Univariate modified Fourier methods for second order boundary value problems

Ben Adcock
DAMTP, Centre for Mathematical Sciences
University of Cambridge
Wilberforce Rd, Cambridge CB3 0WA
United Kingdom

December 12, 2008

Abstract

We develop and analyse a new spectral-Galerkin method for the numerical solution of linear, second order differential equations with homogeneous Neumann boundary conditions. The basis functions for this method are the eigenfunctions of the Laplace operator subject to these boundary conditions. Due to this property this method has a number of beneficial features, including an $\mathcal{O}(N^2)$ condition number and the availability of an optimal, diagonal preconditioner. This method offers a uniform convergence rate of $\mathcal{O}(N^{-3})$, however we show that by the inclusion of an additional $2M$ basis functions, this figure can be increased to $\mathcal{O}(N^{-2M-3})$ for any positive integer M.

Introduction

Univariate modified Fourier series are eigenseries of the Laplace operator equipped with homogeneous Neumann boundary conditions. These were introduced in [8] as an alternative to Fourier series for the approximation of non-periodic functions. Due to the boundary conditions the modified Fourier coefficients of a function decay faster than their Fourier counterparts, meaning more rapid, in fact, uniform convergence.

Also introduced in [8] were quadrature routines to evaluate such coefficients numerically, thus circumventing the need for the Fast Fourier Transform (FFT). The methods proposed offer a number of benefits over this conventional approach, not least being that the coefficients can be calculated one-by-one, and any N coefficients may be calculated in $\mathcal{O}(N)$ operations without the restriction that N need be a highly composite integer or fixed in advance.

Since these functions may be viewed as eigenfunctions of a regular Sturm–Liouville problem, the convergence of the modified Fourier series of a given function is only algebraic in the truncation parameter. Standard expansions in Jacobi polynomials converge spectrally (under the assumption of smoothness). However, the modified Fourier basis offers at least one significant advantage in higher dimensions. The coefficients lie on a hyperbolic cross [6, 19] and, due to the adaptive method for their evaluation, this means that modified Fourier expansions are amenable to sparse approximation methods. The result is that, rather than needing $\mathcal{O}\left(N^d\right)$ coefficients, where d is the dimension, expansions only need to comprise $\mathcal{O}\left(N(\log N)^{d-1}\right)$ terms without unduly affecting the convergence rate. Moreover, there is a simple tool to accelerate the convergence rate of the expansion, namely the polynomial subtraction process, [16]. This has been extended to the multivariate setting in [6].

One obvious application of modified Fourier series is in the spectral-Galerkin approximation of differential equations. Due to the nature of the basis functions, these series are best suited to second order problems with Neumann boundary conditions:

$$\mathcal{L}[u](x) = -\triangle u(x) + a(x) \cdot \nabla u(x) + b(x)u(x) = f(x), \quad x \in \Omega, \quad \frac{\partial u}{\partial n}\Big|_{\partial\Omega} = 0.$$

In this paper we shall present the theory of such approximations for the domain $\Omega = [-1, 1]$.

In this setting the so-called *modified Fourier–Galerkin method* will not generally outperform standard spectral-Galerkin schemes based on Jacobi polynomials. However, there are various situations where this method can be expected to offer advantages, which we now describe.

As alluded to, the modified Fourier method only comprises $\mathcal{O}\left(N(\log N)^{d-1}\right)$ terms for $d \geq 2$. These terms can be found in $\mathcal{O}\left(N^2\right)$ operations, independent of d, [1]. On the other hand, standard spectral approximations involve $\mathcal{O}\left(N^d\right)$ terms which can be found in at best $\mathcal{O}\left(N^{d+1}\right)$ or $\mathcal{O}\left(N^d \log N\right)$ operations, [3]. Thus, despite offering a typically slower convergence rate, the modified Fourier approach has a much reduced complexity.

Furthermore, unlike the univariate case, the solution u to $\mathcal{L}[u] = f$ with $\Omega = [-1,1]^d$ and $d \geq 2$ is only guaranteed H^2-regularity, even if the inhomogeneous term f is smooth, [5]. For such problems, standard spectral schemes offer only algebraic convergence. In this setting the modified Fourier method will also be competitive.

The implementation of the modified Fourier method for $\Omega = [-1,1]^d$, $d \geq 2$, has been studied in [1]. Several numerical examples are presented. These compare this method to standard polynomial schemes, highlighting the aforementioned features.

Unfortunately, when the solution u to such a problem is smooth, polynomial spectral methods will yield better approximations (for sufficiently large N). As mentioned, there is a tool to accelerate the convergence rate of modified Fourier expansions. In this paper we demonstrate how to successfully incorporate it into spectral approximations for univariate boundary value problems. The aim of future study is to do the same for spectral approximations in the d-variate cube, and by doing so create a new method that is competitive for a larger class of problems.

The modified Fourier approach possesses one other significant advantage. The Laplace–Neumann eigenfunctions are known explicitly on the equilateral triangle, [14]. For this reason, they form a suitable basis for the spectral approximation of boundary value problems on such domains. This approach may possess significant advantages over standard spectral schemes in terms of its relative simplicity and warrants future investigation.

The outline of the remainder of this paper is as follows. In Section 1 we introduce the basic properties of modified Fourier expansions, including convergence and numerical evaluation of the coefficients. The modified Fourier method is developed and analysed in Section 2. We provide estimates for the rate of convergence, devise an optimal preconditioner and introduce an iterative scheme for the discretization equations. In Section 3 we apply this method to variable coefficient problems. A device for accelerating convergence is introduced and analysed in Section 4. Finally, in Section 5 we consider how to deal with other boundary conditions.

Notation: We write $(\cdot,\cdot)$ for the standard $L^2(-1,1)$ inner product and $\|\cdot\|$ for the induced norm. We shall denote the H^q, $q > 0$, and uniform norms by $\|\cdot\|_q$ and $\|\cdot\|_\infty$ respectively. N shall be a truncation parameter.

1 Modified Fourier series in the unit interval

1.1 Definition, basic properties and convergence

The modified Fourier basis introduced in [8] is

$$\{\cos n\pi x \ : \ n \geq 0\} \cup \{\sin(n - \tfrac{1}{2})\pi x \ : \ n \geq 1\}. \tag{1.1}$$

This is precisely the set of eigenfunctions of the univariate Laplace operator on $[-1,1]$ with homogeneous Neumann boundary conditions. From elementary spectral theory we deduce:

Lemma 1. *The modified Fourier functions form an orthonormal basis of $L^2(-1,1)$.*

If $f \in L^2(-1,1)$ we define its truncated modified Fourier series by

$$\mathcal{F}_N[f](x) = \sum_{i=0}^{1} \sum_{n=0}^{N} c_n^{[i]} \hat{f}_n^{[i]} \phi_n^{[i]}(x), \quad -1 \leq x \leq 1,$$

where $c_0^{[0]} = \frac{1}{2}$, $c_n^{[i]} = 1$ otherwise, $\phi_n^{[0]}(x) = \cos n\pi x$, $\phi_n^{[1]}(x) = \sin(n - \frac{1}{2})\pi x$ and

$$\hat{f}_n^{[i]} = \int_{-1}^{1} f(x)\phi_n^{[i]}(x)dx.$$

To simplify notation we set $\phi_0^{[1]} = 0$. $\mathcal{F}_N$ is the orthogonal projection from $L^2(-1,1)$ onto the space

$$\mathcal{S}_N = \text{span}\{\phi_n^{[i]} : i = 0, 1,\ n = 0, ..., N\}$$

In particular, we have:

Corollary 1. *Suppose that $f \in L^2(-1,1)$. Then $\mathcal{F}_N[f]$ is the best approximation to f from $\mathcal{S}_N$ in the L^2 norm, $\|f - \mathcal{F}_N[f]\| \to 0$ as $N \to \infty$ and*

$$\|f\|^2 = \sum_{i=0}^{1}\sum_{n=0}^{\infty} c_n^{[i]}|\hat{f}_n^{[i]}|^2. \tag{1.2}$$

Proof. Since $\mathcal{F}_N$ is the orthogonal projection, for $\phi \in \mathcal{S}_N$ we have

$$\|f - \phi\|^2 = \|f - \mathcal{F}_N[f]\|^2 + \|\mathcal{F}_N[f] - \phi\|^2 \geq \|f - \mathcal{F}_N[f]\|^2,$$

so that $\mathcal{F}_N[f]$ is the best approximation. Using Lemma 1 we obtain convergence. To deduce the identity (1.2), we first note that this formula holds for $\mathcal{F}_N[f] \in \mathcal{S}_N$ by orthogonality. L^2 convergence of $\mathcal{F}_N[f]$ to f now gives the result. □

The key difference between the Fourier and modified Fourier bases is summed up in the following theorem:

Theorem 1. *Suppose that $f \in H^1(-1,1)$. Then $\mathcal{F}_N[f]$ is the best approximation to f from $\mathcal{S}_N$ in the H^1 norm, $\|f - \mathcal{F}_N[f]\|_1 \to 0$ as $N \to \infty$ and*

$$\|f\|_1^2 = \sum_{i=0}^{1}\sum_{n=0}^{\infty} c_n^{[i]}(1 + \mu_n^{[i]})|\hat{f}_n^{[i]}|^2,$$

where $\mu_n^{[0]} = n^2\pi^2$ and $\mu_n^{[1]} = (n - \frac{1}{2})^2\pi^2$ are the eigenfunctions corresponding to the modified Fourier basis functions.

Proof. One integration by parts gives

$$\hat{f}_n^{[i]} = \frac{(-1)^{1-i}}{(\mu_n^{[i]})^{\frac{1}{2}}}\int_{-1}^{1} f'(x)\psi_n^{[1-i]}(x)\ dx = \frac{(-1)^{1-i}}{(\mu_n^{[i]})^{\frac{1}{2}}}\breve{f'}_n^{[1-i]},$$

where $\psi_n^{[i]}$ is a Laplace–Dirichlet eigenfunction

$$\psi_n^{[0]}(x) = \cos(n - \tfrac{1}{2})\pi x, \quad \psi_n^{[1]}(x) = \sin n\pi x, \quad n \geq 1,$$

and $\breve{g}_n^{[i]}$ is the coefficient of a function g corresponding to $\psi_n^{[i]}$. Differentiating $\mathcal{F}_N[f]$ and using this formula, we obtain

$$(\mathcal{F}_N[f])'(x) = \sum_{i=0}^{1}\sum_{n=1}^{N}(\mu_n^{[i]})^{\frac{1}{2}}(-1)^{1-i}\hat{f}_n^{[i]}\psi_n^{[1-i]}(x) = \sum_{i=0}^{1}\sum_{n=1}^{N}\breve{f'}_n^{[1-i]}\psi_n^{[1-i]}(x), \quad -1 \leq x \leq 1,$$

so that $(\mathcal{F}_N[f])'$ is the N^{th} truncated Laplace–Dirichlet series of f'. The set of Laplace–Dirichlet eigenfunctions is an orthonormal basis of $L^2(-1,1)$, so an equivalent version of Corollary 1 holds. This gives the result. □

The modified Fourier basis is dense in $H^1(-1,1)$, not merely $L^2(-1,1)$. From this we immediately deduce uniform convergence of $\mathcal{F}_N[f]$ to f:

Theorem 2. *Suppose that $f \in H^1(-1,1)$. Then $\|f - \mathcal{F}_N\|_\infty \to 0$ as $N \to \infty$.*

Proof. This follows immediately from Theorem 1 and Sobolev's inequality

$$\|g\|_\infty \leq \sqrt{\tfrac{5}{2}\|g\|\|g\|_1}, \quad \forall g \in H^1(-1,1). \qquad \square$$

Note that pointwise convergence in $(-1,1)$ of the modified Fourier expansion was originally proved in [8] and uniform convergence in [13]. Theorem 2 presents an alternative approach.

Theorem 1 cannot be extended to other Sobolev spaces unless particular derivative conditions are enforced (the analogue of the periodic Sobolev spaces $H^k(\mathbb{T})$ for Fourier series). To see why this is the case we consider the asymptotic expansion of the coefficients:

Lemma 2. *Suppose that $f \in H^{2k}(-1,1)$, $k \geq 1$. Then, for $n \geq 1$ and $i = 0, 1$,*

$$\hat{f}_n^{[i]} = (-1)^{n+i} \sum_{r=0}^{k-1} \frac{(-1)^r}{(\mu_n^{[i]})^{r+1}} \triangle^{[i]}[f^{(2r+1)}] + \frac{(-1)^k}{(\mu_n^{[i]})^k} \int_{-1}^{1} f^{(2k)}(x)\phi_n^{[i]}(x)\, dx,$$

where $\triangle^{[i]}[g] = g(1) + (-1)^{i+1}g(-1)$. Suppose further that $f^{(2r+1)}(\pm 1) = 0$ for $r = 0, ..., k-1$. Then $|\hat{f}_n^{[i]}| \leq (\mu_n^{[i]})^{-k}\|f\|_{2k}$. If additionally $f \in H^{2k+1}(-1,1)$ or $f \in H^{2k+2}(-1,1)$ then $|\hat{f}_n^{[i]}| \leq (\mu_n^{[i]})^{-k-\frac{1}{2}}\|f\|_{2k+1}$ or $|\hat{f}_n^{[i]}| \leq c(\mu_n^{[i]})^{-k-1}\|f\|_{2k+2}$ respectively, where $c = 2\sqrt{\frac{5}{2}} + 1$.

