

1 **Numerical Methods for Wave Phenomena**

2 by

3 **Fortino Garcia**

4 B.A., Rice University, 2015

5 M.S., University of Colorado Boulder, 2018

6 A thesis submitted to the

7 Faculty of the Graduate School of the

8 University of Colorado in partial fulfillment

9 of the requirements for the degree of

10 Doctor of Philosophy

11 Department of Applied Mathematics

12 2021

13 Committee Members:

14 Daniel Appelö, Chair

15 Prof. Adrianna Gillman

16 Prof. Stephen Becker

17 Prof. Olof Runborg

18 Dr. Anders Petersson

¹⁹ Garcia, Fortino (Ph.D., Applied Mathematics)

²⁰ Numerical Methods for Wave Phenomena

²¹ Thesis directed by Prof. Daniel Appelö

²² This dissertation describes numerical methods for wave phenomena and is divided into two
²³ main sections. The first concerns a new time-domain approach to solving the Helmholtz equation.
²⁴ The second concerns numerical methods for the optimal control of closed quantum systems.

²⁵ The efficient solution of the Helmholtz equation is an active area of research. Traditionally,
²⁶ many methods in the literature take the approach of solving the Helmholtz equation “directly”.
²⁷ By directly we mean solving the Helmholtz equation in the frequency domain, whether by a direct
²⁸ discretization of the PDE via finite differences/elements or by integral equation methods. An
²⁹ alternative approach is to instead solve the Helmholtz equation by seeking time-harmonic solutions
³⁰ in the time-domain. In this thesis we present the WaveHoltz iteration, which is a fixed-point
³¹ iteration for solving the Helmholtz equation by instead solving a sequence of wave equations. We
³² demonstrate that WaveHoltz is amenable to acceleration via Krylov subspace methods. Moreover
³³ we show that WaveHoltz is simple to implement, inherits the memory-leanness and scalability of
³⁴ the underlying wave equation discretization, and that it is possible to remove time-discretization
³⁵ errors from the WaveHoltz solution.

³⁶ The second part of the thesis introduces tools for devising optimal controls to realize logic
³⁷ gates in closed quantum systems. We motivate a novel approximation of control functions via
³⁸ B-spline wavelets with carrier waves that are specifically constructed to trigger the transition fre-
³⁹ quencies of a quantum system. Using the symplectic and time-reversible Störmer-Verlet scheme, we
⁴⁰ take a “discretize-then-optimize” approach to determine a corresponding adjoint partitioned Runge
⁴¹ Kutta scheme. This allows the computation of **exact** discrete gradients for the quantum optimal
⁴² control problem. Finally, we outline a submitted solution to the IBM SWAP Gate Challenge using
⁴³ these methods.

44

Dedication

45

To my mom and my sisters for their unconditional love and support.

46

Acknowledgements

47 First and foremost, I would like to thank my advisor Daniel Appelö for his support and
48 encouragement. I hope to be just as patient and confident in every student I work with in the
49 future as you have been with me. Thank you to both Olof Runborg and Anders Petersson for
50 hosting me at KTH and Lawrence Livermore National Lab, respectively, as well as for continued
51 mentorship and guidance from which I have learned an immense amount. I would also like to
52 thank Adrianna Gillman, without whom I would not have heard of the applied math department
53 at Boulder or the Helmholtz equation.

54 I don't believe I would be successful without the support of the many wonderful friends I have
55 made in the applied math department and in Boulder. Thank you to Caleb, Allen, Alec, Minah,
56 Mingyu, David, Mike, Ryan, Liam, Lyndsey, Lauren, and Roxie. A very special thanks to Caleb,
57 Sara, and Lyndsey for the late game nights that kept me sane during the worst of the pandemic. I
58 would also like to thank all of the graduate students in APPM for letting me be a part of such a
59 kind and supportive environment during this time of my life.

60 To my friends from my time in Texas, including Taylor, Matthew, Neel, Nick, Luis, Sam,
61 Eric, Michael, and all of the members of Steve Cox's Beard, thank you for your friendship all of
62 these years.

63 Finally, I would like to thank the NSF for the support provided by NSF Grant DMS-1913076,
64 as well as to STINT initiation grant IB2019-8154.

65

Contents

66 Chapter

67	1	Introduction	1
68	1.1	Outline of chapters	2
69	2	WaveHoltz: Iterative Solution of the Helmholtz Equation via the Wave Equation	4
70	2.1	WaveHoltz: A New Method for Designing Scalable Parallel Helmholtz Solvers	9
71	2.1.1	Iteration for the Energy Conserving Case	11
72	2.1.2	Analysis of the Discrete Iteration	18
73	2.1.3	Tunable Filters	22
74	2.1.4	Multiple Frequencies in One Solve	23
75	2.1.5	WaveHoltz Iteration for Impedance Boundary Conditions	23
76	2.2	Wave Equation Solvers	24
77	2.2.1	The Energy Based Discontinuous Galerkin Method	24
78	2.2.2	Finite Difference Discretizations	26
79	2.2.3	Time Discretization	27
80	2.3	Numerical Examples	28
81	2.3.1	Examples in One Dimension	28
82	2.3.2	Problems in Two Dimensions	34
83	2.3.3	Problems in Three Dimensions	44
84	2.4	Summary and Future Work	46

85	3	Analysis of an Iterative Solution of the Helmholtz Equation via the Wave Equation for Impedance Boundary Conditions	48
86	3.1	The General Iteration	52
88	3.1.1	Iteration for the Energy Conserving Case for the General WaveHoltz Iteration	53
89	3.1.2	Convergence in the Non-Energy Conserving Case	61
90	3.2	Damped Wave/Helmholtz Equation	67
91	3.3	Analysis of Higher Order Time Stepping Schemes for the Discrete Iteration	69
92	3.4	Wave Equation Solvers	73
93	3.4.1	The Energy Based Discontinuous Galerkin Method	73
94	3.4.2	Symmetric Interior Penalty Discontinuous Galerkin Method	75
95	3.4.3	Finite Difference Discretizations	75
96	3.4.4	Time Discretization	76
97	3.5	Numerical Examples	76
98	3.5.1	Examples in One Dimension	76
99	3.5.2	Examples in Two Dimensions	84
100	3.6	Summary and Future Work	87
101	4	El WaveHoltz Method	89
102	4.1	Governing Equations	94
103	4.1.1	The Time Harmonic Elastic Wave Equation	94
104	4.1.2	The Elastic Wave Equation	95
105	4.2	The El WaveHoltz Iteration	96
106	4.3	Numerical Methods and Discrete Analysis	98
107	4.3.1	El WaveHoltz by Finite Differences	98
108	4.3.2	El WaveHoltz by Symmetric Interior Penalty Discontinuous Galerkin Method	101
109	4.3.3	Explicit Time-Corrected Scheme	102
110	4.3.4	Implicit Time-Corrected Scheme	103

111	4.3.5 Krylov Solution of the El WaveHoltz Iteration	106
112	4.4 Numerical Experiments	108
113	4.4.1 Accuracy of the Finite Difference Method	108
114	4.4.2 Verification of Corrected Time-Steppers	109
115	4.4.3 Accuracy of the Symmetric Interior Penalty Discontinuous Galerkin Method	110
116	4.4.4 Effects on Number of Iterations from Number of Periods and Accuracy . . .	111
117	4.4.5 Iteration Count as a Function of Frequency for Rectangles and Annular Sectors	113
118	4.4.6 Effect of Boundary Conditions in A Cube	115
119	4.4.7 Iteration Count as a Function of Wave Speed Ratio	115
120	4.4.8 Comparison of Explicit and Implicit El WaveHoltz with Direct Discretization of Elastic Helmholtz	117
121	4.4.9 Materials with Spatially Varying Properties	121
122	4.4.10 Vibrations of a Toroidal Shell	122
123	4.5 Conclusion	123
124		
125	5 Optimal Control of Closed Quantum Systems via B-Splines with Carrier Waves	125
126	5.1 Hamiltonian model	130
127	5.1.1 Rotating wave approximation	131
128	5.1.2 Resonant frequencies	132
129	5.2 Quadratic B-splines with carrier waves	134
130	5.3 Real-valued formulation	135
131	5.3.1 Time integration	136
132	5.3.2 Time step restrictions for accuracy and stability	137
133	5.4 Discretizing the objective function and its gradient	139
134	5.4.1 Discretizing the objective function	139
135	5.4.2 The discrete adjoint approach	141
136	5.5 Numerical optimization	144

137	5.5.1 A CNOT gate on a single qudit with guard levels	145
138	5.5.2 The Hessian of the objective function	149
139	5.5.3 Risk-neutral controls	151
140	5.6 Comparing Juqbox with QuTiP/pulse_optim and Grape-TF	154
141	5.6.1 Setup of simulation codes	155
142	5.6.2 Numerical results	156
143	5.7 Conclusions	160
144	6 IBM Open Science Prize – SWAP Gate Challenge	162
145	6.1 Hamiltonian model	163
146	6.2 Optimal control with Juqbox.jl and Quandary	165
147	6.2.1 Open system optimal control	167
148	6.3 Rabi pulse calibrations	168
149	6.4 Gaussian square and DRAG pulses in Qiskit	169
150	6.5 Converting Qiskit pulses to B-splines with carrier waves	170
151	6.6 Reverse model calibration using X- and Cx-gates	171
152	6.6.1 X-gates	172
153	6.6.2 Calibrating a cross-talk model using the Cx gates	174
154	6.7 Implementation of custom pulses in Qiskit	177
155	6.8 Randomized benchmarking results	178
156	6.9 Conclusion	182
157	7 Conclusion	184
158	Bibliography	187
159	Appendix	
160	.1 Proof of Lemma 2.1.1	194

161	.2	Verification of Discrete Solution	196
162	.3	Proof of Lemma 2.1.5	196
163	.4	Proof of Lemma 3.1.1	198
164	.5	Wave Equation Extension	201
165	.6	Verification of Discrete Solution	204
166	.7	Well-definedness of modified frequencies	205
167	.8	Error in discrete Helmholtz frequency	207
168	.9	Verification of Discrete Solution	208
169	.10	Time-step Restriction	208
170	.11	Motivation of Conjecture 1	209
171	.12	Composite quantum systems and essential states	212
172	.13	The Hamiltonian in a rotating frame of reference	215
173	.14	Conditions for resonance	216
174	.15	Derivation of the discrete adjoint scheme	217
175	.16	Proof of Corollary 1	222
176	.17	Computing the gradient of the discrete objective function	223

177

Tables**178 Table**

179	2.1 Maximum error for various combinations of boundary conditions and methods.	30
180	4.1 L_1, L_2 and L_∞ errors of the computed solution with corresponding estimated rates of convergence.	109
181		
182	4.2 Estimated rates of convergence for the spatial discretization.	111
183		
184	4.3 The table displays the number of iterations required and the efficiency of the longer times to reduce the relative residual by a factor 10^{-10} for the two cases (described in the text).	113
185		
186	4.4 The effect on iteration count depending on different combinations of λ and μ	117
187		
188	4.5 Comparison of the number of iterations for the three different methods.	118
189		
190	4.6 The number of right hand side evaluations (estimated) for the three different methods. The top four rows display the actual number of right hand side evaluations and the rows below indicate how many times more the HH and IWH method evaluates the right hand side. An infinity sign indicates that the computation did not converge.	119
191		
192	4.7 The table reports how many times longer a computation with the HH and IWH method takes compared to the explicit WH method.	120
193		
194	5.1 The Frobenius norm of the symmetric and asymmetric parts of the approximate Hessian, L , for the case $\alpha_{max}/2\pi = 3.0$ MHz.	150
195		

196	5.2 Gate duration, number of time steps (M) and total number of control parameters (D) in the $ 0\rangle \leftrightarrow d\rangle$ SWAP gate simulations. The number of time steps and control parameters are the same for pulse_optim and Grape-TF.	157
197		
198		
199	5.3 QuTiP/pulse_optim results for $ 0\rangle \leftrightarrow d\rangle$ SWAP gates. Note the larger infidelity and guard state population for $d = 6$	157
200		
201	5.4 Grape-TF results for $ 0\rangle \leftrightarrow d\rangle$ SWAP gates. Note the very large infidelity for $d = 6$. These simulations used two NVIDIA P-100 GPUs to accelerate TensorFlow.	157
202		
203	5.5 Juqbox results for $ 0\rangle \leftrightarrow d\rangle$ SWAP gates.	158

204

Figures

205 **Figure**

206	2.1 The filter transfer function β for $\omega = 10$	14
207	2.2 Convergence of the residual for the plain WaveHoltz iteration and its accelerated versions using LSQR, QMR, CG and GMRES. The titles of the figures indicate the boundary conditions used to the left and right, e.g. D-N means Dirichlet on the left and Neumann on the right.	29
208		
209		
210		
211	2.3 Left: Number of iterations divided by ω as a function of ω for different boundary conditions. Middle and right: Zoom in around a resonance for the Dirichlet problem when using Krylov acceleration (middle) and when using WHI (right).	31
212		
213		
214	2.4 Left: Convergence history of the near resonant frequency 4.1π for the WaveHoltz filter and a tunable filter, and that of the frequency 1.5π for reference. Middle: The error between successive WaveHoltz iterates with the usual WaveHoltz filter. Right: Convergence of the solution for the CG accelerated WaveHoltz iteration with a point forcing.	32
215		
216		
217		
218		
219	2.5 (Left) The usual WaveHoltz filter (in blue) and updated tunable filter (in red). (Right) Closeup of both the usual WaveHoltz filter and the updated tunable filter near the resonant frequency 4π	34
220		
221		
222	2.6 Typical solutions computed with the GMRES accelerated WHI at $\omega = 77.5$. The thick lines indicate Dirichlet boundary conditions.	35
223		

224	2.7 To the left: number of iterations as a function of frequency to reduce the relative 225 residual below 10^{-7} for problems with no trapped waves. Middle: the same but 226 for problems with trapped waves and for the interior problem. Both are with the 227 GMRES accelerated WHI . To the right: Residuals for the GMRES accelerated 228 WHI , the CG accelerated WHI and for GMRES solution of the directly discretized 229 Helmholtz problem.	36
230	2.8 Left: the speed of sound (squared) used in example 2.3.2.2. Red indicates a rigid 231 wall and black indicates open walls. Middle: Number of iterations as a function of 232 frequency. Right: Compute time normalized by the frequency times the number of 233 degrees of freedom.	38
234	2.9 The magnitude of the Helmholtz solution for, from left to right, $\omega = 25\pi, 50\pi$ and 235 100π	39
236	2.10 The maximum error GMRES residuals as a function of number of iterations for four 237 different mesh sizes. The rates of convergence agree with the order of the method. .	40
238	2.11 Displayed is the base 10 logarithm of the magnitude of the Helmholtz solution 239 ($\log_{10} u $) caused by a point source near the surface. The results are, from top 240 to bottom, for $\omega = 800, 400$ and 200	42
241	2.12 Zoom in of the base 10 logarithm of the magnitude of the Helmholtz solution 242 ($\log_{10} u $) caused by a point source near the surface. The results are, from left 243 to right, for $\omega = 800, 400$ and 200	42
244	2.13 Computation of three Helmholtz problems by one solve. The frequencies are $\omega =$ 245 $15, 30$ and 60 . The material model is also displayed, red is $c^2 = 1$ and dark blue is 246 $c^2 = 0.1$	43
247	2.14 To the left: number of iterations as a function of frequency to reduce the relative 248 residual below $5 \cdot 10^{-5}$ for problems with no trapped waves. Here WHI is accelerated 249 by TFQMR. To the right: the same but for the interior problem. Here WHI is 250 accelerated with either CG or TFQMR.	45

251	2.15 Displayed is the base 10 logarithm of the magnitude of the Helmholtz solution ($\log_{10} u $) caused by a point source near the surface for $\omega = 200$ (left) and $\omega = 300$ (right) at the slice $x = 0.1$	46
254	3.1 (Left, Middle) The initial conditions v_0 and v_1 for a Helmholtz frequency of $\omega = 10\pi$. (Right) The estimate of the quantity $1 - \ \mathcal{S}\ $ with increasing Helmholtz frequency ω	78
256	3.2 The norm of WaveHoltz iterates for increasing Helmholtz frequencies of $\omega = 10\pi, 40\pi$, and 70π for the adversarial example of Figure 3.1.	79
258	3.3 The estimate of the quantity $1 - \ \mathcal{S}\ $ with increasing Helmholtz frequency ω for a radially symmetric initial condition.	80
260	3.4 Convergence of the discrete WaveHoltz solution to the true solution of the discrete Helmholtz problem with fixed spatial discretization. Solid lines indicate relative errors between discrete solutions. The blue and yellow solid lines indicate relative errors between the usual discrete WaveHoltz solution and the true solution, and the red and purple solid lines indicate relative errors for the frequency corrected solution.	82
265	3.5 Number of iterations as a function of ω for different boundary conditions and damp- ing parameters. Left: $\eta = 1/2\omega$, Middle: $\eta = 1/2$, Right: $\eta = \omega/2$. In the above legends each entry is made up of a two letter string where the first letter indicates the boundary condition on the left at $x = -6$, and the second letter indicates the boundary condition on the right at $x = 6$. Here D indicates Dirichlet, N indicates Neumann, and I indicates impedance/Sommerfeld conditions.	83
271	3.6 Computational domain where the mesh in blue corresponds to a wavespeed of $c =$ 2100 , the mesh in green corresponds to a speed of $c = 1000$, and the magenta mesh with $c = 2900$. The solid black line is not physical and is meant to more easily distinguish between regions.	85
275	3.7 Number of iterations to reach a GMRES tolerance of 10^{-10} for the wedge problem in 2D with all Neumann or all impedance boundary conditions.	86

277	3.8 In the above we plot the \log_{10} of the absolute value of the real part of the Helmholtz	
278	solution with frequency $\omega = 40\pi$ for (Left) damping parameter $\eta = 20\pi$ and (Right)	
279	no damping.	87
280	4.1 (Left) Convergence of the discrete WaveHoltz solution to the true solution of the	
281	discrete Helmholtz problem. (Right) Convergence of the discrete WaveHoltz solution	
282	to the true solution of the discrete Helmholtz problem for a manufactured solution.	110
283	4.2 From top to bottom: displacement magnitude, σ_{xx} , σ_{xy} and σ_{yy} . The domain is	
284	$[0, 8] \times [0, 1]$ and the color scales are $[0, 8]$, $[-50, 50]$, $[-15, 15]$ and $[-40, 40]$ respec-	
285	tively.	112
286	4.3 The number of iterations as a function of frequency to reach convergence for (Left)	
287	a rectangle, quarter circle and half circle, and (Right) the unit cube with Dirichlet	
288	or free surface conditions.	114
289	4.4 From top left to bottom right: displacement magnitude, σ_{xy} , σ_{xx} and σ_{yy}	116
290	4.5 The \log_{10} of the magnitude of the displacements for the CG accelerated solution of	
291	WH for the inclusion problem using sixth order polynomials within each element.	
292	(Left) Solution using a grid resolution of at least one element per wavelength, and	
293	(Right) two elements per wavelength.	122
294	4.6 The solution in the toroidal shell for (Left) $\omega = 5.1234$, and (Right) $\omega = 10.2468$. The	
295	projection onto the xy-plane is the magnitude of the displacement on the outermost	
296	free surface $r = 2$. In black we display the (scaled) displaced mesh for $r = 2$	123
297	5.1 The real part of a quadratic B-spline control function, with zero carrier frequency	
298	(dashed black). The solid colored lines are the individual B-spline wavelets.	134
299	5.2 Convergence of the IPOPT iteration for the CNOT gate with the parameter con-	
300	straint $\ \alpha\ _\infty \leq \alpha_{max}$. Here, $\alpha_{max}/2\pi = 4$ MHz (left) and $\alpha_{max}/2\pi = 3$ MHz	
301	(right).	146

302	5.3	The population of the “forbidden” state $ 5\rangle$ as function of time for the four initial conditions of the CNOT gate. Here, $\alpha_{max}/2\pi = 3$ MHz.	147
303			
304	5.4	The rotating frame control functions $p(t)$ (blue) and $q(t)$ (orange) for realizing a CNOT gate with $D_1 = 10$ basis function per carrier wave and three carrier wave frequencies. Here, $\alpha_{max}/2\pi = 3$ MHz.	148
305			
306	5.5	The population of the states $ 0\rangle$ (blue), $ 1\rangle$ (orange), $ 2\rangle$ (green) and $ 3\rangle$ (purple), as function of time, for each initial condition of the CNOT gate. Here, $\alpha_{max}/2\pi = 3$ MHz.	149
307			
309	5.6	The eigenvalues of the symmetric part of the approximate Hessian, $0.5(L + L^T)$, evaluated at the optima for the parameter thresholds $\alpha_{max}/2\pi = 4$ MHz (blue triangles) and $\alpha_{max}/2\pi = 3$ MHz (orange circles). The positive eigenvalues are shown on a log-scale on the left and the small eigenvalues are shown on a linear scale on the right.	151
310			
311	5.7	Infidelity objective (\mathcal{J}_1) and guard level objective (\mathcal{J}_2) as function of ε in $H_s^u(\varepsilon)$. Here ‘NF’ and ‘RN’ correspond to the “Noise-Free” and “Risk-Neutral” cases.	153
314			
315	5.8	Control functions (without carrier waves) for the cases: “noise-free” (top), and “risk-neutral” (bottom). Here, $p_{k,n}(t)$ and $q_{k,n}(t)$ are defined in (5.27).	154
316			
317	5.9	Magnitude of the Fourier spectrum of the laboratory frame control function for the $ 0\rangle \leftrightarrow 5\rangle$ SWAP gate.	159
319			
320	6.1	Results of the Rabi experiment on qubit 5 of the Casablanca hardware.	169
321	6.2	The pulse schedule for an X gate on Casablanca for qubit 5 (left), and qubit 6 (right).	173
322	6.3	The pulse schedule for a CNOT gate on Casablanca where qubit 5 is the control qubit and qubit 6 is the target.	175
323			
324	6.4	The pulse schedule for the first interleaved RB circuit.	179
325	6.5	Simulated results. The observed ground state population as function of the length of the Clifford circuit. Here, the non-interleaved circuits are shown in blue and interleaved ones in red. The estimated error per Clifford (EPC) is 2.36%.	180
327			

328	6.6 Classification results after 1000 shots in two of the randomized circuits with one	
329	interleaved SWAP gate. Ideally the $ 00\rangle$ state should have 100% of the population.	181
330	6.7 Randomized benchmarking results on the Casablanca hardware using the Interleave-	
331	dRBFitter.	182
332	1 Values of $\alpha = \cos(\omega\Delta t)(2 + \omega^2\Delta t^2)$ for values of $\omega\Delta t$ in the interval $[0, 2]$. The red	
333	lines indicate the desired bound on α , and the black line indicates the maximum	
334	allowable value of $\omega\Delta t$ at $r \approx 1.93$	209
335	2 A plot of the discrete filter function using five time-steps $0 \leq r \leq 5/4$. On the left	
336	we plot the full range of values of r , and on the right we zoom in close to resonance,	
337	i.e. $r = 1$	210
338	3 A bound on the gap from resonance that creates problematic modes. The blue curve	
339	is the true gap, $1 - r^*$, and the dotted red curve is a proposed bound.	211

Chapter 1

Introduction

Many problems of practical interest in diverse areas such as acoustics, seismics, and quantum mechanics are governed by equations in which the solution is defined by a wave or a superposition of waves. Waves exist at an incredibly wide range of scales, and their accurate numerical treatment is of great practical importance. In acoustic scattering problems, for instance, it is desirable to obtain time-harmonic solutions to the wave equation. These solutions satisfy the **Helmholtz equation**, the efficient and scalable solution of which is an active area of research (see the review articles [47, 50, 44]). On a much smaller scale, the wave-particle duality is a foundational principle of quantum mechanics which implies that all information about a particle is contained in its wave function. This wave function, which can be interpreted as a probability distribution, evolves according to the **Schrödinger equation** and is key to understanding how we may eventually exploit the nascent power of quantum computers.

In either case, the numerical treatment of wave propagation problems requires high-order accurate and efficient numerical methods. These methods must be able to scale well in many dimensions, and potentially simulate over long distances and/or times. This thesis will develop numerical methods for two different wave problems: (1) an iterative wave equation solution method for Helmholtz problems, and (2) optimizing control functions for realizing logical gates in closed quantum systems, where the evolution of the state vector is governed by the time dependent Schrödinger equation.

The chapters of this thesis are thus separated into two broad sections. The first section,

361 **Chapters 2-4**, concern time-domain methods for Helmholtz problems. The second section, **Chap-**
362 **ters 5-6**, concern both theory and practice of optimal control methods for quantum systems. In
363 each section, there is considerable overlap in notation but this notation is generally restated and
364 redefined within each chapter so that they may be read independently. Each of **Chapters 2-5**
365 has either been published [14], submitted for publication [94], or is to be submitted for publica-
366 tion [51, 13].

367 **1.1 Outline of chapters**

368 The first section of the thesis, **Chapters 2-4**, focus on time-domain methods for solving the
369 Helmholtz equation. In **Chapter 2** we introduce the WaveHoltz iteration, which is a fixed-point
370 iteration that filters the solution of the wave equation with time-periodic forcing and boundary
371 data. We show that the WaveHoltz iteration can be recast as a positive definite linear system
372 of equations which can be solved using Krylov subspace techniques. We additionally present a
373 continuous and discrete analysis for energy-conserving problems.

374 In **Chapter 3**, we extend the analysis of the energy-conserving WaveHoltz iteration to prob-
375 lems with damping and/or impedance boundary conditions. Furthermore, we investigate higher
376 order modified equation timestepping schemes and show that the WaveHoltz solution converges to
377 the discrete Helmholtz solution to the order matching the order of the timestepping scheme. We
378 then present a method to *completely* remove time discretization error from the WaveHoltz solution.

379 In **Chapter 4**, we apply the WaveHoltz iteration to the “elastic” Helmholtz equation (also
380 known as the Navier equation) for energy-conserving problems with Dirichlet and/or free surface
381 boundary conditions. We present a discrete analysis for an implicit timestepping scheme in which
382 time discretization errors are removed.

383 In the second section of the thesis, **Chapters 5-6**, we consider optimal control methods for
384 quantum systems. In **Chapter 5**, we describe an optimal control problem for closed quantum
385 systems governed by Schrödinger’s equation. We motivate and describe the novel use of B-splines
386 with carrier waves to interpolate control functions, which allow the number of parameters to be

387 independent of the number of timesteps used in the simulation. The system is discretized with the
388 Störmer-Verlet scheme, which is a symplectic partitioned Runge-Kutta scheme. Using a “discretize-
389 then-optimize” approach, we derive a discrete timestepping scheme used to compute *exact* discrete
390 gradients at the cost of solving two Schrödinger systems. The methods described in this chapter
391 have been implemented in the Julia programming language, [1], and are made available as the open-
392 source package Juqbox (available through GitHub at <https://github.com/LLNL/Juqbox.jl>).

393 In **Chapter 6**, we apply the methods of **Chapter 5** to the IBM SWAP Gate Challenge. We
394 describe the approach taken for a submission to the SWAP Gate Challenge, and present the results
395 of building a custom gate for a real-word noisy quantum system.

396 Finally in **Chapter 7**, we summarize the results of the thesis. Additionally, future research
397 directions are discussed.

Chapter 2

399 **WaveHoltz: Iterative Solution of the Helmholtz Equation via the Wave
400 Equation**

401 The defining feature of waves are their ability to propagate over large distances without
 402 changing their shape. It is this property that allows them to carry information which underpins all
 403 communication, be it through speech or electronic transmission of data. Waves can also be used to
 404 probe the interior of the earth, the human body or engineering structures like buildings or bridges.
 405 This probing can be turned into images of the interior by the means of solving inverse problems.
 406 Harnessing the nature of waves requires high-order accurate and efficient numerical methods that
 407 are able to simulate wave propagation in three dimensions and over long distances. For cutting
 408 edge problems in scientific and engineering research such simulations must be carried out on parallel
 409 high-performance computing platforms and thus the numerical methods must scale while being easy
 410 to implement and generally applicable.

