Multivariate modified Fourier series and application to boundary value problems

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

December 29, 2009

Abstract

In this paper we analyse the approximation-theoretic properties of modified Fourier series in Cartesian product domains with coefficients from both full and hyperbolic cross index sets. We show that the number of expansion coefficients may be reduced significantly whilst retaining comparable error estimates. In doing so we extend the univariate results of Iserles, Nørsett and S. Olver. We then demonstrate that these series can be used in the spectral-Galerkin approximation of second order Neumann boundary value problems, which offers some advantages over standard Chebyshev or Legendre polynomial discretizations.

1 Introduction

Univariate modified Fourier series—eigenseries of the Laplace operator subject to homogeneous Neumann boundary conditions—were introduced in [17] as an adjustment of Fourier series. Combined with modern quadrature methods (as opposed to the Fast Fourier Transform) to evaluate the coefficients, the benefit of using such series to approximate a non-periodic function f is a faster convergence rate (the convergence is uniform and there is no Gibbs phenomenon on the boundary). Moreover, the coefficients may be calculated adaptively in fewer operations without the restriction that the truncation parameter be a highly composite integer. In [18] these series and quadrature methods were generalised to Cartesian product domains.

In [15], alongside so-called polynomial subtraction (a familiar device for Fourier series [19, 21]), the authors used a hyperbolic cross index set [3, 27] to accelerate convergence. Due to the method of calculating the coefficients, such a device can be readily incorporated into modified Fourier series to produce approximations comprising a far reduced number of terms over approximations based on Fourier series or orthogonal polynomials. Thus in higher dimensions, modified Fourier series become an increasingly attractive option.

The aim of this paper is twofold. In Sections 2–4 we extend the work of [1, 15, 17, 18] and provide convergence analysis for modified Fourier series in various norms using various index sets. For reasons that we make clear, modified Fourier approximations are best analysed in so-called Sobolev spaces of *dominating mixed smoothness* [24]. Using this framework, we prove uniform convergence of such series, and provide estimates for the convergence rate in the L^2, H^q, $q \geq 1$, and uniform norms. We conclude that using a hyperbolic cross index set does not unduly affect the convergence rate, aside from possibly a logarithmic factor, provided additional (mixed) smoothness assumptions are imposed where necessary.

For univariate modified Fourier series it was observed in [17] and proved in [22] that the convergence rate is one order greater inside the interval than at the endpoints. We prove the same result for d-variate cubes using both full and hyperbolic cross index sets. Finally, we demonstrate that the advantage of modified Fourier series over classical Fourier series can be expressed as the observation that the set of modified Fourier eigenfunctions is not only an orthogonal basis for $\mathrm{L}^2(-1,1)^d$, but also for the space $\mathrm{H}^1_{\mathrm{mix}}(-1,1)^d$ (the first Sobolev space of dominating mixed smoothness).

One significant use of Fourier series is the discretization of boundary value problems with periodic boundary conditions. This approach offers numerous benefits, including rapid convergence and low complexity (see

[20] for the application of hyperbolic cross index sets to Fourier methods for periodic boundary value problems). The second aim of this paper is to provide the first steps towards the application of modified Fourier series to the numerical solution of boundary value problems in two or more dimensions (see [1] for the case $d = 1$). Because each modified Fourier basis function satisfies homogeneous Neumann boundary conditions, modified Fourier expansions are best suited to discretizations of non-periodic boundary value problems with the same boundary conditions. In the second half of this paper we consider the application to linear, second order problems defined on d-variate cubes. Much like the Fourier spectral method, this technique possesses a number of beneficial properties, including reasonable conditioning and the availability of an optimal, diagonal preconditioner. Furthermore, the operational cost of this method grows only moderately with the dimension d: the so-called *modified Fourier–Galerkin* approximation comprises $\mathcal{O}\left(N(\log N)^{d-1}\right)$ coefficients which can be found in only $\mathcal{O}\left(N^2\right)$ operations using standard iterative techniques. In comparison, the efficient spectral-Galerkin methods of Shen [12, 25, 26] based on Legendre and Chebyshev polynomials involve $\mathcal{O}\left(N^d\right)$ coefficients that can be found in at best $\mathcal{O}\left(N^{d+1}\right)$ operations.

The modified Fourier basis is best suited to Neumann boundary value problems. It can be applied to problems with other boundary conditions, however techniques for enforcing the boundary conditions are either increasingly complicated for $d \geq 2$ or lead to a loss of accuracy. For this reason, a better approach is to choose basis functions that satisfy the boundary conditions inherently. Given, for example, Robin boundary conditions, we use instead the basis of Laplace eigenfunctions subject to these boundary conditions. Such basis is very similar to the modified Fourier basis (the analysis of convergence is virtually identical), and the resulting Galerkin method possesses many similar features, including mild conditioning and low complexity. 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 (we return to this topic in Section 6). For the task of function approximation, modified Fourier expansions converge faster than expansions based on, for example, Laplace–Dirichlet eigenfunctions (which do not converge uniformly unless the function being approximated also satisfies homogeneous Dirichlet boundary conditions). Hence they are the natural choice from such class of bases. However, for the purposes spectral discretizations (where the exact solution automatically satisfies the boundary conditions), each basis is immediately adapted to a particular problem.

The disadvantage of all such methods is that they converge only algebraically in terms of the truncation parameter N. Standard orthogonal polynomial methods converge spectrally provided the solution is smooth. However, due to the much reduced complexity, for numerous test problems these methods convey an advantage for moderate values of N. In Section 5.4 we present several such examples. Unfortunately, the algebraic convergence rate means that beyond a certain (possibly large) threshold polynomial-based spectral methods will always outperform the modified Fourier method. As we discuss in greater detail in the conclusion of this paper (Section 7), convergence acceleration of the modified Fourier method is a subject of both current and future investigation.

Modified Fourier series provide a promising new approach for the approximation of functions and the numerical solution of partial differential equations. We mention in passing that, to date, modified Fourier series have found application in a number of other areas, including the computation of spectra of Fredholm operators [5]. Several examples in this paper demonstrate their potential over more standard algorithms. We stress that the aim of this paper is to provide a first insight into this topic and application. A great deal of future research is required, beyond the scope of this paper, before such methods become competitive algorithms. Insofar as such challenges are concerned, we highlight a number of open problems in Section 7.

Notation: Throughout we shall write $(\cdot, \cdot)$ for the standard $\mathrm{L}^2(\Omega)$ inner product on some domain Ω. We write $\| \cdot \|$ for the L^2 norm, $\| \cdot \|_q$ for the H^q norm, $q > 0$, and $\| \cdot \|_\infty$ for the uniform norm. N shall be a truncation parameter and I_N some finite index set. For a multi-index $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d$, D^α will correspond to the derivative operator $\partial^{\alpha_1}_{x_1} \ldots \partial^{\alpha_d}_{x_d}$ of order $|\alpha| = \alpha_1 + \ldots + \alpha_d$. If $\alpha = (r, r, \ldots, r)$, $r \in \mathbb{N}$, we also write D^r, and, if $r = 1$, just D.

We define $[d]$ to be the set of ordered tuples of length at most d with entries in $\{1, \ldots, d\}$. For $t \in [d]$ we write $|t|$ for the length (number of elements) in t, so that $t = (t_1, \ldots, t_{|t|})$ and $1 \leq t_1 < \ldots < t_{|t|} \leq d$. If $j \in \{1, \ldots, d\}$ we write $j \in t$ if $j = t_l$ for some $l = 1, \ldots, |t|$. Given $t \in [d]$, we define $\bar{t} \in [d]$ as the ordered tuple of length $d - |t|$ of elements not in t.

2 Modified Fourier series in $[-1,1]^d$

2.1 Definition and basic properties

The modified Fourier basis is the set of eigenfunctions of the Laplace operator subject to homogeneous Neumann boundary conditions. On the domain $\bar{\Omega}$, where $\Omega = (-1,1)^d$, these arise from Cartesian products of the univariate eigenfunctions

$$\phi^{[0]}_0(x) = \frac{1}{\sqrt{2}}, \quad \phi^{[0]}_n(x) = \cos n\pi x, \quad \phi^{[1]}_n(x) = \sin(n-\tfrac{1}{2})\pi x, \quad n \in \mathbb{N}_+ = \mathbb{N}\backslash\{0\}, \quad x \in [-1,1].$$

Given multi-indices $n = (n_1,\ldots,n_d) \in \mathbb{N}^d$ and $i = (i_1,\ldots,i_d) \in \{0,1\}^d$, the d-variate eigenfunctions are

$$\phi^{[i]}_n(x) = \prod_{j=1}^{d} \phi^{[i_j]}_{n_j}(x_j), \quad x = (x_1,\ldots,x_d) \in [-1,1]^d, \tag{2.1}$$

with corresponding eigenvalues $\mu^{[i]}_n = \sum_{j=1}^d \mu^{[i_j]}_{n_j}$, where $\mu^{[0]}_0 = 0$, $\mu^{[0]}_n = n^2\pi^2$ and $\mu^{[1]}_n = (n-\frac{1}{2})^2\pi^2$, $n \in \mathbb{N}_+$. For ease of notation we shall write $\phi^{[i]}_n$ and $\mu^{[i]}_n$ in this way, with the understanding that $\phi^{[i]}_n = 0$ and $\mu^{[i]}_n = 0$ if $i_j = 1$ and $n_j = 0$ for some $j = 1,\ldots,d$.

Concerning the density of such functions, we have the following:

Lemma 2.1. *The set $\{\phi^{[i]}_n : n \in \mathbb{N}^d,\ i \in \{0,1\}^d\}$ is an orthonormal basis of $\mathrm{L}^2(-1,1)^d$.*

Proof. This is a standard result of spectral theory. □

For a function $f \in \mathrm{L}^2(-1,1)^d$, truncation parameter $N \in \mathbb{N}$ and finite index set $I_N \subset \mathbb{N}^d$, we define the truncated modified Fourier series of f as

$$\mathcal{F}_N[f](x) = \sum_{i\in\{0,1\}^d} \sum_{n\in I_N} \hat{f}^{[i]}_n \phi^{[i]}_n(x), \quad \text{where} \quad \hat{f}^{[i]}_n = \int_{(-1,1)^d} f(x)\phi^{[i]}_n(x)\,\mathrm{d}x.$$

In [17, 18] quadrature routines are developed to evaluate the coefficients $\hat{f}^{[i]}_n$ numerically. Using highly oscillatory methods, where applicable, and so-called exotic quadrature elsewhere, any M coefficients can be found in $\mathcal{O}(M)$ operations. We shall not discuss such routines here. Such methods are greatly advantageous for modified Fourier approximations (they facilitate the use of hyperbolic cross index sets). However, there are a number of unresolved issues and open problems associated with their implementation, which we do not intend to address presently. We refer the reader to [18] and references therein for further detail (see also Section 7). For the remainder of this paper we shall assume that the error in approximating the coefficients is insignificant in comparison to the error in approximating f by $\mathcal{F}_N[f]$.

If we define the finite dimensional space $\mathcal{S}_N = \mathrm{span}\{\phi^{[i]}_n : n \in I_N,\ i \in \{0,1\}^d\}$, then $\mathcal{F}_N : \mathrm{L}^2(-1,1)^d \to \mathcal{S}_N$ is the orthogonal projection onto $\mathcal{S}_N$ with respect to the standard Euclidean inner product. We state, without proof, a version of Parseval's lemma for such series:

Lemma 2.2 (Parseval)**.** *Suppose that $f \in \mathrm{L}^2(-1,1)^d$, $\cup_{N\geq 0} I_N = \mathbb{N}^d$ and $I_1 \subset I_2 \subset \ldots \subset \mathbb{N}^d$. 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*

$$\|f\|^2 = \sum_{i\in\{0,1\}^d} \sum_{n\in\mathbb{N}^d} |\hat{f}^{[i]}_n|^2. \tag{2.2}$$

Unlike its Fourier counterpart, the modified Fourier basis is not closed under differentiation. If we differentiate $\phi^{[i]}_n$ with respect to x_1, say, we obtain

$$\partial_{x_1}\phi^{[i]}_n(x) = (-1)^{1+i_1}(\mu^{[i_1]}_{n_1})^{\frac{1}{2}}\psi^{[1-i_1]}_{n_1}(x_1)\prod_{j=2}^{d}\phi^{[i_j]}_{n_j}(x_j),$$

where $\{\psi_n^{[i]} \ : i = 0,1,\ n \in \mathbb{N}_+\}$ is the set of eigenfunctions of the univariate Laplace operator subject to homogeneous Dirichlet boundary conditions:

$$\psi_n^{[0]}(x) = \cos(n - \tfrac{1}{2})\pi x, \quad \psi_n^{[1]}(x) = \sin n\pi x, \quad n \in \mathbb{N}_+.$$

In particular, the Laplace–Neumann and Laplace–Dirichlet operators share eigenvalues (aside from the 0 eigenvalue of the former). We conclude that $\partial_{x_1}\phi_n^{[i]}(x)$ is proportional to an eigenfunction of the Laplace operator on $[-1,1]^d$ which obeys homogeneous Dirichlet boundary conditions on the subset of the boundary Γ_1, where $\Gamma_j = \{x \in [-1,1]^d \ : x_j = \pm 1\}$ for $j = 1,\ldots,d$, and homogeneous Neumann boundary conditions on $\Gamma\backslash\Gamma_1$, where $\Gamma = \partial\Omega = \cup_j \Gamma_j$. Such eigenfunctions are orthogonal and dense in $\mathrm{L}^2(-1,1)^d$. Repeating this argument for various j we obtain:

Lemma 2.3 (Duality)**.** *Suppose that $\alpha = (\alpha_1,\ldots,\alpha_d) \in \mathbb{N}^d$. If we apply the operator D^α to the set of modified Fourier eigenfunctions we obtain, up to scalar multiples, the eigenfunctions of the Laplace operator that obey homogeneous Dirichlet boundary conditions on the faces Γ_j where α_j is odd, and homogeneous Neumann boundary conditions elsewhere. Such eigenfunctions are orthonormal and dense in $\mathrm{L}^2(-1,1)^d$.*

This duality is essential to proving many of the convergence properties of modified Fourier series. As mentioned, the set of modified Fourier eigenfunctions is not only dense and orthogonal in $\mathrm{L}^2(-1,1)^d$, but also in several other Sobolev spaces. Using this lemma, we now show this for the space $\mathrm{H}^1(-1,1)^d$:

Lemma 2.4. *The set $\{\phi_n^{[i]} : n \in \mathbb{N}^d, i \in \{0,1\}^d\}$ is an orthogonal basis of $\mathrm{H}^1(-1,1)^d$. If $f \in \mathrm{H}^1(-1,1)^d$ 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*

$$\|f\|_1^2 = \sum_{i\in\{0,1\}^d} \sum_{n\in\mathbb{N}^d} (1 + \mu_n^{[i]})|\hat{f}_n^{[i]}|^2. \tag{2.3}$$

Proof. Orthogonality follows immediately from the Duality lemma. To establish density it suffices to prove that $\|\partial_{x_j}(f - \mathcal{F}_N[f])\| \to 0$, $N \to \infty$, for each j. By symmetry, it is enough to consider the case $j = 1$. Now,

$$\partial_{x_1}\mathcal{F}_N[f](x) = \sum_{i\in\{0,1\}^d} \sum_{\substack{n\in I_N \\ n_1\neq 0}} \hat{f}_n^{[i]}(-1)^{1+i_1}(\mu_{n_1}^{[i_1]})^{\frac{1}{2}}\tilde{\phi}_n^{[i]}(x),$$

where $\tilde{\phi}_n^{[i]} = \psi_{n_1}^{[1-i_1]}(x_1)\prod_{j=2}^d \phi_{n_j}^{[i_j]}(x_j)$ is an eigenfunction of the type introduced above. For $f \in \mathrm{H}^1(-1,1)^d$ and $n_1 \neq 0$, we obtain, via integration by parts,

$$\begin{aligned}\hat{f}_n^{[i]} &= \int_{(-1,1)^d} f(x)\phi_n^{[i]}(x)\,\mathrm{d}x = (-1)^{i_1}(\mu_{n_1}^{[i_1]})^{-\frac{1}{2}}\int_{(-1,1)^d} f(x)\partial_{x_1}\tilde{\phi}_n^{[i]}(x)\,\mathrm{d}x \\ &= (-1)^{1+i_1}(\mu_{n_1}^{[i_1]})^{-\frac{1}{2}}\int_{(-1,1)^d} \partial_{x_1}f(x)\tilde{\phi}_n^{[i]}(x)\,\mathrm{d}x.\end{aligned}$$

Using the above relation, we see that $\partial_{x_1}\mathcal{F}_N[f](x)$ is precisely the orthogonal projection of $\partial_{x_1}f$ onto the space $\tilde{\mathcal{S}}_N = \mathrm{span}\{\tilde{\phi}_n^{[i]} \ : n \in I_N,\ i \in \{0,1\}^d\}$. By the Duality lemma, the set $\{\tilde{\phi}_n^{[i]} \ : n \in \mathbb{N}^d,\ i \in \{0,1\}^d\}$ is an orthonormal basis is of $\mathrm{L}^2(-1,1)^d$. Since $\partial_{x_1}f \in \mathrm{L}^2(-1,1)^d$, it follows that $\|\partial_{x_1}(f - \mathcal{F}_N[f])\| \to 0$ as $N \to \infty$. Furthermore, using a version of Parseval's lemma for this basis we see that

$$\|\partial_{x_1}f\|^2 = \sum_{i\in\{0,1\}^d} \sum_{n\in\mathbb{N}^d} \mu_{n_1}^{[i_1]}|\hat{f}_n^{[i]}|^2.$$

Replacing 1 by $j = 2,\ldots,d$ in the above formula and summing each contribution gives (2.3). To conclude that $\mathcal{F}_N[f]$ is the best approximation in the H^1 norm, we merely notice that $\mathcal{F}_N : \mathrm{H}^1(-1,1)^d \to \mathcal{S}_N$ is the orthogonal projection with respect to the H^1 inner product. □

Lemma 2.4 provides an equivalent characterisation of the H^1 norm of a function $f \in \mathrm{H}^1(-1,1)^d$ in terms of its modified Fourier coefficients. An identical approach is employed for the periodic spaces $\mathrm{H}^q(\mathbb{T}^d)$, $q \geq 0$, using Fourier coefficients. Likewise, we may do the same in the modified Fourier setting when $q \neq 0,1$

provided we restrict to spaces of functions with vanishing odd derivatives on $\partial\Omega$—the analogue of periodicity for modified Fourier series. We shall not fully adopt this approach. Nonetheless, in the sequel it will be useful to consider the modified Fourier expansion of a function that satisfies a finite number of such derivative conditions on the boundary. For this we have the following result:

Lemma 2.5. *Suppose that $u \in \mathrm{H}^{2k+l}(-1,1)^d$, $l = 0,1$, obeys homogeneous Neumann boundary conditions up to order $k \in \mathbb{N}_+$ on $\partial\Omega$:*

$$\partial^{2r+1}_{x_j} u\big|_{\Gamma_j} = 0, \quad j = 1,\ldots,d, \quad r = 0,\ldots,k-1. \tag{2.4}$$

Then, for $r = 0,\ldots,2k+l$, $\mathcal{F}_N[u]$ 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 the characterisation:

$$\|u\|_r^2 = \sum_{i\in\{0,1\}^d} \sum_{n\in\mathbb{N}^d} \left[\sum_{|\alpha|\leq r} \prod_{j=1}^d (\mu^{[i_j]}_{n_j})^{\alpha_j} \right] |\hat{u}^{[i]}_n|^2. \tag{2.5}$$

Proof. This is very similar to Lemma 2.4. We may show (by repeated integration by parts, noticing that the boundary terms vanish due to (2.4)) that if u obeys the prescribed boundary conditions then $\mathrm{D}^\alpha \mathcal{F}_N[u]$, $|\alpha| \leq 2k+l$, is precisely the orthogonal projection of $\mathrm{D}^\alpha u$ onto the space spanned by Laplace eigenfunctions that satisfy homogeneous Dirichlet boundary conditions on the faces Γ_j when α_j is odd, and Neumann boundary conditions elsewhere. □

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

Corollary 2.6 (Bernstein's Inequality). *For $\phi \in \mathcal{S}_N$ and $r \in \mathbb{N}$ we have $\|\phi\|_r \leq \max_{n\in I_N} \left\{(1+\mu^{[0]}_n)^{\frac{r}{2}}\right\} \|\phi\|$.*

Proof. For $i \in \{0,1\}^d$ and $n \in I_N$, $\mu^{[i]}_n \leq \mu^{[0]}_n$. Furthermore

$$(1+\mu^{[i]}_n)^r = \sum_{|\alpha|\leq r} c_{\alpha,r} \prod_{j=1}^d (\mu^{[i_j]}_{n_j})^{\alpha_j}, \tag{2.6}$$

for some constants $c_{\alpha,r} \geq 1$. Hence, using (2.5) with $\phi \in \mathcal{S}_N$, we obtain

$$\|\phi\|_r^2 \leq \sum_{i\in\{0,1\}^d} \sum_{n\in I_N} (1+\mu^{[i]}_n)^r |\hat{\phi}^{[i]}_n|^2 \leq \max_{n\in I_N} \left\{(1+\mu^{[0]}_n)^r\right\} \sum_{i\in\{0,1\}^d} \sum_{n\in I_N} |\hat{\phi}^{[i]}_n|^2,$$

and Parseval's lemma gives the result. □

An advantage of modified Fourier series is that there is no Gibbs phenomenon on the boundary. Indeed, the modified Fourier expansion of a sufficiently smooth function converges uniformly on $[-1,1]^d$. We shall now prove this. One reason for doing so is to be able to express the error as a convergent infinite series, which in turn will allow us to derive estimates for the pointwise and uniform rates of convergence. This shall require particular choices of the index set I_N, which we defer to the sequel. However, uniform convergence may be established independently of the choice of index set. To do so we must consider Sobolev spaces of dominating mixed smoothness.