Proof. The expansion is derived by repeated integration by parts. The bounds follow immediately from this expansion, using Sobolev's inequality where necessary. $\square$

Evidently, the analogue of periodicity for modified Fourier series are the Neumann conditions

$$f^{(2r+1)}(\pm 1) = 0, \quad r = 0, 1, 2, ...$$

We shall not consider such functions. However, it will be of much use in the sequel to assess the case where the first k such conditions are satisfied. We now do this:

Theorem 3. *Suppose that $u \in H^{2k+1}(-1,1)$ obeys the first k derivative conditions, in other words $u^{(2s+1)}(\pm 1) = 0$, $s = 0, ..., k-1$. Then, for $r = 0, ..., 2k+1$, $\mathcal{F}_N[u]$ is the best approximation to u from $\mathcal{S}_N$ in the H^r norm, $\|u - \mathcal{F}_N[u]\|_r \to 0$ and we have*

$$\|u\|_r^2 = \sum_{i=0}^{1} \sum_{n=0}^{\infty} c_n^{[i]} \sum_{j=0}^{r} (\mu_n^{[i]})^j |\hat{u}_n^{[i]}|^2. \tag{1.3}$$

Proof. This is essentially the same as the proof of Lemma 1. If r is even then $(\mathcal{F}_N[u])^{(r)}$ is the truncated modified Fourier series of $u^{(r)}$, otherwise it is the truncated Laplace–Dirichlet series. Thus we obtain convergence. (1.3) follows in the same manner as before. $\square$

In the sequel we will use a simple version of Bernstein's inequality, which now follows immediately:

Corollary 2. *Suppose that $\phi \in \mathcal{S}_N$. Then, for $r = 0, 1, 2, ...$*

$$\|\phi\|_r \leq (N+1)^r \pi^r \|\phi\|.$$

Proof. We use (1.3). Noting that $\sum_{j=0}^{r} (\mu_n^{[i]})^j \leq (1 + \mu_n^{[i]})^r \leq (N+1)^{2r}\pi^{2r}$ for $n \leq N$ gives the result. $\square$

In a similar manner to Theorem 2 we may assert uniform convergence of the first $2k$ derivatives of $\mathcal{F}_N[u]$:

Theorem 4. *Suppose that u is as in Theorem 3. Then $\|(\mathcal{F}_N[u])^{(r)} - u^{(r)}\|_\infty \to 0$ as $N \to \infty$ for $r = 0, ..., 2k$.*

Proof. From Theorem 3 we have $\|u - \mathcal{F}_N[u]\|_{2k+1} \to 0$ as $N \to \infty$. For $r = 0, ..., 2k$, we have

$$\|u^{(r)} - (\mathcal{F}_N[u])^{(r)}\|_\infty \leq \sqrt{\tfrac{5}{2}\|u^{(r)} - (\mathcal{F}_N[u])^{(r)}\|\|u^{(r)} - (\mathcal{F}_N[u])^{(r)}\|_1} \leq \sqrt{\tfrac{5}{2}}\|u - \mathcal{F}_N[u]\|_{2k+1},$$

which follows from the inequalities $\|u\| \leq \|u\|_1 \leq ... \leq \|u\|_{2k+1}$. $\square$

1.2 Rate of convergence

To provide estimates for the rate of convergence in various norms we shall adopt two approaches. The first uses the characterization of the H^r norm given in Theorems 1 and 3. It is a standard tool of Fourier analysis:

Theorem 5. *Suppose that u is as in Theorem 3. Then*

$$\|u - \mathcal{F}_N[u]\|_s \leq (N\pi)^{s-r}\|u\|_r, \quad s = 0, ..., 2k+1, \quad r = s, ..., 2k+1.$$

Proof. From Theorem 3 we have

$$\|u - \mathcal{F}_N[u]\|_s^2 = \sum_{i=0}^{1}\sum_{n>N}\sum_{j=0}^{s}(\mu_n^{[i]})^j|\hat{u}_n^{[i]}|^2 \leq (N\pi)^{2(s-r)}\sum_{i=0}^{1}\sum_{n=0}^{\infty}c_n^{[i]}\sum_{j=r-s}^{r}(\mu_n^{[i]})^j|\hat{u}_n^{[i]}|^2$$
$$\leq (N\pi)^{2(s-r)}\sum_{i=0}^{1}\sum_{n=0}^{\infty}c_n^{[i]}\sum_{j=0}^{r}(\mu_n^{[i]})^j|\hat{u}_n^{[i]}|^2 = (N\pi)^{2(s-r)}\|u\|_r^2,$$

which gives the result. □

These estimates are not optimal for a function $u \in H^{2k+2}(-1,1)$. The reason being that there is no characterization of $\|u\|_{2k+2}$ in terms of modified Fourier coefficients of u, since $u^{(2k)}$ does not obey any derivative conditions. To deduce rates of convergence in this case we use the bounds derived in Lemma 2:

Theorem 6. *Suppose that $u \in H^{2k+2}(-1,1)$ obeys the first k derivative conditions. Then*

$$\|u - \mathcal{F}_N[u]\|_s \leq c_s(N\pi)^{s-2k-\frac{3}{2}}\|u\|_{2k+2}, \quad s = 0, ..., 2k+1,$$

where $c_s = c\left(2^{s+1}(4k+3-2s)\pi^{-1}\right)^{\frac{1}{2}}$ and c is the constant of Lemma 2.

Proof. From Lemma 2 we have

$$\|u - \mathcal{F}_N[u]\|_s^2 \leq \sum_{i=0}^{1}\sum_{n>N}(1+\mu_n^{[i]})^s|\hat{u}_n^{[i]}|^2 \leq 2^{s+1}c^2\|u\|_{2k+2}^2\sum_{n\geq N}(n\pi)^{2s-4k-4}.$$

Since $\sum_{n\geq N} n^{-t-1} \leq tN^{-t}$ for $t > 0$ we obtain the result. □

In the same manner we obtain a uniform error estimate:

Theorem 7. *Suppose that $u \in H^{2k+2}(-1,1)$ obeys the first k derivative conditions. Then*

$$\|u^{(s)} - (\mathcal{F}_N[u])^{(s)}\|_\infty \leq 2c(r-s)\pi^{-1}(N\pi)^{s-r}\|u\|_{r+1}, \quad s = 0, ..., 2k, \quad r = s+1, ..., 2k+1,$$

where c is the constant of Lemma 2.

Proof. Due to Theorems 2 and 4 and the bound in Lemma 2 we have

$$\|u^{(s)} - (\mathcal{F}_N[u])^{(s)}\|_\infty \leq \sum_{i=0}^{1}\sum_{n>N}|\hat{u}_n^{[i]}|\|(\phi_n^{[i]})^{(s)}\|_\infty \leq 2c\|u\|_{r+1}\sum_{n\geq N}(n\pi)^{s-r-1} \leq 2c(r-s)\pi^{-1}(N\pi)^{s-r}\|u\|_{r+1},$$

which gives the result. □

It turns out that, under some additional smoothness assumptions, the convergence rate of $\mathcal{F}_N[u]$ to u is faster by one power of N inside the interval. This result is due to S. Olver:

Theorem 8 (S. Olver, [13])**.** *Suppose that $u \in C^{2k+2}(-1,1)$, $u^{(2k+2)}$ has bounded variation and u obeys the first k derivative conditions. Then the error*

$$u(x) - \mathcal{F}_N[u](x) = \mathcal{O}\left(N^{-2k-2}\right),$$

uniformly for x in compact subsets of $(-1,1)$.

For the purposes of spectral methods, this result is somewhat superfluous since the Galerkin approximation does not usually offer a faster convergence rate inside the interval. In this case, the estimates in Theorem 7 are better suited.

1.3 Evaluation of modified Fourier coefficients

The standard means to evaluate Fourier coefficients to high accuracy is the Fast Fourier Transform (FFT). Indeed it is possible to do the same for modified Fourier series. However, recent developments in computational techniques for highly oscillatory integrals [7, 12] have been used effectively to evaluate such coefficients, [8, 10]. This approach possesses a number of advantages. Whereas the FFT requires $\mathcal{O}(N \log N)$ operations to evaluate the first N Fourier coefficients, the number of operations for the methods proposed is $\mathcal{O}(N)$. Such techniques are also fully adaptive: unlike the FFT, N does not need to be a highly composite integer, or to be fixed in advance.

The technique described in [8] is as follows. Suppose that $-1 = c_1 < c_2 < ... < c_\nu = 1$ are given quadrature nodes with multiplicities $m_1, ..., m_\nu$ and q is a polynomial such that

$$q^{(2r)}(c_k) = f^{(2r+1)}(c_k), \quad r = 0, ..., m_k - 1, \quad k = 1, 2, ..., \nu.$$

Then, if $p(x) = f(0) + \int_0^x q(x')dx'$, we approximate the modified Fourier coefficients by

$$\hat{f}_n^{[i]} \approx \int_{-1}^{1} p(x)\phi_n^{[i]}(x)\ dx, \tag{1.4}$$

which may be calculated explicitly. The asymptotic error in doing so is $\mathcal{O}(n^{-2s-2})$ where $s = \min\{m_1, m_\nu\}$, and, since the modified Fourier coefficients decay like $\mathcal{O}(n^{-2})$, the relative error is $\mathcal{O}(n^{-2s})$. Note that this convergence rate does not depend on the interpolation at the interior nodes, which acts to lower the error constant.

Unfortunately such a technique may not be used to calculate $\hat{f}_0^{[0]}$ and does not generate the required accuracy for small n. However these integrals may be evaluated by standard quadrature instead. Further details are given in [8, 10].

For the rest of this paper we shall assume that the error in calculating the modified Fourier coefficients of a function f is insignificant in comparison to the error in the numerical methods considered, and thus is of little concern in any corresponding estimates.

2 The modified Fourier–Galerkin method

The main concern of this paper is the numerical approximation of the univariate, second order boundary value problem

$$\mathcal{L}[u] = -u_{xx} + au_x + bu = f, \quad u_x(\pm 1) = 0,$$

by modified Fourier series. For the moment we shall assume that a and b are constant (later we shall address the variable coefficient case). In weak form, if $T : H^1(-1,1) \times H^1(-1,1) \to \mathbb{R}$ is the bilinear form

$$T(u,v) = (u_x, v_x) + a(u_x, v) + b(u,v), \quad \forall u, v \in H^1(-1,1),$$

then we may rewrite this problem as

$$\text{find} \quad u \in H^1(-1,1) : \quad T(u,v) = (f,v), \quad \forall v \in H^1(-1,1).$$

Initially we shall assume that the operator T is continuous and coercive. In other words the exist positive constants γ and ω such that

$$|T(u,v)| \le \gamma \|u\|_1 \|v\|_1, \quad T(u,u) \ge \omega \|u\|_1^2, \quad \forall u, v \in H^1(-1,1).$$

It is readily verified by Young's inequality, [11],

$$xy \le \frac{1}{4\epsilon}x^2 + \epsilon y^2, \quad \forall x, y \in \mathbb{R}, \epsilon > 0,$$

that T is continuous and coercive provided $b - \frac{1}{4}a^2 > 0$. Indeed, under this assumption,

$$T(u,u) = \|u_x\|^2 + a(u_x, u) + b\|u\|^2 \ge (1-\epsilon)\|u_x\|^2 + (b - \tfrac{1}{4\epsilon}a^2)\|u\|^2 \ge \omega \|u\|_1^2,$$

for some appropriately chosen $\omega > 0$.

