Integration preconditioners for differential operators in spectral τ-methods

E. A. Coutsias* T. Hagstrom† J. S. Hesthaven‡ D. Torres§

Abstract

We present simple, banded preconditioners that transform linear ordinary differential operators with polynomial coefficients into banded form. These are applicable to a wide class of Galerkin approximation problems, including expansions in terms of all the classical orthogonal polynomials. The preconditioners are in fact the n-th order integration operators for the polynomial families employed in the Galerkin approximation, with n the order of the differential operator. The resulting matrix problems are algorithmically simpler, as well as better conditioned than the original forms. The good conditioning allows the extension of our ideas even to problems with arbitrary, nonsingular coefficients as well as to certain quasilinear problems by the use of iterative methods. We also present extensions to partial differential operators with polynomial coefficients by considering preconditioners in the form of tensor products of appropriate combinations of integration operators. The origin of the tridiagonal integration operators for arbitrary classical orthogonal polynomial families is shown to lie with the Gauss contiguity relations for Hypergeometric functions.

Key words: spectral methods, orthogonal polynomials, boundary value problems.

AMS subject classifications: 65Q05, 65L60, 65P20, 76-08, 33A45, 33A50, 33A65.

*Dept. of Mathematics and Statistics, University of New Mexico, Albuquerque, NM 87131. E-mail: *vageli@math.unm.edu*

†Dept. of Mathematics and Statistics, University of New Mexico, Albuquerque, NM 87131. E-mail: *hagstrom@math.unm.edu*

‡Association EURATOM - Risø National Laboratory, Optics and Fluid Dynamics Department, P.O. Box 49, DK-4000, Roskilde, Denmark. E-mail: *jsh@risoe.dk*

§Dept. of Mathematics and Statistics, University of New Mexico, Albuquerque, NM 87131. E-mail: *dtorres@math.unm.edu*

1 Introduction

The classical orthogonal polynomial bases most commonly used in Numerical Analysis originate as eigenfunctions of singular Sturm-Liouville problems. The derivatives of such polynomials form an orthogonal basis as well; in fact they are also classical orthogonal polynomials. As a result of the Gauss contiguity relations for Hypergeometric Functions [1], elements of the original basis have a simple expression in terms of elements of the derivative basis, involving at most three terms of contiguous degrees. Consequently, although the matrices representing differentiation to various orders in terms of an orthogonal polynomial basis are almost full upper triangular matrices (only exception are the Hermite polynomials), those representing integration are banded, of bandwidth $2n+1$, with n the order of integration. This, together with the fact that multiplication by a monomial is also a tridiagonal matrix leads to simple, banded preconditioners which simultaneously band classes of matrices of the form PD^m, with P the operator of multiplication by a polynomial $p(x)$ and D^m the operator of m-fold differentiation. This allows the formulation of efficient algorithms for the solution of differential equations with polynomial coefficients in any of the classical polynomial bases. In Sec. 2 we establish the above facts and discuss how they can be used in the context of the Lanczos τ-method to construct a spectrally accurate solution of a differential equation with general polynomial (and rational) coefficients in $\mathcal{O}(M)$ operations with M the truncation order. The appeal of the method is due not only to the efficiency of solution, but also to the excellent conditioning of the resulting matrix problems [6]. This latter property, established under certain general assumptions in Sec. 3 permits the extension to problems with arbitrary nonsingular coefficients as well as certain nonlinear problems by the use of iterative methods, and we present examples in Sec. 4.

In Sec. 5 we show how the ideas can be extended to higher dimensional problems. The pre-conditioning scheme presented here has the advantage that it can be extended easily to treat problems in multidimensions. Additional dimensions are included in the banded matrix formulation through tensor products, which amounts to re-

placing entries in banded matrices by blocks which are themselves banded. A natural way of deriving such block-banded forms for the higher dimensional case is by interpreting integration and differentiation as change of basis transformations among related polynomial families. Effectively, then, we consider expansions in each variable in terms of a basis of derivative polynomials of order equal to the maximum order derivatives in that variable present in the given equation. The numerical solution of several simple test cases is presented. As an alternative to finite-difference based time stepping schemes for time dependent problems, we apply our method to problems in one space dimension plus time. Functions are expanded in a double spectral expansion, in both the space and time variables. The problems are solved by inversion of the block-banded matrix resulting from the application of integration preconditioners of appropriate order in each variable. The operation count for a problem discretized to a $N \times M$ resolution is $min(\mathcal{O}(M^3N), \mathcal{O}(N^3M))$ for a single solution but it is lower if the same matrix is inverted several times, say as a result for solving with different forcing functions etc, in which case the cost is $min(\mathcal{O}(M^2N), \mathcal{O}(N^2M))$. These numbers reflect the alternatives available when deriving the block-banded forms for the differential operator. Analogous estimates also hold for higher dimensional problems. Finally, in Sec. 6 we present some concluding remarks and further connections with previous works.

In a previous paper [6] we examined the use of the spectral integration operators as postconditioners for linear differential operators with polynomial coefficients in arbitrary bases of classical orthogonal polynomials. In that form, the method was a generalization of the method of treating a differential equation by expanding the highest derivative in terms of Chebyshev polynomials originally introduced by Clenshaw [4]. The integration preconditioner for the Chebyshev polynomials is also analyzed by Greengard [11], while the recurrence relation for the derivatives of the Jacobi polynomials has also appeared in a recent review by Fornberg [7].

2 The method

Throughout, we assume that we are working with a family of polynomials $\{Q_k\}_0^\infty$ which are orthogonal and complete over the interval (a, b) (here a and/or b can be infinite) with respect to the weight $w(x)$. In the cases of interest, these are the eigenfunctions of a Sturm-Liouville (SL) problem,

$$(p(x)Q_k')' + \lambda_k w(x) Q_k = 0, \tag{1}$$

so that the Q_k' form an orthogonal family as well, with weight $p(x)$ which satisfies $p(x) \to 0$ as $x \to a, b$. We will assume that the functions under consideration possess sufficient differentiability properties over (a, b) and can be expressed as a series involving the Q_k. See [3] for a discussion of the convergence properties and the introduction of relevant function spaces.

2.1 Integration operators and derivative bases

We write $\langle Q_l, Q_m \rangle_w = \int_a^b Q_m Q_l w(x) dx = h_m \delta_{lm}$, with h_m the norming constants of the Q_m [1]. We set $Q_n(x) = \sum_0^n K_{nm} x^m$. It is well known that all orthogonal polynomial families share a three-term recurrence of the form [1]:

$$\sum_{l=-1}^{1} Q_{k+l} a_{k+l,k} = x Q_k \quad , \quad k = 0, 1, \cdots \tag{2}$$

This follows since, if $l < k - 1$, $deg(xQ_l) = l + 1 < k$ and, by orthogonality, $\langle xQ_k, Q_l \rangle_w = \langle Q_k, xQ_l \rangle_w = 0$. By matching powers we easily see that [14]:

$$a_{n,n-1} = \frac{K_{n-1,n-1}}{K_{n,n}} \quad , \quad a_{n,n+1} = \frac{h_{n+1}}{h_n} a_{n+1,n},$$
$$a_{n,n} = \frac{K_{n,n-1}}{K_{n,n}} - \frac{K_{n+1,n}}{K_{n+1,n+1}} \tag{3}$$

the second relationship following since

$$h_{n+1} a_{n+1,n} = \langle xQ_n, Q_{n+1} \rangle_w = \langle xQ_{n+1}, Q_n \rangle_w = h_n a_{n,n+1}.$$

In many important cases, including the classical orthogonal polynomials (i.e. Jacobi, Chebyshev, Legendre, Gegenbauer, Hermite and Laguerre polynomials) there also holds a relation of the form

$$\sum_{l=-1}^{1} Q_{k+l}' b_{k+l,k} = Q_k \quad , \quad k = 0, 1, \cdots . \tag{4}$$

Here, as well as in (2), we introduced $Q_{-1} = 0$. In fact:

Lemma 2.1 *Suppose that the eigenfunctions of the SL problem (1) are polynomials. Then their derivatives, $\{Q_k'\}_1^\infty$ also constitute an orthogonal family with respect to weight $p(x)$ where it is assumed that $p(x) > 0$ for $x \in (a, b)$ and $p(a) = p(b) = 0$. Moreover, the expression of Q_k $(k > 1)$ in terms of the Q_k' involves at most only $\{Q_k'\}_{k-1}^{k+1}$, i.e. it has the form (4).*

Proof: Integrating by parts we see

$$\begin{aligned} \int_a^b Q_k' Q_l' p(x) dx &= Q_k Q_l' p \Big|_a^b - \int_a^b Q_k (Q_l' p)' dx \\ &= \lambda_k \int_a^b Q_k Q_l w dx = \lambda_k h_k \delta_{kl} \end{aligned}$$

since $p(x)$ vanishes at the end-points, and the orthogonality of the Q'_k follows from that of the Q_k. We introduce $h'_{k-1} = \lambda_k h_k$ for the norming constants of the Q'_k, indexing according to the degree of the corresponding polynomial, namely $\deg(Q'_k) = k-1$. Now, since the $\{Q'_l\}_1^{k+1}$ are independent, we have that $Q_k = \sum_1^{k+1} Q'_l b_{l,k}$ and in order to establish (4) we must show that $\langle Q_k, Q'_l \rangle_p = 0$ for $l = 1, \ldots, k-2$. Clearly

$$\begin{aligned} \langle Q_k, Q'_l \rangle_p &= Q_k Q'_l p x \big|_a^b - \int_a^b \left(x Q'_k Q'_l p + Q_k x (Q'_l p)' \right) dx \\ &= -\langle Q'_k, x Q'_l \rangle_p + \lambda_l \langle Q_k, x Q_l \rangle_w = 0 \end{aligned}$$

where the first integral vanishes due to the orthogonality of the Q'_k since $deg(xQ'_l) = l \leq k-2 < k-1 = deg(Q'_k)$, while the second integral vanishes since $deg(Q_k) = k > k-1 \geq l+1 = deg(xQ_l)$.

