Title
Simulating Nonlinear Faraday Waves on a Cylinder

Author
Qadeer, Saad

Publication Date
2018

Peer reviewed|Thesis/dissertation

Simulating Nonlinear Faraday Waves on a Cylinder

by

Saad Qadeer

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Applied Mathematics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Jon Wilkening, Chair
Professor John Strain
Professor Daniel Tataru
Professor Alexandre Bayen

Spring 2018

Simulating Nonlinear Faraday Waves on a Cylinder

Abstract

Simulating Nonlinear Faraday Waves on a Cylinder

by

Saad Qadeer

Doctor of Philosophy in Applied Mathematics

University of California, Berkeley

Professor Jon Wilkening, Chair

In 1831, Michael Faraday observed the formation of standing waves on the surface of a vibrating fluid body. Subsequent experiments have revealed the existence of a rich tapestry of patterned states that can be accessed by varying the frequency and amplitude of the vibration and have spurred vast research in hydrodynamics and pattern formation. These include linear analyses to determine the conditions for the onset of the patterns, weakly nonlinear studies to understand pattern selection, and dynamical systems approaches to study mode competition and chaos. Recently, there has been some work towards numerical simulations in various three-dimensional geometries. These methods however possess low orders of accuracy, making them unsuitable for nonlinear regimes.

We present a new technique for fast and accurate simulations of nonlinear Faraday waves in a cylinder. Beginning from a viscous potential flow model, we generalize the Transformed Field Expansion to this geometry for finding the highly non-local Dirichlet-to-Neumann operator (DNO) for the Laplace equation. A spectral method relying on Zernike polynomials is developed to rapidly compute the bulk potential. We prove the effectiveness of representing functions on the unit disc in terms of these polynomials and also show that the DNO algorithm possesses spectral accuracy, unlike a method based on Bessel functions.

The free surface evolution equations are solved in time using Picard iterations carried out by left-Radau quadrature. The results are in perfect agreement with the instability thresholds and surface patterns predicted for the linearized problem. The nonlinear simulations reproduce several qualitative features observed experimentally. In addition, by enabling one to switch between various nonlinear regimes, the technique allows a precise determination of the mechanisms triggering various experimental observations.

To

Papa, Mama and Ahsan

Contents

Contents **ii**

List of Figures **iii**

List of Tables **v**

1 Introduction **1**

2 On Zernike Polynomials **6**
2.1 Defining Zernike Polynomials 6
2.2 Accuracy of the Modal Representation 9
2.3 Computational Matters 13

3 Computing the Dirichlet-Neumann Operator on a Cylinder **21**
3.1 Introduction 21
3.2 Two Derivations of the Transformed Field Expansion 23
3.3 Solving the Poisson Equations 30
3.4 A Proof of Convergence 33
3.5 Numerical Results 38
3.6 Conclusion 40

4 Simulating Three-Dimensional Faraday Waves **44**
4.1 The Viscous Model 44
4.2 The Modal Formulation 46
4.3 Time Integration 48
4.4 Linear Results 52
4.5 Nonlinear Results 59
4.6 Conclusion 64

Bibliography **67**

List of Figures

1.1 The stability diagram obtained from the work of [32]. Here, α is the forcing acceleration and $p_{mn} = 4\omega_{mn}^2/\omega^2$ indicates the nature of the excitation: if the integer closest to $\sqrt{p_{mn}}$ is odd (even), the response is subharmonic with frequency $\omega/2$ (harmonic with frequency ω). 2
1.2 Parameter space close to the bimodal critical point for $(4, 3)$ and $(7, 2)$ from [12]. Here, f_0 and A are the frequency and amplitude of the forcing respectively. Mode competition occurs in the shaded regions with both modes co-existing and exchanging energy in the region labelled "periodic" before the onset of chaos. . . 3

2.1 L^2 errors in the Zernike polynomial-based modal representation. Observe that the decay decays exponentially against N, as proven in Theorem 2.2.2. 12
2.2 L^∞ errors in the Zernike representation vs N. The spread of the deviation on the unit disc for the case $k = 2, \alpha = 16.3475$ and $N = 30$ is also shown. 13
2.3 On left, the representation of ζ_{mn} in terms of Bessel functions is algebraic while, on right, the representation of $J_{mn}(a_{mn}\rho)e^{im\theta}$ in terms of Zernike polynomials is exponential. 14

3.1 Convergence of Neumann data vs TFE order K. The parameter choices are $M = 16$, $J = 20$, $N = 32$, $h = 0.5$ and $\epsilon = 0.2$. 39
3.2 Errors in Neumann data vs. TFE order K for different values of ϵ, for case (i) in Fig. 3.1. 40
3.3 Evolution of the interface at $t = 0, 1/80, 2/80, \ldots, 14/80$. We use $M = 4, J = 20, N = 40, K = 2, h = 0.5, \epsilon = 0.01$ and a time-step of $\Delta t = 1/1200$. 41

4.1 $L^\infty([0, 10])$ errors vs Δt for the test problem (4.22) solved by the s-point left-hand Radau Picard refinement scheme. The initial solution was computed by the RK4 method, followed by $(2s - 5)$ refinement sweeps. 51
4.2 Critical amplitudes vs. p. The solid lines are the theoretically calculated values while the circles indicate the results from the simulation. 55
4.3 Temporal and spectral profiles for case (i) in Table 4.2. The solid line is obtained from the damped Mathieu equation and the crosses from the numerical results. Observe that the time profile is almost perfectly sinusoidal with period $2T$. . . . 57

4.4 Temporal and spectral profiles for case (iv) in Table 4.2. Observe that the time profile is almost sinusoidal with period T. There is a constant offset and a small contribution of period $T/2$. 57
4.5 Temporal and spectral profiles for case (v) in Table 4.2. While the period is still $2T$ for this subharmonic response, the dominant contribution has period $2T/3$. . 58
4.6 Temporal and spectral profiles for case (vi) in Table 4.2. Similar to Figure 4.5, the dominant contribution period $T/2$ while the overall period is T. 58
4.7 Temporal and spectral profiles for case (vii) in Table 4.2. The dominant period is $2T/5$ with overall period $2T$. 59
4.8 Numerically computed instability parameters for the mode $(0, 1)$ compared against the nonlinear predictions from [37] and the linear regime from the previous section. 60
4.9 Amplitudes of secondary modes $|\beta_{mn}|$ vs. products of amplitudes $|\beta_{m_1n_1}\beta_{m_2n_2}|$ of the primary modes. On the left, we have $(1, 2) + (3, 1) \to (4, 2)$ and on the right, $(3, 1) + (3, 1) \to (6, 1)$. 62
4.10 Secondary modes may interact with primary or other secondary modes to produce tertiary modes. On the left, we have $(3, 1) + (6, 1) \to (9, 1)$ and on the right, $(6, 1) + (6, 1) \to (12, 1)$. 62
4.11 The time evolution of various modes with their spectral profiles. As expected, the response is sub-harmonic for $(3, 1)$ and $(9, 1)$ and harmonic for $(6, 1)$. 63
4.12 Evolution of modes $(0, 1)$ and $(3, 1)$ along with the amplitude envelopes $|\gamma_{mn}(t)|$. As predicted, the envelope for $(0, 1)$ converges to a non-zero fixed point while that for $(3, 1)$ converges to zero. 64
4.13 The phase-plane trajectory of the amplitude envelopes $\gamma_{mn}(t)$. Observe that it tends towards the stable fixed point $\approx (0.5672, 0)$. 65
4.14 The amplitude envelopes for both $(0, 1)$ and $(3, 1)$ decay with $(\omega/2\pi, \alpha/g) = (19.5971, 0.1056)$. This demonstrates the sensitivity of the results to small changes in parameter values close to the critical thresholds. 66

List of Tables

4.1 Comparison of critical amplitudes for various modes in different regimes. SH stands subharmonic and H for harmonic. 54
4.2 Comparison of critical amplitudes for various types of excitations undergone by the mode $(2, 2)$. SH stands subharmonic and H for harmonic. 56
4.3 Parameters for simultaneous excitation of a pair of modes via the fourth-order nonlinear model. 61

Acknowledgments

There are a great many people I need to thank but none more so than my adviser Jon Wilkening. I was extremely fortunate to be mentored by this incredible person who patiently allowed me to make my own mistakes while steadily expanding my horizons. Apart from opening up many wondrous sides of analysis and numerics to me, he directed a deluge of opportunities my way, including a memorable semester trip to ICERM. His technique, humility and thoughtfulness continue to inspire the budding academic in me.

I am also grateful to John Strain for his insightful comments on my work, for always seeking out the next challenge for me, and for his steady stream of witty one-liners; Daniel Tataru, for being extremely helpful at every stage of the journey and shaping my entire understanding of PDEs; and Alex Bayen, for his superb lectures on control theory and for imparting kind advice and assistance at various points.

I am obliged to Per-Olof Persson, for teaching me all I know about meshing and DG methods; Chris Rycroft, for his help on using POV-Ray; David Nicholls, for his encouragement and insights into the TFE method; Diane Henderson, for suggesting the dissipation mechanism and sharing her thorough understanding of the Faraday phenomenon; Didier Clamond, for helping me better appreciate the scope of my research; and Paul Milewski and Pavel Lushnikov, for their critical observations and suggestions for further development.

At the UC Berkeley mathematics department, I came into contact with many brilliant graduate students who helped me massively and made the journey more exciting. In particular, I would like to mention Chris Wong, Eric Hallman, Danny Hermes, Will Pazner and Rocky Foster. I am also indebted to the department staff members including, among others, Barb Waller, Jennifer Pinney, Marsha Snow, Vicky Lee, Igor Savine and Mark Jenkinson, who made me feel welcome and were a constant source of help and encouragement.

Living away from home for the better part of five years would have been an impossibility without the reassuring presence of my friends Mudassir Moosa, Saad Shaukat, Mohammad Ahmad and Osama Khan. I am grateful to every one of them for teaching me a great deal about life and friendship and for shaping crucial parts of my identity. I will forever treasure our long drives, food-binges and rambling conversations. I am also indebted to my cricket and soccer buddies for providing spaces where I could, for small periods of time, let go of work pressures and give full rein to my athletic chops.

Chapter 1

Introduction

In this section, we provide an introduction to the Faraday wave problem and briefly describe some of the theoretical, experimental and numerical studies that have been undertaken to explore its various facets.

The Linearized Problem

The parametric instability problem in vertically oscillating fluid containers originated with the work of Faraday in 1831 [20]. While observing the formation of standing waves on the surface of the fluid, he noted that the frequency of the standing waves was precisely half the frequency of the applied vibration (i.e., a subharmonic response). Mathiessen [33, 34] reproduced the experiment but found the frequencies to be identical (i.e., a harmonic response). The first theoretical effort to break the impasse was undertaken by Benjamin and Ursell [5]. Using the linearized form of Euler's equations, they showed that the evolution of the amplitude of each mode was governed by a Mathieu equation. The instability regions for this equation take the form of tongues in the forcing acceleration-frequency plane. For frequency ω and infinitesimal forcing, the frequency of the response could be $j\omega/2$, where j is a positive integer indexing the various tongues. In the absence of viscosity, however, the theory predicted exponential growth for larger forcing accelerations, thus limiting the validity of the results due to the linearity assumption.

The linear viscous problem was eventually satisfactorily addressed by Kumar and Tuckerman [32]. They proceeded by linearizing the Navier-Stokes equations and expanding the unknowns in terms of the normal modes of the system. Recognizing that the resulting equations for each modal amplitude were of the Floquet form, they were able to reduce the determination of the excitation parameters to an eigenvalue problem. Their work showed that the neutral stability curves from the inviscid problem were raised and smoothed out, so the accelerations required for the onset of instability had to be necessarily non-zero (see Figure 1.1). In addition, the frequency of the response was limited to $\omega/2$ (for odd-indexed tongues) and ω (for even-indexed tongues). In [7], this technique was extended to multiple-frequency forcing regimes.

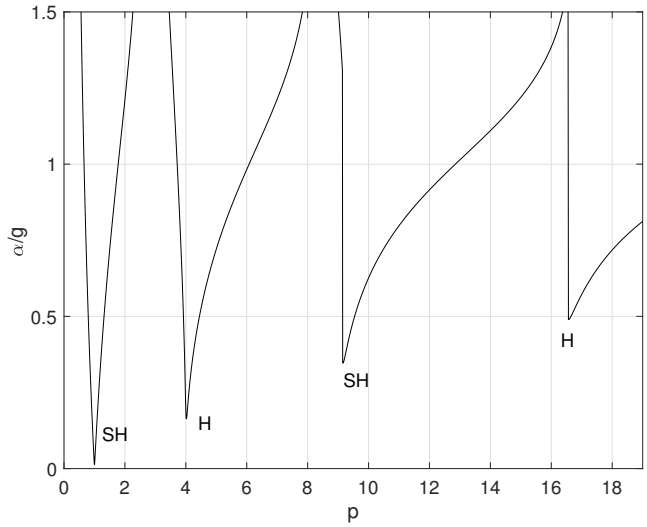

Figure 1.1: The stability diagram obtained from the work of [32]. Here, α is the forcing acceleration and $p_{mn} = 4\omega_{mn}^2/\omega^2$ indicates the nature of the excitation: if the integer closest to $\sqrt{p_{mn}}$ is odd (even), the response is subharmonic with frequency $\omega/2$ (harmonic with frequency ω).

Understanding Nonlinear Interactions

At the same time, considerable interest was drawn by the nonlinear interactions at play. By keeping terms upto the third order, Abramson et al [3] developed equations for the nonlinear self-interactions of small wave-number modes in a cylinder (specifically, the $(0, 1)$ and $(1, 1)$ modes) and obtained fairly good agreement with experiments. In addition, they also demonstrated hysteresis, whereby the shape of the excitation regions changed as the path through the parametric space was varied. Ockendon and Ockendon [47] allowed for a single mode to be excited to a finite but small amplitude. Using a series approximation up to the third order, they performed a stability analysis of the evolution equations for the slow amplitude modulation. Miles [37] developed an averaged Lagrangian approach, keeping terms up to the fourth order, and used it to obtain a Hamiltonian system for the amplitude envelopes. This was used to explore the bifurcation structure of the system. This work also factored in linear dissipation as well as exploring internal resonance between a pair of modes. It was later complemented by careful experiments in [25] that demonstrated the presence of $2 : 1$ resonance between a (sub-harmonically excited) Faraday wave and its sub and super-harmonics.

A remarkable series of experiments was performed in the 1980s. Gollub and Meyer [21] generated Faraday waves on a cylinder and observed the presence of multiple thresholds above the primary linear threshold. When crossed, in turn these led to pattern precession, azimuthal modulation and, finally, chaos. This work was followed by the seminal Ciliberto

and Gollub experiment [12]. They focused extensively on the intersection point of the first instability tongues for the $(4, 3)$ and $(7, 2)$ modes. By choosing parameter values inside the tongues close to the bimodal critical point, they observed amplitude modulations, followed by a bifurcation that doubled the period of the modulations and, finally, chaotic motion (see Figure 1.2). They made the case that the system was an instance of a strange attractor and calculated its dimension. In addition, they also developed a four-dimensional system by describing the amplitude dynamics on the slow time-scale using a pair of coupled Mathieu oscillators [11].

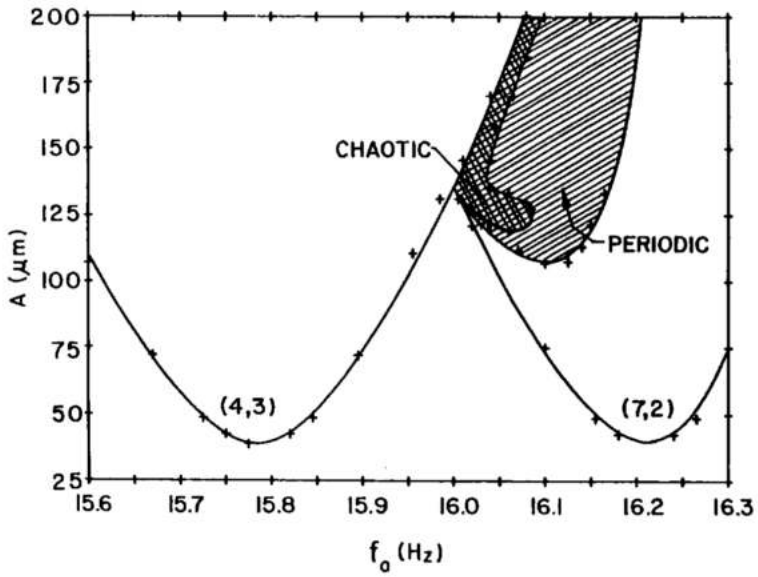

Figure 1.2: Parameter space close to the bimodal critical point for $(4, 3)$ and $(7, 2)$ from [12]. Here, f_0 and A are the frequency and amplitude of the forcing respectively. Mode competition occurs in the shaded regions with both modes co-existing and exchanging energy in the region labelled "periodic" before the onset of chaos.

This experiment generated immense interest in the pattern selection community. Meron and Procaccia [35, 36] used a center-manifold reduction and normal form theory to reduce the hydrodynamic model to a low-dimensional problem. With parameters chosen phenomenologically, they were able to explain various observations on the slow time-scale but were unable to address chaos; this was put down to an inability to enforce the no-flow boundary condition in the experiment. Umeki and Kambe [30] also addressed the problem. By using the averaged Lagrangian approach pioneered by Miles [37, 39], they were able to accurately demonstrate a Hopf bifurcation, a period doubling bifurcation and mixed modes. They also showed that chaotic dynamics were possible. Crawford et al [14] claimed that these findings may not be relevant to this regime and themselves explained the observations as a consequence of the $O(2)$ symmetry and the subharmonic excitation triggered by the Faraday phenomenon.

Numerical Results

Numerical simulations of the Faraday phenomenon are relatively rare in comparison to experimental or weakly nonlinear studies, with the majority of techniques catering to the two-dimensional case [10, 52]. The first three-dimensional simulation was carried out by O'Connor [48] using an arbitrary Lagrangian-Eulerian method on a domain with a square cross-section. A more thorough technique was developed by Perinet et al [49]. Their method relied on meshing the domain, using the projection method to solve the full Navier-Stokes system and advecting the free surface by using the immersed boundary method. Their results exhibited the formation of various nonlinear patterns such as squares and hexagons in a rectangular domain. However, this approach was ill-suited for a cylinder; indeed, an extension to this geometry was briefly addressed in [28] but the results differed vastly from experiments and were unable to exhibit Bessel functions, the primary surface wave pattern on a circular cylinder. Moreover, the low-order accuracy of the method and the Courant-Friedrichs-Levy time-step condition necessitated small step-sizes, increasing the computational cost. As a result, this method is infeasible for validating theoretical predictions and experimental observations.

Numerical studies of this phenomenon in a cylinder were also conducted by Jian et al [56]. They numerically integrated a third-order system for the free-surface amplitude and obtained amplitude modulations. Using a commercial software, Krishnaraja and Das also studied the Faraday wave problem on a cylinder and observed amplitude modulation and period doubling, tripling and quadrupling before wave-breaking [31]. Their results for the thresholds for the axisymmetric mode $(0, 1)$ were in fairly good agreement with the corresponding experiments reported in [15].

In this thesis, we develop a novel technique for the simulation of nonlinear Faraday waves in a circular cylinder. The underlying model is that of Dias et al [16], which incorporates linear damping with a potential flow. The latter property allows us to formulate the problem in terms of surface variables only, with the effects of the bulk accounted for by Laplace's equation. This equation in turn is solved by generalizing the Transformed Field Expansion (TFE) method to a three-dimensional cylinder. A spectrally accurate method is proposed for computing the solutions rapidly. After proving that this technique is highly suitable for interfaces of moderate height, we apply it to the Faraday phenomenon. The results show a high degree of agreement with both linear and nonlinear regimes. We are able to validate the technique by showing it accurately reproduces several theoretical predictions.

The thesis is structured as follows. We begin by introducing Zernike polynomials and present analytical and numerical results that demonstrate their effectiveness as an accurate representational tool on the unit disc. In particular, they are much more accurate than Bessel functions for representing generic smooth functions. We also show that they allow significant computational acceleration. In the next section, we extend the TFE method to our domain and develop the spectral method to solve it, with Zernike polynomials playing a key role. The accuracy and computational cost of this method are explored at length. The next section introduces the viscous potential flow model which is used, with the machinery assembled

earlier, to simulate Faraday waves. We provide details of the time integration method along with the results from various regimes. We conclude by discussing the limitations of the technique and identifying further extensions and applications.

Chapter 2

On Zernike Polynomials

In this chapter, we introduce a special class of functions on the unit disc. Zernike polynomials [2], as these functions are known, are of central importance to the results in the thesis. We begin by defining them conventionally and detailing the key properties that make them so attractive. To gain further insights into their effectiveness, we next present an alternate, sharper, characterization. This leads to useful approximation results that shall feature prominently in the analysis of subsequent algorithms. We also numerically demonstrate that Bessel approximations do not enjoy the same approximation properties. Finally, we take advantage of the structure of these functions to derive some results and optimize certain computations that are ubiquitous in a modal setting.

2.1 Defining Zernike Polynomials

We begin by discussing some key properties of Jacobi polynomials. These are a crucial ingredient in the construction of Zernike polynomials and hence many of their properties are carried forward in some shape or form.

Jacobi Polynomials

The Jacobi polynomials $\{P_n^{(\alpha,\beta)}\}_{n\geq 0}$ are a family of orthogonal polynomials on $[-1,1]$ with respect to the weight $w(x) = (1-x)^\alpha(1+x)^\beta$, for $\alpha, \beta > -1$. One way of defining them is by the Stieltjes procedure, which yields the following recurrence formulas [51]

$$\begin{aligned} P_0^{(\alpha,\beta)}(x) &= 1 \\ P_1^{(\alpha,\beta)}(x) &= \frac{1}{2}(\alpha+\beta+2)x + \frac{1}{2}(\alpha-\beta) \\ P_n^{(\alpha,\beta)}(x) &= (a_n^{(\alpha,\beta)}x - b_n^{(\alpha,\beta)})P_{n-1}^{(\alpha,\beta)}(x) + c_n^{(\alpha,\beta)}P_{n-2}^{(\alpha,\beta)}(x), \quad n \geq 2 \end{aligned} \tag{2.1}$$

where

$$\begin{aligned} a_n^{(\alpha,\beta)} &= \frac{(2n+\alpha+\beta-1)(2n+\alpha+\beta)}{2n(n+\alpha+\beta)} \\ b_n^{(\alpha,\beta)} &= \frac{(\beta^2-\alpha^2)(2n-1+\alpha+\beta)}{2n(n+\alpha+\beta)(2n-2+\alpha+\beta)} \\ c_n^{(\alpha,\beta)} &= \frac{(n-1+\alpha)(n-1+\beta)(2n+\alpha+\beta)}{n(n+\alpha+\beta)(2n-2+\alpha+\beta)}. \end{aligned}$$

These choices result in $P_n^{(\alpha,\beta)}(1) = \binom{n+\alpha}{n}$. This recurrence relation is a particularly efficient way of evaluating these polynomials and hence is used extensively in numerics. It is however not well-suited for deriving approximation estimates, which is why we consider another characterization. Consider the second-order linear differential equation

$$(1-x^2)y''(x) + (\beta-\alpha-(\alpha+\beta+2)x)y'(x) + n(n+\alpha+\beta+1)y(x) = 0.$$

It can be shown that this equation has exactly one non-trivial polynomial solution (up to scaling) for each $n \geq 0$ and that its degree is n [51]. Denote this solution by $y_n^{(\alpha,\beta)}(x)$. The differential equation admits the Sturm-Liouville form

$$-\frac{1}{w(x)}\left[(1-x^2)w(x)y'(x)\right]' = n(n+\alpha+\beta+1)y(x). \tag{2.2}$$

Define the operator Q by $(Qu)(x) = -\frac{1}{w(x)}\left[(1-x^2)w(x)u'(x)\right]'$. It follows from the self-adjoint nature of Q on $L^2([-1,1];w(x))$ that its eigenfunctions corresponding to distinct eigenvalues must be orthogonal. In particular, this holds for the family of polynomials $\{y_n^{(\alpha,\beta)}(x)\}_{n\geq 0}$. The uniqueness of orthogonal polynomials then shows that $y_n^{(\alpha,\beta)}(x)$ and $P_n^{(\alpha,\beta)}(x)$ are identical up to a scaling factor.

Jacobi polynomials also possess the property

$$\partial_x P_n^{(\alpha,\beta)}(x) = \left(\frac{\alpha+\beta+n+1}{2}\right) P_{n-1}^{(\alpha+1,\beta+1)}(x). \tag{2.3}$$

Using this repeatedly allows us to find $\partial_x^k P_n^{(\alpha,\beta)}(x)$ for $k \leq n$. Finally, we also have [2]

$$\int_{-1}^{1} \left[P_n^{(\alpha,\beta)}(x)\right]^2 (1-x)^\alpha (1+x)^\beta \, dx = \frac{2^{\alpha+\beta+1}}{2n+\alpha+\beta+1} \frac{\Gamma(n+\alpha+1)\Gamma(n+\beta+1)}{n!\Gamma(n+\alpha+\beta+1)}. \tag{2.4}$$

Zernike Polynomials and their Properties

Let D be the unit disc in the plane. For $m, n \in \mathbb{Z}$ with $n \geq 0$, set $\mu_{mn} = \sqrt{1+|m|+2n}$ and define

$$\zeta_{mn}(\rho,\theta) = \mu_{mn} P_n^{(0,|m|)}(2\rho^2-1)\rho^{|m|}e^{im\theta}$$

where (ρ, θ) are the polar coordinates. These are known as Zernike polynomials. In [2], these functions are indexed differently but we prefer this form as it leads to simpler expressions.

The weight $\rho^{|m|}$ and the factor μ_{mn} ensure that the family $\{\zeta_{mn}\}_{m\in\mathbb{Z},n\geq 0}$ is orthonormal on the unit disc with respect to the inner product

$$\langle v, w\rangle_{L^2(D)} = \frac{1}{\pi}\int_0^{2\pi}\int_0^1 \overline{v(\rho,\theta)} w(\rho,\theta)\ \rho\ d\rho\ d\theta. \tag{2.5}$$

Indeed,

$$\begin{aligned}
&\langle \zeta_{m_1n_1}, \zeta_{m_2n_2}\rangle_{L^2(D)} \\
= \ & \frac{\mu_{m_1n_1}\mu_{m_2n_2}}{\pi}\int_0^{2\pi}\int_0^1 P_{n_1}^{(0,|m_1|)}(2\rho^2-1)P_{n_2}^{(0,|m_2|)}(2\rho^2-1)\rho^{|m_1|+|m_2|+1}e^{i(m_2-m_1)\theta}\ d\rho\ d\theta \\
= \ & \frac{\mu_{m_1n_1}\mu_{m_2n_2}}{2}\delta_{m_1,m_2}\int_{-1}^1 P_{n_1}^{(0,|m_1|)}(x)P_{n_2}^{(0,|m_2|)}(x)\left(\frac{1+x}{2}\right)^{\frac{|m_1|+|m_2|}{2}}\ dx
\end{aligned}$$

where we used the substitution $x = 2\rho^2 - 1$. Continuing, we replace the m_2's by m_1's to get

$$\begin{aligned}
\langle \zeta_{m_1n_1}, \zeta_{m_2n_2}\rangle_{L^2(D)} &= \delta_{m_1,m_2}\frac{\mu_{m_1n_1}\mu_{m_1n_2}}{2^{|m_1|+1}}\int_{-1}^1 P_{n_1}^{(0,|m_1|)}(x)P_{n_2}^{(0,|m_1|)}(x)\,(1+x)^{|m_1|}\ dx \\
&= \delta_{m_1,m_2}\delta_{n_1,n_2}
\end{aligned}$$

because of the orthogonality of Jacobi polynomials and (2.4).

By the Stone-Weierstrass theorem, the algebra generated by $\{\zeta_{mn}\}$ is dense in $C(D)$, the space of continuous complex-valued functions on D. As $C(D)$ in turn is dense in $L^2(D)$, we conclude that $\{\zeta_{mn}\}$ forms an orthonormal basis for $L^2(D)$. This can be used to define the modal representation for $v \in L^2(D)$ as

$$v(\rho,\theta) = \sum_{m\in\mathbb{Z},n\geq 0}\langle \zeta_{mn}, v\rangle_{L^2(D)}\,\zeta_{mn}(\rho,\theta) \tag{2.6}$$

where the equality is to be interpreted in the L^2 sense. The orthonormality of the basis functions leads to Parseval's identity

$$\|v\|^2_{L^2(D)} = \sum_{m\in\mathbb{Z},n\geq 1}\left|\langle \zeta_{mn}, v\rangle_{L^2(D)}\right|^2.$$

In practice, however, we do not compute the entire modal representation but instead truncate it after a certain number of terms have been included. For positive integers M, N, define the projection

$$\mathcal{P}_{MN}v(\rho,\theta) = \sum_{|m|\leq M,n\leq N}\langle \zeta_{mn}, v\rangle_{L^2(D)}\,\zeta_{mn}(\rho,\theta). \tag{2.7}$$

From the remarks above, it is evident that $\|v - \mathcal{P}_{MN}v\|_{L^2(D)} \to 0$ as $M, N \to \infty$. A precise bound on the rate of convergence is determined in the next section.

2.2 Accuracy of the Modal Representation

The modal representation described in the previous section provides a powerful way to represent and manipulate a large class of functions on the unit disc D. However, its accuracy properties are as yet unclear. In particular, for the approximation error

$$e_{MN}(v) \;=\; \|v - \mathcal{P}_{MN} v\|_{L^2(D)}\,, \tag{2.8}$$

we would like to study the effect of changing the parameters M and N and bound its rate of decay. This is precisely the goal of this section. We first establish a characterization of Zernike polynomials and use it to determine the rate of convergence. This is followed by numerical results that further support the claim.

