Modified Fourier expansions: theory, construction and applications

Benjamin James Stevens Adcock

Trinity Hall
University of Cambridge
1 July 2010

This dissertation is submitted for
the degree of Doctor of Philosophy

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

Ben Adcock
Cambridge
1 July 2010

Abstract

Modified Fourier expansions present an alternative to more standard algorithms for the approximation of nonperiodic functions in bounded domains. This thesis addresses the theory of such expansions, their effective construction and computation, and their application to the numerical solution of partial differential equations.

As the name indicates, modified Fourier expansions are closely related to classical Fourier series. The latter are naturally defined in the d-variate cube, and, in an analogous fashion, we primarily study modified Fourier expansions in this domain. However, whilst Fourier coefficients are commonly computed with the Fast Fourier Transform (FFT), we use modern numerical quadratures instead. In contrast to the FFT, such schemes are adaptive, leading to great potential savings in computational cost.

Standard algorithms for the approximation of nonperiodic functions in d-variate cubes exhibit complexities that grow exponentially with dimension. The aforementioned quadratures permit the design of approximations based on modified Fourier expansions that do not possess this feature. Consequently, such schemes are increasingly effective in higher dimensions. When applied to the numerical solution of boundary value problems, such savings in computational cost impart benefits over more commonly used polynomial-based methods. Moreover, regardless of the dimensionality of the problem, modified Fourier methods lead to well-conditioned matrices and corresponding linear systems that can be solved cheaply with standard iterative techniques.

The theoretical component of this thesis furnishes modified Fourier expansions with a convergence analysis in arbitrary dimensions. In particular, we prove uniform convergence of modified Fourier expansions under rather general conditions. Furthermore, it is known that the notion of modified Fourier expansions can be effectively generalised, resulting in a family of approximation bases sharing many of the features of the modified Fourier case. The purpose of such a generalisation is to obtain both faster rates and higher degrees of convergence. Having detailed the approximation-theoretic properties of modified Fourier expansions, we extend this analysis to the general case and thereby verify this improvement.

A central drawback of these expansions is that their convergence rate is both fixed and typically slow. This makes the construction of effective convergence acceleration techniques imperative. In the final part of this thesis, we design and analyse a robust method, applicable in arbitrary numbers of dimensions, for accelerating convergence of modified Fourier expansions. When employed in the approximation of multivariate functions, this culminates in efficient, high-order approximants comprising relatively small numbers of terms.

Acknowledgements

It is a pleasure to acknowledge the support I have received during my doctoral studies. First, I owe a great deal of thanks to my supervisor Arieh Iserles. His enthusiasm and passion for numerical analysis are virtually unbounded, and it has been a privilege to be his student. His willingness to discuss new ideas and to provide guidance have been a great help during my graduate studies. Outside of mathematics, I will particularly miss our lively coffee break discussions.

My research has greatly benefited from regular conversations and collaborations with Daan Huybrechs, Alfredo Deaño and Sheehan Olver. I have also frequently pestered Euan Spence, and I owe him thanks for first introducing me to some of the topics pursued in this thesis. Syvert Nørsett has also been of great assistance. Finally, I have enjoyed discussions with Chris Budd and David Levin.

It has been a pleasure to be part of the numerical analysis group in Cambridge. I have had many fruitful discussions with Anders Hansen and Alexei Shadrin, in particular. I would also like to thank the other members: Brad Baxter, Alex Benton, Marianna Khanamiryan, Mike Powell, Malcolm Sabin and Tanya Shingel. Outside of the group, I also owe thanks to Guillaume Dujardin, Jennifer Ryan and Carola Schönlieb.

During both my graduate and undergraduate studies in Cambridge, I have found the advice of Tom Körner invaluable. Tadashi Tokieda has also taken a keen interest in my progress. I am particularly thankful to him for his advice and encouragement with regard to postdoctoral positions.

I am indebted to my parents for their continued support and guidance, whilst my brother has always been eager to offer advice. Finally, it is impossible to underestimate the contribution of my wife Tina. For this, I cannot thank her enough.

Ben Adcock
Cambridge
1 July 2010

Contents

Notation

General

$\mathbb{R}$	Set of real numbers
$\mathbb{Z}$	Set of integers
$\mathbb{N}$	Set of positive integers
$\mathbb{N}_0$	Set of non-negative integers
$\mathbb{C}$	Set of complex numbers
$\lfloor x \rfloor$	Integral part of $x \in \mathbb{R}$
i	Imaginary unit
$\mathrm{Re}\, z$, $\mathrm{Im}\, z$	Real and imaginary parts of $z \in \mathbb{C}$
d	Dimension
x	Multivariable $x = (x_1, ..., x_d) \in \mathbb{R}^d$
$x.y$	The dot product $x_1 y_1 + ... + x_d y_d$
n, i	Multi-indices $n = (n_1, ..., n_d) \in \mathbb{N}_0^d$ and $i = (i_1, ..., i_d) \in \{0, 1\}^d$
$\lvert n \rvert$	The quantity $n_1 + ... + n_d$
$\lvert n \rvert_\infty$	The quantity $\max\{n_1, \ldots, n_d\}$
$\bar{n}$	The quantity $\max\{1, n\}$ for $n \in \mathbb{N}_0$
$\lvert n \rvert_0$	The quantity $\bar{n}_1 ... \bar{n}_d$ for $n \in \mathbb{N}_0^d$

Multivariate Notation

$[d]$	Set of ordered tuples with entries in $\{1, ..., d\}$
$\emptyset$	The empty tuple
$[d]^*$	The set $[d] \cup \{\emptyset\}$
$\lvert t \rvert$	Length (number of elements) in $t \in [d]$
$\bar{t}$	Ordered tuple with entries in $\{1, .., d\}$ and not in t
$[t]$	The set of tuples $u \subseteq t$
$\bar{u}$	For $u \in [t]$, the tuple of elements in t but not in u
$t \cap u$	The tuple of elements in both t and u
$t \cup u$	The tuple of elements in t or in u
x_t	Restriction of the variable x to those entries in t

Differential Operators

β	Multi-index $(\beta_1, ..., \beta_d) \in \mathbb{N}_0^d$
D^β	Differential operator $\frac{\partial^{\lvert\beta\rvert}}{\partial_{x_1}^{\beta_1} ... \partial_{x_d}^{\beta_d}} = \partial_{x_1}^{\beta_1} ... \partial_{x_d}^{\beta_d}$
D^r, D	The operators $\partial_{x_1}^r ... \partial_{x_d}^r$, $\partial_{x_1} ... \partial_{x_d}$
D_t^r	The operator $\partial_{x_{t_1}}^r ... \partial_{x_{t_{\lvert t\rvert}}}^r$ for $t \in [d]$
∇	Gradient operator
$\triangle$	Laplace operator

Functional Analysis

Ω	Bounded domain in $\mathbb{R}^d$
$\bar{\Omega}$	Closure of Ω
Γ	Boundary of Ω
$\mathrm{C}^r(\Omega)$, $\mathrm{C}^r(\bar{\Omega})$	Spaces of r-times continuously differentiable functions
$\mathrm{C}^{r,\lambda}(\bar{\Omega})$	Space of Hölder continuous functions
$\mathrm{L}^2(\Omega)$	Space of square integrable functions on Ω
$\mathrm{H}^r(\Omega)$	Sobolev space with index r
$\mathrm{H}_0^r(\Omega)$	Closure of $\mathrm{C}_0^\infty(\Omega)$ in $\mathrm{H}^r(\Omega)$
$\mathrm{H}_{\mathrm{mix}}^r(\Omega)$	Sobolev space of dominating mixed smoothness

Inner Products and Norms

$(\cdot, \cdot)$	$\mathrm{L}^2(\Omega)$ inner product
$\lVert\cdot\rVert$	$\mathrm{L}^2(\Omega)$ norm
$\lVert\cdot\rVert_\infty$	Uniform norm on Ω
$\lVert\cdot\rVert_r$	$\mathrm{H}^r(\Omega)$ norm
$\lVert\cdot\rVert_{r,\mathrm{mix}}$	$\mathrm{H}_{\mathrm{mix}}^r(\Omega)$ norm

Asymptotics

$\mathcal{O}(\cdot)$	Big-O notation
$o(\cdot)$	Little-O notation
$\sim$	Asymptotic expansion symbol

Eigenfunction Expansions

ϕ_n	Eigenfunction corresponding to index $n \in \mathbb{N}_0^d$
$\phi_n^{[i]}$	Eigenfunction corresponding to indices $i \in \{0,1\}^d$, $n \in \mathbb{N}_0^d$
ψ_n, $\psi_n^{[i]}$	Dual eigenfunction
μ_n, $\mu_n^{[i]}$	Eigenvalue
$\hat{f}_n^{[i]}$, $\check{f}_n^{[i]}$	Coefficients of f corresponding to $\phi_n^{[i]}$, $\psi_n^{[i]}$
I_N	Finite index set in $I_N \subseteq \mathbb{N}_0^d$
$\mathcal{S}_N$	Space spanned by $\phi_n^{[i]}$, $n \in I_N$, $i \in \{0,1\}^d$
$\mathcal{F}_N[f]$	Truncated eigenfunction expansion

Linear Operators and Bilinear Forms

$\mathcal{I}$	Identity operator
$\mathcal{L}_0$	Linear, self-adjoint differential operator
$\mathcal{L}$	Arbitrary linear differential operator
$\mathcal{L}_1$	The operator $\mathcal{L} - \mathcal{L}_0$
T, T_0	Bilinear forms associated with $\mathcal{L}$, $\mathcal{L}_0$
γ, ω	Constants of continuity and coercivity
$\mathcal{B}$, $\mathcal{B}_r$	Boundary conditions

Galerkin Approximation

u_N	Galerkin approximation
$\bar{u}_n^{[i]}$, $\bar{u}$	Coeffcients of u_N, vector of coefficients
A_{G}	Galerkin matrix
M_{G}, N_{G}	Matrices of the splitting $A_{\mathrm{G}} = M_{\mathrm{G}} + N_{\mathrm{G}}$
$\kappa(A)$, $\kappa_s(A)$	Condition number and spectral condition number of A
$\rho(A)$	Spectral radius of A

Convergence Acceleration

$\mathcal{A}_{r_t,n_{\bar{t}}}^{[i]}[f]$	Term in asymptotic expansion of $\hat{f}_n^{[i]}$
$\bar{\mathcal{A}}_{r_t,n_{\bar{t}}}^{[i]}[f]$	Approximation of $\mathcal{A}_{r_t,n_{\bar{t}}}^{[i]}[f]$
$p_r^{[i]}$	Cardinal function
$q_r^{[i]}$	Subtraction basis function
g_k	Subtraction function
$\mathcal{F}_{N,k}[f]$	The approximation $\mathcal{F}_N[f - g_k] + g_k$
$m(r)$	Parameters in Eckhoff's approximation
$V^{[i]}$	Matrix of Eckhoff's approximation

Chapter 1

Introduction

This thesis concerns a classical problem in numerical analysis: namely, the practical approximation of smooth, nonperiodic functions defined on bounded domains. This problem lies at the heart of countless methods in computational mathematics, with applications ranging from the numerical approximation of partial differential equations to the reconstruction of images from discrete data.

Ostensibly, the approach we consider is extremely well known. Our approximation scheme is based on expanding a function in eigenfunctions of a suitable differential operator with prescribed boundary conditions. In particular, the majority of this thesis deals with one of the simplest examples of such an approach: expansions in eigenfunctions of the Laplace operator subject to either homogeneous Dirichlet or Neumann boundary conditions. Nonetheless, as we henceforth describe, the cornerstone of the new research into this subject involves the successful development and implementation of a number of novel numerical tools for such expansions. The ensuing analysis requires both new techniques and generalisations of existing results.

The expansion of univariate functions defined on compact intervals in eigenfunctions of general linear differential operators has been the subject of a broad array of literature. There is a well-developed theory of such expansions, including significant contributions from Birkhoff [25, 26], who first studied the topic in its general form, and the volumes of Naimark [127] and Dunford and Schwartz [51]. However, outside of the case of Fourier expansions (which are usually studied independently, and have spawned their own field, harmonic analysis [103, 107]), few attempts have been made to date at performing practical computations with such eigenfunctions. The techniques introduced in this thesis, and elsewhere, bridge the gap between existing theory and practical applications. Moreover, issues arising from practical problems lead to generalisations in new directions, thereby complementing existing literature.

Fourier series present the most elementary example of such an approach. Their immense success and widespread use in a myriad of practical applications, including image and signal processing, electrical engineering and acoustics, can be attributed to several principal ingredients. First, once an analytic and periodic function is expanded in this basis, the expansion converges exponentially fast in the number of approximation terms. Second, provided N is a highly composite integer, the first N approximation coefficients can be computed to within exponentially small error in $\mathcal{O}(N \log N)$ operations using the Fast Fourier Transform [37]. Third, in the context of discretisations of partial differential equations, the Fourier basis leads to stable, well-conditioned algorithms, including diagonal matrices for constant coefficient

problems.[1]

However, once periodicity is no longer present, Fourier series are far less appealing: the error committed by the truncated expansion is $\mathcal{O}\left(N^{-1}\right)$ inside the domain, and there is no uniform convergence. The presence of $\mathcal{O}\left(1\right)$ oscillations near the endpoints—the celebrated Gibbs phenomenon [83]—is a blight in many practical applications of such expansions [96]. For these reasons, the amelioration or, indeed, complete resolution of this phenomenon is an area of continuing research [156].

In view of these shortfalls, the objective of this thesis is the study of approximation schemes for nonperiodic functions that share many of the benefits of Fourier series (in particular, those relating to the approximation of differential equations), whilst offering faster and, in particular, uniform convergence. It turns out that the Fourier basis is amongst the worst (in terms of rate of convergence) eigenfunction basis for approximating nonperiodic functions. A minor modification, replacing the first-order differential operator by the Laplace operator equipped with homogeneous Neumann boundary conditions, leads to uniformly convergent eigenfunction expansions. In practice this corresponds to replacing the standard Fourier basis on $[-1, 1]$, given by

$$\{\cos n\pi x : n \in \mathbb{N}_0\} \cup \{\sin n\pi x : n \in \mathbb{N}\},$$

with a basis containing sine functions with shifted arguments:

$$\{\cos n\pi x : n \in \mathbb{N}_0\} \cup \left\{\sin(n - \tfrac{1}{2})\pi x : n \in \mathbb{N}\right\}.$$

Such a basis, introduced in [94], forms the primary subject of this thesis. To emphasise its proximity to classical Fourier series, we shall refer to expansions in such *Laplace–Neumann* eigenfunctions as *modified Fourier expansions* (a term coined in [94]).

For modified Fourier expansions to enjoy any of the success of Fourier series, they must be practical. In other words, expansions must be accompanied by numerical schemes to calculate coefficients in a manner that mirrors the benefits of the FFT. A central tenet of this thesis is the use of modern numerical quadratures for this task. These stem from recently developed numerical techniques for the computation of highly oscillatory integrals. Such a strategy not only possesses several of the virtues of the FFT, it also offers a number of crucial advantages. Unlike the FFT, this approach is adaptive: coefficients are calculated one by one, and changing N does not require recalculation of any existing values. Moreover, *any* N (not necessarily contiguous) coefficients can be calculated in $\mathcal{O}\left(N\right)$ operations, without the restriction that N be a highly composite integer. As we describe in due course, in addition to their inherent benefits, such features have important consequences for the design of efficient numerical schemes based on modified Fourier expansions. In particular, they facilitate the incorporation of so-called *hyperbolic cross index sets* [13, 158], which greatly reduce the computational cost of constructing and evaluating approximations in multivariate domains.

Regardless of these factors, classical Fourier series are not commonly used to approximate nonperiodic functions. Given an analytic, nonperiodic function in the unit interval, approximation will more routinely be carried out by expanding in certain orthogonal polynomials. Since the relevant coefficients can be calculated using the FFT, Chebyshev polynomials are

[1] In fact, the expansion and subsequent truncation of the solution of a periodic partial differential equation in its Fourier series was the first example of a *spectral method* for numerical solution of such problems [67]. Its success has spawned a large area of computational mathematics based on approximating solutions of differential equations in rapidly convergent orthogonal bases [42, 142, 159].

most typically utilised [31], although other Jacobi polynomials, including Legendre polynomials, are also employed [42, chapter 2]. Such an approach is efficient, simple to implement and provides exponentially accurate approximations. For this reason, orthogonal polynomials are the common starting point for many spectral methods for the discretisation of nonperiodic partial differential equations [142].[2]

Rather than judging modified Fourier expansions in comparison with classical Fourier series, their merits must be viewed in light of polynomial-based methods. However, as we now detail, there are a number of significant instances where such expansions can be expected to offer benefits, therefore motivating their continued study.

1.1 Rationale

A common problem in the design of numerical methods for nonperiodic partial differential equations is as follows. Whilst spectral methods based on orthogonal polynomials possess the great advantage of rapid convergence, they lack both the generality and adaptability of finite element methods. Conversely, despite being versatile and adaptable, finite element methods converge slowly. As we now describe, modified Fourier expansions offer a potential route towards the design of approximation schemes that incorporate both these features—flexibility and high accuracy—thus extending the range of spectral methods to a wider class of problems.

Notwithstanding, the techniques of this thesis will not typically confer significant advantages over polynomial-based methods for the straightforward task of approximating nonperiodic functions of one variable. In spite of this, we mention in passing that Fourier or Fourier-like series, despite their slow convergence, are also commonly used in applications lacking periodicity (in particular, image and signal processing [96]), since Fourier data is more often available in certain applications. A key component therein is the design of algorithms for convergence acceleration. This topic, including the significant adaption and generalisation of certain existing schemes to modified Fourier expansions, also forms an important constituent of this thesis (see Section 1.2.2).

The simplest setting for modified Fourier expansions is the approximation of functions defined on the unit interval. The d-variate cube presents the first extension of this topic. Theoretically speaking, such generalisation is attained relatively easily by means of Cartesian products. Indeed, multivariate Fourier series and Chebyshev polynomials are well established in this domain. However, such approximations typically involve $\mathcal{O}\left(N^d\right)$ coefficients which can be computed in, at best, $\mathcal{O}\left(N^d \log N\right)$ operations via the FFT. These figures grow exponentially with dimension, thereby making such schemes impractical for higher-dimensional problems (even when $d = 3, 4$ significant effort is required to form such approximations).

Many physical problems, when formulated as mathematical models, require the approximation of higher-dimensional functions or the solution of higher-dimensional differential equations. Notably, these include applications ranging from fluid dynamics (the Navier–Stokes equations) to quantum mechanics and computational chemistry (the Schrödinger equation) [41]. In particular, direct modelling of a system of m interacting particles in $\mathbb{R}^d$ in theory involves solving equations in md variables [75].

[2] The term *spectral method* here refers to any approximation basis that delivers so-called *spectral* accuracy: in other words, convergence faster than any algebraic power. Commonly, once the approximated function is endowed with sufficient regularity (for example, analyticity), exponential convergence is witnessed [42].

Amongst the finite elements community, there exist well-developed numerical methods for higher-dimensional problems—so-called *sparse grid* finite element methods—which reduce the aforementioned figure to just $\mathcal{O}\left(N(\log N)^{d-1}\right)$, or even $\mathcal{O}\left(N\right)$ independently of d, without causing a significant deterioration in approximation quality [40, 41]. Theoretically at least, spectral methods based on either Fourier series or orthogonal polynomials can also be designed to reflect this feature [158]. Such approximations exploit a related tool, the aforementioned hyperbolic cross, to reduce computational cost. However, due to the non-adaptivity of the FFT, it becomes significantly more difficult to realise such approximations as practical schemes. Though there exist a number of non-standard variants of FFT to address this situation [18, 60], such techniques are not typically simple to implement. For these reasons, outside of so-called *sparse grid Fourier methods* for periodic partial differential equations [75, 110], few such spectral methods currently exist.

In contrast to the FFT, however, the numerical quadratures outlined previously are adaptive, thus allowing tools such as the hyperbolic cross to be incorporated into modified Fourier expansions in a straightforward manner. For this reason, modified Fourier approximations have potential application to higher-dimensional problems. In Chapter 2 we assess the theory and construction of such approximations in d-variate cubes.

Outside of the approximation of multivariate functions, the numerical solution of boundary value problems in two or more dimensions using modified Fourier expansions is a primary focus of this thesis. As mentioned, the Fourier method for periodic problems is endowed with a number of beneficial features. Due to the similarities between the Fourier and modified Fourier bases, there is reason to expect that such features are inherited. This turns out to be the case, making such techniques eminently suitable for these problems. Moreover, the incorporation of the hyperbolic cross yields a method that possesses a number of advantages over standard polynomial-based spectral methods, as we consider further in Chapter 4.

Though we shall not address the following topic in this thesis, we mention in passing that the modified Fourier basis has also found application in another area: namely, the computation of spectra of highly oscillatory Fredholm operators [38]. Such problems occur naturally in a number of disciplines, including acoustic scattering, laser engineering (in particular, the Fox–Li operator [46]) and electromagnetics. Analysis and numerical examples indicate that the resulting method is both more effective and substantially simpler to implement than polynomial-based approaches.

Aside from the previously discussed advantages of modified Fourier expansions, there is at least one other significant motive for their continued study: the construction of approximation schemes in non-tensor-product domains. Both Fourier series and expansions in Chebyshev polynomials are limited to d-variate cubes. Though orthogonal polynomials can be constructed in, for example, triangular domains (via Koornwinder polynomials [106] and Dubiner's warped tensor-product construction [50]), this approach is by no means straightforward [42]. A particular issue herein is the determination of optimal quadrature nodes. In recent years, there have been a number of attempts to extend classical Chebyshev polynomials beyond tensor-product domains [125]. However, this results in orthogonal polynomials defined on a deltoid, not a triangle, which presents a number of practical issues.

Conversely, Laplace eigenfunctions are known explicitly (as sums of plane waves) in a variety of higher-dimensional simplices, including various triangles and tetrahedra [88, 95, 141]. This raises the possibility of designing multi-domain approximations based on modified Fourier expansions in simplicial elements, therefore introducing an alternative means to tackle

problems defined in complex geometries.

We mention this topic as an appealing direction for modified Fourier approximations. Though a study of such expansions in the equilateral triangle has been initiated in [88], the method remains in its infancy. In particular, we highlight the absence of a theory of convergence of such expansions and the relatively unexplored generalisation to higher-dimensional simplices. Conversely, the majority of this thesis is devoted to tensor-product domains: a proper understanding of this scenario is naturally vital before tackling the aforementioned case. Nevertheless, we give a more detailed exposition of such topic in Chapter 6.

1.2 Developments and extensions

Having developed modified Fourier expansions in their most basic form in the first segment of this thesis, we pursue two extensions of this approach.

1.2.1 Birkhoff expansions

Modified Fourier expansions arise from eigenfunctions of the Laplace operator subject to homogeneous Neumann boundary conditions. It is simple to generalise this notion to Laplace eigenfunctions corresponding to a variety of other boundary conditions. Indeed, as we detail in Chapters 2 and 4 respectively, such generalisation is of both theoretical interest and practical use. However, a significantly more elaborate extension of this approach arises from the consideration of suitable higher-order operators endowed with particular boundary conditions. The purpose of such generalisation is to attain faster rates of convergence, whilst retaining the benefits of the modified Fourier case. It transpires that a judicious choice of operator and boundary conditions results in a one-parameter family of approximation bases [8]. As we establish, the rates of convergence of the corresponding expansions scale with this parameter.

Chapter 3 is devoted to this topic. As we demonstrate therein, many features of the modified Fourier basis scale effectively to this setting. In particular, similar numerical quadratures are employed to evaluate coefficients, and, in two or more dimensions, a hyperbolic cross may be exploited to reduce the number of expansion coefficients. Furthermore, though a classical theory exists for so-called *Birkhoff expansions* (expansions in eigenfunctions of arbitrary univariate differential operators) [127], it falls short of describing the case at hand. In our analysis of convergence of such expansions, we establish a number of new results specific to these particular approximation bases.

1.2.2 Accelerating convergence

The major shortcoming of Laplace eigenfunction expansions, or their aforementioned generalisation, is that their rate of convergence is both fixed and typically slow. In view of this, the final component of this thesis deals with the topic of accelerating convergence. It details the design and analysis of robust methods based on such eigenfunctions possessing arbitrarily fast rates of convergence.

A variety of techniques, readily adaptable to the modified Fourier case, exist for accelerating convergence of univariate Fourier series. Far fewer studies have addressed the multivariate setting. Yet careful analysis of multivariate modified Fourier expansions, carried out in Chapter 2, indicates how faster convergence can be attained. The purpose of Chapter 5 of this

thesis is the generalisation of several known techniques to the d-variate cube, including previously lacking analysis. Moreover, we introduce several important improvements of existing methods, aimed at both increasing numerical stability and lowering computational cost.

The obvious purpose of such study is to render modified Fourier approximations more widely effective in comparison to polynomial-based methods. In Chapter 6, we give some preliminary insight into the application of such approximations to the discretisation of boundary value problems.

1.3 Existing literature

The topic of modified Fourier expansions was introduced by Iserles and Nørsett in [94], including quadrature routines to evaluate coefficients. Generalisations to the d-variate cube and equilateral triangle, along with Huybrechs, were pursued in [95] and [88] respectively. An extension to Birkhoff expansions was considered in [8] and a study of convergence acceleration initiated in [87].

Aside from proofs of suitable versions of the Fejér and de la Vallée Poussin theorems for the univariate modified Fourier basis [94], the aforementioned papers mainly omitted the analysis of convergence of such expansions. This was carried out by S. Olver in [134] and the author in [3]. Multivariate expansions in the d-variate cube were studied by the author in [5] and convergence acceleration was addressed in [4]. As regards applications, the modified Fourier basis was employed in [38] to discretise highly oscillatory integral operators. In [5, 3] this basis was applied to the numerical solution of boundary value problems.

Several review papers have also been written on this topic [7, 90].

1.4 Outline of the thesis

The outline of the remainder of this thesis is as follows. In Chapter 2 we develop modified Fourier expansions in the d-variate cube. Chapter 3 is devoted to the generalisation of this work to certain Birkhoff expansions. The spectral discretisation of boundary value problems is studied in Chapter 4, and in Chapter 5 we assess convergence acceleration. Finally, in Chapter 6 we outline directions for future research.

Chapter 2

Laplace eigenfunction expansions

2.1 Introduction

The subject of this chapter is the expansion of nonperiodic functions defined on d-variate cubes in eigenfunctions of the Laplace operator equipped with either homogeneous Dirichlet or Neumann boundary conditions. As described in Chapter 1, Laplace–Neumann expansions (referred to as modified Fourier expansions) confer an advantage over classical Fourier series for the approximation of nonperiodic functions. An objective of this chapter is to confirm such an advantage through providing convergence analysis for these expansions.

However, we shall also study the Laplace–Dirichlet case. There are several reasons for this. First, Laplace–Neumann and Laplace–Dirichlet expansions are dual to each other in a certain sense. Analysis of the former is reliant on an understanding of the latter, and vice versa. Hence, a concurrent study is necessary. Second, much like the modified Fourier basis, Laplace–Dirichlet eigenfunctions have application to the spectral discretisation of certain boundary value problems (a topic we address in detail in Chapter 4), thus independently motivating their study. Finally, the Laplace–Dirichlet case highlights that the analysis pursued in this chapter is more widely applicable than some classical Fourier analysis techniques. Indeed, it is possible to analyse a raft of Laplace eigenfunctions corresponding to a variety of different boundary conditions with only minor modifications of the approach of this chapter. However, for the sake of clarity, we consider only the Dirichlet and Neumann cases. An indication of the generality of this approach is given at the end of this chapter.

The key results of this chapter are as follows:

1. Expansions in Laplace eigenfunctions are best studied in certain non-classical Sobolev spaces: namely, so-called *Sobolev spaces of dominating mixed smoothness*.
2. The set of Laplace–Neumann eigenfunctions is an orthogonal basis of not just $\mathrm{L}^2(\Omega)$, but also of $\mathrm{H}^1_{\mathrm{mix}}(\Omega)$, the first Sobolev space of dominating mixed smoothness. In particular, the truncated expansion of a function $f \in \mathrm{H}^1_{\mathrm{mix}}(\Omega)$ in Laplace–Neumann eigenfunctions converges uniformly on $\bar{\Omega}$. This result holds for (almost) arbitrary index sets.
3. If the standard *full* index set is employed, then the pointwise error committed by the N^{th} truncated expansion is $\mathcal{O}\left(N^{-2}\right)$ in Ω and $\mathcal{O}\left(N^{-1}\right)$ on $\partial\Omega$.
4. The coefficients $\hat{f}_n$ lie on a hyperbolic cross. Hence, a *hyperbolic cross* index set can be incorporated into the truncated expansion. This greatly reduces computational cost from $\mathcal{O}\left(N^d\right)$ when a full index set is used to just $\mathcal{O}\left(N(\log N)^{d-1}\right)$. Moreover, conver-

gence rates only deteriorate by, at most, a logarithmic factor. A further improvement is offered by exploiting a so-called *optimized hyperbolic cross* index set. In this case, the computational cost is $\mathcal{O}(N)$, a figure which no longer grows with dimension. Convergence rates, when measured in an appropriate norm, are not deteriorated.

5. Laplace–Dirichlet expansions have a virtually identical theory to their Laplace–Neumann counterparts. However, both the degree and rate of convergence are one order lower. In particular, much like classical Fourier expansions, the truncated expansion of a function in Laplace–Dirichlet eigenfunctions does not converge uniformly on $\bar{\Omega}$, and suffers from the Gibbs phenomenon.
6. Both the rate and degree of convergence of Laplace–Dirichlet and Laplace–Neumann expansions are determined by whether the function being approximated satisfies certain derivative conditions on the boundary $\partial\Omega$. If a function satisfies the first k such conditions then all rates of convergence increase by a factor of N^{2k}.
7. Both Laplace–Neumann and Laplace–Dirichlet coefficients can be calculated using combinations of classical and highly oscillatory quadratures. Unlike the FFT, such schemes are adaptive (in that coefficients are calculated one by one), therefore permitting the use of hyperbolic cross index sets.

The material in this chapter is based on the author's papers [5, 3].

2.2 Definition and basic properties

Let $\Omega = (-1, 1)$ be the unit interval.[1] On this domain, the eigenfunctions of the Laplace operator subject to homogeneous Dirichlet and Neumann boundary conditions are given by

$$\psi_0^{[0]} = \psi_0^{[1]} = 0, \quad \psi_n^{[0]}(x) = \cos(n - \tfrac{1}{2})\pi x, \quad \psi_n^{[1]}(x) = \sin n\pi x, \quad n \in \mathbb{N}, \tag{2.1}$$

and

$$\phi_0^{[0]}(x) = \frac{1}{\sqrt{2}}, \quad \phi_0^{[1]}(x) = 0, \quad \phi_n^{[0]}(x) = \cos n\pi x, \quad \phi_n^{[1]}(x) = \sin(n - \tfrac{1}{2})\pi x, \quad n \in \mathbb{N}, \tag{2.2}$$

respectively. Note that, for ease of notation, we define $\psi_0^{[0]}$, $\psi_0^{[1]}$ and $\phi_0^{[1]}$ in this manner. Aside from the zero Neumann eigenvalue, both sets of eigenfunctions share the common eigenvalues

$$\mu_n^{[0]} = (\alpha_n^{[0]})^2 = n^2\pi^2, \quad \mu_n^{[1]} = (\alpha_n^{[1]})^2 = (n - \tfrac{1}{2})^2\pi^2, \quad n \in \mathbb{N}. \tag{2.3}$$

Density of both sets of eigenfunctions is immediately confirmed:

Lemma 2.1. *The eigenfunctions (2.1) and (2.2) form orthonormal bases of* $\mathrm{L}^2(\Omega)$.

Proof. This is a standard result of spectral theory [2]. □

The corresponding multivariate eigenfunctions in the d-variate cube $\Omega = (-1, 1)^d$, $d \in \mathbb{N}$, arise precisely from Cartesian products:

[1] We define Ω in this manner for the purpose of symmetry, at the expense of having two types of eigenfunctions (even and odd respectively).

Lemma 2.2. *The Laplace–Dirichlet and Laplace–Neumann eigenfunctions on* $\Omega = (-1,1)^d$ *are precisely Cartesian products of univariate eigenfunctions (2.1) and (2.2) respectively. Moreover, they form orthonormal bases of* $\mathrm{L}^2(\Omega)$.

Proof. As in Lemma 2.1 it is readily verified that the sets of multivariate Laplace–Dirichlet and Laplace–Neumann eigenfunctions form orthonormal bases of $\mathrm{L}^2(\Omega)$. Trivially, any Cartesian product of univariate eigenfunctions is a multivariate eigenfunction. By standard arguments (see [153, p.193]), the set of Cartesian products of eigenfunctions also forms an orthonormal basis of $\mathrm{L}^2(\Omega)$. By density and orthogonality, no other eigenfunctions are permissible. □

If $x = (x_1, \ldots, x_d) \in [-1,1]^d$, $n = (n_1, \ldots, n_d) \in \mathbb{N}_0^d$ and $i = (i_1, \ldots, i_d) \in \{0,1\}^d$ we write

$$\psi_n^{[i]}(x) = \prod_{j=1}^d \psi_{n_j}^{[i_j]}(x_j), \quad \phi_n^{[i]}(x) = \prod_{j=1}^d \phi_{n_j}^{[i_j]}(x_j),$$

for the multivariate eigenfunctions. The corresponding eigenvalues are $\mu_n^{[i]} = \sum_{j=1}^d \mu_{n_j}^{[i_j]}$. Suppose now that $f \in \mathrm{L}^2(\Omega)$ and $N \in \mathbb{N}$. We define the truncated expansion of f in Laplace–Neumann eigenfunctions by

$$\mathcal{F}_N[f](x) = \sum_{i \in \{0,1\}^d} \sum_{n \in I_N} \hat{f}_n^{[i]} \phi_n^{[i]}(x), \quad x \in [-1,1]^d, \tag{2.4}$$

where $\hat{f}_n^{[i]}$ is the coefficient of f corresponding to the eigenfunction $\phi_n^{[i]}$,

$$\hat{f}_n^{[i]} = \left(f, \phi_n^{[i]}\right) = \int_\Omega f(x) \phi_n^{[i]}(x)\, \mathrm{d}x, \quad i \in \{0,1\}^d, \quad n \in \mathbb{N}_0^d, \tag{2.5}$$

and $(\cdot,\cdot)$ is the standard $\mathrm{L}^2(\Omega)$ inner product. Here $I_N \subseteq \mathbb{N}_0^d$ is some finite index set. Usually, $I_N = \{0, \ldots, N\}$ in the univariate setting, so that

$$\mathcal{F}_N[f](x) = \frac{1}{\sqrt{2}} \hat{f}_0^{[0]} + \sum_{n=1}^N \hat{f}_n^{[0]} \cos n\pi x + \hat{f}_n^{[1]} \sin(n - \tfrac{1}{2})\pi x, \quad x \in [-1,1].$$

For multivariate expansions various different choices of I_N are possible. We consider this further in Sections 2.9 and 2.10. Suppose now that we define the finite dimensional space

$$\mathcal{S}_N = \operatorname{span}\left\{ \phi_n^{[i]} : n \in I_N, i \in \{0,1\}^d \right\}. \tag{2.6}$$

Then the operator $\mathcal{F}_N$, as defined in (2.4), is the the orthogonal projection $\mathrm{L}^2(\Omega) \to \mathcal{S}_N$ with respect to $(\cdot,\cdot)$. Similarly, the truncated expansion of f in Laplace–Dirichlet eigenfunctions

$$\mathcal{F}_N[f](x) = \sum_{i \in \{0,1\}^d} \sum_{n \in I_N} \check{f}_n^{[i]} \psi_n^{[i]}(x), \quad x \in [-1,1]^d, \tag{2.7}$$

where $\check{f}_n^{[i]} = (f, \psi_n^{[i]})$, is the orthogonal projection onto the space $\mathcal{S}_N$ spanned by the functions $\psi_n^{[i]}$ with $n \in I_N$, $i \in \{0,1\}^d$.

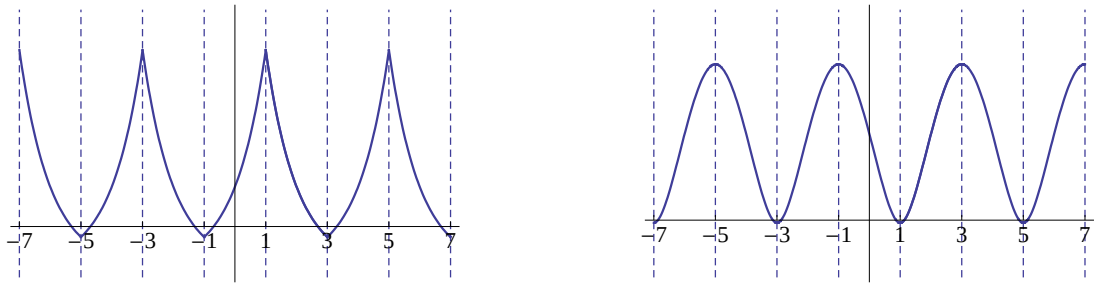


Figure 2.1: Even periodic extensions of the functions $f(x) = \mathrm{e}^x$ (left) and $f(x) = \mathrm{e}^x - x\cosh 1 - \frac{1}{2}x^2 \sinh 1$.

The focus of the first part of this chapter is the convergence of $\mathcal{F}_N[f]$ to f in various norms. The results we prove are essentially independent of the choice of index set I_N, however we will impose the following mild conditions

$$I_1 \subseteq I_2 \subseteq \ldots \subseteq \mathbb{N}_0^d, \quad \bigcup_{N\in\mathbb{N}} I_N = \mathbb{N}_0^d. \tag{2.8}$$

With this assumption to hand, we may now prove a version of Parseval's theorem [107] for these bases:

Theorem 2.3 (Parseval)**.** *Suppose that $\mathcal{S}_N$, $\mathcal{F}_N$ are defined for either Laplace–Neumann or Laplace–Dirichlet eigenfunctions. Suppose further that $f \in \mathrm{L}^2(\Omega)$ and that I_N satisfies (2.8). Then, in both cases, $\mathcal{F}_N[f]$ is the best approximation to f from $\mathcal{S}_N$ in the $\mathrm{L}^2(\Omega)$ norm and $\|f - \mathcal{F}_N[f]\| \to 0$ as $N \to \infty$. Moreover, we have the characterisations*

$$\|f\|^2 = \sum_{i\in\{0,1\}^d} \sum_{n\in\mathbb{N}_0^d} |\hat{f}_n^{[i]}|^2, \quad \|f\|^2 = \sum_{i\in\{0,1\}^d} \sum_{n\in\mathbb{N}^d} |\check{f}_n^{[i]}|^2. \tag{2.9}$$

Proof. Since $f - \mathcal{F}_N[f]$ is orthogonal to $\mathcal{S}_N$, we have

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

for any $\phi \in \mathcal{S}_N$. Hence, $\mathcal{F}_N[f]$ is the best approximation to f from the set $\mathcal{S}_N$. Using Lemmas 2.1 and 2.2 we immediately deduce that $\|f - \mathcal{F}_N[f]\| \to 0$ as $N \to \infty$. To establish (2.9) we first note that the result holds for any $f \in \mathcal{S}_N$. For $f \in \mathrm{L}^2(\Omega)$ we write $f = \mathcal{F}_N[f] + (f - \mathcal{F}_N[f])$ and use orthogonality and the previous result. □

2.3 Comparison to Fourier series

Both Laplace–Dirichlet and Laplace–Neumann expansions exhibit a close relation to classical Fourier series. For example, if a univariate function f defined on $[-1, 1]$ is extended evenly to the real line (see Figure 2.1), then its expansion in Laplace–Neumann eigenfunctions on $[-1, 1]$ is precisely the Fourier series of the periodic extension on $[-2, 2]$ [90]. Similarly, the Laplace–Dirichlet expansion relates to the odd extension of f.

Despite this interpretation, an accurate study of modified Fourier expansions is best achieved without applying known results from Fourier analysis. Moreover, the techniques we develop in this chapter are readily transferrable to a variety of other eigenfunction expansions, most of which do not share this connection to the Fourier basis (see Chapter 3 and Section 2.11).

It is well known that the convergence rate of the Fourier expansion $\mathcal{F}_N[f]$ of a periodic function f is completely governed by its smoothness. In fact, $\|f-\mathcal{F}_N[f]\|_r \leq cN^{r-s}\|f\|_s$ for all $r, s \in \mathbb{N}_0$ [42], where $\|\cdot\|_r$ is the r^{th} classical Sobolev norm. The interpretation of the Laplace–Dirichlet and Laplace–Neumann expansions in this manner demonstrates the barrier to fast convergence. The even extension of f suffers from 'jumps' in its odd derivatives at $x = \pm 1$, whereas the odd extension possesses such jumps in its even derivatives. These *derivative conditions* completely determine the convergence rate of such expansions. As demonstrated in Figure 2.1, if f has no jump in its first derivative, its even periodic extension has higher regularity, thus guaranteeing faster convergence. We explore this issue in greater detail in the forthcoming sections.

2.4 Derivative conditions

Unlike their Fourier counterpart, the bases of Laplace–Dirichlet or Laplace–Neumann eigenfunctions are not closed under differentiation. However, the derivative of a univariate Laplace–Neumann eigenfunction is proportional to a Laplace–Dirichlet eigenfunction and vice-versa:

$$(\phi_n^{[i]})' = (-1)^{1+i}\alpha_n^{[i]}\psi_n^{[1-i]}, \quad (\psi_n^{[i]})' = (-1)^{1+i}\alpha_n^{[i]}\phi_n^{[1-i]}, \quad i \in \{0,1\}, \quad n \in \mathbb{N}_0. \tag{2.10}$$

Suppose that we write $\Gamma = \partial\Omega$ for the boundary of Ω and define the subset Γ_j of Γ by $\Gamma_j = \{x \in \Gamma : x_j = \pm 1\}$, $j = 1, \ldots, d$. We may now generalise this observation to the d-variate cube:

Lemma 2.4. *Suppose that $\beta = (\beta_1, \ldots, \beta_d) \in \mathbb{N}_0^d$ and $\Omega = (-1,1)^d$. Then, if we apply the operator $\mathrm{D}^\beta = \partial_{x_1}^{\beta_1} \ldots \partial_{x_d}^{\beta_d}$ to the set of Laplace–Neumann eigenfunctions on Ω we obtain, up to scalar multiples, the set of Laplace eigenfunctions that satisfy homogeneous Neumann boundary conditions on the faces Γ_j where β_j is even, and homogeneous Dirichlet boundary conditions elsewhere. Such eigenfunctions are orthogonal and dense in $\mathrm{L}^2(\Omega)$.*

Proof. From (2.10) it follows that

$$\mathrm{D}^\beta \phi_n^{[i]} = \prod_{j=1}^d \partial_{x_j}^{\beta_j} \phi_{n_j}^{[i_j]} = \prod_{j=1}^d (-1)^{(1+i_j)\beta_j} (\alpha_{n_j}^{[i_j]})^{\beta_j} \theta_{n_j}^{[i_j]} = \left[\prod_{j=1}^d (-1)^{(1+i_j)\beta_j} (\alpha_{n_j}^{[i_j]})^{\beta_j}\right] \theta_n^{[i]},$$

where $\theta_{n_j}^{[i_j]}$ is a univariate Laplace–Neumann (respectively Laplace–Dirichlet) eigenfunction if β_j is even (odd). Using identical arguments to those given in Section 2.2, we deduce orthogonality and density of the eigenfunctions $\theta_n^{[i]}$. □

This duality is essential to establishing the convergence of such eigenseries. It also underscores why a concurrent study of both sets of eigenfunctions is necessary. Moreover, as we demonstrate in Chapter 5, such duality also has a practical usage.

Though convergence of $\mathcal{F}_N[f]$ to f in the $\mathrm{L}^2(\Omega)$ norm is guaranteed, the rate of convergence may be arbitrarily slow. In the sequel, we verify that this rate improves if the truncated expansion $\mathcal{F}_N[f]$ converges to f in higher-order Sobolev norms. Our goal now is to derive conditions on the function f that ensure such convergence.

At this point, we remark that the primary consideration of this chapter will be the convergence of such expansions in the uniform norm as well as various Sobolev norms with index $p = 2$, i.e. the spaces $\mathrm{H}^r(\Omega)$. Convergence of classical Fourier series in a variety of other norms is the subject of a vast array of literature. We cannot hope to obtain analogous results for modified Fourier expansions within the confines of one chapter. Needless to say, a study of convergence in the spaces $\mathrm{H}^r(\Omega)$ is a natural choice in view of the primary application, boundary value problems, addressed in this thesis.

Returning to derivative conditions, consider a univariate function $f \in \mathrm{C}^\infty[-1,1]$. Since $\phi_n^{[i]}$ is an eigenfunction of the univariate Laplace operator, we have

$$\hat{f}_n^{[i]} = \int_{-1}^{1} f(x)\phi_n^{[i]}(x)\,\mathrm{d}x = -\frac{1}{\mu_n^{[i]}}\int_{-1}^{1} f(x)(\phi_n^{[i]})''(x)\,\mathrm{d}x, \quad n \in \mathbb{N}.$$

Integrating this expression by parts and applying the boundary conditions for $\phi_n^{[i]}$ we obtain

$$\hat{f}_n^{[i]} = \frac{(-1)^{i+1}}{\alpha_n^{[i]}}\breve{f'}_n^{[1-i]}.$$

Integrating by parts once more and iterating the result gives

$$\begin{aligned}\hat{f}_n^{[i]} &= \frac{(-1)^{n+i}}{\mu_n^{[i]}}\left\{f'(1) + (-1)^{i+1}f'(-1)\right\} - \frac{1}{\mu_n^{[i]}}\widehat{f''}_n^{[i]} \\ &= \sum_{r=0}^{k-1}\frac{(-1)^{r+n+i}}{(\mu_n^{[i]})^{r+1}}\left\{f^{(2r+1)}(1) + (-1)^{i+1}f^{(2r+1)}(-1)\right\} + \frac{(-1)^k}{(\mu_n^{[i]})^k}\widehat{f^{(2k)}}_n^{[i]}, \quad k,n \in \mathbb{N}. \quad (2.11)\end{aligned}$$

We immediately observe that the coefficients $\hat{f}_n^{[i]}$ are $\mathcal{O}\left(n^{-2}\right)$ in general and $\mathcal{O}\left(n^{-2k-2}\right)$ provided f satisfies $f^{(2r+1)}(\pm 1) = 0$, $r = 0,\ldots,k-1$. A similar observation, which we explore in greater detail in Section 2.7, holds regarding the coefficients of a multivariate function. If $f \in \mathrm{C}^\infty([-1,1]^d)$, it turns out that $\hat{f}_n^{[i]}$ is $\mathcal{O}(n_1^{-2}\ldots n_d^{-2})$ in general and $\mathcal{O}(n_1^{-2k-2}\ldots n_d^{-2k-2})$ provided f satisfies the first k *Neumann derivative conditions*

$$\partial_{x_j}^{2r+1} f\big|_{\Gamma_j} = 0, \quad j = 1,\ldots,d, \quad r = 0,\ldots,k-1. \quad (2.12)$$

Aside from giving faster decay of the coefficients, such conditions also ensure convergence of the expansion in higher-order Sobolev norms, as we shall demonstrate. Observe that, in particular, the eigenfunctions $\phi_n^{[i]}$ automatically satisfy all derivative conditions.

In the Dirichlet setting we obtain (by identical arguments) the following

$$\check{f}_n^{[i]} = \sum_{r=0}^{k-1}\frac{(-1)^{r+n}}{(\mu_n^{[i]})^{r+\frac{1}{2}}}\left\{f^{(2r)}(1) + (-1)^i f^{(2r)}(-1)\right\} + \frac{(-1)^k}{(\mu_n^{[i]})^{k-\frac{1}{2}}}\widehat{f^{(2k-1)}}_n^{[i]}, \quad k,n \in \mathbb{N}. \quad (2.13)$$

In an analogous manner, we say that a function f satisfies the first k *Dirichlet derivative conditions* if

$$\partial_{x_j}^{2r} f\big|_{\Gamma_j} = 0, \quad j = 1,\ldots,d, \quad r = 0,\ldots,k-1. \quad (2.14)$$

Once more, this ensures convergence of the Dirichlet expansion in higher-order norms.

As discussed in the previous section, these conditions are closely associated with periodic extensions. The standard setting for Fourier analysis arises upon introduction of the periodic spaces $\mathrm{H}^r(\mathbb{T}^d)$, where $\mathbb{T}^d$ is the d-variate torus. Though possible, we shall not adopt a similar approach for Laplace eigenfunction expansions.[2]

Nonetheless, it is of both theoretical interest and practical use to consider functions which satisfy the first k such conditions. From the former standpoint, this governs the rate of convergence of the approximation. For practical purposes, devices for accelerating convergence of such approximations are typically based on interpolating the first k derivative conditions. For example, if this is achieved with a function g_k, then the new approximation to f given by $\mathcal{F}_N[f - g_k] + g_k$ converges at a faster rate. We devote Chapter 5 to this topic.

In Lemma 2.4 we exhibited the duality of the Laplace–Dirichlet and Laplace–Neumann bases. For functions that obey the first k derivative conditions, a similar duality holds for the truncated expansion $\mathcal{F}_N[f]$. We have

Lemma 2.5. *Suppose that $f \in \mathrm{C}^\infty(\bar{\Omega})$, where $\Omega = (-1,1)^d$, and that f satisfies the first k Neumann derivative conditions (2.12) for some $k \in \mathbb{N}_0$.*[3] *Suppose further that $\mathcal{F}_N[f]$ is the truncated expansion of f in Laplace–Neumann eigenfunctions. If $\beta \in \mathbb{N}_0^d$ and $|\beta|_\infty \leq 2k+1$, then $\mathrm{D}^\beta \mathcal{F}_N[f]$ is the truncated expansion of $\mathrm{D}^\beta f$ in Laplace eigenfunctions that obey homogeneous Neumann boundary conditions in the variables x_j when β_j is even and Dirichlet boundary conditions elsewhere.*

Conversely, if f satisfies the first k Dirichlet derivative conditions (2.14) for some $k \in \mathbb{N}$ and $\mathcal{F}_N[f]$ is its truncated expansion in Laplace–Dirichlet eigenfunctions then, for $|\beta|_\infty \leq 2k$, $\mathrm{D}^\beta \mathcal{F}_N[f]$ is the truncated expansion of $\mathrm{D}^\beta f$ in Laplace eigenfunctions that obey homogeneous Dirichlet boundary conditions in the variables x_j when β_j is even and Neumann boundary conditions elsewhere.

Proof. Consider the coefficient $\hat{f}_n^{[i]}$. Since f obeys the first k derivative conditions, we may integrate by parts $\beta_j \leq 2k+1$ times in each variable with vanishing boundary terms. Hence

$$\hat{f}_n^{[i]} = \Big[\prod_{j=1}^{d} (-1)^{(1+i_j)\beta_j} (\alpha_{n_j}^{[i_j]})^{\beta_j} \Big]^{-1} \int_\Omega \mathrm{D}^\beta f(x) \theta_n^{[i]}(x) \, \mathrm{d}x,$$

where $\theta_n^{[i]}$ is the Laplace eigenfunction that obeys homogeneous Neumann boundary conditions in the variables x_j where β_j is even and homogeneous Dirichlet boundary conditions elsewhere. Using this and Lemma 2.4, we obtain

$$\mathrm{D}^\beta \mathcal{F}_N[f](x) = \sum_{i \in \{0,1\}^d} \sum_{n \in I_N} \hat{f}_n^{[i]} \mathrm{D}^\beta \phi_n^{[i]}(x) = \sum_{i \in \{0,1\}^d} \sum_{n \in I_N} \left(\mathrm{D}^\beta f, \theta_n^{[i]} \right) \theta_n^{[i]}(x),$$

which gives the result for the Neumann case. The proof for Laplace–Dirichlet expansions is virtually identical. □

[2] The primary reason for not doing so is that we do not want the number of derivative conditions satisfied to be completely determined by the degree of smoothness. Such approach would not allow us to derive quasi-optimal error estimates as readily.

[3] As a convention, when $k = 0$ we mean that the function satisfies no derivative conditions.

We immediately note one ramification of this lemma: for equal numbers of derivative conditions satisfied, the quantity $\mathrm{D}^\beta \mathcal{F}_N[f]$ is understood in terms of some orthogonal expansion of $\mathrm{D}^\beta f$ for larger values of $|\beta|_\infty$ when $\mathcal{F}_N[f]$ is the expansion of f in Neumann eigenfunctions. In particular, if the function f satisfies no Neumann or Dirichlet conditions then $\mathrm{D}^\beta \mathcal{F}_N[f]$ is known for $|\beta|_\infty \leq 1$ in the Neumann case, but only $\beta = 0$ for the Dirichlet expansion. As we next demonstrate, a consequence of this observation is uniform convergence of the expansion in Laplace–Neumann eigenfunctions. Conversely, the Laplace–Dirichlet expansion suffers from the Gibbs phenomenon.[4]

The result of this lemma also indicates that the classical Sobolev norms are insufficient for a study of multivariate Laplace eigenfunction expansions. By definition, if $f \in \mathrm{H}^{2k+1}(\Omega)$ then $\mathrm{D}^\beta f \in \mathrm{L}^2(\Omega)$ for all $|\beta| \leq 2k+1$. However, in Lemma 2.5 the quantity $\mathrm{D}^\beta \mathcal{F}_N[f]$ is understood for not only such β, but also any value β with $|\beta|_\infty \leq 2k+1$. This warrants the introduction of a new type of Sobolev space, a topic we now consider.

2.5 Sobolev spaces of dominating mixed smoothness

Sobolev spaces of dominating mixed smoothness are the standard setting whenever a hyperbolic cross index set (a device we consider in Section 2.10) or a sparse grid is employed [41, 145, 158]. In the particular case of Laplace eigenfunction expansions, such spaces provide a suitable framework for analysis, even for arbitrary index sets. Subsequently, we shall also see that the associated norms are precisely those required to bound the coefficients $\hat{f}_n^{[i]}$ and $\check{f}_n^{[i]}$ in inverse powers of $n_1 \dots n_d$, which leads to quasi-optimal approximation error estimates.

For $r \in \mathbb{N}_0$ we define the r^{th} *Sobolev space of dominating mixed smoothness*[5] by

$$\mathrm{H}^r_{\text{mix}}(\Omega) = \{f : \mathrm{D}^\beta f \in \mathrm{L}^2(\Omega),\ \forall\ \beta \in \mathbb{N}_0^d : |\beta|_\infty \leq r\}, \tag{2.15}$$

where the derivative D^β is taken in the sense of distributions, with associated norm

$$\|f\|^2_{r,\text{mix}} = \sum_{|\beta|_\infty \leq r} \|\mathrm{D}^\beta f\|^2. \tag{2.16}$$

This space is also commonly denoted by $\mathrm{S}_2^{(r,\dots,r)}\mathrm{H}(\Omega)$ in literature [145, 158].[6] Note that $\mathrm{H}^{rd}(\Omega) \subseteq \mathrm{H}^r_{\text{mix}}(\Omega) \subseteq \mathrm{H}^r(\Omega)$. It is readily seen that the condition $f \in \mathrm{C}^\infty(\bar{\Omega})$ in Lemma 2.5 can be replaced by $f \in \mathrm{H}^{2k+1}_{\text{mix}}(\Omega)$ or $f \in \mathrm{H}^{2k}_{\text{mix}}(\Omega)$ in the Neumann or Dirichlet cases respectively, thus motivating the use of such spaces in this context.

We note in passing the following geometric interpretation: $\mathrm{H}^r_{\text{mix}}(-1,1)^d$ is isomorphic to the tensor-product space $\mathcal{H}^r_{\text{mix}}(-1,1)^d = \mathrm{H}^r(-1,1) \otimes \dots \otimes \mathrm{H}^r(-1,1)$ [85]. Though we shall not make use of this fact directly, we will repeatedly use the following result, which follows immediately from this equivalent definition. This is the standard approximation by smooth functions property: given $f \in \mathrm{H}^r_{\text{mix}}(\Omega)$ and $\epsilon > 0$ there exists $g \in \mathrm{C}^\infty(\bar{\Omega})$ such that $\|f - g\|_{r,\text{mix}} < \epsilon$.

[4] As in [72], we interpret the Gibbs phenomenon as the issue of recovering local information (function values) from global information (coefficients). Several manifestations of this are the slow decay of the coefficients, the lack of uniform convergence and the presence of $\mathcal{O}(1)$ oscillations near the boundary.

[5] In the periodic setting, such spaces are isomorphic to the *Korobov spaces* [158].

[6] This space is a particular example of the spaces $\mathrm{S}_2^\gamma\mathrm{H}(\Omega)$, where $\gamma = (\gamma_1, \dots, \gamma_d) \in \mathbb{N}_0^d$ [145], with $\gamma_1 = \dots = \gamma_d = r$. Though it is possible to study Laplace eigenfunction expansions in such setting, we shall not do this here.

Imbedding theorems for the spaces (2.15) are of central importance to our study. In particular, we require imbeddings in the Hölder spaces $\mathrm{C}^{r,\lambda}(\bar{\Omega})$, $r \in \mathbb{N}_0$, $0 \leq \lambda < 1$. It turns out that, unlike the classical Sobolev spaces (see, for example [2, 56]), imbeddings for the spaces $\mathrm{H}^r_{\mathrm{mix}}(\Omega)$ are essentially independent of the dimension d. Our first result is the following:

Lemma 2.6. *We have the continuous imbedding* $\mathrm{H}^{r+1}_{mix}(\Omega) \hookrightarrow \mathrm{C}^r(\bar{\Omega})$ *for* $r \in \mathbb{N}_0$.

To prove this lemma we require the following observation:

Lemma 2.7. *Suppose that* $f \in \mathrm{C}^\infty(\bar{\Omega})$. *Then*

$$f(x) = \sum_{t \in [d]^*} \int_{-1}^{x_{t_1}} \dots \int_{-1}^{x_{t_{|t|}}} \mathrm{D}_t f(x_t; -1)\, \mathrm{d}x_{t_1} \dots \mathrm{d}x_{t_{|t|}}, \quad x \in \bar{\Omega}, \tag{2.17}$$

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* $(x_t; -1) \in \bar{\Omega}$ *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 (2.17) 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 $[d]^*$ consists of elements t, $(t, d) = (t_1, \dots, t_{|t|}, d)$, where $t \in [d-1]^*$, this expression reduces to (2.17). □

Proof of Lemma 2.6. For $|\beta|_\infty \leq r$, we have $\mathrm{D}^\beta f \in \mathrm{H}^1_{\mathrm{mix}}(\Omega)$. Hence it suffices to derive the result for $r = 0$. To prove this result we first demonstrate that the inequality

$$\|f\|_\infty \leq c \|f\|_{1,\mathrm{mix}}, \tag{2.18}$$

holds for all $f \in \mathrm{C}^\infty(\bar{\Omega})$ and some positive constant c independent of f. To do so, we note that

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

where $\bar{t} \in [d]^* = [d] \cup \{\emptyset\}$ is the tuple of length $d - |t|$ of elements not in t. After an application of Lemma 2.7, we obtain

$$f(x) = \sum_{t \in [d]^*} \int_{-1}^{1} \dots \int_{-1}^{1} \int_{-1}^{x_{t_1}} \dots \int_{-1}^{x_{t_{|t|}}} \mathrm{D} \left(f(x) \prod_{j \notin t} \frac{x_j - 1}{2} \right) \mathrm{d}x_t\, \mathrm{d}x_{\bar{t}}. \tag{2.19}$$

Each integrand involves terms of the form $\mathrm{D}^\beta f$ for some $|\beta|_\infty \leq 1$. Hence, using the Cauchy–Schwarz inequality and replacing suitable upper limits of integration by 1, we obtain (2.18) for $f \in \mathrm{C}^\infty(\bar{\Omega})$.

We now proceed in the standard manner. If $f \in \mathrm{H}^1_{\mathrm{mix}}(\Omega)$ then f is the limit in $\mathrm{H}^1_{\mathrm{mix}}(\Omega)$ of a sequence of functions belonging to $\mathrm{C}^\infty(\bar{\Omega})$. Since (2.18) holds for all $g \in \mathrm{C}^\infty(\bar{\Omega})$, this sequence converges uniformly on $\bar{\Omega}$ to $\tilde{f} \in \mathrm{C}(\bar{\Omega})$. Since $f = \tilde{f}$ a.e. the result follows. □

In fact, it turns out that a stronger result can also be established:

Theorem 2.8. *We have the continuous imbedding* $\mathrm{H}^{r+1}_{mix}(\Omega) \hookrightarrow \mathrm{C}^{r,\frac{1}{2}}(\bar{\Omega})$ *for* $r \in \mathbb{N}_0$.

Proof. Once more it is sufficient to prove this result for $r = 0$. In view of Lemma 2.6, we may assume that $f \in \mathrm{C}(\bar{\Omega})$. Therefore, it suffices to establish that

$$\sup_{\substack{x,y\in\Omega\\ x\neq y}} \frac{|f(x)-f(y)|}{|x-y|^{\frac{1}{2}}} \leq c\|f\|_{1,\mathrm{mix}}, \tag{2.20}$$

for some positive constant c independent of f, where $|x| = |x_1| + \ldots |x_d|$ for $x \in \mathbb{R}^d$. By standard arguments, we may assume that $f \in \mathrm{C}^\infty(\bar{\Omega})$. We have

$$\begin{aligned} f(x) - f(y) =& f(x_1,\ldots,x_d) - f(y_1,x_2,\ldots,x_d) \\ &+ \sum_{j=2}^{d} (-1)^j \left\{ f(y_1,\ldots,y_{j-1},x_j,\ldots,x_d) - f(y_1,\ldots,y_j,x_{j+1},\ldots,x_d) \right\} \\ =& \int_{y_1}^{x_1} \partial_{x_1} f(x_1,\ldots,x_d)\,\mathrm{d}x_1 + \sum_{j=2}^{d} (-1)^j \int_{y_j}^{x_j} \partial_{x_j} f(y_1,\ldots,y_{j-1},x_j,\ldots,x_d)\,\mathrm{d}x_j. \end{aligned}$$

Hence, using the Cauchy–Schwarz inequality and the result of Lemma 2.6 for $d-1$, we obtain

$$|f(x)-f(y)| \leq c\|f\|_{1,\mathrm{mix}} \sum_{j=1}^{d} |x_j - y_j|^{\frac{1}{2}} \leq c\|f\|_{1,\mathrm{mix}} |x-y|^{\frac{1}{2}}.$$

This yields (2.20). □

2.6 Convergence of Laplace eigenseries

We are now in a position to assess the convergence of Laplace–Neumann and Laplace–Dirichlet expansions in various norms. Concerning convergence in the classical Sobolev norms, we have the following result:

Lemma 2.9. *Suppose that* $f \in \mathrm{H}^{2k+l}(\Omega)$, $l = 0, 1$, *obeys the first* $k \in \mathbb{N}_0$ *Neumann derivative conditions (2.12) and that* $\mathcal{F}_N[f]$ *is the truncated expansion of* f *in Laplace–Neumann eigenfunctions. Then, for* $r = 0,\ldots,2k+l$, $\mathcal{F}_N[f]$ *is the best approximation to* f *from* $\mathcal{S}_N$ *in the* $\mathrm{H}^r(\Omega)$ *norm,* $\|f - \mathcal{F}_N[f]\|_r \to 0$ *and we have the characterisation*

$$\|f\|_r^2 = \sum_{i\in\{0,1\}^d} \sum_{n\in\mathbb{N}_0^d} \left[\sum_{|\beta|\leq r} \prod_{j=1}^{d} (\mu_{n_j}^{[i_j]})^{\beta_j} \right] |\hat{f}_n^{[i]}|^2, \quad r = 0,\ldots,2k+l.$$

If $f \in \mathrm{H}^{2k+l-1}(\Omega)$ obeys the first k Dirichlet derivative conditions (2.14) and $\mathcal{F}_N[f]$ is its truncated expansion in Laplace–Dirichlet eigenfunctions, then $\mathcal{F}_N[f]$ is the best approximation to f in the $\mathrm{H}^r(\Omega)$ norm for $r = 0, \ldots, 2k+l-1$, $\|f - \mathcal{F}_N[f]\|_r \to 0$ and

$$\|f\|_r^2 = \sum_{i\in\{0,1\}^d} \sum_{n\in\mathbb{N}^d} \left[\sum_{|\beta|\le r} \prod_{j=1}^{d} (\mu_{n_j}^{[i_j]})^{\beta_j} \right] |\check{f}_n^{[i]}|^2, \quad r = 0, \ldots, 2k+l-1.$$

Proof. Consider the Neumann case. By Lemma 2.5, for each $|\beta| \le 2k+l$, $\mathrm{D}^\beta \mathcal{F}_N[f]$ is the truncated expansion of $\mathrm{D}^\beta f$ in an orthonormal basis of $\mathrm{L}^2(\Omega)$. Hence, by a version of Parseval's theorem for such basis, we have

$$\|\mathrm{D}^\beta(f - \mathcal{F}_N[f])\|^2 = \sum_{i\in\{0,1\}^d} \sum_{n\in\mathbb{N}_0^d} \prod_{j=1}^{d} (\mu_{n_j}^{[i_j]})^{\beta_j} |\hat{f}_n^{[i]}|^2.$$

Summing over $|\beta| \le r$ now gives the result. The Dirichlet case is identical. □

Using an identical method of proof, we also obtain an analogous result for the mixed Sobolev norms:

Lemma 2.10. *Suppose that $f \in \mathrm{H}_{mix}^{2k+l}(\Omega)$, $l = 0, 1$, obeys the first $k \in \mathbb{N}_0$ Neumann derivative conditions (2.12) and that $\mathcal{F}_N[f]$ is the truncated expansion of f in Laplace–Neumann eigenfunctions. Then, for $r = 0, \ldots, 2k+l$, $\mathcal{F}_N[f]$ is the best approximation to f from $\mathcal{S}_N$ in the $\mathrm{H}_{mix}^r(\Omega)$ norm, $\|f - \mathcal{F}_N[f]\|_{r,mix} \to 0$ and we have the characterisation*

$$\|f\|_{r,mix}^2 = \sum_{i\in\{0,1\}^d} \sum_{n\in\mathbb{N}_0^d} \left[\sum_{|\beta|_\infty\le r} \prod_{j=1}^{d} (\mu_{n_j}^{[i_j]})^{\beta_j} \right] |\hat{f}_n^{[i]}|^2, \quad r = 0, \ldots, 2k+l.$$

If $f \in \mathrm{H}_{mix}^{2+l-1}(\Omega)$, $l = 0, 1$, obeys the first k Dirichlet derivative conditions (2.14) and $\mathcal{F}_N[f]$ is its truncated expansion in Laplace–Dirichlet eigenfunctions, then $\mathcal{F}_N[f]$ is the best approximation to f in the $\mathrm{H}_{mix}^r(\Omega)$ norm for $r = 0, \ldots, 2k+l-1$, $\|f - \mathcal{F}_N[f]\|_{r,mix} \to 0$ and

$$\|f\|_{r,mix}^2 = \sum_{i\in\{0,1\}^d} \sum_{n\in\mathbb{N}^d} \left[\sum_{|\beta|_\infty\le r} \prod_{j=1}^{d} (\mu_{n_j}^{[i_j]})^{\beta_j} \right] |\check{f}_n^{[i]}|^2, \quad r = 0, \ldots, 2k+l-1.$$

When $k = 0$, Lemmas 2.9 and 2.10 establish the convergence of the expansion $\mathcal{F}_N[f]$ of a general function f that satisfies no derivative conditions. For Laplace–Neumann eigenfunctions these results may be restated more succinctly: the set $\{\phi_n^{[i]} : n \in \mathbb{N}_0^d, i \in \{0,1\}^d\}$ is an orthogonal basis of not just $\mathrm{L}^2(\Omega)$, but also of $\mathrm{H}^1(\Omega)$ and $\mathrm{H}_{\mathrm{mix}}^1(\Omega)$. The advantage of modified Fourier expansions over both classical Fourier and Laplace–Dirichlet expansions is now evident.

An identical method of proof to that given in Lemma 2.9 can be used to characterise the classical and mixed Sobolev semi-norms[7]. For the sake of brevity we consider only the

[7] In fact, it is easily confirmed that such characterisations exist for any finite collection of derivatives of f.

Neumann case:

$$|f|_r^2 = \sum_{|\beta|=r} \|\mathrm{D}^\beta f\|^2 = \sum_{i\in\{0,1\}^d} \sum_{n\in\mathbb{N}_0^d} \left[\sum_{|\beta|=r} \prod_{j=1}^d (\mu_{n_j}^{[i_j]})^{\beta_j} \right] |\hat{f}_n^{[i]}|^2,$$

$$|f|_{r,\mathrm{mix}}^2 = \sum_{|\beta|_\infty=r} \|\mathrm{D}^\beta f\|^2 = \sum_{i\in\{0,1\}^d} \sum_{n\in\mathbb{N}_0^d} \left[\sum_{|\beta|_\infty=r} \prod_{j=1}^d (\mu_{n_j}^{[i_j]})^{\beta_j} \right] |\hat{f}_n^{[i]}|^2, \quad r = 0, \ldots, 2k+l. \tag{2.21}$$

Such characterisations, along with those given in Lemmas 2.9 and 2.10, can be greatly simplified with a standard tool of Fourier analysis. We first recall the multinomial formula

$$(y_1 + \ldots + y_d)^r = \sum_{|\beta|=r} \frac{r!}{\beta_1! \ldots \beta_d!} \prod_{j=1}^d y_j^{\beta_j}, \quad \forall y \in \mathbb{R}^d, r \in \mathbb{N}_0.$$

Using this we deduce that there are positive constants c_1 and c_2 depending only on r such that

$$c_1 \left(y_1^2 + \ldots + y_d^2\right)^r \leq \sum_{|\beta|=r} \prod_{j=1}^d y_j^{2\beta_j} \leq c_2 \left(y_1^2 + \ldots + y_d^2\right)^r, \quad \forall y \in \mathbb{R}^d.$$

If we now consider the subspace of functions $f \in \mathrm{H}^{2k+l}(\Omega)$ that satisfy the first $k \in \mathbb{N}_0$ Neumann derivative conditions, then an equivalent norm to $\|\cdot\|_r$ on this space is given by

$$(\|f\|_r')^2 = \sum_{i\in\{0,1\}^d} \sum_{n\in\mathbb{N}_0^d} \left(1 + \mu_n^{[i]}\right)^r |\hat{f}_n^{[i]}|^2, \quad r = 0, \ldots, 2k+l. \tag{2.22}$$

The semi-norm $|\cdot|_r$ also has the following equivalent:

$$(|f|_r')^2 = \sum_{i\in\{0,1\}^d} \sum_{n\in\mathbb{N}_0^d} \left(\mu_n^{[i]}\right)^r |\hat{f}_n^{[i]}|^2, \quad r = 0, \ldots, 2k+l.$$

In an identical manner, since there exist constants c_1, c_2 depending only on r such that

$$c_1 \prod_{j=1}^d (1 + y_j^2)^r \leq \sum_{|\beta|_\infty=r} \prod_{j=1}^d y_j^{2\beta_j} \leq c_2 \prod_{j=1}^d (1 + y_j^2)^r, \quad \forall y \in \mathbb{R}^d,$$

we may attain a similar result for the mixed norms. An equivalent norm to $\|\cdot\|_{r,\mathrm{mix}}$ on the subspace of functions $f \in \mathrm{H}_{\mathrm{mix}}^{2k+l}(\Omega)$ that have vanishing Neumann derivative conditions is therefore given by

$$(\|f\|_{r,\mathrm{mix}}')^2 = \sum_{i\in\{0,1\}^d} \sum_{n\in\mathbb{N}_0^d} \prod_{j=1}^d \left(1 + \mu_{n_j}^{[i_j]}\right)^r |\hat{f}_n^{[i]}|^2, \quad r = 0, \ldots, 2k+l. \tag{2.23}$$

Likewise, an equivalent semi-norm is given by

$$(|f|_{r,\mathrm{mix}}')^2 = \sum_{i\in\{0,1\}^d} \sum_{n\in\mathbb{N}_0^d} \prod_{j=1}^d \left(\mu_{n_j}^{[i_j]}\right)^r |\hat{f}_n^{[i]}|^2, \quad r = 0, \ldots, 2k+l.$$

One immediate consequence of such norm equivalences is the following simple version of Bernstein's inequality:

Corollary 2.11 (Bernstein)**.** *Suppose that $\phi \in \mathcal{S}_N$, where $\mathcal{S}_N$ consists of either Laplace–Dirichlet or Laplace–Neumann eigenfunctions. Then*

$$\|\phi\|_r \leq \max_{n\in I_N} \left\{(1+\mu_n^{[0]})^{\frac{r}{2}}\right\}\|\phi\|, \quad \|\phi\|_{r,mix} \leq \max_{n\in I_N} \left\{\prod_{j=1}^{d}(1+\mu_{n_j}^{[0]})^{\frac{r}{2}}\right\}\|\phi\|, \quad r \in \mathbb{N}_0.$$

Proof. Note that $\mu_n^{[i]} \leq \mu_n^{[0]}$ for $n \in \mathbb{N}_0$ and $i \in \{0,1\}$. Since $\phi \in \mathcal{S}_N$ automatically satisfies all derivative conditions, the characterisations (2.22) and (2.23) now provide the result. □

We complete this section by scrutinising the uniform convergence of Laplace eigenfunction expansions:

Theorem 2.12. *Suppose that $f \in \mathrm{H}^{2k+l}_{mix}(\Omega)$, $k \in \mathbb{N}_0$, $l = 0,1$ ($l = 1$ when $k = 0$), obeys the first k Neumann derivative conditions (2.12) and that $\mathcal{F}_N[f]$ is its truncated Laplace–Neumann expansion. Then for $|\beta|_\infty \leq 2k + l - 1$, $\|\mathrm{D}^\beta(f - \mathcal{F}_N[f])\|_\infty \to 0$ as $N \to \infty$.*

If $f \in \mathrm{H}^{2k+l-1}_{mix}(\Omega)$, $k \in \mathbb{N}$, $l = 0,1$, obeys the first k Dirichlet derivative conditions (2.14) and $\mathcal{F}_N[f]$ is its truncated Laplace–Dirichlet expansion, then $\|\mathrm{D}^\beta(f - \mathcal{F}_N[f])\|_\infty \to 0$ as $N \to \infty$ for $|\beta|_\infty \leq 2(k-1) + l$.

Proof. Setting $g = \mathrm{D}^\beta(f - \mathcal{F}_N[f])$ in (2.18) and applying Lemma 2.10 gives the result. □

When $k = 0$ we surmise that the modified Fourier expansion of an arbitrary $f \in \mathrm{H}^1_{\mathrm{mix}}(\Omega)$ converges uniformly on $\bar{\Omega}$. In particular, there is no apparent Gibbs phenomenon. However, the Laplace–Dirichlet expansion—whose convergence mirrors that of a classical Fourier expansion of a nonperiodic function—does exhibit such a phenomenon, unless the function f vanishes on the boundary.

The results of Lemmas 2.9, 2.10 and Theorem 2.12 demonstrate that, for equal number of derivative conditions satisfied, the expansion in Laplace–Neumann eigenfunctions converges in higher-order norms. This equates to faster convergence of the expansion, which we subsequently demonstrate.

No stipulations are made on the index set I_N for the results proved in this section, aside from the mild conditions (2.8). The choice of index set determines the rate of convergence of the expansion, which we consider in the sequel. To do so, we first need to develop bounds for the coefficients $\hat{f}_n^{[i]}$, $\check{f}_n^{[i]}$. This is the content of the next two sections. Such bounds will also be used to obtain hyperbolic cross index sets in Section 2.10.

Returning to Theorem 2.12 briefly, we mention that the classical means to establish uniform convergence of Fourier series is by means of the Féjer and de la Vallée Poussin theorems [107]. A similar approach can be applied for univariate modified Fourier expansions [94]. However, such techniques cannot be used in two or more dimensions unless the index set employed is particularly simple.[8] Conversely, Theorem 2.12 is essentially independent of the index set. Nonetheless, the result requires $\mathrm{H}^1_{\mathrm{mix}}(\Omega)$ smoothness, which is slightly more regularity than that imposed in the classical Fourier result: when $d = 1$, the conditions $f \in C(\mathbb{T})$ and f being of bounded variation ensure uniform convergence of the Fourier series of f (see [103, p. 53]).

[8] Essentially it needs to be either a hypercube in $\mathbb{N}_0^d$ or a sum of hypercubes.

2.7 Asymptotic expansion of coefficients

The aim of this section is to extend the univariate expansions (2.11) and (2.13) to expansions for the multivariate coefficients $\hat{f}_n^{[i]}$ in inverse powers of $n_1, \dots, n_d$.[9] Such expansions not only realise robust bounds for the coefficients, they are also used in Chapter 5 as the starting point for constructing methods to accelerate convergence. From this point onwards, our primary focus is the Neumann case.

To express such expansions we first need some additional notation. Given $j = 1, \dots, d$, $r_j \in \mathbb{N}_0$ and $i_j \in \{0, 1\}$ we define $\mathcal{B}_{r_j}^{[i_j]}[f]$ by

$$\begin{aligned}(-1)^{r_j}\mathcal{B}_{r_j}^{[i_j]}[f](x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_d) &= \partial_{x_j}^{2r_j+1} f(x_1, \dots, x_{j-1}, 1, x_{j+1}, \dots, x_d) \\ &\quad + (-1)^{i_j+1}\partial_{x_j}^{2r_j+1} f(x_1, \dots, x_{j-1}, -1, x_{j+1}, \dots, x_d). \end{aligned} \tag{2.24}$$

For $t \in [d]^* = [d] \cup \{\emptyset\}$, $r_t = (r_{t_1}, \dots, r_{t_{|t|}}) \in \mathbb{N}_0^{|t|}$ and $i_t = (i_{t_1}, \dots, i_{t_{|t|}}) \in \{0, 1\}^{|t|}$ we define $\mathcal{B}_{r_t}^{[i_t]}[f]$ as the composition

$$\mathcal{B}_{r_t}^{[i_t]}[f](x_{\bar{t}}) = \mathcal{B}_{r_{t_1}}^{[i_{t_1}]}\left[\mathcal{B}_{r_{t_2}}^{[i_{t_2}]}\left[\dots\left[\mathcal{B}_{r_{t_{|t|}}}^{[i_{t_{|t|}}]}[f]\right]\dots\right]\right](x_{\bar{t}}), \tag{2.25}$$

with the understanding that when $t = \emptyset$, $\mathcal{B}_{r_t}^{[i_t]}[f] = f$. Note that the operators $\mathcal{B}_{r_j}^{[i_j]}$, $j \in t$, operators commute with each other and with differentiation in the variable $x_{\bar{t}}$. Finally, given $i \in \{0, 1\}^d$, $t \in [d]^*$, $r_t \in \mathbb{N}_0^{|t|}$ with $|r_t|_\infty \leq k - 1$ and $n_{\bar{t}} = (n_{\bar{t}_1}, \dots, n_{\bar{t}_{|\bar{t}|}}) \in \mathbb{N}^{|\bar{t}|}$ we define $\mathcal{A}_{r_t, n_{\bar{t}}}^{[i]}[f] \in \mathbb{R}$ by

$$\mathcal{A}_{r_t, n_{\bar{t}}}^{[i]}[f] = (-1)^{k|\bar{t}|}\prod_{j \notin t}\left(\mu_{n_j}^{[i_j]}\right)^{-k}\int \mathcal{B}_{r_t}^{[i_t]}[\mathrm{D}_{\bar{t}}^{2k} f](x_{\bar{t}})\phi_{n_{\bar{t}}}^{[i_{\bar{t}}]}(x_{\bar{t}})\, dx_{\bar{t}}. \tag{2.26}$$

Observe that the integral is nothing more than the Laplace–Neumann coefficient of the function $\mathcal{B}_{r_t}^{[i_t]}[\mathrm{D}_{\bar{t}}^{2k} f]$ corresponding to indices $i_{\bar{t}}$ and $n_{\bar{t}}$. Moreover, the value $\mathcal{A}_{r_t, n_{\bar{t}}}^{[i]}[f]$ also depends on k, d and $t \in [d]$. However, to simplify notation we will not make this dependence explicit.

Concerning the expansion of the coefficients $\hat{f}_n^{[i]}$, we have the following result[10]:

Lemma 2.13. *Suppose that $f \in \mathrm{H}_{mix}^{2k}(\Omega)$, $k \in \mathbb{N}$, and that $n \in \mathbb{N}^d$. Then*

$$\hat{f}_n^{[i]} = \sum_{t \in [d]^*}\sum_{|r_t|_\infty = 0}^{k-1} \mathcal{A}_{r_t, n_{\bar{t}}}^{[i]}[f](-1)^{|n_t| + |i_t|}\prod_{j \in t}\left(\mu_{n_j}^{[i_j]}\right)^{-(r_j+1)}, \tag{2.27}$$

where $\mathcal{A}_{r_t, n_{\bar{t}}}^{[i]}[f]$ is given by (2.26). Suppose further that f obeys the first k Neumann derivative conditions (2.12). Then the only non-zero term in (2.27) corresponds to $t = \emptyset$. In other words,

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

[9] In the context of Fourier series, such an expansion is occasionally referred to as the *Fourier Coefficient Asymptotic Expansion (FCAE)*, a terminology introduced by Lyness [117, 118, 119].

[10] The bivariate version of this expansion is relatively well known in the context of Fourier series. See, for example [17].

Proof. To prove (2.27) it suffices to consider $f \in \mathrm{C}^\infty(\bar{\Omega})$. To cover the general case we use density, linearity and the bound $|\mathcal{A}^{[i]}_{r_t,n_{\bar{t}}}[f]| \leq c\|f\|_{2k,\mathrm{mix}}$, $\forall f \in \mathrm{H}^{2k}_{\mathrm{mix}}(\Omega)$, for some positive constant c independent of f,$n_{\bar{t}}$, r_t and i (see Lemma 2.16).

We proceed by induction on d. Recalling (2.11) we confirm that the result is valid for $d = 1$. 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

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

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 (2.27).

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 (2.26), any term $\mathcal{A}^{[i]}_{r_t,n_{\bar{t}}}[f]$ with $t \neq \emptyset$ will vanish. This completes the proof. □

As mentioned, the expansion (2.27) has a number of uses. However, it is certainly not unique: provided f is sufficiently smooth, the values $\mathcal{A}^{[i]}_{r_t,n_{\bar{t}}}[f]$ can be re-expanded in inverse powers of $n_{\bar{t}}$. If $f \in \mathrm{C}^\infty(\bar{\Omega})$, this results in a formal asymptotic expansion involving only

inverse powers of $n_1, \dots, n_d$ and values of the function and its partial derivatives at the vertices (see also [95]):

$$\hat{f}_n^{[i]} \sim (-1)^{|n|+|i|} \sum_{r \in \mathbb{N}_0^d} \mathcal{B}_r^{[i]}[f] \prod_{j=1}^{d} \left(\mu_{n_j}^{[i_j]}\right)^{-(r_j+1)}, \quad n \gg 1. \tag{2.28}$$

Note, however, that this expansion is valid only in an asymptotic sense: in general, the right-hand side will not converge for fixed $n \in \mathbb{N}^d$.

Lemma 2.13 does not include those coefficients $\hat{f}_n^{[i]}$ where $n_j = 0$ for some $j = 1, \dots, d$. However, these can be easily handled. Given $n \in \mathbb{N}_0^d$, suppose that $n_t \equiv 0$ for some $t \in [d]$. If

$$f_t(x_{\bar{t}}) = \int_{-1}^{1} \dots \int_{-1}^{1} f(x) \, \mathrm{d}x_t, \tag{2.29}$$

then $\hat{f}_n^{[i]} = \widehat{f}_{t\,n_{\bar{t}}}^{\,[i_{\bar{t}}]}$. Moreover, if $f \in \mathrm{H}^{2k}_{\mathrm{mix}}(-1,1)^d$ then $f_t \in \mathrm{H}^{2k}_{\mathrm{mix}}(-1,1)^{|\bar{t}|}$. Hence, we may now apply Lemma 2.13 to $\widehat{f}_{t\,n_{\bar{t}}}^{\,[i_{\bar{t}}]}$ to give the asymptotic expansion in this case.

As an example, consider the univariate function $f(x) = x\mathrm{e}^x$. A simple calculation yields

$$\begin{aligned}
\hat{f}_0^{[0]} &= \frac{\sqrt{2}}{\mathrm{e}}, \\
\hat{f}_n^{[0]} &= \frac{2(-1)^n(1+\mathrm{e}^2 n^2\pi^2)}{\mathrm{e}(1+n^2\pi^2)^2} = \frac{(-1)^n}{(n\pi)^2}\left\{2\mathrm{e} - \frac{2(2\mathrm{e}-\mathrm{e}^{-1})}{(n\pi)^2}\right\} + \mathcal{O}\left(n^{-6}\right), \\
\hat{f}_n^{[1]} &= \frac{2(-1)^{n+1}(-1+\mathrm{e}^2(n-\frac{1}{2})^2\pi^2)}{\mathrm{e}(1+(n-\frac{1}{2})^2\pi^2)^2} = \frac{(-1)^{n+1}}{(n-\frac{1}{2})^2\pi^2}\left\{2\mathrm{e} - \frac{2(2\mathrm{e}+\mathrm{e}^{-1})}{(n-\frac{1}{2})^2\pi^2}\right\} + \mathcal{O}\left(n^{-6}\right),
\end{aligned} \tag{2.30}$$

from which we immediately deduce that $\hat{f}_n^{[i]} = \mathcal{O}\left(n^{-2}\right)$. Now suppose that we replace f by $g(x) = x\mathrm{e}^x - \mathrm{e}(x + \frac{1}{2}x^2)$. The polynomial term here acts to interpolate the derivatives $f'(\pm 1)$. Another calculation gives

$$\begin{aligned}
\hat{g}_0^{[0]} &= \frac{6-\mathrm{e}^2}{3\sqrt{2}\mathrm{e}}, \\
\hat{g}_n^{[0]} &= \frac{2(-1)^n\left(\mathrm{e}^2(2n^2\pi^2+1) - n^2\pi^2\right)}{\mathrm{e}n^2\pi^2(n^2\pi^2+1)^2} = \frac{2(2\mathrm{e}-\mathrm{e}^{-1})(-1)^{n+1}}{(n\pi)^4} + \mathcal{O}\left(n^{-6}\right), \\
\hat{g}_n^{[1]} &= \frac{2(-1)^n\left((n-\frac{1}{2})^2\pi^2 + \mathrm{e}^2(2(n-\frac{1}{2})^2\pi^2+1)\right)}{\mathrm{e}(n-\frac{1}{2})^2\pi^2\left((n-\frac{1}{2})^2\pi^2+1\right)^2} = \frac{2(2\mathrm{e}+\mathrm{e}^{-1})(-1)^n}{(n-\frac{1}{2})^4\pi^4} + \mathcal{O}\left(n^{-6}\right).
\end{aligned}$$

As predicted, the coefficients $\hat{g}_n^{[i]}$ are $\mathcal{O}\left(n^{-4}\right)$ for large n.

Now suppose that $f(x_1, x_2) = \mathrm{e}^{3(x_2-x_1)}$. The absolute values of the coefficients $\hat{f}_{n_1,n_2}^{[i_1,i_2]}$ are illustrated in Figure 2.2, which highlights the $\mathcal{O}\left(n_1^{-2}n_2^{-2}\right)$ decay. Note that a curve of fixed absolute value (i.e. a curve on which $n_1 n_2$ is constant), is referred to as a *hyperbolic cross* [13], an object we consider in greater detail in the sequel.

Suppose now that we replace f by $g = f - p_0$, where p_0 interpolates the Neumann data of f on the boundary. As predicted by Lemma 2.13 and verified in Figure 2.2, faster decay of the coefficients occurs. If $h = f - p_1$, where p_1 interpolates both the first and second order Neumann data of f on the boundary, then we witness even faster decay.

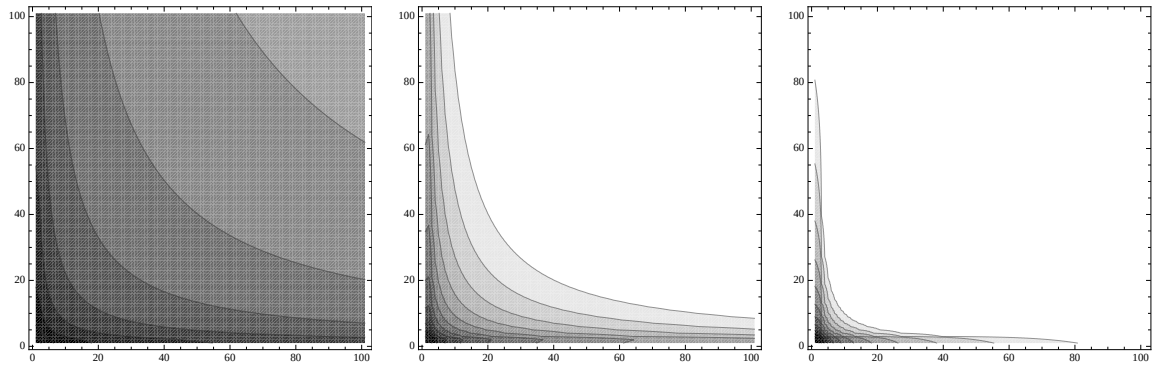


Figure 2.2: Contour plots of the coefficients $\hat{f}^{[0,0]}_{n_1,n_2}$, $\hat{g}^{[0,0]}_{n_1,n_2}$ and $\hat{h}^{[0,0]}_{n_1,n_2}$ (left to right) for $n_1, n_2 = 0, \dots, 100$ with contour lines at 10^{-j} for $j = -2, 1, 0, 1, \dots, 10$.

Similar expansions to (2.27) and (2.28) are obtained in exactly the same manner for the Laplace–Dirichlet coefficients $\check{f}^{[i]}_n$. If we re-define

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

and

$$\mathcal{A}^{[i]}_{r_t,n_{\bar{t}}}[f] = (-1)^{k|\bar{t}|}\prod_{j\notin t}\left(\mu^{[i_j]}_{n_j}\right)^{-k+\frac{1}{2}}\int \mathcal{B}^{[i_t]}_{r_t}[\mathrm{D}^{2k-1}_{\bar{t}} f](x_{\bar{t}})\phi^{[i_{\bar{t}}]}_{n_{\bar{t}}}(x_{\bar{t}})\,dx_{\bar{t}}, \tag{2.31}$$

then we have

Lemma 2.14. *Suppose that $f \in \mathrm{H}^{2k-1}_{mix}(\Omega)$, $k \in \mathbb{N}$, and that $n \in \mathbb{N}^d$. Then*

$$\check{f}^{[i]}_n = \sum_{t\in[d]^*}\sum_{|r_t|_\infty=0}^{k-1}\mathcal{A}^{[i]}_{r_t,n_{\bar{t}}}[f](-1)^{|n_t|}\prod_{j\in t}\left(\mu^{[i_j]}_{n_j}\right)^{-r_j-\frac{1}{2}}, \tag{2.32}$$

where $\mathcal{A}^{[i]}_{r_t,n_{\bar{t}}}[f]$ is given by (2.31). Suppose further that f obeys the first k Dirichlet derivative conditions (2.14). Then the only non-zero term in (2.32) corresponds to $t = \emptyset$. In particular,

$$\check{f}^{[i]}_n = \mathcal{A}^{[i]}_{k-1,n}[f] = (-1)^k\prod_{j=1}^{d}(\mu^{[i_j]}_{n_j})^{-k+\frac{1}{2}}\widehat{\mathrm{D}^{2k-1}} f^{[i]}_n.$$

By means of example, we consider the function $f(x) = x\mathrm{e}^x$ once more. In this case,

$$\begin{aligned}\check{f}^{[0]}_n &= \frac{2(-1)^{n+1}(n-\frac{1}{2})\pi\left((\mathrm{e}^2-1)(n-\frac{1}{2})^2\pi^2+(3+\mathrm{e}^2)\right)}{2\mathrm{e}\left((n-\frac{1}{2})^2\pi^2+1\right)} = \frac{2(-1)^{n+1}\sinh 1}{(n-\frac{1}{2})\pi}+\mathcal{O}\left(n^{-3}\right),\\ \check{f}^{[1]}_n &= \frac{(-1)^{n+1}n\pi\left((1+\mathrm{e}^2)n^2\pi^2+3-\mathrm{e}^2\right)}{\mathrm{e}(n^2\pi^2+1)^2} = \frac{2(-1)^n\cosh 1}{n\pi}+\mathcal{O}\left(n^{-3}\right).\end{aligned}$$

It follows immediately that $\check{f}^{[i]}_n = \mathcal{O}\left(n^{-1}\right)$. This, upon comparison with (2.30), demonstrates the slower decay of such coefficients in comparison to their Laplace–Neumann counterparts.

2.8 Bounds for coefficients

We now seek robust bounds for the coefficients $\hat{f}_n^{[i]}$ and $\check{f}_n^{[i]}$. To do so, it is first useful to define the alternative mixed Sobolev spaces $\mathrm{G}^r_{\mathrm{mix}}(\Omega) = \{f : \mathrm{D}^\alpha f \in \mathrm{L}^1(\Omega),\ \forall\ \alpha : |\alpha|_\infty \leq r\}$ with associated norm $\|\!|f|\!\|_{r,\mathrm{mix}} = \sum_{|\alpha|_\infty \leq r} \|\mathrm{D}^\alpha f\|_{\mathrm{L}^1(\Omega)}$, where $\|g\|_{\mathrm{L}^1(\Omega)} = \int_\Omega |g(x)|\,\mathrm{d}x$.

In the sequel we make use of the following imbedding result:

Lemma 2.15. *The spaces* $\mathrm{G}^r_{mix}(\Omega)$, $\mathrm{H}^r_{mix}(\Omega)$ *satisfy* $\mathrm{H}^r_{mix}(\Omega) \hookrightarrow \mathrm{G}^r_{mix}(\Omega)$ *with imbedding constant*[11] $c = (2r+2)^{\frac{d}{2}}$.[12]

Proof. The existence of an imbedding is a direct consequence of $\mathrm{L}^2(\Omega) \hookrightarrow \mathrm{L}^1(\Omega)$. Furthermore, by the Cauchy–Schwarz inequality,

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

Since there are $(r+1)^d$ choices of $\alpha \in \mathbb{N}_0^d$ with $|\alpha|_\infty \leq r$, we obtain the result. □

Once more, we focus on the modified Fourier coefficients $\hat{f}_n^{[i]}$. To derive a coefficient bound in this case we first require the following lemma:

Lemma 2.16. *Suppose that* $f \in \mathrm{H}^{2k}_{mix}(\Omega)$, $i \in \{0,1\}^d$, $t \in [d]^*$, $r_t \in \mathbb{N}_0^{|t|}$ *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 (2.26). Then*

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

Proof. If $\mathcal{B}^{[i_j]}_{r_j}[f]$ is as in (2.24) then $\mathcal{B}^{[i_j]}_{r_j}[f] = \int_{-1}^1 \partial_{x_j} \left(x_j^{i_j} \partial_{x_j}^{2r_j+1} f(x) \right) \mathrm{d}x_j$. Hence, the composition $\mathcal{B}^{[i_t]}_{r_t}[f]$ defined in (2.25) has integral representation

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

Substituting this into the expression (2.26) for $\mathcal{A}^{[i]}_{r_t,n_{\bar{t}}}[f]$ gives

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

We deduce that

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

Here the final inequality holds since the integral is a sum over derivatives $\mathrm{D}^\beta f$ with $|\beta|_\infty \leq 2k$ each multiplied by $x_1^{\gamma_1} \ldots x_d^{\gamma_d}$ for some suitable multi-index $|\gamma|_\infty \leq 1$. □

[11] By this we mean the constant $c > 0$ such that $\|\!|f|\!\|_{r,\mathrm{mix}} \leq c\|f\|_{r,\mathrm{mix}}$ for all $f \in \mathrm{H}^r_{\mathrm{mix}}(\Omega)$.

[12] This result is essentially independent of Ω (provided Ω is Lipschitz), in which case $c = |\Omega|^{\frac{1}{2}}(r+1)^{\frac{d}{2}}$.

Using this lemma we deduce

Theorem 2.17. *Suppose that $f \in \mathrm{H}^{2k+2}_{mix}(\Omega)$ obeys the first $k \in \mathbb{N}_0$ Neumann derivative conditions (2.12). Then*

$$\left|\hat{f}^{[i]}_n\right| \leq 2^{\chi(n)} \prod_{j:n_j>0} \left(\mu^{[i_j]}_{n_j}\right)^{-(k+1)} |\!|\!| f |\!|\!|_{2k+2,mix}, \quad n \in \mathbb{N}^d_0,$$

where $\chi(n)$, the grade of n, is the number of non-zero entries and the product is taken over those $j = 1, \ldots, d$ with corresponding value $n_j > 0$.