Note: for the classical orthogonal polynomials, the derivatives of arbitrary order are also classical orthogonal polynomials. This follows easily from general properties of the Hypergeometric functions $F(a,b;c;z)$ [1]. For instance, for the Jacobi polynomials we have

$$P_n^{(\alpha,\beta)}(1-2z) = \frac{(\alpha+1)_n}{n!} F(-n, n+\alpha+\beta+1; \alpha+1; z) \ .$$

Since

$$\frac{d}{dz} F(a,b;c;z) = \frac{ab}{c} F(a+1, b+1; c+1; z), \tag{5}$$

it follows that

$$\frac{d}{dx} P_n^{(\alpha,\beta)}(x) = \frac{n+\alpha+\beta+1}{2} P_{n-1}^{(\alpha+1,\beta+1)}(x) \ . \tag{6}$$

Similarly, for the Laguerre and Hermite polynomials we have

$$\frac{d}{dx} L_n^{(\alpha)}(x) = -L_{n-1}^{(\alpha+1)}(x) \ , \ \frac{d}{dx} H_n = 2n H_{n-1}.$$

In all cases, differentiation can be seen as a change of basis, always within the set of classical orthogonal polynomials. We let A, B be the coefficient matrices in (2,4) respectively. The relation between the two sets of coefficients is found from the following:

Lemma 2.2 *The coefficients $b_{m,n}$ in (4) are found from those in (2) as*

$$b_{n+1,n} = \frac{1}{n+1} a_{n+1,n} \ ,$$

$$b_{n-1,n} = \left(1 - \frac{(n-1)\lambda_n}{n\lambda_{n-1}} \right) a_{n-1,n},$$

and

$$b_{n,n} = \frac{1}{n+1} a_{n,n} - \frac{1}{n(n+1)} \left(\sum_{l=1}^{n-1} a_{l,l} - \frac{K_{1,0}}{K_{1,1}} \right).$$

Proof: Since the Q'_k are orthogonal, they must satisfy a relation of form (2):

$$\sum_{l=-1}^{1} Q'_{k+l+1} a'_{k+l,k} = x Q'_{k+1} \quad , \quad k = 0, 1, \cdots \tag{7}$$

Since $Q'_{n+1}(x) = \sum_{m=0}^{n} K'_{nm} x^m = \sum_0^n (m+1) K_{n+1,m+1} x^m$, i.e. $K'_{n,m} = (m+1) K_{n+1,m+1}$, it follows from (3) that

$$a'_{n+1,n} = \frac{n+1}{n+2} a_{n+2,n+1} \ ,$$

$$a'_{n-1,n} = \frac{n}{n+1} \frac{\lambda_{n+1} h_{n+1}}{\lambda_n h_n} a_{n+1,n},$$

$$a'_{n,n} = \frac{n}{n+1} \frac{K_{n+1,n}}{K_{n+1,n+1}} - \frac{n+1}{n+2} \frac{K_{n+2,n+1}}{K_{n+2,n+2}}.$$

Differentiating (2) and using (7) there results:

$$Q_k = \sum_{l=-1}^{1} Q'_{k+l} \left(a_{k+l,k} - a'_{k+l-1,k-1} \right),$$

and the claim follows after some algebra.

Note: The computation of the $b_{i,j}$ from the above expressions is not very practical, and was only given to establish the connection to the $a_{i,j}$. A more direct calculation in the case of the Jacobi polynomials follows from their connection to the hypergeometric functions [14]. Indeed, differentiating (2) for the Jacobis, multiplying (6) by x and expressing the latter in terms of the $P_k^{(\alpha+1,\beta+1)}$ using (2) we find

$$b_{n+1,n}^{(\alpha,\beta)} = a_{n+1,n}^{(\alpha,\beta)} - \frac{n+\alpha+\beta+1}{n+\alpha+\beta+2} a_{n,n-1}^{(\alpha+1,\beta+1)} \ ,$$

$$b_{n,n}^{(\alpha,\beta)} = a_{n,n}^{(\alpha,\beta)} - a_{n-1,n-1}^{(\alpha+1,\beta+1)} \ ,$$

$$b_{n-1,n}^{(\alpha,\beta)} = a_{n-1,n}^{(\alpha,\beta)} - \frac{n+\alpha+\beta+1}{n+\alpha+\beta} a_{n-2,n-1}^{(\alpha+1,\beta+1)} \ .$$

In Appendix A we show that similar relations for the derivatives resulting in a tridiagonal integration operator (as well as a tridiagonal monomial multiplication) hold for all hypergeometric and confluent hypergeometric functions as a result of the Gauss contiguity relations, and this can be used to give an alternative direct derivation of the recurrence coefficients $a_{i,j}$ and $b_{i,j}$.

The nonzero elements of the matrices A, B for the classical orthogonal polynomials are given in Table 1, together with other relevant quantities, using the standard notation [1]. The relations for the Gegenbauer polynomials $C_n^{(\nu)}$ can be constructed from those of the general Jacobis since

$$C_n^{(\nu)} = \frac{\Gamma(\alpha+1)\Gamma(2\alpha+n+1)}{\Gamma(\alpha+n+1)\Gamma(2\alpha+1)} P_n^{(\alpha,\beta)}$$

where $\alpha = \beta = \nu - 1/2$.

2.2 Reduction to banded form via integration preconditioning

Our main result can be stated as follows:

Theorem 2.1 *Consider the orthogonal polynomial family $\{Q_n\}_0^\infty$ and assume that the Q_k satisfy (4) for some matrix $B = (b_{ij})$. Then, the matrix representing the differential operator*

$$L = \sum_{k=0}^{n} p_k(x) D^k \tag{8}$$

with $deg(p_k) = \pi_k$, the degree of the polynomial coefficient of D^k, becomes banded upon left multiplication by $B_{[n]}^n$ with bandwidth $R = \max_k(2\pi_k + 2(n-k)+1)$ where $0 \le k \le n$.

(In the sequel we use L both for the operator and for its matrix representation. Also, arrays are indexed from 0 to N, the maximum order of truncation. An array G whose first k rows have been replaced by zeroes is denoted by $G_{[k]}$. Similarly $\hat{f}_{[i]}$ will denote the vector $\hat{f}$ with its first i components set to zero.)

Example: The following ordinary differential equation arises when one solves the 2-D Helmholtz equation on an annulus:

$$u(x) - \epsilon \left(D^2 + \frac{1}{x+a} D - \frac{k^2}{(x+a)^2} I \right) u(x) = -f$$

The inner radius of the annulus is $a-1 > 0$,k is an integer (representing the Fourier mode), and the range of x is $-1 \le x \le 1$. Multiplying through by the factor $(x+a)^2$ we arrive at an equation with polynomial coefficients which can be transformed to nine-diagonal form via left multiplication by the matrix $B_{[2]}^2$ [5]. This example is further discussed in the next section, where conditioning is considered.

Note: In conjunction with the Lanczos τ-method [10], or alternatively by making use of proper subspaces where differentiation is invertible [6], the above idea can be incorporated into the design of algorithms for the efficient and accurate solution of differential equations with polynomial coefficients. When the τ-method is used, the first n rows which after left multiplication by $B_{[n]}^n$ are null, are replaced by row vectors associated with the τ-constraints. These alter the matrix but do not affect the order of complexity of the solution algorithm, apart from an increase in the bandwidth which remains $\le R + n$.

We establish the above result in a series of lemmas. We define the function evaluation functional at the point x, q_x, as the row vector

$$q_x = (Q_0(x), Q_1(x), \cdots).$$

Also we introduce

$$q_x^{(k)} = \left(Q_0^{(k)}(x), Q_1^{(k)}(x), \cdots \right),$$

the operator of evaluating the k-th derivative. Clearly, the $\{Q_n^{(k)}\}_k^\infty$ form an independent set of polynomials which, for the classical orthogonal polynomials, can be shown to be orthogonal with a weight related simply to $w(x)$. If we write $\hat{f}^{(0)} \equiv \hat{f} = (\hat{f}_0, \hat{f}_1, \cdots)^T$ for the vector of the expansion coefficients of a function $f(x)$ in the given basis and $\hat{f}^{(k)} = (\hat{f}_0^{(k)}, \hat{f}_1^{(k)}, \cdots)^T$ for the vector of the expansion coefficients of the function $f^{(k)}(x)$ (denoting the k'th derivative of the funcion $f(x)$) we have that $f^{(k)}(x) = q_x^{(k)} \hat{f}^{(0)} = q_x \hat{f}^{(k)}$. Then the relations (2), (4) can be written as

$$x q_x = q_x A \quad \text{and} \quad q_x = q_x^{(1)} B = q_x^{(1)} B_{[1]} \tag{9}$$

where A, B are the recursion coefficient matrices for (2), (4) respectively.

The matrix B has the form

$$B = \begin{bmatrix} b_{0,0} & b_{0,1} & 0 & 0 & \cdots \\ b_{1,0} & b_{1,1} & b_{1,2} & 0 & \cdots \\ 0 & b_{2,1} & b_{2,2} & b_{2,3} & 0 \\ 0 & 0 & b_{3,2} & b_{3,3} & \ddots \\ 0 & 0 & 0 & \ddots & \ddots \end{bmatrix}. \tag{10}$$

The very first row of B will always be set to zero. Thus the elements $b_{0,0}$ and $b_{0,1}$ are irrelevant and will never be used. The basic recursion (4) remains unchanged regardless of how the first row of B is defined since $Q_0^{(1)}(x) = 0$ for polynomial families.

We have

Lemma 2.3 *The operator P_n of multiplication by a polynomial $p_n(x)$ is expressed in the basis Q_n by a banded matrix, of bandwidth $2\pi_n + 1$.*

Proof: Consider an expansion $f(x) = q_x \hat{f}^{(0)}$. Multiplying by x, one obtains $xf(x) = x q_x \hat{f}^{(0)}$ where $x q_x$ denotes

$xq_x = (xQ_0(x), xQ_1(x), \cdots)$. Invoking (9) one can substitute for xq_x to obtain $xf(x) = q_x A\hat{f}^{(0)}$. Thus A is the matrix which transforms the vector of expansion coefficients for $f(x)$ into the vector of expansion coefficients for $xf(x)$. The matrix A is tridiagonal since the recurrence relation (2) is a three-term recurrence relation. Extending this argument, it is evident that A^n is the matrix that transforms the vector of expansion coefficients for $f(x)$ into the vector of expansion coefficients for $x^n f(x)$. Since A was tridiagonal A^n is obviously banded with bandwidth $2n+1$. Thus multiplication by a polynomial $p_n(x)$ becomes the operation of multiplication by $p_n(A)$.

Note: If the family has a simple convolution, as is the case for the Chebyshev polynomials for which $2T_mT_n = T_{m+n} + T_{|m-n|}$, then it is convenient to expand the polynomial $p_n(x)$ in terms of the basis thus simplifying the construction of the matrix P_n. Specifically, one first expands $p_n(x) = \sum_{k=0}^{k=n} c_k T_k$. For purposes of implementation, let us consider a truncated Chebyshev expansion of order N, that is $f(x) = T_x^N \hat{f}^{(0)}$ where T_x^N is the row vector $T_x^N = (T_0(x), T_1(x), \cdots, T_N(x))$. The product is $p_n(x)f(x) = \sum_{k=0}^{k=n} c_k T_k T_x^N \hat{f}^{(0)}$ where $T_k T_x^N$ represents the vector $T_k T_x^N = (T_k(x)T_0(x), T_k(x)T_1(x), \cdots, T_k(x)T_N(x))$. The vector $T_k T_x^N$ can be expressed in the form $(T_k T_x^N) = T_x^N A_{T_k}$ (accurate up to the N+1'th coefficient) where A_{T_k} is the $N+1$ by $N+1$ matrix

$$A_{T_k} \equiv \frac{1}{2}\begin{bmatrix} 0 & \cdots & 0 & 1 & 0 & \cdots & \cdots & 0 \\ \vdots & 0 & 1 & 0 & \ddots & \ddots & & \vdots \\ 0 & 1 & 0 & & \ddots & \ddots & \ddots & \vdots \\ 2 & 0 & & & & \ddots & \ddots & 0 \\ 0 & 1 & \ddots & & & & \ddots & 1 \\ \vdots & \ddots & \ddots & \ddots & & & & 0 \\ \vdots & & \ddots & \ddots & \ddots & & & \vdots \\ 0 & \cdots & \cdots & 0 & 1 & 0 & \cdots & 0 \end{bmatrix}$$

where $a_{k0} = 1$, $a_{0k} = 1/2$ etc, which follows from $2T_mT_n = T_{m+n} + T_{|m-n|}$. Note A_{T_0} reduces to the identity matrix. As expected, A_{T_k} has a bandwidth of $2k+1$. Thus $p_n(x)f(x) = \sum_{k=0}^{k=n} c_k T_k T_x^N \hat{f}^{(0)}$ can be written as $p_n(x)f(x) = T_x^N \left(\sum_{k=0}^{k=n} c_k A_{T_k}\right) \hat{f}^{(0)}$ where $\sum_{k=0}^{k=n} c_k A_{T_k}$ is sum of matrices A_{T_k} each with bandwidth less or equal to $2n+1$. This equation illustrates that multiplication by $p_n(x)$ translates into multiplication by a banded matrix for Chebyshev polynomials.

