. Define $[A]_n(\mathbf{J}_m)$ to be the $Nn \times Nn$ constant matrix with the i,j block of size $n \times n$ equal to $[A(i,j)(\mathbf{J}_m)]_{n\times n}$ for $i,j = 0,\ldots,N-1$. Define $[b]_n(\mathbf{J}_m)$ to be the $Nn \times n$ constant matrix with the ith $n \times n$ block equal to $[b_i(\mathbf{J}_m)]_{n\times n}$ for $i = 0,\ldots,N-1$. Then the Galerkin coefficients $\mathbf{X}_g$ satisfy*

$$[A]_n(\mathbf{J}_m)\mathrm{vec}(\mathbf{X}_g^T) = [b]_n(\mathbf{J}_m)\mathbf{e}_0, \tag{4.11}$$

where $\mathbf{e}_0 = [1,0,\ldots,0]^T$ is an n-vector.

Notice that Proposition 4.6 provides a way to compute the exact matrix for the Galerkin computation without any symbolic manipulation, but beware that m depends on both n and the largest degree of polynomial in $A(s)$. Written in this form, we have no trouble taking m to infinity, and we arrive at the main theorem of this section.

THEOREM 4.7. *Using the notation of Proposition 4.6 and Corollary 4.4, the coefficients $\mathbf{X}_g$ of the n-term Galerkin approximation of the solution $x(s)$ to (2.2) satisfy the linear system of equations*

$$[A]_n(\mathbf{J}_\infty)\mathrm{vec}(\mathbf{X}_g^T) = [b]_n(\mathbf{J}_\infty)\mathbf{e}_0, \tag{4.12}$$

where $\mathbf{e}_0 = [1,0,\ldots,0]^T$ is an n-vector.

Proof. Since the elements of $A(s)$ and $b(s)$ are continuous over $[-1,1]$, they admit a uniformly convergent sequence of finite degree polynomial approximations [28, Theorem 1.1]. Let $A^{(m_a)}(s)$ and $b^{(m_b)}(s)$ be best polynomial approximations of order $m_a - 1$ and $m_b - 1$, respectively. Since $A(s)$ is analytic and bounded away from singularity for all $s \in [-1,1]$, there exists an integer M such that $A^{(m_a)}(s)$ is also bounded away from singularity for all $s \in [-1,1]$ and all $m_a > M$ (although the bound may depend on m_a). Assume that $m_a > M$.

Define m as in (4.10). Then by Proposition 4.6, the coefficients $\mathbf{X}_g^{(m_a,m_b)}$ of the n-term Galerkin approximation to the solution of the truncated system satisfy

$$[A^{(m_a)}]_n(\mathbf{J}_m)\mathrm{vec}((\mathbf{X}_g^{(m_a,m_b)})^T) = [b^{(m_b)}]_n(\mathbf{J}_m)\mathbf{e}_0. \tag{4.13}$$

By the definition of m (see (4.10)), (4.13) holds for all integers greater than some minimum value. Therefore, we can take $m \to \infty$ without changing the solution, i.e.,

$$[A^{(m_a)}]_n(\mathbf{J}_\infty)\mathrm{vec}((\mathbf{X}_g^{(m_a,m_b)})^T) = [b^{(m_b)}]_n(\mathbf{J}_\infty)\mathbf{e}_0. \tag{4.14}$$

Next we take $m_a, m_b \to \infty$ to get

$$[A^{(m_a)}]_n(\mathbf{J}_\infty) \to [A]_n(\mathbf{J}_\infty),$$
$$[b^{(m_b)}]_n(\mathbf{J}_\infty) \to [b]_n(\mathbf{J}_\infty),$$

which implies

$$\mathbf{X}_g^{(m_a,m_b)} \to \mathbf{X}_g \tag{4.15}$$

as required. ∎

Theorem 4.7 and Corollary 4.4 reveal the fundamental difference between the Galerkin and pseudospectral approximations. We put them side by side for comparison:

$$[A]_n(\mathbf{J}_\infty)\mathrm{vec}(\mathbf{X}_g^T) = [b]_n(\mathbf{J}_\infty)\mathbf{e}_0, \qquad A(\mathbf{J}_n)\mathrm{vec}(\mathbf{X}_p^T) = b(\mathbf{J}_n)\mathbf{e}_0. \tag{4.16}$$

The difference lies in where the truncation occurs. For the pseudospectral, the infinite Jacobi matrix is first truncated, and then the operator is applied. For the Galerkin, the operator is applied to the infinite Jacobi matrix, and the resulting system is truncated. The question that remains is whether it matters. As we will see in the error estimates in the next section, the interpolating pseudospectral approximation converges at a rate comparable to the Galerkin approximation.