2.1 Galerkin's equations

We seek an approximation $u_N \in \mathcal{S}_N$ to u of the form

$$u_N(x) = \sum_{j=0}^{1} \sum_{m=0}^{N} c_m^{[j]} a_m^{[j]} \phi_m^{[j]}(x), \tag{2.1}$$

with unknown coefficients $a_m^{[j]}$, which satisfies Galerkin's equations $(\mathcal{L}[u_N], \phi) = (f, \phi)$, $\forall \phi \in \mathcal{S}_N$. Setting $\phi = \phi_n^{[i]}$ for $i = 0, 1$ and $n = i, ..., N$, after recalling that the modified Fourier basis functions are orthonormal eigenfunctions of the Laplace operator, we obtain

$$(b + \mu_n^{[i]}) a_n^{[i]} + a \sum_{j=0}^{1} \sum_{m=0}^{N} c_m^{[j]} a_m^{[j]} ((\phi_m^{[j]})', \phi_n^{[i]}) = \hat{f}_n^{[i]}, \quad i = 0, 1, \quad n = i, ..., N.$$

Since $\phi_n^{[0]}$ is even and $\phi_n^{[1]}$ is odd, we have

$$((\phi_m^{[j]})', \phi_n^{[i]}) = \begin{cases} \delta_{n,m}^{[i]} & j = 1 - i, \quad m = 1, ..., N, \\ 0 & \text{otherwise,} \end{cases}$$

where

$$\delta_{n,m}^{[i]} = \int_{-1}^{1} (\phi_m^{[1-i]})'(x) \phi_n^{[i]}(x)\, dx = 2(-1)^{n+m} \frac{\mu_m^{[1-i]}}{\mu_n^{[i]} - \mu_m^{[1-i]}}, \quad i = 0, 1, \quad n, m = 0, ..., N.$$

Hence Galerkin's equations are

$$(b + \mu_n^{[i]}) a_n^{[i]} + a \sum_{m=1}^{N} a_m^{[1-i]} \delta_{n,m}^{[i]} = \hat{f}_n^{[i]}, \quad i = 0, 1, \quad n = i, ..., N. \tag{2.2}$$

If $\hat{a} = (a^{[0]}, a^{[1]})^\top$ is the vector with entries $a_m^{[j]}$ and $\hat{f} = (\hat{f}^{[0]}, \hat{f}^{[1]})^\top$ is the vector of modified Fourier coefficients of the function f then this may be written in matrix form as $A_{\mathrm{G}} \hat{a} = \hat{f}$, where

$$A_{\mathrm{G}} = \begin{pmatrix} D^{[0]} & a\delta^{[0]} \\ a\delta^{[1]} & D^{[1]} \end{pmatrix}. \tag{2.3}$$

Here $\delta^{[i]}$ is the $(N + 1 - i) \times (N + i)$ matrix with entries $\delta_{n,m}^{[i]}$ and $D^{[i]}$ are diagonal matrices with entries $b + \mu_n^{[i]}$, $i = 0, 1$. The diagonal part of A_{G} corresponds to the restriction of the operator $\mathcal{L}_0 = -\partial_{xx} + b\imath$, where $\imath$ is the identity operator, to $\mathcal{S}_N$. The off-diagonal part corresponds to the advection operator $a\mathcal{L}_1$, where $\mathcal{L}_1 = \partial_x$. For future use we define

$$M_{\mathrm{G}} = \begin{pmatrix} D^{[0]} & 0 \\ 0 & D^{[1]} \end{pmatrix}, \quad N_{\mathrm{G}} = \begin{pmatrix} 0 & \delta^{[0]} \\ \delta^{[1]} & 0 \end{pmatrix}, \tag{2.4}$$

which are the matrices of these actions. Note that $A_{\mathrm{G}} = M_{\mathrm{G}} + aN_{\mathrm{G}}$.

2.2 Numerical results

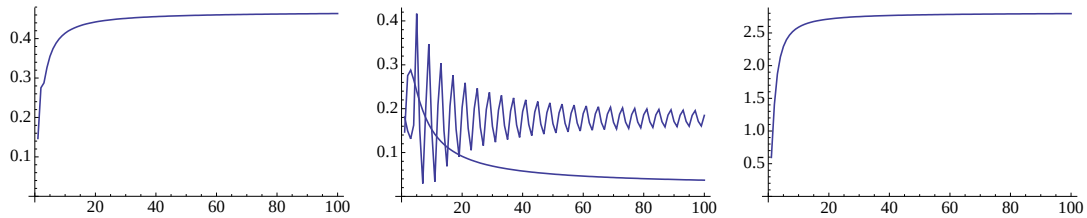
In Figure 1 we give numerical results for the modified Fourier–Galerkin method applied to the problem

$$-u_{xx}(x) + u_x(x) + 2u(x) = x^3 e^{3x}, \quad -1 \le x \le 1, \quad u_x(\pm 1) = 0. \tag{2.5}$$

In this and subsequent examples the modified Fourier coefficients of f are evaluated exactly by direct computation. For the moment Galerkin's equations are solved using the 'LinearSolve' routine in *Mathematica*. However, in the sequel we develop a specific algorithm to solve these equations.

Figure 1(a) displays the scaled error $N^3 E_N$, where

$$E_N = \max \left\{ |u(-1 + \tfrac{j}{N}) - u_N(-1 + \tfrac{j}{N})|, \quad j = 0, ..., 2N \right\}.$$



(a) (b) (c)

Figure 1: Error in the modified Fourier–Galerkin method applied to (2.5). (a): scaled error $N^3 E_N$ for $N = 1, ..., 100$, (b): scaled pointwise error $N^3|u(\frac{1}{2}) - u_N(\frac{1}{2})|$ (top), $N^3|u(1) - u_N(1)|$ (bottom), (c): scaled H^1 error $N^{\frac{5}{2}}\|u - u_N\|_1$.

Since this value remains bounded, it indicates that the uniform error $\|u-u_N\|_\infty$ is $\mathcal{O}\left(N^{-3}\right)$. Thus we observe the same convergence rate for u_N in the uniform and H^1 (see Figure 1(c)) norms as for approximation via $\mathcal{F}_N[u]$ (note that $u_N \neq \mathcal{F}_N[u]$ except when $a = 0$). However, as shown in Figure 1(b), the pointwise convergence rate for the Galerkin method inside the interval is also $\mathcal{O}\left(N^{-3}\right)$, unlike for direct function approximation, for which this value is $\mathcal{O}\left(N^{-4}\right)$.

In the next section we prove that these observations hold in general.

2.3 Analysis of convergence

Under the assumption of coercivity, existence and uniqueness of u_N are guaranteed by the Lax–Milgram Theorem, [15]. We now consider the question of convergence in various norms.

Lemma 3. *Suppose that $u \in H^r(-1, 1)$, $r = 1, 2, 3, 4$. Then we have the estimates*

$$\|u - u_N\|_1 \leq \frac{\gamma}{\omega}(N\pi)^{1-r}\|u\|_r, \quad r = 1, 2, 3, \quad \|u - u_N\|_1 \leq \frac{c_1\gamma}{\omega}(N\pi)^{-\frac{5}{2}}\|u\|_4, \quad r = 4,$$

where c_1 is the constant of Theorem 6.

Proof. From Céa's Lemma, [15], we immediately obtain

$$\|u - u_N\|_1 \leq \frac{\gamma}{\omega} \inf_{\phi \in \mathcal{S}_N} \|u - \phi\|_1.$$

By Theorem 1 this infimum is precisely $\|u - \mathcal{F}_N[u]\|_1$. Theorems 5 and 6 now give the result. □

Note that these estimates are the same as those for spectral-Galerkin methods based on Chebyshev or Legendre polynomials for $u \in H^r(-1, 1)$ where $r = 1, 2, 3$, [3]. When u has higher smoothness such methods offer a superior convergence rate.

These estimates also agree with the numerical results of the previous section. We may derive a bound for the uniform error as follows:

Theorem 9. *Suppose that $u \in H^1(-1, 1)$. Then we have the estimate*

$$\|u - u_N\|_\infty \leq \left(1 + 10|a|\omega^{-1}\right)\|u - \mathcal{F}_N[u]\|_\infty.$$

Proof. We have

$$T(u_N, \phi) = (f, \phi) = T(u, \phi) = T(\mathcal{F}_N[u], \phi) + T(u - \mathcal{F}_N[u], \phi), \quad \forall \phi \in \mathcal{S}_N.$$

If we define $e_N = u_N - \mathcal{F}_N[u]$, then

$$T(e_N, \phi) = T(u - \mathcal{F}_N[u], \phi), \quad \forall \phi \in \mathcal{S}_N.$$

However, since $\phi \in \mathcal{S}_N$ is an eigenfunction of the Laplace operator, we have

$$T(e_N, \phi) = a(u' - (\mathcal{F}_N[u])', \phi), \quad \forall \phi \in \mathcal{S}_N.$$

Setting $\phi = e_N$, applying the coercivity condition on the left hand side and integrating the right hand side by parts gives

$$\omega\|e_N\|_1^2 \leq a(u' - (\mathcal{F}_N[u])', e_N) = a\left[(u - \mathcal{F}_N[u])(1)e_N(1) - (u - \mathcal{F}_N[u])(-1)e_N(-1)\right] - a(u - \mathcal{F}_N[u], e_N').$$

We now use Sobolev's inequality to give

$$\omega\|e_N\|_1^2 \leq 4|a|c\|u - \mathcal{F}_N[u]\|_\infty\|e_N\|_1,$$

so that $\|e_N\|_1 \leq 4|a|c\omega^{-1}\|u - \mathcal{F}_N[u]\|_\infty$. Using the triangle inequality and Sobolev's inequality once more, we obtain

$$\|u - u_N\|_\infty \leq \|e_N\|_\infty + \|u - \mathcal{F}_N[u]\|_\infty \leq c\|e_N\|_1 + \|u - \mathcal{F}_N[u]\|_\infty \leq \left(4|a|c^2\omega^{-1} + 1\right)\|u - \mathcal{F}_N[u]\|_\infty,$$

Since $c^2 = \frac{5}{2}$ we obtain the result. □

Corollary 3. *Suppose that $u \in H^{r+1}(-1,1)$ for $r = 1, 2, 3$. Then*

$$\|u - u_N\|_\infty \leq c_r\|u\|_{r+1}N^{-r}, \quad r = 1, 2, 3,$$

where $c_r = 2c\pi^{-1}(1 + 10|a|\omega^{-1})r$.

Proof. This follows immediately from Theorems 7 and 9. □

The modified Fourier–Galerkin method is sufficiently simple to allow the coercivity condition $b - \frac{1}{4}a^2 > 0$ to be circumvented. We now show that existence, uniqueness and convergence (in particular the same rate of convergence) are maintained under the weaker assumption $b > 0$ irrespective of the value of a. We commence with the following lemma:

Lemma 4. *Suppose that we consider the operator $\mathcal{L}$ with Neumann boundary conditions and $b > 0$. Then*

$$(\mathcal{L}[u], \mathcal{L}[u]) \geq \omega\|u\|_2^2, \quad (\mathcal{L}[u], \mathcal{L}[v]) \leq \gamma\|u\|_2\|v\|_2, \quad \forall u, v \in H^2(-1,1), \tag{2.6}$$

for positive constants γ and ω.

Proof. For any $\epsilon > 0$ we have

$$\|\mathcal{L}[u]\|^2 = \|u_{xx}\|^2 + (a^2 + 2b)\|u_x\|^2 + b^2\|u\|^2 - 2ab(u_x, u) \geq \|u_{xx}\|^2 + (a^2 + 2b - \epsilon b)\|u_x\|^2 + b(b - \epsilon^{-1}a^2)\|u\|^2,$$

Since $b > 0$ we may set $\epsilon = 1 + b^{-1}a^2$ to get

$$\|\mathcal{L}[u]\|^2 \geq \min\{1, b, b^3(b + a^2)^{-1}\}\|u\|_2^2,$$

which gives the second inequality. The first inequality follows from the observation $\|\mathcal{L}[u]\| \leq \max\{1, |a|, b\}\|u\|_2$ for $u \in H^2(-1,1)$. □

Theorem 10. *Suppose that $b > 0$, $u \in H^2(-1,1)$ and $N > |a|\pi^{-1}\omega^{-\frac{1}{2}}$. Then the modified Fourier–Galerkin approximation exists and is unique. Furthermore we have the error estimate*

$$\|u - u_N\|_2 \leq c' \inf_{\phi \in \mathcal{S}_N} \|u - \phi\|_2, \tag{2.7}$$

where $c' = 1 + (1 - a^2\omega^{-1}(N\pi)^{-2})^{-1}\gamma\omega^{-1}$.