We also have the related obvious consequence of Leibnitz's rule:

Lemma 2.4 *The commutator of the operator D^k with the operator P of multiplication by a polynomial $p(x)$ is given by*

$$\left[P, D^k\right] \equiv PD^k - D^k P = \sum_{m=1}^{k} (-1)^m \begin{pmatrix} k \\ m \end{pmatrix} D^{k-m} P^{(m)}.$$

The properties of the $B_{[k]}^k$ are established in lemmas (2.5-2.6):

Lemma 2.5 *Let $f'(x) = \sum_k \hat{f}_k^{(1)} Q_k(x)$; then $B_{[1]}\hat{f}^{(1)} = \hat{f}_{[1]}^{(0)}$ and the first element of $\hat{f}^{(0)}$, $\hat{f}_0^{(0)}$, is undetermined.*

Proof: We have that $f(x) = q_x \hat{f}^{(0)}$ and $f'(x) = q_x^{(1)}\hat{f}^{(0)} = q_x \hat{f}^{(1)}$. Also, by assumption, $q_x^{(1)} B = q_x$. Combining we find

$$q_x^{(1)} \left(\hat{f}^{(0)} - B\hat{f}^{(1)}\right) = 0.$$

Since the Q_k' are independent and orthogonal, the relation claimed follows. Clearly, the first element of $\hat{f}$ remains undetermined, since $Q_0' \equiv 0$.

Corollary 2.1 *A similar relationship holds for the n-th derivative coefficients of $f(x)$. Recall that $q_x^{(1)} B = q_x$ from (9). Similarly, one has $q_x^{(i+1)} B = q_x^{(i)}$. One can generate, using repeated applications of $q_x^{(i+1)} B = q_x^{(i)}$ and recursive substitutions,*

$$q_x^{(n)} B^p = q_x^{(n-p)} \tag{11}$$

where $p \le n$. Setting $n = k$ and $p = k$, in (11) one obtains $q_x^k B^k = q_x$. Arguing as in Lemma 2.5 it is seen that

$$f^{(k)}(x) = q_x^{(k)} \hat{f}^{(0)} = q_x \hat{f}^{(k)} = q_x^{(k)} B^k f^{(k)}$$

using $q_x^k B^k = q_x$ for the last equality. Using the second and last expression in the above equation, one obtains

$$q_x^{(k)} \left(B^k \hat{f}^{(k)} - \hat{f}^{(0)}\right) = 0 \rightarrow \hat{f}_{[k]}^{(0)} - B_{[k]}^k \hat{f}^{(k)} = 0$$

and, again, this relation gives the coefficients of $f(x)$ in terms of those of its k-th derivative, with the components $\hat{f}_0$ through $\hat{f}_{k-1}$ remaining undefined. Recall that the first k components of $q_x^{(k)}$ are zero, since the k'th derivative annihilates the polynomials of order first through $k-1$.

We then have

Lemma 2.6 *If D is the matrix of differentiation in the family, then $B_{[1]}D = I_{[1]}$, the identity with its first row set to zero.*

Proof: Indeed, by definition, $D\hat{f}^{(0)} = \hat{f}^{(1)}$, so that $q_x \hat{f}^{(1)} = q_x D\hat{f}^{(0)}$ from which $q_x^{(1)} B_{[1]}(\hat{f}^{(1)} - D\hat{f}^{(0)}) = 0$ using (9). Thus $q_x^{(1)}(B_{[1]}\hat{f}^{(1)} - B_{[1]}D\hat{f}^{(0)}) = 0$ or $q_x^{(1)}(\hat{f}_{[1]}^{(0)} - B_{[1]}D\hat{f}^{(0)}) = 0$ by Lemma 2.5. One can now write $(I_{[1]} - B_{[1]}D)\hat{f}^{(0)} = 0$ thus concluding the proof.

It then follows that:

Corollary 2.2 *If D^n is the matrix representing n-fold differentiation in the family Q_n, then $B^{l+n}_{[l+n]} D^n = B^l_{[l+n]}$.*

Proof: By the definition of the differentiation matrix D.

$$q_x^{(k)} \hat{f}^{(0)} = q_x D^k \hat{f}^{(0)} \tag{12}$$

Setting $p = n$ in (11), one has $q_x^{(n)} B^n = q_x$. Substituting this expression for q_x into (12), one obtains

$$q_x^{(k)} \hat{f}^{(0)} = q_x^{(n)} B^n D^k \hat{f}^{(0)}. \tag{13}$$

Setting $p = n - k \geq 0$ in (11), one has $q_x^{(n)} B^{n-k} = q_x^{(k)}$. Substituting this expression for $q_x^{(k)}$ in (13), one obtains

$$q_x^{(n)} B^{n-k} \hat{f}^{(0)} = q_x^{(n)} B^n D^k \hat{f}^{(0)}. \tag{14}$$

Now the first n components of $q_x^{(n)}$ are zero. However, the remaining components of $q_x^{(n)}$ are composed of orthogonal polynomials and are therefore independent. Thus, one has $B^{n-k} \hat{f}^{(0)} = B^n D^k \hat{f}^{(0)}$ except for the first n components. Since $f^{(0)}$ is an arbitrary vector, one has $B^{n-k} = B^n D^k$ except for the first n rows. This equation can be stated concisely in the form $B^{n-k}_{[n]} = B^n_{[n]} D^k$ or letting $n = n + l$ and $k = n$, in the form $B^{l+n}_{[l+n]} D^n = B^l_{[l+n]}$.

We now give the proof of our main assertion:

Proof of theorem: Following lemma (2.4) we rewrite the differential operator L as

$$\begin{aligned} L &= \sum_{k=0}^{n} \sum_{m=0}^{k} (-1)^m \begin{pmatrix} k \\ m \end{pmatrix} D^{k-m} P_k^{(m)} \\ &= \sum_{r=0}^{n} D^r \sum_{k=r}^{n} (-1)^{k-r} \begin{pmatrix} k \\ r \end{pmatrix} P_k^{(k-r)} \\ &= \sum_{r=0}^{n} D^r S_r \end{aligned}$$

with $\deg(S_r) = \sigma_r = \max_{r \leq k \leq n}(\pi_k - k + r)$. Then, since by cor.(2.2) $B^n_{[n]} D^r = B^{n-r}_{[n]}$,

$$B^n_{[n]} L = \sum_{r=0}^{n} B^{n-r}_{[n]} S_r$$

and the bandwidth is obviously

$$2 \max_{0 \leq r \leq n} ((n - r) + \sigma_r) + 1 = R \ .$$

Replacing the first n-rows (containing zeroes) by appropriate constraint coefficients, originating from the boundary, initial or other conditions imposed on the solution of the ODE $Lu = f$, will transform the matrix into a banded form with n additional (generally nonzero) rows. This matrix still factors similarly to a banded matrix, with bandwidth $R + n$.

3 Conditioning and Convergence

It is well known that spectral differentiation operators suffer the dual defects of full upper triangular matrix representations with very poor conditioning. Above we have shown how integration preconditioners band spectral differentiation matrices. In this section we show that they produce well-conditioned linear systems, and exploit that fact to give good error estimates for general ordinary differential equations. In the following section we will show that the favorable conditioning properties lead to the rapid convergence of iterative methods for spectral approximations to general variable coefficient and nonlinear equations. In particular we study the preconditioned, discrete system:

$$\mathcal{T}_N \hat{v} = b, \tag{15}$$

$$\left(I_{[n]} + \sum_{j=0}^{n-1} B^{n-j}_{[n]} S_j \right) \hat{v} = B^n_{[n]} \bar{F}_N, \tag{16}$$

under the simplifying assumption that the lead coefficient is 1. Here $\mathcal{T}_N \hat{v} = \mathcal{T} v$ and the matrices S_j are Galerkin approximations to multiplication by the polynomials, $s_j(x)$. Throughout we assume that the polynomial family is one of the symmetric Jacobi (Gegenbauer) families scaled so that:

$$\frac{\sup_k \|Q_k\|_\omega}{\inf_k \|Q_k\|_\omega} = \kappa_Q < \infty. \tag{17}$$

We denote by A_N the coefficient matrix of this system and assume that $\bar{F}_N$ is computed by interpolation of f at the Gauss or Gauss-Lobatto points associated with the truncated orthogonal system.

In [6] a post-conditioning scheme based on the integration operators is analyzed. The main difference here is the inclusion of the $\mathcal{T}$-conditions in the matrix. As the $\mathcal{T}$-conditions typically involve point evaluations of functions and their derivatives, we cannot expect $\mathcal{T}_N$ to be bounded in l_2. Therefore, we introduce the space h_r of infinite vectors satisfying:

$$\|Z\|^2_{h_r} = \sum_{l=0}^{\infty} |Z_l|^2 (l+1)^{2r} < \infty. \tag{18}$$

For finite truncations, the norm is simply defined by the truncated sum. It is associated with the inner product $(Y, Z)_{h_r} = Y^T D_r^2 Z$, where D_r is the diagonal matrix whose jth diagonal entry is j^r. Given any matrix C we have:

$$\|C\|_{h_r} = \|D_r C D_r^{-1}\|_2. \tag{19}$$

Note that for integer $r \geq 0$, and Z the expansion coefficients of a function, z, we have, for positive constants G_0,

G_1, $\bar{G}_1$:

$$G_0\langle z,(\mathcal{L}+1)^r z\rangle \leq \|Z\|_{h_r}^2 \leq G_1\langle z,(\mathcal{L}+1)^r z\rangle \leq \bar{G}_1\|z\|_{H_{r,\omega}}^2, \tag{20}$$

where $\mathcal{L} = -DpD$. (Here, $\langle\cdot,\cdot\rangle$ is the unweighted L^2 inner product.) We now state our assumption on the $\mathcal{T}$-conditions:

Assumption 3.1 *There exists $r_0 \geq 0$ and a constant, $C_{\mathcal{T}}$, such that, for all N, $r \geq r_0$, and $N+1$-vectors V,*

$$|\mathcal{T}_N V| \leq C_{\mathcal{T}}\|V\|_{h_r}.$$

Moreover, for some integer $r \geq r_0$, $r - n \leq 1$.