5. Error estimates. Asymptotic error estimates for polynomial approximation are well established in many contexts, and the theory is now considered classical. Our goal is to apply the classical theory to relate the rate of geometric convergence to some measure of singularity for the solution. We do not seek the tightest bounds in the most appropriate norm as in [7], but instead we offer intuition for understanding the asymptotic rate of convergence. We also present a residual error estimate that may be more useful in practice. We complement the analysis with two representative numerical examples.

To discuss convergence, we need to choose a norm. In the statements and proofs, we will use the standard L^2 and L^∞ norms generalized to $\mathbb{R}^N$-valued functions.

Definition 5.1. *For a function $f : \mathbb{R} \to \mathbb{R}^N$, define the L^2 and L^∞ norms as*

$$\|f\|_{L^2} := \sqrt{\sum_{i=0}^{N-1} \int_{-1}^{1} f_i^2(s) w(s)\, ds}, \tag{5.1}$$

$$\|f\|_{L^\infty} := \max_{0 \le i \le N-1} \left(\sup_{-1 \le s \le 1} |f_i(s)| \right). \tag{5.2}$$

With these norms, we can state error estimates for both Galerkin and pseudospectral methods.

Theorem 5.2 (Galerkin asymptotic error estimate). *Let ρ^* be the sum of the semiaxes of the greatest ellipse with foci at ± 1 in which $x_i(s)$ is analytic for $i = 0, \ldots, N-1$. Then, for $1 < \rho < \rho^*$, the asymptotic error in the Galerkin approximation is*

$$\|x - x_{g,n}\|_{L^2} \le C\rho^{-n}, \tag{5.3}$$

where C is a constant independent of n.

Proof. We begin with the standard error estimate for the Galerkin method [7, section 6.4] in the L^2 norm,

$$\|x - x_{g,n}\|_{L^2} \leq C \|x - R_n x\|_{L^2}. \tag{5.4}$$

The constant C is independent of n but depends on the extremes of the bounded eigenvalues of $A(s)$. Under the consistency hypothesis, the operator R_n is a projection operator such that

$$\|x_i - R_n x_i\|_{L^2} \to 0, \qquad n \to \infty, \tag{5.5}$$

for $i = 0, \ldots, N-1$. For our purpose, we let $R_n x$ be the expansion of $x(s)$ in terms of the Chebyshev polynomials,

$$R_n x(s) = \sum_{k=0}^{n-1} \mathbf{a}_k T_k(s), \tag{5.6}$$

where $T_k(s)$ is the kth Chebyshev polynomial. Since $x(s)$ is continuous for all $s \in [-1, 1]$ and $w(s)$ is normalized, we can bound

$$\|x - R_n x\|_{L^2} \leq \sqrt{N} \|x - R_n x\|_{L^\infty}. \tag{5.7}$$

The Chebyshev series converges uniformly for functions that are continuous on $[-1, 1]$, so we can bound

$$\|x - R_n x\|_{L^\infty} = \left\| \sum_{k=n}^{\infty} \mathbf{a}_k T_k(s) \right\|_{L^\infty} \tag{5.8}$$

$$\leq \left\| \sum_{k=n}^{\infty} |\mathbf{a}_k| \right\|_{\infty} \tag{5.9}$$

since $-1 \leq T_k(s) \leq 1$ for all k. To be sure, the quantity $|\mathbf{a}_k|$ is the componentwise absolute value of the constant vector $\mathbf{a}_k$, and the norm $\|\cdot\|_\infty$ is the standard infinity norm on $\mathbb{R}^N$.

