

A NOVEL NUMERICAL SOLVER FOR NONLINEAR  
BOUNDARY VALUE PROBLEMS, WITH  
APPLICATIONS TO THE FORCED GARDNER  
EQUATION

Andrew C. Cullen

A thesis submitted for the degree of Doctor of Philosophy at  
Monash University

School of Mathematical Sciences  
June, 2018

©Andrew C. Cullen (2018). Except as provisioned by the Copyright Act of 1968, this thesis may not be reproduced in any form without the written permission of the author.

I certify that I have made all reasonable efforts to secure copyright permissions for third-party content included in this thesis, and have not knowingly added copyrighted content to my work without the owner's permission.

## List of publications


The following is a list of published manuscripts resulting from this thesis:

1. Cullen, A.C. and Clarke, S.R. 2017. A Fast and Spectrally Accurate Solver for the Falkner–Skan Equation. *ANZIAM Journal* 58:57-68.

The ideas and algorithms contained within this paper were developed and written up by myself, the candidate, under the supervision of Dr Simon Clarke, who advised upon the progression of the manuscript, and assisted with the editing process. As such, the division of labour was 80% to the candidate, and 20% to the supervisor.

## Declaration

This thesis contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

**Student signature:** 

**Student name:** Andrew Cullen

**Date:** 27-8-18

**Supervisor signature:** 

**Supervisor name:** Simon Clarke

**Date:** 27-8-18

# Abstract

Novel numerical methods are proposed in order to solve variable coefficient nonlinear boundary value problems defined across one dimensional domains. Solutions for these equations are constructed by decomposing nonlinear equations into sequences of linear differential equations through the Homotopy Analysis Method (HAM). These linear equations are then solved sequentially in terms of Gegenbauer polynomials, which are a larger class of polynomials that includes the Chebyshev polynomials as a subset. A robust suite of tools for exploring solutions are then developed around this technique, including several approaches for numerical continuation. In composite, these techniques will be referred to the Gegenbauer Homotopy Analysis Method (GHAM).

As the numerical system that result from implementing GHAM can be described in terms of sparse matrix operators that are invariant with respect to the iteration number, this technique exhibits  $\mathcal{O}(n^{1.305} + mn^{1.05})$  scaling with time, where  $n$  is the number of points in Gegenbauer space required to resolve the solution, and  $m$  is the number of iterations. The numerical properties of this technique, including its computational cost and convergence properties are presented and contrasted with other numerical schemes evaluated within real, Chebychev and Gegenbauer spaces.

These tools are applied to several problems from fluid mechanics, with a specific focus on equations relating to the evolution of atmospheric Rossby waves, which can be treated as steady dispersive waves on the surface of a channel. Mathematically the dynamics of these wave trains can be described by the forced Korteweg de Vries (fKdV) and the forced Gardner equation subject to both local and compact topographic forcing, and it is these equations that will serve as the primary application of the numerical techniques presented within this thesis.



# Acknowledgements

While a PhD may appear to be an individual pursuit, this work would not exist without the generous support of some key people. First and foremost amongst those to whom I owe thanks are my family and friends, especially my mother Robyn, sister Alison, and close friend Peta. Collectively these three have exhibited unwavering support throughout the highs and lows that accompany the pursuit of something like this, and for that I will always be grateful.

I also owe particular thanks to my supervisor Simon Clarke, and more broadly the Monash Mathematics community. Simon and I have been working together since my honours year, and it is credit to his stewardship that I have been able to complete a research project with such breadth. Over the years he has been a constant source of insight and support, even while seeing yet another one of his PhD students wander off from an original motivating problem into the (mathematical) unknown.

Finally to Dave, Laura, Steph and Cass, with whom I shared an office at varying times across the years. Some might call pursuing a PhD to be a mad task, but it was an absolute pleasure each day to come into our little bubble of sanity.

# Contents

Abstract . . . . .	i
Acknowledgements . . . . .	ii
<b>1 Introduction</b>	<b>9</b>
1.1 Properties of the Gardner equation . . . . .	18
1.2 Thesis outline . . . . .	21
<b>2 Numerical Methods</b>	<b>23</b>
2.1 Spectral Methods . . . . .	24
2.1.1 Chebyshev polynomials . . . . .	24
2.1.2 Chebyshev Quadrature Points . . . . .	25
2.1.3 Spectral convergence of Chebyshev approximations . . . . .	26
2.1.4 Differentiation in Chebyshev space . . . . .	26
2.1.5 Integration in Chebyshev space . . . . .	28
2.1.6 Multiplication in Chebyshev space . . . . .	29
2.1.7 Gegenbauer polynomials . . . . .	30
2.2 Spectral Methods for differential equations . . . . .	32
2.2.1 Chebyshev collocation matrices for BVPs . . . . .	32
2.2.2 Gegenbauer polynomials for BVPs . . . . .	33
2.3 Solving nonlinear differential equations . . . . .	42
2.4 Numerical Continuation . . . . .	46
2.5 Domain mappings . . . . .	51
2.6 Discussion . . . . .	53
<b>3 Homotopy Analysis Method</b>	<b>56</b>
3.1 Convergence properties . . . . .	58
3.2 Spectral Homotopy Analysis Method . . . . .	61

3.3	Gegenbauer Homotopy Analysis Method . . . . .	63
3.3.1	Evaluating $R_m$ . . . . .	65
3.3.2	Pseudo-code algorithm for evaluating problems with the GHAM . . .	67
3.3.3	Solving the Falkner–Skan equation . . . . .	68
3.4	Solving problems with multiple solutions through GHAM . . . . .	71
3.4.1	Homotopy based hypersphere continuation . . . . .	75
3.4.2	Homotopic integrated arc-length continuation . . . . .	78
3.5	Computational cost of the GHAM . . . . .	84
3.5.1	Two-dimensional viscous flow in a rectangular domain with porous, moving boundaries . . . . .	91
3.6	A priori estimation of $\hbar$ . . . . .	102
3.6.1	Validation . . . . .	105
3.7	Invariants with respect to $\hbar$ . . . . .	108
3.8	Discussion . . . . .	110
<b>4</b>	<b>Weakly nonlinear waves: solutions of the forced Korteweg–de Vries Equa- tion</b>	<b>113</b>
4.1	Topographic forcing with local support . . . . .	116
4.2	Topographic forcing with compact support . . . . .	122
4.3	Discussion . . . . .	130
<b>5</b>	<b>Asymmetric steady solutions of the forced Korteweg–de Vries equation</b>	<b>133</b>
5.1	Solution space for hydraulic solutions of the fKdV equation . . . . .	141
5.2	Constrained drag minimisation . . . . .	146
5.3	Discussion . . . . .	148
<b>6</b>	<b>Symmetric solutions of the forced Gardner equation</b>	<b>151</b>
6.1	Unforced solutions . . . . .	152
6.2	Positive cubic nonlinearity . . . . .	156
6.3	Negative cubic nonlinearity . . . . .	164
6.4	Discussion . . . . .	170
<b>7</b>	<b>Conclusion</b>	<b>174</b>
	<b>Bibliography</b>	<b>177</b>

# List of Figures

1.1	Schematic of channel flow subject to either a topographic disturbance. . . .	14
2.1	Numerical solution of equation (2.40) using the Gegenbauer method for $\epsilon = 10^{-6}$ , as well as its convergence behaviour with $n$ , where $m = \lceil 1.01n \rceil$ . . . .	42
2.2	Sparsity structure of the second-order linear differential Gegenbauer space operator for equation (2.40), where the non-zero elements are coloured. . . .	43
3.1	Normal flow solutions to the Falkner-Skan equation, and their first and second derivatives, for a) $\beta = 1$ and b) $\beta = -0.1$ . . . . .	69
3.2	The relationship between $\beta$ and $f''(0)$ . In blue are the solutions calculated using the Homotopy based method, and in red are previously calculated solutions by Cebeci and Keller (1971) . . . . .	70
3.3	Error, evaluated as the $L_2$ norm of the residual after 4 iterations, evaluated against $\hbar$ . . . . .	71
3.4	Convergence of the residual at $\beta = 1$ and at the optimal $\hbar$ . The $L_2$ and $L_\infty$ norm of the residual in blue and red respectively. . . . .	72
3.5	Relationship between $\gamma$ and $\phi$ for equation (3.32), calculated both numerically and through the implicit relationship equation (3.36). . . . .	74
3.6	Evaluating the boundary derivative $\frac{du(0)}{dx}$ of solutions of equation (3.37) for $\phi = 0.6$ , and attempting to find solutions where $\frac{du(0)}{dx} = 0$ . . . . .	75
3.7	Analytically determined parameter space (blue), starting point (black) and end point of continuation (magenta), with the corresponding solutions for equation (3.32), calculated using Homotopic integrated arc-length continuation with a step length of 0.33. The continuation path is shown by the dashed red line. . . . .	84

3.8 Parameter space for equation (3.32) in blue, with the positions after using Homotopic integrated arc-length continuation for a step length of 0.13 indicated by red squares (for traversing down the branch) and diamonds (for traversing up the branch). The starting point is given by the orange asterisk. The corresponding solutions of the solutions constructed with the continuation method are included within Figure 3.8b. . . . . 85

3.9 Sparsity of a fourth-order variable coefficient boundary value problem, discretised over 512 points using the GHAM. . . . . 86

3.10  $L$  and  $U$  matrices from a LU decomposition of a  $2^9 \times 2^9$  grid, corresponding to solutions of a nonlinear, variable coefficient viscous fluids problem, subject to the linear operator  $\mathcal{L}_4$  from equation (3.66). . . . . 88

3.11  $L$  and  $U$  matrices from a LUPQR decomposition of a  $2^9 \times 2^9$  grid for the same problem examined in Figure 3.10. The  $P$ ,  $Q$  and  $R$  matrices all correspond to matrices with  $2^9$  nonzero elements located entirely upon the main diagonal. 89

3.12 Solution and error of equation (3.65) calculated at  $\alpha = 1$  and  $R_e = 10$  using  $\mathcal{L}_4$  from equation (3.66). . . . . 92

3.13 Error against  $\hbar$  for solutions of equation (3.65), as calculated using the SHAM (solid lines) and the GHAM (dotted lines). The blue lines correspond to  $\mathcal{L}_1$ , the red to  $\mathcal{L}_2$ , the green to  $\mathcal{L}_3$  and the magenta line corresponds to  $\mathcal{L}_4$ , with all operations truncated after 25 iterations. . . . . 93

3.14 Calculated error at  $\hbar_{\text{opt}}$  for equation (3.65) using the GHAM, as a function of the number of iterations and spatial resolution  $n$ . For these calculations  $H_a = 1$  and  $R_e = 10$ . . . . . 95

3.15 Calculated error at  $\hbar_{\text{opt}}$  for equation (3.65) using the SHAM, as a function of the number of iterations and spatial resolution  $n$ . For these calculations  $H_a = 1$  and  $R_e = 10$ . Note the changed vertical scale, as compared to Figure 3.14. 96

3.16 Impact of the number of iterative steps upon the numerical error when solving equation (3.65) for varying choices of the auxiliary linear operator  $\mathcal{L}$ . Solid lines correspond to the SHAM, dashed denote the GHAM, and the blue dots are Newton Iteration upon a Gegenbauer discretisation. Of the solid lines, blue represents  $\mathcal{L}_1$ , Red is  $\mathcal{L}_2$ , Green is  $\mathcal{L}_3$  and Magenta is  $\mathcal{L}_4$ . All solutions were calculated at the optimal  $\hbar$  for each choice of  $\mathcal{L}$ . . . . . 99

3.17	Computational cost and the error of solving equation (3.65) for varying numbers of iterations, following Figure 3.16 with the inclusion of the computational cost and error using MATLAB’s ‘BVP4C’ routine, as represented by the red circle. . . . .	100
3.18	Scaling of computational time for solutions of equation (3.65) calculated by taking 75 iterations of the SHAM (solid lines) and the GHAM (dotted lines). Results for the GHAM are presented for $n \in [2^6, 2^{14}]$ , whereas the SHAM was only calculated over $n \in [2^6, 2^9]$ , as the Chebyshev collocation matrices within the SHAM become singular after this point. . . . .	100
3.19	Scaling coefficients for the GHAM, assuming that the computational cost scales as $T = CI^S$ . One again, blue, red, green and magenta represent $L_1$ , $L_2$ , $L_3$ and $L_4$ respectively. . . . .	101
3.20	Scaling of computational cost of constructing solutions for equation (3.65) with $n$ . Blue: set up cost for establishing the matrix problem in the context of the GHAM. Red, dashed: LUPQR decomposition, which only needs to occur once at the beginning of the iterative process. Yellow: solving the matrix system using the LUPQR decomposition. Green: transform between real and Chebyshev space. Light blue: calculating derivatives up to fourth-order. Purple, dotted: solving the matrix system using a direct matrix inverse, which is not used within the algorithm. . . . .	103
3.21	Figure 3.21a shows $r$ against $\hbar$ for the $\mathcal{L}_1$ operator of Subsection 3.5.1 in blue, with the red line indicating the demarcation point for the region where $r > 1$ . Figure 3.21b shows the $L_1$ norm of error at $\{2, 10, 50, 100\}$ iterations over a grid of $n = 2^8$ points. The black dashed line indicates the location of $\hat{\hbar}$ , and the green dashed line is marks the point in $\hbar$ that corresponds to $r = 1.107$	
3.22	(a) shows $r$ against $\hbar$ for the $\mathcal{L}_4$ operator of Subsection 3.5.1 in blue, with the red line indicating the demarcation point for the region where $r > 1$ , and (b) shows the $L_1$ norm of error at $\{2, 10, 50, 100\}$ iterations over a grid of $n = 2^8$ points. The black dashed line indicates the location of $\hat{\hbar}$ . . . . .	108
4.1	Phase portrait of the Korteweg–de Vries equation for $\Delta < 0$ , with the +’s corresponding to the locations of the stationary points. . . . .	115
4.2	Parameter space for solutions of equation (4.11) with local support where $f(x) = \delta(x)$ . . . . .	119

4.3	Type I and Type II (in blue and red respectively in Figure (a)) solutions to the fKdV equation with local support when $\gamma = 0.5$ in part (a), with the corresponding locations in the phase space diagram indicated in red in part (b).	120
4.4	Type III, Type IV, and Type V (in blue, red, and yellow respectively in Figure (a)) solutions to the fKdV equation with local support when $\gamma = -0.5$ in part (a), with the corresponding locations in the phase space diagram indicated in red in part (b).	121
4.5	Parameter space for solutions of the fKdV in the form of equation (4.11), subject to the forcing stipulated by equation (4.16) for varying $L$ .	127
4.6	Type I, and Type II (in blue and red respectively in Figure (a)) solutions to the fKdV equation with compact support when $\gamma = 0.5$ and $L = 1$ .	129
4.7	Type III, Type IV, and Type V (in blue, red, and yellow respectively in Figure (a)) solutions to the fKdV equation with compact support when $\gamma = -0.5$ and $L = 1$ .	129
5.1	Plot of $\Delta$ against $\gamma$ , as produced by Ee and Clarke (2007) for the parameter space for asymmetric solutions of equation (5.1).	135
5.2	The locations of $\frac{dA}{dx} = 0$ at $x = 0$ for $\Delta = 0$ are presented in (a) for $-160 \leq \gamma \leq 0$ , and the corresponding locations in log scale in (b), with all solutions constructed based upon equation (5.1).	136
5.3	Figure (a) contains the a subset of the corresponding solitary wave solutions to Figure 5.2 for equation (5.1). The presented solutions correspond to the zeros of Figure 5.2a for $-10^3 \leq \gamma \leq 0$ . The $\gamma = -8$ solitary wave is shown in red, with increasing magnitudes of $\gamma$ corresponding to increased wave amplitudes. Figure (b) shows these same solutions rescaled so that their maximum amplitude is 1.	137
5.4	Parametric relationship between $\Delta$ and $\gamma$ for $-10 \leq \gamma \leq 5$ , subject to $f(x) = \text{sech}^2(x)$ .	142
5.5	Hydraulic, asymmetric solutions of equation (5.1) for $\gamma = \{-8.5, -7.5, -5, -2, 0.5, 5, 10, 25\}$ , ordered from left to right.	144
5.6	Solutions of equation (5.1) for $\gamma = \{0.5, 5, 10, 25\}$ in blue, red, yellow and purple respectively.	145
5.7	Solutions of equation (5.1) for $\gamma = \{-8.5, -7.5, -5, -2\}$ in blue, red, yellow and purple respectively.	145

5.8	$(\gamma, \Delta) = (-2.0285, -2.56)$ in blue and $(\gamma, \Delta) = (-4.1360, -2.56)$ in red. . . .	146
6.1	The form of the two solitary waves of the Gardner equation (in blue and red) and the trivial solution (yellow) for $\Delta = -1.92$ , $r = \frac{3}{2}$ and $q = \frac{3}{2}$ . . . . .	153
6.2	Phase portrait of the Gardner equation for $\Delta < 0$ and $r = q = 3/2$ , with the $c_+$ and $c_-$ denoting the location—in $A$ —of the critical points at $(A, A_x) = \left( \frac{-r}{2q} \pm \frac{\sqrt{r^2 - 4\Delta q}}{2q}, 0 \right)$ . . . . .	155
6.3	Solution space to equation (6.12) subject to the forcing equation (6.13) for varying $\delta$ . Dashed lines correspond to $\delta = \{0, 0.1, 0.2, 0.3\}$ for the blue, red, yellow and purple lines respectively. Solid lines corresponding to $\delta = \{0.4, 0.6, 0.8, 1\}$ for the same colour progression. . . . .	158
6.4	Two perspectives on the solution space to elucidate the structure of equation (6.12) for varying $\delta$ . Dashed lines correspond to $\delta = \{0, 0.1, 0.2, 0.3\}$ for the blue, red, yellow and purple lines respectively. Solid lines corresponding to $\delta = \{0.4, 0.6, 0.8, 1\}$ for the same colour progression. . . . .	159
6.5	Parameter space for the forced Gardner equation in the form equation (6.12) subject to the forcing function equation (6.13) for $\delta = 0.6$ , divided into different colours to cover regions of distinct solution phenomenology. . . . .	161
6.6	Type I (Blue, Red) and II (Yellow, Purple) solutions of equation (6.12) for $\delta = 0.6$ . These solutions correspond to the blue region of Figure 6.5, subject to the forcing function equation (6.13). . . . .	162
6.7	Type III and IV solutions of equation (6.12) for $\delta = 0.6$ . These solutions correspond to the red region of Figure 6.5, subject to the forcing function equation (6.13). . . . .	163
6.8	Type V and VI solutions of equation (6.12) for $\delta = 0.6$ . These solutions correspond to the yellow region of Figure 6.5, subject to the forcing function equation (6.13). . . . .	164
6.9	Type VII and VIII solutions of equation (6.12) for $\delta = 0.6$ . These solutions correspond to the purple region of Figure 6.5, subject to the forcing function equation (6.13). . . . .	165
6.10	Phase portrait of the Gardner equation for $\Delta < 0$ and $r = -q = 3/2$ outside the transition region, with a single critical point at $(A, A_x) = (0, 0)$ . . . . .	167



6.11 Parameter space for the forced Gardner equation (6.14) subject to the forcing function equation (6.13). Figure 6.11a presents solutions within the vicinity of the origin, and Figure 6.11b shows the evolution of a subset of those solutions. Of the solid lines, blue, red, yellow, purple and green are  $\delta = \{1, 0.7, 0.5, 0.3, 0.216095\}$  respectively. Of the dotted lines, yellow, red, blue, burgundy and light blue correspond to  $\delta = \{0.21609, 0.215, 0.214, 0.213, 0.1\}$ . Thus the dotted light blue solution is close to the fKdV equation, and the dark blue solid line corresponds to the mKdV equation where there is only a cubic nonlinearity. . . . . 168

6.12 A representative sample of the solution set for  $\delta = \{0.1, 0.213, 0.215, 0.21609\}$ , corresponding to Figure 6.11a. . . . . 171

6.13 A representative sample of the solution set for  $\delta = \{0.216095, 0.3, 0.5, 1.0\}$ , corresponding to Figure 6.11b. With the exception of the case where  $\delta = 0.216095$ , the blue, red, yellow, purple green and teal plots correspond to solutions at  $A(0) = \{1, 5, 10, -1, -5, -10\}$  respectively. Due to the unique taxonomy of solutions in the  $\delta = 0.216095$  a broader range of solutions for  $-10 \leq A(0) \leq 15$  has been presented. . . . . 172

# Chapter 1

## Introduction

Rossby waves are a large scale tropopause level disturbance to prevailing atmospheric flows that are closely linked with the cyclic interplay of high and low pressure systems (Liebmann and Hendon, 1990), as well as driving in the dynamics of the Gulf Stream currents (Osychny and Cornillon, 2004); zonal winds (Malguzzi and Malanotte-Rizzoli, 1984); ENSO linked climate variability (Ryoo et al., 2013) and broader equatorial waves (Pedlosky, 1979, Boyd, 1980a, Cushman-Roisin and Beckers, 2011). Understanding the factors that influence these waves is a problem of particular importance within geophysical fluid dynamics, and as such this work will focus upon developing the numerical tools necessary to study these topographic effects.

In the most general sense, Rossby waves are seen in fluids upon a  $\beta$ -plane, where the differential rotation rates give rise to shear, which then in turn drives the dynamics of the flow. In the atmosphere, this is seen across the jet stream (Nakamura and Plumb, 1994), at the interface between the stratospheric polar vortices and the sub tropical atmosphere. Along this interface, repeating wave structures are created by the deflection of fluid particles by the Coriolis force, creating the longitudinal advection of potential vorticity.

The dynamics of these planetary scale Rossby waves are not simply the product of shear—they can also be excited in the troposphere, primarily through the effect of topography. In fact, the highest frequency of breaking events is observed over Siberia and Baffin Island (Woollings and Hoskins, 2008)—with both regions being characterised by sharp increases in the gradient of the topography. Other sources of excitation include the latent release of heat, the nonlinear evolution of tropospheric eddies (Scinocca and Haynes, 1998) and

even the evolution of the hole in the ozone layer in the southern hemisphere (Ndarana et al., 2012). As a product of this excitation, the amplitudes can increase until a critical condition is reached, which, in the atmosphere, is generally characterized by an irreversible overturning process which deforms the material contours of the Rossby wave (McIntyre and Palmer, 1983b).

As these waves break a *surf zone* is induced, within which high vorticity air mixes with the ambient air (Jukes and McIntyre, 1987). Vorticity in the surf zone can then roll up into coherent small scale vortices of high-vorticity air, which is then advected away from the region contained within the Rossby wave. This in turn induces trace-gas mixing (Pelly and Hoskins, 2003, Ndarana and Waugh, 2010) and creates cut off lows and blocking highs (Rex, 1950a,b). These are stationary high pressure cells exhibiting significant persistence (Austin, 1980), which is the phenomenological distinction from standard travelling troughs (Lejenas and Okland, 1983). The resulting effects of the *surf zone* can broadly be considered as a bifurcation in the oncoming flow, as the oncoming air currents are deflected around the obstruction. Generally, wave breaking events occur in the equatorial direction, however, in rare cases breaking events have been linked to intrusions of mid-latitude air into the stratospheric polar vortex (Plumb et al., 1994). This asymmetry in the dynamics of breaking events can also be seen in the behaviour of upper tropospheric waves (Wang and Magnusdottir, 2011).

As a consequence of their scale, and their impact on the surrounding atmosphere, a single Rossby wave breaking event can play a significant role in the climate dynamics of a large geographic region (McIntyre and Palmer, 1983b,a, Jukes and McIntyre, 1987), acting as a driver for extreme weather events (Petoukhov et al., 2013). A dramatic example of this was seen in the Pakistani floods and Russian wildfires of 2010, where both events were linked to a single Rossby wave breaking event (Lau and Kim, 2011). In Australia, extreme heat waves can be attributed to overturning Rossby waves (Parker et al., 2014), in which upper level anticyclonic potential vorticity anomalies trap heat, and act as the defining feature of heat waves in the region. While these heat waves have health and agricultural implications on a population level, the more notable risk is of severe fire conditions. The end of heat waves are heralded by the passage of an extreme cold front, and in an Australian context it has been shown that there is a strong link between these cold fronts and high fire risk (Bureau of Meteorology, 1984, Reeder and Smith, 1987, Engel et al., 2013, Parker,

2012, Reeder et al., 2015), with Petoukhov et al. (2013) speculating that the trapping of free waves within the mid-latitude waveguides may well drive these dynamics. This phenomenology has been linked to both the deadly Ash Wednesday and Black Saturday bushfires that swept through south-eastern Australia in 1983 and 2009 respectively. As such, understanding both the formation, evolution and eventual disruption of Rossby wave breaking events has significant implications not just from a weather prediction standpoint, but also when considering community health and well-being.

The study of these dynamics has occurred from both a larger scale meteorological perspective, and one which focusses more on the mathematical physics inherent in these dynamics. This distinction is entirely an artifice, as both approaches are complimentary, but it allows for distinguishing between the different approaches to the overall problem. The meteorological approach has involved authors studying historical and observational data in the context of large scale numerical simulations of climate models; whereas the mathematical approach has generally been to understand the inherent dynamics by considering the key equations, and often reducing the dimensionality of the problem in an attempt to develop greater insights to the overall physics. It is this latter approach that will be focussed upon from this point.

From a mathematical perspective, the study of wave dynamics can be divided further into two main fronts. The first is known as the weakly-nonlinear approach, where perturbations are limited to being only just large enough to give rise to nonlinear effects, which in this context has been approached by Jiong and Liu (1998), Xun et al. (2000), Jiang and Lian-Gui (2009) among others. The second is the finite-amplitude approach, and was pioneered by Long (1953b,a, 1954, 1955, 1959, 1962, 1965, 1970), before being further advanced by Benney (1979). This thesis will be primarily concerned with considering the construction of solutions to weakly-nonlinear models for wave behaviour.

The study of waves has a long history, with a wealth of research that dates back to Isaac Newton and G.G. Stokes in the 17th and 18th centuries. This early work was primarily focussed upon water waves, although the observed dynamics have been shown to have broader implications. Of this work, Stokes was the first to develop a definitive theory of both linear and weakly-nonlinear waves in deep water, as documented by Craik (2004, 2005). While Stokes' theories only applied to periodic, deep-water waves, observations by

Russell (1844) showed that these waves can also exist in shallow water. Constructing a mathematical argument to describe Russell's observations motivated Boussinesq and then later Lord Rayleigh (1876), who examined a steady stationary wave vanishing at infinity, and was able to describe the spatial evolution of the wave through the equation

$$\left(\frac{\partial\eta}{\partial x}\right)^2 + \frac{3}{H^2}\eta^2(\eta - \eta_0) = 0, \quad \eta = \eta(x) \quad (1.1)$$

where  $x$  is the spatial position of a fluid particle,  $\eta$  is the free surface displacement,  $H$  is the height of the water at its equilibrium state and  $\eta_0$  describes the amplitude of the wave. A sketch of these conditions can be seen in Figure 1.1. This equation has an explicit, analytic solution of the form

$$\eta(x) = \eta_0 \operatorname{sech}^2\left(\sqrt{\frac{3\eta_0}{4H^3}}x\right). \quad (1.2)$$

At the end of the 19th century, Korteweg and de Vries (1895) constructed a proof of existence for translational waves that matched the observations of Russell, and resulted in the equation

$$\frac{\partial\eta}{\partial t} = \frac{3}{2}\sqrt{gl}\frac{\partial}{\partial x}\left(\frac{1}{2}\eta^2 + \frac{3}{2}\alpha\eta + \frac{1}{3}\beta\frac{\partial^2\eta}{\partial x^2}\right), \quad (1.3)$$

where  $\eta(x, t)$  is the surface elevation above the equilibrium height,  $l$  is an arbitrary constant related to the fluid velocity,  $g$  is the gravitational constant,  $\alpha$  is a small constant based upon the rate of uniform motion of the fluid and

$$\beta = \frac{1}{3}l^3 - \frac{Tl}{\rho g}$$

is a dimensional parameter, dependent on the surface capillary tension  $T$  and liquid density  $\rho$ . Changing variables to dimensionless distance and velocity, and applying a Galilean transformation results in the canonical, dimensionless form of the Korteweg–de Vries (KdV) equation

$$\eta_t + 6\eta\eta_x + \eta_{xxx} = 0. \quad (1.4)$$

This is a nonlinear equation in terms of a single spatial and temporal variable, that takes a very similar form to the work of Boussinesq (1872) (as translated by Vastano and Mungall (1976)). The initial development of the KdV equation stemmed from the study of the propagation of small amplitude, unidirectional, long wavelength waves in a channel with

a flat bottom and free surface, and balances the effects of weak nonlinearity and weak dispersion. One of the first formalisms of the KdV equation in a modern geophysical flows setting stemmed from Benney (1966), who reached a form of the KdV equation from a first-order expansion in amplitude and dispersion to study long nonlinear waves. This approach has led to higher order extensions by Grimshaw (1997), and Helfrich (2007), the latter of which modified the equations to allow for the presence of strong nonlinearities in the case of weak dispersion. It can also be used to model two-layer flow potential flows, where the flow is two-dimensional, incompressible and irrotational with a uniform density (Baines, 1987, Batchelor, 2000). As such, it can be used to model free-surface waves, marked by the bounding line between two distinct layers of water and air, and can be modified to account for perturbations from both pressure and the channel topography.

These geophysical wave problems are not the limit of applicability for the KdV equation. Beyond problems from geophysical fluid dynamics the KdV equation has also been found to have applications in a number of other fields. Plasma physics is a particularly prominent example of this, where the Gardner–Morikawa transformation (Gardner and Morikawa, 1960) maps the physics of solitary wave propagation in plasma to the KdV equation; the Fermi–Pasta–Ulam problem for longitudinal waves propagating through nonlinear springs coupled together through a lattice of masses (Zabusky and Kruskal, 1965); and acoustic waves in crystalline lattices and other problems from quantum mechanics (Whitham, 1974, Ablowitz and Segur, 1981, Ablowitz and Clarkson, 1991). While the KdV equation has been repeatedly shown to have a range of useful applications, when it comes to geophysical fluid dynamics, caution does have to be taken with regard to the types of problems approached within this framework.

For two-layer flows where both layers had shallow depths, Koop and Butler (1981) showed that the KdV equation was able to accurately model the system for a wide range of amplitude configurations. However, later work by Grue et al. (1999) outlined that the KdV equation deviated from the results from the Euler equations, and experimental data when the wave amplitude increased above 0.4 in the presence of twin layers with moderate depth. As such, caution must be taken when it comes to both modelling physical processes through the KdV equation, as well as using conclusions from the KdV as a proxy for the results from other equations. While caution does need to be applied, the breadth of applications that this weakly-nonlinear equation can be applied towards is sufficient motivation to

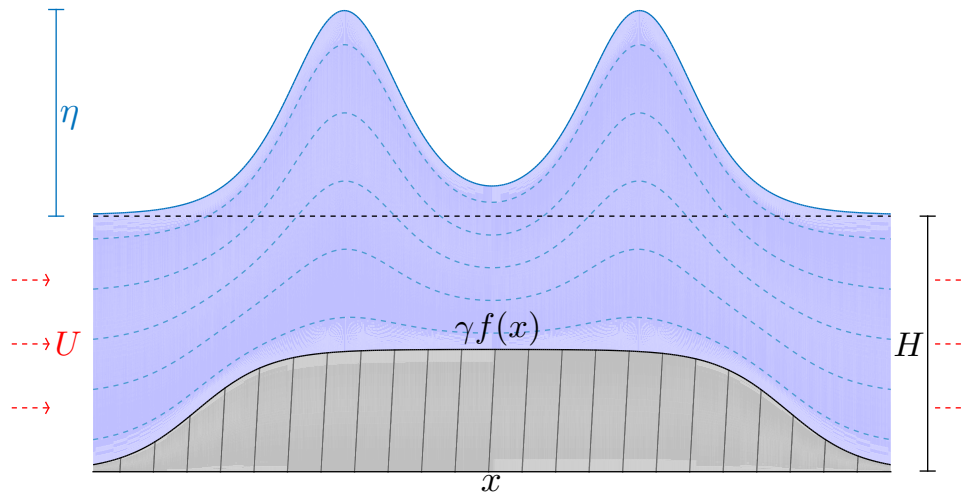


Figure 1.1: Schematic of channel flow subject to either a topographic disturbance.

develop a deeper understanding of the KdV equation and its variants. While these points of caution do need to be accounted for, the advantage of such models is that their greatly reduced dimensionality allows for insight to be gleaned that would otherwise be highly difficult—either in terms of the dimensionality, or the computational complexity required to be resolved—to yield when examining the Euler or Navier–Stokes equations.

From a mathematical perspective, the KdV equation is interesting in that it can take an integrable form, is well posed, can be solved exactly—subject to restricted initial conditions—by inverse-scattering theory Gardner et al. (1967, 1974) and the discovery that solutions of the KdV equation can form what are known as Solitons (Zabusky and Kruskal, 1965, Miura, 1968, Miura et al., 1968, Miura, 1976). These are solitary wave solutions which, while highly nonlinear, exhibit behaviours that are similar to superposition in that multiple solitons could pass through one another. Incorporating forcing to the KdV equation—known as the forced Korteweg–de Vries equation (fKdV)—allows for the influence of topographic and pressure effects to be incorporated within the model. This form of the equation also exhibits an interesting range of dynamics ranging from undular bores and solitary waves to locally steady flows as the Froude number—the ratio between the flow velocity and the wave velocity—progresses from the transcritical to the supercritical regime

(Grimshaw and Smyth, 1986).

When it comes to analysing wave behaviour, the KdV equation does not exist in isolation. Instead, it can be considered to be the weakly–nonlinear limit of what is known as the Gardner equation (Miura et al., 1968). A common derivation of the Gardner equation stems from taking a higher order asymptotic expansion of a problem, relative to the order employed to derive the KdV equation. Through this, the Gardner equation takes the form of the KdV equation, but with the addition of a cubic nonlinearity. This modification to the nonlinear component of the equation allows for saturation effects to be incorporated. This in turn allows the Gardner equation to qualitatively describe the limiting amplitude of internal, large amplitude solitary waves, in contrast to the conjugate flow description (Lamb and Wan, 1998, Rusas and Grue, 2002) which has been used to yield more quantitative results (Stastna and Peltier, 2005). Melville and Helfrich (1987) compared laboratory experiments for a two–layer immiscible flow in the presence of topography to a form of the forced KdV and forced Gardner equations, and found that in the presence of strong forcing the KdV showed poor agreement between the solutions of the KdV equation and relevant experimental results. However, when the experimental conditions were such that the effect of the cubic nonlinearity in the Gardner equation became important, the degree of correlation between the numerical solutions and experimental data was significantly improved. Their work also documented the presence of upstream undular bores and monotonic bores for transcritical Froude numbers, and steady solutions for supercritical flows. This work was reinforced by Grue et al. (1999), who examined the previous results in comparison to a fully nonlinear two–layer numerical solver. Helfrich and Melville (1986) also showed that in the case of inviscid internal waves it was necessary to incorporate the effect of viscous boundary layers into the derivation of the forced Gardner equation, as their contribution to the results was as significant as the nonlinear and dispersive effects present (Kakutani and Matsuuchi, 1975, Miles, 1976, Grimshaw, 1981).

Examining the Gardner equation from a finite–amplitude perspective was pioneered by Benney (Benney, 1966, Benney and Ko, 1978, Benney, 1979), who was the first to link the KdV and Gardner equations together as different limits of geophysical processes (Benney and Ko, 1978). The finite–amplitude limit also connects the Gardner equation to the Camassa–Choi (Choi and Camassa, 1999), Kawahara, Benny–Lin (Biagioni and Linares, 1997), Benney–Luke (Paumond, 2003), Boyd–Dodd (Dodd and Fordy, 1983, Boyd, 1986)



and Kadomtsev-Petviashvili (Kadomtsev and Petviashvili, 1970, Minzoni and Smyth, 1996) equations, and the work of Clarke and Johnson (1997a,b, 1999).

The KdV equation is also tied to the Grimshaw–Yi equation (Grimshaw and Yi, 1991, 1993), which is an integro–differential equation which describes strongly nonlinear waves at near resonance conditions, subject to weak uniform stratification. This equation can describe large amplitude waves until the point of axial flow reversal, which in turn allows for the prediction of wave breaking events. Hanazaki (1993), Rottman et al. (1996) have compared the solution of the Grimshaw–Yi equation to solutions of the governing Navier–Stokes equations, and demonstrated that the equation delivers quantitatively good approximations for both the solutions themselves, and the resulting predictions with regard to the onset of wave breaking. In the case where wave breaking does not occur—which occurs when the flow is not near an exact linear resonance—then Grimshaw and Yi (1991) noted that the solutions behave similarly to those of the fKdV equation, which reinforces the validity of studying the fKdV equation as a proxy for the more complicated Grimshaw–Yi equation.

One advantage of examining the KdV equation numerically is that, due to its numerous applications there are a wealth of results that demonstrate that the solitary wave solutions are well posed, as well as others covering the existence and stability of solitary wave solutions (see for example Miura (1976, 1968), Zhang (1992)). These results have generally been derived for the initial value problem formulation of the KdV equation, however, for the numerical work contained within this thesis, the focus will be upon investigating the steady variant of the equation, formulated as a boundary value problem.

As a result of their potential impact, it is of crucial importance to understand the evolution of breaking events. However, the very nature of these equations, in that they are nonlinear, significantly complicates the process of understanding their solution and parameter spaces. For the majority of nonlinear equations it is impossible to construct a closed–form, analytic solution to the problem, so instead they must be approached through analytic approximations or numerical schemes.

Historically, perturbation based schemes have been one of the most common approaches for constructing solutions to nonlinear equations that do not have an integrable solution. These schemes generally rely upon the equation including small or large parameters known

as the perturbation quantity, about which the perturbation can be constructed. Through manipulating the original equation, the nonlinear problem can be transformed into an infinite series of linear sub-problems, which when solved sequentially can potentially yield a solution to the original nonlinear equation. However, convergence is not guaranteed, and the success of these schemes is strongly dependent upon the size of the perturbation quantities. Beyond this, there is no guarantee that the linear sub-problems will be solvable. As such the applicability of perturbation schemes has generally been limited to weakly–nonlinear problems, where the relative impact of the nonlinearity on the overall solution is small.

Analytic or semi-analytic solutions to some fully nonlinear problems can be constructed with the artificial small–parameter method of Lyapunov (1892) being one of the most widely used approaches (Andrianov et al., 2003, PalTsev, 1967). In Lyapunov’s method, a nonlinear equation  $\mathcal{N}[u(\mathbf{x}, t)] = f(\mathbf{x}, t)$  is decomposed into its strictly linear and strictly nonlinear components, denoted by  $\mathcal{L}_0$  and  $\mathcal{N}_0$  respectively, so that

$$\mathcal{L}_0[u(\mathbf{x}, t)] + \mathcal{N}_0[u(\mathbf{x}, t)] = f(\mathbf{x}, t). \quad (1.5)$$

From here, the motivating nonlinear problem can be modified by the introduction of an artificial small parameter  $q$ , resulting in an equation of the form

$$\mathcal{L}_0[u(\mathbf{x}, t)] + q\mathcal{N}_0[u(\mathbf{x}, t)] = f(\mathbf{x}, t). \quad (1.6)$$

This equation then shares the form of equation (1.5) when  $q = 1$ . By treating  $u$  as a power series in terms of  $q$ , where

$$u(\mathbf{x}, t; q) = \sum_{m=0}^{\infty} u_m(\mathbf{x}, t)q^m \quad (1.7)$$

then equation (1.6) can be recast as an infinite series of linear problems, where

$$\mathcal{L}_0[u_0(\mathbf{x}, t)] = f(\mathbf{x}, t), \quad \mathcal{L}_0[u_1(\mathbf{x}, t)] = -\mathcal{N}_0[u_0(\mathbf{x}, t)] \dots \quad (1.8)$$

Lyapunov’s approach does not give any freedom in the choice of the operators  $\mathcal{L}_0$  and  $\mathcal{N}_0$  and as such, the series of equation (1.8) can easily surpass our capabilities to calculate solutions. Furthermore, the power series expansion in equation (1.7) is not guaranteed to converge.

Beyond Lyapunov there have been several other approaches to constructing solutions to nonlinear equations in terms of artificial small parameters, including the delta–expansion, Adomian decomposition (Adomian, 1986, 1994) and Homotopy perturbation methods (He, 2003). However, these techniques all share similar limitations, namely that there is no (or limited) flexibility in choosing the linear operator and the lack of ability to guarantee that the approximation series converges, nor to control the rate of convergence.

When it comes to constructing numerical solutions to nonlinear differential equations, the literature has a wealth of techniques for solving nonlinear differential equations, including Newton iteration, AiTEM (Yang and Lakoba, 2007), nonlinear Conjugate Gradient (Dhai and Yuan, 1999) and Multigrid based methods (Hemker, 1990), with Newton iteration being the most commonly used. However, none of these approaches are guaranteed to find extant solutions, and initial tests applying some of the more common approaches to the wave problems explored within this work were unsuccessful. As such, we were motivated to develop a new nonlinear solver that exhibited strong convergence control properties and spectral accuracy.

These tools will be applied in order to study the properties of the KdV and Gardner equations within a steady framework, subject to symmetric and asymmetric aperiodic boundary conditions. In order to confirm the validity of this exploration, the analytic properties and symmetries of these two equations will now be considered.

## 1.1 Properties of the Gardner equation

As the KdV equation is a specific case of the Gardner equation, it follows that the parameter spaces of the two equations must be interlinked. Consider the forced generalised Gardner equation for amplitude  $\eta = \eta(x, t)$  and topographic forcing  $g = g(x, t)$ :

$$\frac{\partial \eta}{\partial t} + \Delta \frac{\partial \eta}{\partial x} + r\eta \frac{\partial \eta}{\partial x} + q\eta^2 \frac{\partial \eta}{\partial x} + s \frac{\partial^3 \eta}{\partial x^3} = -\gamma g_x, \quad (1.9)$$

this can clearly be reduced to the form of the fKdV equation by setting  $q = 0$ . To develop an understanding of the properties of this equation, hereafter we will focus on the case where the forcing function  $g$  is independent of  $t$ , so that  $g = g_0 f(x/L)$ . Here  $L$  is the characteristic length scale, where  $0 \leq f \leq 1$  and  $f$  can be considered to be  $\mathcal{O}(1)$ .

As this equation is symmetric in  $x$ , we can impose that  $s$  is strictly positive without loss of generality. From this point, in order to convert the forced Gardner equation into its normalised form, the problem can be rescaled in terms of the parameters

$$t = \sigma\tau, \quad \eta(x, t) = \alpha A(\chi, \tau), \quad x = \beta\chi. \quad (1.10)$$

Making this substitution, and letting that  $\sigma = \beta^2/s$  results in the equation

$$\frac{\partial A}{\partial \tau} + \frac{\Delta\beta^2}{s} \frac{\partial A}{\partial \chi} + \frac{r\alpha\beta^2}{s} A \frac{\partial A}{\partial \chi} + \frac{q\alpha^2}{\beta} A^2 \frac{\partial A}{\partial \chi} + \frac{s}{\beta^3} \frac{\partial^3 A}{\partial \chi^3} = -\frac{\gamma g_0}{\alpha\beta} f_\chi. \quad (1.11)$$

We can simplify this equation by making the substitutions  $c_1 = \Delta\beta^2/s$  and  $f = f(\chi/\hat{L})$ , where  $\hat{L} = \sqrt{\Delta/c_1 s}L$ . As  $s > 0$ , we can say that  $\text{sgn}(c_1) = \text{sgn}(\Delta)$ . Making these changes, the forced Gardner equation is now

$$\frac{\partial A}{\partial \tau} + c_1 \frac{\partial A}{\partial \chi} + \frac{r\alpha c_1}{\Delta} A \frac{\partial A}{\partial \chi} + \frac{q\alpha^2 c_1}{\Delta} A^2 \frac{\partial A}{\partial \chi} + \frac{s}{\beta^3} \frac{\partial^3 A}{\partial \chi^3} = -\frac{\gamma g_0 c_1}{\alpha\Delta} f_\chi. \quad (1.12)$$

The Gardner equation can be considered as the bridge between the Korteweg–de Vries and the modified Korteweg–de Vries (mKdV) equations, which respectively have either a strictly quadratic or strictly cubic nonlinearity, corresponding to either  $q = 0$  or  $r = 0$  in equation (1.9). As such, an additional parameter  $\delta$  can be introduced, that will be a proxy for the transition between the KdV and mKdV equations. By setting that

$$\frac{r\alpha c_1}{2\Delta} = 3(1 - \delta), \quad \frac{q\alpha^2 c_1}{3\Delta} = 3c_2\delta, \quad (1.13)$$

for some  $c_2$  where  $\text{sgn}(c_2) = \text{sgn}(q)$  then it follows that

$$\delta = \frac{r\alpha c_1}{6\Delta}.$$

Equating both parts of equation (1.13) and solving for  $\alpha$  yields

$$\delta = -\phi \pm \sqrt{\phi^2 + 2\phi},$$

where  $\phi = r^2 c_1 c_2 / (8\Delta q)$ . Without loss of generality we can restrict ourself to the positive branch, and then as long as  $\phi > 0$  then  $0 \leq \delta \leq 1$ . Thus equation (1.9) can be reduced to the form

$$\frac{\partial A}{\partial \tau} + c_1 \frac{\partial A}{\partial \chi} + 6(1 - \delta)A \frac{\partial A}{\partial \chi} + 9\delta c_2 A^2 \frac{\partial A}{\partial \chi} + \frac{\partial^3 A}{\partial \chi^3} = -\hat{\gamma} f_\chi, \quad (1.14)$$

where

$$\hat{\gamma} = \frac{\gamma g_0 r c_1^2}{6\Delta^2 \delta}, \quad \delta = -\phi + \sqrt{\phi^2 + 2\phi}, \quad \phi = \frac{r^2 c_1 c_2}{8\Delta q} \quad \hat{L} = \sqrt{\frac{\Delta}{c_1 s}} L.$$

As  $c_2$  is arbitrary, we are free to set that  $c_2 = \pm 1$ , which lets us further simplify equation (1.14) to

$$\frac{\partial A}{\partial \tau} + c_1 \frac{\partial A}{\partial \chi} + 6(1 - \delta)A \frac{\partial A}{\partial \chi} \pm 9\delta A^2 \frac{\partial A}{\partial \chi} + \frac{\partial^3 A}{\partial \chi^3} = -\hat{\gamma} f_\chi. \quad (1.15)$$

To confirm the validity of this substitution, we must confirm that  $\delta$  and  $\hat{\gamma}$  remain finite. For the instance where  $r \rightarrow 0$ , through a Taylor series expansion it follows that

$$\delta \approx \sqrt{2\phi} = |r| \sqrt{\frac{1}{8} \frac{c_1 c_2}{\Delta q}} \quad (1.16)$$

and

$$\hat{\gamma} = \frac{\gamma g_0 r c_1^2}{6\Delta^2 \delta} \approx \frac{\gamma g_0 r c_1^2}{6\Delta^2} \sqrt{4 \frac{\Delta q}{c_1 c_2}}. \quad (1.17)$$

Similarly, as  $q \rightarrow 0$ ,

$$\left. \begin{aligned} \delta &\approx 1 - \frac{1}{2\phi} = 1 + \mathcal{O}(q) \\ \text{and } \hat{\gamma} &\approx -\frac{8}{3} \frac{\gamma g_0 c_1 q}{\Delta r c_2}. \end{aligned} \right\} \quad (1.18)$$

Thus, both  $\delta$  and  $\hat{\gamma}$  remain finite as the solution transitions between the KdV and mKdV equations. Finally, we can implement bounds upon  $\delta$  by noting that  $\delta = -\phi + \sqrt{\phi^2 + 2\phi}$  is only real when  $\delta \geq 0$ . Furthermore, as

$$\frac{d\delta}{d\phi} = \frac{\phi + 1}{\sqrt{\phi^2 + 2\phi}} - 1$$

is both strictly positive and monotonically decreasing for all  $\phi \geq 0$ , then the maximum of  $\delta$  will occur as  $\phi \rightarrow \infty$ . This allows us then to say that  $\delta$  must be bounded to exist within  $[0, 1]$ , where  $\delta = 0$  corresponds to the KdV equation and  $\delta = 1$  to the mKdV equation.

As the results contained here are focussed on the steady variant of the Gardner equation, in that context the aforementioned equation will take the form

$$c_1 A + 3(1 - \delta)A^2 \pm 3\delta A^3 + \frac{d^2 A}{d\chi^2} = -\hat{\gamma}f, \quad \delta \in [0, 1], \quad A = A(\chi). \quad (1.19)$$

Henceforth we will consider this equation as

$$\left. \begin{aligned} \Delta A + rA^2 + qA^3 + \frac{d^2 A}{dx^2} &= -\gamma f, & \delta \in [0, 1], & \quad A = A(x), \\ r &= 3(1 - \delta), \\ q &= \pm 3\delta. \end{aligned} \right\} \quad (1.20)$$

Where the changes from  $\chi \rightarrow x$  and  $c_1 \rightarrow \Delta$  serve to simplify the notation.

## 1.2 Thesis outline

This thesis chronicles the development of a suite of tools for solving nonlinear boundary value problems in one-dimension, with a particular focus upon the forced Gardner equation. The core solver is flexible and highly efficient, exhibiting spectral convergence and quasi-linear scaling in computational cost with the grid size. This solver is built upon repeatedly solving a modified linear system in terms of a sparse linear operator. Crucially for the scheme's numerical performance, all components involving a matrix inverse can be cast to remain constant with the iteration number. Combining this new solver with a wider suite of associated tools will allow us to explore problems from fluid mechanics, with particular focus on equations involving large scale atmospheric waves. All the following chapters contain new contributions to the fields of numerical analysis, wave dynamics and scientific computing.

Chapter 2 is broadly a review of the current state of numerical analysis relevant to linear and nonlinear boundary value problems. Spectrally accurate tools to solve these problems will be introduced, with a particular focus upon numerical discretisations involving Chebyshev and Gegenbauer polynomials. By adding to the work of Olver and Townsend (2013), this chapter will provide the core of the nonlinear solver that is central to numerical work contained within this thesis.

To extend this work into nonlinear boundary value problems, Chapter 3 introduces the Homotopy Analysis Method (HAM), which is an analytic technique developed by Liao (1992). This technique provides an approach for solving nonlinear equations that is based

upon generating a homotopy between a linear equation and the full nonlinear problem, in a manner that allows the nonlinear problem to be solved in terms of a linear perturbation scheme. Within this chapter a new, numerical form of HAM will be introduced, known as the Gegenbauer Homotopy Analysis Method (GHAM). This technique avoids the pitfalls of previously developed nonlinear solvers, in that it exhibits spectral convergence while preserving a sparse, constant linear operator for variable coefficient problems. As the scope of applicability of homotopy based methods has been limited by the lack of associated methods of numerical continuation, several novel approaches for numerical continuation are introduced to resolve this situation. The properties of this suite of new techniques are explored analytically through the development of new convergence theorems; and numerically by constructing solutions to several standard boundary layer equations.

Chapter 4 and Chapter 5 bring together the work of the preceding chapters to examine solutions to the forced Korteweg–de Vries (fKdV) equation, from a symmetric and asymmetric perspective. This equation is a one–dimensional, second–order boundary value problem subject to a single quadratic nonlinearity with broad applications to geophysical fluid dynamics and other fields. By employing GHAM, several new insights are developed with regard to the influence of the topographic length scale upon symmetric solutions; and the impact of the topographic drag upon asymmetric solutions.

Finally in Chapter 6 a systematic exploration of the parameter space of symmetric solutions to the forced Gardner equation will be performed. The Gardner equation is a one–dimensional nonlinear boundary value problem that incorporates a cubic polynomial nonlinearity. As the parameter space of the Gardner equation includes both the fKdV and the modified Korteweg–de Vries (mKdV) equations, understanding the development of solutions within the parameter space of the Gardner equation has broad implications to geophysical fluid dynamics.

# Chapter 2

## Numerical Methods

It is uncommon for complex differential equations to exhibit analytical solutions that hold for anything more than a small subset of conditions. This is particularly true for nonlinear differential equations, and as such the only reliable option for solving such equations is through numerical methods. These schemes involve discretising the equations involved over some space, and then constructing an iterative method to refine a trial solution until the point at which the solution accurately solves the original equations. In broad terms, the numerical solutions to differential equations can be categorised into two main approaches: local and global methods.

Local methods form the framework to the most commonly used numerical tools such as finite difference and finite element methods, and involves the construction of solutions in terms of basis functions that are only locally nonzero. That these basis functions are broadly zero simplifies both the implementation and the solution process for local methods. Global methods, otherwise known as Spectral methods, are more complex to both implement and solve, as they involve basis functions that are nonzero over the majority of the numerical domain. While these methods are more complex, a consequence of the global basis functions is that these methods exhibit exponential convergence, where any increase in the numerical resolution results in a corresponding exponential decrease in the observed error, subject to the restriction that solutions must not exhibit discontinuities. This compares favourably to local methods, for which the error decreases in a polynomial fashion with increases in the numerical resolution.



While spectral methods are more complicated to implement, their enhanced convergence properties make them a valuable tool for the numerical analysis of differential equations. Henceforth this work will focus on the development of novel numerical methods for solving nonlinear differential equations, with a particular focus upon enhancing the exhibited computational performance relative to current techniques. To begin with, this chapter will be devoted to outlining the current state of knowledge for spectral methods in the context of linear and nonlinear differential equations, while also incorporating some novel results.

## 2.1 Spectral Methods

A common approach to constructing spectral methods is to take the Fourier transform of a differential equation (Orszag, 1969). However, this approach is limited to problems where the solution is periodic on the boundaries, or where the boundary conditions can be approximated as being periodic. For aperiodic boundary value problems, a more appropriate set of basis functions are the Chebyshev polynomials, which will serve as the core of the numerical work contained within this thesis. Unlike the Fourier basis functions, Chebyshev polynomials have fixed end points, and as such can be applied to problems subject to aperiodic boundary conditions.

### 2.1.1 Chebyshev polynomials

These Chebyshev polynomials correspond to a family of orthogonal polynomials with numerous important applications in numerical analysis and scientific computing, including in numerical schemes; for polynomial interpolation; and within random matrices. The Chebyshev polynomials of the first kind—denoted by  $T_j$ —correspond to eigenfunctions of the singular Sturm–Liouville problem

$$\frac{d}{dx} \left( \sqrt{1-x^2} \frac{dT_j}{dx}(x) \right) + \frac{j}{\sqrt{1-x^2}} T_j(x) = 0, \quad -1 \leq x \leq 1. \quad (2.1)$$

These eigenfunctions have been well studied (Canuto et al., 1988, Mason and Handscomb, 2002) and admit solutions of the form

$$T_j(x) = \cos(j \cos^{-1} x) \quad x \in [-1, 1], \quad j = 0, 1, \dots \quad (2.2)$$

A useful property of these Chebyshev polynomials is their orthogonality relative to a weighted inner product, which manifests as

$$\int_{-1}^1 \frac{T_i(x)T_j(x)}{\sqrt{1-x^2}} dx = \begin{cases} \pi & \text{if } i = j = 0, \\ \frac{\pi}{2} & \text{if } i = j \geq 1, \\ 0 & \text{if } i \neq j. \end{cases} \quad (2.3)$$

A consequence of this is that the Chebyshev polynomials must satisfy the three-term recurrence relationship

$$T_{j+1}(x) = 2xT_j(x) - T_{j-1}(x), \quad j \geq 1, \quad (2.4)$$

where  $T_0(x) = 1$  and  $T_1(x) = x$  (Canuto et al., 1988). This property in turn can be leveraged to serve as the basis of numerical schemes constructed in terms of the Chebyshev polynomials. These numerical schemes typically involve transforms from functions in real space to a form in terms of a weighted sum of Chebyshev polynomials, which will henceforth be referred to as Chebyshev space. This allows a function  $u(x)$  to be expressed as

$$u(x) = \sum_{j=0}^{\infty} \hat{u}_j T_j(x) \quad (2.5)$$

subject to the restriction that  $x \in [-1, 1]$ , and where  $\hat{u}_j \in \mathbb{R}$  are the Chebyshev coefficients of  $u(x)$ , and  $x$  is evaluated at the Chebyshev collocation points  $x_k$ , the form of which will be discussed within Subsection 2.1.2. However, for the purposes of numerical schemes it is difficult to incorporate a numerical discretisation that involves infinitely many terms, as such we will instead consider the Chebyshev sum truncated to order  $N$

$$u(x) = \sum_{j=0}^N \hat{u}_j T_j(x), \quad u = u(x), \quad (2.6)$$

The transform between  $u(x)$  and  $\hat{u}$  is equivalent to a weighted Discrete Fourier Transform, and the converse operation can also be performed using the Inverse Discrete Fourier Transform (Trefethen, 2013).

### 2.1.2 Chebyshev Quadrature Points

Typically these polynomials in Chebyshev space are represented in terms of collocation (or quadrature) points, with the choice of the distribution of points having implications for integration and differentiation operators. As compared to other orthogonal polynomials—such as the Legendre polynomials—the position and weights of the Chebyshev polynomials

can be analytically determined, and take three forms: *Chebyshev–Gauss*, *Chebyshev–Gauss–Radau* and the *Chebyshev–Gauss–Lobatto* points.

For this work, the collocation points employed are the *Chebyshev–Gauss* points, which take the form

$$x_n = \cos \frac{(2n+1)\pi}{2N} \quad n = 0, \dots, N-1, \quad (2.7)$$

for any integer  $N \in \mathbb{Z}^+$ .