2.2 Sobolev spaces of dominating mixed smoothness

Sobolev spaces of dominating mixed smoothness are the standard setting whenever a hyperbolic cross index set is employed [6, 24, 27]. In the particular case of modified Fourier series, even for full index sets, such spaces provide a suitable framework for analysis.

It turns out that the modified Fourier basis is not just orthogonal and dense in the space $\mathrm{H}^1(-1,1)^d$, but also in the first Sobolev space of dominating mixed smoothness, which we denote $\mathrm{H}^1_{\mathrm{mix}}(-1,1)^d$. This fact ensures uniform convergence of $\mathcal{F}_N[f]$ to f which we prove in the next section. Subsequently we shall also see that the corresponding mixed norms are precisely those required to bound the modified Fourier coefficients $\hat{f}_n^{[i]}$ in inverse powers of $n_1 \dots n_d$. This leads to quasi-optimal error estimates and justifies the use of a hyperbolic cross index set in this context.

For $k \in \mathbb{N}$ we define the k^{th} Sobolev space of dominating mixed smoothness by

$$\mathrm{H}^k_{\text{mix}}(-1,1)^d = \{f : \mathrm{D}^\alpha f \in \mathrm{L}^2(-1,1)^d, \ \forall \ \alpha : |\alpha|_\infty \leq k\}, \tag{2.7}$$

where $|\alpha|_\infty = \max\{\alpha_i\}$, with norm

$$\|f\|^2_{k,\text{mix}} = \sum_{|\alpha|_\infty \leq k} \|\mathrm{D}^\alpha f\|^2. \tag{2.8}$$

This space is also commonly denoted by $\mathrm{S}_2^{(k,\dots,k)}\mathrm{H}(-1,1)^d$ in literature [24, 27].

In an identical manner to Lemma 2.5, we may characterise the space $\mathrm{H}^1_{\text{mix}}(-1,1)^d$ in terms of modified Fourier coefficients. We merely notice (recalling the proof of Lemma 2.4) that $\mathrm{D}^\alpha \mathcal{F}_N[f]$ is an orthogonal projection of $\mathrm{D}^\alpha f$ onto some suitable finite dimensional space not just for $|\alpha| \leq 1$, but also for $|\alpha|_\infty \leq 1$. This yields:

Lemma 2.7. *The set $\{\phi_n^{[i]} : n \in \mathbb{N}^d, i \in \{0,1\}^d\}$ is an orthogonal basis of $\mathrm{H}^1_{mix}(-1,1)^d$. If $f \in \mathrm{H}^1_{mix}(-1,1)^d$ then $\mathcal{F}_N[f]$ is the best approximation to f from $\mathcal{S}_N$ in the H^1_{mix} norm, $\|f - \mathcal{F}_N[f]\|_{1,mix} \to 0$ and*

$$\|f\|^2_{1,mix} = \sum_{i \in \{0,1\}^d} \sum_{n \in \mathbb{N}^d} \left[\sum_{|\alpha|_\infty \leq 1} \prod_{j=1}^d (\mu_{n_j}^{[i_j]})^{\alpha_j} \right] |\hat{f}_n^{[i]}|^2. \tag{2.9}$$

Furthermore, suppose that $u \in \mathrm{H}^{2k+l}_{mix}(-1,1)^d$, $l = 0,1$, satisfies the first $k \in \mathbb{N}_+$ derivative conditions (2.4). Then, for $r = 0,1,\dots,2k+l$, $\mathcal{F}_N[u]$ is the best approximation to u in the H^r_{mix} norm, $\|u - \mathcal{F}_N[u]\|_{r,mix} \to 0$ and

$$\|u\|^2_{r,mix} = \sum_{i \in \{0,1\}^d} \sum_{n \in \mathbb{N}^d} \left[\sum_{|\alpha|_\infty \leq r} \prod_{j=1}^d (\mu_{n_j}^{[i_j]})^{\alpha_j} \right] |\hat{u}_n^{[i]}|^2. \tag{2.10}$$

2.3 Uniform convergence

We commence with the following lemma:

Lemma 2.8. *We have the continuous imbedding $\mathrm{H}^1_{mix}(-1,1)^d \hookrightarrow \mathrm{C}[-1,1]^d$.*

To prove this we need the following lemma:

Lemma 2.9. *Suppose that $f \in \mathrm{C}^\infty[-1,1]^d$. Then*

$$f(x) = \sum_{t \in [d]^*} \int_{-1}^{x_{t_1}} \dots \int_{-1}^{x_{t_{|t|}}} \mathrm{D}_t f(x_t; -1) \, \mathrm{d}x_{t_1} \dots \mathrm{d}x_{t_{|t|}}, \quad x \in [-1,1]^d, \tag{2.11}$$

where $[d]$ is the set of ordered tuples of length at most d with entries in $\{1,\dots,d\}$, $[d]^* = [d] \cup \{\emptyset\}$, $\mathrm{D}_t = \partial_{x_{t_1}} \dots \partial_{x_{t_{|t|}}}$ for $t = (t_1,\dots,t_{|t|}) \in [d]^*$ and $(x_t;-1) \in \mathbb{R}^d$ has j^{th} entry x_j if $j \in t$ and -1 otherwise.

Proof. We use induction on d. For $d = 1$ we have $f(x) = \int_{-1}^x f'(x) \, \mathrm{d}x + f(-1)$, so the result holds. Now assume that (2.11) is valid for $d-1$. Then

$$\begin{aligned} f(x) &= \int_{-1}^{x_d} \partial_{x_d} f(x) \ \mathrm{d}x_d + f(x_1,\dots,x_{d-1},-1) \\ &= \sum_{t \in [d-1]^*} \left\{ \int_{-1}^{x_{t_1}} \dots \int_{-1}^{x_{t_{|t|}}} \int_{-1}^{x_d} \partial_{x_d} \mathrm{D}_t f(x_{(t,d)}; -1) \, \mathrm{d}x_{t_1} \dots \mathrm{d}x_{t_{|t|}} \, \mathrm{d}x_d \right. \\ &\quad \left. + \int_{-1}^{x_{t_1}} \dots \int_{-1}^{x_{t_{|t|}}} \mathrm{D}_t f(x_t; -1) \, \mathrm{d}x_{t_1} \dots \mathrm{d}x_{t_{|t|}} \right\}. \end{aligned}$$

Since the set $[d]^*$ is comprised of elements t and $(t,d) = (t_1,\dots,t_{|t|},d)$, where $t \in [d-1]^*$, this expression reduces to (2.11). Hence the proof is complete. □

Proof of Lemma 2.8. To prove this result we first demonstrate that the inequality

$$\|f\|_\infty \leq c\|f\|_{1,\mathrm{mix}}, \tag{2.12}$$

holds for all $f \in \mathrm{C}^\infty[-1,1]^d$ and some positive constant c independent of f. To do so, we note that

$$f(x_t; -1) = \int_{-1}^{1} \cdots \int_{-1}^{1} \mathrm{D}_{\bar{t}} \left(f(x) \prod_{j \notin t} \frac{x_j - 1}{2} \right) \mathrm{d}x_{\bar{t}_1} \ldots \mathrm{d}x_{\bar{t}_{|\bar{t}|}}, \quad \forall t \in [d]^*,$$

where $\bar{t} \in [d]^*$ is the tuple of length $|\bar{t}| = d - |t|$ of elements not in t. Hence, using Lemma 2.9, we have

$$f(x) = \sum_{t \in [d]^*} \int_{-1}^{1} \cdots \int_{-1}^{1} \int_{-1}^{x_{t_1}} \cdots \int_{-1}^{x_{t_{|t|}}} \mathrm{D} \left(f(x) \prod_{j \notin t} \frac{x_j - 1}{2} \right) \mathrm{d}x_{t_1} \ldots \mathrm{d}x_{t_{|t|}} \, \mathrm{d}x_{\bar{t}_1} \ldots \mathrm{d}x_{\bar{t}_{|\bar{t}|}}.$$

Each integrand involves terms of the form $\mathrm{D}^\alpha f$ for some $|\alpha|_\infty \leq 1$. Hence, using the Cauchy–Schwarz inequality and replacing suitable upper limits of integration by 1, we obtain (2.12) for $f \in \mathrm{C}^\infty[-1,1]^d$.

We now proceed in the standard manner. If $f \in \mathrm{H}^1_{\mathrm{mix}}(-1,1)^d$ then f is the limit in $\mathrm{H}^1_{\mathrm{mix}}(-1,1)^d$ of a sequence of functions belonging to $\mathrm{C}^\infty[-1,1]^d$. Since (2.12) holds for $f \in \mathrm{C}^\infty[-1,1]^d$ this sequence converges uniformly on $[-1,1]^d$ to $\tilde{f} \in \mathrm{C}[-1,1]^d$. Since $f = \tilde{f}$ a.e. the result follows. □

Theorem 2.10. *Suppose that $f \in \mathrm{H}^1_{mix}(-1,1)^d$ and I_N satisfies the conditions of Parseval's lemma. Then, $\mathcal{F}_N[f]$ converges pointwise to f for all $x \in [-1,1]^d$. Moreover, the convergence is uniform.*

Proof. Replacing f by $f - \mathcal{F}_N[f]$ in (2.12) and applying Lemma 2.7 gives the result. □

Prior to considering various different choices of index set and the corresponding error estimates for modified Fourier series, we need to develop bounds for the modified Fourier coefficients. This is the topic of the next section.

2.4 Bounds for modified Fourier coefficients

To obtain robust bounds for the coefficients $\hat{f}^{[i]}_n$ we shall apply Green's theorem to the integral $\hat{f}^{[i]}_n$. For this we need some additional notation. Given $j = 1, \ldots, d$, $r_j \in \mathbb{N}$ and $i_j \in \{0,1\}$ we define $\mathcal{B}^{[i_j]}_{r_j}[f]$ by

$$\begin{aligned}(-1)^{r_j} \mathcal{B}^{[i_j]}_{r_j}[f](x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_d) =& \partial^{2r_j+1}_{x_j} f(x_1, \ldots, x_{j-1}, 1, x_{j+1}, \ldots, x_d) \\ &+ (-1)^{i_j+1} \partial^{2r_j+1}_{x_j} f(x_1, \ldots, x_{j-1}, -1, x_{j+1}, \ldots, x_d).\end{aligned} \tag{2.13}$$

For $r_t = (r_{t_1}, \ldots, r_{t_{|t|}}) \in \mathbb{N}^{|t|}$ and $i_t = (i_{t_1}, \ldots, i_{t_{|t|}}) \in \{0,1\}^{|t|}$ we define $\mathcal{B}^{[i_t]}_{r_t}[f]$ as the composition

$$\mathcal{B}^{[i_t]}_{r_t}[f](x_{\bar{t}}) = \mathcal{B}^{[i_{t_1}]}_{r_{t_1}} \left[\mathcal{B}^{[i_{t_2}]}_{r_{t_2}} \left[\ldots \left[\mathcal{B}^{[i_{t_{|t|}}]}_{r_{t_{|t|}}} [f] \right] \ldots \right] \right], \tag{2.14}$$

with the understanding that when $t = \emptyset$, $\mathcal{B}^{[i_t]}_{r_t}[f] = f$. Note that the operators $\mathcal{B}^{[i_{t_j}]}_{r_{t_j}}$, $j \in t$, commute with each other and with differentiation in the variable $x_{\bar{t}} = (x_{\bar{t}_1}, \ldots, x_{\bar{t}_{|\bar{t}|}})$. Finally, given $i \in \{0,1\}^d$, $t \in [d]^*$, $r_t \in \mathbb{N}^{|t|}$, $|r_t|_\infty \leq k-1$, and $n_{\bar{t}} = (n_{\bar{t}_1}, \ldots, n_{\bar{t}_{|\bar{t}|}}) \in \mathbb{N}^{|\bar{t}|}_+$ we define $\mathcal{A}^{[i]}_{r_t, n_{\bar{t}}}[f] \in \mathbb{R}$ by

$$\mathcal{A}^{[i]}_{r_t, n_{\bar{t}}}[f] = (-1)^{k|\bar{t}|} \prod_{j \notin t} \left(\mu^{[i_j]}_{n_j} \right)^{-k} \int \mathcal{B}^{[i_t]}_{r_t}[\mathrm{D}^{2k}_{\bar{t}} f](x_{\bar{t}}) \phi^{[i_{\bar{t}}]}_{n_{\bar{t}}}(x_{\bar{t}}) \, dx_{\bar{t}}, \tag{2.15}$$

where $\mathrm{D}^{2k}_{\bar{t}} = \partial^{2k}_{x_{\bar{t}_1}} \ldots \partial^{2k}_{x_{\bar{t}_{|\bar{t}|}}}$. Observe that the integral is nothing more than the modified Fourier coefficient of the function $\mathcal{B}^{[i_t]}_{r_t}[\mathrm{D}^{2k}_{\bar{t}} f]$ corresponding to indices $i_{\bar{t}}$ and $n_{\bar{t}}$. We mention that the value $\mathcal{A}^{[i]}_{r_t, n_{\bar{t}}}[f]$ also depends on k and $t \in [d]$. However, to simplify notation we will not make this dependence explicit.

With this in hand we may now deduce the key result of this section:

Theorem 2.11. *Suppose that $f \in \mathrm{H}^{2k}_{mix}(-1,1)^d$, $k \in \mathbb{N}$, and that $n \in \mathbb{N}^d_+$. Then*

$$\hat{f}^{[i]}_n = \sum_{t \in [d]^*} \sum_{|r_t|_\infty = 0}^{k-1} \mathcal{A}^{[i]}_{r_t, n_{\bar{t}}}[f](-1)^{|n_t|+|i_t|} \prod_{j \in t} \left(\mu^{[i_j]}_{n_j}\right)^{-(r_j+1)}, \tag{2.16}$$

where $\mathcal{A}^{[i]}_{r_t,n_{\bar{t}}}[f]$ is given by (2.15). Suppose further that f obeys the first k derivative conditions. Then the only non-zero term in (2.16) corresponds to $t = \emptyset$. In other words

$$\hat{f}^{[i]}_n = (-1)^k \prod_{j=1}^d (\mu^{[i_j]}_{n_j})^{-k} \widehat{\mathrm{D}^{2k} f}^{[i]}_n.$$

Proof. To prove (2.16) it suffices to consider $f \in \mathrm{C}^\infty[-1,1]^d$. To cover the general case we use density, linearity and the bound $|\mathcal{A}^{[i]}_{r_t,n_{\bar{t}}}[f]| \leq c\|f\|_{2k,\mathrm{mix}}$, $\forall f \in \mathrm{H}^{2k}_{\mathrm{mix}}(-1,1)^d$, for some positive constant c independent of f,$n_{\bar{t}}$, r_t and i (see Lemma 2.13). We proceed by induction on d. For $d = 1$ trivial integration by parts verifies the result. Indeed, we have

$$\hat{f}^{[i]}_n = \sum_{r=0}^{k-1} \frac{(-1)^{r+n+i}}{(\mu^{[i]}_n)^{r+1}} \left\{ f^{(2r+1)}(1) + (-1)^{i+1} f^{(2r+1)}(-1) \right\} + \frac{(-1)^k}{(\mu^{[i]}_n)^k} \widehat{f^{(2k)}}^{[i]}_n, \quad n \in \mathbb{N}_+, \quad i \in \{0,1\}. \tag{2.17}$$

Now suppose that the result holds for $d-1$. Then $\hat{f}^{[i]}_n = \widehat{h^{[i_d]}_{n_d}}^{[i']}_{n'}$, where $h^{[i_d]}_{n_d}(x') = \int_{-1}^1 f(x)\phi^{[i_d]}_{n_d} \,\mathrm{d}x_d$ and i', n' and x' are the first $(d-1)$ entries of i, n and x respectively. Using the induction hypothesis we obtain

$$\hat{f}^{[i]}_n = \sum_{u \in [d-1]^*} \sum_{|r_u|_\infty=0}^{k-1} \mathcal{A}^{[i']}_{r_u,n_{\bar{u}}} \left[h^{[i_d]}_{n_d}\right] (-1)^{|n_u|+|i_u|} \prod_{j \in u} \left(\mu^{[i_j]}_{n_j}\right)^{-(r_j+1)}.$$

Applying the result for $d = 1$ to $h^{[i_d]}_{n_d}$ gives

$$\begin{aligned} \hat{f}^{[i]}_n = \sum_{u \in [d-1]^*} \sum_{|r_u|_\infty=0}^{k-1} \Bigg\{ & \sum_{r_d=0}^{k-1} (-1)^{n_d+i_d} \left(\mu^{[i_d]}_{n_d}\right)^{-(r_d+1)} \mathcal{A}^{[i']}_{r_u,n_{\bar{u}}} \left[\mathcal{B}^{[i_d]}_{r_d}[f]\right] \\ & + (-1)^k \left(\mu^{[i_d]}_{n_d}\right)^{-k} \mathcal{A}^{[i']}_{r_u,n_{\bar{u}}} \left[\int_{-1}^1 \partial^{2k}_{x_d} f(x)\phi^{[i_d]}_{n_d}(x_d)\,\mathrm{d}x_d\right] \Bigg\} (-1)^{|n_u|+|i_u|} \prod_{j \in u} \left(\mu^{[i_j]}_{n_j}\right)^{-(r_j+1)}. \end{aligned}$$

Suppose now that $t = (u,d) \in [d]$, where $u \in [d-1]^*$. Then $\mathcal{A}^{[i']}_{r_u,n_{\bar{u}}}\left[\mathcal{B}^{[i_d]}_{r_d}[f]\right] = \mathcal{A}^{[i]}_{r_t,n_{\bar{t}}}[f]$. Furthermore

$$(-1)^k \left(\mu^{[i_d]}_{n_d}\right)^{-k} \mathcal{A}^{[i']}_{r_u,n_{\bar{u}}} \left[\widehat{\partial^{2k}_{x_d} f}^{[i_d]}_{n_d}\right] = \mathcal{A}^{[i]}_{r_u,n_{\bar{u}}}[f],$$

where we consider u as an element of $[d]^*$ on the right hand side of this expression. Hence

$$\hat{f}^{[i]}_n = \sum_{u \in [d-1]^*} \left\{ \sum_{|r_t|_\infty=0}^{k-1} \mathcal{A}^{[i]}_{r_t,n_{\bar{t}}}[f](-1)^{|n_t|+|i_t|} \prod_{j \in t} \left(\mu^{[i_j]}_{n_j}\right)^{-(r_j+1)} + \mathcal{A}^{[i]}_{r_u,n_{\bar{u}}}[f](-1)^{|n_u|+|i_u|} \prod_{j \in u} \left(\mu^{[i_j]}_{n_j}\right)^{-(r_j+1)} \right\}.$$

If $t \in [d]^*$ then either $t = (u,d)$ or $t = u$ for some $u \in [d-1]^*$. The two terms in the above expression correspond to these two possibilities. Hence we obtain (2.16).