We illustrate Assumption 3.1 with the standard examples of Chebyshev approximations and Dirichlet or Neumann boundary conditions. In the former case,

$$\mathcal{T}_N V = \sum_l (-1)^{\eta_l} V_l, \tag{21}$$

where $\eta_l = l$ or $\eta_l = 0$, depending on the boundary. Choosing $r_0 > 1/2$ we have, for all $r \geq r_0$:

$$|\mathcal{T}_N V| \leq \left(\sum_{l=0}^{\infty}(l+1)^{-2r_0}\right)^{1/2}\|V\|_{h_r}. \tag{22}$$

For Neumann conditions,

$$\mathcal{T}_N V = \sum_l (-1)^{\eta_l} l^2 V_l. \tag{23}$$

Choosing $r_0 > 5/2$ we have, for all $r \geq r_0$:

$$|\mathcal{T}_N V| \leq \left(\sum_{l=0}^{\infty}(l+1)^{4-2r_0}\right)^{1/2}\|V\|_{h_r}. \tag{24}$$

We now state a number of results concerning the individual operators in (16). In many cases the proofs can be found in [6, Sec. 4] or constructed in obvious analogy with proofs given there, We will avoid repeating the details of the arguments in [6]. The primary additional facts we will use are:

Lemma 3.1 *For any $r > 0$ and integer k there exists a constant $K(k,r)$ independent of N such that, for any $N \times N$ matrix, C, with bandwidth k,*

$$\|C\|_{h_r} \leq K(k,r)\|C\|_2.$$

Proof: We need only consider the matrix $D_r C D_r^{-1}$. Its nonzero elements are multiplied by $(i+1)^r(j+1)^{-r}$ with $i - j \leq k$. Clearly this factor is uniformly bounded above by a function of k and r, completing the proof.

Lemma 3.2 *For any $r \geq 0$, $s > r$, $Z \in h_s$ and Z_N the infinite vector obtained by setting all but the first $N+1$ components of Z to 0, we have:*

$$\|Z - Z_N\|_{h_r} \leq (N+2)^{r-s}\|Z\|_{h_s}.$$

Proof: We have:

$$\|Z - Z_N\|_{h_r}^2 = \sum_{k=N+1}^{\infty}(k+1)^{2r}|Z_k|^2 \leq (N+2)^{2(r-s)}\sum_{k=N+1}^{\infty}(k+1)^{2s}|Z_k|^2, \tag{25}$$

which implies the stated estimate.

We note that by (20) we can replace the right-hand side by a multiple of the $H_{s,\omega}$ norm of z. This inequality is, then, stronger than can be obtained for the $H_{r,\omega}$ norm of $z = z_N$. (See [3, Ch. 9].) However, we must generally use a larger r than Sobolev's inequality would require to bound $\mathcal{T}$. The example of Neumann conditions illustrates this.

Theorem 3.1

a. *For any $r > 0$ the operators $B_{[n]}^l : h_r \to h_r$, $l = 1,\ldots,n$, are compact and, for some constants, $g_{l,r}$,*

$$\|B_{[n]}^l U\|_{h_r} \leq g_{l,r}\|U\|_{h_{r-l}}.$$

b. *For any $r > 0$, $\|S_j\|_{h_r}$, $j = 0,\ldots,n-1$ are uniformly bounded in N.*

Proof: The matrices $B_{[n]}^l$ and S_j are banded. Bounds on their l_2 norms are given in [6] and, by Lemma 3.1, extend to their h_r norms. As the compactness is proved by approximating $B_{[n]}^l$ by its finite truncations, it can also be extended. We also have an estimate for the entries of $B_{[n]}^l$:

$$|(B_{[n]}^l)_{km}| \leq \alpha k^{-l}. \tag{26}$$

Therefore,

$$\|B_{[n]}^l U\|_{h_r}^2 = \sum_{k=n}^{\infty}(k+1)^{2r}|(B_{[n]}^l U)_k|^2 \leq \bar{g}_{l,r}\sum_{k=0}^{\infty}(k+1)^{2(r-l)}|U_k|^2, \tag{27}$$

from which the desired estimate follows.

We are now in a position to prove:

Theorem 3.2 *Suppose that $w = 0$ is the only solution of the homogeneous system, $Lw = 0$, $\mathcal{T}w = 0$ and that $\mathcal{T}$ satisfies Assumption 3.1. Let r be an integer satisfying $r \geq r_0$ and $r - n \leq 1$. Further suppose that, for some $p \geq 1$, $f \in C^p((a,b))$. Then:*

a. *There exist constants, C_0 and C_1, and an integer, N_0 such that for any $N > N_0$ and vector, y, with $||y||_r = 1$:*

$$C_0 \leq ||A_N y||_r \leq C_1.$$

b. *The matrix $A_N - I$ approaches a compact operator on h_r.*

c. *There exists a constant, C, and an integer N_0, such that, for all $N > N_0$:*

$$\langle (u - v_N), (\mathcal{L}+1)^r (u - v_N) \rangle \leq C N^{-2\mu} ||f||^2_{\omega,p},$$

$$\mu = p - (1/2)\max(0, r-n).$$

Proof: The upper bound on A_N and the compactness of $A_N - I$ follow directly from Theorem 3.1. The lower bound follows from the analysis in [6, Sec. 4], which we outline here for completeness. First, define the compact operator, $\tilde{K} : h_r \to h_r$, in the following way. Given $Z \in h_r$ let $z = \sum_k Z_k Q_k$, $s_j z = \sum_k \bar{Z}^{(j)}_k Q_k$. Then

$$(28)\ (\tilde{K}Z)_i = \begin{cases} (\mathcal{T}z)_i - Z_i, & i = 0, \ldots, n-1 \\ (\sum_{j=0}^{n-1} B^{n-j}_{[n]} \bar{Z}^{(j)})_i, & i \geq n. \end{cases}$$

If, for some $Z \neq 0$, $(I + \tilde{K})Z = 0$, it can be easily shown that a nontrivial solution of the homogeneous problem exists, violating the hypotheses of the theorem. By the Riesz-Schauder theory, $||(I+\tilde{K})^{-1}||_{h_r}$ is bounded. We next show that $A_N - I$ approximates $\tilde{K}$. In particular, let

$$(29)\quad (\bar{K}_N Z)_i = \begin{cases} ((A_N - I)Z_N)_i, & i = 0, \ldots, N, \\ 0, & i > N. \end{cases}$$

Here, Z_N is the $N+1$-vector containing the first $N+1$ components of Z. Let $\epsilon > 0$ be given. Given any vector, Z, $||Z||_{h_r} = 1$, and positive integer M, let Z_M now denote the infinite vector obtained by setting all but the first $M+1$ components of Z to zero. Now, for $M = M(\epsilon)$ sufficiently large, we have, for all N,

$$(30)\ \ ||\tilde{K}(Z - Z_M)||_{h_r} < \frac{\epsilon}{2},\ \ ||\bar{K}_N(Z - Z_M)||_{h_r} < \frac{\epsilon}{2}.$$

Moreover, if $N > M + n + q$, where q is the maximum degree of the polynomials, s_j, $\tilde{K}Z_M = \bar{K}_N Z_M$. Therefore, for $N > M(\epsilon) + n + q$,

$$(31) \qquad ||\tilde{K} - \bar{K}_N||_{h_r} < \epsilon.$$

Choosing ϵ sufficiently small, and, hence, N_0 sufficiently large, we conclude using the Banach lemma that $(A_N)^{-1}$ is uniformly bounded for $N > N_0$.

Standard ode theory implies the existence of a solution, $u \in C^{p+n}((a,b))$. Set $e = u - v_N$, $E_N = U_N - V_N$, $\Delta F = F_N - \bar{F}_N$ and u_N the polynomial whose expansion coefficients are given by U_N. We then have:

$$(32) \qquad A_N E_N = R_N,$$

$$(33) \qquad R_N = \begin{pmatrix} \mathcal{T}_N U_N - Tu \\ B^n_{[n],N}(\Delta F) + W_N \end{pmatrix},$$

$$(34)\quad W_N = -\sum_{j=0}^{n-1} (B^{n-j}_{[n],N} S_{j,N} U_N - (B^{n-j}_{[n]} S_j U)_N).$$

From our bounds on A_N and A_N^{-1} and Assumption 3.1 we conclude,

$$(35) \qquad ||E_N||_{h_r} \leq \frac{1}{C_0} ||R_N||_{h_r},$$

$$(36)\ ||R_N||_{h_r} \leq \bar{C} \left(||B^n_{[n],N}(\Delta F)||_{h_r} + ||U - U_N||_{h_r} \right),$$

where in $U - U_N$, U_N denotes the infinite vector obtained by extending the finite vector by 0. By Lemma 3.2 and (20) we have:

$$(37) \qquad ||U - U_N||_{h_r} \leq G_2 N^{r-p-n} ||f||_{H_{p,\omega}}.$$

Using (20) and the properties of the integration operators we obtain:

$$(38) \qquad ||B^n_{[n],N}(\Delta F)||_{h_r} \leq G_4 ||f_N - \bar{f}||_{H_{l,\omega}},$$

$$(39) \qquad l = \max(0, r-n) \leq 1.$$

By the results of Bernardi and Maday on interpolation error, [2], we have:

$$(40) \qquad ||f_N - \bar{f}||_{H_{l,\omega}} \leq N^{(l/2)-p} ||f||_{H_{p,\omega}}.$$

Therefore, since

$$(41) \qquad ||U - V_N||_{h_r} \leq ||E_N||_{h_r} + ||U - U_N||_{h_r},$$

combining these inequalities and applying (20) yields the error estimate. This completes the proof.

We note that these results fall short of those proved in [6] for the post-conditioning scheme, both in terms of the restrictive assumptions on the coefficients and in the convergence rates. We hope to improve these in future work. Numerical experiments show that the results on conditioning hold for a number of important operators with variable lead coefficients. In the following tables we display the condition numbers in the norms h_0, h_1 and h_2, of truncated approximations to the Dirichlet problem for the cylindrical Poisson, cylindrical Helmholtz and helical Poisson operators:

Truncation	Poisson	Helmholtz	Helical
8	12.6	1264	229
16	20.6	3788	300
32	31.9	9550	399
64	51.2	19580	537
128	90.0	39320	737
256	156.	78700	1020

Table 1: h_0 Condition Numbers: Preconditioned

Truncation	Poisson	Helmholtz	Helical
8	11.4	1321	1110
16	11.7	2777	1108
32	11.9	3844	1108
64	12.0	4084	1108
128	12.1	4133	1108
256	12.1	4154	1108

Table 2: h_1 Condition Numbers: Preconditioned

Poisson Operator:

$$r^2\frac{\partial^2}{\partial r^2}+r\frac{\partial}{\partial r}-k^2,$$

Helmholtz Operator:

$$r^2-\epsilon\big(r^2\frac{\partial^2}{\partial r^2}+r\frac{\partial}{\partial r}-k^2\big),$$

Helical Operator:

$$(\alpha^2r^4+r^2)\frac{\partial^2}{\partial r^2}+(-\alpha^2r^3+r)\frac{\partial}{\partial r}-k^2(1+2\alpha^2r^2+\alpha^4r^4).$$

Here, $k=3$, $\epsilon=.001$, $\alpha=1.5$, and $1\le r\le 3$.