411 In this chapter we focus on approximating solutions to the scalar wave equation in the
 412 frequency domain, i.e. the Helmholtz equation

$$\nabla \cdot (c^2(x)\nabla u) + \omega^2 u = f(x). \quad (2.1)$$

413 However, to obtain such solutions we will use time domain discretizations of the wave equation.
 414 The motivation for developing high order accurate and scalable Helmholtz solvers comes from both
 415 mathematics and applications. On the mathematics side the recent results by Engquist and Zhao
 416 [42] give sharp lower bounds on the number of terms in a separated representation approximation
 417 of the Green's function of the Helmholtz equation as a function of the frequency (wavenumber).

418 These bounds limit the applicability of the state of the art sweeping preconditioners in the high
 419 frequency regime and, for example, for interior and wave guide problems. Motivation also comes
 420 from applications in seismology, optics and acoustics. For example in full waveform inversion the
 421 problems are very large and the robustness of the inversion process can be enhanced by combining
 422 frequency and time domain inversion in a multi-scale fashion to avoid getting trapped in local
 423 minima.

424 Designing efficient iterative solvers for the Helmholtz equation (2.1) is notoriously difficult
 425 and has been the subject of much research (for detailed reviews see Ernst and Gander, [47], Gander
 426 and Zhang [50], and Erlangga, [44]). The main two difficulties in solving the Helmholtz equation
 427 are the resolution requirements and the highly indefinite character of the discretized system of
 428 equations.

429 Assuming that (2.1) has been scaled so that the mean of $c(x)$ is about 1 then the typical
 430 wavelength is $\lambda = 2\pi/\omega$ and the typical wavenumber is $\omega/2\pi$. In order to numerically propagate
 431 solutions to the time dependent wave equation corresponding to (2.1) with small errors it is crucial
 432 to control the dispersion by using high order methods. The basic estimate by Kreiss and Oliger
 433 [75] shows that in order to propagate a wave over J wavelengths with a p th order finite difference
 434 method and with an error no greater than ϵ one must choose the number of points per wavelength
 435 $\text{PPW}(J, p)$ as

$$\text{PPW}(J, p) \geq C(p, \epsilon) J^{\frac{1}{p}}.$$

436 Here $C(p, \epsilon)$ depends on the tolerance ϵ but decreases with increasing order of accuracy p . Con-
 437 sequently, for a problem in d -dimensions and with fixed physical size the number of wavelengths
 438 in the domain will scale as ω^d and to maintain a fixed tolerance the total number of degrees of
 439 freedom needed, $N_p(\omega) = \mathcal{O}(\omega^{d(1+\frac{1}{p})})$, is very large for high frequencies.

440 The dependence on p and ω in $N_p(\omega)$ immediately reveals two fundamental criteria for de-
 441 signing high frequency Helmholtz solvers:

442 1. The solvers must be **parallel**, **memory lean** and they must **scale well**. In 3D the number

443 of degrees of freedom representing the solution cannot be stored on a single computer, and
 444 even on a parallel computer it is important to preserve the sparsity of the discrete version
 445 of (2.1).

- 446 2. The underlying discretizations must be **high order accurate**. At high frequencies and in
 447 3D the extra penalty due to pollution / dispersion errors becomes prohibitive.

448 Further, the linear system matrix, A , resulting from direct discretization of (2.1) is indefinite
 449 so that the robust and easy to implement preconditioned conjugate gradient (PCG) method cannot
 450 be used. Instead the method of necessity becomes the preconditioned generalized minimal residual
 451 method (GMRES). To efficiently precondition GMRES one must exploit the intrinsic properties of
 452 the wave equation. The oscillatory nature of the Helmholtz Green's function and its discrete coun-
 453 terpart A^{-1} can only be well approximated if the (unconditioned) Krylov subspace is allowed to
 454 grow quite large (with “large” scaling adversely with the frequency ω , [47]). The slow growth of the
 455 “spanning power” of the Krylov vectors is due to the underlying local connectivity of the discretiza-
 456 tion, preventing information to propagate rapidly. Efficient preconditioners must thus accelerate
 457 the propagation of information or reduce the cost of each iteration. Without preconditioners the
 458 iteration typically stagnates.

459 Perhaps the first contribution that aimed to improve the propagation of information was the
 460 Analytic Incomplete LU preconditioner (AILU) by Gander and Nataf [49]. The AILU precondi-
 461 tioner finds an LDL^T factorization from an approximation of the same pseudodifferential operators
 462 that are used to construct non-reflecting boundary conditions [39, 6, 66] and sweeps forward then
 463 backward along one of the coordinate directions in a structured grid.

464 The pioneering works on sweeping preconditioners by Engquist and Ying [40, 41] were major
 465 breakthroughs in the solution of the Helmholtz equation. Similar to the AILU, the preconditioners
 466 in [40, 41] use a LDL^T decomposition but exploit the low rank properties of off-diagonal blocks
 467 together with perfectly matched layers to obtain solvers that converge in a small number of GMRES
 468 iterations. The papers [40, 41] were the two first instances of iterative Helmholtz solvers that

469 converge in a small number of iterations that is almost independent of frequency.

470 Once it had been established that low rank approximations, combined with clever use of
 471 sweeping and perfectly matched layers (PML), could be used to find Helmholtz solvers with linear
 472 scaling then many extensions and specializations were constructed. For example, in [100] Stolk
 473 introduced a domain decomposition method with transmission conditions based on the perfectly
 474 matched layer (PML) that is able to achieve near linear scaling. Chen and Xiang, [34], and Vion and
 475 Geuzaine, [105], also considered sweeping domain decomposition method combined with PML and
 476 showed that their methods could be used as efficient preconditioners for the Helmholtz equation.
 477 The method of polarized traces by Zepeda-Núñez, Demanet and co-authors, [113, 112, 111], is a
 478 two step sweeping preconditioner that compresses the traces of the Greens function in an offline
 479 computation and utilizes incomplete Green's formulas to propagate the interface data. See also the
 480 recent review by Gander and Zhang [50] for connections between sweeping methods.

481 Alongside iterative methods there are some attractive direct and multigrid methods. Exam-
 482 ples from the class of direct methods are the Hierarchically Semi-Separable (HSS) parallel multi-
 483 frontal sparse solver by deHoop and co-authors, [107], the spectral collocation solver by Gillman,
 484 Barnett and Martinsson, [57], and the p -FEM approach of Bériot, Prinn and Gabard, [23], which
 485 utilizes an *a priori* error indicator to choose the polynomial order of each element . Notable exam-
 486 ples of multigrid methods are the Wave-ray method by Brandt and Livshits [28, 81] and the shifted
 487 Laplacian preconditioner with multigrid by Erlangga et al. [43].

488 As mentioned previously, the invention of sweeping preconditioners was a breakthrough and
 489 it is likely that they will have lasting and continuing impacts for the solution of the Helmholtz
 490 equation in various settings. There are, however, some limitations. First, in the recent paper [42],
 491 Engquist and Zhao provide precise lower bounds on how the number of terms that are needed
 492 to approximate the Helmholtz Green's function depends on the frequency. In particular, for the
 493 high frequency regime they show that for interior problems and waveguides the rank of the off-
 494 diagonal elements grows fast, rendering sweeping preconditioners less efficient. They also show that
 495 the situation is, in general, worse in 3D than in 2D. This lack of compressibility may, in cases of

496 practical importance, increase the cost of both the factorization and compression as well as the
 497 application of the compressed preconditioner. We note that this loss of compressibility at high
 498 frequency will also prevent direct methods such as [107, 57, 23] from reaching their most efficient
 499 regimes. An additional drawback of direct methods is their memory consumption for 3D problems.

500 Another potential drawback with the sweeping methods is the long setup times before the
 501 solve. Of course all of the algorithms above do not suffer from this deficiency but many of them
 502 do. This may not be problematic when considering a background velocity that does not change but
 503 this is not the case, for example, when inverting for material parameters. In this case the velocity
 504 model will change constantly, necessitating a costly factorization in each update.

505 Finally, the two criterions 1.) and 2.) above are not so easy to meet for sweeping precondi-
 506 tioners. The sweep itself is intrinsically sequential and although there have been at least partially
 507 successful attempts to parallelize the sweeping methods it is hard to say that they are easy to
 508 parallelize in a scalable way. In a similar vein most of the methods use (and some rely on) low
 509 order discretizations. Although it is possible to use higher order accurate discretizations together
 510 with sweeping preconditioners, their scarcity in the literature is noticeable.

511 Another approach that is somewhat popular in the engineering literature is to simply run
 512 the wave equation for a long time to get a Helmholtz solution, see e.g. [70]. The theoretical
 513 underpinning of this approach is the *limiting amplitude principle* which says that every solution to
 514 the wave equation with an oscillatory forcing, in the exterior of a domain with reflecting boundary
 515 conditions tends to the Helmholtz solution. However, since the limiting amplitude principle only
 516 holds for exterior problems this approach does not work for interior problems and becomes very
 517 slow for problems with trapping waves. See e.g. the articles by Ladyzhenskaya [76], Morawetz [88]
 518 and Vainberg [104].

519 An alternative approach, the so called Controllability Method (CM), was originally proposed
 520 by Bristeau et al. [29]. In the CM the solution to the Helmholtz equation is found by solving a
 521 convex constrained least-squares minimization problem where the deviation from time-periodicity
 522 is minimized in the classic wave equation energy. The basic ingredients in an iteration step in CM

523 are: a.) the solution of a forward wave and a backward wave equation over one time-period, and
 524 b.) the solution of a symmetric coercive elliptic (and wave number independent) problem.

525 In [29] and the later spectral element implementations of CM by Heikkola et al. [69, 68] only
 526 sound-soft scatterers were considered. For more general boundary conditions the minimizer of the
 527 cost functional of [29] is not unique but alternative cost functionals that does guarantee uniqueness
 528 (and thus convergence to the Helmholtz solution) were recently proposed by Grote and Tang in
 529 [63]. We also note that if the wave equation is formulated as a first order system it is possible to
 530 avoid solving the elliptic problem [58, 61].

531 In what follows we will present an alternative to the controllability method. Our method,
 532 which we call the WaveHoltz Iteration method (WHI), only requires a single forward wave equation
 533 solve and no elliptic solves but produces a positive definite (and sometimes symmetric) iteration
 534 that can be accelerated by, e.g. the conjugate gradient method or other Krylov subspace methods.
 535 As the WaveHoltz iteration is built from a time domain wave equation solver we claim and hope
 536 to demonstrate that it meets both criterion 1. and 2. above.

537 The rest of the chapter is organized as follows. In Section 2 we present and analyze our
 538 method and its extensions, in Section 3 we briefly outline the numerical methods we use to solve
 539 the wave equation, in Section 4 we present numerical experiments, and in Section 5 we summarize
 540 and conclude.

541 Before proceeding we would like to acknowledge that although our method is distinct from
 542 the controllability method, it was the work by Grote and Tang, [63], that introduced us to CM and
 543 inspired us to derive the method discussed below.

544 **2.1 WaveHoltz: A New Method for Designing Scalable Parallel Helmholtz
 545 Solvers**

546 We consider the Helmholtz equation in a bounded open smooth domain Ω ,

$$\nabla \cdot (c^2(x)\nabla u) + \omega^2 u = f(x), \quad x \in \Omega, \quad (2.2)$$

⁵⁴⁷ with boundary conditions of the type

$$i\alpha\omega u + \beta(c^2(x)\vec{n} \cdot \nabla u) = 0, \quad \alpha^2 + \beta^2 = 1, \quad x \in \partial\Omega. \quad (2.3)$$

⁵⁴⁸ We assume $f \in L^2(\Omega)$ and that $c \in L^\infty(\Omega)$ with the bounds $0 < c_{\min} \leq c(x) \leq c_{\max} < \infty$ a.e. in ⁵⁴⁹ Ω . Away from resonances, this ensures that there is a unique weak solution $u \in H^1(\Omega)$ to (2.2). ⁵⁵⁰ Due to the boundary conditions u is in general complex valued.

⁵⁵¹ We first note that the function $w(t, x) := u(x) \exp(i\omega t)$ is a $T = 2\pi/\omega$ -periodic (in time) ⁵⁵² solution to the forced scalar wave equation

$$\begin{aligned} w_{tt} &= \nabla \cdot (c^2(x)\nabla w) - f(x)e^{i\omega t}, \quad x \in \Omega, \quad 0 \leq t \leq T, \\ w(0, x) &= v_0(x), \quad w_t(0, x) = v_1(x), \\ \alpha w_t + \beta(c^2(x)\vec{n} \cdot \nabla w) &= 0, \quad x \in \partial\Omega, \end{aligned} \quad (2.4)$$

where $v_0 = u$ and $v_1 = i\omega u$. Based on this observation, our approach is to find this w instead of u . We could thus look for initial data v_0 and v_1 such that w is a T -periodic solution to (2.4). However, there may be several such w , see [63], and we therefore impose the alternative constraint that a certain time-average of w should equal the initial data. More precisely, we introduce the following operator acting on the initial data $v_0 \in H^1(\Omega)$, $v_1 \in L^2(\Omega)$,

$$\Pi \begin{bmatrix} v_0 \\ v_1 \end{bmatrix} = \frac{2}{T} \int_0^T \left(\cos(\omega t) - \frac{1}{4} \right) \begin{bmatrix} w(t, x) \\ w_t(t, x) \end{bmatrix} dt, \quad T = \frac{2\pi}{\omega},$$

⁵⁵³ where $w(t, x)$ and its time derivative $w_t(t, x)$ satisfies the wave equation (2.4) with initial data ⁵⁵⁴ v_0 and v_1 . The result of $\Pi[v_0, v_1]^T$ can thus be seen as a filtering in time of $w(\cdot, x)$ around the ⁵⁵⁵ ω -frequency. We will further motivate the choice of time averaging in the analysis below. By ⁵⁵⁶ construction, the solution u of Helmholtz now satisfies the equation

$$\begin{bmatrix} u \\ i\omega u \end{bmatrix} = \Pi \begin{bmatrix} u \\ i\omega u \end{bmatrix}. \quad (2.5)$$

557 The WaveHoltz method then amounts to solving this equation with the fixed point iteration

$$\begin{bmatrix} v \\ v' \end{bmatrix}^{(n+1)} = \Pi \begin{bmatrix} v \\ v' \end{bmatrix}^{(n)}, \quad \begin{bmatrix} v \\ v' \end{bmatrix}^{(0)} \equiv 0. \quad (2.6)$$

558 Provided this iteration converges and the solution to (2.5) is unique, we obtain the Helmholtz
 559 solution as $u = \lim_{n \rightarrow \infty} v^n$.

560 **Remark 2.1.1.** Note that each iteration is inexpensive and that T is reduced by the reciprocal of ω
 561 as ω grows. If we assume that the number of degrees of freedom in each dimension scales with ω and
 562 that we evolve the wave equation with an explicit method this means that the number of timesteps
 563 per iteration is independent of ω . Also note that the iteration is trivial to implement (in parallel or
 564 serial) if there is already a time domain wave equation solver in place. The integral in the filtering
 565 is carried out independently for each degree of freedom and simply amounts to adding up a weighted
 566 sum (e.g. a trapezoidal sum) of the solution one timestep at a time. Finally, note that WHI allows
 567 all the advanced techniques that have been developed for wave equations (e.g. local timestepping,
 568 non-conforming discontinuous Galerkin finite elements h - and p -adaptivity etc.) can be transferred
 569 to the Helmholtz equation and other time harmonic problems.

570 2.1.1 Iteration for the Energy Conserving Case

571 Here we consider boundary conditions of either Dirichlet ($\beta = 0$) or Neumann ($\alpha = 0$) type.
 572 This is typically the most difficult case for iterative Helmholtz solvers when Ω is bounded. The
 573 wave energy is preserved in time and certain ω -frequencies in Helmholtz are resonant, meaning they
 574 equal an eigenvalue of the operator $-\nabla \cdot (c^2(x)\nabla)$. Moreover, the limiting amplitude principle does
 575 not hold, and one can thus not obtain the Helmholtz solution by solving the wave equation over a
 576 long time interval.

577 We start by introducing a simplified iteration for this case. With the given boundary condi-
 578 tions the solution to Helmholtz will be real valued, since f is a real valued function. Without loss
 579 of generality, we may then take $w_t(0, x) = 0$ and $w(t, x) = u(x) \cos(\omega t)$, since for a T -periodic real

580 valued solution there is a time when $w_t(0, x) = 0$. We choose that time as the initial time so that
 581 (2.4) becomes

$$\begin{aligned} w_{tt} &= \nabla \cdot (c(x)^2 \nabla w) - f(x) \cos(\omega t), \quad x \in \Omega, \quad 0 \leq t \leq T, \\ w(0, x) &= v(x), \quad w_t(0, x) \equiv 0, \\ \alpha w_t + \beta(c^2(x) \vec{n} \cdot \nabla w) &= 0, \quad x \in \partial\Omega. \end{aligned} \tag{2.7}$$

582 The simplified iteration is then defined as

$$v^{n+1} = \Pi v^n, \quad v^0 \equiv 0, \tag{2.8}$$

583 where

$$\Pi v = \frac{2}{T} \int_0^T \left(\cos(\omega t) - \frac{1}{4} \right) w(t, x) dt, \quad T = \frac{2\pi}{\omega}, \tag{2.9}$$

584 with $w(t, x)$ solving the wave equation (2.7) with initial data $v = v^n \in H^1(\Omega)$. We now analyze
 585 this iteration.

586 By the choice of boundary conditions the operator $-\nabla \cdot (c^2(x) \nabla)$ has a point spectrum
 587 with non-negative eigenvalues with corresponding eigenfunctions that form an orthonormal basis
 588 of $L^2(\Omega)$. Denote those eigenmodes $(\lambda_j^2, \phi_j(x))$, with $\|\phi_j\|_{L^2(\Omega)} = 1$. We assume that the angular
 589 frequency ω is not a resonance, i.e. $\omega^2 \neq \lambda_j^2$ for all j . The Helmholtz equation (2.2) is then
 590 wellposed.

We recall that for any $q \in L^2(\Omega)$ we can expand

$$q(x) = \sum_{j=0}^{\infty} \hat{q}_j \phi_j(x),$$

for some coefficients \hat{q}_j and

$$\|q\|_{L^2(\Omega)}^2 = \sum_{j=0}^{\infty} |\hat{q}_j|^2, \quad c_{\min}^2 \|\nabla q\|_{L^2(\Omega)}^2 \leq \sum_{j=0}^{\infty} \lambda_j^2 |\hat{q}_j|^2 \leq c_{\max}^2 \|\nabla q\|_{L^2(\Omega)}^2.$$

591 We start by expanding the Helmholtz solution u , the initial data v to the wave equation (2.7), and
 592 the forcing f in this way,

$$u(x) = \sum_{j=0}^{\infty} \hat{u}_j \phi_j(x), \quad v(x) = \sum_{j=0}^{\infty} \hat{v}_j \phi_j(x), \quad f(x) = \sum_{j=0}^{\infty} \hat{f}_j \phi_j(x).$$

⁵⁹³ Then,

$$-\lambda_j^2 \hat{u}_j + \omega^2 \hat{u}_j = \hat{f}_j \quad \Rightarrow \quad \hat{u}_j = \frac{\hat{f}_j}{\omega^2 - \lambda_j^2}.$$

⁵⁹⁴ For the wave equation solution $w(t, x)$ with initial data $w = v$ and $w_t = 0$ we have

$$w(t, x) = \sum_{j=0}^{\infty} \hat{w}_j(t) \phi_j(x), \quad \hat{w}_j(t) = \hat{u}_j \left(\cos(\omega t) - \cos(\lambda_j t) \right) + \hat{v}_j \cos(\lambda_j t). \quad (2.10)$$

The filtering step (2.9) then gives

$$\Pi v = \sum_{j=0}^{\infty} \bar{v}_j \phi_j(x), \quad \bar{v}_j = \hat{u}_j (1 - \beta(\lambda_j)) + \hat{v}_j \beta(\lambda_j),$$

⁵⁹⁵ where

$$\beta(\lambda) := \frac{2}{T} \int_0^T \left(\cos(\omega t) - \frac{1}{4} \right) \cos(\lambda t) dt.$$

⁵⁹⁶ We introduce the linear operator $\mathcal{S} : L^2(\Omega) \rightarrow L^2(\Omega)$,

$$\mathcal{S} \sum_{j=0}^{\infty} \hat{u}_j \phi_j(x) := \sum_{j=0}^{\infty} \beta(\lambda_j) \hat{u}_j \phi_j(x), \quad (2.11)$$

⁵⁹⁷ which gives the filtered solution of the wave equation with $f = 0$, when applied to the initial data

⁵⁹⁸ v . We can then write the iteration as

$$v^{n+1} = \Pi v^n = \mathcal{S}(v^n - u) + u. \quad (2.12)$$

⁵⁹⁹ The operator \mathcal{S} is self-adjoint and has the same eigenfunctions $\phi_j(x)$ as $-\nabla \cdot (c^2(x) \nabla)$ but with the
⁶⁰⁰ (real) eigenvalues $\beta(\lambda_j)$. The convergence properties of the iteration depend on these eigenvalues
⁶⁰¹ and it is therefore of interest to study the range of the filter transfer function β . Figure 2.1 shows
⁶⁰² a plot of β which indicates that the eigenvalues of \mathcal{S} are inside the unit interval, with a few of
⁶⁰³ them being close to 1 (when $\lambda_j \approx \omega$), and most of them being close to zero (when $\lambda_j \gg \omega$). In the
⁶⁰⁴ appendix we show the following lemma about β .

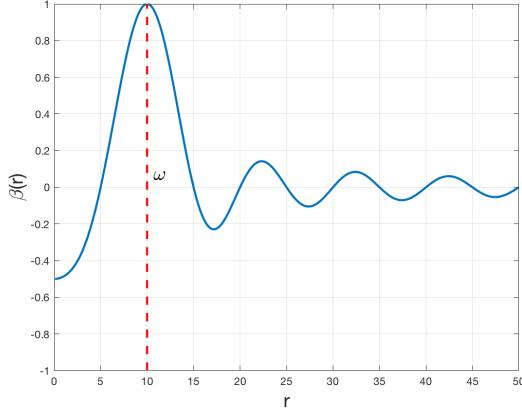


Figure 2.1: The filter transfer function β for $\omega = 10$.

Lemma 2.1.1. *The filter transfer function β satisfies $\beta(\omega) = 1$ and*

$$\begin{aligned} 0 \leq \beta(\lambda) &\leq 1 - \frac{1}{2} \left(\frac{\lambda - \omega}{\omega} \right)^2, & \text{when } \left| \frac{\lambda - \omega}{\omega} \right| \leq \frac{1}{2}, \\ |\beta(\lambda)| &\leq \frac{1}{2}, & \text{when } \left| \frac{\lambda - \omega}{\omega} \right| \geq \frac{1}{2}, \\ |\beta(\lambda)| &\leq b_0 \frac{\omega}{\lambda - \omega}, & \text{when } \lambda > \omega. \end{aligned}$$

605 where $b_0 = \frac{3}{4\pi}$. Moreover, close to ω we have the local expansion

$$\beta(\omega + r) = 1 - b_1 \left(\frac{r}{\omega} \right)^2 + R(r/\omega) \left(\frac{r}{\omega} \right)^3, \quad b_1 = \frac{2\pi^2}{3} - \frac{1}{4} \approx 6.33, \quad \|R\|_\infty \leq \frac{5\pi^3}{6}. \quad (2.13)$$

606 **Remark 2.1.2.** It is easy to see that $\beta(\omega) = 1$ for any constant besides $1/4$. The particular choice
 607 $1/4$ is made to ensure that $\beta'(\omega) = 0$, which is necessary to keep $\beta \leq 1$ in a neighborhood of ω . We
 608 explore other possibilities in Section 2.1.3.

609 From this lemma we can derive some results for the operator \mathcal{S} . To do this we first quantify
 610 the non-resonance condition. We let

$$\delta_j = \frac{\lambda_j - \omega}{\omega},$$

be the relative size of the gap between λ_j and the Helmholtz frequency, and then denote the smallest gap (in magnitude) by δ ,

$$\delta = \delta_{j^*}, \quad j^* = \operatorname{argmin}_j |\delta_j|.$$

611 Then we have

612 **Lemma 2.1.2.** Suppose $\delta > 0$. The spectral radius ρ of \mathcal{S} is strictly less than one, and for small

613 δ ,

$$\rho = 1 - b_1 \delta^2 + O(\delta^3), \quad (2.14)$$

614 with b_1 as in Lemma 2.1.1. Moreover, \mathcal{S} is a bounded linear map from $L^2(\Omega)$ to $H^1(\Omega)$.

Proof. From Lemma 2.1.1 we get

$$\rho = \sup_j |\beta(\lambda_j)| \leq \sup_j \max\left(1 - \frac{1}{2}\delta_j^2, \frac{1}{2}\right) \leq \max\left(1 - \frac{1}{2}\delta^2, \frac{1}{2}\right) < 1.$$

For the more precise estimate when δ is small we will use (2.13). Since $1 > \rho \geq \beta(\omega + \omega\delta) \rightarrow 1$ as $\delta \rightarrow 0$, we can assume that $\rho > 1 - \eta^2/2$, with $\eta := b_1/2\|R\|_\infty$, for small enough δ . Then, since $|\beta(\omega + \omega\delta_j)| \leq 1 - \eta^2/2$ for $|\delta_j| > \eta$ by Lemma 2.1.1, we have

$$\rho = \sup_{|\delta_j| \leq \eta} \beta(\omega + \omega\delta_j) = \beta(\omega + \omega\delta_{k^*}),$$

for some k^* with $|\delta_{k^*}| \leq \eta$. If $\delta_{k^*} = \delta_{j^*}$ (where $\delta = |\delta_{j^*}|$) then (2.13) gives (2.14). If not, we have

$\eta \geq |\delta_{k^*}| \geq \delta$ and by Lemma 2.1.1

$$0 \leq \beta(\omega + \omega\delta_{k^*}) - \beta(\omega + \omega\delta_{j^*}) = -b_1(\delta_{k^*}^2 - \delta^2) + R(\delta_{k^*})\delta_{k^*}^3 - R(\delta_{j^*})\delta_{j^*}^3 \leq -b_1(\delta_{k^*}^2 - \delta^2) + \frac{b_1}{2}(\delta_{k^*}^2 + \delta^2),$$

which implies that $\delta_{k^*}^2 \leq 3\delta^2$ and that

$$0 \leq b_1(\delta_{k^*}^2 - \delta^2) \leq R(\delta_{k^*})\delta_{k^*}^3 - R(\delta_{j^*})\delta_{j^*}^3 \leq \|R\|_\infty(1 + 3\sqrt{3})\delta^3.$$

Therefore,

$$\rho = 1 - b_1 \delta_{k^*}^2 + O(\delta_{k^*}^3) = 1 - b_1 \delta^2 + b_1(\delta^2 - \delta_{k^*}^2) + O(\delta_{k^*}^3) = 1 - b_1 \delta^2 + O(\delta_{k^*}^3 + \delta^3) = 1 - b_1 \delta^2 + O(\delta^3).$$

615 This shows (2.14). For the second statement, we note first that by Lemma

Lemma 1. filterlemma,

$$|\lambda_j \beta(\lambda_j)| \leq \omega \begin{cases} 1, & \lambda_j \leq \omega, \\ \frac{b_0 \lambda_j}{\lambda_j - \omega}, & \lambda_j > \omega, \end{cases} = \omega \begin{cases} 1, & \lambda_j \leq \omega, \\ b_0(1 + 1/\delta_j), & \lambda_j > \omega, \end{cases} \leq \omega \min(1, b_0(1 + 1/|\delta|)) =: D.$$

Suppose now that $g \in L^2(\Omega)$ and

$$g(x) = \sum_{j=0}^{\infty} \hat{g}_j \phi_j(x).$$

Then

$$\|\mathcal{S}g\|_{H^1(\Omega)}^2 \leq \sum_{j=0}^{\infty} |\beta(\lambda_j)|^2 |\hat{g}_j|^2 + \sum_{j=0}^{\infty} \frac{\lambda_j^2}{c_{\min}^2} |\beta(\lambda_j)|^2 |\hat{g}_j|^2 \leq \left(1 + \frac{D^2}{c_{\min}^2}\right) \sum_{j=0}^{\infty} |\hat{g}_j|^2 = \left(1 + \frac{D^2}{c_{\min}^2}\right) \|g\|_{L^2(\Omega)}^2.$$

616 This proves the lemma. \square

Letting $e^n := u - v^n$ we can rearrange (2.12) and obtain

$$e^{n+1} = \mathcal{S}e^n \Rightarrow \|e^{n+1}\|_{L^2(\Omega)} \leq \rho \|e^n\|_{L^2(\Omega)} \Rightarrow \|e^n\|_{L^2(\Omega)} \leq \rho^n \|e^0\|_{L^2(\Omega)} \rightarrow 0,$$

which shows that v^n converges to u in L^2 . By Lemma 2.1.2 all iterates $v^n \in H^1(\Omega)$ since $v^0 = 0$.