Now suppose that f obeys the first k derivative conditions, in other words $\mathcal{B}^{[i_j]}_{r_j}[f] = 0$ for all $i_j \in \{0,1\}$, $r_j = 0,\ldots,k-1$ and $j = 1,\ldots,d$. According to (2.15), any term $\mathcal{A}^{[i]}_{r_t,n_{\bar{t}}}[f]$ with $t \neq \emptyset$ will vanish. □

Theorem 2.11 does not include those coefficients $\hat{f}^{[i]}_n$ where $n_j = 0$ for some $j = 1,\ldots,d$. However, these can be easily handled. Given $n \in \mathbb{N}^d$, suppose that $n_t \equiv 0$ for some $t \in [d]$. If

$$f_t(x_{\bar{t}}) = \int_{-1}^1 \ldots \int_{-1}^1 f(x)\,\mathrm{d}x_{t_1}\ldots x_{t_{|t|}}, \tag{2.18}$$

then $\hat{f}^{[i]}_n = \widehat{f_t}^{[i_{\bar{t}}]}_{n_{\bar{t}}}$. Note that if $f \in \mathrm{H}^{2k}_{\mathrm{mix}}(-1,1)^d$ then $f_t \in \mathrm{H}^{2k}_{\mathrm{mix}}(-1,1)^{|\bar{t}|}$. We may now apply Theorem 2.11 to f_t to give the coefficient expansion in this case.

We now wish to derive bounds for the coefficients $\hat{f}^{[i]}_n$. To do so it is useful to consider the alternate mixed Sobolev spaces $\mathrm{G}^k_{\mathrm{mix}}(-1,1)^d = \left\{ f : \mathrm{D}^\alpha f \in \mathrm{L}^1(-1,1)^d,\ \forall\ \alpha : |\alpha|_\infty \leq k \right\}$, $k \in \mathbb{N}$, with norm $\|\!| f |\!\|_{k,\mathrm{mix}} = \sum_{|\alpha|_\infty \leq k} \| \mathrm{D}^\alpha f \|_{\mathrm{L}^1(-1,1)^d}$, where $\|g\|_{\mathrm{L}^1(-1,1)^d} = \int_{(-1,1)^d} |g(x)| \,\mathrm{d}x$. Regarding such spaces, we have the following result (which we shall use in the sequel):

Lemma 2.12. *The spaces $\mathrm{G}^k_{\mathit{mix}}(-1,1)^d$, $\mathrm{H}^k_{\mathit{mix}}(-1,1)^d$ satisfy $\mathrm{H}^k_{\mathit{mix}}(-1,1)^d \hookrightarrow \mathrm{G}^k_{\mathit{mix}}(-1,1)^d$ with imbedding constant $c = (2k+2)^{\frac{d}{2}}$.*

Proof. The existence of an imbedding is direct consequence of $\mathrm{L}^2(-1,1)^d \hookrightarrow \mathrm{L}^1(-1,1)^d$. For the imbedding constant we use the Cauchy–Schwarz inequality to obtain

$$\|\!| f |\!\|_{k,\mathrm{mix}} \leq 2^{\frac{d}{2}} \sum_{|\alpha|_\infty \leq k} \| \mathrm{D}^\alpha f \| \leq 2^{\frac{d}{2}} \left(\sum_{|\alpha|_\infty \leq k} 1 \right)^{\frac{1}{2}} \| f \|_{k,\mathrm{mix}}.$$

Since there are $(k+1)^d$ choices of $\alpha \in \mathbb{N}^d$ with $|\alpha|_\infty \leq k$ we obtain the result. □

To derive coefficient bounds we first need the following lemma:

Lemma 2.13. *Suppose that $f \in \mathrm{H}^{2k}_{\mathit{mix}}(-1,1)^d$, $i \in \{0,1\}^d$, $t \in [d]^*$, $r_t \in \mathbb{N}^{|t|}$ with $|r_t|_\infty \leq k-1$, $n_{\bar{t}} \in \mathbb{N}^{|\bar{t}|}$ and that $\mathcal{A}^{[i]}_{r_t,n_{\bar{t}}}[f]$ is given by (2.15). Then*

$$\left| \mathcal{A}^{[i]}_{r_t,n_{\bar{t}}}[f] \right| \leq \prod_{j \notin t} \left(\mu^{[i_j]}_{n_j} \right)^{-k} \|\!| f |\!\|_{2k,\mathit{mix}}.$$

Proof. If $\mathcal{B}^{[i_j]}_{r_j}[f]$ is given by (2.13) then $\mathcal{B}^{[i_j]}_{r_j}[f] = \int_{-1}^{1} \partial^{2r_j+2}_{x_j} \left(f(x) x_j^{i_j} \right) \mathrm{d}x_j$. Hence, the composition $\mathcal{B}^{[i_t]}_{r_t}[f]$ defined in (2.14) has integral representation

$$\mathcal{B}^{[i_t]}_{r_t}[f] = \int_{-1}^{1} \cdots \int_{-1}^{1} \mathrm{D}^{2r_t+2}_t \left(f(x) \prod_{j \in t} x_j^{i_j} \right) \mathrm{d}x_t.$$

Substituting this into the expression (2.15) for $\mathcal{A}^{[i]}_{r_t,n_{\bar{t}}}[f]$ gives

$$\mathcal{A}^{[i]}_{r_t,n_{\bar{t}}}[f] = (-1)^{k|\bar{t}|} \prod_{j \notin t} \left(\mu^{[i_j]}_{n_j} \right)^{-k} \int_{(-1,1)^d} \mathrm{D}^{2r_t+2}_t \mathrm{D}^{2k}_{\bar{t}} \left(f(x) \prod_{j \in t} x_j^{i_j} \right) \phi^{[i_{\bar{t}}]}_{n_{\bar{t}}}(x_{\bar{t}}) \,\mathrm{d}x.$$

We deduce that

$$\left| \mathcal{A}^{[i]}_{r_t,n_{\bar{t}}}[f] \right| \leq \prod_{j \notin t} \left(\mu^{[i_j]}_{n_j} \right)^{-k} \int_{(-1,1)^d} \left| \mathrm{D}^{2r_t+2}_t \mathrm{D}^{2k}_{\bar{t}} \left(f(x) \prod_{j \in t} x_j^{i_j} \right) \right| \mathrm{d}x \leq \prod_{j \notin t} \left(\mu^{[i_j]}_{n_j} \right)^{-k} \|\!| f |\!\|_{2k,\mathrm{mix}}.$$

Here the final inequality holds since the integral is a sum over derivatives $\mathrm{D}^\alpha f$ with $|\alpha|_\infty \leq 2k$ each multiplied by $x_1^{\beta_1} \ldots x_d^{\beta_d}$ for some suitable multi-index $|\beta|_\infty \leq 1$. □

Using this lemma we deduce the following:

Theorem 2.14. *Suppose that $f \in \mathrm{H}^{2k+2}_{\mathit{mix}}(-1,1)^d$ obeys the first $k \in \mathbb{N}$ derivative conditions (by convention, when $k = 0$ we mean that the function f obeys no derivative conditions). Then*

$$\left| \hat{f}^{[i]}_n \right| \leq 2^{\chi(n)} \left(\prod_{j : n_j > 0} \mu^{[i_j]}_{n_j} \right)^{-(k+1)} \|\!| f |\!\|_{2k+2,\mathit{mix}}, \quad n \in \mathbb{N}^d,$$

where $\chi(n)$, the grade of n, is the number of non-zero entries.

Proof. Suppose first that $n \in \mathbb{N}_+^d$. Then, using Lemma 2.11 (with k replaced by $k+1$) and the fact that f obeys the first k derivative conditions, we obtain

$$\hat{f}_n^{[i]} = \sum_{t \in [d]^*} \mathcal{A}_{k_t, n_{\bar{t}}}^{[i]}[f] (-1)^{|n_t|+|i_t|} \prod_{j \in t} \left(\mu_{n_j}^{[i_j]} \right)^{-(k+1)},$$

where $k_t = (k, k, \ldots, k) \in \mathbb{N}^{|t|}$. Using the bound for $\mathcal{A}_{k_t, n_{\bar{t}}}^{[i]}[f]$ from Lemma 2.13 we have

$$\left| \hat{f}_n^{[i]} \right| \leq \prod_{j=1}^{d} \left(\mu_{n_j}^{[i_j]} \right)^{-(k+1)} |||f|||_{2k+2,\mathrm{mix}} \sum_{t \in [d]^*} 1.$$

Since $|[d]^*| = 2^d$ and $\chi(n) = d$ in this case, we obtain the result for $n \in \mathbb{N}^d$. Now suppose that $n_t \equiv 0$ for some $t \in [d]$. Then, using the previous result,

$$|\hat{f}_n^{[i]}| = \left| \widehat{f_t}_{n_{\bar{t}}}^{[i_{\bar{t}}]} \right| \leq 2^{|\bar{t}|} \prod_{j : n_j > 0} \left(\mu_{n_j}^{[i_j]} \right)^{-(k+1)} |||f_t|||_{2k+2,\mathrm{mix}},$$

where f_t is defined in (2.18). Moreover,

$$|||f_t|||_{2k+2,\mathrm{mix}} = \sum_{\substack{|\alpha|_\infty \leq 2k+2 \\ \alpha \in \mathbb{N}^{\chi(n)}}} \int_{(-1,1)^{\chi(n)}} \left| \mathrm{D}^\alpha f_t(x) \right| \mathrm{d}x \leq \sum_{\substack{|\alpha|_\infty \leq 2k+2 \\ \alpha \in \mathbb{N}^{\chi(n)}}} \int_{(-1,1)^d} \left| \mathrm{D}^\alpha f(x) \right| \mathrm{d}x \leq |||f|||_{2k+2,\mathrm{mix}},$$

thus completing the proof. □

Using Lemma 2.12 we may also derive a bound for $\hat{f}_n^{[i]}$ in terms of $\|f\|_{2k+2,\mathrm{mix}}$:

Corollary 2.15. *Suppose that $f \in \mathrm{H}_{mix}^{2k+2}(-1,1)^d$ obeys the first $k \in \mathbb{N}$ derivative conditions. Then*

$$\left| \hat{f}_n^{[i]} \right| \leq 2^{\chi(n) + \frac{d}{2}} (2k+3)^{\frac{\chi(n)}{2}} \left(\prod_{j : n_j > 0} \mu_{n_j}^{[i_j]} \right)^{-(k+1)} \|f\|_{2k+2,mix}, \quad n \in \mathbb{N}^d.$$

Proof. If $\chi(n) = d$ the result follows immediately from Theorem 2.14 and Lemma 2.12. Now suppose that $\chi(n) < d$. We have

$$|\hat{f}_n^{[i]}| \leq 2^{\chi(n)} \left(\prod_{j : n_j > 0} \mu_{n_j}^{[i_j]} \right)^{-(k+1)} |||f_t|||_{2k+2,\mathrm{mix}}.$$

Furthermore, $|||f_t|||_{2k+2,\mathrm{mix}} \leq (4k+6)^{\frac{\chi(n)}{2}} \|f_t\|_{2k+2,\mathrm{mix}}$ and $\|\mathrm{D}^\alpha f_t\| \leq 2^{\frac{d}{2} - \frac{\chi(n)}{2}} \|\mathrm{D}^\alpha f\|$, $\alpha \in \mathbb{N}^{\chi(n)}$. Combining these observations we obtain $|||f_t|||_{2k+2,\mathrm{mix}} \leq 2^{\frac{d}{2}} (2k+3)^{\frac{\chi(n)}{2}} \|f\|_{2k+2,\mathrm{mix}}$, completing the proof. □

For the purposes of subsequent sections the following corollary is in fact more useful:

Corollary 2.16 (Coefficient bounds)**.** *Suppose that $f \in \mathrm{H}_{mix}^{2k+2}(-1,1)^d$ obeys the first $k \in \mathbb{N}$ derivative conditions. Then*

$$\left| \hat{f}_n^{[i]} \right| \leq 2^{\chi(n) + \frac{d}{2}} (2k+3)^{\frac{\chi(n)}{2}} (2^{|i|} \pi^{-\chi(n)})^{2(k+1)} (\bar{n}_1 \ldots \bar{n}_d)^{-2(k+1)} \|f\|_{2k+2,mix}, \quad n \in \mathbb{N}^d,$$

where $\bar{m} = \max\{m, 1\}$ for $m \in \mathbb{N}$.

Proof. For $n \in \mathbb{N}_+$ and $i \in \{0,1\}$ it is easily shown that $\mu_n^{[i]} \geq (2^{|i|} \pi^{-1})^{-2} n^2$. The result now follows immediately from Corollary 2.15. □

With these bounds in hand we are able to provide quasi-optimal estimates for the error $f - \mathcal{F}_N[f]$ in various norms using various index sets. We consider this in the next two sections.

3 Full index sets

The results of the previous sections do not make any assumptions regarding the index set I_N aside from the stipulations that I_N be finite, $\cup_N I_N = \mathbb{N}^d$ and $I_1 \subset I_2 \subset \ldots . \subset \mathbb{N}^d$. The size of I_N determines the cost of constructing the approximation $\mathcal{F}_N[f]$: using numerical quadrature, the number of operations required to evaluate the coefficients is $\mathcal{O}\left(|I_N|\right)$. Moreover, such methods are adaptive, making it possible to utilise any index set we choose.

Standard intuition leads to the *full index set*

$$I_N = \left\{n \in \mathbb{N}^d : \max_{j=1,\ldots,d}\{n_j\} \leq N\right\}, \tag{3.1}$$

which is just the hypercube of length $N+1$ in $\mathbb{N}^d$. Indeed, the prevalence of this index set in spectral discretizations is due to the fact that the method of choice for evaluating Fourier or Chebyshev coefficients, namely the FFT, computes all the coefficients in I_N in a non-adaptive way. However, $|I_N| = \mathcal{O}\left(N^d\right)$ and this figure grows exponentially with dimension. To alleviate this problem we employ a hyperbolic cross index set in the sequel. Such index set is viable precisely because it does not deteriorate the convergence rate of the approximation unduly, as we shall prove. To this end, for the purposes of comparison, we consider the approximation properties of modified Fourier series based on (3.1) in the remainder of this section. In the univariate case, this has been thoroughly dealt with in [1], [17] and [22]. We now extend these results to the multivariate setting.

3.1 Pointwise and uniform convergence rates

We first address the rate of pointwise convergence. This generalises the univariate result of S. Olver [22], to d-variate cubes. To do so, we require the following lemma:

Lemma 3.1. *Suppose that $f \in \mathrm{H}^{2k+3+l}_{mix}(-1,1)^d$, $l = 0, 1$, obeys the first $k \in \mathbb{N}$ derivative conditions and that I_N is the full index set (3.1). Then*

$$f(x) - \mathcal{F}_N[f](x) = \sum_{j=1}^{d}\sum_{i_j=0}^{1} \mathcal{B}^{[i_j]}_k[f](x_{\bar{j}})\left(p^{[i_j]}(x_j) - \mathcal{F}_N[p^{[i_j]}](x_j)\right) + \mathcal{O}\left(N^{-2k-2-l}\right), \tag{3.2}$$

where $\mathcal{B}^{[i_j]}_k[\cdot]$ is as in (2.13), $\bar{j} \in [d]$ is the tuple $(1,\ldots,j-1,j+1,\ldots,d)$ and $p^{[i_j]}(x_j)$ is a univariate polynomial of degree $(2(k+1)-i_j)$ satisfying the first k derivative conditions and $\mathcal{B}^{[i_j]}_k[p^{[i_j]}] = 1$.