Decay of Approximation Error

For any integer $r \geq 0$, let $H^r(D)$ denote the corresponding Sobolev space on D.

Theorem 2.2.1 *Define the map*

$$Lu = -\rho^{-1}\partial_\rho\left(\rho(1-\rho^2)\partial_\rho u\right) - \rho^{-2}\partial_\theta^2 u.$$

Then,

(a) L is bounded from $H^{l+2}(D)$ to $H^l(D)$ for any integer $l \geq 0$.

(b) the Zernike polynomials $\{\zeta_{mn}\}$ are eigenfunctions of L with eigenvalues $\lambda_{mn} = (|m| + 2n)(|m| + 2n + 2)$.

Proof:

(a) This follows easily from rewriting

$$Lu = -\Delta u + (\rho^2\partial_\rho^2 + 3\rho\partial_\rho)u$$

and using the fact that both operators above are bounded from $H^{l+2}(D)$ to $H^l(D)$ for any integer $l \geq 0$.

(b) We solve the eigenvalue problem $Lu = \lambda u$ by separation of variables. Writing $u(\rho,\theta) = R(\rho)T(\theta)$ gives

$$\frac{\rho}{R(\rho)}\partial_\rho\left(\rho(1-\rho^2)R'(\rho)\right) + \lambda\rho^2 = -\frac{T''(\theta)}{T(\theta)}.$$

Setting both sides equal to a constant μ gives $T''(\theta) + \mu T(\theta) = 0$. The conditions $T(0) = T(2\pi)$ and $T'(0) = T'(2\pi)$ lead to $\mu = m^2$ for integers $m \geq 0$ and $T_m(\theta) = a_m e^{im\theta} + b_m e^{-im\theta}$. The equation for R_m then reads

$$\frac{1}{\rho}\partial_\rho\left(\rho(1-\rho^2)R_m'(\rho)\right) \;=\; (m^2\rho^{-2} - \lambda)R_m(\rho). \tag{2.9}$$

Next, we use the transformation $x = 2\rho^2 - 1$ and set $X_m(x) = R_m(\rho)$. Note that $\partial_\rho = 4\rho\partial_x$ so we have

$$\begin{aligned} \rho(1-\rho^2)R'_m(\rho) &= 4\rho^2(1-\rho^2)X'_m(x) \\ &= (1-x)^2X'_m(x) \\ \Rightarrow \frac{1}{\rho}\partial_\rho\left(\rho(1-\rho^2)R'_m(\rho)\right) &= 4\partial_x((1-x^2)X'_m(x)). \end{aligned}$$

Plugging this in (2.9) yields

$$\partial_x((1-x^2)X'_m(x)) = \left(\frac{m^2}{2(1+x)} - \frac{\lambda}{4}\right)X_m(x). \tag{2.10}$$

Decomposing

$$\frac{m^2}{2(1+x)} = \left[\frac{m^2}{4}\left(\frac{1-x}{1+x}\right) - \frac{|m|}{2}\right] + \left[\frac{m^2}{4} + \frac{|m|}{2}\right]$$

allows us to write (2.10) as

$$\partial_x\left((1-x^2)X'_m(x)\right) - \frac{|m|}{2}\left(\frac{|m|}{2}\left(\frac{1-x}{1+x}\right) - 1\right)X_m(x) = \left(\frac{m^2}{4} + \frac{|m|}{2} - \frac{\lambda}{4}\right)X_m(x). \tag{2.11}$$

Observe next that

$$\begin{aligned} & \frac{1}{(1+x)^{|m|/2}}\partial_x\left((1-x^2)(1+x)^{|m|}\partial_x\left(\frac{X_m(x)}{(1+x)^{|m|/2}}\right)\right) \\ = \; & \frac{1}{(1+x)^{|m|/2}}\partial_x\left((1-x^2)(1+x)^{|m|/2}X'_m(x) - \frac{|m|}{2}(1-x)(1+x)^{|m|/2}\right) \\ = \; & \partial_x((1-x^2)X'_m(x)) + \frac{|m|}{2}(1-x)X'_m(x) - \frac{|m|}{2(1+x)^{|m|/2}}\left[-(1+x)^{|m|/2}X_m(x) + \right. \\ & \left.(1-x)(1+x)^{|m|/2}X'_m(x) + \frac{|m|}{2}(1-x)(1+x)^{|m|/2-1}X_m(x)\right] \\ = \; & \partial_x\left((1-x^2)X'_m(x)\right) - \frac{|m|}{2}\left(\frac{|m|}{2}\left(\frac{1-x}{1+x}\right) - 1\right)X_m(x) \end{aligned}$$

which is exactly the left hand side of (2.11). Thus,

$$\frac{1}{(1+x)^{|m|/2}}\partial_x\left((1-x^2)(1+x)^{|m|}\partial_x\left(\frac{X_m(x)}{(1+x)^{|m|/2}}\right)\right) = \left(\frac{m^2}{4} + \frac{|m|}{2} - \frac{\lambda}{4}\right)X_m(x).$$

Comparing this with (2.2), we deduce that this is the eigenvalue problem associated with $P_n^{(0,|m|)}(x)(1+x)^{|m|/2}$ with eigenvalues $-n(n+|m|+1)$. This gives

$$\begin{aligned} \lambda &= m^2 + 2|m| + 4n(n+|m|+1) \\ &= (|m|+2n)(|m|+2n+2) \end{aligned}$$

We conclude that $P_n^{(0,|m|)}(x)(1+x)^{|m|/2}e^{im\theta}$, which is precisely ζ_{mn} up to scaling, is an eigenfunction of L with the desired eigenvalue.

Note that the operator L defined in the last theorem is self-adjoint in $L^2(D)$. This fact has a crucial bearing on the following approximation result.

Theorem 2.2.2 *Let s be a positive real number and let $v \in H^s(D)$. For $M, N \geq 0$, let $\mathcal{P}_{MN}v$ be the projection of v on $\{\zeta_{mn}\}$ as defined in the previous section. Then,*

$$e_{MN}(v) = \|v - \mathcal{P}_{MN}v\|_{L^2(D)} \lesssim \min(M, 2N)^{-s} \|v\|_{H^s(D)}.$$

Proof: We follow the standard argument presented in [6]. First suppose that $s = 2k$ for some integer $k \geq 1$ and let $\Lambda_{MN} = \{(m, n) : |m| > M \text{ or } n > N\}$. Note then from (2.6) and (2.7) that

$$v - \mathcal{P}_{MN}v = \sum_{m,n \in \Lambda_{MN}} \langle \zeta_{mn}, v \rangle_{L^2(D)} \zeta_{mn}.$$

We have by Theorem 2.2.1(b) and the self-adjoint nature of L

$$\langle \zeta_{mn}, v \rangle_{L^2(D)} = \lambda_{mn}^{-k} \left\langle L^k \zeta_{mn}, v \right\rangle_{L^2(D)} = \lambda_{mn}^{-k} \left\langle \zeta_{mn}, L^k v \right\rangle_{L^2(D)}.$$

From Parseval's identity, it follows that

$$\begin{aligned}
\|v - \mathcal{P}_{MN}v\|_{L^2(D)}^2 &= \sum_{m,n \in \Lambda_{MN}} |\langle \zeta_{mn}, v \rangle_{L^2(D)}|^2 \\
&= \sum_{m,n \in \Lambda_{MN}} \lambda_{mn}^{-2k} |\left\langle \zeta_{mn}, L^k v \right\rangle_{L^2(D)}|^2 \\
&\lesssim \min(M, 2N)^{-4k} \sum_{m,n \in \Lambda_{MN}} |\left\langle \zeta_{mn}, L^k v \right\rangle_{L^2(D)}|^2 \\
&\leq \min(M, 2N)^{-4k} \left\| L^k v \right\|_{L^2(D)}^2
\end{aligned}$$

where we used the fact that $\lambda_{mn} = (|m|+2n)(|m|+2n+2) \geq (\min(M, 2N))^2$ for $m, n \in \Lambda_{MN}$. By Theorem 2.2.1(a), we have

$$\left\| L^k v \right\|_{L^2(D)} \lesssim \|v\|_{H^{2k}(D)}$$

so we have the desired result in the case that $s = 2k$. In addition, observe that

$$\|v - \mathcal{P}_{MN}v\|_{L^2(D)}^2 = \sum_{m,n \in \Lambda_{MN}} |\langle \zeta_{mn}, v \rangle_{L^2(D)}|^2 \leq \|v\|_{L^2(D)}^2.$$

Finally, let s be a positive real number and choose an integer k such that $2k > s$. The operator $(I - \mathcal{P}_{MN})$ is continuous from $L^2(D)$ to $L^2(D)$ with norm 1 and from $H^{2k}(D)$ to $L^2(D)$ with norm $\lesssim \min(M, 2N)^{-2k}$. Interpolating between these, we deduce that it is bounded from $H^s(D)$ to $L^2(D)$ with norm $\lesssim \min(M, 2N)^{-s}$. Thus,

$$e_{MN}(v) = \|v - \mathcal{P}_{MN}v\|_{L^2(D)} \lesssim \min(M, 2N)^{-s} \|v\|_{H^s(D)}.$$

Theorem 2.2.2 shows that the rate of error decay is faster than any power of $\min(M, 2N)^{-1}$. This is commonly termed spectral accuracy [9]. Also note that if v has a finite highest angular frequency m' so that $\langle \zeta_{mn}, v \rangle_{L^2(D)} = 0$ for $m > m'$, then applying the same argument as above shows that the error decay occurs at rate N^{-s}, provided $M \geq m'$.

Numerical Results

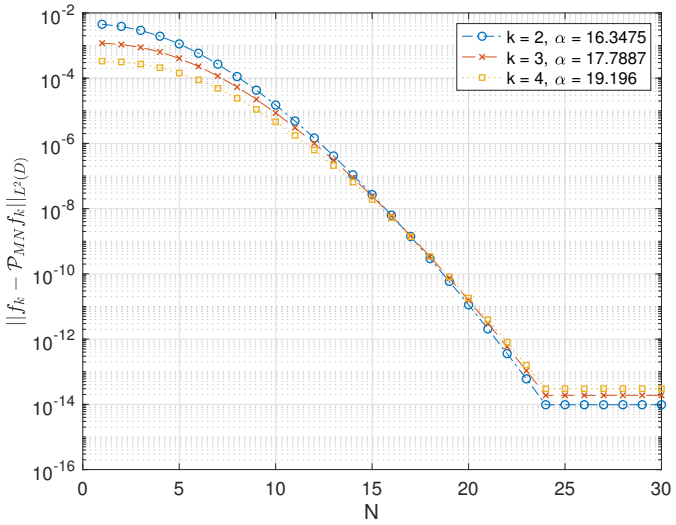

Figure 2.1: L^2 errors in the Zernike polynomial-based modal representation. Observe that the decay decays exponentially against N, as proven in Theorem 2.2.2.

We next confirm the spectral accuracy of the modal representation numerically. For $k \geq 0$, let

$$f(\rho, \theta) = e^{-\alpha\rho^2} \rho^k \cos(k\theta).$$

The coefficients $\langle \zeta_{mn}, f \rangle_{L^2(D)}$ in the Zernike representation of f can computed by using a high-order Jacobi-Gauss quadrature rule. Theorem 2.2.2 predicts that the error $\|f - \mathcal{P}_{MN} f\|_{L^2(D)}$ will decay faster than any power of N^{-1}, provided that $M \geq k$. Figure 2.1 confirms the spectral decay for multiple values of k and α with $1 \leq N \leq 30$ and $M = 16$.

In order to show that this representation avoids spurious behavior, in Figure 2.2 we present the L^∞ errors for the cases considered in Figure 2.1. These errors are computed by sampling the functions on a fine mesh consisting of 14230 points. It is clear that representation is spectrally accurate even in this norm. In addition, we also show that the spread of the deviation on the mesh for one of the cases with $N = 30$ is fairly uniform.

Finally, we demonstrate the superiority of Zernike polynomials over Bessel functions for representational purposes. First, let J_m be the mth Bessel function of order zero and a_{mn}

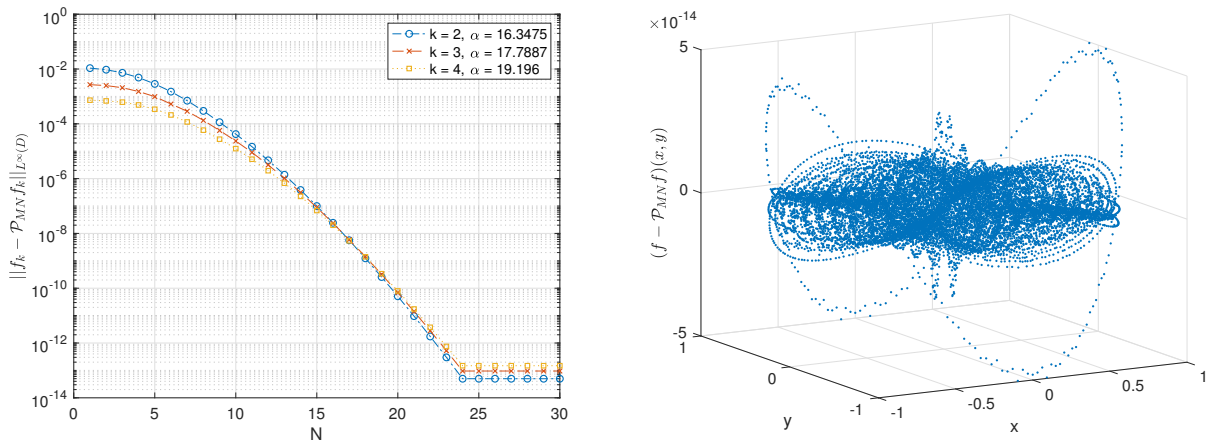

Figure 2.2: L^∞ errors in the Zernike representation vs N. The spread of the deviation on the unit disc for the case $k = 2, \alpha = 16.3475$ and $N = 30$ is also shown.

the nth zero of $J'_m(\cdot)$. It follows from the orthogonality of $\{J_m(a_{mn}\rho)\}_{n\geq 1}$ that any function g on the unit disc can then be represented as

$$g(\rho,\theta) = \sum_{m\geq 0, n\geq 1} \beta_{mn} J_m(a_{mn}\rho) e^{im\theta}$$

where

$$\beta_{mn} = \left[2\pi \int_0^1 (J_m(a_{mn}\rho))^2 \ \rho \ d\rho\right]^{-1} \int_0^{2\pi}\int_0^1 g(\rho,\theta) J_m(a_{mn}\rho) e^{-im\theta} \ \rho \ d\rho \ d\theta.$$

In order to compare the two techniques, we represent functions from one family in terms of the other and vice versa. Figure 2.3 shows the results for the L^∞ norm. The plots show that the error decay for the Bessel function representation is algebraic, as opposed to the spectral accuracy possessed by Zernike polynomials. An intuitive reason for this is that Bessel functions are not as oscillatory as Zernike polynomials near $\rho = 1$ and, as a result, are less accurate close to the outer boundary. A useful analog is the comparison of a Fourier sine-series on $[0, \pi]$ with a Chebyshev expansion. The zeros of the latter cluster near the boundaries like $1/n^2$, where n is the mode number, and yield spectrally accurate representations. Meanwhile, the zeros of the former cluster like $1/n$ and lead to algebraic decay of mode amplitudes.

2.3 Computational Matters

In this section, we derive a result concerning Zernike polynomials that illustrates their particular amenability to Galerkin methods. In addition, we present techniques to efficiently perform certain computations using the modal representation that will help us greatly going forward.

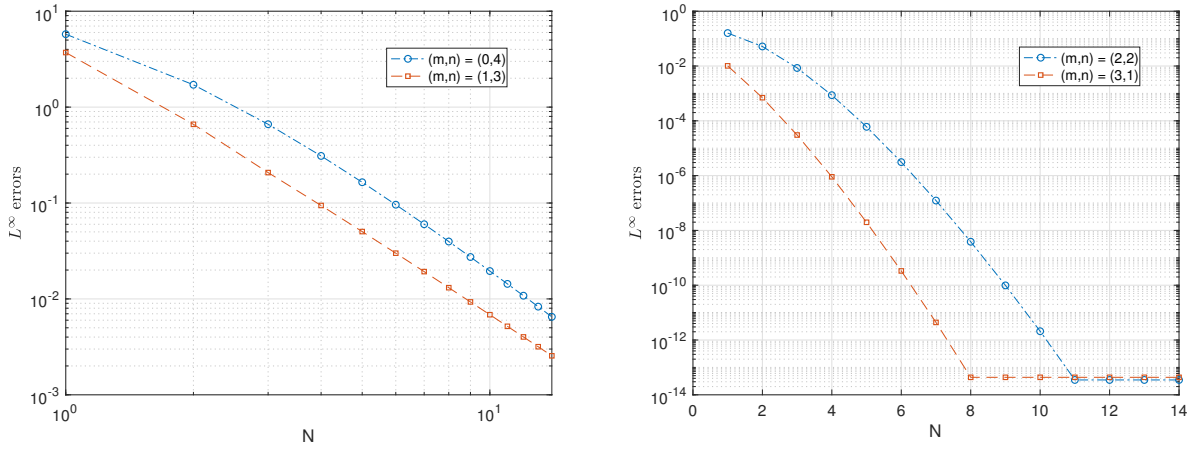

Figure 2.3: On left, the representation of ζ_{mn} in terms of Bessel functions is algebraic while, on right, the representation of $J_{mn}(a_{mn}\rho)e^{im\theta}$ in terms of Zernike polynomials is exponential.

The Stiffness Matrix

In Galerkin methods, one frequently needs to set up the mass and stiffness matrices with respect to the chosen basis functions. We here show that these computations are particularly simple for Zernike polynomials, consequently making them well-suited for use in problems on circular domains.

Note first that the orthonormality of these functions ensures that the mass matrix will merely be the identity. The stiffness matrix, on the other hand, requires some work. Fortunately, as we shall see, the structure of these functions enables us to derive a simple closed form expression for this matrix as well.

Let $m, m', n, n' \in \mathbb{Z}$ with $n, n' \geq 1$ and consider the typical stiffness term

$$A_{m'n',mn} = \frac{1}{\pi}\int_0^{2\pi}\int_0^1 \overline{\nabla\zeta_{m'n'}(\rho,\theta)}\cdot\nabla\zeta_{mn}(\rho,\theta)\ \rho\ d\rho\ d\theta$$

where $\nabla = (\partial_\rho, \rho^{-1}\partial_\theta)$ is the gradient operator in polar coordinates. We have

$$\nabla\zeta_{mn}(\rho,\theta) = \mu_{mn}e^{im\theta}\begin{pmatrix} 4\partial_\rho P_n^{(0,|m|)\prime}(2\rho^2-1)\rho^{|m|+1} + |m|P_n^{(0,|m|)}(2\rho^2-1)\rho^{|m|-1} \\ imP_n^{(0,|m|)}(2\rho^2-1)\rho^{|m|-1}\end{pmatrix}. \quad (2.12)$$

As a result,

$$\begin{aligned}&\frac{1}{\pi}\int_0^{2\pi}\int_0^1 \overline{\nabla\zeta_{m'n'}(\rho,\theta)}\cdot\nabla\zeta_{mn}(\rho,\theta)\ \rho\ d\rho\ d\theta \\ =\ &\mu_{m'n'}\mu_{mn}\frac{1}{\pi}\int_0^{2\pi}e^{i(m-m')\theta}\ d\theta\left\{\int_0^1 \underbrace{16P_{n'}^{(0,|m'|)\prime}(2\rho^2-1)P_n^{(0,|m|)\prime}(2\rho^2-1)\rho^{|m'|+|m|+3}\ d\rho}_{(1)}\right.\end{aligned}$$

$$
\begin{aligned}
&+\underbrace{4\int_0^1 |m'|P_{n'}^{(0,|m'|)}(2\rho^2-1)P_n^{(0,|m|)\prime}(2\rho^2-1)\rho^{|m'|+|m|+1}\,d\rho}_{\text{(2a)}}\\
&+\underbrace{4\int_0^1 |m|P_{n'}^{(0,|m'|)\prime}(2\rho^2-1)P_n^{(0,|m|)}(2\rho^2-1)\rho^{|m'|+|m|+1}\,d\rho}_{\text{(2b)}}\\
&+\underbrace{(|mm'|+mm')\int_0^1 P_{n'}^{(0,|m'|)}(2\rho^2-1)P_n^{(0,|m|)}(2\rho^2-1)\rho^{|m'|+|m|-1}\,d\rho}_{\text{(3)}}\Bigg\}
\end{aligned}
\tag{2.13}
$$

Note first that $\frac{1}{\pi}\int_0^{2\pi} e^{i(m-m')\theta}\,d\theta = 2\delta_{m,m'}$ so we can safely take $m = m'$ in the subsequent calculations. We use the substitution $x = 2\rho^2 - 1$ in (1) to obtain

$$
\begin{aligned}
(1) \quad : \quad & \frac{16}{4(2^{|m|+1})}\int_{-1}^{1} P_{n'}^{(0,|m|)\prime}(x)P_n^{(0,|m|)\prime}(x)(1+x)^{|m|+1}\,dx\\
= \quad & 2^{1-|m|}\left[P_{n'}^{(0,|m|)}(x)P_n^{(0,|m|)\prime}(x)(1+x)^{|m|+1}\right]_{-1}^{1} -\\
& 2^{1-|m|}\int_{-1}^{1} P_{n'}^{(0,|m|)}(x)\underbrace{\left[P_n^{(0,|m|)\prime\prime}(x)(1+x)+(|m|+1)P_n^{(0,|m|)\prime}(x)\right]}_{\text{(A)}}(1+x)^{|m|}\,dx
\end{aligned}
$$

where we integrated by parts. Observe that the degree of the expression (A) above is at most $(n-1)$. Hence, if $n' \geq n$, the above integral evaluates to zero. Similarly, if $n > n'$, we can switch the functions while integrating by parts and still get zero. Thus, the only contributions are from the first term

$$
(1) \quad : \quad 2^{1-|m|}\begin{cases} P_{n'}^{(0,|m|)}(1)P_n^{(0,|m|)\prime}(1)(2)^{|m|+1} & \text{if } n' \geq n \\ P_{n'}^{(0,|m|)\prime}(1)P_n^{(0,|m|)}(1)(2)^{|m|+1} & \text{if } n > n' \end{cases} \tag{2.14}
$$

We also have

$$
P_n^{(0,|m|)}(1) = 1, \quad P_n^{(0,|m|)\prime}(1) = \frac{n(n+|m|+1)}{2}.
$$

Setting $\gamma_{n,n'} = \min(n, n')$ allows us to write (2.14) as $2\gamma_{n,n'}(\gamma_{n,n'} + |m| + 1)$.

Using the same substitution as above in (2a) and (2b) yields

$$
(2a) + (2b) \quad : \quad 2^{-|m|}|m|\int_{-1}^{1}\left[P_{n'}^{(0,|m|)}(x)P_n^{(0,|m|)\prime}(x) + P_{n'}^{(0,|m|)\prime}(x)P_n^{(0,|m|)}(x)\right](1+x)^{|m|}\,dx \tag{2.15}
$$

Leaving this as it is for now, we next turn to (3). Using the same substitution and integrating by parts yields

$$
(3) \quad : \quad \frac{2m^2}{4(2^{|m|-1})}\int_{-1}^{1} P_{n'}^{(0,|m|)}(x)P_n^{(0,|m|)}(x)(1+x)^{|m|-1}\,dx
$$

$$
\begin{aligned}
&= 2^{-|m|}m^2\left[P_{n'}^{(0,|m|)}(x)P_n^{(0,|m|)}(x)\frac{(1+x)^{|m|}}{|m|}\right]_{-1}^{1} - \\
&\quad 2^{-|m|}m^2\int_{-1}^{1}\left[P_{n'}^{(0,|m|)}(x)P_n^{(0,|m|)}(x)\right]'\frac{(1+x)^{|m|}}{|m|}\,dx \\
&= |m| - \underbrace{2^{-|m|}|m|\int_{-1}^{1}\left[P_{n'}^{(0,|m|)}(x)P_n^{(0,|m|)\prime}(x) + P_{n'}^{(0,|m|)\prime}(x)P_n^{(0,|m|)}(x)\right](1+x)^{|m|}\,dx}_{\text{(B)}}
\end{aligned}
$$

Observe that the integrals (2.15) and (B) are identical but have opposite signs. Consequently, when we add (1), (2a), (2b) and (3) as they appear in (2.13), they cancel out and we obtain the concise expression

$$A_{m'n',mn} = 2\delta_{m,m'}\mu_{mn'}\mu_{mn}(2\gamma_{n,n'}(\gamma_{n,n'}+|m|+1)+|m|). \tag{2.16}$$

Optimizing Computations in Modal Space

We have seen that Zernike polynomials are able to represent functions on the disc in an accurate and stable manner. However, their usefulness to spectral methods also depends critically on their amenability to basic operations, such as multiplying two functions whose modal representations are known. The obvious procedure for doing this is to evaluate the modal representations at certain points (chosen according to some quadrature rule), multiply them together and convert those values back. This however proves to be costly. Indeed, a naive implementation of this method could end up requiring $O(M^2N^2)$ operations.

We instead present a technique that will enable us to execute such operations in $O(M(M+N)(N+\log(M)))$ complexity. This will result in a significant speed up and make the use of Zernike polynomials even more appealing.

Let M, N be positive integers and suppose we have functions f and g such that

$$f(\rho,\theta) = \sum_{-M<m\le M}\sum_{0\le n\le N}\alpha_{mn}\zeta_{mn}(\rho,\theta), \quad g(\rho,\theta) = \sum_{-M<m\le M}\sum_{0\le n\le N}\beta_{mn}\zeta_{mn}(\rho,\theta).$$

We shall consider three types of operations: (i) fg ; (ii) $\nabla f\cdot\nabla g$ and (iii) $(\Delta f)g$. By associativity, operations involving more than two functions can be reduced into these types.

For (i), the modal coefficients are given by

$$
\begin{aligned}
\langle\zeta_{mn}, fg\rangle_{L^2(D)} &= \frac{1}{\pi}\sum_{-M<m_1,m_2\le M}\sum_{0\le n_1,n_2\le N}\alpha_{m_1,n_1}\beta_{m_2,n_2}\mu_{m_1n_1}\mu_{m_2n_2}\mu_{mn}\int_0^{2\pi}e^{i(m_1+m_2-m)\theta}\,d\theta \\
&\quad \int_0^1 P_{n_1}^{(0,|m_1|)}(2\rho^2-1)P_{n_2}^{(0,|m_2|)}(2\rho^2-1)P_n^{(0,|m|)}(2\rho^2-1)\rho^{|m_1|+|m_2|+|m|+1}\,d\rho \\
&= \frac{1}{2}\sum_{-M<m_1,m_2\le M}\sum_{0\le n_1,n_2\le N}\alpha_{m_1,n_1}\beta_{m_2,n_2}\mu_{m_1n_1}\mu_{m_2n_2}\mu_{mn}\delta_{m_1+m_2,m}
\end{aligned}
$$

$$\int_{-1}^{1} P_{n_1}^{(0,|m_1|)}(r) P_{n_2}^{(0,|m_2|)}(r) P_n^{(0,|m|)}(r) \left(\frac{1+r}{2}\right)^{\frac{|m_1|+|m_2|+|m|}{2}} \, dr$$

where we used the substitution $r = 2\rho^2 - 1$. Observe that the highest degree in the integrand above is $(3N + M)$. Let $\{(r_j, \sigma_j)\}$ be the $N_g = (3N + M)/2 + 1$ point Gauss-Legendre quadrature scheme on $[-1, 1]$. This scheme is guaranteed to evaluate all polynomials up to degree $(3N + M + 1)$ so it is well-suited for our task. Thus, we have

$$\begin{aligned}\langle \zeta_{mn}, fg \rangle_{L^2(D)} &= \frac{1}{2} \sum_{-M<m_1,m_2\le M} \sum_{0\le n_1,n_2\le N} \alpha_{m_1,n_1}\beta_{m_2,n_2}\mu_{m_1n_1}\mu_{m_2n_2}\mu_{mn}\delta_{m_1+m_2,m} \\ &\quad \sum_{j=1}^{N_g} P_{n_1}^{(0,|m_1|)}(r_j) P_{n_2}^{(0,|m_2|)}(r_j) P_n^{(0,|m|)}(r_j) \left(\frac{1+r_j}{2}\right)^{\frac{|m_1|+|m_2|+|m|}{2}} \sigma_j.\end{aligned}$$

Even though we have a weight function of the form $(1 + r)^\beta$ in the integral, this quadrature rule is preferred to a Gauss-Jacobi quadrature scheme because it allows us to write this expression as

$$\begin{aligned}&\frac{1}{2}\sum_{j=1}^{N_g} \left\{ \sum_{-M<m_1,m_2\le M} \delta_{m_1+m_2,m} \left[\sum_{n_1=0}^{N} \alpha_{m_1,n_1}\mu_{m_1n_1} P_{n_1}^{(0,|m_1|)}(r_j) \left(\frac{1+r_j}{2}\right)^{\frac{|m_1|}{2}} \sigma_j^{1/3}\right] \right. \\ &\left. \left[\sum_{n_2=0}^{N} \beta_{m_2,n_2}\mu_{m_2n_2} P_{n_2}^{(0,|m_2|)}(r_j) \left(\frac{1+r_j}{2}\right)^{\frac{|m_2|}{2}} \sigma_j^{1/3}\right] \right\} \mu_{mn} P_n^{(0,|m|)}(r_j) \left(\frac{1+r_j}{2}\right)^{\frac{|m|}{2}} \sigma_j^{1/3}.\end{aligned}$$

The terms of the form $\mu_{mn} P_n^{(0,|m|)}(r_j) \left(\frac{1+r_j}{2}\right)^{\frac{|m|}{2}} \sigma_j^{1/3}$ can be pre-computed for all the polynomials and quadrature points and weights. Given $\{\alpha_{mn}\}$ and $\{\beta_{mn}\}$, the sums in the square parentheses are evaluated at the quadrature points and their Fast-Fourier Transforms computed; this requires $O(M(N + \log(M)))$ operations for each i. Multiplying them together and taking the inverse FFTs executes the convolution inside the curly braces and requires an additional $O(M \log(M))$ operations. Finally, we can multiply the external factors at cost $O(MN)$ and sum over the quadrature index i to obtain all the modal coefficients of fg. The total complexity therefore comes up to $O(M(M + N)(N + \log(M)))$.