Proof. Suppose first that $n \in \mathbb{N}^d$. Then, using Lemma 2.13 (with k replaced by $k+1$) and the fact that f obeys the first k derivative conditions, we obtain

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

where $k_t = (k, k, \ldots, k) \in \mathbb{N}_0^{|t|}$. Using the bound for $\mathcal{A}^{[i]}_{k_t, n_{\bar{t}}}[f]$ from Lemma 2.16 we obtain

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

Since $|[d]^*| = 2^d$, this gives 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,

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

where f_t is defined in (2.29). Moreover,

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

and the final term is bounded above by $|\!|\!| f |\!|\!|_{2k+2,\mathrm{mix}}$. This completes the proof. □

Using Lemma 2.15 we may also derive a bound for $\hat{f}^{[i]}_n$ in terms of $\|f\|_{2k+2,\mathrm{mix}}$:

Corollary 2.18. *Suppose that $f \in \mathrm{H}^{2k+2}_{mix}(\Omega)$ obeys the first k Neumann derivative conditions (2.12). Then*

$$\left|\hat{f}^{[i]}_n\right| \leq 2^{\chi(n)+\frac{d}{2}} (2k+3)^{\frac{\chi(n)}{2}} \prod_{j:n_j>0} \left(\mu^{[i_j]}_{n_j}\right)^{-(k+1)} \|f\|_{2k+2,mix}, \quad n \in \mathbb{N}^d_0.$$

Proof. If $\chi(n) = d$ the result follows immediately from Theorem 2.17 and Lemma 2.15. Now suppose that $\chi(n) < d$. We have

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

Furthermore, $\|\!|f_t|\!\|_{2k+2,\mathrm{mix}} \leq (4k+6)^{\frac{\chi(n)}{2}} \|f_t\|_{2k+2,\mathrm{mix}}$ and it is simple to show that

$$\|\mathrm{D}^\beta f_t\| \leq 2^{\frac{d}{2}-\frac{\chi(n)}{2}} \|\mathrm{D}^\beta f\|, \quad \beta \in \mathbb{N}_0^{\chi(n)}.$$

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. □

In the sequel the following corollary will in fact be of greater use:

Corollary 2.19. *Suppose that $f \in \mathrm{H}^{2k+2}_{mix}(\Omega)$ obeys the first k Neumann derivative conditions (2.12). Then*

$$\left|\hat{f}_n^{[i]}\right| \leq 2^{\chi(n)+\frac{d}{2}} (2k+3)^{\frac{\chi(n)}{2}} (2^{|i|}\pi^{-\chi(n)})^{2(k+1)} (\bar{n}_1 \ldots \bar{n}_d)^{-2(k+1)} \|f\|_{2k+2,mix}, \quad n \in \mathbb{N}_0^d,$$

where $\bar{m} = \max\{m, 1\}$ for $m \in \mathbb{N}_0$.

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 2.18. □

In particular, Corollary 2.19 provides the aforementioned estimate $\hat{f}_n^{[i]} = \mathcal{O}(n_1^{-2} \ldots n_d^{-2})$ for an arbitrary function f. The Dirichlet case may be addressed in a similar manner. If $f \in \mathrm{H}^{2k+1}_{\mathrm{mix}}(\Omega)$ satisfies the first k Dirichlet derivative conditions (2.14), then

$$|\check{f}_n^{[i]}| \leq 2^{(2k+3)d}\pi^{-(2k+1)d}(k+1)^{\frac{d}{2}}(n_1 \ldots n_d)^{-2k-1}\|f\|_{2k+1,\mathrm{mix}}, \quad n \in \mathbb{N}^d, \quad i \in \{0,1\}^d.$$

When no derivative conditions are satisfied this figure is $\mathcal{O}\left(n_1^{-1} \ldots n_d^{-1}\right)$, hence verifying the slower decay of Laplace–Dirichlet coefficients.

2.9 Full index sets

Thus far we have made no stipulation as regards the index set I_N (aside from the mild conditions (2.8)). The choice of index set determines the computational cost of both forming and evaluating the approximation. As we shall see, the $\mathcal{O}(|I_N|)$ coefficients of the approximation $\mathcal{F}_N[f]$ can be constructed in $\mathcal{O}(|I_N|)$ operations using numerical quadrature (see Section 2.12). Moreover, such schemes are adaptive, making it possible to utilise any index set we choose.

Standard intuition leads to the *full index set*

$$I_N = \left\{n \in \mathbb{N}_0^d : |n|_\infty \leq N\right\}, \tag{2.33}$$

which is just the hypercube of length $N+1$ in $\mathbb{N}_0^d$. Indeed, the prevalence of this index set in applications is due to the fact that the method of choice for evaluating Fourier or Chebyshev coefficients, namely the Fast Fourier Transform (FFT) [37], computes all the coefficients in I_N in a non-adaptive manner.

For univariate expansions, (2.33) is the only obvious choice. However, numerous choices of index set are permissible in the multivariate setting, including the spherical index set

$$I_N = \left\{n \in \mathbb{N}_0^d : n_1^2 + \ldots n_d^2 \leq N\right\}, \tag{2.34}$$

and various polyhedral index sets. Many properties of univariate Fourier expansions are either untrue or unknown for multivariate expansions with coefficients from such index sets [58, 59].

Nonetheless, $|I_N| = \mathcal{O}\left(N^d\right)$ for both (2.33) and (2.34). This figure grows exponentially with dimension, making classical Fourier series unsuitable for higher dimensional problems. To alleviate this problem, we employ various hyperbolic cross index sets in the sequel. Such index sets are viable precisely because they do 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 Laplace eigenfunction expansions based on (2.33) in the remainder of this section.

Throughout the remainder of this chapter the bivariate case will serve as our primary example. If (2.33) is employed, then, for a bivariate function f, the truncated expansion $\mathcal{F}_N[f]$ is given by

$$\begin{aligned}\mathcal{F}_N[f](x_1,x_2) =& \frac{1}{2}\hat{f}^{[0,0]}_{0,0} + \frac{1}{\sqrt{2}}\sum_{n_1=1}^{N-1}\left\{\hat{f}^{[0,0]}_{n_1,0}\cos n_1\pi x_1 + \hat{f}^{[1,0]}_{n_1,0}\sin(n_1-\tfrac{1}{2})\pi x_1\right\}\\ &+\frac{1}{\sqrt{2}}\sum_{n_2=0}^{N-1}\left\{\hat{f}^{[0,0]}_{0,n_2}\cos n_2\pi x_2 + \hat{f}^{[0,1]}_{0,n_2}\sin(n_2-\tfrac{1}{2})\pi x_2\right\}\\ &+\sum_{n_1,n_2=1}^{N-1}\left\{\hat{f}^{[0,0]}_{n_1,n_2}\cos n_1\pi x_1\cos n_2\pi x_2 + \hat{f}^{[0,1]}_{n_1,n_2}\cos n_1\pi x_1\sin(n_2-\tfrac{1}{2})\pi x_2\right.\\ &\left.+\hat{f}^{[1,0]}_{n_1,n_2}\sin(n_1-\tfrac{1}{2})\pi x_1\cos n_2\pi x_2 + \hat{f}^{[1,1]}_{n_1,n_2}\sin(n_1-\tfrac{1}{2})\pi x_1\sin(n_2-\tfrac{1}{2})\pi x_2\right\}.\end{aligned}$$

2.9.1 Uniform and pointwise convergence rates

The modified Fourier expansion of a function $f \in \mathrm{H}^1_{\mathrm{mix}}(\Omega)$ converges uniformly on $\bar{\Omega}$. We now assess the rate of convergence:

Theorem 2.20. *Suppose that $f \in \mathrm{H}^{2k+2}_{mix}(\Omega)$ satisfies the first $k \in \mathbb{N}_0$ derivative conditions. Suppose further that I_N is the full index set (2.33) and $\mathcal{F}_N[f]$ is the truncated modified Fourier expansion of f. Then*

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

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 [1].

Proof. We have

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

Now $\sum_{n=1}^{N} n^{-2(k+1)} \leq \zeta(2(k+1))$ and $\sum_{n>N} n^{-2(k+1)} \leq \frac{1}{2k+1} N^{-(2k+1)}$. Hence

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

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

Though not immediately obvious, this result is quasi-optimal, in the sense that the uniform error is not $o(N^{-2k-1})$ for an arbitrary function f. We demonstrate this fact in the sequel.

For a general function f, Theorem 2.20 verifies the aforementioned $\mathcal{O}\left(N^{-1}\right)$ uniform convergence rate of $\mathcal{F}_N[f]$. It turns out that this approximation converges at a rate one power of N faster inside Ω than on the boundary Γ. Hence, $\mathcal{O}\left(N^{-2}\right)$ for an arbitrary function obeying no derivative conditions and $\mathcal{O}\left(N^{-2k-2}\right)$ when the first k conditions are satisfied. In fact, we may also determine the exact leading order asymptotic behaviour of the error $f(x) - \mathcal{F}_N[f](x)$ at any point $x \in \bar{\Omega}$. The univariate case of this result was demonstrated by S. Olver [134]; for the multivariate extension we generalise the technique used therein.

To do so, we first introduce the *Lerch transcendental function* [152], given by

$$\Phi(z,s,a) = \sum_{n=0}^{\infty} \frac{z^n}{(n+a)^s}, \quad \mathrm{Re}\,(a) > 0, \quad \mathrm{Re}\,(s) > 1, \quad |z| \leq 1. \tag{2.35}$$

With this to hand, we have

Lemma 2.21. *Suppose that $f \in \mathrm{H}^{2k+3+l}_{mix}(\Omega)$, $l = 0, 1$, obeys the first $k \in \mathbb{N}_0$ Neumann derivative conditions. Suppose further that I_N is the full index set (2.33) and $\mathcal{F}_N[f]$ is the truncated modified Fourier expansion of f. Then*

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

where $\bar{j} \in [d]$ is the tuple $(1, \ldots, j-1, j+1, \ldots, d)$,

$$\tilde{\Phi}^{[i]}(N, k+1, x) = \mathrm{Re}\left[(-\mathrm{e}^{\mathrm{i}\pi x})^{N+1-\frac{1}{2}i} \pi^{-2(k+1)} \Phi(-\mathrm{e}^{\mathrm{i}\pi x}, 2k+2, N+1-\tfrac{1}{2}i)\right],$$

and i *is the imaginary unit.*

Proof. Since uniform convergence is guaranteed by Theorem 2.12, we may write

$$f(x) - \mathcal{F}_N[f](x) = \sum_{t\in[d]} \sum_{i\in\{0,1\}^d} \sum_{\substack{n_j > N \\ j\in t}} \sum_{|n_{\bar{t}}|_\infty \leq N} \hat{f}_n^{[i]} \phi_n^{[i]}(x).$$

Because $\hat{f}_n^{[i]} = \mathcal{O}\left(n^{-2k-2}\right)$ by Theorem 2.17, where $n^{-2k-2} = (n_1 \ldots n_d)^{-2k-2}$, the largest contribution occurs when $|t| = 1$. Hence

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

We now expand $\hat{f}_n^{[i]}$ in powers of n_j as in Lemma 2.13. For each j, we obtain

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

Substituting this into the previous expression gives

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

Now

$$\begin{aligned} \sum_{n>N} \frac{(-1)^n}{(\mu_n^{[0]})^{k+1}} \phi_n^{[0]}(x) &= \mathrm{Re}\left[(-\mathrm{e}^{\mathrm{i}\pi x})^{N+1} \pi^{-2(k+1)} \sum_{m=0}^{\infty} \frac{(-1)^m}{(N+1+m)^{2(k+1)}} (-\mathrm{e}^{\mathrm{i}\pi x})^m\right] \\ &= \mathrm{Re}\left[(-\mathrm{e}^{\mathrm{i}\pi x})^{N+1} \pi^{-2(k+1)} \Phi(-\mathrm{e}^{\mathrm{i}\pi x}, 2k+2, N+1)\right] = \tilde{\Phi}^{[0]}(N, k+1, x). \end{aligned}$$

In an identical manner, we can also show that

$$\sum_{n>N} \frac{(-1)^{n+1}}{(\mu_n^{[1]})^{k+1}} \phi_n^{[1]}(x) = \tilde{\Phi}^{[1]}(N, k+1, x).$$

Substituting these results into the previous formula now completes the proof. □

We are now able to determine the leading order asymptotic behaviour of $f(x) - \mathcal{F}_N[f](x)$. This follows immediately from the observation

$$\Phi(-\mathrm{e}^{\mathrm{i}\pi x}, s, a) = \frac{a^{-s}}{1+\mathrm{e}^{\mathrm{i}\pi x}} + \mathcal{O}\left(a^{-(s+1)}\right), \quad -1 < x < 1, \quad a \to \infty.^{13} \tag{2.36}$$

We deduce

Theorem 2.22. *Suppose that f, k, I_N and $\mathcal{F}_N[f]$ are as in Lemma 2.21. Then, for $x \in \Omega$,*

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

In particular, $f(x) - \mathcal{F}_N[f](x) = \mathcal{O}\left(N^{-2k-2}\right)$ uniformly for x in compact subsets of Ω.

The main result of this theorem, faster convergence away from the boundary, is demonstrated in Figure 2.3. The error at the endpoints is approximately 10^{-2}, whereas in the subinterval $[-\frac{1}{2}, \frac{1}{2}]$ this value is much smaller, roughly 10^{-4}. In Figure 2.4 we consider the bivariate case. Once more we observe that the error is much smaller away from the boundary. This figure also highlights that the convergence rate is slower on the whole of the boundary, not just the corners, as may be expected.

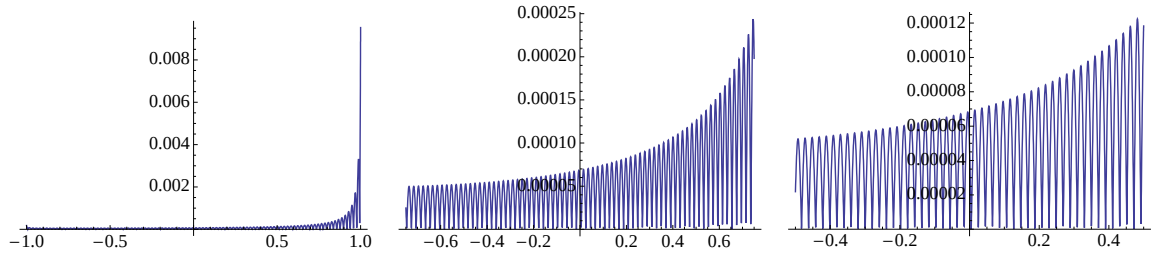


Figure 2.3: Graph of $|f(x) - \mathcal{F}_{50}[f](x)|$ for $-1 \leq x \leq 1$ (left), $-\frac{3}{4} \leq x \leq \frac{3}{4}$ (middle) and $-\frac{1}{2} \leq x \leq \frac{1}{2}$ (right), where $f(x) = \mathrm{Ai}(-3x - 4)$ and Ai is the Airy function [1].

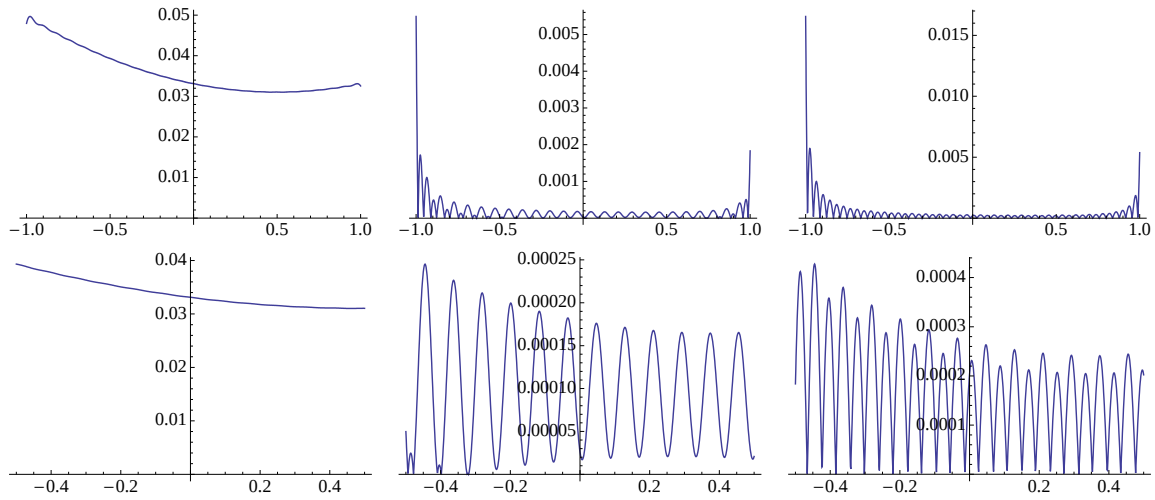


Figure 2.4: Absolute error $|f(x, y_0) - \mathcal{F}_{50}[f](x, y_0)|$, where $f(x_1, x_2) = (x_1^2 - x_2 + 4)\cos 2x_2 \sin 3x_2$, for $-1 \leq x \leq 1$ (top row) and $-\frac{1}{2} \leq x \leq \frac{1}{2}$ (bottom row), where $y_0 = 1, \frac{2}{3}, \frac{1}{3}$ (left to right).

As established in [134], the condition $f \in \mathrm{H}^{2k+3}(-1, 1)$ (in the univariate case) can be replaced with the slightly weaker conditions that $f \in \mathrm{C}^{2k+2}[-1, 1]$ and $f^{(2k+2)}$ has bounded variation. However, since our primary focus is on boundary value problems, we shall continue to present conditions in specific Sobolev spaces.

The two results of this section, Theorems 2.20 and 2.22, can be readily generalised to provide estimates for the error $\mathrm{D}^{\beta}(f - \mathcal{F}_N[f])$, where $|\beta|_\infty \leq 2k$. The corresponding pointwise and uniform convergence rates are $\mathcal{O}\left(N^{|\beta|_\infty - 2k - 1}\right)$ and $\mathcal{O}\left(N^{|\beta|_\infty - 2k - 2}\right)$ respectively. We may also provide analogous versions of these theorems for expansions based on Laplace–Dirichlet eigenfunctions. In this case the respective convergence rates are one power of N slower.

As mentioned, Lemma 2.21 may be used to deduce quasi-optimality of the uniform error estimate given in Theorem 2.20. As described in [134], for $x = \pm 1$ the Lerch function $\Phi(-\mathrm{e}^{i\pi x}, s, a)$ reduces to the Hurwitz zeta function $\zeta(s, a)$ [1], from which the estimate $\Phi(1, s, a) = \mathcal{O}\left(a^{-1}\right)$ is easily deduced. This also verifies the previously made observation that the convergence rate is $\mathcal{O}\left(N^{-1}\right)$ on the whole of the boundary.

[13] All the terms in this asymptotic expansion can in fact be prescribed (see, for example, [134]).

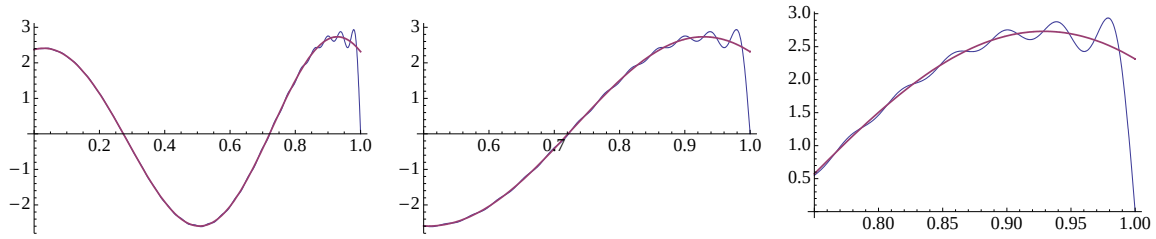


Figure 2.5: Graphs of $f'(x)$ and $(\mathcal{F}_{50}[f])'(x)$ for $0 \leq x \leq 1$ (left), $\frac{1}{2} \leq x \leq 1$ (middle) and $\frac{3}{4} \leq x \leq 1$ (right), where $f(x) = \mathrm{Ai}(-3x - 4)$.

Modified Fourier expansions have no apparent Gibbs phenomenon. Nonetheless, for an arbitrary function f, a weak Gibbs phenomenon occurs in the first partial derivative. The error $\mathrm{D}^\beta f(x) - \mathrm{D}^\beta \mathcal{F}_N[f](x)$, where $|\beta|_\infty = 1$, converges pointwise away from the boundary but not uniformly on $\bar{\Omega}$. This is verified in Figure 2.5, where $\mathcal{O}(1)$ Gibbs oscillations are observed near the endpoint $x = 1$. Since $\mathrm{D}^\beta \mathcal{F}_N[f]$ is nothing more than the Laplace–Dirichlet expansion of $\mathrm{D}^\beta f$, this effect is equivalently stated as the presence of the Gibbs phenomenon in Laplace–Dirichlet expansions.

Non-uniform convergence of Laplace–Dirichlet expansions is easily exhibited by example (e.g. $f(x) = 1$). Much like the Fourier setting, a proof of pointwise convergence away from the boundary can be obtained by developing Féjer and de la Vallée Poussin theorems for this basis (see [107] for details of the Fourier case and [94] for the extension of such results to this basis). However, a simple argument using Lerch functions is also easily provided. The advantage of this approach, as we demonstrate in Chapter 3, is that it can be applied to expansions where the above results are not readily available.

Lemma 2.23. *Suppose that $f \in \mathrm{H}^1_{mix}(\Omega)$ and that $\mathcal{F}_N[f]$ is the Laplace–Dirichlet expansion of f. Then $\mathcal{F}_N[f](x)$ converges to $f(x)$ uniformly in compact subsets of Ω.*

Proof. First suppose that $f \in \mathrm{C}^\infty(\bar{\Omega})$. It is easily demonstrated using the method of Lemma 2.21 and Theorem 2.22 that $\mathcal{F}_N[f](x)$ converges uniformly in a compact subset Ω' of Ω to a function $\tilde{f} \in \mathrm{C}(\Omega')$. Suppose that $f(x_0) \neq \tilde{f}(x_0)$ for some $x_0 \in \Omega'$. Then $f(x) \neq \tilde{f}(x)$ in the closure of some neighbourhood U of x_0. Hence

$$0 < \int_U |f(x) - \tilde{f}(x)|^2 \,\mathrm{d}x \leq \|f - \tilde{f}\|^2 = \lim_{N\to\infty} \|f - \mathcal{F}_N[f]\|^2 = 0,$$

giving a contradiction. Thus $\tilde{f}(x) = f(x)$ for $x \in \Omega'$. Now suppose that $f \in \mathrm{H}^1_{\mathrm{mix}}(\Omega)$. Given $\epsilon > 0$ there exists a function $g \in \mathrm{C}^\infty(\bar{\Omega})$ with $\|f - g\|_{1,\mathrm{mix}} < \epsilon$. Hence

$$|f(x) - \mathcal{F}_N[f](x)| \leq \|f - g\|_\infty + \|\mathcal{F}_N[f - g]\|_\infty + |g(x) - \mathcal{F}_N[g](x)|.$$

Using Lemma 2.6 and Bessel's inequality, we obtain $|f(x) - \mathcal{F}_N[f](x)| < 2c\epsilon + |g(x) - \mathcal{F}_N[g](x)|$. The full result now follows immediately from the previous arguments. □

2.9.2 Convergence rates in other norms

We now turn our attention to providing error estimates for $f - \mathcal{F}_N[f]$ in various Sobolev norms. We commence with the following lemma:

Lemma 2.24. *Suppose that* $f \in \mathrm{H}^{2k+l}(\Omega)$, $l = 0, 1$ *satisfies the first* $k \in \mathbb{N}_0$ *Neumann derivative conditions (2.12). Then*

$$\|f - \mathcal{F}_N[f]\|_s \leq c_{r,s} N^{s-r} |f|_r, \quad r = s, \ldots, 2k+l, \quad s = 0, \ldots, 2k+l, \tag{2.37}$$

for some positive constant $c_{r,s}$ *depending only on* r *and* s.

Proof. An application of (2.22) gives

$$\begin{aligned}
\|f - \mathcal{F}_N[f]\|_s^2 &\leq c_2 \sum_{i \in \{0,1\}^d} \sum_{n \notin I_N} (1 + \mu_n^{[i]})^s |\hat{f}_n^{[i]}|^2 \\
&\leq 2^s c_2 \max_{\substack{n \notin I_N \\ i \in \{0,1\}^d}} \{(\mu_n^{[i]})^{s-r}\} \sum_{i \in \{0,1\}^d} \sum_{n \notin I_N} (\mu_n^{[i]})^r |\hat{f}_n^{[i]}|^2 \\
&\leq 2^s c_2 \max_{\substack{n \notin I_N \\ i \in \{0,1\}^d}} \{(\mu_n^{[i]})^{s-r}\} |f|_r^2.
\end{aligned} \tag{2.38}$$

For $n \notin I_N$, $\mu_n^{[i]} \geq (N\pi)^2$. Using this and the previous expression we obtain the result. □

Lemma 2.24 is an example of a standard type of estimate for approximations in Fourier-like bases [42]. However, its conclusion may lead to the assertion that, for smooth f satisfying the first k Neumann derivative conditions, $\|f - \mathcal{F}_N[f]\|_{2k+1} = \mathcal{O}(1)$, an estimate which, in view of Lemma 2.9, is not optimal.[14] 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 2.8.

Lemma 2.25. *Suppose that* $f \in \mathrm{H}^{2k+2}_{mix}(\Omega)$ *satisfies the first* $k \in \mathbb{N}_0$ *Neumann derivative conditions (2.12). Then, for* $s = 0, \ldots, 2k+1$, *we have*

$$\|f - \mathcal{F}_N[f]\|_s \leq c_{s,k} N^{s-2k-\frac{3}{2}} \|f\|_{2k+2,mix}, \tag{2.39}$$

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

Proof. Using Lemma 2.9 we have

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

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

$$\begin{aligned}
\|f - \mathcal{F}_N[f]\|_s^2 &\leq c_s \sum_{|\beta| \leq s} \sum_{t \in [d]} \sum_{\substack{n_j > N \\ j \in t}} \sum_{|n_{\bar{t}}|_\infty \leq N} \prod_{j=1}^d n^{2\beta_j - 4k - 4} \\
&\leq c_s \sum_{|\beta| \leq s} \sum_{t \in [d]} N^{|\beta| - (4k+3)|t|} \leq c_{s,k} N^{s-(4k+3)},
\end{aligned}$$

as required. □

[14] This may be explained as follows: there is no characterisation of the norms $\|f\|_{2k+2}$, $\|f\|_{2k+2,\mathrm{mix}}$ in terms of modified Fourier coefficients. Equivalently, the periodic extension of f (as in Section 2.3) has only $\mathrm{H}^{2k+1}_{\mathrm{mix}}(\Omega)$-regularity.

As in Theorem 2.20, it is possible to prescribe values for the constants appearing in Lemmas 2.24 and 2.25. However, we shall not do this: numerical results indicate that such constants are not excessively large.

2.10 Hyperbolic cross approximations

The modified Fourier approximation $\mathcal{F}_N[f]$ based on (2.33) satisfies $\|f - \mathcal{F}_N[f]\|_\infty \leq cM^{-\frac{1}{d}}$, where $M = N^d$. When the total number of terms M is fixed, this figure deteriorates exponentially with dimension (equivalently $|I_N| = \mathcal{O}\left(N^d\right)$ grows exponentially with dimension). This observation is commonly referred to as *the curse of dimensionality*, a phrase attributed to Bellman [22]. Such behaviour is typical of orthogonal expansions based on (2.33) [41].

Since Bellman's observation, significant advances have been made in *breaking* the curse of dimensionality. To achieve this, we first assume that the function f possesses mixed Sobolev regularity. We next generate index sets by discarding any term in the approximation whose absolute value is below a certain tolerance (using, for example, the bounds derived in Section 2.8). This leads to so-called *hyperbolic cross* index sets.[15]

We consider two types of hyperbolic cross index sets. The first, our starting point, ameliorates this exponential growth to just a logarithmic factor: $|I_N| = \mathcal{O}\left(N(\log N)^{d-1}\right)$. However, with the introduction of so-called *optimized hyperbolic cross* index sets [75], we are able to completely overcome the curse of dimensionality. One caveat is required: the various constants involved exhibit exponential growth with d, thus limiting such an approach to only moderate numbers of dimensions. This topic is discussed in greater detail in [41].[16]

Once more we shall not prescribe exact values to the various constants appearing in error estimates. With a little effort, and the use of the coefficient bounds of Section 2.8, this can be achieved.

2.10.1 Construction of hyperbolic cross index sets

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

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

whose absolute value in some norm is greater than a tolerance ϵ.[17] To do so, we first require appropriate bounds for the coefficients $\hat{f}_n^{[i]}$ and the functions $\phi_n^{[i]}$. Given a norm $\|\!|\!| \cdot |\!|\!|$, the coefficient bounds of Section 2.8 yield $\|\!|\!| \hat{f}_n^{[i]} \phi_n^{[i]} |\!|\!| \leq c\|f\|_{2,\mathrm{mix}} \bar{n}^{-2} \|\!|\!| \phi_n^{[i]} |\!|\!|$. Next, we define the tolerance ϵ as precisely this upper bound with $n = (N, 0, \ldots, 0)$. In other words

[15] This process is somewhat different to the *sparse grids* approach for (typically) finite element discretizations [41]. However, the end result, the breaking of the curse of dimensionality, is the same. Sparse grids are discussed further in Section 2.10.3.

[16] The state-of-the-art finite element methods described therein can tackle 18 dimensional problems. Nonetheless, the particular structure of high-dimensional functions (specifically, the *concentration of measure phenomenon* [21]), offers a potential route to address such problems. In this sense the curse of *high* dimension is somewhat of a misnomer, the *curse of moderate dimension* being perhaps more apt a phrase.

[17] Due to their faster convergence rate over expansions based on Laplace–Dirichlet eigenfunctions, we consider only modified Fourier expansions throughout. Simple, standard adjustments can be made for the latter.

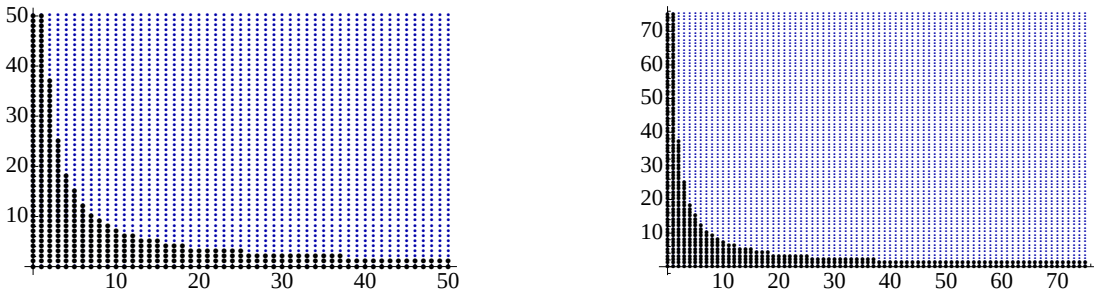


Figure 2.6: Graphs of the index sets (2.33) (small dots) and (2.41) (larger dots) for $N = 50$ (left diagram) and $N = 75$ (right diagram).

$\epsilon = c\|f\|_{2,\mathrm{mix}}|||\phi^{[i]}_{(N,0,\ldots,0)}|||N^{-2}$. Proceeding in this manner, including only those n for which this bound exceeds ϵ, we obtain a *hyperbolic cross* [13, 158] index set:

$$I_N = \{n \in \mathbb{N}^d : \bar{n}^2|||\phi^{[i]}_n|||^{-1} \leq N^2|||\phi^{[i]}_{(N,0,\ldots,0)}|||^{-1}\}. \tag{2.40}$$

This section is devoted to the study of such index sets for several choices of the norm $|||\cdot|||$.

2.10.2 The $\mathrm{L}^2(\Omega)$ norm hyperbolic cross index set

Our first consideration is the index set that originates from the $\mathrm{L}^2(\Omega)$ and uniform norms. In this case $\|\phi^{[i]}_n\|_\infty = \|\phi^{[i]}_n\| = 1$. It follows that $|||\hat{f}^{[i]}_n\phi^{[i]}_n||| \leq c\|f\|_{2,\mathrm{mix}}\bar{n}^{-2}$ and, therefore,

$$I_N = \{n \in \mathbb{N}^d : |n|_0 \leq N\}, \tag{2.41}$$

where $|n|_0 = \bar{n}_1 \ldots \bar{n}_d$, $n \in \mathbb{N}^d_0$.[18] Typical forms of this index set are shown in Figure 2.6.

In the remainder of this section we detail the benefit of this index set. There are two aspects to this study: 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 compared to approximations based on the full index set (2.33). We commence with the former:

Lemma 2.26. *Suppose that $\theta_d(t)$ is the number of terms $n = (n_1, \ldots, n_d) \in \mathbb{N}^d_0$ such that $|n|_0 \leq t$. 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 [48]. A simple inductive argument appears in [87], which we now repeat here, since similar methods will be used in the sequel:

[18] Though $|\cdot|_0$ is not a norm on $\mathbb{N}^d_0$ we shall use this notation.

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

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

Evaluation of this integral completes the proof. □

Corollary 2.27. *The number of terms in the expansion $\mathcal{F}_N[f]$ based on the hyperbolic cross (2.41) is*

$$\frac{2^d}{(d-1)!} N(\log N)^{d-1} + \mathcal{O}\left(N(\log N)^{d-2}\right). \tag{2.42}$$

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

We mention in passing that an upper bound for the number of terms in I_N is also readily established [87]. However, for our purposes, (2.42) will suffice.

We now consider the approximation error $f - \mathcal{F}_N[f]$, where $\mathcal{F}_N[f]$ is based on the hyperbolic cross index set (2.41). As in the case of the full index set, there are two components to this study: estimates based on the characterisations given in Lemmas 2.9 and 2.10 and estimates using the coefficient bounds of Section 2.8. We commence with the former:

Lemma 2.28. *Suppose that $f \in \mathrm{H}^{2k+l}(\Omega)$, $l = 0, 1$, satisfies the first $k \in \mathbb{N}_0$ Neumann derivative conditions (2.12) and that I_N is the hyperbolic cross index set (2.41). Then, for some positive constant $c_{r,s}$ independent of f and N,*

$$\|f - \mathcal{F}_N[f]\|_s \le c_{r,s} N^{\frac{s-r}{d}} |f|_r, \quad r = s, \ldots, 2k+l, \quad s = 0, \ldots, 2k+l. \tag{2.43}$$

If, additionally, $f \in \mathrm{H}^{2k+l}_{mix}(\Omega)$, then, for $s = 0, \ldots, 2k+l$,

$$\|f - \mathcal{F}_N[f]\|_s \le c_{r,s} N^{s-r} |f|_{r,mix}, \quad r = s, \ldots, 2k+l. \tag{2.44}$$

Proof. By a standard inequality $1 + \mu_n^{[i]} \ge c\bar{n}^{\frac{2}{d}}$, and for $n \notin I_N$ we have $1 + \mu_n^{[i]} \ge N^{\frac{2}{d}}$. Hence, using (2.38), we obtain

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

which gives (2.43). Next we consider (2.44). Clearly $\|f - \mathcal{F}_N[f]\|_s \le \|f - \mathcal{F}_N[f]\|_{s,\mathrm{mix}}$. Furthermore

$$\begin{aligned}\|f - \mathcal{F}_N[f]\|_{s,\mathrm{mix}}^2 &\le \sum_{i\in\{0,1\}^d} \sum_{n\notin I_N} |\hat{f}_n^{[i]}|^2 \prod_{j=1}^d (1 + \mu_{n_j}^{[i_j]})^s\\ &\le c_{r,s} N^{2(s-r)} \sum_{i\in\{0,1\}^d} \sum_{n\in\mathbb{N}_0^d} |\hat{f}_n^{[i]}|^2 \prod_{j=1}^d (\mu_{n_j}^{[i_j]})^r \le c_{r,s} N^{2(s-r)} |f|_{r,\mathrm{mix}}^2,\end{aligned}$$

which yields (2.44). □

We now provide estimates using the coefficient bounds of Section 2.8:

Theorem 2.29. *Suppose that $f \in \mathrm{H}^{2k+2}_{mix}(\Omega)$ obeys the first $k \in \mathbb{N}_0$ Neumann derivative conditions and I_N is the hyperbolic cross index set (2.41). Then, for $s = 1, \dots, 2k+1$,*

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

where the constants $c_k, c_{k,s} > 0$ are independent of f and N.

To prove this theorem we require the following lemma:

Lemma 2.30. *Suppose that $\gamma_{r,d}(t) = \sum_{\bar{n}>t} \bar{n}^{-r-1}$ and $r > 0$. Then*

$$\gamma_{r,d}(t) = \frac{t^{-r}(\log t)^{d-1}}{r(d-1)!} + \mathcal{O}\left(t^{-r}(\log t)^{d-2}\right), \quad t \gg 1. \tag{2.45}$$

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

$$\delta_{r,s,d}(t) = \frac{1}{r-s} \left\{1 + \zeta(s+1)\right\}^{d-1} t^{s-r} + \begin{cases} \mathcal{O}\left(t^{-r}(\log t)^{d-1}\right) & 0 < s \leq 1 \\ \mathcal{O}\left(t^{s-r-1}\right) & s > 1. \end{cases} \tag{2.46}$$

Proof. We use 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, as required. Now assume that the result is true up to d. Then

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

where θ_d is as in Lemma 2.26. Thus we obtain (2.45). Next we consider $\delta_{r,s,d}(t)$:

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

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 2.29. This follows immediately from Corollary 2.19 and Lemma 2.30. □

Theorem 2.29 reveals that the convergence rate of the modified Fourier expansion based on the hyperbolic cross (2.41) is comparable to that of the approximation based on the full index set (2.33). Indeed, for the $\mathrm{L}^2(\Omega)$ 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 $\mathrm{H}^s(\Omega)$ rate, $s \geq 1$, remains the same. In an identical manner, we can also show that the uniform error of the derivative $\mathrm{D}^\beta(f - \mathcal{F}_N[f])$ is $\mathcal{O}\left(N^{|\beta|_\infty-2k-1}(\log N)^{d-1}\right)$ for $|\beta|_\infty \leq 2k$. Once more, this is comparable to the estimate for the full index set approximation.

As is necessary for hyperbolic cross approximations, additional (mixed) smoothness is required for the estimates of Lemma 2.28 in comparison to those of Lemma 2.24. If only $\mathrm{H}^r(\Omega)$ regularity is imposed, the hyperbolic cross approximation will converge more slowly than its counterpart based on the full index set (or at a comparable rate if the number of terms $M = |I_N|$ is fixed). 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 2.25 and Theorem 2.29 respectively).

It is also of interest to examine the effect of the hyperbolic cross on the pointwise convergence rate. As we shall see in the sequel, this also only degrades by a factor of $\mathcal{O}\left((\log N)^{d-1}\right)$. Moreover, the smoothness requirement remains the same. To investigate this, we need to introduce a related concept, the so-called *step hyperbolic cross*.

2.10.3 Step hyperbolic cross index sets

Step hyperbolic cross index sets are closely related to the sparse grid technique [41]. The idea is to construct the approximation $\mathcal{F}_N[f]$ from hierarchical bases or subspaces. To this end, we define hypercubes $\rho(\beta) \subseteq \mathbb{N}_0^d$ by

$$\rho(\beta) = \{n \in \mathbb{N}_0^d : \lfloor 2^{\beta_j - 1} \rfloor \leq n_j < 2^{\beta_j}, \quad j = 1, \ldots, d\}, \quad \beta \in \mathbb{N}_0^d,$$

and corresponding basis elements

$$\mathcal{F}_\beta[f](x) = \sum_{i\in\{0,1\}^d}\sum_{n\in\rho(\beta)} \hat{f}_n^{[i]}\phi_n^{[i]}(x), \quad x \in \bar{\Omega}, \quad \beta \in \mathbb{N}_0^d. \tag{2.47}$$

Clearly, $\cup_{\beta\in\mathbb{N}_0^d}\rho(\beta) = \mathbb{N}_0^d$. Hence, for $f \in \mathrm{L}^2(\Omega)$, we may write $f = \sum_{\beta\in\mathbb{N}_0^d}\mathcal{F}_\beta[f]$, with identification in the $\mathrm{L}^2(\Omega)$ sense. Suppose now that $N = 2^m$. We seek a new approximation $\mathcal{F}_N[f]$ based on this decomposition. To this end, we introduce the finite set $W_m \subseteq \mathbb{N}_0^d$ and define

$$\mathcal{F}_N[f](x) = \sum_{\beta\in W_m}\mathcal{F}_\beta[f](x) = \sum_{i\in\{0,1\}^d}\sum_{n\in Q_m}\hat{f}_n^{[i]}\phi_n^{[i]}(x), \tag{2.48}$$

where $Q_m = \cup_{\beta\in W_m}\rho(\beta)$. Note that the approximation $\mathcal{F}_N[f]$ based on the full index set (2.33) is readily recovered by setting $W_m = \{\beta \in \mathbb{N}_0^d : |\beta|_\infty \leq m\}$.

To reduce the number of approximation terms, we now wish to specify W_m by including only those basis elements $\mathcal{F}_\beta[f]$ that have significant contribution to $\mathcal{F}_N[f]$. To do so, we follow the standard approach of [41]. Suppose that $c(\beta)$, the *local cost function*, is proportional to the cost of forming $\mathcal{F}_\beta[f]$. In other words, $c(\beta) = |\rho(\beta)|$. Suppose further that $b(\beta)$, the *local benefit function*, is proportional to an upper bound for $\|\mathcal{F}_\beta[f]\|^2$, where $\|\cdot\|$ is some arbitrary norm. If we introduce the *cost benefit ratio* $cbr(\beta) = c(\beta)b(\beta)^{-1}$, then W_m is given by $\{\beta \in \mathbb{N}_0^d : cbr(\beta) \leq cbr(m, 0, \ldots, 0)\}$. This set is referred to as a *sparse grid* index set [41].

In the context of Fourier series, sparse grids are usually used as a computational tool [110]. Indeed, as we discuss further later, a version of FFT, the so-called *Sparse Grid Fast Fourier Transform (SGFFT)*, can be designed (with considerable effort) for use with such index sets [18, 60]. Somewhat conversely, however, we shall use the sparse grid framework to answer analytical questions regarding hyperbolic cross index sets, namely, the rate of pointwise convergence.

We now return to explicit construction of W_m. First notice that $|\rho(\beta)| = 2^{|\beta|-\chi(\beta)} \leq 2^{|\beta|}$, where $\chi(\beta)$ is the grade of β. Turning our attention to $\mathcal{F}_\beta[f]$, suppose that we consider the uniform norm $\|\cdot\|_\infty$. Using Corollary 2.19, we have

$$\begin{aligned}\|\mathcal{F}_\beta[f]\|_\infty &\leq \sum_{i\in\{0,1\}^d}\sum_{n\in\rho(\beta)}|\hat{f}_n^{[i]}| \leq c\|f\|_{2k+2,\mathrm{mix}}\sum_{n\in\rho(\beta)}\bar{n}^{-2k-2}\\ &= c\|f\|_{2k+2,\mathrm{mix}}\prod_{j=1}^{d}\sum_{n_j=\lfloor 2^{\beta_j-1}\rfloor}^{2^{\beta_j}-1}\bar{n}_j^{-2k-2} \leq c\|f\|_{2k+2,\mathrm{mix}}2^{-(2k+1)|\beta|},\end{aligned}$$

for some constant c independent of f and β. If we now define $c(\beta) = 2^{|\beta|}$ and $b(\beta) = 2^{-(4k+2)|\beta|}$, then $cbr(\beta) = 2^{(4k+3)|\beta|}$ and we obtain

$$W_m = \{\beta \in \mathbb{N}_0^d : |\beta| \leq m\}.^{19} \tag{2.49}$$

As expected, the resultant *step hyperbolic cross* index set $Q_m = \cup_{\beta\in W_m}\rho(\beta)$ is closely related to the hyperbolic cross index set (2.41). The following result is well known (see, for example [110]):

Lemma 2.31. *Suppose that $N = 2^m$, I_N is the hyperbolic cross index set (2.41) and $Q_m = \cup_{\beta\in W_m}\rho(\beta)$ is the step hyperbolic cross index set, where W_m is given by (2.49). Then*

$$Q_m \subseteq I_N \subseteq Q_{m+d}. \tag{2.50}$$

[19] As with the hyperbolic cross (2.41), the same set is obtained upon consideration of the $\mathrm{L}^2(\Omega)$ norm.

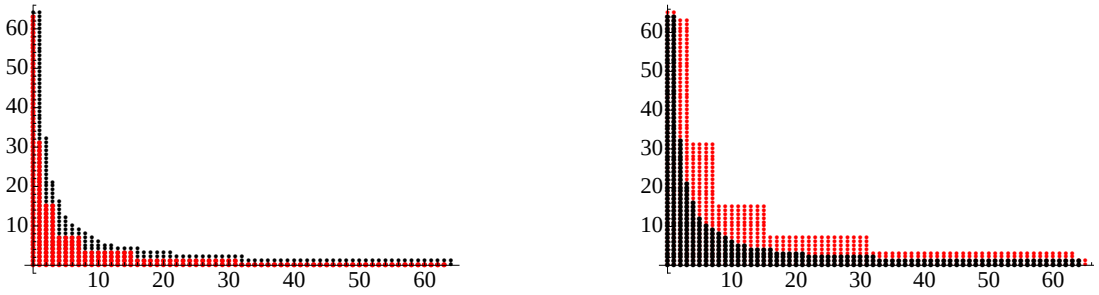


Figure 2.7: Graphs of the hyperbolic cross index set I_{64} (2.41) and step hyperbolic cross index sets Q_6 (left) and Q_8 (right) based on (2.49).

Proof. Suppose that $n \in Q_m$. Then $n_j < 2^{\beta_j}$ for $j = 1, \ldots, d$ and some $|\beta| \leq m$. Hence $|n|_0 = \bar{n}_1 \ldots \bar{n}_d < 2^{|\beta|} \leq 2^m = N$, so $n \in I_N$ as required. Now suppose that $n \in I_N$. Then $\lfloor 2^{\beta_j - 1} \rfloor \leq n_j < 2^{\beta_j}$ for $j = 1, \ldots, d$ and some $\beta \in \mathbb{N}_0^d$. Hence $2^{|\beta|-d} \leq \bar{n}_1 \ldots \bar{n}_d = |n|_0 \leq N$. In particular $|\beta| \leq m + d$ and so $n \in Q_{m+d}$. □

In Figure 2.7 we demonstrate these inclusions. The step hyperbolic cross index set Q_m allows us to scrutinise the pointwise convergence rate of the approximation $\mathcal{F}_N[f]$. Indeed, we have:

Theorem 2.32. *Suppose that $Q_m = \cup_{\beta \in W_m} \rho(\beta)$ is the step hyperbolic cross index set, where W_m is given by (2.49), and that $\mathcal{F}_N[f]$ is given by (2.48). Suppose further that $f \in \mathrm{H}^{2k+3}_{mix}(\Omega)$ obeys the first $k \in \mathbb{N}_0$ derivative conditions. 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 Ω.*

Proof. We first claim that the term $\mathcal{F}_\beta[f]$ defined by (2.47) satisfies $\mathcal{F}_\beta[f](x) = \mathcal{O}\left(2^{-2(k+1)|\beta|}\right)$, $|\beta| \to \infty$. We prove this result by induction on d. For $d = 1$, this follows immediately from known properties of univariate expansions. Suppose now that the result holds for all functions f of at most $(d-1)$ variables. Consider the asymptotic expansion of $\hat{f}_n^{[i]}$. Since f obeys the first k derivative conditions, Lemma 2.13 gives

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

Here $\widehat{p}_k{}^{[i]}_n$ is the modified Fourier coefficient of the univariate polynomial $p_k^{[i]}$ of degree $2k+2$ that satisfies $\mathcal{B}_r^{[i]}[f] = \delta_{r,k}$, $r \in \mathbb{N}_0$, where $\mathcal{B}_r^{[i]}$ is the quantity defined in (2.24). Note that existence of such a polynomial is guaranteed (see Section 5.2.1 of Chapter 5).

The quantity $\mathcal{A}^{[i]}_{k_t, n_{\bar{t}}}[f]$ 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

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

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

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

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

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

which completes the first step of the proof.

Since the main result has already been proved in Theorem 2.22 for the approximation $\mathcal{F}_N[f]$ based on the full index set (2.33), it suffices to consider the difference between this and the approximation based on the step hyperbolic cross Q_m. This difference is precisely

$$\sum_{\substack{|\beta|>m \\ |\beta|_\infty \le m}} \mathcal{F}_\beta[f](x) = \sum_{|\beta'|_\infty=0}^{m} \sum_{\beta_d=m-|\beta'|}^{m} \mathcal{F}_\beta[f](x),$$

where $\beta' = (\beta_1, \ldots, \beta_{d-1})$ contains the first $(d-1)$ entries of β. Hence, using the previous result, it follows that

$$\begin{aligned}\sum_{\substack{|\beta|>m \\ |\beta|_\infty \le m}} \mathcal{F}_\beta[f](x) &= \mathcal{O}\left(\sum_{|\beta'|_\infty=0}^{m} 2^{-2(k+1)|\beta'|} \sum_{\beta_d=m-|\beta'|}^{m} 2^{-2(k+1)\beta_d}\right) \\ &= \mathcal{O}\left(\sum_{|\beta'|_\infty=0}^{m} 2^{-2(k+1)|\beta'|} 2^{-2(k+1)(m-|\beta'|)}\right) = \mathcal{O}\left(m^{d-1} 2^{-2(k+1)m}\right),\end{aligned}$$

which completes the proof. □

The inclusion (2.50) indicates that an analogous result holds for the approximation based on the hyperbolic cross (2.41). A numerical example, demonstrating this faster pointwise rate of convergence, is given in Figure 2.8. We mention in passing that Lemma 2.23, concerning the pointwise convergence of Laplace–Dirichlet expansions based on the full index set (2.33), is also readily extended to this setting.

2.10.4 Optimized hyperbolic cross index sets

Thus far we have considered (step) hyperbolic cross index sets that arise from the uniform or $\mathrm{L}^2(\Omega)$ norms. Such a construct mitigates the curse of dimensionality to a $(\log N)^{d-1}$ factor. However, this effect can be completely removed by introducing so-called *optimized hyperbolic cross* index sets [75, 76].[20]

[20] This approach is a generalisation of the so-called *energy norm* hyperbolic cross considered in [40, 41].

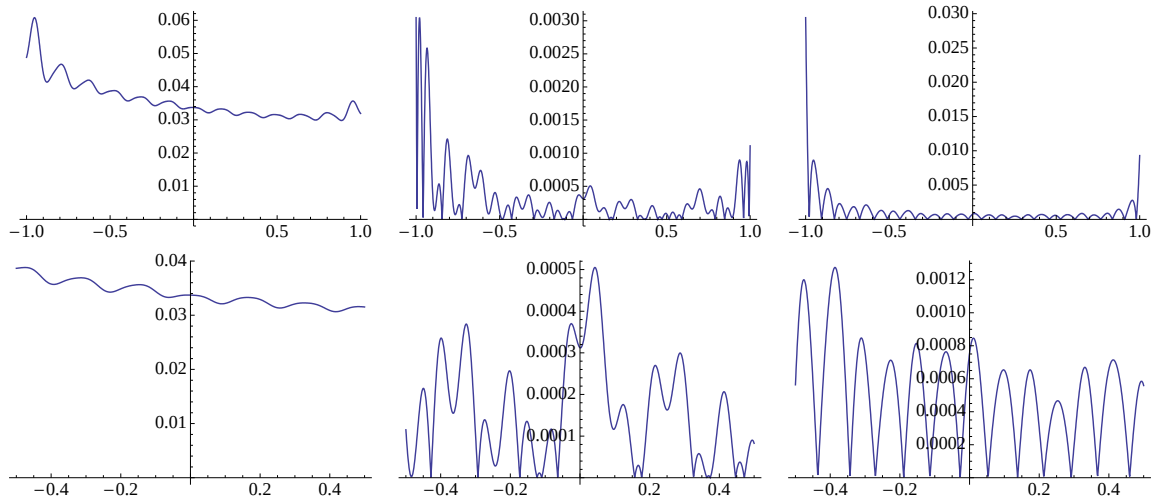


Figure 2.8: Absolute error $|f(x,y_0)-\mathcal{F}_{50}[f](x,y_0)|$, where $f(x_1,x_2)=(x_1^2-x_1+4)\cos 2x_2\sin 3x_2$ and $\mathcal{F}_{50}[f]$ is the Laplace–Neumann approximation based on the hyperbolic cross index set (2.41), for $-1\leq x\leq 1$ (top row) and $-\frac{1}{2}\leq x\leq \frac{1}{2}$ (bottom row) and $y_0=1,\frac{2}{3},\frac{1}{3}$ (left to right).

Such index sets arise from considerations of the $\mathrm{H}^r(\Omega)$ norm for values $r\in\mathbb{R}$ (not necessarily integer). Proceeding as in [75], we obtain the index set

$$I_{N,\sigma}=\left\{n\in\mathbb{N}_0^d : |n|_0|\bar{n}|_\infty^{-\sigma}\leq N^{1-\sigma}\right\}, \tag{2.51}$$

where $-\infty<\sigma\leq 1$ and $\bar{n}=(\bar{n}_1,\ldots,\bar{n}_d)$. Observe that $I_{N,\sigma}\subseteq I_{N,\tau}$ provided $\tau\leq\sigma$.

For $\sigma=-\infty$ or $\sigma=0$, $I_{N,\sigma}$ reduces to the full (2.33) or hyperbolic cross (2.41) index set respectively. Our interest lies with values $0<\sigma\leq 1$, for which $|I_{N,\sigma}|=\mathcal{O}(N)$, as we shall now demonstrate[21]:

Lemma 2.33. *Suppose that $\theta_{\sigma,d}(t)$ is the number of terms $n\in\mathbb{N}_0^d$ such that $|n|_0|\bar{n}|_\infty^{-\sigma}\leq t^{1-\sigma}$. Then, for $0<\sigma<1$ we have*

$$\theta_{\sigma,d}(t)=d\{\zeta((1-\sigma)^{-1})\}^{d-1}t+\textit{lower order terms.}$$

When $\sigma=1$, $\theta_{1,d}(t)=dt$.

Proof. The proof of this result is standard (see [75]). We first note that if $n\in\mathbb{N}_0^d$ with $|n|_0|\bar{n}|_\infty^{-\sigma}\leq t^{1-\sigma}$, then $|n|_\infty\leq t$. Furthermore, if $|n|_\infty=n_d$ then $1\leq n_d\leq t|n'|_0^{-(1-\sigma)^{-1}}$, where $n'=(n_1,\ldots,n_{d-1})$. Hence

$$\begin{aligned}\theta_{\sigma,d}(t)=\sum_{\substack{|n|_\infty\leq t\\ |n|_0|\bar{n}|_\infty^{-\sigma}\leq t^{1-\sigma}}}1&=d\sum_{|n'|_\infty\leq t}\sum_{n_d=1}^{t|n'|_0^{-(1-\sigma)^{-1}}}1+\text{lower order terms}\\&=dt\left(\sum_{n=1}^{t}n^{-(1-\sigma)^{-1}}\right)^{d-1}+\text{lower order terms.}\end{aligned}$$

[21] Estimates for $-\infty<\sigma<0$ can also be established [75]. However, we shall not consider this.

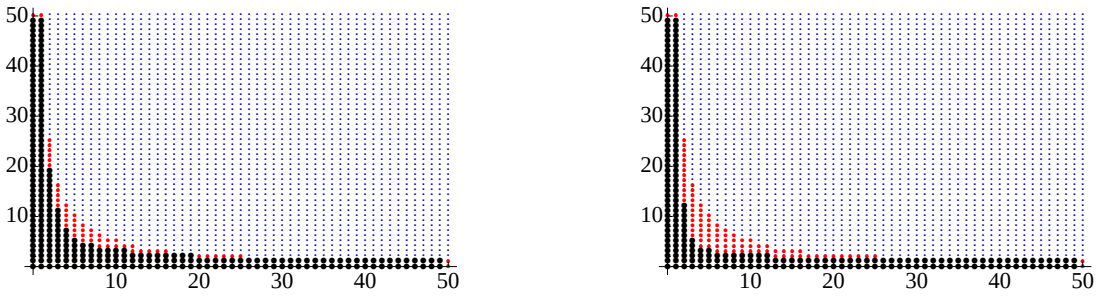


Figure 2.9: Graphs of the index sets (2.33) (small dots), (2.41) (larger dots) and (2.51) (largest dots) for $\sigma = \frac{1}{4}$ (left diagram), $\sigma = \frac{1}{2}$ (right diagram) and $N = 50$.

	$d=2$		$d=3$		$d=4$	
index set	$N=10^2$	$N=10^3$	$N=10^2$	$N=10^3$	$N=10^2$	$N=10^3$
(2.33)	1.02×10^4	1.02×10^6	1.03×10^6	1.00×10^9	1.04×10^8	1.00×10^{12}
(2.41)	6.83×10^2	9.07×10^3	3.22×10^3	5.36×10^4	1.28×10^4	2.57×10^5
$\sigma=\frac{1}{4}$	5.60×10^2	6.59×10^3	2.26×10^3	3.07×10^4	7.86×10^3	1.22×10^5
$\sigma=\frac{1}{2}$	4.69×10^2	5.00×10^3	1.62×10^3	1.84×10^4	4.94×10^3	5.98×10^4

Table 2.1: Comparison of the sizes of the index sets (2.33), (2.41) and (2.51) for $d = 2, 3, 4$ and $N = 10^2, 10^3$. All values to three significant figures.

Since $\sum_{n=1}^{t} n^{-r-1} = \zeta(r+1) + \mathcal{O}\left(t^{-r}\right)$ for $r > 0$, the result follows immediately. □

Typical forms of this index set are given in Figure 2.9. A comparison of the number of terms in this and other index sets is given in Table 2.1. We note that, with $\sigma = \frac{1}{2}$ and $N = 100$, for example, the optimized hyperbolic cross (2.51) contains less than half the number of terms of the hyperbolic cross (2.41).

Next, we address the convergence rate of approximations based on $I_{N,\sigma}$. We have:

Lemma 2.34. *Suppose that $f \in \mathrm{H}^{2k+l}(\Omega)$, $l = 0, 1$, obeys the first $k \in \mathbb{N}_0$ derivative conditions and $\mathcal{F}_N[f]$ is based on the optimized hyperbolic cross (2.51). Then*

$$\|f - \mathcal{F}_N[f]\|_s \leq c_{r,s} N^{\frac{1-\sigma}{d-\sigma}(s-r)} \|f\|_r, \quad r = s, \ldots, 2k+l, \quad s = 0, \ldots, 2k+l,$$

for some positive constant $c_{r,s}$ independent of N and f. Moreover, if $f \in \mathrm{H}^{2k+l}_{mix}(\Omega)$, then

$$\|f - \mathcal{F}_N[f]\|_s \leq c_{r,s} \|f\|_{s,mix} \begin{cases} N^{s-r} & \sigma \leq \frac{s}{r} \\ N^{\frac{1-\sigma}{d-\sigma}(s-dr)} & \sigma > \frac{s}{r} \end{cases}.$$

Proof. As in Lemmas 2.24 and 2.28 we have

$$\|f - \mathcal{F}_N[f]\|_s^2 \leq c \sum_{i \in \{0,1\}^d} \sum_{n \notin I_{N,\sigma}} |\hat{f}_n^{[i]}|^2 |\bar{n}|_\infty^{2s} \leq c \max_{n \notin I_{N,\sigma}} \left\{ |\bar{n}|_\infty^{2(s-r)} \right\} \|f\|_r^2.$$

Note that $|n|_0 \leq |\bar{n}|_\infty^d$. Hence, if $n \notin I_{N,\sigma}$ then $|\bar{n}|_\infty^{d-\sigma} \geq N^{1-\sigma}$. Substituting this into the previous expression now gives the first result.

Now suppose that $f \in \mathrm{H}^{2k+l}_{\mathrm{mix}}(\Omega)$. Then, using standard characterisations and the corresponding result for the full index set, we obtain

$$\begin{aligned}\|f - \mathcal{F}_N[f]\|_s^2 &\leq c \sum_{i\in\{0,1\}^d} \sum_{\substack{n\notin I_{N,\sigma}\\ |n|_\infty \leq N}} |\hat{f}_n^{[i]}|^2 |\bar{n}|_\infty^{2s} + N^{2(s-r)} \|f\|_{r,\mathrm{mix}}^2 \\ &\leq c \max_{\substack{n\notin I_{N,\sigma}\\ |n|_\infty \leq N}} \left\{ |\bar{n}|_\infty^{2s} |n|_0^{-2r} \right\} \|f\|_{r,\mathrm{mix}}^2 + N^{2(s-r)} \|f\|_{r,\mathrm{mix}}^2.\end{aligned}$$

For $n \notin I_{N,\sigma}$ we have $|\bar{n}|_\infty^s |n|_0^{-r} = |\bar{n}|_\infty^{s-\sigma r} \left(|n|_0 |\bar{n}|_\infty^{-\sigma}\right)^{-r} \leq |\bar{n}|_\infty^{s-\sigma r} N^{-(1-\sigma)r}$. If $\sigma \leq \frac{s}{r}$ then the observation that $|\bar{n}|_\infty \leq N$ immediately gives the result. Conversely, if $\sigma > \frac{s}{r}$ then we use the inequality $|\bar{n}|_\infty^{d-\sigma} \geq N^{1-\sigma}$ once more. □

Observe that when $\sigma = 0$ and $\sigma = -\infty$ we recover the results of Lemmas 2.24 and 2.28 respectively. Unsurprisingly, as in previous sections, Lemma 2.34 does not provide an optimal estimate for the convergence rate when the function f has sufficient regularity. To address this scenario, we first require the following lemma:

Lemma 2.35. *Suppose that $0 < \sigma \leq 1$ and $\gamma_{r,\sigma,d}(t) = \sum_{n\notin I_{\sigma,t}} \bar{n}^{-r-1}$ for $r > 0$. Then*

$$\gamma_{r,\sigma,d}(t) = c_{r,\sigma,d} t^{-\frac{d(1-\sigma)}{d-\sigma} r} (\log t)^{d-2} + \textit{lower order terms},$$

for some positive constant $c_{r,\sigma,d}$ independent of f and N. Furthermore, if $r > s > 0$, $j = 1, \ldots, d$ and $\delta_{r,s,\sigma,d}(t) = \sum_{n\notin I_{\sigma,t}} \bar{n}^{-r-1} \bar{n}_j^s$, then

$$\delta_{r,s,\sigma,d}(t) = c_{r,s,\sigma,d} t^{s-r} + \textit{lower order terms},$$

provided $\sigma < \frac{s}{r}$.

As in Lemma 2.30, it is possible to prescribe exact values to such constants. It is also possible to assess $\delta_{r,s,\sigma,d}(t)$ when $\sigma \geq \frac{s}{r}$. However, we shall not pursue this.

Proof. Consider first $\gamma_{r,\sigma,d}(t)$. Without loss of generality we may assume that $|\bar{n}|_\infty = \bar{n}_d$. Since $n \notin I_{N,\sigma}$ we have $n_d \geq t^{\frac{1-\sigma}{d-\sigma}}$. Hence

$$\gamma_{r,\sigma,d}(t) = c \sum_{n_d \geq t^{\frac{1-\sigma}{d-\sigma}}} n_d^{-r-1} \sum_{|n'|_0 \geq (t n_d^{-1})^{1-\sigma}} (\bar{n}')^{-r-1},$$

where $n' = (n_1, \ldots, n_{d-1})$. Using Lemma 2.30 we obtain

$$\begin{aligned}\gamma_{r,\sigma,d}(t) &= c t^{-r(1-\sigma)} (\log t)^{d-2} \sum_{n_d \geq t^{\frac{1-\sigma}{d-\sigma}}} n_d^{-r\sigma-1} + \text{lower order terms} \\ &= c_{r,\sigma,d} t^{-r(1-\sigma)} t^{\frac{-r\sigma(1-\sigma)}{(d-\sigma)}} (\log t)^{d-2} + \text{lower order terms} \\ &= c_{r,\sigma,d} t^{-\frac{d(1-\sigma)}{d-\sigma} r} (\log t)^{d-2} + \text{lower order terms},\end{aligned}$$

as required.

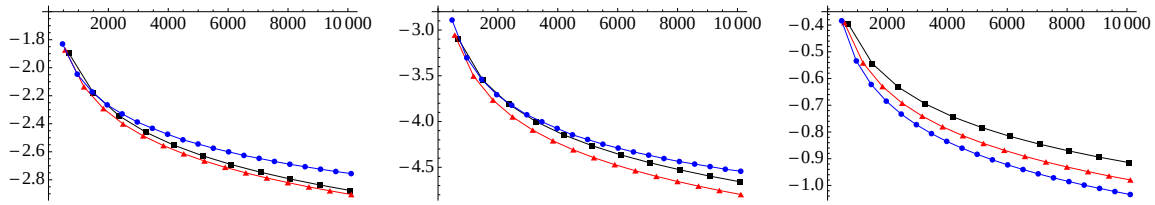


Figure 2.10: Comparison of the hyperbolic cross (2.51) for $\sigma = 0$ (squares), $\sigma = \frac{1}{4}$ (triangles) and $\sigma = \frac{1}{2}$ (circles). Log errors $\log_{10} \|f - \mathcal{F}_N[f]\|_\infty$ (left), $\log_{10} \|f - \mathcal{F}_N[f]\|$ (middle), $\log_{10} \|f - \mathcal{F}_N[f]\|_1$ (right) against number of terms, where $f(x_1, x_2) = (4 + x_1^2 - x_2) \cos 2x_2 \sin 3x_2$.

Next we consider $\delta_{r,s,\sigma,d}(t)$. Without loss of generality, $j = d$. In this case, it suffices to consider only those $n \notin I_{\sigma,t}$ with $|n|_\infty \leq t$ and $|n|_\infty = n_d$. For such n, we have $|n'|_0 |n'|_\infty^{-\sigma} \geq t^{1-\sigma} n_1^{-1}$, where $n' = (n_2, \ldots, n_d)$. We now assume that the result holds for $d - 1$. Then

$$\begin{aligned}
\gamma_{r,\sigma,d}(t) &= c \sum_{n_1=1}^{t} n_1^{-r-1} \sum_{\substack{|n'|_0 |n'|_\infty^{-\sigma} \\ \geq t^{1-\sigma} n_1^{-1}}} |n'|_0^{-r-1} n_d^s + \text{lower order terms} \\
&= c \sum_{n_1=1}^{t} n_1^{-r-1} \left(t n_1^{-(1-\sigma)^{-1}} \right)^{s-r} + \text{lower order terms} \\
&= c t^{s-r} \sum_{n_1=1}^{t} n^{-(s-r)(1-\sigma)^{-1} - r - 1} + \text{lower order terms} \\
&= c_{r,s,\sigma,d} t^{s-r} + \text{lower order terms},
\end{aligned}$$

as required. □

Theorem 2.36. *Suppose that $f \in \mathrm{H}^{2k+2}_{mix}(\Omega)$ obeys the first k Neumann derivative conditions and $I_{N,\sigma}$ is the optimized hyperbolic cross index set (2.51). Then, for $s = 1, \ldots, 2k+1$,*

$$\begin{aligned}
&\|f - \mathcal{F}_N[f]\|_\infty \leq c_k \|f\|_{2k+2,mix} N^{-\frac{d(1-\sigma)(2k+1)}{d-\sigma}} (\log N)^{d-2}, \\
&\|f - \mathcal{F}_N[f]\| \leq c_{k,0} \|f\|_{2k+2,mix} N^{-\frac{d(1-\sigma)(4k+3)}{2(d-\sigma)}} (\log N)^{\frac{d-2}{2}}, \\
&\|f - \mathcal{F}_N[f]\|_s \leq c_{k,s} \|f\|_{2k+2,mix} N^{s-2k-\frac{3}{2}}, \quad \textit{provided } \sigma < \tfrac{2s}{4k+3},
\end{aligned}$$

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

In view of Theorem 2.29, the pertinent observation is that the convergence rate in certain norms of the approximation based on the optimized hyperbolic cross (2.51) is slower than that of the approximation based on the $\mathrm{L}^2(\Omega)$ norm hyperbolic cross (2.41). As described in [41], this is unsurprising: the hyperbolic cross (2.41) is already optimized with respect to the $\mathrm{L}^2(\Omega)$ and uniform norms, so any reduction in size will lead to a deterioration in the convergence rate. Nonetheless, the convergence rate measured in the $\mathrm{H}^r(\Omega)$ norm, $r \in \mathbb{N}$, remains the same, thus making such techniques viable in certain applications, including the discretisation of partial differential equations.

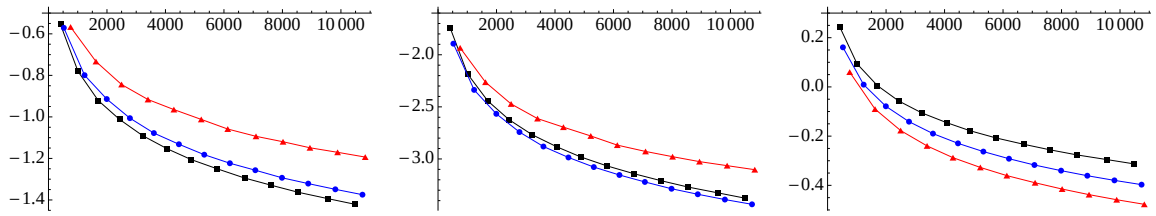


Figure 2.11: Comparison of the hyperbolic cross (2.51) for $\sigma = 0$ (squares), $\sigma = \frac{1}{4}$ (triangles) and $\sigma = \frac{1}{2}$ (circles). Log errors $\log_{10} \|f - \mathcal{F}_N[f]\|_\infty$ (left), $\log_{10} \|f - \mathcal{F}_N[f]\|$ (middle), $\log_{10} \|f - \mathcal{F}_N[f]\|_1$ (right) against number of terms, where $f(x_1, x_2, x_3) = (x_1^2 + 4) \cos x_2 \sin 2x_2 \mathrm{e}^{-\frac{1}{2}x_3}$.

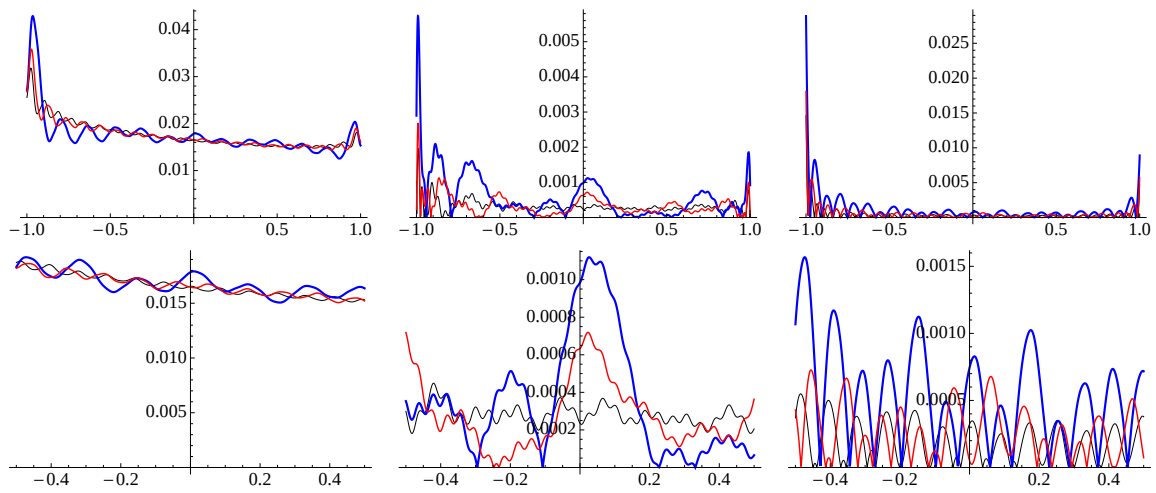


Figure 2.12: Absolute error $|f(x, y_0) - \mathcal{F}_{50}[f](x, y_0)|$, where $f(x_1, x_2) = (x_1^2 - x_1 + 4) \cos 2x_2 \sin 3x_2$ and $\mathcal{F}_{50}[f]$ is the modified Fourier approximation based on the hyperbolic cross (2.51) with $\sigma = \frac{1}{2}$ (thick line), $\sigma = \frac{1}{4}$ (thinner line) or $\sigma = 0$ (thinnest line), for $-1 \leq x \leq 1$ (top row), $-\frac{1}{2} \leq x \leq \frac{1}{2}$ (bottom row) and $y_0 = 1, \frac{2}{3}, \frac{1}{3}$ (left to right).

A comparison of the approximation error for various values of σ is given in Figures 2.10 and 2.11. As established in Theorem 2.29, the optimized hyperbolic cross approximation offers a lower $\mathrm{H}^1(\Omega)$ norm error for the equal number of terms. Conversely, both the uniform and $\mathrm{L}^2(\Omega)$ norm errors are larger.

The pointwise rate of convergence can also be assessed. As in previous scenarios, the convergence rate inside the domain is exactly one power of N faster than on the boundary. This is analysed in an identical manner to the $\mathrm{L}^2(\Omega)$ norm hyperbolic cross case, upon introduction of a suitable step hyperbolic cross. We shall not pursue this further. Numerical results are given in Figure 2.12.

This concludes our discussion of hyperbolic cross approximations. By the introduction of suitable index sets, we have demonstrated how the curse of dimensionality can be broken to a significant extent. We end this chapter with two brief sections. The first addresses Laplace eigenfunctions expansions relating to other boundary conditions. In the second, we briefly describe the numerical quadratures employed to calculate modified Fourier coefficients.

2.11 Other boundary conditions

The focus of this chapter has been the analysis of expansions in Laplace eigenfunctions subject to either homogeneous Dirichlet or Neumann boundary conditions. The key to this study is the duality enjoyed by these bases (see Lemmas 2.4 and 2.5).

Such techniques are applicable to other eigenfunction expansions. As we shall describe in Chapter 3, there is a natural extension to certain higher, even-order differential operators accompanied by suitable boundary conditions. However, the Laplace operator itself can be equipped with numerous other boundary conditions, some of which yield eigenfunction expansions that can be studied in a virtually identical manner.

For example, the univariate eigenfunctions

$$\phi_n^{[0]}(x) = \cos((n-\tfrac{3}{4})\pi x + \tfrac{1}{4}\pi), \quad \phi_n^{[1]}(x) = \cos((n-\tfrac{1}{4})\pi x - \tfrac{1}{4}\pi), \quad n \in \mathbb{N}, \tag{2.52}$$

that arise from the mixed boundary conditions $\phi(1) = \phi'(-1) = 0$ are amenable to such techniques (naturally, so are their multivariate extension). Note that the dual functions in this case are the eigenfunctions that arise from the mixed boundary conditions $\phi'(1) = \phi(-1) = 0$, given by

$$\phi_n^{[0]}(x) = \sin((n-\tfrac{3}{4})\pi x + \tfrac{1}{4}\pi), \quad \phi_n^{[1]}(x) = \sin((n-\tfrac{1}{4})\pi x - \tfrac{1}{4}\pi), \quad n \in \mathbb{N}. \tag{2.53}$$

Eigenfunctions arising from the Robin boundary conditions $\phi'(\pm 1) + \theta\phi(\pm 1) = 0$ can also be studied. Here

$$\begin{aligned} \phi_0^{[0]}(x) &= (\theta^{-1}\sinh(2\theta))^{-\frac{1}{2}} \mathrm{e}^{-\theta x}, \quad \phi_n^{[0]}(x) = (n^2\pi^2 + \theta^2)^{-\frac{1}{2}} \left(n\pi\cos n\pi x - \theta \sin n\pi x\right), \quad n \in \mathbb{N}, \\ \phi_n^{[1]}(x) &= ((n-\tfrac{1}{2})^2\pi^2 + \theta^2)^{-\frac{1}{2}} \left((n-\tfrac{1}{2})\pi \sin(n-\tfrac{1}{2})\pi x + \theta\cos(n-\tfrac{1}{2})\pi x\right), \quad n \in \mathbb{N}. \end{aligned} \tag{2.54}$$

Appropriate duality stems from the action of the operator $\partial_x + \theta\mathcal{I}$, where $\mathcal{I}$ is the identity operator: if $\mathcal{F}_N[f]$ is the truncated expansion of f in Laplace–Robin eigenfunctions, then $(\mathcal{F}_N[f])' + \theta\mathcal{F}_N[f]$ is the truncated expansion of $f' + \theta f$ in Laplace–Dirichlet eigenfunctions.

For the purposes of function approximation, none of these bases will offer a faster convergence rate than modified Fourier expansions. In fact, the expansion in mixed eigenfunctions (2.52) or (2.53) converges at the same rate as the expansion in Laplace–Dirichlet eigenfunction: in other words, one power of N slower. Expansions in the Laplace–Robin eigenfunctions (2.54) converge at the same rate as their Laplace–Neumann counterparts. Moreover, it can be shown that no Laplace eigenfunction expansion will offer a faster convergence rate than the modified Fourier case [38]. Nonetheless, as we demonstrate in Chapter 4, such eigenfunction bases are each well suited to the spectral discretisation of boundary value problems subject to the same boundary conditions.

Unfortunately, this duality technique appears limited to these types of boundary conditions. Given general boundary conditions $a_\pm\phi(\pm 1) + b_\pm\phi(\pm 1) = 0$, of regular, separable type [127], it is not clear how to adapt this approach for the case $a_- \neq a_+$, $b_- \neq b_+$.

We remark in passing that, in the multivariate setting, a great variety of nonseparable boundary conditions can be prescribed to the Laplace operator. However, the corresponding eigenfunctions are themselves nonseparable, rendering them unsuitable for practical purposes. Handling such boundary conditions in, for example, the spectral approximation of partial differential equations is typically a difficult task. We shall consider this briefly in Chapter 4.

2.12 Computation of modified Fourier coefficients

The final issue we address in this chapter is the numerical computation of the modified Fourier[22] coefficients $\hat{f}_n^{[i]}$. This topic was first considered (in the univariate setting) in [94], and generalised to the d-variate cube in [95]. The cornerstone of the schemes developed therein is the observation that the integrand $f(x)\phi_n^{[i]}(x)$ oscillates rapidly for large n. In recent years, great progress has been made in the design of numerical methods for highly oscillatory integrals [89]. Important examples include Filon-type methods [93], Levin-type methods [132] and the method of numerical stationary phase [91]. Rather than high oscillation being a barrier to effective computation, such methods exploit it: as the frequency ω (or in this case n) increases, the error typically decreases. Furthermore, the number of coefficients involved in the approximation is essentially independent of ω. The resulting method is adaptive: changing ω does not require the recalculation of any coefficients. Such behaviour contrasts sharply with classical quadrature schemes—for example, standard Gaussian quadrature—whose accuracy declines with increasing ω.

In the context of modified Fourier coefficients, Filon-type methods have been most widely studied (a Levin-type method is employed in [134]). We now describe this method in greater detail.

2.12.1 Filon-type methods

Suppose first that $d = 1$. The basis for the Filon method is the asymptotic expansion

$$\hat{f}_n^{[i]} = A_{k,n}^{[i]}[f] + \mathcal{O}\left(n^{-2k-2}\right) = \sum_{r=0}^{k-1} \frac{(-1)^{n+i}}{(\mu_n^{[i]})^{r+1}} \mathcal{B}_r^{[i]}[f] + \mathcal{O}\left(n^{-2k-2}\right). \tag{2.55}$$

Truncating this expansion after k terms leads to the so-called *asymptotic method* $\hat{f}_n^{[i]} \approx A_{k,n}^{[i]}[f]$. The asymptotic order[23] of this approximation is $2k+2$ and, since $\hat{f}_n^{[i]} = \mathcal{O}\left(n^{-2}\right)$, the relative asymptotic order is $\mathcal{O}\left(n^{-2k}\right)$. Note that this approach requires explicit calculation of the derivatives $f^{(2r+1)}(\pm 1)$, $r = 0, \ldots, k-1$. However, as described in [92, 94], derivatives can be replaced by finite differences in a straightforward manner.

Unfortunately, the asymptotic method can only be used when n is sufficiently large. In practice, the approximation $A_{k,n}^{[i]}[f]$ is often unacceptable for realistic values of k and n. Regardless, the expansion (2.55) is the starting point for Filon-type methods, which we now describe.

The Filon-type method is very easily defined. Given nodes $-1 = c_1 < c_2 < \ldots < c_\nu = 1$ and multiplicities $m_1, \ldots, m_\nu$ we first construct a polynomial ϕ such that

$$\phi^{(2r)}(c_s) = f^{(2r+1)}(c_s), \quad r = 0, \ldots, m_s - 1, \quad s = 1, 2, \ldots, \nu.$$

If $p(x) = f(0) + \int_0^x \phi(t)\, dt$, then we refer to

$$Q_{m,n}^{[i]}[f] = \int_{-1}^{1} p(x)\phi_n^{[i]}(x)\, \mathrm{d}x,$$

[22] The techniques described in this section are equally applicable (with only minor modifications) to other Laplace eigenfunctions. However, we shall focus on the modified Fourier case.

[23] If an approximation to $\hat{f}_n$ commits an error of $\mathcal{O}\left(n^{-m}\right)$ we say it is of *asymptotic order* m.

as a *Filon-type* approximation based on nodes $c_1, \ldots, c_\nu$ and multiplicities $m = (m_1, \ldots, m_\nu)$. The asymptotic order of this approximation is $2k + 2$, where $k = \min\{m_1, m_\nu\}$.

To relate $Q^{[i]}_{m,n}[f]$ to the asymptotic method, we observe that

$$Q^{[i]}_{m,n}[f] = A^{[i]}_{k,n}[f] + E^{[i]}_{m,n}[f], \tag{2.56}$$

where the residual $E^{[i]}_{m,n}[f]$ is $\mathcal{O}\left(n^{-2k-4}\right)$. This interpretation explains the effect of the internal nodes $c_2, \ldots, c_{\nu-1}$ of the Filon-type method. Such nodes, whilst not increasing the asymptotic order, act to approximate the higher-order terms in the asymptotic expansion (2.55).

In view of (2.56), we may expect the Filon-type method to behave in a similar manner to the $(k+1)^{\text{th}}$ asymptotic method $A^{[i]}_{k+1,n}[f]$. However, Filon-type methods typically offer greatly superior performance. Numerical examples attest to the fact that Filon-type methods yield high accuracy even when n is small [94]. This can be explained as follows: for large n, accuracy is assured by rapid decay of the asymptotic expansion, whereas for small n, the high order of the underlying classical quadrature ensures precision.

The interpretation (2.56) also provides a compelling alternative means to devise Filon-type schemes. If we make the ansatz

$$E^{[i]}_{k,n}[f] = \frac{(-1)^{n+i}}{(\mu^{[i]}_n)^{k+1}} \sum_{s=1}^{\nu} \sum_{r=0}^{m_s-1} b^{[i]}_{r,s} f^{(2r+1)}(c_s),$$

with values $b^{[i]}_{r,s}$ independent of n and f, then high accuracy will occur, provided such values are chosen so that the approximation

$$\sum_{s=1}^{\nu} \sum_{r=0}^{m_s-1} b_{r,s} g^{(2r+1)}(c_s) \approx \mathcal{B}^{[i]}_k[g] = (-1)^k \left[g^{(2k+1)}(1) + (-1)^{i+1} g^{(2k+1)}(-1) \right],$$

is exact for all polynomials f of maximal degree. Hence, the problem of designing Filon-type quadratures is reduced to the approximation of derivatives by finite differences. This methodology typically allows for easier design and construction of efficient schemes [95].

2.12.2 Exotic quadrature

Due to its asymptotic nature, the Filon-type method cannot be used for nonoscillatory integrals, the most pertinent example of this being the coefficient $\hat{f}^{[0]}_0$. Moreover, for small n, the Filon-type approximation will not offer sufficient precision. An alternative is to use Gaussian quadrature in this setting. However, this action requires additional function evaluations and, more importantly, the derivative values computed as part of the Filon-type quadrature are wasted. The idea proposed in [94] is to reuse such values in classical quadrature schemes, an approach termed *exotic quadrature* [8]. In this spirit, we define the quadrature rule

$$Q[h] = 2h(0) + \sum_{r=1}^{\nu} \sum_{s=0}^{m_r-1} b_{r,s} h^{(2s+1)}(c_r) \approx \int_{-1}^{1} h(x)\,\mathrm{d}x, \tag{2.57}$$

with weights $b_{r,s}$ chosen to maximise the order of the scheme. Depending on the coefficient we wish to approximate, we set $h(x) = f(x)$ or $h(x) = f(x)\phi^{[i]}_n(x)$.

We remark in passing that both Filon and exotic quadratures use (Hermite) interpolation at internal nodes. No general theory currently exists pertaining to the optimal location of such nodes. However, numerical examples presented in [8, 94, 95] suggest that these values should be chosen to maximise the order of (2.57), if possible.

2.12.3 Multivariate modified Fourier coefficients

The design of effective quadratures in the multivariate setting is complicated by the fact that, for various values of n, the integrand $f(x)\phi_n^{[i]}(x)$ oscillates rapidly in some variables and not in others.

For parameters $n = (n_1, \ldots, n_d)$ with $\min\{n_j\} \gg 1$, the multivariate asymptotic expansion (2.28), suitably truncated, is once more the starting point. The construction of Filon-type methods based on this expansion is affected by two further issues. First, since only the odd derivatives are used as interpolation conditions, the interpolation problem (an example of a *Birkhoff–Hermite interpolation problem* [115, 116]) may not be solvable. Further, for a particular configuration of nodes, the corresponding interpolation polynomial need not exist. However, both issues can be resolved in the modified Fourier setting. In [95] a Filon-type method was introduced using a so-called *tartan grid* to cover the domain.[24] To construct the Filon-type approximant, the function f and certain partial derivatives are evaluated on this grid.

Along the lines of [94], multivariable exotic quadratures can also be constructed to handle nonoscillatory coefficients. A combination of exotic and Filon quadratures is then used for those coefficients with corresponding high oscillation in only a subset of the variables $n_1, \ldots, n_d$ [95]. As numerical examples demonstrate [87], the Filon-type method is used for the vast majority of coefficients, with exotic quadrature (in one or more variables) used only for those coefficients with at least one very small parameter $n_1, \ldots, n_d$.

The approach outlined above is theoretically clear, but numerous issues remain. For example, robust and accurate error bounds for both Filon-type and exotic quadratures are largely lacking. Recently some advances have been made in the univariate setting [121], however the general picture is far from apparent. On a closely related topic, the stability of such methods is as of yet largely unexplored, and few criteria currently exist for the optimization of quadrature parameters. Consequently, these methods require a great deal of future work before they can be converted into effective algorithms. We refer the reader to [8, 94, 95] for a more thorough discussion of the open problems relating to such schemes. Nonetheless, as we now discuss, such approach provides a compelling alternative to more standard techniques.

2.12.4 Quadrature and the Fast Fourier Transform

Once derivative values are specified (or calculated), we may compute any M coefficients in $\mathcal{O}(M)$ operations using the aforementioned quadratures. Furthermore, this approach is adaptive: we may readily compute any M' additional coefficients in $\mathcal{O}(M')$ operations without recomputing any existing values.[25] This fact permits the use of the hyperbolic cross for modified Fourier expansions, which, as demonstrated, offers significant computational savings.

[24] This particular approach is applicable to tensor-product domains only. Construction of quadratures in the equilateral triangle has been studied in [88].

[25] Having said this, it may be advantageous in practice to recompute existing values to higher accuracy if the truncation parameter N is increased, potentially resulting in a higher computational cost [87].

This approach contrasts sharply with the FFT, which computes all values in the full index set (2.33) in a non-adaptive manner. Moreover, the truncation parameter N must be highly composite. Regardless, in view of Section 2.3, the FFT can be used in conjunction with modified Fourier expansions, provided (2.33) is employed. In particular, the expansion $\mathcal{F}_N[f]$ can be evaluated at N equally spaced nodes in $\mathcal{O}\left(N^d \log N\right)$ operations. Moreover, products and derivatives of modified Fourier sums can be evaluated with the same operational count.

As noted, the classical FFT is unsuitable for hyperbolic cross approximations. In this setting, the SGFFT can be used (provided a step hyperbolic cross is employed) [18, 60]. However, this approach is by no means simple nor straightforward to implement [87]. Nonetheless, in the context of modified Fourier sums, this device can, in theory, be exploited to evaluate products and derivatives, for example, with the resulting operational cost being $\mathcal{O}\left(N(\log N)^d\right)$.

This concludes our study of modified Fourier expansions. We shall return to this theme in Chapters 4 and 5 respectively, where we discuss their applications to boundary value problems and their effective convergence acceleration. Before doing so, however, the next chapter concerns the generalisation of the modified Fourier basis to bases consisting of eigenfunctions of suitable higher-order differential operators.

Chapter 3

Expansions in polyharmonic eigenfunctions

3.1 Introduction

Modified Fourier expansions give the fastest possible convergence rate amongst all Laplace eigenfunction expansions. Had the individual eigenfunctions obeyed additional, higher-order boundary conditions, this rate of convergence would have increased. The aim of this chapter is to demonstrate that modified Fourier expansions and their theory can be successfully generalised to expansions with convergence rates of arbitrary algebraic order. By a judicious choice of both differential operator and boundary conditions, we introduce a one-parameter family of expansions with a uniform convergence rate of $\mathcal{O}\left(N^{-q}\right)$ for any fixed $q \in \mathbb{N}$. The corresponding coefficients decay like $\mathcal{O}\left(n^{-q-1}\right)$. Such expansions share many similar properties with the Laplace case, which corresponds to index $q = 1$. In particular, coefficients can be calculated using similar quadrature methods to those introduced in Section 2.12.

The expansion of a function in Laplace eigenfunctions is one particular example of the much larger field of so-called *Birkhoff expansions* [127]. This topic addresses the expansion of a function in eigenfunctions of an arbitrary linear differential operator with prescribed boundary conditions. The route to generalising modified Fourier expansions lies with first understanding this general scenario. Motivated by such considerations as simplicity of the eigenvalues and eigenfunctions and convergence rate of the expansion, we develop, in this chapter, a family of Birkhoff expansions based on eigenfunctions of a particular class of differential operators (polyharmonic operators) equipped with certain boundary conditions. As we subsequently indicate, such eigenfunctions are optimized for practical computations.

Although Birkhoff expansions have been extensively studied from a theoretical standpoint, few attempts have been made at practical computations (outside of the Fourier setting[1]). Moreover, despite a well-established classical theory for univariate Birkhoff expansions [51, 127], the fundamental characteristics of the particular expansions introduced in this chapter are insufficiently described by such theory. Hence, having described this shortfall in further detail, we provide a full theory of such expansions in the unit interval.

[1]The Fourier basis functions can be viewed as eigenfunctions of the operator $\frac{\mathrm{d}}{\mathrm{d}x}$ equipped with periodic boundary conditions $\phi(1) = \phi(-1)$. We remark, however, that, aside from providing this example, this viewpoint is largely superfluous.

Very little literature currently exists pertaining to Birkhoff expansions for multivariate functions. To this end, after detailing polyharmonic eigenfunction expansions in the unit interval, we next demonstrate an appropriate extension to the d-variate cube. We then present a complete analysis of convergence, thereby generalising the work of Chapter 2 to arbitrary $q \geq 1$ (in particular, the convergence results that we establish make no stipulations regarding the index set employed).

We mention in passing that this topic—the generalisation of modified Fourier expansions to polyharmonic expansions—was originally pursued in [8]. Part of this chapter will summarise salient aspects of that study. The main content, however, builds on this work by both presenting a full convergence theory for such expansions and establishing an extension to the d-variate cube. Elements of this material have recently appeared in the author's paper [6].

3.1.1 Birkhoff expansions

The natural starting point for the generalisation of modified Fourier expansions is the unit interval. An extension to the d-variate cube can only be pursued with sufficient understanding of this case. Our present goal is therefore to determine the univariate differential operator (of fixed, even-order) and boundary conditions with the fastest decay of expansion coefficients (and, correspondingly, the fastest uniform convergence rate). Such a form will not, in general, be unique. Hence, practical considerations, notably simplicity of the eigenfunctions and eigenvalues, will be exploited where possible.

To this end, suppose that $\mathcal{L}_0 = (-1)^q \frac{\mathrm{d}^{2q}}{\mathrm{d}x^{2q}} + \ldots$ is a self-adjoint linear differential operator of order $2q$, $q \in \mathbb{N}$, with smooth coefficients. We could, in theory, drop the assumption of self-adjointness. However, since real eigenvalues are desirable for the purpose of practical computations, it makes sense to enforce this condition. Nothing is gained in terms of convergence or rate of decay of expansion coefficients by considering the non-self-adjoint case[2]. Suppose further that $\mathcal{B}_1[\phi], \ldots, \mathcal{B}_{2q}[\phi]$, $\phi \in \mathrm{C}^{2q-1}[-1,1]$, are $2q$ linearly independent, linear functions of the values $\phi(\pm 1), \phi'(\pm 1), \ldots, \phi^{(2q-1)}(\pm 1)$, giving rise to homogeneous boundary conditions $\mathcal{B}_r[\phi] = 0$, $r = 1, \ldots, 2q$. Such forms can be augmented to form a dual basis $\mathcal{B}_1, \ldots, \mathcal{B}_{4q}$ of the $4q$-dimensional vector space

$$\left\{ \left(\phi(-1), \phi'(-1), \ldots, \phi^{(2q-1)}(-1), \phi(1), \phi'(1), \ldots, \phi^{(2q-1)}(1) \right) : \phi \in \mathrm{C}^{2q-1}[-1,1] \right\},$$

for which the condition

$$\int_{-1}^{1} \mathcal{L}_0[\phi](x)\psi(x)\,\mathrm{d}x = \sum_{r=1}^{4q} \mathcal{B}_r[\phi]\mathcal{B}_{4q+1-r}[\psi] + \int_{-1}^{1} \phi(x)\mathcal{L}_0[\psi](x)\,\mathrm{d}x, \tag{3.1}$$

holds for all $\phi, \psi \in \mathrm{C}^{2q}[-1,1]$.

Under some mild assumptions, the spectrum of $\mathcal{L}_0$ equipped with boundary conditions $\mathcal{B}_r[\phi] = 0$, $r = 1, \ldots, 2q$, is countable, with real eigenvalues $\mu_1, \mu_2, \ldots$ having no finite limit point in $\mathbb{R}$, and orthonormal eigenfunctions $\phi_1, \phi_2, \ldots$ [127]. This indicates that a function

[2] We hasten to add, however, that this situation changes dramatically if odd order operators are considered. In this case, there are a number of prominent non-self-adjoint examples where the classical theory of Birkhoff expansions does not hold. For example, the operator $\frac{\mathrm{d}^3}{\mathrm{d}x^3}$ when equipped with boundary conditions $u(-1) = u(1) = u'(1) = 0$ does not possess a countable spectrum.