Using the classical result stated in [20, section 3], we have

$$\limsup_{k \to \infty} |\mathbf{a}_{k,i}|^{1/k} = \frac{1}{\rho_i^*}, \qquad i = 0, \ldots, N-1, \tag{5.10}$$

where ρ_i^* is the sum of the semiaxes of the greatest ellipse with foci at ± 1 in which $x_i(s)$ is analytic. This implies that, asymptotically, the error in the Chebyshev expansion decays roughly like $\mathcal{O}(\rho_i^{-k})$ for $1 < \rho_i < \rho_i^*$. We take $\rho = \min_i \rho_i$, which suffices to prove the estimate (5.3). ∎

Theorem 5.2 recalls the well-known fact that the convergence of many polynomial approximations (e.g., power series, Fourier series) depends on the size of the region in the complex plane in which the function is analytic. Thus, the location of the singularity nearest the interval $[-1, 1]$ determines the rate at which the approximation converges as one includes higher powers in the polynomial approximation. Next we derive a similar result for the pseudospectral approximation using the fact that it interpolates $x(s)$ at the Gaussian points of the weight function $w(s)$.

THEOREM 5.3 (pseudospectral asymptotic error estimate). *Let ρ^* be the sum of the semiaxes of the greatest ellipse with foci at ± 1 in which $x_i(s)$ is analytic for $i = 0, \ldots, N-1$. Then, for $1 < \rho < \rho^*$, the asymptotic error in the pseudospectral approximation is*

$$\|x - x_{p,n}\|_{L^2} \leq C\rho^{-n}, \tag{5.11}$$

where C is a constant independent of n.

Proof. Recall that $x_{c,n}(s)$ is the Lagrange interpolant of $x(s)$ at the Gaussian points of $w(s)$, and let $x_{c,n,i}(s)$ be the ith component of $x_{c,n}(s)$. We will use the result from [28, Theorem 4.8] that

$$\int_{-1}^{1} (x_i(s) - x_{c,n,i}(s))^2 w(s)\, ds \leq 4E_n^2(x_i), \tag{5.12}$$

where $E_n(x_i)$ is the error of the best approximation polynomial in the uniform norm. We can, again, bound $E_n(x_i)$ by the error of the Chebyshev expansion (5.6). Using Theorem 3.2 with (5.12),

$$\begin{aligned} \|x - x_{p,n}\|_{L^2} &= \|x - x_{c,n}\|_{L^2} \\ &\leq 2\sqrt{N}\, \|x - R_n x\|_{L^\infty}. \end{aligned}$$

The remainder of the proof proceeds exactly as the proof of Theorem 5.2. ∎

We have shown, using classical approximation theory, that the interpolating pseudospectral method and the Galerkin method have the same asymptotic rate of geometric convergence. This rate of convergence depends on the size of the region in the complex plane where the functions $x(s)$ are analytic. The structure of the matrix equation reveals at least one singularity that occurs when $A(s^*)$ is rank deficient for some $s^* \in \mathbb{R}$, assuming the right-hand side $b(s^*)$ does not fortuitously remove it. For a general parameterized matrix, this fact may not be useful. However, for many parameterized systems in practice, the range of the parameter is dictated by the existence and/or stability criteria. The value that makes the system singular is often known and has some interpretation in terms of the model. For example, the solution of an elliptic PDE with parameterized coefficients has a singularity in the parameter space where the coefficients touch zero. In such cases, one may have an upper bound on ρ, which is the sum of the semiaxes of the ellipse of analyticity, and this can be used to estimate the geometric rate of convergence a priori.

We end this section with a residual error estimate—similar to residual error estimates for constant matrix equations—that may be more useful in practice than the asymptotic results.

THEOREM 5.4. *Define the residual $r(y, s)$ as in (3.23), and let $e(y, s) = x(s) - y(s)$ be the $\mathbb{R}^N$-valued function representing the error in the approximation $y(s)$. Then*

$$C_1 \|r(y)\|_{L^2} \leq \|e(y)\|_{L^2} \leq C_2 \|r(y)\|_{L^2} \tag{5.13}$$

for some constants C_1 and C_2, which are independent of $y(s)$.

Proof. Since $A(s)$ is nonsingular for all $s \in [-1, 1]$, we can write

$$A^{-1}(s)r(y, s) \;=\; y(s) - A^{-1}(s)b(s) \;=\; e(y, s) \tag{5.14}$$

so that

$$\begin{aligned}\|e(y)\|_{L^2}^2 &= \left\langle e(y)^T e(y)\right\rangle \\ &= \left\langle r^T(y) A^{-T} A^{-1} r(y)\right\rangle .\end{aligned}$$

Since $A(s)$ is bounded, so is $A^{-1}(s)$. Therefore, there exist constants C_1^* and C_2^* that depend only on $A(s)$ such that

$$C_1^* \left\langle r^T(y) r(y)\right\rangle \le \left\langle e^T(y) e(y)\right\rangle \le C_2^* \left\langle r^T(y) r(y)\right\rangle . \tag{5.15}$$

Taking the square root yields the desired result. ∎

Theorem 5.4 states that the L^2 norm of the residual behaves like the L^2 norm of the error. In many cases, this residual error may be much easier to compute than the true L^2 error. However, as in residual error estimates for constant matrix problems, the constants in Theorem 5.4 will be large if the bounds on the eigenvalues of $A(s)$ are large. We apply these results in the next section with two numerical examples.