Similarly, for operation type (ii), we use (2.12) to obtain

$$\begin{aligned}&\langle \zeta_{mn}, \nabla f \cdot \nabla g \rangle \\ = \; &\frac{1}{\pi} \sum_{-M<m_1,m_2\le M} \sum_{0\le n_1,n_2\le N} \alpha_{m_1,n_1}\beta_{m_2,n_2}\mu_{m_1n_1}\mu_{m_2n_2}\mu_{mn} \int_0^{2\pi} e^{i(m_1+m_2-m)\theta} \, d\theta \\ &\left[\int_0^1 16 P_{n_1}^{(0,|m_1|)\prime}(2\rho^2-1) P_{n_2}^{(0,|m_2|)\prime}(2\rho^2-1) P_n^{(0,|m|)}(2\rho^2-1) \rho^{|m_1|+|m_2|+|m|+3} \, d\rho\right.\end{aligned}$$

$$
\begin{aligned}
&+\int_0^1 4|m_1|P_{n_1}^{(0,|m_1|)}(2\rho^2-1)P_{n_2}^{(0,|m_2|)\prime}(2\rho^2-1)P_n^{(0,|m|)}(2\rho^2-1)\rho^{|m_1|+|m_2|+|m|+1}\,d\rho\\
&+\int_0^1 4|m_2|P_{n_1}^{(0,|m_1|)\prime}(2\rho^2-1)P_{n_2}^{(0,|m_2|)}(2\rho^2-1)P_n^{(0,|m|)}(2\rho^2-1)\rho^{|m_1|+|m_2|+|m|+1}\,d\rho\\
&+(|m_1m_2|-m_1m_2)\int_0^1 P_{n_1}^{(0,|m_1|)}(2\rho^2-1)P_{n_2}^{(0,|m_2|)}(2\rho^2-1)P_n^{(0,|m|)}(2\rho^2-1)\\
&\rho^{|m_1|+|m_2|+|m|-1}\,d\rho\Big]\\
=\;&\frac{1}{2}\sum_{-M<m_1,m_2\le M}\ \sum_{0\le n_1,n_2\le N}\alpha_{m_1,n_1}\beta_{m_2,n_2}\mu_{m_1n_1}\mu_{m_2n_2}\mu_{mn}\delta_{m_1+m_2,m}\\
&\left[\int_{-1}^1 16P_{n_1}^{(0,|m_1|)\prime}(r)P_{n_2}^{(0,|m_2|)\prime}(r)P_n^{(0,|m|)}(r)\left(\frac{1+r}{2}\right)^{\frac{|m_1|+|m_2|+|m|+2}{2}}\,dr\right.\\
&+\int_{-1}^1 4|m_1|P_{n_1}^{(0,|m_1|)}(r)P_{n_2}^{(0,|m_2|)\prime}(r)P_n^{(0,|m|)}(r)\left(\frac{1+r}{2}\right)^{\frac{|m_1|+|m_2|+|m|}{2}}\,dr\\
&+\int_{-1}^1 4|m_2|P_{n_1}^{(0,|m_1|)\prime}(r)P_{n_2}^{(0,|m_2|)}(r)P_n^{(0,|m|)}(r)\left(\frac{1+r}{2}\right)^{\frac{|m_1|+|m_2|+|m|}{2}}\,dr\\
&\left.+(|m_1m_2|-m_1m_2)\int_{-1}^1 P_{n_1}^{(0,|m_1|)}(r)P_{n_2}^{(0,|m_2|)}(r)P_n^{(0,|m|)}(r)\left(\frac{1+r}{2}\right)^{\frac{|m_1|+|m_2|+|m|-2}{2}}\,dr\right]\\
=\;&\frac{1}{2}\sum_{-M<m_1,m_2\le M}\ \sum_{0\le n_1,n_2\le N}\alpha_{m_1,n_1}\beta_{m_2,n_2}\mu_{m_1n_1}\mu_{m_2n_2}\mu_{mn}\delta_{m_1+m_2,m}\\
&\left[\sum_{j=1}^{N_g}16P_{n_1}^{(0,|m_1|)\prime}(r_j)P_{n_2}^{(0,|m_2|)\prime}(r_j)P_n^{(0,|m|)}(r_j)\left(\frac{1+r_j}{2}\right)^{\frac{|m_1|+|m_2|+|m|+2}{2}}\sigma_j\right.\\
&+\sum_{j=1}^{N_g}4|m_1|P_{n_1}^{(0,|m_1|)}(r_j)P_{n_2}^{(0,|m_2|)\prime}(r_j)P_n^{(0,|m|)}(r_j)\left(\frac{1+r_j}{2}\right)^{\frac{|m_1|+|m_2|+|m|}{2}}\sigma_j\\
&+\sum_{j=1}^{N_g}4|m_2|P_{n_1}^{(0,|m_1|)\prime}(r_j)P_{n_2}^{(0,|m_2|)}(r_j)P_n^{(0,|m|)}(r_j)\left(\frac{1+r_j}{2}\right)^{\frac{|m_1|+|m_2|+|m|}{2}}\sigma_j\\
&\left.+(|m_1m_2|-m_1m_2)\sum_{j=1}^{N_g}P_{n_1}^{(0,|m_1|)}(r_j)P_{n_2}^{(0,|m_2|)}(r_j)P_n^{(0,|m|)}(r_j)\left(\frac{1+r_j}{2}\right)^{\frac{|m_1|+|m_2|+|m|-2}{2}}\sigma_j\right]\\
=\;&\frac{1}{2}\sum_{j=1}^{N_g}\left\{16\sum_{-M<m_1,m_2\le M}\delta_{m_1+m_2,m}\left[\sum_{n_1=0}^{N}\alpha_{m_1,n_1}\mu_{m_1n_1}P_{n_1}^{(0,|m_1|)\prime}(r_j)\left(\frac{1+r_j}{2}\right)^{\frac{|m_1|+1}{2}}\sigma_j^{1/3}\right]\right.
\end{aligned}
$$

$$\left[\sum_{n_2=0}^{N}\beta_{m_2,n_2}\mu_{m_2n_2}P_{n_2}^{(0,|m_2|)\prime}(r_j)\left(\frac{1+r_j}{2}\right)^{\frac{|m_2|+1}{2}}\sigma_j^{1/3}\right]$$

$$+4\sum_{-M<m_1,m_2\leq M}\delta_{m_1+m_2,m}\left[\sum_{n_1=0}^{N}\alpha_{m_1,n_1}\mu_{m_1n_1}|m_1|P_{n_1}^{(0,|m_1|)}(r_j)\left(\frac{1+r_j}{2}\right)^{\frac{|m_1|}{2}}\sigma_j^{1/3}\right]$$

$$\left[\sum_{n_2=0}^{N}\beta_{m_2,n_2}\mu_{m_2n_2}P_{n_2}^{(0,|m_2|)\prime}(r_j)\left(\frac{1+r_j}{2}\right)^{\frac{|m_2|}{2}}\sigma_j^{1/3}\right]$$

$$+4\sum_{-M<m_1,m_2\leq M}\delta_{m_1+m_2,m}\left[\sum_{n_1=0}^{N}\alpha_{m_1,n_1}\mu_{m_1n_1}P_{n_1}^{(0,|m_1|)\prime}(r_j)\left(\frac{1+r_j}{2}\right)^{\frac{|m_1|}{2}}\sigma_j^{1/3}\right]$$

$$\left[\sum_{n_2=0}^{N}\beta_{m_2,n_2}\mu_{m_2n_2}|m_2|P_{n_2}^{(0,|m_2|)}(r_j)\left(\frac{1+r_j}{2}\right)^{\frac{|m_2|}{2}}\sigma_j^{1/3}\right]$$

$$+\sum_{-M<m_1,m_2\leq M}\delta_{m_1+m_2,m}\left[\sum_{n_1=0}^{N}\alpha_{m_1,n_1}\mu_{m_1n_1}|m_1|P_{n_1}^{(0,|m_1|)}(r_j)\left(\frac{1+r_j}{2}\right)^{\frac{|m_1|-1}{2}}\sigma_j^{1/3}\right]$$

$$\left[\sum_{n_2=0}^{N}\beta_{m_2,n_2}\mu_{m_2n_2}|m_2|P_{n_2}^{(0,|m_2|)}(r_j)\left(\frac{1+r_j}{2}\right)^{\frac{|m_2|-1}{2}}\sigma_j^{1/3}\right]$$

$$-\sum_{-M<m_1,m_2\leq M}\delta_{m_1+m_2,m}\left[\sum_{n_1=0}^{N}\alpha_{m_1,n_1}\mu_{m_1n_1}m_1P_{n_1}^{(0,|m_1|)}(r_j)\left(\frac{1+r_j}{2}\right)^{\frac{|m_1|-1}{2}}\sigma_j^{1/3}\right]$$

$$\left[\sum_{n_2=0}^{N}\beta_{m_2,n_2}\mu_{m_2n_2}m_2P_{n_2}^{(0,|m_2|)}(r_j)\left(\frac{1+r_j}{2}\right)^{\frac{|m_2|-1}{2}}\sigma_j^{1/3}\right]\Bigg\}\times$$

$$\mu_{mn}P_n^{(0,|m|)}(r_j)\left(\frac{1+r_j}{2}\right)^{\frac{|m|}{2}}\sigma_j^{1/3}.$$

The derivatives appearing in the above expression can be evaluated by using the derivative property (2.3) of Jacobi polynomials. Observe that this decomposition has exactly the same order of complexity as operation (i). Finally, for case (iii), we first note that

$$\begin{aligned}\Delta\zeta_{mn}(\rho,\theta) &= \rho^{-1}\partial_\rho(\rho\partial_\rho\zeta_{mn}(\rho,\theta))+\rho^{-2}\partial_\theta^2\zeta_{mn}(\rho,\theta)\\ &= \mu_{mn}e^{im\theta}\left[16P_n^{(0,|m|)\prime\prime}(2\rho^2-1)\rho^{|m|+2}+8(|m|+1)P_n^{(0,|m|)\prime}(2\rho^2-1)\rho^{|m|}\right].\end{aligned}$$

Thus,

$$\langle\zeta_{mn},(\Delta f)g\rangle$$

$$
\begin{aligned}
&= \frac{1}{\pi} \sum_{-M<m_1,m_2\le M} \sum_{0\le n_1,n_2\le N} \alpha_{m_1,n_1}\beta_{m_2,n_2}\mu_{m_1n_1}\mu_{m_2n_2}\mu_{mn} \int_0^{2\pi} e^{i(m_1+m_2-m)\theta}\, d\theta \\
&\quad \left[\int_0^1 16P_{n_1}^{(0,|m_1|)\prime\prime}(2\rho^2-1)P_{n_2}^{(0,|m_2|)}(2\rho^2-1)P_n^{(0,|m|)}(2\rho^2-1)\rho^{|m_1|+|m_2|+|m|+3}\, d\rho \right. \\
&\quad \left. + \int_0^1 8(|m_1|+1)P_{n_1}^{(0,|m_1|)\prime}(2\rho^2-1)P_{n_2}^{(0,|m_2|)}(2\rho^2-1)P_n^{(0,|m|)}(2\rho^2-1)\rho^{|m_1|+|m_2|+|m|+1}\, d\rho\right] \\
&= \frac{1}{2} \sum_{-M<m_1,m_2\le M} \sum_{0\le n_1,n_2\le N} \alpha_{m_1,n_1}\beta_{m_2,n_2}\mu_{m_1n_1}\mu_{m_2n_2}\mu_{mn}\delta_{m_1+m_2,m} \\
&\quad \left[\int_{-1}^1 16P_{n_1}^{(0,|m_1|)\prime\prime}(r)P_{n_2}^{(0,|m_2|)}(r)P_n^{(0,|m|)}(r)\left(\frac{1+r}{2}\right)^{\frac{|m_1|+|m_2|+|m|+2}{2}} dr \right. \\
&\quad \left. + \int_{-1}^1 8(|m_1|+1)P_{n_1}^{(0,|m_1|)\prime}(r)P_{n_2}^{(0,|m_2|)}(r)P_n^{(0,|m|)}(r)\left(\frac{1+r}{2}\right)^{\frac{|m_1|+|m_2|+|m|}{2}} dr\right] \\
&= \frac{1}{2} \sum_{-M<m_1,m_2\le M} \sum_{0\le n_1,n_2\le N} \alpha_{m_1,n_1}\beta_{m_2,n_2}\mu_{m_1n_1}\mu_{m_2n_2}\mu_{mn}\delta_{m_1+m_2,m} \\
&\quad \left[\sum_{j=1}^{N_g} 16P_{n_1}^{(0,|m_1|)\prime\prime}(r)P_{n_2}^{(0,|m_2|)}(r)P_n^{(0,|m|)}(r)\left(\frac{1+r}{2}\right)^{\frac{|m_1|+|m_2|+|m|+2}{2}} \sigma_j \right. \\
&\quad \left. + \sum_{j=1}^{N_g} 8(|m_1|+1)P_{n_1}^{(0,|m_1|)\prime}(r)P_{n_2}^{(0,|m_2|)}(r)P_n^{(0,|m|)}(r)\left(\frac{1+r}{2}\right)^{\frac{|m_1|+|m_2|+|m|}{2}} \sigma_j\right] \\
&= \frac{1}{2}\sum_{j=1}^{N_g} \left\{ \sum_{-M<m_1,m_2\le M} \delta_{m_1+m_2,m} \left[\sum_{n_1=0}^{N} \alpha_{m_1,n_1}\mu_{m_1n_1} \left(16P_{n_1}^{(0,|m_1|)\prime\prime}(r_j)\left(\frac{1+r_j}{2}\right)^{\frac{|m_1|+2}{2}} + \right.\right.\right. \\
&\quad \left.\left.\left. 8(|m_1|+1)P_{n_1}^{(0,|m_1|)\prime}(r_j)\left(\frac{1+r_j}{2}\right)^{\frac{|m_1|}{2}}\right)\sigma_j^{1/3}\right] \times \\
&\quad \left.\left[\sum_{n_2=0}^{N} \beta_{m_2,n_2}\mu_{m_2n_2}P_{n_2}^{(0,|m_2|)}(r_j)\left(\frac{1+r_j}{2}\right)^{\frac{|m_2|}{2}}\sigma_j^{1/3}\right]\right\} \mu_{mn}P_n^{(0,|m|)}(r_j)\left(\frac{1+r_j}{2}\right)^{\frac{|m|}{2}}\sigma_j^{1/3}.
\end{aligned}
$$

The complexity of this operation is also the same as for the previous types. As we shall see in the following chapters, being able to do these computations rapidly adds greatly to the appeal of Zernike polynomials as basis functions on a circular domain.

Chapter 3

Computing the Dirichlet-Neumann Operator on a Cylinder

In this chapter, we present a key component of our water-wave simulation algorithm. The Dirichlet-Neumann operator for the Laplace equation (referred to as DNO henceforth) is a crucial ingredient in modern studies of the inviscid water-wave problem. Its analysis and computation pose significant challenges, which become even more pronounced in higher dimensions. In what follows, we begin by describing the history of the DNO and how it fits into the larger water-wave problem. We then move on to our take on its computation on a cylindrical domain and develop a rapid and accurate algorithm with the help of the tools introduced in Chapter 2. We also discuss some subtleties in its implementation before demonstrating its effectiveness on a number of test problems.

3.1 Introduction

A History of the DNO

Water-wave equations are notoriously hard to solve numerically because of the nonlinear nature of the problem and the evolving domain, which is itself an unknown quantity. Modern formulations of this problem have focused on the evolution of the boundary variables with the information from the interior of the domain obtained with the help of a Dirichlet-Neumann operator. The non-locality of the DNO has been noted to pose a severe challenge in both numerical and theoretical studies [24, 29, 54].

Traditionally, numerical computations of the DNO have been restricted to the 2D case. Several elegant and robust numerical techniques have been devised including, among others, conformal mapping, finite element and boundary integral methods. However, these techniques do not carry over successfully to higher dimensions since they either rely inherently on the geometry of a 2D space or scale poorly with dimension. We therefore need to consider approaches that may not be widely used in 2D but can be extended to 3D. In [54], the

authors exhaustively analyze a number of these, including the operator expansion of Craig & Sulem [13], the integral equation formulation of Ablowitz, Fokas & Musslimani [1] and the transformed field expansion method (TFE) of Bruno & Reitich [8] and Nicholls & Reitich [44, 45]. The first two are shown to suffer from catastrophic numerical instabilities which severely limits their utility. In particular, they involve significant cancellations of terms or, equivalently, a rapid decay of singular values. As a result, these methods require multiple precision arithmetic to yield accurate solutions.

The TFE method, on the other hand, possesses a straightforward generalization to 3D and yields a numerically stable high-order algorithm. In addition, it is also able to handle artificial dissipation [29]. A careful analysis of this technique as applied to problems from fluid mechanics and acoustics is presented in [46], including proofs of the analytic dependence of the DNO on the domain shape and the convergence of the method. One shortcoming of this approach is that it is unable to capture a surface bending back on itself as it requires the interface to be the graph of a function. For the majority of applications, however, this is not an issue.

In this chapter, we generalize the TFE method to build a solver for the DNO problem for Laplace's equation on a cylinder. We specifically work in this geometry as it poses the most significant challenges out of all the regular geometries. Rectangular geometries with periodic boundary conditions essentially avoid the issue of addressing the interaction of the fluid with a wall and, at any rate, can be treated similarly by an extension of this technique. The method proceeds by assuming the domain is a perturbation of a simpler geometry and expanding the DNO in terms of the perturbation parameter. Naive implementations of this approach however lead to large cancellations, making it unsuitable for numerical procedures. These cancellations can be avoided by first flattening out the domain; this also simplifies the geometry of the problem and the PDE to be solved is replaced by a sequence of related problems. Successful implementations of this technique therefore require the rapid computation of the solutions to the associated problems.

The DNO in the Water-Wave Problem

Consider a cylinder of unit radius with a flat bottom containing an incompressible, irrotational and inviscid fluid. Denote the cylindrical coordinates on this geometry by (ρ', θ', z'). Suppose the fluid at rest has a depth of h (written $z' = -h$) while the interface at the top is given by $z' = \eta(\rho', \theta')$ (we assume that $h > \|\eta\|_\infty$).

The irrotationality of the fluid allows us to express its velocity at any point as the gradient of a potential function ϕ. The evolution of the fluid is then described by Euler's equations:

$$\Delta'\phi = 0 \qquad -h < z' < \eta \tag{3.1}$$

$$\partial_t\phi + \frac{1}{2}|\nabla'\phi|^2 + (g - F(t))\eta = 0 \qquad z' = \eta \tag{3.2}$$

$$\partial_t\eta + \nabla'_H\eta \cdot \nabla'_H\phi = \partial_{z'}\phi \qquad z' = \eta \tag{3.3}$$

where $F(t)$ is an external force. At the lateral and bottom boundaries, we have the no-flow conditions $\frac{\partial \phi}{\partial \mathbf{n}} = 0$ while at the interface, we have the Dirichlet condition $\phi|_{z'=\eta} = q$. Here, ∇'_H represents the horizontal gradient operator (given by $(\partial_{\rho'}, (\rho')^{-1}\partial_{\theta'}, 0)$ in cylindrical coordinates).

These equations can in fact be reformulated as an evolution problem for the surface variables η and q only [13, 57]:

$$\partial_t \eta = G[\eta]q. \tag{3.4}$$

$$\partial_t q = -(g - F(t))\eta - \frac{1}{2}|\nabla'_H q|^2 + \frac{(G[\eta]q + \nabla'_H \eta \cdot \nabla'_H q)^2}{2(1 + |\nabla'_H \eta|^2)}. \tag{3.5}$$

where $G[\eta]q$ is the Dirichlet-Neumann operator (DNO) given by

$$\begin{aligned} G[\eta]q &= [\nabla' \phi]|_{z'=\eta} . (-\nabla'_H \eta, 1) \\ &= [-\nabla'_H \phi \cdot \nabla'_H \eta + \partial_{z'} \phi]_{z'=\eta} \end{aligned} \tag{3.6}$$

and ϕ is the solution of (3.1) with the boundary conditions specified above.

Thus, we only need to solve the first-order system (3.4, 3.5) to completely capture the dynamics of the free surface. The problem, of course, lies in the computation of the highly non-local DNO as it requires, in principle, the solution of Laplace's equation on evolving domain in three dimensions.

3.2 Two Derivations of the Transformed Field Expansion

In this section, we develop the Transformed Field Expansion (TFE) method for Laplace's equation on a cylinder. We present two derivations of this formulation. The first of these employs techniques from differential geometry and is more concise. The second, following the method outlined in [43], uses elementary calculus tools and is therefore more accessible.

The key idea of the TFE method is to flatten the boundary of the domain and obtain, in place of Laplace's equation, a sequence of associated Poisson equations, the solutions of which yield the potential in the bulk. This reformulation allows us to build a spectrally accurate technique for computing the DNO. The first step is the change of variables

$$\rho = \rho', \quad \theta = \theta', \quad z = h\left(\frac{z' - \eta}{h + \eta}\right). \tag{3.7}$$

Observe then that, as $z \in [-h, 0]$, the domain takes the shape of an unperturbed cylinder C in terms of (ρ, θ, z). In addition, we introduce a new symbol for the bulk potential in the new coordinates

$$u(\rho, \theta, z) = \phi(\rho, \theta, h^{-1}(h + \eta)z + \eta) = \phi(\rho', \theta', z'). \tag{3.8}$$

We now need to determine the transformation that Laplace's equation (3.1) undergoes. We present two ways of going about this.

The First Derivation

Denote the Cartesian coordinates by $\tilde{x}^i = (\tilde{x}, \tilde{y}, \tilde{z})$, the cylindrical coordinates by $x'^i = (\rho', \theta', z')$ and the new coordinates in (3.7) by $x^i = (\rho, \theta, z)$. Observe that the metric tensor in the x^i is given by $\mathcal{G} = E_1^T E_1$ where

$$E_1 = \left(\frac{\partial \tilde{x}^i}{\partial x^j}\right) = \left(\frac{\partial \tilde{x}^i}{\partial x'^k}\right)\left(\frac{\partial x'^k}{\partial x^j}\right) = \begin{pmatrix} \cos(\theta) & -\rho\sin(\theta) & 0 \\ \sin(\theta) & \rho\cos(\theta) & 0 \\ \left(1+\frac{z}{h}\right)\eta_\rho & \left(1+\frac{z}{h}\right)\eta_\theta & 1+\frac{\eta}{h} \end{pmatrix}.$$

is the change-of-coordinates matrix. The Laplace-Beltrami operator applied to both sides of (3.8) yields

$$\frac{1}{\sqrt{\det \mathcal{G}}}\frac{\partial}{\partial x^a}\left(\sqrt{\det \mathcal{G}}(\mathcal{G}^{-1})^{ab}u_{x^b}\right) = \Delta\phi = 0. \tag{3.9}$$

We have $\sqrt{\det \mathcal{G}} = \det(E_1) = h^{-1}\rho(h+\eta)$ and $\mathcal{G}^{-1} = E_1^{-1}E_1^{-T} = (h+\eta)^{-2}E_2E_2^T$ where

$$E_2 = \begin{pmatrix} h+\eta & 0 & 0 \\ 0 & \rho^{-1}(h+\eta) & 0 \\ -(h+z)\eta_\rho & -\rho^{-1}(h+z)\eta_\theta & h \end{pmatrix}.$$

Plugging these in (3.9) allows us to write

$$\operatorname{div}\left((h+\eta)^{-1}EE^T\nabla u\right) = 0, \quad \text{where} \quad E = \begin{pmatrix} h+\eta & 0 & 0 \\ 0 & h+\eta & 0 \\ -(h+z)\eta_\rho & -\rho^{-1}(h+z)\eta_\theta & h \end{pmatrix} \tag{3.10}$$

and $\operatorname{div}\mathbf{v} = \rho^{-1}\partial_\rho(\rho v_1) + \rho^{-1}\partial_\theta v_2 + \partial_z v_3$ and $\nabla v = (\partial_\rho v, \rho^{-1}\partial_\theta v, \partial_z v)^T$ are the divergence and gradient operators in cylindrical coordinates respectively. Expanding (3.10) leads to

$$\begin{aligned} (h+\eta)^{-1}\operatorname{div}(EE^T\nabla u) &= -\nabla((h+\eta)^{-1})\cdot(EE^T\nabla u) \\ \Rightarrow \operatorname{div}(EE^T\nabla u) &= (h+\eta)^{-1}(\eta_\rho, \rho^{-1}\eta_\theta, 0)\ EE^T\ \nabla u \\ &= (\eta_\rho, \rho^{-1}\eta_\theta, 0)\ E^T\nabla u \end{aligned} \tag{3.11}$$

where we used the fact that

$$(h+\eta)^{-1}(\eta_\rho, \rho^{-1}\eta_\theta, 0)\ E = (\eta_\rho, \rho^{-1}\eta_\theta, 0).$$

Ostensibly, we have bartered an elementary equation on a challenging domain for a much tougher problem on a simple geometry. This form, however, lends itself to a simplification inspired by boundary perturbation methods. We assume the interface to be a deviation from a flat surface, to wit, $\eta(\rho, \theta) = \epsilon f(\rho, \theta)$ for some ϵ. The conditions under which this assumption leads to a useful solution will be made precise later on but it is worth noting

that the actual value of ϵ is irrelevant. Writing $EE^T = h^2 I + \epsilon A_1(f) + \epsilon^2 A_2(f)$ and the first and second columns of E as $B_0 + \epsilon B_1(f)$ and $C_0 + \epsilon C_1(f)$ respectively in (3.11) yields

$$\operatorname{div}\left[(h^2 I + \epsilon A_1(f) + \epsilon^2 A_2(f))\nabla u\right] = \epsilon f_\rho (B_0 + \epsilon B_1(f)) \cdot \nabla u + \epsilon \rho^{-1} f_\theta (C_0 + \epsilon C_1(f)) \cdot \nabla u$$

Grouping together similar powers of ϵ leads to

$$\begin{aligned} -h^2 \Delta u &= \epsilon \left[\operatorname{div}(A_1(f)\nabla u) - \left(f_\rho B_0 + \frac{f_\theta}{\rho} C_0 \right) \cdot \nabla u \right] + \\ &\quad \epsilon^2 \left[\operatorname{div}(A_2(f)\nabla u) - \left(f_\rho B_1(f) + \frac{f_\theta}{\rho} C_1(f) \right) \cdot \nabla u \right]. \end{aligned} \tag{3.12}$$

The Second Derivation

This method relies mainly on the chain rule. Using $z = h\left(\frac{z'-\eta}{h+\eta}\right)$, we have

$$\begin{aligned} \partial_{\rho'} z &= h\left(\frac{(h+\eta)(-\partial_{\rho'}\eta) - (z'-\eta)(\partial_{\rho'}\eta)}{(h+\eta)^2} \right) \\ &= (-\partial_{\rho'}\eta)\left(\frac{h}{h+\eta} + \frac{h}{h+\eta}\left(\frac{z'-\eta}{h+\eta}\right)\right) \\ &= (-\partial_{\rho'}\eta)\left(\frac{h}{h+\eta} + \frac{z}{h+\eta}\right) \\ &= (-\partial_{\rho'}\eta)\left(\frac{h+z}{h+\eta}\right) \end{aligned} \tag{3.13}$$

and similarly

$$\partial_{\theta'} z = (-\partial_{\theta'}\eta)\left(\frac{h+z}{h+\eta}\right). \tag{3.14}$$

At this stage, we define the following useful quantities

$$\begin{aligned} M(\rho,\theta) &= h + \eta(\rho,\theta) \\ N_1(\rho,\theta,z) &= -(\partial_\rho \eta(\rho,\theta))(h+z) \\ N_2(\rho,\theta,z) &= -\rho^{-1}(\partial_\theta \eta(\rho,\theta))(h+z). \end{aligned}$$

We can therefore write (3.13) and (3.14) as

$$\partial_{\rho'} z = M^{-1} N_1, \quad \partial_{\theta'} z = \rho M^{-1} N_2.$$

Observe then that

$$\partial_{\rho'}\phi = \partial_\rho u + (\partial_{\rho'} z)(\partial_z u) = \partial_\rho u + (M^{-1}N_1)(\partial_z u)$$

$$
\begin{aligned}
\Rightarrow (M\partial_{\rho'})\phi &= (M\partial_\rho + N_1\partial_z)u && (3.15)\\
\partial_{\theta'}\phi &= \partial_\theta u + (\partial_{\theta'}z)(\partial_z u) = \partial_\theta u + (\rho M^{-1}N_2)(\partial_z u)\\
\Rightarrow (M\partial_{\theta'})\phi &= (M\partial_\theta + \rho N_2\partial_z)u && (3.16)\\
\partial_{z'}\phi &= (\partial_{z'}z)(\partial_z u) = (hM^{-1})(\partial_z u)\\
\Rightarrow (M\partial_{z'})\phi &= (h\partial_z)u. && (3.17)
\end{aligned}
$$