$f \in \mathrm{L}^2(-1,1)$ may be expanded in such eigenfunctions:

$$f(x) \sim \sum_{n=1}^{\infty} \hat{f}_n \phi_n(x), \quad \text{where} \quad \hat{f}_n = \int_{-1}^{1} f(x)\phi_n(x)\,\mathrm{d}x.$$

We wish to select an operator $\mathcal{L}_0$ and boundary conditions with both the fastest decay of the expansion coefficients and the simplest eigenfunctions and eigenvalues. Considering the first criterion, let ϕ be an eigenfunction of $\mathcal{L}_0$ with eigenvalue $\mu = \alpha^{2q} \neq 0$. Using (3.1) and applying the boundary conditions $\mathcal{B}_r[\phi] = 0$, $r = 1, \ldots, 2q$, gives

$$\begin{aligned}\int_{-1}^{1} f(x)\phi(x)\,\mathrm{d}x &= \frac{1}{\mu}\int_{-1}^{1} f(x)\mathcal{L}_0[\phi](x)\,\mathrm{d}x \\ &= \frac{1}{\mu}\sum_{r=1}^{4q} \mathcal{B}_r[\phi]\mathcal{B}_{4q+1-r}[f] + \frac{1}{\mu}\int_{-1}^{1} \mathcal{L}_0[f](x)\phi(x)\,\mathrm{d}x \\ &= \frac{1}{\mu}\sum_{r=2q+1}^{4q} \mathcal{B}_r[\phi]\mathcal{B}_{4q+1-r}[f] + \frac{1}{\mu}\int_{-1}^{1} \mathcal{L}_0[f](x)\phi(x)\,\mathrm{d}x.\end{aligned}$$

It is known that $\phi^{(r)}(\pm 1) = \mathcal{O}\left(\alpha^r\right)$ and that the n^{th} value $\alpha_n = \mathcal{O}\left(n\right)$ [127]. Hence

$$\int_{-1}^{1} f(x)\phi(x)\,\mathrm{d}x = \mathcal{O}\left(\alpha^{m-2q}\right),$$

where m is the maximal order of derivative appearing in the forms $\mathcal{B}_{2q+1}, \ldots, \mathcal{B}_{4q}$. We now seek to minimise m over all possible boundary conditions. Since the forms $\mathcal{B}_1, \ldots, \mathcal{B}_{4q}$ are linearly independent, simple arguments demonstrate that $m = q - 1$ is the minimal value. In this case, the highest derivative in both $\mathcal{B}_r$ and $\mathcal{B}_{q+r}$ is of order $q + r - 1$ for $r = 1, \ldots q$ (after a possible reordering). Though numerous different boundary conditions share this property, practical considerations exhort us to choose the simplest. These are the *Neumann boundary conditions*

$$\mathcal{B}_r[\phi] = \phi^{(q+r-1)}(-1), \quad \mathcal{B}_{q+r}[\phi] = \phi^{(q+r-1)}(1), \quad r = 1, \ldots, q.$$

It follows that $\hat{f}_n = \mathcal{O}\left(n^{-q-1}\right)$.

Having prescribed 'optimal' boundary conditions, we turn our attention to the operator $\mathcal{L}_0$. Throughout this derivation, aside from the order q and imposition of self-adjointness, $\mathcal{L}_0$ was arbitrary. Once again, given freedom to choose, we resort to simplicity. This leads naturally to the polyharmonic operator $(-1)^q \frac{\mathrm{d}^{2q}}{\mathrm{d}x^{2q}}$.

For these reasons, the remainder of this chapter is devoted to the study of expansions in the *polyharmonic–Neumann* eigenfunctions:

$$(-1)^q \phi^{(2q)} = \alpha^{2q}\phi, \qquad \phi^{(r)}(\pm 1) = 0, \quad r = q, q+1, \ldots, 2q-1. \tag{3.2}$$

Observe that when $q = 1$ this reduces to the Laplace–Neumann case studied previously.

Incidentally, though considerations of simplicity naturally lead us to (3.2), there is also sound theoretical justification. As described in [127], the spectrum of a general operator $\mathcal{L}_0$ is well understood in the asymptotic regime $|\alpha| \to \infty$. Under some mild assumptions, both the eigenvalues and eigenfunctions of a general $2q^{\text{th}}$ order operator $\mathcal{L}_0$ are asymptotic to those of

the polyharmonic operator (the highest order term in $\mathcal{L}_0$) with the same boundary conditions. In other words, *no advantage is gained* from expansions based on eigenfunctions of a more general operator.[3]

3.1.2 Background

Birkhoff expansions have been extensively studied since their introduction by George Birkhoff [25, 26]. Consequently, they are well-developed theoretically [123]. Much is known about both their convergence and the asymptotic behaviour of the eigenvalues and eigenfunctions [23, 51, 127]. In particular, the phenomenon of *equiconvergence*—where a Birkhoff expansion can be related to a model trigonometric expansion and hence studied with classical tools of Fourier analysis—has been extensively explored [123, 154].

Nevertheless, a number of omissions exist. The apparently obvious statement that Neumann boundary conditions yield uniformly convergent expansions and the fastest possible rate of convergence seems to be lacking. Indeed, classical convergence results typically assume that the function being approximated satisfies the same boundary conditions as those associated to the linear operator [123, 157]. Additionally, the majority of studies pertaining to equiconvergence consider only convergence away from the endpoints. From a practical standpoint, such results are of limited use. Furthermore, a significant proportion of existing theory addresses only the worst case scenario, including, for example, the *Dirichlet boundary conditions*

$$\phi^{(r)}(\pm 1) = 0, \quad r = 0, \ldots, q-1, \tag{3.3}$$

which, in contrast to the Neumann case, lack uniform convergence and possess the slowest convergence rate amongst all possible Birkhoff expansions.

Outside of this context, polyharmonic–Neumann eigenfunctions have been notably considered by Krein [108] and Kolmogorov [105] in the theory of n-widths. As a result of these and subsequent investigations, much is known about the zeros of such eigenfunctions [136]. Regardless, to the best of our knowledge, no attempts have been made outside of [8] to devise practical approximation schemes based on such eigenfunctions.

There are two principal reasons for this omission: construction and computation of the eigenvalues and eigenfunctions and numerical evaluation of the coefficients $\hat{f}_n$. Both problems were addressed in [8], and we shall revisit the principal aspects of that study in the course of this chapter.

3.1.3 Key results

The key results of this chapter are divided into two parts: results relating to the spectrum of (3.2) and its eigenfunctions, and results pertaining to the convergence of eigenfunction expansions. Specifically, as regards the former, we establish the following:

1. The eigenfunction of (3.2) corresponding to index α can be expressed as a finite sum of products of trigonometric and hyperbolic functions with real coefficients given as a

[3] A similar statement can also be made regarding more complicated boundary conditions. Under some rather general assumptions, the eigenfunctions corresponding to general boundary conditions $\mathcal{B}_r[\phi] = 0$ are asymptotic to those eigenfunctions corresponding to boundary conditions arising from only the highest order derivative in $\mathcal{B}_r[\cdot]$ [127]. Thus we obtain a similar conclusion: there is no practical advantage gained from equipping the polyharmonic operator with more general boundary conditions.

solution of a $q \times q$ algebraic eigenproblem. The eigenfunctions occur in two cases: even and odd.

2. The eigenvalues are non-negative and, aside from the q-fold zero eigenvalue, positive. Eigenvalues lie in intervals of exponentially small width and the n^{th} value α_n satisfies

$$\alpha_n = \frac{1}{4}(2n + q - 1)\pi + \mathcal{O}\left(\mathrm{e}^{-\gamma_q n\pi}\right), \quad n \gg 1, \tag{3.4}$$

for some constant $\gamma_q > 0$ depending only on q.

3. The eigenfunctions ϕ_n are exponentially close to regular oscillators in compact subsets of $(-1, 1)$. Specifically, for $-1 < x < 1$ and $n \gg 1$,

$$\phi_n(x) = \cos\left[\tfrac{1}{4}(2n + q - 1)\pi x + \tfrac{1}{2}(n + q - 1)\pi\right] + \mathcal{O}\left(\mathrm{e}^{-\frac{1}{2}\gamma_q(1-|x|)n\pi}\right). \tag{3.5}$$

These results have important consequences for practical computation. Simple construction of eigenfunctions and rapid numerical evaluation of eigenvalues (via standard iterative techniques, e.g. Newton–Raphson) follows from 1 and 2. Result 3 asserts that the coefficients $\hat{f}_n$ can be calculated to high accuracy with highly oscillatory methods. We mention in passing that (3.4) and (3.5) represent significant improvements of classical results for Birkhoff expansions. In general, such estimates are known with only $\mathcal{O}\left(n^{-1}\right)$ remainder terms [51, 127]. These improved estimates, however, are rather specific to the polyharmonic–Neumann case (as we discuss further in Section 3.3.3), and this presents yet another compelling reason to develop expansions in such eigenfunctions, as opposed to arbitrary Birkhoff expansions.

The presence of exponentially small error terms in (3.4) and (3.5) not only justifies statements made previously about the computation of eigenvalues, it also allows for a more accurate study of convergence—the second topic we address in this chapter. In particular, we establish the following:

1. For all $q \in \mathbb{N}$, the basis of polyharmonic–Neumann eigenfunctions is dense and orthogonal in $\mathrm{H}^q(-1, 1)$ with respect to the inner product

$$(f, g)_q = (f, g) + (f^{(q)}, g^{(q)}), \quad \forall f, g \in \mathrm{H}^q(-1, 1). \tag{3.6}$$

2. For $r = 0, \ldots, q$ the truncated expansion of a function $f \in \mathrm{H}^r(-1, 1)$ converges to f in the $\mathrm{H}^r(-1, 1)$ norm.
3. The coefficients $\hat{f}_n$ of a function $f \in \mathrm{H}^{q+1}(-1, 1)$ are $\mathcal{O}\left(n^{-q-1}\right)$ for large n.
4. The expansion of $f \in \mathrm{H}^1(-1, 1)$ converges uniformly, and, provided $f \in H^{q+1}(-1, 1)$, the uniform error is $\mathcal{O}\left(N^{-q}\right)$. If, additionally, $f \in H^{q+2}(-1, 1)$ the rate of convergence is $\mathcal{O}\left(N^{-q-1}\right)$ in compact subsets of $(-1, 1)$. Furthermore, a full asymptotic expansion of the error at any point $x \in [-1, 1]$ can be prescribed.
5. Derivative conditions completely determine the convergence rate. If a function f obeys the conditions

$$f^{((2r+1)q+s)}(\pm 1) = 0, \quad r = 0, \ldots, k - 1, \quad s = 0, \ldots, q - 1,$$

then all convergence rates increase by a factor of N^{2kq}.

6. The theory of univariate polyharmonic–Neumann expansions can be scaled up to the d-variate cube via tensor products. This leads to a family of approximation bases corresponding to eigenfunctions of certain subpolyharmonic operators.

Polyharmonic–Neumann eigenfunctions, in theory, facilitate the design of approximations with convergence rates of arbitrary algebraic order. One caveat is required: as q increases, performing practical computations with such eigenfunctions becomes increasingly cumbersome. Moreover, the presence of round-off error also hampers computations. This is described in further detail in Sections 3.2.4 and 3.7. It is not within the scope of this chapter to properly assess the impact of such issues, nor shall we address the comparison of polyharmonic–Neumann approximations with more mature algorithms. Regardless, in view of the applications of modified Fourier expansions, where such an approach have been found to convey a number of benefits, this particular generalisation warrants further study.

3.2 Polyharmonic eigenfunction bases

The operator $\mathcal{L}_0 = (-1)^q \frac{\mathrm{d}^{2q}}{\mathrm{d}x^{2q}}$ equipped with homogeneous Neumann boundary conditions is semi-positive definite: thus, all eigenvalues are nonnegative. Clearly $\mathcal{L}_0[\phi] = 0$ if and only if ϕ is a polynomial of degree $q-1$, meaning that 0 is a q-fold eigenvalue. The corresponding orthonormal eigenfunctions are $\phi_{0,n}$, $n = 0, \dots, q-1$, where $\phi_{0,n}$ is the n^{th} Legendre polynomial.

All other eigenvalues are positive, and it follows from elementary spectral theory that such eigenvalues are simple, countable and have no finite limit point [114]. The corresponding eigenfunctions ϕ_n, $n \in \mathbb{N}$, in combination with $\phi_{0,n}$, $n = 0, \dots, q-1$, form a dense, orthonormal subset of $\mathrm{L}^2(-1,1)$.

As we exhibit in Section 3.2.3, eigenfunctions occur in two flavours, even and odd. Hence, we will occasionally use the notation $\phi_n^{[i]}$, $\phi_{0,n}^{[i]}$, thereby denoting the even ($i = 0$) and odd ($i = 1$) cases explicitly. More frequently, however, we will write $\phi_{0,n}$, ϕ_n and ignore this fact.

3.2.1 Expansions in polyharmonic eigenfunctions

We define the truncated expansion of a function $f \in \mathrm{L}^2(-1,1)$ in polyharmonic–Neumann eigenfunctions as

$$\mathcal{F}_N[f](x) = \sum_{n=0}^{q-1} \hat{f}_{0,n}\phi_{0,n}(x) + \sum_{n=1}^{N} \hat{f}_n \phi_n(x), \quad x \in [-1,1],$$

where $\hat{f}_{0,n} = \int_{-1}^{1} f(x)\phi_{0,n}(x)\,\mathrm{d}x$ and $\hat{f}_n = \int_{-1}^{1} f(x)\phi_n(x)\,\mathrm{d}x$. Due to $\mathrm{L}^2(-1,1)$ orthogonality and density, $\mathcal{F}_N[f]$ converges to f in the $\mathrm{L}^2(-1,1)$ norm. Moreover, a Parseval-type characterisation holds,

$$\|f\|^2 = \sum_{n=0}^{q-1} |\hat{f}_{0,n}|^2 + \sum_{n=1}^{\infty} |\hat{f}_n|^2, \quad \forall f \in \mathrm{L}^2(-1,1). \tag{3.7}$$

Central to analysis of the approximation $\mathcal{F}_N[f]$ is the duality enjoyed by the polyharmonic basis. For $q = 1$, as demonstrated in Lemmas 2.4 and 2.5, such duality is clear: the derivative of a Laplace–Neumann eigenfunction is a Laplace–Dirichlet eigenfunction and the derivative of $\mathcal{F}_N[f]$ is the truncated expansion of f' in Laplace–Dirichlet eigenfunctions. The following lemma generalises this result to $q \geq 1$:

Lemma 3.1. *If we apply the operator $\frac{\mathrm{d}^q}{\mathrm{d}x^q}$ to the set of polyharmonic–Neumann eigenfunctions ϕ_n, we obtain, up to scalar multiples, the set of polyharmonic eigenfunctions that satisfy*

the Dirichlet boundary conditions (3.3). Such eigenfunctions are dense and orthogonal in $\mathrm{L}^2(-1,1)$*. Moreover, for* $f \in \mathrm{H}^q(-1,1)$*,* $(\mathcal{F}_N[f])^{(q)}$ *is the truncated expansion of* $f^{(q)}$ *in such eigenfunctions.*

Proof. It is clear that q-fold differentiation yields the set of polyharmonic–Dirichlet eigenfunctions (note that the polyharmonic–Dirichlet operator has no zero eigenvalue). Density and orthogonality now follow directly from standard spectral theory.

For the second result, we first note that, for $f \in \mathrm{H}^q(-1,1)$,

$$\int_{-1}^{1} f(x)\phi(x)\,\mathrm{d}x = \frac{(-1)^{q+r}}{\alpha^{2q}} \int_{-1}^{1} f^{(r)}(x)\phi^{(2q-r)}(x)\,\mathrm{d}x, \quad r = 0,\dots,q, \tag{3.8}$$

where ϕ is a polyharmonic–Neumann eigenfunction with corresponding eigenvalue $\mu = \alpha^{2q}$. This follows from the equality $\phi^{(2q)} = (-1)^q\alpha^{2q}\phi$ and repeated integration by parts. Now, suppose that $\phi^{(q)} = c\psi$, where ψ is the corresponding normalised polyharmonic–Dirichlet eigenfunction and c is a constant. Using (3.8) with $r = q$ gives

$$c^2 = c^2 \int_{-1}^{1} \psi(x)\psi(x)\,\mathrm{d}x = \int_{-1}^{1} \phi^{(q)}(x)\phi^{(q)}(x)\,\mathrm{d}x = \alpha^{2q}.$$

Moreover, we have

$$\int_{-1}^{1} f(x)\phi(x)\,\mathrm{d}x = \frac{1}{\alpha^{2q}} \int_{-1}^{1} f^{(q)}(x)\phi^{(q)}(x)\,\mathrm{d}x = \frac{1}{c} \int_{-1}^{1} f^{(q)}(x)\psi(x)\,\mathrm{d}x,$$

so that $(f,\phi)\phi^{(q)}(x) = (f^{(q)},\psi)\psi(x)$. The result now follows immediately. □

Straightaway this lemma provokes the following question: what is the corresponding duality for the derivative operator $\frac{\mathrm{d}^p}{\mathrm{d}x^p}$, $p = 1,\dots,q-1$? Unfortunately, we no longer obtain an orthogonal basis. Instead, as we describe in the next section, we obtain polyharmonic eigenfunctions subject to certain non-self-adjoint boundary conditions.

3.2.2 Biorthogonal pairs of polyharmonic eigenfunctions

To describe the case $p = 1,\dots,q-1$, we first recall some general theory of Birkhoff expansions (see [127] for a more thorough exposition). Given an arbitrary linear differential operator $\mathcal{L}_0$ (not necessarily self-adjoint) of order $2q$ equipped with boundary conditions $\mathcal{B}_r[\phi] = 0$, $r = 1,\dots,2q$, we may define the adjoint operator $\mathcal{L}_0^*$ and boundary conditions $\mathcal{B}_r^*[\psi] = 0$ so that the relation

$$\int_{-1}^{1} \mathcal{L}_0[\phi](x)\bar{\psi}(x)\,\mathrm{d}x = \int_{-1}^{1} \phi(x)\overline{\mathcal{L}_0^*[\psi]}(x)\,\mathrm{d}x$$

holds for all $2q$-times continuously differentiable, complex-valued functions ϕ,ψ, where ϕ satisfies the boundary conditions $\mathcal{B}_r[\phi] = 0$ and ψ satisfies the dual boundary conditions $\mathcal{B}_r^*[\psi] = 0$. Here $\bar{z}$ denotes the complex conjugate of $z \in \mathbb{C}$. We say that an operator is *self-adjoint* if $\mathcal{L}_0 = \mathcal{L}_0^*$ and $\mathcal{B}_r = \mathcal{B}_r^*$, $r = 1,\dots,2q$.

It is well known that if μ is an eigenvalue of $\mathcal{L}_0$ with the aforementioned boundary conditions, then $\bar{\mu}$ is an eigenvalue of the adjoint problem. Moreover, if ϕ and ψ are eigenfunctions of $\mathcal{L}_0$ and $\mathcal{L}_0^*$ respectively, with corresponding eigenvalues μ and ν, then ϕ and ψ are orthogonal unless $\mu = \bar{\nu}$.

Under some mild assumptions, the spectrum of $\mathcal{L}_0$ is countable with eigenvalues $\{\mu_n\}$ and eigenfunctions $\{\phi_n\}$ [127]. If $\{\psi_n\}$ is the corresponding set of eigenfunctions of the adjoint, then $(\phi_n, \psi_m) = \delta_{n,m}$ (after appropriate renormalisation), and we refer to the pair $\{\phi_n, \psi_n\}$ as a *biorthogonal pair* of eigenfunctions. This biorthogonality signals that a function $f \in \mathrm{L}^2(-1,1)$ can be expanded in the formal series

$$f(x) \sim \sum_{n=1}^{\infty} (f, \psi_n)\, \phi_n(x). \tag{3.9}$$

Note that we do not make any assumptions regarding the convergence of the right-hand side of (3.9) at this point.

Our interest lies with the case of the polyharmonic operator $\mathcal{L}_0 = (-1)^q \frac{\mathrm{d}^{2q}}{\mathrm{d}x^{2q}}$. It is evident that, when prescribed either Neumann $\phi^{(q+r)}(\pm 1) = 0$, $r = 0, \dots, q-1$, or Dirichlet $\phi^{(r)}(\pm 1) = 0$ boundary conditions, this operator is self-adjoint. Nonetheless, to describe the duality enjoyed by polyharmonic–Neumann expansions properly, we first need to catalogue the nature of the polyharmonic operator under a variety of other boundary conditions:

Lemma 3.2. *Suppose that $p = 1, \dots, q-1$ and that the polyharmonic operator $\mathcal{L}_0 = (-1)^q \frac{\mathrm{d}^{2q}}{\mathrm{d}x^{2q}}$ is equipped with boundary conditions*

$$\phi^{(q+r-p)}(\pm 1) = 0, \quad r = 0, \dots, q-1. \tag{3.10}$$

Then, the adjoint operator $\mathcal{L}_0^* = \mathcal{L}_0 = (-1)^q \frac{\mathrm{d}^{2q}}{\mathrm{d}x^{2q}}$ and the adjoint boundary conditions are

$$\begin{aligned} \psi^{(r)}(\pm 1) = 0, &\quad r = 0, \dots, p-1, \\ \psi^{(2q-r-1)}(\pm 1) = 0, &\quad r = 0, \dots, q-p-1. \end{aligned} \tag{3.11}$$

In particular, the corresponding pair of polyharmonic eigenfunctions subject to boundary conditions (3.10) and (3.11) are biorthogonal.

Proof. We have

$$\int_{-1}^{1} \mathcal{L}_0[\phi](x)\bar{\psi}(x)\,\mathrm{d}x = (-1)^q \sum_{r=0}^{2q-1} (-1)^{r+1} \phi^{(r)}(x)\bar{\psi}^{(2q-r-1)}(x)\Big|_{-1}^{1} + \int_{-1}^{1} \phi(x)\overline{\mathcal{L}_0[\psi]}(x)\,\mathrm{d}x.$$

If ϕ satisfies boundary conditions (3.10), then this sum vanishes for all ψ precisely when ψ obeys the conditions (3.11). □

Before detailing the duality exhibited by polyharmonic–Neumann eigenfunctions, it is informative to describe the nature of the zero eigenvalue of the polyharmonic operator equipped with boundary conditions (3.10) or (3.11). Recall that the polyharmonic–Neumann operator has a zero eigenvalue of multiplicity q. The corresponding eigenspace is $\mathbb{P}_{q-1}$, the space of polynomials of degree $q-1$, and we write $\{\phi_{0,n} : n = 0, \dots, q-1\}$ for the orthonormal basis of polynomials of this space (note that $\phi_{0,n} = (n+\frac{1}{2})^{\frac{1}{2}} P_n$, where P_n is the n^{th} Legendre polynomial). Trivial calculations verify that the polyharmonic operator with boundary conditions (3.10) or (3.11) has a $(q-p)$-fold zero eigenvalue. The corresponding eigenspaces are $\mathbb{P}_{q-p-1}$ and

$$\left\{ g \in \mathbb{P}_{q+p-1} : g^{(r)}(\pm 1) = 0,\ r = 0, \dots, p-1 \right\},$$

respectively.

With this to hand, we are now in a position to prove the main result of this section:

Theorem 3.3. *If we apply the differentiation operator* $\frac{\mathrm{d}^p}{\mathrm{d}x^p}$, $p = 1, \ldots, q-1$, *to the set of polyharmonic–Neumann eigenfunctions, we obtain, up to scalar multiples, the set of polyharmonic eigenfunctions that satisfy the boundary conditions (3.10). Furthermore, for* $f \in \mathrm{H}^p(-1,1)$, $(\mathcal{F}_N[f])^{(p)}$ *is the truncated expansion of* $f^{(p)}$ *in the biorthogonal pair of polyharmonic eigenfunctions corresponding to boundary conditions (3.10) and (3.11).*

Proof. The first result is trivial. For the second, we proceed exactly as in Lemma 3.1. Suppose that ϕ_n is the n^{th} polyharmonic–Neumann eigenfunction with eigenvalue $\mu_n = \alpha_n^{2q} \neq 0$. Let $\phi_n^{(p)} = c_n \psi_n$ and $\phi_n^{(2q-p)} = d_n \chi_n$ for constants c_n, d_n where $\{\psi_n, \chi_n\}$ is the biorthogonal pair corresponding to boundary conditions (3.10) and (3.11). Assume that such eigenfunctions are normalised so that $(\psi_n, \chi_m) = \delta_{n,m}$. Setting $r = p$, $\phi = \phi_m$ and $f = \phi_n$ in (3.8) immediately gives

$$\delta_{n,m} = \frac{(-1)^{q+p}}{\alpha_m^{2q}} c_n d_m \int_{-1}^{1} \psi_n(x) \chi_m(x)\, \mathrm{d}x.$$

In particular, $c_n d_n = (-1)^{q+p} \alpha_n^{2q}$. Moreover, using (3.8) once more,

$$\hat{f}_n \phi_n^{(p)}(x) = \frac{(-1)^{q+p}}{\alpha_n^{2q}} c_n d_n \int_{-1}^{1} f^{(p)}(x) \chi_n(x)\, \mathrm{d}x \psi_n(x) = \left(f^{(p)}, \chi_n \right) \psi_n(x).$$

It follows that

$$\frac{\mathrm{d}^p}{\mathrm{d}x^p} \sum_{n=1}^{N} \hat{f}_n \phi_n(x) = \sum_{n=1}^{N} \left(f^{(p)}, \chi_n \right) \psi_n(x), \tag{3.12}$$

for any $N \in \mathbb{N}$. To complete the proof, we need to assess the component of $\mathcal{F}_N[f]$ corresponding to the q-fold zero eigenvalue. To this end, suppose that we write $\{\psi_{0,n} : n = 0, \ldots, q-p-1\}$ and $\{\chi_{0,n} : n = 0, \ldots, q-p-1\}$ for the sets of polyharmonic eigenfunctions subject to boundary conditions (3.10) and (3.11) respectively and corresponding to the zero eigenvalue. To prove the full result, it suffices to show that

$$\frac{\mathrm{d}^p}{\mathrm{d}x^p} \sum_{n=0}^{q-1} \hat{f}_{0,n} \phi_{0,n}(x) = \sum_{n=0}^{q-p-1} \left(f^{(p)}, \chi_{0,n} \right) \psi_{0,n}(x). \tag{3.13}$$

Since $\{\psi_{0,n}\}$ is a basis for $\mathbb{P}_{q-p-1}$, it follows that

$$\frac{\mathrm{d}^p}{\mathrm{d}x^p} \sum_{n=0}^{q-1} \hat{f}_{0,n} \phi_{0,n}(x) = \sum_{n=0}^{q-p-1} a_n \psi_{0,n}(x),$$

for some values $a_n \in \mathbb{R}$. Due to the biorthogonality relation $(\psi_{0,n}, \chi_{0,m}) = \delta_{n,m}$, we have

$$a_n = \left(\frac{\mathrm{d}^p}{\mathrm{d}x^p} \sum_{m=0}^{q-1} \hat{f}_{0,m} \phi_{0,m}, \chi_{0,n} \right).$$

In view of (3.12) and the fact that $(\psi_n, \chi_{0,m}) = 0$, we may write, for any $N \in \mathbb{N}$,

$$a_n = \left(\frac{\mathrm{d}^p}{\mathrm{d}x^p} \left\{ \sum_{m=0}^{q-1} \hat{f}_{0,m} \phi_{0,m} + \sum_{m=1}^{N} \hat{f}_n \phi_m \right\}, \chi_{0,n} \right).$$

We now note that, since $\chi_{0,n}^{(r)}(\pm 1) = 0$ for $r = 0, \ldots, p-1$, integration by parts p times gives the relation

$$\left(g^{(p)}, \chi_{0,n}\right) = \left(g, \chi_{0,n}^{(p)}\right) \tag{3.14}$$

for any function $g \in \mathrm{H}^p(-1,1)$. In particular,

$$a_n = \left(\sum_{m=0}^{q-1} \hat{f}_{0,m}\phi_{0,m} + \sum_{m=1}^{N} \hat{f}_n \phi_m, \chi_{0,n}^{(p)}\right) = \left(\mathcal{F}_N[f], \chi_{0,n}^{(p)}\right).$$

Since N was arbitrary and $\mathcal{F}_N[f] \to f$ in the $\mathrm{L}^2(-1,1)$ norm, it follows that $a_n = (f, \chi_{0,n}^{(p)})$. An application of (3.14) now gives $a_n = (f^{(p)}, \chi_{0,n})$, hence verifying (3.13). □

The properties of the biorthogonal pairs of polyharmonic eigenfunctions introduced in this section are well understood within the general context of Birkhoff expansions [51]. Though certain standard results—in particular, $\mathrm{L}^2(-1,1)$ convergence of the expansion (3.9)—could be utilised in our study of polyharmonic expansions, we shall not do this. The reasons for this are twofold. First, such general results insufficiently describe the polyharmonic–Neumann case (as we shall consider in Section 3.3), and second, they do not easily generalise to the d-variate cube. Instead, we develop alternative, simpler techniques to tackle the polyharmonic–Neumann case directly. Incidentally, convergence of the expansion (3.9) will be a by-product of these results.

3.2.3 Construction of polyharmonic eigenfunctions

In this section, we briefly describe the systematic approach developed in [8] for the construction of polyharmonic eigenfunctions. Let ϕ be a polyharmonic–Neumann eigenfunction with eigenvalue $\mu = \alpha^{2q}$. We first note that

$$\phi(x) = \sum_{r=0}^{2q-1} c_r \mathrm{e}^{\lambda_r \alpha x},$$

where the values $\lambda_r \in \mathbb{C}$ satisfy $\lambda_r^{2q} = (-1)^q$, $r = 0, \ldots, 2q-1$ and the parameters $c_r \in \mathbb{C}$ are determined by the boundary conditions. Simplification of this expression requires separately addressing the two cases corresponding to even and odd q.

Even q

The values λ_r are roots of unity in this case, and the eigenfunction ϕ takes one of two possible forms $\phi^{[i]}$ which is even if $i = 0$ and odd if $i = 1$. These are given by

$$\begin{aligned}\phi^{[0]}(x) = &\sum_{r=0}^{\frac{q}{2}} c_r^{[0]} \cos\left(\alpha^{[0]} x \sin\frac{\pi r}{q}\right) \cosh\left(\alpha^{[0]} x \cos\frac{\pi r}{q}\right) \\ &+ \sum_{r=1}^{\frac{q}{2}-1} d_r^{[0]} \sin\left(\alpha^{[0]} x \sin\frac{\pi r}{q}\right) \sinh\left(\alpha^{[0]} x \cos\frac{\pi r}{q}\right),\end{aligned} \tag{3.15}$$

and

$$\phi^{[1]}(x) = \sum_{r=0}^{\frac{q}{2}-1} c_r^{[1]} \cos\left(\alpha^{[1]} x \sin\frac{\pi r}{q}\right) \sinh\left(\alpha^{[1]} x \cos\frac{\pi r}{q}\right) + \sum_{r=1}^{\frac{q}{2}} d_r^{[1]} \sin\left(\alpha^{[1]} x \sin\frac{\pi r}{q}\right) \cosh\left(\alpha^{[1]} x \cos\frac{\pi r}{q}\right), \tag{3.16}$$

respectively. The parameters $c_r^{[i]}$, $d_r^{[i]}$ and $\alpha^{[i]}$ are specified by enforcing the boundary conditions, which results in an algebraic $q \times q$ eigenproblem.

Odd q

The odd case is addressed in an identical manner, resulting in

$$\phi^{[0]}(x) = \sum_{r=0}^{\frac{q-1}{2}} c_r^{[0]} \cos\left(\alpha^{[0]} x \sin\frac{\pi(r+\frac{1}{2})}{q}\right) \cosh\left(\alpha^{[0]} x \cos\frac{\pi(r+\frac{1}{2})}{q}\right) + \sum_{r=0}^{\frac{q-3}{2}} d_r^{[0]} \sin\left(\alpha^{[0]} x \sin\frac{\pi(r+\frac{1}{2})}{q}\right) \sinh\left(\alpha^{[0]} x \cos\frac{\pi(r+\frac{1}{2})}{q}\right),$$

and

$$\phi^{[1]}(x) = \sum_{r=0}^{\frac{q-3}{2}} c_r^{[1]} \cos\left(\alpha^{[1]} x \sin\frac{\pi(r+\frac{1}{2})}{q}\right) \sinh\left(\alpha^{[1]} x \cos\frac{\pi(r+\frac{1}{2})}{q}\right) + \sum_{r=0}^{\frac{q-1}{2}} d_r^{[1]} \sin\left(\alpha^{[1]} x \sin\frac{\pi(r+\frac{1}{2})}{q}\right) \cosh\left(\alpha^{[1]} x \cos\frac{\pi(r+\frac{1}{2})}{q}\right).$$

We conclude the following: for arbitrary $q \geq 1$, the eigenfunctions split into even and odd functions, and in each case they can be expressed as sums of products of trigonometric and hyperbolic functions.

The biharmonic ($q = 2$) case warrants further attention. It presents the first significant extension beyond the modified Fourier case, and highlights several features of general polyharmonic–Neumann expansions. In this setting, the eigenfunctions are given by

$$\phi_n^{[0]}(x) = \frac{1}{\sqrt{2}}\left(\frac{\cos\alpha_n^{[0]} x}{\cos\alpha_n^{[0]}} + \frac{\cosh\alpha_n^{[0]} x}{\cosh\alpha_n^{[0]}}\right), \quad \phi_n^{[1]}(x) = \frac{1}{\sqrt{2}}\left(\frac{\sin\alpha_n^{[1]} x}{\sin\alpha_n^{[1]}} + \frac{\sinh\alpha_n^{[1]} x}{\sinh\alpha_n^{[1]}}\right), \tag{3.17}$$

and the values $\alpha_n^{[i]}$, $i = 0, 1$, $n \in \mathbb{N}$, satisfy the nonlinear equations

$$\tanh\alpha_n^{[0]} + \tan\alpha_n^{[0]} = 0, \quad \tanh\alpha_n^{[1]} - \tan\alpha_n^{[1]} = 0.$$

These values lie in intervals of exponentially small width in n. In fact,

$$\alpha_n^{[0]} \in \left((n-\tfrac{1}{4})\pi, (n-\tfrac{1}{4})\pi + c\mathrm{e}^{-2(n-\frac{1}{4})\pi}\right), \quad \alpha_n^{[1]} \in \left((n+\tfrac{1}{4})\pi - c\mathrm{e}^{-2(n+\frac{1}{4})\pi}, (n+\tfrac{1}{4})\pi\right), \tag{3.18}$$

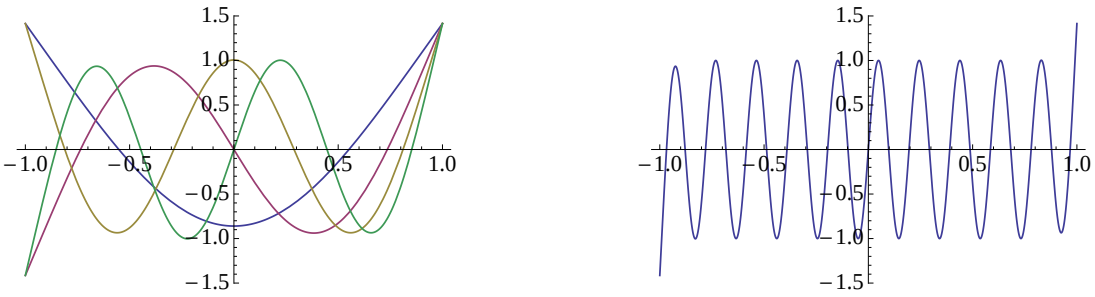


Figure 3.1: (left) the biharmonic eigenfunctions ϕ_n, $n = 1, 2, 3, 4$. (right) the function ϕ_{20}.

where $c = \frac{\cos 1 + \sin 1}{\sin 1}$. This establishes (3.4) in this setting. Likewise, the estimate (3.5) is also easily demonstrated after a brief consideration of (3.17).[4]

In Figure 3.1(a), we plot the first four biharmonic eigenfunctions. Herein another property of polyharmonic eigenfunctions is apparent: namely, the n^{th} eigenfunction has precisely n simple zeros in $(-1, 1)$, and the zeros interlace.[5] Simple arguments, along similar lines to those already given, demonstrate that these observations are valid for all n when $q = 2$. Such behaviour is characteristic of Sturm–Liouville eigenfunctions [114]. Moreover, it is known to hold also for eigenfunctions corresponding to a wide variety of higher-order differential operators, including the polyharmonic operator under current consideration. This result is a by-product of the theory of n-widths [136, chpt. 3].

Figure 3.1(b) plots the eigenfunction ϕ_{20}. From this we surmise that the zeros of polyharmonic eigenfunctions are, in addition, asymptotically uniformly distributed in $[-1, 1]$, a result we shall establish in the sequel. This figure also illustrates that the eigenfunction ϕ_n behaves like a regular oscillator away from the endpoints $x = \pm 1$, as predicted by (3.5).

3.2.4 Computation of polyharmonic–Neumann eigenvalues

The eigenfunctions ϕ_n can be constructed in a systematic manner. Provided the values α_n have been computed, the coefficients of the eigenfunctions can be easily found by solving an algebraic eigenproblem.

It remains to scrutinise the computation of such values. However, as we hypothesised in (3.4) and will prove in the forthcoming section, such values lie in intervals of exponentially small width. For this reason, computation can be carried out extremely easily by means of Newton–Raphson iterations. Moreover, for even moderate n, we may use the approximation $\alpha_n \approx \frac{1}{4}(2n + q - 1)\pi$ instead.

In Table 3.1, we demonstrate the effectiveness of this algorithm for computing the values α_n. For $q = 2, 3, 4$ no more than 4 iterations are required to obtain machine precision. Furthermore, for $n \geq 16$, the approximation $\alpha_n \approx \frac{1}{4}(2n + q - 1)$ can be used without resorting to any iterations at all. We note in passing that, for $q = 3$, the values $\alpha_{2n-1} = n\pi$ are known explicitly [8].

[4] Note that, to relate (3.18) to the general case (3.4), we must reorder the eigenvalues $\alpha_{2n-1} = \alpha_n^{[0]}$ and $\alpha_{2n} = \alpha_n^{[1]}$. Correspondingly, we reorder $\phi_{2n-1} = \phi_n^{[0]}$ and $\phi_{2n} = \phi_n^{[1]}$.

[5] In fact, ϕ_n appears to have $n + 1$ simple zeros in this figure. However, we need to augment the basis by $\phi_{0,0}(x) = \frac{1}{\sqrt{2}}$ and $\phi_{0,1}(x) = \frac{\sqrt{3}}{\sqrt{2}}x$ having 0 and 1 zeros respectively.

	n	1	2	3	4	5	10	15	20	25	30
$q=2$	e_n	2.43	4.00	5.16	6.99	8.44	15.5	22.5	29.5	36.4	43.3
	a_n	3	3	2	2	2	1	0	0	0	0
$q=3$	e_n	—	3.62	—	6.20	—	13.6	—	25.7	—	37.7
	a_n	0	3	0	2	0	1	0	0	0	0
$q=4$	e_n	2.35	4.63	4.42	5.44	6.97	11.6	16.8	21.5	26.5	31.4
	a_n	4	3	3	2	2	1	1	0	0	0

Table 3.1: Numerical computation of α_n for $q = 2, 3, 4$. The value $e_n = -\log_{10}\left(|\alpha_n - \frac{1}{4}(2n+q-1)|/\alpha_n\right)$ measures the number of significant digits (a dash indicates where $\alpha_n = \frac{1}{4}(2n+q-1)$ exactly) and a_n is the number of Newton–Raphson iterations required to obtain machine epsilon.

To connect this discussion to the narrative of Section 3.1.1, we remark that, by choosing both the simplest operator and boundary conditions, we have greatly aided the task of computing the values α_n. If we were to choose a basis of eigenfunctions for which the n^{th} value α_n is known to within only $\mathcal{O}\left(n^{-1}\right)$ accuracy (as is the case for an overwhelming number of operators and boundary conditions), then computation would be considerably more complicated.

Two further remarks regarding practical issues are worthy of mention. First, as q increases, so does the computational cost of constructing and evaluating the eigenfunctions ϕ_n. Moreover, it becomes extremely cumbersome to derive analytic expressions for the coefficients $c_r^{[i]}$, $d_r^{[i]}$ of such eigenfunctions. For $q = 4$, we resorted to a symbolic algebra package for this task. Second, since the eigenfunctions involve increasing numbers of hyperbolic functions for large q, there is increasing susceptibility to round-off error in calculations. As a result, it appears inadvisable to use values of q much beyond $q = 4$. Regardless, the remainder of this chapter will furnish analysis of the general case $q \geq 1$.

3.3 Asymptotic character of polyharmonic–Neumann eigenvalues and eigenfunctions

The aim of this section is to establish the estimates (3.4) and (3.5). As stated, similar estimates with only $\mathcal{O}\left(n^{-1}\right)$ error terms are well known for Birkhoff expansions [51, 127]. Yet, to the best of our knowledge, estimates for the polyharmonic case with exponentially small remainders do not currently exist. Most likely, this is due to the fact that such estimates are only valid under rather specific conditions, a point we discuss further in Section 3.3.3.

Proofs in this section will follow along the same lines as those given in [127]. However, the greatly simplified nature of the linear operator and boundary conditions allows for a more straightforward argument, and in turn, facilitates more precise results. For ease of presentation, we work predominantly on the interval $[0, 1]$, as opposed to $[-1, 1]$, in this section.

3.3.1 Polyharmonic–Neumann eigenvalues

Consider an eigenfunction ϕ with eigenvalue $\mu = \alpha^{2q} \neq 0$. We have

$$(-1)^q \phi^{(2q)}(x) = \alpha^{2q}\phi(x), \quad \phi^{(q)}(0) = \ldots = \phi^{(2q-1)}(0) = \phi^{(q)}(1) = \ldots = \phi^{(2q-1)}(1) = 0.$$

Write $\phi(x) = \sum_{s=0}^{2q-1} c_s \mathrm{e}^{\alpha\lambda_s x}$, where $\lambda_0, \ldots, \lambda_{2q-1}$ are the solutions of $\lambda^{2q} = (-1)^q$ and $c_0, \ldots, c_{2q-1} \in \mathbb{C}$ are constants to be specified. Substituting the boundary conditions yields the system of equations

$$\sum_{s=0}^{2q-1} c_s(\alpha\lambda_s)^{r+q} = \sum_{s=0}^{2q-1} c_s(\alpha\lambda_s)^{r+q}\mathrm{e}^{\alpha\lambda_s} = 0, \quad r = 0, ..., q-1.$$

Hence the values α are the solutions of the equation $g(\alpha) = 0$, where

$$g(\alpha) = \det \begin{pmatrix} \mathrm{e}^{\alpha\lambda_0} & \mathrm{e}^{\alpha\lambda_1} & \cdots & \mathrm{e}^{\alpha\lambda_{2q-1}} \\ \lambda_0\mathrm{e}^{\alpha\lambda_0} & \lambda_1\mathrm{e}^{\alpha\lambda_1} & \cdots & \lambda_{2q-1}\mathrm{e}^{\alpha\lambda_{2q-1}} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_0^{q-1}\mathrm{e}^{\alpha\lambda_0} & \lambda_1^{q-1}\mathrm{e}^{\alpha\lambda_1} & \cdots & \lambda_{2q-1}^{q-1}\mathrm{e}^{\alpha\lambda_{2q-1}} \\ 1 & 1 & \cdots & 1 \\ \lambda_0 & \lambda_1 & \cdots & \lambda_{2q-1} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_0^{q-1} & \lambda_1^{q-1} & \cdots & \lambda_{2q-1}^{q-1} \end{pmatrix}. \tag{3.19}$$

Using Cramer's rule we obtain

$$g(\alpha) = \sum_{\sigma\in S_{2q-1}} \operatorname{sgn}(\sigma)\mathrm{e}^{\alpha\sum_{r=0}^{q-1}\lambda_{\sigma(r)}} \prod_{r=0}^{q-1} \left(\lambda_{\sigma(r)}\lambda_{\sigma(q+r)}\right)^r, \tag{3.20}$$

where S_{2q-1} is the set of permutations of the numbers $\{0, 1, \ldots, 2q-1\}$ and $\operatorname{sgn}(\sigma)$ takes value $+1$ if σ is an even permutation and -1 otherwise.

To analyse the asymptotic behaviour $\alpha \to \infty$, we must first scrutinise the sum $\sum_{r=0}^{q-1}\lambda_{\sigma(r)}$ (note that α is real and positive). To do so, we introduce the following ordering on the values $\lambda_0, \ldots, \lambda_{2q-1}$. We define $\lambda_0 = -\mathrm{i}$ and $\lambda_r = \lambda_0\lambda^r$, where $\lambda = \mathrm{e}^{\frac{\mathrm{i}\pi}{q}}$. In particular, $\lambda_q = \mathrm{i}$. For such an ordering, observe that $\operatorname{Re}\lambda_r \geq 0$ for $r = 0, \ldots, q$, and $\operatorname{Re}\lambda_r < 0$ otherwise.

We now require the following lemma:

Lemma 3.4. *We have* $\max_{\sigma\in S_{2q-1}} \operatorname{Re}\sum_{r=0}^{q-1}\lambda_{\sigma(r)} = \cot\frac{\pi}{2q} = \theta_q$. *This maximum is attained precisely when* $\sigma \in T_{2q-1} = U_q \cup V_q$, *where*

$$U_q = \{\sigma \in S_{2q-1} : \{\sigma(r) : r = 0, \ldots, q-1\} = \{0, \ldots, q-1\}\},$$
$$V_q = \{\sigma \in S_{2q-1} : \{\sigma(r) : r = 0, \ldots, q-1\} = \{1, \ldots, q\}\}.$$

Moreover, the sum $\sum_{r=0}^{q-1}\lambda_{\sigma(r)} = \theta_q - \mathrm{i}$ *for* $\sigma \in U_q$ *and* $\sum_{r=0}^{q-1}\lambda_{\sigma(r)} = \theta_q + \mathrm{i}$ *for* $\sigma \in V_q$. *Conversely, if* $\sigma \notin T_{2q-1}$ *then* $\operatorname{Re}\sum_{r=0}^{q-1}\lambda_{\sigma(r)} \leq \theta_q - \gamma_q$, *where* $\gamma_q = \sin\frac{\pi}{q}$.

Proof. It is clear from the ordering of $\lambda_0, \dots, \lambda_{2q-1}$ that the maximum value is attained only for $\sigma \in T_{2q-1}$. Furthermore

$$\sum_{r=0}^{q-1} \lambda_r = \lambda_0 \sum_{r=0}^{q-1} \lambda^r = \frac{2\mathrm{i}}{\mathrm{e}^{\frac{\mathrm{i}\pi}{q}} - 1} = \theta_q - \mathrm{i},$$

and $\sum_{r=1}^{q} \lambda_r = 2\mathrm{i} + \sum_{r=0}^{q-1} \lambda_r$. For the final part, we merely note that $|\operatorname{Re} \lambda_r| \geq \operatorname{Re} \lambda_1 = \gamma_q$ for $r \neq 0, q$. □

With this lemma to hand, we may provide an estimate for the function g:

Lemma 3.5. *The function $g(\alpha)$ defined by (3.19) satisfies*

$$g(\alpha) = \mathrm{e}^{\theta_q \alpha} \det A_0 \det A_1 \left(\mathrm{e}^{-\mathrm{i}\alpha} - \mathrm{e}^{-\mathrm{i}\pi(q-1)} \mathrm{e}^{\mathrm{i}\alpha} \right) + \mathcal{O}\left(\mathrm{e}^{(\theta_q - \gamma_q)\alpha} \right), \quad \alpha \to \infty,$$

where $A_0, A_1 \in \mathbb{C}^{q \times q}$ have $(r,s)^{\text{th}}$ entries λ_s^r and λ_{q+s}^r respectively for $r, s = 0, \dots, q-1$.

Note that both A_0 and A_1 are Vandermonde matrices, hence their corresponding determinants are known analytically [66]. However, since these exact values are of little relevance to the present discussion, we shall not pursue this further.

Proof of Lemma 3.5. Applying the result of Lemma 3.4 to (3.20) gives

$$\begin{aligned} g(\alpha) =& \mathrm{e}^{(\theta_q - \mathrm{i})\alpha} \sum_{\sigma \in U_q} \operatorname{sgn}(\sigma) \prod_{r=0}^{q-1} \left(\lambda_{\sigma(r)} \lambda_{\sigma(q+r)} \right)^r \\ &+ \mathrm{e}^{(\theta_q + \mathrm{i})\alpha} \sum_{\sigma \in V_q} \operatorname{sgn}(\sigma) \prod_{r=0}^{q-1} \left(\lambda_{\sigma(r)} \lambda_{\sigma(q+r)} \right)^r + \mathcal{O}\left(\mathrm{e}^{(\theta_q - \gamma_q)\alpha} \right), \quad \alpha \to \infty. \end{aligned} \tag{3.21}$$

If $\sigma \in U_q$, we may write

$$\sigma(r) = \begin{cases} \sigma'(r) & r = 0, \dots, q-1 \\ q + \sigma''(r-q) & r = q, \dots, 2q-1, \end{cases}$$

where $\sigma', \sigma'' \in S_{q-1}$. In particular, $\operatorname{sgn}(\sigma) = \operatorname{sgn}(\sigma')\operatorname{sgn}(\sigma'')$. Hence

$$\sum_{\sigma \in U_q} \operatorname{sgn}(\sigma) \prod_{r=0}^{q-1} \left(\lambda_{\sigma(r)} \lambda_{\sigma(q+r)} \right)^r = \sum_{\sigma', \sigma'' \in S_{q-1}} \operatorname{sgn}(\sigma')\operatorname{sgn}(\sigma'') \prod_{r=0}^{q-1} \left(\lambda_{\sigma'(r)} \lambda_{q+\sigma''(r)} \right)^r$$

and this is precisely $\det A_0 \det A_1$. Similar arguments can be applied to $\sigma \in V_q$. Noting that $\lambda_{2q} = \lambda_0$, we write

$$\sigma(r) = \begin{cases} 1 + \sigma'(r) & r = 0, \dots, q-1 \\ q + 1 + \sigma''(r-q) & r = q, \dots, 2q-1. \end{cases}$$

In this case $\operatorname{sgn}(\sigma) = -\operatorname{sgn}(\sigma')\operatorname{sgn}(\sigma'')$, hence

$$\sum_{\sigma \in V_q} \operatorname{sgn}(\sigma) \prod_{r=0}^{q-1} \left(\lambda_{\sigma(r)} \lambda_{\sigma(q+r)} \right)^r = -\det A_2 \det A_3,$$

where $A_2, A_3 \in \mathbb{C}^{q\times q}$ have $(r,s)^{\text{th}}$ entries λ_{1+s}^r and λ_{q+1+s}^r respectively. Observe that $A_2 = DA_0$, $A_3 = DA_1$, where $D \in \mathbb{C}^{q\times q}$ is the diagonal matrix with r^{th} entry λ^r. Hence

$$\det A_2 \det A_3 = (\det D)^2 \det A_0 \det A_1 = \lambda^{q(q-1)} \det A_0 \det A_1 = \mathrm{e}^{-\mathrm{i}\pi(q-1)} \det A_0 \det A_1,$$

Substituting this expression into (3.21) now completes the proof. □

We are now in a position to establish the key result of this section:

Theorem 3.6. *Suppose that $\mu_n = \alpha_n^{2q}$, $n \in \mathbb{N}$, is the n^{th} eigenvalue of the polyharmonic operator subject to homogeneous Neumann boundary conditions on $[0,1]$. Then*

$$\alpha_n = \frac{1}{2}(2n+q-1)\pi + \mathcal{O}\left(\mathrm{e}^{-n\pi\gamma_q}\right), \quad n \to \infty.$$

Proof. For an eigenvalue $\mu = \alpha^{2q}$ we have $g(\alpha) = 0$. Hence, $\mathrm{e}^{2\mathrm{i}\alpha} = \mathrm{e}^{i\pi(q-1)} + \mathcal{O}\left(\mathrm{e}^{-\gamma_q\alpha}\right)$. □

Mapping this result to $[-1,1]$ divides the eigenvalue by 2, thus verifying (3.4). Note that, when $q = 2$, this gives $\alpha_n = \frac{1}{4}(2n+1)\pi + \mathcal{O}(\mathrm{e}^{-\frac{1}{2}n\pi\gamma_2})$. Relabelling n by $2n-1$ or $2n$, we obtain the asymptotic estimates $n \pm \frac{1}{4}$. This corresponds precisely to the known result for biharmonic eigenvalues (see Section 3.2.3).

3.3.2 Polyharmonic–Neumann eigenfunctions

Next, we address the asymptotic nature of the eigenfunctions. We commence by noting that an eigenfunction ϕ corresponding to eigenvalue $\mu = \alpha^{2q} \neq 0$ can be written as

$$\phi(x) = \det \begin{pmatrix} \mathrm{e}^{\alpha\lambda_0 x} & \mathrm{e}^{\alpha\lambda_1 x} & \cdots & \mathrm{e}^{\alpha\lambda_{2q-1} x} \\ \lambda_0^q \mathrm{e}^{\alpha\lambda_0} & \lambda_1^q \mathrm{e}^{\alpha\lambda_1} & \cdots & \lambda_{2q-1}^q \mathrm{e}^{\alpha\lambda_{2q-1}} \\ \lambda_0^{q+1} \mathrm{e}^{\alpha\lambda_0} & \lambda_1^{q+1} \mathrm{e}^{\alpha\lambda_1} & \cdots & \lambda_{2q-1}^{q+1} \mathrm{e}^{\alpha\lambda_{2q-1}} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_0^{2q-1} \mathrm{e}^{\alpha\lambda_0} & \lambda_1^{2q-1} \mathrm{e}^{\alpha\lambda_1} & \cdots & \lambda_{2q-1}^{2q-1} \mathrm{e}^{\alpha\lambda_{2q-1}} \\ \lambda_0^q & \lambda_1^q & \cdots & \lambda_{2q-1}^q \\ \lambda_0^{q+1} & \lambda_1^{q+1} & \cdots & \lambda_{2q-1}^{q+1} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_0^{2q-2} & \lambda_1^{2q-2} & \cdots & \lambda_{2q-1}^{2q-2} \end{pmatrix} = \sum_{s=0}^{2q-1} \mathrm{e}^{\alpha\lambda_s x} (-1)^s \det A^{[s]},$$

where $A^{[s]}$ is the corresponding minor

$$A^{[s]} = \begin{pmatrix} \lambda_0^q \mathrm{e}^{\alpha\lambda_0} & \cdots & \lambda_{s-1}^q \mathrm{e}^{\alpha\lambda_{s-1}} & \lambda_{s+1}^q \mathrm{e}^{\alpha\lambda_{s+1}} & \cdots & \lambda_{2q-1}^q \mathrm{e}^{\alpha\lambda_{2q-1}} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \lambda_0^{2q-1} \mathrm{e}^{\alpha\lambda_0} & \cdots & \lambda_{s-1}^{2q-1} \mathrm{e}^{\alpha\lambda_{s-1}} & \lambda_{s+1}^{2q-1} \mathrm{e}^{\alpha\lambda_{s+1}} & \cdots & \lambda_{2q-1}^{2q-1} \mathrm{e}^{\alpha\lambda_{2q-1}} \\ \lambda_0^q & \cdots & \lambda_{s-1}^q & \lambda_{s+1}^q & \cdots & \lambda_{2q-1}^q \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \lambda_0^{2q-2} & \cdots & \lambda_{s-1}^{2q-2} & \lambda_{s+1}^{2q-2} & \cdots & \lambda_{2q-1}^{2q-2} \end{pmatrix}.$$

We first need to assess the asymptotic behaviour of such values:

Lemma 3.7. *We have*

$$(-1)^s \det A^{[s]} = \begin{cases} c_s \mathrm{e}^{(\theta_q - \lambda_s)\alpha} + \mathcal{O}\left(\mathrm{e}^{(\theta_q - \operatorname{Re}\lambda_s - \gamma_q)\alpha}\right) & s = 0, \dots, q \\ c_s \mathrm{e}^{(\theta_q - \mathrm{i})\alpha} + c'_s \mathrm{e}^{(\theta_q + \mathrm{i})\alpha} + \mathcal{O}\left(\mathrm{e}^{(\theta_q - \gamma_q)\alpha}\right) & s = q+1, \dots, 2q-1, \end{cases}$$

where the values c_s, $s = 0, \dots, 2q-1$, c'_s, $s = q, \dots, 2q-1$, *depend only on* $\lambda_0, \dots, \lambda_{2q-1}$, *and, in particular,*

$$c_0 = \mathrm{e}^{\frac{(q-1)}{2}\mathrm{i}\pi} c_q. \tag{3.22}$$

Proof. Observe first that

$$\det A^{[s]} = \sum_{\sigma \in S_{s,2q-1}} \operatorname{sgn}(\sigma) \mathrm{e}^{\alpha \sum_{r=0}^{q-1} \lambda_{\sigma(r)}} \prod_{r=0}^{q-1} \lambda_{\sigma(r)}^{r+q} \prod_{r=0}^{q-2} \lambda_{\sigma(r+q)}^{r+q}, \tag{3.23}$$

where $S_{s,2q-1}$ is the set of permutations $\sigma : \{0, \dots, 2q-2\} \to \{0, \dots, s-1, s+1, \dots, 2q-1\}$. We first consider the case $s = 0, \dots, q$. For such values, we have

$$\max_{\sigma \in S_{s,2q-1}} \operatorname{Re} \sum_{r=0}^{q-1} \lambda_{\sigma(r)} = \theta_q - \operatorname{Re}\lambda_s,$$

and this is attained precisely when $\{\sigma(r) : r = 0, \dots, q-1\} = \{0, \dots, s-1, s+1, \dots, q\}$. We write

$$\sigma(r) = \begin{cases} \sigma'(r) & r = 0, \dots, q-1 \\ q+1+\sigma''(r-q) & r = q, \dots, 2q-2, \end{cases}$$

for such σ, where $\sigma' \in S_{s,q}$ and $\sigma'' \in S_{q-2}$ is a permutation of $\{0, \dots, q-2\}$. Substituting this into (3.23) now gives

$$\begin{aligned} \det A^{[s]} &= \mathrm{e}^{(\theta_q - \lambda_s)\alpha} \sum_{\substack{\sigma' \in S_{s,q} \\ \sigma'' \in S_{q-2}}} \operatorname{sgn}(\sigma')\operatorname{sgn}(\sigma'') \prod_{r=0}^{q-1} \lambda_{\sigma'(r)}^{r+q} \prod_{r=0}^{q-2} \lambda_{\sigma''(r)+q+1}^{r+q} + \mathcal{O}\left(\mathrm{e}^{(\theta_q - \operatorname{Re}\lambda_s - \gamma_q)\alpha}\right) \\ &= \mathrm{e}^{(\theta_q - \lambda_s)\alpha} \det B \det C^{[s]} + \mathcal{O}\left(\mathrm{e}^{(\theta_q - \operatorname{Re}\lambda_s - \gamma_q)\alpha}\right), \end{aligned}$$

where $B \in \mathbb{C}^{(q-1)\times(q-1)}$ has $(r,s)^{\text{th}}$ entry λ_{s+q+1}^{r+q} and

$$C^{[s]} = \begin{pmatrix} \lambda_0^q & \cdots & \lambda_{s-1}^q & \lambda_{s+1}^q & \cdots & \lambda_q^q \\ \lambda_0^{q+1} & \cdots & \lambda_{s-1}^{q+1} & \lambda_{s+1}^{q+1} & \cdots & \lambda_q^{q+1} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \lambda_0^{2q-1} & \cdots & \lambda_{s-1}^{2q-1} & \lambda_{s+1}^{2q-1} & \cdots & \lambda_q^{2q-1} \end{pmatrix}.$$

This proves the result for $s = 0, \dots, q$. Note that $C^{[0]} = DC^{[q]}$, where $D \in \mathbb{C}^{q\times q}$ is the diagonal matrix with r^{th} entry λ^{r+q}. Therefore, $\det C^{[0]} = \lambda^{q^2 + \frac{1}{2}q(q-1)} \det C^{[q]}$, and this yields (3.22).

Next, we consider $s = q+1, \dots, 2q-1$. In this case

$$\max_{\sigma \in S_{s,2q-1}} \operatorname{Re} \sum_{r=0}^{q-1} \lambda_{\sigma(r)} = \theta_q$$

and this is attained when $\{\sigma(r) : r = 0, \dots, q-1\} = \{0, \dots, q-1\}$ or $\{1, \dots, q\}$. For the former, we write

$$\sigma(r) = \begin{cases} \sigma'(r) & r = 0, \dots, q-1 \\ \sigma''(r) & r = q, \dots, 2q-2, \end{cases}$$

where $\sigma' \in S_q$ and $\sigma'' : \{q, \dots, 2q-2\} \to \{q, \dots, s-1, s+1, \dots, 2q-1\}$. For the latter,

$$\sigma(r) = \begin{cases} \sigma'(r)+1 & r = 0, \dots, q-1 \\ \sigma''(r) & r = q, \dots, 2q-2, \end{cases}$$

where $\sigma'' : \{q, \dots, 2q-2\} \to \{q+1, \dots, s-1, s+1, \dots, 2q\}$. Hence

$$\det A^{[s]} = \mathrm{e}^{\theta_q \alpha} \left\{ \mathrm{e}^{-\mathrm{i}\alpha} \det A_0 \det D_s - \mathrm{e}^{\mathrm{i}\alpha} \det A_2 \det D_s' \right\} + \mathcal{O}\left(\mathrm{e}^{(\theta_q - \gamma_q)\alpha} \right),$$

where A_0, A_2 are the matrices of Lemma 3.5 and D_s and D_s' are given by

$$\begin{pmatrix} \lambda_q^q & \cdots & \lambda_{s-1}^q & \lambda_{s+1}^q & \cdots & \lambda_{2q-1}^q \\ \lambda_q^{q+1} & \cdots & \lambda_{s-1}^{q+1} & \lambda_{s+1}^{q+1} & \cdots & \lambda_{2q-1}^{q+1} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \lambda_q^{2q-1} & \cdots & \lambda_{s-1}^{2q-1} & \lambda_{s+1}^{2q-1} & \cdots & \lambda_{2q-1}^{2q-1} \end{pmatrix}, \quad \begin{pmatrix} \lambda_{q+1}^q & \cdots & \lambda_{s-1}^q & \lambda_{s+1}^q & \cdots & \lambda_{2q}^q \\ \lambda_{q+1}^{q+1} & \cdots & \lambda_{s-1}^{q+1} & \lambda_{s+1}^{q+1} & \cdots & \lambda_{2q}^{q+1} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \lambda_{q+1}^{2q-1} & \cdots & \lambda_{s-1}^{2q-1} & \lambda_{s+1}^{2q-1} & \cdots & \lambda_{2q}^{2q-1} \end{pmatrix}$$

respectively. This completes the proof. □

In view of this lemma, we now renormalise the eigenfunction ϕ by dividing by $\mathrm{e}^{\alpha(\theta_q + \mathrm{i})}$. It follows immediately from Lemmas 3.5 and 3.7 that

$$\begin{aligned} \phi(x) =& c_q \left\{ \mathrm{e}^{\mathrm{i}\alpha x} + \mathrm{e}^{-\frac{q-1}{2}\mathrm{i}\pi} \mathrm{e}^{-\mathrm{i}\alpha x} \right\} \\ &+ \sum_{s=1}^{q-1} c_s \mathrm{e}^{\lambda_s \alpha (x-1)} + \sum_{s=q+1}^{2q-1} c_s \mathrm{e}^{\alpha \lambda_s x} + \mathcal{O}\left(\max\{\mathrm{e}^{-\alpha\gamma_q x}, \mathrm{e}^{-\alpha\gamma_q(1-x)}\} \right). \end{aligned} \tag{3.24}$$

From this we deduce:

Lemma 3.8. *Suppose that ϕ is the polyharmonic–Neumann eigenfunction on $[0,1]$ with corresponding eigenvalue $\mu = \alpha^{2q} \neq 0$. Then*

$$\phi(x) = c_q \cos\left[\alpha x + \tfrac{1}{4}(q-1)\pi\right] + \mathcal{O}\left(\max\{\mathrm{e}^{-\alpha\gamma_q x}, \mathrm{e}^{-\alpha\gamma_q(1-x)}\} \right),$$

where c_q is independent of α and x.

Proof. Let $q = 4t + s$, $s = 0, 1, 2, 3$. It suffices to show that $\phi(x)$ is asymptotic to $\cos(\alpha x - \frac{\pi}{4})$, $\cos \alpha x$, $\cos(\alpha x + \frac{\pi}{4})$ or $\cos(\alpha x + \frac{\pi}{2})$ respectively. First, suppose that $q = 2l + 1$. Then $\mathrm{e}^{-\frac{q-1}{2}\mathrm{i}\pi} = (-1)^l$, and, from (3.24), we deduce that

$$\begin{aligned} \phi(x) &= c_q \left\{ \mathrm{e}^{\mathrm{i}\alpha x} + (-1)^l \mathrm{e}^{-\mathrm{i}\alpha x} \right\} + \mathcal{O}\left(\max\{\mathrm{e}^{-\alpha\gamma_q x}, \mathrm{e}^{-\alpha\gamma_q(1-x)}\} \right) \\ &= c_q (1 + (-1)^l) \cos \alpha x + \mathrm{i} c_q (1 + (-1)^{l+1}) \sin \alpha x + \mathcal{O}\left(\max\{\mathrm{e}^{-\alpha\gamma_q x}, \mathrm{e}^{-\alpha\gamma_q(1-x)}\} \right), \end{aligned}$$

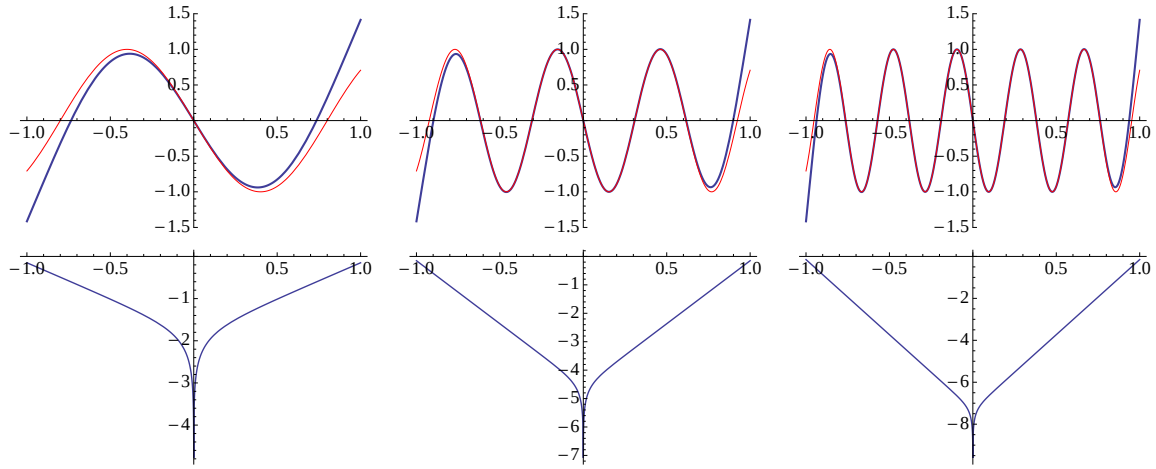


Figure 3.2: Top row: the biharmonic eigenfunctions ϕ_n (thicker line) and approximations $\sin \frac{1}{4}(2n+1)\pi x$ (thinner line) for $n = 2, 6, 10$ (left to right). Bottom row: the error $\log_{10} |\phi_n(x) - \sin \frac{1}{4}(2n+1)\pi x|$.

which completes the proof for $q = 4t+1$ and $q = 4t+3$. Now suppose that $q = 2l$. Then $\mathrm{e}^{-\frac{q-1}{2}\mathrm{i}\pi} = (-1)^l \mathrm{i}$. Since

$$\begin{aligned} \mathrm{e}^{\mathrm{i}\alpha x} + (-1)^l \mathrm{i} \mathrm{e}^{-\mathrm{i}\alpha x} &= \left(\mathrm{i} + (-1)^l\right)\left(\sin \alpha x + (-1)^l \cos \alpha x\right) \\ &= \sqrt{2}\left(\mathrm{i} + (-1)^l\right) \sin\left(\alpha x + (-1)^l \tfrac{\pi}{4}\right), \end{aligned}$$

we also obtain the result in this case. □

Theorem 3.9. *The n^{th} polyharmonic–Neumann eigenfunction on $[-1, 1]$ satisfies*

$$\phi_n(x) = \cos\left[\tfrac{1}{4}(2n+q-1)\pi x + \tfrac{1}{2}(n+q-1)\pi\right] + \mathcal{O}\left(\mathrm{e}^{-\frac{1}{2}\gamma_q(1-|x|)n\pi}\right).$$

Proof. This follows from Lemma 3.8 and the mapping $[0, 1] \to [-1, 1]$, $x \mapsto -1 + 2x$. □

This theorem verifies (3.5): polyharmonic–Neumann eigenfunctions are exponentially close to regular oscillators in $(-1, 1)$. In Figures 3.2 and 3.3 we exhibit this result for $q = 2, 3$. Once more, we observe the very rapid onset of the asymptotic behaviour away from the endpoints. Moreover, the straight lines in the logarithmic error plot highlight the nature of the error term in Theorem 3.9.

Theorem 3.9 also demonstrates the phenomenon of equiconvergence [123]. Polyharmonic–Neumann eigenfunctions are asymptotic to trigonometric functions in $(-1, 1)$. Hence, convergence of polyharmonic expansions can be studied with standard tools of Fourier analysis. This classical approach, however, is unsuitable for an accurate study of polyharmonic–Neumann expansions. As we shall prove through alternative means, classical Fourier series converge much more slowly than such expansions. Hence, relating polyharmonic–Neumann expansions to trigonometric series does little to illuminate their convergence.

To connect Theorem 3.9 to the explicit example of biharmonic eigenfunctions, we note that, when $q = 2$, this result gives

$$\phi_{2n-1}(x) = \cos(n - \tfrac{1}{4})\pi x + \mathcal{O}\left(\mathrm{e}^{-(1-|x|)n\pi\gamma_2}\right), \quad \phi_{2n}(x) = \sin(n + \tfrac{1}{4})\pi x + \mathcal{O}\left(\mathrm{e}^{-(1-|x|)n\pi\gamma_2}\right).$$

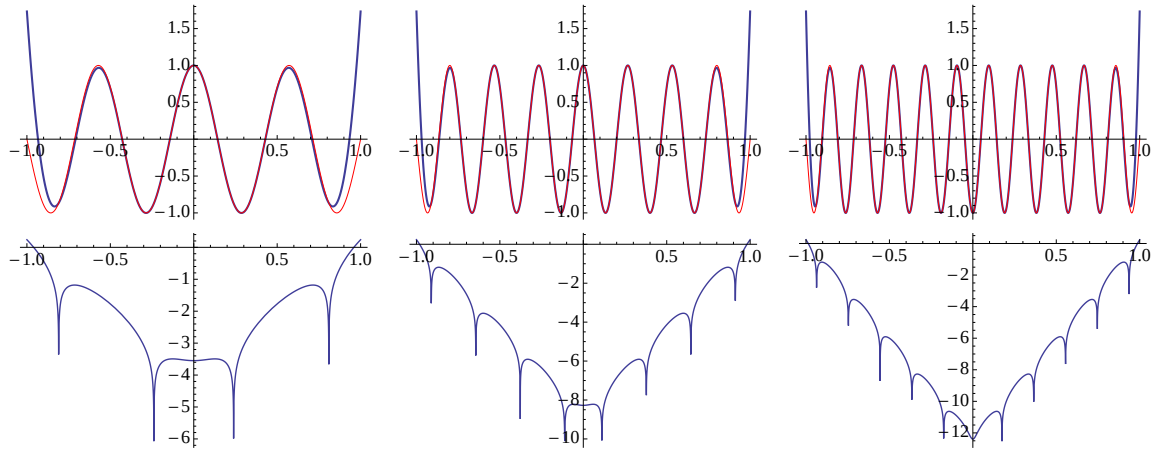


Figure 3.3: Top row: the triharmonic eigenfunctions ϕ_n (thicker line) and approximations $\cos\frac{1}{2}(n+1)\pi x$ (thinner line) for $n = 6, 14, 20$ (left to right). Bottom row: the error $\log_{10}|\phi_n(x) - \cos\frac{1}{2}(n+1)\pi x|$.

A brief comparison with (3.17) verifies this result in the biharmonic case. Note that ϕ_{2n-1}, an even function, corresponds to $\phi_n^{[0]}$. Likewise, ϕ_{2n} coincides with the odd function $\phi_n^{[1]}$.

Returning to the general case, and immediate consequence of Theorem 3.9 concerns the distribution of the zeros of the eigenfunctions ϕ_n in the limit $n \to \infty$. We have:

Corollary 3.10. *The zeros of ϕ_n are asymptotically uniformly distributed as $n \to \infty$.*

Proof. Suppose that $I = [a, b] \subseteq (-1, 1)$ is a closed interval. Let $Z_n(I)$ be the number of zeros of ϕ_n in I. It follows from Theorem 3.9 that $Z_n(I) = \frac{1}{2}(b-a)n + \mathcal{O}(1)$ as $n \to \infty$. Since ϕ_n has precisely $n + q$ simple zeros in $[-1, 1]$ (see [136]), it follows that the proportion of zeros in I is $\frac{1}{2}|I| + \mathcal{O}\left(n^{-1}\right)$ for large n (note that $|I| = 2$ for $I = [-1, 1]$, which explains the factor of $\frac{1}{2}$).

It now remains to show that the same result holds for intervals I containing at least one of the endpoints $x = \pm 1$. For this, we first note that ϕ_n is either even or odd: hence, it suffices to consider $I = [a, 1] \subseteq (-1, 1]$. If $a > 0$, then

$$Z(I) = \frac{1}{2}Z\left([-1, -a] \cup [a, 1]\right) = \frac{1}{2}\left\{Z([-1, 1]) - Z([-a, a])\right\} = \frac{1}{2}(1-a)n + \mathcal{O}(1),$$

as required. If $a < 0$, then $Z(I) = Z[-1, 1] - Z[-a, 1]$, and the result follows. □

To complete this section, we present several results concerning the growth of the derivatives of ϕ_n, both of which will be used later. The first concerns the uniform norm $\|\phi_n^{(r)}\|_\infty$. Intuitively, it feels correct that $\|\phi_n^{(r)}\|_\infty = \mathcal{O}(n^r)$ for large n. This is indeed the case:

Lemma 3.11. *Suppose that ϕ is the polyharmonic–Neumann eigenfunction corresponding to eigenvalue $\mu = \alpha^{2q}$. Then $\|\phi^{(r)}\|_\infty = \mathcal{O}(\alpha^r)$ for large α and any $r \in \mathbb{N}_0$.*

Proof. We work on the interval $[0, 1]$. Recalling (3.24) and the fact that $c_s = \mathcal{O}(1)$ independently of α, it follows that

$$|\phi^{(r)}(x)| \leq \alpha^r \left\{ \sum_{s=0}^{q-1} |c_s| \left| \mathrm{e}^{\lambda_s \alpha (x-1)} \right| + \sum_{s=q}^{2q-1} |c_s| \left| \mathrm{e}^{\lambda_s \alpha x} \right| \right\} \leq c\alpha^r,$$

as required. □

A particular consequence of this result is that $\|\phi\| \leq \|\phi\|_\infty = \mathcal{O}(1)$ for large α. Hence, the eigenfunction ϕ is $\mathrm{L}^2(-1,1)$ normalised independently of α. Next, we present our second result, which gives a pointwise estimate for the derivative $\phi^{(r)}$ evaluated at the endpoints $x = \pm 1$:

Lemma 3.12. *Suppose that ϕ is as in Lemma 3.11. Then $\phi^{(r)}(1) = c_r \alpha^r \mathrm{e}^{2\mathrm{i}\alpha} + \mathcal{O}(\alpha^r \mathrm{e}^{-\gamma_q \alpha})$, where c_r is a constant independent of α.*

Proof. Once more, we work on $[0,1]$. In view of Lemma 3.7 and the normalisation introduced previously, we have

$$\begin{aligned}\phi^{(r)}(1) &= \alpha^r \sum_{s=0}^{2q-1} \lambda_s^r \mathrm{e}^{\alpha\lambda_s} (-1)^s \det A^{[s]} \mathrm{e}^{-\alpha(\theta_q + \mathrm{i})} \\ &= \alpha^r \left\{ \mathrm{e}^{-\mathrm{i}\alpha} \sum_{s=0}^{q} \lambda_s^r c_s + \sum_{s=q+1}^{2q-1} \lambda_s^r \mathrm{e}^{\alpha\lambda_s} c_s \right\} + \mathcal{O}\left(\alpha^r \mathrm{e}^{-\gamma_q \alpha}\right) \\ &= \alpha^r \left\{ \mathrm{e}^{-\mathrm{i}\alpha} \sum_{s=0}^{q} \lambda_s^r c_s \right\} + \mathcal{O}\left(\alpha^r \mathrm{e}^{-\gamma_q \alpha}\right).\end{aligned}$$

In view of Lemma 3.5, $\mathrm{e}^{-\mathrm{i}\alpha} = (-1)^q \mathrm{e}^{\mathrm{i}\alpha} + \mathcal{O}(\mathrm{e}^{-\gamma_q \alpha})$. The result now follows immediately upon recalling that the mapping $[0,1] \to [-1,1]$ scales the value α by $\frac{1}{2}$. □

Note that $\phi(-1) = \pm\phi(1)$ with sign depending on n, where $\alpha = \alpha_n$. Hence, this lemma also establishes the growth of derivatives at the endpoint $x = -1$.

3.3.3 Other boundary conditions

The results proved in this section, detailing the existence of exponentially small remainder terms, are quite specific to the particular operator $\mathcal{L}_0$ and boundary conditions. A complete categorisation of such operators and boundary conditions is beyond the scope of this chapter. However, we now briefly indicate the type of restrictions necessary.

Certainly, such results are not, in general, valid once the operator has non-constant coefficients. For example, the Mathieu eigenvalues and eigenfunctions [120], defined by

$$\phi''(x) + (\mu - 2a\cos 2\pi x)\phi(x) = 0, \quad x \in [-1,1], \quad \phi'(\pm 1) = 0,$$

where a is constant, are indeed asymptotic to the corresponding Laplace–Neumann eigenvalues and eigenfunctions, but only with algebraically decaying remainder.[6] Furthermore, it need not be the case that a constant coefficient operator $\mathcal{L}_0$ with Neumann boundary conditions possesses such a property. For example, if $\mathcal{L}_0 = \partial_x^4 - a\partial_x^2$, where $a > 0$ is constant, then it is also easily seen that the remainder is only algebraically decaying.

Even within the simple realm of the polyharmonic operator, more complicated boundary conditions may destroy such estimates. For example, with $q = 2$ and the boundary conditions

$$u(\pm 1) + au'(\pm 1) = 0, \quad u''(\pm 1) = 0,$$

[6] In fact, a full asymptotic expansion in inverse powers of μ can be prescribed. Moreover, this asymptotic expansion turns out to be convergent for all $\mu \neq 0$ [120].

where $a \neq 0$, the values α_{2n-1} are the roots of the equation $\tanh\alpha - \tan\alpha = 2(a\alpha)^{-1}$, and therefore do not lie within intervals of exponentially small width.

The evidence of these examples indicates that the results of this section hold only in a very restrictive set of cases.[7] Thus far, we have not found another non-trivial operator—that is, an operator without explicitly known eigenvalues and eigenfunctions—with this property. The only generalisation that we have to date is for the polyharmonic operator with boundary conditions $u^{(l_r)}(\pm 1) = 0$, $r = 0, \ldots, q-1$, where $0 \leq l_0 < l_1 < \ldots < l_{q-1} \leq 2q-1$. Note that this includes the case of Neumann boundary conditions established in this section. The proof of this more general statement follows along identical lines to that given previously. We remark in passing that, though such boundary conditions yield eigenfunctions with the same properties as the Neumann case, the related expansion coefficients decay more slowly unless $l_r = q + r$, $r = 0, \ldots, q-1$, in which case we recover Neumann boundary conditions. Hence, such eigenfunctions, though of theoretical interest, are not best suited for practical computations.

3.4 Analysis of polyharmonic–Neumann expansions

Having established a number of key properties of polyharmonic–Neumann eigenfunctions, we now find ourselves in a position to provide a complete convergence analysis. For the remainder of this chapter, we shall use the notation c (or c_r, $c_{r,s}$) for a positive constant, independent of N and f (but dependent on r and s, where appropriate).

3.4.1 Density and convergence

As previously stated, the truncated expansion $\mathcal{F}_N[f]$ converges in the $\mathrm{L}^2(-1,1)$ norm. However, much like the Laplace ($q = 1$) case, a significantly stronger result holds concerning the space $\mathrm{H}^q(-1,1)$. To prove such a result, we must first establish that the bilinear form (3.6) provides an equivalent inner product on this space:

Lemma 3.13. *The bilinear form (3.6) is an inner product on* $\mathrm{H}^q(-1,1)$. *The associated norm* $|\!|\!| \cdot |\!|\!|_q$, *given by* $|\!|\!| f |\!|\!|_q^2 = \|f\|^2 + \|f^{(q)}\|^2$, *is equivalent to* $\|\cdot\|_q$.

Proof. This result follows immediately from the additive interpolation inequality $\|f^{(r)}\| \leq c_r \left(\|f\| + \|f^{(q)}\|\right)$, $r = 0, \ldots, q$, $\forall f \in \mathrm{H}^q(-1,1)$ [2]. □

The first result of this section, the univariate generalisation of Lemma 2.9 to arbitrary $q \geq 1$, is a direct consequence of Lemmas 3.1 and 3.13:

Lemma 3.14. *The set of polyharmonic–Neumann eigenfunctions is dense and orthogonal in* $\mathrm{H}^q(-1,1)$ *with respect to the inner product (3.6). In particular, for* $f \in \mathrm{H}^q(-1,1)$, $\mathcal{F}_N[f]$

[7] Of course, there are many trivial cases with explicitly known eigenfunctions. For example, the eigenfunctions of the polyharmonic operator with boundary conditions $\phi^{(2r+1)}(\pm 1) = 0$, $r = 0, \ldots, q-1$ are precisely the Laplace–Neumann eigenfunctions. Another example involves merely periodic boundary conditions $\phi^{(r)}(1) = \phi^{(r)}(-1)$, $r = 0, \ldots, 2q-1$ and any constant coefficient operator. In this case, eigenfunctions are just the Fourier basis functions. The proper classification of operators and boundary conditions to accurately reflect such examples requires deeper study, along the lines of [127].

converges to f in this norm and we have the characterisation

$$|\!|\!| f |\!|\!|_q^2 = \sum_{n=0}^{q-1} |\hat{f}_{0,n}|^2 + \sum_{n=1}^{\infty} (1+\mu_n)|\hat{f}_n|^2.$$

This lemma immediately presents the following question: for $f \in \mathrm{H}^r(-1,1)$, $r = 1, \ldots, q-1$, does $\mathcal{F}_N[f]$ converge to f in the $\mathrm{H}^r(-1,1)$ norm? This question can be answered almost immediately with standard results. We have established that $(\mathcal{F}_N[f])^{(r)}$ is the truncated expansion of $f^{(r)}$ in a certain biorthogonal pair of polyharmonic eigenfunctions (Theorem 3.3). Hence, the result follows directly from $\mathrm{L}^2(-1,1)$ convergence of this expansion—a well-known fact [51]. However, such results are not readily scalable to higher dimensions with the same level of generality pursued in Chapter 2 (i.e. arbitrary index sets). Therefore, we devote the remainder of this section to providing an alternative, arguably simpler, proof, which can be extended in this way (as we address in Section 3.5). Our method of proof will be based on the known results for $r = 0, q$—consequences of self-adjoint spectral theory—and interpolation therein for the intermediate values $r = 1, \ldots, q-1$.

To derive this result, we first need to establish a number of properties of the truncated expansion $\mathcal{F}_N[f]$ of a function f in biorthogonal pairs of polyharmonic eigenfunctions corresponding to boundary conditions (3.10) and (3.11). Namely, we shall prove a Bessel inequality for such expansions, $\|\mathcal{F}_N[f]\| \leq c\|f\|$, and also demonstrate that the sequence $\{(f,\psi_n)\}_{n=1}^{\infty}$, where ψ_n is a polyharmonic eigenfunction with boundary conditions (3.10) or (3.11), is in $l^2(\mathbb{N})$—the space of square summable sequences—and has corresponding norm bounded by $c\|f\|$.

We first require the following two lemmas:

Lemma 3.15. *Suppose that $a_n = \int_0^1 \mathrm{e}^{znx} f(x)\,\mathrm{d}x$, $n \in \mathbb{N}$, $f \in \mathrm{L}^2(0,1)$, where $z \neq 0$ and $\operatorname{Re} z \leq 0$. Then $\{a_n\} \in l^2(\mathbb{N})$ and $\sum_{n=1}^{\infty} |a_n|^2 \leq c\|f\|^2$.*

Proof. Though this lemma is established in [51, p.2332], we shall repeat the proof, since a d-variate generalisation will be obtained in the sequel.

Suppose first that $z = 2\pi i c$ with $c \in \mathbb{R}$ and, without loss of generality, $c > 0$. Let $m \in \mathbb{N}$ be minimal such that $\frac{m}{c} \geq 1$. Extend f to $[0, \frac{m}{c}]$ by setting $f(x) = 0$ for $1 < x \leq \frac{m}{c}$. Then

$$a_n = \int_0^{\frac{m}{c}} \mathrm{e}^{2\pi \mathrm{i} c n x} f(x)\,\mathrm{d}x = \sum_{i=1}^{m} \int_{\frac{i-1}{c}}^{\frac{i}{c}} \mathrm{e}^{2\pi \mathrm{i} c n x} f(x)\,\mathrm{d}x.$$

The i^{th} integral corresponds to the n^{th} Fourier coefficient of the restriction of the function f to $[\frac{i-1}{c}, \frac{i}{c}]$. Hence, by Parseval's lemma,

$$\sum_{n=1}^{\infty} |a_n|^2 \leq c \sum_{i=1}^{m} \int_{\frac{i-1}{c}}^{\frac{i}{c}} |f(x)|^2\,\mathrm{d}x \leq c\|f\|^2,$$

as required. Suppose now that $\operatorname{Re} z < 0$. Then $|a_n| \leq \int_0^1 \mathrm{e}^{\operatorname{Re} znx} |f(x)|\,\mathrm{d}x$, and we may assume that f is non-negative. Extending f by zero to a function $f \in \mathrm{L}^2(0,\infty)$, it suffices to show that the sequence $\{b_n\}$, where $b_n = \int_0^{\infty} \mathrm{e}^{-nx} f(x)\,\mathrm{d}x$, is in $l^2(\mathbb{N})$ for any $f \in \mathrm{L}^2(0,\infty)$, and has norm bounded by $\|f\|$. Since f is positive, b_n is a decreasing sequence. We deduce that

$$|b_n|^2 \leq \int_{n-1}^{n} |b_t|^2\,\mathrm{d}t.$$

It therefore suffices to prove that $\int_0^\infty |b_t|^2 \,\mathrm{d}t \leq c\|f\|^2$. However, by Fubini's theorem,

$$\begin{aligned}\int_0^\infty |b_t|^2 \,\mathrm{d}t &= \int_0^\infty \int_0^\infty \int_0^\infty \mathrm{e}^{-tx}\mathrm{e}^{-ty} f(x)f(y)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t \\ &= \int_0^\infty \int_0^\infty \frac{f(x)f(y)}{x+y}\,\mathrm{d}x\,\mathrm{d}y = \int_0^\infty f(x)g(x)\,\mathrm{d}x,\end{aligned}$$

where $g(x) = \int_0^\infty \frac{f(y)}{x+y}\,\mathrm{d}y$. Thus, the result is true, provided $g \in L^2(0,\infty)$ with $\|g\| \leq c\|f\|$. Note that

$$g(x) = \int_0^\infty \frac{f(xy)}{1+y}\,\mathrm{d}y.$$

Hence

$$\begin{aligned}\|g\|^2 &= \int_0^\infty g(z)^2\,\mathrm{d}z = \int_0^\infty \int_0^\infty \int_0^\infty \frac{f(xz)f(yz)}{(1+x)(1+y)}\,\mathrm{d}y\,\mathrm{d}x\,\mathrm{d}z \\ &\leq \int_0^\infty \int_0^\infty \frac{1}{(1+x)(1+y)} \left[\int_0^\infty |f(xz)|^2\,\mathrm{d}z\right]^{\frac{1}{2}} \left[\int_0^\infty |f(yz)|^2\,\mathrm{d}z\right]^{\frac{1}{2}} \mathrm{d}x\,\mathrm{d}y \\ &= \|f\|^2 \int_0^\infty \int_0^\infty \frac{1}{\sqrt{xy}(1+x)(1+y)}\,\mathrm{d}x\,\mathrm{d}y \leq c\|f\|^2,\end{aligned}$$

as required. □

Lemma 3.16. *Suppose that* $\{b_n\} \in l^2(\mathbb{N})$. *Then, for any* $\operatorname{Re} z \leq 0$, $z \neq 0$, *the family of all finite sums of terms of the form* $b_n \mathrm{e}^{znx}$ *is uniformly bounded in* $\mathrm{L}^2(0,1)$ *with norm bounded by* $c\left(\sum_{n=1}^\infty |b_n|^2\right)^{\frac{1}{2}}$ *for some* $c > 0$ *independent of* $\{b_n\}$.

Proof. Let $I \subseteq \mathbb{N}$ be finite. By the standard duality pairing,

$$\|g\| = \sup_{\substack{f \in \mathrm{L}^2(\Omega) \\ f \neq 0}} \frac{(g,f)}{\|f\|}, \quad \forall g \in \mathrm{L}^2(\Omega), \tag{3.25}$$

it follows that

$$\left\| \sum_{n \in I} b_n \mathrm{e}^{znx} \right\| = \sup_{\substack{f \in \mathrm{L}^2(0,1) \\ f \neq 0}} \frac{1}{\|f\|} \sum_{n \in I} b_n \int_0^1 \mathrm{e}^{znx} f(x)\,\mathrm{d}x.$$

By Lemma 3.15, $\sum_{n\in I} b_n \int_0^1 \mathrm{e}^{znx} f(x)\,\mathrm{d}x \leq c\left(\sum_{n=1}^\infty |b_n|^2\right)^{\frac{1}{2}} \|f\|$ as required. □

With these lemmas in hand, we now return to the polyharmonic problem:

Corollary 3.17. *Suppose that* $\{\psi_n\}$ *is the set of polyharmonic eigenfunctions subject to boundary conditions (3.10) or (3.11). Then, for* $f \in \mathrm{L}^2(-1,1)$, *the sequence* $\{(f,\psi_n)\} \in l^2(\mathbb{N})$ *with norm bounded by* $c\|f\|$.

Proof. We work on $[0,1]$. Using the result of Section 3.3.2, it follows that

$$\psi(x) = \sum_{s=0}^{q} \mathrm{e}^{\alpha\lambda_s(x-1)} c_s + \sum_{s=q+1}^{2q-1} \mathrm{e}^{\alpha\lambda_s x} c_s + \mathcal{O}\left(\mathrm{e}^{-\gamma_q \alpha}\right), \tag{3.26}$$

where $\psi(x)$ is the eigenfunction corresponding to the eigenvalue $\mu = \alpha^{2q}$ and the c_s are independent of α. Hence, $\psi_n(x)$ is a sum exponentials of the form e^{znx} with $\operatorname{Re} z \leq 0$, $z \neq 0$. The result now follows immediately from Lemma 3.15. □

Corollary 3.18. *Suppose that $\{\psi_n\}$ and $\{\chi_n\}$ are a biorthogonal pair of polyharmonic eigenfunctions subject to boundary conditions (3.10) and (3.11) respectively. Then, the family of all finite sums of terms $(f, \chi_n)\psi_n$ is uniformly bounded in $\mathrm{L}^2(-1,1)$ with norm bounded by $c\|f\|$.*

Proof. By Corollary 3.17, $(f, \chi_n) \in l^2(\mathbb{N})$ with norm bounded by $c\|f\|$. The result now follows upon writing ψ in the form (3.26) once more and applying Lemma 3.16. □

Corollary 3.18 immediately provides a Bessel-type inequality for expansions in polyharmonic eigenfunctions. We have:

Corollary 3.19. *Suppose that $f \in \mathrm{H}^r(-1,1)$, $r = 0, \ldots, q$, and that $\mathcal{F}_N[f]$ is the truncated expansion of f in polyharmonic–Neumann eigenfunctions. Then $\|\mathcal{F}_N[f]\|_r \leq c\|f\|_r$.*

Proof. The function $(\mathcal{F}_N[f])^{(r)}$ is a finite sum of terms of the form $(f^{(r)}, \chi_n)\psi_n$. Hence, an application of Corollary 3.18 gives the result. □

Having established these key properties of polyharmonic expansions, we may now prove the main result of this section and thereby answer the question raised previously:

Theorem 3.20. *Suppose that $f \in \mathrm{H}^r(-1,1)$, $r = 0, \ldots, q$, and that $\mathcal{F}_N[f]$ is the truncated expansion of f in polyharmonic–Neumann eigenfunctions. Then $\mathcal{F}_N[f]$ converges to f in the $\mathrm{H}^r(-1,1)$ norm.*

Proof. Since we have already proved the result for $r = 0, q$, we assume that $r = 1, \ldots, q-1$. In this case, given $\epsilon > 0$, there exists $g \in \mathrm{H}^q(-1,1)$ with $\|f - g\|_r < \epsilon$ [2]. In view of Corollary 3.19, $\|\mathcal{F}_N[f-g]\|_r < c\epsilon$. Hence

$$\|f - \mathcal{F}_N[f]\|_r \leq \|g - \mathcal{F}_N[g]\|_r + \|f - g\|_r + \|\mathcal{F}_N[f-g]\|_r < \|g - \mathcal{F}_N[g]\|_q + (1+c)\epsilon.$$

Since $g \in \mathrm{H}^q(-1,1)$, $\|g - \mathcal{F}_N[g]\|_q < \epsilon$ for sufficiently large N, completing the proof. □

An immediate consequence of this theorem is uniform convergence of polyharmonic–Neumann expansions:

Corollary 3.21. *Suppose that $f \in \mathrm{H}^r(-1,1)$, $r = 1, \ldots, q$, and that $\mathcal{F}_N[f]$ is the truncated expansion of f in polyharmonic–Neumann eigenfunctions. Then $(\mathcal{F}_N[f])^{(s)}$ converges uniformly to $f^{(s)}$ for $s = 0, \ldots, r-1$.*

Proof. This follows immediately from the Sobolev imbedding $\mathrm{H}^s(-1,1) \hookrightarrow \mathrm{C}^{s-1}[-1,1]$, $s \in \mathbb{N}$, and Theorem 3.20. □

In particular, $\mathcal{F}_N[f]$ converges uniformly to $f \in \mathrm{H}^1(-1,1)$. However, provided $f \in \mathrm{H}^q(-1,1)$, the first $q-1$ derivatives of $\mathcal{F}_N[f]$ converge uniformly to the corresponding derivatives of f. This observation succinctly expresses the advantage of increasing q: namely, convergence in higher-order norms. Note that Theorem 3.20 and Corollary 3.21 generalise

Lemma 2.9 and Theorem 2.12 respectively to $q \geq 1$ when $d = 1$. We provide an extension to both $q \geq 1$ and $d \geq 1$ in Section 3.5.

We mention in passing that a direct consequence of Theorem 3.20 is $\mathrm{L}^2(-1,1)$ norm convergence of the expansion of a function $f \in \mathrm{L}^2(-1,1)$ in any biorthogonal pair of polyharmonic eigenfunctions subject to boundary conditions (3.10) or (3.11). This result, as mentioned, is well known in a much wider context. The proof presented above cannot be extended to other differential operators and boundary conditions, aside from simple cases, since it relies both on the particular duality of polyharmonic eigenfunctions[8] (Theorem 3.3) and known results for the Dirichlet and Neumann cases.

The analysis of this section also gives criteria for both the best and worst boundary conditions to prescribe to the polyharmonic operator in terms of the convergence of the truncated expansion $\mathcal{F}_N[f]$, as opposed to the arguments described in Section 3.1.1 based on the decay of the coefficients. It is easily established that the expansion based on polyharmonic eigenfunctions subject to boundary conditions (3.10) converges maximally in the $\mathrm{H}^{q-p}(-1,1)$ norm, $p = 0, \ldots, q$. Correspondingly, for boundary conditions (3.11), only $\mathrm{L}^2(-1,1)$ convergence occurs. Hence, choosing $p = 0$ for the highest possible degree of convergence, we once more arrive at Neumann boundary conditions. Conversely, Dirichlet boundary conditions ($p = q$) give the worst degree of convergence.

Estimates for the rate of convergence in various norms will be established in Section 3.5 for the general case $d \geq 1$. Our final topic concerning univariate expansions addresses pointwise convergence. As in the $q = 1$ case (see Section 2.9.1), both a higher order and faster rate of convergence occurs at the interior points $x \in (-1,1)$.

3.4.2 Pointwise convergence

To establish the degree of pointwise convergence of $\mathcal{F}_N[f]$, and to provide estimates for the rate of convergence, we must first expand the coefficient $\hat{f}_n$ in inverse powers of n. Starting from (3.8) with $r = q$ and integrating by parts, we obtain, for each $p = 0, \ldots, q$,

$$\begin{aligned}\hat{f}_n = &\frac{1}{\alpha_n^{2q}} \sum_{s=0}^{p-1} (-1)^s \left[f^{(q+s)}(1)\phi_n^{(q-s-1)}(1) - f^{(q+s)}(-1)\phi_n^{(q-s-1)}(-1) \right] \\ &+ \frac{(-1)^p}{\alpha_n^{2q}} \int_{-1}^{1} f^{(q+p)}(x)\phi_n^{(q-p)}(x)\,\mathrm{d}x, \end{aligned} \tag{3.27}$$

provided $f \in \mathrm{H}^{q+p}(-1,1)$. Note that, if $f \in \mathrm{H}^{q+1}(-1,1)$, this verifies that $\hat{f}_n = \mathcal{O}\left(n^{-q-1}\right)$ for polyharmonic–Neumann expansion coefficients, as previously stated.