6. Numerical examples. We examine two simple examples of spectral methods applied to parameterized matrix equations. The first is a 2×2 symmetric parameterized matrix, and the second comes from a discretized second order ODE. In both cases, we relate the convergence of the spectral methods to the size of the region of analyticity and verify this relationship numerically. We also compare the behavior of the true error to the behavior of the residual error estimate from Theorem 5.4.

To keep the computations simple, we use a constant weight function $w(s)$. The corresponding orthonormal polynomials are the normalized Legendre polynomials, and the Gauss points are the Gauss–Legendre points.

6.1. A 2×2 parameterized matrix equation. Let $\varepsilon > 0$, and consider the following parameterized matrix equation:

$$\begin{bmatrix} 1+\varepsilon & s \\ s & 1 \end{bmatrix} \begin{bmatrix} x_0(s) \\ x_1(s) \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix} . \tag{6.1}$$

For this case, we can easily compute the exact solution,

$$x_0(s) = \frac{2-s}{1+\varepsilon-s^2}, \qquad x_1(s) = \frac{1+\varepsilon-2s}{1+\varepsilon-s^2}. \tag{6.2}$$

Both of these functions have poles at $s = \pm\sqrt{1+\varepsilon}$, so the sum of the semiaxes of the ellipse of analyticity is bounded, i.e., $\rho < \sqrt{1+\varepsilon}$. Notice that the matrix is linear in s, and the right-hand side has no dependence on s. Thus, Corollary 4.3 implies that the Galerkin approximation is equal to the pseudospectral approximation for all n; there is no need to solve the system (3.29) to compute the Galerkin approximation. In Figure 6.1 we plot both the true L^2 error and the residual error estimate for four values of ε. The results confirm the analysis.

6.2. A parameterized second order ODE. Consider the second order boundary value problem

$$\frac{d}{dt}\left(\alpha(s,t)\frac{du}{dt}\right) = 1, \qquad t \in [0,1], \tag{6.3}$$

$$u(0) = 0, \tag{6.4}$$

$$u(1) = 0, \tag{6.5}$$

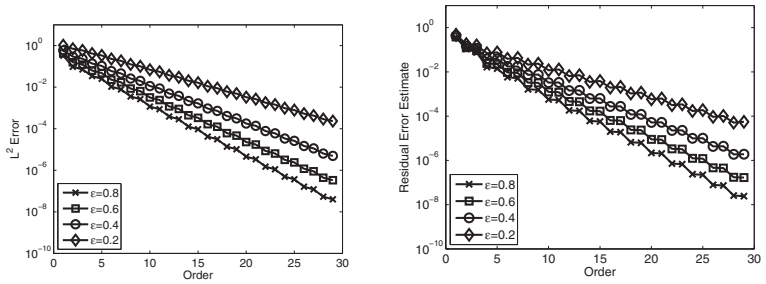

FIG. 6.1. *The convergence of the spectral methods applied to* (6.1). *The figure on the left plots the* L^2 *error as the order of approximation increases, and the figure on the right plots the residual error estimate. The stairstep behavior is related to the fact that* $x_0(s)$ *and* $x_1(s)$ *are odd functions over* $[-1, 1]$.

where, for $\varepsilon > 0$,

$$\alpha(s,t) = 1 + 4\cos(\pi s)(t^2 - t), \qquad s \in [\varepsilon, 1]. \tag{6.6}$$

The exact solution is

$$u(s,t) = \frac{1}{8\cos(\pi s)} \ln\left(1 + 4\cos(\pi s)(t^2 - t)\right). \tag{6.7}$$

The solution $u(s,t)$ has a singularity at $s = 0$ and $t = 1/2$. Notice that we have adjusted the range of s to be bounded away from 0 by ε. We use a standard piecewise linear Galerkin finite element method with 512 elements in the t domain to construct a stiffness matrix parameterized by s, i.e.,

$$(K_0 + \cos(\pi s)K_1)x(s) = b. \tag{6.8}$$

Figure 6.2 shows the convergence of the residual error estimate for both Galerkin and pseudospectral approximations as n increases. (Despite having the exact solution (6.7) available, we do not present the decay of the L^2 error; it is dominated entirely by the discretization error in the t domain.) As ε gets closer to zero, the geometric convergence rate of the spectral methods degrades considerably. Also note that each element of the parameterized stiffness matrix is an analytic function of s, but Figure 6.2 verifies that the less expensive pseudospectral approximation converges at the same rate as the Galerkin approximation.