### 2.1.3 Spectral convergence of Chebyshev approximations

Of particular importance for numerical schemes is how the accuracy of Chebyshev approximations improves with respect to the number of grid points employed in the discretisation. If  $f(\mathbf{x})$  is a  $m$  times differentiable function evaluated upon the Chebyshev–Gauss quadrature points, and the  $m$ -th derivative exhibits bounded total variation then the  $L_\infty$  norm of the difference between  $f(\mathbf{x})$  and its Chebyshev approximation  $p_N$  of the form of equation (2.6) behaves as

$$\|f(\mathbf{x}) - p_N(\mathbf{x})\|_\infty = \mathcal{O}(N^{-m})$$

Furthermore, if  $f(x)$  is infinitely differentiable, then the convergence rate must be faster than  $\mathcal{O}(N^{-m})$ , and as such the approximating function exhibits spectral accuracy.

### 2.1.4 Differentiation in Chebyshev space

For the purposes of constructing Chebyshev methods, it is important to consider the construction of both differential and integral operators. By recasting equation (2.2)—the general form of the Chebyshev polynomials—through the change of variables  $x = \cos(\theta)$ , so that

$$T_j(x) = \cos(j\theta),$$

then it is possible to describe the derivative of a Chebyshev polynomial as

$$\frac{dT_j(x)}{dx} = -n \sin(j\theta) \frac{d\theta}{dx} = \frac{j \sin(j \cos^{-1}(x))}{\sqrt{1-x^2}}. \quad (2.8)$$

As a consequence of this, the derivatives of a function  $u(x)$  expressed in terms of Chebyshev polynomials

$$u = \sum_{n=0}^N \hat{u}_n T_n(x), \quad u = u(x), \quad (2.9)$$

can be evaluated by taking

$$\frac{du}{dx} = \sum_{n=0}^N \hat{u}_n \frac{dT_n(x)}{dx} = \frac{\sum_{n=0}^N \hat{u}_n n \sin(n\theta)}{\sqrt{1-x^2}}. \quad (2.10)$$

Higher order derivatives can be similarly constructed through this approach, by simply applying successive first-order derivatives.

The dependence upon  $(1-x^2)^{-1/2}$  imposes that equation (2.10) must be singular at  $x = \pm 1$ . In general these derivatives exist, however the relationship used to calculate them no longer holds at these extrema of the domain. Instead derivatives can be evaluated at the boundaries of the Chebyshev domain by evaluating

$$T'_j(\pm 1) = (\pm 1)^{j+1} j^2 \quad \text{for} \quad j = 0, \dots, N.$$

This inherent singularity is also present in higher order derivatives, however in a similar vein to the formulation above

$$T''_j(\pm 1) = (\pm 1)^j \left( \frac{j^4 - j^2}{3} \right), \quad (2.11)$$

and when generalised to  $n$ -th order derivatives becomes

$$\frac{d^n T_j}{dx^n}(\pm 1) = (\pm 1)^{j+n} \prod_{k=0}^{n-1} \frac{j^2 - k^2}{2k+1}. \quad (2.12)$$

Just as the Discrete Fourier Transform can be used to calculate Chebyshev coefficients it is similarly possible to use the Fast Fourier Transform to calculate the derivatives of a function  $u(x)$  on *Chebyshev-Gauss-Lobatto* points. Expressing  $u(\mathbf{x})$  as  $U = \{u(x_0), u(x_1), \dots, u(x_N), u(x_{N-1}), \dots, u(x_1)\}$ , then the Chebyshev coefficients  $\hat{u}_k$  can be calculated by taking

$$\hat{u}_k = \frac{\pi}{N} \sum_{j=1}^{2N} e^{-ik\theta_j} U_j, \text{ for } k = -N + 1, \dots, N$$

which is the equivalent of taking the Fast Fourier Transform of  $U$ . Differentiating this gives that the first derivative, expressed in Chebyshev space, must be equivalent to

$$\hat{W}_k = ik\hat{u}_k, \text{ with } \hat{W}_N = 0.$$

Higher order derivatives can be calculated within Chebyshev space by repeatedly applying this process. This derivative can then be transformed back to real space via an operation that is equivalent to taking the Inverse Fast Fourier Transform, so that

$$W_j = \frac{1}{2\pi} \sum_{k=-N+1}^N e^{ik\theta_j} \hat{W}_k, \text{ for } j = 1, \dots, 2N.$$

Rescaling these parameters then returns the solution to physical space

$$\left. \begin{aligned} w_j &= -\frac{W_j}{\sqrt{1-x_j^2}}, \text{ for } j = 1, \dots, N-1 \\ w_0 &= \frac{1}{2\pi} \sum_{n=0}^N n^2 \hat{u}_n, \\ w_N &= \frac{1}{2\pi} \sum_{n=0}^N (-1)^{n+1} n^2 \hat{u}_n. \end{aligned} \right\} \quad (2.13)$$

Here  $w_j$  is shorthand for  $du(x_j)/dx$ . Employing the Fast Fourier Transform to evaluate derivatives exhibits  $\mathcal{O}(n \log n)$  scaling, which is significantly faster than performing the equivalent operations manually, or by calculating the derivatives through the use of equation (2.8).

### 2.1.5 Integration in Chebyshev space

Integrating over Chebyshev quadrature points presents more flexibility, as the integral can be evaluated in terms of its representation in either real or Chebyshev space. In real space the integral of  $u(x)$  over  $x \in [-1, 1]$  can be approximated by the quadrature

$$\int_{-1}^1 u(x)dx \approx \sum_{i=0}^{n-1} w_i \sqrt{1-x_i^2} u(x_i), \quad (2.14)$$

subject to the requirement that  $x_i$  must be defined upon the Chebyshev–Gauss points of Subsection 2.1.2. Here the associated quadrature weights are  $w_i = \frac{\pi}{N+1}$  (Abramowitz and Stegun, 1972).

An alternate approach to evaluating the integral of  $u(x)$  is to consider the problem in terms of its Chebyshev coefficients  $\hat{u}$ . By doing so, the integral over  $x \in [-1, 1]$  can be constructed using Fejér’s rule (Fejér, 1933, Dahlquist and Björck, 2008)—otherwise known as the Clenshaw–Curtis quadrature (Clenshaw, 1972, Daubechies, 1992)—by exploiting the fact that

$$\mu_k = \int_{-1}^1 T_k(x)dx = \begin{cases} 0 & \text{if } k \text{ is odd,} \\ 2/(1-k^2) & \text{if } k \text{ is even.} \end{cases}$$

It then follows that associated quadrature weights of

$$w_i = \begin{cases} \frac{2}{1-i^2} & \text{if } i \text{ is even,} \\ 0 & \text{if } i \text{ is odd,} \end{cases} \quad (2.15)$$

can then in turn be used to calculate

$$\int_{-1}^1 u(x)dx = \sum_{i=0}^{n-1} w_i \hat{u}_i. \quad (2.16)$$

As the form of  $\mu_k$  can be determined without introducing any approximations this quadrature formula does not introduce any numerical error into the construction of the integral, beyond any introduced through the transforms between Chebyshev and real space.

### 2.1.6 Multiplication in Chebyshev space

While addition operations in Chebyshev space can be constructed trivially, evaluating the product of functions within this space is significantly more complicated. For numerical solutions to variable coefficient boundary value problems, any product between the coefficient terms and their corresponding differential operators must be realisable within a matrix formulation. One approach for these products is to treat them as a convolution operation,

however it is not possible to discretise such an operation in terms of matrix products, and as such convolution based approaches are not a viable approach when considering variable coefficient boundary value problems.

An alternate approach can be found by considering the product of two functions  $a(x)$  and  $u(x)$ , both of which are represented in Chebyshev space through

$$a(x) = \sum_{j=0}^{\infty} a_j T_j(x) \quad \text{and} \quad u(x) = \sum_{k=0}^{\infty} u_k T_k(x).$$

Truncating these two polynomials to degree  $N$  allows the product to be represented in terms of the Cauchy product

$$a(x)u(x) = \sum_{j=0}^N \sum_{k=0}^N a_j u_k T_j(x) T_k(x) = \sum_{k=0}^{2N} c_k T_k(x),$$

subject to finding some  $\mathbf{c} = (c_0, c_1, \dots)$ . Following Baszenski and Tasche (1997), for the case of Chebyshev polynomials  $c_k$  has the closed form

$$c_k = \begin{cases} \frac{a_0 u_0}{2} + \sum_{j=1}^{2N} a_j u_j & \text{for } k = 0, \\ \frac{1}{2} \sum_{j=0}^k a_{k-j} u_j + \frac{1}{2} \sum_{l=1}^{N-k} (a_j u_{k+j} + a_{k+j} u_j) & \text{for } k = 1, \dots, N-1. \end{cases}$$

This then allows the matrix multiplication of the vectors  $a$  and  $u$  to be written as a matrix in terms of  $a$  multiplied by the vector  $u$ . However, this process is ostensibly  $\mathcal{O}(N^2)$ —which is to say that for a polynomial truncated to order  $N$ , then the multiplication will require on the order of  $N^2$  operations—and as such, is relatively inefficient. However, we will show in Subsection 2.2.2 that it is possible to exploit the Gegenbauer polynomials to construct matrices to represent such operations in a simpler manner.

### 2.1.7 Gegenbauer polynomials

For all their advantageous numerical properties, Chebyshev polynomials have limited utility for constructing matrix operators for variable coefficient linear boundary value problems, as the resulting matrix operators rapidly become singular with increases in the grid resolution. However, this tendency can be avoided by instead turning to the Gegenbauer (otherwise known as Ultraspherical) polynomials

$$C_j^{(\lambda)} \text{ for } j \geq 0$$

The Chebyshev and Legendre polynomials correspond to limiting cases of these Gegenbauer polynomials, as

$$T_j(x) = \lim_{\lambda \rightarrow 0^+} \frac{j}{2\lambda} C_j^{(\lambda)}(x).$$

The Gegenbauer polynomials exhibit the previously introduced advantageous properties of the Chebyshev polynomials, while also introducing an advantageous sparsity to the matrix operators. These polynomials are orthogonal with respect to the weight

$$(1 - x^2)^{\lambda - \frac{1}{2}},$$

and take the form

$$C_k^{(\lambda)} = \frac{2^k(\lambda + k - 1)!}{k!(\lambda - 1)!} x^k + \mathcal{O}(x^{k-1}).$$

More specifically, the Gegenbauer polynomials can be constructed in terms of the recurrence relationship

$$C_{j+1}^{(\lambda)}(x) = \frac{2(j + \lambda)}{j + 1} x C_j^{(\lambda)}(x) - \frac{j + 2\lambda - 1}{j + 1} C_{j-1}^{(\lambda)}(x), \quad j \geq 1, \quad (2.17)$$

subject to  $C_0^{(\lambda)} = 1$  and  $C_1^{(\lambda)} = 2\lambda x$  for all  $\lambda \in \mathbb{Z}^+$ . One notable advantage of constructing a numerical scheme in terms of the Gegenbauer polynomials is that differentiation in Chebyshev space can be expressed as

$$\frac{d^k T_j}{dx^k} = \begin{cases} 2^{k-1} j(k-1)! C_{n-k}^{(k)} & \text{for } j \geq k, \\ 0 & \text{for } 0 \leq j \leq k-1. \end{cases}$$

Consequently Chebyshev differentiation operators of any order can be represented by a sparse matrix in Gegenbauer space. This compares favourably to the traditional approach of constructing collocation based Chebyshev differentiation matrix operators, which results in dense matrix representations.



## 2.2 Spectral Methods for differential equations

Having now established the basic tools required for representing functions in terms of Chebyshev and Gegenbauer polynomials, we can turn to the application of these basis functions for constructing spectrally accurate solutions to linear boundary value problems. Two such approaches will be presented in the proceeding work, the first of which involves discretising one-dimensional boundary value problems with Chebyshev collocation matrices, which have been suggested to be the most efficient numerical approach for solving variable coefficient problems (Boyd, 2001). However, such methods produce dense, numerically difficult matrices, and generally struggle in the presence of singular points. This latter problem is of particular concern for problems incorporating domain transformations from infinite and semi-infinite spaces to  $x \in [-1, 1]$ .

The second approach is based upon the work of Olver and Townsend (2013), and leverages the properties of both Gegenbauer and Chebyshev polynomials for a scheme that demonstrates super-algebraic convergence through the use of sparse, almost-banded and well-conditioned matrix equations. This technique involves constructing solutions to linear equations within Gegenbauer space, in a manner which is both fast and able to handle highly oscillatory functions.

### 2.2.1 Chebyshev collocation matrices for BVPs

Chebyshev collocation matrices can be used to construct solutions to linear differential equations defined in terms of the arbitrary linear operator  $\mathcal{L}$  :

$$\mathcal{L}[u(x)] = f(x) \text{ subject to } u(-1) = 0, u(1) = 0,$$

by partitioning the unknown solution  $u$  in Chebyshev space in the manner of equation (2.6). By then exploiting the differentiability of the Chebyshev polynomials—following Subsection 2.1.4—the  $m$ -th order derivative at the collocation points can be expressed as

$$\frac{d^m u}{dx^m}(x_j) = \sum_{k=0}^N D_{jk}^m u(x_k), \quad j = 0, \dots, N, \quad (2.18)$$

where  $D$  is the Chebyshev spectral differentiation matrix

$$\left. \begin{aligned}
D_{jk} &= \left(\frac{1}{2}\right) \frac{c_j}{c_k} \frac{(-1)^{(j+k+1)}}{\sin\left(\frac{\pi}{2N}(j+k)\right) \sin\left(\frac{\pi}{2N}(k-j)\right)}, \quad j \neq k, \\
D_{jj} &= \left(\frac{-1}{2}\right) \frac{x_k}{\sin^2\left(\frac{\pi k}{N}\right)}, \quad j \neq 0, N, \\
D_{00} &= -D_{NN} = \frac{2N^2 + 1}{6}.
\end{aligned} \right\} \quad (2.19)$$

Being able to render all derivative operators in  $\mathcal{L}$  in terms of powers of  $D$  makes this Chebyshev collocation approach relatively simple to implement, as the operator

$$\mathcal{L}[u(x)] = \sum_{l=0}^L a_l(x) \frac{d^l u(x)}{dx^l} = \phi(x)$$

can be discretised in terms of the matrix equation

$$\mathbf{A}\mathbf{U} = \mathbf{G},$$

the components of which take the form

$$\left. \begin{aligned}
\mathbf{A} &= \sum_{l=0}^L \mathbf{a}_l D^l, \\
\mathbf{a}_l &= \text{diag}[a_l(x_0), a_l(x_1), \dots, a_l(x_N)], \\
\mathbf{U} &= [u(x_0), u(x_1), \dots, u(x_N)]^T, \\
\mathbf{G} &= [\phi(x_0), \phi(x_1), \dots, \phi(x_N)]^T,
\end{aligned} \right\} \quad (2.20)$$

and the evaluation points  $\{x_0, x_1, \dots, x_N\}$  are the Chebyshev collocation points.

### 2.2.2 Gegenbauer polynomials for BVPs

An alternate approach for constructing spectrally accurate solutions to boundary value problems was recently introduced by Olver and Townsend (2013), based upon the Gegenbauer polynomials. To understand the properties of this scheme, we again consider the linear boundary value problem

$$\left. \begin{aligned}
\mathcal{L}u(x) &= f(x) \quad \text{and} \quad \mathcal{B}u(x) = \mathbf{c} \\
\mathcal{L} &= a_N(x) \frac{d^N}{dx^N} + \dots + a_1(x) \frac{d}{dx} + a_0(x),
\end{aligned} \right\} \quad (2.21)$$

Here  $u(x)$  is the unknown solution, subject to  $K$  boundary conditions expressed by the linear differential operator  $\mathcal{B}$  and where all three of  $\mathbf{c} \in \mathbb{C}^K$ ,  $\{a_i(x)|i = 0, 1, \dots, N\}$  and  $f(x)$  are smooth, non-singular functions.

Olver and Townsend (2013) noted that if the differentiation operators are represented in terms of the Gegenbauer polynomials  $C_n^{(\alpha)}(x)$  then, unlike Chebyshev based techniques, the resulting matrix discretisation will be sparse. This intriguing property is a direct consequence of the inherent sparsity in the mapping between Chebyshev and Gegenbauer spaces. The resulting numerical discretisation will be sparse and banded, even for variable coefficient boundary value problems. As such, solving numerical problems upon a Gegenbauer basis has significant numerical implications, both in terms of the time required to solve these systems, and the memory required to solve the matrix operators. The Olver and Townsend method also has the additional advantage of allowing linear boundary conditions to be imposed in a flexible manner through boundary bordering, which means that Dirichlet, Robin, mixed and integral boundary conditions can all be simply incorporated into the numerical scheme.

To explore the process of solving linear boundary value problems upon a Gegenbauer basis, let us first consider the generalised first-order linear operator

$$\mathcal{L} = \frac{d}{dx} + a_0(x) \tag{2.22}$$

applied to an unknown function partitioned upon a Chebyshev basis in the manner of equation (2.6). Through Subsection 2.1.7 the derivative of  $u$  in terms of the Gegenbauer polynomials is

$$\frac{du}{dx}(x) = \sum_{k=1}^N k \hat{u}_k C_{k-1}^{(1)}(x), \tag{2.23}$$

which can naturally be partitioned to take the form of a matrix operator acting upon  $\hat{u}_k$  for all  $k \in [1, N]$ . As such the first-order linear differential equation can be partitioned as the product of  $\mathcal{D}_1 \mathcal{C}_1 \hat{u}$ , where  $\mathcal{D}_1$  and  $\mathcal{C}_1$  are square matrix operators in  $\mathbb{R}^{N \times N}$  that take the form

$$\mathcal{D}_1 = \begin{pmatrix} 0 & 1 & & & \\ & & 2 & & \\ & & & 3 & \\ & & & & 4 \\ & & & & & \ddots \end{pmatrix} \quad (2.24)$$

and  $\mathcal{C}_1 = C_{k-1}^{(1)}$ . From this equation (2.22) can be discretised as

$$\mathcal{L} := \mathcal{D}_1 \mathcal{C}_1 + \mathcal{M}[a_0], \quad (2.25)$$

where  $\mathcal{M}[a_0] \in \mathbb{R}^{N \times N}$  is a matrix operator that when multiplied by  $\hat{u}$  is equal to the discrete form of  $a_0 u(x)$  in Chebyshev space. The structure of  $\mathcal{M}[a_0]$  can be elucidated by considering the discretisation of  $a_0(x)$  in Chebyshev space, which allows the product to be expressed as

$$a(x)u(x) = \sum_{j=0}^N \sum_{k=0}^N \hat{a}_j \hat{u}_k T_j(x) T_k(x)$$

This sum of sums can be simplified through the results of Subsection 2.1.6 to take the form

$$a(x)u(x) = \sum_{k=0}^N c_k T_k(x)$$

for some  $\mathbf{c} = (c_0, c_1, \dots)$ . Solving for  $\mathbf{c}$  reveals that  $\mathcal{M}_0[a_0]$  takes the form of the sum of a Toeplitz matrix and a modified Hankel operator, so that

$$\mathcal{M}_0[a] = \frac{1}{2} \left[ \begin{pmatrix} 2a_0 & a_1 & a_2 & \cdots \\ a_1 & 2a_0 & a_1 & \ddots \\ a_2 & a_1 & 2a_0 & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 & \cdots \\ a_1 & a_2 & a_3 & \ddots \\ a_2 & a_3 & a_4 & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{pmatrix} \right].$$

While the differential operator acts in Gegenbauer space, as it is entirely constructed in terms of the Gegenbauer polynomials of the second kind  $C_k^{(1)}$ , the same does not hold for the multiplication operator. As this operator is constructed as a product of the Chebyshev polynomials of the first kind  $T_k$ , it must exist in Chebyshev space, rather than Gegenbauer space. This suggests that equation (2.25) must be modified so that both of its terms exist in Gegenbauer space.

This modification can be constructed by introducing an additional operator that maps between  $T_k$  and  $C_k^{(1)}$  must be introduced. Through equation (2.17) the recurrence relationship

$$T_k = \begin{cases} \frac{1}{2}(C_k^{(1)} - C_{k-2}^{(1)}) & k \geq 2, \\ \frac{1}{2}C_1^{(1)} & k = 1, \\ C_0^{(1)} & k = 0, \end{cases}$$

can be derived, and any series in terms of Chebyshev polynomials of the first kind can instead be represented as

$$\begin{aligned} \sum_{k=0}^{\infty} u_k T_k(x) &= u_0 C_0^{(1)}(x) + \frac{1}{2} u_1 C_1^{(1)}(x) + \frac{1}{2} \sum_{k=2}^{\infty} u_k (C_k^{(1)}(x) - C_{k-2}^{(1)}(x)) \\ &= (u_0 - \frac{1}{2} u_2) C_0^{(1)}(x) + \sum_{k=1}^{\infty} \frac{1}{2} (u_k - u_{k+2}) C_k^{(1)}(x). \end{aligned} \quad (2.26)$$

Expressing this conversion process as the application of an operator  $\mathcal{S}_0 \in \mathbb{R}^{N \times N}$

$$\mathcal{S}_0 = \begin{pmatrix} 1 & -\frac{1}{2} & & & \\ & \frac{1}{2} & -\frac{1}{2} & & \\ & & \frac{1}{2} & -\frac{1}{2} & \\ & & & \ddots & \ddots \\ & & & & \ddots & \ddots \end{pmatrix} \quad (2.27)$$

acting on  $\hat{\mathbf{u}}$  allows the operator equation (2.22) to be discretised as

$$\left. \begin{aligned} \mathcal{L}\hat{\mathbf{u}} &= \mathcal{S}_0 \mathbf{f}, \\ \mathcal{L} &:= \mathcal{D}_1 + \mathcal{S}_0 \mathcal{M}_0[a]. \end{aligned} \right\} \quad (2.28)$$

The process for discretising linear operators can be generalised to equations involving higher order derivatives following a modified version of the procedure outlined above. In terms of the Gegenbauer polynomials  $C^{(\lambda)}$ , we can write the general form of the derivative as

$$\frac{dC_k^{(\lambda)}}{dx}(x) = \begin{cases} 2\lambda C_{k-1}^{(\lambda+1)} & k \geq 1, \\ 0 & k = 0. \end{cases}$$

Using the recurrence relationship



$$a(x) = \sum_{j=0}^{\infty} a_j C_j^{(\lambda)}(x) \text{ and } u(x) = \sum_{k=0}^{\infty} \hat{u}_k C_k^{(\lambda)}(x),$$

must be able to be rendered as

$$a(x)u(x) = \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} a_j \hat{u}_k C_j^{(\lambda)}(x) C_k^{(\lambda)}(x). \quad (2.33)$$

To render this as single sum in terms of  $C_j^{(\lambda)}$  Olver and Townsend (2013) noted that, in a manner that follows Carlitz (1961)

$$C_j^{(\lambda)}(x) C_k^{(\lambda)}(x) = \sum_{s=0}^{\min(j,k)} c_s^\lambda(j, k) C_{j+k-2s}^{(\lambda)}(x),$$

for which the coefficients of  $C_{j+k-2s}^{(\lambda)}(x)$  are

$$c_s^\lambda(j, k) = \frac{j+k+\lambda-2s}{j+k+\lambda-s} \frac{(\lambda)_s (\lambda)_{(j-s)} (\lambda)_{(k-s)}}{s! (j-s)! (k-s)!} \frac{(2\lambda)_{(j+k-s)} (j+k-2s)!}{(\lambda)_{(j+k-s)} (2\lambda)_{j+k-2s}}.$$

Within this  $(\lambda)_k$  is shorthand for the Pochhammer symbol

$$(\lambda)_k = (\lambda+k-1)!/(\lambda-1)!$$

The above result in conjunction with equation (2.33) gives

$$a(x)u(x) = \sum_{j=0}^{\infty} \left( \sum_{k=0}^{\infty} \sum_{s=\max(0,k-j)}^k a_{2s+j-k} c_s^\lambda(k, 2s+j-k) u_k \right) C_j^{(\lambda)}(x),$$

which then gives that the multiplication operator in terms of a  $C^{(\lambda)}$  Gegenbauer series must be

$$\mathcal{M}_\lambda [a]_{j,k} = \sum_{s=\max(0,k-j)}^k a_{2s+j-k} c_s^\lambda(k, 2s+j-k), \text{ for } j, k \geq 0.$$

Focusing now on the discrete approximation

$$a(x) = \sum_{j=0}^m a_j C_j^{(\lambda)}(x) \text{ and } u(x) = \sum_{k=0}^m \hat{u}_k C_k^{(\lambda)}(x)$$

then  $\mathcal{M}_\lambda [a]$  can instead be expressed in terms through the recurrence relationship

$$c_{s+1}^\lambda(j, k+2) = c_s^\lambda(j, k) \frac{j+k+\lambda-s}{j+k+\lambda-s+1} \frac{\lambda+s}{s+1} \frac{j-s}{\lambda+j-s-1} \frac{2\lambda+j+k-s}{\lambda+j+k-s} \frac{k-s+\lambda}{k-s+1}, \quad (2.34)$$

where the starting point to this recurrence process can be found by evaluating

$$c_0^\lambda(j, k) = \prod_{t=0}^{j-1} \frac{\lambda+t}{1+t} \times \prod_{t=0}^{j-1} \frac{k+1+t}{k+\lambda+t}. \quad (2.35)$$

Incorporating all of the above results allows the general discretised linear operator for equation (2.21) to be expressed in Gegenbauer space as

$$\mathcal{L} := \mathcal{M}_N[a^N] \mathcal{D}_N + \sum_{\lambda=1}^{N-1} \mathcal{S}_{N-1} \dots \mathcal{S}_\lambda \mathcal{M}_\lambda[a^\lambda] \mathcal{D}_\lambda + \mathcal{S}_{N-1} \dots \mathcal{S}_0 \mathcal{M}_0[a^0]. \quad (2.36)$$

The solution to the linear system incorporating this operator can be found in  $\mathcal{O}(N)$  time, and gives the solution in Chebyshev space, which can then be converted to real space using a discrete cosine transform. This technique is advantageous, as it allows for super-algebraic convergence for variable coefficient linear differential equations of arbitrary order, while only requiring the solution of an almost banded, sparse and well-conditioned linear system. This is due to the product of all the operators being sparse and banded—with the exception of the first  $n$  rows of an  $n$ -th order differential equation, which are allocated to the boundary banding.

Furthermore, while deriving this technique we did not need to place any restrictions on the form of the boundary operator  $\mathcal{B}$ . In fact, for a boundary operator

$$\mathcal{B}u = \mathbf{c},$$

then  $\mathcal{B}$  is free to be any  $K$  linear boundary conditions, subject to the restriction that  $\mathbf{c} \in \mathcal{C}^K$ . This means that we can implement a range of boundary conditions, from the standard mix of Dirichlet and Neumann conditions, to more complicated conditions such as absorbing boundary conditions; imposing that the solution integrates to a constant value or that it evaluates to a fixed constant value over  $(-1, 1)$ . These can be implemented by taking



$$\begin{aligned}\mathcal{B} &= \begin{pmatrix} T_0(-1) & T_1(-1) & T_2(-1) & T_3(-1) & \cdots \\ T_0(1) & T_1(1) & T_2(1) & T_3(1) & \cdots \end{pmatrix} \\ &= \begin{pmatrix} 1 & -1 & 1 & -1 & \cdots \\ 1 & 1 & 1 & 1 & \cdots \end{pmatrix}\end{aligned}$$

which implements Dirichlet conditions. Neumann conditions can similarly be imposed by

$$\begin{aligned}\mathcal{B} &= \begin{pmatrix} T'_0(-1) & T'_1(-1) & T'_2(-1) & \cdots & T'_k(-1) & \cdots \\ T'_0(1) & T'_1(1) & T'_2(1) & \cdots & T'_k(1) & \cdots \end{pmatrix} \\ &= \begin{pmatrix} 0 & 1 & -4 & \cdots & (-1)^{(k+1)}k^2 & \cdots \\ 0 & 1 & 4 & \cdots & k^2 & \cdots \end{pmatrix}\end{aligned}$$

While Olver and Townsend (2013) mentions that the solution can be conditioned, using the Clenshaw–Curtis quadrature as described in Subsection 2.1.5, to integrate to a constant, we can in fact use the tools of the Gegenbauer method in order to condition the solution so that a product of the solution integrates to a constant value. If  $\mathbf{d}$  is a vector of the Clenshaw–Curtis weights from equation (2.15), then

$$\int_{-1}^1 f(x)u(x)dx = \mathbf{d}\mathcal{M}_0[f(x)]\hat{u}, \quad (2.37)$$

where  $\mathbf{d}\mathcal{M}_0[f(x)] \in \mathbb{R}^{1 \times n}$ . This allows a broader range of conditions to be applied to any linear equation (and in turn, any nonlinear equation solved using the Gegenbauer method as a base). Combining these pieces together gives the matrix operator of the form

$$A \begin{pmatrix} \hat{u}_0 \\ \hat{u}_1 \\ \vdots \\ \hat{u}_{n-1} \end{pmatrix} = \begin{pmatrix} \mathbf{c} \\ \mathcal{P}_{n-K}\mathcal{S}_{n-1}\mathcal{S}_{n-2} \cdots \mathcal{S}_0\mathbf{f} \end{pmatrix} \quad (2.38)$$

where

$$A = \begin{pmatrix} \mathcal{B}\mathcal{P}_n^T \\ \mathcal{P}_{n-K}\mathcal{L} \end{pmatrix}. \quad (2.39)$$

Here  $\mathcal{L}$  is the matrix created by equation (2.36), and  $\mathcal{P}_j \in \mathbb{R}^{n \times n}$  is a projection operator ( $I_j, \mathbf{0}$ ) constructed in terms of the identity matrix  $I_j$ , which has size  $\mathbb{R}^{j \times j}$ , and  $K$  is the number of boundary conditions being applied. The resultant matrix operators from this system are sparse, almost-banded and can be solved in  $\mathcal{O}(m^2n)$  operations, and requiring  $\mathcal{O}(mn)$  storage, where  $n$  is the number of Chebyshev modes required to resolve the solution, and  $m = 1 + (\text{number of non-zero superdiagonals}) + (\text{number of non-zero subdiagonals})$ . The number of non-zero diagonals excludes the  $K$  dense rows corresponding to the boundary conditions.

As an example of this, let us consider a linear, variable coefficient singular perturbation problem for  $y(x)$  of the form

$$\left. \begin{aligned} \epsilon y_{xx} + (x^2 - 0.5^2)y &= 0, \\ y(-1) &= 1, \\ y(1) &= 2. \end{aligned} \right\} \quad (2.40)$$

The construction of a solution to this equation using standard perturbation schemes is not viable, due to its lack of  $y_x$  terms. However, WKB analysis does show that the solution should exhibit two highly oscillatory regions. These regions should be contained within  $x \in [-1, -0.5]$  and  $[0.5, 1]$  and oscillate with a frequency proportional to  $1/\sqrt{\epsilon}$ . To explore the behaviour of the Olver and Townsend (2013) solver, this equation was discretised and then solved on a Gegenbauer basis, with the structure for  $\epsilon = 10^{-6}$  shown in Figure 2.1a.

The solution presented herein accurately aligns with the dynamics predicted through WKB analysis. While there is no known exact solution for this problem to analyse the accuracy of this numerical solution, the convergence of the technique can still be assessed by examining the Cauchy error, as seen in Figure 2.1b. This figure presents the  $L_2$  norm (denoted by  $\|\cdot\|_2$ ) of the difference between  $\hat{y}_m$  and  $\hat{y}_n$ , for  $m = \lceil 1.01n \rceil$ . To make the comparison between solutions at different resolutions, the comparison is made between the coefficients in Gegenbauer space, denoted by  $\hat{y}$ , rather than real space, due to the difference in grid point spacing as the spatial resolution increases. For  $n$  between 376 and 512 the error

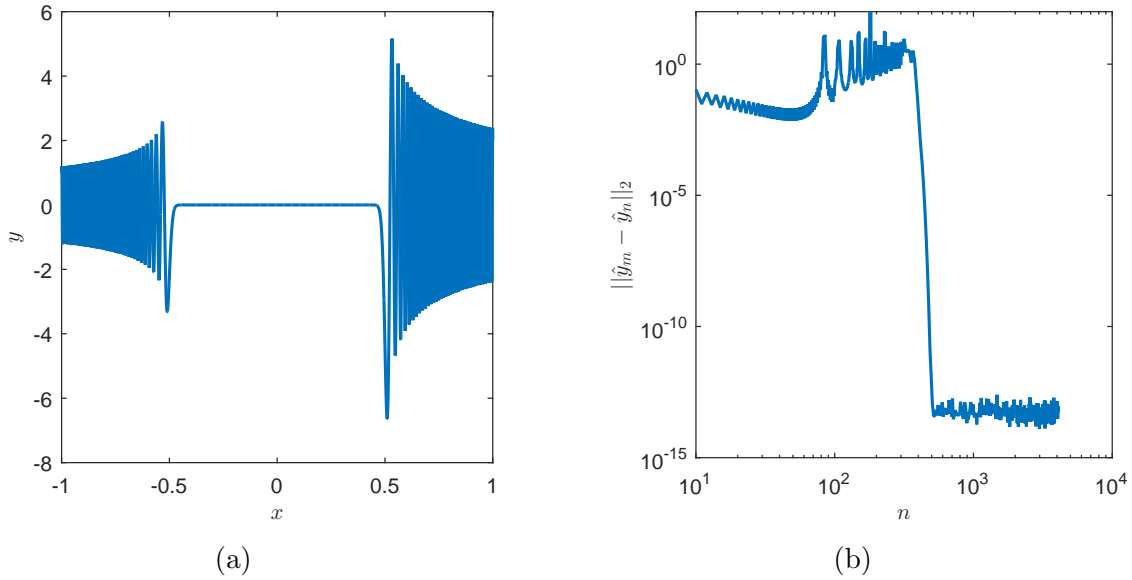


Figure 2.1: Numerical solution of equation (2.40) using the Gegenbauer method for  $\epsilon = 10^{-6}$ , as well as its convergence behaviour with  $n$ , where  $m = \lceil 1.01n \rceil$ .

decreases in a spectral manner, to the point where the Cauchy error confirms that the solution has converged, up to machine precision, and is independent of the number of coefficients for Gegenbauer polynomial expansions of order  $n \geq 512$ . The sparsity of the linear matrix operator for a solution constructed for  $n = 2^{12}$  Gegenbauer coefficients can be seen in Figure 2.2. The entire matrix system has been discretised using five diagonal elements, as well as two rows at the top of the matrix that account for the effect of the boundary conditions. For this discretisation, the number of nonzero elements corresponds to a sparse matrix where only 0.17% of the matrix elements are filled. The banded nature of this matrix, and the low fill-in density it exhibits has significant implications for the computational efficiency of the scheme, due to the reduced interdependence of the equations being solved.

### 2.3 Solving nonlinear differential equations

As useful as these solvers are, they are inherently limited in the types of problems they can approach. To consider the broader set of nonlinear differential equations, many tools can be considered, with Newton iteration being far and away the most commonly employed.

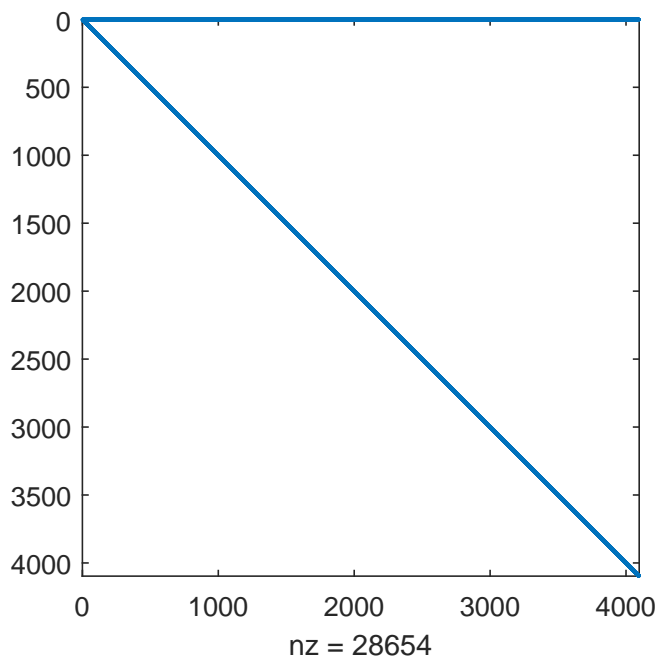


Figure 2.2: Sparsity structure of the second-order linear differential Gegenbauer space operator for equation (2.40), where the non-zero elements are coloured.

Newton iteration uses Taylor series expansion about the zeros of a function to construct an iterative scheme. For a nonlinear problem of the form

$$f(\mathbf{x}) = \mathbf{0} \tag{2.41}$$

then the tangent  $g(\mathbf{x})$  to the function  $f(\mathbf{x})$ , at the point  $\mathbf{x} = \mathbf{x}_n$ , must be

$$g(\mathbf{x}) = f(\mathbf{x}_n) + J(\mathbf{x}_n)(\mathbf{x} - \mathbf{x}_n),$$

which is expressed in terms of the Jacobian matrix  $J(\mathbf{x}_n)$ . To aide in the search for a solution that satisfies  $f(\mathbf{x}) = \mathbf{0}$ , we will impose that  $g(\mathbf{x})$  is similarly equal to zero so that

$$0 = f(\mathbf{x}_n) + J(\mathbf{x}_n)(\mathbf{x} - \mathbf{x}_n).$$

While  $\mathbf{x}$ , the solution to this equation, is not guaranteed to also solve equation (2.41), we can still propose that the above equation is equivalent to

$$0 = f(\mathbf{x}_n) + J(\mathbf{x}_n)(\mathbf{x}_{n+1} - \mathbf{x}_n)$$

which in turn gives rise to the iterative scheme

$$\mathbf{x}_{n+1} = \mathbf{x}_n - J(\mathbf{x}_n)^{-1} f(\mathbf{x}_n). \quad (2.42)$$

This approach can be discretised in terms of Gegenbauer technique of Subsection 2.2.2 by considering the more general case of infinite-dimensional Newton iteration (Birkisson and Driscoll, 2012, Driscoll et al., 2008). Following Olver and Townsend (2013), and considering a nonlinear problem of the form

$$\left. \begin{aligned} \mathcal{L}u + g(u) &= f \\ \mathcal{B}u &= \mathbf{0}, \end{aligned} \right\} \quad (2.43)$$

where  $\mathcal{L}$  and  $g(u)$  are strictly linear and nonlinear functions respectively, then employing infinite-dimensional Newton iteration results in the iterative scheme

$$u_{k+1} = u_k + \begin{pmatrix} \mathcal{B} \\ \mathcal{L} - g'(u_k) \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{0} \\ \mathcal{L}u_k + g(u_k) - f \end{pmatrix}. \quad (2.44)$$

However, since the matrix being inverted is a function of the solution itself, the bandedness of the Gegenbauer multiplication operator is not preserved through this process. As a result, the computational cost for solving nonlinear equations using this technique does not display the same advantages as the Gegenbauer method for linear differential equations, negating the primary advantage of these basis functions. Furthermore, our initial testing encountered significant convergence control issues when implementing infinite-dimensional Newton iteration for nonlinear problems, further underscoring the need to consider alternate techniques.

The poor scaling of infinite-dimensional Newton iteration can be somewhat negated by considering inexact Newton's methods, which relax the requirement to exactly solve the matrix equations in equation (2.42). In doing so, the speed of constructing an approximate solution can be enhanced by replacing the matrix solver with an approximate method, such as Conjugate Gradient (Concus et al., 1976), or Newton-Krylov iteration. By doing so, the solution to the iterative step can be thought to be  $\hat{x}_{n+1}$ , where  $\hat{x}_{n+1} \approx x_{n+1}$ . Any inaccuracies in the calculation  $\hat{x}_{n+1}$  are immaterial, as they will be resolved through continuing the iterative process. Through this process nonlinear systems of differential equations can be solved with a lower computational cost than would be required for solving

the system exactly.

Another alternative to Newton's method is nonlinear relaxation, which involves decomposing the nonlinear equation into a series of smaller sub-problems. If each of these can be solved independently in an iterative loop that employs the previous iterations results, then a solution to the full nonlinear equation may be able to be found. Two commonly employed relaxation based techniques are the Jacobi–Newton and Seidel–Newton algorithms (Ortega and Rheinbolt, 1970).

For a nonlinear differential equation of the form  $f(\mathbf{x}) = \mathbf{0}$ , Jacobi–Newton attempts to construct a solution by solving

$$f_i(x_1^m, \dots, x_{i-1}^m, x_i^{m+1}, x_{i+1}^m, \dots, x_n^m) = 0$$

for  $x_i^{m+1}$ . Here the subscript denotes the position in space and the superscript denotes the iteration number. These equations are solved using Newton's method. In a similar fashion, Seidel–Newton iteration solves the nonlinear differential equation by seeking solutions to

$$f_i(x_1^{m+1}, \dots, x_{i-1}^{m+1}, x_i^{m+1}, x_{i+1}^m, \dots, x_n^m) = 0$$

for  $x_i^{m+1}$  using Newton's method. The fundamental difference between these two methods is that Seidel–Newton uses the  $(m + 1)$ -st iteration at each point if it has been previously calculated, otherwise the  $m$ -th iteration is used.

Another approach for solving steady differential equations is to construct an analogous unsteady differential equation, that should approach the solution in the long time limit. The resulting unsteady nonlinear equations can then be solved using a number of iterative methods, including Runge–Kutta and its variants (Canuto et al., 1988, Hairer and Wanner, 1996); split–step Fourier methods (Atabakan, 1974, Yoshida, 1990); exponential integrator based techniques like Exponential Time Differencing methods (Cox and Matthews, 2002) and the Accelerated Imaginary–Time Evolution method (Yang and Lakoba, 2007). The latter of these involves manipulating the equation so that time occurs in the set of imaginary numbers, and where the iteration is constructed using a preconditioning system that enhances convergence. While all of these methods are able to construct solutions to unsteady nonlinear systems, there is no guarantee that the initial conditions will converge

upon a steady solution. Furthermore, the computational cost of the iterative scheme can be significantly higher than directly solving the ODE, due to the amount of iterative steps required to resolve the time invariant solutions.

The approaches outlined above—Newton methods, inexact Newton methods, nonlinear relaxation and time stepping techniques—represent the most commonly used approaches for constructing numerical solutions of nonlinear differential equations. However, all of these techniques share similar limitations, in that their representations in terms of dense matrices results in systems that require time commitments that scale nonlinearly with respect to the grid resolution employed for the scheme. This in turn introduces an unfavourable interplay between numerical error and the computational time required to solve a system to within a desired tolerance.

## 2.4 Numerical Continuation

Solutions to equations, especially nonlinear equations, often do not exist in isolation, but rather families of solutions can exist, tied together within some parameter space. As such, it is important to consider numerical tools for exploring these families of solutions. For nonlinear problems, the approaches to explore parameter dependence broadly fall under the categories of continuation, embedding and homotopy problems. In a most general sense, these techniques construct continuous deformations between different systems of equations. Continuation methods can be used to extend results obtained at one position in parameter space—where the calculations may be relatively easy—to a position where the solution is much harder to construct. Continuation methods can also be used to compute and contrast solutions of equations where the boundary conditions change (Kawahara et al., 2012).

Mathematically, continuation techniques can be traced back to Poincare (1881), and in their most general form involve a mapping  $\mathbf{F} : \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}^n$  described by the equation

$$\mathbf{F}(\mathbf{x}, \lambda) = \mathbf{0} \tag{2.45}$$

in terms of  $\mathbf{x} \in \mathbb{R}^n$  and a scalar  $\lambda \in \mathbb{R}$ , which defines the continuation problem. Geometrically, this equation defines a one-dimensional curve in  $\mathbb{R}^{n+1}$  subject to the properties of the equation  $\mathbf{F}$ , which, in this context, can be thought of as a mapping. Continuation techniques then seek to follow the one-dimensional curve, by considering that if  $(\mathbf{x}_0, \lambda_0)$  is a regular solution of  $\mathbf{F}(\mathbf{x}, \lambda) = \mathbf{0}$ , then it follows that in the neighbourhood of  $\mathbf{x}_0$

there should be a family of solutions that can be parametrised by  $\mathbf{x}(s)$ —where  $s$  is an arbitrary parameter—subject to  $\mathbf{x}(0) = \mathbf{x}_0$ . If  $\mathbf{x}(s)$  exists for  $s \neq 0$ , then a range of solutions in parameter space should be able to be constructed. Continuation then is the process of gradually deforming the solutions of the problems in parameter space, from a straightforward problem to something more formidable by moving in small parameter increments.

Numerical continuation is a family of numerical algorithms for generating the solution curve, with natural continuation and pseudo-arclength continuation being popular examples from a group of methods known as predictor–corrector methods (Kuznetsov, 1998). These techniques begin by predicting a new point on the solution curve, before iteratively refining it until a solution to equation (2.45) is found, if it exists. While these techniques are somewhat parallelizable, they are fundamentally sequential techniques, with previously calculated points required to predict the next solution point on the curve in parameter space. These methods are heavily dependent upon adaptive routines for adjusting the step size between known points on the solution curve and the next predicted point. Adaptive step–sizing algorithms are often heuristics based upon the rate of convergence (or lack thereof) of the technique at the previous prediction step and the amount of curvature in the solution in parameter space (Allgower and Georg, 2003). When the solution fails to converge, the step size is shortened and a new predictor step is constructed. This, however, gives rise to one of the fundamental problems of numerical continuation: there will be multiple steps where calculation time is wasted on solutions for which the step size is too large for convergence. Beyond this, there is also an interplay between step size and the rate of convergence, which is difficult (if not impossible) to optimise.

To further explore the idea of continuation, let us re-frame equation (2.45) as

$$\mathbf{F}(\mathbf{z}) = \mathbf{0}, \text{ where } \mathbf{z} = (\mathbf{x}, \lambda) \in \mathbb{R}^{n+1}, \quad (2.46)$$

where we have concatenated the  $n$ -length vector  $\mathbf{x}$  and the scalar  $\lambda$  that describes our position in parameter space into a  $\mathbb{R}^{n+1}$  vector  $\mathbf{z}$ , where  $\lambda$  will exist over some bounded domain  $[\lambda_{\min}, \lambda_{\max}] \subseteq \mathbb{R}$ . The above equation will have corresponding Jacobian derivatives, which are matrices of the appropriate size that take the form

$$\mathbf{F}_\lambda \equiv \frac{\partial \mathbf{F}}{\partial \lambda} \in \mathbb{R}^{n+1}, \quad \mathbf{F}_\mathbf{x} \equiv \frac{\partial \mathbf{F}}{\partial \mathbf{x}} \in \mathbb{R}^{n \times n}, \quad \text{and} \quad \mathbf{F}_\mathbf{z} \equiv \frac{\partial \mathbf{F}}{\partial \mathbf{z}} \in \mathbb{R}^{n \times (n+1)}. \quad (2.47)$$



Then, if we assume that we have a known solution  $\mathbf{x}$  at  $\lambda = \lambda_{\min}$ , we will then search for a vector  $\mathbf{x}^*$  at  $\lambda^* = \lambda_{\max}$  that satisfies

$$\mathbf{F}(\mathbf{x}^*, \lambda^*) = \mathbf{0},$$

by deforming from the solution at  $\lambda = \lambda_{\min}$ . The most basic form of continuation is that, given a solution  $(\mathbf{x}, \lambda) \in \mathbb{R}^{n+1}$  then a new point on the curve  $(\mathbf{x}_1, \lambda + h)$  can be found by incrementing  $\lambda$  by a small, strictly positive  $h$  and then iteratively solving the  $n$  equations for

$$\mathbf{F}(\mathbf{x}_1, \lambda + h) = \mathbf{0}$$

for the unknowns  $\mathbf{x}_1 \in \mathbb{R}^n$  using any suitable nonlinear technique. This approach of continuation is known as natural parameter continuation (Govaerts, 2000).

Natural parameter continuation is an effective tool for basic continuation techniques, however it breaks down when there is a fold point—a point in which there is some degree of inflection in the curve in  $\mathbb{R}^{n+1}$  parameter space. Physically this manifests itself as a local or global maxima of  $\lambda$ , where the solutions are no longer one-to-one. Dickson et al. (2007) and others have shown that this occurs when the Jacobian matrix  $\mathbf{F}_{\mathbf{x}}(\mathbf{x}, \lambda)$  becomes singular. In the neighbourhood of a fold point, any increasing perturbation of  $\lambda$  will yield an inconsistent system of nonlinear equations that do not contain a solution, irrespective of the magnitude of the step size perturbation  $h$ .

The psuedo-arclength continuation of Keller (1977) is one of the most common approaches for traversing fold point singularities, which involves parametrising  $\mathbf{x}$  and  $\lambda$  to be functions of the length along the curve  $s$ . Psuedo-arclength continuation then uses the unit-length tangent  $\mathbf{T} \in \mathbb{R}^{n+1}$  to the curve for predicting the point at which the next solution may lie. The tangent direction  $\mathbf{T}$  can be calculated by finding the null vector of the Jacobian matrix  $\mathbf{F}_{\mathbf{z}}$ . However, to this point the system is under-determined, and as such a closure must be constructed in order to solve this system. This can be done by imposing that  $(\mathbf{x}, \lambda)$  must exist in the hyperplane orthogonal to the tangent direction  $\mathbf{T}$  at a distance  $h$  from the previously calculated solution.

To expand upon this, if we have a solution  $(\mathbf{x}_0, \lambda_0)$  of  $\mathbf{F}(\mathbf{x}, \lambda) = \mathbf{0}$ , and a direction vector along the arc length curve  $(\dot{\mathbf{x}}_0, \dot{\lambda}_0)$ , then we can search for new solutions

$$\left. \begin{aligned} \mathbf{F}(\mathbf{x}_1, \lambda_1) &= \mathbf{0} \\ (\mathbf{x}_1 - \mathbf{x}_0)^* \dot{\mathbf{x}}_0 + (\lambda_1 - \lambda_0) \dot{\lambda}_0 - \Delta s &= 0 \end{aligned} \right\} \quad (2.48)$$

subject to the condition that  $\|\dot{\mathbf{x}}_1\|_2 + \dot{\lambda}_1^2 = 1$  and a given step size  $\Delta s$ . The pseudo-arc length continuation equations are then typically solved by employing Newton iteration, which involves solving the iterative matrix equation

$$\begin{pmatrix} \mathbf{F}_{\mathbf{x}_1}^{(\nu)} & \mathbf{F}_{\lambda_1}^{(\nu)} \\ \dot{\mathbf{x}}_0^* & \dot{\lambda}_0 \end{pmatrix} \begin{pmatrix} \Delta \mathbf{x}_1^{(\nu)} \\ \Delta \lambda_1^{(\nu)} \end{pmatrix} = - \begin{pmatrix} \mathbf{F}(\mathbf{x}_1^{(\nu)}, \lambda_1^{(\nu)}) \\ (\mathbf{x}_1^{(\nu)} - \mathbf{x}_0)^* \dot{\mathbf{x}}_0 + (\lambda_1^{(\nu)} - \lambda_0) \dot{\lambda}_0 - \Delta s \end{pmatrix}. \quad (2.49)$$

The new direction vector is calculated through solving

$$\begin{pmatrix} \mathbf{F}_{\mathbf{x}}^1 & \mathbf{F}_{\lambda}^1 \\ \dot{\mathbf{x}}_0^* & \dot{\lambda}_0 \end{pmatrix} \begin{pmatrix} \dot{\mathbf{x}}_1 \\ \dot{\lambda}_1 \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ 1 \end{pmatrix} \quad (2.50)$$

subject to the condition that  $\|\dot{\mathbf{x}}_1\|^2 + \dot{\lambda}_1^2 = 1$ , which can be satisfied by rescaling  $(\mathbf{x}_1, \lambda_1)$ . Following this approach it is impossible to disentangle the numerical continuation process from the Newton iteration, and necessitates that each step of the continuation process requires a significant amount of unique matrix inversions.

Another approach for continuation methods is to construct a homotopy between solutions at two points in parameter space, with the intent to deform from one known solution onto a new, unknown solution (Borisevich and Schullerus, 2012, Gunji et al., 2003). In order to link a known solution  $F(\mathbf{u})$  at one position in parameter space, to an as yet undetermined solution  $G(\mathbf{u})$  in the neighbourhood of  $F$ , then a homotopy  $\mathcal{H} : \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}^n$  can be defined in terms of  $\mathbf{u} = \{\mathbf{x}, \lambda\} \in \mathbb{R}^{n+1}$  where

$$H(\mathbf{u}, 1) = G(\mathbf{u}), \quad H(\mathbf{u}, 0) = F(\mathbf{u}).$$

Typically, homotopy continuation methods are defined by

$$\mathcal{H}(\mathbf{u}, q) = qG(\mathbf{u}) + (1 - q)F(\mathbf{u}),$$

with methods attempting to trace a smooth curve from  $(\mathbf{u}, q) = (\mathbf{u}_0, 0)$  to  $(\mathbf{u}, 1)$ , where  $\mathbf{u}_0$  is a known solution at a defined point in parameter space. This of course leads to the

questions of: if the homotopy curve exists, is it finite and smooth, and how it can be found? Allgower and Georg (2003) outlines that as a consequence of the implicit function theorem, if the Jacobian matrix  $H'(\mathbf{u})$  describing the matrix of partial derivatives of the homotopy has maximal rank, then a smooth, finite curve can be constructed.

Another approach for the arclength continuation is to consider a secant based method. If we know some solution  $\mathbf{u}_0 = (\mathbf{x}_0, \lambda_0)$  to the differential equation, then, we will assume that our differential solver will converge upon a solution  $\mathbf{u}_1 = (\mathbf{x}_1, \lambda_1)$  if it is given the input  $\mathbf{u}^* = (\mathbf{x}_0, \lambda_1)$  for some  $\lambda_1$  where  $|\lambda_1 - \lambda_0|$  is less than a small parameter  $\epsilon$ . If such a solution can be found, then we can construct a derivative  $\dot{\mathbf{u}}_1$  using finite differencing, which can then be used to approximate the path along the arc, so that  $\mathbf{u}_2(s) = \mathbf{u}_1(s) + \Delta s \dot{\mathbf{u}}_1$ . This can be coupled with the condition that  $\|\dot{\mathbf{u}}_n\|^2 = 1$ , as is the case for the traditional pseudo-arclength continuation. More generally,

$$\left. \begin{aligned} \dot{\mathbf{u}}_n &= \frac{\mathbf{u}_n - \mathbf{u}_{n-1}}{\Delta s_n} \\ \|\dot{\mathbf{u}}_n\|^2 &= 1 \\ \mathbf{u}_{n+1}^* &= \mathbf{u}_n \pm \Delta s_{n+1} \dot{\mathbf{u}}_n \end{aligned} \right\} \quad (2.51)$$

where  $\mathbf{u}^* = (\mathbf{x}^*, \lambda^*)$  gives the approximate solution  $\mathbf{u}^*$  at the parameter space position  $\lambda^*$ . From here a new solution along the branch in parameter space can be constructed as long as  $\mathbf{x}^*$  is sufficiently close to the true solution at  $\lambda^*$  for our nonlinear solver to solve  $\mathcal{N}(x, \lambda^*)$  given our initial predicted solution.

Of course, this leads to the question of how we can ensure that  $\mathbf{u}^*$  is close enough to our true solution that the desired convergence will occur. This can be done by controlling the step size  $\Delta s_n$ , and making it inversely proportional to the norm of the residual at the previous step, and the gradient of the curve with respect to the parameter space variable  $\lambda$ .

To this point, the continuation has been considered in terms of  $(\mathbf{x}, \lambda) \in \mathbb{R}^{n+1}$ , where  $\mathbf{x}$  is the solution vector at a position  $\lambda$  in parameter space. However, continuation for a solutions  $\mathbf{x} \in \mathbb{R}^n$  can be expressed in terms of a feature of  $\mathbf{x}$ . Here the feature,  $g(\mathbf{x})$  can be a single point of the vector, or the norm of the vector, or any other metric. This allows for the continuation to be expressed in terms of a  $\mathbb{R}^{m+1}$  vector, where  $m \ll n$ —simplifying the process of solving the system.

## 2.5 Domain mappings

Both the linear and nonlinear solvers presented to this point share one fundamental limitation, in that the numerical domain is limited to the region over which the basis functions are defined. As both the Chebyshev and Gegenbauer basis functions are both defined upon  $x \in [-1, 1]$ , the range of problems that be directly approached using these techniques is inherently limited.

Several methods have been developed in order to expand the range of applicability of spectral methods. These can broadly be categorised as being part of one of two distinct approaches: domain truncation and approximation using orthogonal systems (Boyd, 1982, 2001, Canuto et al., 2006, Shen and Wang, 2009). Orthogonal systems, which can include the sets of Laguerre, Hermite and Jacobi polynomials, typically exist on unbounded domains, and as such these polynomials can be used as the basis of a numerical method for problems that exist on these domains (Boyd, 1980b, Christov, 1982, Guo et al., 2002). While using orthogonal systems has its own unique advantages, the work contained within this thesis will be focussed on domain truncation, as it is a more flexible approach that can be leveraged for a large number of problem types, and can be flexibly integrated into numerical schemes.

Domain truncation is a common and effective strategy for dealing with unbounded domains and involves taking an unbounded domain and excising the component of the domain outside a defined, finite region. This truncated problem can then be solved either directly or with augmented artificial or transparent boundary conditions to approximate the truncated domain component.