As demonstrated, $\mathcal{F}_N[f]$ and its first $q-1$ derivatives converge uniformly to $f \in \mathrm{H}^q(-1,1)$. However, as the following result verifies, the q^{th} derivative of this expansion also converges pointwise away from the endpoints:

Lemma 3.22. *Suppose that $f \in \mathrm{H}^{q+1}(-1,1)$. Then $(\mathcal{F}_N[f])^{(q)}$ converges uniformly to $f^{(q)}$ in compact subsets of $(-1,1)$.*

[8] As discussed in Section 2.11, such duality arguments are only applicable in a small number of cases, even when $q = 1$.

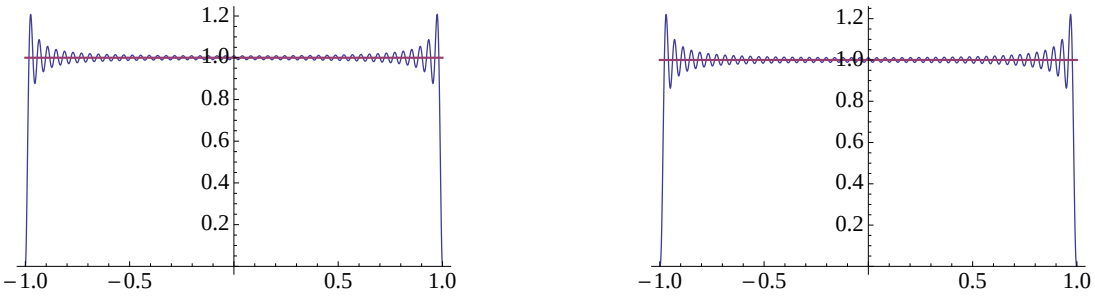


Figure 3.4: The Gibbs phenomenon for polyharmonic–Dirichlet expansions. Graph of $f(x) = 1$ and $\mathcal{F}_{50}[f](x)$ for $-1 \leq x \leq 1$, where $q = 2$ (left), $q = 3$ (right) and $\mathcal{F}_N[f]$ is the expansion of f in polyharmonic–Dirichlet eigenfunctions.

Proof. Suppose first that $f \in \mathrm{C}^\infty[-1,1]$. Then, from (3.27), we have

$$\hat{f}_n \phi_n^{(q)}(x) = \frac{1}{\alpha_n^{2q}} \left[f^{(q)}(1)\phi_n^{(q-1)}(1) - f^{(q)}(-1)\phi_n^{(q-1)}(-1) \right] \phi_n^{(q)}(x) + \mathcal{O}\left(n^{-2}\right), \quad x \in [-1,1].$$

Now suppose that $I \subseteq (-1,1)$ is a compact set. As we prove in the sequel, the partial sums $\sum_{n=1}^{N} \frac{\phi^{(q-1)}(\pm 1)}{\alpha_n^{2q}} \phi_n^{(q)}(x)$ form a Cauchy sequence uniformly for $x \in I$. Hence, the truncated sums $(\mathcal{F}_N[f])^{(q)}$ converge uniformly on I to a continuous function g. Arguments as in Lemma 2.23 now complete the proof. □

Equivalently, this lemma states that the expansion of a function $f \in \mathrm{H}^1(-1,1)$ in polyharmonic eigenfunctions subject to the Dirichlet boundary conditions (3.3) converges pointwise in $(-1,1)$. The lack of convergence at the endpoints is also easily confirmed. Hence, a Gibbs phenomenon occurs for such expansions (likewise, a Gibbs phenomenon occurs in the q^{th} derivative of polyharmonic–Neumann expansions). This is demonstrated in Figure 3.4 for $q = 2, 3$.[9]

Returning to Neumann expansions, we now address the rate of pointwise convergence. In doing so, we derive a full asymptotic expansion for the error $f(x) - \mathcal{F}_N[f](x)$ at any point $x \in [-1,1]$. To facilitate this, however, we must first digress and give a full expansion of the coefficient $\hat{f}_n$ in inverse powers of n. This is based on (3.27). Setting $p = q$ in this formula and iterating the result gives

$$\begin{aligned}
&= \sum_{r=0}^{k-1} \frac{(-1)^{rq}}{\alpha_n^{2(r+1)q}} \sum_{s=0}^{q-1} (-1)^s \left[f^{((2r+1)q+s)}(1)\phi_n^{(q-s-1)}(1) - f^{((2r+1)q+s)}(-1)\phi_n^{(q-s-1)}(-1) \right] \\
&\quad + \frac{(-1)^{kq}}{\alpha_n^{2(k+1)q}} \sum_{s=0}^{p-1} (-1)^s \left[f^{((2k+1)q+s)}(1)\phi_n^{(q-s-1)}(1) - f^{((2k+1)q+s)}(-1)\phi_n^{(q-s-1)}(-1) \right] \\
&\quad + \frac{(-1)^{kq+p}}{\alpha_n^{2(k+1)q}} \int_{-1}^{1} f^{((2k+1)q+p)}(x)\phi_n^{(q-p)}(x)\,\mathrm{d}x, \qquad (3.28)
\end{aligned}$$

[9] As in the classical Fourier case [107], the existence of $\mathcal{O}(1)$ oscillations follows from setting $x = 1 - N^{-1}$. It is well known that the magnitude of the overshoot of a truncated Fourier sum can be explicitly calculated (unsurprisingly, in view of Section 2.3, Laplace–Dirichlet expansions have precisely the same overshoot).

for $f \in \mathrm{H}^{(2k+1)q+p}(-1,1)$, where $p = 0, \ldots, q-1$ and $k \in \mathbb{N}_0$. Observe that this expansion generalises the $q = 1$ result, equation (2.11), to general $q \geq 1$. Moreover, as in the $q = 1$ case, when iterated, (3.28) provides an asymptotic expansion for the coefficients $\hat{f}_n$ of a function $f \in \mathrm{C}^\infty[-1,1]$:

$$\hat{f}_n \sim \sum_{r=0}^{\infty} \sum_{s=0}^{q-1} \frac{(-1)^{rq+s}}{\alpha_n^{2(r+1)q}} \left[f^{((2r+1)q+s)}(1) \phi_n^{(q-s-1)}(1) - f^{((2r+1)q+s)}(-1) \phi_n^{(q-s-1)}(-1) \right]. \quad (3.29)$$

Upon recalling that $\alpha_n = \mathcal{O}(n)$ and $\phi_n^{(q-s-1)}(\pm 1) = \mathcal{O}\left(n^{q-s-1}\right)$, (3.29) is confirmed as an asymptotic expansion for $\hat{f}_n$ in powers of n^{-1}. Note that a power n^{-l} is only included in $\hat{f}_n$ if $l = (2r+1)q + s + 1$ for some $r \in \mathbb{N}_0$ and $s = 0, \ldots, q-1$.

In the particular case $q = 2$, simple algebra confirms that (3.29) reduces to

$$\hat{f}_n^{[i]} \sim \sqrt{2} \sum_{r=0}^{\infty} \left\{ \frac{\gamma_n^{[i]}}{(\alpha_n^{[i]})^{4r+3}} \left[f^{(4r+2)}(1) + (-1)^i f^{(4r+2)}(-1) \right] \right.$$
$$\left. - \frac{1}{(\alpha_n^{[i]})^{4r+4}} \left[f^{(4r+3)}(1) + (-1)^{i+1} f^{(4r+3)}(-1) \right] \right\},$$

where $i = 0$ corresponds to the even eigenfunction ϕ_{2n-1}, $i = 1$ the odd eigenfunction ϕ_{2n}, and $\gamma_n^{[0]} = -\tan \alpha_n^{[0]}$, $\gamma_n^{[1]} = \cot \alpha_n^{[1]}$. By means of example, consider the function $f(x) = \mathrm{e}^x$. In this case

$$\hat{f}_n^{[0]} = -2\sqrt{2} \frac{\alpha_n^{[0]} \tan \alpha_n^{[0]} \cosh 1 + \sinh 1}{(\alpha_n^{[0]})^4 - 1} = \sqrt{2} \left\{ \frac{\gamma_n^{[0]}}{(\alpha_n^{[0]})^3} 2 \cosh 1 - \frac{1}{(\alpha_n^{[0]})^4} 2 \sinh 1 \right\} + \mathcal{O}\left(n^{-7}\right),$$
$$\hat{f}_n^{[1]} = 2\sqrt{2} \frac{\alpha_n^{[1]} \cot \alpha_n^{[1]} \sinh 1 - \cosh 1}{(\alpha_n^{[1]})^4 - 1} = \sqrt{2} \left\{ \frac{\gamma_n^{[1]}}{(\alpha_n^{[1]})^3} 2 \sinh 1 - \frac{1}{(\alpha_n^{[1]})^4} 2 \cosh 1 \right\} + \mathcal{O}\left(n^{-7}\right),$$

thus verifying (3.28). Though not important to our present discussion, we notice that the coefficients $\hat{f}_0^{[0]}$ and $\hat{f}_0^{[1]}$ of the eigenfunctions $\frac{1}{\sqrt{2}}$ and $\frac{\sqrt{3}}{\sqrt{2}}x$ corresponding to the double zero eigenvalue are given by $\sqrt{2}\sinh 1$ and $\sqrt{6}\mathrm{e}^{-1}$ respectively.

The expansions (3.28) and (3.29) give a first insight into a topic we shall consider in greater detail in Section 3.6: the appropriate derivative conditions for polyharmonic–Neumann expansions. To this end, we now introduce some additional notation. We define the set $N_m \subseteq \mathbb{N}_0$ by

$$N_m = \{ l \in \mathbb{N}_0 : l = (2r+1)q + s < m,\ r \in \mathbb{N}_0,\ s = 0, \ldots, q-1 \}, \quad m \in \mathbb{N}_0, \quad (3.30)$$

and, for $p = 0, \ldots, q-1$ and $k \in \mathbb{N}_0$ ($p \neq 0$ when $k = 0$),

$$\rho_{k,p} = \begin{cases} 2kq & p = 0 \\ (2k+1)q + p & p = 1, \ldots, q-1. \end{cases} \quad (3.31)$$

Note that $\rho_{k,p} - 1$ is equal to the highest-order derivative appearing in (3.28) and that the value $f^{(l)}(\pm 1)$ appears in (3.28) if and only if $l \in N_{\rho_{k,p}}$. Additionally, if $f^{(l)}(\pm 1) = 0$ for $l \in N_{\rho_{k,p}}$ then $\hat{f}_n = \mathcal{O}\left(n^{-(2k+1)q-p-1}\right)$.

With this to hand, we now return to the rate of pointwise convergence. Our intention is to demonstrate that the convergence rate is $\mathcal{O}\left(N^{-q}\right)$ at $x = \pm 1$ and $\mathcal{O}\left(N^{-q-1}\right)$ for $x \in (-1,1)$, thereby generalising the $q = 1$ result (Theorem 2.22). To do this, we first use Corollary 3.21 to write $f(x) - \mathcal{F}_N[f](x) = \sum_{n>N} \hat{f}_n \phi_n(x)$. In view of (3.28), it follows that

$$\begin{aligned} &f(x) - \mathcal{F}_N[f](x) \\ &= \sum_{r=0}^{k-1} \sum_{s=0}^{q-1} (-1)^{rq+s} \left\{ f^{((2r+1)q+s)}(1)\Phi^+(r,s,N,x) - f^{((2r+1)q+s)}(-1)\Phi^-(r,s,N,x) \right\} \\ &\quad + \sum_{s=0}^{p-1} (-1)^{kq+s} \left\{ f^{((2k+1)q+s)}(1)\Phi^+(k,s,N,x) - f^{((2k+1)q+s)}(-1)\Phi^-(k,s,N,x) \right\} \\ &\quad + \mathcal{O}\left(N^{-(2k+1)q-p}\right), \end{aligned} \tag{3.32}$$

where the function $\Phi^{\pm}(r,s,N,x)$ is defined by

$$\Phi^{\pm}(r,s,N,x) = \sum_{n>N} \frac{\phi_n^{(q-s-1)}(\pm 1)}{\alpha_n^{2(r+1)q}} \phi_n(x).$$

In particular, if $f \in \mathrm{C}^{\infty}[-1,1]$, then

$$\begin{aligned} f(x) - \mathcal{F}_N[f](x) \sim \sum_{r=0}^{\infty} \sum_{s=0}^{q-1} \Big\{ &(-1)^{rq+s} f^{((2r+1)q+s)}(1)\Phi^+(r,s,N,x) \\ &+ (-1)^{rq+s+1} f^{((2r+1)q+s)}(-1)\Phi^-(r,s,N,x) \Big\}. \end{aligned} \tag{3.33}$$

Though not immediately apparent, (3.33) is an asymptotic expansion of the error in inverse powers of N. To demonstrate this, we must scrutinise the behaviour of the functions $\Phi^{\pm}(r,s,N,x)$ for large N. Intuition arising from the $q = 1$ case suggests that $\Phi^{\pm}(r,s,N,x)$ is $\mathcal{O}\left(N^{-((2r+1)q+s+1)}\right)$ for $x \in (-1,1)$ and $\mathcal{O}\left(N^{-((2r+1)q+s)}\right)$ for $x = \pm 1$, thus confirming (3.33) as an asymptotic expansion. The following two lemmas verify these estimates:

Lemma 3.23. *Suppose that $x \in (-1,1)$. Then, the function $\Phi^{\pm}(r,s,N,x)$ satisfies*

$$\begin{aligned} \Phi^{\pm}(r,s,N,x) = &\operatorname{Re}\left[c_1^{\pm} (-\mathrm{e}^{\mathrm{i}\pi x})^{A_N + \frac{q-1}{4}} \Phi\left(-\mathrm{e}^{\mathrm{i}\pi x}, (2r+1)q+s+1, A_N + \tfrac{q-1}{4}\right) \right] \\ &+ \operatorname{Re}\left[c_2^{\pm} (-\mathrm{e}^{\mathrm{i}\pi x})^{A_N + \frac{q-3}{4}} \Phi\left(-\mathrm{e}^{\mathrm{i}\pi x}, (2r+1)q+s+1, A_N + \tfrac{q-3}{4}\right) \right] \\ &+ \mathcal{O}\left(N^{-(2r+1)q-s} \mathrm{e}^{-\frac{1}{2}\gamma_q(1-|x|)N\pi} \right), \end{aligned}$$

where $A_N = \lfloor \frac{1}{2} N \rfloor$, Φ is the Lerch transcendental function (2.35) and $c_1^{\pm}, c_2^{\pm} \in \mathbb{C}$ are constants independent of N and x. In particular,

$$\begin{aligned} \Phi^{\pm}(r,s,N,x) = &\operatorname{Re}\left\{ \frac{(-\mathrm{e}^{\mathrm{i}\pi x})^{A_N + \frac{q-3}{4}}}{1 + \mathrm{e}^{\mathrm{i}\pi x}} \left[c_1^{\pm} (-\mathrm{e}^{\mathrm{i}\pi x})^{\frac{1}{2}} + c_2^{\pm} \right] \right\} N^{-(2r+1)q-s-1} \\ &+ \mathcal{O}\left(N^{-(2r+1)q-s-2} \right). \end{aligned}$$

Proof. Since $\alpha_n = \frac{1}{4}(2n+q-1)\pi + \mathcal{O}(\mathrm{e}^{-\frac{1}{2}n\pi\gamma_q})$ (Theorem 3.6) it follows that $\mathrm{e}^{2\mathrm{i}\alpha_n} = c(-1)^n + \mathcal{O}(\mathrm{e}^{-\frac{1}{2}n\pi\gamma_q})$ for some constant c independent of n. Moreover, by Lemma 3.12,

$$\begin{aligned}\phi_n^{(q-s-1)}(\pm 1) &= c\alpha_n^{q-s-1}\mathrm{e}^{2\mathrm{i}\alpha_n} + \mathcal{O}\left(\alpha_n^{q-s-1}\mathrm{e}^{-\gamma_q\alpha_n}\right).\\ &= c\left(\tfrac{1}{2}n\pi + \tfrac{q-1}{4}\pi\right)^{q-s-1}(-1)^n + \mathcal{O}\left(n^{q-s-1}\mathrm{e}^{-\frac{1}{2}\gamma_q n\pi}\right).\end{aligned}$$

It follows from Theorem 3.9 that

$$\begin{aligned}\phi_n(x) &= \mathrm{Re}\left[\mathrm{e}^{\mathrm{i}\frac{1}{4}(2n+q-1)\pi x + \mathrm{i}\frac{1}{2}(n+q-1)\pi}\right] + \mathcal{O}\left(\mathrm{e}^{-\frac{1}{2}\gamma_q(1-|x|)n\pi}\right)\\ &= \mathrm{Re}\left[c\mathrm{i}^n\mathrm{e}^{\mathrm{i}\frac{1}{4}(2n+q-1)\pi x}\right] + \mathcal{O}\left(\mathrm{e}^{-\frac{1}{2}\gamma_q(1-|x|)n\pi}\right).\end{aligned}$$

Combining these results, we obtain

$$\begin{aligned}\Phi^{\pm}(r,s,N,x) =& \mathrm{Re}\left[c\mathrm{e}^{\mathrm{i}\frac{q-1}{4}\pi x}\sum_{n>N}\left(\tfrac{1}{2}n+\tfrac{q-1}{4}\right)^{-(2r+1)q-s-1}(-\mathrm{i})^n\mathrm{e}^{\mathrm{i}\frac{n}{2}\pi x}\right]\\ &+ \mathcal{O}\left(N^{-(2r+1)q-s-1}\mathrm{e}^{\frac{1}{2}\gamma_q(1-|x|)N\pi}\right).\end{aligned}$$

Now, for any $a \in \mathbb{R}$ and $b > 1$, we have

$$\begin{aligned}\sum_{n>N}\left(\tfrac{1}{2}n+a\right)^{-b}(-\mathrm{i})^n\mathrm{e}^{\mathrm{i}\frac{n}{2}\pi x} =& \sum_{n>A_N}(n+a)^{-b}(-\mathrm{e}^{\mathrm{i}\pi x})^n\\ &+ \mathrm{i}\mathrm{e}^{-\frac{1}{2}\mathrm{i}\pi x}\sum_{n>A_N}\left(n+a-\tfrac{1}{2}\right)^{-b}(-\mathrm{e}^{\mathrm{i}\pi x})^n.\end{aligned}$$

Rewriting n as $n+A_N$ and using the definition of the Lerch function now gives the first result. For the second we merely use the estimate (2.36). □

In an identical manner, we may also derive an estimate for $x = \pm 1$:

Lemma 3.24. *Suppose that $x = \pm 1$. Then*

$$\Phi^{\pm}(r,s,N,x) = c^{\pm}\zeta\left((2r+1)q+s+1, N+\tfrac{1}{2}(q+1)\right) + \mathcal{O}\left(\mathrm{e}^{-\frac{1}{2}\gamma_q N\pi}\right),$$

where $\zeta(\cdot,\cdot)$ is the Hurwitz zeta function [1]. In particular, $\Phi^{\pm}(r,s,N,x) = \mathcal{O}\left(N^{-(2r+1)q-s}\right)$ for $x = \pm 1$.

With these two lemmas at hand, analysis of the pointwise convergence rate follows straightaway from (3.32):

Theorem 3.25. *Suppose that $p = 0,\ldots,q-1$, $k \in \mathbb{N}_0$ and that $f \in \mathrm{H}^{(2k+1)q+p+2}(-1,1)$ satisfies $f^{(l)}(\pm 1) = 0$ whenever $l \in N_{\rho_{k,p}}$. Then $f(\pm 1) - \mathcal{F}_N[f](\pm 1) = \mathcal{O}\left(N^{-(2k+1)q-p}\right)$ and $f(x) - \mathcal{F}_N[f](x) = \mathcal{O}\left(N^{-(2k+1)q-p-1}\right)$ uniformly in compact subsets of $(-1,1)$.*

In particular, for general functions f (i.e. $k = p = 0$), this theorem provides the stated $\mathcal{O}\left(N^{-q-1}\right)$ pointwise convergence rate estimate away from $x = \pm 1$. Additionally, it verifies that the convergence rate is one power of N faster inside the interval than at the endpoints. This effect is exhibited in Figure 3.5 for $q = 2, 3$ (compare with Figure 2.3 for $q = 1$).

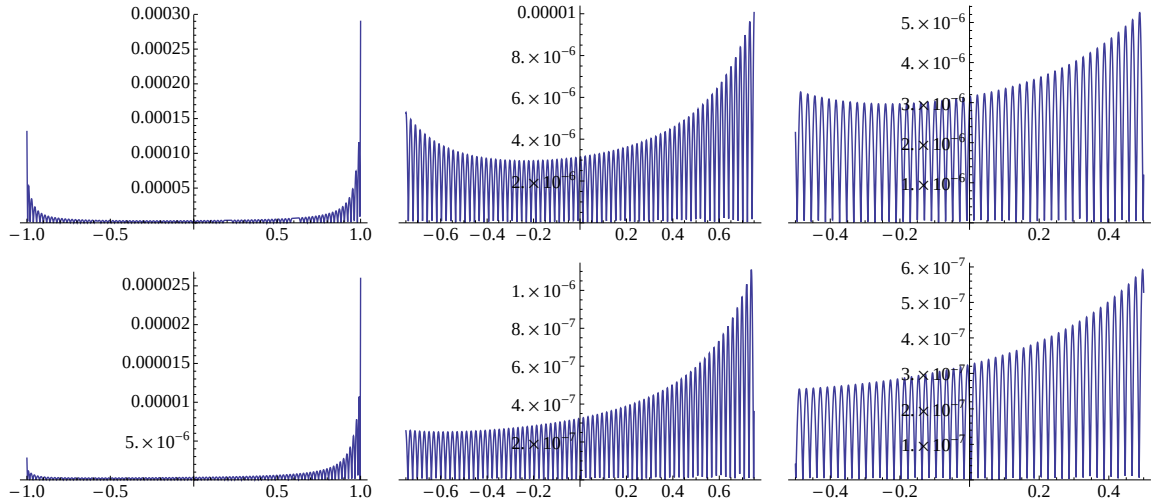


Figure 3.5: Graphs of $|f(x) - \mathcal{F}_{50}[f](x)|$ for $-1 \leq x \leq 1$ (left), $-\frac{3}{4} \leq x \leq \frac{3}{4}$ (middle) and $-\frac{1}{2} \leq x \leq \frac{1}{2}$ (right), where $f(x) = \mathrm{Ai}(-3x - 4)$ and Ai is the Airy function [1], with $q = 2$ (top row) and $q = 3$ (bottom row).

3.5 Many dimensions

The aim of this section is the generalisation of univariate polyharmonic–Neumann expansions to the d-variate cube $\Omega = (-1, 1)^d$, along the same lines as Chapter 2. Immediately, we are confronted by a problem. The obvious extension, via eigenfunctions of the multivariate polyharmonic operator $(-1)^q \triangle^q$, is completely unsuitable. Such eigenfunctions cannot be expressed in terms of simple functions, and thus have limited practical use.

Ideally, we seek separable eigenfunctions, with underlying properties inherited from the univariate case. An obvious approach is merely to form all Cartesian products of the univariate polyharmonic eigenfunctions. Evidently, such functions are no longer eigenfunctions of the polyharmonic operator itself, meaning that standard spectral theory does not immediately apply. Nonetheless, progress can be made. As we shall prove, these functions turn out to be precisely the eigenfunctions of the *subpolyharmonic*[10] operator

$$\mathcal{L}_0 = (-1)^q \sum_{j=1}^{q} \partial_{x_j}^{2q}, \tag{3.34}$$

subject to the homogeneous Neumann boundary conditions

$$\partial_{x_j}^{q+r} \phi \big|_{\partial\Gamma_j} = 0, \quad j = 1, \ldots, d, \quad r = 0, \ldots, q-1. \tag{3.35}$$

Our intuition suggests that $\mathcal{L}_0$ ought to be spectrally equivalent to the polyharmonic operator, and, as we will prove, this turns out to be the case. Hence, the eigenfunctions and eigenvalues

[10] This nomenclature stems from the fact that the subpolyharmonic operator contains fewer derivatives than the multivariate polyharmonic operator $(-1)^q \triangle^q$. Consequently, if T_0 and T are the associated weak forms, then $T_0(f, f) \leq T(f, f)$ for all $f \in \mathrm{H}^q(\Omega)$. The prefix 'sub' should not be confused with the notion of subharmonic functions.

of $\mathcal{L}_0$ subject to boundary conditions (3.35) inherit the spectral properties of the polyharmonic operator, including, for example, $\mathrm{L}^2(\Omega)$ density. Therefore, the remainder of this chapter is devoted to the construction and analysis of expansions in the eigenfunctions of $\mathcal{L}_0$ equipped with Neumann boundary conditions. The results we prove extend the work of Chapter 2 and Section 3.4 to the $q \geq 1$ and $d \geq 1$ setting.

To commence, we must first confirm that the eigenfunctions of $\mathcal{L}_0$ arise precisely from Cartesian products:

Lemma 3.26. *The eigenfunctions of (3.34) subject to boundary conditions (3.35) are precisely the Cartesian products of the eigenfunctions of the univariate polyharmonic operator (3.2).*

Proof. The d-variate eigenfunctions of (3.34) subject to boundary conditions (3.35) form an orthonormal basis of $\mathrm{L}^2(\Omega)$ (see Theorem 3.29). We now proceed exactly as in the $q = 1$ case (Lemma 2.2). □

As in previous studies, duality is central to the convergence analysis of subpolyharmonic–Neumann expansions. To properly address this notion, we first require a new family of Sobolev norms. Such norms generalise the univariate norms $\|\!|\!| \cdot |\!|\!|_q$ introduced in Section 3.4.

3.5.1 Modulo q norms

In the univariate setting, our analysis exploited the duality between the Neumann and Dirichlet eigenfunctions under the action of the q^{th} derivative operator (Lemma 3.1). In light of this, we introduced a new norm $\|\!|\!| \cdot |\!|\!|_q$ involving only the function and its q^{th} derivative. In the same spirit, we now introduce a family of multivariate norms involving only the function and its q^{th} order partial derivatives. We refer to such norms as *modulo q norms*.

Lemma 3.27. *The bilinear form*

$$(f, g)_q = (f, g) + \sum_{j=1}^{d} \left(\partial_{x_j}^q f, \partial_{x_j}^q g \right), \quad f, g \in \mathrm{H}^q(\Omega),$$

is an inner product on $\mathrm{H}^q(\Omega)$. *The associated norm, given by* $|\!|\!| f |\!|\!|_q^2 = \|f\|^2 + \sum_{j=1}^d \|\partial_{x_j}^q f\|^2$, *is equivalent to* $\|\cdot\|_q$.

Proof. Trivially $|\!|\!| f |\!|\!|_q \leq \|f\|_q$. To demonstrate the other inequality, we need to show that $\|\mathrm{D}^\beta f\| \leq c |\!|\!| f |\!|\!|_q$ for $|\beta| \leq q$. In view of the additive interpolation inequality [2],

$$\|\partial_{x_j}^r g\| \leq c \left(\|g\| + \|\partial_{x_j}^q g\| \right), \quad \forall g \in \mathrm{H}^q(\Omega), \quad r = 0, \ldots, q, \tag{3.36}$$

it suffices to consider $|\beta| = q$. We now prove this result by induction on q. For $q = 1$, there is nothing to prove, hence we assume that $\|\mathrm{D}^\beta f\| \leq c |\!|\!| f |\!|\!|_{q-1}$ for all $|\beta| \leq q - 1$. If $|\beta| = q$ we may write $\beta = \gamma + \delta$ with $|\gamma| = q - 1$ and $|\delta| = 1$. Without loss of generality $\delta = (1, 0, ..., 0)$. Therefore, by the result for $q - 1$

$$\|\mathrm{D}^\beta f\|^2 \leq c |\!|\!| \mathrm{D}^\delta f |\!|\!|_{q-1} \leq c \left(\|\partial_{x_1} f\|^2 + \sum_{j=1}^{d} \|\partial_{x_j}^{q-1} \partial_{x_1} f\|^2 \right) \leq c \left(\|f\|^2 + \sum_{j=1}^{d} \|\partial_{x_j}^{q-1} \partial_{x_1} f\|^2 \right).$$

Here the second inequality follows from (3.36) with $r = 1$, $j = 1$ and $g = f$. Now, it can be shown that (see [24, p.171])

$$\|\partial_{x_j}^{q-1}\partial_{x_1} f\| \le c\left(\sum_{j=1}^{d}\|\partial_{x_j}^{q} f\| + \|f\|\right), \quad \forall f \in \mathrm{H}^q(\Omega), \quad j = 1, ..., d, \tag{3.37}$$

with constant $c > 0$ independent of f. Hence $\|D^\beta f\| \le c\|\!|f|\!\|_q$ as required. □

The inequality (3.37) is a simple consequence of a rather extensive body of literature aimed at determining equivalences of function spaces defined in terms of boundedness conditions on various partial derivatives. Simply put, given a collection of partial derivatives of an arbitrary function, which other partial derivatives can be bounded by a linear combination of norms of derivatives from the collection? We shall not dwell on this topic, and refer the reader to [24] for a thorough exposition.

Since the mixed Sobolev spaces $\mathrm{H}^q_{\mathrm{mix}}(\Omega)$ are important in the study of uniform convergence of modified Fourier expansions, we also require so-called *mixed* modulo q norms. We have

Lemma 3.28. *The bilinear form*

$$(f, g)_{q,mix} = \sum_{|\beta|_\infty \le 1} \left(\mathrm{D}^{\beta q} f, \mathrm{D}^{\beta q} g\right), \quad f, g \in \mathrm{H}^q_{mix}(\Omega),$$

is an inner product on $\mathrm{H}^q_{mix}(\Omega)$. *Moreover, the associated norm* $\|\!|\cdot|\!\|_{q,mix}$, *given by* $\|\!|f|\!\|^2_{q,mix} = \sum_{|\beta|_\infty \le 1} \|\mathrm{D}^{\beta q} f\|^2$, *is equivalent to* $\|\cdot\|_{q,mix}$.

Proof. Once more $\|\!|f|\!\|_{q,\mathrm{mix}} \le \|f\|_{q,\mathrm{mix}}$. To prove the other inequality, it suffices to show that $\|\mathrm{D}^\gamma f\| \le c\|\!|f|\!\|_{q,\mathrm{mix}}$, where $|\gamma|_\infty \le q$ and $\gamma \ne \beta q$ for any $|\beta|_\infty \le 1$. For such a γ we write $\gamma = \beta q + \delta$, where $\delta_j = \gamma_j$ and $\beta_j = 0$ if $0 \le \gamma_j \le q - 1$ and $\delta_j = 0$ and $\beta_j = 1$ otherwise. Since $\mathrm{D}^\gamma f = \mathrm{D}^\delta \mathrm{D}^{\beta q} f$, we first consider $\|\mathrm{D}^\delta g\|$ for some function $g \in \mathrm{H}^q_{\mathrm{mix}}(\Omega)$. It follows from repeated application of the interpolation inequality (3.36), where necessary, that

$$\|\mathrm{D}^\delta g\|^2 \le c \sum_{\substack{|\theta|_\infty \le 1 \\ \theta_j = 0 \ \Leftrightarrow \ \delta_j = 0}} \|D^{\theta q} g\|^2.$$

Hence we obtain

$$\|\mathrm{D}^\gamma f\|^2 \le c \sum_{\substack{|\theta|_\infty \le 1 \\ \theta_j = 0 \ \Leftrightarrow \ \delta_j = 0}} \|\mathrm{D}^{(\beta+\theta)q} f\|^2 \le c\|\!|f|\!\|^2_{q,\mathrm{mix}},$$

since $|\beta + \theta|_\infty \le 1$. □

3.5.2 Density and convergence

In this section, we establish results pertaining to the convergence of subpolyharmonic eigenfunction expansions. Our first task is to establish $\mathrm{L}^2(\Omega)$ density of such eigenfunctions:

Theorem 3.29. *The set of subpolyharmonic–Neumann eigenfunctions forms an orthonormal basis of* $\mathrm{L}^2(\Omega)$.

Proof. In view of Lemma 3.27, the bilinear form $(\cdot,\cdot)_q : \mathrm{H}^q(\Omega)\times\mathrm{H}^q(\Omega)\to\mathbb{R}$ is continuous and coercive. Orthogonality and densitynow follow from standard arguments ([56, p.335]). The inverse operator $\mathcal{L}_0^{-1} : \mathrm{L}^2(\Omega)\to\mathrm{L}^2(\Omega)$ is bounded, linear, symmetric and compact; hence, it has a countable orthonormal basis of eigenfunctions. □

Given finite index sets $I_N \subseteq \mathbb{N}_0^d$, $N\in\mathbb{N}$, satisfying (2.8), we define the truncated expansion $\mathcal{F}_N[f]$ of a function $f\in\mathrm{L}^2(\Omega)$ in the standard manner. Convergence of $\mathcal{F}_N[f]$ to f in $\mathrm{L}^2(\Omega)$ follows immediately from Theorem 3.29.

For ease of notation, we henceforth relabel the univariate polyharmonic eigenfunctions so that $\phi_n = \phi_{0,n}$ when $n = 0,\dots,q-1$ and $\phi_{n+q}=\phi_n$ otherwise. Since the multivariate eigenfunctions are formed from Cartesian products, we write $\phi_n(x) = \phi_{n_1}(x_1)\dots\phi_{n_d}(x_d)$, $n\in\mathbb{N}_0^d$, for a multivariate eigenfunction using this labelling. If the corresponding multivariate coefficient is written as $\hat{f}_n$, then we readily obtain:

$$\|f\|^2 = \sum_{n\in\mathbb{N}_0^d} |\hat{f}_n|^2, \quad \forall f\in\mathrm{L}^2(\Omega).$$

Not only can we characterise the $\mathrm{L}^2(\Omega)$ norm: by using multivariate duality arguments, we may also characterise the modulo q-norms introduced in the previous section:

Theorem 3.30. *The subpolyharmonic–Neumann eigenfunctions form an orthogonal basis of the spaces* $\mathrm{H}^q(\Omega)$ *and* $\mathrm{H}^q_{mix}(\Omega)$ *with respect to the inner products* $(\cdot,\cdot)_q$ *and* $(\cdot,\cdot)_{q,mix}$. *Furthermore*

$$|\!|\!| f |\!|\!|_q^2 = \sum_{n\in\mathbb{N}_0^d} (1+\mu_n)|\hat{f}_n|^2, \quad \forall f\in\mathrm{H}^q(\Omega), \tag{3.38}$$

$$|\!|\!| f |\!|\!|_{q,mix}^2 = \sum_{n\in\mathbb{N}_0^d} \left[\prod_{j=1}^d (1+\mu_{n_j})\right] |\hat{f}_n|^2, \quad \forall f\in\mathrm{H}^q_{mix}(\Omega), \tag{3.39}$$

respectively.

Proof. Using similar arguments to the $q=1$, $d\geq 1$ (Lemmas 2.9 and 2.10) and the $q\geq 1$, $d=1$ (Lemma 3.1) cases, it is easily confirmed that $\mathrm{D}^{q\beta}\mathcal{F}_N[f]$, $|\beta|_\infty\leq 1$, is precisely the truncated expansion of $\mathrm{D}^{q\beta}f$ in subpolyharmonic eigenfunctions that obey Dirichlet boundary conditions in the variables x_j when $\beta_j=1$ and Neumann boundary conditions otherwise. Convergence of $\mathrm{D}^{\beta q}\mathcal{F}_N[f]$ to $\mathrm{D}^{\beta q}f$ in the $\mathrm{L}^2(\Omega)$ norm now follows immediately from density and orthogonality of such eigenfunctions. □

As in the univariate setting (Section 3.4.1), the question of convergence of $\mathcal{F}_N[f]$ to f in the intermediate cases $\mathrm{H}^r(\Omega)$, $\mathrm{H}^r_{\mathrm{mix}}(\Omega)$, $r=1,\dots,q-1$, naturally arises. We devote the remainder of this section to this topic. To answer this question, we pursue a similar approach to the $d=1$ case, our first task being the generalisation of Lemma 3.15 to the d-variate cube:

Lemma 3.31. *Suppose that* $n\in\mathbb{N}^d$, $f\in\mathrm{L}^2(0,1)^d$ *and that* $a_n = \int_\Omega \prod_{j=1}^d \mathrm{e}^{z_j n_j x_j} f(x)\,\mathrm{d}x$, *where* $z_j\neq 0$ *and* $\mathrm{Re}\, z_j\leq 0$, $j=1,\dots,d$. *Then* $\{a_n\}\in l^2(\mathbb{N}^d)$ *and* $\sum_{n\in\mathbb{N}^d}|a_n|^2\leq c\|f\|^2$.

Proof. We first assume that $z_j = 2\pi \mathrm{i} c_j$, $j = 1, \dots, d$. Let $m_j \in \mathbb{N}$ be minimal such that $m_j/c_j \geq 1$. Extending f by zero to $[0, \frac{m_1}{c_1}] \times \ldots \times [0, \frac{m_d}{c_d}]$, we have

$$a_n = \sum_{i_1=1}^{m_1} \dots \sum_{i_d=1}^{m_d} \int_{\frac{i_1-1}{c_1}}^{\frac{i_1}{c_1}} \dots \int_{\frac{i_d-1}{c_d}}^{\frac{i_d}{c_d}} \prod_{j=1}^{d} \mathrm{e}^{2\pi \mathrm{i} c_j n_j x_j} f(x) \, \mathrm{d}x.$$

Each integral is the Fourier coefficient of the restriction of f to the corresponding hypercube. Hence, using Parseval's lemma, we immediately obtain the result.

Next we consider the case $\operatorname{Re} z_j < 0$, $j = 1, \dots, d$. As in the univariate setting, it suffices to show that $\sum_{n \in \mathbb{N}^d} |b_n|^2 \leq c\|f\|^2$ for all non-negative functions $f \in \mathrm{L}^2(\mathbb{R}_+^d)$, where

$$b_n = \int_{\mathbb{R}_+^d} \mathrm{e}^{-n.x} f(x) \, \mathrm{d}x.$$

Since $|b_n| \leq |b_m|$ when $m_j \leq n_j$, $j = 1, \dots, d$, the result holds, provided $\int_{\mathbb{R}_+^d} |b_t|^2 \, \mathrm{d}t \leq c\|f\|^2$. Now

$$\begin{aligned} \int_{\mathbb{R}_+^d} |b_t|^2 \, \mathrm{d}t &= \int_{\mathbb{R}_+^d} \int_{\mathbb{R}_+^d} \int_{\mathbb{R}_+^d} f(x) f(y) \mathrm{e}^{-t.(x+y)} \, \mathrm{d}t \, \mathrm{d}x \, \mathrm{d}y \\ &= \int_{\mathbb{R}_+^d} \int_{\mathbb{R}_+^d} f(x) f(y) \prod_{j=1}^{d} (x_j + y_j)^{-1} \, \mathrm{d}x \, \mathrm{d}y. \end{aligned}$$

Hence, as in the univariate case, it suffices to prove that $g \in \mathrm{L}^2(\mathbb{R}_+^d)$ with $\|g\| \leq c\|f\|$, where

$$g(x) = \int_{\mathbb{R}_+^d} f(x_1 y_1, \dots, x_d y_d) \prod_{j=1}^{d} (1 + y_j)^{-1} \, \mathrm{d}y.$$

We have

$$\begin{aligned} \|g\|^2 &= \int_{\mathbb{R}_+^d} \int_{\mathbb{R}_+^d} \int_{\mathbb{R}_+^d} f(x_1 w_1, \dots, x_d w_d) f(y_1 w_1, \dots, y_d w_d) \prod_{j=1}^{d} (1 + x_j)^{-1} (1 + y_j)^{-1} \, \mathrm{d}w \, \mathrm{d}x \, \mathrm{d}y \\ &\leq \|f\|^2 \int_{\mathbb{R}_+^d} \int_{\mathbb{R}_+^d} \prod_{j=1}^{d} (1 + x_j)^{-1} (1 + y_j)^{-1} x_j^{-\frac{1}{2}} y_j^{-\frac{1}{2}} \, \mathrm{d}x \, \mathrm{d}y \leq c\|f\|^2, \end{aligned}$$

as required.

Now suppose that $\operatorname{Re} z_j = 0$ for $j = 1, \dots, l$ and $\operatorname{Re} z_j < 0$ otherwise. We set $z_j = 2\pi \mathrm{i} c_j$, $j = 1, \dots, l$ and define m_j as before. Extending f by zero to $[0, \frac{m_1}{c_1}] \times \ldots \times [0, \frac{m_l}{c_l}] \times [0, 1] \times \ldots \times [0, 1]$ gives

$$a_n = \sum_{i_1=1}^{m_1} \dots \sum_{i_l=1}^{m_l} \int_{\frac{i_1-1}{c_1}}^{\frac{i_1}{c_1}} \dots \int_{\frac{i_l-1}{c_l}}^{\frac{i_l}{c_l}} \prod_{j=1}^{l} \mathrm{e}^{2\pi \mathrm{i} c_j n_j x_j} g_{n_{l+1}, \dots, n_d}(x_1, \dots, x_l) \, \mathrm{d}x_1 \dots dx_l,$$

where

$$g_{n_{l+1}, \dots, n_d}(x_1, \dots, x_l) = \int_0^1 \dots \int_0^1 \prod_{j=l+1}^{d} \mathrm{e}^{z_j n_j x_j} f(x) \, \mathrm{d}x_{l+1} \dots dx_d.$$

Hence, using the first result proved, we have

$$\begin{aligned}\sum_{n\in\mathbb{N}^d} |a_n|^2 &\leq c \sum_{n_{l+1}=1}^{\infty} \dots \sum_{n_d=1}^{\infty} \|g_{n_{l+1},\dots,n_d}\|^2 \\ &= c \int_0^1 \dots \int_0^1 \sum_{n_{l+1}=1}^{\infty} \dots \sum_{n_d=1}^{\infty} |g_{n_{l+1},\dots,n_d}(x_1,\dots,x_l)|^2 \,\mathrm{d}x_1 \dots dx_l.\end{aligned}$$

From the second result, it follows that

$$\sum_{n_{l+1}=1}^{\infty} \dots \sum_{n_d=1}^{\infty} |g_{n_{l+1},\dots,n_d}(x_1,\dots,x_l)|^2 \leq c \int_0^1 \dots \int_0^1 |f(x)|^2 \,\mathrm{d}x_{l+1} \dots dx_d.$$

Combining these observations now completes the proof. □

With this result to hand, generalisations of Lemma 3.16 and Corollaries 3.17, 3.18 and 3.19 now follow immediately. Since the proofs are identical, they are omitted.

Lemma 3.32. *Suppose that* $\{b_n\} \in l_2(\mathbb{N}^d)$. *Then, for any* $\operatorname{Re} z_j \leq 0$, $z_j \neq 0$, $j = 1,\dots,d$, *the family of finite sums of functions* $b_n \prod_{j=1}^d \mathrm{e}^{z_j n_j x_j}$, $n \in \mathbb{N}^d$, *is uniformly bounded in* $\mathrm{L}^2(0,1)$ *with norm bounded by* $c\left(\sum_{n\in\mathbb{N}^d} |b_n|^2\right)^{\frac{1}{2}}$ *for some* $c > 0$ *independent of* $\{b_n\}$.

Corollary 3.33. *Suppose that* $\{\psi_n\}$ *is the set of Cartesian products of the univariate polyharmonic eigenfunctions subject to boundary conditions (3.10) or (3.11) (with* p *not necessarily the same for each variable). Then, for* $f \in \mathrm{L}^2(\Omega)$, *the sequence* $\{(f,\psi_n)\} \in l^2(\mathbb{N}^d)$ *with norm bounded by* $c\|f\|$.

Corollary 3.34. *Suppose that* $\{\psi_n\}$ *and* $\{\chi_n\}$ *are sets of Cartesian products of biorthogonal pairs of univariate polyharmonic eigenfunctions subject to boundary conditions (3.10) and (3.11), respectively. Then the family of all finite sums of terms* $(f,\chi_n)\psi_n$ *is uniformly bounded in* $\mathrm{L}^2(\Omega)$ *with norm bounded by* $c\|f\|$.

Corollary 3.35. *Suppose that* $f \in \mathrm{H}^r(\Omega)$ *or* $f \in \mathrm{H}^r_{mix}(\Omega)$, $r = 0,\dots,q$, *and that* $\mathcal{F}_N[f]$ *is the truncated expansion of* f *in subpolyharmonic–Neumann eigenfunctions. Then* $\|\mathcal{F}_N[f]\|_r \leq c\|f\|_r$ *and* $\|\mathcal{F}_N[f]\|_{r,mix} \leq c\|f\|_{r,mix}$ *respectively.*

With these results to hand, we immediately deduce the key result of this section—the extension of Theorem 3.20 and Corollary 3.21. This is proved in an identical manner:

Theorem 3.36. *Suppose that* $f \in \mathrm{H}^r(\Omega)$, $r = 0,\dots,q$. *Then* $\mathcal{F}_N[f]$ *converges to* f *in the* $\mathrm{H}^r(\Omega)$ *norm. Moreover, if* $f \in \mathrm{H}^r_{mix}(\Omega)$, *then* $\mathcal{F}_N[f]$ *converges to* f *in the* $\mathrm{H}^r_{mix}(\Omega)$ *norm and* $\mathrm{D}^\beta \mathcal{F}_N[f]$ *converges uniformly to* $\mathrm{D}^\beta f$ *for* $|\beta|_\infty \leq r-1$.

In particular, $\mathcal{F}_N[f]$ converges uniformly to $f \in \mathrm{H}^1_{\mathrm{mix}}(\Omega)$, as in the $q = 1$ case. Theorem 3.36 therefore extends the modified Fourier result (Theorem 2.12) to arbitrary $q \geq 1$. This result completes our study of convergence of multivariate expansions. Next, we scrutinise the rate of convergence in various norms. In doing so, we highlight the central advantage of (sub)polyharmonic expansions over modified Fourier expansions: specifically, their faster convergence rate.

3.5.3 Rate of convergence

As in the modified Fourier case, we pursue two approaches: estimates based on the norm characterisations (3.38) and (3.39), and estimates which use the coefficient bound $|\hat{f}_n| \leq c\|f\|_{q+1,\mathrm{mix}}(\bar{n}_1 \dots \bar{n}_d)^{-q-1}$.[11] For both techniques, we will make use of the following multiplicative interpolation inequality for Sobolev norms (see, for example [2]):

$$\|f\|_r \leq c_{r,s}\|f\|^{1-\frac{r}{s}}\|f\|_s^{\frac{r}{s}}, \quad \forall f \in \mathrm{H}^s(\Omega). \tag{3.40}$$

Lemma 3.37. *Suppose that $f \in \mathrm{H}^r(\Omega)$ or $f \in \mathrm{H}^r_{mix}(\Omega)$ for $r = 0, \dots, q$. Then*

$$\|f - \mathcal{F}_N[f]\|_r \leq c \max_{n \notin I_N} \{\mu_n\}^{\frac{r-s}{2q}} \|f\|_s, \quad s = r, \dots, q,$$

and

$$\|f - \mathcal{F}_N[f]\|_r \leq c \max_{n \notin I_N} \left\{ \prod_{j=1}^d \left(1 + \mu_{n_j}\right) \right\}^{\frac{r-s}{2q}} \|f\|_{s,mix}, \quad s = r, \dots, q,$$

respectively.

Proof. We first prove this result for $r = 0$. By (3.38), we have $\|f - \mathcal{F}_N[f]\|^2 = \sum_{n \notin I_N} |\hat{f}_n|^2$. Observe that

$$\sum_{j=1}^d \alpha_{n_j}^{2s} |\hat{f}_n|^2 = \sum_{j=1}^d \left(\partial_{x_j}^s f, \Psi_{j,n}\right)^2,$$

where $\Psi_{j,n}(x) = \phi_{n_1}(x_1) \dots \phi_{n_{j-1}}(x_{j-1}) \psi_{n_j}(x_j) \phi_{n_{j+1}}(x_{j+1}) \dots \phi_{n_d}(x_d)$, ψ_{n_j} is the univariate polyharmonic eigenfunction equipped with boundary conditions (3.11), and $p = q - s$. It follows from Corollary 3.33 that

$$\sum_{n \notin I_N} \sum_{j=1}^d \left(\partial_{x_j}^s f, \Psi_{j,n}\right)^2 \leq c \sum_{j=1}^d \|\partial_{x_j}^s f\|^2 \leq c\|f\|_s^2.$$

We next note that $\sum_{j=1}^d \alpha_{n_j}^{2s} \geq c(\mu_n)^{\frac{s}{q}}$. Using this and the previous assertion, we deduce that

$$\|f - \mathcal{F}_N[f]\|_s^2 \leq c \max_{n \notin I_N} \{\mu_n\}^{-\frac{s}{q}} \sum_{n \notin I_N} \sum_{j=1}^d \alpha_{n_j}^{2s} |\hat{f}_n|^2 \leq c \max_{n \notin I_N} \{\mu_n\}^{-\frac{s}{q}} \|f\|_s^2,$$

which gives the result for $r = 0$. Now, suppose that $r = 1, \dots, s$. By (3.40) and the result for $r = 0$, we have

$$\|f - \mathcal{F}_N[f]\|_r \leq c\|f - \mathcal{F}_N[f]\|^{1-\frac{r}{s}} \|f - \mathcal{F}_N[f]\|_s^{\frac{r}{s}} \leq c \max_{n \notin I_N} \{\mu_n\}^{\frac{r-s}{q}} \|f\|_s^{1-\frac{r}{s}} \|f - \mathcal{F}_N[f]\|_s^{\frac{r}{s}}.$$

By Corollary 3.35, $\|f - \mathcal{F}_N[f]\|_s \leq \|f\|_s + \|\mathcal{F}_N[f]\|_s \leq c\|f\|_s$, which completes the proof for the classical Sobolev regularity case. The case of mixed smoothness is verified in an identical manner. □

[11] As in Section 2.7, an expansion of the multivariate subpolyharmonic coefficients can be found with few conceptual difficulties. An immediate consequence of such expansion is this bound.

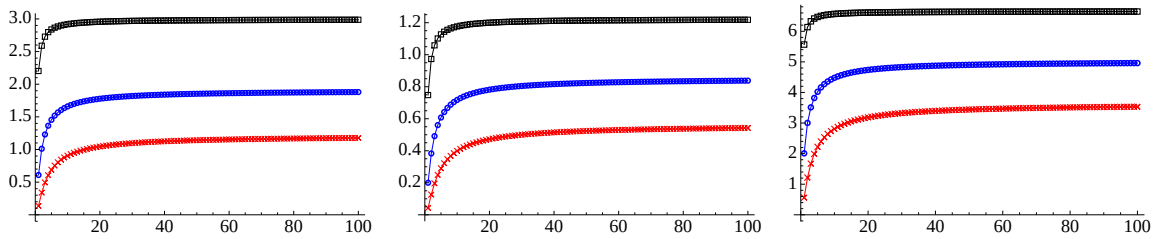


Figure 3.6: Error in approximating $f(x) = \mathrm{e}^{2x}$ by $\mathcal{F}_N[f](x)$ for $q = 1$ (squares), $q = 2$ (circles) and $q = 3$ (crosses). Left: scaled error $N^q \|f - \mathcal{F}_N[f]\|_\infty$ for $N = 1, \ldots, 100$. Middle: scaled error $N^{q+\frac{1}{2}} \|f - \mathcal{F}_N[f]\|$. Right: scaled error $N^{q-\frac{1}{2}} \|f - \mathcal{F}_N[f]\|_1$.

The estimates of this lemma, based on summation techniques, are independent of the index set. Of course, to obtain bounds involving N, we need to specify I_N. If we employ either the full (2.33) or hyperbolic cross (2.41) index sets (or indeed, the optimized hyperbolic cross (2.51)), then such bounds follow exactly as in the $q = 1$ case (see Sections 2.9 and 2.10). Once more, we obtain a now familiar result: given sufficient mixed regularity, the hyperbolic cross approximation converges at a rate comparable to its full counterpart.

Next, we obtain estimates based on coefficient bounds. In view of the advantage conveyed by the hyperbolic cross, we shall not consider the full index set further. Moreover, for the sake of simplicity, we will only consider the $\mathrm{L}^2(-1,1)$ norm hyperbolic cross (2.41), as opposed to the optimized hyperbolic cross (2.51).

Theorem 3.38. *Suppose that $f \in H^{q+1}_{mix}(\Omega)$ and that I_N is the hyperbolic cross (2.41). Then* $\|f - \mathcal{F}_N[f]\| \leq c\|f\|_{q+1,mix} N^{-q-\frac{1}{2}} (\log N)^{\frac{d-1}{2}}$,

$$\|f - \mathcal{F}_N[f]\|_r \leq c\|f\|_{q+1,mix} N^{r-q-\frac{1}{2}}, \quad r = 1, \ldots, q,$$

and $\|\mathrm{D}^\beta (f - \mathcal{F}_N[f])\|_\infty \leq c\|f\|_{q+1,mix} N^{|\beta|_\infty - q} (\log N)^{d-1}$ *for* $|\beta|_\infty \leq q - 1$.

Proof. This follows immediately from the bound $|\hat{f}_n| \leq c\|f\|_{q+1,\mathrm{mix}} (\bar{n}_1 \ldots \bar{n}_d)^{-q-1}$, Lemma 2.30, the norm characterisation (3.38), and the interpolation inequality (3.40). □

In Figure 3.6, we verify the results of Theorem 3.38 for the uniform, $\mathrm{L}^2(\Omega)$ and $\mathrm{H}^1(\Omega)$ norms, and $q = 1, 2, 3$. We remark in passing that estimates for the pointwise convergence rate away from the boundary $\partial\Omega$ can also be established along identical lines as Section 2.10.3. The resulting convergence rate is $\mathcal{O}\left(N^{-q-1} (\log N)^{d-1}\right)$, provided $f \in \mathrm{H}^{q+2}_{\mathrm{mix}}(\Omega)$.

As we address in the next section, such rates of convergence increase, provided the function f satisfies certain derivative conditions.

3.6 Derivative conditions

As in the $q = 1$ case, derivative conditions completely determine both the degree and rate of convergence of expansions in (sub)polyharmonic eigenfunctions. An indication of such derivative conditions was given in Section 3.4.2: the pointwise convergence rate was found to increase, provided

$$f^{(l)}(\pm 1) = 0, \quad \forall l \in N_{\rho_{k,p}}, \tag{3.41}$$

where $\rho_{k,p}$ and $N_{\rho_{k,p}}$ are given by (3.31) and (3.30) respectively. For this reason, we say that a function $f \in \mathrm{H}^{1+\rho_{k,p}}(-1,1)$ obeys *the first $\rho_{k,p}$ derivative conditions*, $k, p \in \mathbb{N}_0$, provided (3.41) holds.[12] A multivariate analogue is easily established. For $f \in \mathrm{H}^{1+\rho_{k,p}}(\Omega)$, the appropriate derivative conditions are given by

$$\partial_{x_j}^l f\big|_{\Gamma_j} = 0, \quad j = 1, \dots, d, \quad \forall l \in N_{\rho_{k,p}}. \tag{3.42}$$

To determine the effect of such derivative conditions on the convergence of $\mathcal{F}_N[f]$ to f, we first need to extend the modulo q norms of Section 3.5.1 to include higher-order derivatives D^β, where each component of β is an arbitrary multiple of q. We have

Lemma 3.39. *Suppose that $k \in \mathbb{N}$. Then the bilinear forms*

$$(f,g)_{k,q} = \sum_{|\beta| \leq k} \left(\mathrm{D}^{\beta q} f, \mathrm{D}^{\beta q} g \right), \quad f, g \in \mathrm{H}^{kq}(\Omega),$$
$$(f,g)_{k,q,mix} = \sum_{|\beta|_\infty \leq k} \left(\mathrm{D}^{\beta q} f, \mathrm{D}^{\beta q} g \right), \quad f, g \in \mathrm{H}^{kq}_{mix}(\Omega),$$

are inner products on the spaces $\mathrm{H}^{kq}(\Omega)$ and $\mathrm{H}^{kq}_{mix}(\Omega)$ respectively. The associated norms $\|\!|\cdot|\!\|_{k,q}$ and $\|\!|\cdot|\!\|_{k,q,mix}$ equivalent to $\|\cdot\|_{kq}$ and $\|\cdot\|_{kq,mix}$ respectively.

Proof. The case $k = 1$ has been established in Lemma 3.27 (note that $\|\!|\cdot|\!\|_{1,q}$ is just $\|\!|\cdot|\!\|_q$). Assume now that the result holds for $k-1$. In view of (3.36), it suffices to prove that $\|\mathrm{D}^\beta f\| \leq c\|f\|_{k,q}$ for all $|\beta| = kq$. For such β, we write $\beta = \gamma + \delta$, where $|\gamma| = (k-1)q$ and $|\delta| = q$. Then, using the induction hypothesis, we have $\|\mathrm{D}^\beta f\| \leq c\|\!|\mathrm{D}^\delta|\!\|_{k-1,q}$. By the result for $k = 1$, we obtain

$$\|\!|\mathrm{D}^\delta|\!\|^2_{k-1,q} = \sum_{|\beta| \leq k-1} \|\mathrm{D}^\delta \mathrm{D}^{\beta q} f\|^2 \leq c \sum_{|\beta| \leq k-1} \sum_{j=1}^d \left(\|\partial_{x_j}^q \mathrm{D}^{\beta q} f\|^2 + \|\mathrm{D}^{\beta q} f\|^2 \right) \leq c\|\!|f|\!\|^2_{k,q},$$

as required. The result for the mixed norms is verified in an identical manner to the proof of Lemma 3.28. □

Our first convergence result generalises Theorem 3.30:

Theorem 3.40. *Suppose that $k \in \mathbb{N}$, $l = 0, 1$, and that $f \in \mathrm{H}^{(2k+l)q}(\Omega)$ or $\mathrm{H}^{(2k+l)q}_{mix}(\Omega)$ obeys the first $\rho_{k,0}$ derivative conditions. Then $\mathcal{F}_N[f]$ converges to f in the $H^{(2k+l)q}(\Omega)$ and $H^{(2k+l)q}_{mix}(\Omega)$ norms respectively. Moreover, for $r = 0, \dots, 2k+l$, we have*

$$\|\!|f|\!\|^2_{r,q} = \sum_{n \in \mathbb{N}_0^d} \left[\sum_{|\beta| \leq q} \prod_{j=1}^d (\mu_{n_j})^{\beta_j} \right] |\hat{f}_n|^2,$$
$$\|\!|f|\!\|^2_{r,q,mix} = \sum_{n \in \mathbb{N}_0^d} \left[\sum_{|\beta|_\infty \leq q} \prod_{j=1}^d (\mu_{n_j})^{\beta_j} \right] |\hat{f}_n|^2,$$

respectively.

[12] There is a slight contradiction in terminology here with the definition (2.12) given in Chapter 2: namely, when $q = 1$, $\rho_{k,p} = 2k$ as opposed to k. However, it is convenient to define $\rho_{k,p}$ in this manner, so we shall proceed with this definition.

Proof. This follows immediately from now standard techniques. Repeated integration by parts and substitution of the boundary conditions give that $\mathrm{D}^{\beta q}\mathcal{F}_N[f]$ is the truncated expansion of $\mathrm{D}^{\beta q}f$ in subpolyharmonic eigenfunctions that obey either Dirichlet or Neumann boundary conditions in each variable x_j. □

As in Section 3.5.2, the result for the modulo q norms forms the basis of the argument in the general case. To obtain such a result, we first require a suitable extension of Bessel's inequality (Corollary 3.35) to arbitrary index $r \in \mathbb{N}_0$ (as opposed to just $r = 0, \ldots, q$):

Corollary 3.41. *Suppose that f obeys the first $\rho_{k,p}$ derivative conditions and $f \in \mathrm{H}^{\rho_{k,p}}(\Omega)$, $p \neq 0$, or $f \in \mathrm{H}^{2kq+l}(\Omega)$, $p = 0$, where $l = 0, \ldots, q$. Then $\|\mathcal{F}_N[f]\|_r \leq c\|f\|_r$ for $r = 0, \ldots, \rho_{k,p}$ or $r = 0, \ldots, 2kq + l$ respectively. The same result holds for the mixed spaces.*

Proof. For such a function f, the derivative $\mathrm{D}^{\beta}\mathcal{F}_N[f]$, where $|\beta| \leq \rho_{k,p}$ or $|\beta| \leq 2kq + l$ respectively, is the truncated expansion of $\mathrm{D}^{\beta}f$ in Cartesian products of polyharmonic eigenfunctions subject to boundary conditions (3.10) or (3.11) in each variable. The result now follows immediately from Corollary 3.34. □

We are now in a position to derive a full convergence result for polyharmonic functions obeying the derivative conditions (3.42). This result generalises Theorem 3.36, and its proof is identical:

Theorem 3.42. *Suppose that f is as in Corollary 3.41. Then $\mathcal{F}_N[f]$ converges to f in the $\mathrm{H}^r(\Omega)$ (respectively $\mathrm{H}^r_{mix}(\Omega)$) norm for $r = 0, \ldots, \rho_{k,p}$ or $r = 0, \ldots, 2kq + l$.*

Analogously, convergence in the uniform norm can also be established along identical lines to Theorem 3.36. We shall not give a full discussion as regards the rate of convergence, aside from mentioning that all convergence rates (given sufficient regularity) increase by factors of N^{2kq+p} over the general case, provided the function satisfies the first $\rho_{k,p}$ derivative conditions. For both the pointwise and uniform convergence rates, this fact was established in Theorem 3.25. Other cases can be proved along the same lines as Lemma 3.5.3 and Theorem 3.38.

This completes our study of convergence of (sub)polyharmonic–Neumann expansions. To finish this chapter, we next briefly detail the computation of polyharmonic–Neumann coefficients, using similar methods to those described in Section 2.12 for the modified Fourier case.

3.7 Quadrature

As discussed, a central reason why Birkhoff expansions have not been more extensively used is the lack of availability of robust means to compute the coefficients $\hat{f}_n$. In particular, unlike the modified Fourier case, the FFT cannot be used to compute polyharmonic–Neumann expansion coefficients. Nonetheless, the quadratures exhibited in Section 2.12 for the $q = 1$ case offer a compelling means to such perform such computations. This topic was pursued in greater detail in [8]. In this section, we describe a number of conspicuous aspects of that study. In particular, we demonstrate how the quadratures of Section 2.12 are successfully generalised. For the sake of simplicity, we focus on the univariate setting throughout.

As ever, our starting point is the asymptotic expansion (3.28). Immediately, it is apparent that only certain derivatives appear in this expansion: namely, those values $l \in N_{\rho_{k,p}}$, where

$N_{\rho_{k,p}}$ is given by (3.30). Any Filon-type quadrature scheme ought to reflect this fact. Hence, to this end, we let $-1 = c_1 < c_2 < \dots c_\nu = 1$ be given quadrature nodes and $m_1, \dots, m_\nu$ be their multiplicities. Moreover, we assume that each $m_j = \rho_{k_j,p_j}$ for some suitable k_j and p_j and that $m_1 = m_\nu = \rho_{k,p}$.

If p is a polynomial such that

$$p^{(l)}(c_s) = f^{(l)}(c_s), \quad \forall l \in N_{m_s}, \quad s = 1, \dots, \nu,$$

we define the Filon-type quadrature by

$$Q_{m,n}[f] = \int_{-1}^{1} p(x)\phi_n(x) \approx \hat{f}_n,$$

where $m = (m_1, \dots, m_\nu)$ is the vector of multiplicities. Upon comparison with (3.28), we note that the asymptotic order of this scheme is $(2k+1)q + p + 1$, and, since $\hat{f}_n = \mathcal{O}\left(n^{-q-1}\right)$, the relative asymptotic order is $2kq + p$.

As in Section 2.12, Filon-type methods for polyharmonic coefficients can be more easily designed as a combination of a truncated asymptotic expansion and a scaled approximation of certain derivatives [8]. However, for the sake of brevity, we shall not dwell on this issue.

Instead, we now briefly mention the design of appropriate exotic quadratures for the lower order coefficients. As in the $q = 1$ setting, the goal is to reuse derivative information in a classical quadrature scheme. In this spirit, we define

$$Q_m[g] = \sum_{r=1}^{\nu} \sum_{l \in N_{m_r}} b_{r,l} g^{(l)}(c_r) \approx \int_{-1}^{1} g(x)\,\mathrm{d}x, \tag{3.43}$$

where the coefficients $b_{r,l}$ are chosen to maximise order. Explicit numerical examples of both exotic and Filon-type quadratures for polyharmonic–Neumann coefficients are given in [8].

We mention in passing that (3.43) is a special case of *Birkhoff quadrature* [29]. Nonetheless, little theory currently exists pertaining to the maximal attainable order and the optimal location of quadrature nodes. As in the $q = 1$ case, a whole raft of questions remain regarding the design and implementation of both Filon and exotic quadratures for calculating polyharmonic–Neumann expansion coefficients.

This concludes our assessment of polyharmonic eigenfunction expansions. In the following two chapters we return primarily to the modified Fourier case. First, we consider the application of such expansions to the numerical solution of boundary value problems.

Chapter 4

Boundary value problems

4.1 Introduction

Orthogonal bases commonly find application in the numerical solution of partial differential equations. In this chapter, we describe in detail the application of Laplace eigenfunction expansions to the boundary value problem

$$\mathcal{L}[u](x) = f(x), \quad x \in \Omega, \quad \mathcal{B}[u] = 0, \tag{4.1}$$

where $\mathcal{L}$ is a linear, even-order differential operator, $\mathcal{B}[u] = 0$ are prescribed (nonperiodic) boundary conditions and $\Omega = (-1, 1)^d$ is the d-variate cube. In particular, the primary concern of this chapter is the linear, second-order advection-diffusion problem, where $\mathcal{L} = -\triangle + a \cdot \nabla + b\mathcal{I}$ and $\mathcal{I}$ is the identity operator (higher-order problems are addressed in Section 4.4.5).

Our approach to discretise (4.1) is a spectral–Galerkin technique: the solution u is expanded in a rapidly convergent series of basis functions, whose coefficients ensure that the residual $\mathcal{L}[u] - f$ is orthogonal to this basis. Standard spectral methods for (4.1) employ orthogonal polynomials (of typically Chebyshev or Legendre type) or, in the special case of periodic boundary conditions, Fourier series. Their principal benefit is so-called spectral convergence (faster than any algebraic power of N), thereby ensuring high accuracy at relatively low computational cost [31, 42].

In contrast, finite element methods—where the solution is expanded in low-order piecewise polynomials—converge only algebraically [45]. However, they are endowed with a number of important advantages over spectral methods, not least their flexibility and adaptability. Whilst high-order orthogonal polynomials and Fourier series are usually restricted to hypercubes, finite elements can be constructed in a wide variety of non-tensor-product domains, thus making such schemes applicable to complex geometries. Moreover, as we henceforth describe, finite elements schemes are more readily applicable to higher-dimensional problems than standard spectral approximations.

A central motivating factor in the development of modified Fourier expansions, as described in Chapter 1, is the design of spectral methods that offer both rapid convergence and flexibility. We defer a discussion of modified Fourier expansions in the equilateral triangle to Chapter 6. As considered in Chapter 2, modified Fourier expansions, when equipped with a hyperbolic cross index set, are well suited to problems in higher dimensions. For this reason,

the focus of this chapter is the discretisation of the boundary value problem (4.1), where $\Omega = (-1, 1)^d$ is the d-variate cube and $d \geq 1$ is arbitrary.

A primary concern in the development of spectral approximations is the question of satisfying boundary conditions. Since modified Fourier basis functions naturally obey homogeneous Neumann boundary conditions, they are most suitable for problems (4.1) endowed with the same boundary conditions. Conversely, given Dirichlet boundary conditions, we employ Laplace–Dirichlet eigenfunctions. Other boundary conditions are thus tackled by Laplace eigenfunctions subject to the same boundary conditions. The result is a family of methods for boundary value problems, each adapted to the particular boundary conditions. As in Chapter 2, the Dirichlet and Neumann problems will be our principal consideration. Other boundary conditions are considered towards the end of the chapter.

There are many alternative means to satisfy boundary conditions. Forming suitable linear combinations of basis functions is an approach commonly employed in Chebyshev or Legendre polynomial discretisations [146, 147]. However, this scheme typically leads to a loss of accuracy for Laplace eigenfunction approximations. Other techniques, based on either interpolating boundary conditions exactly or using so-called penalty schemes [81], become increasingly complicated in two or more dimensions. Instead, we pursue arguably the simplest approach: selecting basis functions that inherently satisfy the prescribed boundary conditions. We mention in passing that, since the solution u of (4.1) obeys the boundary conditions (in the language of Chapter 2, the first derivative condition), the approximation to a Dirichlet problem, for example, will converge uniformly throughout the domain.

The numerical solution of higher-dimensional problems has received significant attention of late. Such problems are a recurrent theme in a wide variety of applications, including fluid dynamics (the Navier–Stokes equations), quantum mechanics and computational chemistry (the Schrödinger equation) [41]. The main stumbling block towards effective discretisation of such problems is the previously mentioned exponential growth in computational cost with dimension (the curse of dimensionality, see Chapter 2). Though the design of so-called *sparse grid finite element methods*, where the numerical approximation consists of only $\mathcal{O}\left(N(\log N)^{d-1}\right)$ or even $\mathcal{O}(N)$ terms, is a mature field, few spectral methods currently exist that exhibit this property.[1] As described in Chapter 2, this shortfall can be attributed to the difficulty in rapidly computing only those coefficients of a function with indices from a hyperbolic cross. Classical spectral approximation schemes for (4.1) based on orthogonal polynomials (typically of Chebyshev or Legendre type [42]) are thus restricted in practice to $d = 1, 2, 3, 4$.

Nevertheless, as documented in Chapter 2, Laplace eigenfunction expansions may, in general, converge slowly. This translates into only algebraic convergence of the corresponding spectral approximation to the solution u of (4.1). However, the reduced computational cost of forming the approximation means that such an approach offers an advantage over more standard polynomial-based methods, as we shall confirm by numerical example. Unrelated to cost considerations, Laplace eigenfunction methods are also endowed with several beneficial features pertaining to the conditioning of the various matrices present, and the ease at which resulting linear systems can be solved. Such attributes are also chronicled in the sequel.

Needless to say, accelerating convergence of Laplace eigenfunction methods for boundary value problems, thereby increasing their effectiveness, is of singular importance. This topic is discussed further in Chapters 5 and 6. First, however, we must assess the methods in their

[1] Aside from the periodic case: the numerical solution of periodic partial differential equations by the so-called *sparse grid Fourier method* has been documented in [110].

most basic forms. This chapter is devoted to this topic.

The key results of the present chapter are as follows:

1. Spectral-Galerkin approximations to second order Neumann boundary value problems based on modified Fourier expansions exhibit $\mathcal{O}(N^{-\frac{5}{2}})$ errors in the $\mathrm{H}^1(\Omega)$ norm. An analogous result, albeit one power of N slower, holds for Laplace–Dirichlet approximations of Dirichlet problems.
2. Much like the Fourier method for periodic problems, methods based on Laplace eigenfunctions are reasonably well conditioned. In particular, discretisation matrices have $\mathcal{O}\left(N^2\right)$ condition numbers, and there exist optimal, diagonal preconditioners.
3. Provided a hyperbolic cross index set is used, Laplace eigenfunction approximations can be constructed in $\mathcal{O}\left(N^2\right)$ operations, *regardless* of d, using standard iterative methods. Due to this greatly reduced figure over the standard $\mathcal{O}\left(N^{d+1}\right)$ estimate for approximations based on full index sets, Laplace eigenfunction methods convey an advantage over standard polynomial-based spectral methods for moderate values of the parameter N.
4. Approximations based on Laplace eigenfunctions can be constructed for a variety of other boundary value problems, including numerous fourth and higher-order problems, for which they possess a number of advantages over more standard techniques.

The application of Laplace eigenfunctions to the numerical solution of boundary value problems was addressed in [3] (the univariate case) and [5] (the multivariate case). This chapter is based on those studies.

We remark in passing that such methods are not restricted solely to boundary value problems (4.1) in the d-variate cube. Other potential applications are outlined in Chapter 6. However, (4.1) presents the first stepping stone towards the design of effective methods based on Laplace eigenfunctions, and therefore remains our consideration throughout.

4.2 Spectral methods for boundary value problems

We commence with a brief review of the salient aspects of boundary value problems, specifically existence and uniqueness of solutions, and their numerical discretisation by spectral–Galerkin methods. There is an abundance of literature on this topic, and we refer the reader to [42] or [142], for example, for further details.

Consider the boundary value problem (4.1), where $f \in \mathrm{L}^2(\Omega)$ and $\mathcal{L}$ is a linear, even-order differential operator. We assume that the problem can be expressed in weak form as

$$\text{find } u \in \mathrm{H}(\Omega): \quad T(u,v) = (f,v), \quad \forall v \in \mathrm{H}(\Omega), \tag{4.2}$$

where $\mathrm{H}(\Omega)$ is some appropriate Hilbert space with norm $\|\cdot\|_{\mathrm{H}}$ and $T : \mathrm{H}(\Omega) \times \mathrm{H}(\Omega) \to \mathbb{R}$ is a bilinear form. Depending on their particular form, boundary conditions are either enforced by the definition of the operator T (so-called *natural* boundary conditions) or the space $\mathrm{H}(\Omega)$ itself (*essential* boundary conditions). Typically, Neumann boundary conditions are enforced in the former manner and Dirichlet boundary conditions by the latter.

Throughout, we assume that boundary conditions are homogeneous. If not, then, given some function g that satisfies $\mathcal{B}[g] = \mathcal{B}[u]$, we may decompose $u = v + g$, where the new function v is the solution of the homogeneous boundary value problem $\mathcal{L}[v] = f - \mathcal{L}[g]$,

$\mathcal{B}[v] = 0$. For the domains that we consider throughout this chapter, i.e. d-variate cubes, construction of an appropriate function g is relatively simple.[2]

Returning to (4.2), we now suppose that the form T satisfies the continuity and coercivity conditions

$$|T(u,v)| \leq \gamma \|u\|_{\mathrm{H}} \|v\|_{\mathrm{H}}, \quad T(u,u) \geq \omega \|u\|_{\mathrm{H}}^2, \quad \forall u, v \in \mathrm{H}(\Omega). \tag{4.3}$$

In this setting, existence and uniqueness of a solution to (4.2) is guaranteed by the well-renowned Lax–Milgram theorem:

Theorem 4.1 (Lax–Milgram)**.** *Suppose that* $\mathrm{H}(\Omega)$ *is a Hilbert space,* $f \in \mathrm{L}^2(\Omega)$ *and that the bilinear form* T *satisfies the continuity and coercivity conditions (4.3). Then there exists a unique solution to (4.2) satisfying the stability condition* $\|u\|_{\mathrm{H}} \leq \gamma\omega^{-1}\|f\|$.

With existence and uniqueness to hand, we now turn our attention to the numerical solution of (4.1). A standard approach is to approximate the solution u in some finite-dimensional space $\mathcal{S}_N = \mathrm{span}\{\phi_n : n \in I_N\}$, where ϕ_n are appropriate basis functions. Suppose that $u_N \in \mathcal{S}_N$ is the approximation to u. To specify u_N, we seek to make the residual $\mathcal{L}[u_N] - f$ small. There are numerous ways to realise this, but we will consider the Galerkin approach throughout: enforce that $\mathcal{L}[u_N] - f$ is orthogonal to $\mathcal{S}_N$.[3] In other words,

$$T(u_N, v) = (f, v), \quad \forall v \in \mathcal{S}_N. \tag{4.4}$$

These are referred to as *Galerkin's equations*. Since the approximation u_N satisfies a discretised version of (4.2), its existence and uniqueness are once more guaranteed by the Lax–Milgram theorem. Moreover, u_N also satisfies the stability estimate $\|u_N\|_{\mathrm{H}} \leq \gamma\omega^{-1}\|f\|$, thus ensuring that there is no blow-up in the numerical approximation, for example.

Convergence of the approximation u_N is guaranteed by Céa's Lemma:

Lemma 4.2 (Céa)**.** *Suppose that* $u_N \in \mathcal{S}_N$ *is the Galerkin approximation to (4.2). Then*

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

Céa's lemma reduces the question of convergence of the approximation u_N to merely a consideration of the approximation properties of the subspace $\mathcal{S}_N$. If $\mathcal{S}_N$ consists of Laplace eigenfunctions, for example, then convergence is therefore governed by the results of Chapter 2.

The Galerkin formulation encompasses both spectral and finite element methods. A key component of the former is to choose a basis $\mathcal{S}_N$ with a high degree of approximation. Basis functions are typically global, thus leading to small, dense matrices. Conversely, finite element basis functions are locally supported, thereby producing sparse, banded matrices. However, as a payoff, slow convergence of finite element approximations necessitates the solution of much larger linear systems to obtain reasonable accuracy [45].

[2] A subtraction function of the type subsequently considered in Chapter 5 can be used, for example.

[3] An alternative approach, leading to so-called *collocation* methods, is to enforce that the residual vanishes at a set of nodes. Typically, node locations are related classical quadrature formulae derived from the particular approximation basis. Collocation schemes are extremely effective in many situations (for example, nonlinear problems). However, for linear problems, at least, Galerkin schemes typically yield simpler discretisations and optimal error estimates [147].

Outside of convergence, the second central facet of Galerkin approximations is the question of computing the approximation u_N. To this end, suppose that $u_N = \sum_{n \in I_N} \bar{u}_n \phi_n$ has coefficients $\bar{u}_n \in \mathbb{R}$. If $\bar{u} \in \mathbb{R}^{|I_N|}$ is the vector of coefficients and $A_G \in \mathbb{R}^{|I_N| \times |I_N|}$ has $(n,m)^{\text{th}}$ entry $T(\phi_m, \phi_n)$, then Galerkin's equations (4.4) can be expressed in matrix form as $A_G \bar{u} = \hat{f}$, where $\hat{f} \in \mathbb{R}^{|I_N|}$ has n^{th} entry $\hat{f}_n$. We refer to A_G as the *Galerkin matrix*.[4]

A number of numerical considerations are of great importance in the computation of the spectral–Galerkin approximation u_N. First, numerical schemes must be available to calculate the entries of both the matrix A_G and the vector $\hat{f}_n$. In some cases (for example, where $\mathcal{L}$ has constant coefficients), the entries of the matrix A_G may be known explicitly. However, in the general case, they need to be approximated by numerical quadrature. If the quadrature used for this task is dependent on the truncation parameter N (for example, if the FFT were used), this leads to a so-called *Galerkin with numerical integration (GNI)* scheme [42]. We shall not pursue this approach. Instead, we use the numerical quadrature outlined in Section 2.12, where necessary, the accuracy of which is not automatically coupled to N.

The second practical consideration is how to solve Galerkin's equations efficiently. As stated, the matrix A_G is typically dense; hence, standard iterative methods can be expensive. Furthermore, the matrix is often ill-conditioned, making the construction of effective preconditioners paramount [42]. Nonetheless, A_G frequently inherits much of the structure of the continuous problem (4.2), thus commonly aiding both these tasks.

Having summarised the principal aspects of spectral–Galerkin schemes, we now address the discretisation of second-order boundary value problems with Laplace eigenfunctions.

4.3 Discretisation of second order boundary value problems

Let

$$\mathcal{L}[u](x) = -\triangle u(x) + a \cdot \nabla u(x) + bu(x) = f(x), \quad x \in \Omega = (-1,1)^d, \tag{4.5}$$

be a boundary value problem equipped with either homogeneous Neumann $\mathcal{B}[u] = \frac{\partial u}{\partial n}|_\Gamma$ or Dirichlet $\mathcal{B}[u] = u|_\Gamma$ boundary conditions (other boundary conditions are discussed in Section 4.4.4). For the sake of simplicity, we assume that $a = (a_1, \dots, a_d)^\top \in \mathbb{R}^d$ and $b \in \mathbb{R}$ are constant; in the sequel, we address the variable-coefficient case.

Both the Dirichlet and Neumann problems share the common bilinear form

$$T(u,v) = (\nabla u, \nabla v) + (a \cdot \nabla u, v) + (bu, v), \tag{4.6}$$

where $(\nabla u, \nabla v) = \int_\Omega \nabla u \cdot \nabla v$. In the Neumann case, since the boundary conditions are essential, we let $\mathrm{H}(\Omega) = \mathrm{H}^1(\Omega)$, so that $T : \mathrm{H}^1(\Omega) \times \mathrm{H}^1(\Omega) \to \mathbb{R}$. Conversely, in the Dirichlet case $T : \mathrm{H}^1_0(\Omega) \times \mathrm{H}^1_0(\Omega) \to \mathbb{R}$, where the space $\mathrm{H}^1_0(\Omega)$ is defined in the usual manner as the closure of $\mathrm{C}^\infty_0(\Omega)$ in $\mathrm{H}^1(\Omega)$.

Continuity and coercivity of these forms are readily established:

Lemma 4.3. *Suppose that T is defined by (4.6). Then T is $\mathrm{H}^1_0(\Omega)$-continuous and coercive (in other words, the restriction of T to $\mathrm{H}^1_0(\Omega) \times \mathrm{H}^1_0(\Omega)$ is continuous and coercive in the $\mathrm{H}^1_0(\Omega)$*

[4] In general, $n = (n_1, \dots, n_d)$ is a multi-index. Hence, for practical purposes, some ordering is given to the index set I_N. The choice of such ordering can impact upon the numerical behaviour of the matrix A_G [42]. However, we shall not address this issue further.

norm) if and only if $b > -C^{-2}$*, where* $C = 2d^{-\frac{1}{2}}\pi^{-1}$ *is the constant of Poincaré's inequality [56]. Specifically,*

$$|T(u,v)| \le \left(1 + C\|a\| + C^2|b|\right)|u|_1|v|_1, \quad T(u,u) \ge \min\{1, 1+bC^2\}|u|_1^2, \quad \forall u, v \in \mathrm{H}_0^1(\Omega),$$

where $|\cdot|_1$ *is the standard norm on* $\mathrm{H}_0^1(\Omega)$ *and* $\|a\|^2 = \sum_{j=1}^d a_j^2$*.*

Proof. We first recall Poincaré's inequality $\|u\| \le C|u|_1$, $\forall u \in \mathrm{H}_0^1(\Omega)$. Note that the constant in this inequality is sharp.[5] We have

$$|T(u,v)| \le \|\nabla u\|\|\nabla v\| + \|a\|\|\nabla u\|\|v\| + |b|\|u\|\|v\| \le |u|_1|v|_1 + C\|a\||u|_1|v|_1 + C^2|b||u|_1|v|_1,$$

which gives the first result. For the second, we first note that $(a\cdot\nabla u, u) = u^2 a\cdot\hat{n}|_\Gamma = 0$, since $u|_\Gamma = 0$. Here $\hat{n}$ is the unit outward normal vector.[6] Hence

$$T(u,u) = |u|_1^2 + b\|u\|^2 \ge \min\{1, 1+bC^2\}|u|_1^2,$$

as required. To show that the condition $b > -C^{-2} = -\frac{1}{4}d\pi^2$ is also necessary, we consider the function $u(x) = \prod_{j=1}^d \cos\frac{1}{2}\pi x_j$. In this case $\|u\|^2 = 1$ and $|u|_1^2 = \|\nabla u\|^2 = \frac{1}{4}d\pi^2 = C^{-2}$, which gives

$$T(u,u) = \|\nabla u\|^2 + b\|u\|^2 = \left(1 + bC^2\right)|u|_1^2.$$

Hence T cannot be coercive if $b \le -C^{-2}$. □

Lemma 4.4. *The bilinear* T *form defined by (4.6) is* $\mathrm{H}^1(\Omega)$*-continuous and coercive if and only if* $b - \frac{1}{4}\|a\|^2 > 0$*. Specifically*

$$|T(u,v)| \le 2\max\{1, b, \|a\|\}\|u\|_1\|v\|_1, \quad T(u,u) \ge \left(b - \tfrac{1}{4}\|a\|^2\right)\min\left\{\left(b + \tfrac{1}{4}\|a\|^2\right)^{-1}, \tfrac{1}{2}\right\}\|u\|_1^2,$$

for all $u, v \in \mathrm{H}^1(\Omega)$*.*

Proof. The proof of continuity is virtually identical to that of the previous lemma. For coercivity we use Young's inequality[7] to give

$$\begin{aligned} T(u,u) = \|\nabla u\|^2 + (a\cdot\nabla u, u) + b\|u\|^2 &\ge \|\nabla u\|^2 - \left(\epsilon\|\nabla u\|^2 + \tfrac{1}{4\epsilon}\|a\|^2\|u\|^2\right) + b\|u\|^2 \\ &= (1-\epsilon)\|\nabla u\|^2 + \left(b - \tfrac{1}{4\epsilon}\|a\|^2\right)\|u\|^2, \end{aligned}$$

for all $\epsilon > 0$. If we set $\epsilon = \|a\|^2(2b + \frac{1}{2}\|a\|^2)^{-1}$ and substitute this into the above expression, we obtain

$$|T(u,u)| \ge \frac{b - \frac{1}{4}\|a\|^2}{b + \frac{1}{4}\|a\|^2}\|\nabla u\|^2 + \frac{b - \frac{1}{4}\|a\|^2}{2}\|u\|^2,$$

as required. To show that the condition $b - \frac{1}{4}\|a\|^2 > 0$ is also necessary, consider the function $u(x) = \mathrm{e}^{\frac{1}{2}(1+\sqrt{2})x.a}$. In this case $T(u,u) = \left(b - \frac{1}{4}\|a\|^2\right)\|u\|^2$. Hence, no other lower bound is permissible. □

[5] This constant is precisely the square root of the reciprocal of the smallest eigenvalue of the Laplace operator on $\Omega = (-1,1)^d$ subject to homogeneous Dirichlet boundary conditions. This is easily proved by expanding an arbitrary function $u \in \mathrm{H}_0^1(\Omega)$ in Laplace–Dirichlet eigenfunctions. Incidentally, this result holds for *any* bounded, convex domain [135]. Note that Poincaré's inequality guarantees that the semi-norm $|\cdot|_1$ is in fact a norm on $\mathrm{H}_0^1(\Omega)$.

[6] We write $u|_\Gamma$ with the understanding that this refers to the trace of u on $\Gamma = \partial\Omega$. Further considerations of trace operators are not necessary for this particular study, and we refer the reader to [56] for details.

[7] The algebraic inequality $xy \le \frac{1}{4\epsilon}x^2 + \epsilon y^2$ for all $\epsilon > 0$, $x, y \in \mathbb{R}$, is referred to as *Young's inequality* [124].

Under the assumptions of Lemma 4.3 and Lemma 4.4 respectively, existence and uniqueness of a weak solution to the homogeneous Dirichlet and Neumann problems are guaranteed by the Lax–Milgram theorem.

We remark in passing that the Dirichlet problem may be reduced to a canonical form by writing

$$u(x) = \mathrm{e}^{\frac{1}{2}\sum_{j=1}^{d} a_j x_j} v(x).$$

The new function $v \in \mathrm{H}_0^1(\Omega)$ is the solution to the boundary value problem

$$-\triangle v(x) + \left(b + \tfrac{1}{4}\|a\|^2\right) v(x) = \mathrm{e}^{-\frac{1}{2}\sum_{j=1}^{d} a_j x_j} f(x), \quad v\big|_{\partial\Omega} = 0.$$

This equation, representing a Helmholtz problem, has no advection term and is therefore simpler. In particular, the discretisation of this problem with Laplace–Dirichlet eigenfunctions possesses a diagonal matrix. However, though theoretically possible, in the variable-coefficient case this transformation requires the computation of indefinite integrals of the functions $a_j(x)$. Further, the Laplace–Dirichlet discretisation now leads to a full matrix. For this reason, we shall consider the non-canonical formulation (4.5) throughout. Note that the Neumann problem cannot readily be reduced to a canonical form in this manner, since the boundary conditions are not preserved by the above transformation.

4.3.1 The Galerkin approximation

We now seek an approximation $u_N \in \mathcal{S}_N$ to the Dirichlet and Neumann problems, where $\mathcal{S}_N$ consists of Laplace–Dirichlet or Laplace–Neumann eigenfunctions respectively. We write

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

in each case, and refer to u_N as the *Laplace–Dirichlet Galerkin* (respectively *Laplace–Neumann Galerkin/modified Fourier–Galerkin*) approximation. Throughout the remainder of this chapter, we assume that either the full (2.33) or hyperbolic cross (2.41) index set is employed. The majority of the results proved can be immediately applied or suitably adapted to other index sets, including the optimized hyperbolic cross (2.51). However, for the sake of simplicity, we shall not pursue this further.

In both the Dirichlet and Neumann cases, the coefficients $\bar{u}_n^{[i]} \in \mathbb{R}$ are specified by Galerkin's equations (4.4). For this, we have the following lemma:

Lemma 4.5. *The coefficients $\bar{u}_n^{[i]}$ of the approximation u_N satisfy*

$$(b + \mu_n^{[i]})\bar{u}_n^{[i]} + \sum_{j=1}^{d} \sum_{\substack{m_j\in\mathbb{N} \\ (n;m_j)\in I_N}} a_j \delta_{n_j,m_j}^{[i_j]} \bar{u}_{(n;m_j)}^{[(i;1-i_j)]} = \hat{f}_n^{[i]}, \quad i \in \{0,1\}^d, \quad n \in I_N, \tag{4.7}$$

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

$$\delta_{n,m}^{[i]} = 2(-1)^{n+m} \frac{\alpha_n^{[1-i]}\alpha_m^{[i]}}{\mu_n^{[1-i]} - \mu_m^{[i]}}, \quad \delta_{n,m}^{[i]} = 2(-1)^{n+m} \frac{\mu_m^{[1-i]}}{\mu_n^{[i]} - \mu_m^{[1-i]}}, \quad i \in \{0,1\}, \quad n,m \in \mathbb{N}_0,$$

in the Dirichlet and Neumann cases respectively.

Proof. Since the Dirichlet and Neumann cases are identical, we consider the latter. Setting $v = \phi_n^{[i]}$, $i \in \{0,1\}^d$, $n \in I_N$, in Galerkin's equations (4.4) gives

$$T(u_N, \phi_n^{[i]}) = (b + \mu_n^{[i]})a_n^{[i]} + \sum_{j=1}^{d} \sum_{l\in\{0,1\}^d} \sum_{m\in I_N} a_j(\partial_{x_j}\phi_m^{[l]}, \phi_n^{[i]})a_m^{[l]}.$$

Here the first term arises as a direct consequence of the fact that the approximation basis consists of orthonormal Laplace eigenfunctions. Now,

$$(\partial_{x_j}\phi_m^{[l]}, \phi_n^{[i]}) = ((\phi_{m_j}^{[l_j]})', \phi_{n_j}^{[i_j]}) \prod_{k\neq j} (\phi_{m_k}^{[l_k]}, \phi_{n_k}^{[i_k]}) = \begin{cases} \delta_{n_j,m_j}^{[i_j]} & l = (i; 1-i_j),\ m_k = n_k,\ k \neq j, \\ 0 & \text{otherwise,} \end{cases}$$

which gives the result. □

In the univariate setting, if $\bar{u} = (\bar{u}^{[0]}, \bar{u}^{[1]})^\top$ is the vector with entries $\bar{u}_n^{[i]}$ and $\hat{f} = (\hat{f}^{[0]}, \hat{f}^{[1]})^\top$ is the vector of coefficients $\hat{f}_n^{[i]}$, Galerkin's equations can be written in matrix form as $A_{\mathrm{G}}\bar{u} = \hat{f}$, where

$$A_{\mathrm{G}} = \begin{pmatrix} D^{[0]} & a\delta^{[0]} \\ a\delta^{[1]} & D^{[1]} \end{pmatrix}. \tag{4.8}$$

Here $\delta^{[i]}$ is the $(N+1-i)\times(N+i)$ matrix with entries $\delta_{n,m}^{[i]}$ and $D^{[i]}$ is the $(N+1-i)\times(N+1-i)$ diagonal matrix with entries $b + \mu_n^{[i]}$. The diagonal blocks of this matrix correspond to the restriction of the operator $\mathcal{L}_0 = -\partial_{xx} + b\mathcal{I}$, where $\mathcal{I}$ is the identity operator, to $\mathcal{S}_N$. The off-diagonal blocks correspond to the advection operator $\mathcal{L}_1 = a\partial_x$. For this reason, we define

$$M_{\mathrm{G}} = \begin{pmatrix} D^{[0]} & 0 \\ 0 & D^{[1]} \end{pmatrix}, \quad N_{\mathrm{G}} = \begin{pmatrix} 0 & a\delta^{[0]} \\ a\delta^{[1]} & 0 \end{pmatrix}, \tag{4.9}$$

as the matrices of these actions. Note that $A_{\mathrm{G}} = M_{\mathrm{G}} + N_{\mathrm{G}}$.

The operator splitting $\mathcal{L} = \mathcal{L}_0 + \mathcal{L}_1$ is fundamental to a number of both practical and analytical aspects of the modified Fourier–Galerkin method. The same splitting is also employed in the multivariate setting upon defining $\mathcal{L}_0 = -\triangle + b\mathcal{I}$ and $\mathcal{L}_1 = a \cdot \nabla$. As before, we write $A_{\mathrm{G}} = M_{\mathrm{G}} + N_{\mathrm{G}}$, where M_{G} and N_{G} are the matrices corresponding to the operators $\mathcal{L}_0$ and $\mathcal{L}_1$. The first term of (4.7) verifies that M_{G} is diagonal once more with n^{th} entry $b + \mu_n^{[i]}$. Note that the matrix M_{G} is symmetric (a fact independent of the discretisation: determined solely by the self-adjointness of the operator $\mathcal{L}_0$), whilst N_{G} is skew-symmetric for Dirichlet problems. Hence, A_{G} is not symmetric in general, aside from the case $a = 0$, i.e. the Helmholtz problem.

When $a = 0$, u_N may be reinterpreted as precisely $\mathcal{F}_N[u]$. Indeed, it follows from (4.7) that $\bar{u}_n^{[i]} = (b+\mu_n^{[i]})^{-1}\hat{f}_n^{[i]}$, which, upon comparison with (4.5), is nothing more than $\hat{u}_n^{[i]}$ (or $\check{u}_n^{[i]}$ for the Dirichlet problem). In this case, the construction and analysis of the approximation u_N is completely accounted for by the techniques of Chapter 2. Nonetheless, when $a \neq 0$ (or when $b = b(x)$ is not constant) this ceases to be the case, and new techniques are required to address the approximation u_N. The next two sections are devoted to the construction of this approximation. Analysis of convergence is considered in Section 4.3.4.

4.3.2 Properties of the discretisation matrix

For an arbitrary nonsingular matrix $A \in \mathbb{R}^{M\times M}$ we define the spectral condition number $\kappa_s(A)$ as the ratio of the largest and smallest eigenvalues in absolute value. We refer to the quantity

$$\kappa(A) = \sqrt{\frac{\lambda_{\max}(A^\top A)}{\lambda_{\min}(A^\top A)}},$$

as the condition number.[8] For the Galerkin matrix A_{G}, the size of these quantities are of singular interest, since they determine the impact of round-off errors in the solution of Galerkin's equations [42]. They also determine the convergence rate of various iterative solution techniques and the necessary step-size restrictions in fully-discretised approximations to time-dependent problems (see Chapter 6). Note that when A_{G} is symmetric (i.e. $a = 0$), the quantities $\kappa_s(A_{\mathrm{G}})$ and $\kappa(A_{\mathrm{G}})$ coincide. However, for $a \neq 0$ this ceases to be the case.

In standard spectral discretisations, the (spectral) condition number is often rather large, so the design of effective preconditioners is imperative. We say that an invertible matrix $P \in \mathbb{R}^{M\times M}$ is an *optimal* (spectral) preconditioner for A if the (spectral) condition number of the matrix AP^{-1} is $\mathcal{O}(1)$ as $M \to \infty$.[9] A fundamental consideration in the design of effective preconditioners is that the linear system $Px = y$ should be 'easier' to solve (i.e. of lower computational cost) than the linear system $Ax = b$.

In this section, we demonstrate that the (spectral) condition number of the Galerkin matrix A_{G} arising from Laplace eigenfunction discretisations is $\mathcal{O}\left(N^2\right)$. Moreover, there exists an optimal, right preconditioner given by the matrix M_{G}. Since this preconditioner is diagonal, it is cheap and simple to apply.