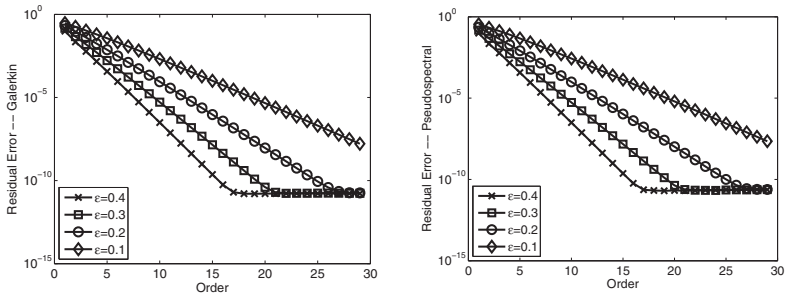

FIG. 6.2. *The convergence of the residual error estimate for the Galerkin and pseudospectral approximations applied to the parameterized matrix equation* (6.8).

7. Summary and conclusions. We have presented an application of spectral methods to parameterized matrix equations. Such parameterized systems arise in many applications. The goal of a spectral method is to construct a global polynomial approximation of the $\mathbb{R}^N$-valued function that satisfies the parameterized system.

We derived two basic spectral methods: (i) the interpolatory pseudospectral method, which approximates the coefficients of the truncated Fourier series with Gaussian quadrature formulas, and (ii) the Galerkin method, which finds an approximation in a finite dimensional subspace by requiring that the residual be orthogonal to the approximation space. The primary work involved in the pseudospectral method is solving the parameterized system at a finite set of parameter values, whereas the Galerkin method requires the solution of a coupled system of equations many times larger than the original parameterized system.

We showed that one can interpret the differences between these two methods as a choice of when to truncate an infinite linear system of equations. Employing this relationship we derived conditions under which these two approximations are equivalent. In this case, there is no reason to solve the large coupled system of equations for the Galerkin approximation.

Using classical techniques, we presented asymptotic error estimates relating the decay of the error to the size of the region of analyticity of the solution; we also derived a residual error estimate that may be more useful in practice. We verified the theoretical developments with two numerical examples: a 2×2 matrix equation and a finite element discretization of a parameterized second order ODE.

The popularity of spectral methods for PDEs stems from their *infinite* (i.e., geometric) order of convergence for smooth functions compared to finite difference schemes. We have the same advantage in the case of parameterized matrix equations, plus the added bonus that there are no boundary conditions to consider. The primary concern for these methods is determining the value of the parameter closest to the domain that renders the system singular.

Acknowledgment. We would like to thank James Lambers for his helpful and insightful feedback.

REFERENCES

[1] I. Babuška, M. K. Deb, and J. T. Oden, *Solution of stochastic partial differential equations using Galerkin finite element techniques*, Comput. Methods Appl. Mech. Engrg., 190 (2001), pp. 6359–6372.

[2] I. Babuška, F. Nobile, and R. Tempone, *A stochastic collocation method for elliptic partial differential equations with random input data*, SIAM J. Numer. Anal., 45 (2007), pp. 1005–1034.

[3] I. Babuška, R. Tempone, and G. E. Zouraris, *Galerkin finite element approximations of stochastic elliptic partial differential equations*, SIAM J. Numer. Anal., 42 (2004), pp. 800–825.

[4] J. P. Boyd, *Chebyshev and Fourier Spectral Methods*, 2nd ed., Dover, New York, 2001.

[5] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, 2nd ed., Springer, New York, 2002.

[6] C. Brezinski and M. Redivo-Zaglia, *The PageRank vector: Properties, computation, approximation, and acceleration*, SIAM J. Matrix Anal. Appl., 28 (2006), pp. 551–575.