In their most general form, a domain mapping can be expressed in the form

$$\left. \begin{aligned}
 x &= g(y; s) && s > 0 \text{ and} \\
 y \in I := (-1, 1), x \in \Lambda := (a, b) & \text{ for } a, b \in [-\infty, \infty], \\
 \frac{dx}{dy} &= g'(y; s) > 0 && s > 0, y \in I, \\
 g(-1; s) = a, g(1, s) = b, &&& \text{for } \Lambda = (a, b).
 \end{aligned} \right\} \quad (2.52)$$

This imposes that the mapping  $g$  must be a one-to-one transform, subject to the positive scaling factor  $s$ , which can be used to control the length of the resolved domain, and the spacing of the collocation points. Furthermore, it must be true that the mapping is explicitly invertible, where the inverse mapping is

$$y = g^{-1}(x; s) := h(x, s), \quad x \in \Lambda, y \in I, s > 0.$$

For the purposes of mapping from a domain, that has either been truncated to or exists for  $x \in (a, b)$ , the most commonly used mapping is the linear mapping

$$x = (b - a) \frac{y + 1}{2} + a, \quad y = \frac{2x - b - a}{b - a}.$$

For mappings on the semi-infinite line, where  $x \in \Lambda = (0, \infty)$  to  $y \in I = (-1, 1)$ , subject to  $s > 0$ , the most common mappings are:

- Algebraic mapping, where:

$$x = \frac{s(1 + y)}{1 - y}, \quad y = \frac{x - s}{x + s}.$$

- Logarithmic mapping:

$$x = s \operatorname{arctanh} \left( \frac{y + 1}{2} \right) = \frac{s}{2} \log_e \frac{3 + y}{1 - y}, \quad y = -1 + 2 \tanh \left( \frac{x}{s} \right).$$

- Exponential mapping:

$$x = \sinh \left( \frac{s(1 + y)}{2} \right), \quad y = \frac{2}{s} \log_e \left( x + \sqrt{x^2 + 1} \right) - 1,$$

where  $x$  exists in the truncated domain  $x \in (0, L_s)$  and  $L_s = \sinh(s)$ .

Similar mappings exist for the infinite line  $x \in \Lambda = (-\infty, \infty)$ , where for  $s > 0$  there exists:

- Algebraic mapping:

$$x = \frac{sy}{\sqrt{1 - y^2}}, \quad y = \frac{x}{\sqrt{x^2 + s^2}}.$$

- Logarithmic mapping:

$$x = s \operatorname{arctanh}(y) = \frac{s}{2} \log \left( \frac{1 + y}{1 - y} \right), \quad y = \tanh \left( \frac{x}{s} \right)$$

- Exponential mapping:

$$x = \sinh(sy), \quad y = \frac{1}{s} \log_e \left( x + \sqrt{x^2 + 1} \right),$$

where, once again,  $x$  exists in the truncated domain  $x \in (-L_s, L_s)$ , where  $L_s = \sinh(s)$ .

The feature that distinguishes these mappings is that as  $y \rightarrow \pm 1$ ,  $x$  varies either algebraically, logarithmically or exponentially for the corresponding mapping. Thus, by carefully selecting the mapping, the distribution of collocation points can be changed to increase the local point density within the centre or boundary regions of the domain. Other options include constructing a mapping which redistributes the point distribution to focus up regions exhibiting large local rates of change or oscillations, or to follow the same distribution as the underlying collocation functions. This in turn can affect the performance of the numerical scheme, as increasing the local point density of the collocation points increases the local numerical resolution, which can affect both the numerical error of the scheme, and confer the ability to resolve high frequency oscillations—such as those shown in Figure 2.1a—without increasing the overall number of points across the domain. As the Chebyshev–Gauss collocation points are distributed so that the frequency of points increases as  $|x| \rightarrow 1$ , it can be advantageous to employ a domain mapping to convert the point distribution to one where points are evenly distributed across the domain, or are concentrated about  $x = 0$ .

It must be noted that in all cases except linear mappings, introducing these transforms to an ODE with constant coefficients will result in a variable coefficient equation subject to potential boundary singularities. These singularities pose a significant numerical hurdle, due to the difficulty of resolving the range of scales within the boundary region. Such problems pose a particular challenge to techniques built upon Chebyshev collocation matrices, which served as further motivation for the development of our novel numerical method in terms of the Gegenbauer basis functions, as will be outlined in Chapter 3.

## 2.6 Discussion

Over the course of this chapter a review of the current state of knowledge for spectral methods has been presented. Due to the motivating interest of solving nonlinear boundary value problems, a particular focus has been placed upon problems defined in terms of both the Chebyshev and Gegenbauer spaces. The latter of which is particularly interesting, as it

encompasses both the Legendre and Chebyshev spaces, while also introducing some advantageous properties. As the Gegenbauer polynomials are relatively unknown as a spectral basis function, a particular focus was placed upon establishing the process for conducting elementary operations within this space, including addition, subtraction, multiplication, differentiation and integration. One of the novel results contained within this chapter is the application of Clenshaw–Curtis quadratures to Gegenbauer methods, as shown in Subsection 2.1.5 and Subsection 2.2.2.

These elementary operations were then extended in order to construct numerical discretisations for solving linear boundary value problems, using matrix methods defined in terms of both the Chebyshev and Gegenbauer polynomials. While solving boundary value problems in terms of these polynomials is intrinsically more complicated than the more common approaches using local basis functions, spectral methods have uniquely advantageous convergence properties. The most notable of these is that increases in the resolution of the numerical discretisation result in exponential decreases in the error for spectral methods, in contrast to the linear decrease in error for local methods.

While there is a wealth of literature considering the construction of numerical solutions for boundary value problems in terms of the Chebyshev basis functions, all the employed approaches share the same limitations, in that the resulting matrix equations are dense, and have large conditions numbers that scale poorly with the resolution. As the condition number grows, so too does the error that results from solving the resulting matrix equations, eventually leading to the systems becoming ill-conditioned. Furthermore, as a consequence of the density of the Chebyshev matrix operators, as the resolution of the numerical discretisation increases so too does the computational cost of both storing the matrices in memory, and for solving the resulting matrix equations.

These properties compare unfavourably to those exhibited by the Gegenbauer methods, broadly based upon the work of Olver and Townsend (2013), which involve constructing linear operators in terms of sparse matrix operators with low fill-in densities. This property alone has significant implications to the computational cost of solving these systems, and their ability to scale towards problems which require high grid resolutions. The sparsity of these Gegenbauer matrices, and the convergence properties of this scheme with respect to the resolution of the numerical discretisation were explored by considering the solution of a

linear, variable coefficient singular perturbation problem in one–dimension. Within Chapter 3 the properties of the Gegenbauer method will be further elucidated, in the context of nonlinear problems, with a specific focus upon the techniques computational advantages.

As the ultimate aim of this work is to consider constructing solutions to nonlinear, variable coefficient boundary value problems, a range of extensions of the aforementioned linear solvers were introduced. These included a range of nonlinear solvers, alongside methods for constructing domain mappings and performing numerical continuation. In aggregate, these tools greatly expand the scope of problems that can be approached numerically within the framework of spectral methods.

However, all the nonlinear solvers discussed within this chapter result in dense matrix operators, that must be repeatedly solved in order to construct solutions for nonlinear boundary value problems. As such, the techniques contained within this chapter will be used as the basis for developing a new, spectrally accurate approach for solving nonlinear boundary value problems in terms of sparse matrix operators, alongside solving innovative approaches for approaching the associated continuation problems.



# Chapter 3

## Homotopy Analysis Method

As a technique for solving nonlinear equations, the Homotopy Analysis Method (HAM) is built around the idea of constructing a homotopy—an idea from differential geometry—that defines a smooth, continuous deformation from one equation onto another. If this homotopy can be found between the two functions, it follows that the same homotopy should also hold for the solutions of the two equations, and thus allowing one known solution to an equation to be deformed onto the unknown solution of another equation. This process was first proposed by Liao (1992), and can be considered an extension of the Lyapunov and Homotopy Perturbation methods.

The process of constructing this homotopy involves partitioning a nonlinear problem into an infinite sequence of linear sub-problems, which can then be solved sequentially. One crucial feature is that the technique introduces very few limitations upon the linear sub-problems being employed in the solution process. Introducing an additional convergence control parameter allows the technique to be applied to problems that have either been considered intractable, or that have been limited to narrow regions of convergence. From a fluid mechanics perspective these enhanced convergence control properties have been used to further explore Blasius boundary layers (Liao and Campo, 2002), Magnetohydrodynamic flow (Hayat and Sajid, 2007), steady-state resonant progressive waves (Xu et al., 2012) and Von Karman swirling flow in the presence of viscosity (Yang and Liao, 2006).

Considering this technique in terms of the nonlinear differential operator  $\mathcal{N}$  where

$$\mathcal{N}[u(\mathbf{x})] = \psi(x), \tag{3.1}$$

the solution to this can be constructed by deforming from a solution of an arbitrary linear differential equation  $\mathcal{L}[u(\mathbf{x})]$  through the homotopy

$$\mathcal{H}[\phi(x); q] = (1 - q)\mathcal{L}[\phi(\mathbf{x}; q) - u_0(\mathbf{x}; q)] - q\hbar H(\mathbf{x}) (\mathcal{N}[\phi(\mathbf{x}; q)] - \psi(x)), \quad (3.2)$$

A specific feature of the HAM is the introduction of the auxiliary convergence control parameters  $\hbar$  and  $H(\mathbf{x})$ , although the latter is almost uniformly set to 1 across all implementations. The homotopy itself is controlled by the deformation parameter  $q \in [0, 1]$  and  $\phi(\mathbf{x}; q)$  is a representation of the solution across  $q$ , corresponding to the deformation from a trial function  $u_0(\mathbf{x}; q)$  to  $u(\mathbf{x})$ . The function  $\phi(\mathbf{x}; q)$  can be considered as a Maclaurin series in terms of  $q$

$$\phi(\mathbf{x}; q) = \phi(\mathbf{x}; 0) + \sum_{m=1}^{\infty} \frac{1}{m!} \frac{\partial^m \phi(\mathbf{x}; 0)}{\partial q^m} q^m \quad (3.3)$$

$$= U_0(\mathbf{x}) + \sum_{m=1}^{\infty} U_m(\mathbf{x}) q^m, \quad (3.4)$$

where

$$U_m(\mathbf{x}) = \frac{1}{m!} \frac{\partial^m \phi(\mathbf{x}; 0)}{\partial q^m}.$$

By imposing, without loss of generality, that  $\mathcal{H} = 0$  then the behaviour of  $\phi$  can be considered further in the context of the limits of  $q$  within equation (3.2). Therefore

$$\left. \begin{aligned} \mathcal{H}[\phi(\mathbf{x}; 0); 0] &= \mathcal{L}[\phi(\mathbf{x}; 0) - u_0(\mathbf{x})], & \phi(\mathbf{x}; 0) &= u_0(\mathbf{x}), \\ \mathcal{H}[\phi(\mathbf{x}; 1); 1] &= \mathcal{N}[\phi(\mathbf{x}; 1)] - \psi(x), & \phi(\mathbf{x}; 1) &= u(\mathbf{x}). \end{aligned} \right\} \quad (3.5)$$

This corresponds to deforming  $\phi(\mathbf{x}; q)$  from an initial guess of  $u_0(\mathbf{x})$  onto the exact solution  $u(\mathbf{x})$ . This is contingent upon  $\phi(\mathbf{x}; q)$  being analytic at  $q = 0$ , and more broadly existing over  $q \in [0, 1]$ .

The existence and convergence of the scheme outlined in equation (3.2) is determined by the choice of the two convergence control parameters  $\hbar$  and  $H(\mathbf{x})$  and the auxiliary linear operator  $\mathcal{L}$ . The latter operator can be defined almost arbitrarily, although our testing has indicated that the most efficient choice of  $\mathcal{L}$  resembles the linearisation of  $\mathcal{N}[u(\mathbf{x})]$ —an idea

that will be explored later in this chapter.

Subject to the choice of  $\mathcal{L}$ , the homotopy formulation of equation (3.2) can be employed to solve an equation of the form equation (3.1) by substituting  $\phi$  from equation (3.3) into the homotopy equation equation (3.2), differentiating  $m$ -times with respect to  $q$  and then setting  $q = 0$ . This in turn results in the sequence of serially dependent equations

$$\left. \begin{aligned} \mathcal{L}[U_m(\mathbf{x}) - \chi_m U_{m-1}(\mathbf{x})] &= \hbar R_m, \\ R_m &= \frac{1}{(m-1)!} \left\{ \frac{\partial^{m-1}}{\partial q^{m-1}} \mathcal{N} \left[ \sum_{n=0}^{m-1} U_n(\mathbf{x}) q^n \right] \right\} \Big|_{q=0} - (1 - \chi_m) \psi(x), \\ \chi_m &= \begin{cases} 0 & \text{if } m \leq 1, \\ 1 & \text{if } m \geq 2. \end{cases} \end{aligned} \right\} \quad (3.6)$$

The sum within  $R_m$  has been truncated to  $n \in [0, m-1]$  as the terms of order  $q^m$  and higher will all vanish after setting  $q = 0$ , irrespective of the form of the nonlinearity. As a consequence of this equation (3.6) must be strictly linear with respect to  $U_m$ . This result means that through careful choice of  $\mathcal{L}$ , the system can be iteratively solved through analytic, semi-analytic or numerical schemes.

As the homotopy parameter  $q$  maps the original equation at  $q = 0$  to the motivating nonlinear equation at  $q = 1$ , it follows that  $u(\mathbf{x}) = \phi(\mathbf{x}; 1)$  in equation (3.3), i.e.

$$u(\mathbf{x}) = U_0 + \sum_{m=1}^{\infty} U_m(\mathbf{x}). \quad (3.7)$$

### 3.1 Convergence properties

While it is possible to construct  $u(\mathbf{x})$  of this form, it is not guaranteed to satisfy the motivating nonlinear problem equation (3.1). As each  $U_m$  is an implicit function of  $\hbar$  and  $\mathcal{L}$ , then we can assume that equation (3.7) will converge to the solution of equation (3.1) for some subset of these two parameters. Typically, implementations of the HAM treat  $\mathcal{L}$  as a constant, as changing it makes fundamental changes to the solution expression of equation (3.6)—although such changes will be explored within Subsection 3.5.1. Subject to a choice of  $\mathcal{L}$ , the HAM implementations typically construct a family of solutions across a range of  $\hbar \in [-2, 0)$ , which are then considered in terms of an additional, introduced

convergence criteria. This criteria can be crafted in response to the problem at hand, but common implementations include evaluating the function, or its derivatives, at a fixed point along the domain; or considering the norm of the solution. This criteria should be constant over a subset of  $\hbar \in [-2, 0)$ , and the solution set corresponding to this region should satisfy equation (3.1).

A secondary approach, that is more typically suited to numerical implementations of the HAM is to consider the  $L^p$  norm (or any other suitable norm) of the residual error. In the case of the  $L^2$  norm this takes the form of

$$E(\hbar) = \left( \int_{\mathcal{D}} |\mathcal{N}[\hat{u}(\mathbf{x}; \hbar)] - \phi(x)|^2 d\mathbf{x} \right)^{1/2}, \quad (3.8)$$

where  $\mathcal{D}$  is the problem domain. Since the integrand is a positive-definite function, then so is the integral of the error, and as such there must be some  $\hbar_{\text{opt}}$  for which

$$\hbar_{\text{opt}} = \min_{\hbar} E(\hbar). \quad (3.9)$$

More formally, the convergence properties of the HAM can be explored through the following theorems. The first three follow Liao (2013), and as such the proofs have been omitted from this work, while the last is unique to this work.

**Theorem 3.1.** *Suppose that  $T \subset \mathbb{R}$  is a Banach space subject to a suitable norm  $\|\cdot\|$ , such as  $\|\cdot\|_{\infty}$ , over which the sequence  $U_m(\mathbf{x})$  is defined for some prescribed value of  $\hbar$ , and that initial approximation  $U_0$  remains sufficiently close to the true solution  $u(\mathbf{x})$  of the original equation. Taking  $r \in \mathbb{R}^{\neq 0}$ , then the following statements must hold:*

- (i) *If  $\|U_{k+1}(\mathbf{x})\| \leq r\|U_k(\mathbf{x})\|$  for some  $\hbar$  and an  $r$  over all  $k \in \mathbb{R}^+$ , then the series solution  $\psi(t; q)$  converges uniformly at  $q = 1$  to  $u(\mathbf{x})$  if and only if  $0 < r < 1$ .*
- (ii) *If  $\|U_{k+1}(\mathbf{x})\| \geq r\|U_k(\mathbf{x})\|$  for some  $\hbar$  and an  $r$  over all  $k \in \mathbb{R}^+$ , then the series solution  $\psi(t; q)$  diverges uniformly at  $q = 1$  to  $u(\mathbf{x})$  if  $r > 1$ .*

**Theorem 3.2.** *If  $S_N(\mathbf{x})$  is a convergent series at  $q = 1$ , then the series as  $N \rightarrow \infty$  converges exactly to the solution of the original nonlinear problem.*

**Theorem 3.3.** *If  $\|U_{k+1}(\mathbf{x})\| \leq r\|U_k(\mathbf{x})\|$  for some  $\hbar$  and a  $r$  over all  $k \in \mathbb{R}^+$ , then the error after  $M$  steps of the HAM process must be bounded by  $\frac{r^{M+1}}{1-r}\|U_0\|$ .*

**Theorem 3.4.** Any sequence  $u(\mathbf{x}) = \sum_{M=0}^{\infty} U_M(\mathbf{x})$  solved through equation (3.2) at  $\mathcal{H} = 0$  can only converge upon the solution of the motivating nonlinear problem if  $\hbar \in [-2, 0)$ .

*Proof.* To show this, let us consider a nonlinear equation formulated as

$$\mathcal{L}[u] + \mathcal{N}[u] = \psi, \quad (3.10)$$

where  $\mathcal{L}$  will also serve as the auxiliary linear operator, and  $\mathcal{N}$  may contain linear components. Decomposing this into an infinite sequence of equations through the HAM yields

$$\left. \begin{aligned} U_m - (\chi_m + \hbar)U_{m-1} &= \hbar\mathcal{L}^{-1} \left[ R_{m-1}[\hat{U}_{m-1}] - (1 - \chi_m)\psi \right] \text{ and,} \\ R_{m-1} &= \frac{1}{m-1!} \frac{\partial^{m-1}}{\partial q^{m-1}} \left( \mathcal{N} \left[ \sum_{n=0}^{m-1} U_n q^n \right] - (1 - \chi_m)\psi \right) \Big|_{q=0}, \end{aligned} \right\} \quad (3.11)$$

where  $\hat{U}_{m-1}$  is the set of previously calculated solutions  $\{U_0, U_1, \dots, U_{m-1}\}$ . It then follows from Theorem 3.1 that a necessary and sufficient condition for convergence is that

$$\frac{\|U_m\|}{\|U_{m-1}\|} \leq 1,$$

where  $\|\cdot\|$  is a suitable norm in Banach space. Incorporating equation (3.11) into the ratio of norms gives that

$$\frac{\|U_m\|}{\|U_{m-1}\|} = \frac{\|(\hbar + 1)U_{m-1} + \hbar\mathcal{L}^{-1}R[\hat{U}_{m-1}]\|}{\|U_{m-1}\|}, \quad (3.12)$$

where  $\hat{U}_{j-1}$  is the set  $\{U_0, U_1, \dots, U_{m-1}\}$ . Thus equation (3.12) must be bounded by

$$\frac{\|(\hbar + 1)U_{m-1} + \hbar\mathcal{L}^{-1}R[\hat{U}_{m-1}]\|}{\|U_{m-1}\|} \leq |(\hbar + 1)| + |\hbar| \frac{\|\mathcal{L}^{-1}R[\hat{U}_{m-1}]\|}{\|U_{m-1}\|}. \quad (3.13)$$

As convergence can only occur when  $\|U_m\|/\|U_{m-1}\| \leq 1$ , it then must be that

$$\|\mathcal{L}^{-1}R[\hat{U}_{m-1}]\| \leq \frac{1 - |\hbar + 1|}{|\hbar|} \|U_m\|. \quad (3.14)$$

As  $\|\cdot\|$  must be strictly positive, this identify can only be satisfied when

$$1 \leq \frac{1 - |\hbar + 1|}{|\hbar|}, \quad (3.15)$$

which can only be satisfied for  $-2 \leq \hbar \leq 0$ . In the case where  $\hbar = 0$  the iterative scheme of equation (3.11) will produce a convergent infinite sequence, however, the scheme will correspond to

$$\left. \begin{aligned} U_0 &= \mathcal{L}^{-1}[\phi], \\ U_1 &= \hbar U_0 \text{ and} \\ U_m &= (1 + \hbar)U_m \text{ for } m \geq 2. \end{aligned} \right\} \quad (3.16)$$

As this iterative scheme is independent of  $\mathcal{N}$ , it follows that the necessary, but not sufficient, condition for any convergence to occur is that  $-2 \leq \hbar < 0$ .

□

## 3.2 Spectral Homotopy Analysis Method

While the original formulation of the HAM was built upon solving equation (3.6) algebraically, the scheme has since evolved to incorporate both computer algebra and numerical discretisations. The latter approach, using spectral methods, falls under the aegis of the Spectral Homotopy Analysis Method (SHAM). These semi-analytic and numerical extensions were driven by one of the fundamental limitations of the HAM—in that the auxiliary linear operator  $\mathcal{L}$  must be chosen in such a way that the resulting equations can be solved. Computer algebra approaches broaden the forms of  $\mathcal{L}$  that can be solved for, with the numerical framework increasing the set of admitted forms of  $\mathcal{L}$  even further, to the point where it is no longer a primary consideration when solving a nonlinear equation.

Motsa et al. (2010, 2012) were the first to consider strictly numerical analogues of the HAM, using Chebyshev collocation matrices to develop what is now known as the SHAM. Initial testing of this scheme demonstrated that the SHAM has the potential to significantly outperform MATLAB’s inbuilt boundary value problem routine ‘BVP4C’ (Nik et al., 2013), although a formal exploration of its numerical scaling has not been conducted.

The initial process for solving a differential equation using the SHAM follows that of the HAM. Beginning with a nonlinear problem of the form

$$\mathcal{N}[u] = \psi(x), \quad (3.17)$$

this equation can be solved by constructing the homotopy

$$(1 - q)\mathcal{L}[\phi(x; q); u_0(x)] = q\hbar (\mathcal{N}[\phi(x; q)] - \psi(x)). \quad (3.18)$$

This homotopy equation can be solved by repeatedly differentiating with respect to  $q$  and then normalising the resultant sequence of equations following the manner outlined previously within this chapter. Rather than solving these equations in an analytic or semi-analytic manner, the solutions can instead be constructed numerically in terms of the Chebyshev collocation matrices described within Subsection 2.2.1. This then yields the matrix equation

Repeating the previously outlined process for solving these homotopy equations, we differentiate with respect to  $q$  and then normalise the equations allows the iterative process of equation (3.6) to be constructed numerically in terms of the Chebyshev collocation matrices of Subsection 2.2.1 to yield the matrix equation

$$\mathbf{U}_m = \chi_m \mathbf{U}_{m-1} + \hbar \mathbf{A}^{-1} [\mathbf{N}_m]. \quad (3.19)$$

Here  $\mathbf{A}$  is the discretised matrix operators for the linear differential equation using the Chebyshev collocation approach, and  $\mathbf{N}_m$  is the matrix representation of  $R_m$ . Subject to a choice of  $\hbar$ , the general numerical solution to the original nonlinear problem is then

$$\mathbf{U} = \sum_{m=0}^{\infty} \mathbf{U}_m. \quad (3.20)$$

This approach to numerically discretise the linear differential equations that result from using the HAM shares many of the same advantages and disadvantages of other collocation based schema. While it does exhibit spectral accuracy in solving the resulting linear problems, it does so at the cost of having to solve dense matrix systems. Beyond this, the scheme is unable to handle variable coefficient differential equations, as the scheme under these conditions has a condition number that grows rapidly with the grid resolution. This final limitation is a particular problem when domain mappings from semi-infinite or infinite domains are incorporated into the problem, in order to solve over the Chebyshev domain  $x \in [-1, 1]$ .

### 3.3 Gegenbauer Homotopy Analysis Method

Motivated by the desire to exploit the advantageous convergence properties of the HAM without being tied to the limitations of Chebyshev collocation matrices, the remainder of this chapter will focus upon the development and testing of a new numerical scheme. This will be built upon the Gegenbauer methods for linear boundary value problems—as described in Subsection 2.2.2—to give a technique that we will call the Gegenbauer Homotopy Analysis Method (GHAM).

This method shares the Gegenbauer techniques advantageous properties of super-algebraic convergence for smooth differential equations, as discussed in Chapter 2. However, in contrast to other nonlinear solvers built upon the Gegenbauer method, the linear operator does not change, and the sparsity of the matrix operator is preserved across the iterative process. In aggregate these two effects have significant implications for the scheme’s computational performance, which will be further explored within Section 3.5.

The nonlinear discretisation for the GHAM replicates the approach taken for the SHAM, with the difference being in the application of the the boundary conditions. To understand this, let us once again consider a nonlinear problem of the form

$$\left. \begin{aligned} \mathcal{N}[u(x)] &= \psi(x), x \in (a, b) \\ \mathcal{B}[u(a), u'(a), \dots, u(b), u'(b), \dots] &= \mathbf{c}. \end{aligned} \right\} \quad (3.21)$$

Inhomogeneous boundary conditions have the potential to significantly complicate the iterative process that will result from applying the HAM to this nonlinear problem. As such, homogeneous boundary conditions can be imposed by solving the problem

$$\left. \begin{aligned} \mathcal{L}[u_0(x)] &= \psi(x), \\ \mathcal{B}[u(x), u'(x), \dots] &= \mathbf{c}, \end{aligned} \right\} \quad (3.22)$$

for an arbitrary auxiliary linear operator  $\mathcal{L}$ , to construct a mapping for equation (3.21) whereby

$$u(x) = v(x) + u_0(x).$$



An alternate approach is to arbitrarily construct a  $u_0(x)$  that satisfies the inhomogeneous boundary conditions, and applying that to the above mapping. In either case, this substitution gives rise to the modified nonlinear equation

$$\mathcal{L}_1[v(x)] + \mathcal{N}_1[v(x)] = \psi_1(x).$$

The process of decomposing  $u(x)$  into an equation with homogeneous boundary conditions is not strictly necessary, however it does greatly simplify the process of enforcing the boundary conditions for each of the steps of the homotopy process.

Within the framework of this modified nonlinear equation, a homotopy can be constructed between the auxiliary linear operator  $\mathcal{L}$  and the new nonlinear equation in the same manner as described at the start of this chapter, resulting in the iterative scheme

$$\left. \begin{aligned} \mathcal{L}[v_m - \chi_m v_{m-1}] &= \hbar \mathcal{L}_1[v_{m-1}] + \hbar R_m, \\ R_m &= \frac{1}{(m-1)!} \frac{\partial^{m-1}}{\partial q^{m-1}} (\mathcal{N}_1 + (1 - \chi_m) \psi_1) \Big|_{q=0}. \end{aligned} \right\} \quad (3.23)$$

This can then in turn be solved numerically using the framework developed within Subsection 2.2.2, which results in the numerical scheme

$$\mathbf{V}_m = \chi_m \mathbf{V}_{m-1} + \hbar \mathbf{A}^{-1} (\mathbf{A}_1 \mathbf{V}_{m-1} + \mathbf{N}_m). \quad (3.24)$$

Within this framework,  $\mathbf{V}_m$  is the vector representation of  $v_m$ ,  $\mathbf{A}$  and  $\mathbf{A}_1$  are the constant matrix operators corresponding to  $\mathcal{L}$  and  $\mathcal{L}_1$  respectively as represented through the Gegenbauer method, and  $\mathbf{N}_m$  represents  $R_m$ . From this, the solution to the motivating nonlinear equation (3.21) is then

$$\mathbf{U} = \mathbf{U}_0 + \sum_{m=0}^{\infty} \mathbf{V}_m. \quad (3.25)$$

The difference between this framework and that expressed by equation (3.19) within Section 3.2, beyond the manner in which the matrices are discretised, stems from the mapping to homogeneous boundary conditions.

### 3.3.1 Evaluating $R_m$

One of the largest hurdles for implementing the HAM is calculating the form that the nonlinear component of  $R_m$  will take. In analytic approaches, or approaches using symbolic computer algebra packages, the chain rule process of evaluating

$$\frac{1}{(m-1)!} \frac{\partial^{m-1} \mathcal{N}}{\partial q^{m-1}} \Big|_{q=0}$$

is fairly trivial, however evaluating it within a numerical context introduces additional hurdles. For polynomial nonlinearities, these derivatives can be described analytically by expanding powers of  $\sum_{i=0}^{\infty} U_i q^i$  and using Cauchy products so that

$$\frac{1}{(m-1)!} \frac{\partial^{m-1} \mathcal{N}}{\partial q^{m-1}} \Big|_{q=0} = \begin{cases} \sum_{i=0}^{m-1} U_i U_{m-1-i} \text{ for } \mathcal{N} = U^2 \\ \sum_{i=0}^{m-1} U_{m-1-i} \sum_{j=0}^i U_j U_{i-j} \text{ for } \mathcal{N} = U^3 \\ \sum_{i=0}^{m-1} U_{m-1-i} \sum_{j=0}^i U_{i-j} \sum_{k=0}^j U_{j-k} U_k \text{ for } \mathcal{N} = U^4 \\ \dots \end{cases} \quad (3.26)$$

More broadly, polynomial nonlinearities involving varying derivatives of  $U$  can be expressed in the same way. For example if  $\mathcal{N} = UU'U''U'''$  (where primes denote successive orders of differentiation), then the derivative of this function with respect to  $q$  will be

$$\frac{1}{(m-1)!} \frac{\partial^{m-1} \mathcal{N}}{\partial q^{m-1}} \Big|_{q=0} = \sum_{i=0}^{m-1} U_{m-1-i} \sum_{j=0}^i U'_{i-j} \sum_{k=0}^j U''_{j-k} U'''_k.$$

However this approach is not extensible to non-polynomial nonlinearities, limiting the potential applications of the numerical analogues of the HAM. One avenue to resolve this is to simply take a power series expansion of  $\mathcal{N}$  in terms of  $U$ , which would allow any nonlinearity to be rendered as a sum of polynomial nonlinearities, the form of which we have shown can be described in a closed form manner using Cauchy products. This relies upon being able to calculate and analytically describe these expansions, which is not always possible, and any ensuing truncation of the polynomial expansion would introduce additional numerical error to the scheme.

An alternate approach for calculating these derivatives involves applying Faà di Bruno's formula evaluating derivatives involving the composition of two functions. Through this

framework, the derivatives that result from methods based upon the HAM can be evaluated as

$$\frac{1}{(m-1)!} \left. \frac{\partial^{m-1} \mathcal{N}}{\partial q^{m-1}} \right|_{q=0} = \sum_{k_1+2k_2+\dots+nk_n=n} \binom{n}{k_1, k_2, \dots, k_n} G^{(r)}(U_0) \prod_{i=1}^n U_i^{k_i}.$$

Here  $(k_1 + 2k_2 + \dots + nk_n = n)$  corresponds to the summation across the set of all non-negative integer solutions  $\{k_1, k_2, \dots, k_n\}$ —which amounts to count of each number in the set of partitions of  $n$ ;  $r = k_1 + k_2 + \dots + k_n$ ; and

$$G^{(r)} = \frac{d^{(r)} \mathcal{N}}{dU^{(r)}}.$$

The Faà di Bruno formula also admits a generalised form (Mishkov, 2000), which allows derivatives to be constructed for nonlinearities involving  $u$  and varying orders of its spatial derivatives by considering the chain rule for a vector argument. If we consider the vector  $\mathbf{V} = [U(q), U'(q), U''(q), \dots, U^{(r)}]$ , then

$$\begin{aligned} \frac{1}{(m-1)!} \frac{d^{m-1} \mathcal{N}[V(q)]}{dq^{m-1}} &= \sum_0 \sum_1 \sum_2 \dots \sum_{m-1} \frac{1}{\prod_{i=1}^{m-1} (i!)^{k_i} \prod_{i=1}^{m-1} \prod_{j=1}^r q_{ij}!} \\ &\quad \times \frac{\partial^k \mathcal{N}[U_0, U'_0, \dots, U_0^{(r)}]}{\partial U^{p_1} \partial (U')^{p_2} \dots \partial (U^{(r)})^{p_r}} \\ &\quad \times \prod_{i=1}^{m-1} U^{q_{i,1}} (U')^{q_{i,2}} \dots (U^{(r)})^{q_{i,r}}. \end{aligned} \tag{3.27}$$

In this, the summation indices correspond to the full set of solutions to the non-negative Diophantine equations

$$\left. \begin{aligned}
\sum_0 &\rightarrow k_1 + 2k_2 + \cdots + nk_n = n, \\
\sum_1 &\rightarrow q_{1,1} + q_{1,2} + \cdots + q_{1,r} = k_1, \\
\sum_2 &\rightarrow q_{2,1} + q_{2,2} + \cdots + q_{2,r} = k_2, \\
&\vdots \\
\sum_{m-1} &\rightarrow q_{m-1,1} + q_{m-1,2} + \cdots + q_{m-1,r} = k_{m-1}.
\end{aligned} \right\} \quad (3.28)$$

The remaining indices  $p_j$  and  $k$  are then

$$\left. \begin{aligned}
p_j &= \sum_{i=1}^{m-1} q_{i,j} \text{ for } j = 1, 2, \dots, r, \\
k &= \sum_{i=1}^r p_i = \sum_{j=1}^{m-1} k_j.
\end{aligned} \right\} \quad (3.29)$$

This then allows any nonlinear operator to be decomposed into powers of its coefficients. One interesting consequence of this formulation is that it is clear that the  $m$ -th derivative will involve sums of products of the powers of  $\{U, U', \dots, U^{(r)}\}$ , multiplied by derivatives of the nonlinear operator solely evaluated at  $\{U_0, U'_0, \dots, U_0^{(r)}\}$ . Through this, we now have the tools to construct the equations for the HAM without the need for symbolic algebra, subject to the requirement that we are able to know and calculate the partial derivatives

$$\frac{\partial^k \mathcal{N}[U_0, U'_0, \dots, U_0^{(r)}]}{\partial U^{p_1} \partial (U')^{p_2} \dots \partial (U^{(r)})^{p_r}}. \quad (3.30)$$

Interestingly, while Faà di Bruno's formula has been used in the context of the HAM by Ali et al. (2010), the multivariate formulation has not.

### 3.3.2 Psuedo-code algorithm for evaluating problems with the GHAM

To outline the process of solving a nonlinear problem using the GHAM, a psuedo-code algorithm is outlined within Algorithm 1. This code can be invoked in several different contexts. This algorithm can be called for a single  $\tilde{h}$ , based either upon previous results or through the process outlined within Section 3.6; or it can be used to approximate  $\tilde{h}_{\text{opt}}$

by calling the algorithm as part of a numerical optimisation process, based upon the error after a small number of iterations.

---

**Algorithm 1** Process for solving using the GHAM for a given  $\hbar$ , error tolerance  $e_t$ , and maximum number of iterations  $n_m$

---

Define the auxiliary linear operator  $\mathcal{L}$   
Define the nonlinear operator  $\mathcal{N}$   
Discretise  $\mathcal{L}$  upon the Gegenbauer basis  
Establish the linear boundary conditions  
Construct matrix operator  $A$  in the manner of Subsection 2.2.2  
Precompute inverse, directly or through LU decomposition  
Solve  $\mathcal{L}(u_0) = \phi$   
Calculate  $R_0$   
Solve  $\mathcal{L}(u_1) = \hbar\mathcal{N}(u_0)$   
Evaluate the error  $e$   
 $n = 2$   
**while**  $e > e_t$  and  $n < n_m$  **do**  
    Evaluate  $R_n$  in the manner of Subsection 3.3.1  
    Solve  $\mathcal{L}(u_n) = (\hbar + 1)\mathcal{L}(u_{n-1}) + \hbar R_n$   
    Evaluate the error  $e$   
     $n \leftarrow n + 1$   
**end while**

---

### 3.3.3 Solving the Falkner–Skan equation

In order to explore the numerical properties of the GHAM, in Cullen and Clarke (2017) we considered the nonlinear Falkner-Skan equation, which models the high Reynolds number flow past a flat plate inclined at an angle of attack by recasting the physical equations in terms of the two-point boundary value problem

$$\left. \begin{aligned} \frac{d^3 f}{d\eta^3} + f \frac{d^2 f}{d\eta^2} + \beta \left[ 1 - \left( \frac{df}{d\eta} \right)^2 \right] &= 0, \\ f(0) = \frac{df}{d\eta}(0) = 0, \frac{df}{d\eta}(\eta) &\rightarrow 1 \text{ as } \eta \rightarrow \infty. \end{aligned} \right\} \quad (3.31)$$

Here  $\eta$  and  $f$  are similarity variables representing the displacement and velocity respectively, and the wedge angle of attack is  $\pi\beta/2$ . A special case of this is the well known Blasius equation for flow past a flat plate, which corresponds to the case where  $\beta = 0$ . The Falkner-Skan equation is known to have a single solution for  $\beta \geq 0$ , and two solutions known as the normal and reverse flow solutions, based upon the sign of the second derivative on

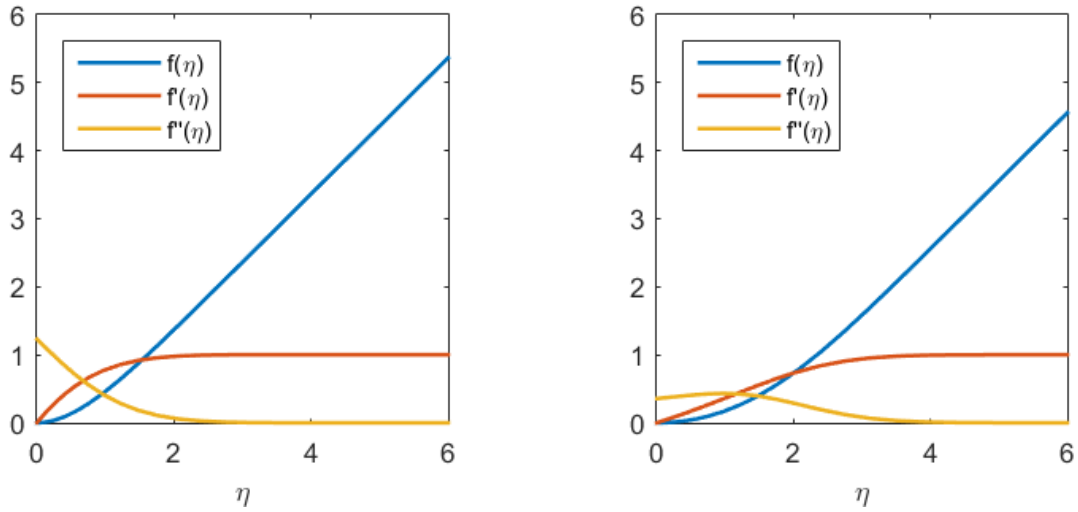


Figure 3.1: Normal flow solutions to the Falkner-Skan equation, and their first and second derivatives, for a)  $\beta = 1$  and b)  $\beta = -0.1$

the boundary—for  $\beta_{\min} \leq \beta < 0$ , which was shown to by Fazio (2013) to be  $\beta_{\min} \approx -0.1988$ .

Implementing the GHAM in terms of a 512 point Chebyshev–Gauss discretisation, subject to a linear mapping between the truncated domain  $x \in [0, 20]$  onto  $[-1, 1]$ , equation (3.31) was solved in MATLAB over 20 steps of the iterative process. This domain truncation was justified by considering the rate in which  $\frac{\partial f}{\partial \eta}(\eta)$  approaches 1 as  $\eta$  increases, which reinforces that the domain truncation is appropriate in the context of the original boundary condition that  $\frac{\partial f}{\partial \eta}(\eta) \rightarrow 1$  as  $\eta \rightarrow \infty$ .

Solutions to the Falkner-Skan equation using the GHAM were found over  $-0.198 < \beta < 5$ , generating a range of solutions, with the solutions for the two indicative cases of  $\beta = -0.1$  and 1 presented in Figure 3.1. A representative sample of these results, and comparison results using the SHAM and MATLAB’s ‘BVP4C’ routine are presented in Table 3.1, where it can be seen that with the exception of  $\beta = -0.195$  there is strong agreement between the solutions generated by all three techniques. Of particular note is the time required to calculate these solutions—for a 512 Chebyshev grid at  $h = -1$  for both the GHAM and the SHAM, and a 50 point grid in ‘bvp4c’, it took an average of 0.039, 0.22 and 2.16 seconds per calculation for the GHAM, the SHAM and ‘bvp4c’ respectively. It must be noted that the disparity between the number of grid points between the homotopy based codes and ‘bvp4c’ is a product of the significantly slower computational performance of the latter.

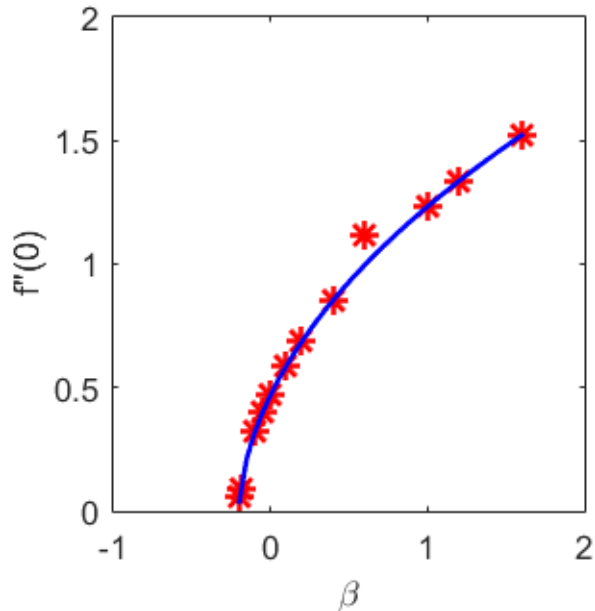


Figure 3.2: The relationship between  $\beta$  and  $f''(0)$ . In blue are the solutions calculated using the Homotopy based method, and in red are previously calculated solutions by Cebeci and Keller (1971)

	$\beta$						
	-0.195	0	0.2	0.6	1.0	1.2	2.0
GHAM	$5.8 \times 10^{-7}$	$4.8 \times 10^{-7}$	$3.8 \times 10^{-7}$	$3.6 \times 10^{-7}$	$2.2 \times 10^{-7}$	$3.1 \times 10^{-7}$	$3.1 \times 10^{-7}$
'bvp4c'	$2.3 \times 10^{-2}$	$3.5 \times 10^{-6}$	$9.4 \times 10^{-6}$	$4.5 \times 10^{-6}$	$2.8 \times 10^{-5}$	$1.8 \times 10^{-5}$	$2.2 \times 10^{-5}$

Table 3.1: Absolute error between calculated values of the skin friction coefficient  $f''(0)$  and the reference case of the SHAM for both the GHAM and MATLAB's 'BVP4C' routine, presented for a range of  $\beta$  values.

A more comprehensive comparison of the relationship between  $\beta$  and the work of Cebeci and Keller (1971) is shown in Figure 3.2. With the exception of  $\beta = 0.6$ , the solutions align to at least 3 significant figures. The disparity between the two appears to be the result of a tabulation mistake in the original paper, as the presented result for  $\beta = 0.6$  aligns much more closely with  $\beta = 0.8$  from the results calculated using the GHAM.

To assess the validity of this new technique, the rate of convergence of the  $L_2$  and  $L_\infty$  norms of the residual was considered for solutions constructed with the GHAM, shown in Figure 3.3 and Figure 3.4. The former figure is particularly noteworthy, as it demonstrates

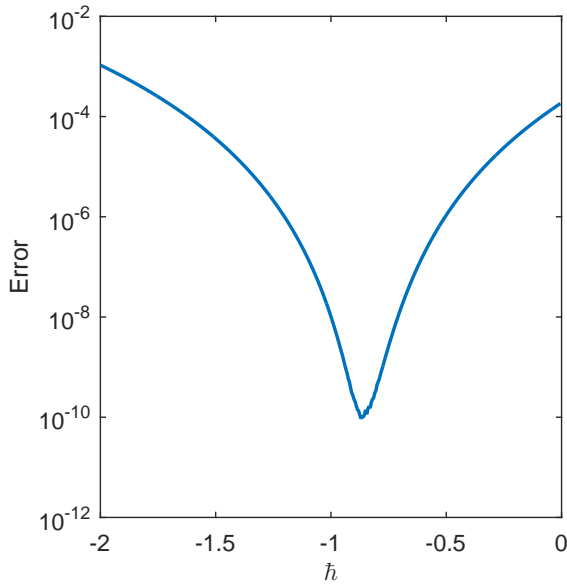


Figure 3.3: Error, evaluated as the  $L_2$  norm of the residual after 4 iterations, evaluated against  $\hat{h}$ .

the convex nature of the error with respect to  $\hat{h}$ , and the clear computational advantage of solving the problem at  $\hat{h}_{\text{opt}}$ . In the case where  $\beta = 1$ , the solution at  $\hat{h}_{\text{opt}} \approx -0.9$  took just 18 iterations for the residual to reach machine precision.

At this point it must be noted that the technique presented was unable to resolve the known secondary solution within the region  $\beta_{\text{min}} \leq \beta \leq 0$ , as to this point there is no flexibility within the scheme to admit multiple solutions. As such, we can now consider extensions to the GHAM in order to consider problems which admit multiple solutions.

### 3.4 Solving problems with multiple solutions through GHAM

Techniques based upon the HAM have previously considered such problems in terms of boundary conditions modified to incorporate an additional free parameter. Multiple distinct solutions can then be identified by varying the free boundary parameter until the problem satisfies the original boundary conditions.



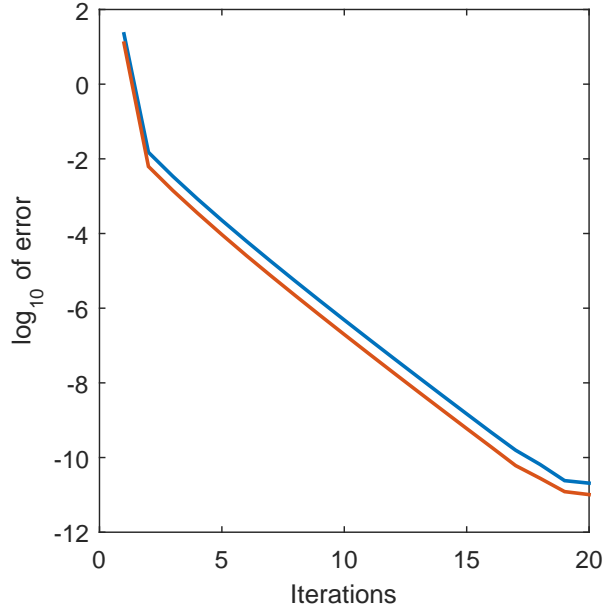


Figure 3.4: Convergence of the residual at  $\beta = 1$  and at the optimal  $\hbar$ . The  $L_2$  and  $L_\infty$  norm of the residual in blue and red respectively.

An example of this can be found in Abbasbandy et al. (2009), who considers a nonlinear boundary value problem that arises in the study of the kinetics of chemical reactions. This problem is the one-dimensional reaction-diffusion model

$$\left. \begin{aligned} \frac{d^2 u(x)}{dx^2} - \phi^2 u^n(x) &= 0 \\ \frac{du(0)}{dx} = 0, \quad u(1) &= 1, \end{aligned} \right\} \quad (3.32)$$

which describes a reaction in the presence of a porous catalyst (Sun et al., 2004). Here  $u$  is the nondimensionalised reactant concentrate,  $x \in (0, 1)$  denotes the position transverse to the catalyst,  $\phi$  is the Thiele modulus and  $n \geq -1$  is the reaction order.

In order to solve this equation using the HAM, Abbasbandy et al. (2009) reconsidered the motivating equation as the initial value problem

$$\left. \begin{aligned} \frac{d^2 u(x)}{dx^2} - \phi^2 u^n(x) &= 0 \\ u(0) = \gamma, \quad \frac{du(0)}{dx} &= 0. \end{aligned} \right\} \quad (3.33)$$

Here  $\gamma$  is the auxiliary boundary parameter, that, in the manner of a shooting problem is then iterated over until the equation satisfies the original boundary condition that  $u(1) = 1$ . In the case when  $n = -1$ , this equation is rare in that it is a nonlinear boundary value problem that admits a known, analytic solution (Magyari, 2008), which, in implicit form is given by

$$x = \frac{\gamma}{i\phi} \sqrt{\frac{\pi}{2}} \text{Erf} \left[ i \sqrt{\ln \left( \frac{u}{\gamma} \right)} \right] \quad (3.34)$$

where  $i = \sqrt{-1}$ , and Erf is the error function. The inverse of this is

$$u = \gamma \exp \left( - \left[ \text{InvErf} \left( \sqrt{\frac{2}{\pi}} \frac{i}{\gamma} \phi x \right) \right]^2 \right). \quad (3.35)$$

Here InvErf is the inverse of the error function, which does not have a closed form description, but can be calculated numerically. Substituting in the condition that  $u(1; \gamma) = 1$ , it follows that the relationship between  $\phi$  and  $\gamma$  is

$$\frac{\gamma}{i\phi} \sqrt{\frac{\pi}{2}} \text{Erf} \left( i \sqrt{\ln \left( \frac{1}{\gamma} \right)} \right) = 1. \quad (3.36)$$

Reconsidering equation (3.32) through the GHAM requires recasting the problem as the boundary value problem

$$\left. \begin{aligned} \frac{d^2 u(x)}{dx^2} - \phi^2 u^n(x) &= 0 \\ u(0) = \gamma, \quad u(1) &= 1. \end{aligned} \right\} \quad (3.37)$$

By then searching through the  $(\hbar, \gamma)$  parameter space, the extra degree of freedom created by the introduction of  $\gamma$  can be accounted for by searching for a solution that satisfies the original boundary conditions from equation (3.32). An example of the results found for  $\phi = 0.6$  can be seen in Figure 3.6. Two local minima can be seen at  $(\hbar, \gamma) = (-1.89, 0.7789)$  and  $(-0.8974, 0.1005)$ , both of which correspond to the equivalent exact solutions given by equation (3.36), and shown in Figure 3.5.

Figure 3.6 demonstrates the utility of varying  $\gamma$  for satisfying the original boundary conditions of equation (3.32). As would be expected, the inherent nonlinearity of this problem manifests in a strong sensitivity to the introduced boundary condition  $\gamma$ . More interestingly

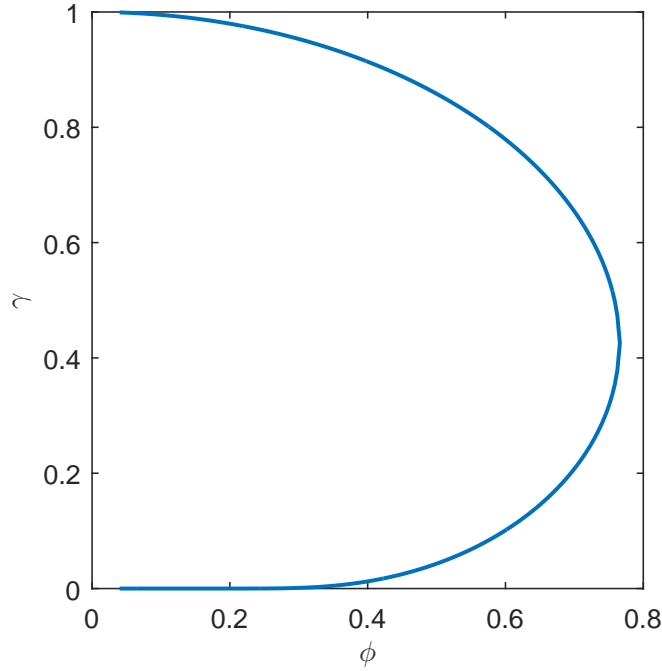


Figure 3.5: Relationship between  $\gamma$  and  $\phi$  for equation (3.32), calculated both numerically and through the implicit relationship equation (3.36).

is how the problem manifests its inherent sensitivity to the convergence control parameter  $\hbar$ . While the results are largely independent of  $\hbar$  for the majority of  $\hbar \in (-2, 0)$ , both exhibit distinct local minima within each band of distinct  $\gamma$ .

Repeating this process across  $\phi \in [0, 0.8]$  perfectly recreates the solution profile of Figure 3.5. However, the computational cost for such an exploration is not trivial, due to the requirement for high resolution sampling across the  $(\hbar, \gamma)$  parameter space for each  $\phi$ . This cost can be somewhat mollified by constructing matrix operators for  $\mathcal{L}$  that are independent of  $\phi$ , so the operators can be reused across all  $\phi$ . However, the primary contributor to the cost is simply the exploring the parameter space in terms of  $\hbar$  and  $\gamma$ . The number of sample points that need to be evaluated can be minimised by utilising a modified sampling criteria—such a Latin hypercubes—or by utilising an optimisation technique to search for the distinct local minima of the parameter space, however there is still, inherently, a significant cost involved.

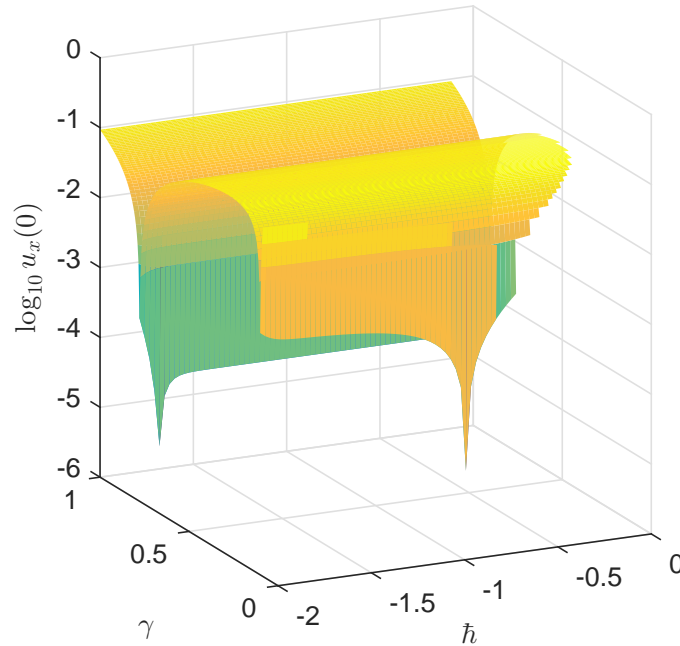


Figure 3.6: Evaluating the boundary derivative  $\frac{du(0)}{dx}$  of solutions of equation (3.37) for  $\phi = 0.6$ , and attempting to find solutions where  $\frac{du(0)}{dx} = 0$ .

While the preceding technique is the standard method for approaching problems with multiple solutions using variants of the HAM, the computational costs associated with it suggest that there is significant scope for improvement. As such, several new tools will be introduced to consider these problems within a numerical continuation framework.

### 3.4.1 Homotopy based hypersphere continuation

Numerical continuation regimes predominantly fall into two taxonomies—arc-length continuation and homotopy based continuation. While the techniques of Section 2.4 can be adapted for nonlinear numerical methods, within this section two new numerical continuation routines will be introduced, one which adapts the precepts of arc-length continuation for solutions constructed using GHAM; and the second, discussed in Subsection 3.4.2, will be a novel combination of both homotopy and arc-length continuation, in a manner that exhibits some distinct advantages over standard techniques.

The first method involves taking a nonlinear problem of the form

$$\mathcal{N}[u(x), \lambda] = f(x), \quad (3.38)$$

where  $\lambda$  defines the position in parameter space to be explored through the continuation process. This formulation also applies to eigenvalue problems. If this equation was to be solved using the GHAM for a fixed value of  $\lambda$ ,  $u(x)$  would be decomposed into the Maclaurin series in terms of  $q$

$$u(x) = u_0(x) + \sum_{m=1}^{\infty} q^m u_m(x) \Big|_{q=0}. \quad (3.39)$$

For the purposes of considering the problem within a continuation framework, the parameter  $\lambda$  can similarly be constructed in terms of the similar Maclaurin series

$$\lambda = \lambda_0 + \sum_{m=1}^{\infty} q^m \lambda_m \Big|_{q=0} \quad (3.40)$$

In the context of a continuation process,  $(u_0(x), \lambda_0)$  can be solutions calculated at a known position in parameter space. The remaining values of  $\lambda_m$  can be solved concurrently with the  $v_j$ , following the method of equation (3.23). Of course, as a consequence of introducing this extra set of unknown parameters, the system of equations will no longer be well posed. In order to provide for a coherent progression from one known solution to a new unknown solution at a different position in parameter space, the form of equation (2.48) can be modified by introducing the closure

$$\left( \frac{dF(u(\mathbf{x}))}{ds} \right)^2 + \left( \frac{d\lambda}{ds} \right)^2 = 1, \quad (3.41)$$

where  $s$  is an imposed step length parameter, and  $F(u(\mathbf{x}))$  is some feature of the solution space, as expressed in Chebyshev space. In this work we have considered several different features, with the  $L_2$  norm  $\|\cdot\|_2$  in both the Chebyshev and real spaces, the mean value in real space, the maximum value of the solution, and the value of the solution at  $x = 0$ . The mean, and the value of the solution at  $x = 0$  have been the most successful metrics tested.