Proof. Since uniform convergence is guaranteed by Theorem 2.10, we may write

$$f(x) - \mathcal{F}_N[f](x) = \sum_{t\in[d]}\sum_{i\in\{0,1\}^d}\sum_{\substack{n_j>N\\ j\in t}}\sum_{\substack{n_j=0\\ j\notin t}}^{N} \hat{f}^{[i]}_n \phi^{[i]}_n(x).$$

By the Coefficients bounds corollary we have $\hat{f}^{[i]}_n = \mathcal{O}\left((n_1\ldots n_d)^{-2k-2}\right)$. The largest contribution thus occurs when $|t| = 1$. Hence

$$f(x) - \mathcal{F}_N[f](x) = \sum_{j=1}^{d}\sum_{i\in\{0,1\}^d}\sum_{n_j>N}\sum_{\substack{n_l=0\\ l\neq j}}^{N} \hat{f}^{[i]}_n \phi^{[i]}_n(x) + \mathcal{O}\left(N^{-4k-4}\right).$$

We now expand the coefficient $\hat{f}^{[i]}_n$ in powers of n_j^{-1}. For each j, $2(k+1)$ integrations by parts give

$$\hat{f}^{[i]}_n = \frac{(-1)^{n_j+i_j}}{(\mu^{[i_j]}_{n_j})^{k+1}}\widehat{\mathcal{B}^{[i_j]}_k[f]}^{[i_{\bar{j}}]}_{n_{\bar{j}}} + \mathcal{O}\left((n_1\ldots n_d)^{-2}n_j^{-2k-1-l}\right) = \widehat{p^{[i_j]}}^{[i_j]}_{n_j}\widehat{\mathcal{B}^{[i_j]}_k[f]}^{[i_{\bar{j}}]}_{n_{\bar{j}}} + \mathcal{O}\left((n_1\ldots n_d)^{-2}n_j^{-2k-1-l}\right).$$

Substituting this into the previous expression we obtain

$$\begin{aligned} f(x) - \mathcal{F}_N[f](x) &= \sum_{j=1}^{d} \sum_{i_j=0}^{1} \mathcal{F}_N \left[\mathcal{B}_k^{[i_j]}[f] \right] (x_{\bar{j}}) \sum_{n_j > N} \widehat{p^{[i_j]}}^{[i_j]}_{n_j} \phi_{n_j}^{[i_j]}(x_j) + \mathcal{O}\left(N^{-2k-2-l}\right) \\ &= \sum_{j=1}^{d} \sum_{i_j=0}^{1} \mathcal{B}_k^{[i_j]}[f](x_{\bar{j}}) \sum_{n_j > N} \widehat{p^{[i_j]}}^{[i_j]}_{n_j} \phi_{n_j}^{[i_j]}(x_j) + \mathcal{O}\left(N^{-2k-2-l}\right). \end{aligned}$$

Since the infinite sum is precisely $p^{[i_j]}(x_j) - \mathcal{F}_N[p^{[i_j]}](x_j)$, the result follows. □

Note that it is not clear *a priori* that such univariate polynomials $p^{[i_j]}$ exist. However, this has been demonstrated in [17]. Using this lemma, we now deduce the following theorem:

Theorem 3.2. *Suppose that $f \in \mathrm{H}^{2k+3}_{mix}(-1,1)^d$ obeys the first k derivative conditions. Then the error $f(x) - \mathcal{F}_N[f](x)$ is $\mathcal{O}\left(N^{-2k-2}\right)$ uniformly in any compact subset of $(-1,1)^d$.*

Proof. This follows immediately from Lemma 3.1 and the univariate result applied to the function $p^{[i_j]}$. □

In fact, using univariate arguments described in [22], we may easily determine the exact leading order asymptotic behaviour of the error $f(x) - \mathcal{F}_N[f](x)$. It can be shown that

$$p^{[i_j]}(x_j) - \mathcal{F}_N[p^{[i_j]}](x_j) = (N\pi)^{-2(k+1)} \mathrm{Re} \left[\frac{(-\mathrm{e}^{\mathrm{i}\pi x_j})^{N+1-\frac{1}{2}i_j}}{1 + \mathrm{e}^{\mathrm{i}\pi x_j}} \right] + \mathcal{O}\left(N^{-2k-3}\right), \quad -1 < x_j < 1,$$

where i is the imaginary unit. Substituting this into (3.2) immediately yields

$$f(x) - \mathcal{F}_N[f](x) = (N\pi)^{-2(k+1)} \sum_{j=1}^{d} \sum_{i_j=0}^{1} \mathcal{B}_k^{[i_j]}[f](x_{\bar{j}}) \mathrm{Re} \left[\frac{(-\mathrm{e}^{\mathrm{i}\pi x_j})^{N+1-\frac{1}{2}i_j}}{1 + \mathrm{e}^{\mathrm{i}\pi x_j}} \right] + \mathcal{O}\left(N^{-2k-2-l}\right),$$

for $x \in (-1,1)^d$. Provided $f \in \mathrm{H}^{2k+4}_{\mathrm{mix}}(-1,1)^d$ this establishes the leading order asymptotics.

We note that Theorem 3.2 excludes subsets of the boundary. Dealing with such regions is much easier, we merely use a bound for the uniform error. For such a bound, we also require lower smoothness:

Theorem 3.3. *Suppose that $f \in \mathrm{H}^{2k+2}_{mix}(-1,1)^d$ satisfies the first $k \in \mathbb{N}$ derivative conditions and I_N is the full index set (3.1). Then*

$$\|f - \mathcal{F}_N[f]\|_\infty \leq \|f\|_{2k+2,mix} \left[2^{\frac{3}{2}} (1 + 4^{k+1}) c_k \right]^d \left[(2k+1) c_k \right]^{-1} N^{-(2k+1)},$$

where $c_k = 1 + 2(2k+3)^{\frac{1}{2}} \pi^{-2(k+1)} \zeta(2(k+1))$ and $\zeta(\cdot)$ is the zeta function.

Proof. We have

$$\begin{aligned} &\|f - \mathcal{F}_N[f]\|_\infty \\ &\qquad \leq \sum_{i \in \{0,1\}^d} \sum_{n \notin I_N} |\hat{f}_n^{[i]}| \\ &\qquad \leq \|f\|_{2k+2,\mathrm{mix}} \sum_{i \in \{0,1\}^d} 2^{2(k+1)|i|} \sum_{t \in [d]} \sum_{\substack{n_j = 0 \\ j \notin t}}^{N} \sum_{\substack{n_j > N \\ j \in t}} 2^{\chi(n) + \frac{d}{2}} (2k+3)^{\frac{\chi(n)}{2}} \pi^{-2(k+1)\chi(n)} (\bar{n}_1 \ldots \bar{n}_d)^{-2k-2} \\ &\qquad = \|f\|_{2k+2,\mathrm{mix}} 2^{\frac{d}{2}} (1 + 4^{k+1})^d \sum_{t \in [d]} \sum_{\substack{n_j = 0 \\ j \notin t}}^{N} \sum_{\substack{n_j > N \\ j \in t}} \left[2(2k+3)^{\frac{1}{2}} \pi^{-2(k+1)} \right]^{\chi(n)} (\bar{n}_1 \ldots \bar{n}_d)^{-2k-2}. \end{aligned}$$

Since $\sum_{n=1}^{N} n^{-2(k+1)} \leq \zeta(2(k+1))$ and $\sum_{n>N} n^{-2(k+1)} \leq \frac{1}{2k+1} N^{-(2k+1)}$ it follows that

$$\|f - \mathcal{F}_N[f]\|_\infty \leq \|f\|_{2k+2,\mathrm{mix}} \left[2^{\frac{1}{2}} (1 + 4^{k+1}) c_k \right]^d \sum_{t \in [d]} \{(2k+1) c_k\}^{-|t|} N^{-(2k+1)|t|}.$$

It is easily shown that $\sum_{t \in [d]} a^{|t|} \leq 2^d a$ for any constant $0 \leq a \leq 1$. Setting $a = [(2k+1)c_k]^{-1} N^{-(2k+1)}$ and substituting into the previous expression now yields the result. □

The result of Theorems 3.2 and 3.3 is that, for a general function f obeying no derivative conditions, the convergence rate of its modified Fourier series $\mathcal{F}_N[f]$ is $\mathcal{O}\left(N^{-1}\right)$ on the boundary and $\mathcal{O}\left(N^{-2}\right)$ inside the domain. Conversely, when f obeys certain derivative conditions, not only are higher degrees of convergence guaranteed (see Section 2), such rates of convergence also increase.

3.2 Estimates in other norms

Concerning the error in the H^s norm, we have:

Lemma 3.4. *Suppose that $f \in \mathrm{H}^{2k+l}(-1,1)^d$, $l = 0, 1$, satisfies the first $k \in \mathbb{N}$ derivative conditions and I_N is the full index set (3.1). Then*

$$\|f - \mathcal{F}_N[f]\|_s \leq c_{r,s} N^{s-r} \|f\|_r, \quad r = s, \ldots, 2k+l, \quad s = 0, \ldots, 2k+l, \tag{3.3}$$

for some positive constant $c_{r,s}$ independent of f and N.

Proof. For $n \notin I_N$, $\mu_n^{[i]} \geq (N\pi)^2$. Using this, Lemma 2.5 and (2.6) we have

$$\begin{aligned}\|f - \mathcal{F}_N[f]\|_s^2 &\leq \sum_{i \in \{0,1\}^d} \sum_{n \notin I_N} (1 + \mu_n^{[i]})^s |\hat{f}_n^{[i]}|^2 \leq (N\pi)^{2(s-r)} \sum_{i \in \{0,1\}^d} \sum_{n \notin I_N} (1 + \mu_n^{[i]})^r |\hat{f}_n^{[i]}|^2 \\ &\leq c_{r,s} N^{2(s-r)} \sum_{i \in \{0,1\}^d} \sum_{n \in \mathbb{N}^d} \sum_{|\alpha| \leq r} \prod_{j=1}^{d} (\mu_{n_j}^{[i_j]})^{\alpha_j} |\hat{f}_n^{[i]}|^2 = c_{r,s} N^{2(s-r)} \|f\|_r^2,\end{aligned}$$

as required. □

The conclusion of Lemma 3.4 may lead to the assertion that, for smooth f satisfying the first k odd derivative conditions, $\|f - \mathcal{F}_N[f]\|_{2k+1} = \mathcal{O}(1)$, an estimate which in view Lemma 2.5 is not optimal. However, it turns out that $\|f - \mathcal{F}_N[f]\|_{2k+1} = \mathcal{O}(N^{-\frac{1}{2}})$ in this case, as we shall now prove. To show this, instead of using the above method of proof, we utilise the coefficient bounds of Section 2.4.

Lemma 3.5. *Suppose that $f \in \mathrm{H}^{2k+2}_{mix}(-1,1)^d$ satisfies the first $k \in \mathbb{N}$ derivative conditions and I_N is the full index set (3.1). Then*

$$\|f - \mathcal{F}_N[f]\|_s \leq c_s N^{s-2k-\frac{3}{2}} \|f\|_{2k+2,mix}, \quad s = 0, \ldots, 2k+1, \tag{3.4}$$

for some positive constant c_s independent of f and N.

Proof. Using Lemma 2.5 we have

$$\|f - \mathcal{F}_N[f]\|_s^2 = \sum_{i \in \{0,1\}^d} \sum_{|\alpha| \leq s} \sum_{t \in [d]} \sum_{\substack{n_j = 0 \\ j \notin t}}^{N} \sum_{\substack{n_j > N \\ j \in t}} |\hat{f}_n^{[i]}|^2 \prod_{j=1}^{d} (\mu_{n_j}^{[i_j]})^{\alpha_j}.$$

Since $\hat{f}_n^{[i]} = \mathcal{O}\left((n_1 \ldots n_d)^{-2k-2}\right)$ it follows that

$$\|f - \mathcal{F}_N[f]\|_s^2 \leq c_s \sum_{|\alpha| \leq s} \sum_{t \in [d]} \sum_{\substack{n_j = 0 \\ j \notin t}}^{N} \sum_{\substack{n_j > N \\ j \in t}} \prod_{j=1}^{d} n_j^{2\alpha_j - 4k - 4} \leq c_s \sum_{|\alpha| \leq s} \sum_{t \in [d]} N^{2|\alpha| - (4k+3)|t|} \leq c_s N^{2s - (4k+3)},$$

as required. □

As in Theorem 3.3, it is possible to prescribe values for the constants appearing in Lemmas 3.4 and 3.5. However, we shall not do this either here or in the remainder of this paper: numerical results indicate that such constants are not unduly large.

4 Hyperbolic cross index sets

A hyperbolic cross index set is obtained by including only those terms in the expansion

$$\sum_{i \in \{0,1\}^d} \sum_{n \in \mathbb{N}^d} \hat{f}^{[i]}_n \phi^{[i]}_n(x),$$

whose absolute value in some norm is greater than some tolerance ϵ. To do so, we need bounds for the coefficients $\hat{f}^{[i]}_n$ and the functions $\phi^{[i]}_n$. Given an arbitrary norm $|||\cdot|||$, we use the bounds of Section 2.4 for the former, to obtain:

$$|||\hat{f}^{[i]}_n \phi^{[i]}_n||| \leq c\|f\|_{2,\mathrm{mix}}(\bar{n}_1 \ldots \bar{n}_d)^{-2}|||\phi^{[i]}_n|||.$$

We shall consider the index set that originates from the L^2 and uniform norms. (It is possible to take a more general viewpoint and consider the index set arising from the H^s norm, $s \geq 0$, leading to a so-called *optimized hyperbolic cross* [11]. This possesses some advantages over the L^2 norm hyperbolic cross. However, though the analysis presented herein can be extended to this setting, for simplicity, we shall not pursue this further.). In this case $\|\phi^{[i]}_n\|_\infty = \|\phi^{[i]}_n\| = 1$, and $|||\hat{f}^{[i]}_n \phi^{[i]}_n||| \leq c\|f\|_{2,\mathrm{mix}}(\bar{n}_1 \ldots \bar{n}_d)^{-2}$. The tolerance ϵ is defined as precisely this upper bound with $n = (N, 0, \ldots, 0)$. In other words $\epsilon = c\|f\|_{2,\mathrm{mix}}N^{-2}$. This yields the *hyperbolic cross* [3, 27] index set

$$I_N = \{n \in \mathbb{N}^d : \bar{n}_1 \ldots \bar{n}_d \leq N\}. \tag{4.1}$$

We devote the remainder of this section to showing the benefit of this index set. There are two aspects to this: the reduced cost in forming the approximation—in other words, the reduced size of the hyperbolic cross index set—and the retention of similar error estimates in comparison to approximations based on the full index set (3.1). We commence with the former:

Lemma 4.1. *Suppose that $\theta_d(t)$ is the number of terms $n = (n_1, \ldots, n_d) \in \mathbb{N}^d$ such that $\bar{n}_1 \ldots \bar{n}_d \leq t$. Then*

$$\theta_d(t) = \frac{t(\log t)^{d-1}}{(d-1)!} + \mathcal{O}\left(t(\log t)^{d-2}\right), \quad t \gg 1.$$

For a proof of this in a more general setting we refer to [8]. A simple inductive argument appears in [15], which we now repeat here, since similar methods will be used in the sequel:

Proof. For $d = 1$, $\theta_1(t) = t$ as required. Suppose now that the result is true for $d-1$. Then

$$\begin{aligned}\theta_d(t) = \sum_{n=1}^{\lfloor t \rfloor} \theta_{d-1}\left(\frac{t}{n}\right) &= \frac{1}{(d-2)!}\sum_{n=1}^{\lfloor t \rfloor} \frac{t}{n}\left[\log\left(\frac{t}{n}\right)\right]^{d-2} + \mathcal{O}\left(t(\log t)^{d-2}\right) \\ &= \frac{1}{(d-2)!}\int_1^t \frac{t}{n}\left[\log\left(\frac{t}{n}\right)\right]^{d-2} \mathrm{d}n + \mathcal{O}\left(t(\log t)^{d-2}\right) \\ &= \frac{1}{(d-2)!}t\int_1^t x^{-1}(\log x)^{d-2}\,\mathrm{d}x + \mathcal{O}\left(t(\log t)^{d-2}\right).\end{aligned}$$

Evaluation of this integral completes the proof. □

Corollary 4.2. *The number of terms in the expansion $\mathcal{F}_N[f]$ based on the hyperbolic cross (4.1) is*

$$\frac{2^d}{(d-1)!}N(\log N)^{d-1} + \mathcal{O}\left(N(\log N)^{d-2}\right). \tag{4.2}$$

Proof. For any n with strictly positive entries there are 2^d choices of $i \in \{0,1\}^d$. The total number of coefficients $\hat{f}^{[i]}_n$ where at least one entry of n is zero is $\mathcal{O}\left(N(\log N)^{d-2}\right)$. □

4.1 Convergence rate in various norms

We now assess the rate of convergence in various norms of the approximation $\mathcal{F}_N[f]$ based on the hyperbolic cross (4.1):

Lemma 4.3. *Suppose that $f \in \mathrm{H}^{2k+l}(-1,1)^d$, $l = 0, 1$, satisfies the first $k \in \mathbb{N}$ derivative conditions and I_N is the hyperbolic cross index set (4.1). Then, for some positive constant $c_{r,s}$ independent of f and N,*

$$\|f - \mathcal{F}_N[f]\|_s \leq c_{r,s} N^{\frac{s-r}{d}} \|f\|_r, \quad r = s, \ldots, 2k+l, \quad s = 0, \ldots, 2k+l. \tag{4.3}$$

If, additionally, $f \in \mathrm{H}^{2k+l}_{\mathit{mix}}(-1,1)^d$, then

$$\|f - \mathcal{F}_N[f]\|_s \leq c_{r,s} N^{s-r} \|f\|_{r,\mathit{mix}}, \quad r = s, \ldots, 2k+l, \quad s = 0, \ldots, 2k+l. \tag{4.4}$$

Proof. Due to Lemma 2.5 and (2.6) we may write

$$\|f - \mathcal{F}_N[f]\|^2_s \leq \sum_{i \in \{0,1\}^d} \sum_{n \notin I_N} |\hat{f}^{[i]}_n|^2 (1 + \mu^{[i]}_n)^s = \sum_{i \in \{0,1\}^d} \sum_{n \notin I_N} |\hat{f}^{[i]}_n|^2 (1 + \mu^{[i]}_n)^r (1 + \mu^{[i]}_n)^{s-r}.$$

By a standard inequality $1 + \mu^{[i]}_n \geq c\,(\bar{n}_1 \ldots \bar{n}_d)^{\frac{2}{d}}$, and, since $n \notin I_N$, this gives $1 + \mu^{[i]}_n \geq N^{\frac{2}{d}}$. Hence

$$\|f - \mathcal{F}_N[f]\|^2_s \leq c_{r,s} N^{\frac{2(s-r)}{d}} \sum_{i \in \{0,1\}^d} \sum_{n \in \mathbb{N}^d} |\hat{f}^{[i]}_n|^2 (1 + \mu^{[i]}_n)^r \leq c_{r,s} N^{\frac{2(s-r)}{d}} \|f\|^2_r,$$

which gives (4.3). Next consider (4.4). Clearly $\|f - \mathcal{F}_N[f]\|_s \leq \|f - \mathcal{F}_N[f]\|_{s,\mathrm{mix}}$ and, by Lemma 2.7,

$$\begin{aligned}
\|f - \mathcal{F}_N[f]\|^2_{s,\mathrm{mix}} &= \sum_{i \in \{0,1\}^d} \sum_{n \notin I_N} \left[\sum_{|\alpha|_\infty \leq s} \prod_{j=1}^d (\mu^{[i_j]}_{n_j})^{\alpha_j} \right] |\hat{f}^{[i]}_n|^2 \\
&\leq \sum_{i \in \{0,1\}^d} \sum_{n \notin I_N} |\hat{f}^{[i]}_n|^2 \prod_{j=1}^d (1 + \mu^{[i_j]}_{n_j})^s \leq c_{r,s} N^{2(s-r)} \sum_{i \in \{0,1\}^d} \sum_{n \in \mathbb{N}^d} |\hat{f}^{[i]}_n|^2 \prod_{j=1}^d (1 + \mu^{[i_j]}_{n_j})^r \\
&\leq c_{r,s} N^{2(s-r)} \sum_{i \in \{0,1\}^d} \sum_{n \in \mathbb{N}^d} \left[\sum_{|\alpha|_\infty \leq r} \prod_{j=1}^d (\mu^{[i_j]}_{n_j})^{\alpha_j} \right] |\hat{f}^{[i]}_n|^2 \leq c_{r,s} N^{2(s-r)} \|f\|^2_{r,\mathrm{mix}}.
\end{aligned}$$

This yields (4.4). □

As with the full index set, this lemma is non-optimal for functions f of sufficient smoothness. For this case, as before, we require the coefficient bounds to derive error estimates:

Theorem 4.4. *Suppose that $f \in \mathrm{H}^{2k+2}_{\mathit{mix}}(-1,1)^d$ obeys the first $k \in \mathbb{N}$ derivative conditions and I_N is the hyperbolic cross index set (4.1). Then*

$$\begin{aligned}
\|f - \mathcal{F}_N[f]\|_\infty &\leq c_k \|f\|_{2k+2,\mathit{mix}} N^{-2k-1} (\log N)^{d-1}, \\
\|f - \mathcal{F}_N[f]\| &\leq c_{k,0} \|f\|_{2k+2,\mathit{mix}} N^{-2k-\frac{3}{2}} (\log N)^{\frac{d-1}{2}}, \\
\|f - \mathcal{F}_N[f]\|_s &\leq c_{k,s} \|f\|_{2k+2,\mathit{mix}} N^{s-2k-\frac{3}{2}}, \quad s = 1, \ldots, 2k+1,
\end{aligned}$$

where c_k, $c_{k,s}$ are positive constants independent of f and N.