We then have from (3.1)

$$
\begin{aligned}
0 &= M^2\Delta'\phi\\
&= \frac{M^2}{\rho'}\partial_{\rho'}(\rho'\partial_{\rho'}\phi) + \frac{M^2}{(\rho')^2}\partial^2_{\theta'}\phi + M^2\partial^2_{z'}\phi\\
&= \left[\frac{M}{\rho'}\partial_{\rho'}(M\rho'\partial_{\rho'}\phi) - M(\partial_{\rho'}\phi)(\partial_{\rho'}M)\right] + \left[\frac{M}{(\rho')^2}\partial_{\theta'}(M\partial_{\theta'}\phi) - \frac{M}{(\rho')^2}(\partial_{\theta'}\phi)(\partial_{\theta'}M)\right] +\\
&\quad M\partial_{z'}(M\partial_{z'}\phi)\\
&= \left[M\partial_{\rho'}(M\partial_{\rho'}\phi) + \frac{M}{\rho'}(M\partial_{\rho'}\phi)\right] + \frac{M}{(\rho')^2}\partial_{\theta'}(M\partial_{\theta'}\phi) + M\partial_{z'}(M\partial_{z'}\phi)\\
&\quad -M(\partial_{\rho'}\phi)(\partial_{\rho'}M) - \frac{M}{(\rho')^2}(\partial_{\theta'}\phi)(\partial_{\theta'}M)\\
&= [M\partial_\rho + N_1\partial_z][M\partial_\rho u + N_1\partial_z u] + \frac{M}{\rho}[M\partial_\rho u + N_1\partial_z u] +\\
&\quad \frac{1}{\rho^2}[M\partial_\theta + \rho N_2\partial_z][M\partial_\theta u + \rho N_2\partial_z u] + [h\partial_z][h\partial_z u]\\
&\quad -(\partial_\rho M)[M\partial_\rho u + N_1\partial_z u] - \frac{(\partial_\theta M)}{\rho^2}[M\partial_\theta u + \rho N_2\partial_z u]\\
&= M\partial_\rho[M\partial_\rho u] + M\partial_\rho[N_1\partial_z u] + N_1\partial_z[M\partial_\rho u] + N_1\partial_z[N_1\partial_z u] + \frac{M}{\rho}[M\partial_\rho u + N_1\partial_z u] +\\
&\quad \frac{1}{\rho^2}(M\partial_\theta[M\partial_\theta u] + \rho N_2\partial_z[M\partial_\theta u] + M\partial_\theta[\rho N_2\partial_z u] + \rho N_2\partial_z[\rho N_2\partial_z u]) +\\
&\quad h\partial_z[h\partial_z u] - (\partial_\rho M)[M\partial_\rho u] - (\partial_\rho M)[N_1\partial_z u] - \frac{(\partial_\theta M)}{\rho^2}[M\partial_\theta u] - \frac{(\partial_\theta M)}{\rho^2}[\rho N_2\partial_z u]\\
&= [\partial_\rho(M^2\partial_\rho u) - (M\partial_\rho u)(\partial_\rho M)] + [\partial_z(MN_1\partial_\rho u) - (N_1\partial_z u)(\partial_\rho M)] +\\
&\quad [\partial_\rho(MN_1\partial_z u) - (N_1\partial_z u)(\partial_\rho M)] + [\partial_z(N_1^2\partial_z u) - (N_1\partial_z u)(\partial_z N_1)] + \frac{M}{\rho}[M\partial_\rho u + N_1\partial_z u]\\
&\quad +\frac{1}{\rho^2}[\partial_\theta(M^2\partial_\theta u) - (M\partial_\theta u)(\partial_\theta M)] + \frac{1}{\rho}[\partial_z(MN_2\partial_\theta u) - (M\partial_\theta u)(\partial_z N_2)]\\
&\quad +\frac{1}{\rho}[\partial_\theta(MN_2\partial_z u) - (N_2\partial_z u)(\partial_\theta M)] + [\partial_z(N_2^2\partial_z u) - (N_2\partial_z u)(\partial_z N_2)]\\
&\quad +\partial_z[h^2\partial_z u] - (\partial_\rho M)[M\partial_\rho u] - (\partial_\rho M)[N_1\partial_z u] - \frac{(\partial_\theta M)}{\rho^2}[M\partial_\theta u] - \frac{(\partial_\theta M)}{\rho}[N_2\partial_z u]
\end{aligned}
$$

$$
\begin{aligned}
&= \left(\partial_\rho + \frac{1}{\rho}\right)\left\{M^2\partial_\rho u + MN_1\partial_z u\right\} + \frac{1}{\rho}\partial_\theta\left\{M^2\left(\frac{\partial_\theta u}{\rho}\right) + MN_2\partial_z u\right\} + \\
&\quad \partial_z\left\{MN_1\partial_\rho u + N_1^2\partial_z u + MN_2\left(\frac{\partial_\theta u}{\rho}\right) + N^2\partial_z u + h^2\partial_z u\right\} \\
&\quad -(M\partial_\rho u + N_1\partial_z u)\left(2\partial_\rho M + \partial_z N_1\right) - \left(M\left(\frac{\partial_\theta u}{\rho}\right) + N_2\partial_z u\right)\left(2\left(\frac{\partial_\theta M}{\rho}\right) + \partial_z N_2\right).
\end{aligned} \tag{3.18}
$$

Observe that

$$2\partial_\rho M + \partial_z N_1 = 2\partial_\rho\eta - \partial_\rho\eta = \partial_\rho\eta$$

and

$$2\left(\frac{\partial_\theta M}{\rho}\right) + \partial_z N_2 = 2\left(\frac{\partial_\theta\eta}{\rho}\right) - \left(\frac{\partial_\theta\eta}{\rho}\right) = \frac{\partial_\theta\eta}{\rho}$$

so, continuing from (3.18), we have

$$
\begin{aligned}
0 &= \operatorname{div}\begin{pmatrix} M^2\partial_\rho u + MN_1\partial_z u \\ M^2\left(\frac{\partial_\theta u}{\rho}\right) + MN_2\partial_z u \\ MN_1\partial_\rho u + N_1^2\partial_z u + MN_2\left(\frac{\partial_\theta u}{\rho}\right) + N^2\partial_z u + h^2\partial_z u \end{pmatrix} - (M\partial_\rho u + N_1\partial_z u)(\partial_\rho\eta) - \\
&\quad \left(M\left(\frac{\partial_\theta u}{\rho}\right) + N_2\partial_z u\right)\left(\frac{\partial_\theta\eta}{\rho}\right). \\
&= \operatorname{div}\left[\begin{pmatrix} M^2 & 0 & MN_1 \\ 0 & M^2 & MN_2 \\ MN_1 & MN_2 & N_1^2 + N_2^2 + h^2 \end{pmatrix}\nabla u\right] - (\partial_\rho\eta)\begin{pmatrix} M \\ 0 \\ N_1 \end{pmatrix}\cdot\nabla u - \left(\frac{\partial_\theta\eta}{\rho}\right)\begin{pmatrix} 0 \\ M \\ N_2 \end{pmatrix}\cdot\nabla u \\
&= \operatorname{div}(A\nabla u) - (\partial_\rho\eta)\, B\cdot\nabla u - \left(\frac{\partial_\theta\eta}{\rho}\right) C\cdot\nabla u
\end{aligned} \tag{3.19}
$$

where A, B, C are the respective matrices in the second-to-last step. Observe now that

$$
\begin{aligned}
A &= \begin{pmatrix} (h+\eta)^2 & 0 & -(\partial_\rho\eta)(h+\eta)(h+z) \\ 0 & (h+\eta)^2 & -\left(\frac{\partial_\theta\eta}{\rho}\right)(h+\eta)(h+z) \\ -(\partial_\rho\eta)(h+\eta)(h+z) & -\left(\frac{\partial_\theta\eta}{\rho}\right)(h+\eta)(h+z) & h^2 + |\nabla_H\eta|^2(h+z)^2 \end{pmatrix} \\
&= h^2 I + A_1(\eta) + A_2(\eta)
\end{aligned}
$$

where

$$
A_1(\eta) = \begin{pmatrix} 2h\eta & 0 & -h(\partial_\rho\eta)(h+z) \\ 0 & 2h\eta & -h\left(\frac{\partial_\theta\eta}{\rho}\right)(h+z) \\ -h(\partial_\rho\eta)(h+z) & -h\left(\frac{\partial_\theta\eta}{\rho}\right)(h+z) & 0 \end{pmatrix}
$$

$$A_2(\eta) = \begin{pmatrix} \eta^2 & 0 & -\eta(\partial_\rho\eta)(h+z) \\ 0 & \eta^2 & -\eta\left(\frac{\partial_\theta\eta}{\rho}\right)(h+z) \\ -\eta(\partial_\rho\eta)(h+z) & -\eta\left(\frac{\partial_\theta\eta}{\rho}\right)(h+z) & |\nabla_H\eta|^2(h+z)^2 \end{pmatrix}.$$

Similarly, we can also write $B = B_0 + B_1(\eta)$ and $C = C_0 + C_1(\eta)$ where

$$B_0 = \begin{pmatrix} h \\ 0 \\ 0 \end{pmatrix}, \quad B_1(\eta) = \begin{pmatrix} \eta \\ 0 \\ -\partial_\rho\eta(h+z) \end{pmatrix}, \quad C_0 = \begin{pmatrix} 0 \\ h \\ 0 \end{pmatrix}, \quad C_1(\eta) = \begin{pmatrix} 0 \\ \eta \\ -\left(\frac{\partial_\theta\eta}{\rho}\right)(h+z) \end{pmatrix}.$$

Note that all these matrices have been chosen so that $D_i(k\eta) = k^i D_i(\eta)$ and that these are precisely the same matrices that appeared in the previous derivation. Using these decompositions in (3.19), we obtain

$$\begin{aligned} 0 &= h^2\mathrm{div}(\nabla u) + \mathrm{div}(A_1(\eta)\nabla u) + \mathrm{div}(A_2(\eta)\nabla u) - (\partial_\rho\eta)\, B_0\cdot\nabla u \\ &\quad - (\partial_\rho\eta)\, B_1(\eta)\cdot\nabla u - \left(\frac{\partial_\theta\eta}{\rho}\right) C_0\cdot\nabla u - \left(\frac{\partial_\theta\eta}{\rho}\right) C_1(\eta)\cdot\nabla u \\ \Rightarrow -h^2\Delta u &= \mathrm{div}(A_1(\eta)\nabla u) + \mathrm{div}(A_2(\eta)\nabla u) - \left[(\partial_\rho\eta)\, B_0 + \left(\frac{\partial_\theta\eta}{\rho}\right) C_0\right]\cdot\nabla u - \\ &\quad \left[(\partial_\rho\eta)\, B_1(\eta) + \left(\frac{\partial_\theta\eta}{\rho}\right) C_1(\eta)\right]\cdot\nabla u. \end{aligned} \tag{3.20}$$

This is the rewritten form of (3.1). For the boundary condition on the lateral walls, we use (3.15) to get

$$((h+\eta)\partial_\rho - (\partial_\rho\eta)(h+z)\partial_z)u|_{\rho=1} = 0.$$

Similarly, at the bottom boundary, we have from (3.17) that $\partial_z u|_{z=-h} = 0$ while the Dirichlet condition at the interface is replaced by $u|_{z=0} = q$.

As in the previous derivation, we now assume the interface to be of the form $\eta(\rho,\theta) = \epsilon f(\rho,\theta)$. Using this in (3.20) and collecting similar powers of ϵ yields

$$\begin{aligned} -h^2\Delta u &= \epsilon\left[\mathrm{div}(A_1(f)\nabla u) - \left(f_\rho B_0 + \frac{f_\theta}{\rho}C_0\right)\cdot\nabla u\right] + \\ &\quad \epsilon^2\left[\mathrm{div}(A_2(f)\nabla u) - \left(f_\rho B_1(f) + \frac{f_\theta}{\rho}C_1(f)\right)\cdot\nabla u\right]. \end{aligned} \tag{3.21}$$

which is identical to (3.12). This derivation moreover allows us to determine the boundary conditions

$$u|_{z=0} = q, \quad \partial_z u|_{z=-h} = 0, \quad h\partial_\rho u|_{\rho=1} = \epsilon[-f\partial_\rho u + (h+z)\partial_\rho f\partial_z u]_{\rho=1}. \tag{3.22}$$

Next, we address the effect of the coordinate change (3.7) on the expression (3.6) for the DNO. We have

$$\begin{aligned} MG[\eta]q &= \left\{-(M\partial_{\rho'}\phi)(\partial_{\rho'}\eta) - (\rho')^{-2}(M\partial_{\theta'}\phi)(\partial_{\theta'}\eta) + (M\partial_{z'}\phi)\right\}\big|_{z'=\eta} \\ &= \left\{-((M\partial_\rho + N_1\partial_z)u)(\partial_\rho\eta) - \rho^{-2}((M\partial_\rho + \rho N_2\partial_z)u)(\partial_\theta\eta) + (h\partial_z u)\right\}\big|_{z=0}. \end{aligned}$$

Since $N_1|_{z=0} = -h\partial_\rho\eta$ and $N_2|_{z=0} = -h\rho^{-1}\partial_\theta\eta$, we get

$$\begin{aligned} hG[\eta]q + \eta G[\eta]q &= -\left[h(\partial_\rho u) + \eta(\partial_\rho u) - h(\partial_\rho\eta)(\partial_z u)\right](\partial_\rho\eta) \\ &\quad - \left[h\frac{\partial_\theta u}{\rho} + \eta\frac{\partial_\theta u}{\rho} - h\left(\frac{\partial_\theta\eta}{\rho}\right)(\partial_z u)\right]\left(\frac{\partial_\theta\eta}{\rho}\right) + h(\partial_z u) \\ \Rightarrow hG[\eta]q &= h(\partial_z u) + H(\rho,\theta;\eta,u) \end{aligned} \tag{3.23}$$

where

$$\begin{aligned} H(\rho,\theta;\eta,u) &= -\eta G[\eta]q - \left[h(\partial_\rho u) + \eta(\partial_\rho u) - h(\partial_\rho\eta)(\partial_z u)\right](\partial_\rho\eta) \\ &\quad - \left[h\frac{\partial_\theta u}{\rho} + \eta\frac{\partial_\theta u}{\rho} - h\left(\frac{\partial_\theta\eta}{\rho}\right)(\partial_z u)\right]\left(\frac{\partial_\theta\eta}{\rho}\right). \\ &= -\eta G[\eta]q - h(\partial_\rho u)(\partial_\rho\eta) - \eta(\partial_\rho u)(\partial_\rho\eta) + h(\partial_\rho\eta)^2(\partial_z u) \\ &\quad -h\left(\frac{\partial_\theta u}{\rho}\right)\left(\frac{\partial_\theta\eta}{\rho}\right) - \eta\left(\frac{\partial_\theta u}{\rho}\right)\left(\frac{\partial_\theta\eta}{\rho}\right) + h\left(\frac{\partial_\theta\eta}{\rho}\right)^2(\partial_z u). \\ &= -\eta G[\eta]q - (h+\eta)\nabla_H\eta\cdot\nabla_H u + h|\nabla_H\eta|^2(\partial_z u) \end{aligned}$$

As this is evaluated at $z = 0$, we can safely replace $\nabla_H u$ by $\nabla_H q$. In addition, using $\eta(\rho,\theta) = \epsilon f(\rho,\theta)$ and collecting like powers gives

$$hG[\epsilon f]q = h(\partial_z u) + \epsilon\left(-fG[\epsilon f]q - h\nabla_H f\cdot\nabla_H q\right) + \epsilon^2\left(-f\nabla_H f\cdot\nabla_H q + h|\nabla_H f|^2\partial_z u\right). \tag{3.24}$$

The Field Expansion

We now expand the transformed field $u(\rho,\theta,z) = \sum_{k=0}^{\infty}\epsilon^k u_k(\rho,\theta,z)$ and plug it in (3.12) (or (3.21)). Comparing coefficients of like powers of ϵ leads to a sequence of Poisson equations

$$-\Delta u_k = r_k \tag{3.25}$$

where

$$\begin{aligned} r_k &= h^{-2}\left\{\mathrm{div}(A_1(f)\nabla u_{k-1}) + \mathrm{div}(A_2(f)\nabla u_{k-2})\right. \\ &\quad \left. -\left(f_\rho B_0 + \rho^{-1}f_\theta C_0\right)\cdot\nabla u_{k-1} - \left(f_\rho B_1(f) + \rho^{-1}f_\theta C_1(f)\right)\cdot\nabla u_{k-2}\right\} \end{aligned} \tag{3.26}$$

The boundary conditions can likewise be obtained from (3.22):

$$u_k|_{z=0} = \delta_{k,0}q, \qquad \partial_z u_k|_{z=-h} = 0, \qquad \partial_\rho u_k|_{\rho=1} = \chi_k \tag{3.27}$$

where $\chi_k = h^{-1}\left[-f\partial_\rho u_{k-1} + (h+z)\partial_\rho f \partial_z u_{k-1}\right]_{\rho=1}$.

Observe that the right sides of both (3.26) and (3.27) depend on lower order terms in the expansion. As a result, we can sequentially solve this three-term recurrence for the u_k, up to a sufficiently high order K, and combine them to obtain an approximation to u.

The solutions u_k at different orders can then be used to compute the Neumann data. Plugging in $G[\epsilon f]q = \sum_{k=0}^{\infty} \epsilon^k G_k[f]q$ and the expansion for u in (3.24) and comparing coefficients yields

$$\begin{aligned} hG_k[f]q &= h(u_k)_z - fG_{k-1}[f]q + h|\nabla_H f|^2 (u_{k-2})_z \\ &\quad -\delta_{k,1}\left(h\nabla_H f \cdot \nabla_H q\right) - \delta_{k,2}\left(f\nabla_H f \cdot \nabla_H q\right). \end{aligned} \tag{3.28}$$

Note that the formulation (3.25) with associated boundary conditions (3.27) is exact and, assuming the expansions converge, yield the true solution to (3.11). In Section 3.4, we shall show that, under certain conditions, these expansions do indeed converge strongly.

3.3 Solving the Poisson Equations

We next present a method to solve the model Poisson equation $-\Delta w = r$ on a flat cylinder C of unit radius and height h with boundary conditions

$$w|_{z=0} = q, \qquad w_z|_{z=-h} = 0, \qquad w_\rho|_{\rho=1} = \chi. \tag{3.29}$$

We begin by building a basis for a function space on C. Let M, N, J be positive integers.

- Let $\{z_j\}_{0\le j\le J}$ be the $(J+1)$ Chebyshev-Lobatto points over $[-h, 0]$ defined by

$$z_j = -\frac{h}{2}\left(1 + \cos\left(\frac{\pi j}{J}\right)\right).$$

  For $0 \le j \le J$, let ℓ_j be the jth Lagrange polynomial with respect to these nodes so that $\ell_j(z_i) = \delta_{ij}$.

- For $-M < m \le M$ and $0 \le n \le N$, let ζ_{mn} be the (m, n) Zernike polynomial introduced in Chapter 2.

As basis functions on C, we consider

$$\psi_{mnj}(\rho, \theta, z) = \zeta_{mn}(\rho, \theta)\ell_j(z). \tag{3.30}$$

This choice of basis functions allows us to replace the unwieldy Bessel functions and hyperbolic functions along the radial and vertical axes respectively by families of polynomials. These polynomials are easy to evaluate and lend themselves to rapid manipulations. This considerably simplifies the resulting formulation and speeds up the subsequent computations.

We also employ the inner product on $L^2(C)$

$$\langle v, w\rangle = \frac{1}{\pi}\int_{-h}^{0}\int_{0}^{2\pi}\int_{0}^{1} \overline{v}w\ \rho\ d\rho\ d\theta\ dz.$$

For the model problem $-\Delta w = r$ with (3.29), we begin by imposing the Galerkin condition

$$\langle \psi_{m'n'j'}, -\Delta w\rangle = \langle \psi_{m'n'j'}, r\rangle$$

for $0 \le n' \le N$, $-M < m' \le M$, $0 \le j' \le J-1$. Integrating by parts gives

$$\iiint_C \overline{\nabla\psi_{m'n'j'}}\cdot\nabla w\ dV = \iiint_C \overline{\psi_{m'n'j'}} r\ dV + \iint_{\partial C}\overline{\psi_{m'n'j'}}\frac{\partial w}{\partial \mathbf{n}}\ dA \tag{3.31}$$

where $\partial C = B_u \cup B_d \cup S$; here, B_u and B_d are the upper and lower ends of the cylinder respectively and S is the curved surface. Note that as $\psi_{m'n'j'}|_{B_u} \equiv 0$, there is no contribution from B_u. Meanwhile, the second condition in (3.29) ensures that the integral over B_d is also zero. As a result, the boundary contributions can be written as

$$\iint_{\partial C}\overline{\psi_{m'n'j'}}\frac{\partial w}{\partial \mathbf{n}}\ dA = \iint_S \overline{\psi_{m'n'j'}} w_\rho\ dA = \iint_S \overline{\psi_{m'n'j'}}\chi\ dA := I_{m'n'j'}(\chi).$$

Next, write $w = \sum_{m,n,j} c_{mnj}\psi_{mnj}$ and decompose $r = \sum_{m,n,j} d_{mnj}\psi_{mnj}$, where

$$d_{mnj} \;=\; \langle \zeta_{mn}, r(\cdot,\cdot,z_j)\rangle_{L^2(D)}\,, \tag{3.32}$$

to obtain

$$\sum_{m,n,j} c_{m,n,j}\iiint_C \overline{\nabla\psi_{m'n'j'}}\cdot\nabla\psi_{mnj}\ dV \;=\; \sum_{m,n,j} d_{mnj}\iiint_C \overline{\psi_{m'n'j'}}\psi_{mnj}\ dV + I_{m'n'j'}(\chi). \tag{3.33}$$

The Dirichlet condition in (3.29) implies that $\{c_{mnj}\}$ is known for $j = J$ so those terms can be moved to the right as well. We therefore have a system of the type $Sc = Td + \kappa$, where S and T are the stiffness and mass matrices respectively and κ is a vector generated by the boundary data χ and $\{c_{mnJ}\}$.

These stiffness and mass integrals can be computed by exploiting the structure of the basis functions. Define the matrices A_m, Σ and $\tilde{\Sigma}$ by

$$\begin{aligned} A_{m,n'n} &= \frac{1}{\pi}\int_0^{2\pi}\int_0^1 \overline{\nabla_H\zeta_{mn'}}\cdot\nabla_H\zeta_{mn}\ \rho\ d\rho\ d\theta \\ \Sigma_{j'j} &= \int_{-h}^{0}\ell_{j'}(z)\ell_j(z)\ dz \\ \tilde{\Sigma}_{j'j} &= \int_{-h}^{0}\ell'_{j'}(z)\ell'_j(z)\ dz. \end{aligned} \tag{3.34}$$

Then,

$$T_{m'n'j',mnj} = \langle \zeta_{m'n'}, \zeta_{mn} \rangle_H \Sigma_{j'j} = \delta_{m'm}\delta_{n'n}\Sigma_{j'j}$$

where we used the orthonormality of the ζ_{mn}. In addition,

$$S_{m'n'j',mnj} = \delta_{m'm}(A_{m,n'n}\Sigma_{j'j} + \delta_{n'n}\tilde{\Sigma}_{j'j}). \tag{3.35}$$

Finally, define the matrices $\Gamma_{m,nj} = c_{mnj}$, $E_{m,nj} = d_{mnj}$ and $K_{m,nj} = \kappa_{mnj}$ and set $G_m = E_m\Sigma + K_m$ to get

$$A_m\Gamma_m\Sigma + \Gamma_m\tilde{\Sigma} = G_m. \tag{3.36}$$

This is a sparse linear system of size $2MJ(N+1)$. Instead of using an iterative or direct solver, we design an alternative, well-conditioned method for this problem. First note that, from (2.16), we have

$$A_{m,n'n} = 2\mu_{mn'}\mu_{mn}(2\gamma_{n'n}(\gamma_{n'n} + |m| + 1) + |m|)$$

where $\gamma_{n'n} = \min\{n', n\}$. The symmetric positive definiteness of A_m allows the eigen-decomposition $A_m = W_m D_m^2 W_m^T$.

Next, let $\{(x_i, \omega_i)\}$ be the $(J+1)$ point Gauss-Legendre quadrature scheme over $[-h, 0]$. Defining the matrices $E_{ij} = \ell_j(x_i)\omega_i^{1/2}$ and $\tilde{E}_{ij} = \ell_j'(x_i)\omega_i^{1/2}$ for $0 \le j \le J-1$ allows us to write

$$\Sigma = E^T E, \quad \tilde{\Sigma} = \tilde{E}^T\tilde{E}.$$

Note that the columns of E (and $\tilde{E}$) must be linearly independent: any linear combination of the columns that equals zero would correspond to a polynomial of degree at most J ($J-1$ for $\tilde{E}$) that has $(J+1)$ zeros (at the quadrature points) so that polynomial must be identically zero; the coefficients in the linear combination must necessarily be all zero since the Lagrange polynomials are linearly independent. Thus, the QR factorizations $E = QR$ and $\tilde{E} = \tilde{Q}\tilde{R}$ yield invertible upper triangular matrices. Plugging these decompositions into (3.36) gives

$$\begin{aligned} (W_m D_m^2 W_m^T)\Gamma_m(R^T R) + \Gamma_m(\tilde{R}^T\tilde{R}) &= G_m \\ D_m^2(W_m^T\Gamma_m)R^T + (W_m^T\Gamma_m)\tilde{R}^T\tilde{R}R^{-1} &= W_m^T G_m R^{-1} \end{aligned}$$

Perform the singular value decomposition $R\tilde{R}^{-1} = U\Lambda V^T$ to obtain $\tilde{R}^T = R^T U\Lambda^{-1}V^T$ and $\tilde{R}R^{-1} = V\Lambda^{-1}U^T$. This gives $\tilde{R}^T\tilde{R}R^{-1} = R^T U\Lambda^{-2}U^T$ and hence

$$\begin{aligned} D_m^2(W_m^T\Gamma_m R^T) + (W_m^T\Gamma_m R^T)U\Lambda^{-2}U^T &= W_m^T G_m R^{-1} \\ D_m^2(W_m^T\Gamma_m R^T U) + (W_m^T\Gamma_m R^T U)\Lambda^{-2} &= W_m^T G_m R^{-1} U \end{aligned} \tag{3.37}$$

As both D_m^2 and Λ are diagonal, we have

$$(W_m^T\Gamma_m R^T U)_{nj} = \frac{(W_m^T G_m R^{-1} U)_{nj}}{D_{m,nn}^2 + \Lambda_{jj}}$$

which can then be used to solve for Γ_m.

From a numerical perspective, using the decompositions above avoids squaring the condition number. For a computational analysis, assume that M and J are $O(N)$ as well. The bulk of the computation essentially involves finding the coefficients $\{d_{mnj}\}$ and performing the matrix multiplications specified in (3.37). The latter require $O(N^3)$ operations for each m. The former requires the computation of the expressions for r_k in (3.26) and the projection-interpolant in (3.32). Upon expanding the formulas for r_k, we obtain

$$\begin{aligned} r_k(\rho,\theta,z) &= -2h^{-1}f\Delta_H u_{k-1} + h^{-1}(h+z)(2\nabla_H f\cdot\nabla_H(\partial_z u_{k-1}) + (\partial_z u_{k-1})\Delta_H f) \\ &\quad -h^{-2}f^2\Delta_H u_{k-2} + h^{-2}f(h+z)(2\nabla_H f\cdot\nabla_H(\partial_z u_{k-2}) + (\partial_z u_{k-2})\Delta_H f) \\ &\quad -h^{-2}(h+z)|\nabla_H f|^2(2(\partial_z u_{k-2}) + (h+z)(\partial_z^2 u_{k-2})). \end{aligned} \tag{3.38}$$

Note that the solutions $\{u_j\}$ for $j<k$ are already calculated in terms of the basis functions; in particular, at each vertical slice indexed by j, these solutions are linear combinations of Zernike polynomials. We can also find the Zernike modal representation for f by (2.6) and (2.7). Hence, carrying out the computation (3.32), for every j, comes down to a sequence of computations of the sort

$$\langle\zeta_{mn}, v_1v_2\rangle_{L^2(D)}, \quad \langle\zeta_{mn}, \nabla_H v_1\cdot\nabla_H v_2\rangle_{L^2(D)} \quad \text{and} \quad \langle\zeta_{mn}, (\Delta_H v_1)v_2\rangle_{L^2(D)}$$

where v_1 and v_2 are functions on D with known Zernike modal representations. In Section 2.3, we presented techniques for the rapid computation of these types of operations and showed that it could be done in $O(M(M+N)(N+\log(M)))$ complexity. Under the assumption that M is $O(N)$, the procedure requires $O(N^3)$ operations for each j. With $J=O(N)$ as well, we conclude that the assembly of the coefficients $\{d_{mnj}\}$ can be executed in $O(N^4)$ steps, which is the total complexity of our Poisson solver on a flat cylinder.

3.4 A Proof of Convergence

In this section, we present a proof of convergence for the transformed field expansions. We begin by establishing the error estimates for the projection-interpolant as defined in (3.32). The subsequent result for the solution from the TFE method provides a precise rate of decay for the errors appearing in the numerical solutions and describes the dependence on the various parameters. In particular, it sheds light on the role of the "perturbation parameter" ϵ.