We can therefore also get convergence in H^1 . Let

$$e^n(x) = \sum_{j=0}^{\infty} \hat{e}_j^n \phi_j(x),$$

and consider similarly

$$\begin{aligned} \sum_{j=0}^{\infty} |\hat{e}_j^{n+1}|^2 \lambda_j^2 &= \sum_{j=0}^{\infty} \beta(\lambda_j)^2 |\hat{e}_j^n|^2 \lambda_j^2 \leq \rho^2 \sum_{j=0}^{\infty} |\hat{e}_j^n|^2 \lambda_j^2 \Rightarrow \\ \|\nabla e^n\|_{L^2(\Omega)}^2 &\leq \frac{1}{c_{\min}^2} \sum_{j=0}^{\infty} |\hat{e}_j^n|^2 \lambda_j^2 \leq \frac{\rho^{2n}}{c_{\min}^2} \sum_{j=0}^{\infty} |\hat{e}_j^0|^2 \lambda_j^2 \leq \rho^{2n} \frac{c_{\max}^2}{c_{\min}^2} \|\nabla e^0\|_{L^2(\Omega)}^2 \rightarrow 0. \end{aligned}$$

617 We conclude that the iteration converges in H^1 with convergence rate ρ . By Lemma 2.1.1 we have

618 $\rho \sim 1 - 6.33\delta^2$ and, not surprisingly, the smallest gap, δ , determines the convergence factor. We

619 have thus showed

620 **Theorem 2.1.3.** *The iteration in (2.8) and (2.9) converges in $H^1(\Omega)$ for the Dirichlet and Neu-
621 mann problems away from resonances to the solution of the Helmholtz equation (2.2). The conver-
622 gence rate is $1 - O(\delta^2)$, where δ is the minimum gap between ω and the eigenvalues of $-\nabla \cdot (c^2(x)\nabla)$.*

623 As discussed in the introduction, the dependence of the convergence rate on ω is often of
624 interest. For the energy conserving case, however, this question is ambiguous as the problem is not
625 well-defined for all ω . As soon as $\omega = \lambda_j$ there are either no or an infinite number of solutions.

626 In higher dimensions, the eigenvalues λ_j get denser as j increases, meaning that in general the
 627 problem will be closer and closer to resonance as ω grows. Therefore, solving the interior undamped
 628 Helmholtz equation for high frequencies, with pure Dirichlet or Neumann boundary conditions, may
 629 not be of great practical interest.

Nevertheless, we can make the following analysis. By the work of Weyl [109] we know that the eigenvalues grow asymptotically as $\lambda_j \sim j^{1/d}$ in d dimensions. The average minimum gap δ when $\omega \approx \lambda_j$ is then

$$\begin{aligned}\delta &\approx \frac{1}{\lambda_{j+1} - \lambda_j} \int_{\lambda_j}^{\lambda_{j+1}} \frac{\min(\lambda - \lambda_j, \lambda_{j+1} - \lambda)}{\omega} d\lambda = \frac{\lambda_{j+1} - \lambda_j}{4\omega} \sim \frac{(j+1)^{1/d} - j^{1/d}}{\omega} \approx \frac{j^{1/d-1}}{d\omega} \\ &\sim \frac{\omega^{1-d}}{\omega} \\ &\sim \omega^{-d}.\end{aligned}$$

630 When the convergence rate is $1 - O(\delta^2)$, the number iterations to achieve a fixed accuracy grows
 631 as $O(1/\delta^2)$. This shows that the number of iterations would grow at the unacceptable rate ω^{2d} for
 632 the iteration.

Fortunately, one can accelerate the convergence by using the conjugate gradient method in the energy conserving case and with any other Krylov method in the general case. The linear system that we actually want to solve is

$$(I - \mathcal{S})v =: \mathcal{A}v = b := \Pi 0.$$

633 Moreover, with $b = \Pi 0$ pre-computed we can easily evaluate the action of \mathcal{A} at the cost of a single
 634 wave solve. Precisely, since $\mathcal{A}v = v - \Pi v + b$ we simply carry out the evaluation of $\mathcal{A}v$ by evolving
 635 the wave equation for one period in time with v as the initial data and then subtract the filtered
 636 solution from the sum of the initial data and the right hand side b .

637 The operator \mathcal{A} is self adjoint and positive, since $-1/2 < \beta(\lambda_j) < 1$, which implies that the
 638 eigenvalues of \mathcal{A} lie in the interval $(0, 3/2)$. The condition number of \mathcal{A} is of the same order as
 639 $1 - \rho$, where ρ is the spectral radius of \mathcal{S} , i.e. by the simple analysis above, $\text{cond}(\mathcal{A}) \sim \omega^{2d}$. If
 640 this system is solved using the (**unconditioned**) conjugate gradient method the convergence rate

641 is $1 - 1/\sqrt{\text{cond}(\mathcal{A})} \sim 1 - 1/\omega^d$, [25]. Thus, then the method just requires $\sim \omega^d$ iterations for fixed
 642 accuracy.

Remark 2.1.3. *The operator \mathcal{A} is self-adjoint and coercive when $\delta > 0$ since*

$$\langle \mathcal{A}u, u \rangle = \langle (I - \mathcal{S})u, u \rangle = \sum_{j=0}^{\infty} (1 - \beta(\lambda_j)) |\hat{u}_j|^2 \geq (1 - \rho) \|u\|^2.$$

643 This should be contrasted with the original indefinite Helmholtz problem, which is not coercive. In
 644 fact, the eigenvalues satisfy the simple relation $\lambda_{\text{WHI}} = 1 - \beta(\lambda_{\text{Helmholtz}} + \omega) \not\approx 0$. The two for-
 645 mulations are however mathematically equivalent for the interior Dirichlet and Neumann problems
 646 away from resonances, as the analysis above shows.

647 The coercivity also implies that the solution to (2.5) for the simplified iteration is unique
 648 since $w = \Pi w$ is equivalent to $\mathcal{A}(w - u) = 0$.

649 **Remark 2.1.4.** A discretization would have approximately a fixed number of grid points per wave-
 650 length, leading to a (sparse) matrix of size $N \times N$ with $N \sim \omega^d$. Hence, the number of iterations
 651 for WHI would be $O(N^2)$ and the total cost $O(N^3)$ since each iteration costs $O(N)$. This should be
 652 compared with a direct solution method which is better than $O(N^3)$ when the matrix is sparse.

653 **Remark 2.1.5.** In the Krylov accelerated case this analysis suggests that the number of iterations
 654 would now be $O(N)$ and the total cost $O(N^2)$. However, in the experiments below we observe slightly
 655 better complexity for interior problems and significantly better complexity for open problems. In fact,
 656 for the open problems we find that, in both two and three dimensions, the number of iterations scale
 657 as $\sim \omega$ which is the required number of iterations for the information to travel through the domain.

658 2.1.2 Analysis of the Discrete Iteration

659 To better understand the effects of discretizations we consider the following discrete version
 660 of the algorithm for the energy conserving case described above in Section 2.1.1. We introduce the
 661 temporal grid points $t_n = n\Delta t$ and a spatial grid with N points together with the vector $w^n \in \mathbb{R}^N$
 662 containing the grid function values of the approximation at $t = t_n$. We also let $f \in \mathbb{R}^N$ hold the

663 corresponding values of the right hand side. The discretization of the continuous spatial operator
 664 $-\nabla \cdot (c^2(x)\nabla)$, including the boundary conditions, is denoted L_h and it can be represented as an
 665 $N \times N$ matrix. The values $-\nabla \cdot (c^2(x)\nabla w)$ are then approximated by $L_h w^n$. As in the continuous
 666 case, we assume L_h has the eigenmodes (λ_j^2, ϕ_j) , such that $L_h \phi_j = \lambda_j^2 \phi_j$ for $j = 1, \dots, N$, where all
 667 λ_j are strictly positive and ordered as $0 \leq \lambda_1 \leq \dots \leq \lambda_N$.

We let the Helmholtz solution u be given

$$-L_h u + \omega^2 u = f.$$

668 The numerical approximation of the iteration operator is denoted Π_h , and it is implemented as
 669 follows. Given $v \in \mathbb{R}^N$, we use the leap frog method to solve the wave equation as

$$w^{n+1} = 2w^n - w^{n-1} - \Delta t^2 L_h w^n - \Delta t^2 f \cos(\omega t_n), \quad (2.15)$$

with time step $\Delta t = T/M$ for some integer M , and initial data

$$w^0 = v, \quad w^{-1} = v - \frac{\Delta t^2}{2}(L_h v + f).$$

670 The trapezoidal rule is then used to compute $\Pi_h v$,

$$\Pi_h v = \frac{2\Delta t}{T} \sum_{n=0}^M \eta_n \left(\cos(\omega t_n) - \frac{1}{4} \right) w^n, \quad \eta_n = \begin{cases} \frac{1}{2}, & n = 0 \text{ or } n = M, \\ 1, & 0 < n < M. \end{cases} \quad (2.16)$$

671 With these definitions we can prove

672 **Theorem 2.1.4.** *Suppose there are no resonances, such that $\delta_h = \min_j |\lambda_j - \omega|/\omega > 0$. Moreover,*
 673 *assume that Δt satisfies the stability and accuracy requirements*

$$\Delta t < \frac{2}{\lambda_N + 2\omega/\pi}, \quad \Delta t \omega \leq \min(\delta_h, 1). \quad (2.17)$$

Then the fixed point iteration $v^{(k+1)} = \Pi_h v^{(k)}$ with $v^{(0)} = 0$ converges to v^∞ which is a solution to the discretized Helmholtz equation with the modified frequency $\tilde{\omega}$,

$$-L_h v^\infty + \tilde{\omega}^2 v^\infty = f, \quad \tilde{\omega} = 2 \frac{\sin(\Delta t \omega / 2)}{\Delta t}.$$

674 The convergence rate is at least $\rho_h = \max(1 - 0.3\delta_h^2, 0.6)$.

Proof. We expand all functions in eigenmodes of L_h ,

$$w^n = \sum_{j=1}^N \hat{w}_j^n \phi_j, \quad f = \sum_{j=1}^N \hat{f}_j \phi_j, \quad u = \sum_{j=1}^N \hat{u}_j \phi_j, \quad v = \sum_{j=1}^N \hat{v}_j \phi_j, \quad v^\infty = \sum_{j=1}^N \hat{v}_j^\infty \phi_j.$$

Then the Helmholtz eigenmodes of u and v^∞ satisfy

$$\hat{u}_j = \frac{\hat{f}_j}{\omega^2 - \lambda_j^2}, \quad \hat{v}_j^\infty = \frac{\hat{f}_j}{\tilde{\omega}^2 - \lambda_j^2}.$$

We note that $\tilde{\omega}$ is not resonant and \hat{v}_j^∞ is well-defined for all j , since by (5) and (3.25)

$$|\tilde{\omega} - \lambda_j| \geq |\omega - \lambda_j| - |\tilde{\omega} - \omega| \geq \omega \delta_h - \frac{\Delta t^2 \omega^3}{24} \geq \omega \left(\delta_h - \frac{1}{24} \min(\delta_h, 1)^2 \right) > 0.$$

675 The wave solution eigenmodes are given by the difference equation

$$\hat{w}_j^{n+1} - 2\hat{w}_j^n + \hat{w}_j^{n-1} + \Delta t^2 \lambda_j^2 \hat{w}_j^n = -\Delta t^2 \hat{f}_j \cos(\omega t_n). \quad (2.18)$$

with initial data

$$\hat{w}_j^0 = \hat{v}_j, \quad \hat{w}_j^{-1} = \hat{v}_j \left(1 - \frac{1}{2} \Delta t^2 \lambda_j^2 \right) - \frac{1}{2} \Delta t^2 \hat{f}_j.$$

By (3.25)

$$|2 - \Delta t^2 \lambda_j^2| < 2,$$

676 and the characteristic polynomial for the equation, $r^2 + (\Delta t^2 \lambda_j^2 - 2)r + 1$, then has two roots on
677 the boundary of the unit circle. The solution is therefore stable and is given by (the verification
678 of which is found in Appendix .2)

$$\hat{w}_j^n = (\hat{v}_j - \hat{v}_j^\infty) \cos(\tilde{\lambda}_j t_n) + \hat{v}_j^\infty \cos(\omega t_n), \quad (2.19)$$

where $\tilde{\lambda}_j$ is well-defined by the relation

$$2 \frac{\sin(\Delta t \tilde{\lambda}_j / 2)}{\Delta t} = \lambda_j.$$

Now, let

$$\Pi_h v = \sum_{j=1}^{\infty} \bar{v}_j \phi_j.$$

Then the numerical integration gives

$$\begin{aligned}\bar{v}_j &= \frac{2\Delta t}{T} \sum_{n=0}^M \eta_n \left(\cos(\omega t_n) - \frac{1}{4} \right) \left((\hat{v}_j - \hat{v}_j^\infty) \cos(\tilde{\lambda}_j t_n) + \hat{v}_j^\infty \cos(\omega t_n) \right) \\ &= (\hat{v}_j - \hat{v}_j^\infty) \beta_h(\tilde{\lambda}_j) + \hat{v}_j^\infty \beta_h(\omega) = \hat{v}_j \beta_h(\tilde{\lambda}_j) + (1 - \beta_h(\tilde{\lambda}_j)) \hat{v}_j^\infty,\end{aligned}$$

where

$$\beta_h(\lambda) = \frac{2\Delta t}{T} \sum_{n=0}^M \eta_n \cos(\lambda t_n) \left(\cos(\omega t_n) - \frac{1}{4} \right),$$

and we used the fact that the trapezoidal rule is exact, and equal to one, when $\lambda = \omega$. (Recall that for periodic functions the trapezoidal rule is exact for all pure trigonometric functions of order less than the number of grid points.) Hence, if $|\beta_h(\tilde{\lambda}_j)| < 1$ the j -th mode in the fixed point iteration converges to \hat{v}_j^∞ . This is ensured by the following lemma, the proof of which is found in Appendix .3.

Lemma 2.1.5. *Under the assumptions of Theorem 2.1.4,*

$$\max_{1 \leq j \leq N} |\beta_h(\tilde{\lambda}_j)| \leq \rho_h =: \max(1 - 0.3\delta_h^2, 0.63). \quad (2.20)$$

Since the bound $|\beta_h(\tilde{\lambda}_j)| \leq \rho_h < 1$ in the lemma is uniform for all j the convergence $v^{(k)} \rightarrow v^\infty$ with rate at least ρ_h follows. This concludes the proof of the theorem.

687

□

Remark 2.1.6. *The discretization above is used as an example to illustrate the impact of going from the continuous to the discrete iteration. For a particular discretization we can improve the iteration further by using the knowledge of how it approximates ω and the eigenvalues of the continuous operator. Indeed, for the discretization above, let us define $\bar{\omega}$ by the relation*

$$\omega = 2 \frac{\sin(\Delta t \bar{\omega}/2)}{\Delta t}.$$

688 Then if we use $f \cos(\bar{\omega} t_n)$ instead of $f \cos(\omega t_n)$ in the time stepping (2.15), the limit will be precisely the Helmholtz solution, $v^\infty = u$. Furthermore, the condition $\Delta t \omega \leq \min(\delta_h, 1)$ can be quite restrictive for problems close to resonance. It is only important to ensure convergence of the iterations. Another way to do that is to slightly change the discrete filter by replacing the constant 1/4

692 in (3.24) by a Δt -dependent number such that $|\beta_h(\lambda)| < 1$ for $\lambda \neq \omega$. Another option is to use a
 693 higher order quadrature rule, which would mitigate the restriction on Δt .

694 **2.1.3 Tunable Filters**

In Lemma 2.1.1 we saw that the filter transfer function satisfies $\beta(\omega) = 1$ and $-1/2 < \beta(r) < 1$ when $r \neq \omega$ and that these conditions guaranteed convergence of the WaveHoltz iteration. To improve convergence when $r \approx \omega$ we now consider a more general filter transfer function

$$\bar{\beta}(\lambda) = \frac{2}{T} \int_0^T (\cos(\omega t) + \alpha(t)) \cos(\lambda t) dt, \quad \alpha(t) = a_0 + \sum_{n=1}^{\infty} a_n \sin(n\omega t), \quad (2.21)$$

where we refer to $\alpha(t)$ as a *time-dependent shift*. As before, necessary conditions for convergence are $\bar{\beta}(\omega) = 1$, $\bar{\beta}'(\omega) = 0$. Straightforward calculations reveal that these conditions require that the two first coefficients must satisfy

$$a_1 = \frac{1}{2\pi} (1 + 4a_0).$$

The remaining terms in the sum are orthogonal to $\cos(\lambda t)$ when $\lambda = \omega$. Carrying out the integration in full for each term yields the general form

$$\bar{\beta}(\lambda) = \frac{\lambda\omega \sin(\lambda T)}{\pi(\lambda^2 - \omega^2)} + a_0 \frac{\omega \sin(\lambda T)}{\pi\lambda} + \sum_{n=1}^{\infty} a_n \frac{n\omega^2}{\pi(\lambda^2 - n^2\omega^2)} (\cos(\lambda T) - 1),$$

from which it follows that another necessary condition is $|a_0| < 1/2$ since $|\bar{\beta}(r)| < 1$ and

$$\bar{\beta}(0) = a_0 \lim_{\lambda \rightarrow 0} \frac{\omega \sin(2\pi\lambda/\omega)}{\pi\lambda} = 2a_0.$$

695 We note that the standard filter, where $a_0 = -1/4$ and $a_1 = 0$, satisfies the necessary conditions.

696 **Remark 2.1.7.** For the remaining coefficients a_n we only need to ensure that $|\bar{\beta}(r)| < 1$ which
 697 leaves large freedom to design $\bar{\beta}$. For example we may try to maximize $|\bar{\beta}''(\omega)|$ (minimize $\bar{\beta}''(\omega)$)
 698 so that $\bar{\beta}(r)$ is sharply peaked around $r = \omega$. We do not pursue a systematic study of this here but
 699 illustrate the utility of the added flexibility of (2.21) with numerical experiments below in Section 2.3.

700 **2.1.4 Multiple Frequencies in One Solve**

We can use the WaveHoltz algorithm to solve for multiple frequencies at once. Suppose we look for the solutions u_i of

$$\nabla \cdot (c^2(x) \nabla u_i) + \omega_i^2 u_i = f_i(x), \quad i = 1, \dots, N,$$

with the same c and boundary condition for all i . To find those solutions we include all frequencies in the wave equation part of the iteration (2.4), and solve

$$w_{tt} = \nabla \cdot (c(x)^2 \nabla w) - \sum_{i=1}^N f_i(x) \cos(\omega_i t). \quad (2.22)$$

701 We then seek a decomposition

$$w(x, t) \equiv \sum_{i=1}^N u_i(x) \cos(\omega_i t), \quad (2.23)$$

of the solution. The filtering part of the WaveHoltz iteration is also updated to reflect the multiple frequencies

$$v_{n+1} = \frac{2}{T} \int_0^T \left(\sum_{i=1}^N \cos(\omega_i t) - \frac{1}{4} \right) w(x, t) dt.$$

702 As before we take $v_0 = 0$ when we deal with energy conserving boundary conditions. To this end
 703 we assume that the frequencies are related by an integer multiple in a way so that the period T
 704 can be chosen based on the lowest frequency.

705 The different $u_i(x)$ in (2.23) can be found as follows. Once we have found the time periodic
 706 solution to (2.22) evolve one more period and sample $w(x, t)$ at N distinct times t_j , $j = 1, \dots, N$.

707 We then have

$$u_i(x) = \sum_{j=1}^N \beta_{ij} w(x, t_j),$$

708 where the coefficients β_{ij} are the elements of A^{-1} with the elements of A being $a_{ij} = \cos(\omega_j t_i)$.

709 **2.1.5 WaveHoltz Iteration for Impedance Boundary Conditions**

710 For impedance and other boundary conditions that leads to a decreasing energy for the wave
 711 equation we cannot make the simplifying assumption in (2.4) that $w_t(0, x) = 0$ but we must seek

⁷¹² both $v_0(x)$ and $v_1(x)$ in (2.4). To do so we define an extended iteration (2.8) where we apply Π to
⁷¹³ both the displacement and the velocity:

$$\begin{bmatrix} v \\ v' \end{bmatrix}^{(n+1)} = \tilde{\Pi} \begin{bmatrix} v \\ v' \end{bmatrix}^{(n)}, \quad \begin{bmatrix} v \\ v' \end{bmatrix}^{(0)} \equiv 0, \quad (2.24)$$

⁷¹⁴ where

$$\tilde{\Pi} \begin{bmatrix} v \\ v' \end{bmatrix} = \frac{2}{T} \int_0^T \left(\cos(\omega t) - \frac{1}{4} \right) \begin{bmatrix} w(t, x) \\ w_t(t, x) \end{bmatrix} dt, \quad T = \frac{2\pi}{\omega}. \quad (2.25)$$

⁷¹⁵ Here $w(t, x)$ and its time derivative $w_t(t, x)$ satisfies the wave equation (2.4) with initial data
⁷¹⁶ $v_0(x) \equiv v^{(n)}$ and $v_1(x) \equiv v'^{(n)}$.

⁷¹⁷ 2.2 Wave Equation Solvers

⁷¹⁸ In this section we briefly outline the numerical methods we use in the experimental section
⁷¹⁹ below. We consider both discontinuous Galerkin finite element solvers and finite difference solvers.
⁷²⁰ In all the experiments we always use the trapezoidal rule to compute the integral in the WaveHoltz
⁷²¹ iteration.

⁷²² 2.2.1 The Energy Based Discontinuous Galerkin Method

⁷²³ Our spatial discretization is a direct application of the formulation described for general
⁷²⁴ second order wave equations in [9, 10]. Here we outline the spatial discretization for the special
⁷²⁵ case of the scalar wave equation in one dimension and refer the reader to [9] for the general case.

⁷²⁶ The energy of the scalar wave equation is

$$H(t) = \int_D \frac{v^2}{2} + G(x, w_x) dx,$$

⁷²⁷ where

$$G(x, w_x) = \frac{c^2(x)w_x^2}{2},$$

⁷²⁸ is the potential energy density, v is the velocity (not to be confused with the iterates v^n above)
⁷²⁹ or the time derivative of the displacement, $v = w_t$. The wave equation, written as a second order

⁷³⁰ equation in space and first order in time then takes the form

$$\begin{aligned} w_t &= v, \\ v_t &= -\delta G - f(x) \cos(\omega t), \end{aligned}$$

⁷³¹ where δG is the variational derivative of the potential energy

$$\delta G = -(G_{w_x})_x = -(c^2(x)w_x)_x.$$

⁷³² For the continuous problem the change in energy is

$$\frac{dH(t)}{dt} = \int_D vv_t + w_t(c^2(x)w_x)_x dx = - \int_D vf(x) \cos(\omega t) dx + [w_t(c^2(x)w_x)]_{\partial D}, \quad (2.26)$$

⁷³³ where the last equality follows from integration by parts together with the wave equation. Now,
⁷³⁴ a variational formulation that mimics the above energy identity can be obtained if the equation
⁷³⁵ $v - w_t = 0$ is tested with the variational derivative of the potential energy. Let Ω_j be an element and
⁷³⁶ $\Pi^s(\Omega_j)$ be the space of polynomials of degree s , then the variational formulation on that element
⁷³⁷ is:

⁷³⁸ **Problem 1.** Find $v^h \in \Pi^s(\Omega_j)$, $w^h \in \Pi^r(\Omega_j)$ such that for all $\psi \in \Pi^s(\Omega_j)$, $\phi \in \Pi^r(\Omega_j)$

$$\int_{\Omega_j} c^2 \phi_x \left(\frac{\partial w_x^h}{\partial t} - v_x^h \right) dx = [c^2 \phi_x \cdot n (v^* - v^h)]_{\partial \Omega_j}, \quad (2.27)$$

$$\int_{\Omega_j} \psi \frac{\partial v^h}{\partial t} + c^2 \psi_x \cdot w_x^h + \psi f(x) \cos(\omega t) dx = [\psi (c^2 w_x)^*]_{\partial \Omega_j}. \quad (2.28)$$

⁷³⁹ Let $[[\zeta]]$ and $\{\zeta\}$ denote the jump and average of a quantity ζ at the interface between two
⁷⁴⁰ elements, then, choosing the numerical fluxes as

$$\begin{aligned} v^* &= \{v\} - \tau_1 [[c^2 w_x]] \\ (c^2 w_x)^* &= \{c^2 w_x\} - \tau_2 [[v]], \end{aligned}$$

⁷⁴¹ will yields a contribution $-\tau_1([[c^2 w_x]])^2 - \tau_2([[v]])^2$ from each element face. To this end we choose
⁷⁴² $\tau_i > 0$ (so called upwind or Sommerfeld fluxes) which together with the choice that the approxi-
⁷⁴³ mation spaces be of the same degree $r = s$ result in methods that are $r + 1$ order accurate in space

744 and measured in the L_2 norm. We note that even in the case of energy conserving numerical fluxes
745 the formulation does not lead to a **symmetric** matrix for the WaveHoltz iteration (it is of course
746 positive definite though).

747 Physical boundary conditions can also be handled by appropriate specification of the numer-
748 ical fluxes, see [9] for details. The above variational formulation and choice of numerical fluxes
749 results in an energy identity similar to (2.26). However, as the energy is invariant to certain trans-
750 formations the variational problem does not fully determine the time derivatives of w^h on each
751 element and independent equations must be introduced. In this case there is one invariant and an
752 independent equation is $\int_{\Omega_j} \left(\frac{\partial w^h}{\partial t} - v^h \right) = 0$.

753 Denoting the degrees of freedom on element Ω_j by v_j and w_j the semi-discretization according
754 to (2.27)-(2.28) on element Ω_j can be written

$$S\left(\frac{\partial w_j}{\partial t} - v_j\right) = L_1(v_{j-1}, v_j, v_{j+1}, w_{j-1}, w_j, w_{j+1}), \quad (2.29)$$

$$M\frac{\partial v_j}{\partial t} + Sw_j + f_j \cos(\omega t) = L_2(v_{j-1}, v_j, v_{j+1}, w_{j-1}, w_j, w_{j+1}), \quad (2.30)$$

755 where the elements of the element matrices M and S are $M_{kl} = \int_{\Omega_j} \phi_k \phi_l dx$ and $S_{kl} = \int_{\Omega_j} c^2(\phi_k)_x (\phi_l)_x dx$
756 respectively and the lift operators L_1 and L_2 represents the numerical fluxes. Note that a conve-
757 nient way to directly enforce the independent equation is to compute the time derivatives of w_j
758 according to

$$\frac{\partial w_j}{\partial t} = v_j = S^\dagger L_1(v_{j-1}, v_j, v_{j+1}, w_{j-1}, w_j, w_{j+1}),$$

759 where S^\dagger is the pseudo inverse of S .

760 2.2.2 Finite Difference Discretizations

761 For the finite difference examples we exclusively consider Cartesian domains $(x, y, z) \in$
762 $[L_x, R_x] \times [L_y, R_y] \times [L_z, R_z]$ discretized by uniform grids $(x_i, y_j, z_k) = (L_x + ih_x, L_y + jh_y, L_z + kh_z)$,
763 with $i = 0, \dots, n_x$ and $h_x = (R_x - L_x)/n_x$, etc.