To establish this theorem we need the following lemma:

Lemma 4.5. *Suppose that $\gamma_{r,d}(t) = \sum_{\bar{n}_1 \ldots \bar{n}_d > t} (\bar{n}_1 \ldots \bar{n}_d)^{-r-1}$, $r > 0$. Then*

$$\gamma_{r,d}(t) = \frac{t^{-r} (\log t)^{d-1}}{r(d-1)!} + \mathcal{O}\left(t^{-r} (\log t)^{d-2}\right), \quad t \gg 1. \tag{4.5}$$

Furthermore, if $\delta_{r,s,d}(t) = \sum_{\bar{n}_1 \ldots \bar{n}_d > t} (\bar{n}_1 \ldots \bar{n}_d)^{-r-1} \bar{n}^s_j$ for $r > s > 0$ and $j = 1, \ldots, d$, then

$$\delta_{r,s,d}(t) = \frac{1}{r-s} \left\{1 + \zeta(s+1)\right\}^{d-1} t^{s-r} + \begin{cases} \mathcal{O}\left(t^{-r} (\log t)^{d-1}\right) & 0 < s \leq 1 \\ \mathcal{O}\left(t^{s-r-1}\right) & s > 1. \end{cases} \tag{4.6}$$

Proof. By induction on d. For $d = 1$ we have $\gamma_{r,1}(t) = \sum_{n>t} n^{-r-1} = \frac{t^{-r}}{r} + \mathcal{O}\left(t^{-r-1}\right)$ for large t. Now assume that the result is true up to d. Then

$$\begin{aligned}
\gamma_{r,d}(t) &= \gamma_{r,d-1}(t) + \sum_{n=1}^{t} n^{-r-1} \gamma_{r,d-1}\left(\frac{t}{n}\right) + \sum_{n>t} n^{-r-1} \gamma_{r,d-1}(1) \\
&= \sum_{n=1}^{t} n^{-r-1} \gamma_{r,d-1}\left(\frac{t}{n}\right) + \mathcal{O}\left(t^{-r} (\log t)^{d-2}\right) \\
&= \frac{t^{-r-1}}{r(d-2)!} \sum_{n=1}^{t} \frac{t}{n} \left[\log\left(\frac{t}{n}\right)\right]^{d-2} + \mathcal{O}\left(t^{-r} (\log t)^{d-2}\right) \\
&= \frac{t^{-r-1}}{r} \theta_d(t) + \mathcal{O}\left(t^{-r} (\log t)^{d-2}\right) = \frac{t^{-r} (\log t)^{d-1}}{r(d-1)!} + \mathcal{O}\left(t^{-r} (\log t)^{d-2}\right),
\end{aligned}$$

where θ_d is as in Lemma 4.1. Thus we obtain (4.5). Next we consider

$$\begin{aligned}
\delta_{r,s,d}(t) &= \delta_{r,s,d-1}(t) + \sum_{n=1}^{t} n^{-r-1} \delta_{r,s,d-1}\left(\frac{t}{n}\right) + \delta_{r,s,d-1}(1) \sum_{n>t} n^{-r-1} \\
&= \delta_{r,s,d-1}(t) + \sum_{n=1}^{t} n^{-r-1} \delta_{r,s,d-1}\left(\frac{t}{n}\right) + \mathcal{O}\left(t^{-r}\right).
\end{aligned}$$

By the induction hypothesis, the first term is

$$\delta_{r,s,d-1}(t) = \frac{1}{r-s} \left\{1 + \zeta(s+1)\right\}^{d-2} t^{s-r} + \begin{cases} \mathcal{O}\left(t^{-r} (\log t)^{d-2}\right) & 0 < s \leq 1 \\ \mathcal{O}\left(t^{s-r-1}\right) & s > 1. \end{cases}$$

For the second term, we have

$$\begin{aligned}
\sum_{n=1}^{t} n^{-r-1} \delta_{r,s,d-1}\left(\frac{t}{n}\right) =& \frac{1}{r-s} \left\{1 + \zeta(s+1)\right\}^{d-2} \sum_{n=1}^{t} n^{-r-1} \left(\frac{t}{n}\right)^{s-r} \\
&+ \begin{cases} \mathcal{O}\left(t^{-r} (\log t)^{d-2} \sum_{n=1}^{t} n^{-1}\right) & 0 < s \leq 1 \\ \mathcal{O}\left(t^{s-r-1} \sum_{n=1}^{t} n^{-s}\right) & s > 1. \end{cases} \\
=& \frac{1}{r-s} \left\{1 + \zeta(s+1)\right\}^{d-2} \zeta(s+1) t^{s-r} + \begin{cases} \mathcal{O}\left(t^{-r} (\log t)^{d-1}\right) & 0 < s \leq 1 \\ \mathcal{O}\left(t^{s-r-1}\right) & s > 1. \end{cases}
\end{aligned}$$

Combining this and the previous result completes the proof. □

Proof of Theorem 4.4. This follows immediately from the Coefficient bounds corollary and Lemma 4.5. □

Theorem 4.4 indicates that the convergence rate of $\mathcal{F}_N[f]$ using the hyperbolic cross (4.1) is comparable to that of the approximation based on the full index set (3.1). Indeed, for the L^2 and uniform rates we only lose factors of $\mathcal{O}\left((\log N)^{d-1}\right)$ and $\mathcal{O}((\log N)^{\frac{d-1}{2}})$ respectively. The H^s rate, $s \geq 1$, remains the same. Moreover, as evidenced by Corollary 4.2, the hyperbolic cross offers a vast saving in computational cost: forming $\mathcal{F}_N[f]$ involves only $\mathcal{O}\left(N(\log N)^{d-1}\right)$ operations as opposed to $\mathcal{O}\left(N^d\right)$.

As is necessary for hyperbolic cross approximations, additional (mixed) smoothness is required for the estimates of Lemma 4.3 in comparison to those of Lemma 3.4. If only H^r-regularity is imposed, the hyperbolic cross approximation will converge more slowly than its counterpart based on the full index set. However, for approximations based on either the full or hyperbolic cross index set the minimal regularity required to obtain an optimal convergence rate is the same (see Lemma 3.5 and Theorem 4.4 respectively).

It is also of interest to consider the affect of the hyperbolic cross on the pointwise rate of convergence. As we shall see in the next section, this also only deteriorates by a factor of $\mathcal{O}\left((\log N)^{d-1}\right)$. Moreover the smoothness assumption remains the same.

4.2 Pointwise convergence rate

To analyse the pointwise convergence rate of $\mathcal{F}_N[f]$ we consider a slight adjustment of the index set (4.1), namely a *step hyperbolic cross.* To this end, we suppose that $N = 2^r$ and define

$$Q_r = \bigcup_{|\alpha| \leq r} \rho(\alpha), \quad \text{where} \quad \rho(\alpha) = \{n \in \mathbb{N}^d : \lfloor 2^{\alpha_j - 1} \rfloor \leq n_j < 2^{\alpha_j},\ j = 1, \ldots, d\}, \quad \alpha \in \mathbb{N}^d. \tag{4.7}$$

We call Q_r the step hyperbolic cross of size r. Note that we have the inclusion $Q_r \subset I_N \subset Q_{r+d}$, (see, for example [20]), where I_N is the hyperbolic cross index set (4.1).

Theorem 4.6. *Suppose that $f \in \mathrm{H}^{2k+3}_{mix}(-1,1)^d$ obeys the first $k \in \mathbb{N}$ derivative conditions. Suppose further that $N = 2^r$ and that $\mathcal{F}_N[f]$ is the truncated modified Fourier expansion of f using the step hyperbolic cross Q_r. Then $f(x) - \mathcal{F}_N[f](x) = \mathcal{O}\left(N^{-2k-2}(\log N)^{d-1}\right)$ uniformly for x in compact subsets of $(-1,1)^d$.*

Proof. Let $\mathcal{F}_\alpha[f](x) = \sum_{i \in \{0,1\}^d} \sum_{n \in \rho(\alpha)} \hat{f}^{[i]}_n \phi^{[i]}_n(x)$, $\alpha \in \mathbb{N}^d$, so that $\mathcal{F}_N[f] = \sum_{|\alpha| \leq r} \mathcal{F}_\alpha[f]$. We first claim that

$$\mathcal{F}_\alpha[f](x) = \mathcal{O}\left(2^{-2(k+1)|\alpha|}\right), \quad |\alpha| \to \infty. \tag{4.8}$$

To prove this result we use induction on d. The case $d = 1$ is trivial, so we now assume that the result holds for all functions f of at most $(d-1)$ variables. We first recall the asymptotic expansion of $\hat{f}^{[i]}_n$. Since f obeys the first k derivative conditions, an application of Theorem 2.11 yields

$$\hat{f}^{[i]}_n = \sum_{t \in [d]} \mathcal{A}^{[i]}_{k_t, n_{\bar{t}}}[f] \prod_{j \in t} \widehat{p^{[i_j]}}^{[i_j]}_{n_j} + \mathcal{O}\left((n_1 \ldots n_d)^{-2k-3}\right),$$

where $p^{[i_j]}$ is the polynomial defined in Lemma 3.1. The term $\mathcal{A}^{[i]}_{k_t, n_{\bar{t}}}[f]$ is the modified Fourier coefficient of a function $\mathcal{H}^{[i]}_{\bar{t}}[f](x_{\bar{t}})$ that satisfies the first k derivative conditions in the variables $x_{\bar{t}}$. Hence

$$\begin{aligned}
\mathcal{F}_\alpha[f](x) &= \sum_{t \in [d]} \sum_{i \in \{0,1\}^d} \sum_{n \in \rho(\alpha)} \mathcal{A}^{[i]}_{k_t, n_{\bar{t}}}[f] \prod_{j \in t} \widehat{p^{[i_j]}}^{[i_j]}_{n_j} \phi^{[i]}_n(x) + \mathcal{O}\left(2^{-2(k+1)|\alpha|}\right) \\
&= \sum_{t \in [d]} \mathcal{F}_{\alpha_{\bar{t}}}\left[\mathcal{H}^{[i]}_{\bar{t}}\right](x_{\bar{t}}) \prod_{j \in t} \sum_{n_j = \lfloor 2^{\alpha_j - 1} \rfloor}^{2^{\alpha_j} - 1} \widehat{p^{[i_j]}}^{[i_j]}_{n_j} \phi^{[i_j]}_{n_j}(x_j) + \mathcal{O}\left(2^{-2(k+1)|\alpha|}\right).
\end{aligned}$$

Since $p^{[i_j]}$ obeys the first k derivative conditions, an application of the univariate result gives

$$\sum_{n_j = \lfloor 2^{\alpha_j - 1} \rfloor}^{2^{\alpha_j} - 1} \widehat{p^{[i_j]}}^{[i_j]}_{n_j} \phi^{[i_j]}_{n_j}(x_j) = \mathcal{O}\left(2^{-2(k+1)\alpha_j}\right), \quad j = 1, \ldots, d.$$

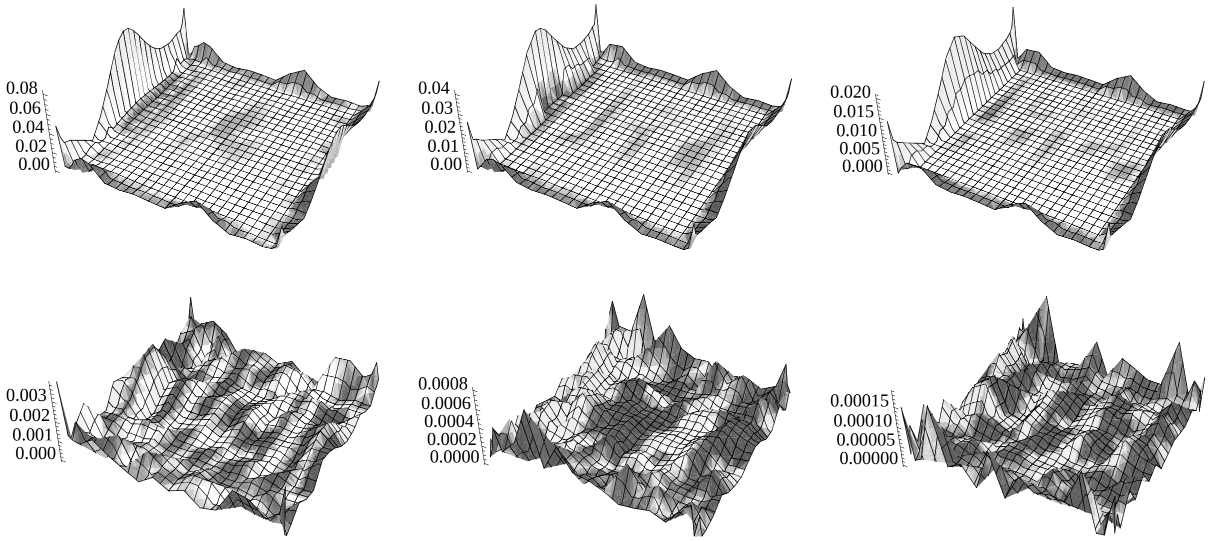
Substituting this into the previous expression and using the induction hypothesis on the term $\mathcal{F}_{\alpha_{\bar{t}}}\left[\mathcal{H}^{[i]}_{\bar{t}}\right](x_{\bar{t}})$ (note that $|\bar{t}| < d$) now gives

$$\mathcal{F}_\alpha[f](x) = \mathcal{O}\left(\sum_{t \in [d]} 2^{-2(k+1)|\alpha_{\bar{t}}|} \prod_{j \in t} 2^{-2(k+1)\alpha_j}\right) = \mathcal{O}\left(2^{-2(k+1)|\alpha|}\right),$$

which completes the first step of the proof.

Since the main result has already been proved in Theorem 3.2 for the approximation $\mathcal{F}_N[f]$ based on the full index set (3.1), it suffices to consider the difference between this and the approximation based on the step hyperbolic cross Q_r. This difference is precisely

$$\sum_{\substack{|\alpha| > r \\ |\alpha|_\infty \leq r}} \mathcal{F}_\alpha[f](x) = \sum_{|\alpha'|_\infty = 0}^{r} \sum_{\alpha_d = r - |\alpha'|}^{r} \mathcal{F}_\alpha[f](x),$$



$N = 50$ $N = 100$ $N = 200$

Figure 1: Absolute error $|f(x, y) - \mathcal{F}_N[f](x, y)|$ for $f(x, y) = \mathrm{e}^{\sin 6x}(y + \tan^2(1 - y^2))$ and $-1 \leq x, y \leq 1$ (top row), $-0.9 \leq x, y \leq 0.9$ (bottom row).

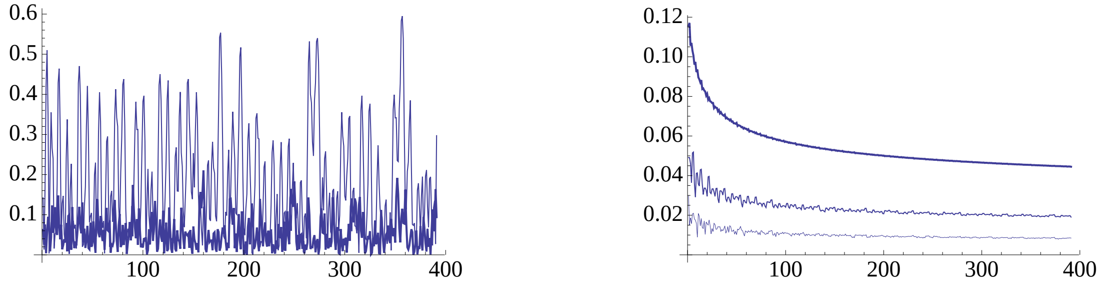


Figure 2: (left) scaled pointwise error $N^2(\log N)^{-1}|f(x, y) - \mathcal{F}_N[f](x, y)|$, $N = 10, \ldots, 400$, for $f(x, y) = \mathrm{e}^{xy}$ and $(x, y) = (-\frac{3}{4}, -\frac{3}{4})$ (thicker line),$(\frac{1}{2}, -\frac{1}{2})$ (thinner line). (right) scaled pointwise error $N(\log N)^{-1}|f(x, y)-\mathcal{F}_N[f](x, y)|$ for $(x, y) = (-1, 1)$, $(-\frac{1}{2}, -1)$, $(-1, -\frac{1}{4})$.

where $\alpha' = (\alpha_1, \ldots, \alpha_{d-1})$ is the first $(d - 1)$ entries of α. Hence, using (4.8), it follows that

$$\sum_{\substack{|\alpha|>r \\ |\alpha|_\infty \leq r}} \mathcal{F}_\alpha[f](x) = \mathcal{O}\left(\sum_{|\alpha'|_\infty=0}^{r} 2^{-2(k+1)|\alpha'|} \sum_{\alpha_d=r-|\alpha'|}^{r} 2^{-2(k+1)\alpha_d}\right)$$

$$= \mathcal{O}\left(\sum_{|\alpha'|_\infty=0}^{r} 2^{-2(k+1)|\alpha'|} 2^{-2(k+1)(r-|\alpha'|)}\right) = \mathcal{O}\left(r^{d-1} 2^{-2(k+1)r}\right),$$

which completes the proof. □

The inclusion $Q_r \subset I_N \subset Q_{r+d}$ indicates that this is also the case for the hyperbolic cross (4.1).

4.3 Numerical Results

In Figures 1 and 2 we give numerical results for the bivariate modified Fourier approximation $\mathcal{F}_N[f](x,y)$ using the hyperbolic cross (4.1). As Figure 1 demonstrates, the error inside the domain is much smaller than on the boundary, as predicted by Theorem 4.6. Observe further that when N doubles the uniform error roughly halves, whereas the error in $[-0.9,-0.9]\times[-0.9,-0.9]$ roughly quarters, again as predicted. Figure 2 verifies the results of Theorems 4.4 and 4.6: namely, the convergence rate is $\mathcal{O}\left(N^{-2}(\log N)\right)$ for $(x,y)\in(-1,1)^2$ and $\mathcal{O}\left(N^{-1}(\log N)\right)$ on the boundary.