In order to implement equation (3.41), a first order forward difference discretisation was used for the derivatives, so that

$$(F(u^{(j)}(\mathbf{x})) - F(u^{(j-1)}(\mathbf{x})))^2 + (\lambda^{(j)} - \lambda^{(j-1)})^2 = \Delta s^2. \quad (3.42)$$

Here  $\Delta s$  is an arbitrarily prescribed step size, that can be varied to control convergence, and the superscript  $(\cdot)^{(j)}$  denotes the  $j$ -th solution in our continuation process (separate to the solution in terms of our homotopy series), with  $(u^{(j-1)}, \lambda^{(j-1)})$  being a pairing of known solutions, and  $(u^{(j)}, \lambda^{(j)})$  corresponding to the unknown set that is being solved for. When expanded, this gives

$$\begin{aligned} -2F(u^{(j-1)}(\mathbf{x}))F(u^{(j)}(\mathbf{x})) - 2\lambda^{(j-1)}\gamma^{(j)} + (F(u^{(j)}(\mathbf{x})))^2 + (F(u^{(j-1)}(\mathbf{x})))^2 + (\lambda^{(j)})^2 \\ = \Delta s^2 - (F(u^{(j-1)}(\mathbf{x})))^2 - (\lambda^{(j-1)})^2. \end{aligned} \quad (3.43)$$

In this form, the first two terms are linear with respect to the unknown functions, the third and fourth terms are nonlinear, and the right hand side is exclusively written in terms of known solutions. As a result of this, we can simply implement this in terms of the GHAM framework, and solve for  $\lambda_m$  concurrently with  $u_m$  at each stage of our iterative process.

The matrix used in the GHAM discretisation to describe the linear system can be modified for the continuation process, with  $A$  able to be considered in the partitioned form

$$A = \left( \begin{array}{c|c} B & C \\ \hline D & E \end{array} \right). \quad (3.44)$$

Here  $A$  is a matrix of size  $(n+1) \times (n+1)$ ;  $(B, C)$  is the  $(n+1) \times 1$  vector corresponding to equation (3.43); and  $E$  is the  $1 \times n$  size matrix that corresponds to the coefficients of the term  $\lambda$  in the nonlinear equation. As the majority of the components of this matrix are invariant with respect to the position in the condition process, the Sherman-Morrison formula (Bartlett, 1951) can be leveraged in order to minimise the computational cost of constructing new matrix inversions. It is also possible to use analogues of this scheme to use the Sherman-Morrison formula to update LU decompositions of the matrix operator (Quintana-Orti and Van De Geijn, 2008).

## Computational Approach

In order to more easily differentiate between multiple solutions, the algorithm considers the starting point  $(u_0, \lambda_0)$  to be perturbations from previously calculated solutions at  $(u^{(j-1)}, \lambda^{(j-1)})$ , with multiple starting points constructed in the neighbourhood of the previously calculated solution. Each of these points is evaluated for a small number of iterative

steps. The resultant  $(\bar{x}, \lambda)$  combinations at each point are then classified into three groups, following the rules:

1. If the calculated points lie upon the previously calculated curve, then these points are discarded.
2. If the solutions converge up to the order of the iteration threshold but are not below the error tolerance, they are placed into group 1.
3. If the calculated points converge and are below the specified error tolerance then they are classified into group 2.
4. If the solutions upon the calculated points diverge, these points are discarded.

If groups 1 and 2 are empty, this suggests that either the solution has reached the end of its branch, or that no solutions can be found at that step size. If so, the step size is decreased and the problem is restarted. If these groups are non-empty, the sets can then be clustered using an adaptive form of K-means clustering, based upon (Arthur and Vassilvitskii, 2007), whereby the number of clusters is increased until there is no meaningful difference between the clusters. For each cluster not in the neighbourhood of the previously calculated curve, the lowest residual solution is selected, and, if it is not below the specified error tolerance is iterated upon. Any distinct points that satisfy the hypersphere constraint equation (3.43) and have a residual where the error is below the defined tolerance are considered to be new points on the solution curve, which can then in turn be iterated upon, repeating this process until either the solution curve rejoins itself, or the step size decreases below a defined threshold, suggesting that the solution branch has no further solutions.

This routine can be parallelised at two points. The first is if the previously described clustering identifies multiple solutions, then these points can be assessed independently. As well as this the set of starting points  $(u_0, \lambda_0)$  about  $(u^{(j-1)}, \lambda^{(j-1)})$  can all be iterated upon independently, giving rise for additional potential to parallelise the solution process.

### 3.4.2 Homotopic integrated arc-length continuation

While homotopy and arc-length based techniques have always been considered to be distinct, separate forms of numerical continuation, it is in fact possible to construct an arc-length continuation scheme based not upon the length between a known solution and

a new solution in parameter space, but rather the length along the curve defining the homotopy that connects these solutions using the Homotopy Analysis Method. This novel method has the distinct advantage in that this path must exist if a solution can be constructed using the HAM, and if it exists then the path is guaranteed to be both smooth and continuous, which means that any issues in navigating folds and bifurcations can be avoided.

To elaborate on this process, let us again consider a generalised nonlinear problem of the form

$$\mathcal{N}[(u(x), \lambda)] = \psi(x), \quad (3.45)$$

which has a known solution at  $(u_0(x), \lambda_0)$ . Within this framework one equation of the vector  $\mathcal{N}$  will correspond to the as yet un-introduced closure equation for the continuation.

The solution to this problem through the GHAM can be found by constructing a homotopy between the auxiliary linear vector operator  $\mathcal{L}[u(x)]$  and  $\mathcal{N}[u(x), \lambda]$  where

$$\mathcal{H} \equiv (1 - q)\mathcal{L}[(U, \Lambda) - (u_0, \lambda_0)] + q\mathcal{h}(\mathcal{N}[(U, \Lambda)] - \psi(x)).$$

In a similar manner to  $\mathcal{N}$ , the operator  $\mathcal{L}$  will include one equation that will be the starting point of the homotopy deformation for the closure equation.

Defining both  $u(x)$  and  $\lambda$  as the Maclaurin series in terms of  $q$

$$\left. \begin{aligned} u(x) &= U(x) \Big|_{q=1} = u_0 + \sum_{j=1}^{\infty} U_j q^j \Big|_{q=1} \\ \lambda &= \Lambda \Big|_{q=1} = \Lambda_0 + \sum_{j=1}^{\infty} \Lambda_j q^j \Big|_{q=1} \end{aligned} \right\} \quad (3.46)$$

allows for the motivating nonlinear problem to once again be partitioned into a series of linear equations. However, the introduction of the extra degree of freedom means that these equations must be ill-posed. To resolve this, an arc-length constraint can be introduced in order to close the system of equations. Rather than using the differential form of the arc-length, the integral arc length

$$\delta s = \int_{q=0}^1 \left( \|u - u_0\|^2 + |\lambda - \lambda_0|^2 \right) dq \quad (3.47)$$



can be introduced, where  $\|\cdot\|$  is the  $L_2$  norm. This equation will form the component of  $\mathcal{N}$  corresponding to the closure.

As calculating this norm in real space would introduce significant computational hurdles, this instead will consider the  $L_2$  norm in Chebyshev space. Substituting in the expansions for  $u(x)$  and  $\lambda$  from equation (3.46) gives

$$|\lambda - \lambda_0|^2 = \left( \sum_{j=1}^{\infty} \Lambda_j q^j \right)^2 = \left( \sum_{j=0}^{\infty} c_j q^j \right) - 2\lambda_0 \left( \sum_{j=0}^{\infty} \Lambda_j q^j \right) + \lambda_0^2, \quad (3.48)$$

where  $c_j = \sum_{k=0}^j \Lambda_k \Lambda_{j-k}$ . From this, it follows that

$$\int_{q=0}^1 |\lambda - \lambda_0|^2 dq = \left( \sum_{j=0}^{\infty} \frac{c_j - 2\Lambda_0 \Lambda_j}{j+1} \right) + \Lambda_0^2. \quad (3.49)$$

Similarly, we can say that

$$\int_{q=0}^1 \|u - u_0\|^2 dq = \sum_{n=0}^N \left( \left( \sum_{j=0}^{\infty} \frac{d_{j,n} - 2u_{0,n} U_{j,n}}{j+1} \right) + u_{0,n}^2 \right) \quad (3.50)$$

where the subscript  $n$  denotes the  $n$ -th Chebyshev mode of  $U_j$ , and similarly for  $c_j$  we have that

$$d_j = \sum_{k=0}^j U_{k,n} U_{k-j,n}. \quad (3.51)$$

This means that the arc-length of the homotopy path is defined by

$$\begin{aligned} \delta s = & \frac{1}{3} \left( \sum_{n=0}^N U_{1,n}^2 + \Lambda_1^2 \right) + \frac{1}{4} \left( \sum_{n=0}^N 2U_{1,n} U_{2,n} + 2\Lambda_1 \Lambda_2 \right) + \frac{1}{5} \left( \sum_{n=0}^N 2U_{1,n} U_{3,n} + U_{2,n}^2 + 2\Lambda_1 \Lambda_3 + \Lambda_2^2 \right) \\ & + \frac{1}{6} \left( \sum_{n=0}^N 2U_{1,n} U_{4,n} + 2U_{2,n} U_{3,n} + 2\Lambda_1 \Lambda_4 + \Lambda_2 \Lambda_3 \right) + \dots \end{aligned} \quad (3.52)$$

This can then be partitioned so that

$$\left. \begin{aligned}
\frac{1}{3} \left( \sum_{n=0}^N U_{1,n}^2 + \Lambda_1^2 \right) &= \delta s \\
\sum_{n=0}^N 2U_{1,n}U_{2,n} + 2\Lambda_1\Lambda_2 &= 0 \\
\sum_{n=0}^N 2U_{1,n}U_{3,n} + 2\Lambda_1\Lambda_3 &= - \sum_{n=0}^N U_{2,n}^2 - \Lambda_2^2 \\
\dots &
\end{aligned} \right\} \quad (3.53)$$

In the context of a perturbation scheme, the  $j$ -th line the left hand side involves  $(U_1, \Lambda_1)$  and  $(U_j, \Lambda_j)$ , and the right hand side involves the set of terms incorporating  $\{U_k, \Lambda_k\}$  for  $1 < k < j$ . As such, each line can be considered to be a linear constraint upon  $(U_j, \Lambda_j)$ , which can be used to condition the solution  $(u, \lambda)$  so that it is of length  $\delta s$  from the previously known solution  $(u_0, \lambda_0)$ . And, beautifully, because the coefficient of the unknowns  $(U_j, \Lambda_j)$  are always  $2U_1$  and  $2\Lambda_1$  respectively, our matrix discretisation will remain constant across all steps of the iterative process, which greatly simplifies the computational cost of solving this nonlinear problem.

The only complication to this iterative process is the first line of equation (3.53), which, crucially, is not a linear constraint upon  $(U_1, \Lambda_1)$ . However, in the context of constructing a solution using the GHAM, when we are solving for  $(U_1, \Lambda_1)$ , our iterative scheme is solving the equation

$$\mathcal{L}[(U_1, \Lambda_1)] = \hbar [\mathcal{N}[(u_0, \lambda_0)] - \psi(x)], \quad (3.54)$$

which is a linear equation. This means that to solve for  $(U_1, \Lambda_1)$ , we can think of

$$\mathcal{N}_2[(U_1, \Lambda_1)] \equiv \frac{1}{3} \left( \sum_{n=0}^N U_{1,n}^2 + \Lambda_1^2 \right) - \delta s = 0$$

as a nonlinear equation to be solved using the GHAM, where equation (3.54) is a linear constraint. Thus if we introduce an auxiliary linear operator  $\mathcal{L}_2$  so that

$$\mathcal{L}_2[U_1, \Lambda_1] = \sum_{n=0}^N U_{1,n} + \Lambda_1$$

then we can construct the homotopy between the linear solution and the nonlinear solution

$$\mathcal{H} \equiv (1 - q)\mathcal{L}_2[V - V_0, \tau - \tau_0] + q\hbar_2\mathcal{N}_2[V, \tau], \quad (3.55)$$

subject to the constraint equation (3.54). Here

$$\left. \begin{aligned} U_1 \equiv V \Big|_{q=1} &= V_0 + \sum_{j=1}^{\infty} V_j q^j \Big|_{q=1}, \\ \lambda_1 \equiv \Lambda_1 \Big|_{q=1} &= \tau_0 + \sum_{j=1}^{\infty} \tau_j q^j \Big|_{q=1}, \end{aligned} \right\} \quad (3.56)$$

and  $(V_0, \tau_0)$  correspond to the solution of  $\mathcal{L}_2[(V_0, \tau_0)] = 0$ .

To this point we have yet to define the component of  $\mathcal{L}$  that corresponds to the closure equation. While this component shares the broader properties of  $\mathcal{L}$ , in that it can be defined almost arbitrarily, testing for this work has revealed that structuring it as

$$\hat{\mathcal{L}} = \sum_{n=0}^N U_{m,n} \pm \Lambda_i \quad (3.57)$$

allows for solutions moving up and down continuation branches to be explored by simply changing the sign of equation (3.57).

The broader details of this continuation approach may appear on first appearance to be almost identical to the continuation approach of Subsection 3.4.1—with the only difference being that this approach using an integral arc-length, as compared to the differential form of arc length in the earlier technique. This difference, while subtle, has profound differences in their respective explorations of the parameter space. Hypersphere continuation approximates the arc-length as being the distance in physical space between the previously calculated solution and the new solution in  $n$ -dimensional space; whereas the homotopy arc-length continuation calculates the distance along the homotopy path between the previously calculated solution and the new solution in Chebyshev space. Because the latter's parametrisation is in terms of the homotopy path, it is guaranteed to be smooth, continuous and one-to-one between the two points, which allows the technique to easily traverse folds and bifurcations. This is particularly advantageous in the case of folds, which typically pose a significant difficulty to continuation schemes, which in turn imposes that the step size of the continuation process must be lowered significantly—increasing the

computational cost.

A secondary advantage of this homotopy arc-length continuation scheme comes from how the homotopy path is defined. While homotopy methods typically only consider the solutions at the extrema of  $q$ —which for this problem correspond to the old and new positions within the continuation process— $\phi(\mathbf{x}; q)$  also encodes a family of solutions for  $q \in (0, 1)$ , that corresponds to different points along the homotopy curve. These points can be considered approximations to the true continuation path between the old and new solutions. As such this approach provides additional information that relative to traditional continuation schemes, which only provide the location of the old and new points along the parameter space. This path of course must have a corresponding solution space, which can be used as starting points for rapidly iterating upon the solution space along the true path through the parameter space.

### Testing

To explore the performance of this proposed technique, we again turn to the nonlinear one-dimensional reaction–diffusion model of equation (3.32), as outlined in Section 3.4. Figure 3.7a presents the starting and end points of the continuation process (in black and magenta), the true solution path (in blue), and the approximation to the solution path (in red) for a single step of  $\delta s = 0.33$  from the upper solution branch at  $(\phi, \gamma) = (0.6, 0.7789)$  towards the lower branch. This step length was chosen to demonstrate the length of the traversal that is possible within this continuation framework, with the step length in this case only being limited by the scope of the parameter space admitted by the original problem.

In a surprising result, rather than the multiple small steps required of previous steps to traverse folds—such as the one at  $\phi \approx 0.7$ —this new technique was able to accurately traverse the jump from the upper branch of solutions to the lower branch in the parameter space in a terms of a single step of the continuation process. Furthermore, while the homotopy path—the red dashed line in Figure 3.7a—does not accurately match the true set of solutions in parameter space, it is both indicative of the true curve, and can serve as the starting point for accurately resolving the solutions along the curve. This can be performed by iterating upon a solution for fixed  $\phi$ , to refine the trial solution on the homotopy curve through parameter space until they converge upon the true curve, without the need of any additional continuation. This further heightens the efficiency advantage of this technique,

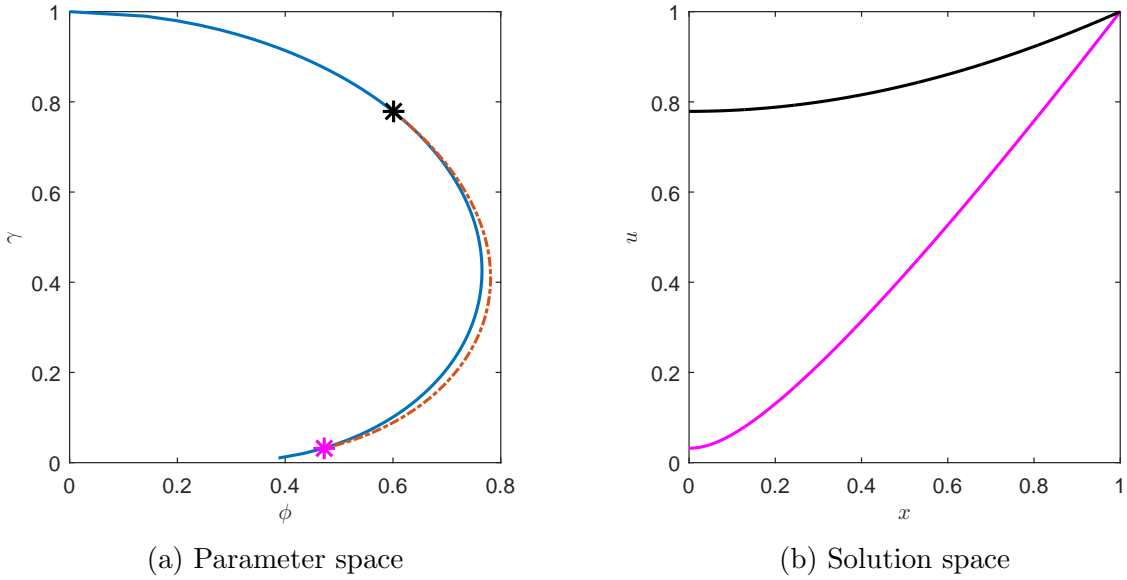


Figure 3.7: Analytically determined parameter space (blue), starting point (black) and end point of continuation (magenta), with the corresponding solutions for equation (3.32), calculated using Homotopic integrated arc-length continuation with a step length of 0.33. The continuation path is shown by the dashed red line.

as it can both accurately traverse large steps as part of the continuation process, but also gives information about the solutions along the continuation path, which can be refined without the more costly continuation process.

Lowering the step length to  $\delta s = 0.3$  allowed traversals both up and down the parameter space to be considered. Moving up the upper branch from the starting point in Figure 3.8 led to the parameter space position represented by the red diamond, while moving down the branch lead to the solutions denoted by the red squares. The multiple lower branch positions correspond to different convergent subsets of  $\hbar$ , and each exhibits its own distinct homotopy curve within the parameter space.

### 3.5 Computational cost of the GHAM

As mentioned in Section 3.3, compared to other methods the GHAM exhibits superior scaling, in terms of the computational time required to solve the resulting equations, due to its sparse, constant matrix operators. This contrasts favourably with other techniques,

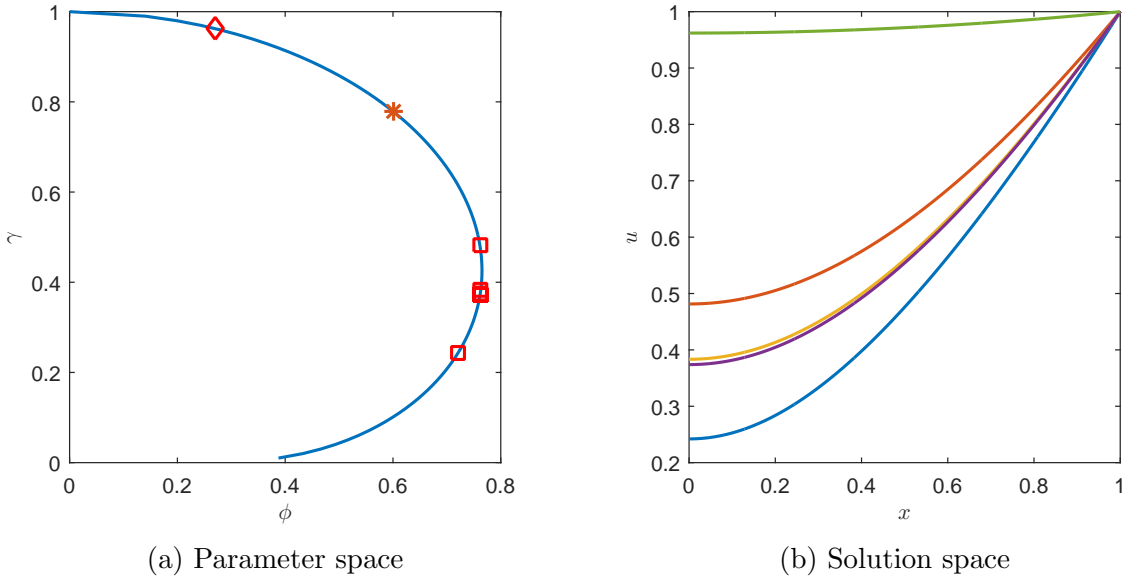


Figure 3.8: Parameter space for equation (3.32) in blue, with the positions after using Homotopic integrated arc-length continuation for a step length of 0.13 indicated by red squares (for traversing down the branch) and diamonds (for traversing up the branch). The starting point is given by the orange asterisk. The corresponding solutions of the solutions constructed with the continuation method are included within Figure 3.8b.

as they generally require a dense discretisation, even when the techniques are expressed in terms of the Gegenbauer basis functions.

These differences can be elaborated by considering the cost of solving the matrix equations that result from these numerical schemes. Through the Coppersmith-Winograd algorithm, the single matrix inversion required for methods based upon the HAM can be constructed with  $\mathcal{O}(n^{2.373})$  operation for a matrix system in  $\mathbb{R}^{n \times n}$  (Davie and Stothers, 2013, Le-Gall, 2014). From this point, each step of the matrix equation can then be solved with an additional  $\mathcal{O}(n^2)$  operations for the requisite matrix-vector product at each step of the homotopy iterative process. However, taking this approach results in a dense matrix operator that corresponds to the inverse matrix, and as such the matrix-vector products are in terms of dense systems—negating the fundamental advantages of considering a Gegenbauer based approach. Furthermore, there is also a secondary cost corresponding to the amount of memory required to store these dense matrices as  $n$  scales.

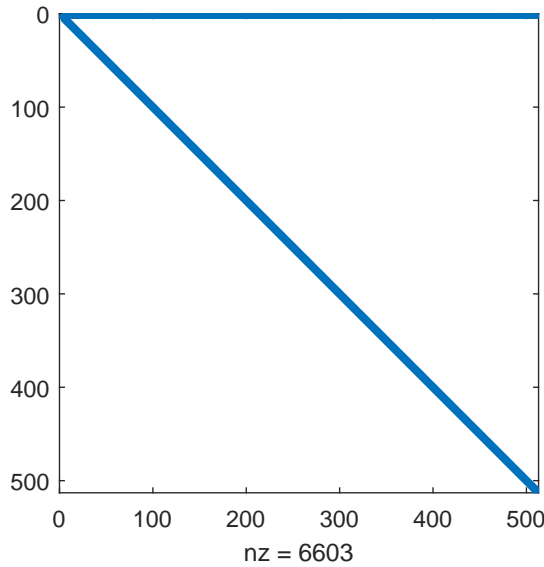


Figure 3.9: Sparsity of a fourth-order variable coefficient boundary value problem, discretised over 512 points using the GHAM.

An alternate approach for the GHAM involves considering the matrix inversion operations in terms of LU decomposition (Horn and Johnson, 1985). This process factors a matrix into a lower and upper triangular matrix, respectively known as the  $L$  and  $U$  matrices. By doing so, the process of Gaussian elimination can essentially be encoded into the  $L$  and  $U$  matrices, which reduces the complexity of solving the matrix systems at each step. In essence, this involves recasting a matrix system

$$A\mathbf{x} = \mathbf{f} \tag{3.58}$$

as  $LU\mathbf{x} = \mathbf{f}$  and solving the equations

$$\left. \begin{array}{l} L\mathbf{b} = \mathbf{f} \\ U\mathbf{x} = \mathbf{b}. \end{array} \right\} \tag{3.59}$$

Crucially, these two matrices are not necessarily dense matrices (as compared to the matrix inversion), resulting in a decrease in the technique’s storage requirements.

To illustrate this, the sparsity of the matrix operator for a fourth-order boundary value problem discretised using the Gegenbauer method over 512 points—shown in Figure 3.9—

can be compared to its equivalent sparsity after LU decomposition in Figure 3.10. The Gegenbauer discretisation clearly results in a sparse, diagonally dominated system for which only 2.5% of the matrix components are non-zero. Across a range of test problems, the total number of non-zero infill points for the LU decomposition was equivalent to between 1% and 46% of the amount of elements in a single equivalent dense  $\mathbb{R}^{n \times n}$  discretisation matrix. While there is an inherent memory overhead incurred in the storing of sparse matrices, as compared to their full equivalents, the decrease in the overall sparsity still induces commensurate effects upon the memory footprint of storing these matrices

Beyond the implications of LU decomposition for the overall memory required to resolve the matrix system, LU decomposition is desirable for its implications in terms of the time required to solve these systems. For a dense matrix system, the cost of constructing the original LU decomposition is  $\mathcal{O}(n^3)$  operations, with solutions of a matrix equation involving the LU decomposition able to be solved in  $\mathcal{O}(n^2)$  operations. For dense matrices these scaling properties make LU decomposition a poor alternative to the Coppersmith-Winograd algorithm, however in the case of sparse matrices the partial preservation of sparsity within LU decomposition significantly improves the scaling of both the decomposition process, and then of solving the resulting matrix equations. For systems where the number of non-zero points is  $\mathcal{O}(n)$ , the theoretical lower bound on the number of operations for both constructing the LU decomposition and evaluating the matrix-vector products is  $\mathcal{O}(n) + \mathcal{O}(nnz)$  (Duff et al., 1986, Gilbert and Peierls, 1988). Numerical evidence of these scaling properties will be discussed further within Subsection 3.5.1.

For the systems that result from using the GHAM, it was found that the most computationally efficient and effective form of LU decomposition involves scaled pivoting—whereby the relative magnitudes of points, rather than their absolute magnitudes, are taken into account in the selection of the pivots. In brief, LU decomposition with scaled pivoting—known as LUPQR decomposition—involves solving for the components  $R$ ,  $P$ ,  $Q$ ,  $L$  and  $U$  for a matrix  $A$ , for which

$$\left. \begin{aligned} RPAQ &= LU, \text{ so that} \\ \mathbf{u} &= QU^{-1}L^{-1}PR^{-1}\mathbf{f}. \end{aligned} \right\} \quad (3.60)$$

The specific details of how these auxiliary matrices can be constructed can be found in Golub and Van Loan (1996). For the GHAM, the specific implementation of this process



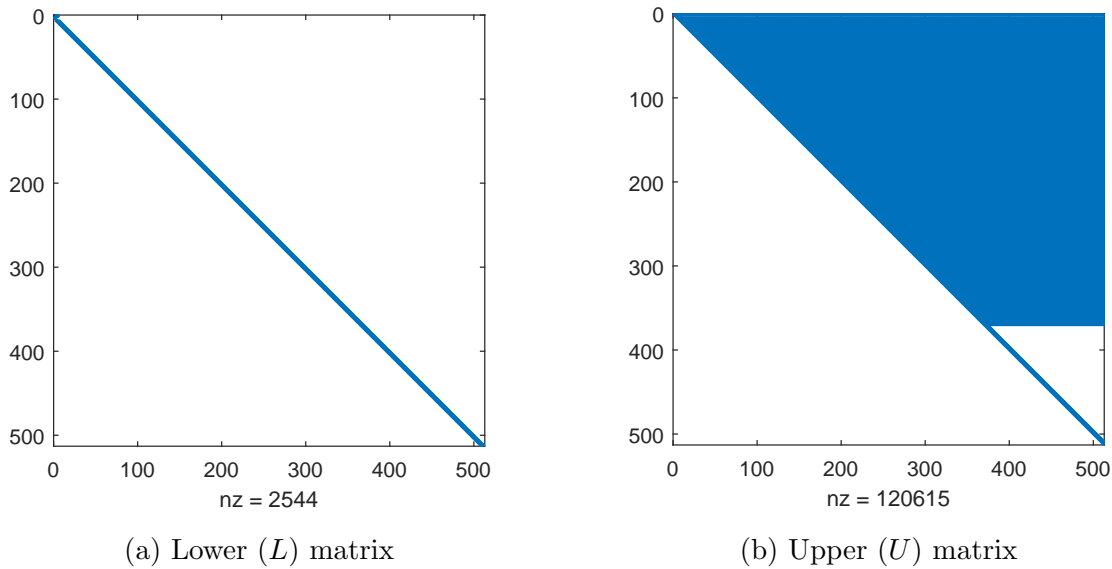


Figure 3.10:  $L$  and  $U$  matrices from a LU decomposition of a  $2^9 \times 2^9$  grid, corresponding to solutions of a nonlinear, variable coefficient viscous fluids problem, subject to the linear operator  $\mathcal{L}_4$  from equation (3.66).

was handled through MATLAB’s inbuilt sparse LU factoriser, which is in turn built upon the Suitesparse UMFPACK package (Davis, 2006, 2011).

In the context of the GHAM, LUPQR decomposition was reliably faster than all other tested methods for solving the system of linear equations. It would be expected that the additional complexity of the scheme would result in further increases of the memory footprint required, relative to a LU decomposition without the incorporation of scaled pivoting. However, for the same fourth-order test problem used above for LU decomposition, the LUPQR decomposition created the set of matrices Figure 3.11, which in aggregate only filled 1% more of the equivalent dense-matrix structure as compared to LU decomposition—a result that held as the size of the numerical discretisation was varied. Thus, for a very minor increase in the memory footprint, incorporating LUPQR decomposition results in noticeable differences to the speed of solving the resulting matrix systems.

Of course, solving the matrix systems is only one component of the overall computational cost of solving a nonlinear equation (or system of equations) within the GHAM framework. Other factors that influence the exhibited scaling include setting up the Gegenbauer

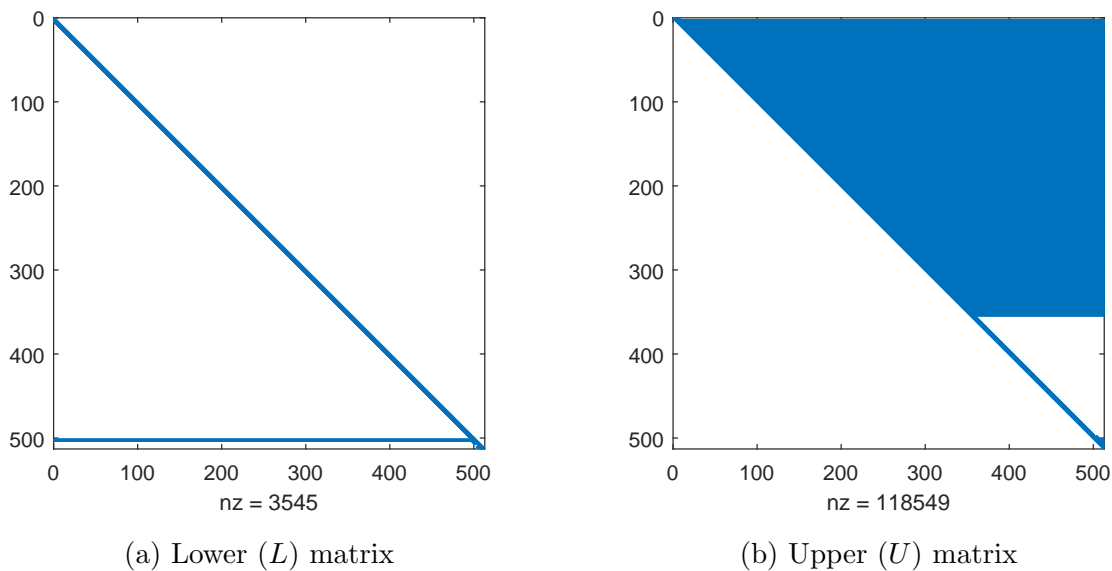


Figure 3.11:  $L$  and  $U$  matrices from a LUPQR decomposition of a  $2^9 \times 2^9$  grid for the same problem examined in Figure 3.10. The  $P$ ,  $Q$  and  $R$  matrices all correspond to matrices with  $2^9$  nonzero elements located entirely upon the main diagonal.

matrix system; evaluating derivatives; the transform between Chebyshev and real space; and evaluating the right hand side term  $R_m$ . The scaling of  $R_m$  is particularly important to understand, due to its scaling with  $m$ , as evidenced by the results within Subsection 3.3.1.

While the computational complexity for evaluating  $R_m$  is heavily problem dependent, an understanding of the scheme's scaling can be obtained by examining the cases of strictly quadratic and cubic nonlinearities. For a problem exhibiting a strictly quadratic nonlinearity solved over  $M$  steps of the GHAM iterative process, evaluating each  $R_m$  for  $m \leq M$  involves  $\lfloor \frac{m^2}{4} \rfloor$  multiplication operations and  $\lfloor \frac{(m-2)^2}{4} \rfloor$  addition operations, where  $\lfloor \cdot \rfloor$  represents the floor function. For a problem discretised over  $n$  quadrature points, the computational complexity of evaluating all the  $R_m$  terms will scale with

$$\mathcal{O}(nm^2). \tag{3.61}$$

In the case of a strictly cubic nonlinearity, the number of multiplication operations required is  $\mathcal{O}\left(m + \lfloor \frac{3(m-1)^2+1}{4} \rfloor\right)$ , and the number of addition operations is  $(m-2)^2$ , and as such will scale in a similar manner to that of equation (3.61), for a strictly quadratic nonlinearity.

If  $G^{(r)}(U_0) \neq 0$  for all  $r$ , then evaluating each  $R_m$  using Faa di Bruno’s formula will involve an additional  $h(m)$  products and  $p(m)$  sums, where  $p(m)$  is the number of partitions of  $m$  and  $h(m)$  is the total number of parts in all partitions of  $m$ . While neither of these two terms have an analytic description, they do respectively correspond to the A000041 and A006128 sequences from Sloane (2018). Our testing shows that the quadratic component of this appears to dominate, and that the scheme still broadly scales as  $\mathcal{O}(nm^2)$ .

As such, the dominant factor for solving a system of equations using the GHAM is the cost of solving the matrix equations, which, ignoring sparsity will behave as

$$\mathcal{O}(nm^2 + n^3 + mn^2), \tag{3.62}$$

where again  $m$  is the number of iterations and  $n$  is the number of grid points. Assuming that  $m < n$ , then the dominant component of the above relationship becomes

$$\mathcal{O}(n^3 + mn^2). \tag{3.63}$$

The cost of the  $\mathcal{O}(n^3)$  operations can also be considered to be amortised across the iterative process, as it is only required once. Furthermore, if we instead impose that the number of non-zero elements of the matrix equations is  $\mathcal{O}(n)$  then as was described previously the LU decomposition also scales with  $\mathcal{O}(n)$ , which leads to equation (3.62) reducing to

$$\mathcal{O}(nm^2). \tag{3.64}$$

Of course these results are for calculations at a single value of  $\hbar$ , and any exploration of the convergence properties across a range of the convergence control parameter will necessitate an increase in the computational cost. However, the number of samples in  $\hbar$  should be small relative to  $n$  and  $m$ , and as such equation (3.64) will still hold.

Subject to the assumption that  $m$  is independent of  $n$ —a result that will be shown in the following section—this process demonstrates quasi-linear  $\mathcal{O}(n)$  scaling with respect to the grid resolution. While spectral multigrid and Krylov methods are frequently touted as being the most efficient of the currently extant numerical methods, a  $d$ –dimensional problem discretised upon  $n$  points in each dimension is still limited to scaling with either  $\mathcal{O}(n^d \log_2 n)$  operations per iteration for Chebyshev methods, and  $\mathcal{O}(n^{d+1})$  operations for Legendre methods (Lottes and Fischer, 2005, Canuto et al., 2006). Outside of spectral

multigrid, other techniques scale at an absolute minimum with  $\mathcal{O}(n^2)$ . Thus for the one-dimensional problems being assessed here, the GHAM has the potential to significantly outperform all other currently extant techniques. These theoretical results will now be confirmed by considering the solutions of a fourth-order boundary value problem.

### 3.5.1 Two-dimensional viscous flow in a rectangular domain with porous, moving boundaries

The two-dimensional flow of a laminar, viscous and incompressible fluid confined within a rectangular domain bounded by moving porous walls can be recast as a nonlinear, variable coefficient boundary value problem, which has applications to mixing processes, as well as for boundary layer control systems. This example follows the work of Motsa (2014), Xu et al. (2010), Rashidi et al. (2014), and considers the fourth-order boundary value problem

$$\left. \begin{aligned} y^{(iv)}(x) + \alpha(xy'''(x) + 3y''(x)) + R_e(y(x)y'''(x) - y'(x)y''(x)) &= 0, \\ y(0) = 0, \quad y''(0) &= 0, \\ y(1) = 1, \quad y'(1) &= 0. \end{aligned} \right\} \quad (3.65)$$

Here  $\alpha$  is the non-dimensional wall dilation rate;  $R_e$  is the permeation Reynolds number, which is positive for fluid injection across the boundary, and negative in the presence of suction on the boundaries; and  $y(x)$  is a dimensionless function describing the mass flow as a function of the wall normal direction.

As was discussed previously, the solution to a nonlinear differential equation using Homotopy based techniques can be constructed in terms of a range of auxiliary linear operators, which will serve as the basis of the iterative scheme. To explore the implications of this freedom, a range of auxiliary linear operators can be constructed and compared, which replicate the original nonlinear differential equation to varying degrees. For this problem the following operators will be used

$$\left. \begin{aligned} \mathcal{L}_1[y] &= y^{(iv)}, \\ \mathcal{L}_2[y] &= y^{(iv)} + \alpha(xy''' + 3y''), \\ \mathcal{L}_3[y] &= y^{(iv)} + \alpha(xy''' + 3y'') + R_e(y''' - y''), \\ \mathcal{L}_4[y] &= y^{(iv)} + \alpha(xy''' + 3y'') + R_e(\hat{y}_0 y''' - \hat{y}'_0 y''). \end{aligned} \right\} \quad (3.66)$$

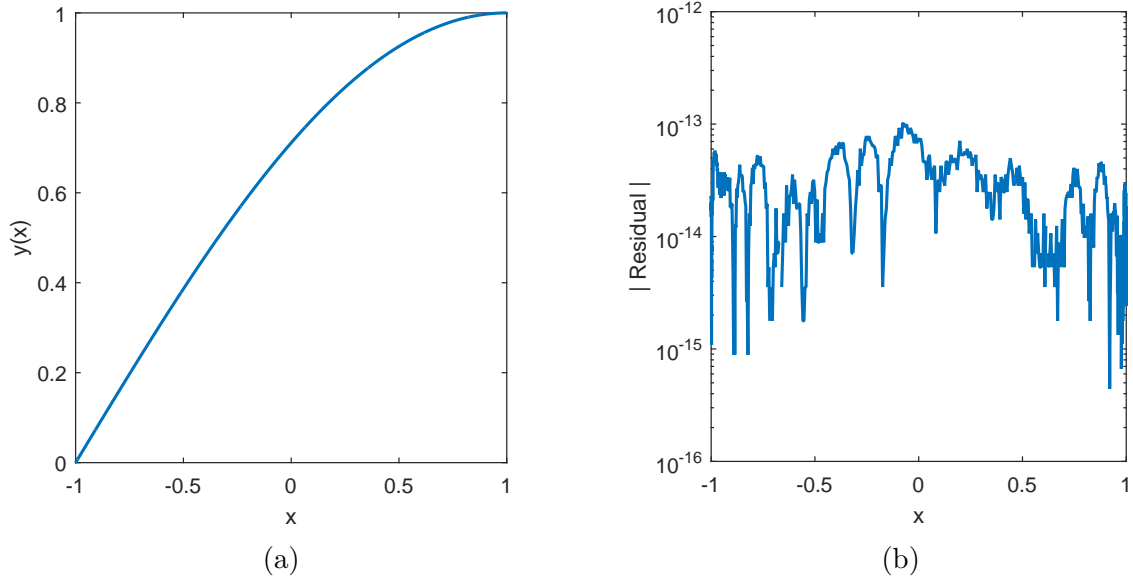


Figure 3.12: Solution and error of equation (3.65) calculated at  $\alpha = 1$  and  $R_e = 10$  using  $\mathcal{L}_4$  from equation (3.66).

Here  $\hat{y}_0$  is the solution to the linear boundary value problem

$$\left. \begin{aligned} \hat{y}_0^{(iv)}(x) &= 0, \\ \hat{y}_0(0) &= 0, \quad \hat{y}_0''(0) = 0, \\ \hat{y}_0(1) &= 1, \quad \hat{y}_0'(1) = 0. \end{aligned} \right\} \quad (3.67)$$

To solve equation (3.65) using these operators, the equations must be mapped to the Chebyshev basis. This can be done by simply using the algebraic mapping  $x = \frac{w+1}{2}$ , so that

$$\left. \begin{aligned} \mathcal{L}_1[y] &= y^{(iv)}, \\ \mathcal{L}_2[y] &= y^{(iv)} + \alpha \left( \frac{w+1}{4} y''' + \frac{3}{4} y'' \right), \\ \mathcal{L}_3[y] &= y^{(iv)} + \alpha \left( \frac{w+1}{4} y''' + \frac{3}{4} y'' \right) + R_e \left( \frac{1}{2} y''' - \frac{1}{4} y'' \right), \\ \mathcal{L}_4[y] &= y^{(iv)} + \alpha \left( \frac{w+1}{4} y''' + \frac{3}{4} y'' \right) + \frac{1}{2} R_e (\hat{y}_0 y''' - \hat{y}_0' y''). \end{aligned} \right\} \quad (3.68)$$

All four choices of linear operator can be used to construct a solution that satisfies this problem, an example of which can be seen in Figure 3.12a, where each operator will exhibit different convergence properties. The primary differences between the techniques

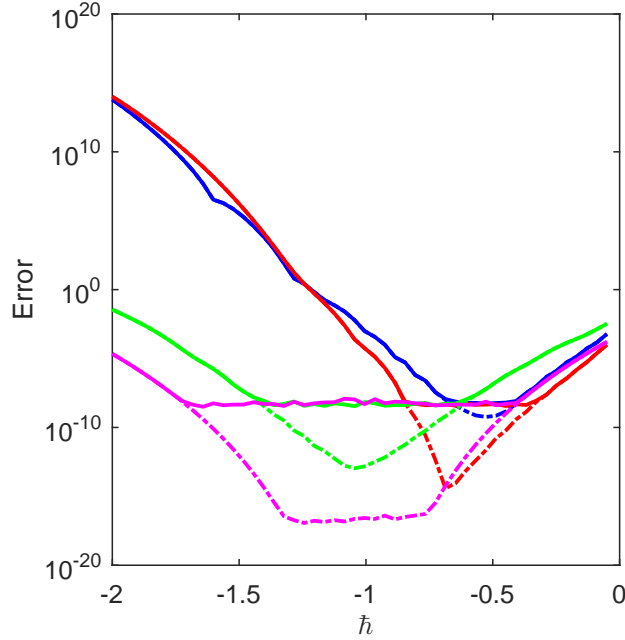


Figure 3.13: Error against  $\bar{h}$  for solutions of equation (3.65), as calculated using the SHAM (solid lines) and the GHAM (dotted lines). The blue lines correspond to  $\mathcal{L}_1$ , the red to  $\mathcal{L}_2$ , the green to  $\mathcal{L}_3$  and the magenta line corresponds to  $\mathcal{L}_4$ , with all operations truncated after 25 iterations.

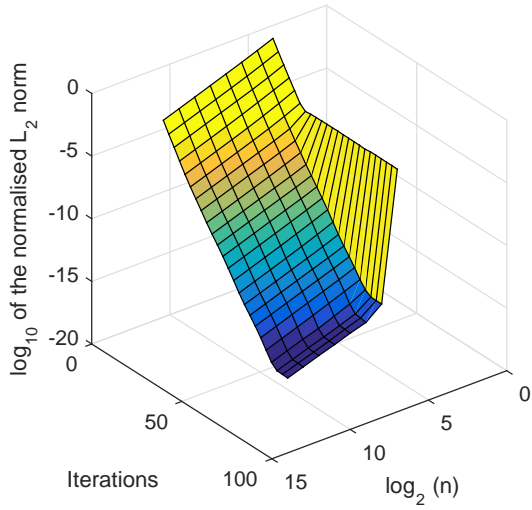
are two fold: firstly the difficulty in constructing solutions numerically and second, their convergence properties. Figure 3.13 allows us to explore the relationship between the choice of  $\bar{h}$  and the error that results for each of the linear operators, where the error is defined as the integral over the residual, i.e.  $\int_0^1 R(x)dx$ . As was discussed in Section 3.1, the optimal value of  $\bar{h}$  can be considered to be the value of  $\bar{h}$  where the error is minimised, for a given number of iterations. However, in this example we see that as the number of iterations increases, the error hits a threshold value, below which the solution from the SHAM will not improve. For solutions of equation (3.65), both the SHAM and the GHAM show that the schemes share the same convergence properties for each linear operator, with the only point of difference being that the implementation of the SHAM used to calculate these results is unable to calculate solutions with greater accuracy than  $\mathcal{O}(10^{-10})$ . As might be expected, as the operator  $\mathcal{L}$  is modified to take a closer form to the full nonlinear equation, the observed convergent region at 25 iterations widens, and, generally, yields higher accuracy solutions. The exception to this is that while  $\mathcal{L}_2$  has a smaller convergent region, at its optimal  $\bar{h}$ —approximately 0.75—the calculated error is markedly lower than

that calculated using  $\mathcal{L}_3$  at its  $\tilde{h}_{\text{opt}}$ . Based upon examining the results from other nonlinear equations, this pattern generally holds—that as long as the auxiliary linear operator  $\mathcal{L}$  reflects, in some basic sense, the original dynamics of the nonlinear equation being solved, then the scheme will be convergent. However by more closely replicating the full nonlinear equation greater convergence properties can be observed—which will have ensuing implications for the computational cost of constructing a solution to the nonlinear problem at hand.

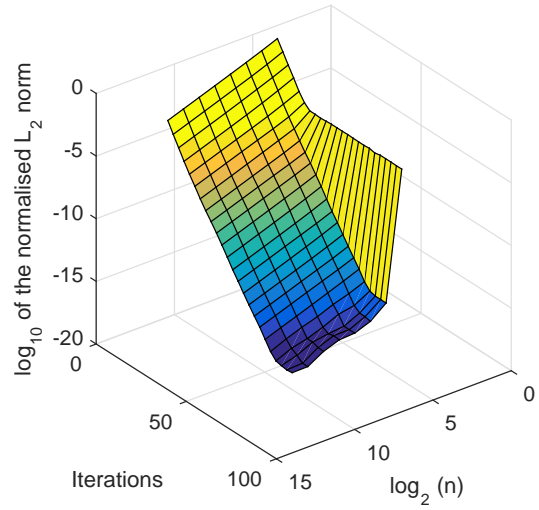
A notable property of the error, as exhibited within Figure 3.13, is that it is broadly convex with respect to  $h$ , a result that has been confirmed through other testing. This convexity is advantageous, as it significantly simplifies the process of searching for  $\tilde{h}_{\text{opt}}$ . This search can be performed at low iterations, with the solution at  $\tilde{h}_{\text{opt}}$  then being iterated upon until a desired error tolerance is satisfied. While there are noticeable regions within Figure 3.13 exhibiting local minima, and as such any optimization technique cannot assume convexity, and be rigorous enough to avoid such small minima.

Figure 3.14 presents an analysis of the error, as driven by both the convergence of the nonlinear problem, and from the linear sub-problems. As the grid resolution increases beyond a threshold value of  $n = 2^4$ , the error of the solution becomes nominally independent of resolution in the manner of Figure 2.1b. This results from the resolution passing the threshold for the spectral convergence of the Gegenbauer method for the linear problem, so that beyond this point changes in the error are driven by the process of solving the nonlinear problem. This can be seen in the change of the convergence of the solution as the number of iterations increases. As long as the threshold resolution has been reached, the primary driver of the convergence of the solution is the number of iterations, up until the point where the numerical solution has converged, within machine precision, to the true solution. While the different auxiliary linear operators of equation (3.68) exhibit different rates of convergence, as will be shown further in Figure 3.16b, the spectral convergence behaviour are still shared across all choices of  $\mathcal{L}$ .

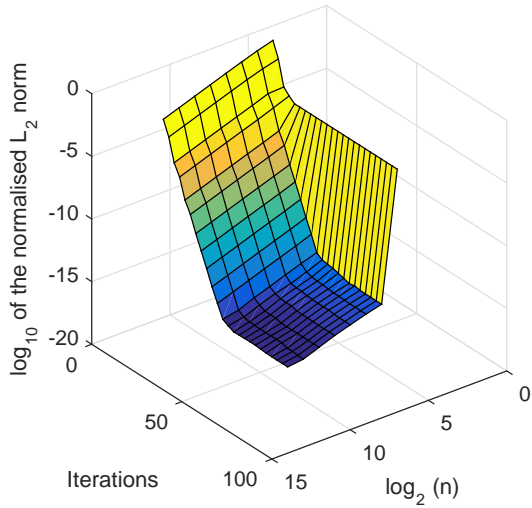
To explore how the dynamics of the GHAM differ from those exhibited by the SHAM, the results from Figure 3.14 were recreated using the SHAM in Figure 3.15, where the linear matrix equations are discretised using Chebyshev-collocation matrices. While the convergence results of the SHAM and the GHAM largely seem comparable, there are some fundamental differences between the performances of the two techniques. Given sufficient



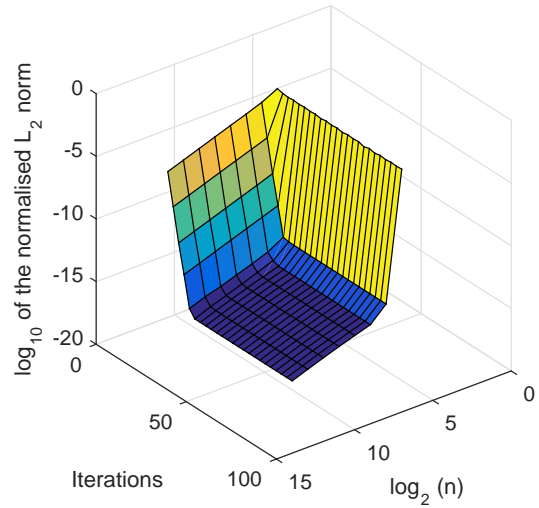
(a)  $\mathcal{L}_1$



(b)  $\mathcal{L}_2$



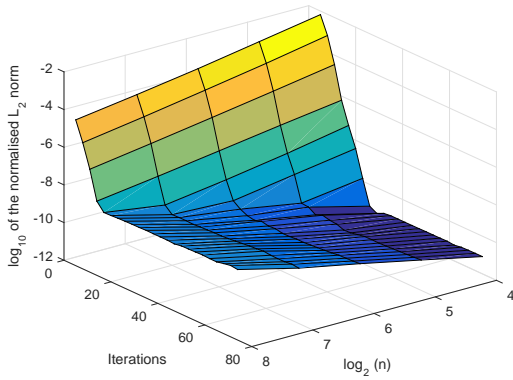
(c)  $\mathcal{L}_3$



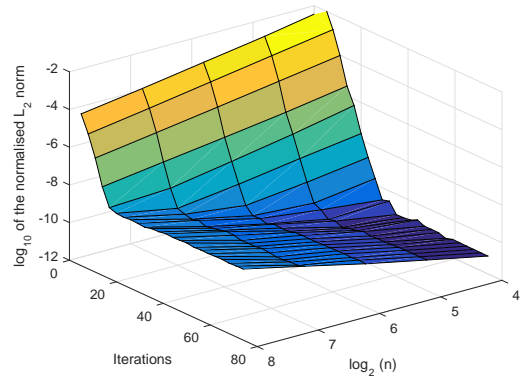
(d)  $\mathcal{L}_4$

Figure 3.14: Calculated error at  $\bar{h}_{\text{opt}}$  for equation (3.65) using the GHAM, as a function of the number of iterations and spatial resolution  $n$ . For these calculations  $H_a = 1$  and  $R_e = 10$ .

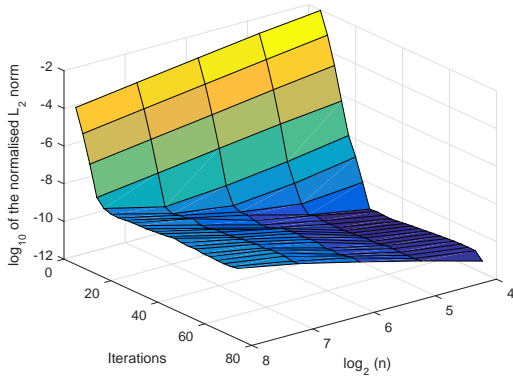




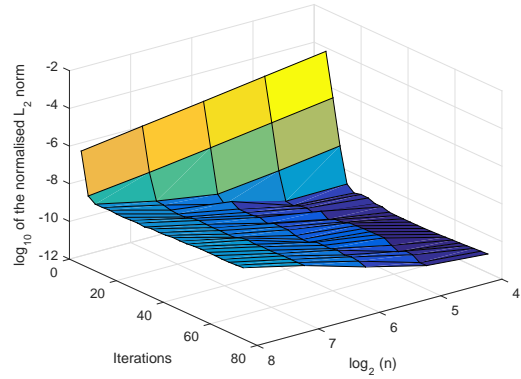
(a)  $\mathcal{L}_1$



(b)  $\mathcal{L}_2$



(c)  $\mathcal{L}_3$



(d)  $\mathcal{L}_4$

Figure 3.15: Calculated error at  $\tilde{h}_{\text{opt}}$  for equation (3.65) using the SHAM, as a function of the number of iterations and spatial resolution  $n$ . For these calculations  $H_a = 1$  and  $R_e = 10$ . Note the changed vertical scale, as compared to Figure 3.14.

iterations, the results from the SHAM converge for any spatial resolution, as compared to the GHAM which requires a threshold number of grid points in order to resolve a solution. As the grid resolution increases the accuracy of the GHAM increases, while, somewhat counter intuitively, the accuracy of the SHAM actually decreases. This can be attributed to the Chebyshev collocation matrices that underpin the SHAM becoming singular as the grid resolution increases, which in turn increases the inherent errors in attempting to solve these systems numerically. The factors underpinning this phenomena have been discussed by Trefethen and Trummer (1989), Rothman (1991), Breuer and Everson (1992), Bayliss et al. (1995), Belfert (1997), Peyret (2002), with the main cause being the accumulation of roundoff errors in the calculation of the differentiation matrices, and the effect of the decreased spacing between grid points, especially as  $|x| \rightarrow 1$ . This also explains why the SHAM is unable to match the accuracy exhibited by the GHAM—the latter of which converges to machine precision.

To this point, the convergence of the GHAM and the SHAM has been considered in the context of their accuracy, and the convergence of the residuals of the calculated solutions as a function of the grid resolution and the number of iterations employed. However, the primary advantage of the sparse matrix formulation of the GHAM is its low theoretical scaling of computational time with respect to the grid resolution. Figure 3.16 considers the performance of both the GHAM and the SHAM at  $h_{\text{opt}}$  for solving equation (3.65), subject to a selection of linear operators.

For the GHAM solver, the solution in terms of  $\mathcal{L}_4$  significantly outperforms all other schemes, converging to machine precision in only 20 iterations. While its pre-eminence is unsurprising, given how closely  $\mathcal{L}_4$  approximates the full nonlinear problem, the degree to which it outperforms the other choices of  $\mathcal{L}$  is surprising. In contrast to this, the remaining choices for  $\mathcal{L}$  all take approximately three times as many iterations in order to converge—however all three still clearly exhibit spectral convergence. Interestingly,  $\mathcal{L}_2$  slightly outperforms  $\mathcal{L}_1$  and  $\mathcal{L}_3$ , even though  $\mathcal{L}_3$  can be considered to be a more accurate linear representation of the nonlinear problem. These differences are maintained when considering the computational time required for all the solutions.

The relative performance of each of the linear operators was maintained within the SHAM. However, there is a marked difference in the minimum error exhibited by the GHAM, as

compared to the SHAM. While the GHAM is able to converge to machine precision, the SHAM cannot resolve solutions beyond a threshold error, and that threshold is not constant as the resolution is changed.

To explore the relative numerical performance of these schemes, two other numerical approaches were taken to solve the motivating equation. The first was Newton Iteration based upon a Gegenbauer discretisation (as described by Olver & Townsend Olver and Townsend (2013)), in order to separate out the contribution to the numerical efficiency from the nonlinear approach, and then the linear solver that both approach is built upon. The second comparison was to MATLAB’s ‘BVP4C’ routine. The latter comparison should theoretically be unfavourable towards GHAM, as BVP4C is a highly optimised routine written in compiled C, as compared to the GHAM codebase which is in terms of uncompiled MATLAB code. However, with both approaches, as shown in Figure 3.17, GHAM for the optimal choice of  $\mathcal{L}$  and  $h$  is multiple orders of magnitude faster than the comparison approaches. Even the dense matrices of SHAM outperform Newton iteration across all tested resolutions, due to Newton iteration necessitating the construction of a new dense matrix operator at each step.

As the resolution of the grid discretisation is doubled, the time required to solve using both the SHAM and Newton–Iteration increases by almost an order of magnitude, whereas there is only an approximate doubling in the computational time required when the calculations are performed with the GHAM. This result has been replicated across other resolutions, and stems from the rate of growth of the number of elements in the sparse matrix operators used within the GHAM, as compared to all the other tested techniques.