We commence with estimates for the spectral condition number:

Lemma 4.6. *Suppose that I_N is either the full (2.33) or the hyperbolic cross (2.41) index set. Then the spectral condition number $\kappa_s(A_{\mathrm{G}})$ is $\mathcal{O}\left(N^2\right)$ for both the Dirichlet and Neumann problems, provided the operator T is coercive. Specifically,*

$$\kappa_s(A_{\mathrm{G}}) \leq \frac{\gamma}{\omega}\left(1 + N^2\pi^2 d\right), \quad \kappa_s(A_{\mathrm{G}}) \leq \frac{\gamma}{\omega}\left(1 + (d - 1 + N^2)\pi^2\right),$$

in the full and hyperbolic cross cases respectively, where γ and ω are the constants of continuity and coercivity.

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 2.11), $\|u\|_1^2 \leq \max_{n\in I_N}\{1 + \mu_n^{[0]}\}\|u\|^2$. Moreover, for $n \in I_N$,

$$1 + \mu_n^{[0]} \leq 1 + N^2\pi^2 d, \quad 1 + \mu_n^{[0]} \leq 1 + (d - 1 + N^2)\pi^2, \tag{4.10}$$

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

We may also prove the same result for the condition number $\kappa(A_{\mathrm{G}})$. To do so, we first require the following lemma:

[8] More precisely, since $\kappa(A) = \|A\|\|A^{-1}\|$, where $\|\cdot\|$ is the L^2 matrix norm, this is the L^2 condition number.

[9] Specifically, P is a *right* preconditioner. If we consider the quantities $P^{-1}A$ or $P^{-\frac{1}{2}}A(P^{-\frac{1}{2}})^\top$ instead, then P is referred to as a *left* or *symmetric* preconditioner respectively [142].

Lemma 4.7. *Suppose that λ is an eigenvalue of $A_{\mathrm{G}}^{\top}A_{\mathrm{G}}$ with associated eigenfunction $u \in \mathcal{S}_N$. Then*

$$(\mathcal{F}_N[\mathcal{L}[u]], \mathcal{F}_N[\mathcal{L}[\phi]]) = \lambda(u, \phi), \quad \forall \phi \in \mathcal{S}_N. \tag{4.11}$$

Proof. Dropping the i superscript for ease of notation, the matrix A_{G} has $(n,m)^{\text{th}}$ entry $(\mathcal{L}[\phi_m], \phi_n)$. Let $u \in \mathcal{S}_N$ be an eigenfunction with eigenvalue λ. Then

$$\sum_{m \in I_N} \left(A_{\mathrm{G}}^{\top}A_{\mathrm{G}}\right)_{n,m} (u, \phi_m) = \lambda(u, \phi_n), \quad \forall n \in I_N.$$

Expanding the left-hand side gives

$$\begin{aligned}\sum_{m \in I_N} \left(A_{\mathrm{G}}^{\top}A_{\mathrm{G}}\right)_{n,m} (u, \phi_m) &= \sum_{m,r \in I_N} (\mathcal{L}[\phi_n], \phi_r)\,(\mathcal{L}[\phi_m], \phi_r)\,(u, \phi_m) \\ &= \sum_{r \in I_N} (\mathcal{L}[\phi_n], \phi_r)\,(\mathcal{L}[\mathcal{F}_N[u]], \phi_r) = (\mathcal{L}[\mathcal{F}_N[u]], \mathcal{F}_N[\mathcal{L}[\phi_n]])\,.\end{aligned}$$

Hence $(\mathcal{L}[\mathcal{F}_N[u]], \mathcal{F}_N[\mathcal{L}[\phi_n]]) = \lambda(u, \phi_n)$ for all $n \in I_N$. Linearity now gives the result. □

Theorem 4.8. *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 $\mathcal{O}\left(N^2\right)$ in both the Dirichlet and Neumann cases, provided the operator $\mathcal{L}$ is coercive. Specifically, if γ' is a positive constant such that $\|\mathcal{L}[u]\|^2 \leq \gamma'\|u\|_2^2$ for all $u \in \mathrm{H}^2(\Omega)$, then we have the bounds*

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

in the full and hyperbolic cross cases respectively, where ω is the coercivity constant of $\mathcal{L}$.

Proof. Setting $\phi = u$ in (4.11) gives $\|\mathcal{F}_N[\mathcal{L}[u]]\|^2 = \lambda\|u\|^2$. Now, by the duality pairing (3.25),

$$\|\mathcal{F}_N[\mathcal{L}[u]]\| = \sup_{g \in \mathrm{L}^2(\Omega)} \frac{(\mathcal{F}_N[\mathcal{L}[u]], g)}{\|g\|} \geq \sup_{g \in \mathcal{S}_N} \frac{(\mathcal{L}[u], g)}{\|g\|}. \tag{4.12}$$

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

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

To derive an upper bound for λ, we note that

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

by Bernstein's Inequality. The result now follows from (4.10). □

The proofs of Lemma 4.6 and Theorem 4.8 highlight that the minimal eigenvalues of both A_{G} and $A_{\mathrm{G}}^{\top}A_{\mathrm{G}}$ are independent of the Galerkin discretisation employed. In particular, they are independent of the index set. The upper bounds, however, rely on Bernstein-type estimates which are dependent on both the discretisation basis and index set employed.

The constant γ' defined in Theorem 4.8 exists regardless of any assumptions on the values a and b (much like the standard continuity constant γ). It represents a continuity constant for the *normal* form $T' : \mathrm{H}^2(\Omega) \times \mathrm{H}^2(\Omega) \to \mathbb{R}$, where $T'(u, v) = (\mathcal{L}[u], \mathcal{L}[v])$. Note that an explicit value for γ' is given by $\gamma' = 3\max\{1, \|a\|, |b|\}^2$.

Next, we assess the preconditioner M_{G} for the matrix A_{G}. To do so, it is first necessary to establish coercivity for a variety of normal forms similar to that introduced above. We have

Lemma 4.9. *Suppose that* $b - \frac{1}{4}\|a\|^2 > 0$ *and that* $\mathcal{L}_0 = -\triangle + b\mathcal{I}$. *Then*

$$(\mathcal{L}[u], \mathcal{L}_0[u]) \geq \omega'\|u\|_2^2, \quad \forall u \in \mathrm{H}^2(\Omega), \quad \frac{\partial u}{\partial n}\Big|_{\partial\Omega} = 0,$$

for some positive constant ω' *given explicitly by*

$$\omega' = \left(b - \tfrac{1}{4}\|a\|^2\right) \min\left\{1, \left(b + \tfrac{1}{4}\|a\|^2\right)^{-1}, b^2\left(b + \tfrac{1}{4}\|a\|^2\right)^{-1}\right\}.$$

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

$$\|\mathcal{L}_0[u]\|^2 = \|\triangle u\|^2 + 2b\|\nabla u\|^2 + b^2\|u\|^2,$$

and

$$|(\mathcal{L}_1[u], \mathcal{L}_0[u])| \leq |(a \cdot \nabla u, \triangle u)| + b|(a \cdot \nabla u, u)| \leq \|a\|\|\nabla u\|\|\triangle u\| + b\|a\|\|\nabla u\|\|u\|.$$

Using Young's inequality we obtain

$$|(\mathcal{L}_1[u], \mathcal{L}_0[u])| \leq \epsilon\|\triangle u\|^2 + \tfrac{\|a\|^2}{2\epsilon}\|\nabla u\|^2 + b^2\epsilon\|u\|^2, \quad \forall \epsilon > 0.$$

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. □

Lemma 4.10. *Suppose that* $b > 0$. *Then*

$$(\mathcal{L}[u], \mathcal{L}[u]) = \|\mathcal{L}[u]\|^2 \geq \omega'\|u\|_2^2, \quad \forall u \in \mathrm{H}^2(\Omega) \cap \mathrm{H}_0^1(\Omega),$$

with positive constant $\omega' = b\min\{1, b, (\|a\|^2 + b)^{-1}\}$.

Proof. We have

$$\|\mathcal{L}[u]\|^2 = \|\triangle u\|^2 + 2b\|\nabla u\|^2 + b^2\|u\|^2 + 2(a\cdot\nabla u, -\triangle u + bu) + \|a\cdot\nabla u\|^2.$$

Since $u \in \mathrm{H}_0^1(\Omega)$, the term $(a\cdot\nabla u, bu)$ vanishes. An application of Young's inequality now gives

$$\begin{aligned}\|\mathcal{L}[u]\|^2 &\geq (1-\epsilon)\|\triangle u\|^2 + 2b\|\nabla u\|^2 + (1-\epsilon^{-1})\|a\cdot\nabla u\|^2 + b^2\|u\|^2\\ &\geq (1-\epsilon)\|\triangle u\|^2 + (2b + (1-\epsilon^{-1})\|a\|^2)\|\nabla u\|^2 + b^2\|u\|^2.\end{aligned}$$

Setting $\epsilon = \|a\|^2(b+\|a\|^2)^{-1}$ completes the proof. □

With these results to hand, we may now confirm M_{G} as an optimal preconditioner for both the Dirichlet and Neumann problems. We commence with the latter:

Theorem 4.11. *Suppose that A_{G} is the Galerkin matrix for the Neumann problem. Then, the right preconditioner M_{G} is optimal for the (spectral) condition number, provided the operator T is coercive. Specifically,*

$$\kappa_s(A_{\mathrm{G}}) \leq \frac{\gamma\max\{1,b\}}{\omega\min\{1,b\}}, \quad \kappa(A_{\mathrm{G}}) \leq \sqrt{\frac{\gamma'\max\{1,2b,b^2\}}{\omega'\min\{1,2b,b^2\}}},$$

where γ' and ω' are the constants of Theorem 4.8 and Lemma 4.9 respectively.

Proof. Suppose that λ is an eigenvalue of $A_{\mathrm{G}}M_{\mathrm{G}}^{-1}$ with eigenfunction $u \in \mathcal{S}_N$. Suppose further that $u = (-\triangle + b\mathcal{I})v$ for some $v \in \mathcal{S}_N$. Then

$$(\mathcal{L}[v],\phi) = \lambda(\mathcal{L}_0[v],\phi), \quad \forall\phi \in \mathcal{S}_N.$$

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

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

which gives the first result. 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 Lemma 4.7, 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 $\|\mathcal{L}_0[u]\|^2 \geq \min\{1,2b,b^2\}\|u\|_2^2$ for all $u \in \mathrm{H}^2(-1,1)^d$ satisfying $\frac{\partial u}{\partial n}|_\Gamma = 0$. Hence

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

which yields $\lambda \leq \gamma'(\min\{1,2b,b^2\})^{-1}$. To provide a lower bound, we use (4.12) with $g = \mathcal{L}_0[v]$ to give

$$\|\mathcal{F}_N[\mathcal{L}[v]]\| \geq \frac{(\mathcal{L}[v],\mathcal{L}_0[v])}{\|\mathcal{L}_0[v]\|}.$$

Applications of Lemma 4.9 and the continuity condition for $\mathcal{L}_0$ now yield

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

Hence $\lambda \geq (\omega')^2(\max\{1,2b,b^2\})^{-1}$ and the proof is complete. □

Theorem 4.12. *Suppose that A_{G} is the Galerkin matrix for the Dirichlet problem. Then the right preconditioner M_{G} is optimal for the spectral condition number, provided the operator T is coercive. It is optimal for the condition number, provided $b > 0$. Specifically, $\kappa_s(A_{\mathrm{G}}) \leq \frac{\gamma}{\omega}$, and, for sufficiently large N,*

$$\kappa(A_{\mathrm{G}}) \leq \sqrt{\frac{2\gamma' \max\{1, 2b, b^2\}}{\omega' \min\{1, 2b, b^2\}}}.$$

Proof. The proof for the spectral condition number is identical to the proof of Theorem 4.11. For the condition number, we once more use the expression $\|\mathcal{F}_N[\mathcal{L}[v]]\|^2 = \lambda\|\mathcal{L}_0[v]\|^2$. An upper bound for λ is provided in the same manner as before. For a lower bound, we first write $\|\mathcal{F}_N[\mathcal{L}[v]]\|^2 = \|\mathcal{L}[v]\|^2 - \|\mathcal{L}[v] - \mathcal{F}_N[\mathcal{L}[v]]\|^2$. Since $v \in \mathcal{S}_N$ and the operators $\triangle$ and $\mathcal{I}$ commute with $\mathcal{F}_N$, we obtain

$$\|\mathcal{F}_N[\mathcal{L}[v]]\|^2 = \|\mathcal{L}[v]\|^2 - \|\mathcal{L}_1[v] - \mathcal{F}_N[\mathcal{L}_1[v]]\|^2 \geq \|\mathcal{L}[v]\|^2 - c\|v - \mathcal{F}_N[v]\|_1^2,$$

for some positive constant c independent of N. An application of (2.38) gives

$$\|v - \mathcal{F}_N[v]\|_1^2 \leq c \max_{\substack{n \notin I_N \\ i \in \{0,1\}^d}} (1 + \mu_n^{[i]})^{-1} \|v\|_2^2.$$

For any index set I_N satisfying (2.8), this maximum must tend to zero as N tends to infinity. Hence, for sufficiently large N, $\|v - \mathcal{F}_N[v]\|_1^2 \leq \frac{1}{2}\omega'\|v\|_2^2$. For such N,we obtain $\|\mathcal{F}_N[\mathcal{L}[v]]\|^2 \geq \frac{1}{2}\omega'\|v\|_2^2$ and this gives a lower bound for λ. □

Note that Theorem 4.11 is independent of the particular discretisation used: the matrix M_{G} is an optimal preconditioner for any choice of approximation basis. However, for Laplace eigenfunctions this preconditioner is diagonal and therefore of practical use. We note in passing that the observation that M_{G} is an optimal preconditioner is equivalent to stating that the set of functions $(b + \mu_n^{[i]})^{-1}\phi_n^{[i]}$ forms an optimally conditioned approximation basis for the problem (4.5).

Theorem 4.11 requires the Neumann coercivity condition $b - \frac{1}{4}\|a\|^2 > 0$. It transpires that, in the univariate setting at least, the weaker condition $b > 0$ may be imposed. Moreover, existence and uniqueness of a solution u to the exact problem (4.1) (respectively, the Galerkin approximation u_N) is also guaranteed under this condition, irrespective of the value of a. We shall not describe this case here, and we refer the reader to [3] for further details. Unfortunately, the extension of this result to the multivariate case remains an open problem.

In Table 4.1, we demonstrate the effect of the preconditioner M_{G}. For example, when $N = 40$, the original Galerkin matrix has a condition number of approximately $5,000$, whereas upon preconditioning, this figure is reduced to around 2.5: roughly $2,000$ times smaller.

4.3.3 Efficient solution techniques

For standard spectral discretisations in Cartesian product domains, Galerkin's equations are normally written in tensor-product form. For example, when $d = 2$, the matrix U of unknowns $\bar{u}_{n,m}^{[i]}$ satisfies the matrix equation $A_{\mathrm{G},x}U + U(A_{\mathrm{G},y})^\top = \hat{F}$, where $A_{\mathrm{G},x}$ and $A_{\mathrm{G},y}$ are the matrices corresponding to the univariate differential operators $-\partial_{xx}^2 + a_1\partial_x + \frac{1}{2}b\mathcal{I}$ and $-\partial_{xx}^2 + a_2\partial_x + \frac{1}{2}b\mathcal{I}$ respectively, and $\hat{F}$ is the matrix of coefficients of the inhomogeneous term f.

	$N = 10$	$N = 20$	$N = 30$	$N = 40$
$\kappa_s(A_\mathrm{G})$	231.521	954.753	2171.47	3881.68
$\kappa(A_\mathrm{G})$	362.676	1443.68	3245.39	5767.78
$\kappa_s(A_\mathrm{G}M_\mathrm{G}^{-1})$	1.16097	1.16099	1.16099	1.16099
$\kappa(A_\mathrm{G}M_\mathrm{G}^{-1})$	2.61680	2.61702	2.61704	2.61705

Table 4.1: Condition numbers for the univariate modified Fourier–Galerkin matrix A_G and the preconditioned matrix $A_\mathrm{G}M_\mathrm{G}^{-1}$ with parameters $a = 3$ and $b = 4$.

The advantage of this approach is that it facilitates the use of novel solution techniques such as the matrix diagonalisation and Schur decomposition methods [42]. Both techniques are solely based on the univariate matrices $A_{\mathrm{G},x}$ and $A_{\mathrm{G},y}$, which, as discussed previously, are reasonably conditioned.

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. Nonetheless, due to the simple nature of the particular equations arising from Laplace eigenfunction discretisations, it turns out that techniques such as these are unnecessary.

Instead, we consider standard iterative methods. In contrast to the aforementioned techniques, these methods are essentially independent of the dimension d. Since the matrix M_G is an optimal preconditioner for A_G, the conjugate gradient algorithm [66] may be applied to the preconditioned normal equations. If we write Galerkin's equations as $A_\mathrm{G}\bar{u} = \hat{f}$, then these equations are precisely

$$(A_\mathrm{G}M_\mathrm{G}^{-1})^\top A_\mathrm{G}M_\mathrm{G}^{-1}\bar{v} = (A_\mathrm{G}M_\mathrm{G}^{-1})^\top \hat{f}, \tag{4.13}$$

where $M_\mathrm{G}^{-1}\bar{v} = \bar{u}$. Since $(A_\mathrm{G}M_\mathrm{G}^{-1})^\top A_\mathrm{G}M_\mathrm{G}^{-1}$ is symmetric and has $\mathcal{O}(1)$ condition number, the conjugate gradient method converges to within a prescribed tolerance in a number of operations independent of the parameter N. Hence, the total cost of solving Galerkin's equations is proportional to the number of operations required to perform matrix-vector multiplications involving A_G (since M_G is diagonal, its contribution can be ignored). The cost of direct evaluation of such matrix-vector products is determined by the number of nonzero matrix entries, for which we have the following lemma:

Lemma 4.13. *Suppose that $N \gg d$. Then the number of non-zero entries of the matrix A_G is*

$$d2^d N^{d+1} + \mathcal{O}\left(N^d\right), \tag{4.14}$$

in the case of the full index set (2.33), and

$$d2^d N^2 \lceil (1+\zeta(2))^{d-1} \rceil + \mathcal{O}\left(N(\log N)^{d-1}\right), \tag{4.15}$$

for the hyperbolic cross (2.41).

Proof. In view of Lemma 4.5 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 (4.14). For the hyperbolic cross (2.41) we have

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

Evaluating this final sum gives (4.15). □

We conclude that Galerkin's equations can be solved in, at most, $\mathcal{O}\left(N^{d+1}\right)$ operations in the full index set case, and only $\mathcal{O}\left(N^2\right)$ operations, *regardless* of d, when a hyperbolic cross is employed. Hence, even for $d = 2$, the hyperbolic cross will offer an advantage over the full index set, provided, of course, that the convergence rate of the approximation remains comparable. As we establish in the next section, this is indeed the case.

Since the action of the matrix $N_{\mathrm{G}} = A_{\mathrm{G}} - M_{\mathrm{G}}$ corresponds to finding Laplace–Dirichlet or Laplace–Neumann coefficients of derivatives of finite sums of Laplace–Dirichlet or Laplace–Neumann eigenfunctions, a variant of the Fast Fourier Transform (FFT) can be employed in the full index set case. In this manner, the aforementioned 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, an analogue of the sparse grid FFT could be employed [60], provided N is highly composite, thereby reducing the figure of $\mathcal{O}\left(N^2\right)$ to just $\mathcal{O}\left(N(\log N)^d\right)$. However, as mentioned in Section 2.10.3, this technique is neither easy nor straightforward to implement.

As is common for spectral methods in tensor-product domains, the matrix A_{G} is increasingly sparse for large d. Indeed, Lemma 4.13 indicates that the *sparsity ratio* is $\mathcal{O}\left(N^{2(d-1)}\right)$ when a hyperbolic cross is employed. In Figure 4.1 we plot the matrix A_{G} corresponding to $d = 2, 3$. Herein we observe both the increasing sparsity and the non-tensor-product structure of A_{G} in the hyperbolic cross case.

4.3.4 Analysis of convergence

In the trivial case $a = 0$, the Galerkin approximation u_N is precisely $\mathcal{F}_N[u]$. For such problems, the error estimates of Chapter 2 are sufficient. When $a \neq 0$, this is no longer the case. In this setting, Céa's lemma is the starting point for our analysis.

Céa's lemma immediately provides an estimate for the convergence rate in the $\mathrm{H}^1(\Omega)$ norm. Since $\mathcal{F}_N[u]$ is the best approximation to u in this norm (see Theorem 2.9), we have $\|u - u_N\|_1 \leq \gamma\omega^{-1}\|u - \mathcal{F}_N[u]\|_1$. The results of Chapter 2 now provide estimates for the rate of convergence using various index sets. Since the solution u satisfies at least the first derivative condition, we have:

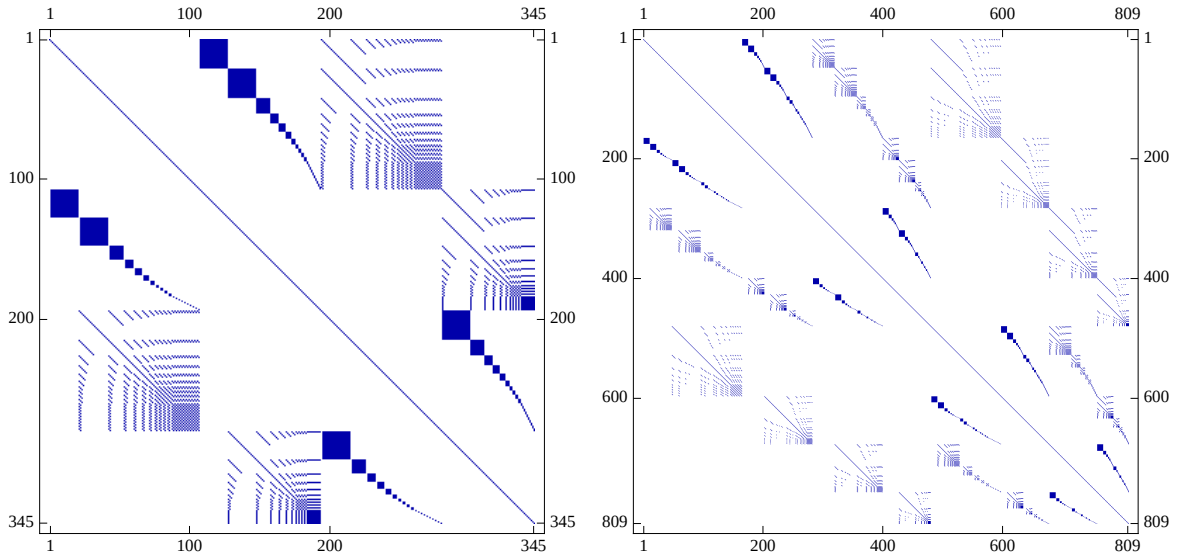


Figure 4.1: Pattern of the $d = 2$ (left) and $d = 3$ (right) modified Fourier–Galerkin matrices A_{G} with $N = 20$ and $N = 10$ respectively.

Theorem 4.14. *Suppose that u_N is the modified Fourier–Galerkin approximation based on the full index set (2.33). Then*

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

where $c_{r,1}$ and c_1 are the constants from Lemmas 2.24 and 2.25 respectively. If u_N is the approximation based on the hyperbolic cross (2.41), then

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

and $\|u - u_N\|_1 \leq \gamma\omega^{-1} c_1 N^{-\frac{5}{2}} \|u\|_{4,mix}$, where $c_{r,1}$ and c_1 are the constants from Lemma 2.28 and Theorem 2.29 respectively.

As in the case of function approximation, when $u \notin \mathrm{H}^4_{\mathrm{mix}}(-1,1)^d$ we require additional regularity for the hyperbolic cross approximation to obtain the same convergence rate as its full index set counterpart. However, provided that $u \in \mathrm{H}^4_{\mathrm{mix}}(-1,1)^d$, the convergence rates in this norm are identical.

Estimates for the error of the Laplace–Dirichlet Galerkin approximation can be obtained in a straightforward manner. For example, if $u \in \mathrm{H}^3_{\mathrm{mix}}(\Omega)$ is the solution to the problem (4.5), then Lemma 2.25 and Theorem 2.29 give that $\|u - u_N\|_1$ is $\mathcal{O}(N^{-\frac{3}{2}})$.

As is evident, all error estimates for Galerkin approximations rely on the smoothness of the solution u. Clearly $\mathrm{H}^1(\Omega)$ regularity is guaranteed, since u is defined as the solution of the weak problem (4.2). However, it can also be shown that $u \in \mathrm{H}^2(\Omega)$ [80]. In other words, u is a classical solution: the residual $\mathcal{L}[u] - f$ vanishes almost everywhere in Ω. Moreover, u also satisfies the stability estimate $\|u\|_2 \leq c\|f\|$, where $c > 0$ depends only on Ω and $\mathcal{L}$.

In the univariate setting, higher regularity is guaranteed by a so-called *shift* theorem. If $f \in \mathrm{H}^r(-1,1)$ then the solution $u \in \mathrm{H}^{r+2}(-1,1)$ for any $r \geq 0$ [77, 80]. In two or

more dimensions, where the boundary Γ is non-smooth, this is no longer the case: even in $f \in \mathrm{C}^\infty(\bar{\Omega})$, examples can be constructed where $u \notin \mathrm{H}^{2+r}(\Omega)$ for any $r > 0$ [77]. In general, the solution u will exhibit weak singularities at the vertices of the domain. Effective numerical treatment of such discontinuities, along the lines of smoothing the function u by subtracting out such singularities, is beyond the scope of this study.[10]

Returning to the problem at hand, we note that, when measured in the $\mathrm{H}^1(\Omega)$ norm, the error $\|u - u_N\|_1$ is asymptotically of the same order as $\|u - \mathcal{F}_N[u]\|_1$, i.e. the error incurred by the best approximation to u from $\mathcal{S}_N$. Hence, we refer to u_N as a *quasi-optimal* approximation in this norm. We now assess the same question for the $\mathrm{L}^2(\Omega)$ norm. We commence with the Dirichlet problem:

Lemma 4.15. *Suppose that u_N is the Laplace–Dirichlet Galerkin approximation. Then*

$$\|u - u_N\| \leq \left(1 + \|a\|\omega^{-1}\right) \|u - \mathcal{F}_N[u]\|. \tag{4.16}$$

Proof. Since u_N is the Galerkin approximation to u, we have $T(u_N, \phi) = (f, v) = T(u, \phi)$ for all $\phi \in \mathcal{S}_N$. Suppose that we write $u = \mathcal{F}_N[u] + (u - \mathcal{F}_N[u])$. Then, setting $\phi = u_N - \mathcal{F}_N[u] \in \mathcal{S}_N$ in the above expression and using the coercivity condition gives

$$\omega\|u_N - \mathcal{F}_N[u]\|_1^2 \leq T(u_N - \mathcal{F}_N[u], u_N - \mathcal{F}_N[u]) = T(u - \mathcal{F}_N[u], u_N - \mathcal{F}_N[u]).$$

Since $u - \mathcal{F}_N[u]$ is orthogonal to any $\phi \in \mathcal{S}_N$, we obtain

$$\omega\|u_N - \mathcal{F}_N[u]\|_1^2 \leq (a \cdot \nabla[u - \mathcal{F}_N[u]], u_N - \mathcal{F}_N[u]).$$

We now note that $(a \cdot \nabla v, w) = -(v, a \cdot \nabla w)$, $\forall v, w \in \mathrm{H}_0^1(\Omega)$. Setting $v = u - \mathcal{F}_N[u]$ and $w = u_N - \mathcal{F}_N[u]$ now yields

$$\omega\|u_N - \mathcal{F}_N[u]\|_1^2 = -(u - \mathcal{F}_N[u], a \cdot \nabla[u_N - \mathcal{F}_N[u]]) \leq \|a\|\|u - \mathcal{F}_N[u]\|\|u_N - \mathcal{F}_N[u]\|_1.$$

This gives $\|u_N - \mathcal{F}_N[u]\|_1 \leq \|a\|\omega^{-1}\|u - \mathcal{F}_N[u]\|$. This result follows straightaway from the decomposition $u - u_N = u - \mathcal{F}_N[u] + \mathcal{F}_N[u] - u_N$. □

In view of this lemma, we deduce that u_N, the approximation to the Dirichlet problem, is also quasi-optimal in the $\mathrm{L}^2(\Omega)$ norm. Note that, given sufficient regularity, it follows that $\|u - u_N\| = \mathcal{O}(N^{-\frac{5}{2}}(\log N)^{\frac{d-1}{2}})$ when a hyperbolic cross index set is employed.

Next we address the Neumann case:

Lemma 4.16. *Suppose that u_N is the modified Fourier–Galerkin approximation. Then*

$$\|u - u_N\| \leq c\left(\|u - \mathcal{F}_N[u]\|_\Gamma + \|u - \mathcal{F}_N[u]\|\right), \tag{4.17}$$

where $\|g\|_\Gamma^2 = \int_\Gamma |g(x)|^2 \, \mathrm{d}x$ for $g \in \mathrm{L}^2(\Gamma)$, and $c > 0$ is independent of u and N.

[10] A more detailed assessment is given, for example, in [31]. We note, however, that it is commonly recommended that such singularities be ignored when designing numerical algorithms, except in cases when it is known *a priori* that the solution u is discontinuous, or if slow convergence of numerical schemes suggests *a posteriori* that singularities are present [31, p.121].

Proof. As in the previous lemma, we have

$$\omega\|u_N - \mathcal{F}_N[u]\|_1^2 \leq (a\cdot\nabla[u-\mathcal{F}_N[u]], u_N - \mathcal{F}_N[u]).$$

We note that $(a\cdot\nabla v, w) = \int_\Gamma \hat{n}.avw - (v, a\cdot\nabla w)$, $\forall v, w \in \mathrm{H}^1(\Omega)$, where $\hat{n}$ is the unit outward normal vector on Γ. Hence,

$$\omega\|u_N - \mathcal{F}_N[u]\|_1^2 \leq c\|u-\mathcal{F}_N[u]\|_\Gamma \|u_N - \mathcal{F}_N[u]\|_\Gamma. \tag{4.18}$$

The result now follows from the trace inequality $\|g\|_\Gamma \leq c\|g\|_1$, $\forall g \in \mathrm{H}^1(\Omega)$ [56]. □

As a result of this lemma, to assess the error $\|u - u_N\|$ we require an estimate for $\|u - \mathcal{F}_N[u]\|_\Gamma$. To provide this, we first derive an improved trace inequality:

Lemma 4.17. *We have* $\|g\|_\Gamma \leq c\sqrt{\|g\|\|g\|_1}$, $\forall g \in \mathrm{H}^1(\Omega)$, *where* $c > 0$ *is independent of* g.

Proof. Consider $g(\pm 1, x_2, \ldots, x_d)$. By the univariate Sobolev interpolation inequality $\|h\|_\infty \leq c\sqrt{\|h\|\|h\|_1}$, $\forall h \in \mathrm{H}^1(-1,1)$, it follows that

$$\begin{aligned}
&\int_{-1}^1 \ldots \int_{-1}^1 g(\pm 1, x_2, \ldots, x_d)^2 \,\mathrm{d}x_2 \ldots \mathrm{d}x_d \\
&\quad \leq c \int_{-1}^1 \ldots \int_{-1}^1 \left[\int_{-1}^1 g(x)^2 \,\mathrm{d}x_1\right]^{\frac{1}{2}} \left[\int_{-1}^1 g(x)^2 + \partial_{x_1} g(x)^2 \,\mathrm{d}x_1\right]^{\frac{1}{2}} \mathrm{d}x_2 \ldots \mathrm{d}x_d \\
&\quad \leq c\|g\|\|g\|_1.
\end{aligned}$$

Hence we obtain the result. □

Corollary 4.18. *Suppose that* u_N *is the modified Fourier–Galerkin approximation. Then*

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

for some constant $c > 0$ *independent of* u *and* N.

From this corollary, we conclude that u_N is no longer quasi-optimal in the $\mathrm{L}^2(\Omega)$ norm. For example, when a hyperbolic cross index set is employed, $\|u - u_N\| = \mathcal{O}(N^{-3}(\log N)^{\frac{d-1}{4}})$, whereas $\|u - \mathcal{F}_N[u]\| = \mathcal{O}(N^{-\frac{7}{2}}(\log N)^{\frac{d-1}{2}})$. This estimate is corroborated by numerical example at the end of this section.

In view of the results of this section, we deduce that the hyperbolic cross approximation converges no more slowly than its full index set counterpart (given sufficient regularity). This, in combination with the arguments of the previous section, establishes the advantage of the hyperbolic cross in this context (i.e. reduced computational cost and comparable error estimates). Hence it shall form our primary consideration from now on.

Our final result of this section assesses the uniform error of the approximation u_N:

Lemma 4.19. *Suppose that* $u \in \mathrm{L}^\infty(\Omega)$ *and that* u_N *is the Laplace–Dirichlet or Laplace–Neumann Galerkin approximation based on the hyperbolic cross (2.41). Then, for* $d = 1$,

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

and for $d \geq 2$,

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

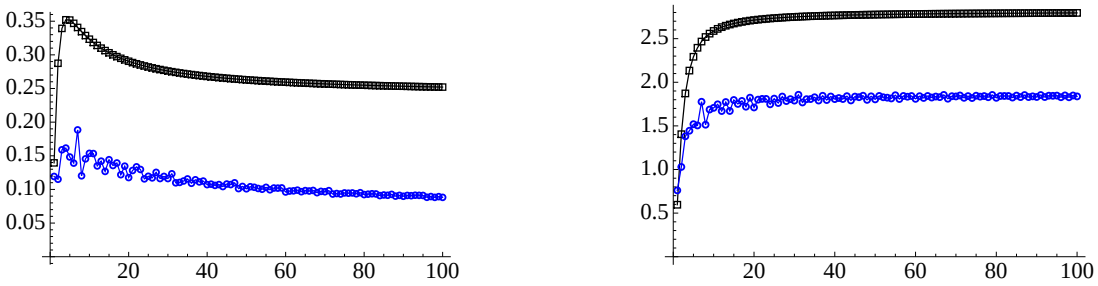


Figure 4.2: Scaled errors $N^3(\log N)^{-\frac{d-1}{4}}\|u-u_N\|$ (left) and $N^{\frac{5}{2}}\|u-u_N\|_1$ (right) for (4.19) (squares) and (4.20) (circles).

Proof. Consider the univariate case first. For $d = 1$, the estimate (4.18) reduces to $\|u_N - \mathcal{F}_N[u]\|_1 \leq c\|u - \mathcal{F}_N[u]\|_\infty$. The result now follows immediately from the decomposition $u - u_N = \mathcal{F}_N[u] - u_N + u - \mathcal{F}_N[u]$ and the imbedding $\mathrm{H}^1(-1,1) \hookrightarrow \mathrm{C}[-1,1]$.

Next we consider the case $d \geq 2$. Once more, we have $\|u-u_N\|_\infty \leq \|\mathcal{F}_N[u]-u_N\|_\infty + \|u-\mathcal{F}_N[u]\|_\infty$. Now suppose that $v \in \mathcal{S}_N$ is arbitrary. We claim that $\|v\|_\infty \leq cN^{\frac{1}{2}}(\log N)^{\frac{d-1}{2}}\|v\|$ for some constant $c > 0$ independent of N and v. In particular, if $v = \mathcal{F}_N[u] - u_N$, then the lemma is verified upon substituting this result into the previous expression (and noting that $\|u - u_N\|^2 = \|\mathcal{F}_N[u] - u_N\|^2 + \|u - \mathcal{F}_N[u]\|^2 \geq \|\mathcal{F}_N[u] - u_N\|^2$ by orthogonality). However, by the Cauchy–Schwarz inequality,

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

as required. □

In the Dirichlet case, this lemma establishes an $\mathcal{O}\left(N^{-2}(\log N)^{d-1}\right)$ estimate for the uniform error, provided $u \in \mathrm{H}^3_{\mathrm{mix}}(\Omega)$; a result which is asymptotically the same order as $\|u - \mathcal{F}_N[u]\|_\infty$. For the Neumann problem, the corresponding estimate is $\mathcal{O}(N^{-\frac{5}{2}}(\log N)^{d-1})$ for $d \geq 2$, which numerical examples indicate is sub-optimal. Conversely, the univariate estimate is quasi-optimal.

In Figure 4.2 we consider the univariate Neumann problem

$$-u''(x) + u'(x) + 2u(x) = x^3\mathrm{e}^x, \quad x \in (-1,1), \quad u'(\pm 1) = 0, \tag{4.19}$$

and the bivariate problem with parameters $a_1 = -1$, $a_2 = 2$, $b = 4$ and exact solution

$$\begin{aligned} u(x_1,x_2) =& \mathrm{e}^{x_1x_2} - \frac{x_2}{4}\left[(1+x_1)^2\mathrm{e}^{x_2} + (1-x_1)^2\mathrm{e}^{-x_2}\right] - \frac{x_1}{4}\left[(1+x_2)^2\mathrm{e}^{x_1} + (1-x_2)^2\mathrm{e}^{-x_1}\right] \\ &+ \frac{\mathrm{e}}{8}\left[(1-x)^2(1-y)^2 + (1+x)^2(1+y)^2\right], \end{aligned} \tag{4.20}$$

with $f(x_1,x_2)$ given accordingly. Both the previously derived $\mathrm{H}^1(\Omega)$ and $\mathrm{L}^2(\Omega)$ error estimates are confirmed for these examples. Moreover, graphs of the pointwise error, given in Figure 4.3, verify both the uniform error estimate of Lemma 4.19 for the case $d = 1$ and the non-optimality

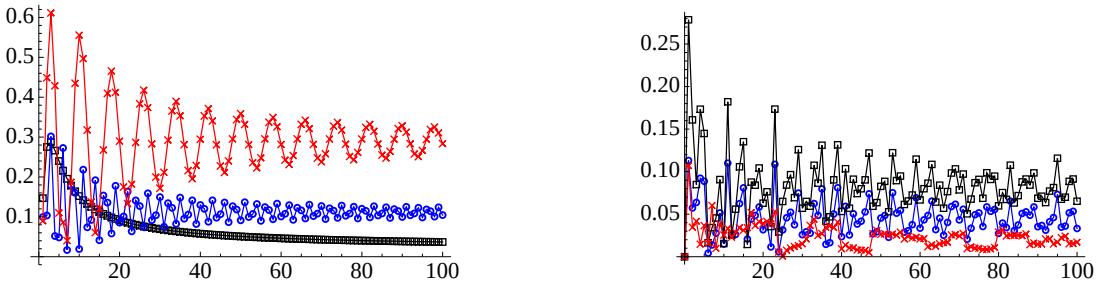
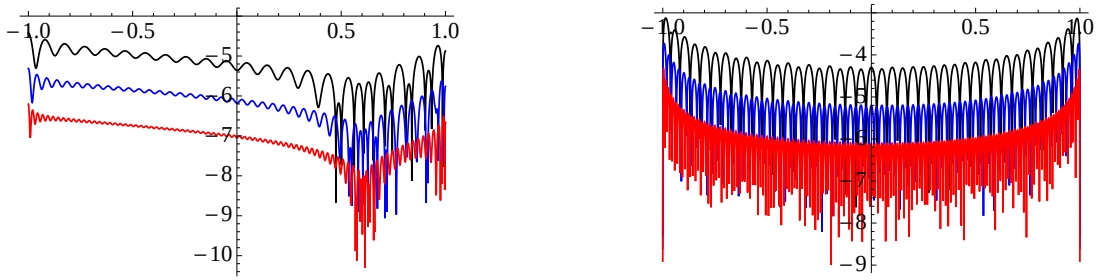


Figure 4.3: Scaled error $N^3(\log N)^{-(d-1)}|u(x) - u_N(x)|$ against $N = 1, \ldots, 100$ for the problems (4.19) (left), where $x = -1$ (squares), $x = \frac{1}{4}$ (circles) and $x = \frac{3}{4}$ (crosses), and (4.20) (right), where $x = (1, -1)$, $x = (-1, 1)$ and $x = (0, -\frac{1}{4})$.



$u(x) = 1 + x^4 - 2x^2$ $\qquad$ $u(x) = x^4 - 1$

Figure 4.4: Log pointwise error $\log_{10}|u(x) - u_N(x)|$ for $-1 \leq x \leq 1$ and $N = 20, 40, 80$ (in descending order), where u_N is the modified Fourier–Galerkin (left) or Laplace–Dirichlet Galerkin (right) approximation to the problem $-u''(x) + u'(x) + 2u(x) = f(x)$ with boundary conditions $u'(\pm 1) = 0$ and $u(\pm 1) = 0$ respectively.

of the corresponding multivariate result. Indeed, the indication given by this example is that the approximation remains quasi-optimal in the uniform norm for $d \geq 2$.

In the trivial case $a = 0$, the pointwise error for both the Neumann and Dirichlet approximations is one power of N faster inside the domain than on the boundary (see Chapter 2). For $a \neq 0$, as we indicate in Figure 4.4, the same effect occurs for the Dirichlet approximation, yet we have no proof of this fact. On the other hand, the Neumann approximation does not offer a faster convergence rate inside the domain when $a \neq 0$, as demonstrated by Figure 4.3.

4.3.5 Numerical comparison

Standard spectral–Galerkin approximations for (4.5) involving Jacobi polynomials, usually of Chebyshev or Legendre type, guarantee spectral convergence provided the solution is smooth. The efficient methods of Shen [78, 146, 147], based on such polynomials, can be optimally preconditioned, and the $\mathcal{O}\left(N^d\right)$ coefficients of the approximation found in $\mathcal{O}\left(N^{d+1}\right)$ operations.

Conversely, both the modified Fourier and Laplace–Dirichlet methods converge slowly unless the solution u obeys higher-order derivative conditions. However, due to their lower complexity (only $\mathcal{O}\left(N(\log N)^{d-1}\right)$ terms which can be found in $\mathcal{O}\left(N^2\right)$ operations), for certain

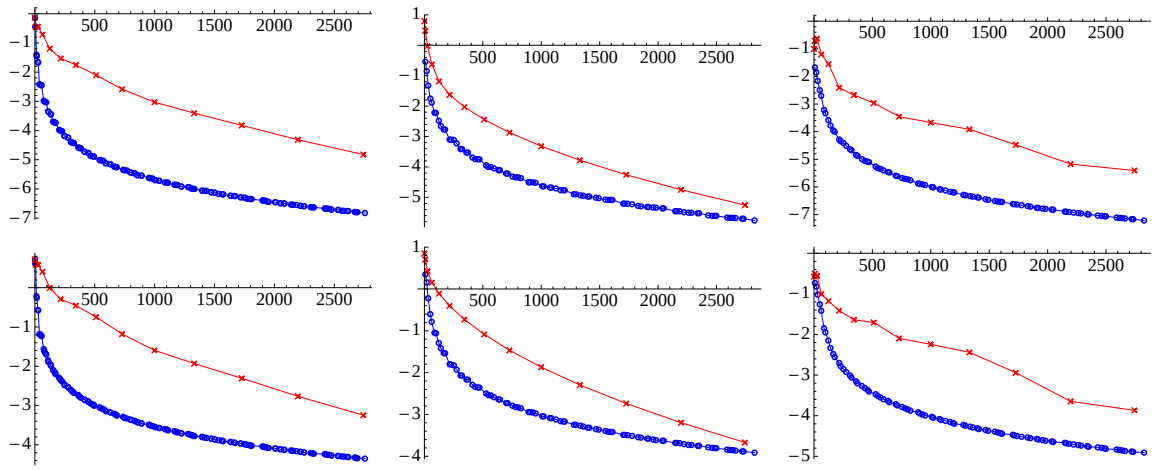


Figure 4.5: Comparison of the modified Fourier (circles) and Legendre–Galerkin (crosses) methods applied to the Neumann problem (4.5) with exact solution (4.21)–(4.23) (left to right). (top) log $\mathrm{L}^2(\Omega)$ error $\log_{10}\|u-u_N\|$ against number of terms, (bottom) log $\mathrm{H}^1(\Omega)$ error $\log_{10}\|u-u_N\|_1$.

examples, these methods offer significantly lower errors for moderate values of the parameter N. We now consider three such examples, all Neumann problems, with parameters $d=3$, $b=2$, $a=0$ and exact solutions

$$u(x,y,z) = \sin(2x(2x^2-2)^2)(\sin y - y\cos 1)(z^5-5z), \tag{4.21}$$

$$u(x,y,z) = \mathrm{e}^{z^2\cos 4y + x^2} - p(x,y,z), \tag{4.22}$$

$$u(x,y,z) = x^2\cos(y\sin 5x)\cosh z - p(x,y,z), \tag{4.23}$$

respectively. Note that in (4.22) and (4.23) 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 4.5 we plot the error against the number of approximation terms for this method and the Legendre–Galerkin method (the Chebyshev–Galerkin method 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 (4.21)–(4.23), 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. Moreover, the techniques to construct Legendre–Galerkin approximations are specific to dimension [146]. Conversely, the coefficients of the modified Fourier approximation are found using only generic iterative techniques, which are essentially independent of d.

Note that these plots do not take into account the operational cost of each method. As discussed, the modified Fourier method is likely to perform even better if we were to take this

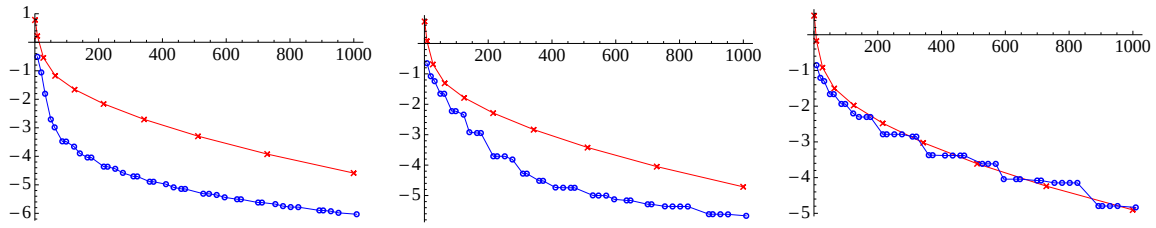


Figure 4.6: Log error $\log_{10}\|u-u_N\|$ against number of terms for the modified Fourier and Legendre–Galerkin methods applied to the problem with $a=0$, $b=2$ and exact solutions u_1 (left), u_2 (middle) and u_3 (right), where $\omega=10$.

factor into account. Having said that, we note that a central issue concerning the modified Fourier method is the computation of the coefficients $\hat{f}_n^{[i]}$, meaning that a direct comparison of the two methods in terms of computational time is premature. The design of efficient, robust algorithms based on the quadratures developed in [94, 95] is a subject of ongoing research, as we discuss briefly in Chapter 6.

Having provided 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 way 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 $\|u_i\|_r=\mathcal{O}(\omega^r)$ and $\|u_i\|_{r,\mathrm{mix}}=\mathcal{O}(\omega^{ir})$ for $i=1,2,3$. Figure 4.6 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 $d=2,3$). To address this issue—thereby making the method effective for a broader range of problems—the topic of convergence acceleration of modified Fourier approximations is broached in Chapter 5. In Chapter 6, we briefly discuss the application of the ensuing techniques to boundary value problems.

4.4 Extensions

The second-order, constant coefficient problem (4.5) presents the simplest setting for Laplace eigenfunction approximations. In this section, we assess the applicability of such techniques to several more general types of problems. In particular, we first consider variable-coefficient, second-order Neumann and Dirichlet problems, and in Section 4.4.5 we scrutinise the application of such methods to higher, even-order boundary value problems.

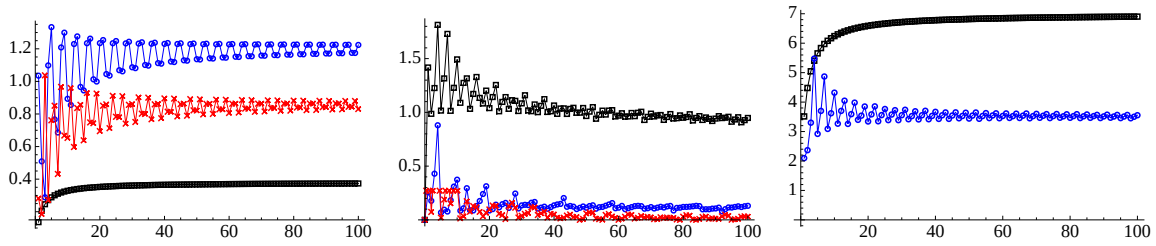


Figure 4.7: (left) scaled pointwise error $N^3|u(x) - u_N(x)|$ where u is given by (4.25) and $x = -1$ (squares), $x = \frac{1}{2}$ (circles), $x = -\frac{1}{4}$ (crosses). (middle) scaled pointwise error $N^3(\log N)^{-1}|u(x)-u_N(x)|$ for (4.26), where $(x_1, x_2) = (1, -1)$, $(1, \frac{1}{4})$ and $(-\frac{1}{2}, -\frac{1}{2})$. (right) scaled H^1 error $N^{\frac{5}{2}}\|u - u_N\|_1$ for (4.25) (squares) and (4.26) (circles).

4.4.1 Variable-coefficient Neumann boundary value problems

The modified Fourier–Galerkin method may be extended in a straightforward manner to the variable-coefficient problem

$$\mathcal{L}[u](x) = -\triangle u(x) + a(x) \cdot \nabla u(x) + b(x)u(x) = f(x), \quad x \in \Omega, \quad \frac{\partial u}{\partial n}\Big|_{\partial\Omega} = 0, \tag{4.24}$$

where $b : \Omega \to \mathbb{R}$ and $a : \Omega \to \mathbb{R}^d$ are (sufficiently smooth[11]) functions of x. Note that coercivity of the bilinear form T is equivalent to the condition $\min_{x\in\bar{\Omega}} \left\{b(x) - \frac{1}{4}\|a(x)\|^2\right\} > 0$. Under this condition, convergence may be analysed in an identical manner to the constant coefficient case previously studied. In particular, the $\mathrm{H}^1(\Omega)$ norm error remains $\mathcal{O}(N^{-\frac{5}{2}})$. Once more, numerical results indicate that the uniform error is $\mathcal{O}(N^{-3}(\log N)^{d-1})$.

In Figure 4.7 we demonstrate these results for the univariate problem with parameters

$$u(x) = \cos 3x + \frac{3x^2}{2} \sin 3, \quad a(x) = x, \quad b(x) = \mathrm{e}^{-x}, \tag{4.25}$$

and the bivariate problem

$$\begin{aligned} u(x_1, y_2) &= \frac{1}{2} \sin 2x_1x_2 - x_1x_2(\cos 2x_1 + \cos 2x_2 + 2\sin 2 - \cos 2), \\ a(x_1, x_2) &\equiv 0, \quad b(x_1, x_2) = \cos(x_1 + x_2). \end{aligned} \tag{4.26}$$

We note in passing that previously derived estimates for the condition number also remain valid in the variable-coefficient setting.

The central question remaining in this case is the computation of the approximation u_N. Once more, we devise a scheme based on an appropriate decomposition of the operator $\mathcal{L}$. To this end, we write $\mathcal{L} = (-\triangle + b_0\mathcal{I}) + (a \cdot \nabla + (b - b_0)\mathcal{I})$, where $b_0 = \max_{x\in\bar{\Omega}} b(x)$, and define M_G and N_G as the matrices corresponding to these operators.[12] Concerning such matrices, we have the following result:

[11] In practice, we shall assume that a and b are continuous on $\bar{\Omega}$. However, lower smoothness conditions may be imposed [142, chapter 6].

[12] This is a standard approach for variable-coefficient discretisations. See, for example, [47, 147].

Lemma 4.20. *Suppose that $\rho(M_{\mathrm{G}}^{-1}N_{\mathrm{G}})$ is the spectral radius of the matrix $M_{\mathrm{G}}^{-1}N_{\mathrm{G}}$. Then*

$$\rho(M_{\mathrm{G}}^{-1}N_{\mathrm{G}}) \leq 1 - \frac{\min_{x\in\bar{\Omega}}\left\{b(x) - \frac{1}{4}\|a(x)\|^2\right\}}{b_0}.$$

In particular, if the operator T is coercive, then $\rho(M_{\mathrm{G}}^{-1}N_{\mathrm{G}}) < 1$.

Proof. Suppose that λ is an eigenvalue of $M_{\mathrm{G}}^{-1}N_{\mathrm{G}}$ with corresponding eigenfunction $u \in \mathcal{S}_N$. Then

$$\lambda = -\frac{((b_0 - b)u, u) + (a\cdot\nabla u, u)}{b_0\|u\|^2 + \|\nabla u\|^2}.$$

Note that $b_0 - b(x) \geq 0$ for all $x \in \bar{\Omega}$. An application of Young's inequality now gives

$$\begin{aligned}
|\lambda| &\leq \frac{\int_\Omega \left(b_0 - b(x) + \frac{1}{4}\|a(x)\|^2\right) u(x)^2\,\mathrm{d}x + \|\nabla u\|^2}{b_0\|u\|^2 + \|\nabla u\|^2} \\
&\leq \frac{\max_{x\in\bar{\Omega}}\left\{b_0 - b(x) + \frac{1}{4}\|a(x)\|^2\right\}\|u\|^2 + \|\nabla u\|^2}{b_0\|u\|^2 + \|\nabla u\|^2} \\
&\leq \frac{\max_{x\in\bar{\Omega}}\left\{b_0 - b(x) + \frac{1}{4}\|a(x)\|^2\right\}\|u\|^2}{b_0} = 1 - \frac{\min_{x\in\bar{\Omega}}\left\{b(x) - \frac{1}{4}\|a(x)\|^2\right\}}{b_0},
\end{aligned}$$

as required. □

As a consequence of this lemma, Galerkin's equations $A_{\mathrm{G}}\bar{u} = \hat{f}$ for the problem (4.24) can be solved using the classical Lanczos iteration $M_{\mathrm{G}}\bar{u}^{k+1} = -N_{\mathrm{G}}\bar{u}^k + \hat{f}$, $k = 0, 1, 2, \ldots$, where $\bar{u}^0$ is arbitrary and $\bar{u}^k$ is the k^{th} iterate [66]. Convergence of this iteration to within a prescribed numerical tolerance is guaranteed by Lemma 4.20. Moreover, since the bound established is independent of N, the number of iterations required is also independent of N. We conclude that the total cost of this approach is determined by the number of operations required to perform matrix-vector multiplications involving N_{G}.

The matrix N_{G} is typically dense; thus, direct evaluation requires $\mathcal{O}\left(|I_N|^2\right)$ operations. However, the action of N_{G} involves finding modified Fourier coefficients of products and derivatives of finite modified Fourier sums. Hence, this figure can be reduced to $\mathcal{O}\left(N^d \log N\right)$ by using the FFT in the case of the full index set (2.33) and, in theory, to $\mathcal{O}\left(N(\log N)^d\right)$ by use of the SGFFT in the hyperbolic cross case (2.41).

Conjugate gradients can also be applied to the preconditioned normal equations.[13] It can be shown that the diagonal matrix M_{G} corresponding to $-\triangle + b_0\mathcal{I}$ is optimal for the spectral condition number. However, though numerical results indicate it to be the case, it is not known whether this holds for the L^2 condition number.

If the FFT or SGFFT are not to be used, it is necessary to express Galerkin's equations explicitly. As expected, the entries of the matrix A_{G} involve both Laplace–Dirichlet and Laplace–Neumann coefficients of the functions $b(x)$ and $a_j(x)$, $j = 1, \ldots, d$, where $a(x) = (a_1(x), \ldots, a_d(x))^\top$. Thus, they may be calculated using the quadrature methods outlined in Section 2.12. The underlying reason for this observation is that the products $\phi_n^{[i]}\phi_m^{[l]}$ and $\phi_n^{[i]}(\phi_m^{[l]})'$, $i, l \in \{0, 1\}$, $n, m \in \mathbb{N}_0$, may be expressed in terms of sums of Laplace–Neumann and Laplace–Dirichlet eigenfunctions.[14]

[13] Commonly, conjugate gradients are preferred over Lanczos iterations [42].

[14] Much like in the classical Fourier case, this observation underpins why fast evaluation of matrix-vector products can be achieved by a variant of the FFT.

To demonstrate this, and hence derive an explicit expression for A_{G}, it is useful to redefine the univariate eigenfunction $\phi_0^{[0]} = 1$ as opposed to $\phi_0^{[0]} = \frac{1}{\sqrt{2}}$. This eigenfunction is no longer normalised; to counter this, we define the scaling parameter $c_n^{[i]}$ by $c_0^{[0]} = \frac{1}{2}$ and $c_n^{[i]} = 1$ otherwise.

With eigenfunctions $\phi_n^{[i]}$ expressed in this manner, for $n, m \in \mathbb{N}_0$ and $i, l \in \{0, 1\}$, we have

$$\begin{aligned}\phi_n^{[i]}\phi_m^{[l]} &= \frac{1}{2}\left\{(-1)^{il}\phi_{n+m-il}^{[i+l]} + \phi_{(-1)^l(n-m)}^{[i+l]}\right\},\\ \phi_n^{[i]}(\phi_m^{[l]})' &= \frac{\alpha_m^{[l]}}{2}\left\{(-1)^{(1-i)(1-l)}\psi_{n+m-il}^{[i+l+1]} + (-1)^{1-l}\psi_{(-1)^{il}(m-n)}^{[1+i+l]}\right\},\end{aligned} \tag{4.27}$$

where $\psi_n^{[i]}$ is the n^{th} univariate Laplace–Dirichlet eigenfunction. Here the sum $i + l$ is taken modulo 2 and $\phi_{-n}^{[0]} = \phi_n^{[0]}$, $\phi_{-n}^{[1]} = -\phi_{n+1}^{[1]}$, $\psi_{-n}^{[0]} = \psi_{n+1}^{[0]}$ and $\psi_{-n}^{[1]} = -\psi_n^{[1]}$ for $n \in \mathbb{N}_0$.

It is now possible to give an explicit expression for the matrix A_{G}. We shall not present the full multivariate case. Instead we focus on the univariate setting. The extension to $d \geq 2$ is conceptually clear, but algebraically convoluted.

Lemma 4.21. *The modified Fourier–Galerkin matrix A_{G} corresponding to the univariate problem (4.24) is given by*

$$A_{\mathrm{G}} = \begin{pmatrix} D^{[0]} & 0 \\ 0 & D^{[1]} \end{pmatrix} + \begin{pmatrix} B^{[0,0]} & B^{[0,1]} \\ B^{[1,0]} & B^{[1,1]} \end{pmatrix} + \begin{pmatrix} C^{[0,0]} & C^{[0,1]} \\ C^{[1,0]} & C^{[1,1]} \end{pmatrix},$$

where $D^{[i]}$ is the diagonal matrix with entries $c_n^{[i]}\mu_n^{[i]}$ and $B^{[i,l]}, C^{[i,l]} \in \mathbb{R}^{(N+1-i)\times(N+1-l)}$ have $(n, m)^{\text{th}}$ entries

$$\begin{aligned}B_{n,m}^{[i,l]} &= \frac{\alpha_m^{[l]}}{2}\left\{(-1)^{(1-i)(1-l)}\check{a}_{n+m-il}^{[i+l+1]} + (-1)^{1-l}\check{a}_{(-1)^{il}(m-n)}^{[1+i+l]}\right\},\\ C_{n,m}^{[i,l]} &= \frac{1}{2}\left\{(-1)^{il}\hat{b}_{n+m-il}^{[i+l]} + \hat{b}_{(-1)^l(n-m)}^{[i+l]}\right\}, \quad i, l \in \{0, 1\}, \quad n, m \in \mathbb{N}_0.\end{aligned}$$

Proof. The entry of A_{G} corresponding to indices $i, l \in \{0, 1\}$ and $n, m \in \mathbb{N}_0$ is

$$(\mathcal{L}[\phi_m^{[l]}], \phi_n^{[i]}) = \int_{-1}^{1}\left\{-(\phi_m^{[l]})''(x) + a(x)(\phi_m^{[l]})'(x) + b(x)\phi_m^{[l]}(x)\right\}\phi_n^{[i]}(x)\,\mathrm{d}x.$$

The result now follows immediately from (4.27). □

4.4.2 General second order Neumann boundary value problems

Unfortunately, the modified Fourier–Galerkin technique is limited to problems of the form (4.24). Following the approach of [142, chapter 6], a significantly more general setting is presented by the following problem:

$$\mathcal{L}[u] = -\sum_{i,j=1}^{d}\partial_{x_i}\left(a_{i,j}\partial_{x_j}u\right) - \sum_{i=1}^{d} b_i\partial_{x_i}u + cu = f, \quad \mathcal{B}[u] = 0, \tag{4.28}$$

where $a_{i,j}, b_i, c : \Omega \to \mathbb{R}$ are given functions of sufficient smoothness and $\mathcal{B}[u]$ are appropriate boundary conditions. In general, the operator (4.28) is nonseparable. Such an operator often occurs when co-ordinate mappings are employed [42].

Associated with the operator $\mathcal{L}$ is the bilinear form

$$T(u,v) = \sum_{i,j=1}^{d} \left(a_{i,j}\partial_{x_i}u, \partial_{x_j}v\right) - \sum_{j=1}^{d} \left(b_i\partial_{x_i}u, v\right) + \left(cu, v\right), \quad \forall u, v \in \mathrm{H}^1(\Omega). \tag{4.29}$$

Appropriate boundary conditions can be assigned by equating the weak form T with the operator $\mathcal{L}$. Since $T(u,v) = (\mathcal{L}[u], v)$ for all u, v of sufficient smoothness, it is readily seen that Dirichlet boundary conditions $\mathcal{B}[u] = u|_{\partial\Omega}$ can be imposed in this setting. However, the appropriate generalisation of Neumann boundary conditions involves the so-called *co-normal* derivative of u:

$$\mathcal{B}[u] = \sum_{i,j=1}^{d} \hat{n}_i a_{i,j} \partial_{x_j} u\big|_{\partial\Omega}, \tag{4.30}$$

where $\hat{n}$ is the unit normal vector. If $a_{i,j} = 0$ whenever $i \neq j$, these readily reduce to the standard Neumann boundary conditions, and Laplace–Neumann eigenfunctions may be used for discretisation. However, such eigenfunctions do not form a suitable basis for approximation of the general problem.

Of course, the boundary conditions (4.30) are natural, and so can be enforced in a weak manner rather than by the choice of approximation basis. However, this approach will yield a poor rate of convergence if the modified Fourier basis is employed (clearly, the co-normal derivative of the Galerkin approximation u_N will not converge uniformly to the corresponding derivative of u).

This raises the question of whether or not a simple basis of eigenfunctions satisfying the boundary conditions (4.30) can be constructed. However, this can be almost immediately disregarded: the boundary conditions are nonseparable, so such eigenfunctions do not arise from Cartesian products, thus limiting their practical use.

4.4.3 General second order Dirichlet boundary value problems

Aside from Neumann boundary conditions involving co-normal derivatives, the operator (4.28) is also frequently endowed with homogeneous Dirichlet boundary conditions: $\mathcal{B}[u] = u|_{\partial\Omega} = 0$. Since such boundary conditions are separable, and identical to those of the simple case (4.5), Laplace–Dirichlet eigenfunctions form a suitable discretisation basis. In this section, we describe some of the salient features of this approach.

It is first necessary to derive a coercivity condition for the operator (4.29). As outlined in [142, chapter 6] we assume that the operator T is *elliptic* on Ω. In other words, there exists a positive constant α such that

$$\sum_{i,j=1}^{d} a_{i,j}(x)\xi_i\xi_j \geq \alpha\|\xi\|^2, \quad \forall \xi = (\xi_1, \ldots, \xi_d) \in \mathbb{R}^d,$$

for almost every $x \in \Omega$. With this in hand, we now consider the other terms of T. Since

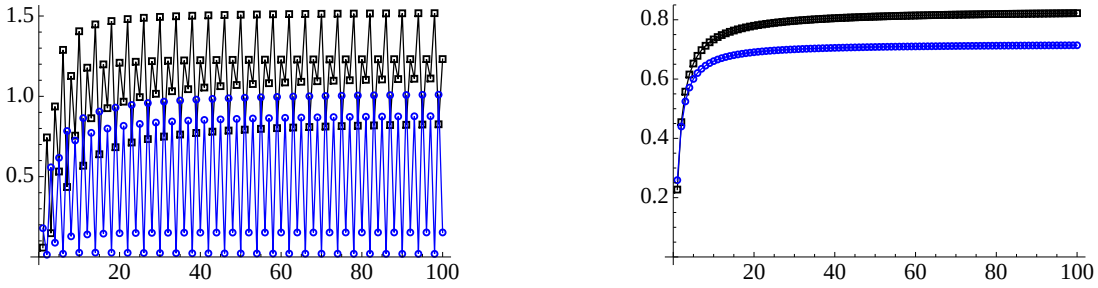


Figure 4.8: Error in the Laplace–Dirichlet Galerkin approximation. (left) scaled error $N^3|u(x_0) - u_N(x_0)|$ for $N = 1, \ldots, 100$, where $x_0 = \frac{1}{2}$ (squares) and $x_0 = -\frac{1}{2}$ (circles) respectively. (right) scaled errors $N^2\|u - u_N\|_\infty$ (squares) and $N^{\frac{5}{2}}\|u - u_N\|$ (circles).

$(b_i\partial_{x_i}u, u) = -\frac{1}{2}\left(\partial_{x_i}b_i, u^2\right)$, the ellipticity condition gives

$$T(u,u) = \sum_{i,j=1}^{d}\left(a_{i,j}\partial_{x_i}u, \partial_{x_j}u\right) + \frac{1}{2}\sum_{i=1}^{d}\left(\partial_{x_i}b_i, u^2\right) + (cu, u).$$

$$\geq \alpha\|\nabla u\|^2 + \min_{x\in\bar{\Omega}}\Big\{\sum_{j=1}^{d}\partial_{x_i}b_i(x) + c(x)\Big\}\|u\|^2.$$

Hence, provided $\min_{x\in\bar{\Omega}}\{\sum_{j=1}^{d}\partial_{x_i}b_i(x) + c(x)\} > -\frac{\alpha}{C^2}$, where C is the constant in Poincaré's inequality, the operator T is coercive.

We may now discretise this problem in the standard manner with Laplace–Dirichlet eigenfunctions. The resulting method admits similar analysis to that given in the constant coefficient case. Once more, the spectral condition number is $\mathcal{O}\left(N^2\right)$ and the matrix M_{G} corresponding to the operator $-\triangle + b_0\mathcal{I}$, where $b_0 \geq 0$ is arbitrary, is an optimal preconditioner. Analysis of the convergence rate of this approximation may also be carried out.

Furthermore, efficient solution of Galerkin's equations can be achieved by the minimax principle [47, 147]. This is based on the splitting $A_{\mathrm{G}} = M_{\mathrm{G}} + N_{\mathrm{G}}$, where the parameter b_0 appearing in M_{G} is chosen appropriately to ensure convergence of the iterative scheme. Alternatively, conjugate gradients can be employed.

Application of the Laplace–Dirichlet Galerkin approximation to the univariate problem $-(a(x)u'(x))' = f(x)$, $u(\pm 1) = 0$, where $u(x) = x^3\mathrm{e}^x - 3\mathrm{e}^{-x} + \left(2\mathrm{e}^{-1} + \mathrm{e}\right) + \left(\mathrm{e}^{-1} - 2\mathrm{e}\right)x$ and $a(x) = 1 + \mathrm{e}^x$, is considered in Figure 4.8. As demonstrated, the uniform and $\mathrm{L}^2(-1,1)$ norm errors are $\mathcal{O}(N^{-2})$ and $\mathcal{O}(N^{-\frac{5}{2}})$ respectively. Moreover, much like the case of the approximation $\mathcal{F}_N[u]$, the pointwise convergence rate is one power of N faster away from the endpoints, i.e. $\mathcal{O}(N^{-3})$. Complementing these results, Table 4.2 verifies both $\mathcal{O}(N^2)$ condition number of the Galerkin matrix and its optimal preconditioning via the diagonal matrix M_{G}.

4.4.4 Other boundary conditions

As discussed in Section 2.11, Laplace eigenfunctions techniques can be applied to a variety of problems with different boundary conditions. The key element is that the boundary conditions

	$N = 10$	$N = 20$	$N = 30$	$N = 40$
$\kappa_($A_G$)$	507.883	2187.79	5113.48	9313.57
$N^{-2}\kappa(A_G)$	5.07883	5.46946	5.68165	5.82098
$\kappa(A_G M_G^{-1})$	2.83349	2.89153	2.90876	2.91682

Table 4.2: Condition numbers for the Laplace–Dirichlet Galerkin matrix A_G and the preconditioned matrix $A_G M_G^{-1}$.

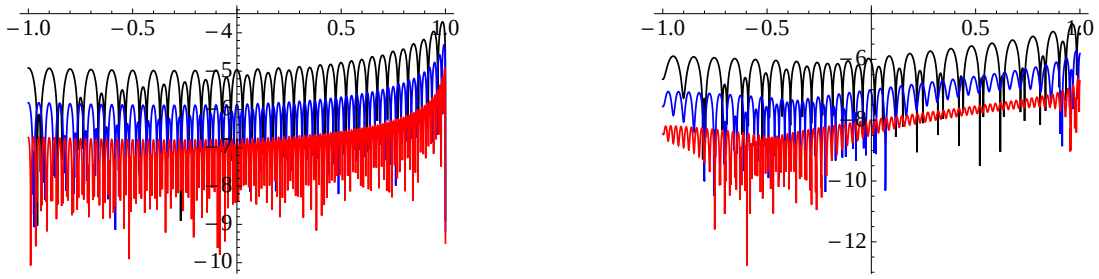


Figure 4.9: Pointwise error $\log_{10}|u(x) - u_N(x)|$ against $x \in [-1, 1]$ for $N = 20, 40, 80$ (in descending order), where u_N is the Galerkin approximation to the problem with mixed (left) and Robin (right) boundary conditions.

be separable, thus endowing the eigenfunctions with a tensor-product structure. As previously suggested, even for second-order problems, there are a whole host of multivariate nonseparable boundary conditions, whose numerical treatment requires considerable care, regardless of the particular approximation scheme used.

To illustrate the application of this approach beyond Neumann or Dirichlet boundary conditions, consider the constant coefficient problem (4.19) with either the mixed $u(-1) = u'(1) = 0$ or Robin $u'(\pm 1) + 3u(\pm 1) = 0$ boundary conditions. Figure 4.9 presents numerical results for the discretisation of these problems based on the Laplace eigenfunctions (2.52) and (2.54) respectively. In correspondence with our expectations (see Section 2.11), the approximation to the Robin problem offers an $\mathcal{O}(N^{-3})$ uniform error, whereas this figure is $\mathcal{O}(N^{-2})$ for the problem with mixed boundary conditions. Note that, as with the Laplace–Dirichlet case, the approximation to the mixed problem converges faster away from the endpoints. Yet, much like in the modified Fourier case, this effect does not occur in the Robin setting.

4.4.5 Higher-order problems

Higher-order boundary value problems often arise in the mathematical modelling of physical phenomena. Even-order differential equations, in particular, arise in astrophysics, structural mechanics and geophysics [55, 79]. Typical examples of such problems include

$$\triangle^2 u - a\triangle u + bu = f, \quad u|_\Gamma = \hat{n} \cdot \nabla u|_\Gamma = 0, \tag{4.31}$$

which serves as a model for the clamped rod problem (and also arises in the time discretisation of various models for flame propagation [147]), and the sixth order problem

$$-(\triangle - b)^3 u = f, \quad u|_\Gamma = \hat{n} \cdot \nabla u|_\Gamma = \triangle u|_\Gamma = 0, \tag{4.32}$$

frequently occurring in astrophysics [15, 30]. Note that both (4.31) and (4.32) are special cases of the general $2q^{\text{th}}$ order Dirichlet problem

$$\mathcal{L}[u] = f, \quad \mathcal{B}_r[u] = 0, \quad r = 0, \ldots, q-1, \tag{4.33}$$

where $\mathcal{B}_{2r}[u] = \triangle^r u|_\Gamma$ and $\mathcal{B}_{2r+1}[u] = \hat{n}.\nabla\triangle^r u|_\Gamma$.

Due to the large number of boundary conditions, finite difference or finite element schemes for (4.33) are typically cumbersome to implement, as are spectral collocation methods [79]. To counter this, basis functions are sought that individually satisfy boundary conditions. Though spectral–Galerkin schemes for such problems can be constructed from generalised Jacobi polynomials [78], there are, in general, far fewer effective methods for higher-order problems than for the second-order case.

We now turn our attention to the discretisation of such problems by Laplace eigenfunction techniques. Consider the univariate $2q^{\text{th}}$ order problem

$$\mathcal{L}[u](x) = f(x), \quad x \in (-1,1), \quad u(\pm 1) = \ldots u^{(q-1)}(\pm 1) = 0. \tag{4.34}$$

To design an approximation scheme, we first seek basis functions that match boundary conditions individually. Thus, given a finite set of Laplace eigenfunctions, we construct new basis functions obeying such conditions by taking appropriate linear combinations. Automatically, Laplace–Dirichlet eigenfunctions satisfy $\lceil \frac{q}{2} \rceil$ boundary conditions and Laplace–Neumann $\lfloor \frac{q}{2} \rfloor$. Linear combinations involving as few eigenfunctions as possible are naturally preferable (they lead to lower bandwidth matrices). This indicates that when q is odd, we should use Laplace–Dirichlet eigenfunctions. Conversely, when q is even, we choose Laplace–Neumann eigenfunctions (due to their faster rate of convergence).[15]

In either case, the remaining boundary conditions are satisfied by forming appropriate linear combinations. We write

$$\Phi_n^{[i]}(x) = \phi_n^{[i]}(x) + \sum_{m=1}^{\lfloor \frac{q}{2} \rfloor} a_{n,m}^{[i]} \phi_{n+m}^{[i]}(x),$$

where $\phi_n^{[i]}$ are univariate Laplace–Dirichlet (respectively Laplace–Neumann) eigenfunctions, and the values $a_{n,m}^{[i]} \in \mathbb{R}$ enforce the remaining boundary conditions. In particular, when $q = 2$,

$$\Phi_n^{[0]}(x) = \cos n\pi x + \cos(n+1)\pi x, \quad \Phi_n^{[1]}(x) = \sin(n-\tfrac{1}{2})\pi x + \sin(n+\tfrac{1}{2})\pi x, \tag{4.35}$$

and, for $q = 3$,

$$\Phi_n^{[0]}(x) = \cos(n-\tfrac{1}{2})\pi x + \frac{2n-1}{2n+1}\cos(n+\tfrac{1}{2})\pi x, \quad \Phi_n^{[1]}(x) = \sin n\pi x + \frac{n}{n+1}\sin(n+1)\pi x.$$

The simple extension to the d-variate cube follows immediately via Cartesian products.

Suppose that we define the finite-dimensional space $X_N = \{\Phi_n^{[i]} : n \in I_N, i \in \{0,1\}^d\}$. Note that $X_N \subseteq \mathcal{S}_{N+\lfloor \frac{q}{2} \rfloor}$. Our Galerkin approximation $u_N \in X_N$ to (4.34) is then given

[15] Polyharmonic–Dirichlet eigenfunctions are seemingly a natural choice for such problems. Indeed, in the constant coefficient case at least, there is no barrier to their use. However, unlike the Laplace case (see Section 4.4.1), the entries of the discretisation matrix corresponding to a variable-coefficient problem are not known explicitly, and, as of this moment, there is no fast method for computing such values.

by the relation $T(u_N, \Phi) = (f, \Phi)$, $\forall \Phi \in X_N$, where T is the weak form corresponding to $\mathcal{L}$. Provided the form T is $\mathrm{H}_0^q(\Omega)$-continuous and coercive, we immediately obtain the error estimate

$$|u - u_N|_q \leq \frac{\gamma}{\omega} \inf_{\Phi \in X_N} |u - \Phi|_q, \tag{4.36}$$

where $|\cdot|_q$ is the inner product on $\mathrm{H}_0^q(\Omega)$. Note that this infimum is attained precisely when $\Phi = \mathcal{H}_N[u]$, where $\mathcal{H}_N : \mathrm{L}^2(-1,1) \to X_N$ is the orthogonal projection onto X_N.

For the problems (4.31) and (4.32), the corresponding weak forms are given by

$$T(u,v) = (\triangle u, \triangle v) + a\,(\nabla u, \nabla v) + b(u,v), \quad \forall u, v \in \mathrm{H}_0^2(\Omega),$$

and

$$T(u,v) = (\nabla\triangle u, \nabla\triangle v) + 3b\,(\triangle u, \triangle v) + 3b^2\,(\nabla u, \nabla v) + b^3(u,v), \quad \forall u, v \in \mathrm{H}_0^3(\Omega),$$

respectively. They are continuous and coercive, provided $a, b \geq 0$ for (4.31) and $b \geq 0$ for (4.32).

Returning to the general setting, we note that, if the operator $\mathcal{L}$ involves only even-order derivatives, as is the case with (4.31) and (4.32), then the corresponding Galerkin matrix A_{G} is banded, with bandwidth $1 + 2\lfloor \frac{q}{2} \rfloor$. In particular, for (4.31) and (4.32), A_{G} is tridiagonal, hence easily solvable.[16] In general, the matrix A_{G} is dense with condition number $\kappa(A_{\mathrm{G}}) = \mathcal{O}(N^{2q})$. Hence, the design of effective preconditioners is of paramount importance. Yet optimal preconditioning, and therefore also fast solution via conjugate gradients, is readily obtained upon considering the banded matrix corresponding to the highest-order derivatives in $\mathcal{L}$.

We now wish to assess the rate of convergence of this approximation. For the sake of simplicity, we restrict ourselves to the univariate, $q = 2$ case. Identification of the projector $\mathcal{H}_N[u]$ for general $2q^{\text{th}}$ order multivariate problems is rather unpleasant, so instead we focus solely on this scenario:

Lemma 4.22. *Suppose that* $X_N = \{\Phi_n^{[i]} : n = 0, \ldots, N, i \in \{0,1\}\}$, *where the functions* $\Phi_n^{[i]}$ *are given by (4.35), and that* $\mathcal{H}_N : \mathrm{L}^2(-1,1) \to X_N$ *is the orthogonal projection. Then*

$$\mathcal{H}_N[f](x) = \mathcal{F}_{N+1}[f](x) - \sum_{i=0}^{1} a_N^{[i]} \sum_{n=0}^{N+1} (-1)^{n+i} \phi_n^{[i]}(x),$$

where $a_N^{[i]} = \frac{1}{2(N+i+1)} \big\{(-1)^i \mathcal{F}_{N+1}[f](1) + \mathcal{F}_{N+1}[f](-1)\big\}$. *In particular, if* $u_N \in X_N$ *is the modified Fourier–Galerkin approximation to the univariate problem (4.34) with* $q = 2$, *then* $|u - u_N|_2 \leq c\|u\|_4 N^{-\frac{3}{2}}$ *for some positive constant* c *independent of* N *and* u.

Proof. For the first part, it suffices to show that $\mathcal{H}_N[f] \in X_N$ and that $(\mathcal{H}_N[f], \Phi_n^{[i]}) = (f, \Phi_n^{[i]})$, $n = 0, \ldots, N$, $i \in \{0,1\}$. Both follow immediately from the properties of the functions $\phi_n^{[i]}$ and $\Phi_n^{[i]}$.

[16] In theory, rather than enforcing a Galerkin condition, we could define the approximation $u_N \in X_N$ via a Petrov–Galerkin criterion: $T(u_N, \phi) = (f, \phi)$, $\forall \phi \in \mathcal{S}_N$. This would lead to a sparser matrix (with only $1 + \lfloor \frac{q}{2} \rfloor$ nonzero diagonals). However, the self-adjointness of the operator $\mathcal{L}$ is now lost when passing to the discrete problems—a less than desirable property.

For the second result, in view of (4.36), an estimate for $|u-u_N|_2$ is provided by $|u-\mathcal{H}_N[u]|_2$. Note that

$$u(x) - \mathcal{H}_N[u](x) = \{u(x) - \mathcal{F}_{N+1}[u](x)\} + \sum_{i=0}^{1} a_N^{[i]} \sum_{n=0}^{N+1} (-1)^{n+i} \phi_n^{[i]}(x),$$

where $a_N^{[i]} = \frac{1}{2(N+i+1)} \left\{(-1)^i \mathcal{F}_{N+1}[u](1) + \mathcal{F}_{N+1}[u](-1)\right\}$. Since $u(\pm 1) = 0$ and $u'(\pm 1) = 0$, we have $\mathcal{F}_N[u](\pm 1) = \mathcal{O}\left(N^{-3}\right)$. Hence $a_N^{[i]} = \mathcal{O}\left(N^{-4}\right)$. It now follows from Lemma 2.25 that

$$\begin{aligned} |u - \mathcal{H}_N[u]|_2^2 &= |u - \mathcal{F}_{N+1}[u]|_2^2 + \sum_{i=0}^{1} (a_N^{[i]})^2 \sum_{n=0}^{N+1} (\mu_n^{[i]})^2 \\ &\leq cN^{-3}\|u\|_4^2 + cN^{-8}\|u\|_4^2 \sum_{n=1}^{N} n^4 \leq c\|u\|^2 N^{-3}\|u\|_4^2, \end{aligned}$$

as required. □

This lemma covers the convergence rate of the modified Fourier–Galerkin approximation to the clamped rod problem (4.31), for example. However, it turns out that we can provide a far more accurate assessment in this case. Specifically, the banded structure of A_{G} allows us to determine an explicit expression for the Galerkin approximation u_N:

Theorem 4.23. *The modified Fourier–Galerkin approximation $u_N \in X_N$ to the univariate problem (4.31) is given by*

$$u_N(x) = \mathcal{F}_{N+1}[u](x) - \sum_{i=0}^{1} \frac{\mathcal{F}_{N+1}[u](1) + (-1)^i \mathcal{F}_{N+1}[u](-1)}{\mathcal{F}_{N+1}[p^{[i]}](1) + (-1)^i \mathcal{F}_{N+1}[p^{[i]}](-1)} \mathcal{F}_{N+1}[p^{[i]}](x), \tag{4.37}$$

where $p^{[i]}$ is the smooth function with modified Fourier coefficients $\widehat{p^{[i]}}_n^{[1-i]} = 0$, $\widehat{p^{[i]}}_n^{[i]} = (-1)^{n+i}(\lambda_n^{[i]})^{-1}$ and $\lambda_n^{[i]} = (\mu_n^{[i]})^2 + a\mu_n^{[i]} + b$. In particular,

$$\|u - u_N\|_\infty \leq c\|u\|_4 N^{-3}, \quad \|u - u_N\|_r \leq c\|u\|_4 \max\{N^{r-\frac{7}{2}}, N^{-3}\}, \quad r \in \mathbb{N}_0,$$

where $c > 0$ is independent of N and u.

Proof. We first verify that $u_N(\pm 1) = 0$. We have

$$\begin{aligned} u_N(1) + (-1)^l u_N(-1) &= \mathcal{F}_{N+1}[u](1) + (-1)^l \mathcal{F}_{N+1}[u](-1) \\ &\quad - \sum_{i=0}^{1} \frac{\mathcal{F}_{N+1}[u](1) + (-1)^i \mathcal{F}_{N+1}[u](-1)}{\mathcal{F}_{N+1}[p^{[i]}](1) + (-1)^i \mathcal{F}_{N+1}[p^{[i]}](-1)} \left\{\mathcal{F}_{N+1}[p^{[i]}](1) + (-1)^l \mathcal{F}_{N+1}[p^{[i]}](-1)\right\}. \end{aligned}$$

Note that $p^{[0]}$ is even, whereas $p^{[1]}$ is odd. Hence $\mathcal{F}_{N+1}[p^{[i]}](1) + (-1)^{1-i}\mathcal{F}_{N+1}[p^{[i]}](-1) = 0$. It now follows that $u_N(1) + (-1)^l u_N(-1) = 0$, $l = 0, 1$, as required.

Next, we must verify that $(\mathcal{L}[u_N], \Phi_n^{[i]}) = (f, \Phi_n^{[i]})$ for $n = 0, \ldots, N$, $i = 0, 1$. Note that $\mathcal{L}[\phi_n^{[i]}] = \lambda_n^{[i]} \phi_n^{[i]}$. Hence

$$\left(\mathcal{L}[u_N], \phi_n^{[i]}\right) = \lambda_n^{[i]} \hat{u}_n^{[i]} - \frac{\mathcal{F}_{N+1}[u](1) + (-1)^i \mathcal{F}_{N+1}[u](-1)}{\mathcal{F}_{N+1}[p^{[i]}](1) + (-1)^i \mathcal{F}_{N+1}[p^{[i]}](-1)} (-1)^{n+i}.$$

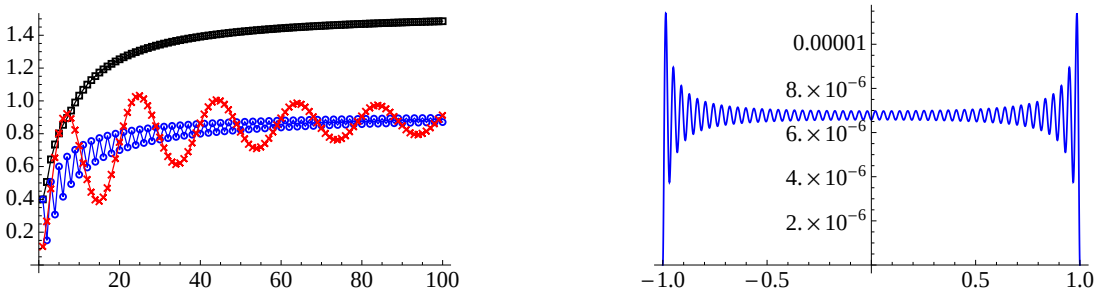


Figure 4.10: Error in the approximation u_N to the problem (4.31) with $a = b = 0$ and exact solution $(\sin 4x - x \sin 4)^2$. (left) scaled errors $N^3 \|u - u_N\|_\infty$ (squares) and $N^3 |u(x_0) - u_N(x_0)|$, where $x_0 = 0$ (circles) and $x_0 = \frac{9}{10}$ (crosses). (right) pointwise error $|u(x) - u_{50}(x)|$ for $-1 \leq x \leq 1$.

Recalling that $\Phi_n^{[i]} = \phi_n^{[i]} + \phi_{n+1}^{[i]}$, we obtain $(\mathcal{L}[u_N], \Phi_n^{[i]}) = \lambda_n^{[i]} \hat{u}_n^{[i]} + \lambda_{n+1}^{[i]} \hat{u}_{n+1}^{[i]}$. Now consider $\hat{u}_n^{[i]}$. Using the asymptotic expansion (2.11), we have

$$\hat{u}_n^{[i]} = -\frac{1}{\mu_n^{[i]}} \left(u'', \phi_n^{[i]} \right) = \frac{(-1)^{n+i+1}}{(\mu_n^{[i]})^2} \left\{ u'''(1) + (-1)^{i+1} u'''(-1) \right\} + \frac{1}{(\mu_n^{[i]})^2} \left(u^{(4)}, \phi_n^{[i]} \right).$$

Since $\mathcal{L}[u] = f$, this gives

$$\begin{aligned} \hat{f}_n^{[i]} = \left(\mathcal{L}[u], \phi_n^{[i]} \right) &= \left(u^{(4)}, \phi_n^{[i]} \right) - a \left(u'', \phi_n^{[i]} \right) + b \hat{u}_n^{[i]} \\ &= \lambda_n^{[i]} \hat{u}_n^{[i]} + (-1)^{n+i} \left\{ u'''(1) + (-1)^{i+1} u'''(-1) \right\}. \end{aligned}$$

It follows immediately that $\lambda_n^{[i]} \hat{u}_n^{[i]} + \lambda_{n+1}^{[i]} \hat{u}_{n+1}^{[i]} = \hat{f}_n^{[i]} + \hat{f}_{n+1}^{[i]} = (f, \Phi_n^{[i]})$, thus completing the first part of the proof. Error estimates are obtained from standard properties of the modified Fourier projector $\mathcal{F}_{N+1}[\cdot]$ and the fact that $u(\pm 1) = u'(\pm 1) = 0$. □

In Figure 4.10 we confirm the result of this theorem. As illustrated, the uniform error is $\mathcal{O}\left(N^{-3}\right)$. Moreover, the pointwise error does not decay at a faster rate in this case, a fact which is easily verified upon scrutinising the expression (4.37).

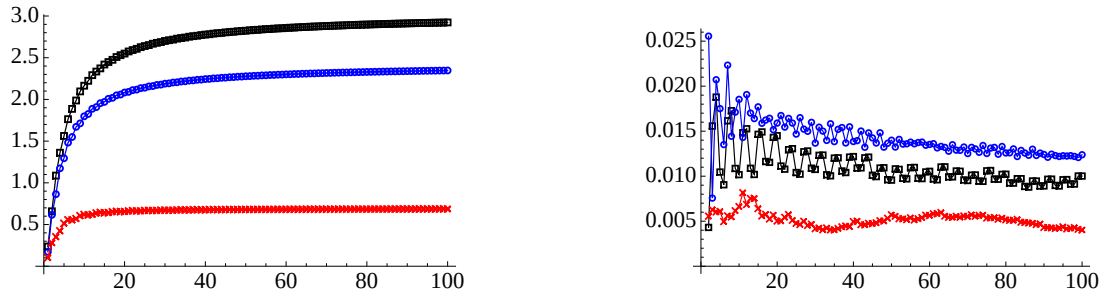
The application of Laplace–Dirichlet eigenfunctions to the sixth order problem (4.32) is considered in Figure 4.11(a). Once more, we observe that the uniform error is $\mathcal{O}\left(N^{-3}\right)$, a result which can be established in the same manner as the biharmonic case studied in Theorem 4.23. Note also that the pointwise convergence rate is not, in general, faster than $\mathcal{O}\left(N^{-3}\right)$.

Figure 4.11(b) gives numerical results for the application of this method to the bivariate clamped rod problem (4.31). As in the univariate cases, the uniform error remains cubic: when a hyperbolic cross is used, $\|u - u_N\|_\infty = \mathcal{O}\left(N^{-3} \log N\right)$.

We mention in passing that, in certain applications, the boundary value problems (4.31) and (4.32) are alternatively specified with so-called *second* boundary conditions of either Dirichlet or Neumann type. In other words,

$$\triangle^r u|_\Gamma = 0, \quad r = 0, \ldots, q-1, \quad \text{or} \quad \hat{n}.\nabla \triangle^r u|_\Gamma = 0, \quad r = 0, \ldots, q-1, \tag{4.38}$$

respectively [15, 30]. Solution of such problems directly via polynomial-based methods is typically unwise. Discretisation matrices suffer from extreme ill-conditioning, and the resultant accumulation of round-off error typically destroys any approximation quality [146].



(*a*) (*b*)

Figure 4.11: (a) error in the approximation u_N to the problem (4.32) with $b = 2$ and exact solution $u(x) = 5(\mathrm{e}^x - \cosh 1 - x \sinh 1)^3$. Scaled errors $N^3\|u - u_N\|_\infty$ (squares) and $N^3|u(x_0) - u_N(x_0)|$, where $x_0 = 0$ (circles) and $x_0 = -\frac{1}{2}$ (crosses). (b) approximation of the bivariate problem (4.31) with $a = b = 0$ and exact solution $u(x_1, x_2) = (\cosh x_1 - \cosh 1)^2 \left(x_2^2 - 1\right)^2$. Scaled errors $N^3(\log N)^{-1}\|u - u_N\|_\infty$ (squares) and $N^3(\log N)^{-1}|u(x_0) - u_N(x_0)|$, where $x_0 = 0$ (circles) and $x_0 = -\frac{1}{2}$ (crosses).

Usually such problems are treated by solving decoupled systems of second order equations. Nonetheless, direct solution of such problems is extremely easy with Laplace eigenfunctions. Laplace–Dirichlet or Laplace–Neumann eigenfunctions automatically satisfy the boundary conditions (4.38), thus immediately permitting discretisation. In particular, for (4.31) or (4.32) the Galerkin matrix is diagonal. All the properties of the $q = 1$ case studied previously are easily generalised to this particular setting.

This section completes our study of modified Fourier–Galerkin methods for boundary value problems. As commented in Section 4.3.5, convergence acceleration is a primary step towards the design of increasingly effective methods based on modified Fourier expansions. The next chapter is devoted to this task.

Chapter 5

Accelerating convergence

5.1 Introduction

A central drawback of expansions in Laplace or polyharmonic eigenfunctions, as demonstrated in previous chapters, is that the rate of convergence may be slow. Nonetheless, the analysis provided in Chapters 2 and 3 highlights the precise criteria that determine such convergence rates: namely, derivative conditions. The purpose of this chapter is to introduce and analyse a technique to accelerate convergence based on these conditions.

There is an abundance of devices for the convergence acceleration of Fourier-like series. We defer a discussion of the relative merits of such methods to Section 5.10. The primary technique we consider in this chapter, the *polynomial subtraction* device, is arguably one of the simplest and best known in its most basic form. However, we present a number of significant generalisations and refinements of this approach, including a full extension to functions defined on the d-variate cube. Outside of function approximation, a central motivation for developing such a technique is its potential for incorporation into spectral discretisations of boundary value problems based on modified Fourier expansions—the main application considered in this thesis. We return briefly to this topic in Chapter 6.

Unfortunately, polynomial subtraction has a number of well-documented drawbacks [54, 62]. Subsequently, we shall describe these issues in greater detail. Nevertheless, through the work of this chapter, we will demonstrate how these issues can be successfully bypassed, leading to a robust, effective method for accelerating convergence. Moreover, the incorporation of a hyperbolic cross index set leads to a highly accurate approximation scheme for multivariate functions comprising relatively small numbers of terms.

For reasons of clarity and simplicity, modified Fourier expansions form the principal concern of this chapter. There are no theoretical barriers to adjusting this device for use with other eigenfunction expansions (including, for example, the polyharmonic expansions of Chapter 3). Needless to say, the vast majority of existing literature on convergence acceleration deals with Fourier series. We mention in passing that both the theory and practical aspects presented in this chapter are equally applicable to this case with only minor adjustments (the main theoretical distinction being a convergence rate one power of N slower than that obtained from modified Fourier expansions, a difference which provided the original motivating factor for study of the latter).

Convergence acceleration of modified Fourier expansions[1] was first considered in [87], where the polynomial subtraction technique (in its most basic form) was extended to multivariate expansions. The key problem with this device is that it requires rather extensive knowledge of the function being approximated, which, in general, is not available. In [4], an extension of this approach, which successfully circumvents this issue, was developed and analysed (based on the work of [54] and [16]). The majority of this chapter originates from the material presented therein.

The key results of this chapter are as follows:

1. If the first $k \in \mathbb{N}_0$ odd derivatives of a function f are known explicitly on the boundary of the d-variate cube, then the k^{th} *polynomial subtraction approximation* of f can be constructed using only the modified Fourier coefficients $\hat{f}_n^{[i]}$ and such derivatives. The corresponding uniform convergence rate is $\mathcal{O}\left(N^{-2k-1}\right)$.
2. Such derivatives can be approximated by linear combinations of the coefficients $\hat{f}_n^{[i]}$. The resulting approximation, the k^{th} *Eckhoff approximation* of f, converges no slower than the corresponding polynomial subtraction approximation. In other words, using only the coefficients $\hat{f}_n^{[i]}$, an approximation can be constructed with a uniform convergence rate of $\mathcal{O}\left(N^{-2k-1}\right)$.
3. Provided certain parameters are selected according to an explicit criterion, the pointwise convergence rate of the k^{th} Eckhoff approximation inside the domain $\Omega = (-1,1)^d$ is $\mathcal{O}\left(N^{-3k-2}\right)$, a full factor of $\mathcal{O}\left(N^{k}\right)$ faster than the corresponding polynomial subtraction approximation.
4. The cost of constructing the standard Eckhoff approximation is $\mathcal{O}\left(k^d N^d\right)$. However, a hyperbolic cross index set can be incorporated into this approximation, thereby reducing this figure to $\mathcal{O}\left(k^d N (\log N)^{d-1}\right)$. The uniform convergence rate is unaffected, aside from a logarithmic factor.
5. Standard implementations of both polynomial subtraction and Eckhoff's method employ certain polynomials to interpolate the requisite derivatives of the function f. This leads to extreme ill-conditioning. However, a vast improvement is obtained by replacing such functions with Laplace–Dirichlet eigenfunctions. Combined with a judicious choice of various parameters and a least squares procedure, this yields a robust, effective method possessing both high accuracy and good numerical stability.

The eventual objective of this chapter is the construction and subsequent analysis of Eckhoff's approximation for functions defined on the d-variate cube. Before doing so, however, we commence with the case of the unit interval. Moreover, since Eckhoff's method is based on the polynomial subtraction technique, we first describe this device in this domain.