The faster convergence rate of modified Fourier expansions away from the boundary indicates a weak Gibbs phenomenon. Indeed, further analysis along the lines of that given in this section demonstrates that, for a function $f\in\mathrm{H}^{2k+1}_{\mathrm{mix}}(-1,1)^d$ obeying the first $k\in\mathbb{N}$ derivative conditions, the derivative $\mathrm{D}^\alpha\mathcal{F}_N[f]$, $|\alpha|_\infty=2k+1$, converges pointwise to $\mathrm{D}^\alpha f$ away from the boundary but not uniformly on $[-1,1]^d$. This indicates a Gibbs phenomenon in the $(2k+1)^{\text{th}}$ derivative. For a general function f obeying no derivative conditions, this translates as pointwise, but nonuniform convergence of any first order partial derivative of $\mathcal{F}_N[f]$.

5 The Modified Fourier–Galerkin method

In this section we consider one particular application of modified Fourier series: namely, the discretization of Neumann boundary value problems (Section 6 deals with the discretization of problems with other boundary conditions). Our aim is to provide analysis and, by both theory and numerical example, establish a number of advantages of this approach over more standard methods. As we discuss further in Section 7, future work is required to develop efficient, robust numerical algorithms based on modified Fourier series.

For the moment, we consider the Neumann boundary value problem

$$\mathcal{L}[u](x)=-\triangle u(x)+a\cdot\nabla u(x)+bu(x)=f(x),\quad x\in[-1,1]^d,\quad \frac{\partial u}{\partial n}\Big|_{\Gamma}=0. \tag{5.1}$$

where $a=(a_1,\ldots,a_d)^\top\in\mathbb{R}^d$ and $b\in\mathbb{R}$ are constants (we consider the general case in the sequel) and f is some given function. Equivalently, in weak form, if $T:\mathrm{H}^1(-1,1)^d\times\mathrm{H}^1(-1,1)^d\to\mathbb{R}$ is the bilinear form

$$T(u,v)=(\nabla u,\nabla v)+(a\cdot\nabla u,v)+b(u,v),\quad \forall u,v\in\mathrm{H}^1(-1,1)^d,$$

where $(\nabla u,\nabla v)=\int_{(-1,1)^d}\nabla u\cdot\nabla v$, we may rewrite (5.1) as

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

We shall use the Lax–Milgram Theorem and Céa's Lemma (see, for example [7, 23]) so it useful to know that the operator T is continuous and coercive if and only if $b-\frac{1}{4}\|a\|^2>0$, where $\|a\|^2=\sum_{i=1}^d a_i^2$. In this case there are positive constants ω and γ such that

$$|T(u,v)|\leq\gamma\|u\|_1\|v\|_1,\quad T(u,u)\geq\omega\|u\|_1^2,\quad \forall u,v\in\mathrm{H}^1(-1,1)^d. \tag{5.2}$$

5.1 Galerkin's equations and iterative solution techniques

We consider the modified Fourier–Galerkin approximation of (5.1). Suppose that we write such approximation $u_N\in\mathcal{S}_N$ as

$$u_N(x)=\sum_{i\in\{0,1\}^d}\sum_{n\in I_N}\bar{u}^{[i]}_n\phi^{[i]}_n(x),\quad x\in[-1,1]^d,$$

where I_N is some appropriate index set. The coefficients $\bar{u}^{[i]}_n\in\mathbb{R}$ enforce Galerkin's equations $T(u_N,\phi)=(f,\phi)$, $\forall\phi\in\mathcal{S}_N$. We have:

Lemma 5.1. *The coefficients $\bar{u}^{[i]}_n$ satisfy*

$$(b+\mu^{[i]}_n)\bar{u}^{[i]}_n+\sum_{j=1}^d\sum_{\substack{m_j\in\mathbb{N},\\(n;m_j)\in I_N}}a_j\delta^{[i_j]}_{n_j,m_j}\bar{u}^{[(i;1-i_j)]}_{(n;m_j)}=\hat{f}^{[i]}_n,\quad i\in\{0,1\}^d,\quad n\in I_N, \tag{5.3}$$

where

$$(n;m_j)=(n_1,\ldots,n_{j-1},m_j,n_{j+1},\ldots,n_d),\quad (i;1-i_j)=(i_1,\ldots,i_{j-1},1-i_j,i_{j+1},\ldots,i_d),$$

and

$$\delta^{[i]}_{n,m}=\int_{-1}^{1}\phi^{[i]}_n(x)(\phi^{[1-i]}_m)'(x)\,\mathrm{d}x=2(-1)^{n+m}\frac{\mu^{[1-i]}_m}{\mu^{[i]}_n-\mu^{[1-i]}_m},\quad i=0,1,\quad n,m\in\mathbb{N}. \tag{5.4}$$

Proof. We set $\phi=\phi^{[i]}_n$, $i\in\{0,1\}^d$, $n\in I_N$ in Galerkin's equations. Due to the Laplace term, we obtain

$$T(u_N,\phi^{[i]}_n)=(b+\mu^{[i]}_n)\bar{u}^{[i]}_n+\sum_{j=1}^{d}\sum_{l\in\{0,1\}^d}\sum_{m\in I_N}a_j(\partial_{x_j}\phi^{[l]}_m,\phi^{[i]}_n)\bar{u}^{[l]}_m.$$

Now

$$(\partial_{x_j}\phi^{[l]}_m,\phi^{[i]}_n)=((\phi^{[l_j]}_{m_j})',\phi^{[i_j]}_{n_j})\prod_{k\neq j}(\phi^{[l_k]}_{m_k},\phi^{[i_k]}_{n_k})=\begin{cases}\delta^{[i_j]}_{n_j,m_j} & l=(i;1-i_j),\quad m_k=n_k,\quad k\neq j,\\ 0 & \text{otherwise},\end{cases}$$

which gives the result. □

For spectral discretizations in Cartesian product domains, Galerkin's equations are normally written in tensor product form. The advantage of this approach is that it facilitates the use of novel solution techniques such as the matrix diagonalization and Schur decomposition methods [7]. Furthermore, the matrices involved, which in this case would correspond to univariate modified Fourier discretizations, are well understood and have a number of beneficial properties [1]. However, we shall not pursue this approach: for approximations using a hyperbolic cross index set, Galerkin's equations do not naturally have a tensor product form.

Instead we consider standard iterative methods. Suppose that we write the discretization matrix as A_{G} and Galerkin's equations as $A_{\mathrm{G}}\bar{u}=\hat{f}$. In addition, we decompose $A_{\mathrm{G}}=M_{\mathrm{G}}+N_{\mathrm{G}}$, where M_{G} is the diagonal matrix corresponding to restriction of the operator $-\triangle+b\mathcal{I}$ to $\mathcal{S}_N$ and $\mathcal{I}$ is the identity operator. As we demonstrate in the sequel, the matrix M_{G} is an optimal preconditioner for A_{G}. Hence Galerkin's equations can be solved via preconditioned conjugate gradients. Moreover, the number of iterations required for convergence within some numerical tolerance is independent of the truncation parameter N.

This fact is independent of the discretization basis. However, in the modified Fourier setting the matrix M_{G} is diagonal (with n^{th} entry $b+\mu^{[i]}_n$), making this scheme practical. The overall cost is thus determined by the number of operations required to perform matrix-vector multiplications involving A_{G}. We have:

Lemma 5.2. *For $N\gg d$ the number of non-zero entries of the matrix A_{G} is $d2^dN^{d+1}+\mathcal{O}\left(N^d\right)$ in the case of the full index set (3.1) and $d2^dN^2\lceil(1+\zeta(2))^{d-1}\rceil+\mathcal{O}\left(N(\log N)^{d-1}\right)$ for the hyperbolic cross (4.1).*

Proof. In view of Lemma 5.1, the number of non-zero matrix entries is

$$\sum_{i\in\{0,1\}^d}\sum_{n\in I_N}\sum_{j=1}^{d}\sum_{\substack{m_j\in\mathbb{N},\\(n;m_j)\in I_N}}1+\mathcal{O}\left(|I_N|\right).$$

If I_N is the full index set, we easily obtain the result. For the hyperbolic cross (4.1) we have

$$\begin{aligned}&\sum_{i\in\{0,1\}^d}\sum_{n\in I_N}\sum_{j=1}^{d}\sum_{\substack{m_j\in\mathbb{N},\\(n;m_j)\in I_N}}1\\ &\qquad=d2^d\sum_{n\in I_N}\sum_{\substack{m_d\in\mathbb{N},\\(n;m_d)\in I_N}}1+\mathcal{O}\left(N(\log N)^{d-1}\right)=d2^d\sum_{n\in I_N}\sum_{m=0}^{N(\bar{n}_1\ldots\bar{n}_{d-1})^{-1}}1+\mathcal{O}\left(N(\log N)^{d-1}\right)\\ &\qquad=d2^d\sum_{n\in I_N}\frac{N}{\bar{n}_1\ldots\bar{n}_{d-1}}+\mathcal{O}\left(N(\log N)^{d-1}\right)=d2^dN^2\sum_{\bar{n}_1,\ldots,\bar{n}_{d-1}=1}^{\infty}\frac{1}{(\bar{n}_1\ldots\bar{n}_{d-1})^2}+\mathcal{O}\left(N(\log N)^{d-1}\right).\end{aligned}$$

Evaluating this final sum completes the proof. □

In light of Lemma 5.2 we conclude that Galerkin's equations can be solved in $\mathcal{O}\left(N^{d+1}\right)$ (full index set) or $\mathcal{O}\left(N^2\right)$ (hyperbolic cross) operations by conjugate gradients and direct evaluation of matrix-vector products. However, since the action of the matrix A_{G} corresponds to finding modified Fourier coefficients of derivatives of finite modified Fourier sums, a variant of the Fast Fourier Transform (FFT) can be employed in the full index set case. In this manner the figure of $\mathcal{O}\left(N^{d+1}\right)$ can easily be reduced to $\mathcal{O}\left(N^d \log N\right)$. For the hyperbolic cross index set, a variant of the sparse grid FFT could be employed [4, 10]. In this manner, the figure of $\mathcal{O}\left(N^2\right)$ could be reduced to $\mathcal{O}\left(N(\log N)^d\right)$. However, this technique is neither easy nor straightforward to implement [15].

5.2 Properties of the discretization matrix

The properties of A_{G}, in particular the L^2 and spectral condition numbers and the existence of effective preconditioners, are of importance in spectral discretizations. In this section, we demonstrate that both the L^2 and spectral condition numbers of the modified Fourier–Galerkin discretization matrix are $\mathcal{O}\left(N^2\right)$ and that the diagonal matrix M_{G} is an optimal preconditioner. The results of this section are extensions of those found in [1].

Lemma 5.3. *Suppose that I_N is either the full or the hyperbolic cross index set. Then the spectral condition number of A_{G} is $\mathcal{O}\left(N^2\right)$ provided the operator $\mathcal{L}$ is coercive. Specifically, if λ is an eigenvalue of A_{G} then*

$$\omega \leq |\lambda| \leq \gamma(1+N^2\pi^2 d), \quad \omega \leq |\lambda| \leq \gamma(1+(d-1+N^2)\pi^2),$$

in the full and hyperbolic cross cases respectively.

Proof. For an eigenvalue λ with eigenfunction $u \in \mathcal{S}_N$ we have $\lambda(u,\phi) = T(u,\phi)$, $\forall \phi \in \mathcal{S}_N$. In particular, $\omega\|u\|^2 \leq |\lambda|\|u\|^2$ and $|\lambda|\|u\|^2 \leq \gamma\|u\|_1^2$. Now, by Bernstein's Inequality (Corollary 2.6), $\|u\|_1^2 \leq \max_{n\in I_N}\{1+\mu_n^{[0]}\}\|u\|^2$. Moreover, for $n \in I_N$,

$$1+\mu_n^{[0]} \leq 1+N^2\pi^2 d, \quad 1+\mu_n^{[0]} \leq 1+(d-1+N^2)\pi^2, \tag{5.5}$$

where I_N is either the full or hyperbolic cross index set respectively. □

We may also prove the same result for the L^2 condition number. To do so we need the following lemma:

Lemma 5.4. *Suppose that λ is an eigenvalue of $A_{\mathrm{G}}^\top A_{\mathrm{G}}$ with associated eigenfunction $u \in \mathcal{S}_N$. Then*

$$(\mathcal{F}_N[\mathcal{L}[u]], \mathcal{F}_N[\mathcal{L}[\phi]]) = \lambda(u,\phi), \quad \forall \phi \in \mathcal{S}_N. \tag{5.6}$$

Proof. This is a trivial generalisation of the proof given in [1], so is not presented here. □

Corollary 5.5. *Suppose that I_N is either the full or the hyperbolic cross index set. Then the* L^2 *condition number of A_{G}, $\kappa(A_{\mathrm{G}})$, is $\mathcal{O}\left(N^2\right)$ provided the operator $\mathcal{L}$ is coercive. Specifically, if $\gamma' > 0$ is such that $\|\mathcal{L}[u]\|^2 \leq (\gamma')^2\|u\|_2^2$ for all $u \in \mathrm{H}^2(-1,1)^d$, then we have the bounds*

$$\kappa(A_{\mathrm{G}}) \leq \omega^{-1}\gamma'(1+N^2\pi^2 d), \quad \kappa(A_{\mathrm{G}}) \leq \omega^{-1}\gamma'(1+(d-1+N^2)\pi^2),$$

in the full and hyperbolic cross cases respectively.

Proof. Setting $\phi = u$ in (5.6) gives $\|\mathcal{F}_N[\mathcal{L}[u]]\|^2 = \lambda\|u\|^2$. Now

$$\|\mathcal{F}_N[\mathcal{L}[u]]\| = \sup_{g\in \mathrm{L}^2(-1,1)^d} \frac{(\mathcal{F}_N[\mathcal{L}[u]], g)}{\|g\|} \geq \sup_{g\in\mathcal{S}_N} \frac{(\mathcal{L}[u], g)}{\|g\|}. \tag{5.7}$$

Suppose that we define $g \in \mathcal{S}_N$ by enforcing the condition $(\mathcal{L}[\phi], g) = (\phi, u)$ for all $\phi \in \mathcal{S}_N$. Note that the coefficients of g are the solution of a linear system involving $A_{\mathrm{G}}^\top$. Hence, existence and uniqueness of g is guaranteed. Furthermore $(\mathcal{L}[u], g) = (u,u) = \|u\|^2$ and, since $\mathcal{L}$ is coercive, $\omega\|g\|_1 \leq \|u\|$. Thus

$$\lambda\|u\|^2 = \|\mathcal{F}_N[\mathcal{L}[u]]\|^2 \geq \left[\frac{(\mathcal{L}[u], g)}{\|g\|}\right]^2 = \frac{\|u\|^4}{\|g\|^2} \geq \omega^2\|u\|^2.$$

To derive an upper bound for λ, we note that

$$\lambda\|u\|^2 = \|\mathcal{F}_N[\mathcal{L}[u]]\|^2 \leq \|\mathcal{L}[u]\|^2 \leq (\gamma')^2\|u\|_2^2 \leq (\gamma')^2 \max_{n\in I_N}\{1+\mu_n^{[0]}\}^2\|u\|^2,$$

by Bernstein's Inequality. The result now follows immediately from (5.5). □

We remark in passing that the lower bounds for the minimal eigenvalues of A_{G} and $A_{\mathrm{G}}^{\top}A_{\mathrm{G}}$ are independent of the Galerkin discretization used. The upper bounds, however, rely on Bernstein-type estimates which are dependent on both the discretization basis and index set employed.

We complete this section by demonstrating that M_{G} is an optimal preconditioner for A_{G}. To do so, we first require the following lemma:

Lemma 5.6. *Suppose that the operator $\mathcal{L}$ is $\mathrm{H}^1(-1,1)^d$-continuous and coercive and that $\mathcal{L}_0 = -\triangle + b\mathcal{I}$. Then, there exists a constant $\omega' > 0$ such that*

$$(\mathcal{L}[u], \mathcal{L}_0[u]) \geq (\omega')^2\|u\|_2^2,$$

for all $u \in \mathrm{H}^2(-1,1)^d$ satisfying $\frac{\partial u}{\partial n}|_\Gamma = 0$.

Proof. We have $(\mathcal{L}[u], \mathcal{L}_0[u]) = \|\mathcal{L}_0[u]\|^2 + (a.\nabla u, \mathcal{L}_0[u])$. Now $\|\mathcal{L}_0[u]\|^2 = \|\triangle u\|^2 + 2b\|\nabla u\|^2 + b^2\|u\|^2$ and

$$|(a.\nabla u, \mathcal{L}_0[u])| \leq |(a.\nabla u, \triangle u)| + b|(a.\nabla u, u)| \leq \|a\|\|\nabla u\|\|\triangle u\| + b\|a\|\|\nabla u\|\|u\|.$$

Using Young's inequality ($xy \leq \epsilon x^2 + \frac{1}{4\epsilon}y^2$, $\forall x, y \in \mathbb{R}, \epsilon > 0$) we obtain

$$|(a.\nabla u, \mathcal{L}_0[u])| \leq \epsilon\|\triangle u\|^2 + \tfrac{\|a\|^2}{2\epsilon}\|\nabla u\|^2 + b^2\epsilon\|u\|^2, \quad \forall \epsilon > 0.$$

Substituting this into the previous expression now gives

$$(\mathcal{L}[u], \mathcal{L}_0[u]) \geq (1-\epsilon)\|\triangle u\|^2 + 2\left(b - \tfrac{\|a\|^2}{4\epsilon}\right)\|\nabla u\|^2 + b^2(1-\epsilon)\|u\|^2.$$

If we set $\epsilon = \|a\|^2(2b + \frac{1}{2}\|a\|^2)^{-1}$ then

$$(\mathcal{L}[u], \mathcal{L}_0[u]) \geq \left(\frac{b - \frac{1}{4}\|a\|^2}{b + \frac{1}{4}\|a\|^2}\right)\|\triangle u\|^2 + \left(b - \tfrac{1}{4}\|a\|^2\right)\|\nabla u\|^2 + b^2\left(\frac{b - \frac{1}{4}\|a\|^2}{b + \frac{1}{4}\|a\|^2}\right)\|u\|^2,$$

which yields the result. □

Theorem 5.7. *Suppose that A_{G} is the modified Fourier–Galerkin matrix. Then the right preconditioner M_{G} is optimal for both the spectral and L^2 condition numbers, provided the operator $\mathcal{L}$ is coercive. Specifically,*

$$\omega'(\max\{1, 2b, b^2\})^{-\frac{1}{2}} \leq \kappa(A_{\mathrm{G}}M_{\mathrm{G}}^{-1}) \leq \gamma'\left(\min\{1, 2b, b^2\}\right)^{-\frac{1}{2}},$$

and if λ is an eigenvalue of $A_{\mathrm{G}}M_{\mathrm{G}}^{-1}$, $\omega(\max\{b,1\})^{-1} \leq |\lambda| \leq \gamma(\min\{b,1\})^{-1}$.