Proof. We defer proof of existence and uniqueness until Section 2.4. We now prove the estimate (2.7). Since $u \in H^2(-1,1)$ the modified Fourier–Galerkin approximation satisfies

$$(\mathcal{L}[u_N], \phi) = (f, \phi) = (\mathcal{L}[u], \phi), \quad \forall \phi \in \mathcal{S}_N. \tag{2.8}$$

Suppose that $\phi \in \mathcal{S}_N$. Using (2.6), we have

$$\|u_N - \phi\|_2^2 \leq \frac{1}{\omega}\|\mathcal{L}[u_N - \phi]\|^2 = \frac{1}{\omega}\left(\|\mathcal{F}_N[\mathcal{L}[u_N - \phi]]\|^2 + \|\mathcal{L}[u_N - \phi] - \mathcal{F}_N[\mathcal{L}[u_N - \phi]]\|^2\right).$$

However, for $\phi \in \mathcal{S}_N$, $\mathcal{L}[\phi] - \mathcal{F}_N[\mathcal{L}[\phi]] = a\,(\phi' - \mathcal{F}_N[\phi'])$. Furthermore, due to Theorem 5,

$$\|\phi' - \mathcal{F}_N[\phi']\|^2 \leq (N\pi)^{-2}\|\phi'\|_1^2 \leq (N\pi)^{-2}\|\phi\|_2^2.$$

Thus

$$\left(1 - a^2\omega^{-1}(N\pi)^{-2}\right)\|u_N - \phi\|_2^2 \leq \omega^{-1}\|\mathcal{F}_N[\mathcal{L}[u_N - \phi]]\|^2.$$

Now, by Corollary 1 and (2.8),

$$\|\mathcal{F}_N[\mathcal{L}[u_N - \phi]]\|^2 = \sum_{i=0}^{1}\sum_{n=0}^{N} c_n^{[i]}\left(\mathcal{L}[u_N - \phi], \phi_n^{[i]}\right)^2 = \sum_{i=0}^{1}\sum_{n=0}^{N} c_n^{[i]}\left(\mathcal{L}[u_N - \phi], \phi_n^{[i]}\right)\left(\mathcal{L}[u - \phi], \phi_n^{[i]}\right).$$

By the Cauchy–Schwarz inequality, Parseval's Theorem and the continuity condition (2.6) we obtain

$$\left(1 - a^2\omega^{-1}(N\pi)^{-2}\right)\|u_N - \phi\|_2^2 \leq \omega^{-1}\|\mathcal{L}[u_N - \phi]\|\|\mathcal{L}[u - \phi]\| \leq \gamma\omega^{-1}\|u_N - \phi\|_2\|u - \phi\|_2.$$

To derive the estimate (2.7) we use the triangle inequality $\|u_N - u\|_2 \leq \|u_N - \phi\|_2 + \|\phi - u\|_2$. □

Using (2.7) we may show that there is no deterioration of the H^1 and uniform error estimates under this weaker assumption:

Lemma 5. *For $b > 0$ and $N > |a|\gamma^{\frac{1}{2}}(\pi\omega)^{-1}$ we have*

$$\|u - u_N\|_1 \leq c''\|u - \mathcal{F}_N[u]\|_1,$$

where $c'' = 1 + (\omega - |a|\gamma^{\frac{1}{2}}(N\pi)^{-1})^{-1}|a|\gamma^{\frac{1}{2}}$.

Proof. In the standard manner, we obtain

$$(\mathcal{L}[e_N], \mathcal{F}_N[\mathcal{L}[e_N]]) = a(u' - (\mathcal{F}_N[u])', \mathcal{F}_N[\mathcal{L}[e_N]]).$$

For the left hand side, simple inequalities yield

$$\begin{aligned}(\mathcal{L}[e_N], \mathcal{L}[e_N]) - (\mathcal{L}[e_N], \mathcal{L}[e_N] - \mathcal{F}_N[\mathcal{L}[e_N]]) &\geq \omega\|e_N\|_2^2 - a(\mathcal{L}[e_N], e_N' - \mathcal{F}_N[e_N'])\\ &\geq \omega\|e_N\|_2^2 - |a|\gamma^{\frac{1}{2}}\|e_N\|_2\|e_N' - \mathcal{F}_N[e_N']\| \geq (\omega - |a|\gamma^{\frac{1}{2}}(N\pi)^{-1})\|e_N\|_2^2.\end{aligned}$$

For the right hand side, we have

$$a(u' - (\mathcal{F}_N[u])', \mathcal{F}_N[\mathcal{L}[e_N]]) \leq |a|\gamma^{\frac{1}{2}}\|u - \mathcal{F}_N[u]\|_1\|e_N\|_2,$$

which gives $\|e_N\|_2 \leq (\omega - |a|\gamma^{\frac{1}{2}}(N\pi)^{-1})^{-1}|a|\gamma^{\frac{1}{2}}\|u - \mathcal{F}_N[u]\|_1$, thus completing the proof. □

This shows that an estimate similar to that of Céa's Lemma is valid under this weaker assumption. We may also reproduce the results of Theorem 9 concerning the uniform convergence rate:

Lemma 6. *For $b > 0$ and $N > \max\{|a|\pi^{-1}\omega^{-\frac{1}{2}}, |a|\gamma^{\frac{1}{2}}(\pi\omega)^{-1}\}$ we have the estimate*

$$\|u - u_N\|_\infty \leq c''\sum_{i=0}^{1}\sum_{n>N}|\hat{u}_n^{[i]}|, \tag{2.9}$$

for some constant c'' dependent on a and b only.

Proof. Since u satisfies homogeneous Neumann boundary conditions, $(\mathcal{F}_N[u])'$ converges uniformly to u'. Hence we may write

$$e_N = a\sum_{i=0}^{1}\sum_{n>N}\hat{u}_n^{[i]}e_n^{[i]},$$

where $e_n^{[i]} \in \mathcal{S}_N$ satisfies

$$(\mathcal{L}[e_n^{[i]}], \phi) = ((\phi_n^{[i]})', \phi), \quad \forall\phi \in \mathcal{S}_N.$$

Now suppose that $v_n^{[i]}$ is the solution to the boundary value problem

$$\mathcal{L}[v_n^{[i]}] = (\phi_n^{[i]})', \quad (v_n^{[i]})'(\pm 1) = 0.$$

Using the previous Lemma, Sobolev's Inequality and Parseval's Theorem, we obtain

$$\|e_n^{[i]}\|_\infty \leq \|v_n^{[i]}\|_\infty + \|e_n^{[i]} - v_n^{[i]}\|_\infty \leq c\|v_n^{[i]}\|_1.$$

It therefore suffices to show that $\|v_n^{[i]}\|_1$ is uniformly bounded for all i and n. To do this we consider the solution of the boundary value problem

$$\mathcal{L}[v](x) = e^{i\omega x}, \quad v_x(\pm 1) = 0.$$

The solution may be expressed as

$$v(x) = \frac{1}{b + ai\omega + \omega^2} e^{i\omega x} + A_+(\omega)v_+(x) + A_-(\omega)v_-(x),$$

where $A_\pm$ enforce the boundary conditions. By direction computation $A_\pm(\omega) = \mathcal{O}\left(\omega^{-1}\right)$. Furthermore, the solutions $v_\pm$ are bounded independently of ω. Hence $\|v\|_k \leq c|\omega|^{k-1}$, $k \geq 1$, which gives the result. □

Evidently, using the coefficient bounds of Lemma 2, we obtain the same estimate for the uniform error as in Theorem 9.

2.4 Eigenanalysis of the method

The eigenvalue ratio of A_{G} and the L^2 condition number

$$\kappa(A_{\mathrm{G}}) = \sqrt{\frac{\lambda_{\max}(A_{\mathrm{G}}^\top A_{\mathrm{G}})}{\lambda_{\min}(A_{\mathrm{G}}^\top A_{\mathrm{G}})}}, \tag{2.10}$$

are of interest since they determine the impact of round-off errors in the solution of the linear system, [3]. It is well-known that these quantities are $\mathcal{O}(N^2)$ for the Fourier spectral method for second order linear operators, whereas they are generally $\mathcal{O}(N^4)$ for Chebyshev or Legendre spectral–Galerkin methods for non-periodic problems (in fact, the methods of Shen, [17, 18], possess $\mathcal{O}(N^2)$ condition numbers, however this is not true in general.). For the modified Fourier–Galerkin method we shall show that the condition number and eigenvalue ratio are both $\mathcal{O}(N^2)$, as in the Fourier case.

Lemma 7. *Suppose that the operator $\mathcal{L}$ is coercive. Then the eigenvalue ratio of the modified Fourier method is $\mathcal{O}(N^2)$, and each eigenvalue has positive real part bounded below by ω, the coercivity constant.*

Proof. Let λ be an eigenvalue of the method with corresponding eigenfunction $u \in \mathcal{S}_N$. Then

$$(\mathcal{L}[u], \phi) = \lambda(u, \phi), \quad \forall \phi \in \mathcal{S}_N.$$

Setting $\phi = u$ gives $\lambda\|u\|^2 = (\mathcal{L}[u], u)$. The lower bound follows directly from the coercivity condition. To derive an upper bound we use the continuity condition and Bernstein's inequality. □

We now move on to the condition number $\kappa(A_{\mathrm{G}})$:

Lemma 8. *Suppose that λ is an eigenvalue of $A_{\mathrm{G}}^\top A_{\mathrm{G}}$ with associated eigenfunction $u \in \mathcal{S}_N$. Then*

$$(\mathcal{F}_N[\mathcal{L}[u]], \mathcal{F}_N[\mathcal{L}[v]]) = \lambda(u, v), \quad \forall v \in \mathcal{S}_N.$$

Proof. Let $\{\phi_n\}$ be an orthonormal system. Then A_{G} has entries $(A_{\mathrm{G}})_{n,m} = (\mathcal{L}[\phi_m], \phi_n)$. If $\hat{a}$ is the eigenvector with coefficients (u, ϕ_n), then

$$\begin{aligned}\lambda(u, \phi_n) = \left((A_{\mathrm{G}}^\top A_{\mathrm{G}})\hat{a}\right)_n &= \sum_k \sum_m (\mathcal{L}[\phi_n], \phi_k)(\mathcal{L}[\phi_m]\hat{a}_m, \phi_k) \\ &= \left(\mathcal{L}[\phi_n], \sum_k (\mathcal{L}[u], \phi_k)\phi_k\right) = \left(\mathcal{F}_N[\mathcal{L}[\phi_n]], \mathcal{F}_N[\mathcal{L}[u]]\right).\end{aligned}$$

The linearity of $\mathcal{L}$ and the fact that u is real-valued give the result. □

Corollary 4. *The maximal eigenvalue of $A_{\mathrm{G}}^{\top}A_{\mathrm{G}}$ is $\mathcal{O}(N^4)$ for large N.*

Proof. We have $\lambda\|u\|^2 = \|\mathcal{F}_N[\mathcal{L}[u]]\|^2 \leq \|\mathcal{L}[u]\|^2$ by Parseval's theorem. However, $(u_{xx}, u_x) = (u_x, u_{xx}) = 0$ since u is real-valued and obeys homogeneous Neumann boundary conditions, and so

$$\|\mathcal{L}[u]\|^2 = \|u_{xx}\|^2 + (a^2+2b)\|u_x\|^2 + b^2\|u\|^2 - 2ab(u_x, u) \leq \|u_{xx}\|^2 + 2(a^2+b)\|u_x\|^2 + 2b^2\|u\|^2.$$

The second derivative term gives the largest contribution, and Bernstein's inequality yields the result. □

Corollary 5. *The minimal eigenvalue of $A_{\mathrm{G}}^{\top}A_{\mathrm{G}}$ is bounded below independently of N provided $b > 0$ and $N > |a|\pi^{-1}\omega^{-\frac{1}{2}}$.*

Proof. We have

$$\|\mathcal{F}_N[\mathcal{L}[u]]\|^2 = \|\mathcal{L}[u]\|^2 - \|\mathcal{L}[u] - \mathcal{F}_N[\mathcal{L}[u]]\|^2 = \|\mathcal{L}[u]\|^2 - a^2\|u' - (\mathcal{F}_N[u])'\|^2.$$

Consequently, we obtain the bound

$$\|\mathcal{F}_N[\mathcal{L}[u]]\|^2 \geq \|\mathcal{L}[u]\|^2 - a^2(N\pi)^{-2}\|u_{xx}\|^2 \geq \left(\omega - a^2(N\pi)^{-2}\right)\|u\|_2^2,$$

as required. □

By simple arguments, this bound for the minimal eigenvalue of $A_{\mathrm{G}}^{\top}A_{\mathrm{G}}$ is also lower bound for the minimal eigenvalue of A_{G}. We conclude:

Corollary 6. *Galerkin's equations have a unique solution for $b > 0$ irrespective of a, provided $N > |a|\pi^{-1}\omega^{-\frac{1}{2}}$. Furthermore, the eigenvalue ratio and condition number of the matrix A_{G} are $\mathcal{O}(N^2)$.*

This also completes the proof of existence and uniqueness of the modified Fourier–Galerkin approximation under the weaker assumption $b > 0$ and $N > |a|\pi^{-1}\omega^{-\frac{1}{2}}$ (see Theorem 10).