764 When we have impedance boundary conditions on the form $w_t \pm \vec{n} \cdot \nabla w = 0$ we evolve the
 765 wave equation as a first order system in time according to the semi-discrete approximation

$$\frac{dv_{ijk}(t)}{dt} = (D_+^x D_-^x + D_+^y D_-^y + D_+^z D_-^z)w_{ijk}, \quad (2.31)$$

$$\frac{dw_{ijk}(t)}{dt} = v_{ijk}, \quad (2.32)$$

766 for all grid points that do not correspond to Dirichlet boundary conditions. On boundaries with
 767 impedance conditions we find the ghost point values by enforcing (here illustrated on the top of
 768 the domain)

$$v_{ijn_z} - D_0^z w_{ijn_z} = 0. \quad (2.33)$$

769 Here we have used the standard forward, backward and centered finite difference operators, for
 770 example $h_x D_+^x w_{i,j,k} = w_{i+1,j,k} - w_{i,j,k}$ etc. For problems with variable coefficients the above dis-
 771 cretization is generalized as in [12].

772 We note that in some of the examples where we require high order accuracy we use the sum-
 773 mation by parts discretization for variable coefficients developed by Mattson in [86] and described
 774 in detail there.

775 2.2.3 Time Discretization

776 In most of the numerical examples we use either an explicit second order accurate centered
 777 discretization of w_{tt} (for finite differences with energy conserving boundary conditions we eliminate
 778 v and time discretize w_{tt} directly as in the analysis in Section 2.1.2) or the classic fourth order
 779 accurate explicit Runge-Kutta method.

780 For some of the DG discretizations we employ Taylor series time-stepping in order to match
 781 the order of accuracy in space and time. Assuming that all the degrees of freedom have been
 782 assembled into a vector \mathbf{w} we can write the semi-discrete method as $\mathbf{w}_t = Q\mathbf{w}$ with Q being a
 783 matrix representing the spatial discretization. Assuming we know the discrete solution at the time

⁷⁸⁴ t_n we can advance it to the next time step $t_{n+1} = t_n + \Delta t$ by the simple formula

$$\begin{aligned}\mathbf{w}(t_n + \Delta t) &= \mathbf{w}(t_n) + \Delta t \mathbf{w}_t(t_n) + \frac{(\Delta t)^2}{2!} \mathbf{w}_{tt}(t_n) \dots \\ &= \mathbf{w}(t_n) + \Delta t Q \mathbf{w}(t_n) + \frac{(\Delta t)^2}{2!} Q^2 \mathbf{w}(t_n) \dots\end{aligned}$$

⁷⁸⁵ The stability domain of the Taylor series which truncates at time derivative number N_T includes
⁷⁸⁶ part of the imaginary axis if $\text{mod}(N_T, 4) = 3$ or $\text{mod}(N_T, 4) = 0$ (see e.g. [72]). However as we use
⁷⁸⁷ a slightly dissipative spatial discretization the spectrum of our discrete operator will be contained
⁷⁸⁸ in the stability domain of all sufficiently large choices of N_T (i.e. the N_T should not be smaller
⁷⁸⁹ than the spatial order of approximation). Note also that the stability domain grows linearly with
⁷⁹⁰ the number of terms.

⁷⁹¹ 2.3 Numerical Examples

⁷⁹² In this section we illustrate the properties of the proposed iteration and its Krylov accelerated
⁷⁹³ version by a sequence of numerical experiments in one, two and three dimensions.

⁷⁹⁴ 2.3.1 Examples in One Dimension

⁷⁹⁵ We begin by presenting some very basic numerical experiments in one dimension.

⁷⁹⁶ 2.3.1.1 Convergence of Different Iterations / Solvers at a Fixed Frequency

⁷⁹⁷ We start by repeating the example described in Section 3.5 in [61]. This example is used in
⁷⁹⁸ [61] to illustrate that the original cost functional from [29] (denoted J in [61]) does not yield the
⁷⁹⁹ correct solution due to the existence of multiple minimizers.

⁸⁰⁰ The example solves the Helmholtz equation with $c = 1$ and with the exact solution

$$u(x) = 16x^2(x - 1)^2, \quad 0 \leq x \leq 1.$$

⁸⁰¹ Here both u (and w_t for the time-dependent problem) and u_x vanish at the endpoints so any
⁸⁰² boundary condition of the form

$$\alpha w_t + \beta(\vec{n} \cdot w_x) = 0, \quad \alpha^2 + \beta^2 = 1,$$

will be satisfied. Dirichlet boundary conditions correspond to $\alpha = 1$ and Neumann boundary conditions correspond to $\alpha = 0$, all other values will be an impedance boundary condition. Here, as in [61], we take the frequency to be $\omega = \pi/4$.

We discretize using the energy based DG method discussed above and use upwind fluxes which adds a small amount of dissipation. For this experiment we use 5 elements with degree $q = 7$ polynomials and we use an 8th order accurate Taylor series method in time. We set Δt so that $n_t \Delta t = T = 2\pi/\omega$ while making the inequality $\Delta t \leq C_{\text{CFL}} \Delta x / (q + 1)$ as sharp as possible (in this experiment we fix $C_{\text{CFL}} = 1/2$). With this resolution in space and time the truncation errors are negligible and we expect that the observed convergence properties should match those of the continuous analysis.

As mentioned above we expect that our method works best when combined with a classical iterative Krylov subspace method. The energy based DG method will produce a matrix A with real eigenvalues in $(0, 3/2)$ but it will not yield a symmetric matrix A . We present results for the WaveHoltz iteration (denoted WHI in figures and tables), and its acceleration with Matlab implementations of LSQR, QMR, CG and GMRES (we use the default unconditioned settings with a tolerance of 10^{-13}). In Figure 2.2 we display the convergence histories for various combinations of boundary conditions.

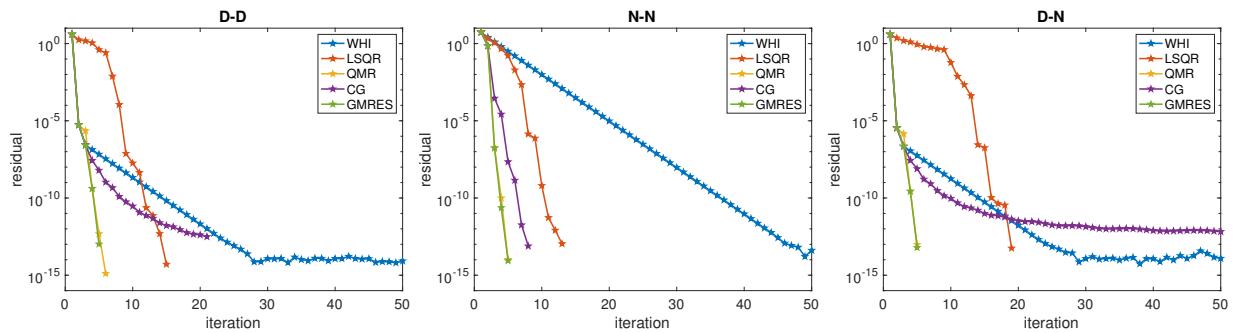


Figure 2.2: Convergence of the residual for the plain WaveHoltz iteration and its accelerated versions using LSQR, QMR, CG and GMRES. The titles of the figures indicate the boundary conditions used to the left and right, e.g. D-N means Dirichlet on the left and Neumann on the right.

820 The residuals for the Krylov accelerated iterations are the ones returned by the Matlab functions
 821 and the residual for the WaveHoltz iteration is simply the L_2 norm of the difference between
 822 two subsequent iterations. As can be seen the convergence behavior for QMR and GMRES are
 823 uniformly the fastest and appears to be insensitive to the type of boundary condition used. Note
 824 that the numerical method used here does not yield a symmetric matrix and CG is not guaranteed
 825 to work. Evidence of this loss or stagnation of convergence can be found in the cases D-D and D-N
 826 in Figure 2.2.

827 The actual errors in the converged solutions can be found in Table 2.1, where it can be seen
 828 that the error for all of the iteration methods are close to the residual tolerance.

Method / b. c.	WHI	LSQR	QMR	CG	GMRES
D-D	94.5(-15)	76.1(-15)	75.9(-15)	151.5(-15)	97.9(-15)
N-N	49.2(-15)	142.8(-15)	144.4(-15)	158.5(-15)	144.1(-15)
D-N	28.3(-15)	55.4(-15)	81.9(-15)	272.3(-15)	67.0(-15)

Table 2.1: Maximum error for various combinations of boundary conditions and methods.

829 **2.3.1.2 Convergence with Increasing Frequency**

830 To study how the number of iterations scale with the Helmholtz frequency ω we solve the
 831 wave equation on the domain $x \in [-6, 6]$ with constant wave speed $c^2(x) = 1$ and with a forcing

$$f(x) = \omega^2 e^{-(\omega x)^2},$$

832 that results in the solution being $\mathcal{O}(1)$ for all ω . The solver is the same as in the previous example.
 833 We keep the number of degrees of freedom per wave length fixed by letting the number of elements
 834 be $5\lceil\omega\rceil$. We always take the polynomial degree to be 7 and the number of Taylor series terms in
 835 the timestepping to be 8. As we now also consider impedance boundary conditions, with $\alpha = 1/2$,
 836 we use WHI accelerated by GMRES.

837 We report the number of iterations it takes to reach a GMRES residual smaller than 10^{-10}
 838 for the six possible combinations of Dirichlet, Neumann and impedance boundary conditions for 50

839 frequencies distributed evenly between 1 and 100. The results are displayed to the left in Figure
 840 2.3 where we plot the number of iterations divided by ω as a function of ω .

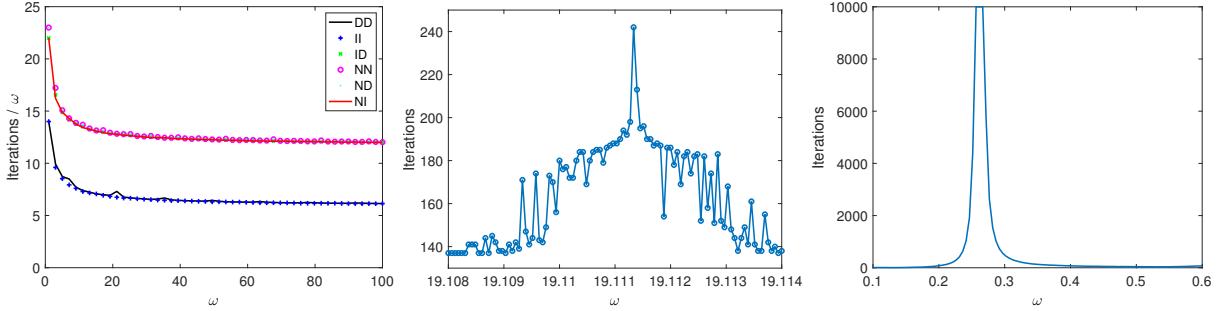


Figure 2.3: Left: Number of iterations divided by ω as a function of ω for different boundary conditions. Middle and right: Zoom in around a resonance for the Dirichlet problem when using Krylov acceleration (middle) and when using WHI (right).

841 It is clear that the asymptotic scaling is linear with growing frequency. Interestingly all the com-
 842 binations of boundary conditions collapse to two different curves with the Dirichlet-Dirichlet and
 843 impedance-impedance conditions converging the fastest.

844 We know from the analysis in Section 2.1.1 that the rate of convergence of the WaveHoltz
 845 iteration deteriorates near resonant frequencies (for non-impedance problems) but from Figure
 846 2.3 it appears that all frequencies converge more or less the same rate. To study the behavior
 847 of the accelerated algorithm for homogenous Dirichlet boundary conditions we zoom in around
 848 $\omega \approx 19.114$ where the continuous problem has a resonance. In the middle graph in Figure 2.3 we
 849 display the required number of iterations around the resonant frequency. As can be seen there
 850 is some deterioration but only in very narrow band around a frequency that is slightly less than
 851 19.114 and probably is the modified resonant frequency discussed in Section 2.1.2. This behavior
 852 can be contrasted to the growth of the number of iterations for the WaveHoltz iteration without
 853 GMRES acceleration, see the right figure in Figure 2.3. Clearly the acceleration of the WaveHoltz
 854 iteration by GMRES improves the robustness of the method near resonances.

855 **2.3.1.3 Multiple Frequencies in One Solve**

856 Here we illustrate the technique described in Section 2.1.4 for finding solutions of multiple
 857 frequencies at once. We set $\omega = 1$ and $\omega_j = 2^{j-1}\omega$ for $j = 1, \dots, 4$, and consider the domain
 858 $x \in [0, 1]$. We use the finite difference discretization discussed in Section 2.2.2 with Dirichlet
 859 boundary conditions . The time evolution is done by a second order centered discretization of w_{tt} ,
 860 as was done in the discrete analysis in Section 2.1.2 . The problem is forced by a point source
 861 centered at $x = 1/2$ for $j = 1, \dots, 4$, and we consider a constant wave speed $c^2(x) = 1$. We
 862 display the convergence with decreasing h in Figure 2.4 on the right, where it can be seen that
 863 each solution u_j converges at a rate of h^2 .

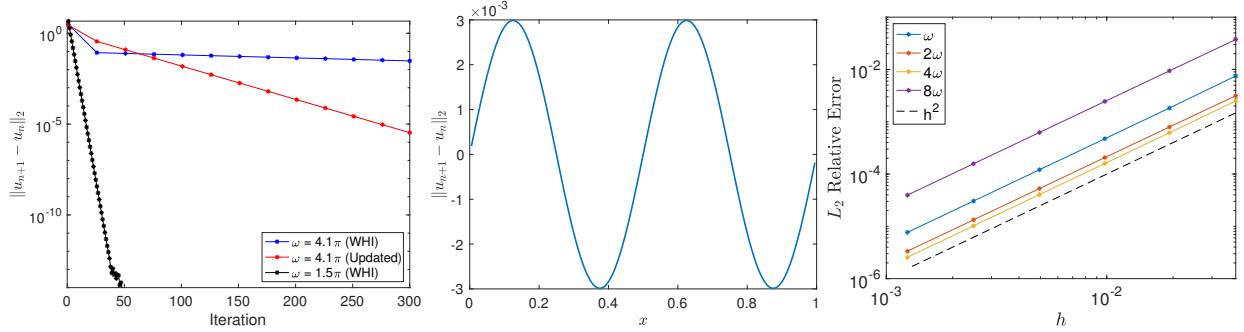


Figure 2.4: Left: Convergence history of the near resonant frequency 4.1π for the WaveHoltz filter and a tunable filter, and that of the frequency 1.5π for reference. Middle: The error between successive WaveHoltz iterates with the usual WaveHoltz filter. Right: Convergence of the solution for the CG accelerated WaveHoltz iteration with a point forcing.

864 **2.3.1.4 Tunable Filters**

865 Here we consider solving a Helmholtz problem in the domain $x \in [0, 1]$ with Dirichlet bound-
 866 ary conditions and constant wave speed $c^2 = 1$. The discretization is the same as in the previous
 867 experiment and we use a point source centered at $x = 1/2$. A straightforward calculation shows
 868 that the resonant frequencies of the problem are integer multiples of π and we specifically consider
 869 solving the Helmholtz problem with frequency $\omega = 4.1\pi$, which has a minimum relative gap to

resonance of $\delta = 1/41 \approx 0.024$. As discussed previously, we expect that the convergence rate of the WaveHoltz iteration will stagnate since ω is close to resonance. We compare the convergence against the problem with frequency $\omega = 1.5\pi$ which has a minimum relative gap to resonance of $\delta = 1/3$. The iteration history is displayed in Figure 2.4 on the left . It can be seen that the usual WaveHoltz iteration converges rapidly for the frequency 1.5π but that of 4.1π stagnates considerably. In the middle of Figure 2.4 we display the difference between successive WaveHoltz iterates for the Helmholtz problem with frequency ω , from which it is clear that the residual is a scaling of the resonant mode $\sin(4\pi x)$.

To improve the rate of convergence close to resonance we leverage a tunable filter as mentioned in Section 2.1.3. To obtain this filter, we consider the filter transfer function (2.21) and truncate the sin expansion of the time-dependent shift $\alpha(t)$ to 12 terms such that $a_n = 0$ for $n > 11$. In this example we take the usual choice of $a_0 = -1/4$ for the constant term in the filter transfer function (2.21) which, as discussed in Section 2.1.3, requires $a_1 = 0$. We then perform a minimization over a discrete set of 3000 equispaced points $r_j \in [0, 16\pi]$ of the empirically constructed functional

$$J(a_2, a_3, \dots, a_{11}) = 10.6\bar{\beta}''(\omega) + 0.1 \sum_{|r_j - \omega| > 0.1} |\bar{\beta}(r_j)|^{20}, \quad (2.34)$$

via 100 steepest descent iterations. The first term in the functional (2.34) minimizes the second derivative at the peak $\omega = 4.1\pi$, while the second weakly enforces that $|\beta(r)| \leq 1$ for all $r > 0$ to ensure convergence of the fixed point iteration.

In Figure 2.5 on the right we see that the updated filter is steeper near $\omega = 4.1\pi$ so that repeated application of the updated filter will more quickly remove the resonant mode with frequency 4π and we thus expect faster convergence.

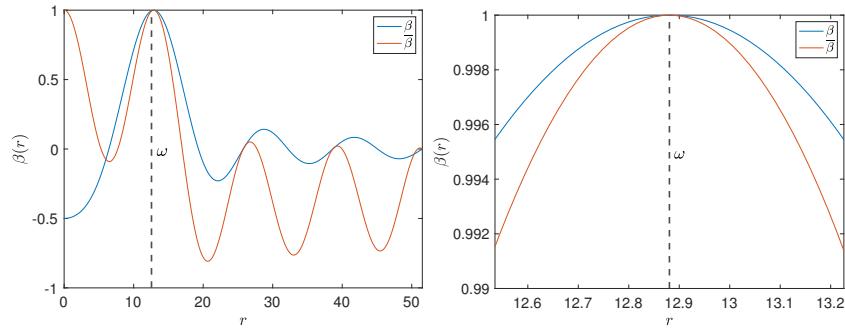


Figure 2.5: (Left) The usual WaveHoltz filter (in blue) and updated tunable filter (in red). (Right) Closeup of both the usual WaveHoltz filter and the updated tunable filter near the resonant frequency 4π .

884 This is confirmed in the resulting iteration history of the updated filter, shown in Figure 2.4 on the
 885 left . The cost of improving convergence behavior near resonance, however, is a larger value of $\bar{\beta}$
 886 for many other modes as shown in Figure 2.5 on the left. A more careful investigation of optimized
 887 filters is left for the future.

888 **2.3.2 Problems in Two Dimensions**

889 In this section we present experiments in two space dimensions.

890 **2.3.2.1 Convergence in Different Geometries**

891 In this example we solve the Helmholtz equation with a constant wave speed, $c^2 = 1$, in the
 892 domain $(x, y) \in [-1, 1]^2$ and with forcing

$$f(x, y) = -\omega^2 e^{-\sigma[(x-0.01)^2 + (y-0.015)^2]},$$

893 where $\sigma = \max(36, \omega^2)$. We vary the frequency according to $\omega = 1/2 + k$, $k = 1, \dots, 100$, and
 894 keep the number of points per wavelength roughly constant by choosing $n_x = n_y = 8\lceil\omega\rceil$. Here we
 895 use the finite difference method outlined in Section 2.2.2 combined with the classic fourth order
 896 Runge-Kutta method in time with a timestep $\Delta t = h_x/c$.

897 For each frequency we solve six different problems consisting of combinations of Dirichlet and
 898 impedance boundary conditions with zero to four open sides and with the two open boundary case
 899 forking into two cases: (1) the open boundaries are opposite each other, or (2) next to each other
 900 forming a corner. In Figure 2.6 we display the real part of the solution for the frequency $\omega = 77.5$
 901 for the six different problems.

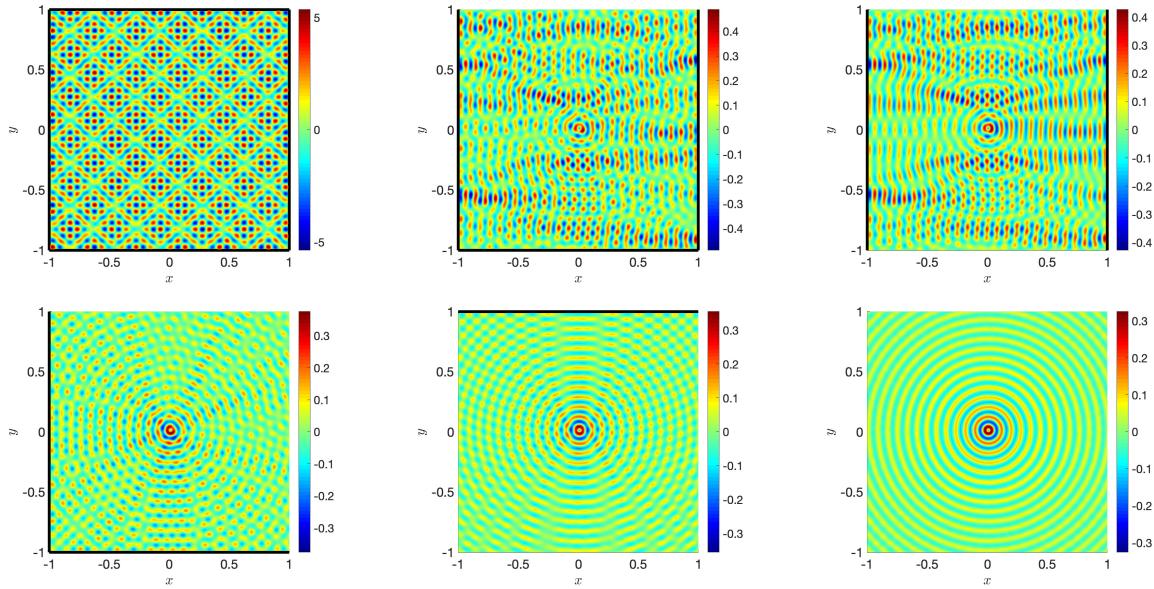


Figure 2.6: Typical solutions computed with the GMRES accelerated WHI at $\omega = 77.5$. The thick lines indicate Dirichlet boundary conditions.

902 In this example, the WaveHoltz iteration is accelerated by GMRES without restarts. Given
 903 that the storage requirement for GMRES grows with the number of iterations, it is often beneficial
 904 (especially for high frequency problems) to integrate and average over several periods to allow
 905 further propagation of information within the domain while mitigating the rapid growth of the
 906 Krylov subspace. For this example we thus choose to perform the WaveHoltz iteration with an
 907 integration time of 10 periods (i.e. we choose $T = 10\frac{2\pi}{\omega}$). In Figure 2.7 we report the number of
 908 iterations needed to reduce the relative residual below 10^{-7} .

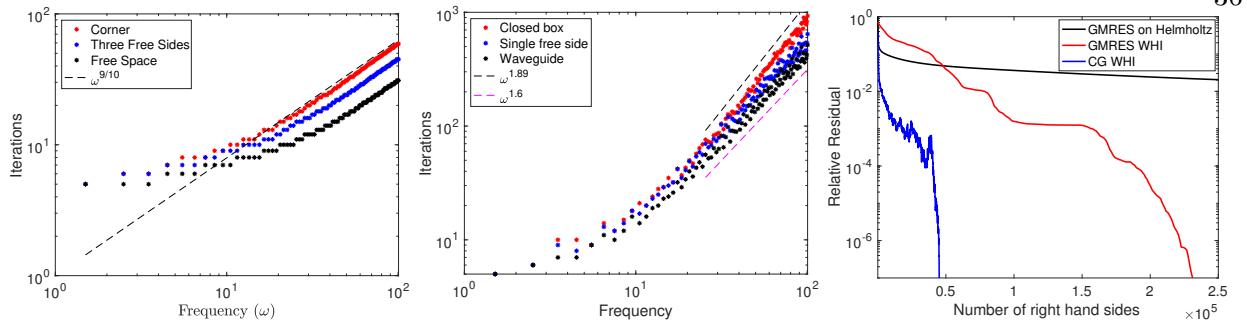


Figure 2.7: To the left: number of iterations as a function of frequency to reduce the relative residual below 10^{-7} for problems with no trapped waves. Middle: the same but for problems with trapped waves and for the interior problem. Both are with the GMRES accelerated WHI . To the right: Residuals for the GMRES accelerated WHI , the CG accelerated WHI and for GMRES solution of the directly discretized Helmholtz problem.

909 It is clear from the results that the geometries where the waves can get trapped are considerably
 910 more difficult and require more iterations. The computational results appear to indicate that the
 911 number of iterations to reach the tolerance scale as $\omega^{1.89}$ for the inner Dirichlet problem and as
 912 $\omega^{1.6}$ for the waveguide and the case with three Dirichlet boundary conditions. On the other hand
 913 for geometries with no trapped waves we see faster convergence (see the left figure in Figure 2.7)
 914 with the number of iterations scaling roughly as $\omega^{9/10}$.

915 To the right in Figure 2.7 we display the residual as a function of the number of right hand
 916 side evaluations (for the wave equation this is equivalent to taking a timestep and for the direct
 917 discretization of Helmholtz this is equivalent to one application of the sparse system matrix, the cost
 918 of these are roughly equivalent) when $\omega = 51.5$ for the pure Dirichlet boundary condition problem.
 919 The three different results are for: 1. the WaveHoltz acceleration, 2. the WaveHoltz iteration
 920 accelerated with conjugate gradient and based on the same spatial discretization but with a second
 921 order accurate centered discretization of w_{tt} using $\Delta t = 0.7h_x$ and, 3. a direct discretization of
 922 the Helmholtz equation (using the spatial discretization described in Section 2.2.2) combined with
 923 GMRES for solving the resulting system of equations. Precisely we use GMRES with restart every

924 100 iterations. For space reasons we only display this for one frequency but note that although the
 925 results may differ a bit between frequencies the trend is similar in the problems we have investigated.

926 It is clear from the residuals that both the GMRES and conjugate gradient accelerated
 927 WaveHoltz iterations are radically faster than applying GMRES to the direct discretization of
 928 Helmholtz. As all the methods use the same spatial discretization this is an indication of the
 929 importance of changing the problem from an indefinite system of equations to a positive definite
 930 and to a symmetric positive definite system.

931 **Remark 2.3.1.** *We note that the problems considered in this experiment can be naturally solved
 932 with integral equation techniques since the the wave speed is constant. In addition as the problem is
 933 posed in two dimensions and can be stored in memory a good sparse solver will also be a very good
 934 alternative. What we want to demonstrate is: 1. The positive definiteness of the accelerated WHI
 935 makes it faster than standard iterative techniques for the direct discretization of Helmholtz, 2. The
 936 complexity is different for open and closed problems as predicted by the theory in [42].*

937 **Remark 2.3.2.** *The middle plot of Figure 2.7 shown here is an updated plot of the published version
 938 found in Figure 4.6 of [14]. We note that the original plot in [14] was computed with a GMRES
 939 implementation that was sensitive to loss of orthogonality in the Arnoldi process, and here we use
 940 a more robust routine for computing the orthonormal basis of the Krylov subspace.*

941 2.3.2.2 Smoothly Varying Wave Speed in an Open Domain

942 In this example we consider a smoothly varying medium in a box $(x, y) \in [-1, 1]^2$. The wave
 943 speed is

$$c^2(x, y) = 1 - 0.4e^{-\left(\frac{x^2+y^2}{0.25^2}\right)^4},$$

944 and is also depicted in Figure 2.8.

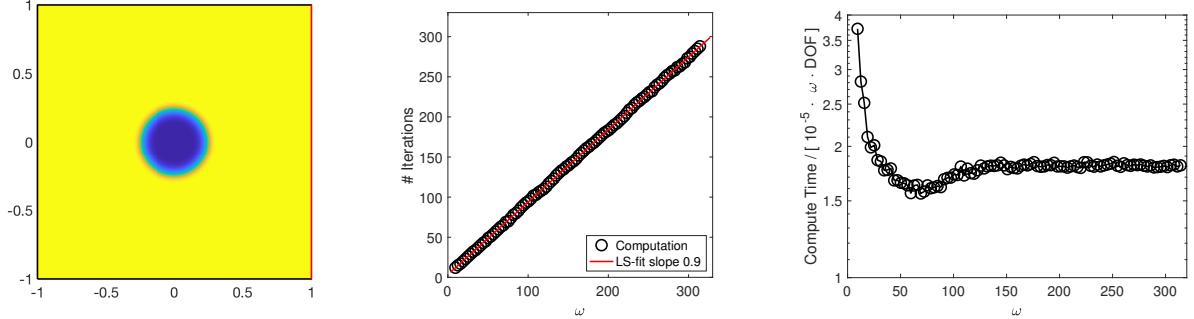


Figure 2.8: Left: the speed of sound (squared) used in example 2.3.2.2. Red indicates a rigid wall and black indicates open walls. Middle: Number of iterations as a function of frequency. Right: Compute time normalized by the frequency times the number of degrees of freedom.