Let $s\geq 1$ be an integer. As in Chapter 2, let $H^s(D)$ be the respective Sobolev space on the unit disc D and let $H^s_\omega((-h,0))$ be the Sobolev space equipped with the norm

$$\|v\|^2_{H^s_\omega((-h,0))} = \int_{-h}^0 \sum_{k=0}^s |v^{(k)}(z)|^2\omega(z)\,dz$$

where $\omega(z) = \frac{1}{2}\left(\frac{-z}{h}\left(1+\frac{z}{h}\right)\right)^{-1/2}$. Note that under the transformation $x=1+2z/h$, the weight function gets changed to the Chebyshev weight $(1-x^2)^{-1/2}$ on $[-1,1]$.

Recall the definition of the polynomials $\{\ell_j\}$: if $\{z_j\}_{0\le j\le J}$ denote the Chebyshev-Lobatto nodes on $[-h,0]$, then ℓ_j is the jth Lagrange interpolating polynomial on these nodes. For a function u defined on $[-h,0]$, setting

$$u_J(z) = \sum_{j=0}^{J} u(z_j)\ell_j(z),$$

gives the Jth Chebyshev interpolant for u. We then have the following result (Statement 5.5.22 from [9]).

Lemma 3.4.1 *Let $s, J \ge 0$ be integers. Take $u \in H^s_\omega((-h,0))$ and let u_J be the Jth Chebyshev interpolant for u. Then,*

$$\|u - u_J\|_{L^2_\omega((-h,0))} \lesssim J^{-s} \|u\|_{H^s_\omega((-h,0))}$$

Recall that $C = D \times (-h,0)$ is the flat cylinder. For $w \in H^s(C) = H^s(D)H^s_\omega((-h,0))$, define the projection-interpolant w_{MNJ} by

$$w_{MNJ}(\rho,\theta,z) = \sum_{0\le j\le J} \mathcal{P}_{MN} w(\rho,\theta,z_j)\ell(z).$$

We next combine the approximation estimate Theorem 2.2.2 for Zernike polynomials on D and Lemma 3.4.1 along the z-axis to obtain approximation estimates for the projection-interpolant on the entire cylinder.

Theorem 3.4.1 *Let $s, M, N, J \ge 0$ be integers. Let $w \in H^s(C)$ and let w_{MNJ} be the corresponding projection-interpolant. Then,*

$$\|w - w_{MNJ}\|_{L^2(C)} \lesssim \left(\min(M,2N)^{-s} + J^{-s}\right) \|w\|_{H^s(C)}.$$

Proof: Observe that

$$\|w - w_{MNJ}\|_{L^2(C)} \le \|w - \mathcal{P}_{MN}w\|_{L^2(D)L^2_\omega((-h,0))} + \|\mathcal{P}_{MN}w - w_{MNJ}\|_{L^2(D)L^2_\omega((-h,0))} \tag{3.39}$$

We have by Theorem 2.2.2

$$\begin{aligned}\|w - \mathcal{P}_{MN}w\|_{L^2(D)L^2_\omega((-h,0))} &\lesssim \min(M,2N)^{-s} \|w\|_{H^s(D)L^2_\omega((-h,0))} \\ &\le \min(M,2N)^{-s} \|w\|_{H^s_\omega(D)H^s((-h,0))}\end{aligned}$$

and by Lemma 3.4.1

$$\begin{aligned}\|\mathcal{P}_{MN}w - w_{MNJ}\|_{L^2_\omega((-h,0))L^2(D)} &\lesssim J^{-s} \|\mathcal{P}_{MN}w\|_{L^2(D)H^s_\omega((-h,0))} \\ &\le J^{-s} \|w\|_{H^s(D)H^s_\omega((-h,0))}.\end{aligned}$$

Putting these together in (3.39) gives the desired result.

We now use this approximation estimate to obtain a convergence result for our TFE-Zernike-Chebyshev algorithm. We closely follow the proof strategy outlined in [46] with a few modifications to account for the difference in domain and numerical method.

We first have the following result from standard elliptic PDE theory [19].

Lemma 3.4.2 *For any integer $s \geq 0$, if $r \in H^s(C)$, $q \in H^{s+3/2}(D)$ and $\chi \in H^{s+1/2}(\partial D \times (-h,0))$, then there exists a unique solution of*

$$-\Delta w = r \text{ in } C, \quad w|_{z=0} = q, \quad \partial_z w|_{z=-h} = 0, \quad \partial_\rho w|_{\rho=1} = \chi,$$

such that for some constant $c_1 > 0$,

$$\|w\|_{H^{s+2}(C)} \leq c_1 \left\{ \|r\|_{H^s(C)} + \|q\|_{H^{s+3/2}(D)} + \|\chi\|_{H^{s+1/2}(\partial D \times (-h,0))} \right\}.$$

Recall the system (3.25,3.27)

$$\Delta u_k = r_k \text{ in } C, \quad u_k|_{z=0} = \delta_{k,0} q, \quad (u_k)_z|_{z=-h} = 0, \quad (u_k)_\rho|_{\rho=1} = \chi_k, \tag{3.40}$$

where

$$\begin{aligned} r_k &= h^{-2} \{ \mathrm{div}(A_1(f)\nabla u_{k-1}) + \mathrm{div}(A_2(f)\nabla u_{k-2}) \\ &\quad - \left(f_\rho B_0 + \rho^{-1} f_\theta C_0\right) \cdot \nabla u_{k-1} - \left(f_\rho B_1(f) + \rho^{-1} f_\theta C_1(f)\right) \cdot \nabla u_{k-2} \}, \quad (3.41) \\ \chi_k &= h^{-1} \left[-f(u_{k-1})_\rho + (h+z) f_\rho (u_{k-1})_z \right]_{\rho=1}. \quad (3.42) \end{aligned}$$

For the $k = 0$ case, the assumptions of Lemma 3.4.2 clearly hold. Next, we need to show that these are satisfied for the Poisson equations for $k \geq 1$. We have the following lemma.

Lemma 3.4.3 *Given integers $s, k \geq 1$ and constants $B, E > 0$ such that*

$$\|u_j\|_{H^{s+2}(C)} \leq E B^j, \quad \forall j < k,$$

there exists $E_1 > 0$ such that r_k and χ_k satisfy

$$\begin{aligned} \|r_k\|_{H^s(C)} &\leq E_1 \left\{ \|f\|_{H^{s+2}(D)} B^{k-1} + \|f\|^2_{H^{s+2}(D)} B^{k-2} \right\} \\ \|\chi_k\|_{H^{s+1/2}(\partial D \times (-h,0))} &\leq E_1 \|f\|_{H^{s+2}(D)} B^{k-1}. \end{aligned}$$

Proof: Let $r_{k,1} = h^{-2}\mathrm{div}(A_1(f)\nabla u_{k-1})$. As

$$A_1(f)\nabla u_{k-1} = \begin{pmatrix} 2hf\nabla_H u_{k-1} - h(h+z)\partial_z u_{k-1} \nabla_H f \\ -h(h+z)\nabla_H f \cdot \nabla_H u_{k-1} \end{pmatrix},$$

we have

$$\|r_{k,1}\|_{H^s(C)} = h^{-2} \|A_1(f)\nabla u_{k-1}\|_{H^{s+1}(C)}$$

$$\begin{aligned}
&\leq 2h^{-1}\left\|f\nabla_H u_{k-1}\right\|_{H^{s+1}(C)} + \left\|\nabla_H f\partial_z u_{k-1}\right\|_{H^{s+1}(C)} + \left\|\nabla_H f\cdot\nabla_H u_{k-1}\right\|_{H^{s+1}(C)}\\
&\leq 2Ah^{-1}\left\|f\right\|_{H^{s+1}(D)}\left\|\nabla_H u_{k-1}\right\|_{H^{s+1}(C)} + 2A\left\|\nabla_H f\right\|_{H^{s+1}(D)}\left\|\nabla_H u_{k-1}\right\|_{H^{s+1}(C)}\\
&\leq 2Ah^{-1}\left\|f\right\|_{H^{s+1}(D)}\left\|u_{k-1}\right\|_{H^{s+2}(C)} + 2A\left\|f\right\|_{H^{s+2}(D)}\left\|u_{k-1}\right\|_{H^{s+2}(C)}\\
&\leq E_{11}\left\|f\right\|_{H^{s+2}(D)}B^{k-1}
\end{aligned}$$

where $E_{11} = 2AE(h^{-1}+1)$. Here, we made use of $\left\|h+z\right\|_{L^\infty(-h,0)} = h$ and the algebra properties of Sobolev spaces

$$\left\|g_1 G\right\|_{H^l(C)} \leq A\left\|g_1\right\|_{H^l(D)}\left\|G\right\|_{H^l(C)}$$

$$\left\|g_1 g_2\right\|_{H^l(D)} \leq A\left\|g_1\right\|_{H^l(D)}\left\|g_2\right\|_{H^l(D)}$$

for $l > 3/2$ and for some constant A. Here, g_1, g_2 are functions defined on D and G is defined on C. As we are using this result for $l = s+1$, we require $s > 1/2$ which is indeed true. (Note that for the second property, we can relax this to $s > -1/2$)

Next, set $r_{k,2} = h^{-2}\mathrm{div}(A_2(f)\nabla u_{k-2})$. We have

$$A_2(f)\nabla u_{k-2} = \begin{pmatrix} f^2\nabla_H u_{k-2} - f(h+z)\partial_z u_{k-2}\nabla_H f\\ -(h+z)f\nabla_H f\cdot\nabla_H u_{k-2} + (h+z)^2|\nabla_H f|^2\partial_z u_{k-2}\end{pmatrix}$$

so that

$$\begin{aligned}
\left\|r_{k,2}\right\|_{H^s(C)} &= h^{-2}\left\|A_2(f)\nabla u_{k-2}\right\|_{H^{s+1}(C)}\\
&\leq h^{-2}\left\|f^2\nabla_H u_{k-2}\right\|_{H^{s+1}(C)} + h^{-1}\left\|f\nabla_H f\partial_z u_{k-2}\right\|_{H^{s+1}(C)} +\\
&\quad h^{-1}\left\|f\nabla_H f\cdot\nabla_H u_{k-2}\right\|_{H^{s+1}(C)} + \left\||\nabla_H f|^2\partial_z u_{k-2}\right\|_{H^{s+1}(C)}\\
&\leq A^2h^{-2}\left\|f\right\|^2_{H^{s+1}(D)}\left\|u_{k-2}\right\|_{H^{s+2}(C)} + A^2\left\|f\right\|_{H^{s+2}(D)}\left\|u_{k-2}\right\|_{H^{s+2}(C)} +\\
&\quad 2A^2h^{-1}\left\|f\right\|_{H^{s+1}(D)}\left\|f\right\|^2_{H^{s+2}(D)}\left\|u_{k-2}\right\|_{H^{s+2}(C)}\\
&= A^2\left(h^{-1}\left\|f\right\|_{H^{s+1}(D)} + \left\|f\right\|_{H^{s+1}(D)}\right)^2\left\|u_{k-2}\right\|_{H^{s+2}(C)}\\
&\leq E_{12}\left\|f\right\|^2_{H^{s+2}(D)}B^{k-2}
\end{aligned}$$

where $E_{12} = A^2E(1+h^{-1})^2$.

Next, set $r_{k,3} = h^{-2}\left(f_\rho B_0 + \rho^{-1}f_\theta C_0\right)\cdot\nabla u_{k-1}$. Observe that we can write

$$r_{k,3} = h^{-1}\nabla_H f\cdot\nabla_H u_{k-1}$$

so that

$$\begin{aligned}
\left\|r_{k,3}\right\|_{H^s(C)} &= h^{-1}\left\|\nabla_H f\cdot\nabla_H u_{k-1}\right\|_{H^s(C)}\\
&\leq Ah^{-1}\left\|f\right\|_{H^{s+1}(D)}\left\|u_{k-1}\right\|_{H^{s+1}(C)}\\
&\leq E_{13}\left\|f\right\|_{H^{s+2}(D)}B^{k-1}
\end{aligned}$$

where $E_{13} = AEh^{-1}$.

Define $r_{k,4} = h^{-2}\left(f_\rho B_1(f) + \rho^{-1} f_\theta C_1(f)\right) \cdot \nabla u_{k-2}$. This can be rewritten as

$$r_{k,4} = h^{-2}\left(f\nabla_H f \cdot \nabla_H u_{k-2} - (h+z)|\nabla_H f|^2 \partial_z u_{k-2}\right)$$

so that

$$\begin{aligned}
\|r_{k,4}\|_{H^s(C)} &= h^{-2}\left\|f\nabla_H f \cdot \nabla_H u_{k-2} - (h+z)|\nabla_H f|^2\partial_z u_{k-2}\right\|_{H^s(C)} \\
&\leq A^2 h^{-2}\|f\|_{H^s(D)}\|f\|_{H^{s+1}(D)}\|u_{k-2}\|_{H^{s+1}(C)} + A^2 h^{-1}\|f\|^2_{H^{s+1}(D)}\|u_{k-2}\|_{H^{s+1}(C)} \\
&\leq E_{14}\|f\|^2_{H^{s+2}(D)} B^{k-2}
\end{aligned}$$

where $E_{14} = A^2 E h^{-1}(1 + h^{-1})$.

Finally, note that

$$\begin{aligned}
\|\chi_k\|_{H^{s+1/2}(\partial D\times(-h,0))} &= h^{-1}\|-f(u_{k-1})_\rho + (h+z)f_\rho(u_{k-1})_z\|_{H^{s+1/2}(\partial D\times(-h,0))} \\
&\leq h^{-1}\|f(u_{k-1})_\rho\|_{H^{s+1/2}(\partial D\times(-h,0))} + \|f_\rho(u_{k-1})_z\|_{H^{s+1/2}(\partial D\times(-h,0))} \\
&\leq Ah^{-1}\|f\|_{H^{s+1/2}(\partial D)}\|(u_{k-1})_\rho\|_{H^{s+1/2}(\partial D\times(-h,0))} + \\
&\quad A\|f_\rho\|_{H^{s+1/2}(\partial D)}\|(u_{k-1})_z\|_{H^{s+1/2}(\partial D\times(-h,0))} \\
&\leq Ah^{-1}\|f\|_{H^{s+1}(D)}\|(u_{k-1})_\rho\|_{H^{s+1}(C)} + A\|f_\rho\|_{H^{s+1}(D)}\|(u_{k-1})_z\|_{H^{s+1}(C)} \\
&\leq Ah^{-1}\|f\|_{H^{s+1}(D)}\|u_{k-1}\|_{H^{s+2}(C)} + A\|f\|_{H^{s+2}(D)}\|u_{k-1}\|_{H^{s+2}(C)} \\
&\leq E_{15}\|f\|_{H^{s+2}(D)} B^{k-1}
\end{aligned}$$

where $E_{15} = AE(h^{-1}+1)$. Choose $E_1 = \max\{E_{11}, E_{12}, E_{13}, E_{14}, E_{15}\}$ to complete the result.

∎

Next, we combine Lemmas 3.4.2 and 3.4.3 to obtain an analyticity result for the transformed field expansion.

Theorem 3.4.2 *Given an integer $s \geq 1$, if $q \in H^{s+3/2}(D)$ and $f \in H^{s+2}(D)$, then there exist constants $E_2, E_3 > 0$ such that*

$$\|u_k\|_{H^{s+2}(C)} \leq E_2\|q\|_{H^{s+3/2}(D)} B^k$$

for any constant $B > E_3\|f\|_{H^{s+2}(D)}$.

Proof: We prove this by induction on k. For $k = 0$, the result follows from Lemma 3.4.2:

$$\|u_0\|_{H^{s+2}(C)} \leq c_1\|q\|_{H^{s+3/2}(D)}.$$

Set $E_2 = c_1$. Assume now that the result holds for all $j < k$. Using this with Lemma 3.4.3 and $E = E_2\|q\|_{H^{s+3/2}(D)}$ yields

$$\|r_k\|_{H^s(C)} \leq E_1\left\{\|f\|_{H^{s+2}(D)} B^{k-1} + \|f\|^2_{H^{s+2}(D)} B^{k-2}\right\}$$

$$\|\chi_k\|_{H^{s+1/2}(\partial D\times(-h,0))} \leq E_1 \|f\|_{H^{s+2}(D)} B^{k-1}.$$

where E_1 is the constant from Lemma 3.4.3. Note that it depends linearly on E so, for some $\alpha > 0$, we have $E_1 = \alpha E = \alpha E_2 \|q\|_{H^{s+3/2}(D)}$. Combining these estimates with Lemma 3.4.2 gives

$$\begin{aligned}
\|u_k\|_{H^{s+2}(C)} &\leq c_1 \left\{ \|r_k\|_{H^s(C)} + \|\chi_k\|_{H^{s+1/2}(\partial D\times(-h,0))} \right\} \\
&\leq E_2 E_1 \left\{ 2\|f\|_{H^{s+2}(D)} B^{k-1} + \|f\|^2_{H^{s+2}(D)} B^{k-2} \right\} \\
&= \alpha E_2^2 \|q\|_{H^{s+3/2}(D)} \left\{ 2\|f\|_{H^{s+2}(D)} B^{k-1} + \|f\|^2_{H^{s+2}(D)} B^{k-2} \right\} \\
&\leq E_2 \|q\|_{H^{s+3/2}(D)} B^k
\end{aligned}$$

if $B > E_3 \|f\|_{H^{s+2}(D)}$ where $E_3 = \alpha E_2 + \sqrt{\alpha^2 E_2^2 + \alpha E_2}$. ∎

This result demonstrates that the transformed field expansion $\sum_{k=0}^{\infty} \epsilon^k u_k$ converges for $B\epsilon < 1$. Hence, this technique is guaranteed to yield the exact solution u of (3.10). More critically, as B scales linearly with $\|f\|_{H^{s+2}(D)}$, this result proves that the true determinant of convergence is $\|\eta\|_{H^{s+2}(D)}$. As a result, the precise value of ϵ is actually irrelevant to the success of this method.

Finally, we show that our numerical solution converges in an appropriate sense to u. Let u^k_{MNJ} be the solutions to the Poisson problems (3.25) obtained from the spectral method and let

$$u_{KMNJ} = \sum_{k=0}^{K} \epsilon^k u^k_{MNJ}$$

denote the numerical approximation to u. We then have the following convergence result.

Theorem 3.4.3 *Assume that for some integer $s \geq 3$, $f \in H^s(D)$ and $q \in H^{s-1/2}(D)$. Then,*

$$\|u - u_{KMNJ}\|_{L^2(C)} \lesssim (B\epsilon)^{K+1} + (\min(M, 2N)^{-s} + J^{-s}) \|q\|_{H^{s-1/2}(D)}$$

for any constant $B \geq E_3 \|f\|_{H^s(D)}$ such that $B\epsilon < 1$, where E_3 is the constant from Theorem 3.4.2.

The proof combines the analyticity result from Theorem 3.4.2 and the approximation estimate on the cylinder from Theorem 3.4.1. See the proof of Theorem 2.1 of [46] for details on how to combine these.

3.5 Numerical Results

In order to test the DNO algorithm, we consider a case where the Laplace equation can be analytically solved and we have a closed form for the Neumann data. Let (ρ, θ, z') be

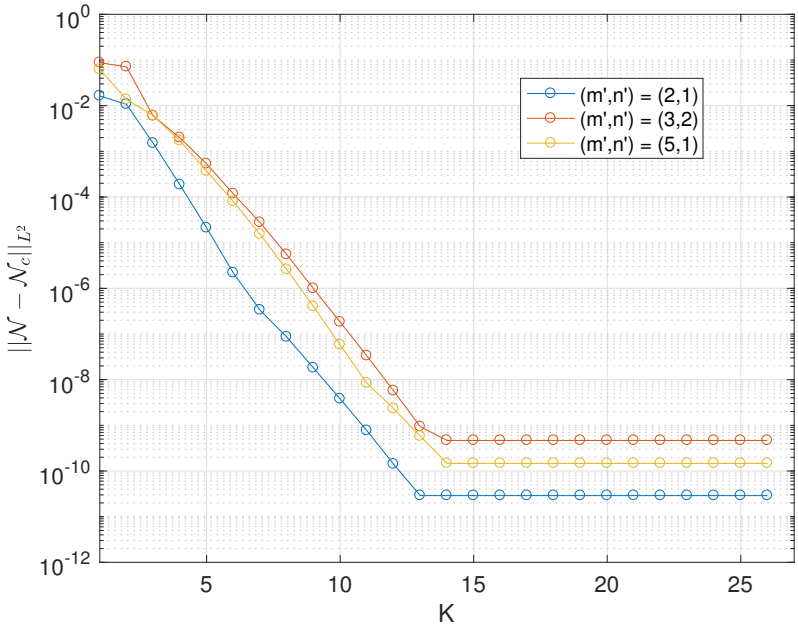

Figure 3.1: Convergence of Neumann data vs TFE order K. The parameter choices are $M = 16$, $J = 20$, $N = 32$, $h = 0.5$ and $\epsilon = 0.2$.

the cylindrical coordinates for the unflattened cylinder (so the interface is $z' = \eta$). The general solution of (3.1) on a cylinder with homogeneous Neumann boundary conditions on the lateral and bottom walls is

$$\phi(\rho, \theta, z') = \sum_{m \in \mathbb{Z}, n \geq 1} b_{m,n} J_{|n|}(a_{|m|n}\rho) e^{im\theta} \cosh(a_{|m|n}(z' + h)) \tag{3.43}$$

where J_m is the mth Bessel function of order zero and a_{mn} is the nth zero of $J'_m(\cdot)$. Note that as ϕ is real valued, we must have $b_{-m,n} = \overline{b_{m,n}}$ for all m, n. Fix $m' \geq 0, n' > 0$ and suppose that for a given interface $\eta(\rho', \theta')$, we have

$$q(\rho, \theta) \quad = \quad J_{m'}(a_{m'n'}\rho) \cos(m'\theta) \cosh(a_{m'n'}(\eta + h)).$$

The particular solution of (3.43) then is

$$\phi(\rho, \theta, z') = J_{m'}(a_{m'n'}\rho) \cos(m'\theta) \cosh(a_{m'n'}(z' + h))$$

which can be used to compute the the Neumann data $\mathcal{N}$ explicitly by (3.6). Figure 3.1 displays the decay in the L^2 errors in the computed Neumann data $\mathcal{N}_c$ for $f(\rho, \theta) = J_1(a_{1,1}\rho)\cos(\theta)$ and (i) $(m', n') = (1, 3)$, (ii) $(m', n') = (3, 2)$ and (iii) $(m', n') = (5, 1)$.

An indication of the role played by the size of the interface η can be garnered by comparing the rates of convergence for different values of ϵ. Figure 3.2 shows that the rate slows down

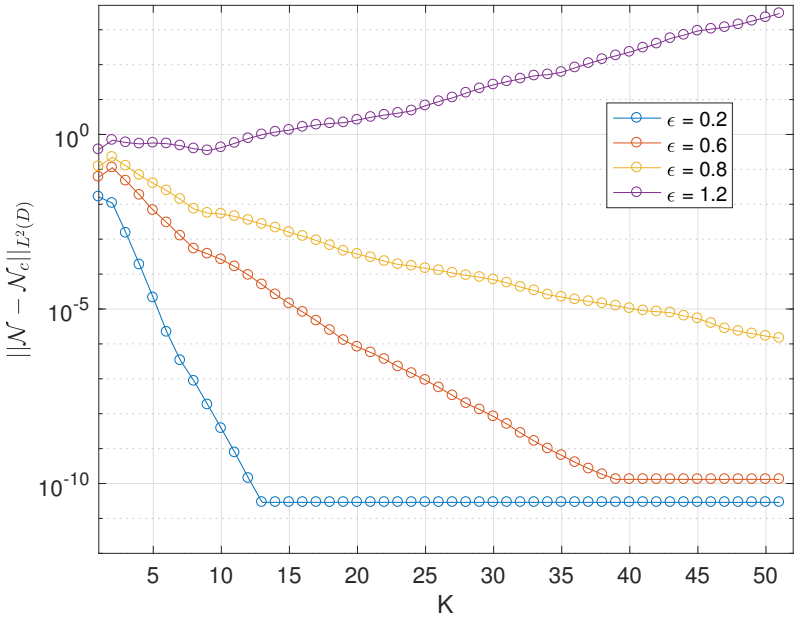

Figure 3.2: Errors in Neumann data vs. TFE order K for different values of ϵ, for case (i) in Fig. 3.1.

for larger values, as suggested by Theorem 3.4.3. For too large a value, the method fails to converge. This poses a limitation on the applicability of this technique, in that it is only guaranteed to work for interfaces of moderate amplitude. When the criterion in Theorem 3.4.3 does not hold, the bulk potential, and hence the Neumann data, may blow up. It is important to realize, however, that the choice of ϵ itself is immaterial and that the true determinant of convergence is $\|\eta\|_{H^s(D)}$.

However, for systems that are supposed to be stable, this technique serves as an extremely efficient simulation tool. We present a qualitative validation of our technique. Consider a fluid at rest with the initial profile $\eta(\rho, \theta) = 0.05\rho e^{-15\rho^2}\cos(\theta)$. Using our DNO solver in conjunction with the time integration scheme described in Section 4.3, we can numerically evolve the system and validate it qualitatively. Figure 3.3 shows various stages in the progression of the fluid. In particular, one can note the collapse of the crest and trough, their outward dispersion and reflection after striking the lateral boundaries.

3.6 Conclusion

We have presented a new technique for computing the DNO for the Laplace equation on a cylinder of finite depth. Its novelty lies in the fact it is primarily tailored for a 3D geometry and does not rely on periodic boundary conditions to avoid dealing with the fluid-boundary

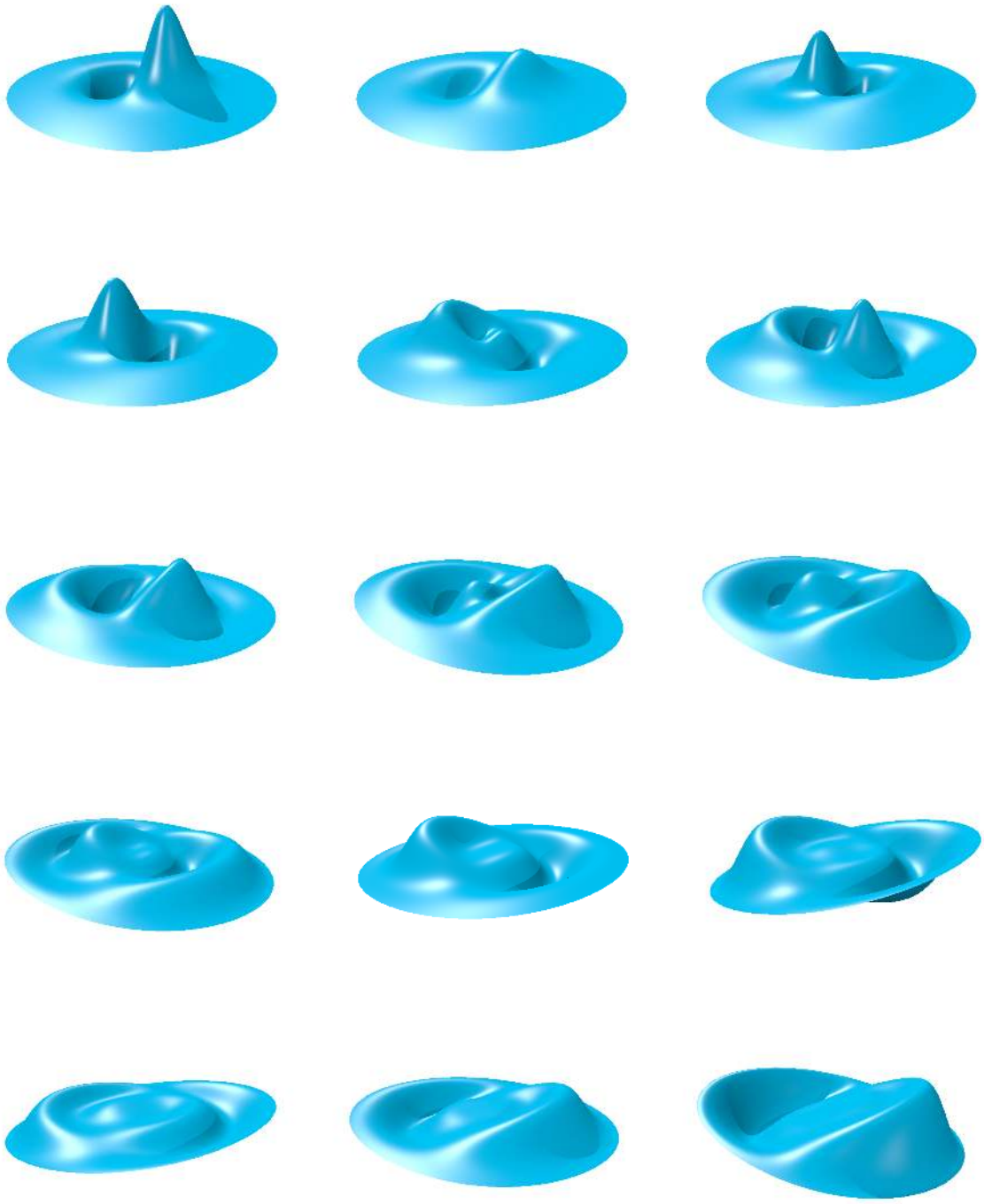

Figure 3.3: Evolution of the interface at $t = 0, 1/80, 2/80, \ldots, 14/80$. We use $M = 4, J = 20, N = 40, K = 2, h = 0.5, \epsilon = 0.01$ and a time-step of $\Delta t = 1/1200$.

interactions. Hence, this method represents a major step-up from the methods that are currently in use. In addition, it is easily applicable to other regular domains in 3D, i.e., prisms, parallelepipeds, etc. and may also find use in domains with structured irregularities (for instance [42]).