2.5 Optimal preconditioning of Galerkin's equations

The modified Fourier–Galerkin matrix admits an optimal, right preconditioner, namely the matrix M_{G}^{-1}:

Theorem 11. *The right preconditioner M_{G}^{-1} is optimal for the eigenvalue ratio, i.e. the eigenvalue ratio of the preconditioned matrix $A_{\mathrm{G}}M_{\mathrm{G}}^{-1}$ is $\mathcal{O}(1)$, for all N provided the operator is coercive. Moreover, if $N > |a|\pi^{-1}\omega^{-\frac{1}{2}}$, under the weaker assumption $b > 0$ it is optimal for both the condition number and eigenvalue ratio.*

Proof. We commence with the eigenvalue ratio in the coercive case. Suppose that $A_{\mathrm{G}}M_{\mathrm{G}}^{-1}\hat{x} = \lambda\hat{x}$. Then, setting $\hat{x} = M_{\mathrm{G}}\hat{y}$ we obtain $A_{\mathrm{G}}\hat{y} = \lambda M_{\mathrm{G}}\hat{y}$. If $\hat{x}$ corresponds to $u \in \mathcal{S}_N$, then $\hat{y}$ corresponds to $v \in \mathcal{S}_N$, where $u = \mathcal{L}_0[v]$ (note that $\mathcal{L}_0[v] \in \mathcal{S}_N$ for $v \in \mathcal{S}_N$). In particular, we have

$$(\mathcal{L}[v], v) = \lambda\left(\mathcal{L}_0[v], v\right),$$

and using the continuity and coercivity conditions we obtain

$$\frac{\gamma}{\min\{b,1\}} \geq \lambda \geq \frac{\omega}{\max\{b,1\}} > 0,$$

which gives the result.

Now consider the L^2 condition number. If λ is an eigenvalue of $(A_{\mathrm{G}}M_{\mathrm{G}}^{-1})^{\top}A_{\mathrm{G}}M_{\mathrm{G}}^{-1}$ with eigenvector $\hat{x}$ corresponding $u \in \mathcal{S}_N$ then

$$M_{\mathrm{G}}^{-1}A_{\mathrm{G}}^{\top}A_{\mathrm{G}}M_{\mathrm{G}}^{-1}\hat{x} = \lambda\hat{x}.$$

Setting $\hat{x} = M_{\mathrm{G}}\hat{y}$ as before we obtain

$$(A_{\mathrm{G}}\hat{y})^{\top}(A_{\mathrm{G}}\hat{y}) = \lambda(M_{\mathrm{G}}\hat{y})^{\top}(M_{\mathrm{G}}\hat{y}),$$

which is equivalent to

$$(\mathcal{F}_N[\mathcal{L}[v]], \mathcal{F}_N[\mathcal{L}[v]]) = \lambda\,(\mathcal{L}_0[v], \mathcal{L}_0[v])\,.$$

Now, for $N > |a|\pi^{-1}\omega^{-\frac{1}{2}}$ and $b > 0$ we have the inequalities

$$\omega''\|v\|_2^2 \leq \|\mathcal{F}_N[\mathcal{L}[v]]\|^2 \leq \gamma''\|v\|_2^2,$$

for some positive constants ω'', γ''. Furthermore, it is easy to deduce that

$$\omega'''\|v\|_2^2 \leq \|\mathcal{L}_0[v]\|^2 \leq \gamma'''\|v\|_2^2$$

for suitable positive γ''' and ω'''. Hence λ is $\mathcal{O}\,(1)$. As before, this result also gives bounds for the maximal and minimal eigenvalues of $A_{\mathrm{G}}M_{\mathrm{G}}^{-1}$, which completes the proof. □

Evidently the preconditioner M_{G}^{-1} is diagonal, so it is simple and cheap to apply.

2.6 Iterative solution of Galerkin's equations

The linear systems arising in modified Fourier–Galerkin discretizations can be solved efficiently using the most basic of iterative procedures. This is based on the decomposition $A_{\mathrm{G}} = M_{\mathrm{G}} + aN_{\mathrm{G}}$, with matrices M_{G} and N_{G} defined in (2.4).

The iterative procedure to solve $A_{\mathrm{G}}\hat{a} = \hat{f}$ is

$$M_{\mathrm{G}}\hat{a}^{k+1} = -aN_{\mathrm{G}}\hat{a}^k + \hat{f}, \tag{2.11}$$

which converges for all choices of $\hat{a}^0$ and $\hat{f}$ if and only if the spectral radius $\rho(aM_{\mathrm{G}}^{-1}N_{\mathrm{G}}) < 1$, [4]. For modified Fourier series this is indeed the case:

Lemma 9. *The spectral radius $\rho(aM_{\mathrm{G}}^{-1}N_{\mathrm{G}})$ is bounded above by $\sqrt{a^2/(4b)}$.*

Proof. Suppose that λ is an eigenvalue of $aM_{\mathrm{G}}^{-1}N_{\mathrm{G}}$ with corresponding normalized eigenfunction u. Then

$$a(u_x, \phi) = \lambda(bu - u_{xx}, \phi), \quad \forall \phi \in \mathcal{S}_N.$$

Thus

$$|\lambda| \leq |a|\frac{|(u_x, u)|}{b + \|u_x\|^2} \leq |a|\frac{\|u_x\|}{b + \|u_x\|^2} \leq \sqrt{\frac{a^2}{4b}},$$

where the last inequality follows by noting that the function $g(y) = y(c+y^2)^{-1}$, $c > 0$, has a unique maximum in $[0, \infty)$ at $y = \sqrt{c}$ and takes value $(2\sqrt{c})^{-1}$ there. □

We see that, provided the operator $\mathcal{L}$ is coercive, the iteration (2.11) converges. Further the bound derived is independent of N, meaning that the number of iterations required to reduce the error $\|\hat{a}^k - \hat{a}\|$ to below a prescribed tolerance is also independent of N. Thus, since M_{G} is diagonal, the operational cost of the scheme is determined by the number of operations required to evaluate $N_{\mathrm{G}}\hat{a}^k$; in other words $\mathcal{O}(N^2)$.

We may reduce this figure to $\mathcal{O}(N \log N)$ by using a version of the FFT to evaluate the matrix-vector multiplications. We include this mainly for interest. Because of the hyperbolic cross, the FFT cannot be easily used for modified Fourier approximations in the d-variate cube. However, it turns out that the operational cost of direct evaluation remains $\mathcal{O}\left(N^2\right)$ for all d, [1].

Galerkin's equations arising from Chebyshev polynomials can be solved at best in $\mathcal{O}(N \log N)$ operations using the FFT. For methods based on Legendre polynomials this quantity is $\mathcal{O}(N^2)$. Hence this method is no more expensive than other techniques for $d = 1$. For $d \geq 2$ it becomes increasingly cheaper, [1].

The iterative scheme presented is unsuitable when the operator $\mathcal{L}$ is not coercive, however in that case an alternative approach based on pre-multiplication by $A_{\mathrm{G}}^{\top}$ can be used.

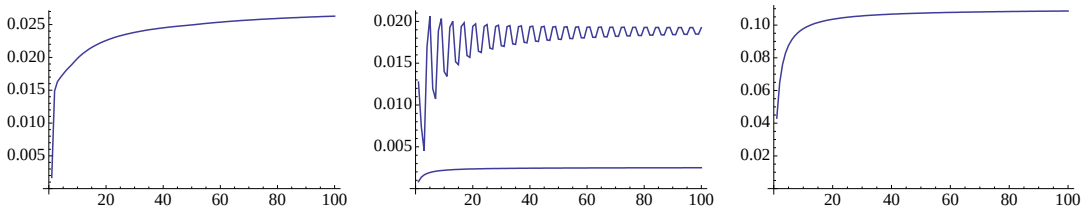


Figure 2: Error in the modified Fourier–Galerkin method applied to the problem with parameters $a(x) = x$, $b(x) = e^{-x}$ and exact solution (3.2). (a): scaled error $N^3 E_N$ for $N = 1, ..., 100$, (b): scaled pointwise error $N^3|u(\frac{1}{2}) - u_N(\frac{1}{2})|$ (top), $N^3|u(1) - u_N(1)|$ (bottom), (c): scaled H^1 error $N^{\frac{5}{2}}\|u - u_N\|_1$.

3 Variable coefficient problems

We may apply the modified Fourier–Galerkin method to problems where $a = a(x)$ and $b = b(x)$ are given functions of sufficient smoothness. Note that in this case the operator $\mathcal{L}$ is coercive provided

$$\min_{x\in[-1,1]}\{b(x)\} - \frac{1}{4}\max_{x\in[-1,1]}\{a(x)\} > 0. \tag{3.1}$$

To formulate Galerkin's equations we need to evaluate the integrals

$$\int_{-1}^{1} b(x)u_N(x)\phi_n^{[i]}(x)\,dx = \sum_{j=0}^{1}\sum_{m=0}^{N} c_m^{[j]}a_m^{[j]}\int_{-1}^{1} b(x)\phi_m^{[j]}(x)\phi_n^{[i]}(x)\,dx, \quad i = 0,1, \quad n = i, ..., N,$$

$$\int_{-1}^{1} a(x)\partial_x u_N(x)\phi_n^{[i]}(x)\,dx = \sum_{j=0}^{1}\sum_{m=0}^{N} c_m^{[j]}a_m^{[j]}\int_{-1}^{1} a(x)\phi_n^{[i]}(x)\partial_x\phi_m^{[j]}(x)\,dx, \quad i = 0,1, \quad n = i, ..., N.$$

For $i, j = 0, 1$, $n = i, ..., N$ and $m = j, ..., N$ we consider

$$B_{n,m}^{[i,j]} = c_m^{[j]}\int_{-1}^{1} b(x)\phi_m^{[j]}(x)\phi_n^{[i]}(x)\,dx, \quad C_{n,m}^{[i,j]} = c_m^{[j]}\int_{-1}^{1} a(x)\phi_n^{[i]}(x)\partial_x\phi_m^{[j]}(x)\,dx.$$

After simple manipulations we obtain:

$$B_{n,m}^{[0,0]} = \frac{c_m^{[0]}}{2}\left(\hat{b}_{n+m}^{[0]} + \hat{b}_{n-m}^{[0]}\right), \quad B_{n,m}^{[0,1]} = \frac{c_m^{[1]}}{2}\left(\hat{b}_{n+m}^{[1]} + \hat{b}_{m-n}^{[1]}\right)$$

$$B_{n,m}^{[1,0]} = \frac{c_m^{[0]}}{2}\left(\hat{b}_{n+m}^{[1]} + \hat{b}_{n-m}^{[1]}\right), \quad B_{n,m}^{[1,1]} = \frac{c_m^{[1]}}{2}\left(\hat{b}_{n-m}^{[0]} - \hat{b}_{n+m-1}^{[0]}\right),$$

$$C_{n,m}^{[0,0]} = -\frac{c_m^{[0]}m\pi}{2}\left(\check{a}_{n+m}^{[1]} + \check{a}_{m-n}^{[1]}\right), \quad C_{n,m}^{[0,1]} = \frac{c_m^{[1]}(m-\frac{1}{2})\pi}{2}\left(\check{a}_{m+n}^{[0]} + \check{a}_{m-n}^{[0]}\right),$$

$$C_{n,m}^{[1,0]} = -\frac{c_m^{[0]}m\pi}{2}\left(\check{a}_{n-m}^{[0]} - \check{a}_{n+m}^{[0]}\right), \quad C_{n,m}^{[1,1]} = \frac{c_m^{[1]}(m-\frac{1}{2})\pi}{2}\left(\check{a}_{n+m-1}^{[1]} + \check{a}_{n-m}^{[1]}\right),$$

where $\check{a}_n^{[i]}$ is the Laplace–Dirichlet coefficient of the function a. If we define the matrices

$$D_{\mathrm{G}} = \begin{pmatrix} D^{[0]} & 0 \\ 0 & D^{[1]} \end{pmatrix}, \quad B_{\mathrm{G}} = \begin{pmatrix} B^{[0,0]} & B^{[0,1]} \\ B^{[1,0]} & B^{[1,1]} \end{pmatrix}, \quad C_{\mathrm{G}} = \begin{pmatrix} C^{[0,0]} & C^{[0,1]} \\ C^{[1,0]} & C^{[1,1]} \end{pmatrix},$$

where $D^{[i]}$ is the diagonal matrix with entries $\mu_n^{[i]}$, then the discretization matrix is $A_{\mathrm{G}} = D_{\mathrm{G}} + B_{\mathrm{G}} + C_{\mathrm{G}}$.

In Figure 2 we consider the modified Fourier–Galerkin approximation to the problem with parameters $a(x) = x$, $b(x) = e^{-x}$, exact solution

$$u(x) = \cos 3x - \frac{1}{2}x + \frac{3\sin 3}{2}x^2, \quad -1 \le x \le 1, \tag{3.2}$$

and f given accordingly.