The relative difference between the numerical performance of these schemes is maintained across varying resolutions for the grid discretisation, with the GHAM significantly outperforming the SHAM at nearly all resolutions tested within Figure 3.18. For small resolutions, the cost of establishing the matrix system dominates, which results in the SHAM exhibiting a relatively lower computational time. However, as  $n$  increases, the cost of solving the system begins to dominate, leading to the GHAM starting to significantly outperforming the SHAM. An additional consideration of note is that unlike the GHAM, the SHAM could not be considered for grid discretisations larger than  $2^9$  points, as the matrices beyond this point become singular. More broadly, as  $n$  increases, the condition number of the SHAM

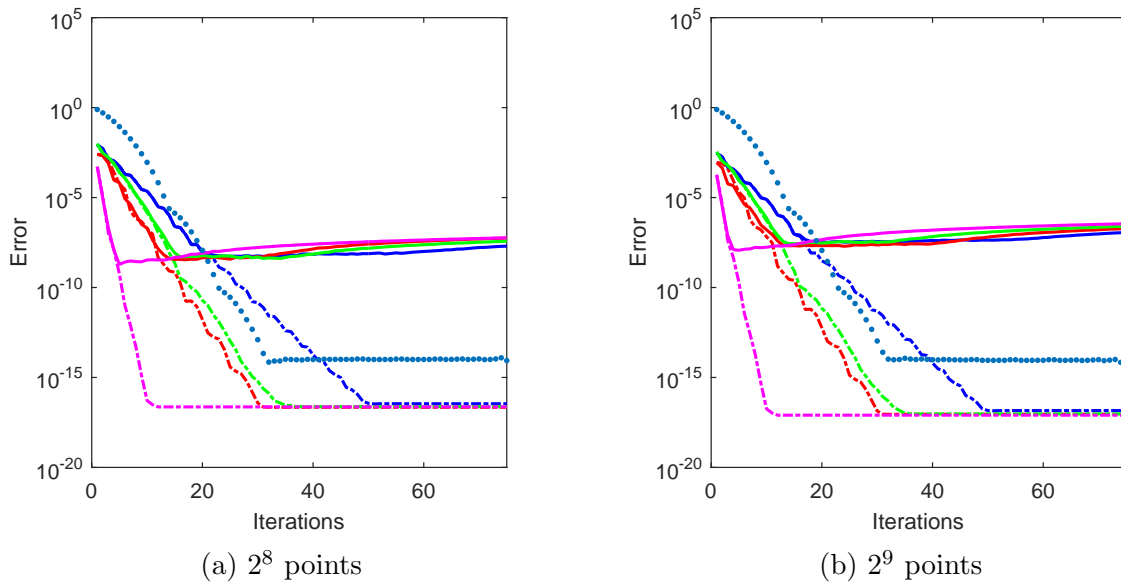


Figure 3.16: Impact of the number of iterative steps upon the numerical error when solving equation (3.65) for varying choices of the auxiliary linear operator  $\mathcal{L}$ . Solid lines correspond to the SHAM, dashed denote the GHAM, and the blue dots are Newton Iteration upon a Gegenbauer discretisation. Of the solid lines, blue represents  $\mathcal{L}_1$ , Red is  $\mathcal{L}_2$ , Green is  $\mathcal{L}_3$  and Magenta is  $\mathcal{L}_4$ . All solutions were calculated at the optimal  $\hbar$  for each choice of  $\mathcal{L}$ .

and GHAM matrices both increase quadratically, but for our test problems the SHAM condition number was consistently larger than those seen by the GHAM discretisation by a factor of between  $10^3$  and  $10^4$ . The differences between the condition numbers exhibited by these two schemes broadly explains the growth in the error shown within Figure 3.18b.

To delve into the factors that drive the computational cost for the GHAM, the process of solving equation (3.65) can be separated out into its constituent parts. The process first involves discretising the system using Gegenbauer polynomials, followed by solving the matrix system at each iteration, converting between Chebyshev space and real space and evaluating the derivatives  $\{y^{(1)}, y^{(2)}, y^{(3)}, y^{(4)}\}$ .

To verify the previously outlined theoretical results for how the GHAM scales with the number of iterations, a regression to the form of  $t = CI^S$  was conducted, where  $I$  is the number of iterations and  $C$  and  $S$  are proportionality constants. This regression was based upon the computational time to calculate a solution between 40 and 80 iterations,

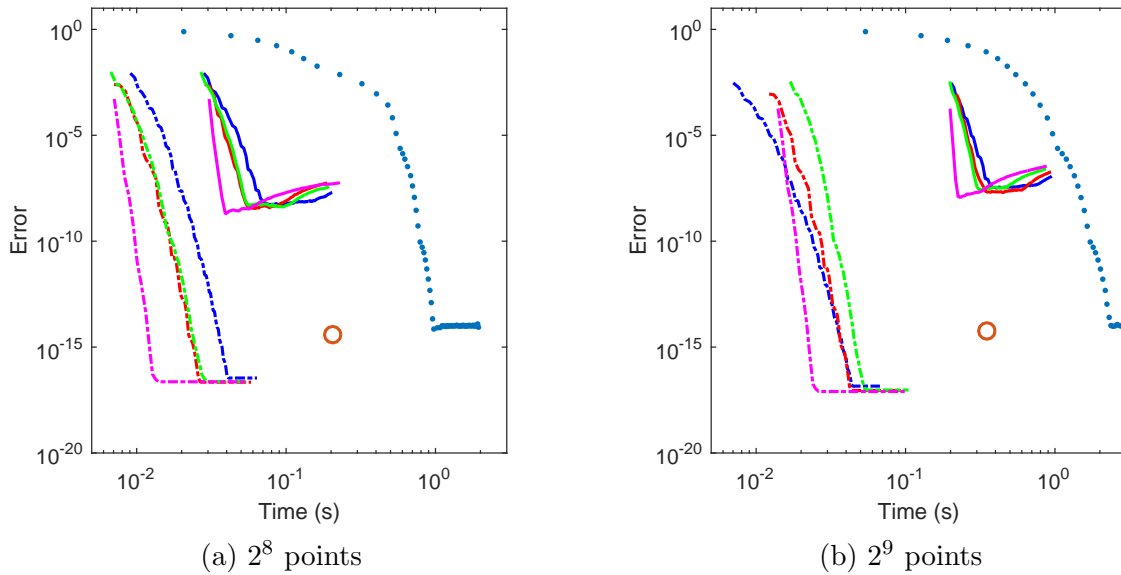


Figure 3.17: Computational cost and the error of solving equation (3.65) for varying numbers of iterations, following Figure 3.16 with the inclusion of the computational cost and error using MATLAB's 'BVP4C' routine, as represented by the red circle.

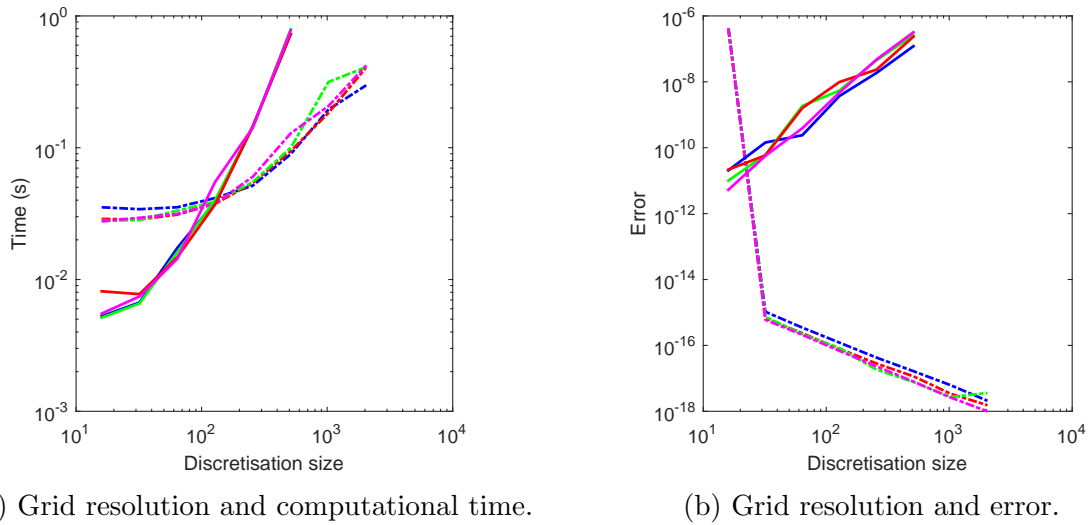


Figure 3.18: Scaling of computational time for solutions of equation (3.65) calculated by taking 75 iterations of the SHAM (solid lines) and the GHAM (dotted lines). Results for the GHAM are presented for  $n \in [2^6, 2^{14}]$ , whereas the SHAM was only calculated over  $n \in [2^6, 2^9]$ , as the Chebyshev collocation matrices within the SHAM become singular after this point.

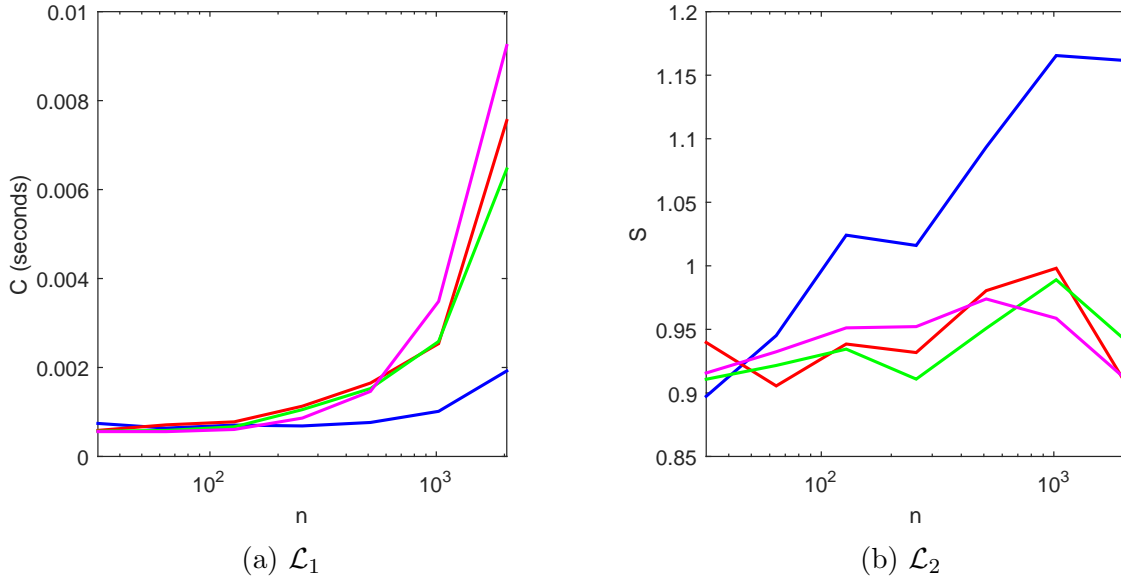


Figure 3.19: Scaling coefficients for the GHAM, assuming that the computational cost scales as  $T = CI^S$ . One again, blue, red, green and magenta represent  $L_1$ ,  $L_2$ ,  $L_3$  and  $L_4$  respectively.

in order to isolate the effect of the iterative process from the cost of evaluating the first step of the scheme, and to examine the scaling properties of the cost of constructing the inhomogeneous component of the matrix equations. This latter point is important to understand, as it should, theoretically, become more costly as  $m$ , the iteration number, increases. For each linear operator  $C$  scales with  $n$ , which reflects the cost of solving the matrix operations after  $LU$  decomposition, and how that grows with the number of grid points in the discretisation. However,  $S$  is approximately 1 for all linear operators, across all grid resolutions, reflecting that the cost of the scheme is entirely driven by the  $LU$  matrix operations, rather than any other components of the iterative process.

Clearly the dominant component of the computational cost stems from the grid resolution, rather than the number of iterations required to solve the problem. To further explore the scaling of computational cost, the results for the GHAM in Figure 3.18a can be decomposed into its constituent parts, the results for which are presented in Figure 3.20. It must be emphasised in this image that the scaling of the purple line, corresponding to solving the matrix inverse directly, has been presented for the purposes of comparison, and is not required for constructing solutions with the GHAM. The dominant contributors to solving

this numerical problem are the cost of performing the  $LUPQR$  decomposition; solving the matrix system after the  $LU$ -decomposition has been employed; evaluating the first four spatial derivatives; and the transform between real and Chebyshev space and visa-versa. While the cost of performing the  $LUPQR$  decomposition would appear to be unfavourably high with respect to the direct cost of solving the matrix systems, as shown in Figure 3.20, it must be stressed that directly solving the matrix would be required at every step of the iterative process, whereas the  $LUPQR$  decomposition only has to be performed once, requiring  $\mathcal{O}(n^{1.305})$  time at the beginning of the iterative process.

The other components of Figure 3.20 do need to be performed at each step of the iterative process, and it is their cost that influences how the scheme scales with the number of iterations—behaving in aggregate as an  $\mathcal{O}(n^{1.05})$  process. This quasi-linear scaling is entirely a product of the low fill-in density of the matrix operators that make up the GHAM.

Based upon the results from equation (3.65) it can be said that—at least for the tested fourth-order systems—the scheme scales with  $\mathcal{O}(n^{1.305} + mn^{1.05})$ , where  $n$  is the spatial resolution and  $m$  is the total number of iterative steps required to both find  $\tilde{h}_{\text{opt}}$ , and then to refine the solution to a prescribed error tolerance. This compares favourably to the SHAM, as solving dense matrix systems is limited by the theoretical scaling  $\mathcal{O}(n^{2.373})$ . Interestingly, while it was predicted in the preceding section that there would be nonlinear scaling with the number of iterations, stemming from the  $\mathcal{O}(m^2n)$  multiplication operations, in practice this term is dominated by the costs involved in solving the matrix system, which scale linearly with iterations. This result has been tested up to 150 iterations, at which point the system still appears to scale linearly with  $m$ .

### 3.6 A priori estimation of $\tilde{h}$

To this point, both the HAM and its numerical variants have been framed as optimisation problems in  $\tilde{h}$ , for which the error, defined by equation (3.8), must be minimised. The corresponding value of the convergence control parameter for this optimisation process is  $\tilde{h}_{\text{opt}}$ , which is the value of  $\tilde{h}$  which converges the fastest.

Alternate approaches have previously been proposed, which consider other metrics to serve as proxies for the convergence of the error, however irrespective of the approach taken there

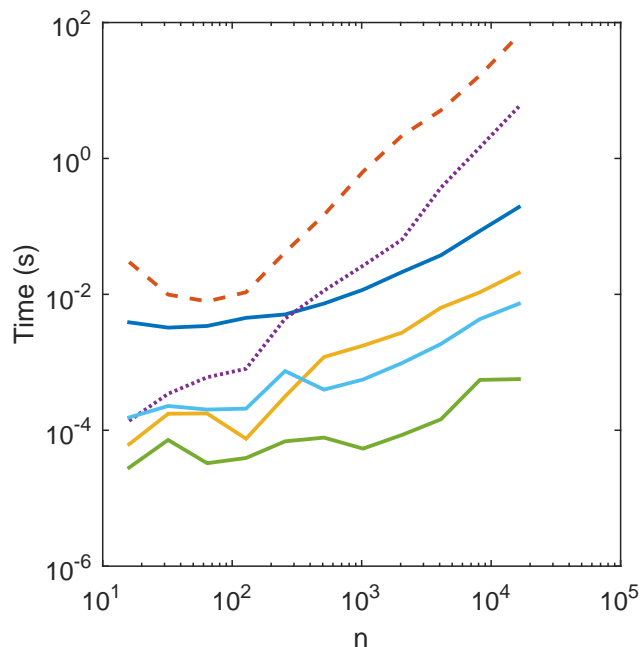


Figure 3.20: Scaling of computational cost of constructing solutions for equation (3.65) with  $n$ . Blue: set up cost for establishing the matrix problem in the context of the GHAM. Red, dashed: LUPQR decomposition, which only needs to occur once at the beginning of the iterative process. Yellow: solving the matrix system using the LUPQR decomposition. Green: transform between real and Chebyshev space. Light blue: calculating derivatives up to fourth-order. Purple, dotted: solving the matrix system using a direct matrix inverse, which is not used within the algorithm.

must be some degree of consideration of the influence of  $\hbar$ . While Theorem 3.4 presents novel results demonstrating why  $\hbar$  must be bounded within  $[-2, 0)$ , the iterative process to construct solutions to a nonlinear equation still necessitates repeatedly solving for varying  $\hbar$ . This process creates an impost for the HAM and, more significantly its numerical analogues.

To try to further limit the bounds of the convergence control parameter for a particular problem, the implications for the choice of  $\hbar$  can be considered with respect to its influence upon the rate of convergence of the sequence  $S_m = \sum_{i=0}^m U_i$ . As was discussed in Theorem 3.1, a necessary condition for convergence of this sequence was that for all  $m$  the inequality

$$\|U_m\| \leq r \|U_{m-1}\|$$



must hold for some  $r$  bounded between 0 and 1, subject to some suitable norm  $\|\cdot\|$ . The parameter  $r$  can then be determined by evaluating

$$r = \max \left( \frac{\|U_1\|}{\|U_0\|}, \frac{\|U_2\|}{\|U_1\|}, \dots, \frac{\|U_m\|}{\|U_{m-1}\|}, \dots \right). \quad (3.69)$$

This parameter can also be tied to the convergence properties of methods based upon the HAM through Theorem 3.3, as the error  $\epsilon_m$  of the partial sum  $S_m = \sum_{i=0}^m u_i$  must behave as

$$\epsilon_m = \|S_m - u\| \leq \frac{r^{(m+1)}}{1-r} \|U_0\|.$$

Thus, it follows that  $\hat{h}_{\text{opt}}$  should correspond to the minima of  $\hat{r} = r^{(m+1)}/(1-r)$ , which will occur at the minima of  $r$ . As such it follows that

$$\hat{h}_{\text{opt}} \approx \hat{h} = \min_{h \in (-2,0)} r. \quad (3.70)$$

However, searching for the minima of  $\hat{r}$  displays the same inherent problems as minimising the error, as  $U_m$  is a nonlinear function that incorporates  $\{\hbar, U_0, U_1, \dots, U_{m-1}\}$ , and as such, we have simply replaced a minimisation problem with a functionally equivalent maximisation. To circumvent this, let us for the moment consider equation (3.69) in terms of the triplet  $\{U_0, U_1, U_2\}$ , so that

$$r \approx \max \left( \frac{\|U_1\|}{\|U_0\|}, \frac{\|U_2\|}{\|U_1\|} \right). \quad (3.71)$$

In the context of a problem formulated as

$$\mathcal{N}[u] = \phi \quad (3.72)$$

then the decomposition of this problem to the infinite set of linear sub-problems for the HAM, subject to an auxiliary linear operator  $\mathcal{L}$  gives

$$\left. \begin{aligned} U_0 &= \mathcal{L}^{-1}\phi \\ U_1 &= \hbar \mathcal{N}_0 \\ U_2 &= U_1 + \hbar \mathcal{L}^{-1}[\mathcal{N}_1] \\ &\vdots \\ U_m &= U_{m-1} + \hbar \mathcal{L}^{-1}[\mathcal{N}_{m-1}]. \end{aligned} \right\} \quad (3.73)$$

In the interests of brevity, the notational shorthand

$$\mathcal{N}_m = \frac{1}{m!} \frac{d^m \mathcal{N}[u]}{dq^m},$$

has been introduced. Through this, equation (3.71) will take the form

$$r = \max \left( \frac{\|\hbar \mathcal{L}^{-1}[\mathcal{N}_0]\|}{\|U_0\|}, \frac{\|U_1 + \hbar \mathcal{L}^{-1}[\mathcal{N}_1]\|}{\|\hbar \mathcal{L}^{-1}[\mathcal{N}_0]\|} \right). \quad (3.74)$$

This relationship does not give any particular insight into the relationship between  $r$  and  $\hbar$ . However, by virtue of its construction ( $U_1/\hbar$ ) must be invariant of  $\hbar$ , and following a similar logic  $\mathcal{N}_1$  must also be linear with respect to  $\hbar$ , so that

$$\mathcal{N}_1 = U_1 \hat{\mathcal{N}}_1 = \hbar \left( \frac{U_1}{\hbar} \right) \hat{\mathcal{N}}_1.$$

The crucial observation here is that  $\mathcal{N}_1$  must be strictly a function of  $U_0$ . As such, equation (3.74) can be re-expressed in terms of these invariants, so that

$$r = \max \left( \frac{\|\hbar \mathcal{L}^{-1}[\mathcal{N}_0]\|}{\|U_0\|}, \frac{\|(\frac{U_1}{\hbar}) + \hbar \mathcal{L}^{-1}[(\frac{U_1}{\hbar}) \hat{\mathcal{N}}_1]\|}{\|(\frac{U_1}{\hbar})\|} \right). \quad (3.75)$$

From our testing almost all choices of  $\mathcal{L}$  yield  $\|U_1\|/\|U_0\| \ll \|U_2\|/\|U_1\|$ , which allows equation (3.71) to be reduced to

$$r \approx \frac{\|(\frac{U_1}{\hbar}) + \hbar \mathcal{L}^{-1}[(\frac{U_1}{\hbar}) \hat{\mathcal{N}}_1]\|}{\|(\frac{U_1}{\hbar})\|}. \quad (3.76)$$

As all of these terms are invariant with respect to  $\hbar$  the minima of  $r$  can be found based upon solving for  $U_1$  at a single choice of  $\hbar$ . As the only contribution with respect to the convergence control parameter is from  $\hbar \mathcal{L}^{-1}[(\frac{U_1}{\hbar}) \hat{\mathcal{N}}_1]$ , which is linear with respect to  $\hbar$ , and as such this approximation to  $r$  will also be strictly linear.

This minima of  $r$  corresponds to the location of  $\hat{\hbar}$ , which approximates  $\hbar_{\text{opt}}$ . Furthermore, this value of  $r$  allows the upper bound of  $\epsilon_m$  to be calculated, under the assumption that equation (3.69) is broadly determined by equation (3.71).

### 3.6.1 Validation

To explore the validity of the algorithm outlined within the previous section, we can return to the problem of a two-dimensional viscous flow within a rectangular domain subject to

porous, moving boundaries—as presented in Subsection 3.5.1. This specific comparison will be considered by constructing solutions with the GHAM in terms of the  $\mathcal{L}_1$  and  $\mathcal{L}_4$  operators.

Both of these test problems showed a strong alignment between the convergence of the error as a function of  $\hbar$ , across a range of iterations, and the predicted location of  $\hbar_{\text{opt}}$ . Figure 3.21a shows the specific results for the  $\mathcal{L}_1$  case of equation (3.68), with the predicted relationship between  $r$  and  $\hbar$  shown in blue, and with the demarcation line for convergence at  $r = 1$  shown with the dashed red line. Using these results,  $\hat{\hbar}$  was found to be approximately  $-0.5327$ . Finding the minima of  $r$ , and thus  $\hat{\hbar}$  was, as was expected, a trivial process, as the relationship between these two variables was linear with respect to  $\hbar$ . More broadly, the convergent region of  $r < 1$  was found for  $\hbar \in [-1.05, 0]$ .

These results were then validated in Figure 3.21b by superimposing the location of  $\hat{\hbar}$ —in black—and the bounds of the convergent region of  $r$ —in green—upon the error of the  $\mathcal{L}_1$  solutions, as constructed at 2, 10, 50 and 100 iterations. While  $\hat{\hbar}$  is an approximation that only considers the triptych of  $\{U_0, U_1, U_2\}$ , the approximation  $\hat{\hbar}$  is strongly correlated with the location of  $\hbar_{\text{opt}}$  for all iterations, with the difference between the two being minuscule down to the point where the profiles are dominated by machine precision error—the point at which the profiles within Figure 3.21b become saturated.

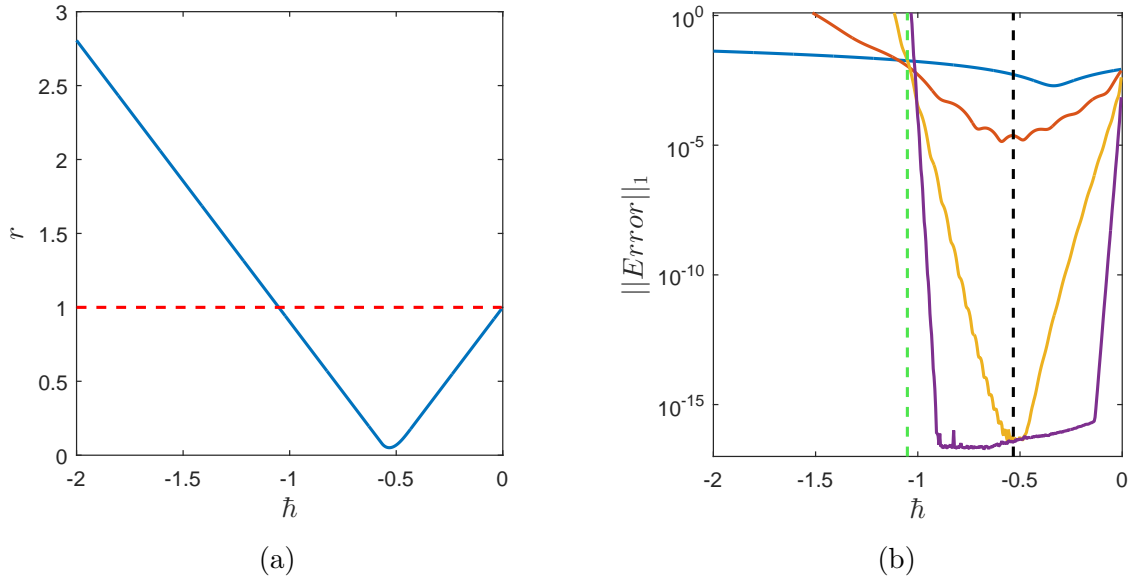


Figure 3.21: Figure 3.21a shows  $r$  against  $\hat{h}$  for the  $\mathcal{L}_1$  operator of Subsection 3.5.1 in blue, with the red line indicating the demarcation point for the region where  $r > 1$ . Figure 3.21b shows the  $L_1$  norm of error at  $\{2, 10, 50, 100\}$  iterations over a grid of  $n = 2^8$  points. The black dashed line indicates the location of  $\hat{h}$ , and the green dashed line is marks the point in  $\hat{h}$  that corresponds to  $r = 1$ .

In the case of the  $\mathcal{L}_4$  operator, Figure 3.22a shows that  $r$  is strictly bounded to lie within  $[0, 1]$  for all  $\hat{h}$ , suggesting that the GHAM solutions constructed with this linear operator should be convergent across all  $\hat{h}$ . This conclusion is borne out by the results within Figure 3.22b. As the number of iterations increases, there is a small but visible difference between  $\hat{h}_{\text{opt}}$  and  $\hat{h}$ , however the difference between the errors at these two values of  $\hat{h}$  is still negligible.

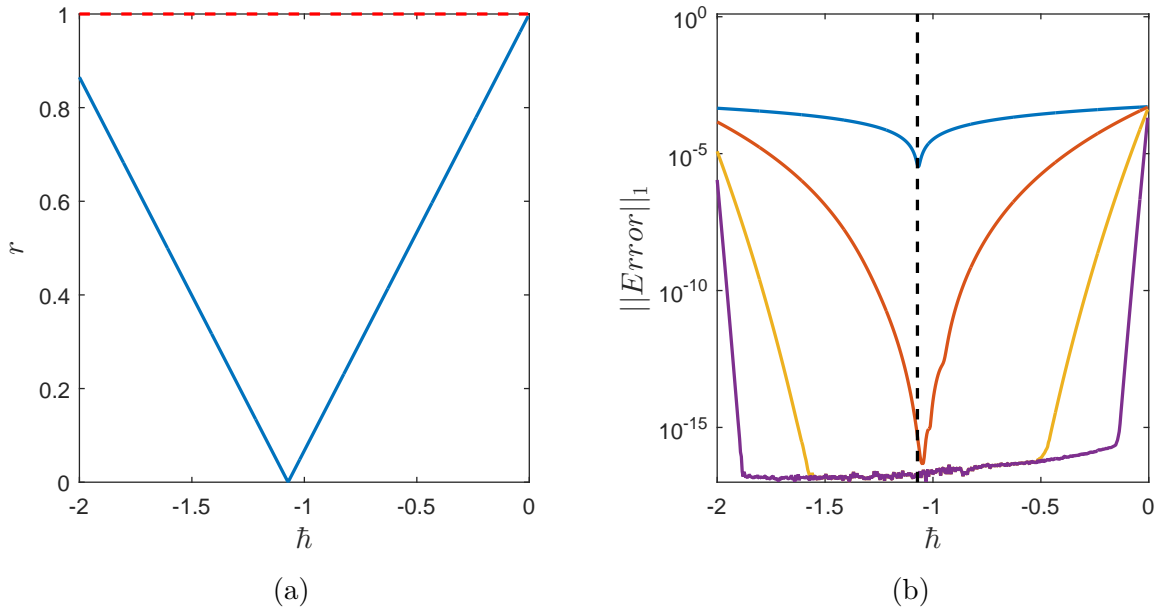


Figure 3.22: (a) shows  $r$  against  $\hat{h}$  for the  $\mathcal{L}_4$  operator of Subsection 3.5.1 in blue, with the red line indicating the demarcation point for the region where  $r > 1$ , and (b) shows the  $L_1$  norm of error at  $\{2, 10, 50, 100\}$  iterations over a grid of  $n = 2^8$  points. The black dashed line indicates the location of  $\hat{h}$ .

The degree of agreement between  $\hat{h}$  and  $\hat{h}_{\text{opt}}$  is remarkable, given that  $\hat{h}$  is constructed in terms of calculations of  $U_0, U_1$  and  $U_2$  conducted at a single  $\hat{h}$ . That  $\hat{h}_{\text{opt}}$  can be accurately approximated has significant implications for numerical techniques based upon the HAM, as there is no longer the need to consider solutions across all  $\hat{h} \in [-2, 0)$ , greatly reducing the amount of calculations required.

This result was the product of realising that it is possible to consider  $U_m$  in terms of invariants of  $\hat{h}$ , an idea that will now be explored further.

### 3.7 Invariants with respect to $\hat{h}$

Up until this point, we have only considered constructing up to  $U_2$  in terms of quantities that are invariant with respect to  $\hat{h}$ . To extend this further, we can consider  $U_3$ :

$$\begin{aligned}
U_3 &= U_2 + \hbar \mathcal{L}^{-1}[\mathcal{N}_2] \\
&= \hbar \left( \frac{U_1}{\hbar} \right) + \hbar^2 \mathcal{L}^{-1} \left[ \left( \frac{U_1}{\hbar} \right) \hat{\mathcal{N}}_1 \right] + \mathcal{L}^{-1}[\mathcal{N}_2] \\
&= \hbar U_2^{(1)} + \hbar^2 U_2^{(2)} + \mathcal{L}^{-1}[\mathcal{N}_2].
\end{aligned}$$

Here we have introduced the notational shorthand  $U_2^{(1)}$  and  $U_2^{(2)}$  to denote the components of  $U_2$  that depend upon  $\hbar$  and  $\hbar^2$  respectively.

Since  $\mathcal{N}_2$  is no longer guaranteed to be linear with respect to  $U_2$  or  $U_1$ , we cannot simply partition these terms out as was done for  $U_2$ . However, by considering  $\mathcal{N}_2$  in the form

$$\mathcal{N}_2 = f \left( U_0, \hbar \left( \frac{U_1}{\hbar} \right), \hbar U_2^{(1)} + \hbar^2 U_2^{(2)} \right),$$

then it follows that  $\mathcal{N}_2$  should be able to be decomposed into terms proportional to  $\hbar$  and  $\hbar^2$ , which in turn means that  $U_3$  can be decomposed into

$$U_3 = \hbar U_3^{(1)} + \hbar^2 U_3^{(2)} + \hbar^3 U_3^{(3)}, \tag{3.77}$$

and more generally that

$$U_m = \sum_{i=1}^m \hbar^i U_m^{(i)}. \tag{3.78}$$

While each  $U_m$  must still be constructed sequentially, they can all be constructed in terms of a sequence of invariant terms. While numerical implementations of this formulation involve a significantly larger memory footprint to store all the invariant terms—from  $\mathcal{O}(mn)$  terms in memory to solve up to  $U_m$  on a grid involving  $n$  points to  $\mathcal{O}(m^2n)$ —it does allow for the solution space across all  $\hbar$  to be explored based upon only the calculations at a single chosen  $\hbar$ . To this point a numerical implementation of this approach has not been finalised to the point where the results warrant inclusion within this work, however initial testing has verified that this technique of constructing solutions in terms of invariant terms is a practical and realistic way of exploring the solution space as a function of  $\hbar$ , without necessitating the computational cost of calculating solutions at a range of values for  $\hbar$ .

### 3.8 Discussion

Across the spectrum of available techniques for solving nonlinear equations, the Homotopy Analysis Method is a recently developed technique that has some uniquely advantageous properties, which were outlined within this chapter. These include a large degree of inherent flexibility for the manner in which the equations are solved, which, in concert with its strong convergence control properties make for a particularly powerful tool for solving nonlinear problems. In its original form as a semi-analytic technique the HAM involves partitioning a nonlinear equation into an infinite series of dependent linear equations. Taking this approach has allowed the technique to successfully solve a range of highly nonlinear problems from fluid mechanics, and has been able to resolve solutions outside the range of traditional techniques.

One of the primary limitations of the HAM is that the linear partitions must be able to be solved algebraically, or through symbolic computer algebra packages. However, this limitation can be avoided by considering the HAM within a numerical framework, where the linear partitions are discretised in terms of spectral matrix operators. Solving the linear equations numerically greatly expands the scope of problems that can be considered within the HAM framework, as the technique is no longer limited to equations that can be solved analytically, or through semi-analytic approaches.

To date, all numerical implementations of the HAM have been in terms of Chebyshev collocation matrices, which are both dense and often become singular, especially when considering variable coefficient boundary value problems. However, by extending the HAM in terms of the Gegenbauer polynomials, we have been able to create a novel numerical technique that involves solving a sequential sequence of linear boundary value problems, discretised in terms of sparse, spectrally accurate matrix operators. When tested on a range of second, third and fourth-order boundary value problems from fluid mechanics, the GHAM was able to accurately and rapidly resolve solutions, while maintaining the flexibility and convergence control properties of the HAM.

Over the course of developing the GHAM, a number of interesting properties of the technique were explored, which have broader applicability within the oeuvre of homotopy based methods. The first of these was what appears to be the first presented result to detail why the convergence control parameter  $\hbar$  must be bounded between  $(-2, 0)$ . While

this is a fundamental bound for the HAM and its numerical analogues, there has to this point not been a full justification for why this bound exists. A second result with significant ramifications is the demonstration within Section 3.6 that the optimal value of  $\hbar$ —for which the HAM based solutions converge the fastest—can be estimated based upon calculations conducted at a single, arbitrary choice of  $\hbar$ , due to reformulating the HAM in terms of invariants with respect to the convergence control parameter. This latter result is particularly interesting, as it allows the HAM and its extensions to be re-framed from being optimisation problems in terms of  $\hbar$  to a single iterative process conducted at  $\hbar_{\text{opt}}$ .

However, the most striking result presented within this chapter relates to how the performance of the GHAM scales with respect to the number of iterations, and the resolution of the numerical discretisation. From a theoretical perspective, solving a nonlinear problem with the GHAM scales with  $\mathcal{O}(n)$  operations, for a problem defined upon a Chebyshev grid of  $n$  points. The performance of the HAM is a product of being able to define all the matrix equations in terms of a single, constant matrix operator, which is inverted repeatedly across the process of solving the motivating nonlinear equation. This in turn allows the cost of solving the matrix equation to be amortised within the overall cost of the iterative scheme, by incorporating techniques such as LU decomposition. That the matrix operator is able to remain constant, while still rapidly iterating towards a spectrally accurate solution is unique within the field of numerical analysis.

These theoretical results were broadly confirmed by comparing the scaling properties of the GHAM for both the Falkner–Skan equation and a second, fourth–order boundary value problem with solutions constructed via the SHAM, Newton–iteration upon a Gegenbauer discretisation and MATLAB’s inbuilt ‘BVP4C’ routine. The GHAM significantly outperformed all the other tested methods, exhibiting a computational complexity which scaled with  $\mathcal{O}(n^{1.305} + mn^{1.05})$ , where  $m$  is the number of iterations required. Additional testing of a second–order boundary value problem—not contained within this work—exhibited  $\mathcal{O}(n^{1.105})$  scaling, reinforcing the results contained within this chapter.

In contrast to other numerical techniques for solving nonlinear equations, employing the HAM as a the framework for a numerical technique allows for the motivating problem to be solved by iterating upon a single, constant matrix operator. As such, the computational cost inherent in solving the matrix equation can be amortised across the overall cost of the



iterative scheme by incorporating techniques like LU decomposition.

To our knowledge, the quasi-linear scaling of the GHAM is unheralded among solvers for nonlinear differential equations. Other schemes typically have a lower bound on scaling of  $\mathcal{O}(n^{2.373})$ , as they require solving a dense  $\mathbb{R}^{n \times n}$  matrix system multiple times across an iterative process, as compared to the GHAM which, through careful construction, only requires a sparse matrix system to be solved for once.

The utility of the GHAM was augmented through the introduction of a suite of novel numerical continuation approaches, which can also be applied to eigenvalues problems. Based upon the initial tests presented within this chapter, Homotopic integrated arc-length continuation appeared to be the most promising. This technique unifies homotopy continuation methods and arc-length based methodologies by smooth paths within the bifurcation diagram that approximate the true path. The most striking feature of Homotopic integrated arc-length continuation was its ability to take significantly larger steps than other schemes, even in the presence of folds. This in turn means that far fewer steps are required to fully traverse the parameter space of any tested problem.

While the scheme was able to traverse folds, further testing will be required to explore the schemes properties in the presence of bifurcations. Another open question is the relative cost of the numerical continuation schemes presented within this chapter, as compared to more standard methods. Such comparisons are difficult due to inherent sensitivity to the choice of step length, and to the incorporation of any adaptive step sizing. While such comparisons have been attempted, they have been omitted due to the difficulty of comparing the computational cost of continuation regimes.

Having now developed the GHAM sufficiently as a suite of numerical tools, and having established its validity as a numerical technique for solving a range of boundary value problems from fluid dynamics, we can now turn to considering its application to the forced Korteweg-de Vries and Gardner equations.

# Chapter 4

## Weakly nonlinear waves: solutions of the forced Korteweg–de Vries Equation

The applicability of the numerical techniques introduced within Chapter 3 will now be assessed by turning to the steady, symmetric variant of the fKdV equation, as was introduced in Chapter 1. As this equation has been well studied, it is a particularly suitable problem for considering the validity of the GHAM for solving boundary value problems within the field of geophysical wave problems. This is not to say that there are no open problems for this particular equation, and as such this chapter will test the applicability of a widely held assumption regarding the influence of the topographic length scale upon the solution space admitted by the fKdV equation.

To begin with, we can consider solutions  $A(x, t)$  of the fKdV equation of the form

$$\left. \begin{aligned} A_t + \Delta A_x + 2rAA_x + A_{xxx} &= -\gamma f_x(x), \\ A(\pm\infty) &= 0. \end{aligned} \right\} \quad (4.1)$$

Specifically, the steady variant of the fKdV equation—where  $A_t = 0$ —will be considered, and thus by integrating equation (4.1)  $A(x)$  can be determined by solving

$$\left. \begin{aligned} \Delta A + rA^2 + A_{xx} &= -\gamma f(x) \\ A(\pm\infty) &= 0, \end{aligned} \right\} \quad (4.2)$$

which has been the subject of rigorous study by (Ee and Clarke, 2007, 2008, Wade, 2015), amongst others, making it a perfect candidate for both validating the GHAM and as the basis for further exploration of the behaviour of the fKdV equation. When examined in terms of the fundamental non-dimensional variables of this problem,  $\Delta$  can be related to the Froude number, which represents the balance between the freestream velocity and the linear wave speed. The most commonly used variant is the depth based Froude number

$$F = \frac{U}{\sqrt{gH}}$$

in which the wave speed is calculated in terms of the gravity  $g$  and unperturbed channel depth  $H$ . When the Froude number is less than 1, the flow is considered subcritical; when it is equal to 1 then it is transcritical; and the case where  $F$  is greater than unity falls under what is known as the supercritical regime. Linear theory predicts that for supercritical flows the equation will only admit symmetric solutions (Stoker, 1957), and as such for our study we will begin by restricting ourselves to this regime. For simplicity's sake the fKdV equation will be interpreted in terms of  $\Delta$ , which can be related to its equivalent Froude number through the mapping  $\Delta = -6(F - 1)$ .

This specific flow regime of symmetric supercritical solutions of the fKdV equation has been approached using Boundary-Integral methods to solve for perturbations to the uniform freestream flow induced by a bump (Forbes and Schwartz, 1982, Vanden-Broeck, 1987); a triangular obstruction (Dias and Vanden-Broeck, 1989); and for flows past trenches of arbitrary geometry (Shen, 1991) among others. Our approach presented herein considers constructing numerical solutions for the fKdV equation, and using these solutions, through the use of numerical continuation, to explore the parameter space of solutions, as well as the degree of qualitatively different solutions that can be admitted.

However, before beginning to approach this problem numerically, a qualitative understanding of the solution dynamics can be developed by considering its homogeneous case, the KdV equation. As this equation is integrable, the equation can be considered in context of the reduced form

$$A_x^2 = -\Delta A^2 - \frac{2}{3}rA^3 + C. \tag{4.3}$$

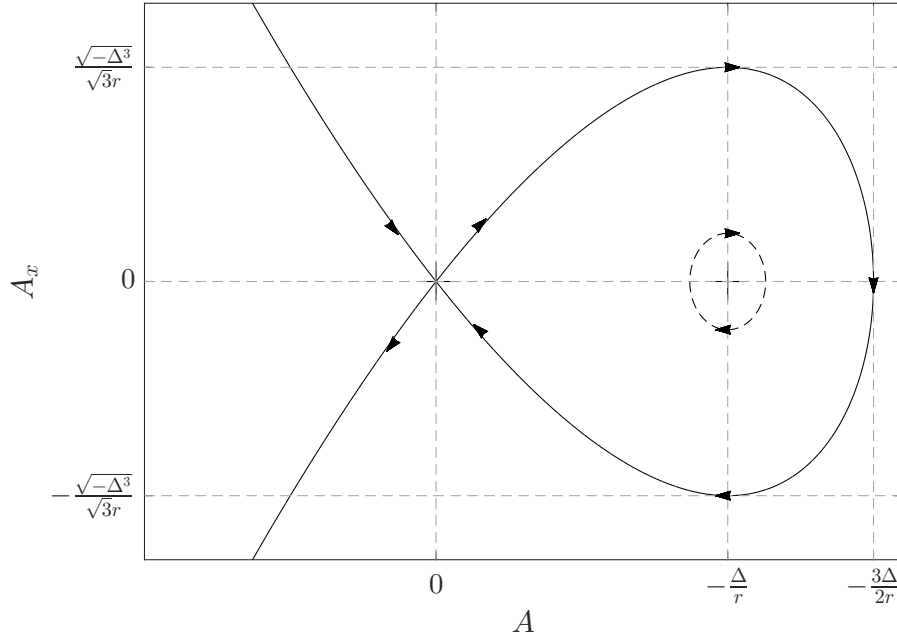


Figure 4.1: Phase portrait of the Korteweg–de Vries equation for  $\Delta < 0$ , with the +’s corresponding to the locations of the stationary points.

The fundamental dynamics of equation (4.2) can be understood by examining the phase portrait of the homogeneous form, as shown in Figure 4.1. This portrait contains two critical points at  $(A, A_x) = (0, 0)$  and  $(-\Delta/r, 0)$ . If  $\Delta < 0$  then these are a saddle and a stable centre respectively, or a centre and a saddle if  $\Delta > 0$ . As was discussed in Section 1.1, the symmetries inherent within the fKdV equation allows the parameter space to be reduced, and as such  $\Delta < 0$  can be imposed without any loss of generality.

While the fKdV equation can admit asymmetric and periodic solutions, we will begin by considering the case of aperiodic symmetric solitary wave solutions to equation (4.2), for which the phase portrait only admits the boundary conditions

$$A(\pm\infty) = A_x(\pm\infty) = 0.$$

## 4.1 Topographic forcing with local support

One approach to deriving the fKdV equation in the presence of topographic forcing is to scale the equation by the height and the length scale of the topographic obstruction. Assuming that the length scale of the height is small relative to the horizontal scaling, then the forcing function  $f(x)$  can be reduced to the Dirac delta function  $\delta(x)$  centred about  $x = 0$ , which we will define as being subject to local support. As a result of this, it can be imposed that  $f(x) = 0$  for all  $x$  except  $x = 0$ , and from this the fKdV equation can be integrated over the region where  $x < 0$  and  $x > 0$ , to give solutions of the form

$$\left. \begin{aligned} A(x) &= -\frac{3\Delta}{2r} \operatorname{sech}^2 \sqrt{\frac{-\Delta}{4}}(x - L_+), & x \geq 0, \\ A(x) &= -\frac{3\Delta}{2r} \operatorname{sech}^2 \sqrt{\frac{-\Delta}{4}}(x - L_-), & x < 0. \end{aligned} \right\} \quad (4.4)$$

Given that solutions to the fKdV equation must be continuous, it must be that

$$\lim_{x \rightarrow 0^+} A(x) = \lim_{x \rightarrow 0^-} A(x),$$

and so in turn it follows that the offset parameter  $L$  must take the form

$$L_+ = \pm L_-.$$

While delta function forcing does not introduce any discontinuities to  $A(x)$ , the discontinuous derivatives for such forcing give rise to the jump condition  $A'(0^+) - A'(0^-) = \gamma$ . In order to satisfy this condition, it must be that solutions of equation (4.2) of the form of equation (4.4) must be subject to

$$A(0) \left( \tanh\left(\frac{\sqrt{-\Delta}4s}{L_-}\right) - \tanh\left(\frac{\sqrt{-\Delta}4s}{L_+}\right) \right) = -\frac{\gamma}{2\sqrt{-\Delta}}. \quad (4.5)$$

The above equation can only be satisfied for non-zero  $\gamma$  if  $L_+ = -L_- = L_0$ . This in turn allows equation (4.5) to be written as

$$f^3 - f = c \quad (4.6)$$

where

$$\left. \begin{aligned} f &= \tanh\left(\sqrt{\frac{-\Delta}{4}}L_0\right) \text{ and} \\ c &= -\frac{\gamma r}{3\Delta} \frac{1}{\sqrt{-\Delta}}. \end{aligned} \right\} \quad (4.7)$$

Solving this equation for  $L_0$  gives rise to a range of solutions of equation (4.4). If  $L_0 > 0$  then this solution denotes a cusped solitary wave where the cusp is concave up, and if  $L_0 < 0$  then the cusp is concave down. The number of real, distinct solutions for  $L_0$  is an implicit function of  $f$ , and as such as equation (4.6) is a cubic equation in terms of  $f$  there are at most three distinct solutions for  $L_0$ . In the case where  $|c| < \frac{2}{3\sqrt{3}}$  then there will be two distinct roots for  $f$  within  $(-1, 1)$ ; if  $|c| > \frac{2}{3\sqrt{3}}$  then there will be only one distinct root for  $f$  within  $(-1, 1)$ ; and when  $|c| = \frac{2}{3\sqrt{3}}$  then there will be a double-root for  $f$  within  $(-1, 1)$ . These values of  $c$  will correspond to what can be labelled as the critical values for  $\gamma$ , as the transition of  $|c|$  from being less than  $\frac{2}{3\sqrt{3}}$  to greater than this value will produce a critical change in the the calculated solutions. In fact, these values will correspond to the location of turning points in the bifurcation diagram. These critical values will occur at

$$\Delta_C = \left(\frac{-3r^2\gamma^2}{4}\right)^{\frac{1}{3}} \quad (4.8)$$

which can be expressed in terms of  $\gamma$  to give that

$$\gamma_C = \pm \frac{2\Delta\sqrt{-\Delta}}{r\sqrt{3}} \quad (4.9)$$

These results suggest that any bifurcation diagram generated in terms of  $\gamma$  should exhibit self-similarity with respect to linear variations in  $\Delta$ .

However, these solutions only admit  $A(0) > 0$  for all  $|\gamma| < \gamma_C$ , and yet previous numerical work by Wade (2015) has shown that it is possible to find solutions which are strictly negative across all  $x$ . This in turn suggests that there is a solution to the fKdV equation, subject to  $\delta(x)$  forcing which admits a strictly negative solution. It has been previously shown that cosech<sup>2</sup> solutions exist for the fKdV equation, and using these we can show that the fKdV equation also admits soliton solutions of the form

$$\left. \begin{aligned} A(x) &= \frac{3\Delta}{2r} \operatorname{cosech}^2 \sqrt{\frac{-\Delta}{4}}(x - L_+), x \geq 0, \\ A(x) &= \frac{3\Delta}{2r} \operatorname{cosech}^2 \sqrt{\frac{-\Delta}{4}}(x - L_-), x < 0. \end{aligned} \right\} \quad (4.10)$$

To solve for  $L$  we proceed in a similar manner to the  $\operatorname{sech}^2$  solutions, so that by continuity we have that  $L_+ = \pm L_-$ . By then applying the jump condition we find that

$$-2\sqrt{\frac{-\Delta}{4}} \left( \coth \left( -\sqrt{\frac{-\Delta}{4}}L_+ \right) - \coth \left( -\sqrt{\frac{-\Delta}{4}}L_- \right) \right) A(0) = -\gamma.$$

Once again this only admits solutions when  $L_+ = -L_- = L_0$ , which in turn gives that

$$6\frac{\Delta}{r} \sqrt{\frac{-\Delta}{4}} \cosh \left( \sqrt{\frac{-\Delta}{4}}L_0 \right) = -\gamma \sinh^3 \left( \sqrt{\frac{-\Delta}{4}}L_0 \right).$$

Solving this allows for an analytic description of all the possible solutions for the fKdV equation with local support to be found. For the remainder of the chapter, all the solutions presented will be for the specific supercritical case of  $A(x)$  where

$$A_{xx} - 1.7454A + \frac{9}{2}A^2 = -\gamma f(x) \quad (4.11)$$

This choice of parameter space can be more broadly generalised through the results contained within Section 1.1. Changing the sign of  $\Delta$ , so that  $\Delta \rightarrow -\Delta'$  provokes a change in  $\eta$  to  $\eta' - \Delta/r$ , thus rescaling the solution space, changing the nature of the stationary points of the system. More broadly, making the change from  $\Delta$  to  $c\Delta'$  requires the rescalings  $r \rightarrow cr'$ ,  $x \rightarrow \sqrt{c}x'$  and  $\gamma \rightarrow \gamma'/c$ . As such, the phase diagram in terms of  $(\gamma, A(0))$  for any  $\Delta$  can be rescaled to another by simply rescaling  $\gamma$ . Performing the same sign inversion on  $r$  by introducing the transform  $r \rightarrow -r'$ , necessitates changing  $\eta \rightarrow -\eta'$  and  $\gamma \rightarrow -\gamma'$ , so that the  $(\gamma, A(0))$  phase diagram is reflected about the  $\gamma$  axis.

The solutions of the fKdV equation, subject to a single topographic obstruction with local support—be that a trough or a rise—can be broadly categorised into five different classes of solutions based upon their respective patterns of traversal around Figure 4.2. The first two types exist for  $\gamma > 0$ , and have been well documented by Miles (1986), Wade (2015). The Type I solutions are perturbations from the uniform stream solution

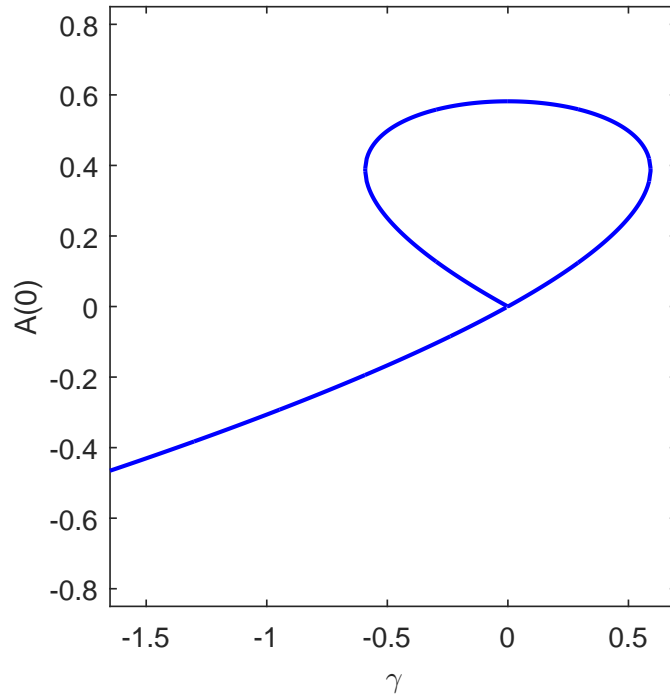


Figure 4.2: Parameter space for solutions of equation (4.11) with local support where  $f(x) = \delta(x)$ .

which, in phase space correspond to starting at the saddle point at  $(A, A_x) = (0, 0)$  and traversing the homoclinic orbit in a clockwise direction in the upper half of the phase plane. As was discussed earlier, the forcing function introduces a discontinuous jump in the derivative of the solution at  $x = 0$ , as a consequence of the discontinuity in the derivative of the delta function. This introduces a distinct downwards jump in the phase plane at the point in the solution path corresponding to  $x = 0$ , which in turn manifests itself as the discontinuous gradients observed in the calculated solutions in Figure 4.3 and Figure 4.4.

The Type II solutions space broadly replicates the form seen in the Type I solutions, with the difference in form a product of the larger amplitude solutions being perturbations from the solitary wave solution, rather than from the trivial solution. In the phase plane, the distinction between the Type I and Type II solutions is that in the former, the jump occurs to the left of the centre in the phase plane at  $(A, A_x) = (-2\Delta/9, 0)$ , whereas the Type II solutions are confined to the right of the centre. The transition from the Type I to Type II



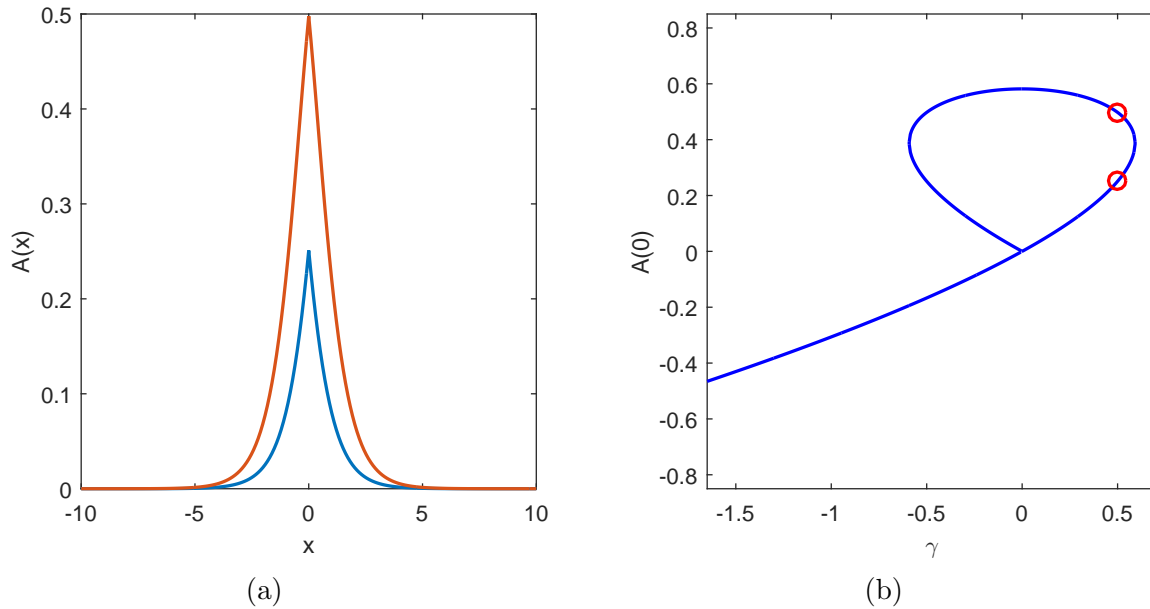


Figure 4.3: Type I and Type II (in blue and red respectively in Figure (a)) solutions to the fKdV equation with local support when  $\gamma = 0.5$  in part (a), with the corresponding locations in the phase space diagram indicated in red in part (b).

solutions occurs occurs at the largest positive value of  $\gamma$ .

The Type III, IV and V solutions correspond to different components of the solutions where  $\gamma < 0$ . In the case of the Type III and Type IV solutions, both begin by traversing the homoclinic orbit in the same manner as the Type I and II solutions – beginning at the saddle and traversing the orbit in a clockwise direction. However, unlike the previously described solutions, the discontinuous jump is only imposed after the solutions have already traversed the orbit past  $(A, A_x) = (-2\Delta/9, 0)$ , before jumping back up to the region of the homoclinic orbit where  $A_x > 0$  again. The solutions then traverse the homoclinic clockwise a second time, until returning the saddle. These solutions transition to the Type III solutions at the solitary wave located at  $\gamma = 0$ . In the case of the Type III solutions, the jump occurs prior to passing the centre for the second time, while the Type IV solutions pass the centre twice before jumping to the upper half of  $A_x$ . These solutions are then, respectively, perturbations to a single solitary wave and a system of two solitary waves. The Type IV solutions exist in the left most branch of Figure 4.2, for  $A(0)$  greater than 0 and  $\gamma$  between 0 and its lowest value for positive  $A(0)$ .