The development of the technique is a generalization of the three-term recurrence formulation presented in [43]. However, using tools from differential geometry, we were able to significantly cut down on the tedious algebra. We allied the formulation with a particular choice of basis functions on the cylinder that are amenable to various operations and hence yield a fast algorithm. The preference for Jacobi polynomials over Bessel functions in the radial direction is born out of the need for faster manipulations and spectral accuracy. It is notable that the infinite series representation for the latter is of the form

$$J_m(r) = r^{|m|}(c_0 + c_1 r^2 + c_2 r^4 + \ldots)$$

and that the Jacobi polynomials, with the orthogonality weight, also possess a similar form, namely, $r^{|m|}P_n^{(0,|m|)}(2r^2-1)$. Thus, these polynomials serve as effective substitutes for Bessel functions but with superior approximation properties for representing functions on the disc. In addition, Jacobi polynomials come with quadrature rules, orthogonality, recurrence relations, derivatives and closed forms of eigenvalues, among other desirable properties. These are harder to obtain for Bessel functions. For instance, orthogonality only occurs when the radial axis is scaled by the zeros of J_m (or its derivatives), which in turn need to be computed beforehand.

A similar advantage is gained by the use of Lagrange polynomials with respect to the Chebyshev-Lobatto nodes across the z-axis in place of hyperbolic functions. These polynomials possess a fast transformation to Chebyshev polynomials, which can in turn be differentiated and evaluated accurately at arbitrary points; this comes in handy when, for instance, setting up the quadrature matrices. While not affecting the computational cost and accuracy, they allow us to apply the boundary conditions more easily than would be possible for other function families. The structure of the basis functions therefore yields a fast, well-conditioned solver and ensures that implementation boils down to a sequence of linear algebra operations that can be performed rapidly using BLAS and LAPACK routines.

The approximation estimate for Zernike polynomial established in Chapter 2 leads to a rigorous convergence proof for the TFE method. In particular, it establishes that the convergence hinges only on $\|\eta\|_{H^s(D)}$: the strength of the error decay, as well as possible divergence, is controlled wholly by this value. The result also shows that the parameter ϵ is merely a book-keeping device to help group together terms of the same order when deriving the recurrence formulas. We can conclude that this technique yields a fast and accurate solver for nonlinear water-wave equations when the amplitude does not grow too large.

Since this approach relies on the potential form of the water-wave equations, it disallows dissipation as it appears in the Navier-Stokes equations. To counter this, one can use models of potential viscous flows that artificially introduce dissipation. These have been noted to lead to correct results in the linear wave limit [29] and have found use in various applications

[40]. The TFE technique lends itself to these models in a fairly straightforward manner, as shown in the next chapter.

Chapter 4

Simulating Three-Dimensional Faraday Waves

In this section, we use the machinery developed in the earlier sections to numerically simulate Faraday waves on a cylinder. For a number of reasons, the tools developed so far lend themselves particularly well to this task. Firstly, as established in Theorem 3.4.3, the Dirichlet-to-Neumann solver is highly accurate for moderate interface heights. As Faraday waves fall primarily in this category, our solver is well-suited to simulating them. Secondly, the Transformed Field Expansion approach allows us to precisely control the degree of nonlinearity in the system. Indeed, by varying the index of the highest allowed order K, we can transition between linear, weakly nonlinear and full nonlinear regimes. Finally, as we shall see, the Zernike basis on the unit disc provides an accurate and computationally efficient framework for evolving the system in time.

4.1 The Viscous Model

In Chapter 3, we introduced the Euler equations for an incompressible, irrotational and inviscid fluid. These assumptions are necessary to be able to describe the fluid equations as a potential flow with the potential satisfying Laplace's equation. Indeed, a viscous fluid cannot be irrotational and it is the irrotationality that allows us to define the potential in the first place. Hence, these assumptions are crucial in order to use the Euler equations and the equivalent surface formulation.

However, viscosity is an essential part of a Faraday system. It is needed to balance the effect of resonance produced by the externally applied vibrations and yield stable standing waves. In order then to marry the two seemingly incompatible requirements of a potential flow and viscosity, we need to consider models that have been termed viscous potential flows [27]. The most popular of these is the artificial dissipation model of Dias et. al. introduced in [16]. Its well-posedness has been studied in [4] while it has also appeared extensively in numerical studies [29, 40, 42]. Similar linear dissipative models have been studied in [38,

55].

We begin from the same point as in Section 3.1. Let ϕ denote the velocity potential of a fluid in a unit cylinder with the interface shape described by $z' = \eta(\rho', \theta')$ and a flat bottom at $z' = -h$. The Dias et. al. model is then given by

$$\begin{aligned} \Delta'\phi &= 0 && -h < z' < \eta && (4.1)\\ \partial_t\phi + \frac{1}{2}|\nabla'\phi|^2 + (g - F(t))\eta + 2\nu\partial_{z'}^2\phi &= 0 && z' = \eta && (4.2)\\ \partial_t\eta + \nabla'_H\eta\cdot\nabla'_H\phi - 2\nu\Delta'_H\eta &= \partial_{z'}\phi && z' = \eta && (4.3) \end{aligned}$$

where $F(t)$ is the externally applied force and ν is the kinematic viscosity. Here, ∇'_H and Δ'_H represent the horizontal gradient and Laplacian operators respectively. At the lateral and bottom boundaries of the cylinder, we impose the no-flow conditions $\frac{\partial\phi}{\partial\mathbf{n}} = 0$ while at the interface, we have the Dirichlet condition $\phi|_{z'=\eta} = q$. Note that this model differs from (3.1) – (3.3) only by the inclusion of an extra term each in (4.2) and (4.3).

As in Section 3.1, this system can be rewritten in terms of the surface variables η and q. Here, however, we present the full derivation since we also need to determine how the additional viscosity terms turn out. Defining the Dirichlet-Neumann operator (DNO)

$$G[\eta]q = [-\nabla'_H\phi\cdot\nabla'_H\eta + \partial_{z'}\phi]_{z'=\eta} \quad (4.4)$$

gives us, by (4.3),

$$\partial_t\eta = G[\eta]q + 2\nu\Delta'_H\eta. \quad (4.5)$$

We also impose the boundary condition $\partial_\rho\eta|_{\rho=1} = 0$. Note that this additional condition is required for the problem to be well-posed on a bounded domain since this model contains a higher order spatial derivative of η. In addition, as $q = \phi|_{z'=\eta}$, we have

$$\nabla'_H q = \nabla'_H\phi + (\partial_{z'}\phi)\nabla'_H\eta \quad (4.6)$$

which, combined with (4.4), gives

$$\partial_{z'}\phi|_{z'=\eta} = \frac{G[\eta]q + \nabla'_H\eta\cdot\nabla'_H\cdot q}{1 + |\nabla'_H\eta|^2}. \quad (4.7)$$

We also have $\partial_t q = \partial_t\phi + (\partial_{z'}\phi)\partial_t\eta$ which, together with (4.2), (4.6) and (4.7), yields

$$\begin{aligned} \partial_t q &= -\frac{1}{2}|\nabla'_H\phi|^2 - \frac{1}{2}(\partial_{z'}\phi)^2 - (g - F(t))\eta - 2\nu\partial_{z'}^2\phi + (\partial_{z'}\phi)\partial_t\eta\\ &= -\frac{1}{2}\left(|\nabla'_H q|^2 - (\partial_{z'}\phi)^2|\nabla'_H\eta|^2 - 2(\partial_{z'}\phi)\nabla'_H\phi\cdot\nabla'_H\eta\right) - \frac{1}{2}(\partial_{z'}\phi)^2 -\\ &\quad (g - F(t))\eta - 2\nu\partial_{z'}^2\phi + (\partial_{z'}\phi)[\partial_{z'}\phi - \nabla'_H\phi\cdot\nabla'_H\eta + 2\nu\Delta'_H\eta]\\ &= -(g - F(t))\eta - \frac{1}{2}|\nabla'_H q|^2 + \frac{1}{2}(\partial_{z'}\phi)^2(1 + |\nabla'_H\eta|^2) + 2\nu[\partial_{z'}\phi\Delta'_H\eta - \partial_{z'}^2\phi] \end{aligned}$$

$$\Rightarrow \partial_t q = -(g - F(t))\eta - \frac{1}{2}|\nabla'_H q|^2 + \frac{(G[\eta]q + \nabla'_H \eta \cdot \nabla'_H q)^2}{2(1 + |\nabla'_H \eta|^2)} + 2\nu V[\eta]q \tag{4.8}$$

where $V[\eta]q$ is the viscosity operator defined as

$$V[\eta]q = \left[\partial_{z'}\phi \Delta'_H \eta - \partial^2_{z'}\phi\right]_{z'=\eta} \tag{4.9}$$

As before, the first-order equations (4.5) and (4.8) capture the complete dynamics of the free surface, provided we can compute (4.4) and (4.9). Observe that the operator V is similar to G in the sense that it requires knowledge of the potential function ϕ in the bulk of the fluid. As a result, the techniques employed in Chapter 3 for the computation of Neumann data will also come in handy when we need to compute the dissipation terms.

4.2 The Modal Formulation

The subject of evaluating the DNO was covered extensively in Chapter 3. We present a short summary here. We first perform the change of coordinates (3.7) so that the problem is posed on a flat cylinder $C = D \times (-h, 0)$ instead of a domain deformed by the interface; here, D is the unit disc in the plane. The Laplacian of the potential in the new coordinates does not necessarily vanish, as in (4.1), since the change of coordinates may not be a conformal map. As a result, we obtain an equation of the form

$$-\Delta u = H(u, \eta) \tag{4.10}$$

for some operator H. Here, u is the symbol for the bulk potential in the new coordinates. Next, we write $\eta = \epsilon f$ and expand $u = \sum_{k\geq 0} \epsilon^k u_k$ as a formal power series; as established in Section 3.4, the value of ϵ is irrelevant to the success of this technique. Plugging these decompositions into the Poisson equation (4.10) and comparing coefficients of similar powers of ϵ yields a system of Poisson equations of the form

$$-\Delta u_k = \tilde{H}(u_{k-1}, u_{k-2}, f) \tag{4.11}$$

for $k \geq 0$ with associated boundary conditions. Since the right hand sides of (4.11) depend only on the lower order terms, the $\{u_k\}$ can be solved for sequentially up to a sufficiently high order K and combined to yield an approximation to u.

The procedure can also be used to get a recurrence relation for the DNO. Writing $G[\epsilon f]q = \sum_{k\geq 0} \epsilon^k G_k[f]q$ gives the formula (3.28)

$$\begin{aligned} G_k[f]q = {} & \partial_z u_k - h^{-1} f G_{k-1}[f]q + |\nabla_H f|^2 (\partial_z u_{k-2}) \\ & -\delta_{k,1} (\nabla_H f \cdot \nabla_H q) - \delta_{k,2} \left(h^{-1} f \nabla_H f \cdot \nabla_H q\right). \end{aligned} \tag{4.12}$$

We can similarly apply the technique for determining the viscosity operator $V[\eta]q$. Using (3.17) gives

$$(h + \eta)^2 V[\eta]q = h(h + \eta)(\partial_z u)(\Delta_H \eta) - h^2 \partial_z^2 u.$$

Using the expansion for u with $V[\epsilon f]q = \sum_{k\geq 0} \epsilon^k V_k[f]q$ and collecting similar powers of ϵ results in the recurrence formulas

$$\begin{aligned} V_k[f]q &= -\partial_z^2 u_k - 2h^{-1} f V_{k-1}[f]q - h^{-2} f^2 V_{k-2}[f]q + \\ &\quad (\Delta_H f)(\partial_z u_{k-1}) + h^{-1}(f\Delta_H f)(\partial_z u_{k-2}). \end{aligned} \tag{4.13}$$

In order to solve the equations (4.11) and hence evaluate the operators (4.12) and (4.13), we developed a fast spectral method in Section 3.3. This method operates by using the polynomial-based basis functions on C as defined in (3.30). The corresponding modal-nodal representation is extremely accurate and is computationally feasible. Indeed, the framework is so appealing for computational purposes that we extend their use to the integration of the first-order system (4.5) and (4.8).

We begin by expanding the unknowns η and q in terms of their Zernike-polynomial modal representations (2.6)

$$\eta(\rho,\theta,t) = \sum_{m,n} \alpha_{mn}(t)\zeta_{mn}(\rho,\theta), \quad q(\rho,\theta,t) = \sum_{m,n} \beta_{mn}(t)\zeta_{mn}(\rho,\theta) \tag{4.14}$$

where $\alpha_{mn}(t) = \langle \zeta_{mn}, \eta(\cdot,\cdot,t)\rangle_{L^2(D)}$ and $\beta_{mn}(t) = \langle \zeta_{mn}, q(\cdot,\cdot,t)\rangle_{L^2(D)}$. Projecting the equations (4.5) and (4.8) on ζ_{mn} leads to

$$\alpha'_{mn}(t) = \langle \zeta_{mn}, G[\eta]q\rangle_{L^2(D)} + 2\nu \langle \zeta_{mn}, \Delta_H \eta\rangle_{L^2(D)} \tag{4.15}$$

$$\begin{aligned} \beta'_{mn}(t) &= -(g - F(t)) \langle \zeta_{mn}, \eta\rangle_{L^2(D)} - \frac{1}{2}\left\langle \zeta_{mn}, |\nabla_H q|^2 \right\rangle_{L^2(D)} + \\ &\quad \frac{1}{2}\left\langle \zeta_{mn}, \frac{(G[\eta]q + \nabla_H \eta \cdot \nabla_H q)^2}{1 + |\nabla_H \eta|^2} \right\rangle_{L^2(D)} + 2\nu \langle \zeta_{mn}, V[\eta]q\rangle_{L^2(D)}. \end{aligned} \tag{4.16}$$

The modal representations of η and q allow us to compute the terms appearing in these equations. The orthonormality of Zernike polynomials yields $\langle \zeta_{mn}, \eta\rangle_{L^2(D)} = \alpha_{mn}$. Similarly, the computation of $\langle \zeta_{mn}, |\nabla_H q|^2\rangle_{L^2(D)}$ can be carried out by using the second operation type outlined in Section 2.3. Meanwhile, the term $\langle \zeta_{mn}, \Delta_H \eta\rangle_{L^2(D)}$ in (4.15) can be integrated by parts to yield

$$\langle \zeta_{mn}, \Delta_H \eta\rangle_{L^2(D)} = \frac{1}{\pi}\int_{\partial D} \zeta_{mn} \frac{\partial \eta}{\partial \mathbf{n}}\, ds - \frac{1}{\pi}\iint_D \nabla_H \zeta_{mn} \cdot \nabla_H \eta\, dA$$

where $\mathbf{n}$ is the unit outward normal vector to ∂D. Using the condition $\frac{\partial \eta}{\partial \mathbf{n}}\big|_{\partial D} = \eta_\rho|_{\rho=1} = 0$ and the expansion for η, we obtain

$$\langle \zeta_{mn}, \Delta_H \eta\rangle_{L^2(D)} = -\frac{1}{\pi}\sum_{m_1,n_1} \alpha_{m_1 n_1} \iint_D \nabla_H \zeta_{mn} \cdot \nabla_H \zeta_{m_1 n_1}\, dA. \tag{4.17}$$

The integrals appearing above are precisely the stiffness terms for which a closed form formula was worked out in Section 2.3.

The algorithm for the computation of the bulk potential u finds the unknown in terms of the basis functions $\{\psi_{mnj}\}$. Thus, the results for $G[\eta]q$ and $V[\eta]q$ are already known in terms of their Zernike-representation since they are evaluated at $j = J$. As the Zernike polynomials are orthonormal, the terms $\langle \zeta_{mn}, G[\eta]q \rangle_{L^2(D)}$ and $\langle \zeta_{mn}, V[\eta]q \rangle_{L^2(D)}$ are simply the coefficients in these representations. Finally, in order to find the modal representation of $\frac{(G[\eta]q+\nabla_H\eta\cdot\nabla_H q)^2}{1+|\nabla_H\eta|^2}$, we express it as

$$\frac{(G[\eta]q + \nabla_H\eta \cdot \nabla_H q)^2}{1 + |\nabla_H\eta|^2} = (G[\eta]q + \nabla_H\eta \cdot \nabla_H q)^2 \left(1 - |\nabla_H\eta|^2 + |\nabla_H\eta|^4 - |\nabla_H\eta|^6 + \ldots\right). \tag{4.18}$$

This expansion is valid as long as $|\nabla_H\eta| < 1$. Observe that this condition is precisely of the type required in Theorem 3.4.2. Also note that the products in (4.18) can be computed by using the first operation type described in Section 2.3. Thus, the entire right-hand side of the first order system can be computed efficiently in terms of the modal representation. As a result, we can express the system to be solved as

$$\begin{cases} \vec{\alpha}'(t) = P\left(t, \vec{\alpha}(t), \vec{\beta}(t)\right) \\ \vec{\beta}'(t) = Q\left(t, \vec{\alpha}(t), \vec{\beta}(t)\right) \\ \vec{\alpha}(0) = \vec{\alpha}_0, \quad \vec{\beta}(0) = \vec{\beta}_0 \end{cases} \tag{4.19}$$

4.3 Time Integration

For integrating the system (4.19), we employ a modified Spectral Deferred Corrections method. In this section, we provide a brief overview of the technique before describing our modifications and its application to our problem.

Spectral Deferred Correction Methods

Spectral deferred correction (SDC) methods were introduced in [17] as a solution refinement technique. An extension of classical deferred correction methods, this method replaces the unbounded and unstable differentiation operators by integration operators based on quadrature rules. An initial solution obtained from a simple method (say, a Runge-Kutta or a multi-step scheme) is refined by performing several sweeps. The added stability promises that the order of convergence is improved from the original $O((\Delta t)^p)$ to $O((\Delta t)^{(J+1)p})$, where J is the number of sweeps performed. In [22, 23], proofs of this fact are presented for both the cases where one-step and multi-step schemes are used for the initial solution. Much progress has been made on developing this technique further, such as its acceleration and its extension to semi-implicit methods [26, 41].

Classical deferred corrections operate by developing, at each stage of a refinement sweep, a differential equation satisfied by the residual $\delta(t)$ between the true solution and the numerical solution. This equation uses the Lagrange interpolating polynomial in place of the

numerical solution at some s points. Using the same pth order method on this equation gives an approximation to $\delta(t)$ which, when added to the previous numerical solution, increases the order of accuracy from $O(h^m)$ to $O(h^{m+p})$. In principle, one can keep performing these corrections to get solutions that are accurate to any arbitrary order as long as the interpolation is accurate. In practice, however, one is constrained by the instabilities introduced by the differentiation operator which is required in the differential equation above. This discourages the use of a large number of interpolation points and hence limits the accuracy of the procedure.

This problem was solved by [17] by replacing the differential equation for δ by a Picard equation. The integrals required in this equation were computed by integrating the corresponding Lagrange interpolating polynomials; in practice, this was represented as a linear operator acting on the function sampled at s points. The spectral accuracy of the integration matrix supplied the "spectral" in the title of the technique. Even this method (and all its extensions), however, used a primitive Forward Euler technique to solve the Picard equation (and Backward Euler for stiff equations). Thus, the only point at which the highly accurate and stable integration operator was made use of was while prescribing the equation obeyed by the residual.

Picard Refinement Scheme

Here, we consider a modification of this method where the refinements are carried out entirely by application of the integration matrix. In addition, we completely do away with the definition of the residual $\delta(t)$ and consider directly the Picard equation.

Consider the first order scalar problem

$$\begin{cases} y'(t) = f(t,y), & a \le t \le b \\ y(a) = y_a. \end{cases}$$

We then have the equivalent Picard equation

$$y(t) = y_a + \int_a^t f(\tau)\ d\tau \tag{4.20}$$

for any $t \in [a,b]$. We refine this interval by introducing equi-spaced points $\{t_i\}_{0\le i\le N}$ with step-size $\Delta t = (b-a)/N$ and $t_i = a + i\Delta t$. Furthermore, we chop up every sub-interval $[t_i, t_{i+1}]$ into

$$t_i = t_i^{(0)} < t_i^{(1)} < \ldots < t_i^{(s-1)} < t_i^{(s)} = t_{i+1}$$

where $\{t_i^{(j)}\}_{0\le j\le s-1}$ are the s left-handed Radau quadrature points on $[t_i, t_{i+1}]$.

We obtain $\phi_0(t)$, as above, from a Runge-Kutta or a multi-step scheme. Let $\phi_k(t)$ be the numerical solution after k refinements on $[t_i, t_{i+1}]$ and let $\phi(t)$ be the accepted numerical solution for $t \le t_i$. In a pure Picard iteration, the next refined solution is obtained by

$$\phi_{k+1}(t) = \phi(t_i) + \int_{t_i}^t f(\tau, \phi_k(\tau))\ d\tau. \tag{4.21}$$

Observe that (4.21) can be seen as a fixed-point iteration for solving the corresponding Picard equation. This can also be used to furnish a simple convergence proof: assume that $\phi(t_i) = y(t_i)$ and that, if L is the Lipschitz constant for f in the y variable, then there exists some K such that $(\Delta t)L \leq K < 1$. Then

$$\begin{aligned}
|\phi_{k+1}(t) - y(t)| &= \left| \int_{t_i}^{t} f(\tau, \phi_k(\tau)) - f(\tau, y(\tau)) \, d\tau \right| \\
&\leq \int_{t_i}^{t} |f(\tau, \phi_k(\tau)) - f(\tau, y(\tau))| \, d\tau \\
&\leq \int_{t_i}^{t} L|\phi_k(\tau) - y(\tau)| \, d\tau \\
\Rightarrow \|\phi_{k+1} - y\|_{L^\infty([t_i, t_{i+1}])} &\leq L \|\phi_k - y\|_{L^\infty([t_i, t_{i+1}])} \int_{t_i}^{t_{i+1}} 1 \, d\tau \\
&\leq L(\Delta t) \|\phi_k - y\|_{L^\infty([t_i, t_{i+1}])} \\
&\leq K \|\phi_k - y\|_{L^\infty([t_i, t_{i+1}])} .
\end{aligned}$$

As $K < 1$, we conclude that $\lim_{k\to\infty} \|\phi_k - y\|_{L^\infty([t_i, t_{i+1}])} = 0$.

For a stronger statement about (4.21), let $\phi_k(t) - y(t) = O((\Delta t)^r)$ for $t \in [t_i, t_{i+1}]$ and for some integer $r \geq 0$. The above calculation then shows that $\phi_k(t) - y(t) = O((\Delta t)^{r+1})$ so the Picard refinement improves the order of accuracy by one at each sweep.

In practice, we evaluate (4.21) only at the points $\{t_i^{(j)}\}_{0\leq j\leq s-1}$ since that aids us in the iterative process. Only when we have the most refined solution do we compute the value at $t = t_i^{(s)} = t_{i+1}$ and the process is repeated for the next sub-interval.

In order to describe how the integration in (4.21) is carried out, let g be a function with values $g_j = g\left(t_i^{(j)}\right)$ for $0 \leq j \leq s-1$ and let $G_j = \int_{t_i}^{t_i^{(j)}} g(t) \, dt$. We can approximate the integrals by replacing g by the Lagrange interpolating polynomial constructed from $\{g_j\}_{0\leq j\leq s-1}$:

$$G_j \approx \int_{t_i}^{t_i^{(j)}} \sum_{l=0}^{s-1} g_l L_l(t) \, dt = \sum_{l=0}^{s-1} g_l \int_{t_i}^{t_i^{(j)}} L_l(t) \, dt = \sum_{l=0}^{s-1} g_l \int_{t_i}^{t_i^{(j)}} \prod_{0\leq m\leq s-1, m\neq l} \frac{t - t_i^{(m)}}{t_i^{(l)} - t_i^{(m)}} \, dt.$$

Define the $s \times s$ matrix S by

$$S_{jl} = \int_{t_i}^{t_i^{(j)}} \prod_{0\leq m\leq s-1, m\neq l} \frac{t - t_i^{(m)}}{t_i^{(l)} - t_i^{(m)}} \, dt.$$

We then have $\mathbf{G} \approx S\mathbf{g}$. These matrix entries can be calculated by breaking them up into

$$S_{jl} = \sum_{r=1}^{j} \int_{t_i^{(r-1)}}^{t_i^{(r)}} \prod_{0\leq m\leq s-1, m\neq l} \frac{t - t_i^{(m)}}{t_i^{(l)} - t_i^{(m)}} \, dt$$

and evaluating the integrals $\int_{t_i^{(r-1)}}^{t_i^{(r)}} \prod_{0\le m\le s-1, m\neq l} \frac{t-t_i^{(m)}}{t_i^{(l)}-t_i^{(m)}}\, dt$ for $1 \le r \le s-1$ by using the Radau quadrature rule.

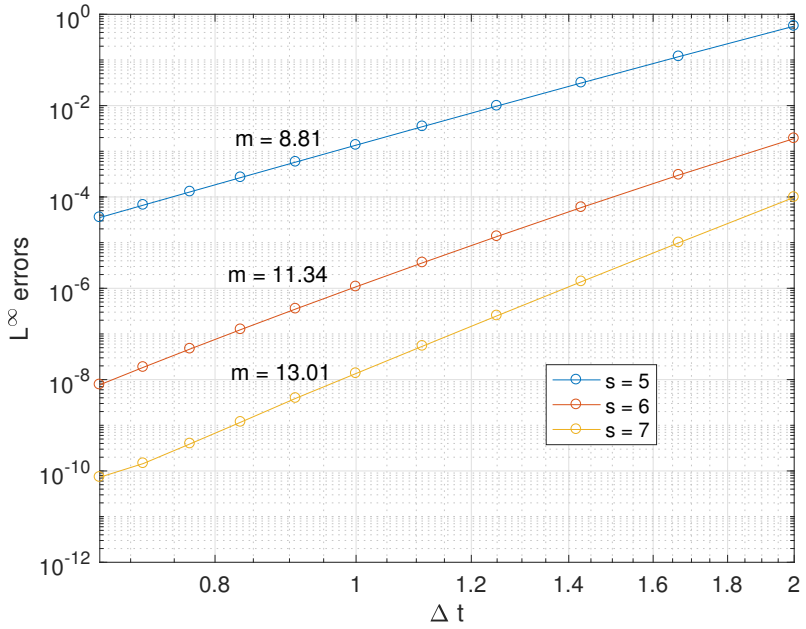

Figure 4.1: $L^\infty([0, 10])$ errors vs Δt for the test problem (4.22) solved by the s-point left-hand Radau Picard refinement scheme. The initial solution was computed by the RK4 method, followed by $(2s-5)$ refinement sweeps.

While the order of accuracy is increased by one at each step of the Picard refinement scheme, it is bounded by the accuracy of the integration procedure. The s-point left-handed Radau quadrature scheme is accurate for polynomials up to degree $2s-2$; thus, it is able to produce solutions accurate to $O((\Delta t)^{2s-1})$. If a pth order method is used for the initial solution, we need to perform $(2s-1-p)$ refinements to attain the maximal order. We use $p=4$ as it is the highest order for which the number of stages equals the order of the method. Figure 4.1 shows the convergence plots for the test problem

$$\begin{cases} y'(t) = t + y(t), & 0 \le t \le 10 \\ y(0) = 1 \end{cases} \tag{4.22}$$

with the exact solution $y(t) = -1 - t + (y(0)+1)e^t$. As claimed, the rates of convergence are fairly close to $O((\Delta t)^{2s-2})$. In practice, one can stop refining the solution once the norm of the corrections falls below a pre-determined tolerance.

This method can be suitably extended to solving the system (4.19). For every i, we first find the coarse solution $\left(\vec{\alpha}^{(0)}(t_i^{(j)}), \vec{\beta}^{(0)}(t_i^{(j)})\right)_{0\le j\le s-1}$ by the Runge-Kutta-4 method. Here,

we are interpreting the vectors $\vec{\alpha}$ etc. as row vectors with j indexing the rows. Having obtained an acceptable solution $\left(\vec{\alpha}(t_i), \vec{\beta}(t_i)\right)_{0\le j\le s-1}$ at the end of the $(i-1)$st step, the solutions are successively refined by computing, for $k \ge 0$,

$$\left(\vec{\alpha}^{(k+1)}(t_i^{(j)}), \vec{\beta}^{(k+1)}(t_i^{(j)})\right) = \left(\vec{\alpha}(t_i), \vec{\beta}(t_i)\right) + S\left(P\left(t_i^j, \vec{\alpha}^{(k)}(t_i^{(j)})\right), Q\left(t_i^j, \vec{\beta}^{(k)}(t_i^{(j)})\right)\right).$$

4.4 Linear Results

We now apply the machinery we have assembled to the Faraday wave problem. Note that simply setting the external forcing to $F(t) = \alpha\cos(\omega t)$ takes us into the Faraday regime. The aim of the next two sections is to simulate Faraday waves in this setting and compare them with theory and experiments.

In this section, we present the results in the linear regime. Note that our approach makes it extremely easy to control the strength of nonlinearity by fixing the shape of the interface f and varying ϵ. In addition, we can specify the largest order of nonlinear interactions by fixing K. For the linear regime, we set $K = 0$ and $\epsilon = 10^{-6}$. In addition, we use $g = 981$ cm s^{-2}, $\nu = 3.0 \times 10^{-3}$ cm^2 s^{-1} and $h = 0.5$ cm and set the computational parameters to $M = 8$, $N = 10$ and $J = 20$.

We begin by determining the theoretical critical acceleration thresholds for given values of ω and compare them with the numerical simulations. We also study the variation in the nature of the resonances by performing a temporal-spectral analysis of the waveforms.