In the coercive case the estimates of Section 2.3 regarding the convergence rate can easily be extended to variable coefficient problems. Hence we obtain a uniform error estimate of $\mathcal{O}\left(N^{-3}\right)$ for example, which confirms the numerical results of Figure 2. The estimates in the case where $\mathcal{L}$ satisfies

$$\|\mathcal{L}[u]\|^2 \geq \omega \|u\|_2^2, \quad \forall u \in H^2(-1,1),$$

can also be extended. Unfortunately, it is not immediately obvious whether this presents a more or less restrictive condition than the variable coefficient coercivity condition (3.1). Under this assumption, however, we also deduce the same $\mathcal{O}\left(N^2\right)$ estimate for the condition number.

Efficient solution of Galerkin's equations can be achieved in $\mathcal{O}\left(N^2\right)$ (or $\mathcal{O}\left(N \log N\right)$ using the FFT) operations as follows. If we define the parameter $b_0 = \max_x b(x)$ and decompose the matrix $A_{\mathrm{G}} = M_{\mathrm{G}} + N_{\mathrm{G}}$, where M_{G} corresponds to the action of the operator $-\partial_x^2 + b_0 \imath$ on $\mathcal{S}_N$ and N_{G} corresponds to $(b - b_0)\imath + a\partial_x$, then the resulting iteration scheme based on this splitting converges.

At each stage we need to evaluate matrix-vector products involving N_{G}. This is equivalent to finding the first N modified Fourier coefficients of products and derivatives of modified Fourier sums. This can be done in $\mathcal{O}\left(N \log N\right)$ operations using the FFT.

4 Accelerating Convergence

The method presented does not possess spectral accuracy. In this section we develop and analyse a technique that exhibits more rapid convergence.

The question of convergence acceleration for modified Fourier series has been addressed in [6]. The approach uses a familiar device of Fourier series, namely polynomial subtraction, [16], which is the following. If f is decomposed as $(f - p) + p$, where p interpolates the first k odd derivatives of f at the endpoints, then we form the modified Fourier series $\mathcal{F}_N[f - p]$ and approximate f by $\mathcal{F}_N[f - p] + p$. Since $f - p$ obeys the first k derivative conditions, the faster convergence of this approximation is guaranteed by the results of Section 1.2.

The method we now present is based on this device.

4.1 A method with increased convergence

We commence with the following observation. If a solution u to $\mathcal{L}[u] = f$ with the boundary conditions $u'(\pm 1) = 0$ also satisfies $u \in H^{2M+4}(-1,1)$ and $u^{(2k+1)}(\pm 1) = 0$ for $k = 1, ..., M$ then the H^1 and uniform errors in the modified Fourier–Galerkin approximation are $\mathcal{O}(N^{-\frac{5}{2}-2M})$ and $\mathcal{O}(N^{-3-2M})$ respectively. This follows immediately from the Céa's Lemma and Theorems 6, 7 and 9.

Suppose now that u does not have this property. In the trivial case $a = 0$ these derivatives are known explicitly: we merely differentiate the relation $\mathcal{L}[u] = f$ to obtain

$$u^{(2k+1)}(\pm 1) = -\sum_{j=0}^{k} b^j f^{(2(k-j)-1)}(\pm 1).$$

In this case we interpolate these values with a polynomial p and apply the modified Fourier–Galerkin method to the 'smoothed' problem $\mathcal{L}[v] = g$, where $v = u - p$ and $g = f - \mathcal{L}[p]$.

Unfortunately, in the general case, these boundary values are not known explicitly. However the equation $\mathcal{L}[u] = f$ gives a two-term recurrence relation for the values $u^{(k)}(\pm 1)$ in terms of f and its derivatives:

$$-u^{(2k+2)}(\pm 1) + a u^{(2k+1)}(\pm 1) + b u^{(2k)}(\pm 1) = f^{(2k)}(\pm 1), \quad k = 0, 1, ...$$

This yields a relation between $u^{(2k+1)}(\pm 1)$, $k = 1, 2, ...$, and $u(\pm 1)$. For $k = 1$ we have

$$u^{(3)}(\pm 1) = abu(\pm 1) - \left[af(\pm 1) + f'(\pm 1)\right],$$

and in general

$$u^{(2k+1)}(\pm 1) = c_k u(\pm 1) - F_k(\pm 1), \quad k = 1, 2, ..., \tag{4.1}$$

where c_k is a constant depending on a and b and F_k is a function of f and its first $2k-1$ derivatives.

To design a new method we now seek to enforce a finite number of these conditions. We proceed as follows. Suppose that p is a function that obeys the boundary conditions. Letting $u = v + p$, we have

$$\mathcal{L}[v] + \mathcal{L}[p] = f, \tag{4.2}$$

and we augment this system with the constraints

$$v^{(2k+1)}(\pm 1) = 0, \quad k = 1, ..., M, \quad \Leftrightarrow \quad c_k u(\pm 1) - p^{(2k+1)}(\pm 1) = F_k(\pm 1), \quad k = 1, ..., M. \tag{4.3}$$

Seeking an approximation $v_N \in \mathcal{S}_N$ to v that satisfies Galerkin's equations for the auxiliary problem (4.2), we introduce $2M$ unknowns $b_0, ..., b_{2M-1}$ and let

$$p(x) = \sum_{r=0}^{2M-1} b_r p_r(x), \tag{4.4}$$

where p_r, $r = 0, ..., 2M-1$, are known functions that satisfy the boundary conditions. Our approximation u_N to u is of the form

$$u_N(x) = p(x) + v_N(x) = \sum_{r=0}^{2M-1} b_r p_r(x) + \sum_{n=0}^{N} \sum_{i=0}^{1} c_n^{[i]} a_n^{[i]} \phi_n^{[i]}(x), \tag{4.5}$$

and satisfies (4.2)–(4.3) approximately:

$$\begin{aligned} (\mathcal{L}[u_N], \phi) &= (f, \phi), \quad \forall \phi \in \mathcal{S}_N, \\ c_k u_N(\pm 1) - u_N^{(2k+1)}(\pm 1) &= F_k(\pm 1), \quad k = 1, ..., M. \end{aligned} \tag{4.6}$$

This is a linear system for the coefficients b_k, $a_n^{[i]}$ which may be expressed in matrix form as

$$\begin{pmatrix} A_{\mathrm{G}} & \widehat{\mathcal{L}[p_0]} & ... & \widehat{\mathcal{L}[p_{2M-1}]} \\ c_1 C & Q_1[p_0] & ... & Q_1[p_{2M-1}] \\ ... & ... & ... & ... \\ c_M C & Q_M[p_0] & ... & Q_M[p_{2M-1}] \end{pmatrix} \begin{pmatrix} a^{[0]} \\ a^{[1]} \\ b_0 \\ \vdots \\ b_{2M-1} \end{pmatrix} = \begin{pmatrix} \hat{f}^{[0]} \\ \hat{f}^{[1]} \\ f_1 \\ \vdots \\ f_M \end{pmatrix}, \tag{4.7}$$

where $\widehat{\mathcal{L}[p_r]}$ is the vector of modified Fourier coefficients of $\mathcal{L}[p_r]$, C is the $2 \times (2N+1)$ matrix

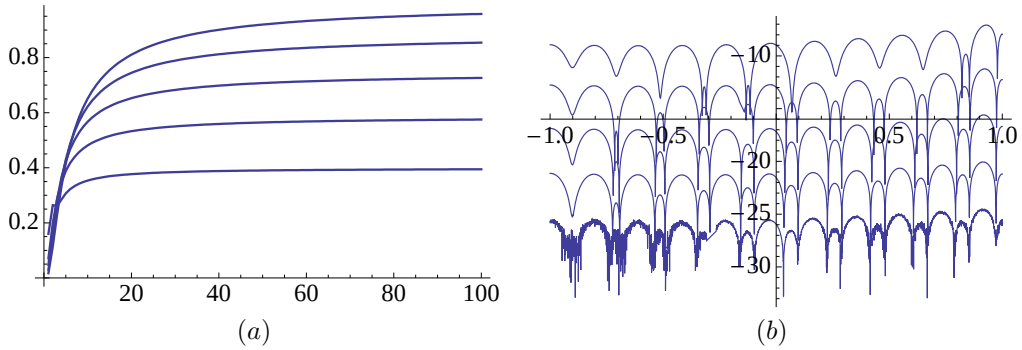
$$C = \begin{pmatrix} \frac{1}{2} & -1 & 1 & ... & (-1)^N & 1 & -1 & ... & (-1)^{N+1} \\ \frac{1}{2} & -1 & 1 & ... & (-1)^N & -1 & 1 & ... & (-1)^N \end{pmatrix},$$

and $Q_k[p_r]$, $f_k \in \mathbb{R}^2$ are the vectors

$$Q_k[p_r] = \begin{pmatrix} c_k p_r(1) - p_r^{(2k+1)}(1) \\ c_k p_r(-1) - p_r^{(2k+1)}(-1) \end{pmatrix}, \quad f_k = \begin{pmatrix} F_k(1) \\ F_k(-1) \end{pmatrix}.$$

One viewpoint of this method is that it forces the modified Fourier coefficients of the residual $\mathcal{L}[u_N] - f$ to decay at an increased rate. However the higher boundary values $u^{(2k+1)}(\pm 1)$ are written in terms of $u(\pm 1)$ to stop the system becoming ill-conditioned for large N. In fact, from the matrix form (4.7), we see that the n^{th} element in the additional rows or columns is $\mathcal{O}(1)$ (provided the functions p_r are chosen sensibly). Hence we would expect, and it turns out to be the case, that the eigenvalue ratio would remain $\mathcal{O}(N^2)$. If we did not write $u^{(2k+1)}(\pm 1)$ in this manner then the n^{th} entry in the $2k^{\text{th}}$ additional row would be $\mathcal{O}(n^{2k})$, leading to very poor conditioning.

Figure 3 gives numerical results for this method with $M = 0, 1, 2, 3, 4$ applied to the problem (2.5), where $M = 0$ refers to the original Galerkin method. As we observe, the uniform error in this example is



(a) (b)

Figure 3: Error in the method (4.6) applied to the problem (2.5). (a): Scaled error $N^{2M+3}E_N$ for $M = 0, 1, 2, 3, 4$ from bottom to top. (b) Log error $\log_e(u(x) - u_{10}(x))$, $-1 \le x \le 1$, for $M = 0, 1, 2, 3, 4$ from top to bottom.

$\mathcal{O}\left(N^{-2M-3}\right)$. As before, these correspond to the known results from function approximation with modified Fourier series and polynomial subtraction. Figure 3(b) gives a comparison for the log error $\log_e(u(x)-u_{10}(x))$ for $M = 0, 1, 2, 3, 4$. Using $M = 4$ and $N = 10$ we are able to obtain numerical precision.

Note that in these examples we use the additional basis functions

$$p_r(x) = \frac{1}{r+3}x^{r+3} - \frac{1}{r+1}x^{r+1}, \quad r = 0, 1, ..., 2M.$$

4.2 Analysis of the method

For the remainder of this paper we use the notation c for a positive constant, independent of the parameters unless specified otherwise. In this section we analyse this method. To this end we rewrite (4.6) as

$$\text{find} \quad u_N \in X_N : \quad (\mathcal{L}[u_N], \phi) = (f, \phi), \quad \forall \phi \in \mathcal{S}_N, \tag{4.8}$$

where $Z_N = \text{span}\{p_0, ..., p_{2M-1}\} + \mathcal{S}_N$ and

$$X_N = \{v \in Z_N : c_k v(\pm 1) - v^{(2k+1)}(\pm 1) = F_k(\pm 1),\ k = 1, ..., M\}. \tag{4.9}$$

It is tempting to consider the method as a type of Petrov–Galerkin method, [3, 15], and use standard results therein. Unfortunately this is not the case (X_N is not a linear space). Nonetheless, under the assumption of coercivity of the operator $\mathcal{L}$, we have:

Lemma 10. *Suppose that $p_0, ..., p_{2M-1}$ are chosen so that $\|p_r\|_4$ is bounded independently of N for all r and so that the interpolation problem*

$$\textit{find} \quad b_r : \quad \sum_{r=0}^{2M-1} b_r p_r^{(2k+1)}(\pm 1) = \theta_{k,\pm}, \quad k = 1, ..., M,$$

has a unique solution for all $\theta_{k,\pm} \in \mathbb{R}$. Then, provided N is sufficiently large, there is a unique solution u_N to (4.8).