Truncation	Poisson	Helmholtz	Helical
8	47.9	3,568	10,870
16	47.9	7,678	10,700
32	47.9	10,070	10,700
64	47.9	10,570	10,700
128	47.9	10,610	10,700
256	47.9	10,620	10,700

Table 3: h_2 Condition Numbers: Preconditioned

Truncation	Poisson	Helmholtz	Helical
8	3.774×10^3	9.584	5.886×10^4
16	5.533×10^4	59.98	8.948×10^5
32	8.427×10^5	922.4	1.374×10^7
64	1.314×10^7	14,400	2.148×10^8
128	2.075×10^8	2.273×10^5	3.393×10^9
256	3.298×10^9	3.613×10^6	5.393×10^{10}

Table 4: h_0 Condition Numbers: Unpreconditioned

Truncation	Poisson	Helmholtz	Helical
8	4.293×10^3	80.81	9.098×10^4
16	4.978×10^4	66.20	1.102×10^6
32	6.480×10^5	867.4	1.450×10^7
64	9.225×10^6	12,400	2.070×10^8
128	1.387×10^8	1.865×10^5	3.114×10^9
256	2.150×10^9	2.891×10^6	4.827×10^{10}

Table 5: h_1 Condition Numbers: Unpreconditioned

Clearly, the condition numbers grow with N in the h_0 norm but remain bounded in h_1 and h_2. This is consistent with the analysis above. Finally, we display the growth of the condition numbers in these norms for the unpreconditioned systems, demonstrating dramatic effects of the preconditioning.

4 General Variable Coefficients

A well-known difficulty with spectral methods is that multiplication by arbitrary functions is represented by full matrices. Therefore, direct solution of the linear systems following from the Galerkin approximation to general differential equations may be expensive. Similar considerations apply to the solution of nonlinear equations. On the other

Truncation	Poisson	Helmholtz	Helical
8	6.409×10^3	2,269	2.335×10^5
16	6.469×10^4	906.4	2.362×10^6
32	7.802×10^5	3,887	2.885×10^7
64	1.068×10^7	53,750	3.961×10^8
128	1.575×10^8	7.931×10^5	5.844×10^9
256	2.417×10^9	1.217×10^7	8.968×10^{10}

Table 6: h_2 Condition Numbers: Unpreconditioned

hand, multiplication in point space may be accomplished in $O(\bar{N})$ operations, where $\bar{N}$ is the number of points where the product is required. This fact is exploited by pseudospectral methods. In this section we show how to combine the integration preconditioners with pseudospectral approximations to multiplication by smooth functions to iteratively compute approximate solutions to variable coefficient and nonlinear equations. For families with a fast interpolation algorithm, such as the Chebyshev family, the complexity of the algorithm will be $O(N \ln N)$, where N is spectral truncation order.

We consider:

$$Lu \equiv \left(D^n + \sum_{j=0}^{n-1} D^j c_j(x) \right) u = f, \quad x \in (a,b), \tag{42}$$

subject to the constraints,

$$\mathcal{T} u = d. \tag{43}$$

Here, the functions c_j are assumed to be smooth (C^∞ for convenience). As before, we approximate u by a finite expansion,

$$u \approx v = \sum_{i=0}^{N} \hat{v}_i Q_i(x). \tag{44}$$

Multiplication by $c_j(x)$ is approximated, in spectral space, by the following recipe: first, evaluate the expansion at some interpolation points, x_k, $k = 0, \ldots, \bar{N}$. Second, multiply at the interpolation points by $c_j(x_k)$. Finally, use the new data at the interpolation points to construct expansion coefficients. We denote by $C_{j,N}$ the matrix representing this process. Note that $C_{j,N}$ is usually never formed. For our examples, its action is computed by fast transforms in $O(\bar{N} \ln \bar{N})$ operations. We may choose $\bar{N} > N$ to avoid aliasing errors, but always $\bar{N} \leq \gamma N$ for some fixed γ as $N \to \infty$. Interpolation points will be at Gauss or Gauss-Lobatto points associated with the family. We assume the following result on the uniform boundedness of the matrices, $C_{j,N}$:

Assumption 4.1 *There exists a constant, G, and an integer $r > r_0$, $r - n \leq 1$, such that:*

$$\sup_{j,N} \|C_{j,N}\|_r \leq G.$$

We expect that this assumption can be proven under appropriate assumptions on the functions c_j and the choice of nodes.

Our final specification of the discrete system is:

$$\mathcal{T}_N \hat{v} = d, \tag{45}$$

$$\left(I_{[n]} + \sum_{j=0}^{n-1} B_{[n]}^{n-j} C_{j,N} \right) \hat{v} = B_{[n]}^n \hat{F}_N. \tag{46}$$

Let A_N denote the coefficient matrix of the system above. Note that its first n rows contain approximations to the constraints and its final $N + 1 - n$ rows contain the approximation to the differential equation. Using Lemma 4.1, we can prove the following result on conditioning and convergence. As the proof is essentially identical to the proof of Theorem 3.2, we omit it here.

Theorem 4.1 *Suppose that $w = 0$ is the only solution of the homogeneous system, $Lw = 0$, $\mathcal{T} w = 0$ and that $\mathcal{T}$ satisfies Assumption 3.1. Let r be an integer satisfying $r \geq r_0$ and $r - n \leq 1$ and suppose that Assumption 4.1 holds for this choice of r. Further suppose that, for some $p \geq 1$, $f \in C^p((a,b))$. Then:*

a. *There exist constants, C_0 and C_1, and an integer, N_0 such that for any $N > N_0$ and vector, y, with $\|y\|_r = 1$:*

$$C_0 \leq \|A_N y\|_r \leq C_1.$$

b. *The matrix $A_N - I$ approaches a compact operator on h_r.*

c. *There exists a constant, C, and an integer N_0, such that, for all $N > N_0$:*

$$\langle (u - v_N), (\mathcal{L} + 1)^r (u - v_N) \rangle C \leq N^{-2\mu} \|f\|_{\omega,p}^2,$$

$$\mu = p - (1/2) \max(0, r - n).$$

4.1 Solution by Iteration

Although Theorem 4.1 establishes the good conditioning of the discretization matrices and the rapid convergence of the approximations for smooth f, the matrix A_N is full so its factorization requires $O(N^3)$ operations. However, if a fast transform is available, multiplication by A_N can be carried out in $O(N \ln N)$ operations. In this section we exploit this feature along with the conclusions of Theorem 4.1 to develop an efficient iterative solution algorithm.

Here we consider Broyden's method, due to its ease of implementation for both linear and nonlinear problems and to the availability of convergence results which are directly applicable to our problem [15]. For completeness we list the algorithm as we use it, which involves only storage of and computation with a small number of vectors of dimension $N + 1$ [16]:

Broyden's Method for the Linear System, $A_N \hat{v} = \hat{Y}_N$:

1. Initialize: $\hat{v}_1 = s_1 = r_0 = \hat{Y}_N$, $\gamma_1 = \langle s_1, s_1 \rangle$.

2. Until $\sqrt{\langle r_k, r_k\rangle} < \epsilon$ do:

$$r_k = \hat{Y}_N - A_N \hat{v}_k, \quad z_{k+1}^{(1)} = r_k,$$

For $j = 2, \ldots, k$ do:

$$z_{k+1}^{(j)} = (I + \gamma_{j-1}^{-1} s_j \otimes s_{j-1}) z_{k+1}^{(j-1)},$$

$$\nu_{k+1} = \langle s_k, z_{k+1}^{(k)} \rangle.$$

Choose $\theta_{k+1} \in (0,2)$ such that $\gamma_k - \theta_{k+1}\nu_{k+1} \neq 0$,

$$s_{k+1} = \left(\frac{\gamma_k}{\gamma_k - \theta_{k+1}\nu_{k+1}} \right) z_{k+1}^{(k)},$$

$$\gamma_{k+1} = \theta_{k+1}^{-1} \langle s_{k+1}, s_{k+1} \rangle,$$

$$\hat{v}_{k+1} = \hat{v}_k + s_{k+1}.$$

Here, $\langle \cdot, \cdot \rangle$ denotes some inner product and $\otimes$ is the outer product of vectors defined by the inner product. We choose $\theta_{k+1} = 1$ unless $\gamma_k - \nu_{k+1}$ is small. Note that if p iterations are needed the total work is $O(pN \ln N + p^2 N)$.

Hwang and Kelley [15] have shown that if an operator, A, is such that $A - I$ is compact, then Broyden's method as described above produces a q-superlinearly convergent sequence of iterates. By part b of Theorem 4.1, this applies (uniformly in N) to our operators A_N if we use the h_r inner product with $r \geq r_0$. Therefore, for any $\epsilon > 0$, the number of iterations required to produce a residual with h_r norm smaller than ϵ is bounded independent of N. Hence, the system can be solved in $O(N \ln N)$ operations.

To illustrate this result we use Chebyshev expansions to solve:

$$D^2 u + \sin x \cdot u = f(x), \quad x \in (-1, 1). \tag{47}$$

The function f and the Dirichlet boundary conditions are chosen so that,

$$u = \frac{1}{\sqrt{\delta}} e^{-(x-x_0)^2/\delta}, \tag{48}$$

is an exact solution. No dealiasing was used. The results tabulated are for $\delta = 5 \times 10^{-4}$ and $x_0 = 1/2$ and, for $N \geq 128$, $\epsilon = 10^{-14}$. (The solution for $N = 64$ was so large that an absolute residual of 10^{-14} was unattainable. In that case only we use $\epsilon = 5 \times 10^{-14}$.)

Clearly, the number of iterations is independent of N while the error rapidly decreases. We note that the results presented are for the l_2 inner product. The convergence does not follow directly from the theory discussed above, because the $\mathcal{T}$-conditions are unbounded in this norm. However, due to their low rank, this did not harm the convergence. Tests with the h_1 inner product show similar behavior. We expect that the properties of A_N will lead to rapid convergence of other iterative schemes. For GMRES, this follows from [17].

N	No. of Its.	Max. Error
64	33	1.1×10^{3}
128	23	1.2
256	23	2.5×10^{-5}
512	23	1.0×10^{-11}

Table 7: Linear Test Problem

4.2 Nonlinear Problems

The method may also be generalized to solve semilinear equations. In particular we consider:

$$D^n u + F(u, Du, \ldots, D^{n-1}u, x) = 0, \quad x \in (a, b), \tag{49}$$

with nonlinear constraints:

$$\mathcal{T}(u) = 0. \tag{50}$$

The discrete equations are formulated in spectral space by approximating F via the same recipe as above. That is, $\hat{F}(\hat{v})$ is computed by first evaluating v and its derivatives in point space, then evaluating F at these points, and finally interpolating the point values to obtain $\hat{F}$. Similarly, $\mathcal{T}_N$ is computed by evaluating $\mathcal{T}(v)$. The discrete system after preconditioning is given by:

$$\mathcal{T}_N(\hat{v}) = 0, \tag{51}$$

$$I_{[n]}\hat{v} + B_{[n]}^n \hat{F}(\hat{v}) = 0. \tag{52}$$

Broyden's method may be applied to the nonlinear discrete problem by simply defining the residuals, r_k, using equations (51-52). Generally, a good initial approximation, $\hat{v}_0$, is needed. Linearizing about a smooth solution, the discrete system has the same properties as for the linear variable coefficient equations discussed above. Therefore, the results of Hwang and Kelley [15] imply local, q-superlinear convergence of the iterates with the number of iterations required to attain a given tolerance bounded independent of N. That is, the nonlinear discrete problem can be solved in $O(N \ln N)$ operations for a sufficiently good initial approximation and assuming that I is a sufficiently good approximation to the Jacobian. (Of course, different initial approximations to the Jacobian, which are low rank perturbations to I, could also be used.)