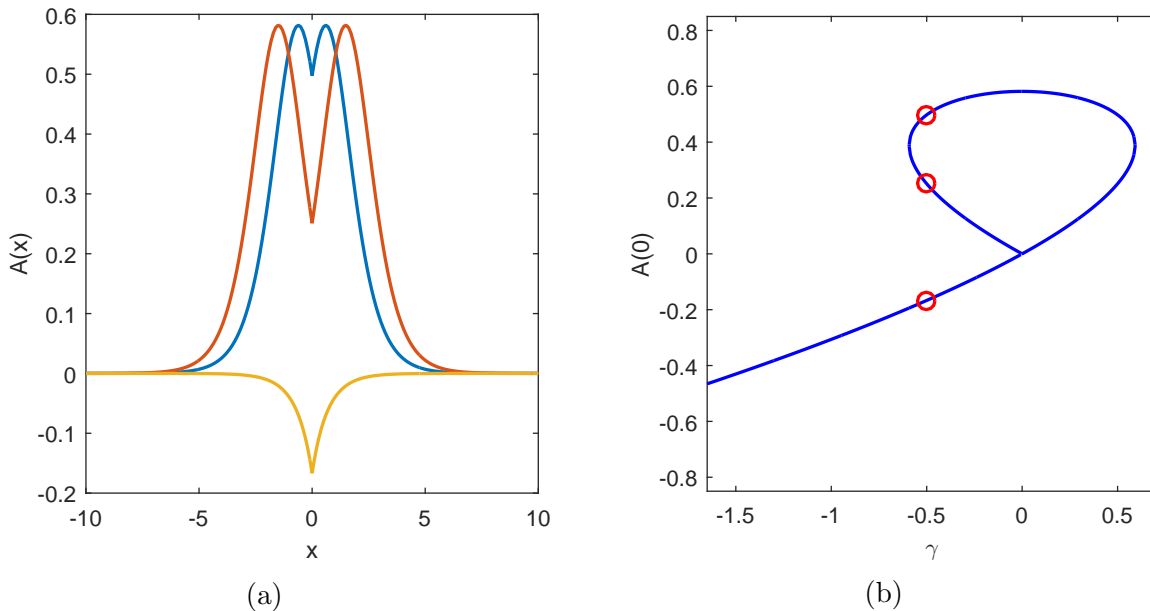


Figure 4.4: Type III, Type IV, and Type V (in blue, red, and yellow respectively in Figure (a)) solutions to the fKdV equation with local support when  $\gamma = -0.5$  in part (a), with the corresponding locations in the phase space diagram indicated in red in part (b).

The final distinct case involves the Type V solutions, which exist only for  $\gamma < 0$  and  $A(0) < 0$ , and unlike the rest of the solution types, can be described using cosech<sup>2</sup> functions, rather than the sech<sup>2</sup> solution forms seen in Types I-IV. These solutions correspond to a path that begins at the saddle, and progressing to the left along the negative  $A_x$  unbounded trajectory in phase-space, before jumping up to the unbounded trajectory that exists for  $A_x > 0$  and returning to the saddle point, in a manner that corresponds to a perturbation from the uniform stream.

While these results only apply for the case of forcing with compact support, Baines (1995), Shen (1995), Dias and Vanden-Broeck (2004), Binder et al. (2005) have proposed that these results can be extended to solutions with compact support (hereafter in the interests of brevity this will be referred to as the Baines (1995) assumption). These authors have shown that under certain conditions the governing flow equations can be reduced so that the effect of any topographic forcing function with compact support—that  $f(x) \neq 0$  for  $x$  in some  $[x^-, x^+]$  can be approximated by a combination of Dirac delta functions. This result is a product of long wavelength asymptotics, whereby if  $L$  is the horizontal length scale and  $D$  is the constant depth as  $x \rightarrow -\infty$ , where  $(\frac{D}{L})^2 \ll 1$ , and all forcing is contained in a region

with local support centred around points  $x = 0$ , then the topographic forcing function  $\gamma f(x)$  can be approximated by

$$\gamma f(x) \approx \tau \delta(x), \text{ where } \int_{-\infty}^{\infty} \gamma f(x) = \tau. \quad (4.12)$$

As a result, equation (4.2) can be solved using the approximation of local support, with

$$\left. \begin{aligned} \Delta A + rA^2 + sA_{xx} &= 0 \\ A_x(x_i^+) - A_x(x_i^-) &= -\tau. \end{aligned} \right\} \quad (4.13)$$

## 4.2 Topographic forcing with compact support

In order to extend the applicability of these results, we shall now consider the case when the topographic forcing can no longer be considered local. This occurs when the relative length scale of the topography can no longer be considered insignificant relative to that of the overall length scale, then it is no longer appropriate to approximate the effect of the topographic forcing through the use of a Dirac delta function. Numerous analytic and numerical studies have been conducted for solutions for the fKdV equation with a single topographic disturbance (Shen et al., 1989, Choi et al., 2008, 2010), although the majority of this work has been focussed on varying the linear phase speed  $\Delta$  and keeping the forcing amplitude constant.

In order to solve for topographic forcing with compact support, several semi-analytic approaches have been developed to solve for  $L_0$ . We shall briefly outline one of the more common techniques, as presented by Shen (1993). For the fKdV equation subject to a forcing function defined with compact support over some region  $x \in [x^-, x^+]$ , then in a similar manner to the locally forced fKdV equation it can be said that

$$A(x) = -\frac{3\Delta}{2r} \operatorname{sech}^2 \left( \sqrt{\frac{-\Delta}{2}} (x - L_0) \right), \quad x \leq x^-$$

where again,  $L_0$  corresponds to a phase shift imparted by the forcing function  $f(x)$ . This phase shift can be solved for by introducing the parameter  $B_\Delta$ , where

$$\begin{aligned}
B_{\Delta}(L_0) &= \int_{x^-}^{x^+} f(x)A_x(x)dx \\
&= \frac{s}{2}(A'(x^+))^2 + \left(\frac{\Delta}{2} + \frac{r}{3}A(x^+)\right)A^2(x^+)
\end{aligned} \tag{4.14}$$

Multiplying the fKdV equation by  $A_x(x)$  and integrating from  $x^-$  to  $\hat{x} > x^+$ , it can be shown that

$$\frac{1}{2}(A'(\hat{x}))^2 + \left(\frac{\Delta}{2} + \frac{r}{3}A(\hat{x})\right)A^2(\hat{x}) = B_{\Delta}(L_0), \quad \hat{x} \geq x^+$$

For this equation to satisfy the solitary wave boundary condition  $A(\infty) = 0$ , then it must be that  $B_{\Delta}(L_0) = 0$ . In this way, it follows that  $L_0$  must be an implicit function of  $\Delta$ , and that  $L_0$  is not necessarily unique.

To solve for this,  $B_{\Delta}(L_0) = 0$ , a trial value of  $L_0$  is chosen, which then gives rise to the initial value problem

$$\left. \begin{aligned}
A_{xx} + \Delta A + rA^2 &= 0, \quad x > x^-, \\
A(x^-) &= -\frac{3\Delta}{2r} \operatorname{sech}^2\left(\sqrt{\frac{-\lambda}{4}}(x^- - L_0)\right), \\
A_x(x^-) &= -\sqrt{-\Delta}A(x^-) \tanh\left(\sqrt{\frac{-\Delta}{4}}(x^- - L_0)\right).
\end{aligned} \right\} \tag{4.15}$$

Solving this boundary value problem over  $x \in (x^-, x^+)$  allows the integral  $B_{\Delta}(L_0) = \int_{x^-}^{x^+} f(x)A_x(x)$  to be computed. From here, a thorough search of space of  $B_{\Delta}$  as a function of  $L_0$  can be used to find the zeros of this function, which will in turn give the viable values of the phase shift  $L_0$ .

It should be clear that this matching process requires a significant degree of computationally intensive iteration, and makes computing the solution to the fKdV equation subject to topographic forcing with compact support both difficult and computationally taxing. As such, with the aim of testing the Baines (1995) assumption, the numerical solution to equation (4.2) will be constructed using the GHAM with the hypersphere continuation of Subsection 3.4.1. Here the spheres will be constructed in terms of the amplitude of the

solution at  $x = 0$ , and the amplitude of the forcing,  $\gamma$ .

To explore the applicability of the Baines (1995) assumption, and following Section 4.1, the parameter space of the fKdV equation will be considered by varying the forcing amplitude  $\gamma$  of the single, symmetrical topographic bump, defined by

$$f(x) = \begin{cases} \left(\frac{4}{3L}\right) \cos^4\left(\frac{\pi x}{2L}\right), & \text{for } |x| \leq L \\ 0, & \text{for } |x| > L, \end{cases} \quad (4.16)$$

will be examined. The choice of this specific forcing function was driven by two specific factors. The first is that the scaling factor— $\left(\frac{4}{3L}\right)$ —rescales  $f(x)$  so that the area under  $f(x)$  is equivalent to  $\delta(x)$  for all  $L$ , which gives the condition

$$\int_{-\infty}^{\infty} f(x) = \int_{-\infty}^{\infty} \delta(x) = 1.$$

This means that all solutions of the fKdV equation, subject to this forcing, should be directly comparable to the local support case. Secondly, in the limit as  $L \rightarrow 0$  the forcing function deforms onto  $\delta(x)$ , and as such varying  $L$  can be thought of as varying the distance between the cases of local and compact support, in order to test the extent to which the Baines (1995) assumption will hold.

To construct numerical solutions for the fKdV equation on a Chebyshev basis, solutions that exist on  $x \in (-\infty, \infty)$ , the problem domain will be truncated onto  $x \in (-B, B)$ , prior to being transformed onto the numerical domain  $(-1, 1)$  through a Logarithmic mapping. Absorbing boundary conditions will be implemented to account for the effect of the region within  $B < |x| < \infty$  on the overall solution. These boundary conditions can be constructed based upon the knowledge that the observed solutions will begin at the saddle located at  $(A, A_x) = (0, 0)$  and traverse the phase plane before returning to the saddle, which means that in the far field  $A(x)$  will decrease as  $x \rightarrow \pm\infty$ , and as such it follows that  $A(x)^2 \ll A(x)$  in this region. Under the assumption that the forcing  $f(x)$  has compact support, we can also state that  $f(x) = 0$  as  $x \rightarrow \pm\infty$ —which means that the fKdV equation can be reduced to

$$\Delta A + A_{xx} = 0. \quad (4.17)$$

If  $\Delta < 0$ , then it follows that

$$A(x) = c_1 \exp\sqrt{|\Delta|x} + c_2 \exp^{-\sqrt{|\Delta|x}}. \quad (4.18)$$

As it has been assumed that  $A(x) \rightarrow 0$  as  $x \rightarrow \pm\infty$ , it follows that  $c_1 = 0$  as  $x \rightarrow \infty$  and  $c_2 = 0$  for  $x \rightarrow -\infty$ . As a consequence of these solutions, the fKdV equation can be recast as

$$\left. \begin{aligned} \Delta A + rA^2 + A_{xx} &= -\gamma f, & x \in (-B, B) \\ A_x(B) &= -\sqrt{|\Delta|}A(B), \\ A_x(-B) &= \sqrt{|\Delta|}A(-B), \end{aligned} \right\} \quad (4.19)$$

where the last two conditions are the absorbing boundary conditions, which can be implemented within the GHAM through boundary bordering.

As detailed in Section 2.5, after applying the Logarithmic mapping the fKdV equation subject to Absorbing boundary conditions becomes

$$\left. \begin{aligned} \hat{\Delta}A + \hat{r}A^2 + \hat{k}A_y + A_{yy} &= -\hat{\gamma}f, & y \in (-1, 1) \\ A_y(1) &= \frac{-\sqrt{|\Delta|}}{\sqrt{\sinh^2(B) + 1}}A(1), \\ A_y(-1) &= \frac{\sqrt{|\Delta|}}{\sqrt{\sinh^2(B) + 1}}A(-1), \end{aligned} \right\} \quad (4.20)$$

where

$$\left. \begin{aligned} \hat{\Delta} &= B^2 (\sinh^2(By) + 1) \Delta \\ \hat{r} &= B^2 (\sinh^2(By) + 1) r \\ \hat{\gamma} &= B^2 (\sinh^2(By) + 1) \gamma \\ \hat{k} &= -\frac{\sinh(By)}{\sqrt{\sinh^2(By) + 1}}. \end{aligned} \right\} \quad (4.21)$$

This form of the fKdV equation can now be solved using the GHAM, in order to explore the progression of the solution space, and its corresponding parameter space as  $\gamma$  is varied.

To construct a continuation regime for exploring the parameter space of the fKdV, the solution at  $\gamma$  will be partitioned into a component stemming from the perturbation in  $\gamma$ , and the component resulting from a previously calculated  $\gamma_0$ , expressed as

$$A(x, \gamma) = \beta(x, \gamma) + \phi(x, \gamma_0).$$

This is then substituted into equation (4.19) to give

$$\left. \begin{aligned} (\hat{\Delta} + 2r\phi)\beta + \hat{r}\beta^2 + \hat{k}\beta_y + \beta_{yy} &= -\hat{\gamma}f - \hat{\Delta}\phi - \hat{r}\phi^2 - \hat{k}\phi_y - \phi_{yy}, \\ \beta_y(1) + \frac{\sqrt{|\Delta|}}{\sqrt{\sinh^2(B) + 1}}\beta(1) &= \frac{-\sqrt{|\Delta|}}{\sqrt{\sinh^2(B) + 1}}\phi(1) - \phi_y(1), \\ \beta_y(-1) + \frac{-\sqrt{|\Delta|}}{\sqrt{\sinh^2(B) + 1}}\beta(-1) &= \frac{\sqrt{|\Delta|}}{\sqrt{\sinh^2(B) + 1}}\phi(-1) - \phi_y(-1), \\ y &\in (-1, 1). \end{aligned} \right\} \quad (4.22)$$

This form of the fKdV equation is now suitable for solving through the GHAM, in the manner described in Section 3.3. The inherent freedom in the choice of  $\mathcal{L}$  in

$$(1 - q)\mathcal{L}[\beta(y; q) - \beta_0(y; q)] = q\hbar\{\mathcal{N}[\beta(y; q)]\}$$

gives rise to several different options for parametrising this problem. The first would be to set that

$$\left. \begin{aligned} \mathcal{L}[\beta] &= (\hat{\Delta} + 2r\phi)\beta + \hat{r}\beta^2 + \hat{k}\beta_y + \beta_{yy} \\ \mathcal{N}[\beta] &= (\hat{\Delta} + 2r\phi)\beta + \hat{r}\beta^2 + \hat{k}\beta_y + \beta_{yy} + \hat{\gamma}f + \hat{\Delta}\phi + \hat{r}\phi^2 + \hat{k}\phi_y + \phi_{yy} \\ \beta &= \beta(y; q). \end{aligned} \right\} \quad (4.23)$$

This scheme involves a single LUPQR decomposition at each step of the exploration of parameter space. The alternate parametrisation involves modifying  $\mathcal{L}$  so that it instead takes the form

$$\mathcal{L}[\beta] = \hat{\Delta}\beta + \hat{r}\beta^2 + \hat{k}\beta_y + \beta_{yy}. \quad (4.24)$$

Once again this is a variable coefficient boundary value problem as  $\hat{\Delta}$ ,  $\hat{r}$  and  $\hat{k}$  are all functions of  $y$  as a result of the domain mapping. However, by omitting the  $2r\phi\beta$  term from the equation, the linear operator  $\mathcal{L}$  is invariant, and as such the LUPQR decomposition only

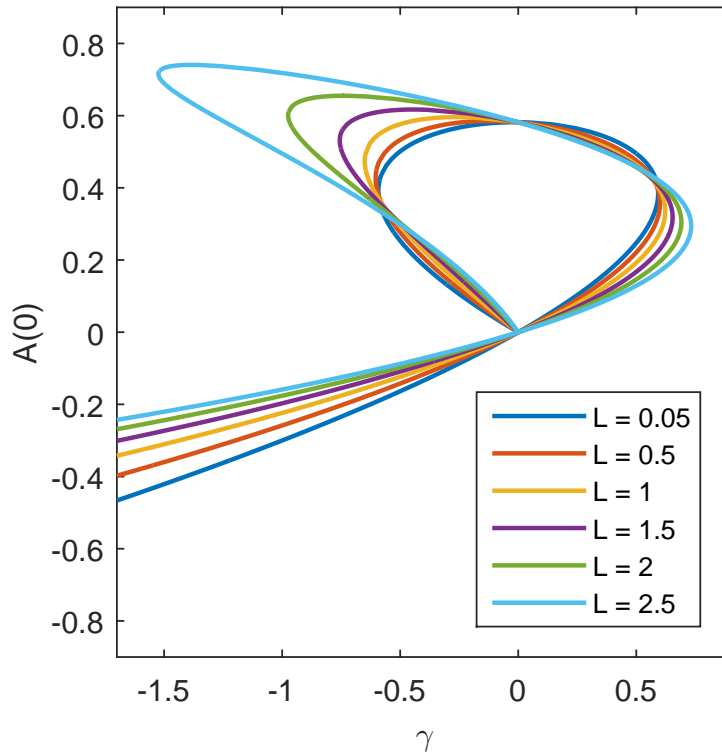


Figure 4.5: Parameter space for solutions of the fKdV in the form of equation (4.11), subject to the forcing stipulated by equation (4.16) for varying  $L$ .

needs to be performed a single time to traverse the entire parameter space, as compared to creating a new decomposition for each new position in parameter space. As was shown in Subsection 3.5.1, the LUPQR decomposition is the dominant component of the computational cost of this scheme, so converting the scheme to only requiring one decomposition for the entire parameter space has obvious implications for the computational efficiency of this scheme.

To explore the applicability of the Baines (1995) assumption Figure 4.5 presents the parameter space (in terms of  $(\gamma, A(0))$ ) for the case where  $f(x)$  is of compact support. In the limit as  $L$  approaches 0, the observed solutions subject to compact collapse down upon to those for local support, as shown in Figure 4.2. Even at  $L = 0.05$ , the solutions subject to local and compact support are indistinguishable.



As  $L$  increases, corresponding to increasing the width of the region of compact support, the correlation of the solutions calculated using compact support and those with local support decreases, to the point where when  $L = \mathcal{O}(1)$  the  $\delta(x)$  function solution can no longer be considered to be a reasonable approximation to the dynamics of the compact forcing case, which emphasises the limits to the Baines (1995) assumption. This decoupling in the behaviour of the solutions is particularly pronounced in the  $\gamma < 0$  region, wherein the solitary wave solution decomposes into two separate solitary wave profiles. These solitary waves then become more distinct as  $L$  increases, to the point where the stationary solution is effectively two separate standing solitary waves, centred symmetrically about  $x = 0$ .

As was discussed previously, the solutions of the fKdV equation can be broadly categorised into five different classes of solutions, as was documented for  $\delta(x)$  forcing in Section 4.1. The Type I and II solutions shown in Figure 4.6, correspond to perturbations from the uniform stream and solitary wave solutions respectively, and are similar to their  $\delta(x)$  function forcing analogues with a similar jump in the phase plane. However, the deformation is not a sharp jump, but rather a smooth transition from the positive  $A_x$  component of the homoclinic orbit to the negative. This results in a smooth profile to the wave across  $x = 0$ .

In contrast to the  $\delta(x)$  solutions is the behaviour of the Type II solutions around  $\gamma = 0$ . For  $\delta(x)$  function forcing, the transition from Type II to Type III solutions occurs around the upper solitary wave solution at  $\gamma = 0$ . In contrast, the Type II to III transition occurs at the peak of the phase portrait which occurs, in this case, at around  $\gamma \approx -0.5$  for  $L = 1$ . This delayed transition, as a function of  $L$ , can be clearly seen in Figure 4.5, where as  $L$  increases the peak of the profile shifts further to the negative, denoting the delayed transition to Type III as the characteristic width of the topography increases.

Once again, the Type III, IV, and V solutions of Figure 4.7 are similar to their equivalent forms for  $\delta(x)$  function forcing, with each solution corresponding to perturbations from either a single solitary wave, pair of solitary waves, and a strictly negative perturbation from the uniform freestream respectively. Similar to the Type I and II solutions, the Type III–V solutions all demonstrate smooth and continuous deformations across all  $x$ .

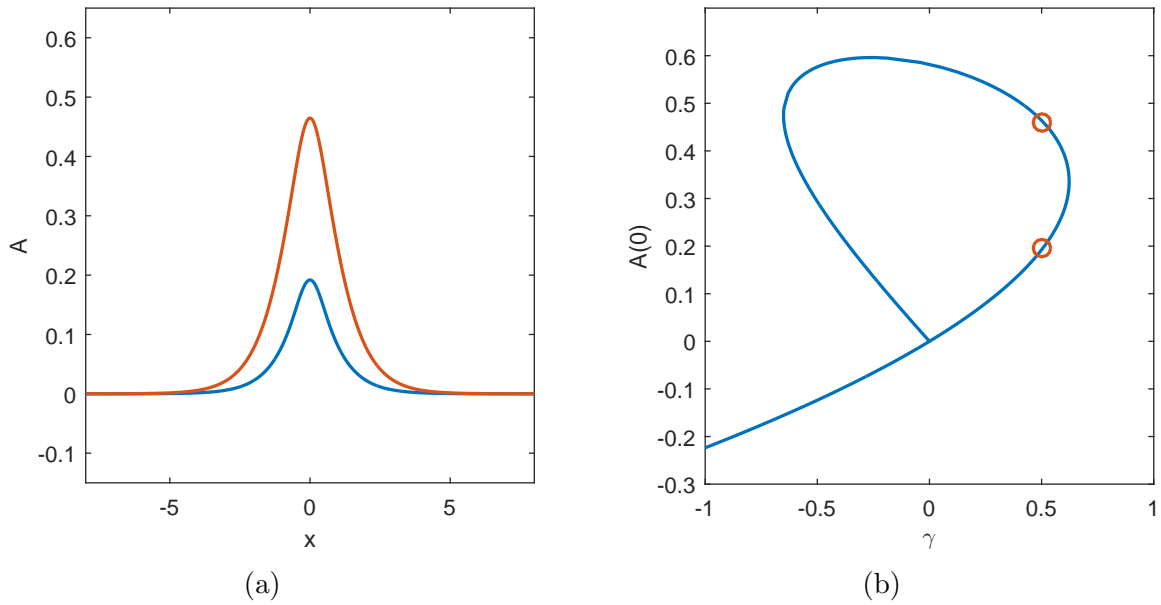


Figure 4.6: Type I, and Type II (in blue and red respectively in Figure (a)) solutions to the fKdV equation with compact support when  $\gamma = 0.5$  and  $L = 1$ .

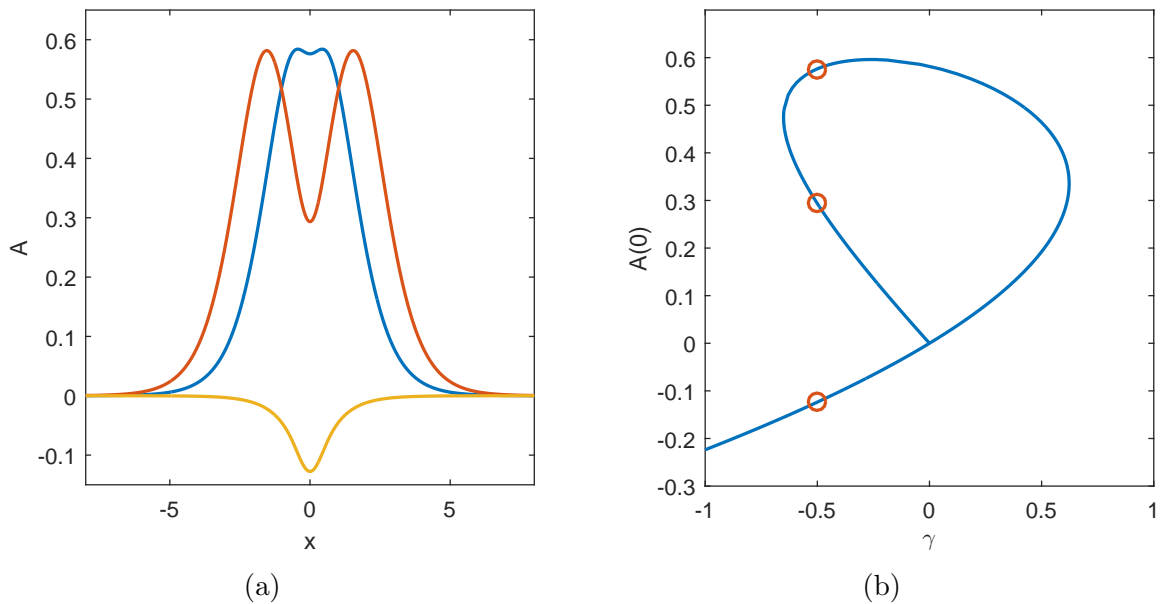


Figure 4.7: Type III, Type IV, and Type V (in blue, red, and yellow respectively in Figure (a)) solutions to the fKdV equation with compact support when  $\gamma = -0.5$  and  $L = 1$ .

### 4.3 Discussion

Examining these results, and the general correlation between the analytic solution of the locally forced case, and the numerical solution of the forcing with compact support, it is clear that the results conform with previously published results, and that the numerical continuation schemes proposed in the prior chapter is accurate and applicable to finding solutions to the fKdV equation. At this point, we now have enough information to return to the initial hypothesis—that solutions with compact support can be substituted by the equivalent equations with local support. Returning to Figure 4.5, it is clear that as the length scale of the topography changes, so too does the accuracy of the local forcing approximation. This should not necessarily be surprising, as the difference, in terms of both height and width, between the two forcing functions is a function of  $L$ . What is surprising, however, is just how sensitive the assumption is. When  $L = 1$ , there are already marked differences in how the solution space evolves, particularly with regard to the points of transition between the Type II and III solutions, and the proportion of Type III and IV solutions in the solution space, as the part of the diagram where these solutions exists becomes more prominent as  $L$  increases. This is more striking given that the chosen forcing function outlined within equation (4.16) closely resembles the form of the  $\delta(x)$  forcing function. For forcing topographies that do not share these topological similarities, it must be surmised that the local forcing hypothesis is unlikely to be satisfied.

The dynamics of these Type I-V solutions is broadly similar between solutions with local and compact support, with the distinction that the Type III and IV solutions become predominant as  $L$  becomes larger, as can be seen with the significant displacement of the turning point in the upper-left quadrant with  $L$ , where  $\gamma < 0$  and  $A(0) > 0$ . This also manifests as increasing the radius of curvature in this region, which in turn increases the complexity of traversing the corner using branch following, as more iterations are required to traverse the corner. A further by-product of this is that as  $L$  approaches approximately 2 (from below), the Type IV branch begins to transition from taking a convex shape as  $\gamma \rightarrow 0^-$  and  $A(0) \rightarrow 0^+$ , to having an inversion point and becoming concave as  $L$  increases. The solutions after the inversion point—as can be seen at around  $\gamma = -0.5$  in Figure 4.5 do not change type, rather this inversion simply reflects the predominance of solutions in the region where  $A(0) \approx > 0.35$ , and the increased sensitivity to changes in  $\gamma$  when  $\gamma$  is approximately smaller than  $-0.5$ . As  $L$  increases beyond 2.5, the magnitude of the concave region becomes more marked, which in turn has an affect upon the radius of curvature at

the Type III to Type IV transition point. This decrease in the radius of curvature, and the nonlinear growth in the magnitude of the location (in terms of  $\gamma$ ) for the Type III to Type IV transition point has meant that we have been unable to traverse the Type IV branch for  $L$  much larger than 2.5. In light of the difficulty of traversing this branch, and the shift from the profile being locally convex to concave indicates that the solution at  $L = 2.5$  may in fact correspond to a transition regime from a flow configuration where the Type IV solutions exist, to one where they do not. While some preliminary asymptotic work has been conducted, which suggests that the Type IV branch does exist as  $L \rightarrow \infty$ , there are still some remaining questions about the validity of the approach taken to derive these results, and as such they have been omitted from this work.

Overall, the dynamics of these Type I-V solutions are broadly similar between solutions with local and compact support, with the primary differences between solutions being the smoothness across  $x = 0$ . However, while previous authors have used local forcing as an approximation for the dynamics of solutions with compact support, this work clearly shows that any conclusions based upon this should be limited to bell-shaped forcing topography with small length scales, and that particular caution should be made in the region of the transition from the Type II to Type III solutions. However, the broad agreement between the solutions with local and compact support; as well as between the solutions with compact support and previously published works further validates the applicability of the GHAM and its associated continuation methods for exploring the parameter space of nonlinear boundary value problems. As such, the technique should be well suited for exploring the forced Gardner equation.

These results also broadly resemble solutions to the fully nonlinear problem, which has been solved using Boundary Integral methods by Tam et al. (2015). Unlike the weakly nonlinear case, the fully nonlinear solutions are not strictly positive or strictly negative, and extra branches of solutions occur, corresponding to the appearance of trapped secondary waves within main waveform. More specifically, the Type I and V, strictly negative solutions exhibit strong agreement between the two types of solutions. However as the forcing amplitude increases the fully nonlinear solutions do not transition into Type II solutions. Instead, the central amplitude of the Type I fully nonlinear solutions continues to grow as the height of the topographic forcing is increased. Of most interest are the Type IV solutions, which in the fully nonlinear case still exist as  $\gamma$  decreases below zero, with the

the centre of the cusping crossing the horizontal axis and becoming negative.

Having now validated the use of the GHAM to solve nonlinear wave problems, and confirming that it shares the advantageous computational properties outlined in Chapter 3, we can now turn to less understood problems from wave dynamics—specifically the asymmetric solutions of the fKdV equation, and symmetric solutions of the forced Gardner equation.

## Chapter 5

# Asymmetric steady solutions of the forced Korteweg–de Vries equation

The solution space of the fKdV equation is not just limited to the symmetric solutions discussed to this point. Rather, a multitude of solutions exist, including symmetry breaking solutions that exist on an orbit connecting the unforced problem's two critical points. These solutions are topographically controlled (Killworth, 1992), and involve a transition from a subcritical upstream flow to a supercritical downstream solution, or a subcritical downstream condition to a supercritical upstream, in a form of solution known as a hydraulic fall. From this point we will solely consider the case where the saddle is the downstream limit, as the contrasting case is unstable.

In contrast to the solution space of the symmetric fKdV equation, which correspond to traversing the homoclinic orbit from the saddle at  $(A, A_x) = (0, 0)$  to itself, the asymmetric solution space of the steady forced KdV equation

$$A_{xx} + \Delta A + 3A^2 = -\gamma f(x), \quad (5.1)$$

for  $\Delta > 0$  involves traversing the phase plane from the upstream critical point  $(A, A_x) = (0, 0)$  to the downstream critical point at  $(A, A_x) = (-\Delta/3, 0)$ . The case where  $\Delta < 0$  reverses the order of the critical points, however, the change in order can be reverted through the addition of a mean level, and as such it can be assumed that  $\Delta > 0$  without loss of generality. The existence and behaviour of these asymmetric solutions has been well documented by Grimshaw and Smyth (1986), Belward and Forbes (1993), Dias

and Vanden-Broeck (2002), Shen et al. (2002) and Donahue and Shen (2010), among others.

The parameter space of these asymmetric hydraulically controlled solutions was mapped in terms of  $(\gamma, \Delta)$  by Ee and Clarke (2007, 2008), for a bell-shaped topographic forcing of the form  $f(x) = \text{sech}^2(x)$ . Their exploration of this space was built upon a consideration of the Hamiltonian

$$H(A, A_x) = \gamma f A + \frac{1}{2} \Delta A^2 + A^3 + \frac{1}{2} A_x^2, \quad (5.2)$$

which, when evaluated at the upstream critical points is  $H_{c_1} = H(0, 0) = 0$ , and at the downstream limit  $H_{c_2} = H(-\Delta/3, 0) = \Delta^3/54$ . The downstream limit of any extant hydraulically controlled solution will lie at the maxima of  $H$ , for which the downstream amplitude will lie within, and ideally on, the homoclinic orbit in the phase plane. By stipulating that

$$\left. \begin{aligned} H(H - H_{c_2}) &\leq 0 \\ -\frac{\Delta}{3} &\leq A \leq \frac{\Delta}{6} \end{aligned} \right\} \quad (5.3)$$

the downstream solution can be restricted to lie within the homoclinic orbit. This in turn corresponds to imposing that the Hamiltonian must lie between its local minima and maxima, located at  $H_{c_1}$  and  $H_{c_2}$  respectively.

With the aim of constructing a parametric representation of the relationship between  $\gamma$  and  $\Delta$  Ee and Clarke (2007) implemented a branch following routine to search for the maximum of  $H$ , for which  $|H - H_{c_2}| \leq \epsilon$  for some small  $\epsilon$ , and conditioned upon satisfying equation (5.3). Through this process, the authors were able to construct the relationship shown in Figure 5.1.

This parameter space can be confirmed through their agreement with an asymptotic analysis of the problem conducted by Grimshaw and Smyth (1986), which stipulated that as  $\gamma \rightarrow 0^\pm$

$$\left(\frac{\Delta}{3}\right)^3 = (K\gamma)^2. \quad (5.4)$$

Here  $K$  is the integral of the forcing  $\int_{-\infty}^{\infty} f dx$  (Shen, 1991). A separate asymptotic limit corresponding to large, positive values of  $\gamma$  can also be found, predicated upon the assump-

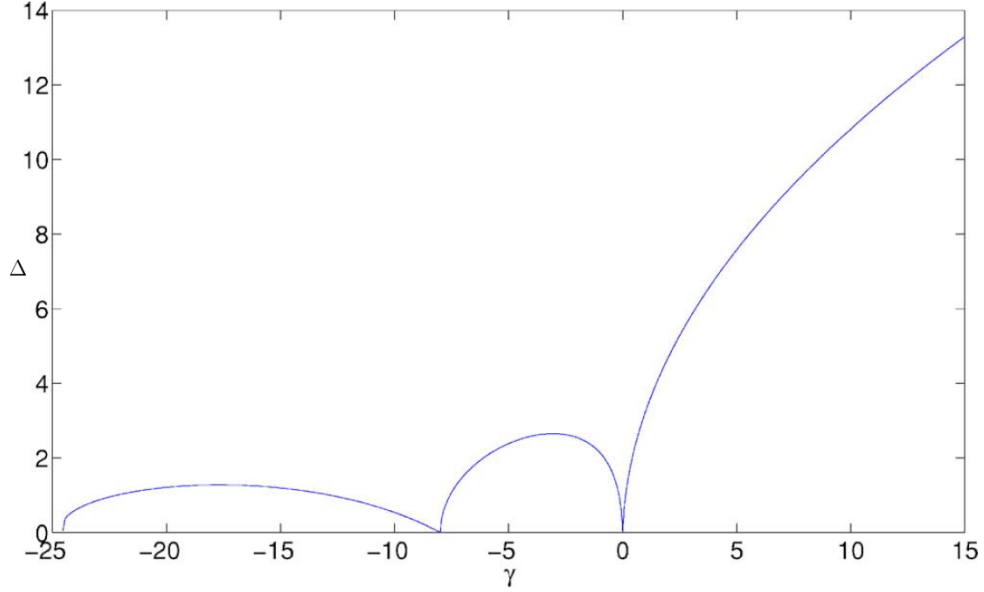


Figure 5.1: Plot of  $\Delta$  against  $\gamma$ , as produced by Ee and Clarke (2007) for the parameter space for asymmetric solutions of equation (5.1).

tion that in this limit the dispersive term is negligible. Then in this limit equation (5.1) will only exhibit solutions when

$$\Delta = \sqrt{12\gamma}. \quad (5.5)$$

As solutions where  $\Delta < 0$  are the equivalent to those with  $\Delta > 0$  with the addition of a rescaling of the mean height, all solutions can be restricted to  $\Delta \geq 0$ . All three of these asymptotic approaches accurately match the numerical results of Ee and Clarke (2007).

The basic form of the parameter space can be inferred by examining the point where the hydraulic solutions collapse onto symmetric solitary waves, which occurs when  $\Delta = 0$ . Camassa and Wu (1991) showed that by imposing that  $f(x) = \text{sech}^2(x)$ , the general solution for solitary wave solutions to the fKdV equation can be expressed as

$$A = a \text{sech}^2(x), \quad (5.6)$$

in terms of a wave amplitude  $a$ . Imposing this form of solution expression within equation (5.1) yields



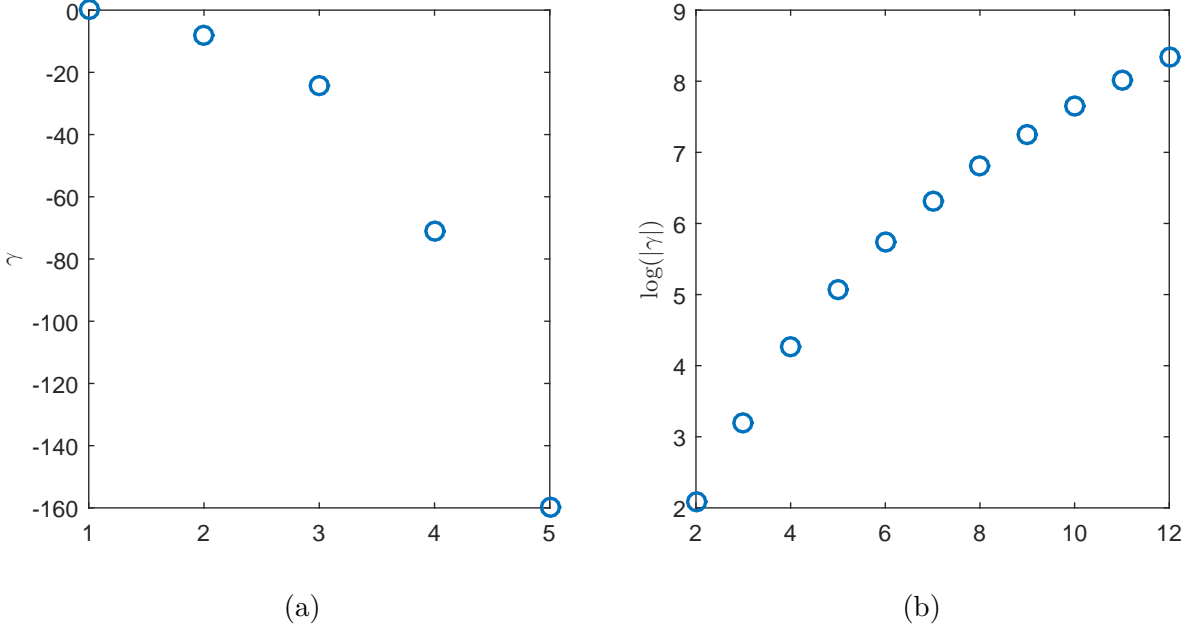


Figure 5.2: The locations of  $\frac{dA}{dx} = 0$  at  $x = 0$  for  $\Delta = 0$  are presented in (a) for  $-160 \leq \gamma \leq 0$ , and the corresponding locations in log scale in (b), with all solutions constructed based upon equation (5.1).

$$(3a^2 - 6a) \operatorname{sech}^4(x) + (4a + \gamma) \operatorname{sech}^2(x) = 0, \quad (5.7)$$

which can be satisfied either by setting that  $(\gamma, a) = (0, 0)$ , or  $(-8, 2)$ . These align with the zeros of  $\Delta$  in Figure 5.1, although, curiously, this figure also includes a third solitary wave at  $\gamma \approx -24$ . This gives rise to questions about its form, as it is clearly not that of a  $\operatorname{sech}^2(x)$  wave. A shooting approach was implemented in order to explore this solution, and to test the hypothesis that other solitary wave solutions may exist beyond the solution at  $\gamma \approx -24$ . This was done by shooting from the downstream condition to the centre of the domain at  $x = 0$ , and selecting solutions for which  $dA/dx$  is zero at  $x = 0$ —a sufficient and necessary condition for solitary wave solutions.

In fact Figure 5.2a suggests that there is an infinite set of solitary wave solutions of the fKdV equation for  $\Delta = 0$ , although the spacing between these solutions grows exponentially. This in turn suggests that the parametric curves in Figure 5.1 extends into  $\gamma < -25$ , however there may be issues with regard to numerical stability as  $\gamma$  decreases further.

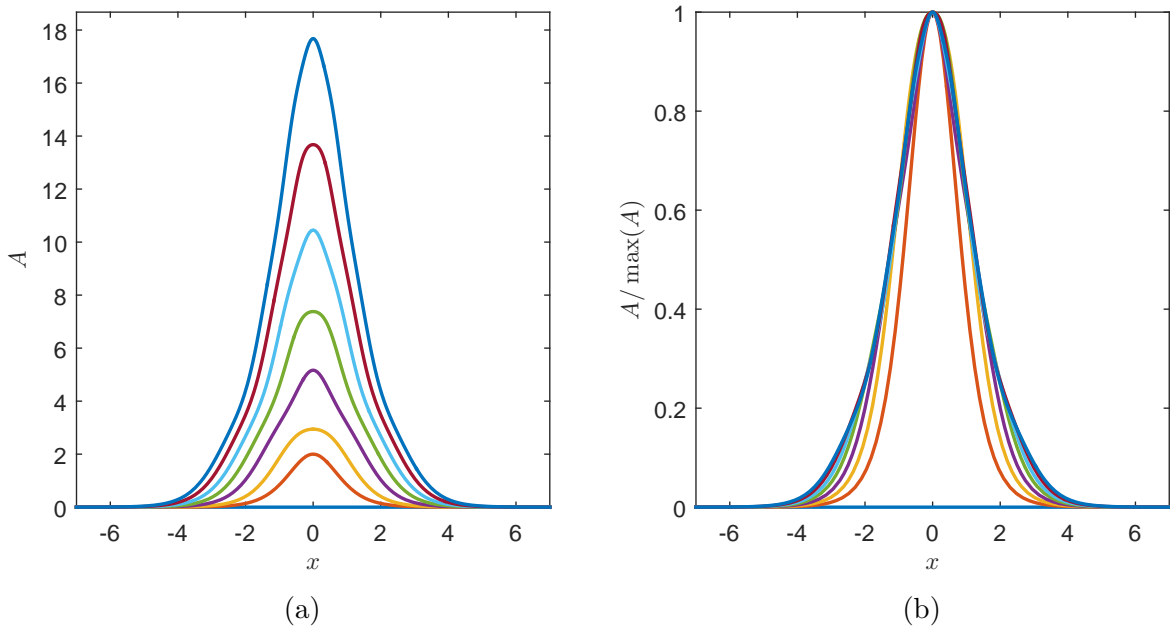


Figure 5.3: Figure (a) contains the a subset of the corresponding solitary wave solutions to Figure 5.2 for equation (5.1). The presented solutions correspond to the zeros of Figure 5.2a for  $-10^3 \leq \gamma \leq 0$ . The  $\gamma = -8$  solitary wave is shown in red, with increasing magnitudes of  $\gamma$  corresponding to increased wave amplitudes. Figure (b) shows these same solutions rescaled so that their maximum amplitude is 1.

The qualitative differences between the  $\text{sech}^2(x)$  solitary wave profile at  $\gamma = -8$  and the broader range of solitary wave solutions admitted by the fKdV equation are explored in Figure 5.3 by considering both the wave profiles and their normalised forms. Unsurprisingly, as  $\gamma$  increases so too does the amplitude of these waves, but with this also comes notable structural changes to the solutions, with a broadening of the wave profile with  $x$ , especially in the region where  $A \rightarrow 0$ . Furthermore, the waves often show a small inflection point at approximately 75% of the wave height—a feature best seen in the  $\gamma = -24.2$  (purple) and  $\gamma = -159.5$  (light blue) solutions from Figure 5.2a, and the peaks alternate between flat, low radius of curvature regions (seen in the yellow, green and burgundy plots of Figure 5.3a), and more tightly cusped, higher radius of curvature profiles (seen in purple, light blue and dark blue of Figure 5.3a).

Returning to the parametric relationship between  $\Delta$  and  $\gamma$ , while the approach of Ee and Clarke (2007) was able to traverse the parameter space for  $\gamma \geq -24$ , the calculations amounted to a search routine over a large domain conducted with only minimal constraints,

which was computationally taxing. To resolve this, inspired by the physics of the flow it was hypothesised that the topographic drag exerted should become constant in the case of steady solutions. This approach may provide a suitable manner to constrain the solution space, and allow the replication of the work of Ee and Clarke (2007, 2008) without the reliance upon their search heuristic.

Returning to the Hamiltonian equation (5.2), a conservative form of the fKdV equation can be constructed by considering

$$\left. \begin{aligned} \frac{\partial H}{\partial A_x} &= A_x \\ \frac{\partial H}{\partial A} &= \Delta A + 3A^2 + \gamma f \\ \frac{\partial H}{\partial x} &= \gamma f_x A. \end{aligned} \right\} \quad (5.8)$$

As a consequence of this, the net change in the Hamiltonian between  $x = \pm\infty$ , which is equivalent to traversing from the upstream condition to the downstream critical point will be

$$\delta H = H(\infty) - H(-\infty) = \int_{-\infty}^{\infty} \frac{dH}{dx} dx = \int_{-\infty}^{\infty} \gamma f_x A dx. \quad (5.9)$$

That the net change in the Hamiltonian  $\delta H$  is in terms of this integral is interesting as the term

$$C_w = - \int_{-\infty}^{\infty} \gamma f_x A dx. \quad (5.10)$$

has previously been identified as the wave-resistance coefficient (Wu, 1987, Camassa and Wu, 1991), which is a metric for the transference from the forcing to the energy of the system. Thus  $C_w$  is functionally equivalent to the drag  $\mathcal{D}$  exerted by the topography, so that

$$\delta H = H(\infty) - H(-\infty) = \mathcal{D}. \quad (5.11)$$

The form of  $H(\pm\infty)$  can be considered by partitioning  $H$  so that

$$H(A, A_x) = V(A) + \frac{1}{2} A_x^2,$$

Evaluating the potential  $V(A)$  and its derivatives at the the critical points  $A = 0$  and  $-\Delta/3$

$$\left. \begin{aligned} V(0) = 0 & \quad V(-\Delta/3) = \frac{\Delta^3}{54} \\ V'(0) = 0 & \quad V'(-\Delta/3) = 0 \\ V''(0) = \Delta & \quad V''(-\Delta/3) = 0. \end{aligned} \right\} \quad (5.12)$$

Thus for  $\Delta > 0$  the saddle at  $A = -\Delta/3$  corresponds to a local minima of the orbits contained within or on the manifold, and the centre at  $A = 0$  will correspond to a local minima over the same region. This in turn means that in the case of hydraulic solutions—for which  $A_x(\pm\infty) = 0$ —it must be that asymmetric solutions must correspond to a minima of  $\mathcal{D}$ , where

$$\int_{-\infty}^{\infty} \gamma f_x A dx = \frac{\Delta^3}{54}. \quad (5.13)$$

To validate this derivation, and to further explore the ramifications of this result, we can returning to the unsteady fKdV equation as introduced in Chapter 4

$$A_t + \Delta A_x + 6AA_x + A_{xxx} = -\gamma f_x. \quad (5.14)$$

The fKdV equation admits a conservation form, which can be constructed by multiplying with  $A(x)$  and integrating with respect to  $x$  to give

$$\frac{dP}{dt} + \frac{1}{2}\Delta A^2 \Big|_{-\infty}^{\infty} + 2A^3 \Big|_{-\infty}^{\infty} + \left( AA_{xx} - \frac{1}{2}A_x^2 \right) \Big|_{-\infty}^{\infty} = - \int_{-\infty}^{\infty} \gamma f_x A dx. \quad (5.15)$$

Here  $P = \int_{-\infty}^{\infty} (1/2)A^2 dx$  is the wave momentum (Camassa and Wu, 1991). In the context of the fKdV equation, it must be true that this equation must become homogeneous as  $|x|$  exceeds the region of support for the topographic forcing function. Thus, considering the case of aperiodic solutions for which  $\Delta > 0$ , then then any asymmetric solution should traverse from the saddle at  $(A, A_x) = (-\Delta/3, 0)$  to the centre at  $(A, A_x) = (0, 0)$ , or the centre to the saddle. We are free to impose that  $A(-\infty) = 0$  and  $A(\infty) = -\Delta/3$ , so that the solutions will be travelling from the centre to the saddle.

In the far field, the solutions to the steady KdV equation can take two forms—either exponentially decaying solutions, or periodic oscillations on the boundary. When  $x \rightarrow -\infty$ , the pattern of approach to the saddle imposes that  $(A, A_x) \rightarrow (-\Delta/3, 0)$ , which in turn

gives that  $A_{xx} \rightarrow 0$ , with numerical experiments confirming that these conditions are always satisfied, and that the solution does decay exponentially. In the region where  $x \rightarrow \infty$  it must be that  $A \rightarrow 0$ , similar to approaching the saddle, however, we can not similarly assume that  $(A_x, A_{xx})$  also trend towards zero. As such, the downstream condition can either correspond to exponentially decaying solutions or periodic oscillations on the boundary. As we are searching for hydraulic solutions, where  $A(x)$  exhibits exponential decay towards the imposed boundary conditions at the saddle and centre, the uncertainty about the upstream behaviour must be accounted for within the derivation.

Using the above results equation (5.15) can be reduced to

$$\frac{dP}{dt} + \frac{1}{2}A_x^2(\infty) = \frac{\Delta^3}{54} - \int_{-\infty}^{\infty} \gamma f_x A dx. \quad (5.16)$$

In the context of steady solutions, where  $\frac{dP}{dt} = 0$ , and as  $A_x^2(\infty)$  must be strictly positive then

$$\int_{-\infty}^{\infty} \gamma f_x A dx \leq \frac{\Delta^3}{54}, \quad (5.17)$$

and that as  $\int_{-\infty}^{\infty} \gamma f_x A dx$  approaches  $\Delta^3/(6r^2)$ ,  $A_x(\infty)$  must in turn approach zero. As such, hydraulic solutions can be found by searching for

$$\int_{-\infty}^{\infty} \gamma f_x A dx = \frac{\Delta^3}{54}, \quad (5.18)$$

thus confirming the previously derived result.

While we refer to the process of solving for these hydraulic solutions as minimising the drag, the question may be asked as to why this needs to be a minimisation process at all, given that equation (5.13) gives an integral condition on the solution that should be able to be satisfied—however, there are two issues with this approach. The first is that equation (5.13) is obviously a nonlinear equation in terms of  $\Delta$ ,  $A$  and  $\gamma$ , where  $A$  is inherently a nonlinear function of  $\Delta$ , due to its influence on both the fKdV equation itself, and on the boundary conditions of the asymmetric problem. Thus incorporating this condition into a solver presents particular hurdles.

The second issue is potentially even more problematic, and it stems from the potential presence of downstream wavetrains in solutions that are approaching the hydraulic limit. By truncating the numerical domain, and imposing linear boundary conditions, nonlinear wave are introduced in the downstream limit. In order to resolve a waveless, hydraulic fall solution downstream,  $L$  must be chosen from a range on the order of the wavelength of the nonlinear waves so that the drag is minimised.

As such, the problem must be considered in the context of an optimisation problem in terms of both the domain length  $L$  and  $\Delta$ , in order to satisfy equation (5.13). This is the equivalent to searching for

$$\min_{L \geq 0, \Delta \geq 0} \int_{-\infty}^{\infty} -\gamma f_x A dx, \quad (5.19)$$

for a given  $\gamma$ , and for  $x$  defined over  $[-L, L]$ .

## 5.1 Solution space for hydraulic solutions of the fKdV equation

The minimisation criteria for the drag can be solved for using a Genetic Algorithms (GA), which is a technique inspired by a classical view of mathematical selection, and employs a fitness criteria (Fisher, 1958) to search for the minima of equation (5.19) in terms of  $\Delta$  and  $L$ . The GAs are a derivative free optimisation heuristic that, by minimising the number of required function evaluations significantly reduces the computational cost of finding a solution, relative to that of multivariate iteration (Holland, 1995).

For each  $\gamma$ , the GA will propose trial doublets of  $(\Delta, L)$  that are passed to the fKdV equation solver developed within the GHAM, which considers the problem subject to Dirchlet boundary conditions subject to a logarithmic domain mapping. Due to the relative computational cost of solving the fKdV equation, the ability of GAs to minimise the number of evaluations is of significant importance. To enhance the convergence properties of the scheme, a simplified form of hypersphere continuation inspired by natural parameter continuation is employed to deform a previously calculated solution at  $\gamma_0$  onto the solution at the new choice of  $\gamma$ . Attempting to increase the performance of the scheme by employing a more considered and complex continuation scheme was found to be an impractical

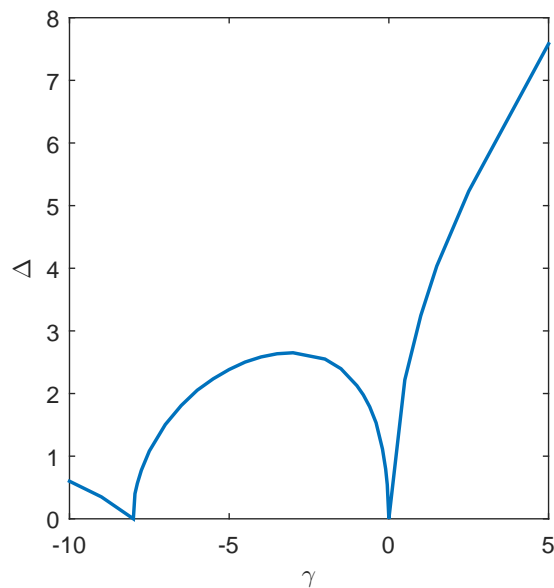


Figure 5.4: Parametric relationship between  $\Delta$  and  $\gamma$  for  $-10 \leq \gamma \leq 5$ , subject to  $f(x) = \text{sech}^2(x)$ .

solution, as changing  $L$  within the GA resulted in a change to the domain length, which in turn created difficulties in interpolating the previously calculated solution over some truncated domain, to the domain corresponding to the new guess of  $L$ . The utility of this simplified continuation approach, and its resulting solutions will be tested by considering the results of Ee and Clarke (2007)—shown in Figure 5.1—as a validation case.

Employing the GHAM and a GA subject to equation (5.19) allowed the parametric relationship in  $(\Delta, \gamma)$  to be constructed, as shown in Figure 5.4. A snapshot of the corresponding solution space is also outlined in Figure 5.5. Both of these perspectives on the case of hydraulic solutions of the fKdV equation align closely with Ee and Clarke (2007) and Figure 5.1. For the region in  $\gamma \geq -8$  the solutions and parametric relationship align exactly, however the proposed methodology did struggle to resolve solutions in the region where  $\gamma \leq -8$ . Outside of this divergent region, the calculations exhibited rapid convergence to machine precision accuracy, and produced results that closely replicated the asymptotic results, as outlined within the previous section.

However, progressing to the region wherein  $\gamma < -8$ , the step size in  $\gamma$  required for the continuation process rapidly diminished to the point where traversing the domain became an intractable problem. Theoretically, the solutions contained within this region are resonant solutions, which consist primarily of a forced symmetric solitary wave subject to a topographic perturbation coupled with a weak downstream shelf. By examining the eigenvalues of these solutions, Ee and Clarke (2007) found this form of solution to be unstable, which may explain the difficulties in constructing solutions within the simplified continuation regime. This result is especially true in the neighbourhood of  $\gamma = -8$ , which may explain the difficulty of resolving solutions within this region.

The numerical instability is not the only contributing factor to the difficulty in resolving these solutions, as around  $\gamma = -8$ , the corresponding value of  $\Delta$  for the hydraulic solution will be small, and thus close to  $\Delta = 0$ . It is known that there are a large range of numerically stable solutions on  $\Delta = 0$  for all  $\gamma$  (Keeler et al., 2017), and all of these  $\Delta = 0$  solutions will also satisfy the drag relationship. This means that the solver may be struggling to distinguish the desired hydraulic asymmetric solutions from the symmetric solutions on  $\Delta = 0$ . This problem may be alleviated as  $\gamma$  becomes more negative, as the expected values of  $\Delta$  along the parametric line become more distinct from zero as  $\gamma$  decreases further. Attempts were made to rectify this problem by more careful parameter selection within the GA, and adaptively refining bounds of the search region in  $\Delta$  to reflect the last calculated solution. However the solutions have not been presented within this work as there are some uncertainties about if the agreement between the results and the work of Ee and Clarke (2007) is a product of already having domain knowledge gained through a consideration of the previously constructed work.

Turning now from the parametric relationship between  $\Delta$  and  $\gamma$  to the corresponding solution space. First, focussing on the region wherein  $\gamma$  is positive the solutions all correspond to unperturbed hydraulic falls centred about  $x = 0$ , where the fall starts at  $A = 0$  upstream, and drops to  $-\Delta/3$  within the downstream region. As  $\gamma$  becomes larger, the effect of dispersion upon the overall solution is minimised, and in the solutions this manifests as the region where  $A_x$  is non-zero becoming restricted to the domain of the topographic perturbation. As  $\Delta$  monotonically increases with  $\gamma$  within this region, the magnitude of both the jump, and the slope must also increase monotonically. This is borne out by the results in Figure 5.6, where the jump between the upstream and downstream



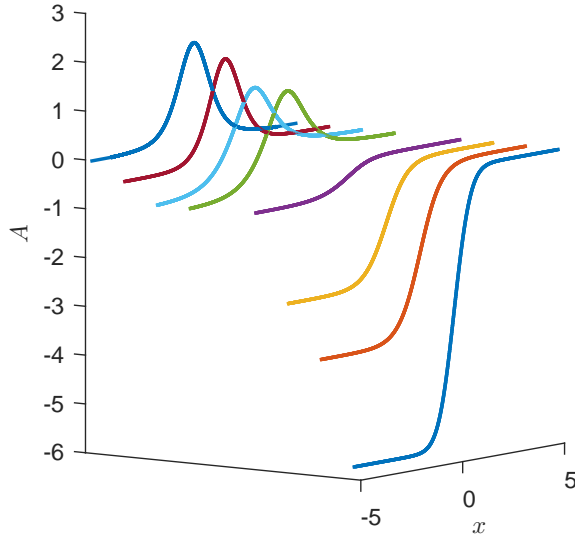


Figure 5.5: Hydraulic, asymmetric solutions of equation (5.1) for  $\gamma = \{-8.5, -7.5, -5, -2, 0.5, 5, 10, 25\}$ , ordered from left to right.

shelves becomes both larger, due to the increasing magnitude of the downstream boundary condition as  $\gamma$  increases, and the jump becomes more distinct and confined to within the region of support for the topographic forcing function.