5.2 Univariate polynomial subtraction

As established in Chapter 2, if a univariate function $f \in \mathrm{H}^{2k+2}(-1,1)$ obeys the first k Neumann derivative conditions $f^{(2r+1)}(\pm 1) = 0$, $r = 0, \ldots, k-1$, then its modified Fourier expansion converges uniformly at a rate of $\mathcal{O}\left(N^{-2k-1}\right)$, as opposed to $\mathcal{O}\left(N^{-1}\right)$ (see Theorem 2.20). Suppose now that f does not satisfy such conditions. We write f in terms of its

[1] We stress that this refers to *modified* Fourier expansions. Convergence acceleration of classical Fourier series has a much more extensive history, as we describe in further detail in Section 5.10.

Lanczos representation $f = (f - g_k) + g_k$ [112, 118], where the function g_k is chosen so that $f^{(2r+1)}(\pm 1) = g_k^{(2r+1)}(\pm 1)$, $r = 0, \ldots, k-1$. Since $f - g_k$ obeys the first k derivative conditions, the new approximation

$$\mathcal{F}_{N,k}[f] = \mathcal{F}_N[f - g_k] + g_k, \tag{5.1}$$

converges uniformly to f at the faster rate of N^{-2k-1}. This is the *polynomial subtraction* technique. We refer to $\mathcal{F}_{N,k}$, as defined in (5.1), as the k^{th} *polynomial subtraction approximation* of f (for convenience, we interpret $\mathcal{F}_{N,0}[f]$ as $\mathcal{F}_N[f]$). This idea dates back to Krylov [109], and was studied more formally in [102, 112]. Since then, it has been widely considered in the context of Fourier series [19, 67, 97, 118]. Its application to modified Fourier expansions was originally considered in [87, 94].

Faster convergence of $\mathcal{F}_{N,k}[f]$ to f in various norms is guaranteed by Theorem 2.20 and Lemma 2.25. For clarity, we now restate these results explicitly in terms of $\mathcal{F}_{N,k}[f]$:

Theorem 5.1. *Suppose that* $k \in \mathbb{N}_0$, $f \in \mathrm{H}^{2k+2}(-1,1)$ *and that* $\mathcal{F}_{N,k}[f]$ *is given by (5.1). Then, the error* $\|f^{(r)} - (\mathcal{F}_{N,k}[f])^{(r)}\|_\infty$ *is* $\mathcal{O}\left(N^{r-2k-1}\right)$ *for* $r = 0, \ldots, 2k$. *If, additionally,* $f \in \mathrm{H}^{2k+3}(-1,1)$, *then the convergence rate of* $(\mathcal{F}_{N,k}[f])^{(r)}$ *to* $f^{(r)}$ *is* $\mathcal{O}\left(N^{r-2k-2}\right)$ *uniformly in compact subsets of* $(-1,1)$ *for* $r = 0, \ldots, 2k+1$.

Theorem 5.2. *Suppose that* $f \in \mathrm{H}^{2k+2}(-1,1)$ *and that* $\mathcal{F}_{N,k}[f]$ *is as in Theorem 5.1. Then* $\|f - \mathcal{F}_{N,k}[f]\|_r$ *is* $\mathcal{O}(N^{r-2k-\frac{3}{2}})$ *for* $r = 0, \ldots, 2k+1$.

5.2.1 Construction of the subtraction function

Vital to the application of this technique is the construction of the function g_k. To accomplish this, it is convenient to first recall the values

$$\mathcal{A}_r^{[i]}[f] = (-1)^r \left[f^{(2r+1)}(1) + (-1)^{i+1} f^{(2r+1)}(-1) \right], \quad i \in \{0,1\}, \quad r \in \mathbb{N}_0, \tag{5.2}$$

which were introduced in Section 2.7. Note that $f^{(2r+1)}(\pm 1) = 0$ if and only if $\mathcal{A}_r^{[i]}[f] = 0$.

The slow convergence of modified Fourier expansions stems from such values being non-zero. In other words, there are 'jumps' in the odd derivatives of f at the endpoints $x = \pm 1$. For this reason, the values $\mathcal{A}_r^{[i]}[f]$ are often referred to as *jump values*.[2] Additionally (and intimately related), such values also determine the rate of decay of the modified Fourier coefficients $\hat{f}_n^{[i]}$, as considered in Chapter 2. Since

$$\hat{f}_n^{[i]} = \sum_{r=0}^{k-1} \frac{(-1)^{n+i}}{(\mu_n^{[i]})^{r+1}} \mathcal{A}_r^{[i]}[f] + \frac{(-1)^k}{(\mu_n^{[i]})^k} \widehat{f^{(2k)}}_n^{[i]}, \quad i \in \{0,1\}, \quad n \in \mathbb{N}, \tag{5.3}$$

it is apparent that $\hat{f}_n^{[i]} = \mathcal{O}\left(n^{-2k-2}\right)$, provided the first k jump values vanish. Equivalent statements can also be constructed regarding the smoothness of the periodic extension of the function f (see Section 2.3).

The definition of the function g_k may be restated in terms of such values. That is to say, we seek a function g_k such that

$$\mathcal{A}_r^{[i]}[g_k] = \mathcal{A}_r^{[i]}[f], \quad r = 0, \ldots, k-1, \quad i = 0, 1. \tag{5.4}$$

[2] This interpretation and terminology is standard practice in the area of computational Fourier analysis that deals with the resolution of the Gibbs phenomenon [72, 156].

To construct g_k, we introduce (smooth) functions $p_0^{[i]}, \dots, p_{k-1}^{[i]}$, where $p_r^{[i]}$ is even (respectively odd) if $i = 0$ ($i = 1$), that satisfy the conditions

$$\mathcal{A}_r^{[i]}\left[p_s^{[i]}\right] = \delta_{r,s}, \quad r, s = 0, \dots, k-1, \quad i \in \{0, 1\}. \tag{5.5}$$

We say that $p_0^{[i]}, \dots, p_{k-1}^{[i]}$ are *cardinal functions for the first k derivative conditions*. With this in hand, we define g_k by

$$g_k(x) = \sum_{i=0}^{1} \sum_{r=0}^{k-1} \mathcal{A}_r^{[i]}[f] p_r^{[i]}(x), \quad x \in [-1, 1]. \tag{5.6}$$

In standard implementations of this device, the r^{th} cardinal function $p_r^{[i]}$ is specified to be a polynomial of degree $2(r+1)-i$ [16, 54, 112]. This explains the name 'polynomial subtraction'. In this case, we refer to the functions $\{p_r^{[i]}\}$ as *cardinal polynomials*.

Having said this, a little care is necessary. It is not immediately obvious that such polynomials exist. As discussed in [94], the condition (5.5) is an example of a so-called *Birkhoff–Hermite* interpolation problem [116] (the interpolation of non-consecutive derivatives by a polynomial). In general, such problems are not guaranteed to have a solution. Despite this warning, however, it is easily confirmed that this particular problem is uniquely solvable [94].[3]

The first few cardinal polynomials are given by

$$\begin{aligned}
p_0^{[0]}(x) &= \frac{1}{4}x^2, \quad p_0^{[1]}(x) = \frac{1}{2}x, \\
p_1^{[0]}(x) &= \frac{1}{48}x^2(x^2-2), \quad p_0^{[1]}(x) = \frac{1}{12}x(x^2-3), \\
p_2^{[0]}(x) &= \frac{1}{1440}x^2\left(x^4 - 5x^2 + 7\right), \quad p_2^{[1]}(x) = \frac{1}{240}x(x^2-5)^2,
\end{aligned}$$

from which it can be explicitly verified that the modified Fourier coefficient of $p_n^{[i]}$, which we write $\widehat{p_r}_n^{[i]}$ (note that, since $p_r^{[i]}$ is even (respectively odd) for $i = 0$ ($i = 1$), the coefficient of $p_r^{[i]}$ corresponding to $\phi_n^{[1-i]}$ is zero), satisfies $\widehat{p_r}_n^{[i]} = (-1)^{n+i}(\mu_n^{[i]})^{-r-1}$. In particular, recalling the expansion (2.11) of the coefficient $\hat{f}_n^{[i]}$,

$$\hat{f}_n^{[i]} = \sum_{r=0}^{k-1} \mathcal{A}_r^{[i]}[f]\widehat{p_r}_n^{[i]} + \frac{(-1)^k}{(\mu_n^{[i]})^k}\widehat{f^{(2k)}}_n^{[i]} = \widehat{g_k}_n^{[i]} + \frac{(-1)^k}{(\mu_n^{[i]})^k}\widehat{f^{(2k)}}_n^{[i]} = \widehat{g_k}_n^{[i]} + \mathcal{O}\left(n^{-2k-2}\right). \tag{5.7}$$

Hence, the function g_k can be viewed as an approximation to f which replicates its modified Fourier coefficients to high order. This viewpoint is the basis for the generalisation of the polynomial subtraction technique that we consider in the sequel, namely Eckhoff's method.

Though the cardinal functions $p_r^{[i]}$ provide the most simple interpretation, there is no need for this restriction. Suppose that the (smooth) functions $q_0^{[i]}, \dots, q_{k-1}^{[i]}$ have the property that the interpolation problem

$$\begin{aligned}
&\text{find} \quad \{a_r^{[i]} : i \in \{0, 1\}, r = 0, \dots, k-1\} \\
&\text{such that} \quad \sum_{s=0}^{k-1} a_s^{[i]} \mathcal{A}_r^{[i]}\left[q_s^{[i]}\right] = b_r^{[i]}, \quad i \in \{0, 1\}, \quad r = 0, \dots, k-1,
\end{aligned} \tag{5.8}$$

[3] The resultant polynomials are shifted Bernoulli polynomials [112]. For this reason, polynomial subtraction is also referred to as the *Bernoulli method* [62, 67].

has a unique solution for all choices $b_r^{[i]} \in \mathbb{R}$. Then we may construct g_k as a linear combination of such functions:

$$g_k(x) = \sum_{i=0}^{1} \sum_{r=0}^{k-1} \tilde{\mathcal{A}}_r^{[i]}[f] q_r^{[i]}(x), \quad x \in [-1, 1],$$

where the values $\tilde{\mathcal{A}}_r^{[i]}[f]$ enforce (5.4).[4] We refer to $\{q_r^{[i]} : i \in \{0,1\}, r = 0, \ldots, k-1\}$ as a *subtraction basis*. Given such a basis, appropriate cardinal functions $p_r^{[i]}$ can always be constructed by taking suitable linear combinations of the functions $q_r^{[i]}$. For example, the aforementioned cardinal polynomials are derived from the subtraction basis consisting of the monomials $q_r^{[i]}(x) = x^{2(r+1)-i}$.

The resulting approximations $\mathcal{F}_{N,k}[f]$ based on either the cardinal or non-cardinal formulations are identical. However, as we demonstrate in Section 5.9, a significant advantage is gained by allowing this general form (an idea which was suggested in [54]). To this end, we briefly introduce two other subtraction bases. The first is

$$q_r^{[i]}(x) = T_{2(r+1)-i}(x), \quad i \in \{0,1\}, \quad r \in \mathbb{N}_0, \tag{5.9}$$

where T_m is the m^{th} Chebyshev polynomial.[5] The second consists of Laplace–Dirichlet eigenfunctions,

$$q_r^{[0]}(x) = \cos(r + \tfrac{1}{2})\pi x, \quad q_r^{[1]}(x) = \sin(r+1)\pi x, \quad r = 0, \ldots, k-1. \tag{5.10}$$

It is readily verified that both sets of functions form subtraction bases:

Lemma 5.3. *Suppose that the functions $q_r^{[i]}$ are given by (5.9) or (5.10). Then the interpolation problem (5.8) has a unique solution for all choices $b_r^{[i]} \in \mathbb{R}$.*

Proof. Suppose that the functions $q_r^{[i]}$ are given by (5.10). Then

$$\mathcal{A}_r^{[i]}\left[q_s^{[i]}\right] = 2(-1)^{s+1}\left(s + \tfrac{i+1}{2}\right)^{2r+1} \pi^{2r+1}.$$

The $k \times k$ matrix with $(r,s)^{\text{th}}$ entry $\mathcal{A}_r^{[i]}[q_s^{[i]}]$ is of the form $(V^{[i]})^\top D^{[i]}$, where $V^{[i]}$ is a $k \times k$ Vandermonde matrix with entries $(r + \frac{i+1}{2})^{2s}\pi^{2s}$ and $D^{[i]}$ is a $k \times k$ diagonal matrix with entries $2(-1)^{s+1}(s + \frac{i+1}{2})\pi$. Hence, the result for (5.10) now follows immediately.

For the Chebyshev polynomials (5.9), existence and uniqueness is a direct consequence of the fact that the related Birkhoff–Hermite interpolation problem is uniquely solvable with a polynomial of degree $2k$. □

We scrutinise the bases (5.9) and (5.10) in greater detail in Section 5.9. As we demonstrate numerically, the use of Laplace–Dirichlet eigenfunctions yields greatly superior numerical results over approximations based on subtraction bases derived from polynomials. Hence, the duality enjoyed by the Laplace–Dirichlet and Laplace–Neumann bases (in the sense of Lemmas 2.4 and 2.5) is not only of theoretical interest, it also has practical consequences.

[4] In general, the values $\tilde{\mathcal{A}}_r^{[i]}[f]$ will also depend on k, unlike the jump values $\mathcal{A}_r^{[i]}[f]$. However, this presents few analytical or computational difficulties. For this reason, we do make this dependence explicit.

[5] We could also use Legendre polynomials with very similar results.

5.3 Eckhoff's method for univariate expansions

There are three well-documented problems with the polynomial subtraction technique, which we now describe. First, the method requires exact jump values. Ordinarily, such values are unknown. In many applications, only the modified Fourier coefficients of a given function may be specified. Moreover, even if arbitrary pointwise values of the function can be calculated, approximation via finite differences is not recommended for this purpose [118]. For the particular application considered in this thesis, the spectral approximation of boundary value problems, modified Fourier coefficients can be calculated, but both derivatives and pointwise values are not explicitly available. To address the convergence acceleration of the modified Fourier–Galerkin method introduced in Chapter 4 (a task we discuss briefly in Chapter 6), we must first develop a more sophisticated polynomial subtraction technique for the related task of function approximation.

The second drawback of polynomial subtraction, as we demonstrate in Section 5.4, relates to higher dimensions. In addition to pointwise jump values, various partial derivatives need to be known over $(d-1)$-dimensional subsets of the boundary. In practice, these may be approximated by lower dimensional techniques. However, to do so requires exact knowledge of $\mathcal{O}\left((k+N)^d\right)$ particular values (modified Fourier coefficients of derivatives evaluated on the boundary), where N is the truncated parameter used. Thus, the situation is even worse in higher dimensions, making the need to develop techniques to approximate such values even more pressing.

A final handicap of the polynomial subtraction device is that, for practical purposes, empirical evidence suggests that the parameter k must remain small [62]. Larger k (i.e. higher derivatives) often leads to a loss of accuracy, even if exact jump values are used. This restricts the potential convergence rate of the approximation, thus limiting its applicability.

As noted in [54], the previous lack of robust methods for the approximation of jump values is the central reason why the polynomial subtraction technique has not been more extensively utilised (see also [118, p.101] and [62]). In this chapter, to circumvent these aforementioned problems, we adapt Eckhoff's method to this task [52, 53, 54]. This approach is based on the observation that the modified Fourier coefficients themselves contain sufficient information to reconstruct the jump values. Hence, such values can be approximated to sufficient accuracy using only coefficients and suitably constructed extrapolation techniques.

After introducing the univariate version of Eckhoff's method for modified Fourier expansions, we next establish an extension to the d-variate cube (Sections 5.4—5.8). In Section 5.9 we address numerical issues—the third drawback mentioned previously—and, as a result, demonstrate how to obtain both high accuracy and improved numerical stability.

5.3.1 Eckhoff's method for the approximation of jump values

Before introducing a technique to approximate the exact jump values (5.2), we must first justify why inexact values can be used without deteriorating the convergence rate of the approximation $\mathcal{F}_{N,k}[f]$. We do this as follows. First, recalling that $\mathcal{F}_N[f]$[6] converges uniformly

[6]For convenience, throughout this chapter, replace N by $N-1$ in the definition of $\mathcal{F}_N[f]$ as given in (2.4).

to f, it follows from the expansion (5.7) that

$$f(x) - \mathcal{F}_N[f](x) = \sum_{i=0}^{1} \sum_{r=0}^{k-1} \mathcal{A}_r^{[i]}[f] \left(p_r^{[i]}(x) - \mathcal{F}_N[p_r^{[i]}](x) \right) + \mathcal{O}\left(N^{-2k-1}\right), \quad x \in [-1,1].$$

Now, suppose that the values $\mathcal{A}_r^{[i]}[f]$ are approximated by values $\bar{\mathcal{A}}_r^{[i]}[f]$ and that g_k is constructed as in (5.6) using these approximate values. Then, it follows from (5.1) and the above expression that

$$f(x) - \mathcal{F}_{N,k}[f](x) = \sum_{i=0}^{1} \sum_{r=0}^{k-1} \left(\mathcal{A}_r^{[i]}[f] - \bar{\mathcal{A}}_r^{[i]}[f] \right) \left(p_r^{[i]}(x) - \mathcal{F}_N[p_r^{[i]}](x) \right) + \mathcal{O}\left(N^{-2k-1}\right).$$

Consider, for example, the uniform error. Since $\|p_r^{[i]} - \mathcal{F}_N[p_r^{[i]}]\|_\infty = \mathcal{O}\left(N^{-2r-1}\right)$, to obtain an $\mathcal{O}\left(N^{-2k-1}\right)$ uniform error with the values $\bar{\mathcal{A}}_r^{[i]}[f]$, we require that

$$\bar{\mathcal{A}}_r^{[i]}[f] = \mathcal{A}_r^{[i]}[f] + \mathcal{O}\left(N^{2(r-k)}\right), \quad r = 0, \ldots, k-1, \quad i \in \{0,1\}. \tag{5.11}$$

In other words, rather than using exact jump values, it suffices to employ sufficiently accurate approximations. To accomplish this prescribed accuracy, we adapt Eckhoff's method [52, 53, 54], as we now describe.

Eckhoff's method is based on (5.7). In essence, we seek values $\bar{\mathcal{A}}_r^{[i]}[f]$ that satisfy this relation approximately. To do so, suppose that $N \leq m(0) < \ldots < m(k-1) \leq aN$, $m(r) \in \mathbb{N}$ are given values and $a \geq 1$ is constant. We define $\bar{\mathcal{A}}_r^{[i]}[f]$ as the solution of the $2k \times 2k$ linear system

$$\sum_{s=0}^{k-1} \widehat{p}_{s\,m(r)}^{[i]} \bar{\mathcal{A}}_s^{[i]}[f] = \hat{f}_{m(r)}^{[i]}, \quad r = 0, \ldots, k-1, \quad i \in \{0,1\}. \tag{5.12}$$

From a practical standpoint, this linear system decouples into two $k \times k$ linear systems corresponding to $i = 0$ and $i = 1$, which can be solved in parallel. Henceforth, we write $V^{[i]}$ for the $k \times k$ matrix with $(r,s)^{\text{th}}$ entry $\widehat{p}_{s\,m(r)}^{[i]}$. Note that the choice of the values $m(r)$ is essentially arbitrary. However, particular choices lead to better numerical stability and a so-called *auto-correction phenomenon* [138], an issue that we address in Section 5.7.

Nonsingularity of the linear system (5.12) can be immediately guaranteed:

Lemma 5.4. *For sufficiently large N, the linear system (5.12) is nonsingular. Moreover, if $p_0^{[i]}, \ldots, p_{k-1}^{[i]}$ are cardinal polynomials or arise from the subtraction basis (5.10), then (5.12) is nonsingular for all N.*

Proof. Suppose first that $P_0^{[i]}, \ldots, P_{k-1}^{[i]}$ (for the sake of clarity, we use this notation) are cardinal polynomials. Then, since $\widehat{P_s}_{m(r)}^{[i]} = (-1)^{m(r)+i} (\mu_{m(r)}^{[i]})^{-s-1}$, $V^{[i]} = D^{[i]} \tilde{V}^{[i]}$, where $D^{[i]}$ is the diagonal matrix with entries $(-1)^{m(r)} (\mu_{m(r)}^{[i]})^{-1}$ and $\tilde{V}^{[i]}$ is the Vandermonde matrix with entries $(\mu_{m(r)}^{[i]})^{-s}$. Nonsingularity (for all N) now follows immediately.

Suppose now that $p_0^{[i]}, \ldots, p_{k-1}^{[i]}$ are arbitrary cardinal functions. Then, since $p_r^{[i]} = P_r^{[i]} + (p_r^{[i]} - P_r^{[i]})$, we may write $V^{[i]} = W^{[i]} + (V^{[i]} - W^{[i]})$, where $W^{[i]}$ is the matrix with $(r,s)^{\text{th}}$

entry $\widehat{P}^{[i]}_{sm(r)}$. To prove the result, it suffices to show that $\|(W^{[i]})^{-1}(V^{[i]} - W^{[i]})\| = o(1)$ for large N, where $\|\cdot\|$ is any matrix norm.

Note that the s^{th} column of $V^{[i]} - W^{[i]}$ has entries $\widehat{(p_s - P_s)}^{[i]}_{m(r)}$. Moreover, $p_s - P_s$ obeys the first k derivative conditions. Hence, it can be shown that $(W^{[i]})^{-1}$ applied to this vector, which is just the vector of Eckhoff's approximation to the jump values of the function $p_s - P_s$, is $o(1)$ (see Theorem 5.5). Using this, we deduce the result.

Suppose now that the functions $q^{[i]}_r$ are given by (5.10). Then, due to (5.8) and Lemma 5.3, it suffices to prove nonsingularity of the matrix with $(r,s)^{\text{th}}$ entries

$$\widehat{q}^{[i]}_{sm(r)} = \frac{2(-1)^{m(r)+s+1}(s+\frac{1-i}{2})}{\left[(m(r)-\frac{i}{2})^2 - (s+\frac{1-i}{2})^2\right]\pi}.$$

After appropriate multiplication by a nonsingular diagonal matrix, we obtain the matrix with entries

$$\left[(m(r)-\tfrac{i}{2})^2 - (s+\tfrac{1-i}{2})^2\right]^{-1}.$$

This is a Cauchy matrix: hence, nonsingularity follows immediately. □

The fundamental aspect of Eckhoff's method, the linear system (5.12), is rather familiar. A similar idea is used in the Richardson extrapolation process [148]. However, the key difference herein is that we seek not just the limiting value of the sequence $\hat{f}^{[i]}_n$, but also the first k terms of its asymptotic expansion. Despite misgivings (see [148, p.33], where the instability of such a process is discussed), this can be done in a reasonably robust manner, as we demonstrate in Section 5.9.

With values $\tilde{\mathcal{A}}^{[i]}_r[f]$ given as the solutions of (5.12), we refer to the resulting approximation $\mathcal{F}_{N,k}[f] = \mathcal{F}_N[f - g_k] + g_k$ as the k^{th} *Eckhoff approximation* of f. In the forthcoming section, we study the convergence rate of this approximation. As we discuss further in Section 5.9, linear systems involving the matrices $V^{[i]}$ can be solved in $\mathcal{O}\left(k^2\right)$ operations. The overall cost of forming Eckhoff's approximation is therefore $\mathcal{O}\left(\max\{k^2, kN\}\right)$. Typically, $k \ll N$, so this figure reduces to $\mathcal{O}\left(kN\right)$.

Standard implementations of Eckhoff's method employ cardinal polynomials [16, 54]. For previously described reasons, we have presented Eckhoff's method in a more general form involving arbitrary cardinal functions. However, though the cardinal function formulation is the simplest version to consider in analysis, for computational purposes, it is often more convenient to present the method in terms of the subtraction basis $q^{[i]}_r$. In this case

$$g_k(x) = \sum_{i=0}^{1}\sum_{r=0}^{k-1} \tilde{\mathcal{A}}^{[i]}_r[f] q^{[i]}_r(x), \quad x \in [-1,1],$$

and the values $\tilde{\mathcal{A}}^{[i]}_r[f]$ are specified by the linear system

$$\sum_{s=0}^{k-1} \widehat{q}^{[i]}_{sm(r)} \tilde{\mathcal{A}}^{[i]}_s[f] = \hat{f}^{[i]}_{m(r)}, \quad r = 0,\ldots,k-1, \quad i \in \{0,1\}. \tag{5.13}$$

The resulting approximation is identical to the cardinal function formulation, but typically exhibits improved numerical behaviour (see Section 5.9).

Throughout this chapter, we study the continuous version of Eckhoff's method (5.12) based on modified Fourier coefficients. An analogous version can also be developed for discrete modified Fourier data. In the Fourier setting, this has been studied in [54, 129].

5.3.2 Convergence rate of Eckhoff's approximation

Eckhoff's method was originally presented for univariate Fourier series in [52, 53, 54]. Analysis of convergence was carried out in [16]. The key result demonstrates that the values $\bar{\mathcal{A}}_r^{[i]}[f]$ approximate the true values $\mathcal{A}_r^{[i]}[f]$ to the accuracy prescribed in (5.11). For the modified Fourier case, we have:

Theorem 5.5. *Suppose that $m(r) = c(r)N + \mathcal{O}(1)$, where $c(r) \geq 1$ and that at most $l \leq k$ of the $c(r)$ are equal. Suppose further that $f \in \mathrm{H}^{2k+l+1}(-1,1)$. Then the coefficients $\bar{\mathcal{A}}_r^{[i]}[f]$ obtained by Eckhoff's method satisfy (5.11).*

A proof of this result was originally given in [16]. Adaption to the modified Fourier case of the techniques used therein presents few conceptual challenges. Nonetheless, since there are several key differences, we now present the salient aspects of the proof for modified Fourier expansions.

Proof of Theorem 5.5. For the sake of brevity, we assume that cardinal polynomials are used (extension to arbitrary subtraction bases is simple). In this case, as in the proof of Lemma 5.4, we write $V^{[i]} = D^{[i]}\tilde{V}^{[i]}$, where $\tilde{V}_{r,s}^{[i]} = x_r^s$ and $x_r = (\mu_{m(r)}^{[i]})^{-1}$.

Upon replacing the coefficient in the right-hand side of (5.12) by its asymptotic expansion (5.7) and rearranging, we obtain the linear system of equations

$$\sum_{r=0}^{k-1} \tilde{V}_{r,s}^{[i]} \left(\mathcal{A}_s^{[i]}[f] - \bar{\mathcal{A}}_s^{[i]}[f] \right) = (-1)^{m(r)+i+k} x_r^{k-1} \widehat{f^{(2k)}}_{m(r)}^{[i]}, \quad r = 0, \ldots, k-1.$$

We now expand the right-hand side once more a total of K times, where $2K \leq l+1$, to give

$$\sum_{s=0}^{k-1} \tilde{V}_{r,s}^{[i]} \left(\mathcal{A}_s^{[i]}[f] - \bar{\mathcal{A}}_s^{[i]}[f] \right) = \sum_{s=k}^{k+K-1} x_r^s \mathcal{A}_s^{[i]}[f] + (-1)^{m(r)+i+k+K} x_r^{k+K-1} \widehat{f^{(2(k+K))}}_{m(r)}^{[i]}. \quad (5.14)$$

The entries of the inverse of a Vandermonde matrix can be exactly prescribed. In fact, the $(r,s)^{\text{th}}$ entry of $(\tilde{V}^{[i]})^{-1}$ is precisely

$$- x_s^{-(r+1)} \prod_{\substack{j=0 \\ j \neq s}}^{k-1} (x_s - x_j)^{-1} \sum_{j=0}^{r} \gamma_j x_s^j, \quad (5.15)$$

where the values $\gamma_0, \ldots, \gamma_k$ are symmetric polynomials of $x_0, \ldots, x_{k-1}$, defined by the relation $\sum_{r=0}^{k} \gamma_r x^r = \prod_{r=0}^{k-1}(x - x_r)$ [16]. For future use, we note that $\gamma_r = \mathcal{O}\left(N^{2(r-k)}\right)$. Suppose now that we define the parameter

$$\omega_r = - \sum_{s=0}^{k-1} x_s^r \prod_{\substack{j=0 \\ j \neq s}}^{k-1} (x_s - x_j)^{-1}.$$

Then, upon inverting the linear system (5.14) using (5.15) and substituting ω_r, we obtain

$$\mathcal{A}_r^{[i]}[f] - \bar{\mathcal{A}}_r^{[i]}[f] = \sum_{s=k}^{k+K-1} \sum_{j=0}^{r} \gamma_j \omega_{s+j-r-1} + (-1)^{i+k+K} \sum_{s=0}^{k-1} (-1)^{m(s)} \widehat{f^{(2(k+K))}}_{m(s)}^{[i]} x_s^{k+K-r-2} \prod_{\substack{j=0 \\ j\neq s}}^{k-1} (x_s - x_j)^{-1} \sum_{j=0}^{r} \gamma_j x_s^j. \tag{5.16}$$

We estimate the two terms of (5.16) separately. For the first, we recall from [16] that $\omega_r = \mathcal{O}\left(N^{2(k-r-1)}\right)$ for all $r \in \mathbb{N}_0$. Hence

$$\sum_{s=k}^{k+K-1} \sum_{j=0}^{r} \gamma_j \omega_{s+j-r-1} = \mathcal{O}\left(\sum_{s=k}^{k+K-1} \sum_{j=0}^{r} N^{2(j-k)} N^{2(k-s-j+r)} \right) = \mathcal{O}\left(N^{2(r-k)}\right),$$

as required. Now consider the second term of (5.16). If $m(r) = c(r)N + \mathcal{O}(1)$, then $\mu_{m(r)}^{[i]} = c(r)^2(N - \frac{i}{2})^2\pi^2 + \mathcal{O}(1)$. Since $x_r = (\mu_{m(r)}^{[i]})^{-1}$, simple arguments demonstrate that

$$\prod_{\substack{j=0 \\ j\neq s}}^{k-1} (x_s - x_j)^{-1} = \mathcal{O}\left(N^{2k+l-3}\right),$$

where l is the number of equal values $c(r)$. Hence, the second term of (5.16) is of order

$$\widehat{f^{(2(k+K))}}_{m(s)}^{[i]} \sum_{s=0}^{k-1} N^{2(r+2-k-K)} N^{2k+l-3} \sum_{j=0}^{r} N^{2(j-k)} N^{-2j} = \mathcal{O}\left(\widehat{f^{(2(k+K))}}_{m(s)}^{[i]} N^{2(r-k)} N^{l+1-2K} \right).$$

If l is odd, $2K = l + 1$ and the result follows immediately. For even values $l = 2K$, we have $f^{2(k+K)} \in \mathrm{H}^1(-1,1)$, so the coefficient $\widehat{f^{(2(k+K))}}_{m(s)}^{[i]} = \mathcal{O}\left(N^{-1}\right)$. Hence the result is also obtained in this case. □

There are several key differences between this result for modified Fourier expansions and the corresponding Fourier case. Most notably, since modified Fourier sine and cosine coefficients both have asymptotic expansion in even powers of n^{-1}, we do not require different regularity for even and odd values of the parameter l. Moreover, Theorem 5.5 only establishes the estimate (5.11). As described in [16], the imposition of two additional degrees of smoothness would have allowed us to determine the exact coefficient of $N^{2(r-k)}$ in this equation. This facilitates the derivation of precise asymptotic estimates for the $\mathrm{L}^2(-1,1)$ norm error of Eckhoff's approximation, as studied in [16]. However, since our interest lies with minimal regularity, we shall not pursue this further.

With this in hand, we may now provide estimates for the rate of convergence:

Theorem 5.6. *Suppose that l and f are as in Theorem 5.5, and that $\mathcal{F}_{N,k}[f]$ is the k^{th} Eckhoff approximation of f. Then $\|f - \mathcal{F}_{N,k}[f]\|_r$ is $\mathcal{O}(N^{r-2k-\frac{3}{2}})$ for $r = 0, \ldots, 2k+1$.*

Proof. Suppose that we write $\mathcal{F}^e_{N,k}[f]$ and $\mathcal{F}_{N,k}[f]$ for the approximations based on the exact jump values $\mathcal{A}^{[i]}_r[f]$ and their approximations $\bar{\mathcal{A}}^{[i]}_r[f]$ respectively. In view of Theorem 5.2, it suffices to consider the difference $\mathcal{F}^e_{N,k}[f] - \mathcal{F}_{N,k}[f]$. We have

$$\|\mathcal{F}^e_{N,k}[f] - \mathcal{F}_{N,k}[f]\|_r \leq \sum_{i=0}^{1} \sum_{r=0}^{k-1} \left|\mathcal{A}^{[i]}_r[f] - \bar{\mathcal{A}}^{[i]}_r[f]\right| \left\|p^{[i]}_r - \mathcal{F}_N[p^{[i]}_r]\right\|_r. \tag{5.17}$$

Now suppose that an arbitrary function $h \in \mathrm{C}^\infty[-1,1]$ satisfies the first $s \in \mathbb{N}_0$ derivative conditions. We claim that

$$\left\|h - \mathcal{F}_N[h]\right\|_r = \mathcal{O}\left(N^{r-2s-\frac{3}{2}}\right), \quad \forall r \in \mathbb{N}_0. \tag{5.18}$$

When $r \leq 2s+1$, this result follows immediately from Lemma 2.25. Now suppose that $r > 2s+1$. Since $\|h - \mathcal{F}_N[h]\|_r \leq \|h\|_r + \|\mathcal{F}_N[h]\|_r$ it suffices to consider $\|\mathcal{F}_N[h]\|_r$. Moreover, recalling that $\hat{h}_n = \mathcal{O}\left(n^{-2s-2}\right)$, we obtain

$$\|\mathcal{F}_N[h]\|^2_r \leq \sum_{i=0}^{1} \sum_{n=0}^{N-1} (1+\mu^{[i]}_n)^r |\hat{h}^{[i]}_n|^2 \leq c \sum_{n=0}^{N-1} \bar{n}^{2r-4s-4} = \mathcal{O}\left(N^{2r-4s-3}\right),$$

which gives (5.18). Substituting (5.18) with $h = p^{[i]}_r$ into (5.17) and using Theorem 5.5 immediately completes the proof. □

Theorem 5.7. *Suppose that f and $\mathcal{F}_{N,k}[f]$ are as in Theorem 5.6. Then, the error $\|f^{(r)} - (\mathcal{F}_{N,k}[f])^{(r)}\|_\infty$ is $\mathcal{O}\left(N^{r-2k-1}\right)$ for $r = 0, \ldots, 2k$.*

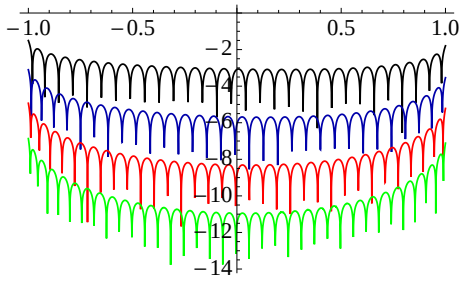
Proof. This follows immediately from Theorem 5.6 and the Sobolev interpolation inequality $\|h\|_\infty \leq c\sqrt{\|h\|\|h\|_1}$, $\forall h \in \mathrm{H}^1(-1,1)$, where $c = \sqrt{\frac{5}{2}}$. □

In Figure 5.1, we demonstrate the improvement offered by Eckhoff's method over the original ($k = 0$) modified Fourier approximation. For both functions considered, using only $N = 15$ and $k = 3$ we obtain 8 digits of accuracy. In comparison, less than two digits are witnessed for the original modified Fourier approximation. We mention in passing that, for the numerical examples presented in this and all sections up to Section 5.9, we use additional precision where necessary. The practical implementation of Eckhoff's method requires further study to lessen the impact of numerical instability, a topic we consider further in Section 5.9.

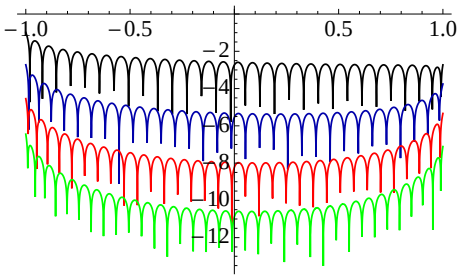
Theorems 5.6 and 5.7, in comparison to Theorems 5.1 and 5.2, demonstrate that Eckhoff's method leads to no deterioration in the convergence rate of the approximation over polynomial subtraction. In [16], the authors also compare the size of the error constants in $\|f - \mathcal{F}^e_{N,k}[f]\|$ and $\|f - \mathcal{F}_{N,k}[f]\|$. They establish that approximating the jump values in this manner not only leads to the same convergence rate, but also does not increase the error constant unduly. For this reason, we address only the asymptotic order of convergence throughout the remainder of this chapter.

Another consequence of Theorems 5.5–5.7 is that, for certain choices of $m(r)$, Eckhoff's method requires additional smoothness to obtain the same convergence rate as the approximation based on the exact jump values. A remedy for this is to employ distinct values $c(r)$. Typical choices of such values are

$$m(r) = (r+1)N, \quad r = 0, \ldots, k-1, \tag{5.19}$$



$f(x) = x^2 \sin 5x + \cos 6x$



$f(x) = \mathrm{e}^{-x} \cos 6x$

Figure 5.1: Eckhoff's method based on $m(r) = N + r$ and cardinal polynomials $p_r^{[i]}$. Log error $\log_{10} |f(x) - \mathcal{F}_{15,k}[f](x)|$ for $-1 \le x \le 1$ and $k = 0, 1, 2, 3$ (in descending order).

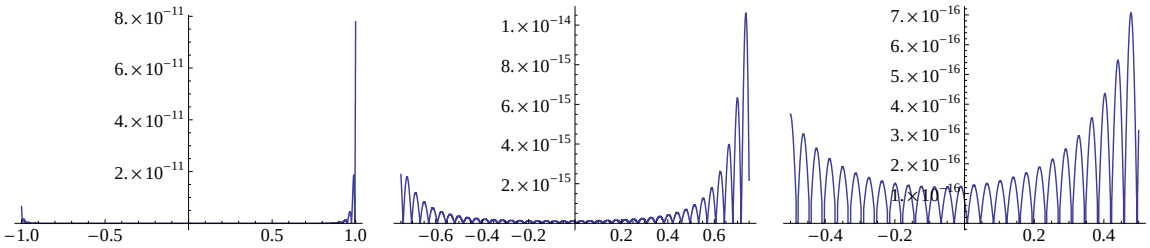


Figure 5.2: Graph of $|f(x) - \mathcal{F}_{25,4}[f](x)|$ for $-1 \le x \le 1$ (left), $-\frac{3}{4} \le x \le \frac{3}{4}$ (middle) and $-\frac{1}{2} \le x \le \frac{1}{2}$ (right), where $f(x) = \mathrm{Ai}(-3x - 4)$ and Ai is the Airy function [1].

or, given some arbitrary value $\omega = 2, 3, \ldots$,

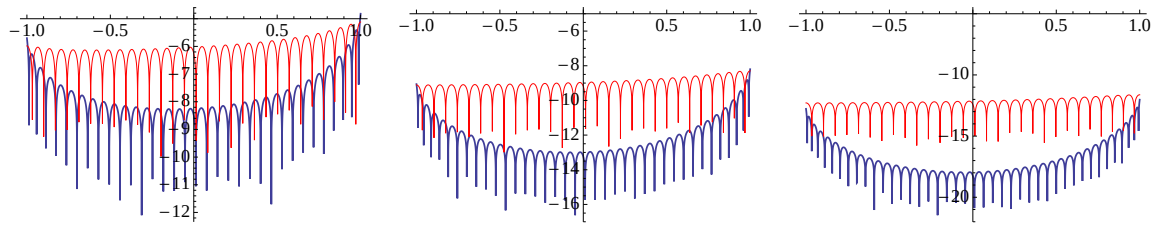
$$m(r) = \omega^r N, \quad r = 0, \ldots, k - 1. \tag{5.20}$$

However, Eckhoff's method, when based on equal values $c(r)$, and, in particular, $m(r) = N + r$, $r = 0, \ldots, k - 1$, conveys a significant advantage over polynomial subtraction, despite necessitating higher regularity. It turns out that, even though the uniform and $\mathrm{H}^r(\Omega)$ norm errors remaining of the same order, the error in compact subsets of Ω is much smaller. In fact, if the values $m(r) = N + r$ are chosen, the error is $\mathcal{O}\left(N^{-3k-2}\right)$ in comparison to $\mathcal{O}\left(N^{-2k-2}\right)$, a full factor of N^k smaller.

In Figure 5.2, this *auto-correction phenomenon* is demonstrated numerically. In this example, we use the values $m(r) = N + r$ and cardinal polynomials $p_r^{[i]}$. As we observe, the error away from the endpoints $x = \pm 1$ is much smaller. Figure 5.3 highlights the advantage offered by Eckhoff's method over polynomial subtraction in this respect; near the endpoints, both approximations offer similar error, whereas inside the interval, Eckhoff's method greatly outperforms the latter.

In reference to Eckhoff's method, the auto-correction phenomenon was observed numerically in [129] and demonstrated theoretically in the univariate, Fourier case in [138]. In Section 5.7, we extend this result to the multivariate modified Fourier setting.

This completes our discussion of the univariate version of Eckhoff's method. We devote the next part of this chapter to the significant generalisation of this method to functions defined



$k = 2$ $k = 4$ $k = 6$

Figure 5.3: Graph of $\log_{10}|f(x) - \mathcal{F}_{15,k}[f](x)|$ for $-1 \le x \le 1$ and $k = 2, 4, 6$, where $\mathcal{F}_{N,k}[f]$ is either the k^{th} polynomial subtraction (thin line) or Eckhoff (thick line) approximation and $f(x) = \text{Ai}(-3x - 4)$.

in the d-variate cube. As in the univariate setting, our first consideration is the polynomial subtraction technique and its multivariate extension.

5.4 Multivariate polynomial subtraction

To accelerate the convergence rate of multivariate modified Fourier expansions, it suffices to interpolate the first k odd partial derivatives of the function f on the whole of the boundary Γ (see Chapter 2). If this is achieved with a function g_k, then the k^{th} polynomial subtraction approximation $\mathcal{F}_{N,k}[f] = \mathcal{F}_N[f - g_k] + g_k$ will converge to f at a faster rate. For ease of reference, we now restate the result of Chapter 2 regarding the convergence rate of this approximation. For the moment, we assume that a full index set (2.33) is employed (we consider hyperbolic cross approximations in Section 5.8).

Theorem 5.8. *Suppose that $f \in \text{H}^{2k+2}_{mix}(\Omega)$ and that $\mathcal{F}_{N,k}[f]$ is the k^{th} polynomial subtraction approximation of f. Then $\|f - \mathcal{F}_{N,k}[f]\|_r$ is $\mathcal{O}(N^{r-2k-\frac{3}{2}})$ for $r = 0, \ldots, 2k+1$ and $\|\text{D}^\beta(f - \mathcal{F}_{N,k}[f])\|_\infty$ is $\mathcal{O}\left(N^{|\beta|_\infty - 2k - 1}\right)$ for $|\beta|_\infty \le 2k$. If, additionally, $f \in \text{H}^{2k+3}_{mix}(\Omega)$, then $\text{D}^\beta\{f(x) - \mathcal{F}_{N,k}[f](x)\}$ is $\mathcal{O}\left(N^{|\beta|_\infty - 2k - 2}\right)$ uniformly in compact subsets of Ω for $|\beta|_\infty \le 2k+1$.*

In the univariate setting, as demonstrated in Section 5.2.1, it is simple to construct the function g_k. With a little effort, this can also be accomplished in higher dimensions. Relevant techniques originate in the field of computer-aided geometric design and are related to so-called *Coons patches* [57]. In the context of modified Fourier expansions, this concept was first studied in [87].[7] A general construction was given in [4], and we shall closely follow this approach.

To this end, suppose that $p_r^{[i]}$, $r \in \mathbb{N}_0$, are the univariate cardinal functions of Section 5.2.1. Given $t \in [d]$, $i_t \in \{0,1\}^{|t|}$ and $r_t \in \mathbb{N}_0^{|t|}$, we define $p_{r_t}^{[i_t]}$ as the composition

$$p_{r_t}^{[i_t]}(x_t) = \prod_{j \in t} p_{r_j}^{[i_j]}(x_j).$$

[7] Bivariate polynomial subtraction for Fourier series has been the subject of a number of studies, including [17]. However, to the best of our knowledge, the extension to arbitrary $d \ge 2$ has not been studied rigorously. Once more, we mention that the work of this chapter is easily adapted to Fourier series with only minor modifications.

We also recall the definition (2.25) of the operator $\mathcal{B}_{r_t}^{[i_t]}[f] = \mathcal{B}_{r_{t_1}}^{[i_{t_1}]}[\ldots[\mathcal{B}_{r_{t_{|t|}}}^{[i_{t_{|t|}}]}[f]]\ldots]$, where

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

for $i_j \in \{0,1\}$, $r_j \in \mathbb{N}_0$ and $j = 1,\ldots,d$.

A suitable function g_k can now be immediately specified:

Lemma 5.9. *Suppose that $f \in \mathrm{H}^{2k}(\Omega)$ and that*

$$g_k(x) = \sum_{t\in[d]}\sum_{i_t\in\{0,1\}^{|t|}}\sum_{|r_t|_\infty=0}^{k-1}(-1)^{|t|+1}\mathcal{B}_{r_t}^{[i_t]}[f](x_{\bar{t}})p_{r_t}^{[i_t]}(x_t), \quad x\in\bar{\Omega}, \tag{5.21}$$

Then $f - g_k$ satisfies the first k Neumann derivative conditions (2.12). Equivalently,

$$\mathcal{B}_{r_j}^{[i_j]}[g_k] = \mathcal{B}_{r_j}^{[i_j]}[f], \quad r_j = 0,\ldots,k-1, \quad j = 1,\ldots,d. \tag{5.22}$$

Proof. It suffices to prove that g_k satisfies (5.22) with $j = 1$, $i_j = 0$ and $r_j = s$. We split the terms of (5.21) corresponding to different $t \in [d]$ into the three following cases: (i) $t = (1)$, (ii) $t = (1,u)$, where $u \in [d]$, $1 \notin u$, and (iii) $t = u$.

Consider case (i). The contribution of the corresponding term to $\mathcal{B}_s^{[0]}[g_k]$ is

$$\sum_{i_1=0}^{1}\sum_{r_1=0}^{k-1}\mathcal{B}_s^{[0]}\left[\mathcal{B}_{r_1}^{[i_1]}[f](x_2,\ldots,x_d)p_{r_1}^{[i_1]}(x_1)\right](x_2,\ldots,x_d) = \mathcal{B}_s^{[0]}[f](x_2,\ldots,x_d),$$

where equality follows directly from the properties of the cardinal functions $p_r^{[i]}$. It now suffices to prove that the contributions of cases (ii) and (iii) cancel. For case (ii), this is

$$\begin{aligned}&\sum_{i_u\in\{0,1\}^{|u|}}\sum_{|r_u|_\infty=0}^{k-1}\sum_{i_1=0}^{1}\sum_{r_1=0}^{k-1}(-1)^{|u|}\mathcal{B}_s^{[0]}\left[\mathcal{B}_{r_t}^{[i_t]}[f](x_{\bar{t}})p_{r_t}^{[i_t]}(x_t)\right](x_2,\ldots,x_d)\\ &\qquad=\sum_{i_u\in\{0,1\}^{|u|}}\sum_{|r_u|_\infty=0}^{k-1}(-1)^{|u|}\mathcal{B}_{(s,r_u)}^{[(0,i_u)]}[f](x_{\bar{u}})p_{r_u}^{[i_u]}(x_u),\end{aligned}$$

where $(0,i_u) = (0,i_{u_1},\ldots,i_{u_{|u|}})$ and $(s,r_u) = (s,r_{u_1},\ldots,r_{u_{|u|}})$. It is readily seen that this term is precisely the negative of the contribution of case (iii). □

As in Chapter 2, the bivariate case will serve as our primary example. In this setting, the function g_k is given explicitly by

$$\begin{aligned}g_k(x) =& \sum_{i_1=0}^{1}\sum_{r_1=0}^{k-1}p_{r_1}^{[i_1]}(x_1)\mathcal{B}_{r_1}^{[i_1]}[f](x_2) + \sum_{i_2=0}^{1}\sum_{r_2=0}^{k-1}\mathcal{B}_{r_2}^{[i_2]}[f](x_1)p_{r_2}^{[i_2]}(x_2)\\ &-\sum_{i_1,i_2=0}^{1}\sum_{r_1,r_2=0}^{k-1}\mathcal{B}_{r_1}^{[i_1]}\left[\mathcal{B}_{r_2}^{[i_2]}[f]\right]p_{r_1}^{[i_1]}(x_1)p_{r_2}^{[i_2]}(x_2).\end{aligned} \tag{5.23}$$

We remark in passing that the phrase 'polynomial subtraction' is a misnomer in the multivariate case: as evidenced by (5.23), the function g_k is no longer a polynomial for $d \geq 2$. Herein lies the main problem with this device. Computation of the function g_k, as given by (5.21), requires knowledge of the exact derivatives of the function f over $(d-1)$-dimensional subsets of the boundary. Classically, this has been considered, somewhat erroneously, as a limitation of this technique [62]. The intention of this chapter is to show that this is not, in fact, the case: approximations of arbitrary order can be constructed using only multivariate modified Fourier coefficients.

An obvious means to alleviate this problem (which we now introduce since it will be used in the sequel) is to approximate such lower dimensional functions using polynomial subtraction (an approach mentioned briefly, but not analysed, in [87]). Doing so requires knowledge of functions over $(d-2)$-dimensional subsets of the boundary. However, we may repeat the same process, replacing exact functions by polynomial subtraction approximations, until we obtain an approximation that uses only derivative values over the 0-dimensional subsets of the boundary consisting of the vertices $(\pm 1, \pm 1, \ldots, \pm 1)$ and the modified Fourier coefficients of higher dimensional derivative functions.

To differentiate between the two approaches, we refer to the approximation based on (5.21) as *exact* polynomial subtraction and the approximation obtained by the above process as *approximate* polynomial subtraction. We write g_k^e, $\mathcal{F}_{N,k}^e[f]$ and g_k^a, $\mathcal{F}_{N,k}^a[f]$ respectively. Note that for $d = 1$ both approximations coincide.

In the bivariate setting, we merely replace the univariate functions $\mathcal{B}_{r_1}^{[i_1]}[f]$ and $\mathcal{B}_{r_2}^{[i_2]}[f]$ by their k^{th} polynomial subtraction approximations. This yields the new function g_k^a given by

$$\begin{aligned} g_k^a(x) = &\sum_{i_1=0}^{1} \sum_{r_1=0}^{k-1} p_{r_1}^{[i_1]}(x_1) \mathcal{F}_{N,k}\left[\mathcal{B}_{r_1}^{[i_1]}[f]\right](x_2) + \sum_{i_2=0}^{1} \sum_{r_2=0}^{k-1} \mathcal{F}_{N,k}\left[\mathcal{B}_{r_2}^{[i_2]}[f]\right](x_1) p_{r_2}^{[i_2]}(x_2) \\ &- \sum_{i_1,i_2=0}^{1} \sum_{r_1,r_2=0}^{k-1} \mathcal{B}_{r_1}^{[i_1]}\left[\mathcal{B}_{r_2}^{[i_2]}[f]\right] p_{r_1}^{[i_1]}(x_1) p_{r_2}^{[i_2]}(x_2). \end{aligned}$$

For $d \geq 3$, we define the new approximation inductively. If $\mathcal{F}_{N,k}^a[\cdot]$ is known for $d-1$, we define the d-variate approximate polynomial subtraction function g_k^a by

$$g_k^a(x) = \sum_{t \in [d]} \sum_{i_t \in \{0,1\}^{|t|}} \sum_{|r_t|_\infty = 0}^{k-1} (-1)^{|t|+1} \mathcal{F}_{N,k}^a\left[\mathcal{B}_{r_t}^{[i_t]}[f]\right](x_{\bar{t}}) p_{r_t}^{[i_t]}(x_t), \quad x \in \bar{\Omega}. \tag{5.24}$$

In its present form, this subtraction function is not fit for practical purposes. Instead, we seek a version of g_k^a that is not inductively defined. The derivation of such a form indicates the appropriate generalisation of Eckhoff's method to the multivariate case, as we subsequently consider.

To obtain the function g_k^a explicitly, it is first necessary to reintroduce the values $\mathcal{A}_{r_t,n_{\bar{t}}}^{[i]}[f]$, which occur in the asymptotic expansion of the coefficients $\hat{f}_n^{[i]}$ (see Section 2.7 and, in particular, (2.26)). Given such values, it will also be of use to notice that we may re-write the expansion (2.27) in terms of the functions $p_{r_t}^{[i_t]}$ (as in the univariate case (5.7)). If $\widehat{p_{r_t}}_{n_t}^{[i_t]}$ is the modified Fourier coefficient of $p_{r_t}^{[i_t]}$ corresponding to indices i_t and n_t, then

$$\hat{f}_n^{[i]} = \sum_{t \in [d]^*} \sum_{|r_t|_\infty = 0}^{k-1} \mathcal{A}_{r_t,n_{\bar{t}}}^{[i]}[f] \widehat{p_{r_t}}_{n_t}^{[i_t]} + \mathcal{O}\left(n^{-2k-2}\right).$$

Once more, the remainder term vanishes, provided $p_{r_t}^{[i_t]}$ consists of cardinal polynomials.

With this in hand, we may now give an explicit expression for g_k^a:

Lemma 5.10. *The approximate polynomial subtraction function g_k^a is given by*

$$g_k^a(x) = \sum_{i\in\{0,1\}^d} \sum_{t\in[d]} \sum_{|r_t|_\infty=0}^{k-1} \sum_{|n_{\bar{t}}|_\infty=0}^{N-1} \mathcal{A}_{r_t,n_{\bar{t}}}^{[i]}[f] p_{r_t}^{[i_t]}(x_t)\phi_{n_{\bar{t}}}^{[i_{\bar{t}}]}(x_{\bar{t}}). \tag{5.25}$$

To prove this lemma, we require the following notation. Given $t \in [d]$, we write $[t]$ for the set of tuples $u \in [d]$ with $u \subseteq t$ (in other words, if $j \in u$ then $j \in t$). We write $[t]^* = [t] \cup \{\emptyset\}$ and $\bar{u} \in [t]^*$ for the tuple of elements in t, but not in u. Further, given $t, u \in [d]^*$, we write $t \cup u \in [d]^*$ for the ordered tuple of elements $j = 1, \dots, d$ in t or in u, $t \cap u$ for the tuple of elements in both t and u and $t\backslash u$ for the tuple of elements in t, but not in u.

Proof of Lemma 5.10. We prove this result by induction on d. For $d = 1$, since $g_k^a = g_k^e$ and $\mathcal{A}_r^{[i]}[f] = \mathcal{B}_r^{[i]}[f]$, there is nothing to prove. Now assume that the result holds for $d-1$. Then, by definition

$$g_k^a(x) = \sum_{t\in[d]} \sum_{i_t\in\{0,1\}^{|t|}} \sum_{|r_t|_\infty=0}^{k-1} (-1)^{|t|+1} \mathcal{F}_{N,k}^a \left[\mathcal{B}_{r_t}^{[i_t]}[f]\right](x_{\bar{t}}) p_{r_t}^{[i_t]}(x_t). \tag{5.26}$$

Since $\mathcal{B}_{r_t}^{[i_t]}[f]$ is a function of at most $(d-1)$ variables, we may use the induction hypothesis to derive an expression for $\mathcal{F}_{N,k}^a \left[\mathcal{B}_{r_t}^{[i_t]}[f]\right](x_{\bar{t}})$. To do so, we require several observations. First, we note that

$$\mathcal{A}_{r_u,n_{\bar{u}}}^{[i_{\bar{t}}]} \left[\mathcal{B}_{r_t}^{[i_t]}[f]\right] = (-1)^{k|\bar{u}|} \prod_{j\in\bar{u}} (\mu_{n_j}^{[i_j]})^{-k} \int \mathcal{B}_{r_u}^{[i_u]} \left[\mathrm{D}_{\bar{u}}^{2k} \mathcal{B}_{r_t}^{[i_t]}[f]\right] \phi_{n_{\bar{u}}}^{[i_{\bar{u}}]}(x_{\bar{u}})\, \mathrm{d}x_{\bar{u}}, \quad \forall u \in [\bar{t}]^*.$$

Since $\bar{u} = \bar{t}\backslash u = \overline{t \cup u}$, and the operators $\mathcal{B}_{r_u}^{[i_u]}$ and $\mathcal{B}_{r_t}^{[i_t]}$ commute with each other and with differentiation in the independent variables, we have

$$\mathcal{A}_{r_u,n_{\bar{u}}}^{[i_{\bar{t}}]} \left[\mathcal{B}_{r_t}^{[i_t]}[f]\right] = (-1)^{k|\bar{u}|} \prod_{j\in\bar{u}} (\mu_{n_j}^{[i_j]})^{-k} \int \mathcal{B}_{r_{t\cup u}}^{[i_{t\cup u}]} \left[\mathrm{D}_{\bar{u}}^{2k} f\right] \phi_{n_{\bar{u}}}^{[i_{\bar{u}}]}(x_{\bar{u}})\, \mathrm{d}x_{\bar{u}} = \mathcal{A}_{r_{t\cup u},n_{\overline{t\cup u}}}^{[i]}[f].$$

Our next observation is as follows: if h is a function of at most $(d-1)$ variables, and g_k^a is the approximate polynomial subtraction function for h, then

$$\mathcal{F}_N[h - g_k^a](x) = \sum_{i\in\{0,1\}^{d-1}} \sum_{|n|_\infty=0}^{N-1} \mathcal{A}_n^{[i]}[h]\phi_n^{[i]}(x), \quad x \in [-1,1]^{d-1},$$

where $\mathcal{A}_n^{[i]}[h]$ is the value $\mathcal{A}_{r_t,n_{\bar{t}}}^{[i]}[h]$ defined in (2.26) corresponding to $t = \emptyset$. This follows immediately from the induction hypothesis and the expressions (2.27) and (5.25).

Returning to $\mathcal{B}_{r_t}^{[i_t]}[f]$, and using these observations, we obtain

$$\begin{aligned}\mathcal{F}_{N,k}^a \left[\mathcal{B}_{r_t}^{[i_t]}[f]\right](x_{\bar{t}}) = &\sum_{i_{\bar{t}}\in\{0,1\}^{|\bar{t}|}} \sum_{|n_{\bar{t}}|_\infty=0}^{N-1} \mathcal{A}_{r_t,n_{\bar{t}}}^{[i]}[f]\phi_{n_{\bar{t}}}^{[i_{\bar{t}}]}(x_{\bar{t}}) \\ &+ \sum_{i_{\bar{t}}\in\{0,1\}^{|\bar{t}|}} \sum_{u\in[\bar{t}]} \sum_{|r_u|_\infty=0}^{k-1} \sum_{|n_{\bar{u}}|_\infty=0}^{N-1} \mathcal{A}_{r_{t\cup u},n_{\overline{t\cup u}}}^{[i]}[f] p_{r_u}^{[i_u]}(x_u)\phi_{n_{\bar{u}}}^{[i_{\bar{u}}]}(x_{\bar{u}}).\end{aligned}$$

Substituting this into (5.26) gives

$$g_k^a(x) = \sum_{i\in\{0,1\}^d}\sum_{t\in[d]}(-1)^{|t|+1}\Bigg\{\sum_{|r_t|_\infty=0}^{k-1}\sum_{|n_{\bar{t}}|_\infty=0}^{N-1}\mathcal{A}^{[i]}_{r_t,n_{\bar{t}}}[f]p^{[i_t]}_{r_t}(x_t)\phi^{[i_{\bar{t}}]}_{n_{\bar{t}}}(x_{\bar{t}}) + \sum_{u\in[\bar{t}]}\sum_{|r_{t\cup u}|_\infty=0}^{k-1}\sum_{|n_{\bar{u}}|_\infty=0}^{N-1}\mathcal{A}^{[i]}_{r_{t\cup u},n_{\overline{t\cup u}}}[f]p^{[i_{t\cup u}]}_{r_{t\cup u}}(x_{t\cup u})\phi^{[i_{\bar{u}}]}_{n_{\bar{u}}}(x_{\bar{u}})\Bigg\}. \tag{5.27}$$

To complete the proof, it suffices to show that, for any $v\in[d]$, the coefficient of the corresponding term $\mathcal{A}^{[i]}_{r_v,n_{\bar{v}}}[f]p^{[i_v]}_{r_v}(x_v)\phi^{[i_{\bar{v}}]}_{n_{\bar{v}}}(x_{\bar{v}})$ in (5.27) is precisely 1. The first term of (5.27) gives a contribution of $(-1)^{|v|+1}$. For the second, the terms that give contributions are those with $t\cup u=v$. Since $t,u\neq\emptyset$, and there are $\binom{|v|}{l}$ possible choices of such u with $|u|=l$, this gives a total contribution of

$$(-1)^{|v|}\binom{|v|}{1}+\ldots+\binom{|v|}{|v|-1}=(-1)^{|v|+1}\sum_{l=1}^{|v|-1}\binom{|v|}{l}(-1)^l=1-(-1)^{|v|+1}.$$

Summing together this and the previous contribution now yields the result. □

The result of Lemma 5.10 not only gives an explicit way to compute the k^{th} approximate polynomial subtraction function, it also demonstrates that faster convergence can be achieved by suitable approximation of the values $\mathcal{A}^{[i]}_{r_t,n_{\bar{t}}}[f]$. We consider this task in the sequel. This is a far less daunting prospect than the exact polynomial subtraction standpoint, where functions of $(d-1)$ variables need to be approximated. For this reason, the values $\mathcal{A}^{[i]}_{r_t,n_{\bar{t}}}[f]$ can be viewed as the appropriate generalisation of the univariate jumps $\mathcal{A}^{[i]}_r[f]$.

For approximate polynomial subtraction to serve as a potential alternative to its exact counterpart (and since we shall require such estimates in the sequel), it remains to demonstrate that the convergence rate is not affected. We have

Theorem 5.11. *Suppose that $f\in \mathrm{H}^{2k+2}_{mix}(\Omega)$ and that $\mathcal{F}^a_{N,k}[f]$ is the k^{th} approximate polynomial subtraction approximation of f. Then $\|f-\mathcal{F}^a_{N,k}[f]\|_r$ is $\mathcal{O}(N^{r-2k-\frac{3}{2}})$ for $r=0,\ldots,2k+1$ and $\|\mathrm{D}^\beta(f-\mathcal{F}^a_{N,k}[f])\|_\infty$ is $\mathcal{O}\left(N^{|\beta|_\infty-2k-1}\right)$ for $|\beta|_\infty\leq 2k$. If, additionally, $f\in\mathrm{H}^{2k+3}_{mix}(\Omega)$, then $\mathrm{D}^\beta\{f(x)-\mathcal{F}^a_{N,k}[f](x)\}$ is $\mathcal{O}\left(N^{|\beta|_\infty-2k-2}\right)$ uniformly in compact subsets of Ω for $|\beta|_\infty\leq 2k+1$.*

Proof. By Theorem 5.8, it suffices to consider the difference $\mathcal{F}^e_{N,k}[f]-\mathcal{F}^a_{N,k}[f]$. We shall use induction on d. For $d=1$ there is nothing to prove. Now suppose that the result holds for $d-1$. We have

$$\mathcal{F}^e_{N,k}[f](x)-\mathcal{F}^a_{N,k}[f](x)=g^e_k(x)-g^a_k(x)-\mathcal{F}_N\left[g^e_k-g^a_k\right](x).$$

Since $\widehat{\mathcal{F}^a_{N,k}[h]}^{[i]}_n=\widehat{\mathcal{F}^e_{N,k}[h]}^{[i]}_n=\hat{h}^{[i]}_n$ for all $i\in\{0,1\}^d$, $n\in I_N$ and arbitrary functions h, it follows that $\mathcal{F}_N[g^e_k-g^a_k]=0$. Hence

$$\begin{aligned}\mathcal{F}^e_{N,k}[f](x)-\mathcal{F}^a_{N,k}[f](x)&=g^e_k(x)-g^a_k(x)\\&=\sum_{t\in[d]}\sum_{i_t\in\{0,1\}^{|t|}}\sum_{|r_t|_\infty=0}^{k-1}(-1)^{|t|+1}\left(\mathcal{B}^{[i_t]}_{r_t}[f](x_{\bar{t}})-\mathcal{F}^a_{N,k}\left[\mathcal{B}^{[i_t]}_{r_t}[f]\right](x_{\bar{t}})\right)p^{[i_t]}_{r_t}(x_t).\end{aligned}$$

For $f \in \mathrm{H}^{2k+2}_{\mathrm{mix}}(\Omega)$, we have $\mathcal{B}^{[i_t]}_{r_t}[f] \in \mathrm{H}^{2k+2}_{\mathrm{mix}}(-1,1)^{d-|t|}$ (see Section 2.5 or [5]). Since $|t| \geq 1$, we may use the induction hypothesis on each such term to obtain the result. □

In view of Theorem 5.8, we deduce that the various convergence rates remain the same. However, although the approximate polynomial subtraction process achieves a significant improvement over exact polynomial subtraction, it still requires explicit knowledge of the values $\mathcal{A}^{[i]}_{r_t,n_{\bar{t}}}[f]$. In general, these are unknown. Since there are

$$2^d \sum_{j=1}^{d} \binom{d}{j} k^j N^{d-j} = 2^d \left\{ (k+N)^d - N^d \right\} = \mathcal{O}\left(kN^{d-1} \right), \quad k \ll N,$$

such values in total, the need for a method of approximation becomes more vital as d increases. We shall achieve this by a suitable extension of Eckhoff's method, which we now introduce.

5.5 Eckhoff's method for multivariate expansions

We now extend Eckhoff's method to the multivariate setting. The bivariate version of this method was originally developed, without analysis, in [128, 130, 131]. In this section, we first establish an extension for general d, and in Section 5.6, we provide pertinent analysis.

As indicated by the form of approximate polynomial subtraction function g^a_k, we seek approximations $\bar{\mathcal{A}}^{[i]}_r[f]$ to the values $\mathcal{A}^{[i]}_{r_t,n_{\bar{t}}}[f]$. To this end, we define the subtraction function

$$g_k(x) = \sum_{i \in \{0,1\}^d} \sum_{t \in [d]} \sum_{|r_t|_\infty = 0}^{k-1} \sum_{|n_{\bar{t}}|_\infty = 0}^{N-1} \bar{\mathcal{A}}^{[i]}_{r_t,n_{\bar{t}}}[f] p^{[i_t]}_{r_t}(x_t) \phi^{[i_{\bar{t}}]}_{n_{\bar{t}}}(x_{\bar{t}}), \tag{5.28}$$

and the approximation $\mathcal{F}_{N,k}[f] = \mathcal{F}_N[f - g_k] + g_k$. In the univariate setting, it follows from (5.12) that the function g_k satisfies the condition

$$\widehat{g_k}^{[i]}_n = \hat{f}^{[i]}_n, \quad n = m(0), \ldots, m(k-1), \quad i \in \{0,1\}. \tag{5.29}$$

For the d-variate extension, we enforce a similar condition. Suppose that we define the finite index set $M_k \subseteq \mathbb{N}^d$ by

$$M_k = \bigcup_{t \in [d]} \left\{ n = (n_1, \ldots, n_d) \in \mathbb{N}^d : n_j = m(r_j),\ r_j = 0, \ldots, k-1,\ j \in t, |n_{\bar{t}}|_\infty < N \right\}. \tag{5.30}$$

We now impose the condition

$$\widehat{g_k}^{[i]}_n = \hat{f}^{[i]}_n, \quad \forall n \in M_k, \quad i \in \{0,1\}^d. \tag{5.31}$$

For $d = 1$, (5.31) reduces to (5.29). For $d = 2$, we obtain the following system of equations

$$\begin{aligned}
\widehat{g_k}^{[i]}_{m(r_1),m(r_2)} &= \widehat{f}^{[i]}_{m(r_1),m(r_2)}, \quad r_1, r_2 = 0, \ldots, k-1, \quad i \in \{0,1\}^2, \\
\widehat{g_k}^{[i]}_{m(r_1),n_2} &= \widehat{f}^{[i]}_{m(r_1),n_2}, \quad r_1 = 0, \ldots, k-1, \quad n_2 = 0, \ldots, N-1, \quad i \in \{0,1\}^2, \\
\widehat{g_k}^{[i]}_{n_1,m(r_2)} &= \widehat{f}^{[i]}_{n_1,m(r_2)}, \quad n_1 = 0, \ldots, N-1, \quad r_2 = 0, \ldots, k-1, \quad i \in \{0,1\}^2.
\end{aligned} \tag{5.32}$$

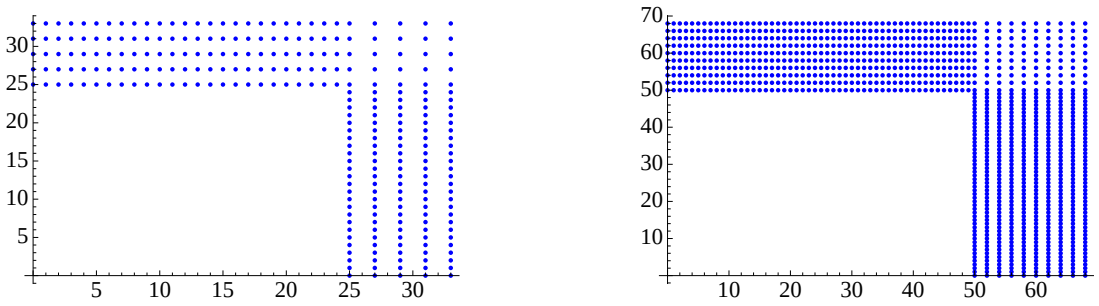


Figure 5.4: Left diagram: the index set M_5 with $N = 25$ and $m(r) = N + 2r$. Right diagram: the index set M_{10} with $N = 50$ and $m(r) = N + 2r$.

Figure 5.4 shows the typical form of the index set M_k for $d = 2$. Note that, as in the univariate case, the system of equations (5.31) completely decouples for different values of $i \in \{0,1\}^d$.

For both practical and analytical purposes, we need to expand the left-hand side of (5.31). Given $u \in [d]$, $s_u \in \{0,\dots,k-1\}^{|u|}$ and $n_{\bar{u}} \in \{0,\dots,N-1\}^{|\bar{u}|}$, the corresponding term of $g_k(x)$ is

$$\bar{\mathcal{A}}^{[i]}_{s_u,n_{\bar{u}}}[f]p^{[i_u]}_{s_u}(x_u)\phi^{[i_{\bar{u}}]}_{n_{\bar{u}}}(x_{\bar{u}}) = \bar{\mathcal{A}}^{[i]}_{r_u,n_{\bar{u}}}[f]\prod_{j\in u}p^{[i_j]}_{s_j}(x_j)\prod_{j\notin u}\phi^{[i_j]}_{n_j}(x_j).$$

This term gives a non-zero contribution to the left-hand side of (5.31) precisely when $t \subseteq u$, where $t \in [d]$ is the tuple associated to $n \in M_k$. Hence

$$\begin{aligned}\widehat{g_k}^{[i]}_n &= \sum_{\substack{u\in[d]\\ t\subseteq u}}\sum_{|s_u|_\infty=0}^{k-1}\bar{\mathcal{A}}^{[i]}_{s_u,n_{\bar{u}}}[f]\prod_{j\in u}\widehat{p_{s_j}}^{[i_j]}_{n_j}\\ &= \sum_{|s_t|_\infty=0}^{k-1}\prod_{j\in t}V^{[i_j]}_{r_j,s_j}\left\{\sum_{t\subseteq u}\sum_{|s_{u\setminus t}|_\infty=0}^{k-1}\bar{\mathcal{A}}^{[i]}_{s_u,n_{\bar{u}}}[f]\prod_{j\in u\setminus t}\widehat{p_{s_j}}^{[i_j]}_{n_j}\right\}.\end{aligned} \tag{5.33}$$

Here r_j, $j \in t$ is the index used in the definition (5.30) of $n \in M_k$, and $V^{[i]}$ is the matrix introduced in Section 5.3.1.

For $d = 2$, substituting (5.33) into (5.32) gives

$$\begin{aligned}&\sum_{s_1,s_2=0}^{k-1}V^{[i_1]}_{r_1,s_1}V^{[i_2]}_{r_2,s_2}\bar{\mathcal{A}}^{[i]}_{s_1,s_2}[f] = \hat{f}^{[i]}_{m(r_1),m(r_2)}, \quad r_1,r_2 = 0,\dots,k-1,\\
&\sum_{s_1=0}^{k-1}V^{[i_1]}_{r_1,s_1}\left\{\bar{\mathcal{A}}^{[i]}_{s_1,n_2}[f] + \sum_{s_2=0}^{k-1}\bar{\mathcal{A}}^{[i]}_{s_1,s_2}[f]\widehat{p_{r_2}}^{[i_2]}_{n_2}\right\} = \hat{f}^{[i]}_{m(r_1),n_2}, \quad r_1 = 0,\dots,k-1,\\
&\qquad\qquad n_2 = 0,\dots,N-1,\\
&\sum_{s_2=0}^{k-1}V^{[i_2]}_{r_2,s_2}\left\{\bar{\mathcal{A}}^{[i]}_{s_2,n_1}[f] + \sum_{s_1=0}^{k-1}\bar{\mathcal{A}}^{[i]}_{s_1,s_2}[f]\widehat{p_{r_1}}^{[i_1]}_{n_1}\right\} = \hat{f}^{[i]}_{n_1,m(r_2)}, \quad n_1 = 0,\dots,N-1,\\
&\qquad\qquad r_2 = 0,\dots,k-1.\end{aligned}$$

In this case, it is obvious how to solve these equations. We first obtain $\bar{\mathcal{A}}^{[i]}_{r_1,r_2}[f]$ from the first equation, and then substitute the result into the second and third equations, solving for $\bar{\mathcal{A}}^{[i]}_{r_1,n_2}[f]$ and $\bar{\mathcal{A}}^{[i]}_{r_2,n_1}[f]$ respectively. The same can be done in $d \geq 3$ dimensions. Starting from the equation corresponding to $t = (1, 2, \ldots, d)$, we find $\bar{\mathcal{A}}^{[i]}_{r_t}[f]$. Using this, we solve the equations corresponding to $|t| = d - 1$, $|t| = d - 2$, and so on. Continuing in this manner, we obtain all the coefficients $\bar{\mathcal{A}}^{[i]}_{r_t,n_{\bar{t}}}[f]$.

Alternatively, it may be easier to use the following explicit expression for such values:

$$\bar{\mathcal{A}}^{[i]}_{r_t,n_{\bar{t}}}[f] = \sum_{\substack{u \in [d] \\ t \subseteq u}} (-1)^{|t|+|u|} \sum_{|r_{u\setminus t}|_\infty = 0}^{k-1} \sum_{|s_u|_\infty = 0}^{k-1} \hat{f}^{[i]}_{(m(s_u);n_{\bar{u}})} \prod_{j \in u} (V^{[i_j]})^{-1}_{r_j,s_j} \prod_{j \in u\setminus t} \widehat{p_{r_j}}^{[i_j]}_{n_j}, \tag{5.34}$$

where $(m(s_u); n_{\bar{u}})$ has j^{th} entry $m(s_j)$ if $j \in u$ and n_j otherwise. We defer a proof of this formula until Section 5.8 (see Lemma 5.36).

Though straightforward in theory, it is fair to mention that the construction of Eckhoff's approximation becomes increasingly cumbersome to implement for large d. However, as we demonstrate in the sequel by numerical example, it is certainly practical for $d = 2, 3$.

Observe that, to find the coefficients $\bar{\mathcal{A}}^{[i]}_{r_t,n_{\bar{t}}}[f]$, we have to solve linear systems involving the matrix $V^{[i]}$. One immediate benefit of Eckhoff's approach is that the coefficients can be found by solving essentially one-dimensional linear systems. Since we need to solve many such systems, it may be easiest to find $(V^{[i]})^{-1}$ first and use (5.34). On a related topic, we note that the existence and uniqueness of a solution to the linear system (5.31) are completely determined by the nonsingularity of the matrix $V^{[i]}$ (see Lemma 5.4).

In the univariate setting, the operational cost of forming Eckhoff's approximation is $\mathcal{O}\left(\max\{k^2, kN\}\right)$. For the multivariate case, this value is $\mathcal{O}\left(\max\{k^{d+1}, k^d N^d\}\right)$. When $k \ll N$, this figure reduces to $k^d N^d$. In comparison, forming the approximation $\mathcal{F}_N[f]$ involves $\mathcal{O}\left(N^d\right)$ operations, so the increase in complexity is relatively mild for moderate values of k. Nonetheless, the value N^d grows exponentially with d. In Section 5.8, we assess how this figure can be dramatically reduced by the use of hyperbolic cross index sets.

Having obtained Eckhoff's approximation for arbitrary $d \geq 2$, we now turn our attention to the convergence analysis of this approximation.

5.6 Analysis of Eckhoff's method

Here, and for the remainder of this chapter, we use the notation $A \lesssim B$ to mean that there exists a constant c independent of N such that $A \leq cB$.

We commence our analysis of the multivariate version of Eckhoff's method with the following lemmas, the first of which is a generalisation of Theorem 5.5:

Lemma 5.12. *Suppose that $h \in \mathrm{H}^{2k+l+1}_{mix}(\Omega)$ obeys the first k Neumann derivative conditions (2.12), where l is the number of equal values $c(r)$, and that $t \in [d]$. Suppose further that the values $\mathcal{E}^{[i_t]}_{r_t,n_{\bar{t}}}$, $r_t \in \{0, \ldots, k-1\}^{|t|}$, $n_{\bar{t}} \in \mathbb{N}_0^{|\bar{t}|}$ are defined by*

$$\sum_{|s_t|_\infty = 0}^{k-1} \prod_{j \in t} V^{[i_j]}_{r_j,s_j} \mathcal{E}^{[i_t]}_{s_t,n_{\bar{t}}} = \hat{h}^{[i]}_n,$$

where $n_j = m(r_j)$, $r_j = 0, \dots, k-1$, *when* $j \in t$ *and* $n_j \in \mathbb{N}_0$ *otherwise. Then*

$$\left|\mathcal{E}^{[i_t]}_{r_t, n_{\bar{t}}}\right| \lesssim N^{2(|r_t|-k|t|)} \bar{n}_{\bar{t}}^{-2k-2},$$

where $\bar{n}_{\bar{t}}^{-2k-2} = \prod_{j \notin t} \bar{n}_j^{-2k-2}$.

To prove this result, we first require the following trivial lemma:

Lemma 5.13. *Suppose that the coefficients* $a_n \in \mathbb{R}$, $n \in \mathbb{N}^d$, *satisfy* $|a_n| \lesssim n^{-2k-l-1} = \prod_{j=1}^d n_j^{-2k-l-1}$, *where* l *is the number of equal values* $c(r)$. *Then*

$$\left| \sum_{|s|_\infty = 0}^{k-1} \prod_{j=1}^{d} (V^{[i_j]})^{-1}_{r_j, s_j} a_n \right| \lesssim N^{2(|r|-kd)}, \quad \text{where } n = (m(s_1), \dots, m(s_d)).$$

Proof. Using (5.15) and elements of the proof of Theorem 5.5, it is easily seen that $(V^{[i_j]})^{-1}_{r_j, s_j} = \mathcal{O}\left(N^{2r_j+l+1}\right)$. The result now follows immediately. □

Proof of Lemma 5.12. We first establish this result for $t = (1, 2, \dots, d)$. In this case, we write

$$\mathcal{E}^{[i]}_r = \sum_{|s|_\infty = 0}^{k-1} \prod_{j=1}^{d} (V^{[i_j]})^{-1}_{r_j, s_j} \hat{h}^{[i]}_n, \quad \text{where } n = (m(s_1), \dots, m(s_d)).$$

We now prove this result by induction on d. For $d = 1$, the result follows immediately from Theorem 5.5. Now suppose that the result holds for $d-1$. We first construct the subtraction function $g^e_{k,K}$ so that $g^e_{k,K}$ interpolates those odd derivatives of h on the boundary Γ with indices $r = k, k+1, \dots, k+K-1$, where $2K \leq l+1$. It follows that

$$\hat{h}^{[i]}_n = \widehat{g^e_{k,K}}^{[i]}_n + \mathcal{R}^{[i]}_n[f], \tag{5.35}$$

where

$$\mathcal{R}^{[i]}_n[h] = \prod_{j=1}^{d} (\mu^{[i_j]}_{n_j})^{-k-K} \widehat{\mathrm{D}^{k+K} h}^{[i]}_n.$$

Note that $|\mathcal{R}^{[i]}_n[h]| \lesssim \bar{n}^{-2k-l-1}$. In particular, if $n_j = m(r_j)$ for $j = 1, \dots, d$, then $|\mathcal{R}^{[i]}_n[h]| \lesssim N^{-(2k+l+1)d}$. The bound

$$\left| \sum_{|s|_\infty = 0}^{k-1} \prod_{j=1}^{d} (V^{[i_j]})^{-1}_{r_j, s_j} \mathcal{R}^{[i]}_n[h] \right| \lesssim N^{2(|r|-kd)},$$

now follows immediately from Lemma 5.13. Hence, to establish the result, it suffices to consider the term

$$\prod_{j=1}^{d} (V^{[i_j]})^{-1}_{r_j, s_j} \widehat{g^e_{k,K}}^{[i]}_n.$$

By definition, the subtraction function $g^e_{k,K}$ is a sum of separable functions of the form $g(x) = g_1(x_u)g_2(x_{\bar{u}})$, where $u \in [d]$ with $|u| < d$ and $g_1 \in \mathrm{H}^{2k+l+1}_{\mathrm{mix}}(-1,1)^{|u|}$, $g_2 \in \mathrm{H}^{2k+l+1}_{\mathrm{mix}}(-1,1)^{d-|u|}$ (see Lemma 5.21). Hence, applying the induction hypothesis to the functions g_1, g_2 gives

$$\sum_{|s|_\infty=0}^{k-1}\prod_{j=1}^{d}(V^{[i_j]})^{-1}_{r_j,s_j}\widehat{g}^{[i]}_n = \sum_{|s_u|_\infty=0}^{k-1}\prod_{j\in u}(V^{[i_j]})^{-1}_{r_j,s_j}\widehat{g_1}^{[i_u]}_{n_u}\sum_{|s_{\bar{u}}|_\infty=0}^{k-1}\prod_{j\in \bar{u}}(V^{[i_j]})^{-1}_{r_j,s_j}\widehat{g_2}^{[i_{\bar{u}}]}_{n_{\bar{u}}}$$
$$\lesssim N^{2(|r_u|-k|u|)}N^{2(|r_{\bar{u}}|-k|\bar{u}|)} = N^{2(|r|-kd)},$$

which completes the proof in the case $t = (1,\dots,d)$.

Now suppose that $t \in [d]$ is arbitrary. The corresponding result is proved in an almost identical manner. Once more, we use induction on d and write $\hat{h}^{[i]}_n$ as in (5.35). As before, $|\mathcal{R}^{[i]}_n[h]| \lesssim N^{-(2k+l+1)|t|}\bar{n}_{\bar{t}}^{-2k-2}$. Hence, an application of Lemma 5.13 gives the requisite bound for this term.

To bound the contribution of $g^e_{k,K}$, we again write $g^e_{k,K}$ as a sum of terms $g(x) = g_1(x_u)g_2(x_{\bar{u}})$. The induction hypothesis therefore gives

$$\Bigg|\sum_{|s|_\infty=0}^{k-1}\prod_{j\in t}(V^{[i_j]})^{-1}_{r_j,s_j}\widehat{g}^{[i]}_n\Bigg| \lesssim N^{2(|r_{t\cap u}-k|t\cap u|)}\bar{n}_{u\setminus t}^{-2k-2}N^{2(|r_{t\cap\bar{u}}-k|t\cap\bar{u}|)}\bar{n}_{\bar{u}\setminus t}^{-2k-2}$$
$$= N^{2(|r_t|-k|t|)}\bar{n}_{\bar{t}}^{-2k-2},$$

as required. □

Lemma 5.12 may appear somewhat abstract. Yet, its relevance will become apparent in due course. Returning to Eckhoff's method, we now require the following:

Lemma 5.14. *Suppose that* $t \in [d]$, $r_t \in \{0,\dots,k-1\}^{|t|}$, $n_{\bar{t}} \in \{0,\dots,N-1\}^{|\bar{t}|}$ *and*

$$\mathcal{E}^{[i]}_{r_t,n_{\bar{t}}}[f] = \sum_{\substack{u\in[d]\\ t\subseteq u}}\sum_{|r_{u\setminus t}|_\infty=0}^{k-1}\Big(\mathcal{A}^{[i]}_{r_u,n_{\bar{u}}}[f] - \bar{\mathcal{A}}^{[i]}_{r_u,n_{\bar{u}}}[f]\Big)\prod_{j\in u\setminus t}\widehat{p_{r_j}}^{[i_j]}_{n_j}, \tag{5.36}$$

where the values $\bar{\mathcal{A}}^{[i]}_{r_t,n_{\bar{t}}}[f]$ *are the coefficients of Eckhoff's approximation. Then*

$$\sum_{|s_t|_\infty=0}^{k-1}\prod_{j\in t}V^{[i_j]}_{r_j,s_j}\mathcal{E}^{[i]}_{s_t,n_{\bar{t}}}[f] = -\sum_{\substack{u\in[d]^*\\ t\not\subseteq u}}\sum_{|s_u|_\infty=0}^{k-1}\mathcal{A}^{[i]}_{s_u,n_{\bar{u}}}[f]\widehat{p_{s_u}}^{[i_u]}_{n_u}, \quad n\in M_k, \quad i\in\{0,1\}^d. \tag{5.37}$$

Proof. Consider the right-hand side of (5.31). Using the expansion (5.4) gives

$$\hat{f}^{[i]}_n = \sum_{\substack{u\in[d]\\ t\subseteq u}}\sum_{|s_u|_\infty=0}^{k-1}\mathcal{A}^{[i]}_{s_u,n_{\bar{u}}}[f]\widehat{p_{s_u}}^{[i_u]}_{n_u} + \sum_{\substack{u\in[d]^*\\ t\not\subseteq u}}\sum_{|s_u|_\infty=0}^{k-1}\mathcal{A}^{[i]}_{s_u,n_{\bar{u}}}[f]\widehat{p_{s_u}}^{[i_u]}_{n_u}, \quad n\in M_k, \quad i\in\{0,1\}^d.$$

Equating this with (5.33) and rearranging yields the result. □

The terms $\mathcal{E}^{[i]}_{r_t,n_{\bar{t}}}[f]$ may, at first glance, appear obscure. However, as the following lemma attests, they play a fundamental role in the analysis of convergence of Eckhoff's approximation. Before stating this lemma, we first mention that, to estimate the convergence rate of the multivariate Eckhoff approximation $\mathcal{F}_{N,k}[f]$, it suffices to consider the difference $\mathcal{F}^a_{N,k}[f] - \mathcal{F}_{N,k}[f]$, where $\mathcal{F}^a_{N,k}[f]$ is the approximate polynomial subtraction approximation introduced in Section 5.4 (see Theorem 5.11). In view of this, we have

Lemma 5.15. *The difference $\mathcal{F}^a_{N,k}[f](x) - \mathcal{F}_{N,k}[f](x)$ is given by*

$$\begin{aligned}&\mathcal{F}^a_{N,k}[f](x) - \mathcal{F}_{N,k}[f](x)\\ &\qquad = \sum_{i\in\{0,1\}^d}\sum_{t\in[d]}\sum_{|r_t|_\infty=0}^{k-1}\sum_{|n_{\bar{t}}|_\infty=0}^{N-1}\mathcal{E}^{[i]}_{r_t,n_{\bar{t}}}[f]\phi^{[i_{\bar{t}}]}_{n_{\bar{t}}}(x_{\bar{t}})\prod_{j\in t}\left\{p^{[i_j]}_{r_j}(x_j) - \mathcal{F}_N[p^{[i_j]}_{r_j}](x_j)\right\}. \quad (5.38)\end{aligned}$$

Proof. We may write

$$\mathcal{F}^a_{N,k}[f](x) - \mathcal{F}_{N,k}[f](x) = h_k(x) - \mathcal{F}_N[h_k](x), \quad (5.39)$$

where h_k is the smooth function

$$h_k(x) = \sum_{i\in\{0,1\}^d}\sum_{t\in[d]}\sum_{|r_t|_\infty=0}^{k-1}\sum_{|n_{\bar{t}}|_\infty=0}^{N-1}\left(\mathcal{A}^{[i]}_{r_t,n_{\bar{t}}}[f] - \bar{\mathcal{A}}^{[i]}_{r_t,n_{\bar{t}}}[f]\right)p^{[i_t]}_{r_t}(x_t)\phi^{[i_{\bar{t}}]}_{n_{\bar{t}}}(x_{\bar{t}}).$$

To prove the result, it suffices to demonstrate that the right-hand sides of (5.38) and (5.39) have equal modified Fourier coefficients for all indices $i \in \{0,1\}^d$ and $n \in \mathbb{N}_0^d$. It is readily shown that both have vanishing coefficients whenever $n \in I_N$, so we consider the case $n \notin I_N$. In this setting, there is some $u \in [d]$ such that $n_j \geq N$ whenever $j \in u$, and $n_j = 0, \ldots, N-1$ otherwise. Using identical arguments to those used to obtain (5.33), it can be shown that the coefficient of the right-hand side of (5.39) is

$$\widehat{h_k}^{[i]}_n = \sum_{|r_u|_\infty=0}^{k-1}\widehat{p_{r_u}}^{[i_u]}_{n_u}\mathcal{E}^{[i]}_{r_u,n_{\bar{u}}}[f]. \quad (5.40)$$

We now consider the corresponding coefficient of (5.38). For each $t \in [d]$, due to the function $\phi^{[i_{\bar{t}}]}_{n_{\bar{t}}}$, we must have $u \subseteq t$; otherwise, the corresponding term vanishes. However, due to the product, it follows that $t \subseteq u$. Hence, $t = u$, and the modified Fourier coefficient of (5.38) reduces to (5.40), completing the proof. □

We are now in a position to provide analysis for Eckhoff's approximation. Somewhat counterintuitively, we first perform our analysis for a function f that satisfies the first k derivative conditions. This is the content of the following two lemmas. The general case is the verified upon writing $f = (f - g^e_k) + g^e_k$, where g^e_k interpolates the first k odd derivatives of f on the boundary Γ.

Lemma 5.16. *Suppose that $f \in \mathrm{H}^{2k+l+1}_{mix}(\Omega)$ satisfies the first k derivative conditions and that $\mathcal{E}^{[i]}_{r_t,n_{\bar{t}}}[f]$ is given by (5.36). Then $\left|\mathcal{E}^{[i]}_{r_t,n_{\bar{t}}}[f]\right| \lesssim N^{2(|r_t|-k|t|)}\bar{n}_{\bar{t}}^{-2k-2}$.*

Proof. Since f obeys the first k derivative conditions, $\mathcal{A}^{[i]}_{r_t,n_{\bar{t}}}[f] = 0$ when $t \neq \emptyset$. Hence

$$\sum_{|s_t|_\infty=0}^{k-1} \prod_{j\in t} V^{[i_j]}_{r_j,s_j} \mathcal{E}^{[i]}_{s_t,n_{\bar{t}}}[f] = -\mathcal{A}^{[i]}_{n}[f], \quad n \in M_k, \quad i \in \{0,1\}^d.$$

Since $\mathcal{A}^{[i]}_n[f] = \hat{f}^{[i]}_n$ in this case, an application of Lemma 5.12 now yields the result. □

We are now able to provide an error estimate for such a function f:

Lemma 5.17. *Suppose that $f \in \mathrm{H}^{2k+l+1}_{mix}(\Omega)$ obeys the first k derivative conditions, where l is the number of equal values $c(r)$, and that $\mathcal{F}_{N,k}[f]$ is the multivariate Eckhoff approximation of f. Then $\|\mathrm{D}^\beta(f - \mathcal{F}_{N,k}[f])\|_\infty$ is $\mathcal{O}\left(N^{|\beta|_\infty - 2k - 1}\right)$ for $|\beta|_\infty \leq 2k$ and $\|(f - \mathcal{F}_{N,k}[f])\|_r$ is $\mathcal{O}(N^{r-2k-\frac{3}{2}})$ for $r = 0, \ldots, 2k+1$.*

Proof. It suffices to consider the difference $\mathcal{F}^a_{N,k}[f] - \mathcal{F}_{N,k}[f]$. Using Lemma 5.15, the bound derived in Lemma 5.14, and the fact that $\|(p^{[i]}_r - \mathcal{F}_N[p^{[i]}_r])^{(s)}\|_\infty = \mathcal{O}\left(N^{s-2r-1}\right)$, $r, s \in \mathbb{N}_0$ (see (5.18)), we deduce that

$$\begin{aligned}
&\|\mathrm{D}^\beta(\mathcal{F}^a_{N,k}[f] - \mathcal{F}_{N,k}[f])\|_\infty \\
&\qquad \lesssim \sum_{i\in\{0,1\}^d} \sum_{t\in[d]} \sum_{|r_t|_\infty=0}^{k-1} \sum_{|n_{\bar{t}}|_\infty=0}^{N-1} |\mathcal{E}^{[i]}_{r_t,n_{\bar{t}}}[f]| \left\|\mathrm{D}^{\beta_{\bar{t}}}\phi^{[i_{\bar{t}}]}_{n_{\bar{t}}}\right\|_\infty \prod_{j\in t} \left\|\left(p^{[i_j]}_{r_j} - \mathcal{F}_N[p^{[i_j]}_{r_j}]\right)^{(\beta_j)}\right\|_\infty \\
&\qquad \lesssim \sum_{t\in[d]} \sum_{|r_t|_\infty=0}^{k-1} \sum_{|n_{\bar{t}}|_\infty=0}^{N-1} \bar{n}_{\bar{t}}^{\beta_{\bar{t}}-2k-2} N^{2(|r_t|-k|t|)} \prod_{j\in t} N^{\beta_j - 2r_j - 1}.
\end{aligned}$$

Since $|\beta|_\infty \leq 2k$, we have $\bar{n}_{\bar{t}}^{\beta_{\bar{t}}-2k-2} \leq \bar{n}_{\bar{t}}^{-2}$. Hence

$$\|\mathrm{D}^\beta(\mathcal{F}^a_{N,k}[f] - \mathcal{F}_{N,k}[f])\|_\infty \lesssim \sum_{t\in[d]} N^{2(|r_t|-k|t|)} N^{|\beta_t|_\infty - 2|r_t| - 2|t|} \lesssim N^{|\beta|_\infty - 2k - 1},$$

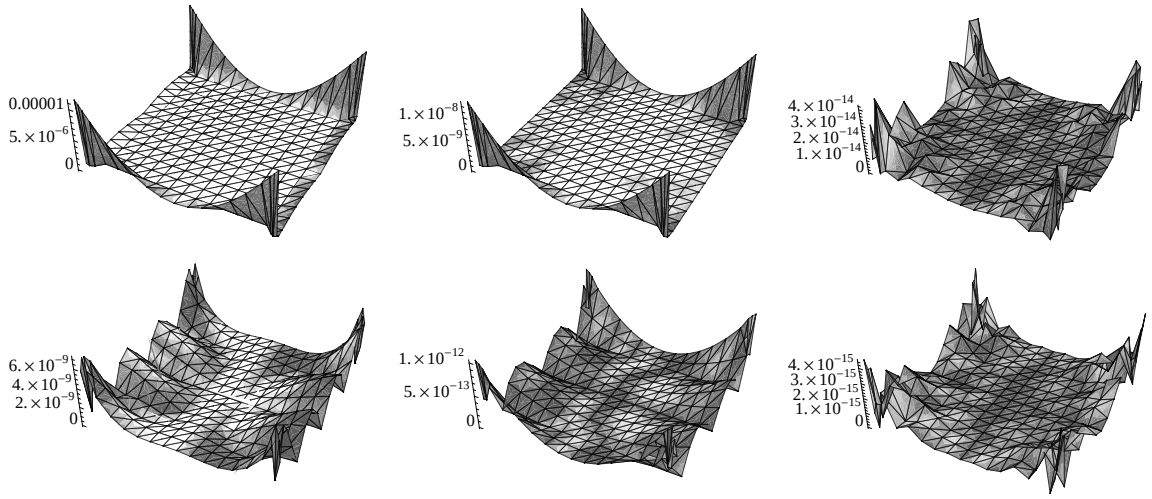
which gives the result for the uniform error. The result for the $\mathrm{H}^r(\Omega)$ norm is established in an identical manner. □

With this in hand, we are able to deduce the main result of this section:

Theorem 5.18. *Suppose that $f \in \mathrm{H}^{2k+l+1}_{mix}(\Omega)$, where l is the number of equal values $c(r)$, and that $\mathcal{F}_{N,k}[f]$ is the multivariate Eckhoff approximation of f. Then $\|\mathrm{D}^\beta(f - \mathcal{F}_{N,k}[f])\|_\infty$ is $\mathcal{O}\left(N^{|\beta|_\infty - 2k - 1}\right)$ for $|\beta|_\infty \leq 2k$ and $\|f - \mathcal{F}_{N,k}[f]\|_r$ is $\mathcal{O}(N^{r-2k-\frac{3}{2}})$ for $r = 0, \ldots, 2k+1$.*

Proof. We proceed by induction on d. Since the univariate result has been proved, we assume that the result holds for $d-1$. Suppose that g^e_k is the exact polynomial subtraction function (5.21), so that $f - g^e_k$ satisfies the first k derivative conditions. Writing $f = (f - g^e_k) + g^e_k$ and using the linearity of $\mathcal{F}_{N,k}[\cdot]$, it follows from Lemma 5.17 that it suffices to consider the difference $g^e_k - \mathcal{F}_{N,k}[g^e_k]$.