To illustrate these results we solve the well-known reaction-diffusion equation:

$$D^2 u + \lambda e^u = 0, \quad x \in (-1, 1), \quad u(\pm 1) = 0. \tag{53}$$

For $\lambda < \lambda_c$ two solutions exist. Here, $.87 < \lambda_c < .88$. In our example we chose $\lambda = .87$ and an initial guess of

$\hat{v} = 0$. The exact solution which we are approximating is given by:

$$u = \ln\left(A\left(1 - \left(\tanh\sqrt{\frac{\lambda A}{2}}x\right)^2\right)\right), \tag{54}$$

$$A = 2.801710482773216533343\ldots. \tag{55}$$

We solved the problem for $N = 16, 32, 64, 128, 256, 512$ with $\epsilon = 10^{-14}$. Again, no dealiasing was employed. In all cases the iterates converged in 67 to 70 iterations, confirming the N-independence of the iterative scheme. As the solution is smooth, the error was already 1.8×10^{-9} for $N = 16$ and on the order of 10^{-14} for finer discretizations.

5 Higher dimensional problems

The complexity of the spectral differentiation operator has made the direct use of spectral methods in more than one dimension impractical. Standard treatments of commonly occuring problems, such as the Poisson [12] and Helmholtz [13] problems have been approached through diagonalization methods. These techniques perform well but have the disadvantage that an expensive matrix multiplication must be performed to transform from eigenvectors of the operators back to physical variables which is necessary, e.g. for the solution of nonlinear problems. Also, the treatment of time dependent problems is typically pursued through a finite-difference discretization in time, which intoduces stability problems and limits the time-accuracy of the method. As an exeption to the latter we must mention the work of Tal-Ezer et al. [18] who employ a Chebyshev discretization of the time-evolution operator. As discussed in the monograph by Canuto et al. [3], the extension to multidimensions is open for several interesting problems in more than two space dimensions.

The simplicity and generality of the integration preconditioner method can be exploited to produce block-banded forms and improve the conditioning of problems in higher dimensions treated by spectral τ-methods. The straightforward extension is based on the use of preconditioners constructed by tensor products of integration operators in each variable, and it allows for the use of different basis functions in each variable. We consider a problem in a rectangle in $\mathbf{R}^m$. For simplicity we will only consider boundary conditions of Dirichlet type. We let $\mathbf{x} = (x_1, \ldots, x_m)$ be a coordinate system such that the sides of the domain under consideration are parallel to coordinate planes. We will consider expansions of the form

$$f(\mathbf{x}) = \sum_{\mathbf{i}} f_{\mathbf{i}} \mathcal{Q}_{\mathbf{i}},$$

where the expansion basis is formed as a product of (possibly different) m orthogonal polynomial bases and $\mathbf{i} = (i_1, \ldots, {}_{1m})$ is a multi-index,

$$\mathcal{Q}_{\mathbf{i}}(\mathbf{x}) \equiv Q_{i_1,1}(x_1)\cdots Q_{i_m,m}(x_m)\ .$$

We now let $\mathcal{L} = \sum_k \mathcal{L}_k$ be a linear differential operator in the x_i , $i = 1, \ldots, m$ with polynomial coefficients in the independent variables. As before, rational function coefficients can also be allowed, provided no singularities are present in the domain and we reduce to polynomials by multiplying by the least common multiple of the denominators.

The $\mathcal{L}_k$ have the form

$$\mathcal{L}_k = \bigotimes_{i=1}^{m} L_{ki},$$

with the L_{ki} a linear differential operator with polynomial coefficients in the variable x_i, i.e. an operator of the form assumed in Theorem (2.1). Let

$$n_i = \max_k \text{order}(L_{ki})\ .$$

Then the extension of theorem (2.1) to multidimensions can be stated as follows:

Theorem 5.1 *The Galerkin representation of the differential operator $\mathcal{L}$ in the basis $\mathcal{Q}$ is transformed into block-banded form via left multiplication by the operator*

$$\mathcal{B} = \bigotimes_{i=1}^{m} B^{n_i}_{[n_i],i},$$

where $B_{[1],i}$ is the operator of integration for the family $\{Q_{ki}\}_{k=0}^{\infty}$. The resullting operator has the form

$$\mathcal{B}\mathcal{L} = \sum_k \bigotimes_{i=1}^{m} B^{n_i}_{[n_i],i} L_{ki}\ .$$

The proof follows from repeated application of theorem (2.1).

We illustrate the use of theorem (5.1) by some simple examples. We limit the discussion to two space dimensions or one space dimension plus time, as we are basically interested in demonstrating the tensor product technique. Questions of conditioning and efficient implementation by the use of sparse matrix solvers will be pursued elsewhere.

In the following examples we use exclusively Chebyshev polynomial expansions, again for simplicity of exposition. Other bases could have been employed in principle, and the only added complication would have been the loss of the fast cosine transform. Thus, the preconditioners employed will be tensor products based on powers of the Chebyshev integration operator

$$B_{[1]} = \begin{bmatrix} 0 & 0 & 0 & 0 & & \dots & 0 \\ 1 & 0 & -\frac{1}{2} & 0 & & \dots & 0 \\ 0 & \frac{1}{4} & 0 & -\frac{1}{4} & 0 & \dots & 0 \\ \vdots & & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & & \frac{1}{2i} & 0 & -\frac{1}{2i} & \vdots \\ \vdots & & & & \ddots & \ddots & \vdots \\ 0 & & \dots & & & \frac{1}{2M} & 0 \end{bmatrix} .$$

Here $B_{[1]}$ is a $M \times M$ matrix.

Example 1: the uni-directional wave equation
We consider the problem

$$\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = 0 \ , \quad (x,t) \in [-1,1]^2 \ , \tag{56}$$

where $u = u(x,t)$ and the boundary conditions are given as

$$u(x,-1) = h(x) \ , \quad u(-1,t) = v(t) \ . \tag{57}$$

Here $\mathcal{L}_1 = L_{1,1}L_{1,2}$ and $\mathcal{L}_2 = L_{2,1}L_{2,2}$ assume the forms $L_{1,1} = I$, $L_{1,2} = D_t$, $L_{2,1} = D_x$, and $L_{2,2} = I$. The integrator for $\mathcal{L}$ is $B_{[1]} \bigotimes B_{[1]}$.

The matrix $\mathcal{BL}$ for the wave equation is

$$\begin{bmatrix} 0 & 0 & 0 & 0 & & \dots & 0 \\ I_{[1]} & B_{[1]} & \frac{-I_{[1]}}{2} & 0 & & \dots & 0 \\ 0 & \frac{I_{[1]}}{4} & B_{[1]} & \frac{-I_{[1]}}{4} & 0 & \dots & 0 \\ \vdots & & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & & \frac{I_{[1]}}{2i} & B_{[1]} & \frac{-I_{[1]}}{2i} & \vdots \\ \vdots & & & & \ddots & \ddots & \vdots \\ 0 & & \dots & & & \frac{I_{[1]}}{2M} & B_{[1]} \end{bmatrix} .$$

Note that the entries are $(N+1)\times(N+1)$ blocks. The same will apply to the matrix operators given in both subsequent examples.

The tau conditions are

$$\sum_{i=0}^{M} (-1)^i u_{ij} = v_j \ , \ j = 0, \dots, N,$$

$$\sum_{j=0}^{N} (-1)^j u_{ij} = h_i \ , \ i = 0, \dots, M,$$

Truncation	Abs error
8×8	1.2×10^{-2}
16×16	4.6×10^{-6}
32×32	1.0×10^{-14}

Table 8: Wave equation - Exact solution: $e^{(2(x-t))^2}$

with one redundant condition at the point $(-1,-1)$.

Table 5 lists the absolute error for the uni-direction wave equation for various truncations. In this, as in the two subsequent examples, a homogeneous solution is chosen, and the boundary conditions are constructed by evaluating that function at the appropriate boundaries.

Example 2: the Laplace equation in a rectangle
Consider now the problem

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \ , \quad (x,y) \in [-1,1]^2 \ , \tag{58}$$

where $u = u(x,y)$, $f = f(x,y)$ and the boundary conditions are given as

$$u(x,\pm 1) = h^{\pm}(x) \ , \quad u(\pm 1, y) = v^{\pm}(y) \ . \tag{59}$$

For Laplace's equation $\mathcal{L}_1 = L_{1,1}L_{1,2}$ and $\mathcal{L}_2 = L_{2,1}L_{2,2}$ become $L_{1,1} = I$, $L_{1,2} = D_x^2$, $L_{2,1} = D_y^2$, and $L_{2,2} = I$ and integrator is $B_{[2]} \bigotimes B_{[2]}$.

For Laplace's equation $\mathcal{BL}$ is the sum of the following two matrices

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ \frac{I_{[2]}}{4} & 0 & \frac{-I_{[2]}}{2(1\cdot 3)} & 0 & \frac{I_{[2]}}{4\cdot 6} & 0 & 0 \\ 0 & \frac{I_{[2]}}{4\cdot 6} & 0 & \frac{-I_{[2]}}{2(2\cdot 4)} & 0 & \frac{I_{[2]}}{6\cdot 8} & 0 \\ 0 & 0 & \frac{I_{[2]}}{(2i)(2i+2)} & 0 & \frac{-I_{[2]}}{2(i)(i+2)} & 0 & \frac{I_{[2]}}{(2i+2)(2i+4)} \\ 0 & 0 & 0 & \ddots & 0 & \ddots & 0 \end{bmatrix}$$

$$+ \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & B_{[2]}^2 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & B_{[2]}^2 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & B_{[2]}^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & B_{[2]}^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & B_{[2]}^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ddots \end{bmatrix} .$$

The tau conditions can now be inserted:

$$\sum_{n=0}^{n=N} u_{mn}(-1)^n = \hat{h}_m^-,$$

Truncation	Abs error (k = 2)	Abs error (k = 8)
8×8	3.2×10^{-5}	84.6
16×16	5.8×10^{-14}	8.5×10^{-2}
32×32	4.3×10^{-14}	1.8×10^{-11}

Table 9: Laplace's equation - Exact solution: $e^{ky}sin(kx)$

Truncation	$(c = 2,\ \nu = 1.3)$	$(c = 5,\ \nu = .5)$
8×8	5.6×10^{-3}	.12
16×16	3.6×10^{-5}	2.7×10^{-3}
32×32	1.5×10^{-9}	2.1×10^{-7}

Table 10: Advection-diffusion with $x_0 = -0.8$, $t_0 = -1.05$ Exact solution: $exp(-\frac{(x-x_0-c(t-t_0))^2}{4\nu(t-t_0)})/(t-t_0)^{\frac{1}{2}}$

$$\sum_{n=0}^{n=N} u_{mn} = \hat{h}_m^+,$$

$$\sum_{m=0}^{m=M} u_{mn}(-1)^m = \hat{v}_n^-,$$

$$\sum_{m=0}^{m=M} u_{mn} = \hat{v}_n^+.$$

Again, the number of tau conditions exceeds the number of zero rows in $\mathcal{BL}$. However, the tau conditions are not all independent, and four of them need to be discarded, corresponding to redundant specifications at the four corners of the domain. This leaves $2(M+1)+2(N+1)-4$ conditions which matches the number of zero rows in BL.