Proof. Suppose that λ is an eigenvalue of $A_{\mathrm{G}}M_{\mathrm{G}}^{-1}$ with eigenfunction $u \in \mathcal{S}_N$. Suppose that $u = (-\triangle + b\mathcal{I})v$ for some $v \in \mathcal{S}_N$. Then $(\mathcal{L}[v], \phi) = \lambda(\mathcal{L}_0[v], \phi)$, $\forall \phi \in \mathcal{S}_N$. Setting $\phi = v$ gives $(\mathcal{L}[v], v) = \lambda(\mathcal{L}_0[v], v)$. It is trivial to show that the operator $\mathcal{L}_0$ is continuous and coercive provided $b > 0$, with constants $\max\{b,1\}$ and $\min\{b,1\}$ respectively. Hence $\omega(\max\{b,1\})^{-1} \leq |\lambda| \leq \gamma(\min\{b,1\})^{-1}$, as required.

Now suppose that λ is an eigenvalue of $(A_{\mathrm{G}}M_{\mathrm{G}}^{-1})^{\top}(A_{\mathrm{G}}M_{\mathrm{G}}^{-1})$ with eigenfunction $u \in \mathcal{S}_N$. Then, using (5.6) we obtain $\|\mathcal{F}_N[\mathcal{L}[v]]\|^2 = \lambda\|\mathcal{L}_0[v]\|^2$, where $u = (-\triangle + b\mathcal{I})v$ once more. Note that $\|\mathcal{L}_0[u]\|^2 \leq \max\{1, 2b, b^2\}\|u\|_2^2$ and $\mathcal{L}_0[u]\|^2 \geq \min\{1, 2b, b^2\}\|u\|$, for all $u \in \mathrm{H}^2(-1,1)^d$ satisfying $\frac{\partial u}{\partial n}|_\Gamma = 0$. Hence

$$\min\{1, 2b, b^2\}\lambda\|v\|_2^2 \leq \|\mathcal{F}_N[\mathcal{L}[v]]\|^2 \leq \|\mathcal{L}[v]\|^2 \leq (\gamma')^2\|v\|_2^2,$$

which yields $\kappa(A_{\mathrm{G}}M_{\mathrm{G}}^{-1}) \leq \gamma'(\min\{1, 2b, b^2\})^{-\frac{1}{2}}$. To provide a lower bound we use (5.7) with $g = \mathcal{L}_0[v]$ to give $\|\mathcal{F}_N[\mathcal{L}[v]]\| \geq \frac{(\mathcal{L}[v], \mathcal{L}_0[v])}{\|\mathcal{L}_0[v]\|}$. We now use Lemma 5.6 to give

$$\|\mathcal{F}_N[\mathcal{L}[v]]\|^2 \geq \frac{(\omega')^2}{\max\{1, 2b, b^2\}}\|v\|_2^2.$$

Hence $\kappa(A_{\mathrm{G}}M_{\mathrm{G}}^{-1}) \geq \omega'(\max\{1, 2b, b^2\})^{-\frac{1}{2}}$ and the proof is complete. □

5.3 Convergence rate and numerical results

The results of Sections 2–4 allow us to immediately provide estimates for the convergence rate of the approximation u_N in the H^1 norm. From Céa's lemma, $\|u-u_N\|_1 \leq \gamma\omega^{-1}\inf_{\phi\in\mathcal{S}_N}\|u-\phi\|_1$. For modified Fourier series, in light of Lemma 2.4, this infimum is precisely $\|u-\mathcal{F}_N[u]\|_1$. Since the solution u automatically obeys the first derivative condition, we obtain

Theorem 5.8. *Suppose that u_N is the modified Fourier–Galerkin approximation based on the full index set (3.1). Then*

$$\|u-u_N\|_1 \leq \gamma\omega^{-1}c_{r,1}N^{1-r}\|u\|_r, \quad r=1,2,3, \quad \|u-u_N\|_1 \leq \gamma\omega^{-1}c_1 N^{-\frac{5}{2}}\|u\|_{4,\mathrm{mix}},$$

where $c_{r,1}$ and c_1 are the constants from Lemmas 3.4 and 3.5 respectively. If u_N is the approximation based on the hyperbolic cross (4.1) then

$$\|u-u_N\|_1 \leq \gamma\omega^{-1}c_{r,1}N^{\frac{1-r}{d}}\|u\|_r, \quad \|u-u_N\|_1 \leq \gamma\omega^{-1}c_{r,1}N^{1-r}\|u\|_{r,\mathrm{mix}}, \quad r=1,2,3,$$

and $\|u-u_N\|_1 \leq \gamma\omega^{-1}c_1N^{-\frac{5}{2}}\|u\|_{4,\mathrm{mix}}$, where $c_{r,1}$ and c_1 are the constants from Lemma 4.3 and Theorem 4.4 respectively.

As in Section 4, when u does not have sufficient regularity the method based on the hyperbolic cross converges at a slower rate than its full index set counterpart. However, provided the solution $u \in \mathrm{H}^4_{\mathrm{mix}}(-1,1)^d$, both the full and hyperbolic cross index sets offer the same convergence rate. Since the latter involves far reduced complexity, we shall focus on it in the remainder of this paper.

In Figure 3 we give numerical results for the modified Fourier–Galerkin method applied to the problems:

$$\text{(P1)} \quad d=2, a_1=-1, a_2=2, b=4,$$
$$u(x,y) = \mathrm{e}^{xy} - \frac{y}{4}\left[(1+x)^2\mathrm{e}^{y}+(1-x)^2\mathrm{e}^{-y}\right] - \frac{x}{4}\left[\mathrm{e}^{x}(1+y)^{+}(1-y)^2\mathrm{e}^{-x}\right] + \frac{\mathrm{e}}{8}\left[(1-x)^2(1-y)^2+(1+x)^2(1+y)^2\right],$$

$$\text{(P2)} \quad d=3, a_1=-1, a_2=2, a_3=1, b=5,$$
$$u(x,y,z) = \frac{1}{8}\left[3+\mathrm{e}^{\frac{1}{8}(1+x)} - \tfrac{1}{32}(x+1)\left(3-x+\mathrm{e}^{\frac{1}{4}}(x+1)\right)\right] \times \left[\sin\tfrac{1}{2}(y+1)-\tfrac{1}{8}(y+1)\left(3+\cos 1+(\cos 1-1)y\right)\right](z-2)(z+1)^2.$$

Figure 3(c) confirms Theorem 5.8 for these examples. Figures 3(a),(b) indicate that the uniform error of this method is $\mathcal{O}\left(N^{-3}(\log N)^{d-1}\right)$, precisely the same as for function approximation using modified Fourier series (note that, unless $a=0$, $u_N \neq \mathcal{F}_N[u]$, so the results of Section 4 do not apply directly). However, unlike the latter, the modified Fourier–Galerkin method does not offer faster convergence inside the domain.

Concerning the rate of uniform convergence, we have:

Theorem 5.9. *Suppose that $u \in \mathrm{L}^\infty[-1,1]^d \cap \mathrm{H}^1(-1,1)^d$ and that u_N is the modified Fourier–Galerkin approximation based on the hyperbolic cross index set (4.1). Then, for some positive constant c independent of u and N,*

$$\|u-u_N\|_\infty \leq cN^{\frac{1}{2}-\frac{1}{d}}(\log N)^{\frac{d-1}{2}}\|u-u_N\|_1 + \|u-\mathcal{F}_N[u]\|_\infty.$$

Proof. Theorem 4.3 gives that $\|v-\mathcal{F}_N[v]\|_1 \leq cN^{-\frac{1}{d}}\|v\|_2$ for any $v \in \mathrm{H}^2(-1,1)^d$ satisfying the first derivative condition. By a standard duality argument [13, p.190], we thus have $\|u-u_N\| \leq cN^{-\frac{1}{d}}\|u-u_N\|_1$. Writing $e_N = \mathcal{F}_N[u]-u_N$, this result yields $\|e_N\| \leq cN^{-\frac{1}{d}}\|u-u_N\|_1$. Further, for any $\phi \in \mathcal{S}_N$ we have

$$\|\phi\|_\infty \leq \sum_{i\in\{0,1\}^d}\sum_{n\in I_N}|\hat{\phi}^{[i]}_n| \leq \left(\sum_{i\in\{0,1\}^d}\sum_{n\in I_N}1\right)^{\frac{1}{2}}\left(\sum_{i\in\{0,1\}^d}\sum_{n\in I_N}|\hat{\phi}^{[i]}_n|^2\right)^{\frac{1}{2}}$$
$$\leq c|I_N|^{\frac{1}{2}}\|\phi\| \leq cN^{\frac{1}{2}}(\log N)^{\frac{d-1}{2}}\|\phi\|.$$

Since $e_N \in \mathcal{S}_N$ we obtain $\|e_N\|_\infty \leq cN^{\frac{1}{2}-\frac{1}{d}}(\log N)^{\frac{d-1}{2}}\|u-u_N\|_1$. A simple application of the triangle inequality $\|u-u_N\|_\infty \leq \|e_N\|_\infty + \|u-\mathcal{F}_N[u]\|_\infty$ now yields the result. □

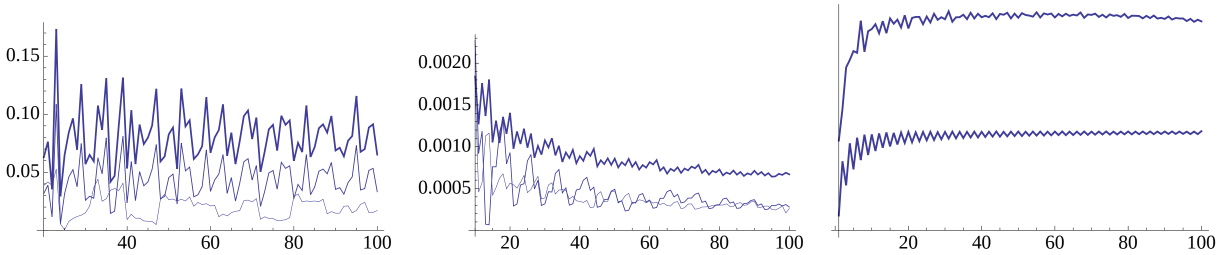


Figure 3: (left) scaled pointwise error $N^3(\log N)^{-1}|u(x,y)-u_N(x,y)|$ for the problem (P1), where $(x,y)=(1,-1)$ (thickest line), $(-1,1)$, $(0,-\frac{1}{4})$ (thinnest line). (middle) scaled pointwise error $N^3(\log N)^{-2}|u(x,y,z)-u_N(x,y,z)|$ for (P2), where $(x,y,z)=(1,1,1)$ (thickest line), $(\frac{3}{4},\frac{3}{4},\frac{3}{4})$, $(\frac{11}{20},\frac{11}{20},\frac{11}{20})$ (thinnest line). (right) Scaled H^1 error $N^{\frac{5}{2}}\|u-u_N\|_1$ for (P1) (top) and (P2) (bottom).

Corollary 5.10. *Suppose that u_N is the modified Fourier–Galerkin approximation based on the hyperbolic cross (4.1) and that $u\in\mathrm{H}^4_{mix}(-1,1)^d$. Then $\|u-u_N\|_\infty\leq cN^{-2-\frac{1}{d}}(\log N)^{\frac{d-1}{2}}\|u\|_{4,mix}$ for some positive constant c independent of u and N.*

When $d=1$, as observed in [1], this result conforms with numerical examples. However, in light of Figure 3, this result is non-optimal for $d\geq 2$.

5.4 Numerical comparison

Standard methods based on Chebyshev or Legendre polynomials yield spectral convergence whenever the solution is smooth. Conversely, the modified Fourier method converges slowly unless the solution u obeys higher order derivative conditions. However, due to its lower complexity, for certain examples the modified Fourier method offers a lower error for moderate values of N. We now consider three such examples, with parameters $d=3$, $b=2$, $a=0$ and exact solutions

$$u(x,y,z)=\sin(2x(2x^2-2)^2)(\sin y-y\cos 1)(z^5-5z), \tag{5.8}$$

$$u(x,y,z)=\mathrm{e}^{z^2\cos 4y+x^2}-p(x,y,z), \tag{5.9}$$

$$u(x,y,z)=x^2\cos(y\sin 5x)\cosh z-p(x,y,z), \tag{5.10}$$

respectively. Note that in (5.9) and (5.10) the function p interpolates the Neumann data of the functions $v(x,y,z)=x^2\cos(y\sin 5x)\cosh z$ and $v(x,y,z)=\mathrm{e}^{z^2\cos 4y+x^2}$:

$$\begin{aligned}p(x,y,z)=&\frac{1}{2}\left[v_x(1,y,z)x^2+v_y(x,1,z)y^2+v_z(x,y,1)z^2\right]\\&-\frac{1}{4}\left[v_{xy}(1,1,z)x^2y^2+v_{xz}(1,y,1)x^2z^2+v_{yz}(x,1,1)y^2z^2\right]+\frac{1}{8}v_{xyz}(1,1,1)x^2y^2z^2.\end{aligned}$$

In Figure 4 we plot the error against number of terms for this method and the Legendre–Galerkin approximation [12, 26] (the Chebyshev–Galerkin approximation [25] gives similar results). As is evident, the modified Fourier method offers a smaller error until the number of approximation coefficients is moderately large. In particular, at least 3375 terms are required before the Legendre approximations to (5.8)–(5.10), which involve $\mathcal{O}\left(N^3\right)$ coefficients in comparison to $\mathcal{O}\left(N(\log N)^2\right)$, become superior. For $d>3$, this effect will become more pronounced: due to its $\mathcal{O}\left(N^d\right)$ terms and $\mathcal{O}\left(N^{d+1}\right)$ complexity, the Legendre method becomes impractical for such higher dimensional problems.

Note that these plots do not take into account the operational cost of each method. As we know from the previous discussion, constructing the modified Fourier–Galerkin approximation involves $\mathcal{O}\left(N^2\right)$ operations, whereas for the Legendre method, even if the coefficients of f are known exactly, this value is $\mathcal{O}\left(N^4\right)$ [26]. Thus the modified Fourier method is likely to perform even better if we were to take this into account. Having said this, we note that a central issue concerning the modified Fourier method is the computation

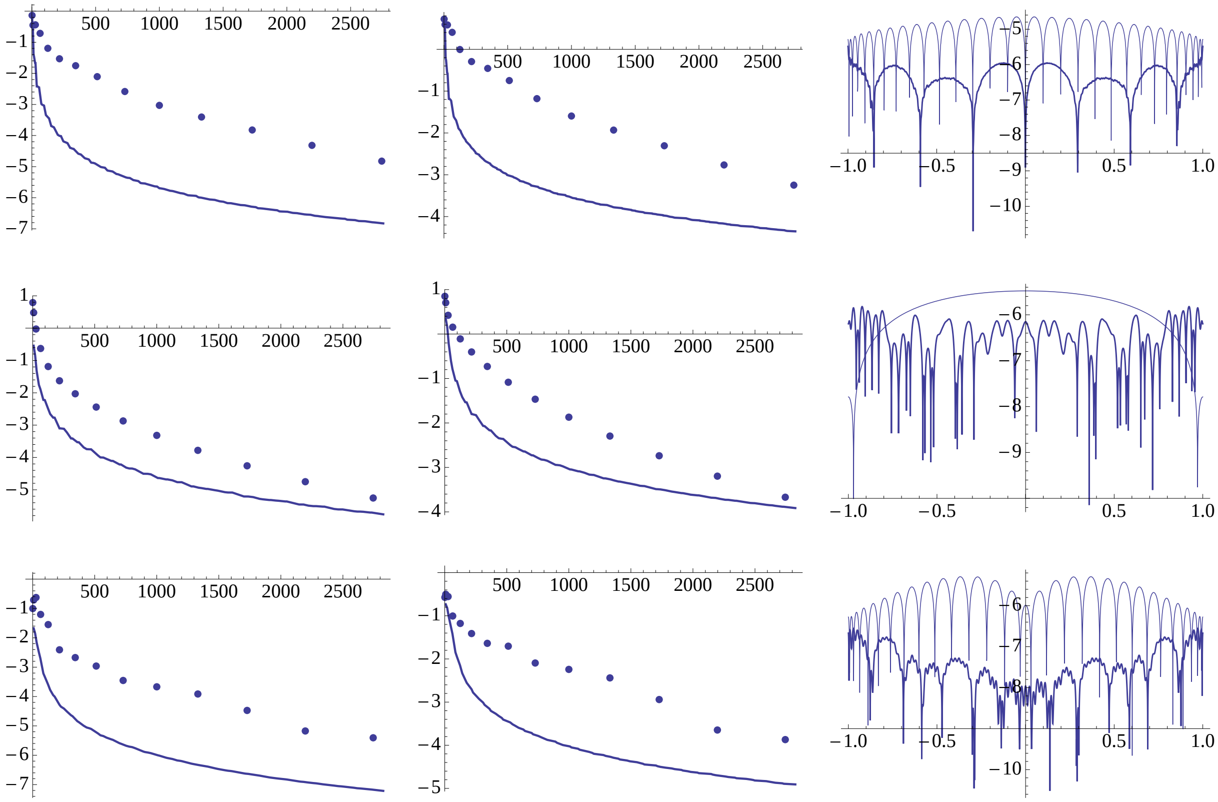


Figure 4: Comparison of the modified Fourier (thick line) and Legendre–Galerkin (dots/thin line) methods applied to (5.1) with exact solution (5.8)–(5.10) (top to bottom). (left) log L^2 error $\log_{10}\|u-u_N\|$ against number of terms, (middle) log H^1 error $\log_{10}\|u-u_N\|_1$ against number of terms, (right) log pointwise error $\log_{10}|u(x,1,1)-u_N(x,1,1)|$ for $-1\leq x\leq 1$, where N is chosen so that the number of terms for each method is approximately 2750.

of the coefficients $\hat{f}^{[i]}_n$, meaning that a direct comparison of the two methods in terms computational time is premature. The design of efficient, robust algorithms based on the quadratures developed in [17, 18] is a subject of future research, as we discuss briefly in Section 7.

Having demonstrated examples where the modified Fourier method is advantageous, it should be noted that such improvement is certainly not in evidence for all problems. In particular, whenever the solution u has large mixed derivative in comparison to its classical derivative, the Legendre–Galerkin approach (which is based on a full index set) will outperform the modified Fourier method (which utilises the hyperbolic cross). This feature is common to all hyperbolic cross/sparse grid methods. By means of example, consider the functions

$$u_1(x,y,z)=v_\omega(x)v_1(y)v_1(z),\quad u_2(x,y,z)=v_\omega(x)v_\omega(y)v_1(z),\quad u_3(x,y,z)=v_\omega(x)v_\omega(y)v_\omega(z),$$

where $v_t(x)=\frac{\cosh[t(1-x^2)]}{\cosh t}$ for $t\in\mathbb{R}$, which satisfy $\|u_i\|_r=\mathcal{O}\left(\omega^r\right)$ and $\|u_i\|_{r,\mathrm{mix}}=\mathcal{O}\left(\omega^{ir}\right)$ for $i=1,2,3$. Figure 5 compares the two methods for these example. For u_1, the modified Fourier method outperforms the Legendre method for moderate values of N. However, this effect is less pronounced for u_2 and does not occur at all for u_3.

Even for problems where the modified Fourier method outperforms polynomial-based methods for moderate N, this regime may be rather small (especially for $d=2,3$). The aim of future work, namely the design of techniques to accelerate the convergence rate, is to address this issue by realising more rapid convergence, thereby making the method effective for a broader range of problems (see Section 7).