The function g^e_k is a finite sum of functions $h(x)$ of the form $h_1(x_t)h_2(x_{\bar{t}})$, $t \in [d]$, $|t| < d$, where $h_1 \in \mathrm{H}^{2k+l+1}_{\mathrm{mix}}(-1,1)^{|t|}$ and $h_2 \in \mathrm{H}^{2k+l+1}_{\mathrm{mix}}(-1,1)^{|\bar{t}|}$. Using linearity once more, it is sufficient to prove the result for h. In the usual manner, we consider the difference $\mathcal{F}^a_{N,k}[h] -$



$k = 1$ $k = 2$ $k = 4$

Figure 5.5: Pointwise error $|f(x_1, x_2) - \mathcal{F}_{25,k}[f](x_1, x_2)|$ for $-1 \leq x_1, x_2 \leq 1$ (top row) and $-\frac{1}{2} \leq x_1, x_2 \leq \frac{1}{2}$ (bottom row), where $\mathcal{F}_{N,k}[f]$ is Eckhoff's approximation based on values $m(r) = N + r$ and $f(x_1, x_2) = (1 - \cos x_1) \sin 3x_2$.

$\mathcal{F}_{N,k}[h]$, where $\mathcal{F}^a_{N,k}[h]$ is the approximate polynomial subtraction approximation of h. A simple argument verifies that

$$\mathcal{F}^a_{N,k}[h] = \mathcal{F}^a_{N,k}[h_1]\mathcal{F}^a_{N,k}[h_2], \quad \mathcal{F}_{N,k}[h] = \mathcal{F}_{N,k}[h_1]\mathcal{F}_{N,k}[h_2].$$

Noting that $a_1b_1 - a_2b_2 = (a_1 - a_2)b_1 + a_2(b_1 - b_2)$ for arbitrary $a_1, a_2, b_1, b_2 \in \mathbb{R}$, we write

$$\mathcal{F}^a_{N,k}[h] - \mathcal{F}_{N,k}[h] = \left(\mathcal{F}^a_{N,k}[h_1] - \mathcal{F}_{N,k}[h_1]\right)\mathcal{F}^a_{N,k}[h_2] + \mathcal{F}_{N,k}[h_1]\left(\mathcal{F}^a_{N,k}[h_2] - \mathcal{F}_{N,k}[h_2]\right).$$

By induction

$$\left\|\mathrm{D}^{\beta_t}\left(\mathcal{F}^a_{N,k}[h_1] - \mathcal{F}_{N,k}[h_1]\right)\right\|_\infty \lesssim N^{|\beta_t|_\infty - 2k - 1}, \quad \left\|\mathrm{D}^{\beta_t}\mathcal{F}_{N,k}[h_1]\right\|_\infty \lesssim 1,$$
$$\left\|\mathrm{D}^{\beta_{\bar{t}}}\left(\mathcal{F}^a_{N,k}[h_2] - \mathcal{F}_{N,k}[h_2]\right)\right\|_\infty \lesssim N^{|\beta_{\bar{t}}|_\infty - 2k - 1}, \quad \left\|\mathrm{D}^{\beta_{\bar{t}}}\mathcal{F}^a_{N,k}[h_2]\right\|_\infty \lesssim 1.$$

Hence

$$\left\|\mathrm{D}^{\beta}\left(\mathcal{F}^a_{N,k}[h] - \mathcal{F}_{N,k}[h]\right)\right\|_\infty \lesssim N^{|\beta_t|_\infty - 2k - 1} + N^{|\beta_{\bar{t}}|_\infty - 2k - 1} \lesssim N^{|\beta|_\infty - 2k - 1},$$

as required. The result for the $\mathrm{H}^r(\Omega)$ norm is verified in an identical manner. □

As in the univariate setting, we arrive at the following conclusion: approximating jump values with Eckhoff's method does not lead to any deterioration in the convergence rate. In Figure 5.5 we demonstrate the benefit offered by the bivariate version of Eckhoff's method. For the function considered, we obtain 13 digits of accuracy using only $k = 4$ and $N = 25$.

Once more, additional smoothness is required for the multivariate version of Eckhoff's method over approximation by polynomial subtraction unless the values $c(r)$, $r = 0, \ldots, k-1$,

are distinct. However, as we consider next, there is an advantage to choosing equal values $c(r)$: namely, a much faster convergence rate inside the domain Ω.

To sum up, a multivariate version of Eckhoff's method can be easily developed for functions defined on the d-variate cube. The resulting approximation attains a uniform convergence rate of $\mathcal{O}\left(N^{-2k-1}\right)$ using only modified Fourier coefficients. In contrast to previous assertions (see, for example, [62]), we conclude that Eckhoff's method is not limited to functions of one variable.

5.7 The auto-correction phenomenon

As established in Theorems 5.8 and 5.11, polynomial subtraction (both exact and approximate) has a convergence rate one power of N faster inside the domain than on the boundary. It transpires that, for the specific choice of the values $m(r) = N + r$ (as mentioned in Section 5.3.2), Eckhoff's approximation possesses the much faster convergence rate of $\mathcal{O}\left(N^{-3k-2}\right)$ away from the boundary—a full $\mathcal{O}\left(N^{k}\right)$ faster than the corresponding approximation based on exact jump values. This auto-correction phenomenon was observed numerically in [129] and proved in the univariate, Fourier case in [138].[8] The aim of this section is to extend this result to the multivariate modified Fourier setting. Furthermore, we shall extend existing theory of the auto-correction phenomenon in the following manner. We will establish that the auto-correction phenomenon manifests itself not just in the convergence rate, but also in the degree of convergence. In other words, as we shall prove, derivatives $\mathrm{D}^{\beta}\mathcal{F}_{N,k}[f](x)$ converge to $\mathrm{D}^{\beta}f(x)$ away from the boundary Γ for higher values of $|\beta|_{\infty}$ than those guaranteed by Theorem 5.18 (i.e. $|\beta|_{\infty} \leq 2k$).

In previous sections, we noted that Eckhoff's approximation decouples into terms corresponding to each particular value of i. The analysis of each term can be undertaken separately, and, since each case is virtually identical, it suffices to consider only one particular value. For the remainder of this section, we assume that f has only non-zero modified Fourier coefficients when $i = (0, 0, \ldots, 0)$. Accordingly, we drop the $[i]$ superscript.

Since uniform convergence of Eckhoff's approximation on $\bar{\Omega}$ is assured by Theorem 5.18, we may write

$$f(x) - \mathcal{F}_{N,k}[f](x) = \sum_{n \notin I_N} \hat{v}_n \phi_n(x) = \sum_{t \in [d]} \sum_{\substack{n_j \geq N \\ j \in t}} \sum_{|n_{\bar{t}}|_{\infty}=0}^{N-1} \hat{v}_n \phi_n(x), \quad x \in \bar{\Omega}, \tag{5.41}$$

where $v(x) = f(x) - g_k(x)$ and g_k is given by (5.28). Following the same method of proof as in [138], we first seek to expand the right-hand side of (5.41) using the so-called *Abel transformation*. Given a sequence $a_m \in \mathbb{R}$, $m \in \mathbb{N}$, we define the operator $\triangle_{r,n}$, $r, n \in \mathbb{N}$, by

$$\triangle_{0,n}[a_m] = a_n, \quad \triangle_{r+1,n}[a_m] = \triangle_{r,n}[a_m] + \triangle_{r,n+1}[a_m], \quad r, n \in \mathbb{N}.$$

It is easily seen that

$$\triangle_{r,n}[a_m] = \sum_{s=0}^{r} \binom{r}{s} a_{n+s}, \quad r, n \in \mathbb{N}. \tag{5.42}$$

[8] The version of Eckhoff's method based on discrete Fourier data also exhibits an auto-correction phenomenon. In fact, the phenomenon is more pronounced in this case: the convergence rate away from the endpoints is a factor of $\mathcal{O}(N^{2(k+1)})$ faster [139].

Now suppose that $a_m \in \mathbb{R}$, $m \in \mathbb{N}^d$. We write $\triangle^j_{r,n}$, $j = 1, \dots, d$, for the above operator acting on the j^{th} entry of n. Further, given $t \in [d]$, $r \in \mathbb{N}^{|t|}$ and $n \in \mathbb{N}^{|t|}$ we define $\triangle^t_{r,n}$ by the composition of $|t|$ such operators:

$$\triangle^t_{r,n}[a_m] = \triangle^{t_1}_{r_{t_1},n_{t_1}}\left[\triangle^{t_2}_{r_{t_2},n_{t_2}}\left[\dots \triangle^{t_{|t|}}_{r_{t_{|t|}},n_{t_{|t|}}}[a_m]\right]\right].$$

It follows from (5.42) that

$$\triangle^t_{r,n}[a_m] = \sum_{s_{t_1}=0}^{r_{t_1}} \dots \sum_{s_{t_{|t|}}=0}^{r_{t_{|t|}}} \binom{r_{t_1}}{s_{t_1}} \dots \binom{r_{t_{|t|}}}{s_{t_{|t|}}} a_{(n+s;m)}, \tag{5.43}$$

where $(n+s;m)$ has j^{th} entry $n_j + s_j$ if $j \in t$ and m_j otherwise.

Before using this transform, we need some additional notation. Given $x, y \in \mathbb{R}^d$, we write $x.y$ for the dot product $x_1y_1 + \dots x_dy_d$, and, if $y = (c, c, \dots, c)$ has equal entries, just $x.c$. Moreover, given $u \in [t]^*$, $r_u \in \mathbb{N}_0^{|u|}$ and $k \in \mathbb{N}_0$, we define $(r_u; k) \in \mathbb{N}_0^{|t|}$ by the condition that the j^{th} entry of $(r_u; k)$, which we write $(r_u; k)_j$, takes value r_j if $j \in u$ and k otherwise.

Lemma 5.19. *Suppose that $g \in \mathrm{H}^1_{mix}(\Omega)$, $t \in [d]$ and that $x \in \Omega$. Then, for $k \in \mathbb{N}_0$ and $n_{\bar{t}} \in \mathbb{N}_0^{|\bar{t}|}$, we have*

$$\sum_{\substack{n_j \geq N \\ j \in t}} \hat{g}_n \phi_{n_t}(x_t) =$$

$$\mathrm{Re}\left\{ \sum_{u \in [t]^*} \sum_{|r_u|_\infty = 0}^{k} \mathrm{e}^{\mathrm{i}\pi x_u.(N-1)} \prod_{j \in t} (1 + \mathrm{e}^{-\mathrm{i}\pi x_j})^{-(r_u;k)_j - 1} \sum_{\substack{n_j \geq N \\ j \in \bar{u}}} \triangle^t_{(r_u;k+1),(n_{\bar{u}};N)}[\hat{g}_m] \mathrm{e}^{\mathrm{i}\pi n_{\bar{u}}.x_{\bar{u}}} \right\},$$

where $(n_{\bar{u}}; N) \in \mathbb{N}_0^{|t|}$ has j^{th} entry n_j if $j \in \bar{u}$ and N otherwise, and the values $m_{\bar{t}} = n_{\bar{t}}$.

Proof. We proceed by induction on $|t|$. Suppose first that $|t| = 1$ and, without loss of generality, that $d = 1$. The verification of the lemma in this case is standard (see also [138]). We have

$$\begin{aligned} \sum_{n \geq N} \hat{g}_n \mathrm{e}^{\mathrm{i}n\pi x} &= \sum_{n \geq N} (\triangle_{1,n}[\hat{g}_m] - \hat{g}_{n+1})\, \mathrm{e}^{\mathrm{i}n\pi x} \\ &= \sum_{n \geq N} \triangle_{1,n}[\hat{g}_m] \mathrm{e}^{\mathrm{i}n\pi x} - \mathrm{e}^{-\mathrm{i}\pi x} \sum_{n \geq N} \hat{g}_n \mathrm{e}^{\mathrm{i}n\pi x} + \hat{g}_N \mathrm{e}^{\mathrm{i}(N-1)\pi x}. \end{aligned}$$

Rearranging gives

$$\sum_{n \geq N} \hat{g}_n \mathrm{e}^{\mathrm{i}n\pi x} = \frac{\mathrm{e}^{\mathrm{i}(N-1)\pi x}}{1 + \mathrm{e}^{-\mathrm{i}\pi x}} \hat{g}_N + \frac{1}{1 + \mathrm{e}^{-\mathrm{i}\pi x}} \sum_{n \geq N} \triangle_{1,n}[\hat{g}_m] \mathrm{e}^{\mathrm{i}n\pi x},$$

which provides the result for $k = 0$. Iterating this process yields the result for general k.

Now let $t \in [d]$ be of length $|t| \geq 2$. Write $t = (t_1, t_2, \dots, t_{|t|})$ and $\tau = (t_2, \dots, t_{|t|}) \in [d]$. Then

$$\sum_{\substack{n_j \geq N \\ j \in t}} \hat{g}_n \phi_{n_t}(x_t) = \sum_{n_{t_1} \geq N} \phi_{n_{t_1}}(x_{t_1}) \sum_{\substack{n_j \geq N \\ j \in \tau}} \hat{g}_n \phi_{n_\tau}(x_\tau).$$

By induction hypothesis, we have

$$\begin{aligned}\sum_{\substack{n_j\geq N\\ j\in t}}\hat{g}_n\phi_{n_t}(x_t) =&\mathrm{Re}\sum_{n_{t_1}\geq N}\mathrm{e}^{\mathrm{i}n_{t_1}\pi x_{t_1}}\Bigg\{\sum_{u\in[\tau]^*}\mathrm{e}^{\mathrm{i}\pi x_u.(N-1)}\\ &\times\sum_{|r_u|_\infty=0}^{k}\prod_{j\in\tau}(1+\mathrm{e}^{-\mathrm{i}\pi x_j})^{-(r_u;k)_j-1}\sum_{\substack{n_j\geq N\\ j\in\bar{u}}}\triangle^{\tau}_{(r_u;k+1),(n_{\bar{u}};N)}[\hat{g}_m]\mathrm{e}^{\mathrm{i}\pi n_{\bar{u}}.x_{\bar{u}}}\Bigg\}\\ =&\mathrm{Re}\sum_{u\in[\tau]^*}\mathrm{e}^{\mathrm{i}\pi x_u.(N-1)}\sum_{|r_u|_\infty=0}^{k}\prod_{j\in\tau}(1+\mathrm{e}^{-\mathrm{i}\pi x_j})^{-(r_u;k)_j-1}\\ &\times\sum_{\substack{n_j\geq N\\ j\in\bar{u}}}\mathrm{e}^{\mathrm{i}\pi n_{\bar{u}}.x_{\bar{u}}}\sum_{n_{t_1}\geq N}\mathrm{e}^{\mathrm{i}n_{t_1}\pi x_{t_1}}\triangle^{\tau}_{(r_u;k+1),(n_{\bar{u}};N)}[\hat{g}_m].\end{aligned}\tag{5.44}$$

Using the result for $|t|=1$ gives

$$\begin{aligned}&\sum_{n_{t_1}\geq N}\mathrm{e}^{\mathrm{i}n_{t_1}\pi x_{t_1}}\triangle^{\tau}_{(r_u;k+1),(n_{\bar{u}};N)}[\hat{g}_n]\\ &\quad=\sum_{r_{t_1}=0}^{k}\mathrm{e}^{\mathrm{i}\pi x_{t_1}(N-1)}(1+\mathrm{e}^{\mathrm{i}\pi x_{t_1}})^{-r_{t_1}-1}\triangle^{t_1}_{r_{t_1},N}\left[\triangle^{\tau}_{(r_u;k+1),(n_{\bar{u}};N)}[\hat{g}_m]\right]\\ &\quad\quad+\sum_{n_{t_1}\geq N}(1+\mathrm{e}^{-\mathrm{i}\pi x_{t_1}})^{-k-1}\triangle^{t_1}_{k+1,n_{t_1}}\left[\triangle^{\tau}_{(r_u;k+1),(n_{\bar{u}};N)}[\hat{g}_m]\right].\end{aligned}\tag{5.45}$$

Substituting (5.45) into (5.44) now completes the proof. Note that if $v\in[t]^*$, then either $v\in[\tau]^*$ or $v=(t_1,u)$ for some $u\in[\tau]^*$. The two terms of (5.45) correspond respectively to these scenarios. □

The crux of the auto-correction phenomenon is the following trivial observation:

Lemma 5.20. *Suppose that $v=f-g_k$, where g_k is given by (5.28), and that $m(r)=N+r$, $r=0,\dots,k-1$. Then $\triangle^t_{r_t,n_t}[\hat{v}_m]=0$ for all $|r_t|_\infty\leq k-1$, $|n_t|_\infty,|m_{\bar{t}}|_\infty\leq N$ and $t\in[d]$.*

Proof. Since $\hat{v}_n=0$ for $|n|_\infty\leq N+k-1$, the result follows directly from (5.43). □

We may now re-write (5.41) as

$$f(x)-\mathcal{F}_{N,k}[f](x)=\sum_{t\in[d]}\sum_{|n_{\bar{t}}|_\infty=0}^{N-1}h_{n_{\bar{t}}}(x_t)\phi_{n_{\bar{t}}}(x_{\bar{t}}),\tag{5.46}$$

where $h_{n_{\bar{t}}}(x_t)$ is obtained from the expansion derived in Lemma 5.19:

$$\begin{aligned}&h_{m_{\bar{t}}}(x_t)\\ &=\mathrm{Re}\left\{\sum_{u\in[t]^*}\sum_{|r_u|_\infty=0}^{k}\mathrm{e}^{\mathrm{i}\pi x_u.(N-1)}\prod_{j\in t}(1+\mathrm{e}^{-\mathrm{i}\pi x_j})^{-(r_u;k)_j-1}\sum_{\substack{n_j\geq N\\ j\in\bar{u}}}\triangle^{t}_{(r_u;k+1),(n_{\bar{u}};N)}[\hat{v}_m]\mathrm{e}^{\mathrm{i}\pi n_{\bar{u}}.x_{\bar{u}}}\right\}.\end{aligned}$$

Consider the term of $h_{m_{\bar{t}}}$ corresponding to $u = t$ separately. This is

$$\mathrm{e}^{\mathrm{i}\pi x_t.(N-1)} \sum_{|r_t|_\infty=0}^{k} \prod_{j\in t} (1+\mathrm{e}^{-\mathrm{i}\pi x_j})^{-r_j-1} \triangle^t_{r_t,N}[\hat{v}_m],$$

where we write $\triangle^t_{r_t,N}$ instead of the full expression $\triangle^t_{r_t,(N,N,\ldots,N)}$. By Lemma 5.20, all terms of this expression where $|r_t|_\infty < k$ are zero. Hence, we define

$$H_{m_{\bar{t}}}(x_t) = \mathrm{e}^{\mathrm{i}\pi x_t.(N-1)} \sum_{|r_t|_\infty=k} \prod_{j\in t} (1+\mathrm{e}^{-\mathrm{i}\pi x_j})^{-r_j-1} \triangle^t_{r_t,N}[\hat{v}_m], \tag{5.47}$$

and

$$G_{m_{\bar{t}}}(x_t) = \sum_{\substack{u\in[t]^* \\ u\neq t}} \left\{ \sum_{|r_u|_\infty=0}^{k} \mathrm{e}^{\mathrm{i}\pi x_u.(N-1)} \prod_{j\in t} (1+\mathrm{e}^{-\mathrm{i}\pi x_j})^{-(r_u;k)_j-1} \right.$$
$$\left. \times \sum_{\substack{n_j\geq N \\ j\in\bar{u}}} \triangle^t_{(r_u;k+1),(n_{\bar{u}};N)}[\hat{v}_m] \mathrm{e}^{\mathrm{i}\pi n_{\bar{u}}.x_{\bar{u}}} \right\}, \tag{5.48}$$

so that the function h_{n_t} may be expressed as $h_{n_{\bar{t}}}(x_t) = \mathrm{Re}\left\{G_{n_{\bar{t}}}(x_t) + H_{n_{\bar{t}}}(x_t)\right\}$. To derive an estimate for the error $f(x) - \mathcal{F}_{N,k}[f](x)$, we first require bounds for the functions $G_{n_{\bar{t}}}$ and $H_{n_{\bar{t}}}$. We derive such bounds in the sequel. First, however, it is useful to consider the case $d = 1$, to demonstrate elements of the multivariate proof. This is given in a similar form in [138].

5.7.1 The univariate case

Using (5.41) and the characterisation given in Lemma 5.19 with $t = (1)$, we may write

$$\begin{aligned} f(x) - \mathcal{F}_{N,k}[f](x) &= \sum_{n\geq N} \hat{v}_n \phi_n(x) \\ &= \mathrm{Re}\left\{ \sum_{r=0}^{k} \frac{\mathrm{e}^{\mathrm{i}(N-1)\pi x}}{(1+\mathrm{e}^{-\mathrm{i}\pi x})^{r+1}} \triangle_{r,N}[\hat{v}_m] + \frac{1}{(1+\mathrm{e}^{-\mathrm{i}\pi x})^{k+1}} \sum_{n\geq N} \triangle_{k+1,n}[\hat{v}_m] \mathrm{e}^{\mathrm{i}n\pi x} \right\}. \end{aligned}$$

In light of Lemma 5.20, $\triangle_{r,N}[\hat{v}_m] = 0$ for $r = 0, \ldots, k-1$, so this reduces to

$$\begin{aligned} f(x) - \mathcal{F}_{N,k}[f](x) &= \mathrm{Re}\left\{ \frac{\mathrm{e}^{\mathrm{i}(N-1)\pi x}}{(1+\mathrm{e}^{-\mathrm{i}\pi x})^{k+1}} \triangle_{k,N}[\hat{v}_m] + \frac{1}{(1+\mathrm{e}^{-\mathrm{i}\pi x})^{k+1}} \sum_{n\geq N} \triangle_{k+1,n}[\hat{v}_m] \mathrm{e}^{\mathrm{i}n\pi x} \right\} \\ &= \mathrm{Re}\left[H(x) + G(x)\right], \end{aligned} \tag{5.49}$$

where $G(x)$ and $H(x)$ are the univariate versions of $G_{n_{\bar{t}}}$ and $H_{n_{\bar{t}}}$. Note that for $d = 1$ there is only one $t \in [d]$, namely, $t = (1)$, and trivially $\bar{t} = \emptyset$.

We seek bounds for G and H. To do so, we require the following two lemmas:

Lemma 5.21. *Suppose that* $h \in \mathrm{L}^2(-1,1)$, $r \in \mathbb{N}_0$ *and* $n \in \mathbb{N}_0$. *Then*

$$\triangle_{r,n}[\hat{h}_m] = \begin{cases} \widehat{\mathcal{G}_r[h]}_{n+\frac{r}{2}} & r \text{ even} \\ \widehat{\mathcal{G}_r[h]}_{n+\frac{r+1}{2}} & r \text{ odd} \end{cases},$$

where $\mathcal{G}_r[h](x) = 2^{-r}(\cos\frac{1}{2}\pi x)^r h(x)$ *and* $\check{g}_n$ *is the coefficient of a function* g *corresponding to the Laplace–Dirichlet cosine function* $\cos(n-\frac{1}{2})\pi x$.

Proof. We proceed by induction on r. For $r = 0$, the result is trivial. Now suppose that the result holds for $r-1$. From the definition of $\triangle_{r,n}[\cdot]$, we have

$$\begin{aligned}\triangle_{r,n}[\hat{h}_m] &= \triangle_{r-1,n}[\hat{h}_m] + \triangle_{r-1,n+1}[\hat{h}_m] \\ &= 2^{-(r-1)} \int_{-1}^{1} (\cos\tfrac{1}{2}\pi x)^{r-1} h(x) \left\{\cos(n+\tfrac{1}{2}(r-1))\pi x + \cos(n+1+\tfrac{1}{2}(r-1))\pi x\right\} dx \\ &= 2^{-r} \int_{-1}^{1} (\cos\tfrac{1}{2}\pi x)^{r} h(x) \cos(n+\tfrac{r}{2})\pi x \,\mathrm{d}x,\end{aligned}$$

as required. □

The function $\mathcal{G}_r[h]$ has the following property, vital to our subsequent analysis:

Lemma 5.22. *Suppose that* $h \in \mathrm{H}^{2k+r}(-1,1)$ *obeys the first* k *Neumann derivative conditions* $h^{(2s+1)}(\pm 1) = 0$, $s = 0,\dots,k-1$. *Then* $\mathcal{G}_r[h]$ *obeys the first* $k+\frac{r}{2}$ *Neumann derivative conditions (2.12) if* r *is even, and the first* $k+\frac{r+1}{2}$ *Dirichlet derivative conditions (2.14) otherwise.*

Proof. This follows from the definition $\mathcal{G}_r[h](x) = 2^{-r}(\cos\frac{1}{2}\pi x)^r h(x)$, Leibniz's rule, and the fact that all even derivatives of $\cos\frac{1}{2}\pi x$ vanish at $x = \pm 1$. □

Corollary 5.23. *Suppose that* h *is as in Lemma 5.22. Then* $|\triangle_{r,n}[\hat{h}_m]| \lesssim n^{-2k-r-2}$.

Proof. This follows immediately from Lemmas 5.21, 5.22 and standard properties of Laplace–Neumann and Laplace–Dirichlet coefficients. □

We are now able to provide bounds for $G(x)$ and $H(x)$. We have

Lemma 5.24. *Suppose that* $f \in \mathrm{H}^{3(k+1)}(-1,1)$ *and that* $G(x)$, $H(x)$ *are as in (5.49). Then* $|G(x)|, |H(x)| \lesssim N^{-3k-2}$ *uniformly for* $x \in (-1,1)$.

Proof. If $v(x) = f(x) - g_k(x)$, then

$$\hat{v}_n = \sum_{r=0}^{k-1} \left(\mathcal{A}_r[f] - \bar{\mathcal{A}}_r[f]\right)\widehat{p}_{rn} + \sum_{r=k}^{k+K-1} \mathcal{A}_r[f]\widehat{p}_{rn} + \mathcal{O}\left(n^{-2(k+K)-l}\right), \tag{5.50}$$

where $K \in \mathbb{N}_0$, $l = 1, 2$ and $2(k+K)+l = 3(k+1)$. Hence

$$\triangle_{s,n}[\hat{v}_m] = \sum_{r=0}^{k-1} \left(\mathcal{A}_r[f] - \bar{\mathcal{A}}_r[f]\right)\triangle_{s,n}[\widehat{p}_{rm}] + \sum_{r=k}^{k+K-1} \mathcal{A}_r[f]\triangle_{s,n}[\widehat{p}_{rm}] + \mathcal{O}\left(n^{-2(k+K)-l}\right).$$

Using Theorem 5.5 and Corollary 5.23, we obtain

$$|\triangle_{s,n}[\hat{v}_m]| \lesssim \sum_{r=0}^{k-1} N^{2(r-k)} \bar{n}^{-2r-s-2} + \bar{n}^{-2k-s-2} + \bar{n}^{-2(k+K)-l}.$$

Now, if k is even, we set $2K = k+2$ and $l = 1$. Conversely, if k is odd, we define $2K = k+1$ and $l = 2$. Substituting such values into the above expression, we obtain

$$|\triangle_{k,N}[\hat{v}_m]| \lesssim N^{-3k-2}, \quad |\triangle_{k+1,n}[\hat{v}_m]| \lesssim N^{-3k-1} n^{-2}, \quad n \geq N.$$

Recalling the definitions of G and H given in (5.49), this give the result. □

A proof of the univariate auto-correction phenomenon now follows immediately:

Theorem 5.25. *Suppose that $f \in \mathrm{H}^{3(k+1)}(-1,1)$ and that $\mathcal{F}_{N,k}[f]$ is the univariate Eckhoff approximation based on the values $m(r) = N + r$, $r = 0, \ldots, k-1$. Then $f(x) - \mathcal{F}_{N,k}[f](x)$ is $\mathcal{O}\left(N^{-3k-2}\right)$ uniformly for x in compact subsets of $(-1,1)$.*

5.7.2 Bounds for $G_{n_{\bar{t}}}$ and $H_{n_{\bar{t}}}$

With the univariate result to hand, we now return to the multivariate case. To derive bounds for $G_{n_{\bar{t}}}$ and $H_{n_{\bar{t}}}$, we commence with the following preliminary result:

Lemma 5.26. *Suppose that $t \in [d]$, $r_t \in \mathbb{N}_0^{|t|}$, $2K \geq k+1$ and that the function $h \in \mathrm{H}^{2(k+K)+1}_{mix}(\Omega)$, satisfies the first $s_j \leq k$ derivative conditions in each variable x_j, $j = 1, \ldots, d$. Then*

$$\left|\triangle_{r_t,n_t}[\hat{h}_n]\right| \lesssim \prod_{j=1}^{d} \bar{n}_j^{2s_j-2} \prod_{j \in t} \bar{n}_j^{-r_j} = \bar{n}^{-2s-2} \bar{n}_t^{-r_t}.$$

Proof. This result follows immediately after applications of Lemmas 5.21, 5.22 and Corollary 5.23 in each variable x_j, $j \in t$. □

With this in hand, we may estimate the functions $G_{n_{\bar{t}}}$ and $H_{n_{\bar{t}}}$. For the latter, we have:

Lemma 5.27. *Suppose that $f \in \mathrm{H}^{3(k+1)}_{mix}(\Omega)$. Then the function $H_{n_{\bar{t}}}$ defined by (5.47) satisfies $\left|H_{n_{\bar{t}}}(x_t)\right| \lesssim N^{-3k-2} \bar{n}_{\bar{t}}^{-2}$ uniformly for x_t in compact subsets of $(-1,1)^{|t|}$.*

Proof. For $n \in \mathbb{N}_0^d$ with $n_j \geq N$ whenever $j \in t$ and $n_j = 0, \ldots, N-1$ otherwise, the coefficient $\hat{v}_n$ satisfies

$$\hat{v}_n = \sum_{|s_t|_\infty = 0}^{k-1} \mathcal{E}_{s_t,n_{\bar{t}}}[f] \widehat{p_{s_t n_t}} + \sum_{\substack{v \in [d]^* \\ t \not\subseteq v}} \sum_{|s_v|_\infty = 0}^{k-1} \mathcal{A}_{s_v,n_{\bar{v}}}[f] \widehat{p_{s_v n_v}}. \tag{5.51}$$

We now substitute the two terms of (5.51) into the definition of $H_{n_{\bar{t}}}$ given by (5.47) and consider them separately. For the first term, we have

$$\triangle^t_{r_t,N}\left[\mathcal{E}_{s_t,n_{\bar{t}}}[f] \widehat{p_{s_t n_t}}\right] = \mathcal{E}_{s_t,n_{\bar{t}}}[f] \triangle^t_{r_t,N}\left[\widehat{p_{s_t n_t}}\right] = \mathcal{E}_{s_t,n_{\bar{t}}}[f] \prod_{j \in t} \triangle^j_{r_j,N}[\widehat{p_{s_j n_j}}].$$

Using Lemmas 5.14 and 5.26, we obtain the bound

$$\left|\triangle^t_{r_t,N}\left[\mathcal{E}_{s_t,n_{\bar{t}}}[f]\widehat{p_{s_t n_t}}\right]\right| \lesssim N^{2(|s_t|_\infty - k)} \prod_{j\in t} N^{-2s_j - r_j - 2} \bar{n}_{\bar{t}}^{-2} \lesssim N^{-2k-|r_t|-2|t|} \bar{n}_{\bar{t}}^{-2}.$$

Since $|r_t| \geq |r_t|_\infty = k$ and $|t| \geq 1$, this gives the required estimate for the first term.

Now consider the second term of (5.51) substituted into (5.47). For $v \in [d]^*$ with $t \not\subseteq v$, either (i) $v \cap t \neq \emptyset$ or (ii) $v \cap t = \emptyset$. Consider case (i) first. We have

$$\triangle^t_{r_t,N}\left[\mathcal{A}_{s_v,n_{\bar{v}}}[f]\widehat{p_{s_v n_v}}\right] = \triangle^{t\backslash v}_{r_{t\backslash v},N}\left[\mathcal{A}_{s_v,n_{\bar{v}}}[f]\right] \triangle^{t\cap v}_{r_{t\cap v},N}[\widehat{p_{s_v n_v}}].$$

Since $\mathcal{A}_{s_v,n_{\bar{v}}}[f] = \hat{h}_{n_{\bar{v}}}$, where h is a function of $x_{\bar{v}}$ that obeys the first k derivative conditions, we may apply Lemma 5.26 to give

$$\begin{aligned}\left|\triangle^t_{r_t,N}\left[\mathcal{A}_{s_v,n_{\bar{v}}}[f]\widehat{p_{s_v n_v}}\right]\right| &\lesssim \prod_{j\in t\cap v} N^{-2s_j - r_j - 2} \prod_{j\in t\backslash v} N^{-2k-r_j-2} \bar{n}_{v\backslash t}^{-2s_{v\backslash t}-2} \bar{n}_{\overline{t\cup v}}^{-2k-2} \\ &\lesssim N^{-|r_t|-2|t\cap v|-2(k+1)(|t\backslash v|)} \bar{n}_{\bar{t}}^{-2} \lesssim N^{-3k-2} \bar{n}_{\bar{t}}^{-2}.\end{aligned}$$

Here the final inequality follows since, by assumption, $|t \cap v|, |t\backslash v| \geq 1$. Now consider case (ii). Since $t \cap v = \emptyset$, we have

$$\triangle^t_{r_t,N}\left[\mathcal{A}_{s_v,n_{\bar{v}}}[f]\widehat{p_{s_v n_v}}\right] = \triangle^t_{r_t,N}\left[\mathcal{A}_{s_v,n_{\bar{v}}}[f]\right] \widehat{p_{s_v n_v}}.$$

Using Lemma 5.26, we obtain

$$\begin{aligned}\left|\triangle^t_{r_t,N}\left[\mathcal{A}_{s_v,n_{\bar{v}}}[f]\widehat{p_{s_v n_v}}\right]\right| &\lesssim \prod_{j\in t} N^{-r_j-2k-2} \prod_{j\notin v\cup t} \bar{n}_j^{-2k-2} \prod_{j\in v} \bar{n}_j^{-2s_j-2} \\ &\lesssim N^{-|r_t|_\infty - 2k-2} \bar{n}_{\bar{t}}^{-2} \lesssim N^{-3k-2} \bar{n}_{\bar{t}}^{-2}.\end{aligned}$$

This completes the proof. □

Next we derive a bound for $G_{n_{\bar{t}}}$:

Lemma 5.28. *Suppose that $f \in \mathrm{H}^{3(k+1)}_{mix}(\Omega)$. Then the function $G_{n_{\bar{t}}}$ defined by (5.48) satisfies $\left|G_{n_{\bar{t}}}(x_t)\right| \lesssim N^{-3k-2}\bar{n}_{\bar{t}}^{-2}$, uniformly for x_t in compact subsets of $(-1,1)^{|t|}$.*

Proof. Since $x_t \in (-1,1)^{|t|}$, it suffices to bound

$$\sum_{\substack{u\in[t]^*\\ u\neq t}} \sum_{|r_u|_\infty=0}^{k} \sum_{\substack{n_j\geq N\\ j\in\bar{u}}} \left|\triangle^t_{(r_u;k+1),(n_{\bar{u}};N)}[\hat{v}_m]\right|, \tag{5.52}$$

by $N^{-3k-2}\bar{n}_{\bar{t}}^{-2}$ (where $m_{\bar{t}} = n_{\bar{t}}$). To do so, we substitute the two terms of (5.51) into (5.52) and consider them separately. For the first term, we have

$$\sum_{\substack{u\in[t]^*\\ u\neq t}} \sum_{|r_u|_\infty=0}^{k} \sum_{\substack{n_j\geq N\\ j\in\bar{u}}} \sum_{|s_t|_\infty=0}^{k-1} \left|\triangle^t_{(r_u;k+1),(n_{\bar{u}};N)}\left[\mathcal{E}_{s_t,n_{\bar{t}}}[f]\widehat{p_{s_t n_t}}\right]\right|. \tag{5.53}$$

Since $u \subseteq t$, we observe that

$$\triangle^t_{(r_u;k+1),(n_{\bar{u}};N)} \left[\mathcal{E}_{s_t,n_{\bar{t}}}[f]\widehat{p_{s_t n_t}}\right] = \mathcal{E}_{s_t,n_{\bar{t}}}[f] \prod_{j\in u} \triangle^j_{r_j,N}[\widehat{p_{s_j n_j}}] \prod_{j\in t\setminus u} \triangle^j_{k+1,n_j}[\widehat{p_{s_j n_j}}].$$

Using Lemmas 5.14 and 5.26, we deduce that

$$\left|\triangle^t_{(r_u;k+1),(n_{\bar{u}};N)} \left[\mathcal{E}_{s_t,n_{\bar{t}}}[f]\widehat{p_{s_t n_t}}\right]\right| \lesssim N^{2(|s_t|_\infty-k)}\bar{n}_{\bar{t}}^{-2} \prod_{j\in u} N^{-2s_j-r_j-2}\bar{n}_{\bar{u}}^{-2s_{\bar{u}}-k-3}.$$

Substituting this into (5.53), we obtain

$$\begin{aligned}
&\sum_{\substack{u\in[t]^*\\ u\neq t}} \sum_{|r_u|_\infty=0}^{k} \sum_{\substack{n_j\geq N\\ j\in\bar{u}}} \sum_{|s_t|_\infty=0}^{k-1} \left|\triangle^t_{(r_u;k+1),(n_{\bar{u}};N)} \left[\widehat{p_{s_t n_t}}\mathcal{E}_{s_t,n_{\bar{t}}}[f]\right]\right| \\
&\quad\lesssim \sum_{\substack{u\in[t]^*\\ u\neq t}} \sum_{|r_u|_\infty=0}^{k} \sum_{\substack{n_j\geq N\\ j\in\bar{u}}} \sum_{|s_t|_\infty=0}^{k-1} N^{2(|s_t|_\infty-k)} \prod_{j\in u} N^{-2s_j-r_j-2}\bar{n}_{\bar{t}}^{-2}\bar{n}_{\bar{u}}^{-2s_{\bar{u}}-k-3} \\
&\quad\lesssim \bar{n}_{\bar{t}}^{-2} \sum_{\substack{u\in[t]^*\\ u\neq t}} \sum_{|r_u|_\infty=0}^{k} \sum_{|s_t|_\infty=0}^{k-1} N^{2(|s_t|_\infty-k)} \prod_{j\in u} N^{-2s_j-r_j-2} \prod_{j\in t\setminus u} N^{-2s_j-k-2} \\
&\quad\lesssim \bar{n}_{\bar{t}}^{-2} \sum_{\substack{u\in[t]^*\\ u\neq t}} \sum_{|r_u|_\infty=0}^{k} N^{-2k-|r_u|-2|u|-(k+2)(|t|-|u|)} \lesssim N^{-3k-2}\bar{n}_{\bar{t}}^{-2},
\end{aligned}$$

as required. Here the last inequality follows by noting that $|t| - |u| \geq 1$.

We now consider the second term of (5.51) substituted into (5.52):

$$\sum_{\substack{v\in[d]^*\\ t\not\subseteq v}} \sum_{|s_v|_\infty=0}^{k-1} \sum_{\substack{u\in[t]^*\\ u\neq t}} \sum_{|r_u|_\infty=0}^{k} \sum_{\substack{n_j\geq N\\ j\in\bar{u}}} \left|\triangle^t_{(r_u;k+1),(n_{\bar{u}};N)} \left[\mathcal{A}_{s_v,n_{\bar{v}}}[f]\widehat{p_{s_v n_v}}\right]\right|. \tag{5.54}$$

As in the proof of Lemma 5.27, we split this into two cases: either (i) $v\cap t \neq \emptyset$ or (ii) $v\cap t = \emptyset$. Suppose that we consider case (i). Since $v \cap t \neq \emptyset$, we have

$$\triangle^t_{(r_u;k+1),(n_{\bar{u}};N)} \left[\mathcal{A}_{s_v,n_{\bar{v}}}[f]\widehat{p_{s_v n_v}}\right] = \triangle^{t\cap v}_{(r_{u\cap v};k+1),(n_{\bar{u}\cap v};N)} \left[\widehat{p_{s_v n_v}}\right] \triangle^{t\cap\bar{v}}_{(r_{u\cap\bar{v}};k+1),(n_{\bar{u}\cap\bar{v}};N)} \left[\mathcal{A}_{s_v,n_{\bar{v}}}[f]\right].$$

Note that

$$\left|\triangle^{t\cap v}_{(r_{u\cap v};k+1),(n_{\bar{u}\cap v};N)} \left[\widehat{p_{s_v n_v}}\right]\right| \lesssim \prod_{j\in u\cap v} N^{-2s_j-r_j-2}\bar{n}_{\bar{u}\cap v}^{-2s_{\bar{u}\cap v}-k-3}\bar{n}_{v\setminus t}^{-2s_{v\setminus t}-2}.$$

Furthermore

$$\left|\triangle^{t\cap\bar{v}}_{(r_{u\cap\bar{v}};k+1),(n_{\bar{u}\cap\bar{v}};N)} \left[\mathcal{A}_{s_v,n_{\bar{v}}}[f]\right]\right| \lesssim \prod_{j\in u\cap\bar{v}} N^{-2k-r_j-2}\bar{n}_{\bar{u}\cap\bar{v}}^{-3k-3}\bar{n}_{\overline{t\cup v}}^{-2k-2}.$$

Combining these two estimates yields

$$\begin{aligned}\Big|\triangle^t_{(r_u;k+1),(n_{\bar u};N)}\ \big[\mathcal{A}_{s_v,n_{\bar v}}[f]\widehat{p_{s_v n_v}}\big]\Big| \\ \lesssim \prod_{j\in u\cap v} N^{-2s_j-r_j-2}\prod_{j\in u\cap\bar v} N^{-2k-r_j-2}\bar n_{\bar u\cap v}^{-2s_{\bar u\cap v}-k-3}\bar n_{\bar u\cap\bar v}^{-3k-3}\bar n_{\bar t}^{-2}.\end{aligned}$$

Hence

$$\begin{aligned}&\Big|\sum_{|s_v|_\infty=0}^{k-1}\sum_{\substack{u\in[t]^*\\ u\neq t}}\sum_{|r_u|_\infty=0}^{k}\sum_{\substack{n_j\geq N\\ j\in\bar u}}\triangle^t_{(r_u;k+1),(n_{\bar u};N)}\ \big[\mathcal{A}_{s_v,n_{\bar v}}[f]\widehat{p_{s_v n_v}}\big]\Big| \\ &\quad\lesssim \sum_{|s_v|_\infty=0}^{k-1}\sum_{\substack{u\in[t]^*\\ u\neq t}}\sum_{|r_u|_\infty=0}^{k}\Big\{\prod_{j\in u\cap v}N^{-2s_j-r_j-2}\prod_{j\in u\cap\bar v}N^{-2k-r_j-2} \\ &\qquad\times\prod_{j\in\bar u\cap v}N^{-2s_j-k-2}\prod_{j\in\bar u\cap\bar v}N^{-3k-2}\bar n_{\bar t}^{-2}\Big\} \\ &\quad\lesssim \sum_{\substack{u\in[t]^*\\ u\neq t}}N^{-2(k+1)|u\cap\bar v|}N^{-(k+2)|\bar u\cap v|}N^{-(3k+2)|\bar u\cap\bar v|}\bar n_{\bar t}^{-2}.\end{aligned}$$

We claim that this term is $\lesssim N^{-3k-2}\bar n_{\bar t}^{-2}$. For each u, we have two possibilities: either $\bar u\cap\bar v\neq\emptyset$ or $\bar u\cap\bar v=\emptyset$. If $\bar u\cap\bar v\neq\emptyset$, then the result follows immediately. Now suppose that $\bar u\cap\bar v=\emptyset$. Since $t\not\subseteq v$ and $u\subset t$, we have

$$\emptyset\neq t\cap\bar v=(\bar u\cap\bar v)\cup(u\cap\bar v)=\emptyset\cup(u\cap\bar v)=u\cap\bar v.$$

Similarly, since $u\neq t$, it follows that $\emptyset\neq\bar u=(\bar u\cap v)\cup(\bar u\cap\bar v)=\bar u\cap v$. Hence, $u\cap\bar v,\bar u\cap v\neq\emptyset$, and the result follows in this case. This completes case (i).

Next, consider case (ii). Since $v\cap t=\emptyset$, we have

$$\triangle^t_{(r_u;k+1),(n_{\bar u};N)}\ \big[\mathcal{A}_{s_v,n_{\bar v}}[f]\widehat{p_{s_v n_v}}\big]=\triangle^t_{(r_u;k+1),(n_{\bar u};N)}\ \big[\mathcal{A}_{s_v,n_{\bar v}}[f]\big]\,\widehat{p_{s_v n_v}}.$$

In the standard manner, we obtain

$$\begin{aligned}\Big|\triangle^t_{(r_u;k+1),(n_{\bar u};N)}\ \big[\mathcal{A}_{s_v,n_{\bar v}}[f]\widehat{p_{s_v n_v}}\big]\Big| &\lesssim \prod_{j\in u}N^{-2k-r_j-2}\bar n_{\bar u}^{-3k-3}\bar n_{\bar v\setminus t}^{-2k-2}\bar n_v^{-2s_v-2} \\ &\lesssim N^{-2(k+1)|u|-|r_u|_\infty}\bar n_{\bar u}^{-3k-3}\bar n_{\bar t}^{-2}.\end{aligned}$$

Hence

$$\sum_{|s_v|_\infty=0}^{k-1}\sum_{\substack{u\in[t]^*\\ u\neq t}}\sum_{|r_u|_\infty=0}^{k}\sum_{\substack{n_j\geq N\\ j\in\bar u}}\Big|\triangle^t_{(r_u;k+1),(n_{\bar u};N)}\ \big[\mathcal{A}_{s_v,n_{\bar v}}[f]\widehat{p_{s_v n_v}}\big]\Big|\lesssim\prod_{j\in\bar u}N^{-3k-2}\bar n_{\bar t}^{-2}\lesssim N^{-3k-2}\bar n_{\bar t}^{-2},$$

where the final inequality follows, since $|\bar u|\geq 1$. This completes the proof. □

(x_1, x_2)	$N = 10$	$N = 20$	$N = 30$	$N = 40$	$N = 50$
$(1, 1)$	4.958×10^{-8}	1.307×10^{-10}	3.799×10^{-12}	3.022×10^{-13}	4.202×10^{-14}
$(-1, -1)$	6.341×10^{-8}	1.372×10^{-10}	3.723×10^{-12}	2.861×10^{-13}	3.898×10^{-14}
$(\frac{1}{2}, \frac{2}{3})$	1.189×10^{-12}	4.293×10^{-15}	2.039×10^{-19}	4.673×10^{-19}	1.485×10^{-20}
$(0, 0)$	9.542×10^{-13}	1.885×10^{-16}	9.473×10^{-19}	2.037×10^{-20}	1.002×10^{-21}

Table 5.1: Pointwise error $|f(x_1, x_2) - \mathcal{F}_{N,k}[f](x_1, x_2)|$ for various values of (x_1, x_2) and N, where $k = 4$ and $f(x_1, x_2) = \left(\mathrm{e}^{3x_1} + \mathrm{e}^{-4x_1}\right)\left(\sin 5x_2 + \frac{1}{2}\right)$. Results to 4 significant digits.

5.7.3 Analysis of the auto-correction phenomenon and numerical results

We may now prove the key result of this section:

Theorem 5.29. *Suppose that $\mathcal{F}_{N,k}[f]$ is the multivariate Eckhoff approximation of the function $f \in \mathrm{H}^{3(k+1)}_{mix}(\Omega)$ based on the values $m(r) = N+r$, $r = 0, \ldots, k-1$. Then $f(x) - \mathcal{F}_{N,k}[f](x)$ is $\mathcal{O}\left(N^{-3k-2}\right)$ uniformly for x in compact subsets of Ω.*

Proof. Substituting the bounds derived in Lemmas 5.27 and 5.28 into the expansion (5.46) immediately yields the result. □

Though the analysis in this section was carried out for the approximation based on cardinal polynomials, it is a simple exercise to extend it to the general subtraction bases described in Section 5.2.1. Hence, we have established the existence of an auto-correction phenomenon for arbitrary dimension d and arbitrary subtraction basis $q_r^{[i]}$.[9]

For general values $m(r)$, it can be shown (by identical methods) that an auto-correction phenomenon is present, provided the first $l \leq k$ values are chosen so that $m(r) = N + r$, $r = 0, \ldots, l-1$. In this case, the convergence rate away from the boundary is $\mathcal{O}\left(N^{-2k-l-2}\right)$. In particular, if $m(0) = N$, as is the case with the choices (5.19) and (5.20), then the convergence rate is $\mathcal{O}\left(N^{-2k-3}\right)$.

The univariate auto-correction phenomenon was demonstrated numerically in Figure 5.2 of Section 5.3.2. For the particular choice of function and parameters, the error at the endpoints is roughly 10^{-8}. Conversely, in the interval $[-0.5, 0.5]$ this value is much smaller, approximately 10^{-12}. In Table 5.1, we present numerical results for the auto-correction phenomenon in the bivariate setting. Once more, we observe that the error inside the domain is much smaller than on the boundary. In particular, at the two points sampled inside the domain, machine epsilon is achieved with $k = 4$ and $N < 30$. Another bivariate example is considered in Figure 5.6. This figure also highlights that the convergence rate is slower on the whole of the boundary, not just at the corners, as may be expected.

Numerical results aside, the theory of the auto-correction phenomenon warrants further scrutiny. It is now apparent that Eckhoff's method offers a significantly faster pointwise convergence rate away from the boundary. Standard arguments demonstrate that this convergence rate is not uniform in $\bar{\Omega}$. For example, consider the univariate scenario. If $x_N = 1 - \frac{1}{N}$, then using the expression (5.49), the fact that $\mathrm{e}^{-i\pi x_N} = -1 - i\pi N^{-1} + \mathcal{O}\left(N^{-2}\right)$, and the estimates of Section 5.7.1, it follows that $f(x_N) - \mathcal{F}_{N,k}[f](x_N) = \mathcal{O}\left(N^{-2k-1}\right)$. This observation

[9] Incidentally, the auto-correction phenomenon is also exhibited by the error $f - \mathcal{F}_{N,k}[f]$ measured in the $\mathrm{L}^2(\Omega')$ norm, where Ω' is some set compactly contained in Ω. This has been studied in the univariate, Fourier case in [138]. The extension to the multivariate, modified Fourier setting is straightforward.

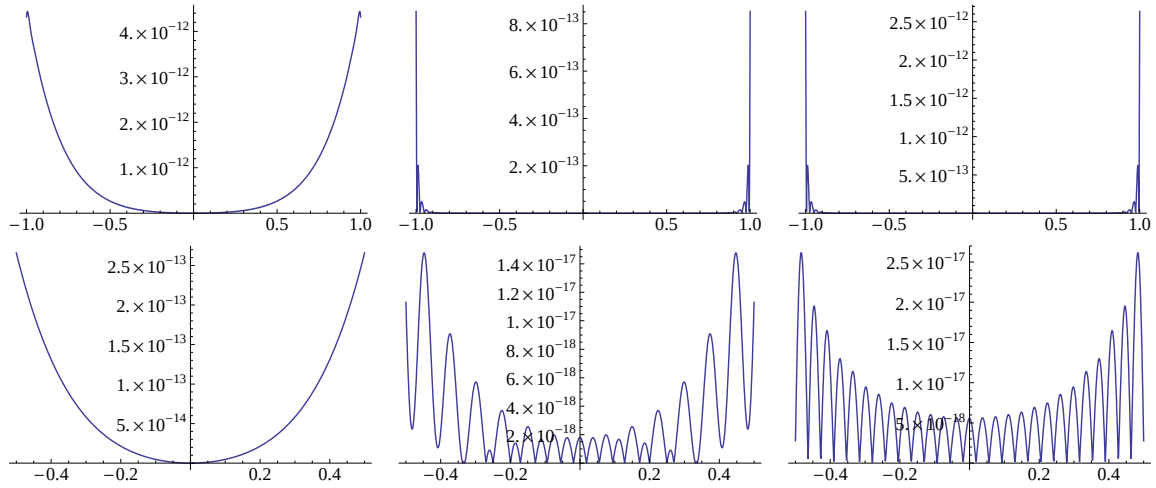


Figure 5.6: Absolute error $|f(x,y_0) - \mathcal{F}_{25,4}[f](x,y_0)|$, where $f(x_1,x_2) = x_1^2 \cosh 3x_1 \cos 2x_2 \sin 3x_2$, for $-1 \le x \le 1$ (top row) and $-\frac{1}{2} \le x \le \frac{1}{2}$ (bottom row) and $y_0 = 1, \frac{2}{3}, \frac{1}{3}$ (left to right).

not only verifies the nonuniformality of the auto-correction phenomenon, it also establishes quasi-optimality of the uniform error estimate of Theorem 5.18.

Since the auto-correction phenomenon concerns faster convergence inside the domain, it is worth examining whether or not this corresponds to a higher degree of pointwise convergence: in other words, whether higher-order derivatives of $\mathcal{F}_{N,k}[f]$ converge to the corresponding derivatives of f. Uniform convergence of any partial derivative with index $|\beta|_\infty \le 2k$ is guaranteed by Theorem 5.18. However, as we assess in the next section, pointwise convergence away from the boundary occurs for any derivative β of the increased order $|\beta|_\infty \le 3k+1$.

5.7.4 Degree of convergence of Eckhoff's approximation

We consider the univariate case. For the sake of simplicity we also assume that $f \in \mathrm{C}^\infty[-1,1]$ throughout this section. Simple adjustments can be made to address functions with lower regularity.

Recall that

$$f(x) - \mathcal{F}_{N,k}[f](x) = \mathrm{Re} \sum_{n \ge N} \hat{v}_n \mathrm{e}^{\mathrm{i}n\pi x} = \mathrm{Re}\,[H(x) + G(x)]\,,$$

where $H(x) = \mathrm{e}^{\mathrm{i}\pi(N-1)x}(1+\mathrm{e}^{-\mathrm{i}\pi x})^{-k} \triangle_{k,N}[\hat{v}_n]$ and

$$G(x) = (1+\mathrm{e}^{-\mathrm{i}\pi x})^{-(k+1)} \sum_{n \ge N} \triangle_{k+1,n}[\hat{v}_n] \mathrm{e}^{\mathrm{i}n\pi x}.$$

We wish to estimate the l^{th} derivative of $f(x) - \mathcal{F}_{N,k}[f](x)$ for $l \in \mathbb{N}_0$ and $x \in (-1,1)$. Trivially, using the bound for $\triangle_{k,N}[\hat{v}_n]$ derived in Lemma 5.24, we have

$$\left| H^{(l)}(x) \right| \lesssim N^{l-3k-2}, \quad \forall l \in \mathbb{N}_0. \tag{5.55}$$

We now turn our attention to $G(x)$. To attain a similar bound to (5.55), we seek an expression for G that does not involve an infinite sum. This is given by the following lemma:

Lemma 5.30. *The function $G(x)$ satisfies*

$$\operatorname{Re}\,[G(x)] = \begin{cases} \Phi(x)\left\{\mathcal{G}_{k+1}[v](x) - \mathcal{F}_{N+\frac{k+1}{2}}[\mathcal{G}_{k+1}[v]](x)\right\} & k \text{ odd} \\ \Phi(x)\left\{\mathcal{G}_{k+1}[v](x) - \tilde{\mathcal{F}}_{N+\frac{k}{2}+1}[\mathcal{G}_{k+1}[v]](x)\right\} & k \text{ even}, \end{cases}$$

where the function $\mathcal{G}_{k+1}[v](x) = 2^{-(k+1)}(\cos\frac{1}{2}\pi x)^{k+1}v(x)$, $\Phi(x) = (2\cos\frac{1}{2}\pi x)^{-(k+1)}$, and $\mathcal{F}_N$ and $\tilde{\mathcal{F}}_N$ are the Laplace–Neumann and Laplace–Dirichlet projection operators respectively.

Proof. Both cases are similar, so we assume that k is odd. In this setting, Lemma 5.21 gives that $\triangle_{k+1,n}[\hat{v}_n] = \widehat{\mathcal{G}_{k+1}[v]}_{n+\frac{k+1}{2}}$. Hence

$$\begin{aligned} G(x) &= (1+\mathrm{e}^{-\mathrm{i}\pi x})^{-(k+1)} \sum_{n\geq N} \triangle_{k+1,n}[\hat{v}_n]\mathrm{e}^{\mathrm{i}n\pi x} \\ &= (1+\mathrm{e}^{-\mathrm{i}\pi x})^{-(k+1)}\mathrm{e}^{-\mathrm{i}\frac{k+1}{2}\pi x} \sum_{n\geq N} \widehat{\mathcal{G}_{k+1}[v]}_{n+\frac{k+1}{2}}\mathrm{e}^{\mathrm{i}(n+\frac{k+1}{2})\pi x}. \end{aligned}$$

Since $(1+\mathrm{e}^{-\mathrm{i}\pi x})^{-(k+1)}\mathrm{e}^{-\mathrm{i}\frac{k+1}{2}\pi x} = (2\cos\frac{1}{2}\pi x)^{-(k+1)}$, it follows that

$$\operatorname{Re}\,[G(x)] = \big(2\cos\tfrac{1}{2}\pi x\big)^{-(k+1)} \sum_{n\geq N+\frac{k+1}{2}} \widehat{\mathcal{G}_{k+1}[v]}_n \cos n\pi x,$$

which completes the proof. □

We next require the following standard lemma:

Lemma 5.31. *Suppose that $h \in \mathrm{C}^\infty[-1,1]$ obeys the first k Neumann or Dirichlet derivative conditions. Then, for all $l \in \mathbb{N}_0$,*

$$\left|h^{(l)}(x) - (\mathcal{F}_N[h])^{(l)}(x)\right| \lesssim N^{l-2k-2}, \quad \left|h^{(l)}(x) - (\tilde{\mathcal{F}}_N[h])^{(l)}(x)\right| \lesssim N^{l-2k-1}, \quad x \in (-1,1),$$

respectively.

Proof. Both cases are identical, so we consider the former. If $l \leq 2k+1$, then the result follows immediately from Theorem 2.22. Now suppose that $l \geq 2k+2$. By assumption, $\hat{h}_n = \mathcal{A}_k[h](-1)^n(n\pi)^{-2k-2} + \mathcal{O}\left(n^{-2k-4}\right)$. Hence

$$\left|h^{(l)}(x) - (\tilde{\mathcal{F}}_N[h])^{(l)}(x)\right| \lesssim \left|h^{(l)}(x)\right| + \left|\mathcal{A}_k[h]\right| \left|\operatorname{Re}\sum_{n=1}^{N-1}(-1)^n(n\pi)^{l-2k-2}\mathrm{e}^{\mathrm{i}n\pi x}\right| + N^{l-2k-2}.$$

Now

$$\begin{aligned} \sum_{n=1}^{N-1}(-1)^n(n\pi)^{l-2k-2}\mathrm{e}^{\mathrm{i}n\pi x} &= \mathrm{i}^{2(k+1)-l}\frac{\mathrm{d}^{l-2k-2}}{\mathrm{d}x^{l-2k-2}}\sum_{n=1}^{N-1}(-\mathrm{e}^{\mathrm{i}\pi x})^n \\ &= \mathrm{i}^{2(k+1)-l}\frac{\mathrm{d}^{l-2k-2}}{\mathrm{d}x^{l-2k-2}}\left(\mathrm{e}^{\mathrm{i}\pi x}\frac{(-\mathrm{e}^{\mathrm{i}\pi x})^{N-1}-1}{\mathrm{e}^{\mathrm{i}\pi x}+1}\right), \end{aligned}$$

provided $x \in (-1,1)$. It is now readily seen that

$$\left| \sum_{n=1}^{N-1} (-1)^n (n\pi)^{l-2k-2} \mathrm{e}^{\mathrm{i}n\pi x} \right| \lesssim N^{l-2k-2},$$

thus completing the proof. □

We are now able to provide a bound for $G(x)$:

Lemma 5.32. *The function $G(x)$ satisfies $|G^{(l)}(x)| \lesssim N^{l-3k-2}$ for all $l \in \mathbb{N}_0$.*

Proof. We write

$$v(x) = f(x) - g_k(x) = \sum_{r=0}^{k-1} \left(\mathcal{A}_r[f] - \bar{\mathcal{A}}_r[f] \right) p_r(x) + \left[f(x) - g_k^e(x) \right],$$

where $g_k^e(x) = \sum_{r=0}^{k-1} \mathcal{A}_r[f] p_r(x)$ is the subtraction function based on exact jump values. This gives

$$\begin{aligned} G(x) = \Phi(x) \Bigg\{ & \sum_{r=0}^{k-1} \left(\mathcal{A}_r[f] - \bar{\mathcal{A}}_r[f] \right) \left(\mathcal{G}_{k+1}[p_r](x) - \mathcal{F}_N[\mathcal{G}_{k+1}[p_r]](x) \right) \\ & + \left(\mathcal{G}_{k+1}[f - g_k^e](x) - \mathcal{F}_N[\mathcal{G}_{k+1}[f - g_k^e]](x) \right) \Bigg\}, \end{aligned} \tag{5.56}$$

where $\mathcal{F}_N$ is either the Laplace–Dirichlet projector when k is even or the Laplace–Neumann projector otherwise. Since $f - g_k^e$ obeys the first k derivative conditions, the result now follows immediately from Theorems 5.1 and 5.5, as well as Lemmas 5.22 and 5.31. □

The key result of this section is an immediate consequence of this lemma:

Theorem 5.33. *Suppose that $f \in \mathrm{C}^\infty[-1,1]$ and that $\mathcal{F}_{N,k}[f]$ is the Eckhoff approximation of f based on the values $m(r) = N + r$, $r = 0, \ldots, k-1$. Then $(\mathcal{F}_{N,k}[f])^{(l)}$ converges uniformly to $f^{(l)}$ in compact subsets of $(-1,1)$ for $l = 0, \ldots, 3k+1$. Moreover, for all $l \in \mathbb{N}_0$, the error $f^{(l)}(x) - (\mathcal{F}_{N,k}[f])^{(l)}(x)$ is $\mathcal{O}\left(N^{l-3k-2}\right)$ for such x.*

This theorem establishes the existence of a higher degree of pointwise convergence of the univariate Eckhoff approximation. Using similar arguments to those given previously, we may also furnish the multivariate version of Eckhoff's method with an analogous result. We shall not do this. Instead, we state:

Theorem 5.34. *Suppose that $f \in \mathrm{C}^\infty(\bar{\Omega})$, where $\Omega = (-1,1)^d$, and that $\mathcal{F}_{N,k}[f]$ is the Eckhoff approximation of f based on the values $m(r) = N + r$, $r = 0, \ldots, k-1$. Then $\mathrm{D}^\beta \mathcal{F}_{N,k}[f]$ converges uniformly to $\mathrm{D}^\beta f$ in compact subsets of Ω for $|\beta|_\infty \leq 3k+1$. Moreover, for all $\beta \in \mathbb{N}_0^d$, the error $\mathrm{D}^\beta f(x) - \mathrm{D}^\beta \mathcal{F}_{N,k}[f](x)$ is $\mathcal{O}\left(N^{|\beta|_\infty - 3k - 2}\right)$ for such x.*

Aside from verifying a higher degree of pointwise convergence, these results demonstrate the existence of a *super-Gibbs phenomenon* for $(2k+1)^{\text{th}}$ order derivatives of Eckhoff's approximation. Consider the univariate setting. As demonstrated, the derivative $(\mathcal{F}_{N,k}[f])^{(2k+1)}$ does not converge uniformly to $f^{(2k+1)}$, yet, away from the endpoints, the error $f^{(2k+1)}(x) - (\mathcal{F}_{N,k}[f])^{(2k+1)}(x) = \mathcal{O}\left(N^{-k-1}\right)$. One facet of the standard Gibbs phenomenon is an $\mathcal{O}(1)$ uniform approximation error, but a pointwise convergence rate of $\mathcal{O}\left(N^{-1}\right)$ away from the endpoints. Evidently, for Eckhoff's approximation, this effect is far more pronounced. Figure 5.7 demonstrates this phenomenon.

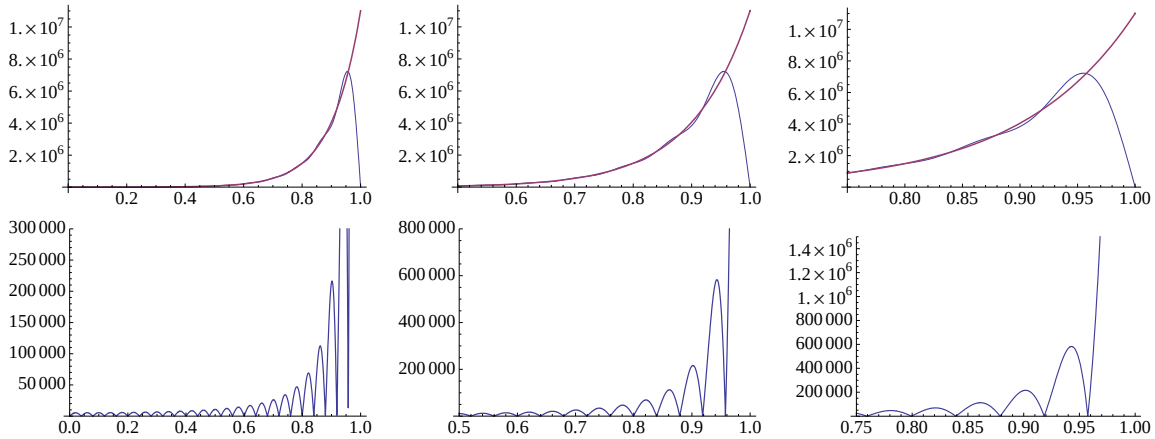


Figure 5.7: Top row: graphs of $f^{(3)}(x)$ and $(\mathcal{F}_{25,1}[f])^{(3)}(x)$ for $0 \leq x \leq 1$ (left), $\frac{1}{2} \leq x \leq 1$ (middle), and $\frac{3}{4} \leq x \leq 1$ (right), where $f(x) = x^2 \sin 5x + \cos 6x$. Bottom row: absolute error $|f^{(3)}(x) - (\mathcal{F}_{25,1}[f])^{(3)}(x)|$.

5.8 Eckhoff's method and the hyperbolic cross

As in Chapters 2–4, a vast reduction in the number of approximation terms (and, as a direct consequence, in the computational effort involved in forming the approximation) can be effectuated by replacing the full index set (2.33) in Eckhoff's approximation by the hyperbolic cross (2.41).[10] There are two aspects of this. First, we replace the index set used in the operator $\mathcal{F}_N[\cdot]$. However, to take full advantage of the hyperbolic cross, we also suitably amend the subtraction function g_k. Instead of (5.28), we employ the function

$$g_k(x) = \sum_{i \in \{0,1\}^d} \sum_{t \in [d]} \sum_{|r_t|_\infty = 0}^{k-1} \sum_{|n_{\bar{t}}|_0 = 0}^{N-1} \bar{\mathcal{A}}^{[i]}_{r_t, n_{\bar{t}}}[f] p^{[i_t]}_{r_t}(x_t) \phi^{[i_{\bar{t}}]}_{n_{\bar{t}}}(x_{\bar{t}}), \tag{5.57}$$

where the expression $|n|_0 = \bar{n}_1 \dots \bar{n}_d$ is as in Chapter 2.[11] The new function g_k satisfies the conditions

$$\widehat{g_k}^{[i]}_n = \hat{f}^{[i]}_n, \quad n \in M_k, \tag{5.58}$$

where M_k is the index set

$$M_k = \bigcup_{t \in [d]} \big\{ n = (n_1, \dots, n_d) \in \mathbb{N}^d : n_j = m(r_j),\ r_j = 0, \dots, k-1, j \in t,\ |n_{\bar{t}}|_0 < N \big\}.$$

The function g_k and index set M_k differ from their 'full' counterparts (5.28) and (5.30) only in the sense that $|n_{\bar{t}}|_\infty$ is replaced by $|n_{\bar{t}}|_0$. With these definitions to hand, we now construct

[10] Naturally, we could also consider the optimized hyperbolic cross (2.51) with similar results. However, for simplicity, we use (2.41) throughout.

[11] In this discussion, as well as subsequent analysis, we revert to full generality once more, neither assuming that the values $m(r) = N + r$ nor that the function f has only non-zero coefficients for one particular value of $i \in \{0,1\}^d$. Having said this, numerical results will be presented for the choice $m(r) = N + r$, in accordance with previous examples.

k	10^{-2}	10^{-4}	10^{-6}	10^{-8}	10^{-10}	10^{-12}	10^{-14}	10^{-16}
1	121	1521	31329	—	—	—	—	—
	89	513	3053	17461	97241	—	—	—
2	49	121	561	1849	10201	60025	—	—
	49	105	297	841	2269	6269	17501	48485
3	81	121	169	441	1225	3969	13689	47089
	81	117	193	353	697	1333	2773	5585
4	81	121	169	289	529	1089	2401	5929
	81	121	165	257	397	593	1005	1649
5	121	121	169	289	361	625	1089	2025
	121	121	169	273	329	493	789	1145

Table 5.2: Number of terms in the full (top value) and hyperbolic cross (bottom value) versions of Eckhoff's approximation required to obtain an accuracy of $\|f - \mathcal{F}_{N,k}[f]\|_\infty < 10^{-2j}$ for $j = 1, 2, \ldots, 8$ and $f(x_1, x_2) = e^{2x_1}(\cos 3x_2 + \sin 2x_2)$ (the dash indicates where in excess of $100,000$ terms are required to obtain the prescribed tolerance).

the approximation $\mathcal{F}_{N,k}[f]$ in the standard manner: $\mathcal{F}_{N,k}[f] = \mathcal{F}_N[f - g_k] + g_k$, where $\mathcal{F}_N[\cdot]$ is the modified Fourier projection operator based on the index set (2.41).

We remark in passing that, when $d = 2$, there is no difference between the subtraction functions (5.28) and (5.57). The only difference between the two resulting approximations arises from the index set used in $\mathcal{F}_N[\cdot]$. However, for $d \geq 3$, the functions (5.28) and (5.57) are distinct, leading to further savings in computational cost.

5.8.1 Cost reduction and numerical results

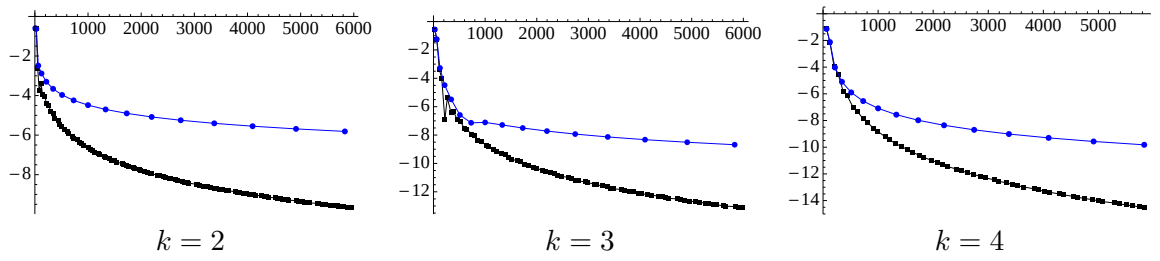
The operational cost of forming the hyperbolic cross version of Eckhoff's approximation is $\mathcal{O}\left(\max\{k^{d+1}, k^d N (\log N)^{d-1}\}\right)$. For $k \ll N$, this represents a significant reduction over the full index set version, where the corresponding figure, as mentioned in Section 5.5, is $\mathcal{O}\left(\max\{k^{d+1}, k^d N^d\}\right)$. No specific techniques are required either: as in the previous case, we repeatedly solve one-dimensional linear systems involving the matrix $V^{[i]}$ (or alternatively, compute $(V^{[i]})^{-1}$ and use (5.34)).

In Table 5.2, we highlight the improvement offered by the bivariate version of this method. For example, when $k = 3$, around $14,000$ terms are required to obtain an error of approximately 10^{-14} with the full index set Eckhoff method. The hyperbolic cross approximation obtains the same accuracy using only $2,800$ terms.

When $d = 3$, the improvement offered is more substantial. In Figure 5.8, we compare the error of the full and hyperbolic cross versions of Eckhoff's method applied to the function

$$f(x_1, x_2, x_3) = \left(x_1^2 \cos 5x_1 + \tfrac{46}{125} \sin 5 - \tfrac{4}{25} \cos 5\right) \times (\cosh 2x_2 - \cosh 1 \sinh 1) \left(x_3 \sin 2x_3 + \tfrac{1}{2} \cos 2 - \tfrac{1}{4} \sin 2\right). \tag{5.59}$$

For $k = 3$, using roughly $5,000$ terms, the hyperbolic cross version yields an error roughly 10^4 times smaller than the full version. For $k = 4$, the hyperbolic cross approximation obtains an error of 10^{-10} using only $1,500$ terms. The full index set approximation does not reach this value until the number of terms exceeds $6,000$.



$k = 2$ $k = 3$ $k = 4$

Figure 5.8: Log error $\log_{10} \|f - \mathcal{F}_{N,k}[f]\|_\infty$ against number of approximation terms for the full (circles) and hyperbolic cross (squares) versions of Eckhoff's method applied to (5.59).

The combination of Eckhoff's method and the hyperbolic cross yields highly accurate approximations comprising only a relatively small number of approximation coefficients. Figure 5.8 also demonstrates the advantage offered by this approach over the standard ($k = 0$) modified Fourier expansion. To obtain an accuracy of 10^{-10} with $k = 4$ requires less than 2000 terms, whereas to do the same with the original approximation $\mathcal{F}_N[f]$ would require in excess of 10^{12} terms—a completely infeasible value.

Key to the supremacy of the hyperbolic cross version over its full counterpart is that k remains small in comparison to N. As k grows, the relative improvement lessens. Asymptotically, at least, the hyperbolic cross method will always outperform the corresponding full index set approximation. However, for larger k, the function under consideration will often be very accurately resolved before the onset of this regime. This effect is demonstrated in Figure 5.8: when $k = 4$, as opposed to $k = 2$, the two graphs only begin to diverge once the error is much smaller. Nonetheless, since no additional effort is required to devise the hyperbolic cross version, its continued consideration is justified. However, it is only fair to warn the reader that the relative improvement may not be as substantial as expected.

5.8.2 Analysis of the hyperbolic cross version of Eckhoff's method

The framework introduced in Section 5.6 forms the basis for the analysis of the hyperbolic cross version of Eckhoff's method—a task we now pursue. As in previous cases, our intention is to demonstrate that there is no deterioration of the convergence rate over polynomial subtraction. To this end, we restate the following result, proved in Chapter 2:

Theorem 5.35. *Suppose that $f \in \mathrm{H}^{2k+2}_{mix}(\Omega)$ and that $\mathcal{F}^e_{N,k}[f]$ is the k^{th} exact polynomial subtraction approximation of f based on the hyperbolic cross (2.41). Then*

$$\|f - \mathcal{F}^e_{N,k}[f]\| \lesssim N^{-2k-\frac{3}{2}} (\log N)^{\frac{d-1}{2}},$$
$$\|f - \mathcal{F}^e_{N,k}[f]\|_r \lesssim N^{r-2k-\frac{3}{2}}, \quad r = 1, \ldots, 2k+1,$$

and $\|\mathrm{D}^\beta (f - \mathcal{F}^e_{N,k}[f])\|_\infty \lesssim N^{|\beta|_\infty - 2k - 1} (\log N)^{d-1}$ for $|\beta|_\infty \leq 2k$.

As in the full index set case (Section 5.4), it is possible to introduce the notion of approximate polynomial subtraction based on the hyperbolic cross. However, since the resulting approximation is not necessary to our subsequent analysis, we shall not pursue this further.

Our approach to analyse convergence is based on estimating the difference between the full and hyperbolic cross versions of Eckhoff's method. To this end, we write g_k, $\mathcal{F}_{N,k}[f]$ and

g_k^h, $\mathcal{F}_{N,k}^h[f]$ for the full and hyperbolic cross versions of Eckhoff's approximation respectively. Fundamental to this approach, and easily confirmed by a brief study of (5.31) and (5.58), is that the coefficients $\bar{\mathcal{A}}_{r_t,n_{\bar{t}}}^{[i]}[f]$ of g_k and g_k^h are identical. The sole difference between the two functions is that we employ only those coefficients with $|n_{\bar{t}}|_0 < N$ in the latter, whereas in the former, we use all coefficients with $|n_{\bar{t}}|_\infty < N$. We make use of this fact later. For now, we first notice that

$$\begin{aligned}\mathcal{F}_{N,k}[f] - \mathcal{F}_{N,k}^h[f] &= (g_k + \mathcal{F}_N[f - g_k]) - \left(g_k^h - \mathcal{F}_N^h[f - g_k^h]\right) \\ &= \left\{\mathcal{F}_N[f - g_k] - \mathcal{F}_N^h[f - g_k]\right\} + \left\{\left(g_k - g_k^h\right) - \mathcal{F}_N^h[g_k - g_k^h]\right\}, \end{aligned} \tag{5.60}$$

where $\mathcal{F}_N[\cdot]$ and $\mathcal{F}_N^h[\cdot]$ are the modified Fourier projection operators based on the full and hyperbolic cross index sets respectively. It therefore suffices to analyse each bracket separately, a task we shall now perform.

An estimate for $\mathcal{F}_N[f - g_k] - \mathcal{F}_N^h[f - g_k]$

Since

$$\mathcal{F}_N[f - g_k](x) - \mathcal{F}_N^h[f - g_k](x) = \sum_{i \in \{0,1\}^d} \sum_{\substack{|n|_\infty < N \\ |n|_0 \geq N}} \left(\hat{f}_n^{[i]} - \widehat{g_k}_n^{[i]}\right) \phi_n^{[i]}(x), \tag{5.61}$$

we first seek an expression for $\hat{f}_n^{[i]} - \widehat{g_k}_n^{[i]}$. By definition, for $|n|_\infty < N$, we have

$$\hat{f}_n^{[i]} - \widehat{g_k}_n^{[i]} = \sum_{t \in [d]} \sum_{|r_t|_\infty = 0}^{k-1} \left(\mathcal{A}_{r_t,n_{\bar{t}}}^{[i]}[f] - \bar{\mathcal{A}}_{r_t,n_{\bar{t}}}^{[i]}[f]\right) \widehat{p_{r_t}}_{n_t}^{[i_t]} + \mathcal{A}_n^{[i]}[f], \tag{5.62}$$

where $\mathcal{A}_n^{[i]}[f]$ is the coefficient $\mathcal{A}_{r_t,n_{\bar{t}}}^{[i]}[f]$ corresponding to the tuple $t = \emptyset$. In other words,

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

Our task now is to derive an expression for $\hat{f}_n^{[i]} - \widehat{g_k}_n^{[i]}$, based on (5.62), that does not involve the values $\bar{\mathcal{A}}_{r_t,n_{\bar{t}}}^{[i]}[f]$. To do so, we first obtain an expression for such values in terms of modified Fourier coefficients of f only. This is provided by the following lemma:

Lemma 5.36. *The values* $\bar{\mathcal{A}}_{r_t,n_{\bar{t}}}^{[i]}[f]$ *satisfy*

$$\bar{\mathcal{A}}_{r_t,n_{\bar{t}}}^{[i]}[f] = \sum_{\substack{u \in [d] \\ t \subseteq u}} (-1)^{|t|+|u|} \sum_{|r_{u\setminus t}|_\infty = 0}^{k-1} \sum_{|s_u|_\infty = 0}^{k-1} (V^{[i_u]})_{r_u,s_u}^{-1} \hat{f}_{(m(s_u);n_{\bar{u}})}^{[i]} \widehat{p_{r_{u\setminus t}}}_{n_{u\setminus t}}^{[i_{u\setminus t}]}, \tag{5.63}$$

where $(V^{[i_u]})_{r_u,s_u}^{-1} = \prod_{j \in u} (V^{[i_j]})_{r_j,s_j}^{-1}$ *and* $(m(s_u); n_{\bar{u}})$ *has* j^{th} *entry* $m(s_u)$ *if* $j \in u$ *and* n_u *otherwise.*

Proof. Trivial calculations based on (5.31) and (5.33) verify the result for $|t| = d$. Now suppose that (5.63) is true for all tuples of length at least $l+1$, and let $t \in [d]$ with $|t| = l$. Then, in view of (5.31) and (5.33), we have

$$\bar{\mathcal{A}}^{[i]}_{r_t,n_{\bar{t}}}[f] + \sum_{\substack{u\in[d]\\ t\subseteq u\\ t\neq u}} \sum_{|r_{u\setminus t}|_\infty=0}^{k-1} \bar{\mathcal{A}}^{[i]}_{r_u,n_{\bar{u}}}[f]\widehat{p_{r_{u\setminus t}}}^{[i_{u\setminus t}]}_{n_{u\setminus t}} = \sum_{|s_t|_\infty=0}^{k-1} (V^{[i_t]})^{-1}_{r_t,s_t}\hat{f}^{[i]}_{(m(s_t);n_{\bar{t}})}.$$

Since $|u| \geq |t|+1$, we may substitute (5.63) into this formula and simplify, to give

$$\begin{aligned}\bar{\mathcal{A}}^{[i]}_{r_t,n_{\bar{t}}}[f] = &\sum_{|s_t|_\infty=0}^{k-1} (V^{[i_t]})^{-1}_{r_t,s_t}\hat{f}^{[i]}_{(m(s_t);n_{\bar{t}})}\\ &- \sum_{\substack{u\in[d]\\ t\subseteq u\\ t\neq u}} \sum_{\substack{v\in[d]\\ u\subseteq v}} (-1)^{|u|+|v|} \sum_{|r_{v\setminus t}|_\infty=0}^{k-1} \sum_{|s_v|_\infty=0}^{k-1} (V^{[i_v]})^{-1}_{r_v,s_v}\widehat{f}^{[i]}_{(m(s_v);n_{\bar{v}})}\widehat{p_{r_{v\setminus t}}}^{[i_{v\setminus t}]}_{n_{v\setminus t}}.\end{aligned}$$

After carefully rearranging the order of the outer two summations of the second term, we obtain

$$\begin{aligned}\bar{\mathcal{A}}^{[i]}_{r_t,n_{\bar{t}}}[f] = &\sum_{|s_t|_\infty=0}^{k-1} (V^{[i_t]})^{-1}_{r_t,s_t}\hat{f}^{[i]}_{(m(s_t);n_{\bar{t}})}\\ &- \sum_{\substack{v\in[d]\\ t\subseteq v\\ t\neq v}} (-1)^{|v|} \sum_{|r_{v\setminus t}|_\infty=0}^{k-1} \sum_{|s_v|_\infty=0}^{k-1} (V^{[i_v]})^{-1}_{r_v,s_v}\widehat{f}^{[i]}_{(m(s_v);n_{\bar{v}})}\widehat{p_{r_{v\setminus t}}}^{[i_{v\setminus t}]}_{n_{v\setminus t}} \sum_{\substack{u\in[d]\\ t\subseteq u\subseteq v\\ u\neq t}} (-1)^{|u|}.\end{aligned} \tag{5.64}$$

We now wish to determine the value of the final sum for each such v. If $|v| = |t| + l$, then it is readily seen that there are $\binom{l}{j}$ choices of u with $|u| = |t| + j$. Hence,

$$\sum_{\substack{u\in[d]\\ t\subseteq u\subseteq v\\ u\neq t}} (-1)^{|u|} = \sum_{j=1}^{l} \binom{l}{j} (-1)^{|t|+j} = (-1)^{|t|+1}$$

Upon substitution of this into (5.64), we obtain the result (observe that the first term of (5.64) corresponds to $v = t$). □

Note that this lemma establishes the previously stated formula (5.34). Not only is this formula vital to the current analysis, it also presents a useful means by which to calculate the coefficients of Eckhoff's approximation, as mentioned in Section 5.5.

With this to hand, we may now provide a formula for $\hat{f}^{[i]}_n - \widehat{g_k}^{[i]}_n$:

Lemma 5.37. *For $|n|_\infty < N$ we have*

$$\hat{f}^{[i]}_n - \widehat{g_k}^{[i]}_n = \sum_{t\in[d]^*} \sum_{|r_t|_\infty=0}^{k-1} \sum_{|s_t|_\infty=0}^{k-1} (-1)^{|t|} (V^{[i_t]})^{-1}_{r_t,s_t} \mathcal{A}^{[i]}_{(m(s_t);n_{\bar{t}})}[f]\widehat{p_{r_t}}^{[i_t]}_{n_t}. \tag{5.65}$$

Proof. Suppose that we write $f - g_k = (g_k^a - g_k) + (f - g_k^a)$, where g_k^a is the approximate polynomial subtraction function for f based on the full index set. Then $g_k^a - g_k$ is a subtraction function of the form (5.28), with coefficients $\tilde{\mathcal{A}}^{[i]}_{r_t,n_{\bar{t}}}[f] = \mathcal{A}^{[i]}_{r_t,n_{\bar{t}}}[f] - \bar{\mathcal{A}}^{[i]}_{r_t,n_{\bar{t}}}[f]$. Since the n^{th} modified Fourier coefficient of $f - g_k^a$ is $\mathcal{A}_n^{[i]}[f]$, the coefficients $\tilde{\mathcal{A}}^{[i]}_{r_t,n_{\bar{t}}}[f]$ satisfy the relation (5.31) with right-hand side $-\mathcal{A}_n^{[i]}[f]$. It now follows from Lemma 5.36 that

$$\tilde{\mathcal{A}}^{[i]}_{r_t,n_{\bar{t}}}[f] = \sum_{\substack{u \in [d] \\ t \subseteq u}} (-1)^{|t|+|u|+1} \sum_{|r_{u\setminus t}|_\infty = 0}^{k-1} \sum_{|s_u|_\infty = 0}^{k-1} (V^{[i_u]})^{-1}_{r_u,s_u} \mathcal{A}^{[i]}_{(m(s_u);n_{\bar{u}})} \widehat{p_{r_{u\setminus t}}}^{\,[i_{u\setminus t}]}_{n_{u\setminus t}}.$$

Substituting this into (5.62) gives

$$\begin{aligned}
\hat{f}_n^{[i]} - \widehat{g_k}_n^{[i]} &= \mathcal{A}_n^{[i]}[f] + \sum_{t \in [d]} \sum_{\substack{u \in [d] \\ t \subseteq u}} \sum_{|r_u|_\infty = 0}^{k-1} \sum_{|s_u|_\infty = 0}^{k-1} (-1)^{|t|+|u|+1} (V^{[i_u]})^{-1}_{r_u,s_u} \mathcal{A}^{[i]}_{(m(s_u);n_{\bar{u}})} \widehat{p_{r_u}}^{\,[i_u]}_{n_u} \\
&= \mathcal{A}_n^{[i]}[f] + \sum_{u \in [d]} (-1)^{|u|+1} \left\{ \sum_{t \in [u]} (-1)^{|t|} \right\} \sum_{|r_u|_\infty = 0}^{k-1} \sum_{|s_u|_\infty = 0}^{k-1} (V^{[i_u]})^{-1}_{r_u,s_u} \mathcal{A}^{[i]}_{(m(s_u);n_{\bar{u}})} \widehat{p_{r_u}}^{\,[i_u]}_{n_u}.
\end{aligned}$$

Evaluating this sum now gives the result. □

Having derived an expression for $\hat{f}_n^{[i]} - \widehat{g_k}_n^{[i]}$, we now seek a bound for its absolute value:

Lemma 5.38. *Suppose that $f \in \mathrm{H}^{2k+l+1}_{mix}(\Omega)$, where l is the number of equal values $c(r)$. Then the coefficients $\widehat{g_k}_n^{[i]}$ satisfy*

$$\left| \hat{f}_n^{[i]} - \widehat{g_k}_n^{[i]} \right| \lesssim \sum_{t \in [d]^*} \sum_{|r_t|_\infty = 0}^{k-1} N^{2(|r_t|-k|t|)} \bar{n}_t^{-2r_t-2} \bar{n}_{\bar{t}}^{-2k-2}.$$

Proof. Recall that $\mathcal{A}_n^{[i]}[f]$ is the modified Fourier coefficient of the function $f - g_k^e$, where g_k^e is the exact polynomial subtraction function for f. Since $f - g_k^e$ satisfies the first k derivative conditions, the result follows from (5.65) and Lemma 5.12. □

This bound, upon substitution into (5.61), now provides the key result of this section:

Lemma 5.39. *Suppose that f is as in Lemma 5.38. Then*

$$\begin{aligned}
\|\mathcal{F}_N[f - g_k] - \mathcal{F}_N^h[f - g_k]\| &\lesssim N^{-2k-\frac{3}{2}} (\log N)^{\frac{d-1}{2}}, \\
\|\mathcal{F}_N[f - g_k] - \mathcal{F}_N^h[f - g_k]\|_r &\lesssim N^{r-2k-\frac{3}{2}}, \quad r = 1, \dots, 2k+1,
\end{aligned}$$

and $\|\mathrm{D}^\beta(\mathcal{F}_N[f - g_k] - \mathcal{F}_N^h[f - g_k])\|_\infty \lesssim N^{|\beta|_\infty - 2k-1} (\log N)^{d-1}$ for $|\beta|_\infty \leq 2k$.

Proof. Since all cases are similar, we consider only the uniform norm. Using (5.61) and the bound derived in Lemma 5.38, we obtain

$$\|\mathrm{D}^\beta(\mathcal{F}_N[f - g_k] - \mathcal{F}_N^h[f - g_k])\|_\infty \lesssim \sum_{t \in [d]^*} \sum_{|r_t|_\infty = 0}^{k-1} N^{2(|r_t|-k|t|)} \sum_{\substack{|n|_\infty < N \\ |n|_0 \geq N}} \bar{n}_t^{\beta_t - 2r_t - 2} \bar{n}_{\bar{t}}^{\beta_{\bar{t}} - 2k - 2}. \tag{5.66}$$

Consider the inner sum. For fixed $t \in [d]^*$, suppose that we define the tuple $u \in [t]^*$ by the criterion that $j \in u$ provided $\beta_j - 2r_j - 2 \geq -1$. Then

$$\begin{aligned}\sum_{\substack{|n|_\infty<N\\ |n|_0\geq N}} \bar{n}_t^{\beta_t-2r_t-2}\bar{n}_{\bar{t}}^{\beta_{\bar{t}}-2k-2} &\leq \sum_{|n_u|_\infty<N} \bar{n}_u^{\beta_u-2r_t-2} \sum_{\substack{|n_{\bar{u}}|_\infty<N\\ |n_{\bar{u}}|_0\geq N}} n_{\bar{u}}^{-2}\\ &\lesssim N^{-1} \sum_{|n_u|_\infty<N} \bar{n}_u^{\beta_u-2r_t-2} = N^{-1}\prod_{j\in u}\sum_{n_j=0}^{N} \bar{n}_j^{\beta_j-2r_j-2}\\ &\lesssim N^{|\beta_u|-2|r_u|-|u|-1}(\log N)^{d-1}.\end{aligned}$$

Substituting this into (5.66), we obtain

$$\|\mathrm{D}^\beta(\mathcal{F}_N[f-g_k]-\mathcal{F}_N^h[f-g_k])\|_\infty \lesssim (\log N)^{d-1}\sum_{t\in[d]^*}\sum_{|r_t|_\infty=0}^{k} N^{2(|r_t|-k|t|)+|\beta_u|-2|r_u|-|u|-1},$$

where, for each $t \in [d]^*$, u is as defined previously. Now, since $|\beta_u| \leq |u||\beta|_\infty \leq 2k(|u|-1)+|\beta|_\infty$, we have

$$\begin{aligned}2(|r_t|-k|t|)+|\beta_u|-2|r_u|-|u|-1 &\leq 2|r_{t\setminus u}|+2k(|u|-|t|-1)+|\beta|_\infty-|u|-1\\ &\leq 2k(|t|-|u|)+2k(|u|-|t|-1)+|\beta|_\infty-|u|-1\\ &\leq -2k+|\beta|_\infty-1.\end{aligned}$$

This completes the proof. □

We next progress to the second term of (5.60).