The Damped Mathieu Equation

We first simplify the system (4.5) and (4.8) by keeping only the linear terms and dropping all quadratic terms and higher. We obtain

$$\partial_t \eta = \partial_{z'}\phi|_{z'=\eta} + 2\nu\Delta'_H\eta \tag{4.23}$$

$$\partial_t q = -(g - \alpha\cos(\omega t))\eta - 2\nu\partial^2_{z'}\phi|_{z'=\eta} \tag{4.24}$$

where ϕ solves Laplace's equation (4.1) with the no-flow boundary conditions at the walls and $\phi|_{z'=\eta} = q$. The general solution to ϕ is

$$\phi(\rho', \theta', z', t) = \sum_{m\ge 0, n\ge 1} J_m(a_{mn}\rho')\left(b^{(1)}_{mn}(t)\cos(m\theta') + b^{(2)}_{mn}(t)\sin(m\theta')\right)\frac{\cosh(a_{mn}(z'+h))}{\cosh(a_{mn}h)} \tag{4.25}$$

where J_m is the mth Bessel function of order zero and a_{mn} is the nth zero of $J'_m(\cdot)$. Let us focus on the evolution of a normal mode of the system (i.e. an eigenfunction of Δ'_H with homogeneous Neumann conditions) $X(\rho', \theta') = J_m(a_{mn}\rho')\cos(m\theta')$. Write

$$\eta(\rho', \theta', t) = A(t)X(\rho', \theta') \text{ and } q(\rho', \theta', t) = B(t)X(\rho', \theta').$$

As $\phi|_{z'=\eta} = \phi|_{z'=0}$ to leading order, we have

$$\phi(\rho', \theta', z', t) = q(\rho', \theta', t)\frac{\cosh(a_{mn}(z'+h))}{\cosh(a_{mn}h)}.$$

Using these decompositions in (4.23) and (4.24) yields

$$\begin{aligned} A'(t) &= a_{mn}\tanh(a_{mn}h)B(t) - 2\nu a_{mn}^2 A(t), \\ B'(t) &= -(g - \alpha\cos(\omega t))A(t) - 2\nu a_{mn}^2 B(t). \end{aligned}$$

Combining them gives

$$A''(t) + 2\delta_{mn}A'(t) + \left[\omega_{mn}^2\left(1 - \frac{\alpha}{g}\cos(\omega t)\right) + \delta_{mn}^2\right]A(t) = 0 \tag{4.26}$$

where $\omega_{mn}^2 = a_{mn}g\tanh(a_{mn}h)$ is the dispersion relation for finite-depth waves and $\delta_{mn} = 2\nu a_{mn}^2$ is the damping coefficient. This is referred to as the damped Mathieu equation.

The amplitude of every mode therefore satisfies an equation of this form. For a given value of ω, the α vs. a_{mn} plane is split into regions depending on whether the flat interface $\eta \equiv 0$ is stable or unstable in that parameter regime. The boundaries of the unstable regions indicate the critical amplitude α_c/g: for $\alpha < \alpha_c/g$, we have exponential decay while $\alpha > \alpha_c/g$ leads to exponential growth, causing the small amplitude assumption to be violated.

In order to determine the critical value α_c/g of the forcing amplitude, we use the technique developed in [32]. This method relies on the observation that the explicit time dependence in (4.26) occurs via a periodic function so its solutions must be of the Floquet form

$$A(t) = e^{(\mu+i\xi)t}\sum_{n\in\mathbb{Z}} A_n e^{in\omega t}$$

where $(\mu + i\xi)$ is the Floquet exponent. If $\xi = 0$, the response is harmonic with the realty conditions $A_{-n} = \overline{A_n}$ while $\xi = \frac{\omega}{2}$ is yields a subharmonic response with $A_{-(n+1)} = \overline{A_n}$. Plugging this ansatz into (4.26), decomposing $\cos(\omega t) = \frac{e^{i\omega t}+e^{-i\omega t}}{2}$ and comparing coefficients of $e^{in\omega t}$ yields

$$\frac{\alpha}{g}\left(A_{n-1} + A_{n+1}\right) = \frac{2}{\omega_{mn}^2}\left[\omega_{mn}^2 + (\mu + i(\xi + n\omega) + \delta_{mn})^2\right]A_n. \tag{4.27}$$

By restricting $n \leq N$ for some large enough N, this is expressible as an eigenvalue problem for α/g. Choosing $\mu = 0$ allows us to search for steady solutions and the corresponding critical accelerations α_c/g. Setting $\xi = 0$ or $\frac{\omega}{2}$ switches between a harmonic or subharmonic response respectively.

Comparing the Critical Amplitudes

In order to numerically determine the rate of growth for a certain α, we need to measure the energy of the system over time. Define the energy function

$$E(t) = \frac{1}{2}\iiint_V |\nabla'\phi|^2 \, dV + \frac{1}{2}g\iint_D \eta^2 \, dA.$$

It can be shown that in the inviscid case with no forcing, this function is constant. Hence, it serves as an ideal device to study the energy content of the system. Note that we can write

$$\iiint_V |\nabla'\phi|^2 \, dV = \iint_{\partial V} \phi\frac{\partial\phi}{\partial\mathbf{n}} \, dS - \iiint_V \phi\Delta'\phi \, dV = \iint_D qG[\eta]q \, dA$$

because of (4.1) and the boundary conditions. Thus, we have

$$E(t) = \frac{1}{2}\iint_D qG[\eta]q + g\eta^2 \, dA.$$

Mode	Type	$\omega/2\pi$ (Hz)	$\omega_{mn}/2\pi$ (Hz)	α_c/g (theor.)	α_c/g (comp.)	Rel. Error
$(0,1)$	SH	18.6368	9.54854	0.0960226	0.0960248	2.34×10^{-5}
$(3,1)$	SH	20.6540	10.0655	0.104835	0.104833	-1.77×10^{-5}
$(2,1)$	H	8.26893	8.31017	0.157309	0.157312	2.08×10^{-5}
$(1,3)$	H	14.4892	14.5615	0.200724	0.200725	1.90×10^{-6}
$(1,2)$	SH	7.24446	11.4545	0.625639	0.625639	6.55×10^{-12}
$(0,3)$	SH	10.0555	15.8991	0.629509	0.629509	-1.11×10^{-8}

Table 4.1: Comparison of critical amplitudes for various modes in different regimes. SH stands subharmonic and H for harmonic.

This formula is preferred to the earlier one since it uses the surface quantities only, which are exactly what we are explicitly updating at each time-step. In addition, using the Zernike modal representation considerably simplifies the computation of the integral since it is replaced by

$$E(t) = \frac{\pi}{2}\sum_{m,n} \langle q, \zeta_{mn}\rangle_{L^2(D)} \langle \zeta_{mn}, G[\eta]q\rangle_{L^2(D)} + g\left|\langle \zeta_{mn}, \eta\rangle_{L^2(D)}\right|^2.$$

Given any time interval $[t_i, t_{i+1}]$, we can estimate the growth rate at t_i by

$$\tilde{\mu} = \frac{1}{t_{i+1} - t_i}\log\left(\frac{E(t_{i+1})}{E(t_i)}\right).$$

For a given wave-number a_{mn}, determining the critical amplitude α_c/g is carried out by calculating the growth rates for a number of different amplitudes and interpolating between them to find the value that results in zero growth. Initially, the interface is at rest and set in the shape of the appropriate mode with noise of the same magnitude added in. The noise eventually dissipates away, leaving only the desired mode, which in turn dictates the growth rate. Table 4.1 compares the theoretical and computed critical thresholds for a few modes for various excitation types. Note that the errors are much smaller than the corresponding values calculated in [49].

The primary difficult in studying linear excitation is that a mode other than the desired one may also be excited and eventually grow uncontrollably. Essentially, this is because the solutions are super-positions of various modes and, as nonlinear effects are neglected, the unconstrained growth cannot be reined in. For this reason, we have carefully chosen the background frequency ω so that only the desired mode is excited.

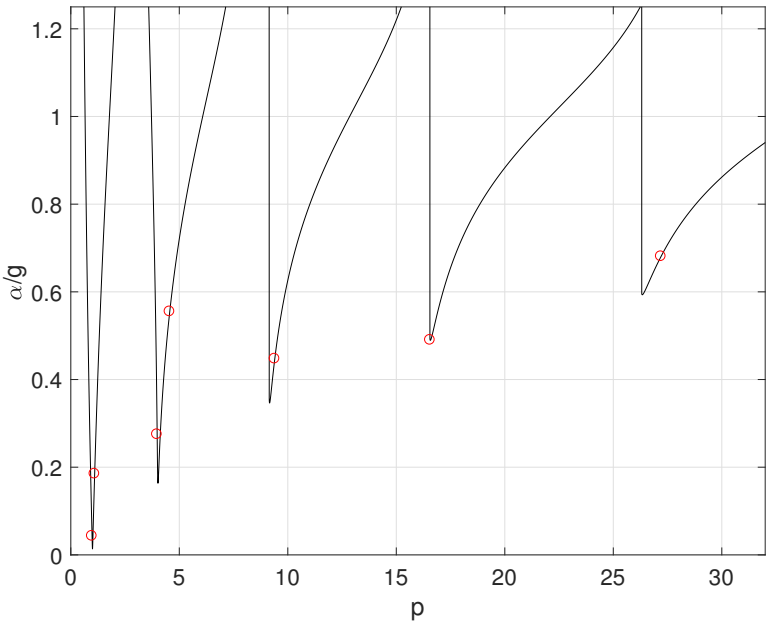

Figure 4.2: Critical amplitudes vs. p. The solid lines are the theoretically calculated values while the circles indicate the results from the simulation.

We run the simulations for 600 forcing periods; the pattern typically stabilizes much earlier than that but we find an extended run to be useful for predicting the critical accelerations more accurately. Each forcing period is chopped into 20 equal intervals; the equations are solved by RK4 and refined by the 8-point left-Radau Picard scheme described earlier. The tolerance for the Picard refinement is set at 10^{-12}. The run-time is around 20 minutes on a single core (the performance in fact suffered upon parallelization).

	Type	p_{22}	$\omega/2\pi$ (Hz)	α_c/g (theor.)	α_c/g (comp.)	Rel. Error
(i)	SH	0.98	26.0481	0.0428505	0.0428505	-7.45×10^{-7}
(ii)	SH	1.10	24.5863	0.184652	0.184652	1.93×10^{-8}
(iii)	H	3.98	12.9255	0.274609	0.274609	2.46×10^{-11}
(iv)	H	4.56	12.0756	0.554786	0.554786	-1.09×10^{-12}
(v)	SH	9.40	8.41057	0.447046	0.447046	-2.22×10^{-12}
(vi)	H	16.56	6.33664	0.489857	0.489857	-1.44×10^{-8}
(vii)	SH	27.20	4.94430	0.680603	0.680606	3.69×10^{-6}

Table 4.2: Comparison of critical amplitudes for various types of excitations undergone by the mode $(2, 2)$. SH stands subharmonic and H for harmonic.

We next focus on a particular mode to study the possible variation in the types of excitations. As in [5], define

$$p_{mn} = \frac{4w_{mn}^2}{w^2} \tag{4.28}$$

for the mode (m, n). This quantity is an extremely useful way of determining the nature of the excitation that the mode (m, n) undergoes for a given ω. For infinitesimal forcing in the inviscid case, the instability occurs only when p_{mn} is a perfect square with period $\frac{\sqrt{p_{mn}}}{2}\omega$. This precise relationship is somewhat lost in the presence of viscosity and finite forcing. However, the parity of the integer β closest to $\sqrt{p_{mn}}$ determines the nature of the excited wave: subharmonic for odd values of β and harmonic for even. Moreover, even though the standing wave contains multiple time modes, the dominant frequency is $\frac{\beta}{2}\omega$. The proximity of β to $\sqrt{p_{mn}}$ suggests that p_{mn} is still useful in illustrating the type of excitation. In Table 4.2, we present the results for the mode $(2, 2)$ (for this mode, we have wavenumber $a_{22} = 6.70613$ cm^{-1} and natural frequency $\omega_{mn}/2\pi = 12.8932$ Hz). In addition, Figure 4.2 gives the critical acceleration against p_{22}, comparing the theoretical and numerically calculated values.

Figures 4.3 – 4.7 show the temporal profiles $f(t)$ for some of the cases shown in Table 4.2. The numerical results are calculated by using the computed critical acceleration values. The temporal profiles are obtained by evaluating the interface at the point $(-0.3531, -0.3016)$ as that leads to significant variation. Note the high degree of agreement between theoretical and numerical results. Letting T stand for the period of forcing, it can be seen that the period of a subharmonic (harmonic) excitation is $2T$ (T). Moreover, the profiles change from mostly sinusoidal to ones containing many different sinusoidal modes. Also shown are the (absolute valued) Fourier spectra for these profiles; ω_1 stands for the frequencies appearing in the profile. Note that the peaks occur only at odd (even) multiples of $\omega/2$ for a subharmonic (harmonic) excitation. Moreover, the largest peaks occur at $\omega_1 = j\omega/2$, where j is the index of the corresponding stability tongue in Figure 4.2.

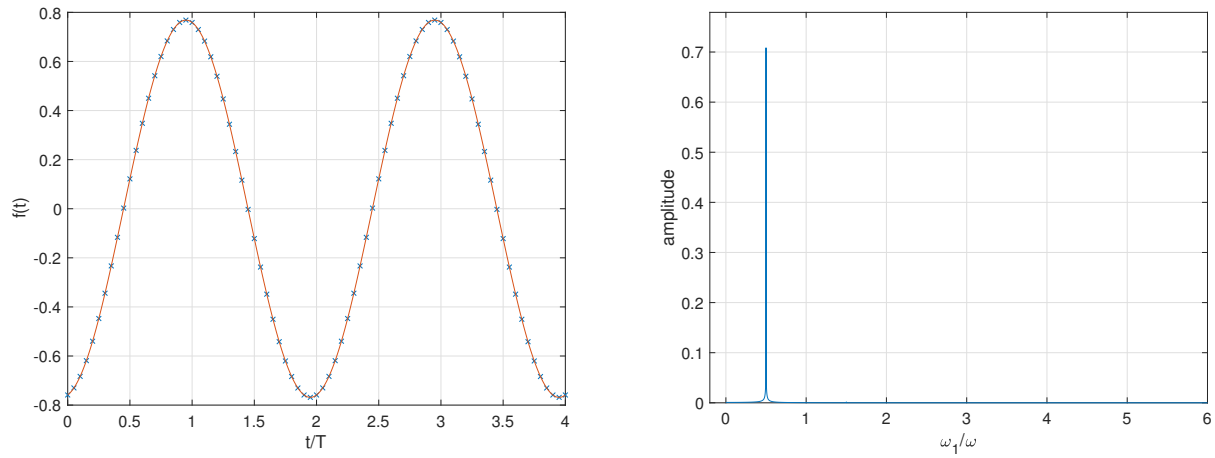

Figure 4.3: Temporal and spectral profiles for case (i) in Table 4.2. The solid line is obtained from the damped Mathieu equation and the crosses from the numerical results. Observe that the time profile is almost perfectly sinusoidal with period $2T$.

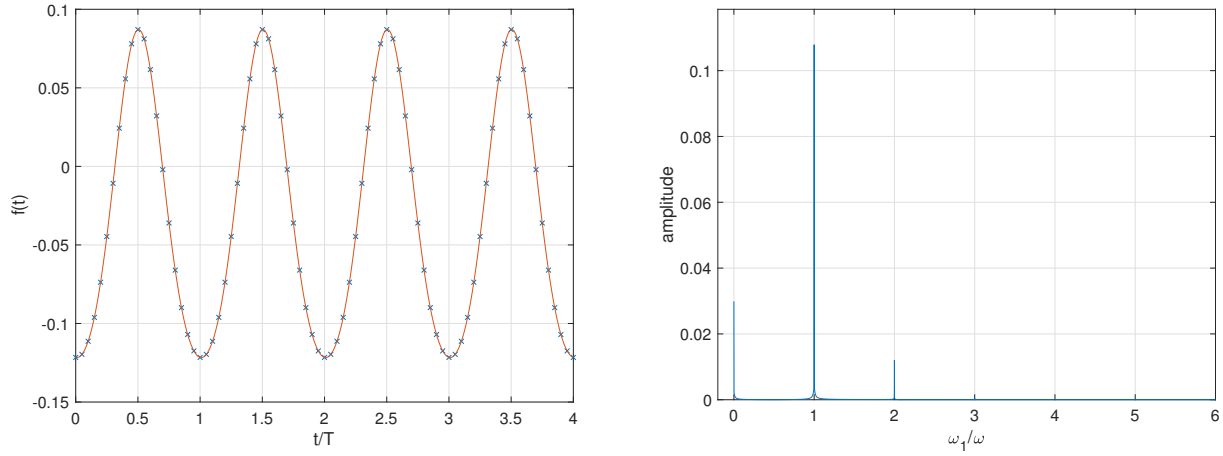

Figure 4.4: Temporal and spectral profiles for case (iv) in Table 4.2. Observe that the time profile is almost sinusoidal with period T. There is a constant offset and a small contribution of period $T/2$.

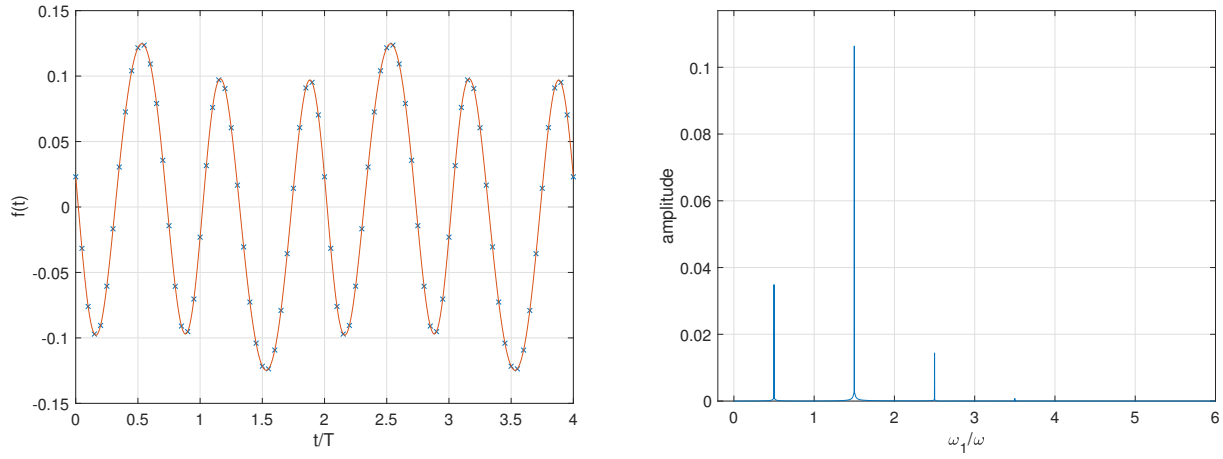

Figure 4.5: Temporal and spectral profiles for case (v) in Table 4.2. While the period is still $2T$ for this subharmonic response, the dominant contribution has period $2T/3$.

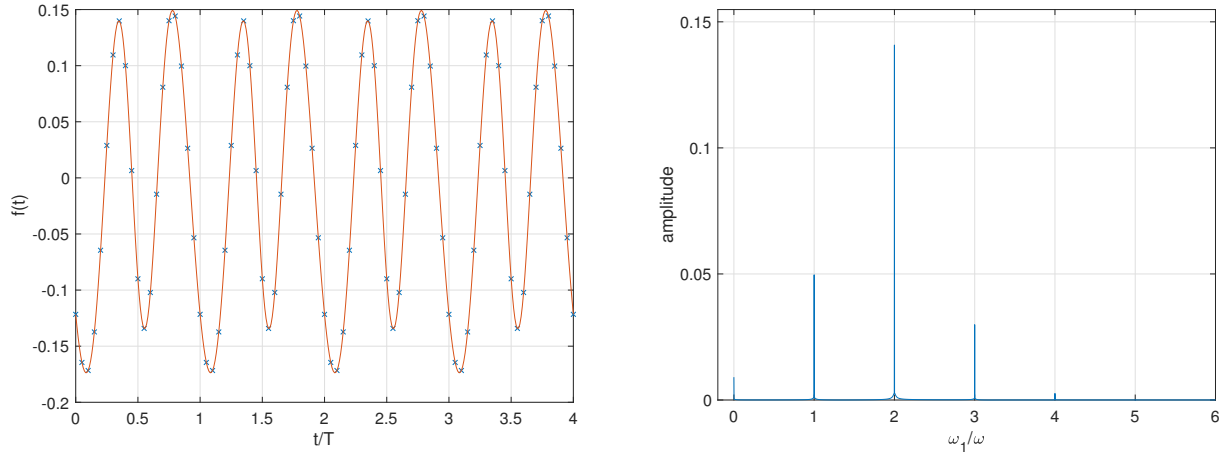

Figure 4.6: Temporal and spectral profiles for case (vi) in Table 4.2. Similar to Figure 4.5, the dominant contribution period $T/2$ while the overall period is T.

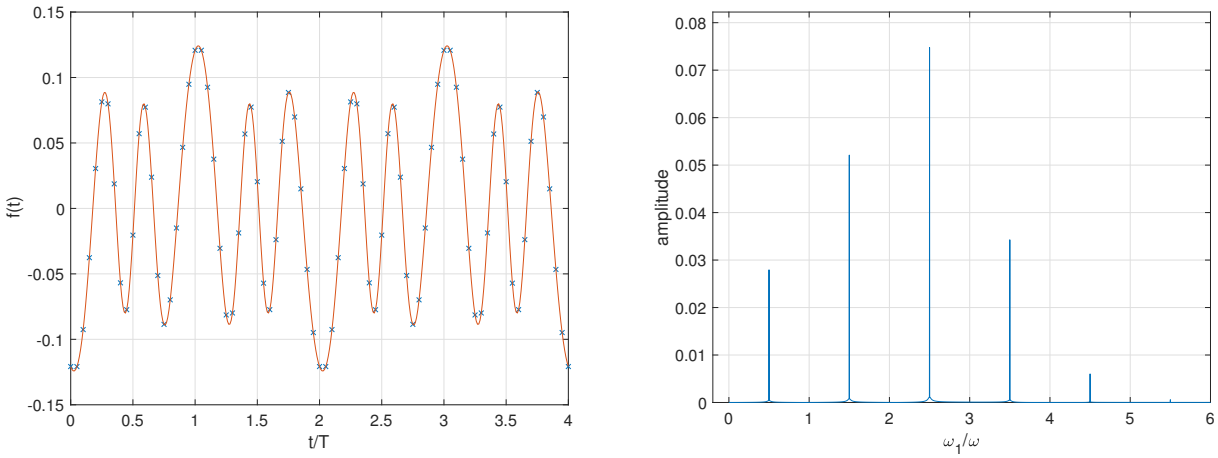

Figure 4.7: Temporal and spectral profiles for case (vii) in Table 4.2. The dominant period is $2T/5$ with overall period $2T$.

4.5 Nonlinear Results

Having seen that our algorithm reproduces the linear results extremely accurately, we can now apply it to the nonlinear problem. As mentioned at the beginning of the last section, our formulation allows us to precisely control the degree of nonlinearity. Firstly, by choosing ϵ to be relatively large, the quadratic terms in (4.8) become relevant. In addition, the highest order of allowed nonlinear interactions is given by $(K+1)$. For instance, by choosing $K = 3$, at most fourth-order interactions are allowed and hence we are in the weakly nonlinear regime. Since theoretical results so far have only been worked out under these assumptions, we shall compare our algorithm against them. Setting K to higher values leads to fully nonlinear Faraday waves. In these regimes, therefore, our algorithm breaks fresh ground and aims, in the future, to serve as a benchmark for the validation of theoretical results, and to help guide future experiments.

Stability Curves

The parameter values that lead to the onset of instability in the nonlinear regime were first predicted by Miles in [37] and improved in [39] to account for capillarity. Their method assumed that a primary mode was excited sub-harmonically and that the forcing acceleration was small. Keeping terms up to the fourth order and factoring in linear dissipation, they determined the dependence of the response on the forcing parameters. While the resulting curves are qualitatively similar to those obtained in the linear regime in the previous section, the presence of the nonlinear terms induces a minor shift (see Figure 4.8).

As before, we can use our simulation to determine the instability parameters. The earlier approach of calculating the growth rates for various acceleration values and interpolating

between them to obtain zero growth is still applicable since it takes into account the entire energy of the system. However, as some higher-order modes may be produced by quadratic interactions and retain some energy, the discretization parameters M and N need to be sufficiently large to account for them.

Figure 4.8 compares the theoretical and numerical critical accelerations for the axisymmetric mode $(0, 1)$. In order to reproduce the theoretical assumptions, we set $K = 3$ and $\epsilon = 0.1$. Observe that the results are in close agreement with the nonlinear predictions. Moreover, the deviations from the curve corresponding to the linear regime show that the nonlinear corrections are indeed accounted for correctly.

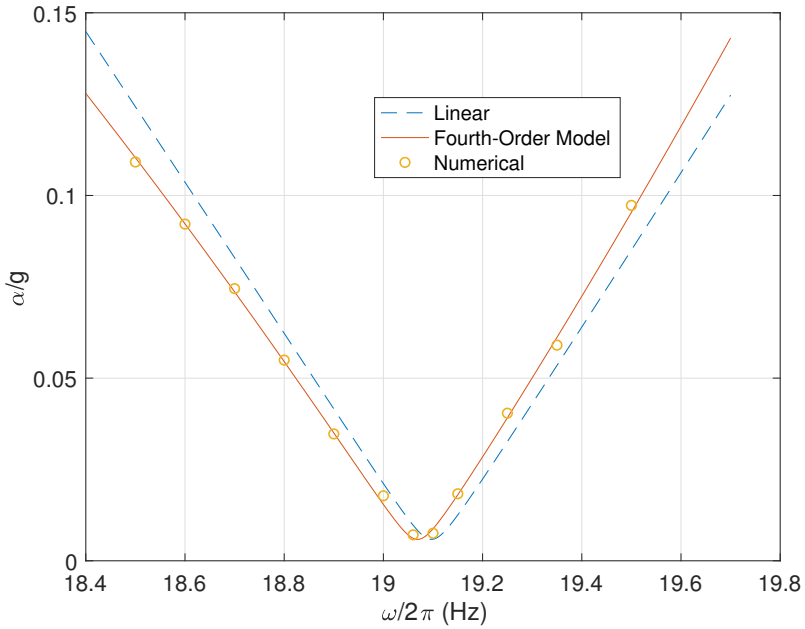

Figure 4.8: Numerically computed instability parameters for the mode $(0, 1)$ compared against the nonlinear predictions from [37] and the linear regime from the previous section.

The comparison is however limited by the fact that the theoretical predictions are only applicable for small values of the forcing acceleration and frequency deviation $(\omega_{mn} - \omega/2)$. In addition, none of the secondary modes that may be produced by quadratic interactions should be resonantly unstable. This requires that the forcing frequency be chosen extremely carefully.

Secondary Modes

Quadratic interactions between primary modes may give rise to secondary modes. Let

$$X_1(\rho, \theta, t) = \beta_{m_1 n_1}(t) J_{m_1}(a_{m_1 n_1}\rho) e^{i m_1 \theta}, \quad X_2(\rho, \theta, t) = \beta_{m_2 n_2}(t) J_{m_2}(a_{m_2 n_2}\rho) e^{i m_2 \theta}$$

describe the evolution of a pair of primary modes. Since we can write

$$J_{m_1}(a_{m_1 n_1}\rho)J_{m_2}(a_{m_2 n_2}\rho) = \sum_{n\geq 1} c_n J_{m_1+m_2}(a_{m_1+m_2,n}\rho)$$

for some c_n, it follows that

$$[X_1 X_2](\rho,\theta,t) = \beta_{m_1 n_1}(t)\beta_{m_2 n_2}(t)\sum_{n\geq 1} c_n J_{m_1+m_2}(a_{m_1+m_2,n}\rho)e^{i(m_1+m_2)\theta}.$$

Thus, we expect that modes with angular frequency (m_1+m_2) will appear. Moreover, the amplitudes of these new modes should be proportional to the product $\beta_{m_1 n_1}(t)\beta_{m_2 n_2}(t)$ of the amplitudes of the primary modes. This process of generating new modes can be expressed concisely by the notation $(m_1, n_1) + (m_2, n_2) \to (m_1 + m_2, n)$.

In order to realize this phenomenon, we choose the forcing parameters such that two modes may be excited simultaneously. This is achieved by superimposing the stability plots of α vs. ω for different modes, identifying pairs of modes whose curves intersect in a relatively sparse region (so that spurious modes are not excited) and finding their points of intersection. We use the stability curves from the fourth-order model above, despite the fact that it is derived under the assumption that exactly one primary mode is excited. In a sense, our numerical experiments serve as tests to determine the widest applicability of the theory. Table 4.3 lists the parameter choices for various pairs of modes.

	Modal Pairs	$\omega/2\pi$ (Hz)	α_c/g
(i)	$(2,1)$ & $(0,1)$	17.542225	0.259107
(ii)	$(0,1)$ & $(3,1)$	19.535932	0.103774
(iii)	$(3,1)$ & $(1,2)$	21.169971	0.242101

Table 4.3: Parameters for simultaneous excitation of a pair of modes via the fourth-order nonlinear model.

Figure 4.9 shows the results for case (iii) from Table 4.3 (with $K = 3$ and $\epsilon = 0.01$). The amplitudes of the newly generated modes are plotted against the products of parent amplitudes and the correlations computed. The high correlation coefficients indicate that the mechanism for the excitation of these modes is precisely the one suggested above. The secondary modes may in turn interact with other primary or secondary modes to generate further modes. Figure 4.10 demonstrates this for two such pairs.