945 Here we use the energy based DG solver and impose a right going plane wave $e^{i\omega(t-x)}$ through
 946 impedance boundary conditions on the left, bottom and top faces of the domain. On the right
 947 boundary we impose a zero Dirichlet condition. In all the computations we use degree 5 polynomials
 948 and a 6th order Taylor series method. The elements used form a Cartesian structured grid and
 949 we scale the number of elements so that we have 8 degrees of freedom per wavelength. The WHI
 950 is applied with an integration time of 5 periods and is accelerated by GMRES with a termination
 951 tolerance 10^{-7} on the relative residual.

952 We solve the Helmholtz problem with $\omega = k\pi, k = 3, 4, \dots, 100$ and measure the total time
 953 from start to time of solution and we also measure the number of iterations needed to converge.
 954 The results, displayed in Figure 2.8, again show that for this type of open problem the iteration
 955 appears to require $N_{\text{iter}} \sim \mathcal{O}(\omega^{0.9})$ iterations to converge to a fixed tolerance. In terms of total
 956 computational time we observe $T_{\text{Total}} \sim \mathcal{O}(\omega N_{\text{DOF}})$ which is slightly higher than what would be
 957 expected from the $\mathcal{O}(\omega^{0.9})$ behavior.

958 However, as the distance traveled by the wave solution is proportional to $cT = 2\pi c/\omega$ and
 959 the information must travel through the domain at least once the time to solution is as good as
 960 can be expected. To reduce the computational complexity further we would need to propagate the
 961 solution faster than the speed of sound by applying a preconditioner or some type of multi-level

strategy. Although we believe this is possible we leave such attempts to future work.

The magnitude of the solutions with $\omega = 25\pi, 50\pi$ and 100π are plotted in Figure 2.9.

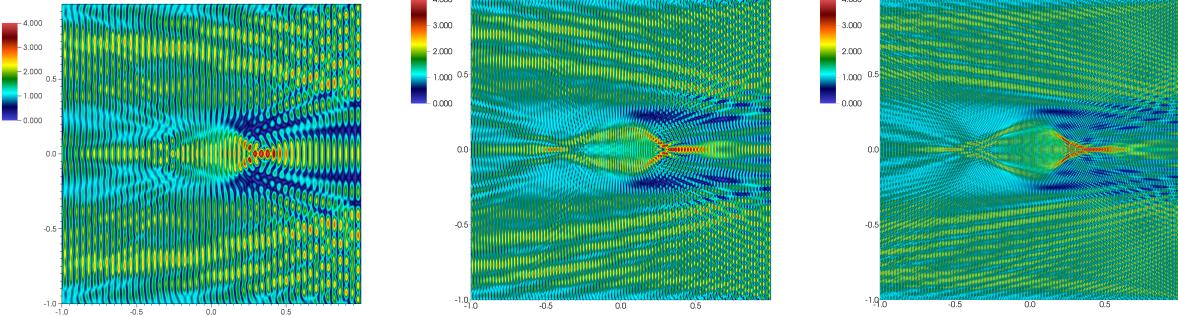


Figure 2.9: The magnitude of the Helmholtz solution for, from left to right, $\omega = 25\pi, 50\pi$ and 100π .

2.3.2.3 Convergence of the Approximation Error and the Residual

As our iteration leads to a linear system of equations (and consequently a different residual) we should check that the residual is still a suitable proxy for the discretization error. Although we have no reason to believe this would not be the case we note that we have not yet performed a detailed analysis and resort to checking this numerically. We consider the same computational domain and method as above but with speed of sound $c = 1$ and with zero Dirichlet boundary conditions. We set $\omega = 2$ and choose the forcing so that the solution is

$$u = -(x^2 - 1)^2(y^2 - 1)^2,$$

and compute the solution using polynomials of degree three in the energy DG method and a fourth order accurate Taylor time stepper. In Figure 2.10 we display the maximum errors in u and the residuals for each GMRES iteration for Cartesian grids with grid spacings $1/2, 1/4, 1/8$ and $1/16$.

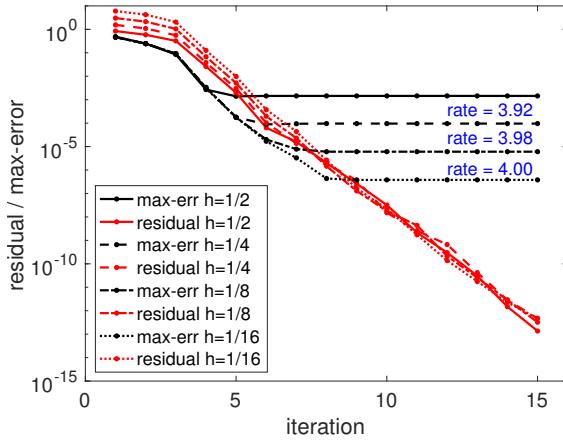


Figure 2.10: The maximum error GMRES residuals as a function of number of iterations for four different mesh sizes. The rates of convergence agree with the order of the method.

As can be seen the residuals and the errors track well until the errors saturate. To the right in the figure we also indicate the rates of convergence based on the subsequent grid refinements. As expected they are very close to four.

2.3.2.4 The Marmousi2 Model

In the last two examples in this section we use the sixth order summation-by-parts finite difference operators developed by Mattson in [86]. Here we use the classic fourth order Runge-Kutta method for timestepping. In this example we simulate the solution caused by a point source placed in a material model where the speed of sound is taken from P-wave velocity in the Marmousi2 model¹. We discretize the full model which consists of 13601×2801 grid points and covers a domain that is roughly 17×3.5 kilometers. On the top surface we prescribe a zero Dirichlet condition and on the remaining three sides we add a 50 grid point wide supergrid layer (see [8]) that is terminated by zero Dirichlet boundary conditions. We accelerate the WHI by the transpose free quasi minimal residual (TFQMR) method and terminate the iteration when the relative residual is below 10^{-5} . We perform each iteration over 8 periods and take 500 timesteps per iteration. The time periodic point forcing is applied near the surface in grid point (6750, 2600) and we perform computations

¹ <http://www.agl.uh.edu/downloads/downloads.htm>

989 with $\omega = 200, 400$ and 800 .

990 As the number of unknowns is relatively large, $\sim 76 \cdot 10^6$, we parallelize the finite difference
991 solver by a straightforward domain decomposition with the communication handled by MPI. The
992 simulations were carried out on Maneframe II at the Center for Scientific Computation at Southern
993 Methodist University using 60 dual Intel Xeon E5-2695v4 2.1 GHz 18-core Broadwell processors
994 with 45 MB of cache each and 256 GB of DDR4-2400 memory. The results displayed in Figure
995 2.11 and 2.12 illustrate the ability of the method to find solutions to large problems and at high
996 frequencies.

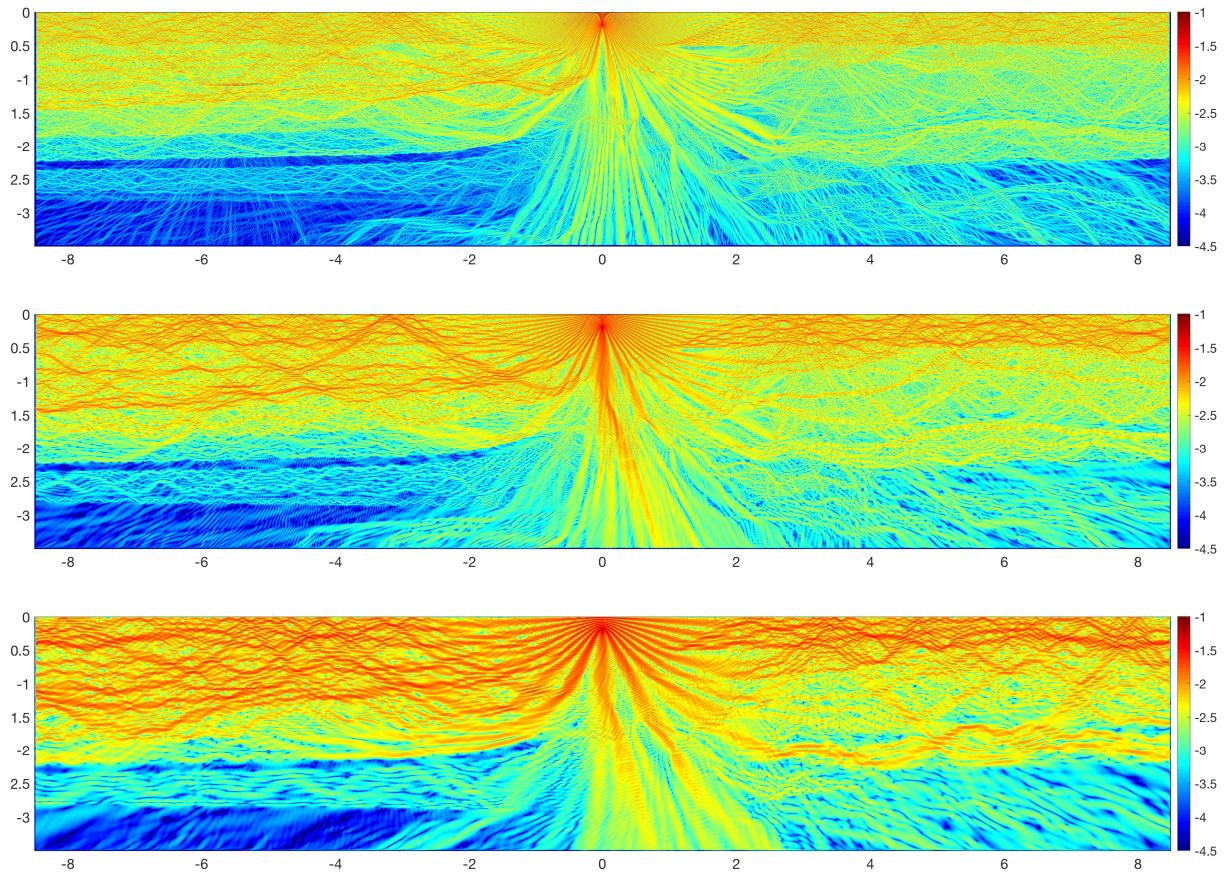


Figure 2.11: Displayed is the base 10 logarithm of the magnitude of the Helmholtz solution ($\log_{10} |u|$) caused by a point source near the surface. The results are, from top to bottom, for $\omega = 800, 400$ and 200 .

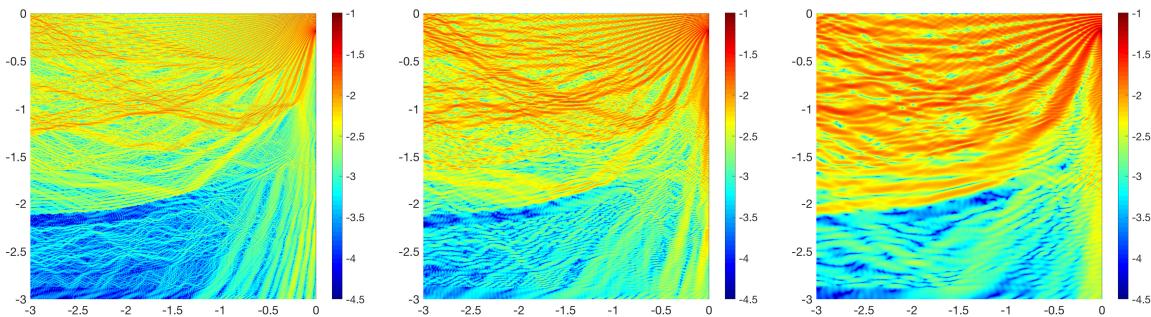


Figure 2.12: Zoom in of the base 10 logarithm of the magnitude of the Helmholtz solution ($\log_{10} |u|$) caused by a point source near the surface. The results are, from left to right, for $\omega = 800, 400$ and 200 .

⁹⁹⁷ **2.3.2.5 Multiple Frequencies**

In this final example in two dimensions we again use the sixth order accurate summation-by-parts discretization from [86] with homogenous Dirichlet boundary conditions on the domain $(x, y) \in [-1, 1]^2$. The spatial discretization size is the same in both coordinates and is taken to be 2/300. The velocity model is taken to be smoothly varying. Precisely we have that

$$c^2(x, y) = 1 - 0.9 \left(e^{-\left(\frac{(x^2 + (y-0.4)^2 - 0.4^2)}{0.2^2}\right)^4} + e^{-\left(\frac{(x^2 + (y+0.4)^2 - 0.3^2)}{0.2^2}\right)^4} \right),$$

⁹⁹⁸ see also Figure 2.13.

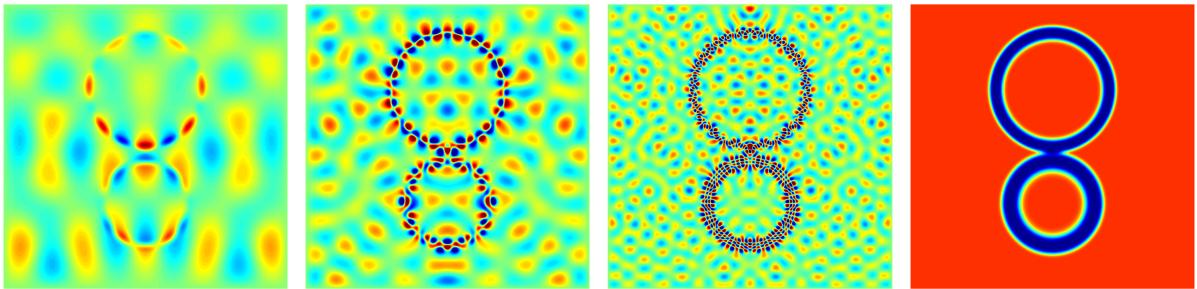


Figure 2.13: Computation of three Helmholtz problems by one solve. The frequencies are $\omega = 15, 30$ and 60 . The material model is also displayed, red is $c^2 = 1$ and dark blue is $c^2 = 0.1$.

We consider three frequencies, $\omega = 15, 30, 60$, and use the same forcing in Helmholtz for all frequencies,

$$f(x, y) = \frac{\sigma}{\pi} e^{-\sigma(x^2+y^2)}, \quad \sigma = (4\omega)^2.$$

⁹⁹⁹ Here we use the WaveHoltz iteration over three periods of the lowest frequency, accelerated by
¹⁰⁰⁰ GMRES (with tolerance 10^{-8}). We time step using a centered second order approximation to w_{tt}
¹⁰⁰¹ with a timestep $\Delta t = 1/600$. Since we solve for three frequencies at once we adjust the filter as
¹⁰⁰² described in Section 2.1.3 and extract all three solutions at once. Those solutions along with the
¹⁰⁰³ material model are displayed in Figure 2.13.

1004 **2.3.3 Problems in Three Dimensions**

1005 In this section we present experiments in three dimensions.

1006 **2.3.3.1 Convergence in Different Geometries**

1007 We solve the wave equation in a box $(x, y, z) \in [-1, 1]^3$ with the smoothly varying medium

$$c^2(x, y, z) = 1 + \frac{1}{10}e^{-(x^2+y^2+z^2)}.$$

1008 We use a uniform grid $(x_i, y_j, z_k) = (-1 + ih, -1 + jh, -1 + kh)$ with grid spacing $h = 1/n$ and
 1009 choose $n = \max(\lceil 10\omega \rceil, 20)$ to keep the resolution fixed. The Helmholtz problem is forced by

$$F(x, y, z) = \omega^3 e^{-36\omega^2((x-x_0)^2+(y-y_0)^2+(z-z_0)^2)}, \quad (2.35)$$

1010 where $x_0 = 1/100$, $y_0 = 3/250$, and $z_0 = 1/200$. We impose a mixture of boundary conditions
 1011 consisting of homogenous Dirichlet and/or impedance boundary conditions: (1) impedance on all
 1012 sides, (2) Dirichlet at $z = 1$ and impedance on all other sides, (3) Dirichlet at $z = -1$, $y = 1$, and
 1013 $x = 1$ with impedance on all other sides, and (4) Dirichlet on all sides. We solve the equations in
 1014 first order form in time and use the semi-discrete approximation described in Section 2.2.2.

1015 In this example the WaveHoltz iteration is performed over 5 periods, numerically integrated
 1016 in time with the classic Runge Kutta method of order four, and accelerated by TFQMR method.
 1017 For the pure Dirichlet problem we also use CG but note that although the spatial discretization
 1018 leads to a symmetric WHI matrix when combined with a centered finite difference approximation in
 1019 time the matrix is only close to symmetric when combined with the slightly dissipative Runge Kutta
 1020 method. The experiments indicate that this slight non-symmetry does not destroy the convergence
 1021 iteration of CG in this case.

1022 In Figure 2.14 we report the number of iterations needed to reduce the relative residual below
 1023 $5 \cdot 10^{-5}$.

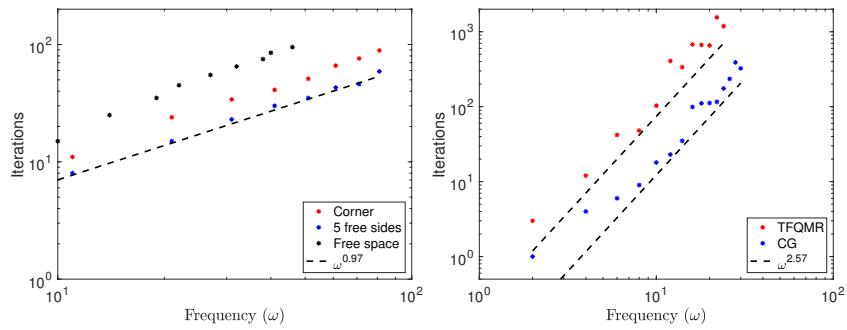


Figure 2.14: To the left: number of iterations as a function of frequency to reduce the relative residual below $5 \cdot 10^{-5}$ for problems with no trapped waves. Here WHI is accelerated by TFQMR. To the right: the same but for the interior problem. Here WHI is accelerated with either CG or TFQMR.

1024 As was seen before in the 2D case, the fully Dirichlet case is notably more difficult and requires
 1025 more iterations than the other problems considered. The computational results indicate that the
 1026 number of iterations to reach the tolerance scale as $\omega^{2.57}$ for the inner Dirichlet problem with either
 1027 CG or TFQMR, with the former taking fewer overall iterations than the latter. By comparison,
 1028 the set of problems with boundary conditions (1)-(3) listed above appear to converge in a number
 1029 of iterations that scales as $\omega^{0.97}$, i.e. close to linear in the frequency ω .

1030 2.3.3.2 Scattering from a Plate

1031 For our final example, we again consider the box $(x, y, z) \in [-1, 1]^3$ with smoothly varying
 1032 medium

$$c^2(x, y, z) = \frac{1}{2} [3 + \sin(16\pi z) \sin(4\pi(x + y))],$$

1033 and impose Dirichlet boundary conditions at $z = 1$ and impedance boundary conditions on all
 1034 other sides. We use a uniform grid $(x_i, y_j, z_k) = (-1 + ih, -1 + jh, -1 + kh)$ with grid spacing
 1035 $h = 1/n$ where n is the number of gridpoints along a single dimension. The discretization in space
 1036 and in time is exactly as in the previous example and the problem is forced by $F(x, y, z)$ as in
 1037 (2.35) with $x_0 = 1/100$, $y_0 = 3/250$, and $z_0 = 4/5$. As in the Marmousi example in the previous

1038 section, we parallelize the finite difference solver by a straightforward domain decomposition with
 1039 the communication handled by MPI. This simulation was carried out on Maneframe II at the Center
 1040 for Scientific Computation at Southern Methodist University using 64 dual Intel Xeon E5-2695v4
 1041 2.1 GHz 18-core Broadwell processors with 45 MB of cache each and 256 GB of DDR4-2400 memory.
 1042 The magnitude of the solution with $\omega = 200$ and $\omega = 300$ is plotted in Figure 2.15.

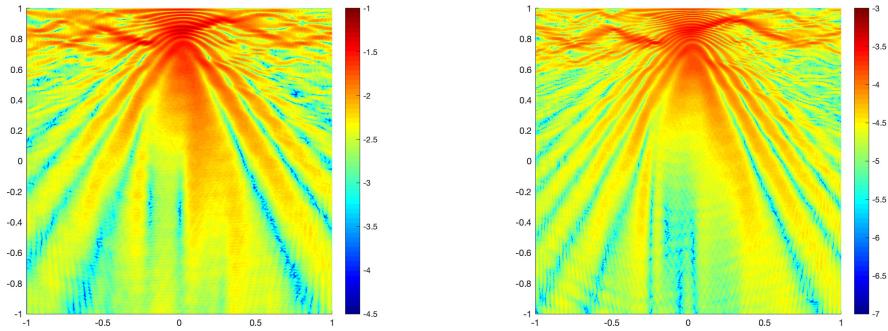


Figure 2.15: Displayed is the base 10 logarithm of the magnitude of the Helmholtz solution ($\log_{10} |u|$) caused by a point source near the surface for $\omega = 200$ (left) and $\omega = 300$ (right) at the slice $x = 0.1$.

1043 We use $n = 1000$ for a total of 10^9 gridpoints in the first case, and $n = 1500$ for a total of $3.375 \cdot 10^9$
 1044 gridpoints in the second for roughly 15-16 points per wavelength.

1045 2.4 Summary and Future Work

1046 We have presented and analyzed the WaveHoltz iteration, a new iterative method for solving
 1047 the Helmholtz equation. The iteration results in positive definite and sometimes symmetric matrices
 1048 that are more amenable for iterative solution by Krylov subspace methods. In choosing a Krylov
 1049 subspace method we note that CG is the most efficient and memory lean choice when the resulting
 1050 system is symmetric positive definite, otherwise GMRES generally outperforms other methods such
 1051 as QMR, LSQR, and TFQMR. As the iteration is based on solving the wave equation it naturally
 1052 parallelizes and can exploit techniques and spatial discretizations that have been developed for the

1053 time dependent problem. Numerical experiments indicate that our iteration appears to converge
1054 significantly faster than when the Helmholtz equation is discretized directly and solved iteratively
1055 with GMRES.

1056 We believe that the numerical and theoretical results above are promising and note that
1057 there are many possible avenues for future exploration. For example we have exclusively used
1058 unconditioned Krylov solvers here but the spectral properties of the operator \mathcal{S} indicate that
1059 preconditioning should be possible. Further, we have not tried to exploit adaptivity in space or
1060 time or any type of sweeping ideas here and we have only briefly touched on the possibilities for
1061 more advanced filter design. We hope to study both the numerical and theoretical properties of
1062 these in the future.

1063 Finally, here we only analyzed the energy conserving problem. In the following chapter we
1064 will analyze problems with energy loss, via either damping or impedance boundary conditions.

1065

Chapter 3

1066 **Analysis of an Iterative Solution of the Helmholtz Equation via the Wave
1067 Equation for Impedance Boundary Conditions**

1068 In this chapter, we continue analyzing time-domain methods for the numerical solution of
1069 the Helmholtz equation

$$\nabla \cdot (c^2(x) \nabla u) + \omega^2 u = f(x), \quad x \in \Omega, \quad (3.1)$$

1070 for a domain Ω , frequency ω , and sound speed $c^2(x)$. The Helmholtz equation (both acoustic
1071 and elastic) is useful for seismic, acoustic, and optics applications. The numerical solution of the
1072 Helmholtz equation is especially difficult due to the resolution requirements and the indefinite
1073 nature of the Helmholtz operator for large frequencies.

1074 In the previous chapter, we introduced a time-domain approach for solving the Helmholtz
1075 equation (3.1). Given the Helmholtz solution, $u(x)$, the time-harmonic wave field $\text{Re}\{u(x)e^{-i\omega t}\}$
1076 satisfies the wave equation

$$w_{tt} = \nabla \cdot (c^2(x) \nabla w) - f(x) \cos(\omega t), \quad x \in \Omega, \quad 0 \leq t \leq T,$$

$$w(0, x) = v_0(x), \quad w_t(0, x) = v_1(x),$$

1077 where $v_0 = \text{Re}\{u(x)\}$ and $v_1 = \omega \text{Im}\{u(x)\}$. In Chapter 2, we introduced an integral operator that
1078 time-averaged the wave solution resulting from initial data v_0^n, v_1^n . The time-averaging generates
1079 new iterates v_0^{n+1}, v_1^{n+1} leading to a fixed-point iteration we named the WaveHoltz iteration. The
1080 convergence of the fixed-point iteration for interior problems with Dirichlet/Neumann boundary
1081 conditions (i.e. energy conserving problems) was proven in the continuous and discrete settings.

1082 For such problems, the WaveHoltz iteration can be reformulated as a symmetric and positive-
 1083 definite system which can be accelerated with Krylov subspace methods such as the conjugate
 1084 gradient method and GMRES. Numerical experiments using the WaveHoltz iteration indicated
 1085 promising scaling with frequency for problems with outflow boundary conditions common in seismic
 1086 applications, though no theoretical proof was given for the convergence of the method in that case.

1087 In this chapter, we extend the analysis of the previous chapter to problems with impedance
 1088 boundary conditions. In addition, we analyze the WaveHoltz iteration when applied to the damped
 1089 Helmholtz equation and prove that the iteration converges for problems with damping or
 1090 impedance conditions. Numerical results verify that for a sufficiently large damping, the num-
 1091 ber of iterations for the WaveHoltz iteration to reach convergence for damped Helmholtz equations
 1092 is independent of frequency. We thus can guarantee convergence of the method to the Helmholtz
 1093 solution via impedance conditions and/or damping.

1094 We also investigate the effect of choice of timestepper used for the WaveHoltz iteration. In
 1095 the previous chapter, we noted that in the discrete case the WaveHoltz iteration converged to the
 1096 solution of a discrete Helmholtz problem with modified frequency. We provided the modification for
 1097 a centered second-order timestepping scheme which would recover the original discrete Helmholtz
 1098 solution. Here we consider higher order modified equation (ME) timestepping schemes and show
 1099 that the fixed-point of the discrete WaveHoltz iteration converges to the discrete Helmholtz solution
 1100 with the order of the timestepper chosen. We additionally show that, as in the case for EM-
 1101 WaveHoltz [93], it is possible to *completely remove* time discretization error from the WaveHoltz
 1102 solution through careful analysis of the discrete iteration and updated quadrature formulas.

1103 The efficient solution of the Helmholtz equation (3.1) via iterative methods is incredibly
 1104 difficult, especially for high-frequency problems of practical interest, and has been the subject
 1105 of much research. We refer to the paper [14] for a more in-depth overview of the literature on
 1106 techniques for solving the Helmholtz equation, as well as the review articles [47, 50, 44]. We focus
 1107 on the literature that is closely related to the methods and approach used here.

1108 The theoretical justification for working in the time-domain comes from the *limiting amplitude*

1109 principle (see [88, 76, 104]) which states that every solution to the wave equation with a time-
1110 harmonic forcing in the exterior of a domain with reflecting boundary conditions tends to the
1111 Helmholtz solution.

1112 Rather than evolving a wave equation forward in time to reach a steady state by appealing
1113 to the limiting amplitude principle, it is possible to cast the problem as a constrained convex least-
1114 squares minimization problem. This approach, originally proposed by Bristeau et al. [29], is the
1115 so-called Controllability Method (CM) which seeks to accelerate the convergence to the steady-state
1116 limit by minimizing the deviation from time-periodicity of the time-domain solution in second-order
1117 form.

1118 In the original CM, along with later work by Heikkola et al. [68, 69], only sound-soft scatterers
1119 were considered as the original cost functional of [29] did not generally yield unique minimizers for
1120 other types of boundary conditions. An alternative functional, J_∞ , proposed by Bardos and Rauch
1121 in [18], however, did yield uniqueness of the minimizer at the cost of requiring the storage of the
1122 entire history of the computed solution to the wave equation which could be prohibitive for large
1123 problems.