Table 5 lists the computed errors using this matrix formulation of Poisson's equation.

Example 3: the advection-diffusion equation
Finally we consider the problem

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = \nu\frac{\partial^2 u}{\partial x^2} \ , \quad (x,t) \in [-1,1]^2 \ , \tag{60}$$

where $u = u(x,y)$, $f = f(x,y)$ and the boundary conditions are given as

$$u(x,-1) = h(x) \ , \quad u(\pm 1, t) = v^{\pm}(t) \ . \tag{61}$$

For this equation $\mathcal{L}_1 = L_{1,1}L_{1,2}$ and $\mathcal{L}_2 = L_{2,1}L_{2,2}$ become $L_{1,1} = I$, $L_{1,2} = D_t$, $L_{2,1} = cD_x - \nu D_x^2$, and $L_{2,2} = I$ and integrator is $B_{[1]} \bigotimes B_{[2]}$.

The composite matrix $\mathcal{BL}$ is the sum of the following three matrices

$$\begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
0 & \frac{cB_{[1]}}{4} & 0 & \frac{-cB_{[1]}}{4} & 0 & 0 & 0 & \cdots \\
0 & 0 & \frac{cB_{[1]}}{6} & 0 & \frac{-cB_{[1]}}{6} & 0 & 0 & \cdots \\
0 & 0 & 0 & \frac{cB_{[1]}}{8} & 0 & \frac{-cB_{[1]}}{8} & 0 & 0 \\
0 & 0 & 0 & 0 & \frac{cB_{[1]}}{10} & 0 & -\frac{cB_{[1]}}{10} & 0 \\
0 & 0 & 0 & 0 & 0 & \ddots & 0 & \ddots \\
0 & 0 & 0 & 0 & 0 & 0 & \ddots & 0
\end{bmatrix}$$

$$-\nu \begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
0 & 0 & B_{[1]} & 0 & 0 & 0 & 0 & 0 & \cdots \\
0 & 0 & 0 & B_{[1]} & 0 & 0 & 0 & 0 & \cdots \\
0 & 0 & 0 & 0 & B_{[1]} & 0 & 0 & 0 & \cdots \\
0 & 0 & 0 & 0 & 0 & B_{[1]} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & B_{[1]} & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & B_{[1]} & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ddots
\end{bmatrix} +$$

$$\begin{bmatrix}
0 & 0 & 0 & 0 & 0 & \cdots & 0 \\
0 & 0 & 0 & 0 & 0 & \cdots & 0 \\
\frac{I_{[1]}}{4} & 0 & \frac{-I_{[1]}}{2(1\cdot 3)} & 0 & \frac{I_{[1]}}{4\cdot 6} & 0 & 0 \\
0 & \frac{I_{[1]}}{4\cdot 6} & 0 & \frac{-I_{[1]}}{2(2\cdot 4)} & 0 & \frac{I_{[1]}}{6\cdot 8} & 0 \\
0 & 0 & \frac{I_{[1]}}{(2i)(2i+2)} & 0 & \frac{-I_{[1]}}{2(i)(i+2)} & 0 & \frac{I_{[1]}}{(2i+2)(2i+4)} \\
0 & 0 & 0 & \ddots & 0 & \ddots & 0
\end{bmatrix}.$$

The tau conditions that must be imposed are:

$$\sum_{n=0}^{n=N} u_{mn}(-1)^n = \hat{h}_m,$$

$$\sum_{m=0}^{m=M} u_{mn}(-1)^m = \hat{v}_n^-,$$

$$\sum_{m=0}^{m=M} u_{mn} = \hat{v}_n^+,$$

and, again, there are two redundant tau conditions at the points $x = \pm 1$, $t = -1$.

Table 5 gives the absolute errors computed for the advection-diffusion equation.

6 Conclusions

The methods discussed in this article are quite useful in deriving efficient, spectrally accurate algorithms for the treatment of initial-boundary value problems in simple geometries with more than one nonperiodic directions. Separation of variables e.g. for the Laplace operator, leads

to equations, which can be easily transformed to form (8). As one may require the repeated solution of such equations high accuracy and efficiency are clearly essential. The bad conditioning of spectral differentiation operators is avoided by the integration preconditioning method, and this permits the treatment of problems at very high order of truncation that may otherwise be impractical. More complex geometries may be accessible as well: if a rational map to a rectangle is available, then the essential features of the method are preserved. Even if that is not feasible, the good conditioning of the resulting problems allows efficient iterative treatments to be applicable.

We must remark, however, that the preconditioners discussed in this note, although quite general, might prove inappropriate for certain problems with singular behavior. The specific structure of a given differential operator might lead to simpler preconditioners and to more natural reduced forms for the system. An example is offered by the Laplace operator in disk geometry; indeed, in solving $\triangle u = f$ in $0 \leq r \leq 2$, $0 \leq \theta \leq 2\pi$, using a Fourier/Chebyshev expansion in the azimuthal and radial directions respectively (with $-1 \leq x = r - 1 \leq 1$), we are led to the equation for the n-th Fourier mode $\left[((x+1)D)^2 - n^2\right] \hat{u} = (x+\hat{1})^2 f_n$. The method discussed above would lead to a pentadiagonal, ill-conditioned operator. However, closer examination of the matrix elements reveals that under left-multiplication by a certain tridiagonal preconditioner [19] (see also [3]) we get a tridiagonal system which can be solved quite naturally by using techniques developed for the study of 3-term recurrence relations [9], and difficulties relating to the coordinate singularity at $x = -1$ are easily bypassed.

We note that Tuckerman [19] gives a theorem on the transformation of matrices into banded form through left multiplication by preconditioners whose form depends on certain properties of the matrix elements. As is also mentioned in [3], preconditioners that lead to banded form have not been readily available, and have had to be searched for in a case-by-case basis. The main appeal of the method presented here is its generality, achieved through the construction of the preconditioner from the basic recursions of a family, and its identification with integration operators. Indeed, the preconditioner depends only on the basis used and the order of the differential operator $\mathbf{L}$, not on its special explicit form, which can be quite complex. Also, if the coefficients (or the solution) exhibit rapid variation over small neighborhoods, a rational coordinate mapping can be introduced to handle the situation with no substantial increase in algorithmic complexity while avoiding the need for considering very high-order truncations. In [6] we employed a variant of the present method, using integration postconditioning, to efficiently resolve shock-layer behavior through a low-order rational map. Naturally, as the Poisson equation in the disk suggests, problems with an underlying singularity may necessitate exploiting further properties of a given problem and special, tailor-made methods may need to be invented in place of the general-purpose technique presented here.

Acknowledgment

Partially supported by DOE Grant DE-FG03-92ER25128; part of this work was performed at Risø National Laboratory, DK–4000 Roskilde, Denmark. The work of the second author was partially supported by NSF Grant DMS-9304406. The authors would like to acknowledge many helpful conversations with J.P. Lynov.

References

[1] M. Abramowitz and I.A. Stegun, *Handbook of Mathematical Functions*, Dover, New York (1965).

[2] C. Bernardi and Y. Maday, *Polynomial interpolation results in Sobolev space*, J. Comput. and Appl. Math., 43 (1992), pp. 53-82.

[3] C. Canuto, M.Y. Hussaini, A. Quarteroni and T.A. Zang, *Spectral Methods in Fluid Dynamics*, Springer-Verlag, New York (1988).

[4] C.W. Clenshaw, *The numerical solution of linear differential equations in Chebyshev series*, Proc. Camb. Phil. Soc., 53 (1957), pp. 134-149.

[5] E.A. Coutsias, F.R. Hansen, T. Huld, G. Knorr and J.P. Lynov. *Spectral Methods in Numerical Plasma Simulation*, Phys. Scripta 40 (1989), pp. 270-279.

[6] E.A. Coutsias, T. Hagstrom and D. Torres, *An algorithm for the efficient spectral solution of linear ordinary differential equations with rational function coefficients*, NASA Technical Memorandum 106762, ICOMP-94-25, (1994).

[7] B. Fornberg, *A review of pseudospectral methods for solving partial differential equations*, Acta Numerica (1994), pp. 203-267.

[8] D.G. Fox and I.B. Parker (1968), *Chebyshev Polynomials in Numerical Analysis*, Oxford University Press, London.

[9] W. Gautschi, *Computational aspects of 3-term recurrence relations*, SIAM Rev. 9 (1967), pp. 24-82.

[10] D. Gottlieb and S. Orszag, *Numerical Analysis of Spectral Methods*, SIAM, Philadelphia (1977).

[11] L. Greengard, Spectral integration and two-point boundary value problems, *SIAM J. Numer. Anal.*, 28, (1991), 1071–1080.

[12] D.B. Haidvogel and T. Zang, *The accurate solution of Poisson's equation by expansion in Chebyshev polynomials*, J. Comp. Phys. 30 (1979), pp. 167-180.

[13] P. Haldenwang, G. Labrosse, S. Abboudi and M. Deville, *Chebyshev 3-D spectral and 2-D pseudospectral solvers for the Helmholtz equation*, J. Comp. Phys. 55 (1984), pp. 115-128.

[14] D. Hochstadt, *Special Functions of Mathematical Physics*, Dover, NY (1975).

[15] D.M. Hwang and C.T. Kelley, *Convergence of Broyden's method in Banach spaces*, SIAM J. Optim., 2 (1992), pp. 505-532.

[16] C.T. Kelley, *Iterative Methods for Linear and Nonlinear Equations*, SIAM, Philadelphia, 1995, to appear.

[17] C.T. Kelley and Z.Q. Xue, *Collective compactness and mesh independence of GMRES*, (1995), in preparation.

[18] H. Tal-Ezer, J.M. Carcione and D. Kosloff, *An accurate and efficient scheme for wave propagation in linear viscoelastic media*, Geophys. 55 (1990), pp. 1366-1379.

[19] L.S. Tuckerman, *Transformations of matrices into banded form*, J. Comp. Phys. 84 (1989), pp. 360-376.

A Hypergeometric Recurrence Relations

Hypergeometric functions satisfy the differential equation

$$z(1-z)\frac{d^2F}{dz^2} + [c-(a+b+1)z]\frac{dF}{dz} - abF = 0 \quad (62)$$

where a, b, and c are parameters and F represents $F(a,b;c;z)$. We show that the hypergeometric family $\bigcup_k F(a+k, b-k, c; z)$ satisfies the following recurrence relationships:

$$\begin{aligned} F(a,b;c;z) &= \alpha F'(a+1,b-1;c;z) \\ +\beta F'(a,b;c;z) &+ \gamma F'(a-1,b+1;c;z), \end{aligned} \quad (63)$$

$$\begin{aligned} [2(1-z)-1]F(a,b;c;z) &= \tilde{\alpha} F(a+1,b-1;c;z) \\ +\tilde{\beta} F(a,b;c;z) &+ \tilde{\gamma} F(a-1,b+1;c;z). \end{aligned} \quad (64)$$

Here $F' = \frac{dF}{dz}$ and α, β, γ, $\tilde{\alpha}$, $\tilde{\beta}$, and $\tilde{\gamma}$, are all coefficients that depend on a,b, and c.