Proof. We shall show that the minimal eigenvalue of the discretization matrix is bounded away from zero for sufficiently large N. If λ is the minimal eigenvalue with eigenfunction $u_N \in X_N$ which has coefficients $a_n^{[i]}$ and b_k, then

$$\begin{aligned}
(\mathcal{L}[u_N], \phi) &= \lambda(v_N, \phi), \quad \forall \phi \in \mathcal{S}_N, \\
c_{k+1}u_N(1) - u_N^{(2k+3)}(1) &= \lambda b_{2k}, \quad k = 0, ..., M-1, \\
c_{k+1}u_N(-1) - u_N^{(2k+3)}(-1) &= \lambda b_{2k+1}, \quad k = 0, ..., M-1,
\end{aligned}$$

where $u_N = p + v_N$. In particular, since $v_N \in \mathcal{S}_N$,

$$\lambda(v_N, \mathcal{F}_N[u_N]) = (\mathcal{L}[u_N], \mathcal{F}_N[u_N]) = (\mathcal{L}[u_N], u_N) - (\mathcal{L}[u_N], p - \mathcal{F}_N[p]). \tag{4.10}$$

Now suppose that $\lambda \to 0$ as $N \to \infty$. The second set of equations, written in matrix form, is

$$(\lambda I + P)\hat{b} = \{c_k u_N(\pm 1)\}_{k=1}^{M},$$

where P is the $2M \times 2M$ non-singular matrix with entries $P_{k,r}$ given by

$$P_{2k,r} = p_r^{(2k+3)}(1), \quad P_{2k+1,r} = p_r^{(2k+3)}(-1), \quad k = 0, ..., M-1, \quad r = 0, ..., 2M-1,$$

and I is the $2M \times 2M$ identity matrix. For sufficiently large N the matrix $\lambda I + P$ is invertible with bounded inverse. Hence, after an application of Sobolev's inequality, we obtain

$$|b_k| \leq c\|u_N\|_1, \quad k = 0, ..., 2M-1.$$

Since $p = \sum b_k p_k$, under the assumptions on each p_k Theorem 6 gives

$$\|p - \mathcal{F}_N[p]\|_1 \leq c\|u_N\|_1 N^{-\frac{5}{2}}.$$

Returning to (4.10), using the continuity and coercivity conditions we obtain

$$|\lambda||(v_N, \mathcal{F}_N[u_N])| \geq \omega\|u_N\|_1^2 - \gamma\|u_N\|_1\|p - \mathcal{F}_N[p]\|_1 \geq (\omega - cN^{-\frac{5}{2}})\|u_N\|_1^2.$$

We also have

$$|(v_N, \mathcal{F}_N[u_N])| = |\|\mathcal{F}_N[u_N]\|^2 - (p, \mathcal{F}_N[u_N])| \leq \|u_N\|^2 + \|p\|\|u_N\|_1 \leq c\|u_N\|_1^2,$$

where the final inequality follows from the bound $\|p\| \leq c\|u_N\|_1$. Thus we obtain

$$c\|u_N\|_1^2|\lambda| \geq \|u_N\|_1^2(\omega - cN^{-\frac{5}{2}}).$$

so that λ is bounded away from zero, giving a contradiction. □

Immediately we deduce the following:

Corollary 7. *The eigenvalue ratio of the discretization matrix is $\mathcal{O}\left(N^2\right)$ for sufficiently large N.*

Proof. For sufficiently large N the minimal eigenvalue of the discretization matrix is bounded away from 0. To bound the maximal eigenvalue we use Gerschgorin's Theorem and the matrix form (4.7). Since maximal diagonal entry is $\mathcal{O}(N^2)$ and the $(n,m)^{\text{th}}$ entry, $n \neq m$, is $\mathcal{O}(m)$, we obtain the result. □

The method is also stable in the following sense:

Corollary 8. *Given $\epsilon > 0$, $\|u_N\|_1 \leq c\|f\| + \epsilon\|f\|_{2k}$ for sufficiently large N.*

Proof. Setting $\phi = \mathcal{F}_N[u_N]$ in (4.8), and using the continuity and coercivity conditions we obtain

$$\omega\|u_N\|_1^2 - \gamma\|u_N\|_1\|u_N - \mathcal{F}_N[u_N]\|_1 \leq \|f\|\|u_N\|_1. \tag{4.11}$$

We have

$$u_N^{(2k+1)}(\pm 1) = c_k u_N(\pm 1) - F_k(\pm 1), \quad k = 1, ..., M.$$

Hence, if u_N has coefficients $a_n^{[i]}$ and b_k, then $|b_k| \leq c(\|u_N\|_1 + \|f\|_{2k})$, which means that

$$\|u_N - \mathcal{F}_N[u_N]\|_1 \leq cN^{-\frac{5}{2}}\left(\|u_N\|_1 + \|f\|_{2k}\right).$$

Rearranging (4.11) now gives the result. □

With existence and uniqueness in hand, we turn our attention to providing an error estimate. To do so we first need to introduce an appropriate projection map $\mathcal{H}_N : H^{2M+4}(-1,1) \to X_N$. We define such a map by enforcing the condition

$$\mathcal{F}_N[\mathcal{H}_N[u]] = \mathcal{F}_N[u]. \tag{4.12}$$

Since $\mathcal{H}_N[u] \in X_N$ this uniquely defines $\mathcal{H}_N$, given the assumptions of Lemma 10:

Lemma 11. *We have*

$$\mathcal{H}_N[u](x) = \mathcal{F}_N\left[u - \sum_{r=0}^{2M-1} b_r p_r\right](x) + \sum_{r=0}^{2M-1} b_r p_r(x), \tag{4.13}$$

where the coefficients b_r are given by the equation $P^N \hat{b} = \hat{U}$, where $\hat{U} \in \mathbb{R}^{2M}$ has entries

$$\hat{U}_{2k} = u^{(2k+3)}(1) - c_{k+1}(u - \mathcal{F}_N[u])(1), \quad \hat{U}_{2k+1} = u^{(2k+3)}(-1) - c_{k+1}(u - \mathcal{F}_N[u])(-1), \quad k = 0, ..., M-1,$$

and P^N is the $2M \times 2M$ matrix with entries

$$P^N_{r,2k} = p_r^{(2k+3)}(1) - c_{k+1}(p_r(1) - \mathcal{F}_N[p_r](1)), \quad k = 0, ..., M-1,$$
$$P^N_{r,2k+1} = p_r^{(2k+3)}(-1) - c_{k+1}(p_r(-1) - \mathcal{F}_N[p_r](-1)), \quad k = 0, ..., M-1.$$

In particular $\mathcal{H}_N[u]$ is uniquely defined for sufficiently large N provided the functions p_r satisfy the assumptions of Lemma 10.

Proof. Suppose that we write

$$\mathcal{H}_N[u](x) = \sum_{r=0}^{2M-1} b_r p_r(x) + \sum_{i=0}^{1} \sum_{n=0}^{N} c_n^{[i]} a_n^{[i]} \phi_n^{[i]}(x).$$

Then, from (4.12) we have

$$a_n^{[i]} = \hat{u}_n^{[i]} - \sum_{r=0}^{2M-1} b_r \widehat{p_r}_n^{[i]},$$

and rearranging we obtain (4.13). Since $\mathcal{H}_N[u] \in X_N$ we observe that

$$c_k \mathcal{H}_N[u](\pm 1) - (\mathcal{H}_N[u])^{(2k+1)}(\pm 1) = F_k(\pm 1), \quad k = 1, ..., M.$$

However by definition of F_k, the right hand side is just $c_k u(\pm 1) - u^{(2k+1)}(\pm 1)$. Further, since $\mathcal{H}_N[u]$ is a sum of the functions p_r and a modified Fourier series of degree N, we obtain

$$c_k \mathcal{H}_N[u](\pm 1) - \sum_{r=0}^{2M-1} b_r p_r^{(2k+1)}(\pm 1) = c_k u(\pm 1) - u^{(2k+1)}(\pm 1).$$

Rearranging and using (4.13) gives the result. Note that uniqueness follows from the non-singularity of P since $P^N = P + \mathcal{O}\left(N^{-3}\right)$. □

Lemma 12. *Suppose that the conditions of Lemma 10 hold. Then, for sufficiently large N, the method admits the following error estimate:*

$$\|u - u_N\|_1 \leq c\|u - \mathcal{H}_N[u]\|_1. \tag{4.14}$$

Proof. We first note that $(\mathcal{L}[u_N], \phi) = (\mathcal{L}[u], \phi)$ for all $\phi \in \mathcal{S}_N$. Setting $\phi = \mathcal{F}_N[u_N - \mathcal{H}_N[u]]$ in (4.8), we obtain

$$(\mathcal{L}[u_N - \mathcal{H}_N[u]], \mathcal{F}_N[u_N - \mathcal{H}_N[u]]) = (\mathcal{L}[u - \mathcal{H}_N[u]], \mathcal{F}_N[u_N - \mathcal{H}_N[u]]) \leq \gamma \|u - \mathcal{H}_N[u]\|_1 \|u_N - \mathcal{H}_N[u]\|_1,$$

using the continuity condition and Parseval's Theorem. Suppose that $\tilde{u} = u_N - \mathcal{H}_N[u]$. Then $\tilde{u} \in Z_N$ and furthermore $c_k \tilde{u}(\pm 1) - \tilde{u}^{(2k+1)}(\pm 1) = 0$ for $k = 1, ..., M$. Note that the left hand side of the above equation is just $(\mathcal{L}[\tilde{u}], \mathcal{F}_N[\tilde{u}])$. We have

$$(\mathcal{L}[\tilde{u}], \mathcal{F}_N[\tilde{u}]) \geq \omega \|\tilde{u}\|_1^2 - \gamma \|\tilde{u}\|_1 \|\tilde{u} - \mathcal{F}_N[\tilde{u}]\|_1.$$

However, since $c_k\tilde{u}(\pm 1) = \tilde{u}^{(2k+1)}(\pm 1)$, we obtain $\|\tilde{u} - \mathcal{F}_N[\tilde{u}]\|_1 \leq cN^{-\frac{5}{2}}\|\tilde{u}\|_1$ in the same manner as before. This gives

$$(\omega - cN^{-\frac{5}{2}})\|u_N - \mathcal{H}_N[u]\|_1^2 \leq (\mathcal{L}[u_N - \mathcal{H}_N[u]], \mathcal{F}_N[u_N - \mathcal{H}_N[u]]) \leq \gamma\|u - \mathcal{H}_N[u]\|_1\|u_N - \mathcal{H}_N[u]\|_1,$$

and the proof is complete. □

We may also provide an estimate for the uniform error:

Corollary 9. *Suppose that u satisfies the conditions of Lemma 10. Then*

$$\|u - u_N\|_\infty \leq c\|u - \mathcal{H}_N[u]\|_\infty,$$

for sufficiently large N.

Proof. In a similar manner to Theorem 9, we have

$$\omega\|e_N\|_1^2 - \gamma\|e_N\|_1\|p - \mathcal{F}_N[p]\|_1 \leq (\mathcal{L}[e_N], \mathcal{F}_N[e_N]) = a\left((u - \mathcal{H}_N[u])', \mathcal{F}_N[e_N]\right) \leq c\|u - \mathcal{H}_N[u]\|_\infty\|e_N\|_1,$$

where $e_N = u_N - \mathcal{H}_N[u] = v_N + p$. It is easily verified that $c_k e_N(\pm 1) = e_N^{(2k+1)}(\pm 1)$ for $k = 1, ..., M$, thus we obtain $\|p - \mathcal{F}_N[p]\|_1 \leq c\|e_N\|_1 N^{-\frac{5}{2}}$ and, for sufficiently large N,

$$\|e_N\|_1 \leq c\|u - \mathcal{H}_N[u]\|_\infty.$$

This gives

$$\|u - u_N\|_\infty \leq \|e_N\|_\infty + \|u - \mathcal{H}_N[u]\|_\infty \leq c\|u - \mathcal{H}_N[u]\|_\infty,$$

so we obtain the result. □

We are now in a position to provide estimates for the convergence rate in the H^1 and uniform norms:

Theorem 12. *Suppose that $u \in H^{4+2M}(-1, 1)$ and the functions $p_0, ..., p_{2M-1} \in H^{4+2M}(-1, 1)$ satisfy the conditions of Lemma 10. Then we have the error estimates*

$$\|u - \mathcal{H}_N[u]\|_1 \leq cN^{-\frac{5}{2}-2M}\|u\|_{4+2M}, \quad \|u - \mathcal{H}_N[u]\|_\infty \leq cN^{-3-2M}\|u\|_{4+2M}, \tag{4.15}$$

for sufficiently large N.