The frequency of the response for a new mode can be worked out by considering the frequencies of the parent modes: if $\beta_{m_j n_j}(t) \sim e^{ir_j\omega t} + e^{-ir_j\omega t}$ for $j = 1, 2$, then

$$\beta_{m_1 n_1}(t)\beta_{m_2 n_2}(t) \sim e^{i(r_1+r_2)\omega t} + e^{-i(r_1+r_2)\omega t} + e^{i(r_1-r_2)\omega t} + e^{-i(r_1-r_2)\omega t}.$$

Thus, the frequency of the resulting mode should be $|r_1 - r_2|\omega$ if $r_1 \neq r_2$ and $|r_1 + r_2|\omega$ otherwise. This is confirmed by the time and spectral profiles of the various modes in Figure

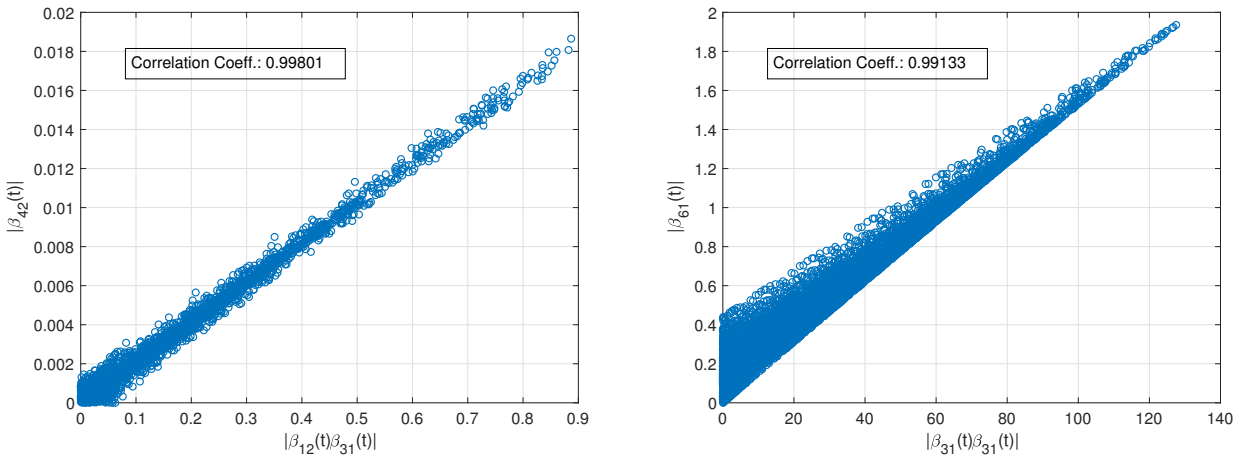

Figure 4.9: Amplitudes of secondary modes $|\beta_{mn}|$ vs. products of amplitudes $|\beta_{m_1n_1}\beta_{m_2n_2}|$ of the primary modes. On the left, we have $(1,2)+(3,1) \to (4,2)$ and on the right, $(3,1)+(3,1) \to (6,1)$.

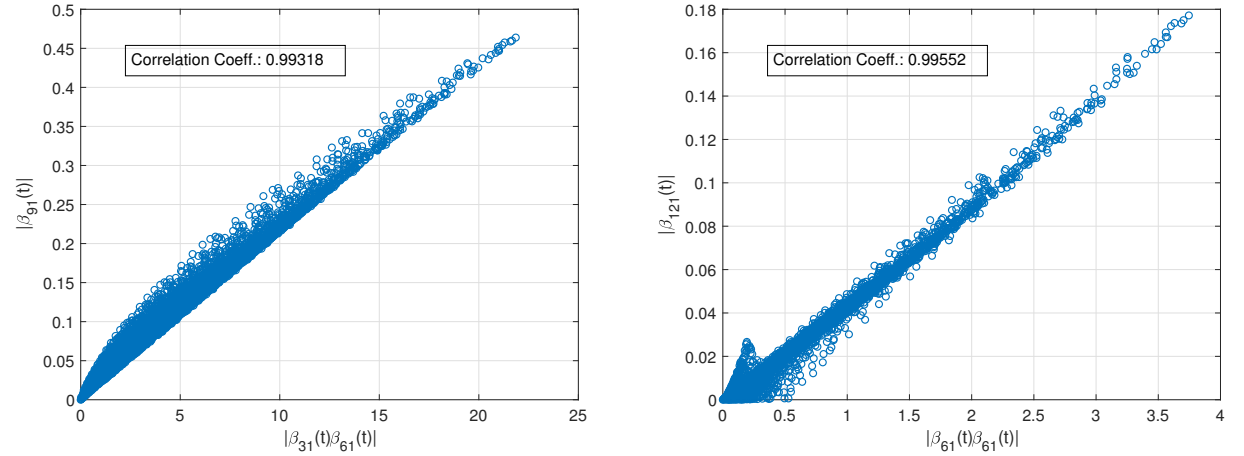

Figure 4.10: Secondary modes may interact with primary or other secondary modes to produce tertiary modes. On the left, we have $(3,1)+(6,1) \to (9,1)$ and on the right, $(6,1)+(6,1) \to (12,1)$.

4.11. The mode $(3,1)$ is excited sub-harmonically; as $(3,1)+(3,1)\rightarrow(6,1)$, the response for this mode should be harmonic. Similarly, $(3,1)+(6,1)\rightarrow(9,1)$ implies that $(9,1)$ is sub-harmonic.

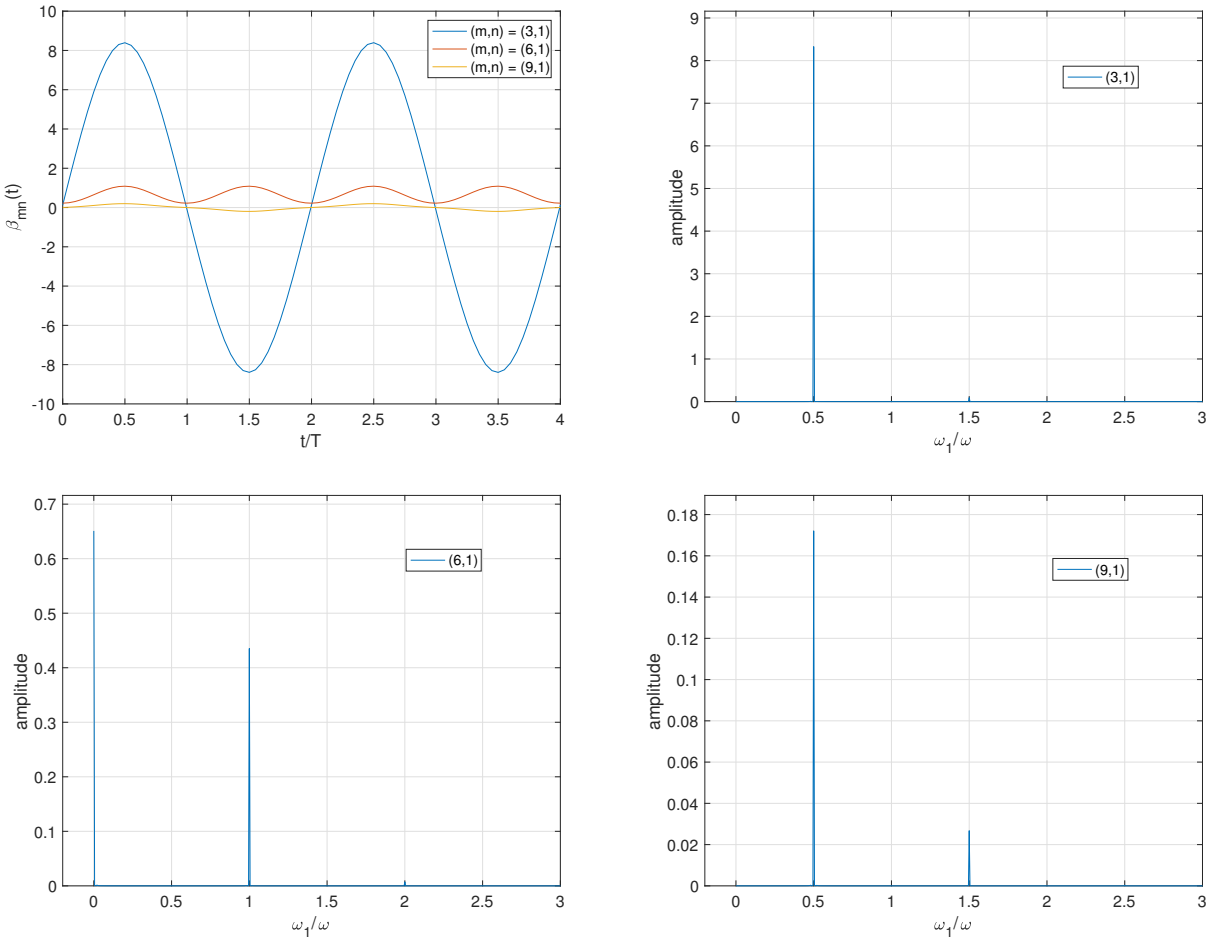

Figure 4.11: The time evolution of various modes with their spectral profiles. As expected, the response is sub-harmonic for $(3,1)$ and $(9,1)$ and harmonic for $(6,1)$.

Amplitude Modulation

Another consequence of nonlinearity is that the amplitudes of the constituent modes may be modulated on a slow time-scale. Thus, we can write $\beta_{mn}(t)=\gamma_{mn}(t)\tilde{\beta}_{mn}(t)$ where $\tilde{\beta}_{mn}$ is periodic (with frequency $\omega/2$ or ω, depending on the response) and $\gamma_{mn}(t)$ is the envelope. These modulations may be periodic, indicating that the envelopes underwent a Hopf bifurcation as the stability thresholds are crossed [12]. By moving appropriately through the parameter space, the period of the modulations may double before leading to chaos (see

Figure 1.2). However, an approximation up to the fourth-order, as we have been considering, is not rich enough to yield even limit cycles. In the presence of damping, all envelopes converge to stable fixed points [37].

In order to further validate our algorithm, we simulate case (ii) from Table 4.3 with $K = 3$ and $\epsilon = 0.1$. According to the bifurcation structure determined in [37], the zero fixed point for the mode $(0, 1)$ loses its stability at this point so its envelope should tend towards a non-zero value. On the other hand, the zero fixed point for $(3, 1)$ becomes stable so we would expect it to gradually vanish. Figure 4.12 shows that this is indeed the case. Another notable aspect is the small phase difference between the two curves. This was noted in [12] to be indicative of energy transfer between the modes. Finally, Figure 4.13 shows that the phase-plane trajectory for the envelopes of the competing modes indeed tends to a stable fixed point at approximately $(0.5672, 0)$.

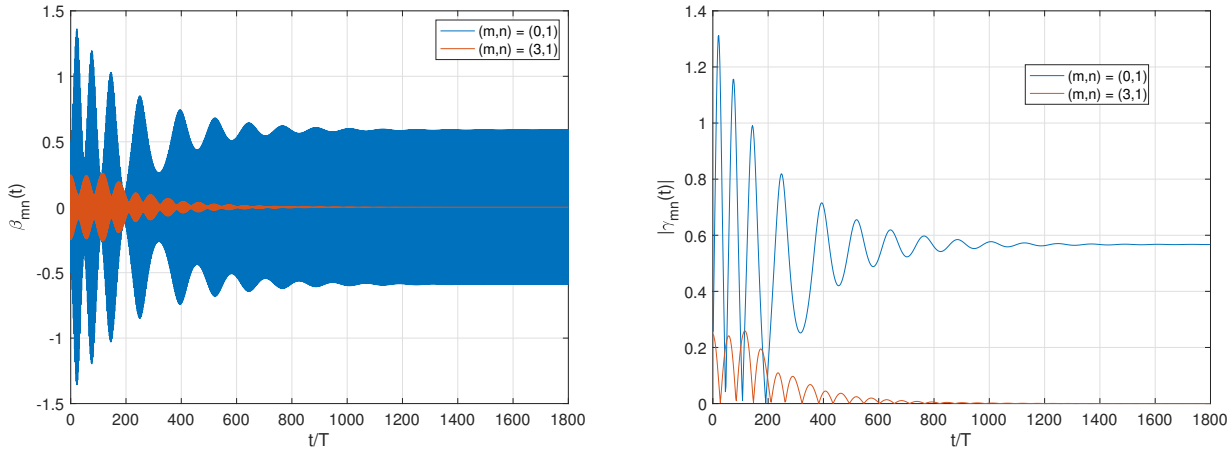

Figure 4.12: Evolution of modes $(0, 1)$ and $(3, 1)$ along with the amplitude envelopes $|\gamma_{mn}(t)|$. As predicted, the envelope for $(0, 1)$ converges to a non-zero fixed point while that for $(3, 1)$ converges to zero.

The ability of our algorithm to accurately reproduce these predictions from the fourth-order model suggests that it should be similarly effective for fully nonlinear simulations. It would be interesting to see if we are able to capture limit cycles, period-doubling and chaos for the envelopes as the parameters are adjusted. However, this must be preceded by a complete determination of the neutral stability curves for the nonlinear problem so we can navigate accurately through the parameter space.

4.6 Conclusion

We have developed a novel technique for simulating nonlinear Faraday waves in a circular cylinder. As opposed to [28], in which a low-order Navier-Stokes solver was developed, we

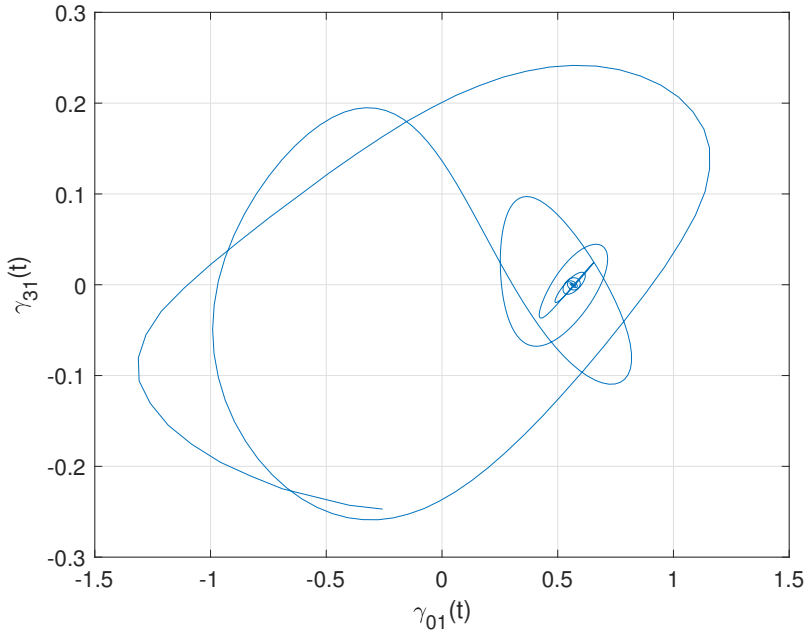

Figure 4.13: The phase-plane trajectory of the amplitude envelopes $\gamma_{mn}(t)$. Observe that it tends towards the stable fixed point $\approx (0.5672, 0)$.

opted for the viscous potential flow model of Dias et al. This model incorporates artificial dissipation into the ideal fluid equations. The damping occurs at the rate $2\nu a^2$ for each normal mode with wave-number a and, is in that sense, referred to as linear. This choice simplifies the problem in that Neumann conditions at the boundaries are sufficient for its solvability. We are, as a result, unable to study the effects of boundary layers on the subsequently generated waves, as proposed in [53]. Nevertheless, this choice of damping mechanism has been used successfully to explain many experimental observations [37, 30].

Among its advantages, this model admits a surface formulation in terms of the Dirichlet-to-Neumann and viscosity operators. Allied with the Transformed Field Expansion, this approach allows one to precisely control the extent of nonlinearity. As established in Sections 4.4 and 4.5, we can easily transition from linear to weakly nonlinear regimes by varying the highest expansion order. The high degree of agreement between theoretical predictions and our results over various regimes indicates that our algorithm is well-suited for identifying the particular mechanisms that trigger experimental observations.

Further challenges for this technique include studying fully nonlinear mode competition, analyzing multiple-frequency forcing regimes and investigating the formation of richer patterns on the surface. Determining the neutral stability curves at a particular order of nonlinearity is a necessary first step to reproducing the conditions for mode competition and chaos onset, as observed by [12]. By necessity, the region of the parameter space that leads to these observations is close to several bifurcation boundaries and, consequently, the results

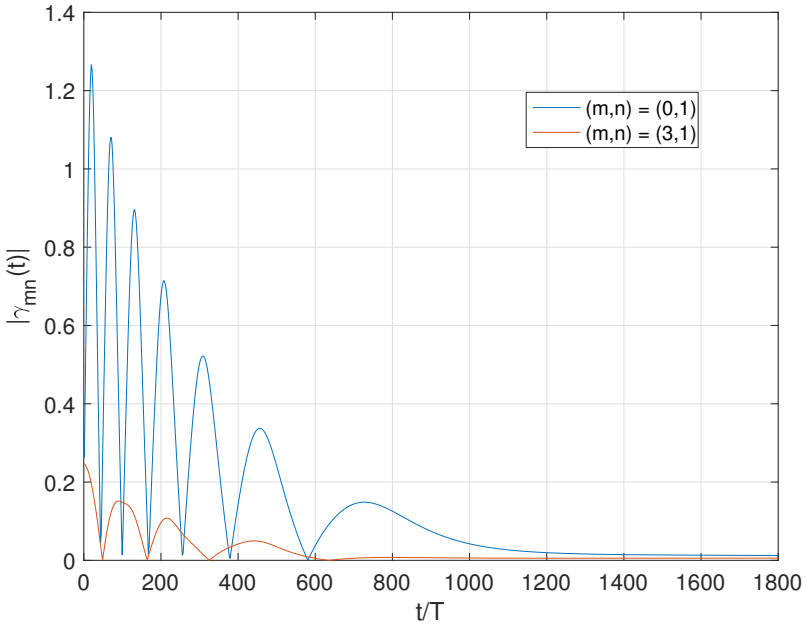

Figure 4.14: The amplitude envelopes for both $(0,1)$ and $(3,1)$ decay with $(\omega/2\pi, \alpha/g) = (19.5971, 0.1056)$. This demonstrates the sensitivity of the results to small changes in parameter values close to the critical thresholds.

are extremely sensitive to small changes in parameter values. Figure 4.14 displays the evolution of the amplitude envelopes with the parameters chosen at the point of intersection of the stability curves from the linearized problem. The decay for both envelopes, as opposed to one envelope converging to a non-zero value in Figure 4.12, illustrates the hazard in using even slightly incorrect parameter values.

The linear theory for multiple-frequency forcing has been developed in a similar manner to the single frequency case [7]. Such regimes serve as logical extensions of the cases that have been studied in Section 4.4. In addition, two forcing frequencies have been noted to generate elaborate patterns, termed super-lattices, in rectangular domains [18]. Applying this idea to a cylinder and testing it numerically would be an interesting application of our algorithm. Another striking illustration would be the numerical reproduction of unconventional patterns, such as pentagons and stars, that have been reported for a cylinder [50].

Bibliography

[1] MJ Ablowitz, AS Fokas, and ZH Musslimani. "On a new non-local formulation of water waves". In: *Journal of Fluid Mechanics* 562 (2006), pp. 313–343.

[2] Milton Abramowitz and Irene A Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*. Vol. 55. Courier Corporation, 1964.

[3] HN Abramson, FT Dodge, and D Kana. "Liquid surface oscillations in longitudinally excited rigid cylindrical containers". In: *Aiaa Journal* 3.4 (1965), pp. 685–695.

[4] David M Ambrose, Jerry L Bona, and David P Nicholls. "Well-posedness of a model for water waves with viscosity". In: *Discrete Contin. Dyn. Syst. B* 17 (2012), pp. 1113–1137.

[5] T Brooke Benjamin and F Ursell. "The stability of the plane free surface of a liquid in vertical periodic motion". In: *Proc. R. Soc. Lond. A* 225.1163 (1954), pp. 505–515.

[6] Christine Bernardi and Yvon Maday. "Spectral methods". In: *Handbook of numerical analysis* 5 (1997), pp. 209–485.

[7] Thomas Besson, W Stuart Edwards, and Laurette S Tuckerman. "Two-frequency parametric excitation of surface waves". In: *Physical Review E* 54.1 (1996), p. 507.

[8] Oscar P Bruno and Fernando Reitich. "Solution of a boundary value problem for the Helmholtz equation via variation of the boundary into the complex domain". In: *Proceedings of the Royal Society of Edinburgh Section A: Mathematics* 122.3-4 (1992), pp. 317–340.

[9] C. Canuto et al. *Spectral Methods: Fundamentals in Single Domains*. Scientific Computation. Springer-Verlag Berlin, 2006. ISBN: 9783540307266.

[10] Peilong Chen. "Nonlinear wave dynamics in Faraday instabilities". In: *Physical Review E* 65.3 (2002), p. 036308.

[11] S Ciliberto and Jerry P Gollub. "Phenomenological model of chaotic mode competition in surface waves". In: *Il Nuovo Cimento D* 6.4 (1985), pp. 309–316.

[12] S Ciliberto and JP Gollub. "Chaotic mode competition in parametrically forced surface waves". In: *Journal of Fluid Mechanics* 158 (1985), pp. 381–398.

[13] Walter Craig and Catherine Sulem. "Numerical simulation of gravity waves". In: *Journal of Computational Physics* 108.1 (1993), pp. 73–83.

[14] John David Crawford, Edgar Knobloch, and Hermann Riecke. "Period-doubling mode interactions with circular symmetry". In: *Physica D: Nonlinear Phenomena* 44.3 (1990), pp. 340–396.

[15] SP Das and EJ Hopfinger. "Parametrically forced gravity waves in a circular cylinder and finite-time singularity". In: *Journal of Fluid Mechanics* 599 (2008), pp. 205–228.

[16] F Dias, AI Dyachenko, and VE Zakharov. "Theory of weakly damped free-surface flows: a new formulation based on potential flow solutions". In: *Physics Letters A* 372.8 (2008), pp. 1297–1302.

[17] Alok Dutt, Leslie Greengard, and Vladimir Rokhlin. "Spectral deferred correction methods for ordinary differential equations". In: *BIT Numerical Mathematics* 40.2 (2000), pp. 241–266.

[18] W Stuart Edwards and Stéphane Fauve. "Parametrically excited quasicrystalline surface waves". In: *Physical Review E* 47.2 (1993), R788.

[19] L.C. Evans. *Partial Differential Equations*. Graduate studies in mathematics. American Mathematical Society, 2010. ISBN: 9780821849743.

[20] M Faraday. "M. Faraday, Philos. Trans. R. Soc. London 121, 299 (1831)." In: *Philos. Trans. R. Soc. London* 121 (1831), p. 299.

[21] JP Gollub and Christopher W Meyer. "Symmetry-breaking instabilities on a fluid surface". In: *Physica D: Nonlinear Phenomena* 6.3 (1983), pp. 337–346.

[22] Anders C Hansen and John Strain. "On the order of deferred correction". In: *Applied numerical mathematics* 61.8 (2011), pp. 961–973.

[23] Anders Christian Hansen and John Strain. "Convergence theory for spectral deferred correction". In: *University of California at Berkeley* (2005).

[24] Benjamin Harrop-Griffiths, Mihaela Ifrim, and Daniel Tataru. "Finite depth gravity water waves in holomorphic coordinates". In: *Annals of PDE* 3.1 (2017), p. 4.

[25] Diane M Henderson and John W Miles. "Faraday waves in 2: 1 internal resonance". In: *Journal of fluid mechanics* 222 (1991), pp. 449–470.

[26] Jingfang Huang, Jun Jia, and Michael Minion. "Accelerating the convergence of spectral deferred correction methods". In: *Journal of Computational Physics* 214.2 (2006), pp. 633–656.

[27] Daniel D Joseph and Terrence Y Liao. "Potential flows of viscous and viscoelastic fluids". In: *Journal of Fluid Mechanics* 265 (1994), pp. 1–23.

[28] L Kahouadji et al. "Numerical simulation of supersquare patterns in Faraday waves". In: *Journal of Fluid Mechanics* 772 (2015).

[29] Maria Kakleas and David P Nicholls. "Numerical simulation of a weakly nonlinear model for water waves with viscosity". In: *Journal of Scientific Computing* 42.2 (2010), pp. 274–290.

[30] T Kambe and M Umeki. "Nonlinear dynamics of two-mode interactions in parametric excitation of surface waves". In: *Journal of Fluid Mechanics* 212 (1990), pp. 373–393.

[31] D Krishnaraja and SP Das. "Numerical simulation of parametrically forced gravity waves in a circular cylindrical container". In: *Journal of Physics: Conference Series*. Vol. 822. 1. IOP Publishing. 2017, p. 012073.

[32] Krishna Kumar and Laurette S Tuckerman. "Parametric instability of the interface between two fluids". In: *Journal of Fluid Mechanics* 279 (1994), pp. 49–68.

[33] Ludwig Matthiessen. "Akustische Versuche, die kleinsten Transversalwellen der Flüssigkeiten betreffend". In: *Annalen der Physik* 210.5 (1868), pp. 107–117.

[34] Ludwig Matthiessen. "Ueber die Transversalschwingungen tönender tropfbarer und elastischer Flüssigkeiten". In: *Annalen der Physik* 217.11 (1870), pp. 375–393.

[35] Ehud Meron and Itamar Procaccia. "Low-dimensional chaos in surface waves: Theoretical analysis of an experiment". In: *Physical Review A* 34.4 (1986), p. 3221.

[36] Ehud Meron and Itamar Procaccia. "Theory of chaos in surface waves: The reduction from hydrodynamics to few-dimensional dynamics". In: *Physical review letters* 56.13 (1986), p. 1323.

[37] John W Miles. "Nonlinear Faraday resonance". In: *Journal of Fluid Mechanics* 146 (1984), pp. 285–302.

[38] John W Miles. "Surface-wave damping in closed basins". In: *Proc. R. Soc. Lond. A* 297.1451 (1967), pp. 459–475.

[39] John Miles and Diane Henderson. "Parametrically forced surface waves". In: *Annual Review of Fluid Mechanics* 22.1 (1990), pp. 143–165.

[40] Paul A Milewski et al. "Faraday pilot-wave dynamics: modelling and computation". In: *Journal of Fluid Mechanics* 778 (2015), pp. 361–388.

[41] Michael L Minion et al. "Semi-implicit spectral deferred correction methods for ordinary differential equations". In: *Communications in Mathematical Sciences* 1.3 (2003), pp. 471–500.

[42] André Nachbin, Paul A Milewski, and John WM Bush. "Tunneling with a hydrodynamic pilot-wave model". In: *Physical Review Fluids* 2.3 (2017), p. 034801.

[43] David P Nicholls. "High-Order Perturbation of Surfaces Short Course: Boundary Value Problems". In: *Lectures on the Theory of Water Waves* 426 (2016), p. 1.

[44] David P Nicholls and Fernando Reitich. "A new approach to analyticity of Dirichlet-Neumann operators". In: *Proceedings of the Royal Society of Edinburgh Section A: Mathematics* 131.6 (2001), pp. 1411–1433.

[45] David P Nicholls and Fernando Reitich. "Stable, high-order computation of traveling water waves in three dimensions". In: *European Journal of Mechanics-B/Fluids* 25.4 (2006), pp. 406–424.

[46] David P Nicholls and Jie Shen. "A rigorous numerical analysis of the transformed field expansion method". In: *SIAM Journal on Numerical Analysis* 47.4 (2009), pp. 2708–2734.

[47] JR Ockendon and H Ockendon. "Resonant surface waves". In: *Journal of Fluid Mechanics* 59.2 (1973), pp. 397–413.

[48] Nicholas L O'Connor. "The complex spatiotemporal dynamics of a shallow fluid layer". PhD thesis. Virginia Tech, 2008.

[49] Nicolas Perinet, Damir Juric, and Laurette S Tuckerman. "Numerical simulation of Faraday waves". In: *Journal of Fluid Mechanics* 635 (2009), pp. 1–26.

[50] Jean Rajchenbach, Didier Clamond, and Alphonse Leroux. "Observation of star-shaped surface gravity waves". In: *Physical review letters* 110.9 (2013), p. 094502.

[51] Gabor Szegö. *Orthogonal polynomials*. Vol. 23. American Mathematical Soc., 1939.

[52] S Ubal, MD Giavedoni, and FA Saita. "A numerical analysis of the influence of the liquid depth on two-dimensional Faraday waves". In: *Physics of Fluids* 15.10 (2003), pp. 3099–3113.

[53] José M Vega, Edgar Knobloch, and Carlos Martel. "Nearly inviscid Faraday waves in annular containers of moderately large aspect ratio". In: *Physica D: Nonlinear Phenomena* 154.3-4 (2001), pp. 313–336.

[54] Jon Wilkening and Vishal Vasan. "Comparison of five methods of computing the Dirichlet-Neumann operator for the water wave problem". In: *Contemp. Math* 635 (2015), pp. 175–210.

[55] Guangyu Wu, Yuming Liu, and Dick KP Yue. "A note on stabilizing the Benjamin–Feir instability". In: *Journal of Fluid Mechanics* 556 (2006), pp. 45–54.

[56] Jian Yong-jun, E Xue-quan, and Bai Wei. "Nonlinear Faraday waves in a parametrically excited circular cylindrical container". In: *Applied Mathematics and Mechanics* 24.10 (2003), pp. 1194–1207.

[57] Vladimir E Zakharov. "Stability of periodic waves of finite amplitude on the surface of a deep fluid". In: *Journal of Applied Mechanics and Technical Physics* 9.2 (1968), pp. 190–194.