The properties of the hypergeometric function that we use are

$$\frac{d}{dz}F(a,b;c;z) = \frac{ab}{c}F(a+1,b+1;c+1;z) \quad (65)$$

$$\begin{aligned} (b-a)(1-z)F(a,b;c;z) - (c-a)F(a-1,b;c;z) \\ +(c-b)F(a,b-1;c;z) = 0 \end{aligned} \quad (66)$$

and the Gauss contiguity relations

$$\begin{aligned} (c-a-1)F(a,b;c;z) + aF(a+1,b;c;z) \\ -(c-1)F(a,b;c-1;z) = 0 \end{aligned} \quad (67)$$

$$(68) \quad (b-a)F(a,b;c;z) + aF(a+1,b;c;z) - bF(a,b+1;c;z) = 0 \quad .$$

The following equations are required in the derivation of equation (63): Equation (68) evaluated the points (a,b,c), $(a+1,b,c)$, and $(a,b+1,c)$; Equation (65) evaluated at the points $(a+1,b-1,c-1)$, $(a,b,c-1)$, and $(a-1,b+1,c-1)$; and Equation (67). After involved algebra, one obtains the first recurrence relation (63) with

$$\alpha = \frac{a(c-b)}{(b-1)(b-a)(b-a-1)},$$

$$\beta = \frac{(a+b+1)-2c}{(b-a-1)(b-a+1)},$$

$$\gamma = \frac{b(c-a)}{(a-1)(b-a)(b-a+1)}.$$

The derivation of the second recurrence relation (64) requires the following equations: Equation (68) evaluated at the points $(a-1,b,c)$ and $(a,b-1,c)$ and Equation (66).

After some involved algebra, one can generate the second recurrence relation (64) with

$$\tilde{\alpha} = \frac{2(c-b)a}{(b-a)(b-a-1)},$$

$$\tilde{\beta} = \frac{(a+b-1)(a+b+1-2c)}{(b-a+1)(b-a-1)},$$

$$\tilde{\gamma} = \frac{2(c-a)b}{(b-a)(b-a+1)}.$$

B Confluent Hypergeometric Recurrence Relations

Confluent hypergeometric functions satisfy the differential equation

$$z\frac{d^2u}{dz^2} + (\gamma - z)\frac{du}{dz} - \alpha u = 0$$

We show that the confluent hypergeometric functions satisfy recurrence relations analogous to the recurrence relations for hypergeometric functions. There are two types of confluent hypergeometric functions. Each one is treated separately.

We start with the first confluent hypergeometric function. It satisfies the following equations:

$$(69) \quad \frac{d\Phi}{dz}(\alpha,\gamma;z) = \frac{\alpha}{\gamma}\Phi(\alpha+1,\gamma+1;z),$$

$$(70) \quad (\gamma-\alpha-1)\Phi(\alpha,\gamma;z) + \alpha\Phi(\alpha+1,\gamma;z) - (\gamma-1)\Phi(\alpha,\gamma-1;z) = 0 \quad .$$

Using Equation (70) evaluated at $(\alpha+1,\gamma+1)$, $(\alpha+1,\gamma)$, and (α,γ), and Equation (69), one can derive

$$(71) \quad \Phi(\alpha,\gamma;z) = \Phi'(\alpha,\gamma;z) + \frac{\gamma-\alpha}{\alpha-1}\Phi'(\alpha-1,\gamma;z).$$

A similar recurrence relation can be derived for confluent hypergeometric functions of the second kind. These functions satify the following two equations

$$(72) \quad \frac{d\Psi}{dz}(\alpha,\gamma;z) = -\alpha\Psi(\alpha+1,\gamma+1;z),$$

$$(73) \quad \Psi(\alpha,\gamma;z) - \alpha\Psi(\alpha+1,\gamma;z) - \Psi(\alpha,\gamma-1;z) = 0.$$

Using (73) evaluated at points $(\alpha+1,\gamma+1)$ and $(\alpha+1,\gamma)$ and Equation (72), one can generate

$$(74) \quad \Psi(\alpha,\gamma;z) = \frac{1}{1-\alpha}\Psi'(\alpha-1,\gamma;z) + \Psi'(\alpha,\gamma;z).$$

In addition to the recurrence equations developed above, both types of confluent hypergeometrics satisfy a recurrence of the form

$$\begin{aligned} zf(\alpha,\gamma;z) &= c_1 f(\alpha-1,\gamma,z) \\ &+ c_2 f(\alpha,\gamma,z) + c_3 f(\alpha+1,\gamma,z). \end{aligned}$$

The first confluent hypergeometric functions satisfy

$$\begin{aligned} (75)\,(\gamma-\alpha)\Phi(\alpha-1,\gamma;z) &+ (2\alpha-\gamma+z)\Phi(\alpha,\gamma;z) \\ -\alpha\Phi(\alpha+1,\gamma;z) &= 0. \end{aligned}$$

This equation can be rearranged into the form

$$\begin{aligned} (76) \quad z\Phi(\alpha,\gamma;z) &= (\alpha-\gamma)\Phi(\alpha-1,\gamma;z) + \\ (\gamma-2\alpha)\Phi(\alpha,\gamma;z) &+ \alpha\Phi(\alpha+1,\gamma;z). \end{aligned}$$

The confluent hypergeometric functions of the second kind satisfy

$$\begin{aligned} \Psi(\alpha-1,\gamma;z) &- (2\alpha-\gamma+z)\Psi(\alpha,\gamma;z) \\ (77) \qquad &+ \alpha(\alpha-\gamma+1)\Psi(\alpha+1,\gamma;z) = 0, \end{aligned}$$

which can be rearranged into the form

$$\begin{aligned} (78) \quad z\Psi(\alpha,\beta;z) &= \Psi(\alpha-1,\gamma;z) - \\ (2\alpha-\gamma)\Psi(\alpha,\gamma;z) &+ \alpha(\alpha-\gamma+1)\Psi(\alpha+1,\gamma;z). \end{aligned}$$

Family	Chebyshev T_k	Legendre P_k	Gegenbauer $C_k^{(\nu)}$	Jacobi $P_k^{(\alpha,\beta)}$	Hypergeometric $F(a+k,b-k;c)$
Q_0	1	1	1	1	-
Q_1	x	x	$2\nu x$	$\frac{1}{2}((\alpha-\beta)+(\alpha+\beta+2)x)$	-
$a_{k-1,k}$	$\frac{1}{2}$	$\frac{k}{2k+1}$	$\frac{k+2\nu-1}{2(k+\nu)}$	$\frac{2(k+\alpha)(k+\beta)}{(2k+\alpha+\beta+1)(2k+\alpha+\beta)}$	$\frac{2(c-a-k)(b-k)}{(b-a-2k)(b-a-2k+1)}$
$a_{k,k}$	0	0	0	$-\frac{(\alpha^2-\beta^2)}{(2k+\alpha+\beta+2)(2k+\alpha+\beta)}$	$\frac{(a+b-1)(a+b+1-2c)}{(b-a-2k+1)(b-a-2k-1)}$
$a_{k+1,k}$	$\frac{1}{2}$	$\frac{k+1}{2k+1}$	$\frac{k+1}{2(k+\nu)}$	$\frac{2(k+1)(k+\alpha+\beta+1)}{(2k+\alpha+\beta+2)(2k+\alpha+\beta+1)}$	$\frac{2(c-b+k)(a+k)}{(b-a-2k)(b-a-2k-1)}$
$b_{k-1,k}$	$-\frac{1}{2(k-1)}$	$-\frac{1}{2k+1}$	$\frac{-1}{2(k+\nu)}$	$\frac{-a_{k-1,k}}{k+\alpha+\beta}$	$\frac{-a_{k-1,k}}{k+a-1}$
$b_{k,k}$	0	0	0	$\frac{-2a_{k,k}}{\alpha+\beta}$	$\frac{-2a_{k,k}}{a+b-1}$
$b_{k+1,k}$	$\frac{1}{2(k+1)}$	$\frac{1}{2k+1}$	$\frac{1}{2(k+\nu)}$	$\frac{a_{k+1,k}}{k+1}$	$\frac{a_{k+1,k}}{k+1-b}$
$w(x)$	$(1-x^2)^{-1/2}$	1	$(1-x^2)^{\nu-\frac{1}{2}}$	$(1-x)^\alpha(1+x)^\beta$	$(1+x)^{c-1}(1-x)^{a+b-c}$
$p(x)$	$(1-x^2)^{1/2}$	$(1-x^2)$	$(1-x^2)^{\nu+\frac{1}{2}}$	$(1-x^2)w(x)$	$-(1+x)^c(1-x)^{a+b-c+1}$
(a,b)	$(-1,1)$	$(-1,1)$	$(-1,1)$	$(-1,1)$	$(-1,1)$
h_k	$\pi/2(\pi,k=0)$	$\frac{2}{2k+1}$	$\frac{\pi 2^{1-2\nu}\Gamma(k+2\nu)}{k!(k+\nu)[\Gamma(\nu)]^2}$	$\frac{2^{\alpha+\beta+1}\Gamma(k+\alpha+1)\Gamma(k+\beta+1)}{(2k+\alpha+\beta+1)k!\Gamma(k+\alpha+\beta+1)}$	-
λ_k	k^2	$k(k+1)$	$k(k+2\nu)$	$k(k+\alpha+\beta+1)$	$(a+k)(b-k)$

Table 11: *Recursions for polynomial families of Hypergeometric type ($a_{0,k}=0$; for the T_k, $a_{1,0}=b_{1,0}=1$)*

Family	Hermite H_k	Laguerre $L_k^{(\alpha)}$	Confluent (first) $\Phi(a,b)$	Confluent (second) $\Psi(a,b)$
Q_0	1	1	-	-
Q_1	$2x$	$1+\alpha-x$	-	-
$a_{k-1,k}$	k	$-(k+\alpha)$	$a-b+k$	1
$a_{k,k}$	0	$2k+\alpha+1$	$b-2a-2k$	$b-2a-2k$
$a_{k+1,k}$	$\frac{1}{2}$	$-(k+1)$	$a+k$	$(a+k)(a-b+k+1)$
$b_{k-1,k}$	0	0	$\frac{b-a-k}{a+k-1}$	$\frac{1}{1-a-k}$
$b_{k,k}$	0	1	1	1
$b_{k+1,k}$	$\frac{1}{2(k+1)}$	-1	0	0
$w(x)$	e^{-x^2}	$x^\alpha e^{-x}$	$x^{b-1}e^{-x}$	$x^{b-1}e^{-x}$
$p(x)$	e^{-x^2}	$x^{\alpha+1}e^{-x}$	$x^b e^{-x}$	$x^b e^{-x}$
(a,b)	$(-\infty,\infty)$	$(0,\infty)$	$(0,\infty)$	$(0,\infty)$
h_k	$\sqrt{\pi}2^k k!$	$\frac{\Gamma(\alpha+k+1)}{k!}$	-	-
λ_k	$2k$	k	$-(a+k)$	$-(a+k)$

Table 12: *Recursions for Confluent Hypergeometric functions*