1124 For the wave equation in second-order form, the initial condition lies in $H^1 \times L^2$, requiring
1125 the solution of a coercive elliptic problem to find a Riesz representative for gradient calculations.
1126 Glowinski and Rossi [58] presented an update to the CM by considering the wave equation in first-
1127 order form, allowing the initial conditions to lie in a reflexive space and thus removing the need for
1128 an elliptic solve each iteration. The discretization chosen in this case, however, had the drawback
1129 of requiring inversion of a mass-matrix at each timestep.

1130 In more recent work by Grote and Tang, [63], the use of an alternative functional (or post-
1131 processing via a compatibility condition) restored uniqueness of the minimizer of CM. In a follow-
1132 up paper, [85], Grote et al. proposed a HDG discretization of the first-order form wave equation
1133 which allowed the scheme to be fully explicit and therefore fully parallel. Moreover, they extend
1134 CM to general boundary conditions for the first-order formulation and additionally proposed a
1135 filtering procedure which allows the original energy functional to be used regardless of the boundary

1136 condition.

1137 The above work has inspired other time-domain methods outside of CM and WaveHoltz.
1138 Work by Stolk [101] leverages time-domain approaches as a preconditioner for a GMRES accelerated
1139 preconditioner for direct Helmholtz discretizations yielding a hybrid time-frequency domain
1140 method. Arnold et al. [15] propose a time-domain method for scattering problems which leverages
1141 the compact support of incident field plane wavelets together with a front-tracking adaptive mesh-
1142 ing algorithm to reduce the cost of computing a Fourier transform of the wave solution to obtain
1143 Helmholtz solutions.

1144 Another important class of methods for solving the Helmholtz equation are the so-called
1145 shifted Laplacian preconditioners. The use of the Laplacian as a preconditioner for Helmholtz
1146 problems emerged with the initial work of Bayliss et al. [20]. In [20], the normal equations of the
1147 discrete Helmholtz equation were iteratively solved using conjugate gradient, with a Symmetric
1148 Successive Over-Relaxation (SSOR) sweep of the discrete Laplacian as a preconditioner. Giles
1149 and Laird then extended the previous preconditioner to instead solve the Helmholtz system with a
1150 flipped sign in front of the Helmholtz term using multigrid [77]. Erlangga, Vuik and Osterlee [46, 43]
1151 further generalized the previous work to use a complex-valued shift of the Laplacian leading to the
1152 shifted Laplacian preconditioner. For a review of the class of shifted Laplacian preconditioners we
1153 refer the reader to the review article by Erlangga [44].

1154 The rest of this chapter is organized as follows. In Section 2 we present analysis for the
1155 general WaveHoltz iteration and prove convergence in the case for impedance boundary conditions.
1156 In Section 3 we present a brief analysis for the case in which damping is present. Section 4 outlines
1157 a discrete analysis of higher order modified equation (ME) schemes, and we additionally present
1158 a method to *completely* remove time discretization error from the discrete WaveHoltz solution.
1159 Finally, in Section 5 we describe our numerical methods, Section 6 present our numerical examples,
1160 and summarize the chapter in Section 7.

₁₁₆₁ **3.1 The General Iteration**

₁₁₆₂ We consider the Helmholtz equation in a bounded open smooth domain Ω ,

$$\nabla \cdot (c^2(x) \nabla u) + \omega^2 u = f(x), \quad x \in \Omega, \quad (3.2)$$

₁₁₆₃ with boundary conditions of the type

$$i\alpha\omega u + \beta(c(x)\vec{n} \cdot \nabla u) = 0, \quad \alpha^2 + \beta^2 = 1, \quad x \in \partial\Omega. \quad (3.3)$$

₁₁₆₄ We assume $f \in L^2(\Omega)$ and that $c \in L^\infty(\Omega)$ with the bounds $0 < c_{\min} \leq c(x) \leq c_{\max} < \infty$ a.e. in

₁₁₆₅ Ω . Away from resonances, this ensures that there is a unique weak solution $u \in H^1(\Omega)$ to (3.2).

₁₁₆₆ Due to the boundary conditions u is in general complex-valued.

₁₁₆₇ We first note that the function $w(t, x) := \operatorname{Re}\{u(x) \exp(-i\omega t)\}$ is a $T = 2\pi/\omega$ -periodic (in
₁₁₆₈ time) solution to the real-valued forced scalar wave equation

$$\begin{aligned} w_{tt} &= \nabla \cdot (c^2(x) \nabla w) - \operatorname{Re}\{f(x)e^{-i\omega t}\}, \quad x \in \Omega, \quad 0 \leq t \leq T, \\ w(0, x) &= v_0(x), \quad w_t(0, x) = v_1(x), \\ \alpha w_t + \beta(c(x)\vec{n} \cdot \nabla w) &= 0, \quad x \in \partial\Omega, \end{aligned} \quad (3.4)$$

₁₁₆₉ where $v_0 = \operatorname{Re}\{u\}$ and $v_1 = \omega \operatorname{Im}\{u\}$. Based on this observation, our approach is to find this w
₁₁₇₀ instead of u . We could thus look for initial data v_0 and v_1 such that w is a T -periodic solution
₁₁₇₁ to (3.4). However, there may be several such w , see [63], and we therefore impose the alternative
₁₁₇₂ constraint that a certain time-average of w should equal the initial data. More precisely, we
₁₁₇₃ introduce the following operator acting on the initial data $v_0 \in H^1(\Omega)$, $v_1 \in L^2(\Omega)$,

$$\Pi \begin{bmatrix} v_0 \\ v_1 \end{bmatrix} = \frac{2}{T} \int_0^T \left(\cos(\omega t) - \frac{1}{4} \right) \begin{bmatrix} w(t, x) \\ w_t(t, x) \end{bmatrix} dt, \quad T = \frac{2\pi}{\omega}, \quad (3.5)$$

₁₁₇₄ where $w(t, x)$ and its time derivative $w_t(t, x)$ satisfies the wave equation (3.4) with initial data
₁₁₇₅ v_0 and v_1 . The result of $\Pi[v_0, v_1]^T$ can thus be seen as a filtering in time of $w(\cdot, x)$ around the
₁₁₇₆ ω -frequency. By construction, the solution u of Helmholtz now satisfies the system of equations

$$\begin{bmatrix} \operatorname{Re}\{u\} \\ \omega \operatorname{Im}\{u\} \end{bmatrix} = \Pi \begin{bmatrix} \operatorname{Re}\{u\} \\ \omega \operatorname{Im}\{u\} \end{bmatrix}. \quad (3.6)$$

1177 The WaveHoltz method then amounts to solving this system of equations with the fixed point
1178 iteration

$$\begin{bmatrix} v \\ v' \end{bmatrix}^{(n+1)} = \Pi \begin{bmatrix} v \\ v' \end{bmatrix}^{(n)}, \quad \begin{bmatrix} v \\ v' \end{bmatrix}^{(0)} \equiv 0. \quad (3.7)$$

1179 Provided this iteration converges and the solution to is unique, we obtain the Helmholtz solution
1180 as $u = \lim_{n \rightarrow \infty} v^n$.

1181 **3.1.1 Iteration for the Energy Conserving Case for the General WaveHoltz Iter-**
1182 **ation**

1183 Here we consider boundary conditions of either Dirichlet ($\beta = 0$) or Neumann ($\alpha = 0$) type
1184 in (3.4). This is typically the most difficult case for iterative Helmholtz solvers when Ω is bounded.
1185 The wave energy is preserved in time and certain ω -frequencies in Helmholtz are resonant, meaning
1186 they equal an eigenvalue of the operator $-\nabla \cdot (c^2(x)\nabla)$. Moreover, the limiting amplitude principle
1187 does not hold, and one can thus not obtain the Helmholtz solution by solving the wave equation
1188 over a long time interval.

1189 By the choice of boundary conditions the operator $-\nabla \cdot (c^2(x)\nabla)$ has a point spectrum
1190 with non-negative eigenvalues. Denote those eigenmodes $(\lambda_j^2, \phi_j(x))$. We assume that the angular
1191 frequency ω is not a resonance, i.e. $\omega^2 \neq \lambda_j^2$ for all j . The Helmholtz equation (3.1) is then
1192 wellposed.

We recall that for any $q \in L^2(\Omega)$ we can expand

$$q(x) = \sum_{j=0}^{\infty} \hat{q}_j \phi_j(x),$$

for some coefficients \hat{q}_j and

$$\|q\|_{L^2(\Omega)}^2 = \sum_{j=0}^{\infty} |\hat{q}_j|^2, \quad c_{\min}^2 \|\nabla q\|_{L^2(\Omega)}^2 \leq \sum_{j=0}^{\infty} \lambda_j^2 |\hat{q}_j|^2 \leq c_{\max}^2 \|\nabla q\|_{L^2(\Omega)}^2.$$

1193 We start by expanding the Helmholtz solution $u = u^R + iu^I$, the initial data v_0, v_1 to the wave

¹¹⁹⁴ equation (3.4), and the forcing $f = f^R + if^I$ in this way,

$$u^R(x) = \sum_{j=0}^{\infty} \hat{u}_j^R \phi_j(x), \quad v_0(x) = \sum_{j=0}^{\infty} \hat{v}_{0,j} \phi_j(x), \quad v_1(x) = \sum_{j=0}^{\infty} \hat{v}_{1,j} \phi_j(x), \quad f^R(x) = \sum_{j=0}^{\infty} \hat{f}_j^R \phi_j(x),$$

¹¹⁹⁵ with analogous expansions for the imaginary parts of u and f , u^I and f^I , respectively. Then,

$$-\lambda_j^2 \hat{u}_j^R + \omega^2 \hat{u}_j^R = \hat{f}_j^R \quad \Rightarrow \quad \hat{u}_j^R = \frac{\hat{f}_j^R}{\omega^2 - \lambda_j^2},$$

and similarly for the imaginary parts \hat{u}_j^I and \hat{f}_j^I . For the wave equation solution $w(t, x)$ with initial data $w = v_0$ and $w_t = v_1$ we have

$$w(t, x) = \sum_{j=0}^{\infty} \hat{w}_j(t) \phi_j(x),$$

$$\hat{w}_j(t) = \hat{u}_j^R [\cos(\omega t) - \cos(\lambda_j t)] + \hat{u}_j^I \left[\sin(\omega t) - \frac{\omega}{\lambda_j} \sin(\lambda_j t) \right] + \hat{v}_{0,j} \cos(\lambda_j t) + \frac{\hat{v}_{1,j}}{\lambda_j} \sin(\lambda_j t),$$

with

$$\hat{w}_0^N(t) = \hat{u}_0^R [\cos(\omega t) - 1] + \hat{u}_0^I [\sin(\omega t) - \omega t] + \hat{v}_{0,0} + \hat{v}_{1,0}t,$$

if $\lambda_0 = 0$, as is the case for Neumann boundary conditions (a special case which we denote via the superscript N in the following analysis). The filtering step then gives

$$\Pi \begin{bmatrix} v_0 \\ v_1 \end{bmatrix} = \sum_{j=0}^{\infty} \begin{bmatrix} \bar{v}_j \\ \bar{v}'_j \end{bmatrix} \phi_j(x),$$

where

$$\bar{v}_j = \hat{u}_j^R (1 - \beta(\lambda_j)) - \hat{u}_j^I \frac{\omega}{\lambda_j} \gamma(\lambda_j) + \hat{v}_{0,j} \beta(\lambda_j) + \frac{\hat{v}_{1,j}}{\lambda_j} \gamma(\lambda_j),$$

$$\bar{v}'_j = \hat{u}_j^R \lambda_j \gamma(\lambda_j) + \omega \hat{u}_j^I (1 - \beta(\lambda_j)) - \hat{v}_{0,j} \lambda_j \gamma(\lambda_j) + \hat{v}_{1,j} \beta(\lambda_j),$$

¹¹⁹⁶ and

$$\beta(\lambda) := \frac{2}{T} \int_0^T \left(\cos(\omega t) - \frac{1}{4} \right) \cos(\lambda t) dt, \quad \gamma(\lambda) := \frac{2}{T} \int_0^T \left(\cos(\omega t) - \frac{1}{4} \right) \sin(\lambda t) dt.$$

By definition we have

$$\left| \frac{\gamma(\lambda_j)}{\lambda_j} \right| \leq \frac{2}{T} \int_0^T \left| \left(\cos(\omega t) - \frac{1}{4} \right) \right| \left| t \frac{\sin(\lambda t)}{\lambda t} \right| dt \leq \frac{2}{T} \int_0^T \frac{5}{4} t dt = \frac{5\pi}{2\omega}, \quad (3.8)$$

1197 since $|\sin(x)/x| \leq 1$, which ensures the boundedness of the coefficients \bar{v}_j, \bar{v}'_j for small eigenvalues
1198 λ_j .

Letting $v_{0,j}, v_{1,j}$ denote the coefficients of v_0, v_1 in the eigenbasis of the Laplacian, we can write the iteration as

$$\begin{bmatrix} v_{0,j}^{n+1} \\ v_{1,j}^{n+1} \end{bmatrix} = \left(\Pi \begin{bmatrix} v_0^n \\ v_1^n \end{bmatrix} \right)_j = (I - B_j) \begin{bmatrix} u_j^R \\ \omega u_j^I \end{bmatrix} + B_j \begin{bmatrix} v_{0,j}^n \\ v_{1,j}^n \end{bmatrix}, \quad (3.9)$$

where if we define $\beta_j = \beta(\lambda_j)$ and $\gamma_j = \gamma(\lambda_j)$ then

$$B_j = \begin{pmatrix} \beta_j & \gamma_j/\lambda_j \\ -\lambda_j \gamma_j & \beta_j \end{pmatrix}, \quad B_0^N = \begin{pmatrix} -1/2 & -\pi/2\omega \\ 0 & -1/2 \end{pmatrix},$$

Moreover, the eigenvectors and eigenvalues of B_j are

$$\xi_j^\pm = \begin{pmatrix} \pm i/\lambda \\ 1 \end{pmatrix}, \quad \xi_0^N = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \mu_j = \beta_j \pm i\gamma_j.$$

Introducing the linear operator $\mathcal{S} : L^2(\Omega) \times L^2(\Omega) \rightarrow L^2(\Omega) \times L^2(\Omega)$,

$$\mathcal{S} \sum_{j=0}^{\infty} \begin{bmatrix} \hat{u}_j^R \\ \hat{u}_j^I \end{bmatrix} \phi_j(x) = \sum_{j=0}^{\infty} B_j \begin{bmatrix} \hat{u}_j^R \\ \hat{u}_j^I \end{bmatrix} \phi_j(x), \quad (3.10)$$

we may write the iteration as

$$\boxed{\begin{bmatrix} v \\ v' \end{bmatrix}^{(n+1)} = \Pi \begin{bmatrix} v \\ v' \end{bmatrix}^{(n)} = \begin{bmatrix} u^R \\ \omega u^I \end{bmatrix} + \mathcal{S} \left(\begin{bmatrix} v \\ v' \end{bmatrix}^{(n)} - \begin{bmatrix} u^R \\ \omega u^I \end{bmatrix} \right).}$$

1199 We note that, in contrast to the simplified iteration analyzed in [14], the operator \mathcal{S} is not symmetric
1200 for the general iteration. Despite this we may identify the eigenmodes of \mathcal{S} from the eigenvectors
1201 of B_j via $\xi_j^\pm \phi_j$ with eigenvalues $\mu_j = \beta_j \pm i\gamma_j$ and $\xi_0^N = \xi_0^N \phi_0$ with eigenvalue $\mu_0^N = -1/2$.

From (3.9), we see that the iteration for each mode takes the form

$$\begin{bmatrix} v_{0,j}^{n+1} \\ v_{1,j}^{n+1} \end{bmatrix} = \left(\Pi \begin{bmatrix} v_0^n \\ v_1^n \end{bmatrix} \right)_j = (I - B_j^n) \begin{bmatrix} u_j^R \\ \omega u_j^I \end{bmatrix} + B_j^n \begin{bmatrix} v_{0,j}^n \\ v_{1,j}^n \end{bmatrix}$$

so that

$$\begin{bmatrix} v_{0,j}^{n+1} - u_j^R \\ v_{1,j}^{n+1} - \omega u_j^I \end{bmatrix} = B_j^n \begin{bmatrix} v_{0,j}^0 - u_j^R \\ v_{1,j}^0 - \omega u_j^I \end{bmatrix}. \quad (3.11)$$

We thus require that $B_j^n \rightarrow 0$ to ensure convergence of the fixed-point iteration to the solution, which is true if and only if the spectral radius of B_j is less than unity uniformly in j . That is, we require that $|\mu_j| < 1$ uniformly in j . Defining the filter function $\mu(\lambda) := \beta(\lambda) + i\gamma(\lambda)$, we may show (with a proof in Appendix .4) the following lemma

Lemma 3.1.1. *The complex-valued filter function μ satisfies $\mu(\omega) = 1$ and*

$$\begin{aligned} 0 \leq |\mu(\lambda)| &\leq 1 - \frac{15}{32} \left(\frac{\lambda - \omega}{\omega} \right)^2, \quad \text{when } \left| \frac{\lambda - \omega}{\omega} \right| \leq \frac{1}{2}, \\ |\mu(\lambda)| &\leq \frac{7}{3\pi} \approx 0.74, \quad \text{when } \left| \frac{\lambda - \omega}{\omega} \right| \geq \frac{1}{2}, \\ |\mu(\lambda)| &\leq b_0 \frac{\omega}{\lambda - \omega}, \quad \text{when } \lambda > \omega, \end{aligned}$$

where $b_0 = 3/2\pi$. Moreover, close to ω we have the local expansion

$$\begin{aligned} |\mu(\omega + r)| &= 1 - b_1 \left(\frac{r}{\omega} \right)^2 + R(r/\omega) \left(\frac{r}{\omega} \right)^3, \\ b_1 &= \frac{\pi^2}{6} - \frac{1}{4} \approx 1.39, \quad \|R\|_\infty \leq \frac{25\pi^4}{4} (36 + 20\pi + 250\pi^2 + 75\pi^3). \end{aligned} \quad (3.12)$$

We denote

$$\delta_j = \frac{\lambda_j - \omega}{\omega},$$

the relative size of the gap between λ_j and the Helmholtz frequency, and then denote the smallest gap (in magnitude) by δ ,

$$\delta = \delta_{j^*}, \quad j^* = \operatorname{argmin}_j |\delta_j|.$$

Then we have the following lemma

Lemma 3.1.2. Suppose $\delta > 0$. Then, the spectral radius ρ of \mathcal{S} is strictly less than one, and for small δ ,

$$\rho = 1 - b_1 \delta^2 + \mathcal{O}(\delta^3), \quad (3.13)$$

with b_1 as in Lemma 3.1.1. Moreover, \mathcal{S} is a bounded linear map from $L^2(\Omega) \times L^2(\Omega)$ to $H^1(\Omega) \times L^2(\Omega)$, and from $H^1(\Omega) \times L^2(\Omega)$ to $H^1(\Omega) \times H^1(\Omega)$.

Proof. From Lemma 3.1.1 we get

$$\rho = \sup_j |\mu(\lambda_j)| \leq \sup_j \max \left(1 - \frac{15}{32} \delta_j^2, \frac{7}{3\pi} \right) \leq \max \left(1 - \frac{15}{32} \delta^2, \frac{7}{3\pi} \right) < 1.$$

For the more precise estimate when δ is small we will use (3.12). Since $1 > \rho \geq |\mu(\omega + \omega\delta)| \rightarrow 1$ as $\delta \rightarrow 0$, we can assume that $\rho > 1 - \eta^2/2$, with $\eta := b_1/2\|R\|_\infty$, for small enough δ . Then, since $|\mu(\omega + \omega\delta_j)| \leq 1 - \eta^2/2$ for $|\delta_j| > \eta$ by Lemma 3.1.1, we have

$$\rho = \sup_{|\delta_j| \leq \eta} |\mu(\omega + \omega\delta_j)| = |\mu(\omega + \omega\delta_{k^*})|,$$

for some k^* with $|\delta_{k^*}| \leq \eta$. If $\delta_{k^*} = \delta_{j^*}$ (where $\delta = |\delta_{j^*}|$) then (3.12) gives (3.13). If not, we have $\eta \geq |\delta_{k^*}| \geq \delta$ and by Lemma 3.1.1

$$0 \leq |\mu(\omega + \omega\delta_{k^*})| - |\mu(\omega + \omega\delta_{j^*})| = -b_1(\delta_{k^*}^2 - \delta^2) + R(\delta_{k^*})\delta_{k^*}^3 - R(\delta_{j^*})\delta_{j^*}^3 \leq -b_1(\delta_{k^*}^2 - \delta^2) + \frac{b_1}{2}(\delta_{k^*}^2 + \delta^2),$$

which implies that $\delta_{k^*}^2 \leq 3\delta^2$ and therefore

$$\rho = 1 - b_1 \delta_{k^*}^2 + \mathcal{O}(\delta_{k^*}^3) = 1 - b_1 \delta^2 + b_1(\delta^2 - \delta_{k^*}^2) + \mathcal{O}(\delta_{k^*}^3) = 1 - b_1 \delta^2 + \mathcal{O}(\delta_{k^*}^3 + \delta^3) = 1 - b_1 \delta^2 + \mathcal{O}(\delta^3).$$

From which (3.13) follows.

Letting $D := \omega \min(1, b_0(1 + 1/|\delta|))$, we note that by Lemma 3.1.1,

$$\begin{aligned} |\lambda_j \mu(\lambda_j)| &\leq \omega \leq D, \quad \lambda_j \leq \omega, \\ |\lambda_j \mu(\lambda_j)| &\leq \omega \frac{b_0 \lambda_j}{\lambda_j - \omega} = \omega b_0(1 + 1/\delta_j) \leq D, \quad \lambda_j > \omega. \end{aligned}$$

Moreover, triangle inequality gives that $|\beta(\lambda_j)|, |\gamma(\lambda_j)| \leq |\mu(\lambda_j)|$, which implies both $\lambda_j |\beta(\lambda_j)| \leq D$ and $\lambda_j |\gamma(\lambda_j)| \leq D$.

Suppose now that $g, h \in L^2(\Omega)$ and

$$g(x) = \sum_{j=0}^{\infty} \hat{g}_j \phi_j(x), \quad h(x) = \sum_{j=0}^{\infty} \hat{h}_j \phi_j(x).$$

Letting $z(x) = [g(x), h(x)]^T$, $\|z\|_{L^2 \times L^2} = 1$, we may split the norm of $\mathcal{S}z$ into

$$\|\mathcal{S}z\|_{H^1(\Omega) \times L^2(\Omega)}^2 = \|\mathcal{S}z\|_{L^2(\Omega) \times L^2(\Omega)}^2 + \|\nabla \mathcal{S}z\|_{L^2(\Omega) \times L^2(\Omega)}^2. \quad (3.14)$$

Letting $C := \max\{D, |\gamma(\lambda_j)|/\lambda_j\}$, which is bounded via the estimate (3.8), straightforward algebra gives the bound

$$\begin{aligned} \|\mathcal{S}z\|_{L^2(\Omega) \times L^2(\Omega)}^2 &= \sum_{j=0}^{\infty} |\beta(\lambda_j) \hat{g}_j + \frac{\gamma(\lambda_j)}{\lambda_j} \hat{h}_j|^2 + |\lambda_j \gamma(\lambda_j) \hat{g}_j - \beta(\lambda_j) \hat{h}_j|^2 \\ &\leq \sum_{j=0}^{\infty} \left(|\hat{g}_j| + \frac{|\gamma(\lambda_j)|}{\lambda_j} |\hat{h}_j| \right)^2 + \left(\lambda_j |\gamma(\lambda_j)| |\hat{g}_j| + |\hat{h}_j| \right)^2 \\ &\leq \sum_{j=0}^{\infty} (1 + C^2) (|\hat{g}_j|^2 + |\hat{h}_j|^2) + 4C |\hat{g}_j| |\hat{h}_j| \\ &\leq \sum_{j=0}^{\infty} (1 + C^2) (|\hat{g}_j|^2 + |\hat{h}_j|^2) + 4C (|\hat{g}_j|^2 + |\hat{h}_j|^2) \\ &= (1 + C^2 + 4C) \|z\|_{L^2(\Omega) \times L^2(\Omega)}^2, \end{aligned}$$

since $ab \leq a^2 + b^2$ for $a, b \in [0, 1]$ and z has unit norm. For the second term of (3.14) we find

$$\|\nabla \mathcal{S}z\|_{L^2(\Omega) \times L^2(\Omega)}^2 \leq \underbrace{\sum_{j=0}^{\infty} \frac{\lambda_j^2}{c_{\min}^2} |\beta(\lambda_j) \hat{g}_j + \frac{\gamma(\lambda_j)}{\lambda_j} \hat{h}_j|^2}_{S_1} + \underbrace{\sum_{j=0}^{\infty} \frac{\lambda_j^2}{c_{\min}^2} |-\gamma(\lambda_j) \lambda_j \hat{g}_j + \beta(\lambda_j) \hat{h}_j|^2}_{S_2}.$$

For $g, h \in L^2(\Omega)$, it follows that

$$\begin{aligned} S_1 &= \sum_{j=0}^{\infty} \frac{\lambda_j^2}{c_{\min}^2} |\beta(\lambda_j) \hat{g}_j + \frac{\gamma(\lambda_j)}{\lambda_j} \hat{h}_j|^2 \leq \sum_{j=0}^{\infty} \frac{D^2}{c_{\min}^2} (|\hat{g}_j|^2 + |\hat{h}_j|^2) + \frac{2\lambda}{c_{\min}^2} |\beta(\lambda_j)| |\gamma(\lambda_j)| |\hat{g}_j| |\hat{h}_j| \\ &\leq \sum_{j=0}^{\infty} \frac{D^2}{c_{\min}^2} (|\hat{g}_j|^2 + |\hat{h}_j|^2) + \frac{2D}{c_{\min}^2} (|\hat{g}_j|^2 + |\hat{h}_j|^2) \\ &= \frac{D^2 + 2D}{c_{\min}^2} \|z\|_{L^2(\Omega) \times L^2(\Omega)}^2, \end{aligned}$$

which gives

$$\|\mathcal{S}\|_{H^1(\Omega) \times L^2(\Omega)}^2 = \sup_{\|z\|=1} \|\mathcal{S}z\|_{H^1(\Omega) \times L^2(\Omega)}^2 \leq \left(1 + C^2 + 4C + \frac{D^2 + 2D}{c_{\min}^2} \right),$$

1213 showing that \mathcal{S} is a bounded linear map from $L^2(\Omega) \times L^2(\Omega)$ to $H^1(\Omega) \times L^2(\Omega)$.