An estimate for $\left(g_k - g_k^h\right) - \mathcal{F}_N^h[g_k - g_k^h]$

Upon recalling that the coefficients $\bar{\mathcal{A}}_{r_t,n_{\bar{t}}}^{[i]}[f]$ of g_k and g_k^h are identical, we notice that

$$g_k(x)-g_k^h(x) = \sum_{i\in\{0,1\}^d}\sum_{\substack{t\in[d]\\ |t|<d}}\sum_{|r_t|_\infty=0}^{k-1}\sum_{\substack{|n_{\bar{t}}|_0\geq N\\ |n_{\bar{t}}|_\infty<N}} \bar{\mathcal{A}}_{r_t,n_{\bar{t}}}^{[i]}[f]p_{r_t}^{[i_t]}(x_t)\phi_{n_{\bar{t}}}^{[i_{\bar{t}}]}(x_{\bar{t}}). \tag{5.67}$$

From this, we immediately deduce that $\mathcal{F}_N^h[g_k - g_k^h] \equiv 0$. Hence, it suffices to estimate (5.67). To provide such an estimate, we first need a bound for $\bar{\mathcal{A}}_{r_t,n_{\bar{t}}}^{[i]}[f]$. To establish such a bound, we require some additional notation. We define the operator $\mathcal{C}_n^{[i]}[\cdot] : \mathrm{L}^2(-1,1) \to \mathbb{R}$, $i \in \{0,1\}$, $n \in \mathbb{N}$, by

$$\mathcal{C}_n^{[i]}[g] = \hat{g}_n^{[i]} - \sum_{r=0}^{k-1}\widehat{p}_r{}_n^{[i]}\sum_{s=0}^{k-1}(V^{[i]})_{r,s}^{-1}\hat{g}_{m(s)}^{[i]} = \int_{-1}^{1} g(x)\left\{\phi_n^{[i]}(x) - \sum_{r=0}^{k-1}\widehat{p}_r{}_n^{[i]}\sum_{s=0}^{k-1}(V^{[i]})_{r,s}^{-1}\phi_{m(s)}^{[i]}(x)\right\}\mathrm{d}x.$$

If $g \in \mathrm{L}^2(-1,1)^d$, write $\mathcal{C}_{n_j}^{[i_j]}[g]$ for the above operator acting on the j^{th} entry of g. Furthermore, given $t \in [d]$, we define

$$\begin{aligned}\mathcal{C}_{n_t}^{[i_t]}[g](x_{\bar{t}}) &= \mathcal{C}_{n_{t_1}}^{[i_{t_1}]}\left[\dots\left[\mathcal{C}_{n_{t_{|t|}}}^{[i_{t_{|t|}}]}[g]\right]\dots\right](x_{\bar{t}}) \\ &= \int_{-1}^{1}\dots\int_{-1}^{1} g(x)\prod_{j\in t}\left\{\phi_{n_j}^{[i_j]}(x_j) - \sum_{r_j=0}^{k-1}\widehat{p_{r_j}}_{n_j}^{[i_j]}\sum_{s_j=0}^{k-1}(V^{[i_j]})^{-1}_{r_j,s_j}\phi_{m(s_j)}^{[i_j]}(x_j)\right\}\mathrm{d}x_t,\end{aligned}$$

in the standard manner. Note that $\mathcal{C}_{n_t}^{[i_t]}[g]$ is a function of $x_{\bar{t}}$. Finally, we let

$$\mathcal{D}_{r_t,n_{\bar{t}}}^{[i]}[g] = \widehat{\mathcal{C}_{n_{\bar{t}}}^{[i_{\bar{t}}]}[g]}_{m(r_t)}^{[i_t]}, \quad \text{where } m(r_t) = (m(r_{t_1}),\dots,m(r_{t_{|t|}})).$$

This notation permits the following succinct expression for $\bar{\mathcal{A}}_{r_t,n_{\bar{t}}}^{[i]}[f]$:

Lemma 5.40. *The value* $\bar{\mathcal{A}}_{r_t,n_{\bar{t}}}^{[i]}[f]$ *satisfies*

$$\bar{\mathcal{A}}_{r_t,n_{\bar{t}}}^{[i]}[f] = \sum_{|s_t|_\infty=0}^{k-1}(V^{[i_t]})^{-1}_{r_t,s_t}\mathcal{D}_{s_t,n_{\bar{t}}}^{[i]}[f].$$

Proof. Given a_j, b_j, $j \in \bar{t}$, we observe that

$$\prod_{j\in\bar{t}}(a_j+b_j) = \sum_{u\in[\bar{t}]^*}\prod_{j\in u}b_j\prod_{j\in\bar{u}}a_j = \sum_{\substack{u\in[d]\\ t\subseteq u}}\prod_{j\in u\setminus t}b_j\prod_{j\notin u}a_j.$$

Suppose now that a_j and b_j are the operators $\mathrm{L}^2(-1,1)^d \to \mathrm{L}^2(-1,1)^{d-1}$ defined by

$$g \longmapsto \int_{-1}^{1} g(x)\phi_{n_j}^{[i_j]}(x_j)\,\mathrm{d}x_j, \quad g \longmapsto -\sum_{r_j=0}^{k-1}\widehat{p_{r_j}}_{n_j}^{[i_j]}\sum_{s_j=0}^{k-1}(V^{[i_j]})^{-1}_{r_j,s_j}\int_{-1}^{1} g(x)\phi_{m(s_j)}^{[i_j]}(x_j)\,\mathrm{d}x_j,$$

respectively. Then, using the above expression, we obtain the formula

$$\begin{aligned}\mathcal{C}_{n_{\bar{t}}}^{[i_{\bar{t}}]}[f] = \sum_{\substack{u\in[d]\\ t\subseteq u}}(-1)^{|t|+|u|}\sum_{|r_{u\setminus t}|_\infty=0}^{k-1}\sum_{|s_{u\setminus t}|_\infty=0}^{k-1}(V^{[i_{u\setminus t}]})^{-1}_{r_{u\setminus t},s_{u\setminus t}}\widehat{p_{r_{u\setminus t}}}_{n_{u\setminus t}}^{[i_{u\setminus t}]} \\ \times\int_{-1}^{1}\dots\int_{-1}^{1} f(x)\phi_{(m(s_{u\setminus t});n_{\bar{u}})}^{[i_{\bar{t}}]}(x_{\bar{t}})\,\mathrm{d}x_{\bar{t}}.\end{aligned}$$

Hence

$$\mathcal{D}_{s_t,n_{\bar{t}}}^{[i]}[f] = \sum_{\substack{u\in[d]\\ t\subseteq u}}(-1)^{|t|+|u|}\sum_{|r_{u\setminus t}|_\infty=0}^{k-1}\sum_{|s_{u\setminus t}|_\infty=0}^{k-1}(V^{[i_{u\setminus t}]})^{-1}_{r_{u\setminus t},s_{u\setminus t}}\widehat{p_{r_{u\setminus t}}}_{n_{u\setminus t}}^{[i_{u\setminus t}]}\hat{f}_{(m(s_u);n_{\bar{u}})}^{[i]},$$

and therefore

$$\sum_{|s_t|_\infty=0}^{k-1}(V^{[i_t]})^{-1}_{r_t,s_t}\mathcal{D}_{s_t,n_{\bar{t}}}^{[i]}[f] = \sum_{\substack{u\in[d]\\ t\subseteq u}}(-1)^{|t|+|u|}\sum_{|r_{u\setminus t}|_\infty=0}^{k-1}\sum_{|s_u|_\infty=0}^{k-1}(V^{[i_u]})^{-1}_{r_u,s_u}\hat{f}_{(m(s_u);n_{\bar{u}})}^{[i]}\widehat{p_{r_{u\setminus t}}}_{n_{u\setminus t}}^{[i_{u\setminus t}]}.$$

Comparing this with the result of Lemma 5.36 now completes the proof. □

The operator $\mathcal{C}_{n_t}^{[i_t]}[g]$ possesses the following property, central to our analysis of $\bar{\mathcal{A}}_{r_t,n_{\bar{t}}}^{[i]}[f]$.

Lemma 5.41. *Suppose that $t \in [d]^*$, $f \in \mathrm{H}_{mix}^{2k+l+1}(\Omega)$, where l is as in Lemma 5.38, and that $g_{k,t}^e$ is the k^{th} subtraction function of f in the variables x_j where $j \in t$. In other words,*

$$\mathcal{B}_{r_j}^{[i_j]}[f] = \mathcal{B}_{r_j}^{[i_j]}\left[g_{k,t}^e\right], \quad r_j = 0, \ldots, k-1, \quad j \in t.$$

Then $\mathcal{C}_{n_t}^{[i_t]}[f] = \mathcal{C}_{n_t}^{[i_t]}\left[f - g_{k,t}^e\right]$.

Proof. Consider the univariate case first. The operator $\mathcal{C}_n^{[i]}[f]$ is precisely the difference between the n^{th} modified Fourier coefficient of f and the coefficient of the corresponding Eckhoff subtraction function g_k. It is easily verified that Eckhoff's method is exact for any function of the same form as g_k. In particular, it is exact for g_k^e. Hence, $\mathcal{C}_n^{[i]}[g_k^e] = 0$, and the result for $d = 1$ follows from linearity. The extension to $d \geq 2$ is attained by applying the univariate result in each variable. □

We are now in a position to derive a bound for $\bar{\mathcal{A}}_{r_t,n_{\bar{t}}}^{[i]}[f]$:

Lemma 5.42. *Suppose that $g \in \mathrm{H}_{mix}^{2k+l+1}(\Omega)$, where l is as in Lemma 5.38. Then*

$$\left|\bar{\mathcal{A}}_{r_t,n_{\bar{t}}}^{[i]}[f]\right| \lesssim \sum_{\substack{u \in [d] \\ t \subseteq u}} \sum_{|r_{u\setminus t}|_\infty = 0}^{k-1} N^{2(|r_{u\setminus t}| - k|u\setminus t|)} \bar{n}_{u\setminus t}^{-2r_{u\setminus t}-2} \bar{n}_{\bar{u}}^{-2k-2}. \tag{5.68}$$

Proof. Using Lemmas 5.40 and 5.41, we may write

$$\bar{\mathcal{A}}_{r_t,n_{\bar{t}}}^{[i]}[f] = \sum_{|s_t|_\infty = 0}^{k-1} (V_{r_t,s_t}^{[i_t]})^{-1} \mathcal{D}_{s_t,n_{\bar{t}}}^{[i]}\left[f - g_{k,\bar{t}}^e\right].$$

Hence, by reversing the arguments of Lemma 5.40, we obtain

$$\bar{\mathcal{A}}_{r_t,n_{\bar{t}}}^{[i]}[f] = \sum_{\substack{u \in [d] \\ t \subseteq u}} (-1)^{|t|+|u|} \sum_{|s_u|_\infty = 0}^{k-1} \sum_{|r_{u\setminus t}|_\infty = 0}^{k-1} (V_{r_u,s_u}^{[i_u]})^{-1} \widehat{(f - g_{k,\bar{t}}^e)}_{(m(s_u);n_{\bar{u}})}^{[i]} \widehat{p_{r_{u\setminus t}}}_{n_{u\setminus t}}^{[i_{u\setminus t}]}.$$

Since $f - g_{k,\bar{t}}^e$ obeys the first k derivative conditions in the variables $x_{\bar{t}}$, an application of Lemma 5.12 now gives

$$\left|\bar{\mathcal{A}}_{r_t,n_{\bar{t}}}^{[i]}[f]\right| \lesssim \sum_{\substack{u \in [d] \\ t \subseteq u}} \sum_{|r_{u\setminus t}|_\infty = 0}^{k-1} N^{2(|r_{u\setminus t}| - k|u\setminus t|)} \bar{n}_{u\setminus t}^{-2r_{u\setminus t}-2} \bar{n}_{\bar{u}}^{-2k-2},$$

as required. □

With this in hand, we now deduce the key result:

Lemma 5.43. *Suppose that f is as in Lemma 5.38. Then*

$$\|(g_k - g_k^h) - \mathcal{F}_N^h[g_k - g_k^h]\| \lesssim N^{-2k-\frac{3}{2}}(\log N)^{\frac{d-1}{2}},$$
$$\|(g_k - g_k^h) - \mathcal{F}_N^h[g_k - g_k^h]\|_r \lesssim N^{r-2k-\frac{3}{2}}, \quad r = 1, \dots, 2k+1,$$

and $\|\mathrm{D}^\beta\left(g_k - g_k^h\right) - \mathrm{D}^\beta \mathcal{F}_N^h[g_k - g_k^h]\|_\infty \lesssim N^{|\beta|_\infty - 2k-1}(\log N)^{d-1}$ *for* $|\beta|_\infty \le 2k$.

Proof. Once more, we consider only the uniform norm. Using (5.67) and the bound derived in Lemma 5.42, we deduce that

$$\begin{aligned}
&\|\mathrm{D}^\beta\left(g_k - g_k^h\right) - \mathrm{D}^\beta \mathcal{F}_N^h[g_k - g_k^h]\|_\infty \\
&\qquad \lesssim \sum_{\substack{t\in[d]\\|t|<d}} \sum_{\substack{u\in[d]\\t\subseteq u}} \sum_{|r_{u\setminus t}|_\infty=0}^{k-1} \sum_{\substack{|n_{\bar{t}}|_0\ge N\\|n_{\bar{t}}|_\infty<N}} N^{2(|r_{u\setminus t}|-k|u\setminus t|)} \bar{n}_{u\setminus t}^{\beta_{u\setminus t}-2r_{u\setminus t}-2} \bar{n}_{\bar{u}}^{\beta_{\bar{u}}-2k-2} \\
&\qquad = \sum_{\substack{t\in[d]\\|t|<d}} \sum_{\substack{|n_{\bar{t}}|_0\ge N\\|n_{\bar{t}}|_\infty<N}} \bar{n}_{\bar{t}}^{\beta_{\bar{t}}-2k-2} \\
&\qquad\quad + \sum_{\substack{t\in[d]\\|t|<d}} \sum_{\substack{u\in[d]\\t\subseteq u\\t\ne u}} \sum_{|r_{u\setminus t}|_\infty=0}^{k-1} N^{2(|r_{u\setminus t}|-k|u\setminus t|)} \sum_{\substack{|n_{\bar{t}}|_0\ge N\\|n_{\bar{t}}|_\infty<N}} \bar{n}_{u\setminus t}^{\beta_{u\setminus t}-2r_{u\setminus t}-2} \bar{n}_{\bar{u}}^{\beta_{\bar{u}}-2k-2}.
\end{aligned} \tag{5.69}$$

Since $|\beta|_\infty \le 2k$ and $|t| \ge 1$, the first term is bounded by $N^{|\beta|_\infty-2k-1}(\log N)^{d-1}$ (recall Lemma 2.30 regarding sums of inverse powers of n for values of n lying outside a hyperbolic cross). Hence, it suffices to establish the same bound for the second term. If v is the tuple $t \subseteq v \subseteq u$ such that $j \in v$ if either $j \in t$ or $\beta_j - 2r_j - 2 \ge -1$, then, much as in Lemma 5.39, the second term of (5.69) is bounded by

$$\sum_{\substack{t\in[d]\\|t|<d}} \sum_{\substack{u\in[d]\\t\subseteq u\\t\ne u}} \sum_{|r_{u\setminus t}|_\infty=0}^{k-1} N^{2(|r_{u\setminus t}|-k|u\setminus t|)} N^{|\beta_{v\setminus t}|-2|r_{v\setminus t}|-|v\setminus t|-|\bar{v}|}(\log N)^{d-1}. \tag{5.70}$$

Consider the exponent of N. Since $|\beta|_\infty \le 2k$, we have

$$\begin{aligned}
2(|r_{u\setminus t}| - k|u\setminus t|) + |\beta_{v\setminus t}| - 2|r_{v\setminus t}| - |v\setminus t| - |\bar{v}| \\
&= 2(|r_{u\setminus v}| - k|u\setminus t|) + |\beta_{v\setminus t}| - |\bar{t}| \\
&\le 2|r_{u\setminus v}| - 2k(|u| - |t|) + |\beta|_\infty(|v| - |t| - 1) + |\beta|_\infty - |\bar{t}| \\
&\le 2|r_{u\setminus v}| + 2k(|v| - |u|) + |\beta|_\infty - 2k - |\bar{t}| \\
&\le |\beta|_\infty - 2k - 1,
\end{aligned}$$

where the final inequality follows from the fact that $|r_{u\setminus v}|_\infty \le k-1$ and $|t| < d$. Substituting this into (5.70) now completes the proof. □

With Lemmas 5.39 and 5.43 to hand, we are now able to provide error estimates for the hyperbolic cross version of Eckhoff's approximation. In view of the decomposition (5.60), such estimates follow immediately:

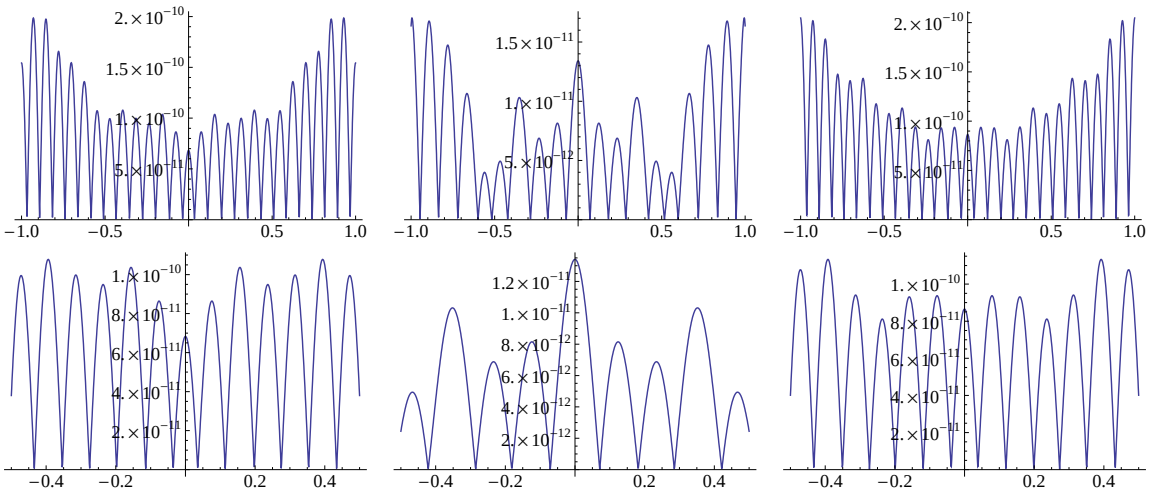


Figure 5.9: Absolute error $|f(x,y_0) - \mathcal{F}_{25,4}[f](x,y_0)|$, where $f(x_1,x_2) = x_1^2 \cosh 3x_1 \cos 2x_2 \sin 3x_2$, for $-1 \leq x \leq 1$ (top row), $-\frac{1}{2} \leq x \leq \frac{1}{2}$ (bottom row), and $y_0 = 1, \frac{2}{3}, \frac{1}{3}$ (left to right).

Theorem 5.44. *Suppose that $f \in \mathrm{H}^{2k+l+1}_{mix}(\Omega)$, where l is the number of equal values $c(r)$, and that $\mathcal{F}_{N,k}[f]$ is the hyperbolic cross version of Eckhoff's approximation to f. Then*

$$\|f - \mathcal{F}_{N,k}[f]\| \lesssim N^{-2k-\frac{3}{2}} (\log N)^{\frac{d-1}{2}},$$
$$\|f - \mathcal{F}_{N,k}[f]\|_r \lesssim N^{r-2k-\frac{3}{2}}, \quad r = 1, \ldots, 2k+1,$$

and $\|\mathrm{D}^\beta (f - \mathcal{F}_{N,k}[f])\|_\infty \lesssim N^{-2k-1} (\log N)^{d-1}$ for $|\beta|_\infty \leq 2k$.

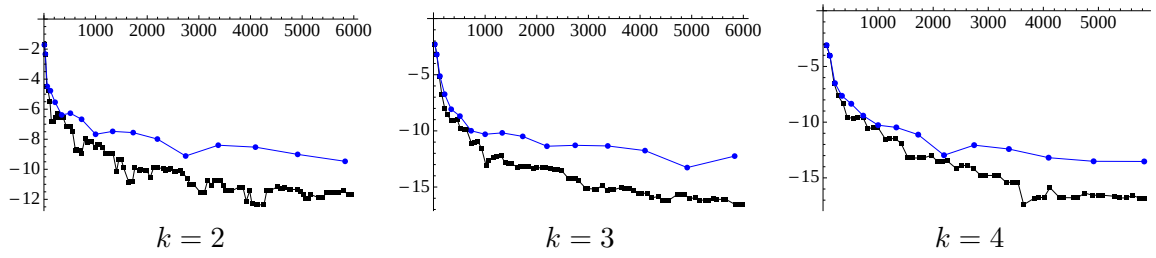
The key result of this section is now apparent: hyperbolic cross index sets may be incorporated into Eckhoff's approximation with only minor deterioration of the convergence rate of the approximation. Furthermore, convergence rates are identical to those of exact polynomial subtraction based on the hyperbolic cross (see Theorem 5.35). As is now familiar, approximating jump values has no effect on the quality of the approximation.

Unfortunately, the hyperbolic cross version of Eckhoff's approximation exhibits no autocorrection phenomenon inside the domain. Much like polynomial subtraction, the pointwise convergence rate away from the boundary is only one power of N faster.[12] The lack of an autocorrection phenomenon is exhibited in Figure 5.9. Though the error lessens away from the boundary, the difference is much smaller than in corresponding results for the approximation based on the full index set (see Figure 5.6).

This somewhat tempers the claims of the previous section. Measured in the uniform norm, the hyperbolic cross approximation greatly outperforms its counterpart based on the full index set. However, away from the boundary, the difference in errors is more marginal. This is demonstrated in Figure 5.10 (using the same function as in Figure 5.8). We observe that, for example, with $k = 4$ and roughly 5,000 terms, the hyperbolic cross approximation yields a pointwise error only 10^3 times smaller than the full version. In comparison, the difference in the uniform error is much larger: roughly 10^5.[13]

[12] This is readily determined upon exploiting the splitting (5.60) once more.

[13] We note, however, that the hyperbolic cross approximation will, asymptotically at least, always outperform



$k = 2$ $k = 3$ $k = 4$

Figure 5.10: Log error $\log_{10} |f(\frac{3}{4}, \frac{3}{4}, \frac{3}{4}) - \mathcal{F}_{N,k}[f](\frac{3}{4}, \frac{3}{4}, \frac{3}{4})|$ against number of approximation terms for the full (circles) and hyperbolic cross (squares) versions of Eckhoff's method applied to (5.59).

5.9 Practical considerations

Thus far, we have concentrated on the analytical aspects of Eckhoff's method. However, computing Eckhoff's approximation brings forth a number of important numerical considerations, which must be properly addressed. Previous studies have indicated that such numerical factors present a significant barrier to the construction of effective approximations based on this approach [62, 118]. In particular, empirical evidence suggests that the value k must remain small, thereby limiting the convergence rate of the approximation. In this section, we demonstrate how such issues can be resolved, including this particular restriction, leading to both an accurate and robust approximation scheme.

Our primary focus in this section is the univariate setting. As demonstrated in Section 5.5, construction of the multivariate Eckhoff approximation involves essentially one-dimensional techniques. Hence, a proper understanding of this case is first necessary. At the end of this section, we present several multivariate examples.

5.9.1 Ill-conditioning

Recall that the construction of Eckhoff's approximation mandates solving a linear system involving the $k \times k$ matrix $V^{[i]}$. The entries of this matrix are either $\widehat{p}^{\,[i]}_{s_{m(r)}}$ or $\widehat{q}^{\,[i]}_{s_{m(r)}}$, depending on whether the cardinal formulation is used or not (see Section 5.3.1). Our first result is of a negative nature: this matrix is extremely ill-conditioned.

Lemma 5.45. *Suppose that $V^{[i]}$ is the matrix with $(r,s)^{\text{th}}$ entry $\widehat{p}^{\,[i]}_{s_{m(r)}}$ or $\widehat{q}^{\,[i]}_{s_{m(r)}}$, where $p^{[i]}_r$ is any cardinal basis and $q^{[i]}_r$ is any subtraction basis. Suppose further that at most $l \leq k$ of the values $c(r)$ are equal, but otherwise the values $m(r)$ are chosen arbitrarily. Then the L^∞ condition number of $V^{[i]}$, $\kappa_\infty(V^{[i]})$, is $\mathcal{O}\left(N^{2k+l-3}\right)$.*

Proof. Due to (5.8), it suffices to consider the cardinal basis formulation. Since $\hat{p}^{\,[i]}_{s_{m(r)}} = \mathcal{O}\left(N^{-2}\right)$, we immediately deduce that $\|V^{[i]}\|_\infty = \mathcal{O}\left(N^{-2}\right)$. Next we consider $\|(V^{[i]})^{-1}\|_\infty$.

its full counterpart. Suppose that the number of approximation terms M is fixed. Then, since $M = \mathcal{O}\left(N^d\right)$ or $M = \mathcal{O}\left(N(\log N)^{d-1}\right)$ for the full or hyperbolic cross approximations respectively, this gives effective pointwise convergence rates of $\mathcal{O}(M^{-\frac{1}{d}(3k+2)})$ and $\mathcal{O}\left(M^{-2k-2}\right)$ (ignoring the $\log N$ term). Since $2k+2 > \frac{1}{d}(3k+2)$ for $d \geq 2$ and $k \geq 1$, we deduce faster convergence of the hyperbolic cross approximation.

As in Lemma 5.4, we write $p_r^{[i]}$ for arbitrary cardinal functions and $P_r^{[i]}$ for cardinal polynomials. This gives $V^{[i]} = W^{[i]} + (V^{[i]} - W^{[i]})$, where $W^{[i]}$ is the matrix with $(r,s)^{\text{th}}$ entry $\widehat{P}^{[i]}_{sm(r)}$. Since $(W^{[i]})^{-1}(V^{[i]} - W^{[i]}) = o(1)$ (see Lemma 5.4), it suffices to consider $\|(W^{[i]})^{-1}\|_\infty$. Recall that $W^{[i]}_{r,s} = \widehat{P}^{[i]}_{sm(r)} = (-1)^{m(r)}(\mu^{[i]}_{m(r)})^{-(s+1)}$. Hence $W^{[i]} = D^{[i]}\tilde{W}^{[i]}$, where $D^{[i]}$ is the diagonal matrix with entries $(-1)^{m(r)}(\mu^{[i]}_{m(r)})^{-1}$ and $\tilde{W}^{[i]}$ is the Vandermonde matrix with entries $(\mu^{[i]}_{m(r)})^{-s}$. For a general $k \times k$ Vandermonde matrix V with entries x_r^s, it is known that

$$\max_{r=0,\dots,k-1} \prod_{\substack{s=0 \\ s\neq r}}^{k-1} \frac{\max\{1,|x_s|\}}{|x_r - x_s|} \leq \|V^{-1}\|_\infty \leq \max_{r=0,\dots,k-1} \prod_{\substack{s=0 \\ s\neq r}}^{k-1} \frac{1+|x_s|}{|x_r - x_s|} \tag{5.71}$$

with equality on the right if all the x_r have the same sign (see, for example, [61]). For the matrix $\tilde{W}^{[i]}$, we have $|x_r - x_s| = \mathcal{O}\left(N^{-2}\right)$ when $c(r) \neq c(s)$ and $|x_r - x_s| = \mathcal{O}\left(N^{-3}\right)$ otherwise. Hence, using (5.71), we deduce that $\|(\tilde{W}^{[i]})^{-1}\|_\infty = \mathcal{O}\left(N^{2k+l-3}\right)$. Since $\|(D^{[i]})^{-1}\|_\infty = \mathcal{O}\left(N^2\right)$, we obtain the result. □

Since the values $m(r)$ are essentially arbitrary, it may appear possible that a judicious choice of such values leads to improved conditioning. However, this is not the case:

Lemma 5.46. *Suppose that $V^{[i]}$ is as in Lemma 5.45. Then there is no choice of distinct values $m(r)$ satisfying $N \leq m(r) \leq aN$ such that the condition number of $V^{[i]}$ is $o(N^{2(k-1)})$.*

Proof. This result follows immediately from the lower bound in (5.71). □

Regardless of these lemmas, when cardinal polynomials are used, reasonably accurate numerical results can often be obtained using the Björk–Pereyra algorithm for Vandermonde matrices [27]. In this manner, the values $\bar{\mathcal{A}}^{[i]}_r[f]$ can be found in $\mathcal{O}\left(k^2\right)$ operations. As is well known, this algorithm produces surprisingly good accuracy even when the underlying matrix is ill-conditioned (see [84, chapter 22]).[14] However, as we now demonstrate, a vast improvement in the condition number can be attained with the non-cardinal formulation.

5.9.2 Subtraction bases and improved conditioning

Since $V^{[i]}$ is essentially a Vandermonde matrix, ill-conditioning is unsurprising: such matrices have a well-earned reputation in this respect. A standard tool to diminish this effect is to replace the original Vandermonde system with a so-called *generalised Vandermonde* system [84]. This equates to replacing the cardinal basis of polynomials with, for example, the Chebyshev polynomials (5.9).[15] Although the condition number of the resulting matrix remains of the same order (as demonstrated in Lemma 5.45), the constant of proportionality is typically greatly reduced.

[14] In fact, in some cases, the magnitude of the numerical error in the Björk–Pereyra algorithm depends only on the machine precision used. In particular, it is independent of the condition number of the matrix [84]. Unfortunately, the corresponding matrix of Eckhoff's method does not satisfy the necessary assumptions for this to be the case, and so the condition number plays a role in the numerical error.

[15] As described in [84], we could also use Legendre or, in general, Jacobi polynomials with similar results.

	$k=4$	$k=6$	$k=8$	$k=10$	$k=12$
(a)	1.6×10^{2}	7.9×10^{2}	1.8×10^{3}	2.5×10^{3}	2.2×10^{3}
(b)	4.0×10^{0}	6.0×10^{-1}	3.2×10^{-2}	8.5×10^{-4}	1.3×10^{-5}
(c)	1.9×10^{-2}	3.5×10^{-6}	6.3×10^{-11}	2.1×10^{-16}	2.0×10^{-22}

Table 5.3: Condition number constant for the matrix $V^{[0]}$ based on values $m(r)=N+r$ and using (a) cardinal polynomial basis, (b) Chebyshev subtraction basis (5.9), and (c) Laplace–Dirichlet basis (5.10). All values to 2 significant digits.

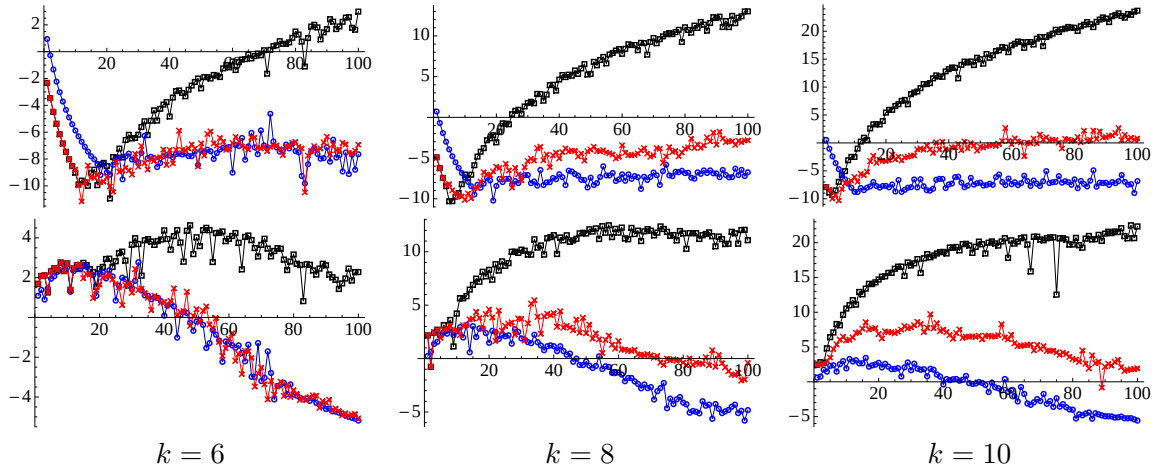


Figure 5.11: Log error $\log_{10}\|f-\mathcal{F}_{N,k}[f]\|_\infty$ against $N=1,\dots,100$ for Eckhoff's approximation using three different bases: cardinal polynomial basis (squares), Chebyshev polynomial basis (crosses) and the Laplace–Dirichlet basis (circles). Here $f(x)=\cosh 6x$ (top diagrams), $f(x)=5e^{\cos 5\pi(1-x^2)}$ (bottom diagrams), and $m(r)=N+r$, $r=0,\dots,k-1$. Numerical results obtained in standard precision, using the *LinearSolve* routine in *Mathematica*.

Having said this, replacement of the cardinal polynomial formulation with a linear system based on Laplace–Dirichlet eigenfunctions (5.10) induces an even more substantial improvement in the condition number. In Table 5.3, we give values for the condition number constant for the matrix $V^{[i]}$ formulated using the cardinal polynomial, Chebyshev (5.9) and Laplace–Dirichlet (5.10) bases. The advantage offered by (5.10) is dramatic. For example, when $k=10$, this constant is roughly 10^{-16}. In comparison, for the cardinal polynomial or Chebyshev bases, these figures are 10^{3} and 10^{-4} respectively, the former being roughly 10^{19} times larger.[16]

This effect is perhaps not surprising: the underlying matrix of the linear system (5.13) is a Cauchy matrix (see Lemma 5.4). Typically such matrices, though often ill-conditioned themselves, are less poorly conditioned than (generalised) Vandermonde matrices [84]. Note that such a linear system can also be solved in $\mathcal{O}\left(k^2\right)$ operations.

The effect of this improvement in conditioning is manifested in numerous examples. In Figure 5.11, we give numerical results for Laplace–Dirichlet, Chebyshev and cardinal polynomial bases applied to several functions. Undeniably, the approximation based on (5.10) offers

[16]The idea of using trigonometric functions as subtraction bases was suggested in [126]. However, no description of their superior numerical behaviour was given.

$m(r)$	$N = 25$	$N = 50$	$N = 100$	$N = 150$	$N = 200$
$N + r$	1.215×10^{24}	1.808×10^{31}	7.398×10^{38}	2.784×10^{43}	5.335×10^{46}
$(r+1)N$	8.688×10^{30}	2.147×10^{36}	5.552×10^{41}	8.185×10^{44}	1.451×10^{47}
$2^r N$	2.933×10^{42}	7.206×10^{47}	1.861×10^{53}	2.742×10^{56}	4.859×10^{58}

Table 5.4: L^∞ condition number of the linear system (5.13) using the functions (5.10) with $k = 10$ and values $m(r)$ given by (5.72)–(5.74). All values to 4 significant digits.

the smallest error. Moreover, unlike the cardinal polynomial basis, the error remains bounded for all N. Note that the functions used here exhibit two features, large derivatives and high oscillation, making their approximation prone to numerical errors. However, simply by selecting an appropriate subtraction basis, we are able to obtain vastly superior approximations. In view of these examples, we arrive at a surprising conclusion: *polynomial subtraction is best achieved without using polynomials!*

5.9.3 Choice of the values $m(r)$

The values $m(r) \geq N$ can be chosen arbitrarily, provided they are distinct and satisfy $m(r) = c(r)N + \mathcal{O}(1)$ for some $c(r) \geq 1$. Numerous choices are possible, including the aforementioned values

$$m(r) = N + r, \quad r = 0, \dots, k-1. \tag{5.72}$$

In this case $c(r) = 1$ for all r, so the function f being approximated must have $\mathrm{H}^{3k+1}(-1,1)$-regularity to ensure an $\mathcal{O}\left(N^{-2k-1}\right)$ uniform convergence rate. Other possibilities that require only $\mathrm{H}^{2k+2}(-1,1)$-regularity are also permitted, including

$$m(r) = (r+1)N, \quad r = 0, \dots, k-1, \tag{5.73}$$

$$m(r) = \omega^r N, \quad r = 0, \dots, k-1. \tag{5.74}$$

One immediate disadvantage of these choices is that they do not lead to a full auto-correction phenomenon (see Section 5.7). Moreover, the values $\hat{f}_n^{[i]}$, $n = 0, \dots, N-1$, $n = m(r)$, $r = 0, \dots, k-1$, required to form the approximation are not contiguous, in contrast to (5.72). Finally, as we now demonstrate, (5.73) and (5.74) both lead to inferior numerical stability in comparison to (5.72).

In all numerical results thus far, we have employed the values (5.72). Seemingly, the condition number of the linear system (5.12) can be vastly improved from $\mathcal{O}\left(N^{3(k-1)}\right)$ to $\mathcal{O}\left(N^{2(k-1)}\right)$ by using the values (5.73) or (5.74) instead (see Lemma 5.45). However, though true in theory, in practice, the constant is so overbearingly large that it nullifies this effect. In Table 5.4, we give numerical results for the condition number of this linear system using the values (5.72)–(5.74). We observe that N must exceed 200 before the values (5.73) begin to offer an advantage (for the values (5.74), the scenario is much worse). However, since $k = 10$ in this example, any reasonable function will be well resolved by Eckhoff's approximation for a much smaller value of N.

A theoretical explanation of this effect is readily provided. Suppose that $W^{[i]}$ is the matrix

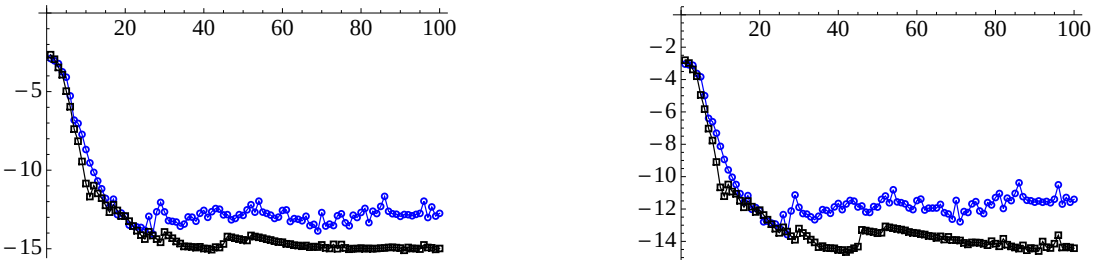


Figure 5.12: L^2 (left) and uniform (right) errors against $N = 1, \ldots, 100$ for Eckhoff's approximation with $k = 8$ applied to the function $f(x) = \mathrm{Ai}(-6x-4)$. Coefficients are calculated by the *Mathematica* routines *LeastSquares* (squares) and *LinearSolve* (circles).

with $(r,s)^{\text{th}}$ entry $(\mu^{[i]}_{m(r)})^{-s}$. In this case, since $\mu^{[i]}_{m(r)} > 0$, the estimate (5.71) gives

$$\|(W^{[i]})^{-1}\|_\infty = \max_{r=0,\ldots,k-1} \prod_{\substack{s=0 \\ s\neq r}}^{k-1} \frac{1+(\mu^{[i]}_{m(s)})^{-2}}{\left|(\mu^{[i]}_{m(r)})^{-2} - (\mu^{[i]}_{m(s)})^{-2}\right|}.$$

Now suppose that $m(r) = c(r)N + \mathcal{O}(1)$ with all the values $c(r)$ distinct. Then $(\mu^{[i]}_{m(r)})^{-2} - (\mu^{[i]}_{m(s)})^{-2} = (N\pi)^{-2}\left(c(r)^{-2} - c(s)^{-2}\right) + \mathcal{O}\left(N^{-3}\right)$, and we see that

$$\|(W^{[i]})^{-1}\|_\infty = c(N\pi)^{2(k-1)} + \mathcal{O}\left(N^{2(k-2)}\right),$$

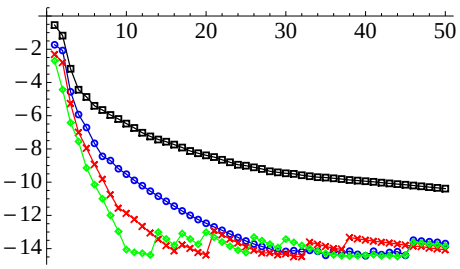
where the constant c is determined by the condition number of the Vandermonde matrix based on nodes $x_r = c(r)^{-2}$. Conversely, when all the values $c(r)$ are equal and $m(r) = c(r)N + d(r)$ for distinct $d(r) \in \mathbb{N}_0$, an identical argument demonstrates that $\|(W^{[i]})^{-1}\|_\infty = c(N\pi)^{3(k-1)} + \mathcal{O}\left(N^{3(k-2)}\right)$, with constant determined by the Vandermonde matrix based on the interpolation points $x_r = d(r)$.

For (5.72)–(5.20), we conclude that the condition number constant is determined by the Vandermonde matrix based on nodes $x_r = r$, $x_r = (r+1)^{-2}$ and $x_r = \omega^{-2r}$ respectively. For the latter two, interpolation nodes become clustered near the origin as k increases, leading to ill-conditioning. Conversely, for (5.72), there is no such clustering.

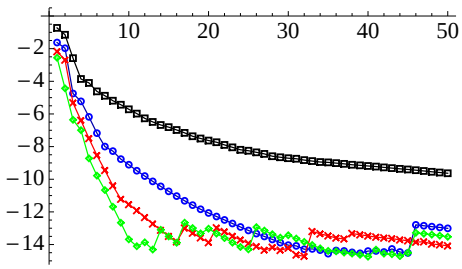
5.9.4 Least squares

Numerical results can be further improved by replacing (5.12) with an overdetermined linear system and using least squares. This approach is fairly standard [16, 54]. We illustrate the improvement offered by this approach in Figure 5.12. In this and all subsequent examples, we overdetermine by a factor of two, leading to two $2k \times k$ linear systems (corresponding to $i = 0, 1$) that are solved in parallel.

As exhibited, the approximation obtained from the least squares procedure offers a lower error (by several orders of magnitudes) than the approximation formed by solving a square linear system. This is somewhat predictable: it has been widely reported that least squares can improve the performance of both this and related techniques for convergence acceleration



$f(x) = x^2 \sin 5x + \cos 6x$



$f(x) = \mathrm{e}^{-x} \cos 6x$

Figure 5.13: The error $\log_{10} \|f - \mathcal{F}_{N,k}[f]\|$ against $N = 1, \ldots, 100$ for $k = 2, 4, 6, 8$ (in descending order).

of Fourier series (see [32, 39, 86] for the use of similar techniques in so-called *Fourier extension methods*[17]).

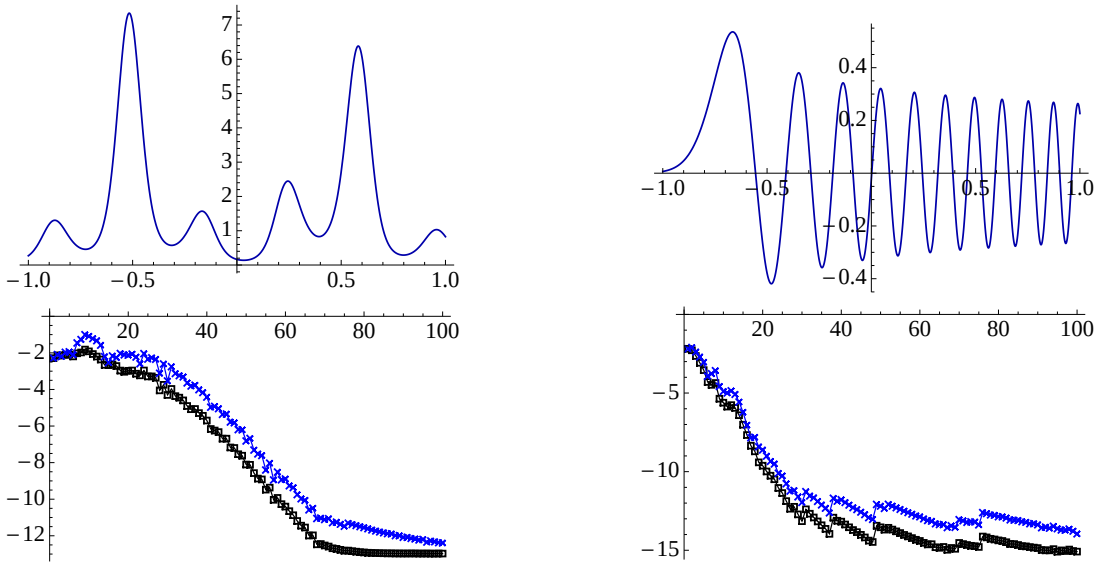
Nonetheless, the application of a least squares procedure to compute Eckhoff's approximation yields an effective numerical approximation. Consider, for example, the two functions approximated in Figure 5.13. Using only $k = 8$ and $N = 10$, we obtain $\mathrm{L}^2(-1, 1)$ errors of approximately 10^{-14}. In comparison, when $k = 2$, the corresponding errors are only 10^{-6}, a factor of 10^8 times larger. Furthermore, Eckhoff's method, in conjunction with least squares, can now be successfully applied to a raft of poorly behaved functions. For example, the functions considered in Figure 5.14 both exhibit (irregular) oscillations inside the interval, thus making their approximation prone to numerical errors. However, once again, Eckhoff's approximation obtains good accuracy using only moderate parameter values (note that the first function approximated in this figure was used as an example in [39] to test the robustness of the Fourier extension method).

As previously noted, common consensus is that the parameter k ought to be kept small in order to mitigate the effect of numerical instability when approximating functions with Eckhoff's method [62]. However, with the approach developed in this section, it is permissible to take much larger values. In Figure 5.13, for example, we used $k = 20$, giving a theoretical uniform convergence rate of N^{-41}!

We are therefore led to a somewhat surprising conclusion: although Eckhoff's method requires the solution of an incredibly ill-conditioned linear system, extremely high accuracy can be obtained through the use of appropriately chosen subtraction bases, parameters $m(r)$, and the solution of an overdetermined least squares system. Furthermore, the practical techniques developed in this section are readily carried over to functions of two or more variables. This is demonstrated in Figure 5.15. For example, when $k = 8$ and $N = 10$, we obtain at least 11 digits of accuracy, in comparison to only 1 for the original ($k = 0$) modified Fourier approximation.

This section completes our study of Eckhoff's method and its multivariate generalisation. We conclude this chapter with a discussion of the Gibbs phenomenon and techniques for its resolution (into which the topic of convergence acceleration of Fourier-like series naturally falls). In doing so, we discuss a number of different approaches and their comparative aspects in relation to Eckhoff's method.

[17] We discuss such methods in greater detail in Section 5.10.3.



$f(x) = \mathrm{e}^{\sin(5.4\pi x - 2.7\pi) - \cos 2\pi x}$ $\qquad$ $f(x) = \mathrm{Ai}(-12x - 9)$

Figure 5.14: Top row: plot of the function $f(x)$. Bottom row: the errors $\log_{10} \|f - \mathcal{F}_{N,20}[f]\|$ (circles) and $\log_{10} \|f - \mathcal{F}_{N,20}[f]\|_\infty$ (crosses) against $N = 1, \ldots, 100$.

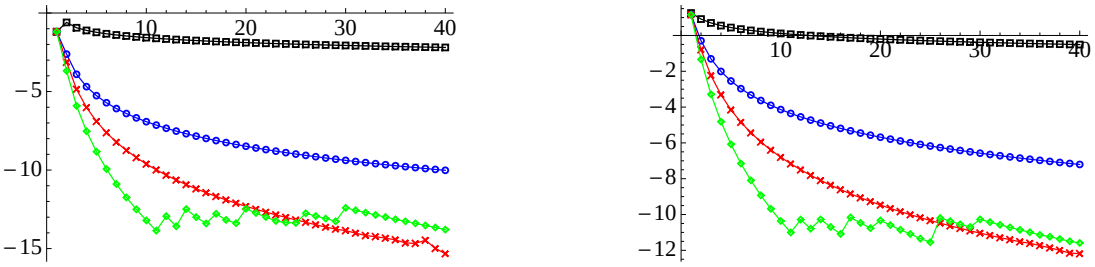
5.10 The Gibbs phenomenon and its resolution

The Gibbs phenomenon has a rich and interesting history. It was observed by Euler in 1755 that the function $f(x) = x$ could be represented on the interval $[-1, 1]$ as an infinite sum of trigonometric functions. Almost a century later, Wilbraham analysed this series, including a description of the overshoot near the endpoints [160]. Forgotten for half a century, this phenomenon was reconsidered by Michelson [122]. The ensuing debate regarding convergence, or lack thereof, between Michelson and Love, carried out in *Nature*, was eventually settled by Gibbs [64, 65] in 1899, with the arbitration of Poincaré. Gibbs' contribution to this problem was first recognised by Bôcher in 1906 [28], who introduced the term *the Gibbs phenomenon*.[18] A detailed and fascinating review of the Gibbs phenomenon and its history is provided in [83], with shorter summaries appearing in [43, 72].

The Gibbs phenomenon is certainly not restricted to Fourier expansions. As we have seen, it appears not only in various Laplace eigenfunction expansions, but also in expansions in eigenfunctions of polyharmonic operators. Numerous other instances have also been recorded (see [72] and references therein).

The original viewpoint of the Gibbs phenomenon focuses on the non-uniform convergence of Fourier series and, in particular, the nature of the overshoot and oscillations near the boundary of the domain. As discussed in [72], a slightly different point of view is that the Gibbs phenomenon concerns *the issue of recovering a function from a finite number of its Fourier coefficients*. In other words, the recovery of local information (pointwise values) from global information (Fourier coefficients). This standpoint raises the question of how the Gibbs

[18] To acknowledge the contribution of Wilbraham, it is also referred to as *the Gibbs–Wilbraham phenomenon*.



$f(x_1, x_2) = (1 - \cos x_1) \sin 3x_2$ $\qquad$ $f(x_1, x_2) = \mathrm{e}^{2x_1 - x_2}$

Figure 5.15: The error $\log_{10} \|f - \mathcal{F}_{N,k}\|_\infty$ against $N = 1, \ldots, 40$ for $k = 0, 2, 4, 8$.

phenomenon can be circumvented.

5.10.1 Resolution of the Gibbs phenomenon

The resolution of the Gibbs phenomenon was first considered by Fejér in 1900. His discovery of uniform convergence of Cesàro means can be viewed as the first construction of a Fourier filter. Indeed, the Cesáro sum is equivalent to a first-order filter [156]. The topic of filtering has been extensively studied since this point, and we refer the reader to [156] for a substantial review of the subject.[19]

Filters successfully enhance the accuracy of Fourier approximations. One immediate advantage is that the approximation remains a sum of trigonometric functions, and hence can be evaluated rapidly with the FFT. However, this increase in accuracy only occurs away from the discontinuity of the function (in other words, the endpoints, for a smooth, nonperiodic function).

In view of this fact, a re-projection method to fully resolve the Gibbs phenomenon was suggested by Gottlieb [73]. The basic idea is to re-expand the Fourier sum of a function in a particular orthonormal basis, the so-called *Gibbs complementary basis*. For a suitably chosen basis, the convergence of the re-projection is exponential (assuming analyticity of the function in some region containing the prescribed interval). Hence, the Gibbs phenomenon can be completely resolved. For Fourier approximations, a suitable basis consists of Gegenbauer polynomials [20], whose parameter λ is varied with the truncation parameter N of the Fourier sum. This process is commonly referred to as *Gegenbauer reconstruction* [72].

This idea has been extended in various ways to include, for example, re-projections of Fourier interpolation approximations [70] and re-projections for expansions in other bases (in particular, bases of Chebyshev and Legendre polynomials) [69, 71]. A review is given in [72] and a general framework in [74]. To date, the most pertinent applications of this method have been in image processing [9, 10] and the spectral approximation of partial differential equations with discontinuous solutions [82].

[19] We mention in passing that, in general, there are two components to resolving the Gibbs phenomenon: detection (location of singularities) and reconstruction. We do not address the former in this thesis. Indeed, we assume that the approximated function is smooth and nonperiodic: in other words, singularities only occur on the boundary of the domain (in the sense of the Fourier extension). There are numerous techniques for singularity detection, and these form an central component in many practical applications, including signal processing. We refer the reader to [156] for further details.

Unfortunately, there are a number of problems associated with the implementation of this technique. In particular, the method is liable to round-off error [11, 62], and care must be taken to ensure convergence of the re-projection [33].[20] This approach also has a number of inherent disadvantages: it is reasonably computationally expensive (the number of operations is $\mathcal{O}\left(N^2\right)$), the resolution power is rather poor[21], and the final approximation, being a sum of Gegenbauer polynomials, cannot be evaluated rapidly using, for example, the FFT.

5.10.2 Polynomial subtraction

Around the same time as Fejér, Krylov introduced the idea of polynomial subtraction. This was later formalised in [102] and [112]. Eckhoff's approach, which we have extended in this chapter, was originally detailed in [52], but the idea behind it has much older origins. The foundation, as commented by Lax [113], is that the Fourier coefficients themselves contain sufficient information to produce accurate representations of the function f. We note that this rather general viewpoint also forms the basis of Gegenbauer reconstruction, yet the path taken to extract such information is different.

As stated, a classical recommendation is that the parameter k should remain small [62], thus *ameliorating* the Gibbs phenomenon rather than resolving it completely. Through the work of this chapter, however, we have demonstrated how much larger values can be taken. Indeed, though we have focused on the case of finite, fixed k in this study, there is no reason not to consider the choice $k = N$, leading to *exponentially accurate* approximations. However, as we now describe, this particular parameter value naturally lends itself to a different interpretation, thereby relating this version of Eckhoff's method to an alternative (and relatively unfamiliar) convergence acceleration device.

5.10.3 Fourier extension methods

Consider, for example, the univariate case. Eckhoff's method, when constructed from Laplace–Neumann and Laplace–Dirichlet eigenfunctions with parameter $k = N$, computes an approximation to a function f from the set

$$\mathcal{S}_N = \left\{\cos \tfrac{1}{2}n\pi x : \ n = 0, \dots, N\right\} \cup \left\{\sin \tfrac{1}{2}n\pi x : \ n = 1, \dots, N\right\}. \tag{5.75}$$

Written in this form, we are immediately struck by the following observation. The set $\mathcal{S}_N$ is precisely the set of classical Fourier basis functions (with index $n \leq N$) on the extended domain $[-2, 2]$. Hence, to find an accurate approximation to f, we seek a smooth function, periodic on $[-2, 2]$, that matches f on $[-1, 1]$. We then approximate f by the truncated Fourier expansion of this function.

The question of computing a periodic extension of f is known as the *Fourier extension problem*. A simple criterion to devise a suitable extension was introduced in [32, 39]. We

[20] Nonetheless, a substantially more resilient technique has recently been suggested in [63]. The so-called inverse Gegenbauer reconstruction method [98, 99] also appears to offer some advantages, albeit at additional computational expense.

[21] As discussed in [11, 68], at least 22.2 modes per wavelength are needed to ensure exponential convergence of the Gegenbauer method. In comparison, methods based on Fourier series or Chebyshev polynomials require 2 and π modes respectively. Eckhoff's method performs similarly to the Fourier method in this respect, requiring only 2 modes per wavelength, once more, provided $k \ll N$.

define

$$f_N = \arg\min_{g \in \mathcal{S}_N} \|f - g\|, \tag{5.76}$$

where $\|\cdot\|$ is the standard $\mathrm{L}^2(-1,1)$ norm. Explicit computation of f_N is now easily achieved. In fact, in the language of spectral methods, the optimization criterion (5.76) is identical to the problem

$$\text{find } f_N \in \mathcal{S}_N: \quad (f_N, \phi) = (f, \phi), \quad \phi \in \mathcal{S}_N. \tag{5.77}$$

Hence, the *Fourier extension method* is a Galerkin method for computing the approximation f_N from the (albeit non-orthogonal) set $\mathcal{S}_N$.

Note that f_N, as defined by (5.76), is not the Eckhoff approximation of f with $k = N$. In the same language as above, Eckhoff's method can be viewed as a *Petrov–Galerkin* method for computing f with trial space $\mathcal{S}_N$ and test space being the set of Laplace–Neumann eigenfunctions with index $n \leq 2N$.[22] Hence, Eckhoff's method with this particular choice of parameters can be viewed as a novel approach to computing the Fourier extension. We remark in passing, however, that although there exists a standard theory for Petrov–Galerkin approximations [14], this does little to illuminate the analysis of Eckhoff's approximation.

Returning to Fourier extension methods, it transpires that the approximation f_N has a rather elegant interpretation in terms of orthogonal polynomials, as demonstrated by Huybrechs [86]. Indeed, f_N can be viewed as the expansion of f in certain *half-range Chebyshev* polynomials with arguments $\cos\frac{1}{2}\pi x$ or $\sin\frac{1}{2}\pi x$. Analysis of convergence therefore follows from standard polynomial approximation results. Indeed, assuming sufficient analyticity of f in a complex neighbourhood of $[-1,1]$, exponential convergence is now witnessed: $f(x) - f_N(x) \sim (3 + 2\sqrt{2})^{-N}$.

We shall return to Fourier extension methods briefly in Chapter 6. It remains to be seen whether such methods possess significant benefits over this version of Eckhoff's method. Certainly, their simple interpretation in terms of orthogonal polynomials makes such techniques immediately attractive (at the expense, however, of having to know both the Laplace–Dirichlet and Laplace–Neumann coefficients of a given function f). Yet this formulation is lost once other eigenfunction bases and subtraction functions are employed. We mention in passing that a brief comparison of the two approaches was carried out in [7], in which both methods behaved in a roughly similar manner.

To conclude this discussion of convergence acceleration techniques, we remark that Eckhoff's method, Fourier extension methods, Gegenbauer reconstruction and filtering are just some of a virtually endless number of techniques to accelerate convergence of Fourier-like series. There are numerous alternative approaches, which we do not intend to discuss in greater detail, including Fourier–Padé methods [49] and techniques from sequence acceleration [36], to name but two (for a more detailed list, see [34] and references therein). Certainly, different techniques are more suitable for particular problems, yet a thorough comparison of such methods would require a lengthy review. A particular motivation for developing Eckhoff's method here is due to its potential application to boundary value problems, a topic we discuss briefly in Chapter 6. Moreover, its simple generalisation to the d-variate cube, as well as its potential application in more complex geometries, signal it as an appropriate choice for both current and future purposes.

[22] A Petrov–Galerkin method is similar to a Galerkin method, except that the *trial space* (the space of functions to which the solution belongs) and *test space* (the space of functions with respect to which inner products are taken) can be distinct [142].

Chapter 6

Conclusions and future work

6.1 Summary of the thesis

The intent of this thesis was the development and analysis of approximation schemes based on certain eigenfunction bases, with particular application to the numerical solution of boundary value problems. Such an approach incorporates a number of novel numerical techniques, including a mixture of classical and highly oscillatory quadratures to evaluate coefficients, as well as the use of hyperbolic cross index sets to considerably decrease computational cost.

Chapter 2 introduced a theory for so-called modified Fourier expansions. Key results included a proof of uniform convergence and estimates for the rate of pointwise convergence. Explicit criteria that determine both the rate and degree of convergence were derived in terms of odd derivatives evaluated on the boundary of the domain.

Modified Fourier expansions were generalised in Chapter 3 to expansions based on eigenfunctions of univariate polyharmonic operators. This led to a one-parameter family of approximation bases with a convergence rate that scaled with the parameter. A thorough convergence analysis was provided. As a by-product, several new results concerning the asymptotic nature of the eigenfunctions and eigenvalues were established. A generalisation to the d-variate cube via Cartesian products was then investigated, culminating in expansions in eigenfunctions of certain subpolyharmonic operators.

In Chapter 4, we assessed the application of Laplace eigenfunctions to the spectral–Galerkin discretisation of boundary value problems defined in the d-variate cube. This approach results in well-conditioned matrices with corresponding linear systems that can be solved inexpensively using generic iterative techniques. The ensuing method possesses several advantages over standard polynomial-based techniques, as substantiated by numerical examples.

Finally, the topic of convergence acceleration was broached in Chapter 5. Using only the modified Fourier coefficients of a function, we constructed approximations with arbitrary rates and degrees of convergence. When combined with a hyperbolic cross, this facilitated the construction of accurate approximations comprising relatively small numbers of terms. Numerical stability was also markedly improved by the appropriate selection of various parameters and the use of a least squares procedure, thus effecting an efficient numerical method possessing both robustness and high accuracy.

There are numerous avenues to pursue in order to extend this work, as we henceforth describe.

6.2 Expansions in the equilateral triangle and higher dimensional simplices

Eigenfunctions of the Laplace operator subject to either homogeneous Neumann or Dirichlet boundary conditions are known to have explicit representations (as sums of trigonometric functions) in a variety of non-tensor-product domains. In the plane, the list includes ellipses, annuli and three types of triangles: the equilateral and right isosceles triangles, and the triangle with angles $\frac{\pi}{2}$, $\frac{\pi}{3}$ and $\frac{\pi}{6}$. This has important consequences for practical applications of modified Fourier expansions. Triangular elements can be used to decompose complex, often polygonal geometries, and possess far more flexibility than rectangular elements.

The current dearth of high-order approximation schemes in triangular domains is a compelling motive for the continued development of modified Fourier expansions. As described in Chapter 1, the lack of a simple high-order scheme based on orthogonal polynomials necessitates the introduction of other techniques. In this regard, the particularly simple nature of Laplace eigenfunctions presents a significant advantage of modified Fourier expansions.

Though a study of modified Fourier expansions in triangular domains has been initiated in [88], including techniques to evaluate coefficients numerically, there remain many open problems and challenges. New hurdles that appear as a consequence of the non-tensor-product structure require a great deal of further insight before such expansions can be converted into effective approximations.

To highlight this, we now present the following (inexhaustive) list of open problems and future challenges within this topic:

1. *Uniform convergence.* A key question we have addressed in this thesis is the uniform convergence of multivariate modified Fourier expansions in the d-variate cube. Intuition and numerical examples suggest that modified Fourier expansions defined in triangular elements behave in a similar manner to corresponding expansions in the unit square. Specifically, expansions converge uniformly throughout the domain. As of this moment, we have no proof of this fact.[1]
2. *Rate of convergence.* Numerical examples suggest that expansions in triangles also mirror expansions in tensor-product domains in terms of their rates of convergence. In particular, faster convergence occurs inside the domain than on the boundary. Estimates for rates of convergence in various norms have previously been obtained in [155]. However, these results are restricted to classes of functions with vanishing Neumann data on the boundary (in analogy with the standard periodic spaces $\mathrm{H}^k(\mathbb{T})$), and therefore fail to describe the approximation error for an arbitrary function.
3. *Mixed Sobolev spaces for triangular domains.* In the d-variate cube, the Sobolev spaces $\mathrm{H}^k_{\mathrm{mix}}(-1,1)^d$ are fundamental to the analysis of modified Fourier approximations. We may define mixed spaces for triangular domains in an identical manner. However, the spaces $\mathrm{H}^k_{\mathrm{mix}}(-1,1)^d$ have a tensor-product structure (see Section 2.5), a property which is lost when passing to the triangle. This indicates that new spaces are necessary for an accurate study of expansions in triangles.

[1] Nonetheless, when the function f has vanishing normal derivative on the whole of the boundary, this result is easily established. In this case, Stokes' theorem verifies that $\mathcal{F}_N[\triangle f] = \triangle \mathcal{F}_N[f]$, where $\mathcal{F}_N[f]$ is the modified Fourier expansion of $f \in \mathrm{H}^2(\Omega)$ and Ω is the triangle. Hence, $\mathcal{F}_N[f] \to f$ in the $\mathrm{H}^2(\Omega)$ norm. Uniform convergence follows at once from the continuous embedding $\mathrm{H}^2(\Omega) \hookrightarrow \mathrm{C}(\bar{\Omega})$. The general case, however, remains unproven.

4. *Gibbs phenomena.* Classical Gibbs phenomena are extremely well understood in the unit interval and d-variate cube. As far as we can ascertain, nothing is known about corresponding (weak) Gibbs phenomena for expansions in triangular domains. Intuition suggests that such phenomena will possess a more complex structure than the simple tensor-product case. Yet, at present, we have no results in this respect.
5. *The hyperbolic cross.* In [88], a hyperbolic cross was derived for modified Fourier coefficients in the equilateral triangle. Little has been established, however, regarding the potential advantage of the resulting index set. In particular, issues concerning rates of convergence are largely unexplored.
6. *Convergence acceleration.* Polynomial subtraction for modified Fourier expansions in multivariate domains with tensor-product structure is now well established. The first steps towards such a construction for the equilateral triangle were undertaken in [88], where a subtraction function was derived for the first derivative condition. Many questions remain, however, as regards this technique. These include the as of yet undetermined convergence rate that results from this device, and how such a construction can be generalised to arbitrary numbers of derivatives.

   Polynomial subtraction also requires explicit derivative information. Having analysed this device, it is logical to consider the approximation of such derivatives by an Eckhoff-type approach. The eventual aim, as in the case of the d-variate cube, is to obtain rapid approximations using only the modified Fourier coefficients of a given function.

   Future study in this topic need not be restricted to Eckhoff-type methods, however. Numerous other devices, including filters and Fourier extension methods[2], can, in theory at least, be generalised to expansions in the equilateral triangle. Naturally, practical schemes will incorporate only the most effective convergence acceleration strategy. Proper extension of a variety of techniques, and a thorough comparative study therein, are both important avenues for future research.

The classification of domains for which Laplace eigenfunctions are explicitly known is an interesting mathematical problem. Already in the 1800s Laplace eigenfunctions had been determined for the equilateral triangle [111]. Since this time, alternate constructions via the *method of images* [104] have been used repeatedly to obtain eigenfunctions [137, 141] (for a more detailed review, see [151]).

A close connection with group theory is revealed, however, upon realising the equilateral triangle (as well as the square, right isosceles triangle, etc) as a member of the family of domains consisting of so-called *fundamental regions of root systems* [7]: that is to say, those domains that can be repeatedly reflected across their boundaries to tile $\mathbb{R}^d$. For such domains, Laplace eigenfunctions can be derived by applying the symmetries described by the corresponding root system to the classical Fourier basis. Future work will also aim to incorporate this theory into the design of modified Fourier approximations in a variety of higher-dimensional simplices.

In [100, 101] families of Laplace eigenfunctions corresponding to Dirichlet, Neumann, Robin and certain Poincaré boundary conditions were obtained using the so-called *Fokas method*. Such results may have direct impact on the understanding of modified Fourier expansions in triangular domains. Evidence suggests that any duality (in the sense of Chapter 2) enjoyed by Laplace–Neumann eigenfunctions in these domains, as opposed to the unit cube,

[2]This is also currently under investigation by D. Huybrechs [7].

will involve families of Laplace eigenfunctions corresponding to more complicated boundary conditions. This work may provide the key to such an understanding.

6.3 Accelerating convergence of modified Fourier–Galerkin approximations

Numerous attempts have been previously made towards the rapid approximation of solutions of boundary value problems of the form (4.5) in one, two, or three dimensions using Fourier or Fourier-like series. Most previous techniques use a variant of the polynomial subtraction process [12, 35, 67, 126, 144, 149], and suffer from the dual restrictions of being commonly limited to the constant coefficient Helmholtz ($a = 0$) problem and requiring exact knowledge of the derivative information of the inhomogeneous term f.

However, bearing in mind the work of Chapter 5, effective treatment of the Helmholtz problem is now straightforward: when $a = 0$, the modified Fourier coefficients of the solution u are known explicitly (in terms of the coefficients of f, see Section 4.3.1). Hence, the rapid approximation of u without derivatives is easily acquired via Eckhoff's method [126].

Outside of this trivial case, accurate approximations can be designed for the solution of problems of the form (4.5) with arbitrary (not necessarily constant) coefficients. Theoretically, this is very simple. For example, if $u_{N,k}$ is an Eckhoff-type approximation, then we specify the coefficients of $u_{N,k}$ by the relation

$$T\left(u_{N,k}, \phi_n^{[i]}\right) = \hat{f}_n^{[i]}, \quad \forall n \in I_N \cup M_k, i \in \{0,1\}^d,$$

where M_k is the index set (5.30) and T is the bilinear form appearing in the weak formulation of the problem (4.5). Note that, upon defining the space $Y_N = \operatorname{span}\{\phi_n^{[i]} : n \in I_N \cup M_k, i \in \{0,1\}^d\}$, this approach can be immediately interpreted as the Petrov–Galerkin method

$$\text{find } u_{N,k} \in X_N\text{:} \quad T\left(u_{N,k}, v\right) = (f, v), \quad \forall v \in Y_N, \tag{6.1}$$

where X_N is space of Eckhoff-type approximants (i.e. functions consisting of a truncated modified Fourier sum and a subtraction function). Observe the generality of this approach: no stipulations have been made in (6.1) regarding either the operator or boundary conditions.[3]

Several key issues immediately present themselves. First, ill-conditioning that was originally confined to a $k \times k$ matrix now permeates throughout the $K \times K$ discretisation matrix, where $K = |I_N| + |M_k| = \mathcal{O}\left((N+k)^d\right)$. Second, since this matrix is dense, it is not yet clear how to rapidly compute the approximation $u_{N,k}$. Noting that Eckhoff's approximation to the Helmholtz problem can be easily constructed suggests that the operator splitting $\mathcal{L} = \mathcal{L}_0 + \mathcal{L}_1$ could be employed once more (as in Chapter 4). However, it remains to be seen whether this approach is effective or not. We remark in passing that no analysis of the approximation (6.1) has yet been devised[4], although numerical experiments demonstrate the increase in convergence.

[3] Of course, such issues are of paramount importance as regards the implementation of such schemes.

[4] Petrov–Galerkin methods have a standard analysis, including a statement analogous to Céa's lemma, provided the finite-dimensional spaces X_N and Y_N satisfy a so-called *inf-sup condition* [14, 142]. Verification of this condition for these particular spaces would immediately lead to error estimates for $u_{N,k}$.

In the univariate case, at least, the significant issue of ill-conditioning can be circumvented. For the problem $-u''(x) + a(x)u'(x) + b(x)u(x) = f(x)$, $x \in [-1,1]$, $u'(\pm 1) = 0$, the odd derivatives of the solution u satisfy a linear relation of the form $u^{(2r+1)}(\pm 1) - c_r^{\pm} u(\pm 1) = d_r^{\pm}$, where the values $c_r^{\pm} \in \mathbb{R}$ depend only on the functions $a(x)$ and $b(x)$ and their first $2r$ derivatives evaluated at $x = \pm 1$, and the $d_r^{\pm}$ depend only on f and its first $2r-1$ derivatives (such derivatives can, of course, be approximated by applying Eckhoff's method in turn to a, b and f). The method proposed in [3] is to approximate u by a function $u_{N,k} \in X_N$ that satisfies both Galerkin's equations and these relations:

$$\begin{aligned} T(u_{N,k}, \phi_n^{[i]}) &= \hat{f}_n^{[i]}, \quad n = 0, \ldots, N, \quad i \in \{0,1\}, \\ u_{N,k}^{(2r+1)}(\pm 1) - c_r^{\pm} u_{N,k}(\pm 1) &= d_r^{\pm}, \quad r = 0, \ldots, k-1. \end{aligned} \tag{6.2}$$

As proved in [3], this approximation not only accelerates convergence, the error $\|u - u_{N,k}\|_\infty$ is $\mathcal{O}\left(N^{-2k-3}\right)$, but the condition number of the linear system also remains $\mathcal{O}\left(N^2\right)$, and the solution $u_{N,k}$ can be constructed in $\mathcal{O}\left(N^2\right)$ operations (provided $k \ll N$). It remains to be seen, however, whether this approach scales to higher dimensions. Certainly, the linear relations obeyed by the partial derivatives of u will become increasingly complicated for problems in two or more variables, thus potentially limiting the scope of this approach. Conversely, Eckhoff's approach, whilst exhibiting ill-conditioning, is, theoretically at least, much simpler to construct for a wide variety of problems.

An alternative to an Eckhoff-type approach is to employ Fourier extension methods.[5] This leads to a Galerkin approximation, with ensuing simple analysis. In fact, if $u_N \in \mathcal{S}_N$ is the approximant, where $\mathcal{S}_N$ is defined in (5.75), then u_N satisfies the equations $T(u_N, \phi) = (f, \phi)$, $\forall \phi \in \mathcal{S}_N$. As with the Eckhoff-type approach, issues of ill-conditioning and computational cost persist. However, since this is a Galerkin approximation (in contrast to the more complicated Petrov–Galerkin setting), convergence can be immediately guaranteed. In fact, exponential convergence is observed, provided u is analytic.

Future work will address the continuing development of fast approximations to partial differential equations based on the aforementioned techniques. As discussed in Chapter 4, by designing effective approximations based on modified Fourier expansions, we aim to extend the range of applicability of modified Fourier methods to a wider variety of problems.

Naturally, this work is not necessarily restricted to tensor-product domains. Upon development of convergence acceleration techniques for expansions in the equilateral triangle, for example, the next step for future research will be the application to boundary value problems in such domains. Combined with a suitable domain decomposition strategy [44, 143], the eventual intent of this work is the construction of high-order approximations in complex geometries. Needless to say, applications are potentially broad-reaching. They include the numerical solution of many problems in fluid dynamics and electromagnetism, for example, more commonly tackled by low-order finite element methods.

6.4 Numerical evaluation of coefficients

Combinations of highly oscillatory and nonstandard classical quadratures form the mainstay of techniques to evaluate the modified Fourier coefficients of a given function. As previously

[5] Methods for differential equations based on univariate Fourier extensions have been studied in [39].

alluded to, the development of efficient, robust techniques based on such quadratures is the main bottleneck towards the development of effective algorithms based on modified Fourier expansions. Herein numerous questions and open problems remain.

Currently, there are few accurate and reliable bounds for either Filon-type or exotic quadratures, nor has the stability of such methods for large numbers of nodes and multiplicities been established.[6] Standard classical quadrature has an extremely well-understood theory, including simple criteria for selecting optimal node locations to obtain highest possible orders. No such theory yet exists for exotic quadrature: as described in [8], an optimal choice of internal nodes was possible in some examples, whereas in others, no choice would increase order. The Peano kernel theorem [140] was proposed as a potential means to tackle such issues in [8]. One intent of future work is to scrutinise this option.

6.5 Other open problems and challenges

Within the topics considered in this thesis themselves, there remain numerous areas for future research, as we now detail.

6.5.1 Polyharmonic expansions

The main stumbling block in the practical application of polyharmonic eigenfunctions involves issues relating to the computation of eigenfunctions for moderate values of the parameter q. Both increased computational cost and susceptibility to round-off error may limit the scope of such techniques. Nonetheless, future work will aim to determine the impact of such issues, and establish potential means for fast, accurate computation for a larger range of q.

Herein we highlight one potential option for further scrutiny. As described in Section 3.2.4, the exponentially accurate estimates for eigenvalues provide sufficiently good approximations for even moderate values of the index n. A closer study may reveal higher-order terms in this expansion, thus circumventing the need for any iterative techniques (outside the first handful of values $n = 1, 2, \ldots$). Moreover, improved estimates for the coefficients of the individual eigenfunctions may provide an effective means to construct such functions without having to solve as many $q \times q$ linear systems.

Aside from this topic, it is of both practical and theoretical interest to determine whether subpolyharmonic eigenfunctions can be constructed in, for example, the equilateral triangle. Evidently, the lack of $\mathrm{L}^2(\Omega)$-orthogonal approximation bases in such a domain motivates this particular endeavour. In [155], Laplace eigenfunctions in the equilateral triangle were obtained by directly solving a particular boundary value problem. Such an approach may also be applicable in this setting.

6.5.2 Eckhoff's method

Eckhoff's method is extremely general in the sense that it can be applied to a large variety of orthogonal expansions with subtraction bases that can be chosen almost arbitrarily. In Chapter 5, we presented numerical results indicating how to best choose such a basis for modified Fourier expansions, yet we have no firm theory establishing this as the optimal choice.

[6] Some recent progress has been made in [121] as regards these issues. A number of suboptimal bounds were also given in [133].

This has potential practical consequences: increasing (or, indeed, guaranteeing) numerical stability renders the resulting methods more effective for a wider range of problems.

On a related topic, the numerical experiments of Section 5.9 exhibit at least one common feature: namely, the error levels off at some particular value. Often, especially for univariate functions, this value is close to machine precision. However, in some cases, it is several orders of magnitude larger (this feature is reasonably common for such approximations [32]). Addressing this barrier is a topic of future research. Least squares routines for ill-conditioned problems are amenable to a whole host of numerical tricks—including cut-offs and iterative refinement [32]—and thus offer a potential solution to this problem.

6.5.3 Applications

Aside from the boundary value problems studied in Chapter 4 and the integral equations of [38], modified Fourier expansions may have applications in a variety of other problems. For example, standard spectral methods for nonperiodic time-dependent partial differential equations often suffer from severe time-step restrictions [42, 142], thus necessitating the use of expensive implicit time-stepping routines. Conversely, Fourier methods for periodic problems offer better stability. Correspondingly, due to the similarity with the modified Fourier basis, there is reason to expect that modified Fourier methods may have application to nonperiodic problems. Needless to say, future work will not only consider time-dependent problems in tensor-product spatial domains, but also the development of modified Fourier–Galerkin approximations for triangular regions.

Outside of differential equations, the convergence acceleration techniques of Chapter 5 may also have application in image and signal processing. The Gegenbauer method (as discussed in Section 5.10) has been successfully applied to such problems [9, 10] (see also [33] and references therein). Yet its drawbacks, as discussed in Section 5.10, indicate that other methods may be better suited for such problems.

Univariate Laplace–Dirichlet and Laplace–Neumann expansions have also been considered in [150], where a numerical method was developed for the solution of Laplace and modified Helmholtz problems defined in convex polygonal domains. The fundamental component of this method is the so-called *global relation*, formulated in the complex plane, which, when discretised, provides a finite collection of Fourier coefficients from which the solution is recovered. In essence, this component of the method is a reconstruction problem: given the first N Fourier (or Fourier-like) coefficients of a function f, recover f to high accuracy. It is eminently possible that both the theory of modified Fourier expansions and the techniques for convergence acceleration have useful application in this area. This remains an object of future research.

6.6 Concluding thoughts

The steadily growing list of papers on the topic of modified Fourier expansions marks a significant attempt to provide new numerical methods for the approximation of functions in bounded domains and their applications, including the numerical solution of differential and integral equations. Many competitive algorithms exist for such problems, including finite element methods, spectral methods and wavelets, to name but a few. However, modified Fourier expansions have thus far proved fruitful in conferring a number of advantages over these more

standard techniques. Clearly, neither modified Fourier nor more established methods present a panacea for all problems. At the same time, in view of the potential benefits outlined previously, these are only the first promising steps towards the development of robust algorithms with a large range of potential applications.

Bibliography

[1] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions*. Dover, new edition, 1974.

[2] R. A. Adams. *Sobolev Spaces*. Academic Press, 1975.

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

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

[5] B. Adcock. Multivariate modified Fourier series and application to boundary value problems. *Num. Math.*, 115(4):511–552, 2010.

[6] B. Adcock. On the convergence of expansions in polyharmonic eigenfunctions. *Technical report NA2010/06, DAMTP, University of Cambridge*, 2010.

[7] B. Adcock and D. Huybrechs. Multivariate modified Fourier expansions. In E. Rønquist et al, editor, *Proceedings of the International Conference on Spectral and High Order Methods (to appear)*, 2010.

[8] B. Adcock, A. Iserles, and S. P. Nørsett. From high oscillation to rapid approximation II: Expansions in Birkhoff series. *Technical report NA2010/02, DAMTP, University of Cambridge*, 2010.

[9] R. Archibald, K. Chen, A. Gelb, and R. Renault. Improving tissue segmentation of human brain MRI through preprocessing by the Gegenbauer reconstruction method. *NeuroImage*, 20(1):489–502, 2003.

[10] R. Archibald and A. Gelb. A method to reduce the Gibbs ringing artifact in MRI scans while keeping tissue boundary integrity. *IEEE Transactions on Medical Imaging*, 21(4):305–319, 2002.

[11] A. Averbuch, M. Israeli, and L. Vozovoi. Analysis and applicaion of Fourier–Gegenbauer method to stiff differential equations. *SIAM J. Num. Anal.*, 33(5):1844–1863, 1996.

[12] A. Averbuch, M. Israeli, and L. Vozovoi. On a fast elliptic solver by a modified Fourier method. *Numer. Algorithms*, 15:287–313, 1997.

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

[14] I. Babuška and A. K. Aziz. *Survey lectures on the mathematical foundation of the finite element method.* In *Foundation of the Finite Element Method with Application to Partial Differential Equations*, ed. by A. K. Aziz, Academic Press, London, New York, pp. 3–359, 1972.

[15] P. Baldwin. Asymptotic estimates of the eigenvalues of a sixth order boundary value problem obtained by using global phase-integral methods. *Phil. Trans. Roy. Soc. London A*, 322:281–305, 1987.

[16] A. Barkhudaryan, R. Barkhudaryan, and A. Poghosyan. Asymptotic behavior of Eckhoff's method for Fourier series convergence acceleration. *Anal. Theory Appl.*, 23(3):228–242, 2007.

[17] G. Baszenski and F.-J. Delvos. *Accelerating the rate of convergence of bivariate Fourier expansions.* In: Chou CK et al (eds) Approximation theory IV. Academic Press, New York, pp 335–340, 1983.

[18] 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.

[19] G. Baszenski, F.-J. Delvos, and M. Tasche. A united approach to accelerating trigonometric expansions. *Comput. Math. Appl.*, 30(3–6):33–49, 1995.

[20] H. Bateman. *Higher Transcendental Functions.* Vol. 2, McGraw–Hill, New York, 1953.

[21] B. Baxter and A. Iserles. On the foundations of computational mathematics. *In Handbook of Numerical Analysis (F. Cucker ed.), Elsevier*, 11(3–35), 2003.

[22] R. E. Bellman. *Adaptive Control Processes: A Guided Tour.* Princeton University Press, 1961.

[23] H. E. Benzinger. The L^p behavior of eigenfunction expansions. *Trans. Amer. Math. Soc.*, 174:333–344, 1972.

[24] O.V. Besov, V.P. Il'in, and S.M. Nikol'skiĭ. *Integral Representations of Functions and Imbedding Theorems, Vol. I.* Scripta Series in Mathematics, edited by Mitchell H. Taibleson (Winston & Sons, Washington, D. C.); (Halsted Press [John Wiley & Sons], New York - Toronto, Ontario - London) (English translation from the Russian), 1979.

[25] G. D. Birkhoff. Boundary value and expansion problems of ordinary linear differential equations. *Trans. Amer. Math. Soc.*, 9(4):373–395, 1908.

[26] G. D. Birkhoff. On the asymptotic character of the solutions of certain linear differential equations containing a parameter. *Trans. Amer. Math. Soc.*, 9(2):219–231, 1908.

[27] A. Björck and V. Pereyra. Solution of Vandermonde systems of equations. *Math. Comp.*, 24:893–903, 1970.

[28] M. Bôcher. Introduction to the theory of Fourier's series. *Ann. Math.*, 7:81–152, 1906.

[29] B. D. Bojanov and G. Nikolov. Comparison of Birkhoff type quadrature formulae. *Math. Comp.*, 54:627–648, 1990.

[30] A. Boutayeb and E. Twizell. Numerical methods for the solution of special and general sixth-order boundary value problems with applications to benard layer eigenvalue problems. *Proc. Roy. Soc. London A*, 431:433–450, 1990.

[31] J. P. Boyd. *Chebyshev and Fourier Spectral Methods.* Springer–Verlag, 1989.

[32] J. P. Boyd. A comparison of numerical algorithms for Fourier Extension of the first, second, and third kinds. *J. Comput. Phys.*, 178:118–160, 2002.

[33] J. P. Boyd. Trouble with Gegenbauer reconstruction for defeating Gibbs phenomenon: Runge phenomenon in the diagonal limit of Gegenbauer polynomial approximations. *J. Comput. Phys.*, 204(1):253–264, 2005.

[34] J. P. Boyd. Acceleration of algebraically-converging Fourier series when the coefficients have series in powers of $1/n$. *J. Comput. Phys.*, 228:1404–1411, 2009.

[35] E. Braverman, B. Epstein, M. Israeli, and A. Averbuch. A fast spectral subtractional solver for elliptic equations. *J. Sci. Comput.*, 21(1):91–128, 2004.

[36] C. Brezinski. Extrapolation algorithms for filtering series of functions, and treating the Gibbs phenomenon. *Numer. Algorithms*, 36:309–329, 2004.

[37] E. O. Brigham. *The Fast Fourier Transform.* Prentice–Hall, Englewood Cliffs, NJ, 1974.

[38] H. Brunner, A. Iserles, and S. P. Nørsett. The computation of the spectra of highly oscillatory Fredholm integral operators. *J. Int. Eqn Appl. (to appear)*, 2010.

[39] O. P. Bruno, Y. Han, and M. M. Pohlman. Accurate, high-order representation of complex three-dimensional surfaces via Fourier continuation analysis. *J. Comput. Phys.*, 227(2):1094–1125, 2007.

[40] H. Bungartz and M. Griebel. A note on the complexity of solving Poisson's equation for spaces of bounded mixed derivatives. *J. Complexity*, 15:167–199, 1999.

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

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

[43] H. S. Carlshaw. A historical note on Gibbs' phenomenon in Fourier's series and integrals. *Bull. Amer. Math. Soc*, 8:420–424, 1925.

[44] T. Chan and T. P. Mathew. Domain Decomposition algorithms. *Acta Numerica*, pages 61–143, 1994.

[45] P. G. Ciarlet. *The finite element method for elliptic problems.* SIAM, 2002.

[46] J. A. Cochran and E. W. Hinds. Eigensystems associated with the complex-symmetric kernels of laser theory. *SIAM J. Appl. Math.*, 26:776–786, 1974.

[47] P. Concus and G. H. Golub. Use of fast direct methods for the efficient numerical solution of nonseparable elliptic equations. *SIAM J. Num. Anal.*, 10(6):1103–1120, 1973.

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

[49] T. A. Driscoll and B. Fornberg. A Padé-based algorithm for overcoming the Gibbs phenomenon. *Numer. Algorithms*, 26:77–92, 2001.

[50] M. Dubiner. Spectral methods on triangles and other domains. *J. Sci. Comput.*, 2:3–31, 1991.

[51] N. Dunford and J. T. Schwartz. *Linear Operators. Part III: Spectral Operators.* Wiley, 1971.

[52] K. S. Eckhoff. Accurate and efficient reconstruction of discontinuous functions from truncated series expansions. *Math. Comp.*, 61(204):745–763, 1993.

[53] K. S. Eckhoff. Accurate reconstructions of functions of finite regularity from truncated Fourier series expansions. *Math. Comp.*, 64(210):671–690, 1995.

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

[55] M. El-Gamel, J. R. Cannon, and A. I. Zayed. Sinc–Galerkin method for solving linear sixth-order boundary value problems. *Math. Comp.*, 73:1–19, 2004.

[56] L. C. Evans. *Partial Differential Equations.* AMS, 1998.

[57] G. Farin. *Curves and Surfaces for CAGD: A Practical Guide.* Academic Press, San Diego, 4th edition, 1997.

[58] C. Fefferman. On the convergence of multiple Fourier series. *Bull. Amer. Math. Soc.*, 77(5):744–745, 1971.

[59] C. Fefferman. On the divergence of multiple Fourier series. *Bull. Amer. Math. Soc.*, 77(2):191–195, 1971.

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

[61] W. Gautschi. Norm estimates for inverses of Vandermonde matrices. *Numer. Math.*, 23:337–347, 1975.

[62] A. Gelb and D. Gottlieb. The resolution of the Gibbs phenomenon for "spliced" functions in one and two dimensions. *Computers Math. Applic.*, 33(11):35–58, 1997.

[63] A. Gelb and J. Tanner. Robust reprojection methods for the resolution of the Gibbs phenomenon. *Appl. Comput. Harmon. Anal.*, 20:3–25, 2006.

[64] J. Gibbs. Fourier's series. *Letter in Nature*, 59:200, 1898.

[65] J. Gibbs. Fourier's series. *Letter in Nature*, 59:606, 1899.

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

[67] D. Gottlieb and S. A. Orszag. *Numerical Analysis of Spectral Methods: Theory and Applications.* Society for Industrial and Applied Mathematics, 1st edition, 1977.

[68] D. Gottlieb and C-W. Shu. On the Gibbs phenomenon II: Resolution properties of the Fourier method for discontinuous waves. *Comput. Methods Appl. Mech. Engrg*, 116:27–37, 1994.

[69] D. Gottlieb and C-W. Shu. On the Gibbs phenomenon IV: Recovering exponential accuracy in a subinterval from a Gegenbauer partial sum of a piecewise analytic function. *Math. Comp.*, 64(211):1081–1095, 1995.

[70] D. Gottlieb and C-W. Shu. On the Gibbs phenomenon V: Recovering exponential accuracy from collocation point values of a piecewise analytic function. *Numer. Math.*, 71(4):511–526, 1995.

[71] D. Gottlieb and C-W. Shu. On the Gibbs phenomenon III: Recovering exponential accuracy in a subinterval from a spectral partial sum of a piecewise analytic function. *SIAM J. Num. Anal.*, 33(1):280–290, 1996.

[72] D. Gottlieb and C-W. Shu. On the Gibbs' phenomenon and its resolution. *SIAM Rev*, 39(4):644–668, 1997.

[73] D. Gottlieb, C-W. Shu, A. Solomonoff, and H. Vandeven. On the Gibbs phenomenon I: Recovering exponential accuracy from the Fourier partial sum of a nonperiodic analytic function. *J. Comput. Appl. Math.*, 43(1–2):91–98, 1992.

[74] D. W. Gottlieb and C. W. Shu. General theory for the resolution of the Gibbs phenomenon. *Academia Nazionale Dei Lincey, ATTI Dei Convegni Lincey*, 147:39–48, 1998.

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

[76] M. Griebel and S. Knapek. Optimized tensor-product approximation spaces. *Constructive Approximation*, 16(4):525–540, 2000.

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

[78] 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.

[79] B.-Y. Guo, J. Shen, and L.-L. Wang. Generalized Jacobi polynomials/functions and their applications. *Appl. Numer. Math.*, 59:1011–1028, 2009.

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

[81] J. S. Hesthaven. Spectral penalty methods. *Appl. Numer. Math.*, 33:23–41, 2000.

[82] J. S. Hesthaven, S. Gottlieb, and D. Gottlieb. *Spectral methods for time–dependent problems*. Cambridge University Press, 2007.

[83] E. Hewitt and R. Hewitt. The Gibbs–Wilbraham phenomenon: An episode in Fourier analysis. *Hist. Exact. Sci.*, 21:129–160, 1979.

[84] N. J. Higham. *Accuracy and stability of numerical algorithms*. SIAM, 2nd edition, 2002.

[85] R. Hochmuth, S. Knapek, and G. Zumbusch. Tensor products of Sobolev spaces and applications. *Technical Report 685, SFB 256, Univ. Bonn*, 2000.

[86] D. Huybrechs. On the Fourier extension of non-periodic functions. *SIAM J. Numer. Anal.*, 47(6):4326–4355, 2010.

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

[88] D. Huybrechs, A. Iserles, and S. P. Nørsett. From high oscillation to rapid approximation V: The equilateral triangle. *IMA J. Num. Anal. (to appear)*, 2010.

[89] D. Huybrechs and S. Olver. Highly oscillatory quadrature. In B. Engquist, A. Fokas, E. Hairer, and A. Iserles, editors, *Highly Oscillatory Problems*, pages 25–50, Cambridge, 2009. Cambridge University Press.

[90] D. Huybrechs and S. Olver. Rapid function approximation by modified Fourier series. In B. Engquist, A. Fokas, E. Hairer, and A. Iserles, editors, *Highly Oscillatory Problems*, pages 51–72, Cambridge, 2009. Cambridge University Press.

[91] D. Huybrechs and S. Vandewalle. On the evaluation of highly oscillatory integrals by analytic continuation. *SIAM J. Num. Anal.*, 44(3):1026–1048, 2006.

[92] A. Iserles and S. P. Nørsett. On quadrature methods for highly oscillatory integrals and their implementation. *BIT*, 44:755–772, 2004.

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

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

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

[96] A. J. Jerri. *The Gibbs Phenomenon in Fourier Analysis, Splines and Wavelet Approximations*. Kluwer Academic, Kordrecht, The Netherlands, 1998.

[97] W. B. Jones and G. Hardy. Accelerating convergence of trigonometric approximations. *Math. Comp.*, 2(111):547–560, 1970.

[98] J.-H. Jung and B. D. Shizgal. Towards the resolution of the Gibbs phenomena. *J. Comput. Appl. Math.*, 161(1):41–65, 2003.

[99] J.-H. Jung and B. D. Shizgal. Generalization of the inverse polynomial reconstruction method in the resolution of the Gibbs phenomenon. *J. Comput. Appl. Math.*, 172(1):131–151, 2004.

[100] K. Kalimeris. *Initial and boundary value problems in two and three dimensions.* PhD thesis, University of Cambridge, 2010.

[101] K. Kalimeris and AS Fokas. The heat equation in the interior of an equilateral triangle. *Studies in Applied Mathematics*, 124(3):283–305, 2010.

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

[103] Y. Katznelson. *An Introduction to Harmonic Analysis.* Dover, 2nd edition, 1976.

[104] J. Keller. The scope of the image method. *Comm. Pure Appl. Math.*, 6:505–512, 1953.

[105] A. N. Kolmogorov. Über die bester Annäherung von Funktionen einer gegebenen Funktionenklasse. *Ann. Maths*, 37:107–110, 1936.

[106] T. Koornwinder. Two-variable analogues of the classical orthogonal polynomials. In R. A. Askey, editor, *Theory and Application of Special Functions*, pages 435–495, New York, 1975. Academic Press.

[107] T. W. Körner. *Fourier Analysis.* Cambridge University Press, 1988.

[108] M. G. Krein. On a special class of differential operators. *Doklady AN USSR*, 2:345–349, 1935.

[109] A. Krylov. On approximate calculations. *Lectures delivered in 1906 (in Russian). St Petersburg*, 1907.

[110] 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.

[111] G. Lamé. Mémoire sur la propagation de la chaluer dans les polyédres. *Journal de École Polytechnique*, 22:194–251, 1833.

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

[113] P. D. Lax. *Accuracy and resolution in the computation of solutions of linear and nonlinear equations.* Recent Advances in Numerical Analysis, Proc. Symposium Univ. of Wisconsin-Madison (C. de Boor and G. H. Golub, eds.), Academic Press, New York, 1978, pp. 107–117.

[114] B. M. Levitan and I. S. Sargsjan. *Introduction to Spectral Theory.* Number 39 in Translations of Mathematical Monographs. Amer. Math. Soc., Providence, RI, 1975.

[115] G. G. Lorentz, K. Jetter, and S. D. Riemenschneider. *Birkhoff Interpolation.* Addison–Wesley, London, 1983.

[116] G. G. Lorentz and K. L. Zeller. Birkhoff interpolation. *SIAM J. Num. Anal.*, 8(1):43–48, 1971.

[117] J. N. Lyness. Adjusted forms of the Fourier Coefficient Asymptotic Expansion and applications in numerical quadrature. *Math. Comp.*, 25:87–104, 1971.

[118] J. N. Lyness. Computational techniques based on the Lanczos representation. *Math. Comp.*, 28(125):81–123, 1974.

[119] J. N. Lyness. The calculation of trigonometric Fourier coefficients. *J. Comput. Phys.*, 54:57–73, 1984.

[120] N. W. McLachlan. *Theory and application of Mathieu functions.* Dover, 1964.

[121] J. M. Melenk. On the convergence of Filon quadrature. *J. Comput. Appl. Math. (to appear)*, 2009.

[122] A. Michelson. Fourier's series. *Letter in Nature*, 58:544–545, 1898.

[123] A. M. Minkin. Equiconvergence theorems for differential operators. *J. Math. Sci.*, 96:3631–3715, 1999.

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

[125] H.Z. Munthe-Kaas. On group Fourier analysis and symmetry preserving discretizations of PDEs. *J. Phys. A*, 39:5563, 2006.

[126] O. F. Naess and K. S. Eckhoff. A modified Fourier–Galerkin method for the Poisson and Helmholtz equations. *J. Sci. Comput.*, 17(1–4):529–539, 2002.

[127] M. A. Naimark. *Linear differential operators.* Harrap, 1968.

[128] A. Nersessian and A. Poghosyan. Bernoulli method in multidimensional case. *Preprint in ArmNIINTI 09.03.2000 N20 Ar-00 (in Russian)*, 2000.

[129] A. Nersessian and A. Poghosyan. Fast convergence of a polynomial-trigonometric interpolation. *Preprint in ArmNIINTI 07.07.2000 N45 Ar-00 (in Russian)*, 2000.

[130] A. Nersessian and A. Poghosyan. Asymptotic errors of accelerated two-dimensional trigonometric approximations. *Proceedings of the ISAAC Fourth Conference on Analysis. Yerevan, Armenia (G. A. Barsegian, H. G. W. Begehr, H. G. Ghazaryan, A. Nersessian eds), Yerevan, pp 70–78*, 2004.

[131] A. Nersessian and A. Poghosyan. The convergence acceleration of two-dimensional Fourier interpolation. *Armenian Journal of Mathematics*, 1:50–63, 2008.

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

[133] S. Olver. *Numerical Approximation of Highly Oscillatory Integrals.* PhD thesis, University of Cambridge, 2008.

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

[135] L. E. Payne and H. F. Weinberger. An optimal Poincaré inequality for convex domains. *Arch. Rat. Mech. Anal.*, 5:286–292, 1960.

[136] M. Pinkus. *N-widths in approximation theory.* Springer–Verlag, Berlin, 1968.

[137] M. A. Pinsky. The eigenvalues of an equilateral triangle. *SIAM J. Math. Anal.*, 11:819–827, 1980.

[138] A. Poghosyan. On an autocorrection phenomenon of the Krylov–Gottlieb–Eckhoff method. *Submitted*, 2006.

[139] A. Poghosyan. On an autocorrection phenomenon of the Eckhoff interpolation. *Submitted*, 2009.

[140] M. J. D. Powell. *Approximation theory and methods.* Cambridge University Press, 1981.

[141] M. Práger. Eigenvalues and eigenfunctions of the Laplace operator on an equilateral triangle. *Appl. Math.*, 43(4):311–320, 1998.

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

[143] A. Quarteroni and A. Valli. *Domain Decomposition Methods for Partial Differential Equations.* Oxford University Press, 1999.

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

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

[146] 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.

[147] 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.

[148] A. Sidi. *Practical extrapolation methods. Theory and applications. Cambridge Monographs on Applied and Computational Mathematics, 10.* Cambridge University Press, 2003.

[149] G. Sköllermo. A Fourier method for the numerical solution of Poisson's equation. *Math. Comp.*, 29:697–711, 1975.

[150] S. A. Smitheman, E. A. Spence, and A. S. Fokas. A spectral collocation method for the Laplace and modified Helmholtz equations in a convex polygon. *IMA J. Num. Anal. (to appear)*, 2009.

[151] E. A. Spence. *Boundary Value Problems for Linear Elliptic PDEs.* PhD thesis, University of Cambridge, 2010.

[152] H.M. Srivastava and J. Choi. *Series associated with the zeta and related functions.* Kluwer Academic, Kordrecht, The Netherlands, 2001.

[153] I. Stakgold. *Boundary value problems of mathematical physics. Volume 1.* Macmillan, 1968.

[154] M. H. Stone. A comparison of the series of Fourier and Birkhoff. *Trans. Amer. Math. Soc.*, 28(4):695–761, 1926.

[155] J. Sun and H. Li. Generalized Fourier transform on an arbitrary triangular domain. *Adv. Comput. Math.*, 22:223–248, 2005.

[156] E Tadmor. Filters, mollifiers and the computation of the Gibbs' phenomenon. *Acta Numerica*, 16:305–378, 2007.

[157] J. Tamarkin. Some general problems of the theory of ordinary linear differential equations and expansion of an arbitrary function in series of fundamental functions. *Math. Zeit.*, 27:1–54, 1928.

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

[159] L. N. Trefethen. *Spectral Methods in Matlab.* SIAM, 2000.

[160] H. Wilbraham. On a certain periodic function. *Cambridge and Dublin Math. J.*, 3:198–201, 1848.