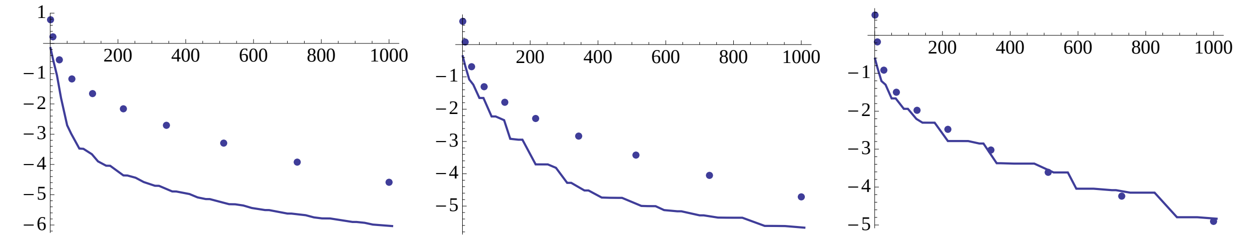


Figure 5: Log error $\log_{10}\|u-u_N\|$ against number of terms for the modified Fourier and Legendre–Galerkin methods applied to the problem with $a=0$, $b=2$ and exact solutions u_1 (left), u_2 (middle) and u_3 (right), where $\omega=10$.

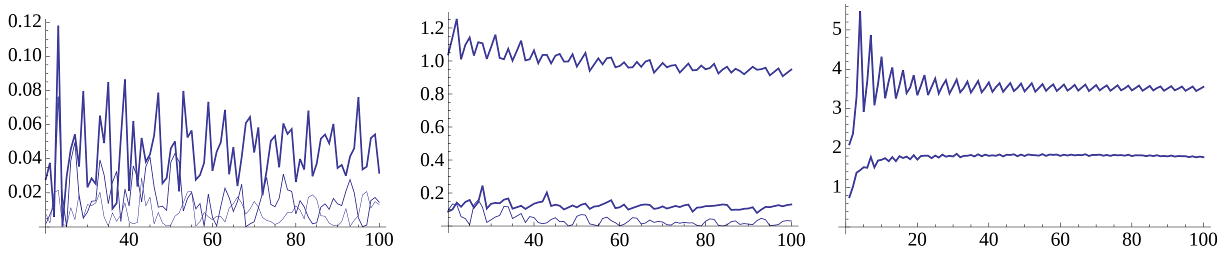


Figure 6: (left) scaled pointwise error $N^3(\log N)^{-1}|u(x,y)-u_N(x,y)|$ where u is given by (5.11) and $(x,y)=(1,-1)$ (thickest line), $(1,\frac{1}{4})$, $(-\frac{1}{2},-\frac{1}{2})$ (thinnest line). (middle) scaled pointwise error for (5.12). (right) scaled H^1 error $N^{\frac{5}{2}}\|u-u_N\|_1$ for (5.11) (bottom) and (5.12) (top).

5.5 Variable coefficient problems

The extension of this method to variable coefficient problems, where $a_j:(-1,1)^d\to\mathbb{R}$ and $b:(-1,1)^d\to\mathbb{R}$ are (sufficiently smooth) functions of x, can be achieved in a straightforward manner. In particular, simple generalisations of Theorems 5.8 and 5.9 are easily established. This fact is demonstrated numerically in Figure 6 for examples with parameters $d=2$, $a\equiv 0$, $b(x,y)=x^2y$ and $b(x,y)=\cos(x+y)$ respectively, and exact solutions

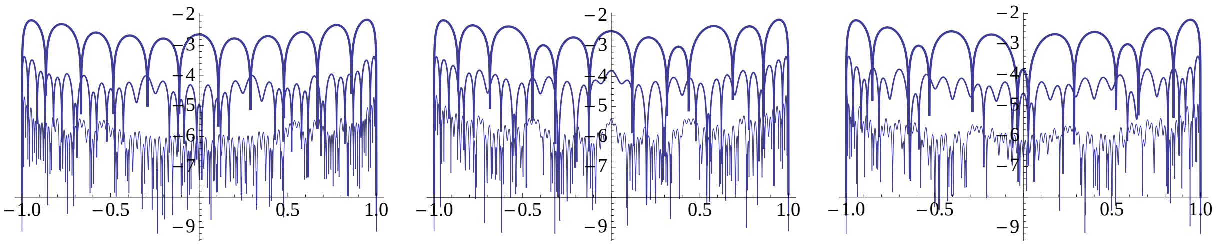
$$\begin{aligned} u(x,y)=&\mathrm{e}^{xy}-\frac{y}{4}\left[(1+x)^2\mathrm{e}^{y}+(1-x)^2\mathrm{e}^{-y}\right]-\frac{x}{4}\left[(1+y)^2\mathrm{e}^{x}+(1-y)^2\mathrm{e}^{-x}\right]\\ &+\frac{\mathrm{e}}{8}\left[(1-x)^2(1-y)^2+(1+x)^2(1+y)^2\right], \qquad (5.11)\\ u(x,y)=&\frac{1}{2}\sin 2xy-xy(\cos 2x+\cos 2y+2\sin 2-\cos 2). \qquad (5.12)\end{aligned}$$

We mention in passing that previously derived estimates for the condition number remain valid in the variable coefficient setting. An optimal, diagonal preconditioner can also be obtained from the discretization of the operator $-\triangle+b_0\mathcal{I}$, where $b_0=\max_{x\in(-1,1)^d}b(x)$.

Much like the Fourier case, the matrix A_{G} has entries that involve modified Fourier (and Laplace–Dirichlet) coefficients of the functions a_j and b. As with the inhomogeneous term f, these may be calculated by numerical quadrature. Efficient solution of Galerkin's equations can be achieved once more by conjugate gradients. The matrix A_{G} is typically dense, thus direct evaluation of matrix-vector products requires $\mathcal{O}\left(N^2(\log N)^{2(d-1)}\right)$ operations. However, since the action of N_{G} corresponds to finding modified Fourier coefficients of products and derivatives of finite modified Fourier sums, this figure could be reduced to $\mathcal{O}\left(N(\log N)^d\right)$ as in the constant coefficient case.

6 Discretization of Dirichlet and Robin boundary value problems

As mentioned in Section 1, the modified Fourier basis is best suited to the spectral discretization of homogeneous Neumann boundary value problems. Analogously, for a homogeneous Dirichlet boundary value



$y_0 = -\frac{1}{4}$ $y_0 = 0$ $y_0 = \frac{1}{2}$

Figure 7: Log pointwise error $\log_{10} |u(x, y_0) - u_N(x, y_0)|$, $-1 \leq x \leq 1$, for the Laplace–Dirichlet Galerkin approximation with $N = 10$ (thickest line), $N = 40$ and $N = 160$ (thinnest line).

problem (for example), we discretize using Laplace–Dirichlet eigenfunctions $\psi_n^{[i]}$ (see Section 2.1). The resulting method exhibits many similar properties to the modified Fourier–Galerkin method (unsurprisingly, given the duality enjoyed by the two bases). In particular, the equations may be solved in $\mathcal{O}\left(N^2\right)$ operations, and there is an optimal, diagonal preconditioner.

In Figure 7 we plot the error for the Galerkin approximation based on Laplace–Dirichlet eigenfunctions applied to the boundary value problem

$$-\triangle u + a.\nabla u + bu = f, \quad u|_\Gamma = 0,$$

with $d = 2$, parameters $a_1 = 1$, $a_2 = -1$, $b = 3$ and exact solution $u(x, y) = (x^2 - 1)^2 (y^2 - 1)$. Observe that doubling N reduces the error by roughly a factor of 4. This indicates an $\mathcal{O}\left(N^{-2}\right)$ uniform error. Inside the domain—unlike the modified Fourier–Galerkin approximation—the error is a full power of N faster, i.e. $\mathcal{O}\left(N^{-3}\right)$. Numerical results indicate that an analogue of Theorem 4.6 holds for Laplace–Dirichlet Galerkin approximations. Note that, due to the boundary conditions, the error on the boundary is in fact zero. However, as is typical for (weak) Gibbs-type phenomena, the maximal error occurs at an $\mathcal{O}\left(N^{-1}\right)$ distance away from the boundary.

For Robin boundary conditions $\frac{\partial u}{\partial n} + \theta u|_\Gamma = 0$, where $\theta \in \mathbb{R}$, we employ a similar approach. The relevant Laplace eigenfunctions subject to such boundary conditions are precisely Cartesian products of the univariate eigenfunctions given explicitly by

$$\phi_0^{[0]}(x) = (\theta^{-1} \sinh(2\theta))^{-\frac{1}{2}} \mathrm{e}^{-\theta x}, \quad \phi_n^{[0]}(x) = (n^2\pi^2 + \theta^2)^{-\frac{1}{2}} (n\pi \cos n\pi x - \theta \sin n\pi x), \quad n \in \mathbb{N}_+,$$
$$\phi_n^{[1]}(x) = ((n - \tfrac{1}{2})^2 \pi^2 + \theta^2)^{-\frac{1}{2}} \left((n - \tfrac{1}{2})\pi \sin(n - \tfrac{1}{2})\pi x + \theta \cos(n - \tfrac{1}{2})\pi x\right), \quad n \in \mathbb{N}_+.$$

As in the Dirichlet case, the resulting method shares many features with the modified Fourier method.

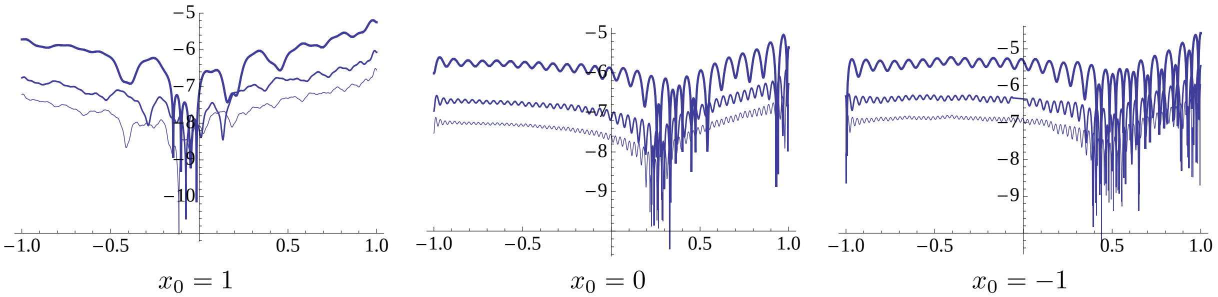
In Figure 8 we plot the error for the Galerkin approximation based on these eigenfunctions applied to the boundary value problem with parameters $a_1 = 2$, $a_2 = 3$, $b = 5$ subject to homogeneous Robin boundary conditions with $\theta = 3$ and exact solution

$$u(x, y) = \frac{1}{26} \left[16\mathrm{e}^{\frac{1-x}{2}} - 8e(x + 1) + (3e - 1)(1 + x)^2\right] (y + 1)^2 (2y - 3). \tag{6.1}$$

These results indicate an $\mathcal{O}\left(N^{-3}\right)$ uniform error, as was observed for the modified Fourier method. Moreover, unlike the Dirichlet approximation, the rate of convergence away from the boundary is not of higher order.

7 Conclusions

We have developed the approximation-theoretic properties of modified Fourier series in Cartesian product domains using both full and hyperbolic cross index sets. In particular we have proved uniform convergence and extended the results of [1, 17, 22] concerning the rate of convergence. In the second half of this paper we have applied such series to the spectral-Galerkin approximation of Neumann boundary value problems.



$x_0 = 1$ $x_0 = 0$ $x_0 = -1$

Figure 8: Log pointwise error $\log_{10} |u(x_0, y) - u_N(x_0, y)|$, $-1 \le y \le 1$, for the Laplace–Robin Galerkin approximation to (6.1) with $N = 50$ (thickest line), $N = 100$ and $N = 150$ (thinnest line).

We have shown that the resulting approximation consists of $\mathcal{O}\left(N(\log N)^{d-1}\right)$ terms which may be found in $\mathcal{O}\left(N^2\right)$ operations using conjugate gradients. Moreover, the discretization matrix has an $\mathcal{O}\left(N^2\right)$ condition number and there is an optimal, diagonal preconditioner. Despite offering only algebraic convergence, we have shown that these methods are more effective than standard polynomial-based approaches for moderate values of the truncation parameter in certain problems. Finally, we have demonstrated how very similar methods can be developed for problems with Dirichlet or Robin boundary conditions.

Numerous avenues remain for future research, as we now describe:

1. *Numerical evaluation of modified Fourier coefficients.* It was not the intent of this paper to address the calculation of modified Fourier coefficients using the methods outlined in [17, 18]. Herein a number of open problems and challenges remain. Robust error estimates for the quadratures employed are largely lacking and criteria for selecting optimal parameters do not yet exist. It is also fair to point out that such schemes become increasingly complicated to implement in higher dimensions. However, this approach allows us to immediately exploit hyperbolic cross index sets, leading to the favourable properties outlined in this paper. Clearly these issues are a barrier to the design of competitive algorithms based on modified Fourier series, with particular application to boundary value problems, and future work will address this shortcoming.
2. *Accelerating convergence.* The results of this paper suggest that, despite converging only algebraically, the modified Fourier method exhibits some benefits over polynomial-based methods. To make the method competitive for a larger range of problems, future work will address the issue of accelerating convergence. For the task of function approximation, a number of convergence acceleration techniques are known. In [2] Eckhoff's method (a well-known technique in the univariate Fourier setting [9]) was generalised to modified Fourier expansions in d-variate cubes. Development of spectral-Galerkin methods utilising this device is a subject of both current and future investigation. In [1] an alternative ad-hoc technique was introduced to accelerate convergence of the univariate modified Fourier–Galerkin method. Future work shall also investigate the extension of this technique to the multivariate setting.
3. *Other boundary value problems.* As demonstrated in Section 6, closely related techniques can be developed for second order Dirichlet and Robin boundary value problems. Unfortunately more complicated boundary conditions cannot be tackled so easily. For example, so-called co-normal boundary conditions are outside the scope of this approach, since the relevant Laplace eigenfunctions cannot be expressed as simple Cartesian products. However, certain higher, even-order differential operators also have simple eigenfunctions [16]. These may have application in the spectral discretization of higher-order problems.
4. *Non-tensor product domains.* Laplace–Neumann eigenfunctions are known explicitly in certain triangles and higher dimensional simplices (see [14] for the case of the equilateral triangle), allowing for the construction of approximation schemes in more complicated geometries. Existing spectral algorithms for triangular domains are complicated to implement, so the modified Fourier approach may offer benefits in this respect.
5. *Fast solvers.* The iterative techniques presented to solve the modified Fourier–Galerkin equations are generic and work well precisely because of the simple nature of the modified Fourier basis. As mentioned, a variant of the sparse grid FFT can be used to increase efficiency. However, it would be

preferable to avoid the use of the FFT altogether and develop ad-hoc methods instead. The quadrature methods introduced in [17, 18]—presented as an alternative to the FFT for the related task of evaluating modified Fourier coefficients—provide a potential means to do this.

Acknowledgements

The author would like to thank Alfredo Deaño (DAMTP, University of Cambridge), Daan Huybrechs (Katholieke Universiteit Leuven), Friedrich Kupka (MPA Garching) and Arieh Iserles (DAMTP, University of Cambridge). He would also like to thank the two anonymous referees for their helpful comments and suggestions.

References

[1] B. Adcock. Univariate modified Fourier methods for second order boundary value problems. *BIT*, 49(2):249–280, 2009.

[2] B. Adcock. Convergence acceleration of modified Fourier series in one or more dimensions. *Math. Comp. (to appear)*, 2010.

[3] K. I. Babenko. Approximation of periodic functions of many variables by trigonometric polynomials. *Soviet Math. Dokl.*, 1:513–516, 1960.

[4] G. Baszenski and F.-J. Delvos. A discrete Fourier transform scheme for Boolean sums of trigonometric operators. In C.K. Chui, W. Schempp, and K. Zeller, editors, *Multivariate Approximation Theory IV, ISNM 90*, pages 15–24, Basel, 1989. Birkhauser.

[5] H. Brunner, A. Iserles, and S. P. Nørsett. The computation of the spectra of highly oscillatory Fredholm integral operators. *Technical report NA2009/03, DAMTP, University of Cambridge*, 2009.

[6] H.-J. Bungartz and M. Griebel. Sparse grids. *Acta Numerica*, 13:147–269, 2004.

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

[8] N. M. Dobrovol'skii and A. L. Roshchenya. Number of lattice points in the hyperbolic cross. *Math. Notes*, 63:319–324, 1998.

[9] K. S. Eckhoff. On a high order numerical method for functions with singularities. *Math. Comp.*, 67(223):1063–1087, 1998.

[10] M. Fenn, S. Kunis, and D. Potts. Fast evaluation of trigonometric polynomials from hyperbolic crosses. *Numer. Algorithms*, 41(4):339–352, 2006.

[11] M. Griebel and J. Hamaekers. Sparse grids for the Schrödinger equation. *Math. Model. Numer. Anal.*, 41:215–247, 2007.

[12] B.-Y. Guo, J. Shen, and L.-L. Wang. Optimal spectral-Galerkin methods using generalized Jacobi polynomials. *J. Sci. Comput.*, 27(1–3):305–322, 2006.

[13] W. Hackbusch. *Elliptic Differential Equations.* Springer–Verlag, 1992.

[14] D. Huybrechs, A. Iserles, and S. P. Nørsett. From high oscillation to rapid approximation V: The equilateral triangle. *Technical report NA2009/04, DAMTP, University of Cambridge*, 2009.

[15] D. Huybrechs, A. Iserles, and S. P. Nørsett. From high oscillation to rapid approximation IV: Accelerating convergence. *IMA J. Num. Anal. (to appear)*, 2010.

[16] A. Iserles and S. P. Nørsett. From high oscillation to rapid approximation II: Expansions in polyharmonic eigenfunctions. *Technical report NA2006/07, DAMTP, University of Cambridge*, 2006.

[17] A. Iserles and S. P. Nørsett. From high oscillation to rapid approximation I: Modified Fourier expansions. *IMA J. Num. Anal.*, 28:862–887, 2008.

[18] A. Iserles and S. P. Nørsett. From high oscillation to rapid approximation III: Multivariate expansions. *IMA J. Num. Anal.*, 29:882–916, 2009.

[19] L. V. Kantorovich and V. I. Krylov. *Approximate Methods of Higher Analysis.* Interscience, New York, 3rd edition, 1958.

[20] F. Kupka. *Sparse grid spectral methods for the numerical solution of partial differential equations with periodic boundary conditions.* PhD thesis, Institut für Mathematik, Universität Wien, 1997.

[21] C. Lanczos. *Discourse on Fourier series.* Hafner, New York, 1966.

[22] S. Olver. On the convergence rate of a modified Fourier series. *Math. Comp.*, 78:1629–1645, 2009.

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

[24] H.-J. Schmeißer and H. Triebel. *Topics in Fourier analysis and function spaces.* Wiley, 1987.

[25] J. Shen. Efficient spectral-Galerkin method I. direct solvers of second and fourth-order equations using Legendre polynomials. *SIAM J. Sci. Comput.*, 15(6):1489–1505, 1994.

[26] J. Shen. Efficient spectral-Galerkin method II. direct solvers of second and fourth-order equations using Chebyshev polynomials. *SIAM J. Sci. Comput.*, 16(1):74–87, 1995.

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