If instead $g \in H^1(\Omega)$ and $h \in L^2(\Omega)$, then

$$\begin{aligned} S_2 &= \sum_{j=0}^{\infty} \frac{\lambda_j^2}{c_{\min}^2} | -\gamma(\lambda_j)\lambda_j \hat{g}_j + \beta(\lambda_j)\hat{h}_j |^2 \\ &\leq \sum_{j=0}^{\infty} \frac{1}{c_{\min}^2} \left(\lambda_j^4 \gamma^2(\lambda_j) |\hat{g}_j|^2 + 2\lambda_j^3 |\beta(\lambda_j)| |\gamma(\lambda_j)| |\hat{g}_j| |\hat{h}_j| + \lambda_j^2 \beta^2(\lambda_j) |\hat{h}_j|^2 \right) \\ &\leq \sum_{j=0}^{\infty} \frac{D^2}{c_{\min}^2} \left(\lambda_j^2 |\hat{g}_j|^2 + 2\lambda_j |\hat{g}_j| |\hat{h}_j| + |\hat{h}_j|^2 \right). \end{aligned}$$

We note that Hölder's inequality gives

$$\sum_{j=0}^{\infty} \lambda_j |\hat{g}_j| |\hat{h}_j| \leq c_{\max} \|\nabla g\|_{L^2(\Omega)} \|h\|_{L^2(\Omega)},$$

so that

$$S_2 \leq \frac{D^2}{c_{\min}^2} \left(c_{\max}^2 \|\nabla g\|_{L^2(\Omega)}^2 + 2c_{\max} \|\nabla g\|_{L^2(\Omega)} \|h\|_{L^2(\Omega)} + \|h\|_{L^2(\Omega)}^2 \right),$$

and thus

$$\|\mathcal{S}\|_{H^1(\Omega) \times H^1(\Omega)}^2 \leq \|\mathcal{S}\|_{H^1(\Omega) \times L^2(\Omega)}^2 + \sup_{\|z\|=1} S_2 \leq \|\mathcal{S}\|_{H^1(\Omega) \times L^2(\Omega)}^2 + \frac{D^2}{c_{\min}^2} (c_{\max}^2 + 2c_{\max} + 1),$$

1214 which shows that \mathcal{S} is a bounded linear map from $H^1(\Omega) \times L^2(\Omega)$ to $H^1(\Omega) \times H^1(\Omega)$, proving the
1215 lemma. \square

Further, denoting $e^n := [\operatorname{Re}\{u\} - v_0^n, \omega \operatorname{Im}\{u\} - v_1^n]^T = [e_0^n, e_1^n]^T$, from (3.11) we obtain

$$e^n = \mathcal{S}[\operatorname{Re}\{u\} - v_0^{n-1}, \omega \operatorname{Im}\{u\} - v_1^{n-1}]^T = \mathcal{S}^n [\operatorname{Re}\{u\} - v_0^0, \omega \operatorname{Im}\{u\} - v_1^0]^T = \mathcal{S}^n e^0,$$

which shows that $e^n \rightarrow 0$ since $S^n \rightarrow 0$. Thus the iterates $[v_0^n, v_1^n]^T$ converge to $[\operatorname{Re}\{u\}, \omega \operatorname{Im}\{u\}]^T$ in $L^2(\Omega) \times L^2(\Omega)$. Since $v_0^0 = v_1^0 = 0$, by Lemma 3.1.2 it follows that the iterates $[v_0^n, v_1^n]^T \in H^1(\Omega) \times L^2(\Omega)$ for $n > 0$. Additionally we have that the iterates $[v_0^n, v_1^n]^T \in H^1(\Omega) \times H^1(\Omega)$ for $n > 1$. We can therefore also get convergence in $H^1(\Omega) \times H^1(\Omega)$. To show this, let

$$\beta_j + i\gamma_j = r_j \exp(i\phi_j), \quad r_j^2 = |\beta_j|^2 + |\gamma_j|^2, \quad \phi_j = \arctan(\gamma_j/\beta_j).$$

It can then be shown that powers of the operator B_j can be written as

$$B_j^n = r_j^n \begin{pmatrix} \cos(n\phi_j) & \sin(n\phi_j)/\lambda_j \\ -\lambda_j \sin(n\phi_j) & \cos(n\phi_j) \end{pmatrix},$$

where each entry is bounded and goes to zero in the limit as $n \rightarrow \infty$ since the spectral radius of B_j is less than one. From the Hölder inequality it follows that

$$\sum_{j=0}^{\infty} \lambda_j |\hat{e}_{j,0}^0| |\hat{e}_{j,1}^1| \leq c_{\max} \|\nabla e_0^0\|_{L^2(\Omega)} \|e_1^0\|_{L^2(\Omega)},$$

so that

$$\begin{aligned} \|\nabla \mathcal{S}^n e^0\|_{L^2(\Omega) \times L^2(\Omega)}^2 &\leq \sum_{j=0}^{\infty} \frac{\lambda_j^2 |r_j|^{2n}}{c_{\min}^2} \left(\left| \cos(n\phi_j) \hat{e}_{j,0}^0 + \frac{\sin(n\phi_j)}{\lambda_j} \hat{e}_{j,1}^1 \right|^2 + \left| -\lambda_j \sin(n\phi_j) \hat{e}_{j,0}^0 + \cos(n\phi_j) \hat{e}_{j,1}^1 \right|^2 \right) \\ &\leq \sum_{j=0}^{\infty} \frac{|\mu_j|^{2n}}{c_{\min}^2} (\lambda_j^2 |\hat{e}_{j,0}^0|^2 + \lambda_j |\hat{e}_{j,0}^0| |\hat{e}_{j,1}^1| + |\hat{e}_{j,1}^1|^2) + \frac{D^2 |\mu_j|^{2n-2}}{c_{\min}^2} (\lambda_j^2 |\hat{e}_{j,0}^0|^2 + \lambda_j |\hat{e}_{j,0}^0| |\hat{e}_{j,1}^1| + |\hat{e}_{j,1}^1|^2) \\ &\leq \left(\sup_j |\mu_j| \right)^{2n-2} \frac{1+D^2}{c_{\min}^2} \sum_{j=0}^{\infty} (\lambda_j^2 |\hat{e}_{j,0}^0|^2 + \lambda_j |\hat{e}_{j,0}^0| |\hat{e}_{j,1}^1| + |\hat{e}_{j,1}^1|^2) \\ &\leq \rho^{2n-2} \frac{1+D^2}{c_{\min}^2} (c_{\max}^2 \|\nabla e_0^0\|_{L^2(\Omega)}^2 + c_{\max} \|\nabla e_0^0\|_{L^2(\Omega)} \|e_1^0\|_{L^2(\Omega)} + \|e_1^0\|_{L^2(\Omega)}^2) \rightarrow 0. \end{aligned}$$

1216 We conclude that the iteration converges in $H^1(\Omega) \times H^1(\Omega)$ with convergence rate ρ . By Lemma
1217 3.1.2 we have $\rho \sim 1 - 1.39\delta^2$ so that the smallest gap, δ , determines the convergence rate. We thus
1218 have proven the following theorem

1219 **Theorem 3.1.3.** *The iteration in (3.7) and (3.5) converges in $H^1(\Omega) \times H^1(\Omega)$ for the Dirichlet
1220 and Neumann problems away from resonances to the solution of the Helmholtz equation (3.1).
1221 The convergence rate is $1 - \mathcal{O}(\delta^2)$, where δ is the minimum gap between ω and the eigenvalues of
1222 $-\nabla \cdot (c^2(x)\nabla)$.*

As seen in [14], we may reformulate the iteration as the linear system

$$(I - \mathcal{S}) v =: \mathcal{A}v = b := \Pi 0,$$

1223 which allows the convergence to be accelerated by a Krylov method. We note that the evaluation
1224 of \mathcal{A} can be done by evolving the wave equation for one period in time with initial data v without
1225 the need to explicitly form the matrix \mathcal{A} .

1226 **Remark 3.1.1.** *The operator \mathcal{A} for the general iteration is not symmetric unlike the simplified
1227 iteration for energy-conserving problems where $v_1 = 0$. For interior, energy-conserving problems
1228 we recommend the use of the simplified iteration so that the conjugate gradient method may be
1229 used to accelerate convergence. For other boundary conditions, the general WaveHoltz iteration is
1230 required and a more versatile Krylov method, such as GMRES, should be used.*

1231 **3.1.2 Convergence in the Non-Energy Conserving Case**

With Theorem 3.1.3 providing convergence of the general WaveHoltz iteration in the energy-conserving case, we turn toward proving convergence for problems with impedance boundary conditions. For simplicity we prove convergence in a single spatial dimension. This is not restrictive, as it is possible for the following technique to be extended to higher dimensions for particular problems. Consider now the following Helmholtz problem with impedance boundary conditions

$$\begin{aligned} [c^2(x)u'(x)]' + \omega^2 u(x) &= f(x), \quad a \leq x \leq b, \\ i\alpha\omega u(a) - \beta c(x)u_x(a) &= 0, \\ i\alpha\omega u(b) + \beta c(x)u_x(b) &= 0, \end{aligned} \tag{3.15}$$

1232 where $\alpha, \beta \neq 0$. Here we assume $1/c \in L^1_{\text{loc}}([a, b])$, $f \in L^2([a, b])$ is compactly supported away from
1233 the boundary, and that $c(a) = c_a$, $c(b) = c_b$ where c is constant in a neighborhood of the endpoints.
1234 We reformulate this in the time domain as

$$\begin{aligned} w_{tt} &= \frac{\partial}{\partial x} \left[c^2(x) \frac{\partial}{\partial x} w \right] - f(x)e^{-i\omega t}, \quad a \leq x \leq b, \quad 0 \leq t \leq T, \\ w(0, x) &= v_0(x), \quad w_t(0, x) = v_1(x), \\ \alpha w_t(t, a) - \beta c(x)w_x(t, a) &= 0, \\ \alpha w_t(t, b) + \beta c(x)w_x(t, b) &= 0. \end{aligned}$$

1235 In general, the solution of the above equation will yield complex-valued solutions and so we take the
1236 real part of the equation as shown earlier and use the general iteration (3.7). Note that in 1D the
1237 impedance boundary conditions with $\alpha = \beta = 1/\sqrt{2}$ are equivalent to outflow/radiation conditions

when the initial data is compactly supported in the interval $[a, b]$. If $\alpha \neq \beta$ with $\alpha, \beta \neq 0$, then in addition to outgoing waves at the boundary there will be reflections due to the impedance boundary condition. In either case, if we let $\tilde{a} < a - c_a T/2$ and $\tilde{b} > b + c_b T/2$, then w is equal to \tilde{w} on $[a, b]$ for $t \in [0, T]$ if \tilde{w} solves the following Neumann problem (with an outline of the construction in Appendix .5)

$$\begin{aligned}\tilde{w}_{tt} &= \frac{\partial}{\partial x} \left[\tilde{c}^2(x) \frac{\partial}{\partial x} \tilde{w} \right] - \operatorname{Re}\{\tilde{f}(x)e^{-i\omega t}\}, \quad \tilde{a} \leq x \leq \tilde{b}, \quad 0 \leq t \leq T, \\ \tilde{w}(0, x) &= \tilde{v}_0(x), \quad \tilde{w}_t(0, x) = \tilde{v}_1(x), \\ \tilde{w}_x(t, a) &= 0, \quad \tilde{w}_x(t, b) = 0,\end{aligned}\tag{3.16}$$

where \tilde{v}_0 and \tilde{c} are the constant extensions (with $\gamma = \alpha/\beta$)

$$\tilde{v}_0(x) = \begin{cases} v_0(a_0), & a \leq x < a_0, \\ v_0(x), & a_0 \leq x \leq b_0, \\ v_0(b_0), & b_0 < x \leq b, \end{cases} \quad \tilde{c}(x) = \begin{cases} \gamma c_a, & a \leq x < a_0, \\ c(x), & a_0 \leq x \leq b_0, \\ \gamma c_b, & b_0 < x \leq b, \end{cases}$$

and \tilde{v}_1, \tilde{f} are zero extensions of v_1 and f , respectively.

That is, we extend the domain such that traveling waves may reflect off of the Neumann boundary but not re-enter the domain of interest, $a \leq x \leq b$, within a period T . Let Π be the WaveHoltz integral operator (3.5) on the original domain Ω with impedance boundary conditions. We recall that iterates generated by Π at a given point, $x \in \Omega$, are the time-average of the wave solution at x generated by the input data. Since the extended wave solution $\tilde{w}(t, x) = w(t, x)$ for $0 \leq t \leq T$, we may write $\Pi = P\tilde{\Pi}E$ where P is a projection operator onto the initial interval, i.e. $Pv(x) = v(x)|_{a \leq x \leq b}$, E is the extension operator such that $[v_0, v_1]^T \rightarrow [\tilde{v}_0, \tilde{v}_1]^T$, and $\tilde{\Pi}$ is the WaveHoltz operator on the domain $\tilde{\Omega}$. If it can be guaranteed that $\omega^2 \neq \lambda_j^2$ where λ_j^2 is an eigenvalue of the operator $-\partial_x(\tilde{c}^2(x)\partial_x)$, then we may prove convergence as was done for Theorem 3.1.3.

To show this, results on the continuity of eigenvalues of the Laplacian from [74] will be used.

We present the framework of [74] needed here and consider the following differential equation

$$-(c^2 y')' = \lambda y, \quad x \in (a', b'), \quad -\infty \leq a' < b' \leq \infty, \quad \lambda \in \mathbb{R},\tag{3.17}$$

where $c^2 : (a', b') \rightarrow \mathbb{R}$ and $1/c^2 \in L^1_{\text{loc}}(a', b')$. Leting $I = [a, b]$, $a' < a < b < b'$ and additionally imposing the Neumann conditions $y'(a) = 0 = y'(b)$, the above Sturm-Liouville (SL) problem is such that all eigenvalues are real, simple, and can be ordered to satisfy

$$0 \leq \lambda_0^2 < \lambda_1^2 < \lambda_2^2 < \dots; \quad \lim_{n \rightarrow \infty} \lambda_n^2 = +\infty. \quad (3.18)$$

1253 We thus immediately have that the eigenvalues of the Laplacian are countable.

1254 Under the above assumptions, we state the following theorem that is proven in [74].

Theorem 3.1.4 (Kong & Zettle). *Let $1/c^2 \in L^1_{\text{loc}}(a', b')$, fix a', b' , and suppose b is such that $a' < b < b'$. Let $\lambda_n(b)$ be an eigenvalue of the SL problem (3.17) with homogeneous Neumann boundary conditions with corresponding eigenfunction $u_n(x; b)$. Then the eigenvalue $\lambda(b) \in C^1([a', b'])$ satisfies the following differential equation:*

$$\lambda'_n(b) = -\lambda_n(b)u_n^2(b; b).$$

1255 That is, the eigenvalues of the SL problem (3.17) are differentiable functions of the endpoint

1256 b . This gives us the following useful corollary.

1257 **Corollary 3.1.4.1.** *For $n = 1, 2, \dots$, $\lambda_n(b)$ is a strictly decreasing function of b on $[a', b']$.*

Proof. For homogeneous Neumann conditions, we have that $u'_n(b; b) = 0$. It follows that $u_n(b; b) \neq 0$ as otherwise $u_n(b; b) \equiv 0$ since u_n satisfies a linear, homogeneous second order ODE. As $\lambda_n(b) > 0$ for $n > 0$ we then have

$$\lambda'_n(b) = -\lambda_n(b)u_n^2(b; b) < 0,$$

1258 so that $\lambda_n(b)$ is a strictly decreasing function of the endpoint b . \square

1259 As a consequence of Theorem 3.1.4, we have

1260 **Lemma 3.1.5.** *Suppose $\tilde{c} \geq 0$ a.e. and $\omega > 0$. Fix a , and consider the Neumann eigenvalues $\lambda_n(b)$*
1261 *for $b \in (b_0 + c_b T/2, r)$ where $r > b_0 + c_b T/2$. Then there exists an endpoint $\tilde{b} \in (b_0 + c_b T/2, r)$ such*
1262 *that $\omega^2 \neq \lambda_n(\tilde{b})$ for each $n \in \mathbb{N}_0$.*

Proof. Clearly we have $\lambda_0(b) = 0$ for every b , and since $\omega > 0$ we have $\omega^2 \neq \lambda_0(b)$. Suppose now that b is such that $\omega^2 = \lambda_n(b)$ for some $n \in \mathbb{N}$. Recall that by (3.18) we have that $\omega^2 = \lambda_n(b) < \lambda_{n+1}(b)$. Since $\lambda_n(b), \lambda_{n+1}(b)$ are continuous, decreasing functions of the endpoint by Corollary 3.1.4.1, there necessarily exists $\delta > 0$ such that

$$\lambda_n(b + \delta) < \omega^2 < \lambda_{n+1}(b + \delta),$$

1263 Letting $\tilde{b} = b + \delta$ we thus have that $\omega^2 \neq \lambda_n(\tilde{b})$ for each $n \in \mathbb{N}_0$, as desired. \square

1264 From this we can prove the following theorem, in which we note we demonstrate convergence
1265 in $H^1(\Omega) \times L^2(\Omega)$ rather than $H^1(\Omega) \times H^1(\Omega)$.

1266 **Theorem 3.1.6.** *Let the 1D domain $\Omega = [a, b]$ be a bounded interval. Suppose $f \in L^2(\Omega)$ is
1267 compactly supported in Ω away from the boundary, $1/c^2 \in L^1_{loc}(\Omega)$, and $c(a) = c_a$, $c(b) = c_b$, with
1268 c constant near the endpoints. Under these conditions, the iteration (3.7) and (3.5) converges in
1269 $H^1(\Omega) \times L^2(\Omega)$ for the Helmholtz problem with impedance boundary conditions to the solution of
1270 the Helmholtz equation (3.15).*

1271 *Proof.* By Lemma 3.1.5, there exists an extended wave equation (3.16) on the domain $\tilde{\Omega} = [\tilde{a}, \tilde{b}]$
1272 with homogeneous Neumann boundary conditions such that the eigenvalues $\tilde{\lambda}_j$ of the Laplacian,
1273 $-\partial_x(\tilde{c}^2 w_x)$, on $\tilde{\Omega}$ are not in resonance. Defining $\tilde{\beta}_j = \beta(\tilde{\lambda}_j)$, $\tilde{\gamma}_j = \gamma(\tilde{\lambda}_j)$, and $\tilde{\mu}_j = \mu(\tilde{\lambda}_j)$, this
1274 immediately gives that the spectral radius of the WaveHoltz operator, $\tilde{\rho} = \sup_j |\tilde{\mu}_j|$, is smaller than
1275 one. Moreover, the extended wave solution \tilde{w} on $\tilde{\Omega}$ coincides with the interior impedance wave
1276 solution w on Ω for $t \in [0, T]$.

1277 Letting u be the solution of the Helmholtz equation (3.15), we define $q(t, x) = \cos(\omega t)[\text{Re}\{u\}, \omega \text{Im}\{u\}]^T$
1278 the time-harmonic Helmholtz solution in Ω and $\tilde{w}^n(t, x)$ the solution of (3.16) with initial data
1279 v_0^n, v_1^n . Letting the error be $e^n := [\text{Re}\{u\} - v_0^n, \omega \text{Im}\{u\} - v_1^n]^T = [e_0^n, e_1^n]^T$, it is clear that the

difference $d(t, x) = q(t, x) - w(t, x)$ satisfies the unforced, homogeneous wave equation

$$\begin{aligned} d_{tt} &= \frac{\partial}{\partial x} \left[c^2(x) \frac{\partial}{\partial x} d \right], \quad a \leq x \leq b, \quad 0 \leq t \leq T, \\ d(0, x) &= e_0(x), \quad d_t(0, x) = e_1(x), \\ \alpha d_t(t, a) - \beta c(x) d_x(t, a) &= 0, \\ \alpha d_t(t, b) + \beta c(x) d_x(t, b) &= 0. \end{aligned}$$

It follows that the WaveHoltz iteration applied to the error is of the form

$$\begin{bmatrix} e_0^{n+1} \\ e_1^{n+1} \end{bmatrix} = \Pi \begin{bmatrix} e_0^n \\ e_1^n \end{bmatrix} = P \tilde{\mathcal{S}} E \begin{bmatrix} e_0^n \\ e_1^n \end{bmatrix} = (P \tilde{\mathcal{S}} E)^{n+1} \begin{bmatrix} e_0^0 \\ e_1^0 \end{bmatrix}, \quad (3.19)$$

where $\tilde{\mathcal{S}}$ is the representation of the operator $\tilde{\Pi}$, as defined in (3.10), with respect to the eigenbasis of the extended Laplacian. We may rearrange the above iteration as

$$(P \tilde{\mathcal{S}} E)^{n+1} e^0 = P(\tilde{\mathcal{S}} E P)^n \tilde{\mathcal{S}} E e^0 = P(\tilde{\mathcal{S}} E P)^n \tilde{e}^0,$$

where $\tilde{e}^0 = \tilde{\mathcal{S}} E e^0 \in H^1(\tilde{\Omega}) \times L^2(\tilde{\Omega})$ since $e^0 \in L^2(\Omega) \times L^2(\Omega)$ and $\tilde{\mathcal{S}}$ is a bounded linear map from

$L^2(\tilde{\Omega}) \times L^2(\tilde{\Omega})$ to $H^1(\tilde{\Omega}) \times L^2(\tilde{\Omega})$ by Lemma 3.1.2. We then have that $(\tilde{\mathcal{S}} E P)^n : H^1(\tilde{\Omega}) \times L^2(\tilde{\Omega}) \rightarrow$

$H^1(\tilde{\Omega}) \times L^2(\tilde{\Omega})$ and we may obtain convergence if $\|(\tilde{\mathcal{S}} E P)^n \tilde{e}^0\|_c$ goes to zero.

Let $\tilde{z} = [\tilde{v}_0, \tilde{v}_1]^T$ with $\tilde{z} \in H^1(\tilde{\Omega}) \times L^2(\tilde{\Omega})$. We define the energy semi-norm $\|\cdot\|_c$ on $H^1(\tilde{\Omega})$,

$$\|\tilde{v}_0\|_c^2 = \left\| \tilde{c} \frac{\partial}{\partial x} \tilde{v}_0 \right\|_{L^2(\tilde{\Omega})}^2 = \sum_{j=0}^{\infty} \tilde{\lambda}_j^2 |\tilde{v}_{0,j}|^2,$$

with the associated semi-norm on $H^1(\tilde{\Omega}) \times L^2(\tilde{\Omega})$

$$\|\tilde{z}\|_c^2 = \|\tilde{v}_0\|_c^2 + \|\tilde{v}_1\|_{L^2(\tilde{\Omega})}^2 = \left\| \tilde{c} \frac{\partial}{\partial x} \tilde{v}_0 \right\|_{L^2(\tilde{\Omega})}^2 + \|\tilde{v}_1\|_{L^2(\tilde{\Omega})}^2.$$

Note that in this semi-norm we have that

$$\begin{aligned} \|EP\tilde{z}\|_c^2 &= \int_{\tilde{\Omega}} \left| \tilde{c} \frac{\partial}{\partial x} EP\tilde{v}_0 \right|^2 + |EP\tilde{v}_1|^2 dx = \int_{\Omega} \left| \tilde{c} \frac{\partial}{\partial x} \tilde{v}_0 \right|^2 + |\tilde{v}_1|^2 dx \leq \int_{\tilde{\Omega}} \left| \tilde{c} \frac{\partial}{\partial x} \tilde{v}_0 \right|^2 + |\tilde{v}_1|^2 dx \\ &\leq \|\tilde{z}\|_c^2. \end{aligned}$$

We define $\tilde{y} = EP\tilde{z}$, where \tilde{y} has the form

$$\tilde{y} = EP \sum_{j=0}^{\infty} \begin{bmatrix} \tilde{v}_{0,j} \\ \tilde{v}_{1,j} \end{bmatrix} \phi_j = \sum_{j=0}^{\infty} \begin{bmatrix} \tilde{y}_{0,j} \\ \tilde{y}_{1,j} \end{bmatrix} \phi_j.$$

It follows that

$$\tilde{\mathcal{S}}EP\tilde{z} = \sum_{j=0}^{\infty} B_j \begin{bmatrix} \tilde{y}_{0,j} \\ \tilde{y}_{1,j} \end{bmatrix} \phi_j = \sum_{j=0}^{\infty} \begin{bmatrix} \tilde{\beta}_j \tilde{y}_{0,j} + \tilde{\gamma}_j \tilde{y}_{1,j} / \tilde{\lambda}_j \\ -\tilde{\lambda}_j \tilde{y}_{0,j} + \tilde{\beta}_j \tilde{y}_{1,j} \end{bmatrix} \phi_j,$$

so that

$$\|\tilde{\mathcal{S}}EP\tilde{z}\|_c^2 = \sum_{j=0}^{\infty} \tilde{\lambda}_j^2 \left(\tilde{\beta}_j \tilde{y}_{0,j} + \frac{\tilde{\gamma}_j}{\tilde{\lambda}_j} \tilde{y}_{1,j} \right)^2 + \sum_{j=0}^{\infty} \left(-\tilde{\lambda}_j \tilde{\gamma}_j \tilde{y}_{0,j} + \tilde{\beta}_j \tilde{y}_{1,j} \right)^2.$$

Since $\tilde{\beta}_j^2 + \tilde{\gamma}_j^2 = |\tilde{\mu}_j|^2 \leq \tilde{\rho}^2 < 1$, a simple expansion shows that

$$\|\tilde{\mathcal{S}}EP\tilde{z}\|_c^2 = \sum_{j=0}^{\infty} (\tilde{\beta}_j^2 + \tilde{\gamma}_j^2) (\tilde{\lambda}_j^2 |\tilde{y}_{0,j}|^2 + |\tilde{y}_{1,j}|^2) \leq \left(\sup_j |\tilde{\mu}_j|^2 \right) \sum_{j=0}^{\infty} \tilde{\lambda}_j^2 |\tilde{y}_{0,j}|^2 + |\tilde{y}_{1,j}|^2 \leq \tilde{\rho}^2 \|\tilde{z}\|_c^2.$$

It follows that

$$\|(\tilde{\mathcal{S}}EP)^n \tilde{e}^0\|_c^2 \leq \tilde{\rho}^2 \|(\tilde{\mathcal{S}}EP)^{n-1} \tilde{e}^0\|_c^2 \leq \dots \leq \tilde{\rho}^{2n} \|\tilde{e}^0\|_c^2 \rightarrow 0,$$

¹²⁸⁵ so that $\|\nabla e_0^n\|_{L^2(\Omega)}, \|e_1^n\|_{L^2(\Omega)} \rightarrow 0$.

With $\tilde{e}_0^n = (\tilde{\mathcal{S}}EP)^n \tilde{e}^0$, an application of the triangle and Poincaré inequality now gives

$$\begin{aligned} \|\tilde{e}_0^n\|_{H^1(\tilde{\Omega})}^2 &= \|\nabla \tilde{e}_0^n\|_{L^2(\tilde{\Omega})}^2 + \|\tilde{e}_0^n\|_{L^2(\tilde{\Omega})}^2 \leq \|\nabla \tilde{e}_0^n\|_{L^2(\tilde{\Omega})}^2 + \|\tilde{e}_0^n - \frac{1}{2} \tilde{e}_{0,0}^n \phi_0\|_{L^2(\tilde{\Omega})}^2 + \frac{1}{2} \|\tilde{e}_{0,0}^n \phi_0\|_{L^2(\tilde{\Omega})}^2 \\ &\leq \|\nabla \tilde{e}_0^n\|_{L^2(\tilde{\Omega})}^2 + C \|\nabla \tilde{e}_0^n\|_{L^2(\tilde{\Omega})}^2 + \frac{1}{2} \|\tilde{e}_{0,0}^n \phi_0\|_{L^2(\tilde{\Omega})}^2 \\ &\leq (C+1) \|\nabla \tilde{e}_0^n\|_{L^2(\tilde{\Omega})}^2 + \frac{1}{2} \|\tilde{e}_{0,0}^n \phi_0\|_{L^2(\tilde{\Omega})}^2, \end{aligned} \quad (3.20)$$

¹²⁸⁶ where $\phi_0 \in \tilde{\Omega}$ is a constant eigenfunction of the Laplacian (and thus of $\tilde{\mathcal{S}}$) with eigenvalue $\lambda_0 = 0$.

¹²⁸⁷ It follows that to obtain convergence in $H^1(\tilde{\Omega})$ of the error \tilde{e}_0^n we must examine the convergence of

¹²⁸⁸ $\tilde{e}_{0,0}$ separately.