[7] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang, *Spectral Methods: Fundamentals in Single Domains*, Springer, New York, 2006.

[8] S. Chandrasekaran, G. H. Golub, M. Gu, and A. H. Sayed, *Parameter estimation in the presence of bounded data uncertainties*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 235–252.

[9] J. Chung and J. G. Nagy, *Nonlinear least squares and super resolution*, J. Phys. Conf. Ser., 124 (2008), p. 012019.

[10] P. G. Constantine and D. F. Gleich, *Random Alpha PageRank*, Internet Math., to appear.
[11] L. Dieci and L. Lopez, *Lyapunov exponents of systems evolving on quadratic groups*, SIAM J. Matrix Anal. Appl., 24 (2003), pp. 1175–1185.
[12] H. C. Elman, O. G. Ernst, D. P. O'Leary, and M. Stewart, *Efficient iterative algorithms for the stochastic finite element method with application to acoustic scattering*, Comput. Methods Appl. Mech. Engrg., 194 (2005), pp. 1037–1055.
[13] O. G. Ernst, C. E. Powell, D. J. Silvester, and E. Ullmann, *Efficient solvers for a linear stochastic Galerkin mixed formulation of diffusion problems with random data*, SIAM J. Sci. Comput., 31 (2009), pp. 1424–1447.
[14] P. Frauenfelder, C. Schwab, and R. A. Todor, *Finite elements for elliptic problems with stochastic coefficients*, Comput. Methods Appl. Mech. Engrg., 194 (2005), pp. 205–228.
[15] W. Gautschi, *The interplay between classical analyses and (numerical) linear algebra—A tribute to Gene Golub*, Electron. Trans. Numer. Anal., 13 (2002), pp. 119–147.
[16] W. Gautschi, *Orthogonal Polynomials: Computation and Approximation*, Clarendon Press, Oxford, 2004.
[17] R. Ghanem and P. Spanos, *Stochastic Finite Elements: A Spectral Approach*, Springer, New York, 1991.
[18] G. H. Golub and G. Meurant, *Matrices, Moments and Quadrature with Applications*, Princeton University Press, Princeton, NJ, 2010.
[19] G. H. Golub and C. F. VanLoan, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, MD, 1996.
[20] D. Gottlieb and S. A. Orszag, *Numerical Analysis of Spectral Methods: Theory and Applications*, SIAM, Philadelphia, 1977.
[21] J. S. Hesthaven, S. Gottlieb, and D. Gottlieb, *Spectral Methods for Time Dependent Problems*, Cambridge University Press, Cambridge, 2007.
[22] T. Kato, *Perturbation Theory for Linear Operators*, 2nd ed., Springer, New York, 1980.
[23] K. Meerbergen and Z. Bai, *The Lanczos method for parameterized symmetric linear systems with multiple right-hand sides*, J. Matrix Anal. Appl., 31 (2010), pp. 1642–1662.
[24] C. D. Meyer, *Matrix Analysis and Applied Linear Algebra*, SIAM, Philadelphia, 2000.
[25] M. Oliver and R. Webster, *Kriging: A method of interpolation for geographical information systems*, Internat. J. Geograph. Inform. Sci., 4 (1990), pp. 313–332.
[26] L. Page, S. Brin, R. Motwani, and T. Winograd, *The pagerank citation ranking: Bringing order to the web*, Technical report 1999-66, Stanford University, Stanford, CA, 1999.
[27] C. E. Powell and H. C. Elman, *Block-diagonal preconditioning for spectral stochastic finite-element systems*, IMA J. Numer. Anal., 29 (2009), pp. 350–375.
[28] T. J. Rivlin, *An Introduction to the Approximation of Functions*, Blaisdell Publishing, New York, 1969.
[29] G. Szegö, *Orthogonal Polynomials*, American Mathematical Society, Providence, RI, 1939.
[30] Q. Wang, P. Moin, and G. Iaccarino, *A rational interpolation scheme with superpolynomial rate of convergence*, SIAM J. Numer. Anal., 47 (2010), pp. 4073–4097.
[31] D. Xiu and J. S. Hesthaven, *High-order collocation methods for differential equations with random inputs*, SIAM J. Sci. Comput., 27 (2005), pp. 1118–1139.
[32] D. Xiu and G. E. Karniadakis, *The Wiener–Askey polynomial chaos for stochastic differential equations*, SIAM J. Sci. Comput., 24 (2002), pp. 619–644.