In the region where  $\gamma < 0$ , the solutions can broadly be considered as perturbations from the solitary wave at  $\gamma = -8$ , with the perturbation taking the form of a downstream shelf. A representative sample of these solutions are shown in Figure 5.7. Unlike  $\gamma > 0$ , in this region the effect of dispersion is always present, and as such the solutions are not confined to the region of support for the topographic forcing.

While every value of  $\Delta$  within the region  $-8 \leq \gamma \leq 0$  has either one or two corresponding values of  $\gamma$ , each  $\gamma$  still corresponds to a unique solution. If we focus on two solutions for  $\Delta = -2.56$ —detailed in Figure 5.8—it is apparent that the hydraulic fall component of the solution remains constant, as it is broadly determined by the upstream and downstream boundary conditions. However, the height and width of the superimposed solitary wave increases with the magnitude of  $\gamma$ . These characteristics appear to continue for  $\gamma < -8$ , however the conclusions that can be made about the solution space in this region are

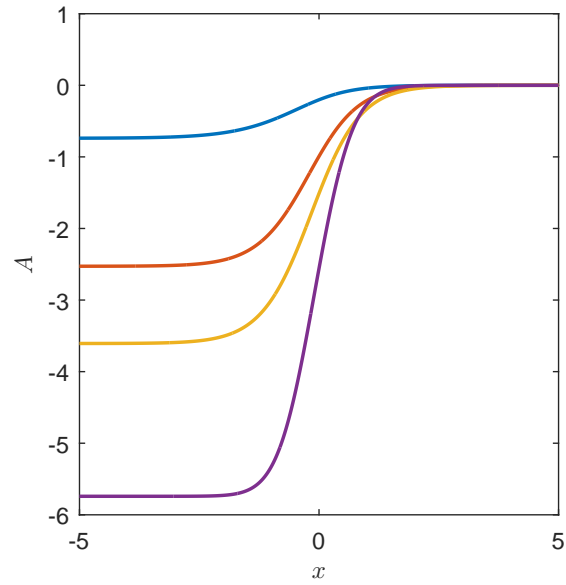


Figure 5.6: Solutions of equation (5.1) for  $\gamma = \{0.5, 5, 10, 25\}$  in blue, red, yellow and purple respectively.

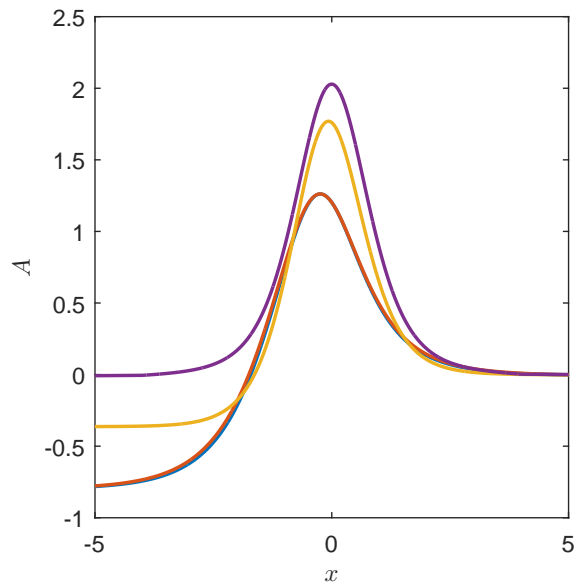


Figure 5.7: Solutions of equation (5.1) for  $\gamma = \{-8.5, -7.5, -5, -2\}$  in blue, red, yellow and purple respectively.

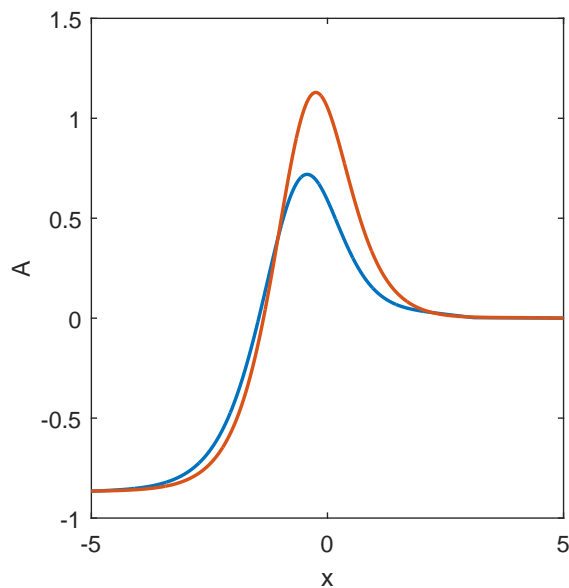


Figure 5.8:  $(\gamma, \Delta) = (-2.0285, -2.56)$  in blue and  $(\gamma, \Delta) = (-4.1360, -2.56)$  in red.

limited due to the lack of a comprehensive exploration of this region of  $\gamma$ . At some point there must be a transition back towards the solitary wave solution as  $\gamma$  approaches the solitary wave located at  $\gamma \approx -24.5$ , so this trend of increasing the amplitude and width of the wave component cannot continue unabated.

## 5.2 Constrained drag minimisation

While the results presented above for asymmetric solutions of the fKdV equation require solving an optimisation problem in terms of  $\Delta$  and the domain length, it may in fact be able to remove the influence of  $L$ . As a consequence of this, the problem would be able to be reduced from a multidimensional optimisation problem to an ODE with an additional free parameter  $\Delta$ , subject to the closure equation (5.13).

The influence of the domain length can be removed if the numerical domain is restricted to only considering the region of support of the topographic forcing function. This can be attempted by considering the fKdV equation in its conservative form, integrated across  $x \in (\hat{x}, \infty)$

$$\int_{\hat{x}}^{\infty} A_{xx} + \Delta AA_x + 3A^2 A_x dx = \int_{\hat{x}}^{\infty} -\gamma f A_x dx. \quad (5.20)$$

By imposing that  $f(x)$  must have compact support over  $x \in [x_-, x_+]$ , then outside the region of the forcing it follows that

$$\frac{1}{2}A_x^2 + \frac{\Delta}{2}A^2 + \frac{r}{3}A^3 = \begin{cases} 0 & \text{for } x \leq x_-, \\ -\frac{\Delta^3}{6r^2} & \text{for } x \geq x_+. \end{cases} \quad (5.21)$$

Both of these equations are now integrable, the solutions of which are

$$\left. \begin{aligned} A &= -\frac{\Delta}{r} + \frac{1}{2}\Delta \operatorname{sech}^2\left(\frac{\sqrt{|\Delta|}}{2}(x - L_+)\right) & \text{for } x \leq x_-, \\ A &= \frac{1}{2}\Delta \operatorname{sech}^2\left(\frac{\sqrt{|\Delta|}}{2}(x - L_+)\right) & \text{for } x \geq x_+. \end{aligned} \right\} \quad (5.22)$$

To reduce the problem from solving over  $x \in (-\infty, \infty)$  to  $x \in [x_-, x_+]$  we could introduce the two equations above, evaluated at  $x_+$  and  $x_-$ . In doing so, two new free parameters  $L_+$  and  $L_-$  are introduced, for which there is no viable closure. Instead, the problem can be considered as a series of coupled nonlinear equations

$$\left. \begin{aligned} A_{xx} + \Delta A + rA^2 &= -\gamma f, \text{ for } A = A(x) \text{ and } f = f(x), \\ \int_{-\infty}^{\infty} \gamma f_x A dx - \frac{\Delta^3}{6r^2} &= 0, \\ \frac{1}{2}A_x^2(x_+) + \frac{\Delta}{2}A^2(x_+) + \frac{r}{3}A^3(x_+) &= -\frac{\Delta^3}{6r^2}, \\ \frac{1}{2}A_x^2(x_-) + \frac{\Delta}{2}A^2(x_-) + \frac{r}{3}A^3(x_-) &= 0, \end{aligned} \right\} \quad (5.23)$$

where the first equation is solved over  $x_- < x < x_+$ . Similarly, the integral can be reduced to

$$\int_{x_-}^{x_+} \gamma f_x A dx$$

as  $x \in [x_-, x_+]$  is the region of support of the topographic forcing function  $f(x)$ , and as such the integrand will be zero outside this region.

As a result of equation (5.23), the fKdV equation in the case where  $\Delta$  is a free parameter can now be considered to be well posed, subject to implementing the nonlinear boundary conditions. The integral equation for the drag can be evaluated within the GHAM through the use of the Gegenbauer Clenshaw-Curtis quadrature from Subsection 2.1.5 and equation (2.37).

An as of yet unresolved problem using this approach stems from the structure of  $f_x$ . As any even forcing function will result in  $f_x$  being odd, then any symmetric solution for  $A$  will satisfy the drag if  $\Delta = 0$ . As was discussed previously, for  $\gamma < 0$  there are an infinite set of symmetric solutions that exist for  $\Delta = 0$ , that appear to be more stable than their corresponding asymmetric solutions. As such, approaching the problem with the preceding methodology converges to the symmetric solutions where  $\Delta = 0$ , rather than the desired asymmetric solutions. It is possible that this problem can be alleviated through careful choice of the auxiliary linear operators used for the GHAM, and as such this technique warrants further exploration, even if it has not produced viable results to this point.

### 5.3 Discussion

While the parametric relationship for the fKdV equation in terms of  $(\gamma, \Delta)$  has previously been elucidated by Ee and Clarke (2007), their process required solving a numerically complicated optimisation process without a clear objective function. In contrast to this, the work contained within this chapter considered the parametric relationship in the context of the drag induced by the topography. In doing so, we were able to produce an integral condition upon  $u$  that greatly reduces the complexity of resolving the parametric curve between  $\gamma$  and  $\Delta$ .

Through this technique we were able to replicate the results of Ee and Clarke (2007) for  $\gamma > -8$ , confirming the validity of the previously published results. However, the presented technique did display some difficulties in resolving the region where  $\gamma < -8$ , with the step size of the numerical continuation regime rapidly decreasing to the point where further explorations of the domain became impractical.

The solutions in the region where  $\gamma < -8$  have previously been shown to be, at best, only marginally stable; and the parametric curve in this region is close to  $\Delta = 0$ , at which point the integral constraint equation (5.13) is satisfied by any symmetric solution. These two

factors likely drive the difficulties in resolving the solutions over this component of the parameter space.

Resolving the issue of solutions within the region where  $\gamma < -8$  will be a particular focus of future work, as through an examination of the properties of the asymmetric fKdV we were able to find that the parametric relationship of Ee and Clarke (2007) will likely continue for  $\gamma$  less than the lower limit presented in the original work, which only found solutions within the region of  $\gamma > -24$ .

There are several potential avenues to modify the technique in a manner that would allow the region for  $\gamma < -8$  to be reliably resolved. The first potential methodological change would be to update the continuation approach used for this problem, as the solutions calculated within this chapter were based upon a simplified parameter deformation approach using adaptive step length. This amounted to using a previously calculated solution in the neighbourhood of  $\gamma$  as the starting point of the homotopy deformation process within the GHAM. This approach was chosen as the Genetic Algorithm optimisation requires changing the length scale  $L$  of the numerical domain, which introduces significant complications to the implementation of more complex, and more rigorous numerical continuation algorithms.

A secondary approach would be to change the manner in which the domain was explored. Currently the algorithm to trace out the parametric curve started at  $\gamma = 0$ , and then progressed across the domain in the positive and negative directions. However, over the course of this experimentation it was found that the solver was far more stable in the region of  $-8 \leq \gamma \leq 0$  when the starting point was taken as the solitary wave solution at  $\gamma = -8$  rather than at 0. While there is no analytic description of the solution at  $\gamma \approx -24$ , in Figure 5.3 it was shown that the solution at this point (and at all other points where  $\Delta = 0$ ) has a simple numerical description. This could then be used as the starting point for attempting to traverse the region where  $-24 \leq \gamma \leq 0$ .

Finally, while initial testing of the approach from Section 5.2 proved unsuccessful, if the technique can be modified so that it does not admit solutions where  $\Delta = 0$  then there is no reason that it should not prove successful. This technique is particularly promising in that it removes the dependence upon the domain length by reducing the solution to one over the region of topographic support. This in turn would mean that the problem could be

removed from falling under the aegis of optimisation problems, which would both simplify the solution process and open up more options for the numerical continuation approach employed to explore the parametric relationship between  $\gamma$  and  $\Delta$ .

The very fact that the presence of hydraulic solutions must correspond to minimas of the drag, and that this drag must converge to a fixed value is also an important result. Prior to deriving this result it had been assumed that in the long time limit unsteady, asymmetric solutions of the fKdV equation would exhibit nonlinear wavetrains that moved towards the boundary, and that these wave trains would never fully cease to exist. However, the constraints upon the drag suggest that the amplitudes of these nonlinear wavetrains will likely decay as time becomes large.

While the work on resolving the full parametric relationship between  $\Delta$  and  $\gamma$  remains ongoing, the overall contribution of this chapter to the broader understanding of hydraulic solutions to the fKdV equation is that the dynamics are significantly controlled by the drag upon the perturbation to the uniform flow. In light of the success of employing the GHAM for exploring the parameter and solution spaces of the fKdV equation, the more general case of the symmetric forced Gardner equation can now be considered.

## Chapter 6

# Symmetric solutions of the forced Gardner equation

As was discussed in Chapter 1, the fKdV equation can be considered as the limit of the more general forced Gardner equation. Rather than just incorporating a quadratic nonlinearity, the forced Gardner equation involves a balance between dispersion and both quadratic and cubic nonlinearities, and takes the form

$$A_{xx} + \Delta A + rA^2 + qA^3 = -\gamma f(x). \quad (6.1)$$

The addition of the cubic nonlinearity significantly enhances the scope of potential applications of this equation, relative to that of the fKdV equation, and increases the breadth of expression of the solution and parameter spaces. As was mentioned in Section 1.1 the forced Gardner equation can be expressed in terms of the introduced parameter  $\delta \in [0, 1]$ , which represents the relative balance between the quadratic and cubic nonlinearities contained within the equation. The limit where  $\delta \rightarrow 0$  involves the equation approaching the fKdV equation, while the case where  $\delta \rightarrow 1$  approaches the modified Korteweg–de Vries (mKdV) equation.

The incorporation of the additional nonlinearity necessitates an extension of the parameter space from that of the fKdV equation. As such, the parameter space of the forced Gardner equation exists as a three–dimensional manifold the equivalent spaces  $(\Delta, \gamma, A(0), \delta)$  or  $(L, \gamma, A(0), \delta)$ . This contrasts to the two–dimensional manifold in  $(L, \gamma, A(0))$  of the fKdV



equation.

To explore this parameter space, the specific case of fixed  $L$  will be considered, and the triptych of free variables  $(\gamma, A(0), \delta)$  will be explored in terms of a series of slices across  $\delta$ . In doing so, the three-dimensional manifold is reduced to covering two-dimensions, a choice that was taken to both enhance the clarity of the presented parameter spaces, and to aide in the comparison between the parameter spaces of the fKdV equation and the forced Gardner equation.

In contrast to the fKdV equation, the forced Gardner equation does not admit any direct symmetries, except in the cases where  $\delta = 0$  and  $\delta = 1$ , which again correspond to the fKdV and mKdV equations respectively. As the symmetries of the fKdV equation were already discussed within Chapter 4, we turn to the case of the mKdV equation. This equation admits the symmetry that as  $A \rightarrow -A'$ , then  $\gamma \rightarrow -\gamma'$ . In terms of the  $(A(0), \gamma)$  parameter space, this symmetry corresponds to a rotation around both axis. While the forced Gardner equation may not admit any similar symmetries across all possible values of the doublet  $(r, q)$ , in the case where the forced Gardner equation approaches the fKdV and mKdV limits, then the parameter space will approach the respective symmetries of these equations.

## 6.1 Unforced solutions

Due to the similarities between these two equations, it follows that the basic features of the forced Gardner equation, of the form equation (6.1), and its parameter space can be elucidated in a similar manner to that of the fKdV equation by examining its unforced case.

Subject to the condition that  $\Delta < 0$ , the Gardner admits the bright solitons (Kamchatnov et al., 2012)

$$A = \frac{-3\Delta\mu + 1}{2} \frac{1}{r \mu + (1 - \mu) \cosh^2(\frac{1}{2}\sqrt{|\Delta|x})}, \quad (6.2)$$

where the equivalent solitons for  $\Delta > 0$  can be constructed through the mapping

$$\hat{A} = A + \frac{r \pm \sqrt{r^2 - 4\Delta q}}{2q}. \quad (6.3)$$

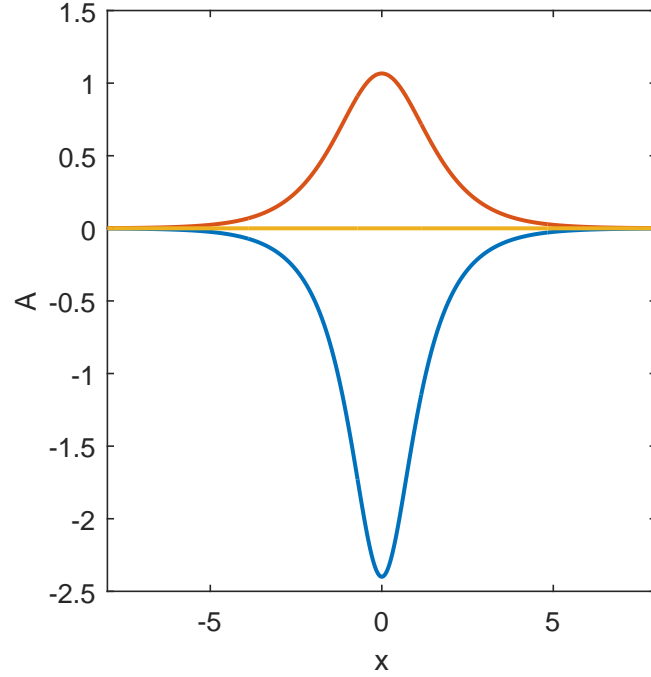


Figure 6.1: The form of the two solitary waves of the Gardner equation (in blue and red) and the trivial solution (yellow) for  $\Delta = -1.92$ ,  $r = \frac{3}{2}$  and  $q = \frac{3}{2}$ .

Within equation (6.2) the soliton parameter  $\mu$  is any extant root of the equation

$$(1 + \mu) \left( \mu^2 + 2 \left( 1 - \frac{4r^2}{9q\Delta} \right) \mu + 1 \right) = 0. \quad (6.4)$$

As this is a cubic equation in terms of  $\mu$ , there must be at most three distinct values of  $\mu$ , and in turn three corresponding solutions, for which one must always be a trivial solution, with the form of all three demonstrated within Figure 6.1. The introduced parameter  $\mu$  must also be restricted to lie within  $(-\infty, 1)$  due to the emergence of a pole for  $\mu \geq 1$  (Kakutani and Yamasaki, 1978, Miles, 1979). This equation admits three distinct solutions for  $\mu$  when  $r^2/(q\Delta) < 0$  or  $r^2/(q\Delta) > 9/2$ ; two distinct solutions when  $r^2/(q\Delta) = 0$  or  $9/2$ , and only admits the trivial solution for  $A$  when  $0 < r^2/(q\Delta) < 9/2$ .

Further insight into the form of the solution space of the Gardner equation can be gleaned from its phase portrait. Subject to  $q \neq 0$ , the Gardner equation admits the three fixed points

$$\left. \begin{aligned} (A, A_x) &= (0, 0) \text{ and} \\ (A, A_x) &= \left( -\frac{r}{2q} \pm \frac{\sqrt{r^2 - 4\Delta q}}{2q}, 0 \right). \end{aligned} \right\} \quad (6.5)$$

The stability of these stationary points can be considered in terms of the Hamiltonian of the forced Gardner equation—

$$\left. \begin{aligned} H(A, A_x) &= V(A) + \frac{1}{2}A_x^2 \text{ where} \\ V(A) &= \frac{1}{2}\Delta A^2 + \frac{r}{3}A^3 + \frac{q}{4}A^4. \end{aligned} \right\} \quad (6.6)$$

By considering the derivatives of  $V(A)$

$$\left. \begin{aligned} V'(A) &= \Delta A + rA^2 + qA^3 \\ V''(A) &= \Delta + 2rA + 3qA^2 \end{aligned} \right\} \quad (6.7)$$

at the first critical point—where  $(A, A_x) = (0, 0)$ —then  $V''(0) = \Delta$  will be strictly positive if  $\Delta > 0$ , and thus this critical point will correspond to a centre in this case. Alternatively, if  $\Delta < 0$  then so too will be  $V''(0)$ , and as such this critical point will be a saddle—behaviour that mimics that seen within the KdV equation.

For the second and third stationary points at

$$(A, A_x) = \left( -\frac{r}{2q} + \frac{\sqrt{r^2 - 4\Delta q}}{2q}, 0 \right)$$

the positive branch stationary point will correspond to  $V'' > 0$ , and thus the critical point will be a stable centre when

$$\alpha(2r - 3q\alpha) < \Delta \text{ where } \alpha = \frac{qr}{2} - \frac{\sqrt{r^2 - 4\Delta q}}{2q}.$$

The alternate case is where  $V''$  will be strictly negative, and thus a saddle for

$$\alpha(2r - 3q\alpha) > \Delta.$$

Similarly, the negative branch stationary point will be a stable centre for

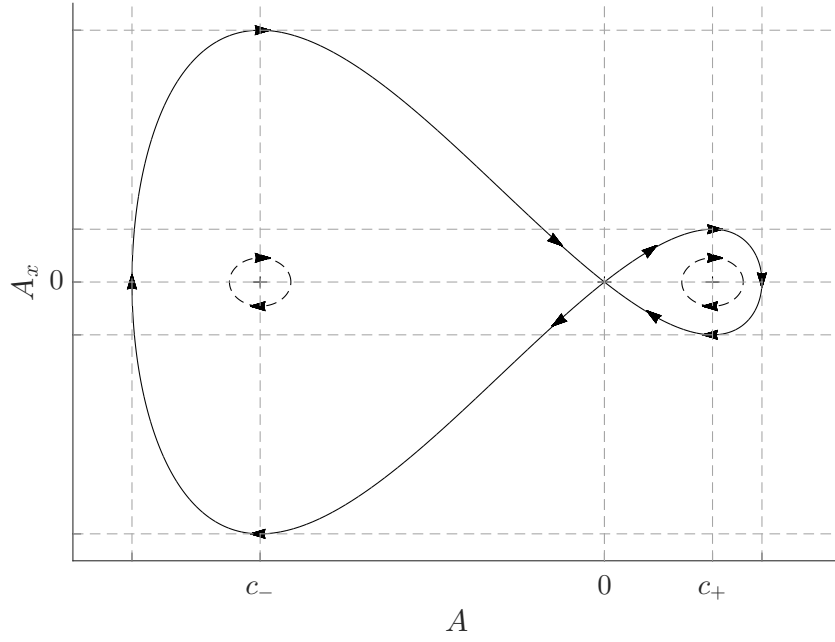


Figure 6.2: Phase portrait of the Gardner equation for  $\Delta < 0$  and  $r = q = 3/2$ , with the  $c_+$  and  $c_-$  denoting the location—in  $A$ —of the critical points at  $(A, A_x) = \left(\frac{-r}{2q} \pm \frac{\sqrt{r^2 - 4\Delta q}}{2q}, 0\right)$ .

$$\beta(2r - 3q\beta) < \Delta \text{ where } \beta = \frac{qr}{2} + \frac{\sqrt{r^2 - 4\Delta q}}{2q}$$

The phase portrait admitted by these stationary points can be seen in Figure 6.2.

The two solitons from equation (6.2) can be extended to incorporate forcing with local support in the same manner as Section 4.1, where the fKdV equation was solved subject to a  $\delta(x)$  forcing function. Beginning by assuming the solitons under the local forcing regime can be described by

$$A(x) = \begin{cases} -\frac{3\Delta}{2} \frac{\mu + 1}{r} \frac{1}{\mu + (1 - \mu) \cosh^2(\frac{1}{2}\sqrt{|\Delta|x - L^+})} & \text{for } x \geq 0, \\ -\frac{3\Delta}{2} \frac{\mu + 1}{r} \frac{1}{\mu + (1 - \mu) \cosh^2(\frac{1}{2}\sqrt{|\Delta|x - L^-})} & \text{for } x < 0, \end{cases} \quad (6.8)$$

then by incorporating the continuity and jump conditions from the  $\delta(x)$  forcing

$$\left. \begin{aligned} A(0^+) = A(0^-) &\equiv A(0) \text{ and} \\ A'(0^+) - A'(0^-) &= -\gamma, \end{aligned} \right\} \quad (6.9)$$

it follows that  $L^+ = \pm L^-$ . The jump condition then imposes that

$$\frac{\sinh(L^+) \cosh(L^+)}{(\mu + (1 - \mu) \cosh^2(L^+))^2} - \frac{\sinh(L^-) \cosh(L^-)}{(\mu + (1 - \mu) \cosh^2(L^-))^2} = \frac{-\gamma r}{3(1 + \mu)(1 - \mu)} |\Delta|^{-3/2}. \quad (6.10)$$

This can only be satisfied if  $\gamma = 0$ , or if  $L^+ \neq L^-$ , and as such setting  $L^+ = -L^- \equiv L^0$  gives rise to the nonlinear equation in  $L^0$

$$\frac{\sinh(L^0) \cosh(L^0)}{(\mu + (1 - \mu) \cosh^2(L^0))^2} = \frac{-\gamma r}{6(1 + \mu)(1 - \mu)} |\Delta|^{-3/2}. \quad (6.11)$$

While there is no closed form analytic solution to  $L^0$ , it should be clear that there can be at most two solutions of  $L^0$  for every  $\mu \neq -1$ . This in turn suggests a change in the manifestation of the bifurcation diagram of the forced Gardner equation, relative to that of the fKdV equation. Within the context of the fKdV equation, that there was only one solitary wave, that existed for  $A(0) > 0$ , meant that there was no possibility for a second zero crossing at  $\gamma = 0$ , thus restricting the overall potential structure of the bifurcation diagram. In contrast, that there is one solitary wave for each of  $A(0)$  positive and negative for the Gardner equation suggests that there may be some degree of symmetry, with respect to the form of the solutions within the upper and lower halves of the Gardner equation bifurcation diagram. Of course, the upper and lower halves of the bifurcation diagram cannot be identical, as the respective absolute value of  $A(0)$  for both solitary waves differs.

This concept of symmetry was confirmed by exploring the bifurcation diagram in the case of local support. However, as was the case of the fKdV equation, the solutions for local support match the solutions for compact support, and as such rather than presenting this limiting case we shall now turn to examining the set of solutions subject to a topographic forcing function with compact support.

## 6.2 Positive cubic nonlinearity

For  $q > 0$ , the solution space for positive cubic nonlinearities of

$$\Delta A + 3(1 - \delta)A^2 + 3\delta A^3 + A_{xx} = -\gamma f(x) \quad (6.12)$$

will be explored for the supercritical case where  $\Delta = -1.92$  and subject to the bell-shaped topographic forcing function of the form

$$f(x) = \begin{cases} \cos^4\left(\frac{\pi}{2}x\right), & \text{for } x \in [-1, 1] \\ 0, & \text{for } |x| > 1. \end{cases} \quad (6.13)$$

The position in parameter space  $\Delta = -1.92$  was chosen based upon previously published work on the Gardner equation, although it must be reiterated that the results from Section 1.1 show that this position in parameter space can freely be mapped to others positions. For ease of comparison the forcing function employed for the forced Gardner equation is—with the exception of the scaling term, which has been omitted—identical to that used for the fKdV equation in Section 4.2.

The numerical solution will be constructed upon the truncated domain  $x \in [-25, 25]$  subject to a Logarithmic mapping to the Chebyshev domain  $[-1, 1]$ , The auxiliary linear operator  $\mathcal{L}$  for the GHAM was simply

$$\mathcal{L}[A] = A_{xx} + \Delta A,$$

as this choice of linear operator was able to resolve the broader parameter space.

Solving across  $(\gamma, A(0))$  for varying  $\delta$  in Figure 6.3 demonstrates the evolution of the parameter space from the fKdV equation—shown with the blue dotted line—to a range of relative balances between the quadratic and cubic coefficients. As  $\delta$  increases the topology of the parameter space changes rapidly, from a transitional regime that bears some expected similarities to the parameter space of the fKdV equation, to a heretofore unseen phase space that resembles a figure eight. This transition between these two systems occurs at  $\delta \approx 0.3725$ .

One point of curiosity is the terminal end of the lower branch solutions for  $\delta < 0.3725$ . The turn towards positive  $\gamma$ —shown in more detail in Figure 6.4—must occur as the Gardner equation must exhibit a second unforced solitary wave, but the question remains with regard to why the tail terminates within this transitional regime. The step size required to

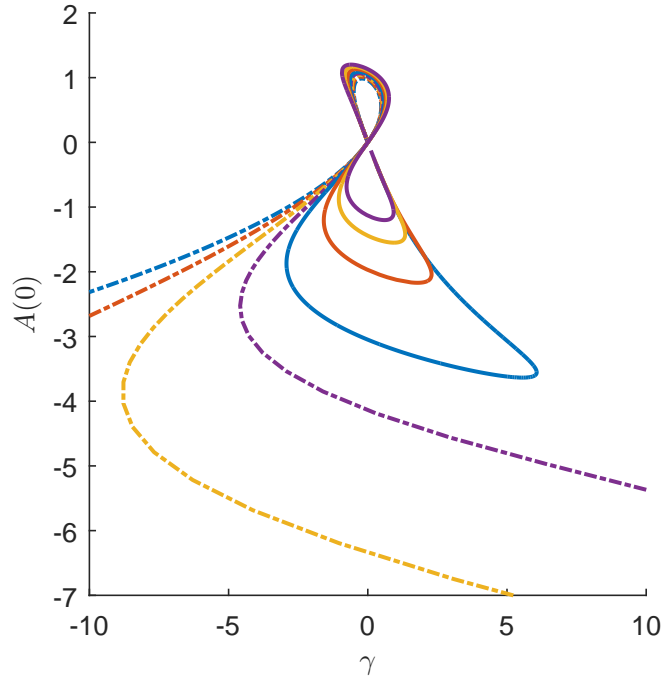


Figure 6.3: Solution space to equation (6.12) subject to the forcing equation (6.13) for varying  $\delta$ . Dashed lines correspond to  $\delta = \{0, 0.1, 0.2, 0.3\}$  for the blue, red, yellow and purple lines respectively. Solid lines corresponding to  $\delta = \{0.4, 0.6, 0.8, 1\}$  for the same colour progression.

resolve the numerical continuation along the lower branch of Figure 6.4 decreases rapidly as the terminal end is approached, which suggests a decline in the numerical stability of the solutions within this region, or of the applicability of the continuation regime. It is also true that all the nonzero  $\delta$  figure-eight shaped curves in parameter space have a finite length, in contrast to the infinite width Type V tails of the fKdV equation. As such, there must be some point in  $\delta$  at which the manifestation of the parameter space to the forced Gardner equation transitions from an infinite length path through parameter space, to a finite length path.

This question can be resolved by considering the asymptotic limit of this problem. In the limit where  $\gamma \rightarrow 0$  it can be assumed that  $\gamma f(x) \approx \gamma \delta(x)$ , and as such as long as a solution exists it can be constructed analytically. By considering  $\gamma = \pm\epsilon$  within equation (6.8) and equation (6.10) for some  $\epsilon \ll 1$ , the location of  $A(0)$  can be considered for varying  $\delta$ . In this limit there are four extant solutions for each of  $\gamma = \pm\epsilon$ , and as such the terminal

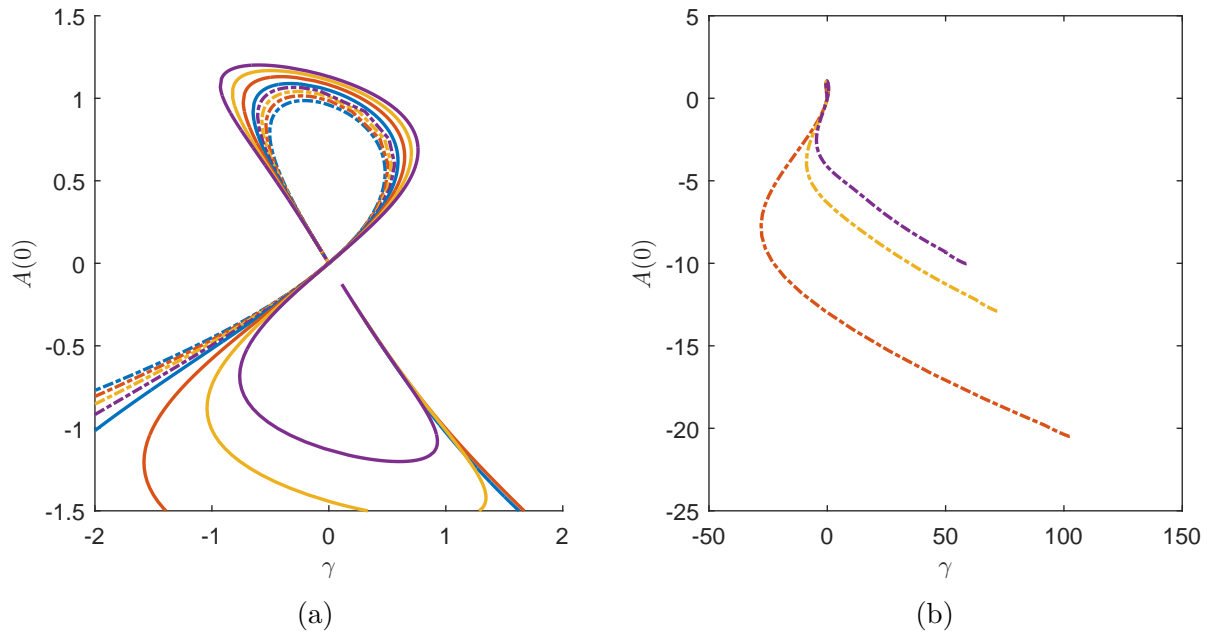


Figure 6.4: Two perspectives on the solution space to elucidate the structure of equation (6.12) for varying  $\delta$ . Dashed lines correspond to  $\delta = \{0, 0.1, 0.2, 0.3\}$  for the blue, red, yellow and purple lines respectively. Solid lines corresponding to  $\delta = \{0.4, 0.6, 0.8, 1\}$  for the same colour progression.

ends should, in fact, turn back towards  $(\gamma, A(0)) = (0, 0)$ . That the parameter spaces presented above do not behave in this manner is most likely an artefact of the simplified arc-length continuation technique used for this problem. The simple stepping regime may be struggling to resolve regions with high radius of curvature. As such, it may be possible to find the final branch of solutions for small  $\delta$  by changing to one of the more numerically rigorous continuation regimes developed within Chapter 3.

Furthermore, this asymptotic approach also indicates that the gap between the lower branch and  $(\gamma, A(0)) = (0, 0)$ —as is visible in Figure 6.4a—should be filled, and that the solution branches should all converge upon the origin. That they do not in turn suggests that there are as yet unresolved issues with the solutions in this region, stemming from either the numerical stability of the problem or the stability of the solutions themselves. While it may not be possible to resolve the latter issue, the former may be addressed by more carefully selecting the auxiliary linear operation  $\mathcal{L}$  and the homotopy parameter  $\hbar$ .



Under the assumption that the solution branch in the fourth quadrant of the  $(\gamma, A(0))$  domain exists and should progress towards the origin, we considered the possibility of self-similarity between the solutions across the domain. While there is a degree of self-similarity in both the region above and below  $A(0) = 0$ , there is no scaling that ties together the entirety of the parameter space, and the corresponding solution space.

As the solution space exhibits topological similarity across  $\delta$ , with the primary differences within the solution space driven by the differential scaling of the amplitudes, the form of the solutions exhibited by the forced Gardner equation can be considered in terms of a single  $\delta$ . As such, we will focus upon the case of  $\delta = 0.6$ , which involves an approximate balance between the quadratic and the cubic nonlinearities.

In a similar manner to the fKdV equation, the variance in the topological manifestation of the solution space correspond to distinct regions of the parameter space. As such, to explore the solution space Figure 6.5 is partitioned into four regions of varying colours. Within the blue region—characterised by  $A(0) > 0$  and  $\gamma_c < \gamma$  for  $\gamma_c \approx 0.1$ —two distinct forms of solutions can be identified in Figure 6.6. Labelling the first form of solutions as Type I, these correspond to perturbations from the unperturbed freestream solution, the characteristics of which present as small amplitude bell-shaped solitons. The Type II solutions exhibit broadly similar characteristics, although they are instead perturbations from the positive  $A(0)$  solitary wave solution of the Gardner equation. Both of these categories of solution involve smoothly varying solitary waves. In the context of the phase plane for the forced Gardner equation, the Type I and II solutions involve traversing the homoclinic orbit for  $A > 0$  from the saddle in a clockwise direction, with a jump corresponding to the topographic forcing prior to passing the centre for a Type I solution, and after for the corresponding Type II.

The next subset of Figure 6.5 is the red region, which contains the Type III and Type IV solutions. Type III solutions contain minor cusping, with the dual peaks varying in both their position and height with  $\gamma$ . The Type IV solutions, while broadly similar, involve sharper cusping. While these solutions similarly exhibit variance in the location of the dual peaks, and with the depth of the cusped region with  $\gamma$ , the magnitude of the dual peaks remains constant throughout the Type IV solutions, differentiating them from the Type III solutions. In terms of their manifestation in the phase plane, both solutions

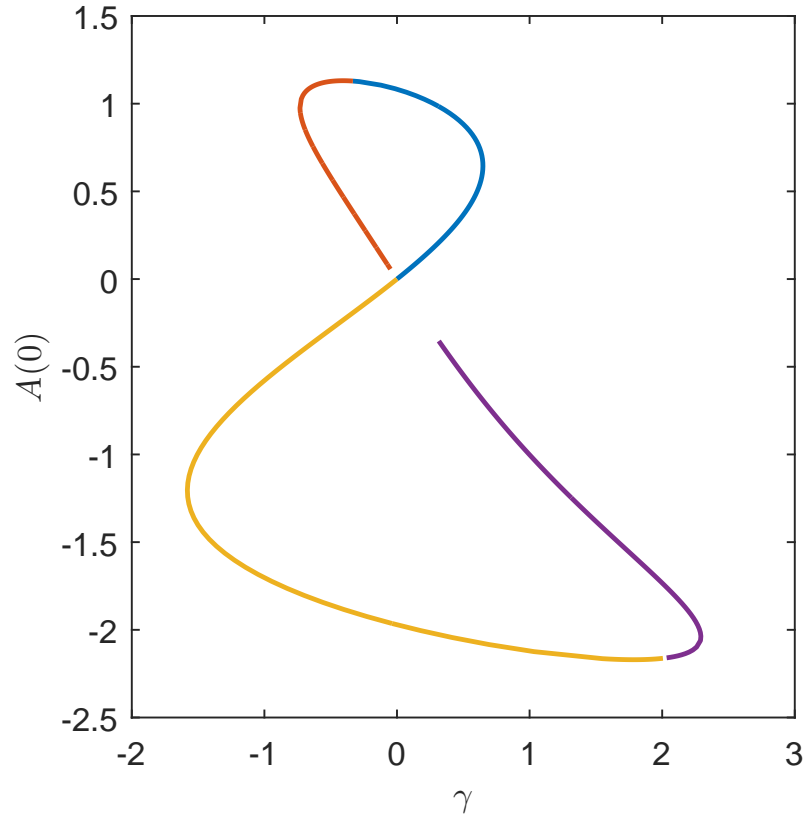


Figure 6.5: Parameter space for the forced Gardner equation in the form equation (6.12) subject to the forcing function equation (6.13) for  $\delta = 0.6$ , divided into different colours to cover regions of distinct solution phenomenology.

begin at the saddle at  $(A, A_x) = (0, 0)$  before traversing clockwise around the homoclinic orbit from the saddle past the point of zero crossing in  $A_x$ . As the solution reaches the region of support for the forcing function, there is a jump from the lower half of the homoclinic orbit back onto the upper half again, before again traversing the homoclinic orbit in the clockwise direction until the orbit returns to the saddle. The Type III and IV solutions can be distinguished by the location of the jump—occurring before the passing the centre for a second time in the Type III solutions, and after in the Type IV solutions.

Type V and VI solutions can be found in the yellow region of Figure 6.5, and broadly share the same characteristics as the Type I and II solution. The Type V solutions are perturbations from the uniform freestream flow, but exist with  $A(x) \leq 0$  for all  $x$ . Similarly

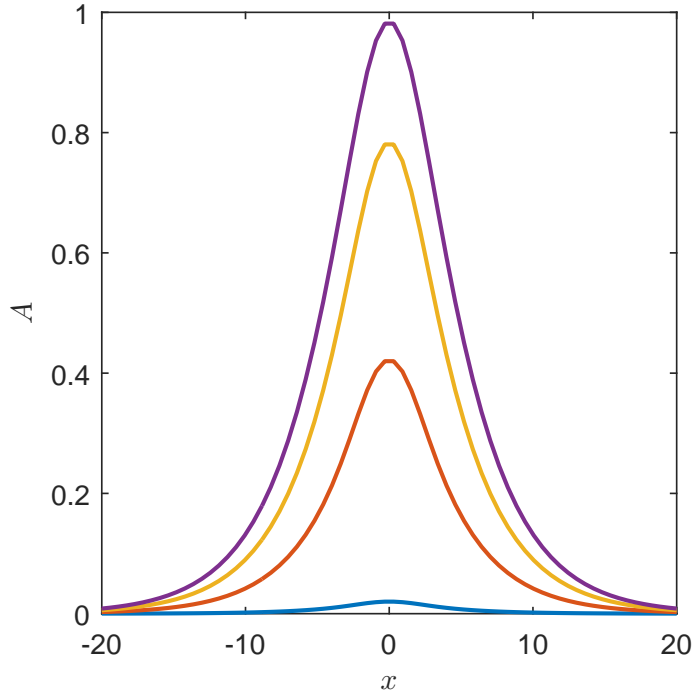


Figure 6.6: Type I (Blue, Red) and II (Yellow, Purple) solutions of equation (6.12) for  $\delta = 0.6$ . These solutions correspond to the blue region of Figure 6.5, subject to the forcing function equation (6.13).

to the Type II solutions, the Type VI solutions are perturbations for the negative solitary wave of the forced Gardner equation, and correspond to the first region of the solution space of the forced Gardner equation that does not have a corresponding form within the solutions of the fKdV equation. The existence of this solution should be of no surprise, as it is only logical that the addition of an extra solitary wave solution would engender solutions that are perturbations of that soliton. In terms of the phase plane, the Type V and VI solutions mirror the form seen in the Type I and Type II solutions, with the exception that the latter solutions correspond to traversing around the homoclinic orbit in a counter-clockwise direction through the region where  $A < 0$ .

The difference between the Type I and II solutions and their corresponding Type V and Type VI solutions, in terms of both their solution spaces and their manifestation in the parameter space of Figure 6.5 can almost entirely be described in terms of the phase plane. As the centres of the phase plane have an asymmetric distribution—with the two centres

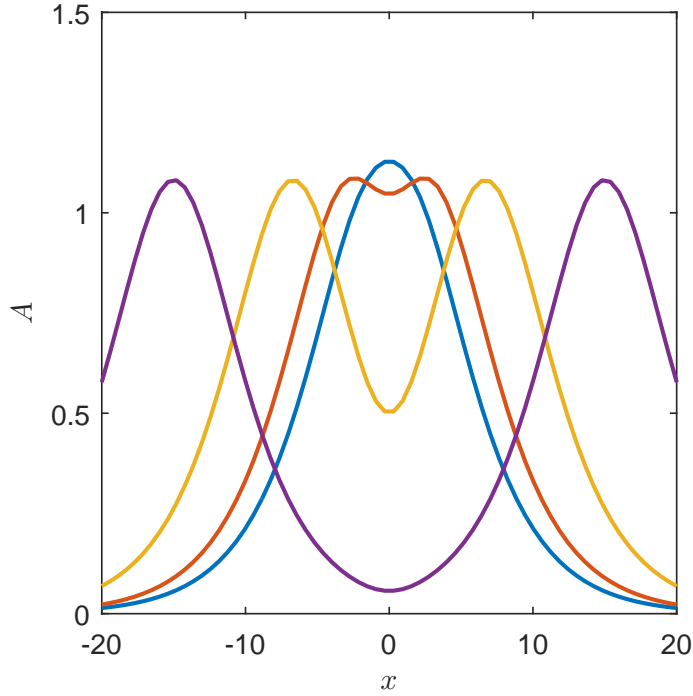


Figure 6.7: Type III and IV solutions of equation (6.12) for  $\delta = 0.6$ . These solutions correspond to the red region of Figure 6.5, subject to the forcing function equation (6.13).

located at  $(A, A_x) = (-r/2q \pm \sqrt{r^2 - 4\delta q/2q})$ —then the orbits that traverse from the saddle and around these centres must have a corresponding change in the manifestation of their amplitudes.

The final region of the solution space to be explored corresponds to the purple region of Figure 6.5, which covers the Type VII and Type VIII solutions. Like the Type V and VII solutions mirroring the first two forms of observed solutions, the Type VII and VIII solutions are effectively mirrors the evolution of the Type III and Type IV solutions. As was the case with the Type V and Type VI solutions, the change in the manifestation of the Type VII and Type VIII relative to those of the Type III and Type IV is a product of the asymmetry of the phase space and the ensuing homoclinic orbits.

While the above results are presented in the case where  $\delta = 0.6$ —which corresponds to a state where the contribution of the cubic and quadratic nonlinearities is roughly even—the results contained within are reflected for all  $\delta \in (0, 1]$ . The case corresponding to the fKdV

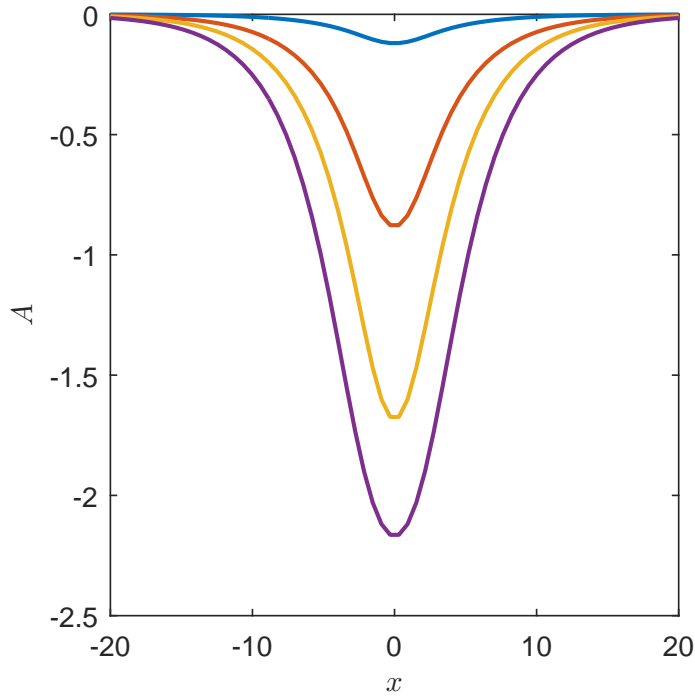


Figure 6.8: Type V and VI solutions of equation (6.12) for  $\delta = 0.6$ . These solutions correspond to the yellow region of Figure 6.5, subject to the forcing function equation (6.13).

equation where  $\delta = 0$  is, of course, an exception, although it does share similarities in the observed characteristics of the Type I–IV solutions.

### 6.3 Negative cubic nonlinearity

Unlike the fKdV equation, the forced Gardner equation does not contain any inherent symmetries that can be used to restrict the parameter space that needs to be explored. As such, to explore the parameter space of the forced Gardner equation we also need to consider the case where the coefficients of the cubic and quadratic coefficients are of opposite signs, i.e.

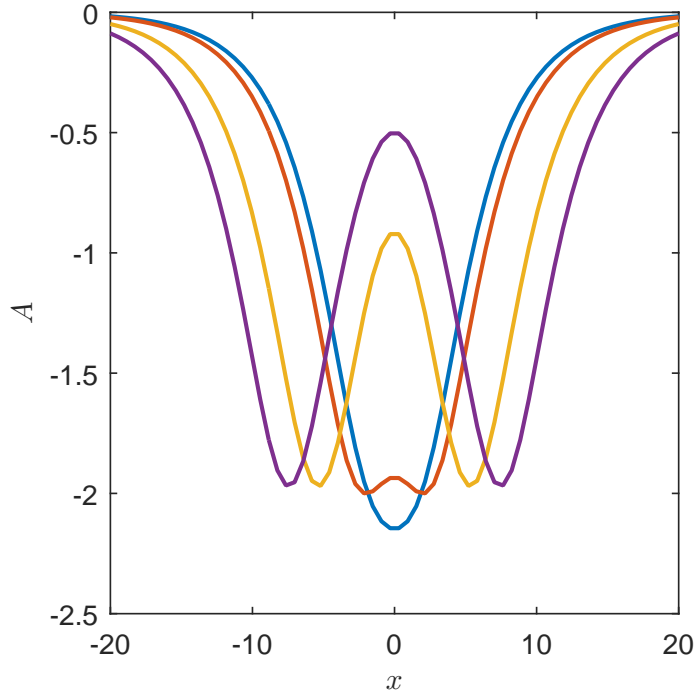


Figure 6.9: Type VII and VIII solutions of equation (6.12) for  $\delta = 0.6$ . These solutions correspond to the purple region of Figure 6.5, subject to the forcing function equation (6.13).

$$\left. \begin{aligned}
 \Delta A + rA^2 + qA^3 + sA_{xx} &= -\gamma f(x), \\
 r(\delta) &= 3(1 - \delta), \\
 q(\delta) &= -3\delta, \\
 \delta &\in [0, 1].
 \end{aligned} \right\} \quad (6.14)$$

Exploring this equation in the context of solutions with local forcing—where the forcing is determined by a  $\delta(x)$  function—suggests that the parameter space of the forced Gardner equation for  $q < 0$  can be partitioned into two distinct phenomenological regions with respect to  $\delta$ , with the demarcation defined by existence of a solitary wave solution. For  $\Delta = -1.92$ , one solitary waves exist for  $\delta \in [0, 1 - \frac{6\sqrt{86-36}}{25}]$ ; whereas in the region where  $\delta \in [1 - \frac{6\sqrt{86-36}}{25}, 1]$  the solitary wave vanishes and the entire solution space exists as perturbations of the zero flow solution.

This transition can also be explained in the context of the eigenvalues of the unforced system. When  $\Delta = 3(\delta^2 - 2\delta + 1)/4\delta$  there is a bifurcation marking the transition of one of the roots from a saddle to a centre, and when  $\Delta = -3(\delta^2 - 2\delta + 1)/4\delta$  the second root experiences the same bifurcation. However for a given  $\Delta$  only one of these transitions will exist for  $\delta \in \mathbb{R}$ . For the chosen  $\Delta$ , and restricting  $\delta$  to lie within  $[0, 1]$  this transition occurs at  $\delta \approx 0.231$ .

Another explanation of the transition can be derived by reconsidering the Hamiltonian equation (6.6) in the context of equation (6.14), so that

$$\left. \begin{aligned} H(A, A_x) &= V(A) + \frac{1}{2}A_x^2 \text{ where} \\ V(A) &= \frac{1}{2}\Delta A^2 + (1 - \delta)A^3 - \frac{3}{4}\delta A^4. \end{aligned} \right\} \quad (6.15)$$

Considering the forced Gardner equation in terms of its Hamiltonian potential  $V(A)$  allows for some insights to be gained into the form of extant solutions. As the number of zeros of  $V(A)$  changes, so too must the form of the waves admitted by the solutions. When the term

$$9/4(1 - \delta)^2 + 27/8\Delta\delta$$

is negative,  $V(A)$  does not contain any roots. However, when it is zero there will be one extant root, and when it is positive there will be two distinct roots. For  $\Delta = -1.92$ , this transition point occurs at

$$\delta = \frac{61 - 6\sqrt{86}}{25}$$

or  $\delta = \delta_c \approx 0.2143$ . This suggests that  $\delta > \delta_c$  should correspond to a distinct phenomenological region, with the transition occurring at  $\delta_c$ .

Prior to the transition taking place, the phase portrait of the Gardner equation subject to a negative cubic nonlinearity, Figure 6.10, broadly resembles that of the KdV equations phase portrait—shown in Figure 4.1—with the only difference being that as  $\delta$  increases the size of the homoclinic orbit in the region where  $A > 0$  decreases. Further increasing  $\delta$  heralds a completely new form of the phase portrait, which is radically different to what was observed for both the KdV equation and the equivalent case for  $q > 0$ . As there is only one critical point outside the transition region, shown in Figure 6.10 there are no closed

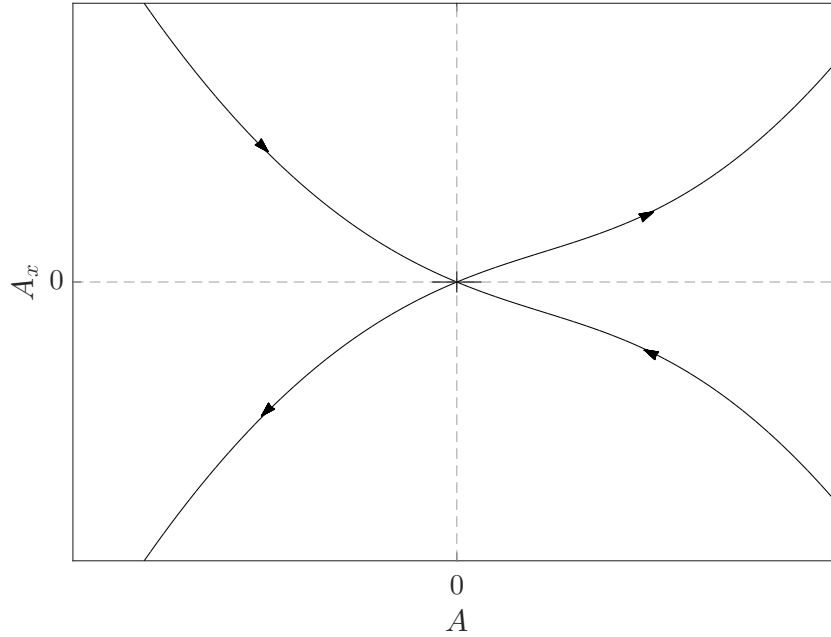


Figure 6.10: Phase portrait of the Gardner equation for  $\Delta < 0$  and  $r = -q = 3/2$  outside the transition region, with a single critical point at  $(A, A_x) = (0, 0)$ .

loops within the phase plane, so the solution space of this equation should, theoretically, only admit perturbations to the zero flow case with unbounded amplitudes.

To explore the ramifications of these transitions, the numerically determined solution and parameter spaces subject to the topographic forcing function equation (6.13) is explored in Figure 6.11. Beginning by considering the case where  $\delta = 0.1$ , at which point quadratic nonlinearity dominates the solution, relative to that of the cubic nonlinearity, the parameter space broadly mimics the progression seen within the fKdV equation. This similarity is reflected in the solution set shown in Figure 6.12a as well. As  $\delta$  increases to 0.213 the shape of the parameter space broadly resembles that of fKdV equation, with the only difference being the vertical scale of the region where  $A(0) > 0$ . In this region the Type II and III solutions—to return to the classification from the fKdV equation—become a more predominant component of the parameter space. In terms of the solution space, this manifests in a wider range of maximum amplitudes for the Type II and III solutions, as evidenced in Figure 6.12b. There also appears to be a slight flattening of some of the large amplitude solutions in the neighbourhood of  $x = 0$ , however this is more likely to be



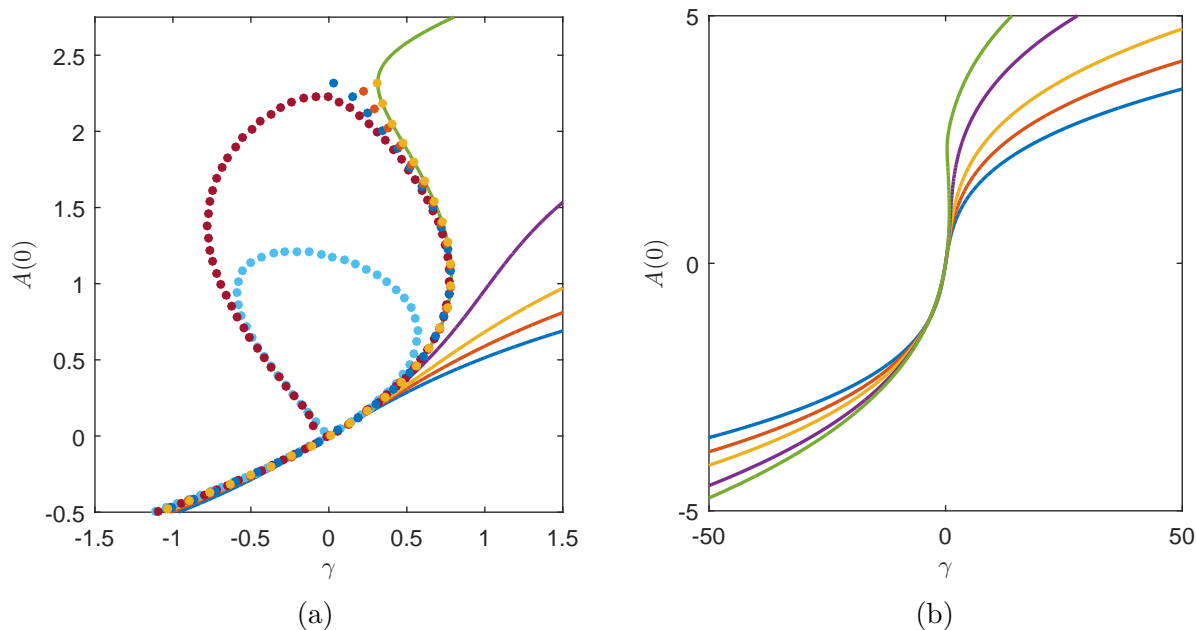


Figure 6.11: Parameter space for the forced Gardner equation (6.14) subject to the forcing function equation (6.13). Figure 6.11a presents solutions within the vicinity of the origin, and Figure 6.11b shows the evolution of a subset of those solutions. Of the solid lines, blue, red, yellow, purple and green are  $\delta = \{1, 0.7, 0.5, 0.3, 0.216095\}$  respectively. Of the dotted lines, yellow, red, blue, burgundy and light blue correspond to  $\delta = \{0.21609, 0.215, 0.214, 0.213, 0.1\}$ . Thus the dotted light blue solution is close to the fKdV equation, and the dark blue solid line corresponds to the mKdV equation where there is only a cubic nonlinearity.

evidence of having selected a solution just after the transition from Type II (which exhibit a single solitary wave) to Type III solutions (typified by two interfering solitons with a small cusping between them), rather than an effect of saturation.