Proof. From (4.13) we see that

$$u - \mathcal{H}_N[u] = \left(u - \sum_{r=0}^{2M-1} b_r p_r\right) - \mathcal{F}_N\left[u - \sum_{r=0}^{2M-1} b_r p_r\right],$$

hence it suffices to show that the modified Fourier coefficients of $v = u - \sum_{r=0}^{2M-1} b_r p_r$ are sufficiently small. To this end let $s_k = \max\{|v^{(2k+1)}(\pm 1)|\}$ so that

$$|\hat{v}_n^{[i]}| \leq c\left(\frac{s_1}{n^4} + \frac{s_2}{n^6} + ... + \frac{s_{M+1}}{n^{4+2M}} + \frac{1}{n^{4+2M}}\|v\|_{4+2M}\right).$$

Using Lemma 11 we see that $v^{(2k+1)}(\pm 1) = c_k(v - \mathcal{F}_N[v])(\pm 1)$ so that

$$s_k = \max\left\{|c_k|\right\}\|v - \mathcal{F}_N[v]\|_\infty.$$

Since u and $p_0, ..., p_{2M-1}$ satisfy the boundary conditions, and $p_0, ..., p_{2M-1}$ are sufficiently smooth, we immediately obtain

$$s_k \leq cN^{-3}\|v\|_4, \quad k = 1, ..., M+1,$$

thus

$$\|v - \mathcal{F}_N[v]\|_\infty \leq cN^{-5}\left(\|v\|_{4+2M} + \|u\|_{4+2M}\right).$$

In turn we now see that $s_k \leq cN^{-5}\left(\|v\|_{4+2M} + \|u\|_{4+2M}\right)$ which means that

$$\|v - \mathcal{F}_N[v]\|_\infty \leq cN^{-7}\left(\|v\|_{4+2M} + \|u\|_{4+2M}\right).$$

We continue this process until we obtain $s_k \leq cN^{-3-2M}\left(\|v\|_{4+2M} + \|u\|_{4+2M}\right)$, meaning that

$$|\hat{v}_n^{[i]}| \leq c\left(n^{-4}N^{-3-2M} + n^{-4-2M}\right)\left(\|v\|_{4+2M} + \|u\|_{4+2M}\right).$$

To relate $\|v\|_{4+2M}$ to $\|u\|_{4+2M}$ we need to bound the coefficients b_r by $\|u\|_{4+2M}$. We have $P^N\hat{b} = \hat{U}$ and, for sufficiently large N, we obtain

$$\|\hat{b}\|_\infty \leq c\|\hat{U}\|_\infty \leq c\|u\|_{4+2M},$$

since $\hat{U}$ involves derivatives of u of order at most $2M+3$ evaluated at ± 1. Returning to v we have

$$|\hat{v}_n^{[i]}| \leq c\left(n^{-4}N^{-3-2M} + n^{-4-2M}\right)\|u\|_{4+2M}.$$

Using simple arguments we obtain (4.15). □

This method is equally applicable to general linear problems. The only difference being that the additional constraints now involve values of $a(x)$ and $b(x)$ and their derivatives at the endpoints. Provided these functions are sufficiently smooth, the results of this section may be generalized.

4.3 Iterative solution of the equations

The discretization equations (4.7) may be solved iteratively in a similar manner to before. We write

$$\tilde{M}_{\mathrm{G}} = \begin{pmatrix} M_{\mathrm{G}} & \widehat{\mathcal{L}_0[p_0]} & \ldots & \widehat{\mathcal{L}_0[p_{2M-1}]} \\ 0 & Q_1'[p_0] & \ldots & Q_1'[p_{2M-1}] \\ \ldots & \ldots & \ldots & \ldots \\ 0 & Q_M'[p_0] & \ldots & Q_M'[p_{2M-1}] \end{pmatrix}, \quad \tilde{N}_{\mathrm{G}} = \frac{1}{a}\begin{pmatrix} N_{\mathrm{G}} & \widehat{\mathcal{L}_1[p_0]} & \ldots & \widehat{\mathcal{L}_1[p_{2M-1}]} \\ c_1 C & Q_1''[p_0] & \ldots & Q_1''[p_{2M-1}] \\ \ldots & \ldots & \ldots & \ldots \\ c_M C & Q_M''[p_0] & \ldots & Q_M''[p_{2M-1}] \end{pmatrix},$$

where

$$Q_k[p_r] = -\begin{pmatrix} p_r^{(2k+1)}(1) \\ p_r^{(2k+1)}(-1) \end{pmatrix} + \begin{pmatrix} c_k p_r(1) \\ c_k p_r(-1) \end{pmatrix} = Q_k'[p_r] + Q_k''[p_r],$$

so that $\tilde{A}_{\mathrm{G}} = \tilde{M}_{\mathrm{G}} + a\tilde{N}_{\mathrm{G}}$ is the discretization matrix.

Lemma 13. *Suppose that the conditions of Lemma 10 hold and the operator $\mathcal{L}$ is coercive. Then, for sufficiently large N, the iteration scheme based on the above splitting converges.*

Proof. It suffices to show that $\rho(a\tilde{M}_{\mathrm{G}}^{-1}\tilde{N}_{\mathrm{G}}) < 1$. If λ is an eigenvalue of $a\tilde{M}_{\mathrm{G}}^{-1}N_{\mathrm{G}}$ with eigenfunction $u_N \in Z_N$, then

$$\begin{aligned} \lambda(\mathcal{L}_0[u_N], \phi) &= (\mathcal{L}_1[u_N], \phi), \quad \phi \in \mathcal{S}_N, \\ \lambda u_N^{(2k+1)}(\pm 1) &= c_k u_N(\pm 1), \quad k = 1, ..., M. \end{aligned} \tag{4.16}$$

Setting $\phi = \mathcal{F}_N[u_N]$ and rearranging, we obtain, after some simple inequalities,

$$|\lambda| \leq \frac{|a|\|u_N'\|\|u_N\|}{b\|u_N\|^2 + \|u_N'\|^2 - c\|u_N - \mathcal{F}_N[u_N]\|_1^2},$$

for some positive constant c. Now suppose that λ is the maximal eigenvalue in absolute value and that $|\lambda| \geq \sqrt{a^2/(4b)} + \theta$, for some $\theta > 0$, for some sufficiently large N. Then, if u_N has coefficients $a_n^{[i]}$ and b_k, using the second condition in (4.16), we have $|b_k| \leq c\|u_N\|_1$, so that $\|u_N - \mathcal{F}_N[u_N]\|_1^2 \leq c\|u_N\|_1^2 N^{-5}$ and

$$|\lambda| \leq \frac{|a|\|u_N'\|\|u_N\|}{(b - cN^{-5})\|u_N\|^2 + (1 - cN^{-5})\|u_N'\|^2}.$$

In the same manner as Lemma 9, we obtain

$$|\lambda| \leq \left(\frac{a^2}{4\,(b - cN^{-5})(\,1 - cN^{-5})}\right)^{\frac{1}{2}} \leq \sqrt{\frac{a^2}{4b}} + \frac{1}{2}\theta$$

for sufficiently large N, giving a contradiction. □

The result of this lemma shows that the spectral radius is bounded by

$$\rho(a\tilde{M}_{\mathrm{G}}^{-1}\tilde{N}_{\mathrm{G}}) \leq \frac{1}{2}\left(1+\sqrt{a^2/(4b)}\right).$$

for sufficiently large N. Hence, using this iterative scheme, Galerkin's equations may be solved in $\mathcal{O}\left(N^2\right)$ operations (or $\mathcal{O}\left(N\log N\right)$ using the FFT).

5 Other boundary conditions

The modified Fourier basis is naturally applicable to Neumann boundary value problems. It can be applied to problems with other boundary conditions, provided we introduce two additional basis functions to interpolate these conditions. This leads to a technique of similar form to that considered in Section 4. Unfortunately this approach becomes increasingly complicated for $d \geq 2$.

Another potential approach is to take linear combinations of basis functions to form a new basis that satisfies the boundary conditions. Unfortunately this method exhibits low accuracy: the new basis functions also satisfy homogeneous Neumann boundary conditions, and in general the exact solution will not.

For these reasons, a better approach is to choose basis functions that satisfy the boundary conditions inherently. To select such functions we proceed as follows. Given a problem $\mathcal{L}[u] = f$ with boundary conditions $\mathcal{B}[u] = 0$, we use the eigenfunctions of the Laplace operator $\triangle$ subject to these conditions. For Neumann boundary conditions we recover the modified Fourier basis. For Dirichlet boundary conditions $u(1) = u(-1) = 0$ we obtain the basis

$$\cos(n-\tfrac{1}{2})\pi x, \quad \sin n\pi x, \quad n \geq 1,$$

and for mixed boundary conditons $u(1) = u'(-1) = 0$ we have

$$\cos((n-\tfrac{3}{4})\pi x + \tfrac{1}{4}\pi), \quad \cos((n-\tfrac{1}{4})\pi x - \tfrac{1}{4}\pi), \quad n \geq 1.$$

Such eigenfunctions can also be derived for Robin and more general boundary conditions.

These sets of eigenfunctions share many properties of modified Fourier basis, and the resulting Galerkin methods possess many similar features, including mild conditioning. Moreover, a device for convergence acceleration can easily be developed, as in Section 4. For this reason the modified Fourier–Galerkin method can be viewed as a particular example of a class of methods for second order boundary value problems, each with basis functions determined by the boundary conditions.

This approach is not restricted to second order boundary value problems. Given a differential operator $\mathcal{L}$ on some domain Ω with certain boundary conditions, we consider the decomposition $\mathcal{L} = \mathcal{L}_0 + \mathcal{L}_1$, where $\mathcal{L}_0$ contains the highest order derivatives, and use the eigenfunctions of this operator as a basis for approximation. For practical purposes $\mathcal{L}_0$ should be self-adjoint, linear and have constant coefficients and Ω should be either the d-variate cube or equilateral triangle. Many of the properties of the modified Fourier method are inherited in this more general setting. One particular example of interest is the application eigenfunctions of the operator $\sum_{j=1}^{d} \partial_{x_j}^{2q}$, $q \geq 1$, which have been introduced in [9], to $2q^{\text{th}}$ order problems. This is an area for future consideration.

Acknowledgements

The author would like to thank Alfredo Deaño (DAMTP, University of Cambridge), Anders Hansen (DAMTP, University of Cambridge), Daan Huybrechs (Katholieke Universiteit Leuven), his supervisor Arieh Iserles (DAMTP, University of Cambridge), Sheehan Olver (University of Oxford) and Jie Shen (Purdue University).

References

[1] B. Adcock, *Sparse grid modified Fourier series and application to boundary value problems*, Technical report NA2008/08, DAMTP, University of Cambridge, (2008).

[2] H.-J. Bungartz and M. Griebel, *Sparse grids*, Acta Numerica, 13 (2004), pp. 147–269.

[3] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang, *Spectral methods: Fundamentals in Single Domains*, Springer, 2006.

[4] G. H. Golub and C. F. Van Loan, *Matrix Computations*, John Hopkins University Press, Baltimore, 2nd ed., 1989.

[5] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.

[6] D. Huybrechs, A. Iserles, and S. P. Nørsett, *From high oscillation to rapid approximation IV: Accelerating convergence*, Technical report NA2007/07, DAMTP, University of Cambridge, (2007).

[7] A. Iserles and S. P. Nørsett, *Efficient quadrature of highly oscillatory integrals using derivatives*, Proc. Royal Soc. A, 461 (2005), pp. 1383–1399.

[8] ———, *From high oscillation to rapid approximation I: Modified Fourier expansions*, Technical report NA2006/05, DAMTP, University of Cambridge, (2006).

[9] ———, *From high oscillation to rapid approximation II: Expansions in polyharmonic eigenfunctions*, Technical report NA2006/07, DAMTP, University of Cambridge, (2006).

[10] ———, *From high oscillation to rapid approximation III: Multivariate expansions*, Technical report NA2007/01, DAMTP, University of Cambridge, (2007).

[11] D. S. Mitrinovic, *Analytic Inequalities*, Springer–Verlag, 1970.

[12] S. Olver, *Moment-free numerical integration of highly oscillatory functions*, IMA J. Num. Anal., 26 (2006), pp. 213–227.

[13] ———, *On the convergence rate of a modified Fourier series*, Technical report NA2007/02, DAMTP, University of Cambridge, (2007).

[14] M. Práger, *Eigenvalues and eigenfunctions of the Laplace operator on an equilateral triangle*, Appl. Math., 43 (1998), pp. 311–320.

[15] A. Quarteroni and A. Valli, *Numerical Approximation of Partial Differential Equations*, Springer–Verlag, 1994.

[16] P. J. Roache, *A pseudo-spectral FFT technique for non-periodic problems*, J. Comput. Phys., 27 (1978), pp. 204–220.

[17] J. Shen, *Efficient spectral-Galerkin method. I. direct solvers of second and fourth-order equations using Legendre polynomials*, SIAM J. Sci. Comput., 15 (1994), pp. 1489–1505.

[18] ———, *Efficient spectral-Galerkin method. II. direct solvers of second and fourth-order equations using Chebyshev polynomials*, SIAM J. Sci. Comput., 16 (1995), pp. 74–87.

[19] V. Temlyakov, *Approximation of Periodic Functions*, Nova Sci., New York, 1993.