At  $\delta = 0.213$  the phase portrait begins to transition away the form of the phase portrait admitted by the fKdV equation, with the first signs of the upper loop unfolding being visible as  $(\gamma, A(0))$  approaches  $(0, 0)$  from the upper left side. As  $\delta$  increases beyond this point there is a clear bifurcation in the dynamics of the parameter space, corresponding to a transition from the region where the positive  $A(0)$  unforced solitary wave is extant, to where it is not. The location of this bifurcation point closely aligns with the analytic predictions outlined earlier in this section, which suggested that the transition would occur when  $\delta$  was in the neighbourhood of 0.214 and 0.231. Solutions for  $\delta$  larger than 0.213 show

a rapid progression from a transitional scheme that resembles an unfolding of the upper region of the fKdV equations phase portrait to the strictly s-shaped profiles of Figure 6.11b. The sensitivity of the phase portrait to  $\delta$  in this region is evidenced by the rapid change as  $\delta$  progresses from 0.21609 to 0.216095. The profiles corresponding to two points in  $\delta$  space are practically indistinguishable in Figure 6.11a up until the point where  $A(0) \approx 2.25$ , at which point the  $\delta = 0.21609$  solutions terminate, whereas the solutions corresponding to 0.216095 exhibit an inflection point and continue off for  $\gamma \rightarrow \infty$ .

This phenomenological difference is, of course, reflected in the solution space for these two values of  $\delta$ . Figure 6.12d presents solutions that are perturbations from the zero flow solutions (in both the positive and negative direction), and from the standard solitary wave: solutions that are analogous to the Type I, Type V and Type II solutions from the fKdV equation. The solutions at  $\delta = 0.216095$  present a far more unique point of the solution space, and take a form seen in Figure 6.13a that only vaguely corresponds to the preceding work.

For the case of the solutions corresponding to a negative forcing amplitude, the solutions are all perturbations from the zero flow solution, and present in a similar manner to the Type V solutions from the fKdV equation, although their growth—shown in Figure 6.11—does progress at a different rate. What is more interesting is the positive amplitude solutions, which progress from the unforced solution up until the point where the amplitude approaches one, at which point the solutions begin to grow faster in width than they do in height, and in doing so show the first signs of saturation with the solution space of the forced Gardner equation. In the context of Figure 6.11a, this saturation process begins at the point where the parameter space begins to fold back on itself. Interestingly, the difference between the values of  $\delta$  at 0.21609 and 0.216095 are incredibly small at this point, yet they are enough to warrant this marked shift in the topology of the solution space.

As the amplitude increases further, the saturated shelf-like solutions begin to exhibit an inflection point around  $x = \pm 1$ , which presents as a secondary wave structure superimposed upon the central region of the shelf. With increases in  $A(0)$  it is this secondary wave structure that drives the change in the solution space, as the shelf-like region retains an almost constant width and height, with only minor changes in both, relative to the growth

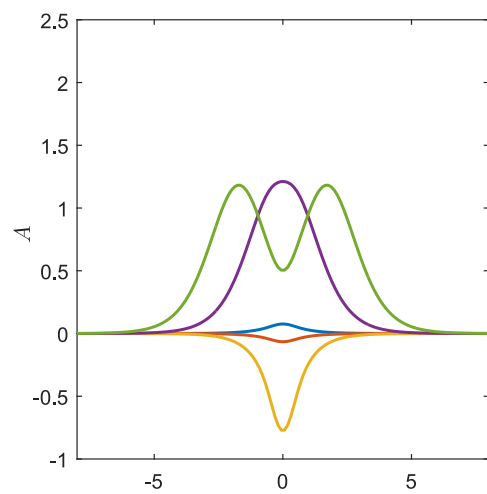
of the secondary wave.

These solutions, and in fact all of the solutions in Figure 6.13, solely represent perturbations to the zero flow solutions, rather than from the solitary wave—even if they do share a strong resemblance to the solitary wave solutions. The difference between the solutions being perturbations from the solitary wave, to being perturbations from the unforced solution is more readily apparent when considering the symmetry of the presented solutions. In the case of the fKdV equation, the form of a solution around the solitary wave with a wave amplitude of  $a$ , would be markedly different to the corresponding wave with an amplitude of  $-a$ , with the positive amplitude solution being broader across all amplitudes. Whereas in Figure 6.12, the solutions at  $a$  and  $-a$  are reflections of each other. This is not to say that these results are a product of symmetries in the equation, but rather that both the positive and negative branches are the product of the same perturbations away from the unforced flow solution. The exception to this is the case where  $\delta = 1$ , which does have an inherent symmetry within its parameter space.

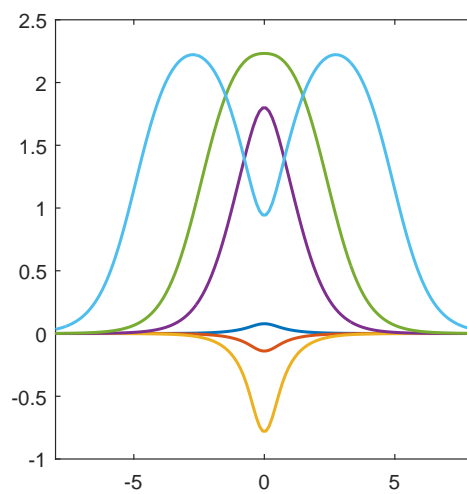
## 6.4 Discussion

Here we have generated the first rigorous exploration of the parameter space of symmetric solutions to the forced Gardner equations, as it transitions from the pure quadratic non-linearity of the fKdV equation to the purely cubic mKdV equation. In the case where the coefficients of the cubic and quadratic nonlinearities are of the same sign, the discoveries regarding the behaviour for positive  $A(0)$  are broadly similar to those seen within the fKdV equation. However, the presence of the second distinct solitary wave solution within the forced Gardner equation perturbs the evolution of the negative region of  $A(0)$ , which then broadly mirrors the positive  $A(0)$  region. The exceptions to this come from the solutions wherein the effects of the cubic coefficient are small, relative to that of the quadratic coefficient.

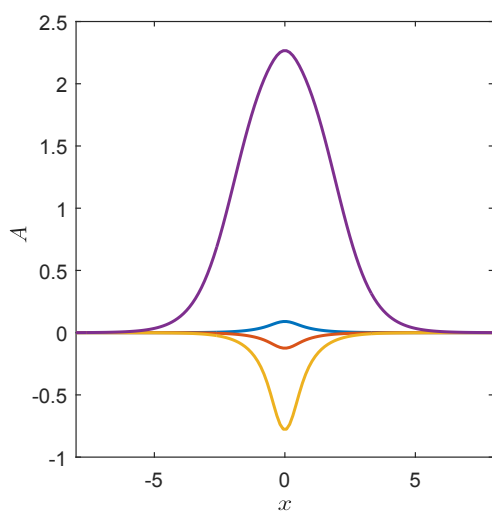
For negative  $A(0)$ , these solutions progressed from  $(\gamma, A(0)) = (0, 0)$  down to the corresponding negative amplitude solitary wave solution, however, unlike the solutions where the coefficients were of the same order, or when the cubic term dominates, they did not exhibit a return to  $(\gamma, A(0)) = (0, 0)$ . By considering the asymptotic case of small  $\gamma$ , it was able to be determined that these solutions likely do exist, however the small radius of curvature within



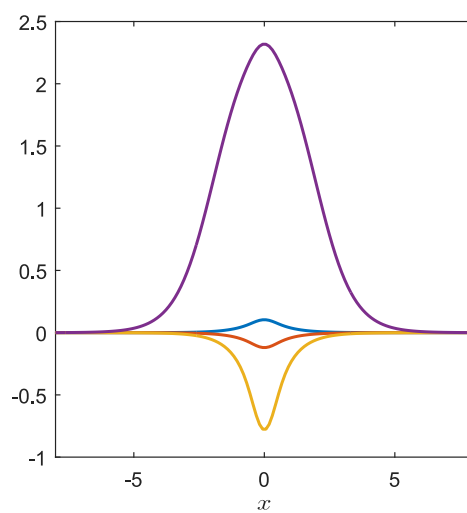
(a)  $\delta = 0.1$



(b)  $\delta = 0.213$



(c)  $\delta = 0.215$



(d)  $\delta = 0.21609$

Figure 6.12: A representative sample of the solution set for  $\delta = \{0.1, 0.213, 0.215, 0.21609\}$ , corresponding to Figure 6.11a.

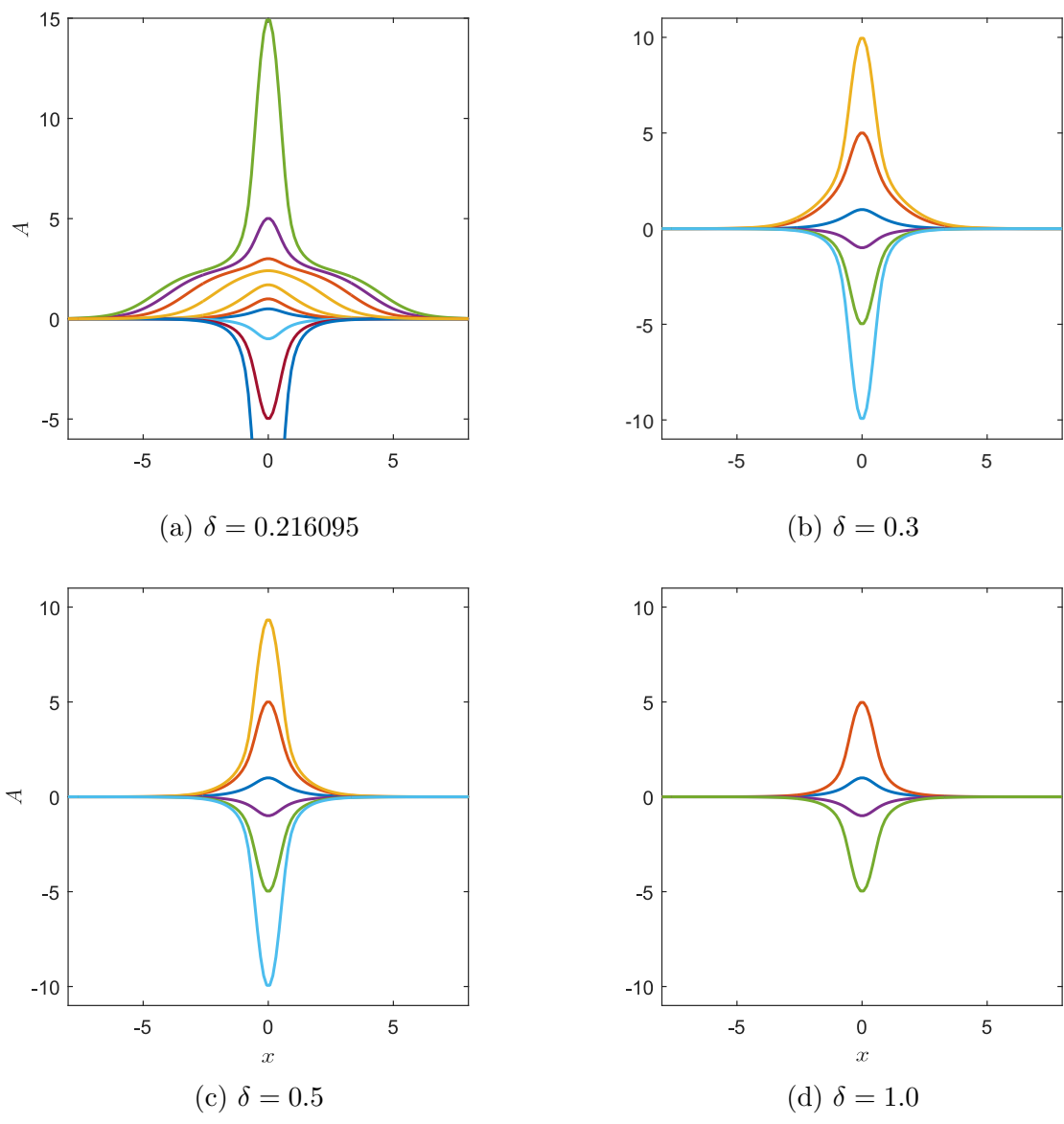


Figure 6.13: A representative sample of the solution set for  $\delta = \{0.216095, 0.3, 0.5, 1.0\}$ , corresponding to Figure 6.11b. With the exception of the case where  $\delta = 0.216095$ , the blue, red, yellow, purple green and teal plots correspond to solutions at  $A(0) = \{1, 5, 10, -1, -5, -10\}$  respectively. Due to the unique taxonomy of solutions in the  $\delta = 0.216095$  a broader range of solutions for  $-10 \leq A(0) \leq 15$  has been presented.

fold point marking the transition region makes resolving this branch difficult. It is hoped that these folds will be able to be traversed by incorporating the new numerical continuation regimes presented in this work, however to this point this still remains an open problem.

In contrast to this, the region of parameter space wherein the cubic and quadratic coefficients are of mixed sign appears to be fully explored. This parameter space presents some very interesting phenomenological behaviour, with marked changes in the regions where each of the nonlinearities dominates. When the quadratic coefficient dominates, the problem exhibits one positive solitary wave solutions, and the solution space exists in terms of perturbations from the zero flow solution and from the solitary wave. The solitary wave perturbations give rise to cusped solutions, of the like seen within the solution space for the fKdV equation.

As was shown analytically the existence of the second solitary wave does not persist as the balance between the two coefficients shifts toward the cubic coefficient, and this drives a marked change from a parameter space that replicates the fKdV equation, to hitherto unseen S-shaped profile. Within the transition between these two different presentations of the parameter space, the Gardner equation admits the most interesting parameter space, with a broad proportion of the solution space being dominated with saturated waves presenting with quasi-tabletop forms. Outside the transition region of  $\delta$ , the solution space exclusively take the form of bell-shaped waves, where the magnitude of amplitude increases with the magnitude of  $\gamma$ .

# Chapter 7

## Conclusion

This project was motivated by the study of the effects of topography upon Rossby wave breaking events, with a particular eye to the work of Benney (1979). Within the Benney framework, our initial approach resulted in a sequence of nonlinear integrodifferential equations, which initially appeared to be computationally intractable. This problem served as the genesis for developing the GHAM, as it was clear that there was a need for a numerically efficient variable coefficient solver for nonlinear boundary value problems, which also both flexibility and strong convergence control properties .

The utility of this solver was preserved as the motivation for this work transitioned from the Benney framework to considering Rossby waves under the aegis of the forced Gardner equation and its variants. As such, this thesis has focussed upon the mathematical advances that lead to the development of the GHAM and its associated tools; testing and analysing the properties of GHAM in the context of boundary layer flows; and considering solutions finite amplitude nonlinear wave problems, specifically the fKdV and forced Gardner equations.

Chapter 2 outlined the numerical advances within the field of spectral methods, that in turn formed the basis of the techniques developed within this thesis. This was presented with a particular focus on Chebyshev and Gegenbauer methods for solving linear boundary value problems. In Chapter 3 we then extended these techniques to nonlinear boundary value problems by incorporating a numerical analogue of the Homotopy Analysis Method. This chapter also introduced a number of novel results regarding the convergence properties of the HAM, and introduced new tools for significantly enhancing the behaviour of

homotopy based methods. Of particular importance was Section 3.6, as it allows numerical analogues of the Homotopy Analysis Method to approximate  $\hbar_{\text{opt}}$ , and in doing so avoid the optimisation problem on the residual that, to date, has been intrinsic to these techniques.

One of the most important discoveries with the GHAM framework was that every step of the iterative process could be expressed in terms of the inversion operation of a sparse matrix, that is invariant with respect to the iteration number. As a direct consequence of this, we were able to show that the GHAM could exhibit a quasi-linear scaling of the computational cost with respect to the number of points used within the Gegenbauer discretisation, while also minimising the memory required to store the resulting matrix equations. In doing so, we were able to significantly outperform other techniques for solving steady, nonlinear boundary value problems.

After having established the properties of the GHAM, we sought to apply it to weakly nonlinear wave problems, specifically the fKdV equation and the forced Gardner equations. Chapter 4 served to both validate the performance of the GHAM and its associated continuation techniques for this form of wave problem, and to test the hypothesis that a rescaled  $\delta(x)$  forcing function can accurately approximate any topographic forcing function with compact support. This hypothesis was tested by considering the parameter space of the fKdV equation for both forcing cases in terms of  $\Delta$ , which relates to the Froude number of the system, and  $\gamma$ , the forcing amplitude. While the local forcing hypothesis has a well established place within the published literature regarding the fkdV equation, however when it was tested it failed to hold for even the most trivial of topographic length scales.

Extending the solver for the fKdV equation onto the case of asymmetric, hydraulically controlled solutions, we were able to investigate the parametric relationship between  $\Delta$  and  $\gamma$  again, which lead to both numerical results and physical insights. The specific approach taken was built upon the hypothesis that in the case of a steady, hydraulic solution the topographic drag acting upon the perturbation to the uniform flow must reach either a minimum or a maximum. By consideration of both the Hamiltonian, and the conservative form of the fKdV equation we were able to verify that the drag must indeed reach a global minima in the case of hydraulically controlled solutions. This in turn allowed the drag condition to be used as an optimisation criteria to elucidate the structure of the parametric relationship between  $\Delta$  and  $\gamma$ . The very fact that asymmetric solutions are conditioned



upon the drag, and that this drag has a finite, nonzero value also has implications for understanding the evolution of the nonlinear downstream wavetrains that exist within unsteady solutions of the asymmetric fKdV equation.

The final numerical result of this thesis was to consider the forced Gardner equation, which is a generalised nonlinear equation subject to both cubic and quadratic polynomials. In a similar manner to the fKdV equation, the parameter space of the forced Gardner equation was explored in terms of  $\Delta$  and  $\gamma$ , but to account for the extra degree of freedom within the Gardner equation we introduced an additional parameter that described the relative balance between the quadratic and cubic nonlinearities. The additional complexity inherent in the forced Gardner equation manifested itself in some interesting modifications to the parameter and solution spaces previously seen for the fKdV equation. This was particularly true for the case of mixed nonlinearities, wherein a bifurcation in the parameter space was able to be identified, and it was within the transition region of this bifurcation that we were able to uncover a solution space that included significant saturation effects.

While this thesis contains a number of novel results, it has also introduced a number of open questions and challenges that will continue to be pursued. On the numerical side of this, work the GHAM and its associated continuation techniques will be consolidated to form a single package for solving variable coefficient boundary value problems. Many opportunities exist to potentially generalise these tools to encompass a broader array of problems. Of particular focus will be considering solving problems in higher dimensions through GHAM; assessing the performance of Homotopic integrated arc-length continuation for problems that exhibit more complex bifurcations; and testing and extending the viability of the ideas regarding estimating  $\hat{h}_{\text{opt}}$  espoused within Section 3.6 and Section 3.7.

With regards to wave dynamics, the numerical results in both the asymmetric fKdV equation and the symmetric Gardner equations has led to additional questions about the existence of solutions in certain limits of the parameter space. Beyond this, additional work must be conducted to relate the numerical results to original motivation of assessing the physics of Rossby wave breaking events. As such, there are multiple exciting prospects for future research relating to the work contained within this thesis, both with regards to the numerical tools that have been developed, and for interpreting solutions generated by these tools within a geophysical fluid dynamics context.

# Bibliography

- ABBASBANDY, S., MAGYARI, E., AND SHIVANIAN, E. 2009. The Homotopy Analysis Method for multiple solutions of Nonlinear Boundary Value Problems. *Communications in Nonlinear Science and Numerical Simulations* 14:3530–3536.
- ABLOWITZ, M. AND CLARKSON, P. 1991. Solitons, Nonlinear Evolution Equations and Inverse Scattering. London Mathematical Society Lecture Note Series, Vol. 149, Cambridge University Press, Cambridge.
- ABLOWITZ, M. J. AND SEGUR, H. 1981. Solitons and the Inverse Scattering Transform. *SIAM Studies in Applied Mathematics* 4.
- M. Abramowitz and I. A. Stegun (eds.) 1972. Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables. New York: Dover, 9th edition.
- ADOMIAN, G. 1986. Nonlinear Stochastic Operator Equations. Kluwer Academic Publishers.
- ADOMIAN, G. 1994. Solving Frontier problems of Physics: The decomposition method. Kluwer Academic Publishers.
- ALI, J., ISLAM, S., ISLAM, S., AND ZAMAN, G. 2010. The Solution of Multipoint Boundary Value Problems by the Optimal Homotopy Asymptotic Method. *Journal of Computers and Mathematics with Applications* 59:2000–2006.
- ALLGOWER, E. L. AND GEORG, K. 2003. Introduction to Numerical Continuation Methods. *Classics in Applied Mathematics* 45.
- ANDRIANOV, I. V., AWREJCEWICZ, J., AND BARANTSEV, R. 2003. Asymptotic approaches in mechanics: New parameters and procedures. *Applied Mechanics Review* 56:87–110.
- ARTHUR, D. AND VASSILVITSKII, S. 2007. k-means++: The Advantages of Careful Seeding. *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. *SIAM* pp. 1027–1035.
- ATABAKAN, Z. 1974. Numerical solutions of the Korteweg-de Vries equation and its generalizations by the split-step Fourier Method. *Lectures in Applied Mathematics* 15:215–216.
- AUSTIN, J. 1980. The blocking of Middle Latitude Westerly Winds by Planetary Waves. *Quarterly Journal of the Royal Meteorological Society* 106:327–350.
- BAINES, P. G. 1987. Upstream Blocking and Airflow Over Mountains. *Annual Review of Fluid Mechanics* 19:75–97.

- BAINES, P. G. 1995. Topographic Effects in Stratified Flows. Cambridge University Press, New York.
- BARTLETT, M. S. 1951. An Inverse Matrix Adjustment Arising in Discriminant Analysis. *The Annals of Mathematical Statistics* 22:107–111.
- BASZENSKI, G. AND TASCHE, M. 1997. Fast Polynomial Multiplication and Convolutions Related to the Discrete Cosine Polynomial transform. *Linear Algebra and its applications* 252:1–25.
- BATCHELOR, G. 2000. An introduction to Fluid Dynamics. Cambridge University Press.
- BAYLISS, A., CLASS, A., AND MATKOWSKY, B. J. 1995. Roundoff Error in Computing Derivatives Using the Chebyshev Differentiation Matrix. *Engineering Sciences and Applied Mathematics* 116:380–38.
- BELFERT, B. D. 1997. Generation of Psuedospectral Differentiation Matrices I. *SIAM Journal on Numerical Analysis* 34:1640–1657.
- BELWARD, S. R. AND FORBES, L. K. 1993. Fully Non-Linear Two-Layer flow over Arbitrary Topography. *Journal of Engineering Mathematics* .
- BENNEY, D. J. 1966. Long Non-Linear Waves in Fluid Flows. *Studies in Applied Mathematics* 45:52–63.
- BENNEY, D. J. 1979. Large Amplitude Rossby Waves. *Studies in Applied Mathematics* 60:1–10.
- BENNEY, D. J. AND KO, D. R. S. 1978. The Propagation of Long Large Amplitude Internal Waves. *studies in Applied Mathematics* 59:187–199.
- BIAGIONI, H. A. AND LINARES, F. 1997. On the Benney-Lin and Kawahara Equations. *Journal of Mathematical Analysis and Applications* 211.
- BINDER, B. J., VANDEN-VANDEN-BROECK, J.-M., AND DIAS, F. 2005. Forced Solitary Waves and fronts past submerged obstacles. *Chaos* 15.
- BIRKISSON, A. AND DRISCOLL, T. A. 2012. Automatic Frechet Differentiation for the numerical solution of Boundary-Value Problems. *ACM Transactions on Mathematical Software* 38.
- BORISEVICH, A. AND SCHULLERUS, G. 2012. Switching strategy based on Homotopy Continuation for Non-Regular Affine systems with application in Induction Motor Control. *arXiv ref:1203.5919v3* .
- BOUSSINESQ, J. 1872. Theorie des ondes et des remous qui se propagent le long d’un canal rectangulaire horizontal, en communiquant au liquide contenu dans ce canal des vitesses sensiblement pareilles de la surface au fond. *Journal de Mathematiques Pures et Appliquees* .
- BOYD, J. P. 1980a. Equatorial Solitary Waves. Part I: Rossby Solitons. *Journal of Physical Oceanography* 10.
- BOYD, J. P. 1980b. The rate of convergence of Hermite function series. *Mathematical Computing* 35:1309–1316.

- BOYD, J. P. 1982. The optimization of convergence for Chebyshev Polynomial Methods in an unbounded domain. *Journal of Computational Physics* 45:43–79.
- BOYD, J. P. 1986. Solitons from Sine Waves: Analytical and Numerical Methods of Non-Integrable Solitary and Cnoidal Waves. *Physica D* 21.
- BOYD, J. P. 2001. Chebyshev and Fourier Spectral Methods. Dover Publications.
- BREUER, K. S. AND EVERSON, R. M. 1992. On the Errors Incurred Calculating Derivatives Using Chebyshev Polynomials. *Journal of Computational Physics* .
- BUREAU OF METEOROLOGY 1984. Report on the Meteorological Aspects of the Ash Wednesday Fires - 16 February, 1983. Australian Bureau of Meteorology 17, pp. 143.
- CAMASSA, R. AND WU, T. Y.-T. 1991. Stability of Forced Steady Solitary Waves. *Philosophical Transactions of the Royal Society of London* 10:429–466.
- CANUTO, C., HUSSAINI, M. Y., QUARTERONI, A., AND ZANG, T. A. 1988. Spectral method in Fluid Dynamics. Springer-Verlag.
- CANUTO, C., HUSSAINI, M. Y., QUARTERONI, A., AND ZANG, T. A. 2006. Spectral Methods, Fundamentals in Single Domains. Springer-Verlag, Berlin.
- CARLITZ, L. 1961. The product of two Ultraspherical Polynomials. *Proceedings of the Glasgow Mathematics Association* 5.
- CEBECI, T. AND KELLER, H. B. 1971. Shooting and Parallel Shooting methods for solving the Falkner-Skan Boundary-Layer equation. *Journal of Computational Physics* 7:289–300.
- CHOI, J. W., LIN, T., SUN, S., AND WHANG, S. 2010. Supercritical Surface Waves generated by Negative or Oscillatory Forcing. *Discrete and Continuous Dynamical Systems - Series B* 14:1313–1335.
- CHOI, J. W., SUN, S. M., AND WHANG, S. I. 2008. Supercritical Surface Gravity Waves Generated by a Positive Forcing. *European Journal of Mechanics - B/Fluids* 27:750–770.
- CHOI, W. AND CAMASSA, R. 1999. Fully Nonlinear Internal Waves in a Two-Fluid System. *Journal of Fluid Mechanics* 96.
- CHRISTOV, C. I. 1982. A complete Orthonormal System of functions in  $l^2(-\infty, \infty)$  space. *SIAM Journal of Applied Mathematics* 42:1337–1344.
- CLARKE, S. R. AND JOHNSON, E. R. 1997a. Topographically-Forced Long Waves on a Sheared Coastal Current. I: The Weakly Nonlinear Response. *Journal of Fluid Mechanics* 343.
- CLARKE, S. R. AND JOHNSON, E. R. 1997b. Topographically-Forced Long Waves on a Sheared Coastal Current. II: Finite-Amplitude Waves. *Journal of Fluid Mechanics* 343.
- CLARKE, S. R. AND JOHNSON, E. R. 1999. Finite-Amplitude Topographic Rossby Waves in a channel. *Physics of Fluids* 11.
- CLENSHAW, C. W. 1972. A note on the summation of Chebyshev Series. *Mathematics of Computation* 9:118–120.
- CONCUS, P., GOLUB, G. H., AND O’LEARY, D. P. 1976. A generalized conjugate gradient method for the numerical solution of elliptic partial differential equations, in: Sparse Matrix Computations. New York: Academic Press.

- COX, S. M. AND MATTHEWS, P. C. 2002. Exponential Time Differencing for Stiff Systems. *Journal of Computational Physics* 176:430–455.
- CRAIK, A. D. D. 2004. The Origin of Water Wave Theory. *Annual Review of Fluid Mechanics* 36:1–28.
- CRAIK, A. D. D. 2005. George Gabriel Stokes on Water Wave Theory. *Annual Review of Fluid Mechanics* 37:23–42.
- CULLEN, A. C. AND CLARKE, S. R. 2017. A fast, spectrally accurate solver for the falkner–skan equation. *ANZIAM Journal* 58:57–68.
- CUSHMAN-ROISIN, B. AND BECKERS, J.-M. 2011. Introduction to Geophysical Fluid Dynamics: Physical and Numerical Aspects. Academic Press, 2 edition.
- DAHLQUIST, G. AND BJÖRCK, . 2008. Numerical Methods in Scientific Computing, volume 1. SIAM: Philadelphia.
- DAUBECHIES, I. 1992. Ten Lectures on Wavelets. SIAM, Philadelphia.
- DAVIE, A. M. AND STOTHERS, A. J. 2013. Improved bound for complexity of Matrix Multiplication. *Proceedings of the Royal Society of Edinburgh* 143A:351–370.
- DAVIS, T. A. 2006. Direct Methods for Sparse Linear Systems. SIAM, Philadelphia.
- DAVIS, T. A. 2011. Algorithm 915, Suisespqr: Multifrontal Multithreaded Rank-Revealing Sparse QR Factorization. *ACM Transactions on Mathematical Software* 38:8:1–8:22.
- DHAI, Y.-H. AND YUAN, Y. 1999. A Nonlinear Conjugate Gradient Method with a Strong Global Convergence Property. *SIAM Journal of Optimization* 1:177–182.
- DIAS, F. AND VANDEN-BROECK, J.-M. 1989. Open Channel Flows with Submerged Obstructions. *Journal of Fluid Mechanics* .
- DIAS, F. AND VANDEN-BROECK, J.-M. 2002. Generalised Critical Free Surface Flows. *Journal of Engineering Mathematics* 42.
- DIAS, F. AND VANDEN-BROECK, J.-M. 2004. Trapped Waves Between Submerged Obstacles. *Journal of Fluid Mechanics* .
- DICKSON, K., KELLEY, C., IPSEN, I., AND KEVREKIDIS, I. 2007. Condition Estimates for Psuedo-Arclength Continuation. *SIAM Journal of Numerical Analysis* 45, 1:263–276.
- DODD, R. AND FORDY, A. 1983. The Prolongation Structures of Quasi-Polynomial Flows. *Proceedings of the Royal Society A* 385.
- DONAHUE, A. S. AND SHEN, S. S.-P. 2010. Stability of Hydraulic Fall and Sub-Critical Cnoidal Waves in Water Flows over a Bump. *Journal of Engineering Mathematics* 68:197–205.
- DRISCOLL, T., BORNEMANN, F., AND TREFETHEN, L. 2008. The Chebop System for Automatic Solution of Differential Equations. *Numerical Mathematics* 48:701–723.
- DUFF, I., ERISMAN, A., AND REID, J. 1986. Direct Methods for Sparse Matrices. Clarendon Press, Oxford.
- EE, B. K. AND CLARKE, S. R. 2007. Weakly Dispersive Hydraulic flows in a Contraction: Parametric Solution and Linear Stability. *Physics of Fluids* 19:1–11.

- EE, B. K. AND CLARKE, S. R. 2008. Weakly Dispersive Hydraulic Flows in a Contraction - Nonlinear Stability Analysis. *Wave Motion* 45:927–939.
- ENGEL, C. B., LANE, T. P., REEDER, M. J., AND REZNY, M. 2013. The Meteorology of Black Saturday. *Quarterly Journal of the Royal Meteorological Society* .
- FAZIO, R. 2013. Blasius problem and Falkner-Skan model: Topfer’s algorithm and its extension. *Computers and Fluid* 75:202–209.
- FEJÉR, L. 1933. On the infinite sequences arising in the theorie of Harmonic Analysis of Interpolation, and of Mechanical Quadratures. *Bulletin of the American Mathematical Society* 39.
- FISHER, R. A. 1958. The Genetical Theory of Natural Selection. Oxford University Press.
- FORBES, L. K. AND SCHWARTZ, L. W. 1982. Free-Surface Flow over a Semicircular Obstruction. *Journal of Fluid Mechanics* 114:299–314.
- GARDNER, C. S., GREENE, J. M., KRUSKAL, M. D., AND MIURA, R. M. 1967. Method for solving the Korteweg-de Vries Equation. *Physical Review Letters* 19:1095–1097.
- GARDNER, C. S., GREENE, J. M., KRUSKAL, M. D., AND MIURA, R. M. 1974. Korteweg-de Vries Equation and Generalizations. VI: Methods for Exact Solution. *Communications on Pure and Applied Mathematics* 27:97–133.
- GARDNER, C. S. AND MORIKAWA, G. K. 1960. Similarity in the Asymptotic Behaviour of Collision Free Hydromagnetic Waves and Water Waves. Technical Report NYU-9082, Courant Institute of Mathematical Sciences, New York University, New York, 1-30.
- GILBERT, J. AND PEIERLS, T. 1988. Sparse Partial Pivoting in Time Proportional to Arithmetic Operations. *SIAM Journal of Scientific and Statistical Computing* 9(5):862–874.
- GOLUB, G. H. AND VAN LOAN, C. F. 1996. Matrix Computations. Johns Hopkins University Press, Baltimore, MD, USA.
- GOVAERTS, W. 2000. Numerical Methods for Bifurcation of Dynamic Equilibria. SIAM, Philadelphia.
- GRIMSHAW, R. H. 1981. Evolution Equations for Long, Nonlinear Internal Waves in Stratified Shear Flows. *Studies in Applied Mathematics* 65:159–188.
- GRIMSHAW, R. H. 1997. Advances in Coastal and Oceanographic Engineering, chapter Internal Solitary Waves. World Scientific Press, Singapore.
- GRIMSHAW, R. H. AND SMYTH, N. 1986. Resonant Flow of a Stratified Fluid Over Topography. *Journal of Fluid Mechanics* 169:429–464.
- GRIMSHAW, R. H. AND YI, Z. 1991. Resonant Generation of Finite-Amplitude Waves by the flow of a Uniformly Stratified Fluid over Topography. *Journal of Fluid Mechanics* 229:603–628.
- GRIMSHAW, R. H. AND YI, Z. 1993. Resonant Generation of Finite-Amplitude Waves by the Uniform Flow of a Uniformly Rotating Fluid Past an Obstacle. *Mathematika* 40:30–50.

- GRUE, J., JENSEN, A., RUSAS, P.-O., AND KRISTIAN SVEEN, J. 1999. Properties of Large-Amplitude Internal Waves. *Journal of Fluid Mechanics* 380:257–278.
- GUNJI, T., KIM, S., KOJIMA, M., TAKEDA, A., FUJISAWA, K., AND MIZUTANI, T. 2003. PHoM - a Polyhedral Homotopy Continuation Method for Polynomial Systems. *Research Reports on Mathematical and Computing Sciences Series B: Operations Research* .
- GUO, B.-Y., SHEN, J., AND WANG, Z.-Q. 2002. Chebyshev Rational Spectral and Pseudospectral Methods on a Semi-Infinite Interval. *International Journal of Numerical Methods* 53:65–84.
- HAIRER, E. AND WANNER, G. 1996. Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems, volume 14 of *Springer Series in Computational Mathematics*. Springer, Berlin, second revised edition edition.
- HANAZAKI, H. 1993. On the Nonlinear Internal Waves Excited in the flow of a Boussinesq Fluid. *Physics of Fluids A* 5:1201–1205.
- HAYAT, T. AND SAJID, M. 2007. Homotopy Analysis of MHD Boundary Layer Flow of an Upper-Convected Maxwell fluid. *International Journal of Engineering Science* 45:393–401.
- HE, J.-H. 2003. Homotopy Perturbation Method: A New Nonlinear Analytic Technique. *Journal of Applied Mathematics and Computation* 135:73–79.
- HELFRICH, K. R. 2007. Decay and Return of Internal Solitary Waves with Rotation. *Physics of Fluids* 19.
- HELFRICH, K. R. AND MELVILLE, W. K. 1986. On Long Nonlinear Internal Waves over Slope-Shelf Topography. *Journal of Fluid Mechanics* 167:285–308.
- HEMKER, P. W. 1990. A Nonlinear Multigrid Method for One-Dimensional Semiconductor Device Simulation. *Journal of Computational and Applied Mathematics* 30:117–126.
- HOLLAND, J. H. 1995. Hidden Order: How Adaptation Builds Complexity. Addison-Wesley.
- HORN, R. A. AND JOHNSON, C. R. 1985. Matrix Analysis. Cambridge University Press.
- JIANG, S. AND LIAN-GUI, Y. 2009. Modified KdV equation for Solitary Rossby Waves with  $\beta$  effect in Barotropic Fluids. *Chinese Physics B* 18:2873–2877.
- JIONG, C. AND LIU, S.-K. 1998. The Solitary Waves of the Barotropic Quasi-Geostrophic Model with the Large-scale Orography. *Advances in Atmospheric Sciences* 15:404–411.
- JUCKES, M. N. AND MCINTYRE, M. E. 1987. A High Resolution, One-Layer Model of Breaking Planetary Waves in the Stratosphere. *Nature* 328:590–596.
- KADOMTSEV, B. B. AND PETVIASHVILI, V. I. 1970. On the Stability of Solitary Waves in Weakly Dispersive Media. *Soviet Physics Doklady* 15.
- KAKUTANI, T. AND MATSUUCHI, K. 1975. Effect of Viscosity on Long Gravity Waves. *Journal of the Physical Society of Japan* 39:237–246.
- KAKUTANI, T. AND YAMASAKI, N. 1978. Solitary Waves on a Two-Layer Fluid. *Journal of the Physical Society of Japan* 45:674–679.

- KAMCHATNOV, A. M., KUO, Y.-H., LIN, T.-C., HORNG, T.-L., GOU, S.-C., CLIFT, R., EL, G. A., AND GRIMSHAW, R. H. 2012. Undular Bore Theory for the Gardner Equation. *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics* 86.
- KAWAHARA, G., UHLMANN, M., AND VAN VEEN, L. 2012. The Significance of Simple Invariant Solutions in Turbulent Flows. *Annual Review of Fluid Mechanics* 44:203–225.
- KEELER, J. S., BINDER, B. J., AND BLYTH, M. G. 2017. On the Critical Free-Surface Flow over Localised Topography. *Journal of Fluid Mechanics* 832:73–96.
- KELLER, H. B. 1977. Applications of Bifurcation Theory, chapter Numerical Solution of Bifurcation and Nonlinear Eigenvalue Problems. Academic Press, New York.
- KILLWORTH, P. D. 1992. On Hydraulic Control in a Stratified Fluid. *Journal of Fluid Mechanics* 237:605–626.
- KOOP, C. G. AND BUTLER, G. 1981. An Investigation of Internal Solitary Waves in a Two-Fluid system. *Journal of Fluid Mechanics* 112:225–251.
- KORTEWEG, D. J. AND DE VRIES, G. 1895. On the change of form of Long Waves advancing in a Rectangular Canal, and on a new type of long Solitary Waves. *Philosophy Magazine* 39:422–443.
- KUZNETSOV, Y. 1998. Elements of Applied Bifurcation Theory. Springer-Verlag, New York.
- LAMB, K. AND WAN, B. 1998. Conjugate Flows and Flat Solitary Waves for a Cocontinuous Stratified Fluid. *Physics of Fluids* 10.
- LAU, W. AND KIM, K.-M. 2011. The 2010 Pakistan Flood and Russian Heat Wave: Teleconnection of Hydrometeorological Extremes. *Journal of Hydrometeorology* 13:392–402.
- LE-GALL, F. 2014. Powers of Tensors and Fast Matrix Multiplication. *Proceedings of the 39th International Symposium on Symbolic and Algebraic Computation (ISSAC 2014)*, *arXiv: 1401.7714* .
- LEJENAS, H. AND OKLAND, H. 1983. Characteristics of Northern Hemisphere Blocking as determined from a long time series of observational data. *Tellus* 35A:350–362.
- LIAO, S. J. 1992. The proposed Homotopy Analysis Technique for the solution of Nonlinear problems. PhD thesis, Shanghai Jiao Tong University.
- LIAO, S. J. 2013. Advances in the Homotopy Analysis Method. World Scientific Publishing, New York.
- LIAO, S. J. AND CAMPO, A. 2002. Analytic solutions of the Temperature Distribution in Blasius Viscous Flow Problems. *Journal of Fluid Mechanics* 453:411–425.
- LIEBMANN, B. AND HENDON, H. 1990. Synoptic-Scale Disturbances near the Equator. *J. Atmos. Sci* 47:1463–1479.
- LONG, R. R. 1953a. A Laboratory Model Resembling the "Bishop-Wave" Phenomenon. *Bulletin of the American Meteorological Society* 34:205–211.
- LONG, R. R. 1953b. Some aspects of the flow of Stratified Fluids. I: A Theoretical Investigation. *Tellus* 5:42–57.



- LONG, R. R. 1954. Some aspects of the flow of Stratified Fluids. II: Experiments with a Two-Fluid System. *Tellus* 6:97–115.
- LONG, R. R. 1955. Some aspects of the flow of Stratified Fluids. III: Continuous Density Gradients. *Tellus* 7:342–357.
- LONG, R. R. 1959. The Motion of Fluids with Density Stratification. *Journal of Geophysical Research* 64:2151–2163.
- LONG, R. R. 1962. Velocity Concentrations in Stratified Fluids. *Journal of Hydrodynamics Division of the American Society of Civil Engineering* 88:9–26.
- LONG, R. R. 1965. On the Boussinesq Approximation and its role in the Theory of Internal Waves. *Tellus* 17:46–52.
- LONG, R. R. 1970. A Theory of Turbulence in Stratified Fluids. *Journal of Fluid Mechanics* 42:349–365.
- LORD RAYLEIGH 1876. On Waves. *Philosophy Magazine* 1:257–271.
- LOTTE, J. AND FISCHER, P. 2005. Hybrid Multigrid/Schwarz Algorithm for the Spectral Element Method. *Journal of Scientific Computing* .
- LYAPUNOV, A. M. 1892. The General Problem of the Stability of Motion. Taylor & Francis, London.
- MAGYARI, E. 2008. Exact Ananalytic Solution of a Nonlinear Reaction-Diffusion Model in Porous Catalysts. *Chemical Engineering Journal* .
- MALGUZZI, P. AND MALANOTTE-RIZZOLI, P. 1984. Nonlinear Stationary Rossby Waves on Nonuniform Zonal Winds and Atmospheric Blocking. Part I: Analytical Theory. *Journal of Atmospheric Science* 41:2620–2628.
- MASON, J. AND HANDSCOMB, D. 2002. Chebyshev Polynomials. CRC Press.
- MCINTYRE, M. E. AND PALMER, T. N. 1983a. A Note on the General Concept of Wave Breaking for Rossby and Gravity Waves. *Pure Appl. Geophys.* 123:964–975.
- MCINTYRE, M. E. AND PALMER, T. N. 1983b. Breaking Planetary Waves in the Stratosphere. *Nature* 305:593–600.
- MELVILLE, W. K. AND HELFRICH, K. R. 1987. Transcritical Two-Layer Flow Over Topography. *Journal of Fluid Mechanics* 178:31–52.
- MILES, J. W. 1976. Korteweg-de Vries Equation Modified by Viscosity. *Physics of Fluids* .
- MILES, J. W. 1979. On Internal Solitary Waves. *Tellus* 31:456–462.
- MILES, J. W. 1986. Stationary, Transcritical Channel Flow. *Journal of Fluid Mechanics* 162:489–499.
- MINZONI, A. A. AND SMYTH, N. F. 1996. Evolution of Lump Solutions for the KP Equation. *Wave Motion* .
- MISHKOV, R. L. 2000. Generalization of the Formula of Faa di Bruno for a Composite Function with a Vector Argument. *International Journal of Mathematics and Mathematical Science* 24:481–491.

- MIURA, R. M. 1968. Korteweg-de Vries Equation and Generalizations. I: A Remarkable Explicit Nonlinear Transformation. *Journal of Mathematical Physics* .
- MIURA, R. M. 1976. The Korteweg-de Vries Equation: A Survey Of Results. *SIAM Review* .
- MIURA, R. M., GARDNER, C. S., AND KRUSKAL, M. D. 1968. Korteweg-de Vries Equation and Generalizations. II: Existence of Conservation Laws and Constants of Motion. *Journal of Mathematical Physics* .
- MOTSA, S., SHATEYI, S., MAREWO, G., AND SIBANDA, P. 2012. An improved Spectral Homotopy Analysis Method for MHD Flow in a Semi-Porous Channel. *Numerical Algorithms* 60:463–481.
- MOTSA, S., SIBANDA, P., AND SHATEYI, S. 2010. A new Spectral-Homotopy Analysis Method for Solving a Nonlinear Second Order BVP. *Communications in Nonlinear Science and Numerical Simulation* 15:2293–2302.
- MOTSA, S. S. 2014. On the Optimal Auxiliary Linear Operator for the Spectral Homotopy Analysis Method Solution of Nonlinear Ordinary Differential Equations. *Mathematical Problems in Engineering* 2014.
- NAKAMURA, M. AND PLUMB, R. 1994. The Effects of Flow Asymmetry on the Direction of Rossby Wave Breaking. *Journal of the Atmospheric Sciences* 51:2031–2045.
- NDARANA, T. AND WAUGH, D. 2010. The link between Cut-Off Lows and Rossby wave breaking in the Southern Hemisphere. *Quarterly Journal of the Royal Meteorological Society* 136:869–885.
- NDARANA, T., WAUGH, D. W., POLVANI, L. M., CORREA, G. J. P., AND GERBER, E. P. 2012. Antarctic Ozone Depletion and Trends In Tropopause Rossby Wave Breaking. *Atmospheric Science Letters* 13:164–168.
- NIK, H., EFFATI, S., MOTSA, S. S., AND SHIRAZIAN, M. 2013. Spectral Homotopy Analysis Method and its Convergence for solving a class of Nonlinear Optimal Control Problems. *Numerical Algorithms* .
- OLVER, S. AND TOWNSEND, A. 2013. A fast and well-conditioned Spectral Method. *SIAM Review* 55:462–489.
- ORSZAG, S. A. 1969. Numerical Methods for the Simulation of Turbulence. *Physical Fluids Supplement II* 12:250–257.
- ORTEGA, J. M. AND RHEINBOLT, W. C. 1970. Iterative Solutions of Nonlinear Equations in Several Variables. Computer Science and Applied Mathematics, New York: Academic Press.
- OSYCHNY, V. AND CORNILLON, P. 2004. Properties of Rossby Waves in the North Atlantic Estimated from Satellite Data. *Journal of Physical Oceanography* 34:61–76.
- PALTSEV, B. V. 1967. Small-Parameter Method in the Boundary Value Problem for an Oseen system. *Mathematical Physics* 7:1144–1166.
- PARKER, T. J. 2012. The Dynamics of Blocking Anticyclones and the connection to Heatwaves in Southern Australia and Rainfall in Northeastern Australia. PhD thesis,

Monash University.

- PARKER, T. J., BERRY, G. J., AND REEDER, M. J. 2014. The structure and evolution of Heat Waves in Southeastern Australia. *Journal of Climate* 27:5768–5785.
- PAUMOND, L. 2003. A Rigorous Link Between KP and a Benney-Luke Equations. *Differential Integral Equations* 16:1039–1064.
- PEDLOSKY, J. 1979. *Geophysical Fluid Dynamics*. Springer-Verlag, New York.
- PELLY, J. L. AND HOSKINS, B. J. 2003. A New Perspective on Blocking. *Journal of the Atmospheric Sciences* 60:743–755.
- PETOUKHOV, V., RAHMSTORF, S., PETRI, S., AND SCHELLNHUBER, H. J. 2013. Quasiresonant Amplification of Planetary Waves and recent Northern Hemisphere Weather Extremes. *Proceedings of the National Academy of Sciences of the United States of America* .
- PEYRET, R. 2002. *Spectral Methods for Incompressible Viscous Flow*. Springer.
- PLUMB, R. A., WAUGH, D. W., ATKINSON, R. J., NEWMAN, P. A., LAIT, L. R., SCHOEBERL, M. R., BROWELL, E. V., SIMMONS, A. J., AND LIOWENSTEIN, M. 1994. Intrusions into the Lower Stratospheric Arctic Vortex during the Winter of 1991/1992. *Journal of Geophysical Research* 99:1071–1088.
- POINCARÉ, H. 1881. Sur les courbes définies par une équation différentielle. Ouvres, I. Gauthier-Villars, Paris.
- QUINTANA-ORTI, E. S. AND VAN DE GEIJN, R. A. V. 2008. Updating an LU Factorization with Pivoting. *ACM Transactions on Mathematical Software* .
- RASHIDI, M. M., ERFANI, E., AND ROSTAMI, B. 2014. Optimal Homotopy Analysis Method for solving Viscous Flow through expanding or contracting gaps with permeability. *Transactions on IoT and Cloud Computing* 2:85–111.
- REEDER, M. J. AND SMITH, R. K. 1987. A Study of Frontal Dynamics with Application to the Australia Summer Cool Change. *Journal of Atmospheric Science* .
- REEDER, M. J., SPENGLER, T., AND MUSGRAVE, T. 2015. Rossby Waves, Extreme Fronts and Wildfires in Southeastern Australia. *Geophysical Research Letters* .
- REX, D. 1950a. Blocking Action in the Middle Troposphere and its Effect upon Regional Action I. An Aerological Study of Blocking Action. *Tellus* 2.
- REX, D. 1950b. Blocking Action in the Middle Troposphere and its Effect upon Regional Climate II. The Climatology of Blocking Action. *Tellus* 2.
- ROTHMAN, E. E. 1991. Reducing Round-Off Error in Chebyshev Pseudospectral Computations. Technical report.
- ROTTMAN, J. W., BROUTMAN, D., AND GRIMSHAW, R. 1996. Numerical Simulation of the flow of a Uniformly Stratified, Inviscid Boussinesq Fluid over Long Topography in a Channel of Finite Depth. *Journal of Fluid Mechanics* .
- RUSAS, P. AND GRUE, J. 2002. Solitary Waves and Conjugate Flows in a Three-Layer Fluid. *European Journal of Mechanics - B/Fluids* 21:185–206.

- RUSSELL, J. S. 1844. Report on Waves. *In* Fourteenth meeting of the British Association for the Advancement of Science.
- RYOO, J.-M., KASPI, Y., WAUGH, D. W., KILADIS, G. N., WALISER, D. E., FETZER, E. J., AND KIM, J. 2013. Impact of Rossby Wave Breaking on U.S. West Coast Winter Precipitation during ENSO Events. *Journal of Climate* 26:6380–6382.
- SCINOCCHA, J. F. AND HAYNES, P. H. 1998. Dynamical Forcing of Stratospheric Planetary Waves by Tropospheric Baroclinic Eddies. *Journal of the Atmospheric Sciences* 55.
- SHEN, J. AND WANG, L.-L. 2009. Some Recent Advances on Spectral Methods for Unbounded Domains. *Communications in Computational Physics* 5:195–241.
- SHEN, S. S.-P. 1991. Locally Forced Critical Surface Waves in Channels of Arbitrary Cross Section. *Journal of Applied Mathematics and Physics (ZAMP)* 42:122–138.
- SHEN, S. S.-P. 1993. A Course on Nonlinear Waves. Kluwer Academic Publishers.
- SHEN, S. S.-P. 1995. On the Accuracy of the Stationary Forced Korteweg-de Vries Equation as a Model Equation for Flows Over a Bump. *Quarterly of Applied Mathematics* 53:707–719.
- SHEN, S. S.-P., SHEN, B., ONG, C. T., AND XU, Z. T. 2002. Collision of Uniform Soliton Trains in Asymmetric Systems. *Discrete and Continuous Dynamical Systems - Series B* 9:131–138.
- SHEN, S. S.-P., SHEN, M. C., AND SUN, S. M. 1989. A Model Equation for Steady Surface Waves Over a Bump. *Journal of Engineering Mathematics* 23:315–323.
- SLOANE, N. J. A. 2018. The On-Line Encyclopedia of Integer Sequences. *published at* <http://oeis.org>.
- STASTNA, M. AND PELTIER, W. 2005. On the Resonant Generation of Large-Amplitude Internal Solitary and Solitary-like Waves. *Journal of Fluid Mechanics* 543:267–292.
- STOKER, J. J. 1957. Water Waves: The Mathematical Theory with Applications. Interscience Publishers.
- SUN, Y. P., LIU, S. B., AND KEITH, S. 2004. Approximate Solution for the Nonlinear model of Diffusion and Reaction in Porous Catalysts by the Decomposition Method. *Chemical Engineering Journal* 102:1–10.
- TAM, A. T., YU, Z., KELSO, R. M., AND BINDER, B. 2015. Predicting Channel Bed Topography in Hydraulic Falls. *Physics of Fluids* 27:112–126.
- TREFETHEN, L. N. 2013. Approximation Theory and Approximation Practice. SIAM.
- TREFETHEN, L. N. AND TRUMMER, M. R. 1989. An Instability Phenomenon in Spectral Methods. *SIAM Journal of Numerical Analysis*.
- VANDEN-BROECK, J.-M. 1987. Free Surface Flow over an Obstruction in a Channel. *Physics of Fluids* 30:2315–2317.
- VASTANO, A. C. J. AND MUNGALL, J. C. H. 1976. Theory of Waves and Surges Which Propagate the Length of a Horizontal Rectangular Canal. Technical report, College of Geosciens, Texas A&M University.

- WADE, S. 2015. Very Steep Solitary Waves in Two-Dimensional Free Surface Flow. PhD thesis, The University of Adelaide.
- WANG, Y.-H. AND MAGNUSDOTTIR, G. 2011. Tropospheric Rossby Wave Breaking and the SAM. *Journal of Climate* 24:2134–2146.
- WHITHAM, G. B. 1974. Linear and Nonlinear Waves. J. Wiley & Sons.
- WOOLLINGS, T. AND HOSKINS, B. 2008. A New Rossby Wave-Breaking Interpretation of the North Atlantic Oscillation. *Journal of the Atmospheric Sciences* 65:609–626.
- WU, T. Y.-T. 1987. Generation of Upstream Advancing Solitons by Moving Disturbances. *Journal of Fluid Mechanics* 184:75–99.
- XU, D. L., LIN, Z. L., LIAO, S. J., AND STIASSNIE, M. 2012. On the Steady-State Fully Resonant Progressive Waves in Water of Finite Depth. *Journal of Fluid Mechanics* 207:1–40.
- XU, H., LIN, Z., LIAO, S., WU, J., AND MAJDALANI, J. 2010. Homotopy Based Solutions of the Navier-Stokes equations for a Porous Channel with Orthogonally Moving Walls. *Physics of Fluids* 22:1–18.
- XUN, J., CHEN, J., AND LIU, S.-K. 2000. A Barotropic Quasi-Geostrophic Model with Large-Scale Topography. *Journal of Tropical Meteorology* 6:66–74.
- YANG, C. AND LIAO, S. J. 2006. On the Explicit Purely Analytic Solution of Von Karman Swirling Viscous Flow. *Communication in Nonlinear Science: Numerical Simulations* 11:83–93.
- YANG, J. AND LAKOBA, T. I. 2007. Accelerated Imaginary-time Evolution Methods for the Computation of Solitary Waves. *arXiv ref:0711.3434v1* .
- YOSHIDA, H. 1990. Construction of Higher Order Symplectic Integrators. *Physics Letters A* 150:262–268.
- ZABUSKY, N. J. AND KRUSKAL, M. D. 1965. Interaction of "Solitons" in a Collisionless Plasma and the Recurrence of Initial States. *Physical Review Letters* 15:240–243.
- ZHANG, B.-Y. 1992. Unique Continuation for the Korteweg-de Vries Equation. *SIAM Journal of Maths* 23:55–71.