It is clear that any constant function on $\tilde{\Omega}$ is an eigenfunction of the operator EP . With

$\tilde{e}_0 = \tilde{\mathcal{S}}Ee_0$ we have

$$\tilde{\mathcal{S}}EP \begin{bmatrix} \tilde{e}_{0,0} \\ 0 \end{bmatrix} \phi_0 = \tilde{\mathcal{S}} \begin{bmatrix} \tilde{e}_{0,0} \\ 0 \end{bmatrix} \phi_0 = B_0^N \begin{bmatrix} \tilde{e}_{0,0} \\ 0 \end{bmatrix} \phi_0 = -\frac{1}{2} \begin{bmatrix} \tilde{e}_{0,0} \\ 0 \end{bmatrix} \phi_0,$$

so that with $\tilde{z}_0 = [\tilde{e}_{0,0}, 0]^T \phi_0$ then

$$\|(\tilde{\mathcal{S}}EP)^n \tilde{z}_0\|_{L^2(\tilde{\Omega}) \times L^2(\tilde{\Omega})} \leq 2^{-1} \|(\tilde{\mathcal{S}}EP)^{n-1} \tilde{z}_0\|_{L^2(\tilde{\Omega}) \times L^2(\tilde{\Omega})} \leq \dots \leq 2^{-n} \|\tilde{z}_0\|_{L^2(\tilde{\Omega}) \times L^2(\tilde{\Omega})} \rightarrow 0.$$

It then follows that

$$\lim_{n \rightarrow \infty} (P\tilde{\mathcal{S}}E)^{n+1} e^0 = \lim_{n \rightarrow \infty} P(\tilde{\mathcal{S}}EP)^n \tilde{e}^0 = 0 \implies \lim_{n \rightarrow \infty} \frac{1}{2} \|\tilde{e}_{0,0}^n \phi_0\|_{L^2(\tilde{\Omega})}^2 = 0.$$

1289 Thus taking the limit of (3.20) gives $\|\tilde{e}_0^n\|_{H^1(\tilde{\Omega})}^2 \rightarrow 0$, so that we obtain convergence of the iteration
1290 in $H^1(\Omega) \times L^2(\Omega)$. \square

1291 **Remark 3.1.2.** *The above analysis is for a single spatial dimension, but we note that it in certain
1292 situations it may be extended to higher dimensions. For instance, interior impedance problems
1293 with constant coefficients may be extended by an appropriate enclosing box from which the above
1294 arguments can give convergence. In general, it is difficult to prove convergence in higher dimensions
1295 as care needs to be taken to make appropriate wavespeed extensions that avoid reflections due to
1296 potentially discontinuous wavespeeds close to boundaries with impedance conditions.*

1297 **3.2 Damped Wave/Helmholtz Equation**

As mentioned in the introduction, a popular preconditioning approach for solving Helmholtz problems is to introduce a damping term as in the shifted Laplacian preconditioners [44]. In this section we similarly consider the complex-valued damped wave equation

$$w_{tt} + \eta w_t = \nabla \cdot [c^2(x) \nabla w] - f(x) e^{-i\omega t},$$

from which we note that if $w(t, x) = u(x) e^{-i\omega t}$ then

$$\nabla \cdot [c^2(x) \nabla u] + (\omega^2 + i\eta\omega) u = f(x),$$

so that we essentially have added a purely imaginary shift of the Laplacian

$$\mathcal{L} = -\nabla \cdot [c^2(x) \nabla] - i\eta\omega.$$

1298 While for the sake of simplicity we consider the complex-valued problem in this section, in practice
 1299 we solve the real-valued problem as presented in Section 3.1 with the filter (3.5). For the above
 1300 complex-valued problem, we may then similarly prove an analogous result to Theorem 3.1.3

Theorem 3.2.1. *The iteration (3.7) with the complex-valued filter*

$$\Pi \begin{bmatrix} v_0 \\ v_1 \end{bmatrix} = \frac{1}{T} \int_0^T e^{i\omega t} \begin{bmatrix} w(t, x) \\ w_t(t, x) \end{bmatrix} dt, \quad T = \frac{2\pi}{\omega},$$

1301 converges for every $\eta > 0$ with a convergence rate bounded by $2(1 - e^{-\eta T/2})/\eta T$.

Proof. Suppose (λ_j^2, ϕ_j) are the eigenmodes of the real-valued Laplacian in the domain Ω . We note that the shifted Laplacian now has a spectrum that is $\lambda_j^2 - i\eta\omega$. Expanding in terms of this basis and taking inner products, we can see that

$$(\omega^2 + i\eta\omega - \lambda_j^2)\hat{u}_j = \hat{f}_j,$$

where we expand the real and imaginary parts of u and f as $\hat{u}_j = u_j^R + iu_j^I$ and $\hat{f}_j = f_j^R + if_j^I$. Let the damped wave equation solution have the form

$$\sum_{n=0}^{\infty} w_j(t) \phi_j(x).$$

Defining $\alpha_j = \sqrt{4\lambda_j^2 - \eta^2}/2$, then the solution can be shown to be given by

$$\begin{aligned} w_j(t) = \hat{u}_j & \left(e^{-i\omega t} - e^{-\frac{\eta t}{2}} \left[\cos(\alpha_j t) + \frac{\eta - 2i\omega}{2\alpha_j} \sin(\alpha_j t) \right] \right) + \hat{v}_{0,j} e^{-\frac{\eta t}{2}} \left[\cos(\alpha_j t) + \frac{\eta}{2\alpha_j} \sin(\alpha_j t) \right] \\ & + \frac{\hat{v}_{1,j}}{\alpha_j} e^{-\frac{\eta t}{2}} \sin(\alpha_j t), \end{aligned}$$

1302 from which we note that we arrive at exactly the same set of coefficients as in the previous analysis
 1303 if $\eta = 0$ and the real part of the solution is taken. Using the complex-valued filters

$$\hat{\beta}(\alpha) := \frac{1}{T} \int_0^T e^{(i\omega - \eta/2)t} \cos(\alpha t) dt, \quad \hat{\gamma}(\alpha) := \frac{1}{T} \int_0^T e^{(i\omega - \eta/2)t} \sin(\alpha t) dt,$$

we can write the iteration as

$$\begin{pmatrix} v_{0,j}^{n+1} \\ v_{1,j}^{n+1} \end{pmatrix} = \Pi \begin{pmatrix} v_{0,j}^n \\ v_{1,j}^n \end{pmatrix} = \left(I - \hat{B}_j \right) \begin{pmatrix} u_j \\ i\omega u_j \end{pmatrix} + \hat{B}_j \begin{pmatrix} v_{0,j}^n \\ v_{1,j}^n \end{pmatrix}, \quad (3.21)$$

where if $\hat{\beta}_j = \hat{\beta}(\alpha_j)$ and $\hat{\gamma}_j = \hat{\gamma}(\alpha_j)$ then

$$\hat{B}_j = \begin{pmatrix} \hat{\beta}_j + \frac{\eta}{2\alpha_j} \hat{\gamma}_j & \hat{\gamma}_j / \alpha_j \\ -(\alpha_j + \frac{\eta^2}{4\alpha_j}) \hat{\gamma}_j & \hat{\beta}_j - \frac{\eta}{2\alpha_j} \hat{\gamma}_j \end{pmatrix}.$$

As in the previous analysis, we require that the spectral radius of \hat{B}_j be less than one. The eigenvalues are given by $\hat{\mu}_j = \hat{\beta}_j \pm i\hat{\gamma}_j$ so that by definition

$$|\hat{\mu}_j| = |\hat{\beta}_j \pm i\hat{\gamma}_j| = \left| \frac{1}{T} \int_0^T e^{i(\omega \pm \alpha_j)t} e^{-\eta t/2} dt \right| \leq \frac{2}{\eta T} (1 - e^{-\eta T/2}) < 1, \quad (3.22)$$

given that $\eta > 0$. □

Thus the iteration **always** converges in the damped case. From (3.22) we see that for a desired fixed rate of convergence the damping parameter η must grow proportionally to ω , and that frequency-independent convergence is achieved by choosing $\eta = \mathcal{O}(\omega)$.

3.3 Analysis of Higher Order Time Stepping Schemes for the Discrete Iteration

We introduce the temporal grid points $t_n = n\Delta t$ and a spatial grid with N points together with the vector $w^n \in \mathbb{R}^N$ containing the grid function values of the approximation at $t = t_n$. We also let $f \in \mathbb{R}^N$ hold the corresponding values of the right hand side. The discretization of the continuous spatial operator $-\nabla \cdot (c^2(x)\nabla)$, including the boundary conditions, is denoted L_h and it can be represented as an $N \times N$ matrix. The values $-\nabla \cdot (c^2(x)\nabla w)$ are then approximated by $L_h w^n$. As in the continuous case, we assume L_h has the eigenmodes (λ_j^2, ϕ_j) , such that $L_h \phi_j = \lambda_j^2 \phi_j$ for $j = 1, \dots, N$, where all λ_j are strictly positive and ordered as $0 \leq \lambda_1 \leq \dots \leq \lambda_N$.

We let the discrete Helmholtz solution u be defined through

$$-L_h u + \omega^2 u = f.$$

The numerical approximation of the iteration operator is denoted Π_h , and it is implemented as follows. Given $v \in \mathbb{R}^N$, we use the leap frog method to solve the wave equation and add in higher

order corrections as in the Modified Equation (ME) approach [99, 5]. For a general $2m$ scheme, recall that via Taylor expansion

$$\frac{w^{n+1} - 2w^n + w^{n-1}}{\Delta t^2} = w_{tt} + 2 \sum_{k=2}^{\infty} \frac{\Delta t^{2(k-1)}}{(2k)!} \frac{\partial^{2k}}{\partial t^{2k}} w^n.$$

Then using the PDE to convert time derivatives to spatial derivatives we get the expression

$$\frac{\partial^{2k}}{\partial t^{2k}} w^n = L_h^k w^n + \cos(\omega t_n) \sum_{\ell=0}^{k-1} (-1)^{k+\ell} \omega^{2(k-\ell-1)} L_h^\ell f,$$

for $k = 1, 2, \dots$. Then for a $2m$ order scheme we have

$$\frac{w^{n+1} - 2w^n + w^{n-1}}{\Delta t^2} - 2 \sum_{k=2}^m \frac{\Delta t^{2(k-1)}}{(2k)!} \left[L_h^k w^n + \cos(\omega t_n) \sum_{\ell=0}^{k-1} (-1)^{k+\ell} \omega^{2(k-\ell-1)} L_h^\ell f \right] = L_h w^n - f \cos(\omega t_n), \quad (3.23)$$

with time step $\Delta t = T/M$ for some integer M , and initial data

$$w^0 = v, \quad w^{-1} = v + \sum_{k=1}^m \frac{(-1)^k \Delta t^{2k}}{(2k)!} \left[-L_h^k v + \sum_{\ell=0}^{k-1} (-1)^\ell \omega^{2(k-\ell-1)} L_h^\ell f \right].$$

¹³¹⁷ The trapezoidal rule is then used to compute $\Pi_h v$,

$$\Pi_h v = \frac{2\Delta t}{T} \sum_{n=0}^M \eta_n \left(\cos(\omega t_n) - \frac{1}{4} \right) w^n, \quad \eta_n = \begin{cases} \frac{1}{2}, & n = 0 \text{ or } n = M, \\ 1, & 0 < n < M. \end{cases} \quad (3.24)$$

¹³¹⁸ We may then prove the following theorem that is a generalization of Theorem 2.4 of [14].

¹³¹⁹ **Theorem 3.3.1.** *Suppose there are no resonances, such that $\delta_h = \min_j |\lambda_j - \omega|/\omega > 0$. Moreover,*
¹³²⁰ *assume that Δt satisfies the stability and accuracy requirements*

$$\Delta t < \frac{2}{\lambda_N + 2\omega/\pi}, \quad \Delta t \omega \leq \min(\delta_h, 1). \quad (3.25)$$

Then the fixed point iteration $v^{(k+1)} = \Pi_h v^{(k)}$ with $v^{(0)} = 0$ converges to v^∞ which is a solution to the discretized Helmholtz equation with the modified frequency $\tilde{\omega}$,

$$-L_h v^\infty + \tilde{\omega}^2 v^\infty = f, \quad \sin^2(\omega \Delta t / 2) = \sum_{j=1}^m \frac{(-1)^{j+1} (\Delta t \tilde{\omega})^{2j}}{2(2j)!} = \sin^2(\tilde{\omega} \Delta t / 2) + \mathcal{O}(\Delta t^{2m+2}),$$

¹³²¹ *where $2m$ is the order of the ME time stepping scheme. Moreover, $|\omega - \tilde{\omega}| = \mathcal{O}(\Delta t^{2m})$, $\|u - v^\infty\| =$
¹³²² $\mathcal{O}(\Delta t^{2m})$, and the convergence rate is at least $\rho_h = \max(1 - 0.3\delta_h^2, 0.6)$.*

Proof. We expand all functions in eigenmodes of L_h ,

$$w^n = \sum_{j=1}^N \hat{w}_j^n \phi_j, \quad f = \sum_{j=1}^N \hat{f}_j \phi_j, \quad u = \sum_{j=1}^N \hat{u}_j \phi_j, \quad v = \sum_{j=1}^N \hat{v}_j \phi_j, \quad v^\infty = \sum_{j=1}^N \hat{v}_j^\infty \phi_j.$$

Then the Helmholtz eigenmodes of u and v^∞ satisfy

$$\hat{u}_j = \frac{\hat{f}_j}{\omega^2 - \lambda_j^2}, \quad \hat{v}_j^\infty = \frac{\hat{f}_j}{\tilde{\omega}^2 - \lambda_j^2}.$$

We note that $\tilde{\omega}$ is well-defined, with a verification in Appendix .7. Moreover, $\tilde{\omega}$ is not resonant and \hat{v}_j^∞ is well-defined for all j , since by (11) and (3.25)

$$|\tilde{\omega} - \lambda_j| \geq |\omega - \lambda_j| - |\tilde{\omega} - \omega| \geq \omega \delta_h - \frac{\Delta t^{2m} \omega^{2m+1}}{(2m+2)!} \geq \omega \left(\delta_h - \frac{1}{(2m+2)!} \min(\delta_h, 1)^{2m} \right) > 0.$$

The wave solution eigenmodes to (3.23) are given by the difference equation

$$\hat{w}_j^{n+1} - 2\hat{w}_j^n + \hat{w}_j^{n-1} + 2 \left[\sum_{k=1}^m \frac{(-1)^{k+1} \Delta t^{2k} \lambda_j^{2k}}{(2k)!} \right] \hat{w}_j^n = 2 \left[\sum_{k=1}^m \frac{(-1)^k \Delta t^{2k}}{(2k)!} \sum_{\ell=0}^{k-1} \omega^{2(k-\ell-1)} \lambda_j^{2\ell} \right] \hat{f}_j \cos(\omega t_n), \quad (3.26)$$

with initial data

$$\hat{w}_j^0 = \hat{v}_j, \quad \hat{w}_j^{-1} = \hat{v}_j \left(1 + \sum_{k=1}^m \frac{(-1)^k \Delta t^{2k}}{(2k)!} \lambda_j^{2k} \right) + \hat{f}_j \left(\sum_{k=1}^m \frac{(-1)^k \Delta t^{2k}}{(2k)!} \sum_{\ell=0}^{k-1} \omega^{2(k-\ell-1)} \lambda_j^{2\ell} \right).$$

By (3.25), the discrete solution is stable and given by (the verification of which is found in Appendix .6)

$$\hat{w}_j^n = (\hat{v}_j - \hat{v}_j^\infty) \cos(\tilde{\lambda}_j t_n) + \hat{v}_j^\infty \cos(\omega t_n), \quad (3.27)$$

where $\tilde{\lambda}_j$ is well-defined by the relation (verification in (10) of Appendix .7)

$$\sin^2(\tilde{\lambda}_j \Delta t / 2) = \sum_{k=1}^m \frac{(-1)^{k+1} (\Delta t \lambda_j)^{2k}}{2(2k)!}. \quad (3.28)$$

¹³²³ Since $|\omega - \tilde{\omega}| \leq \Delta t^2 \omega^3 / 24$, the following lemma (restated from [14]) gives convergence of the discrete iteration.

¹³²⁵ **Lemma 3.3.2.** *Under the assumptions of Theorem 3.3.1,*

$$\max_{1 \leq j \leq N} |\beta_h(\tilde{\lambda}_j)| \leq \rho_h =: \max(1 - 0.3\delta_h^2, 0.6). \quad (3.29)$$

From (11), it follows that $|\omega - \tilde{\omega}| = \mathcal{O}(\Delta t^{2m})$. Letting $e = u - v^\infty$ be the error in the discrete solutions, the components of the error in the basis of the Laplacian satisfy

$$\begin{aligned} |e_j| = |\hat{u}_j - \hat{v}_j^\infty| &= \left| \hat{f}_j \left(\frac{1}{\omega^2 - \lambda_j^2} - \frac{1}{\tilde{\omega}^2 - \lambda_j^2} \right) \right| = \left| \hat{f}_j \left(\frac{\tilde{\omega}^2 - \omega^2}{(\tilde{\omega}^2 - \lambda_j^2)(\omega^2 - \lambda_j^2)} \right) \right| \\ &= \left| \hat{f}_j \left(\frac{(\tilde{\omega} - \omega)(\tilde{\omega} + \omega)}{(\tilde{\omega} - \lambda_j)(\tilde{\omega} + \lambda_j)(\omega - \lambda_j)(\omega + \lambda_j)} \right) \right| \\ &= \left| \hat{f}_j \left(\frac{(\tilde{\omega} - \omega)(\tilde{\omega} + \omega)}{\tilde{\omega} \delta_j (\tilde{\omega} + \lambda_j) \omega \delta_j (\omega + \lambda_j)} \right) \right| \\ &\leq \left| \hat{f}_j(\tilde{\omega} - \omega) \right| \left(\frac{(\tilde{\omega} + \omega)}{\tilde{\omega} \delta^*(\tilde{\omega} + \lambda_1) \omega \delta^*(\omega + \lambda_1)} \right), \end{aligned}$$

where $\delta^* = \min_j \delta_j = (\omega - \lambda_j)/\omega$ and $\tilde{\delta}^* = \min_j \tilde{\delta}_j = (\tilde{\omega} - \lambda_j)/\omega$. This gives

$$\|u - v^\infty\|_2 = \|e\|_2 \leq |\tilde{\omega} - \omega| \left(\frac{(\tilde{\omega} + \omega)}{\tilde{\omega} \delta^*(\tilde{\omega} + \lambda_1) \omega \delta^*(\omega + \lambda_1)} \right) \|f\|_2 \approx \mathcal{O}(\Delta t^{2m}),$$

1326 since $\delta_*, \tilde{\delta}_* > 0$, concluding the proof of the theorem. □

Remark 3.3.1. As alluded to in Remark 6 of [14], knowledge of how a particular discretization approximates the eigenvalues of the continuous operator can be used to improve the iteration. In fact, the above error due to time discretization can be removed by defining $\bar{\omega}$ by the relation

$$\sin^2(\bar{\omega}\Delta t/2) = \sum_{k=1}^m \frac{(-1)^{k+1} (\Delta t \omega)^{2k}}{2(2k)!}.$$

Then using $f \cos(\bar{\omega}t_n)$ instead of $f \cos(\omega t_n)$ in the time stepping (3.23), in addition to the modified trapezoidal quadrature rule (first introduced in [93])

$$\Pi_h v = \frac{2\Delta t}{T} \sum_{n=0}^M \eta_n \frac{\cos(\omega t_n)}{\cos(\bar{\omega} t_n)} \left(\cos(\omega t_n) - \frac{1}{4} \right) w^n, \quad \eta_n = \begin{cases} \frac{1}{2}, & n = 0 \text{ or } n = M, \\ 1, & 0 < n < M, \end{cases}$$

1327 gives that the limit will be precisely the discrete Helmholtz solution, $v^\infty = u$, as long as the time step
 1328 size is chosen so that $\cos(\bar{\omega}t_n) \neq 0$. Moreover, the first timestep restriction of (3.25) arising from
 1329 the usual CFL condition for the second order scheme may be relaxed (expressions for which may be
 1330 found in [56]) though the condition $\Delta t \omega \leq \min(\delta_h, 1)$ may be more restrictive for problems close to
 1331 resonance. We additionally note that in [101] an alternative approach to remove time-discretization

1332 error was presented, however the approach modified the timestepping scheme whereas we modify the
1333 frequency of the forcing and update our quadrature rule.

1334 **3.4 Wave Equation Solvers**

1335 In this section we briefly outline the numerical methods we use in the experimental section
1336 below. We consider both discontinuous Galerkin finite element solvers and finite difference solvers.
1337 In all the experiments we always use the trapezoidal rule to compute the integral in the WaveHoltz
1338 iteration.

1339 **3.4.1 The Energy Based Discontinuous Galerkin Method**

1340 Our spatial discretization is a direct application of the formulation described for general
1341 second order wave equations in [9, 10]. Here we outline the spatial discretization for the special
1342 case of the scalar wave equation in one dimension and refer the reader to [9] for the general case.

1343 The energy of the scalar wave equation is

$$H(t) = \int_D \frac{v^2}{2} + G(x, w_x) dx,$$

1344 where

$$G(x, w_x) = \frac{c^2(x)w_x^2}{2},$$

1345 is the potential energy density, v is the velocity (not to be confused with the iterates v^n above)
1346 or the time derivative of the displacement, $v = w_t$. The wave equation, written as a second order
1347 equation in space and first order in time then takes the form

$$w_t = v,$$

$$v_t = -\delta G,$$

1348 where δG is the variational derivative of the potential energy

$$\delta G = -(G_{w_x})_x = -(c^2(x)w_x)_x.$$

1349 For the continuous problem the change in energy is

$$\frac{dH(t)}{dt} = \int_D vv_t + w_t(c^2(x)w_x)_x dx = [w_t(c^2(x)w_x)]_{\partial D}, \quad (3.30)$$

1350 where the last equality follows from integration by parts together with the wave equation. Now,
1351 a variational formulation that mimics the above energy identity can be obtained if the equation
1352 $v - w_t = 0$ is tested with the variational derivative of the potential energy. Let Ω_j be an element and
1353 $\Pi^s(\Omega_j)$ be the space of polynomials of degree s , then the variational formulation on that element
1354 is:

1355 **Problem 2.** Find $v^h \in \Pi^s(\Omega_j)$, $w^h \in \Pi^r(\Omega_j)$ such that for all $\psi \in \Pi^s(\Omega_j)$, $\phi \in \Pi^r(\Omega_j)$

$$\int_{\Omega_j} c^2 \phi_x \left(\frac{\partial w_x^h}{\partial t} - v_x^h \right) dx = [c^2 \phi_x \cdot n (v^* - v^h)]_{\partial \Omega_j}, \quad (3.31)$$

$$\int_{\Omega_j} \psi \frac{\partial v^h}{\partial t} + c^2 \psi_x \cdot w_x^h dx = [\psi (c^2 w_x)^*]_{\partial \Omega_j}. \quad (3.32)$$

1356 Let $[[f]]$ and $\{f\}$ denote the jump and average of a quantity f at the interface between two
1357 elements, then, choosing the numerical fluxes as

$$\begin{aligned} v^* &= \{v\} - \tau_1 [[c^2 w_x]] \\ (c^2 w_x)^* &= \{c^2 w_x\} - \tau_2 [[v]], \end{aligned}$$

1358 will yields a contribution $-\tau_1([[c^2 w_x]])^2 - \tau_2([[v]])^2$ from each element face to the change of the
1359 discrete energy

$$\frac{dH^h(t)}{dt} = \frac{d}{dt} \sum_j \int_{\Omega_j} \frac{(v^h)^2}{2} + G(x, w_x^h).$$

1360 Physical boundary conditions can also be handled by appropriate specification of the numerical
1361 fluxes, see [9] for details. The above variational formulation and choice of numerical fluxes results
1362 in an energy identity similar to (3.30). However, as the energy is invariant to certain transformations
1363 the variational problem does not fully determine the time derivatives of w^h on each element and
1364 independent equations must be introduced. In this case there is one invariant and an independent
1365 equation is $\int_{\Omega_j} \left(\frac{\partial w^h}{\partial t} - v^h \right) = 0$. For the general case and for the elastic wave equation see [9] and
1366 [10].

1367 In this chapter we always choose $\tau_i > 0$ (so-called upwind or Sommerfeld fluxes) and we
1368 always choose the approximation spaces to be of the same degree $r = s$. These choices result in
1369 methods that are $r + 1$ order accurate in space.

1370 **3.4.2 Symmetric Interior Penalty Discontinuous Galerkin Method**

In addition to the above energy DG method, we also consider the Symmetric Interior Penalty DG (SIPDG) discretization, [62], for examples in two dimensions. The bilinear form in this case is

$$a_h(u, v) = \sum_{K \in \mathcal{T}_h} \int_K c^2 \nabla u \cdot \nabla v \, dx - \sum_{f \in \mathcal{F}_h} \int_F [[u]] \cdot \{c^2 \nabla v\} \, ds - \sum_{f \in \mathcal{F}_h} \int_F [[v]] \cdot \{c^2 \nabla u\} \, ds \\ + \sum_{f \in \mathcal{F}_h} \int_F \gamma h_F^{-1} c^2 [[u]] \cdot [[v]] \, ds,$$

1371 where \mathcal{T}_h is a collection of triangular elements, \mathcal{F}_h is the collection of element faces, h_F is the
1372 diameter of the edge or face F , and γ is the interior penalty stabilization parameter which must
1373 be chosen to be sufficiently large to ensure the system is positive-definite.

1374 **3.4.3 Finite Difference Discretizations**

1375 For the finite difference examples in a single dimension, we consider discretizations by uniform
1376 grids $x_i = x_L + ih_x$, with $i = -1, \dots, n+1$ and $h_x = (x_R - x_L)/n$. To impose impedance boundary
1377 conditions of the form $w_t \pm \vec{n} \cdot \nabla w = 0$ we evolve the wave equation as a first order system in time
1378 according to the semi-discrete approximation

$$\frac{dv_i(t)}{dt} = (D_+ D_-) w_i, \quad (3.33)$$

$$\frac{dw_i(t)}{dt} = v_i, \quad (3.34)$$

1379 and for the boundaries we find the ghost point values by enforcing

$$v_0 - D_0 w_0 = 0, \quad v_n - D_0 w_n = 0. \quad (3.35)$$

1380 Here we have used the standard forward, backward and centered finite difference operators, for
1381 example $hD_+ w_i = w_{i+1} - w_i$ etc.

1382 **3.4.4 Time Discretization**

1383 For some of the numerical examples in a single dimension, we use either an explicit second
 1384 order accurate centered discretization of w_{tt} or use the higher order corrected ME methods described
 1385 in Section 3.3.

1386 For the DG discretizations we employ Taylor series time-stepping in order to match the order
 1387 of accuracy in space and time. Assuming that all the degrees of freedom have been assembled into
 1388 a vector \mathbf{w} we can write the semi-discrete method as $\mathbf{w}_t = Q\mathbf{w}$ with Q being a matrix representing
 1389 the spatial discretization. Assuming we know the discrete solution at the time t_n we can advance
 1390 it to the next time step $t_{n+1} = t_n + \Delta t$ by the simple formula

$$\begin{aligned}\mathbf{w}(t_n + \Delta t) &= \mathbf{w}(t_n) + \Delta t \mathbf{w}_t(t_n) + \frac{(\Delta t)^2}{2!} \mathbf{w}_{tt}(t_n) \dots \\ &= \mathbf{w}(t_n) + \Delta t Q \mathbf{w}(t_n) + \frac{(\Delta t)^2}{2!} Q^2 \mathbf{w}(t_n) \dots\end{aligned}$$

1391 The stability domain of the Taylor series which truncates at time derivative number N_T includes
 1392 the imaginary axis if $\text{mod}(N_T, 4) = 3$ or $\text{mod}(N_T, 4) = 0$. However as we use a slightly dissipative
 1393 spatial discretization the spectrum of our discrete operator will be contained in the stability domain
 1394 of all sufficiently large choices of N_T (i.e. the N_T should not be smaller than the spatial order of
 1395 approximation). Note also that the stability domain grows linearly with the number of terms.

1396 **3.5 Numerical Examples**

1397 In this section we illustrate the properties of the proposed iteration and its Krylov accelerated
 1398 version by a sequence of numerical experiments in one and two spatial dimensions.

1399 **3.5.1 Examples in One Dimension**

1400 **3.5.1.1 Convergence Rate for Impedance Boundary Conditions**

1401 In [14], an application of Weyl asymptotics [109] revealed that the minimal relative gap to
 1402 resonance, $\delta = \min_j |\omega - \lambda_j|/\omega$ where λ_j^2 are the eigenvalues of the Laplacian, shrinks as ω^{-d} where

1403 d is the spatial dimension of the Helmholtz problem of interest. Analysis of the symmetric, positive
 1404 definite formulation of the iteration then yielded a convergence rate of $1 - \mathcal{O}(\delta^2) \approx 1 - \mathcal{O}(\omega^{-2d})$.
 1405 However, numerical experiments with Helmholtz problems with certain open/outflow boundary
 1406 conditions suggest a much more attractive convergence rate than the unacceptable $1 - \mathcal{O}(\omega^{-2d})$
 1407 rate. A natural question then is whether or not this seemingly pessimistic convergence rate can be
 1408 observed for outflow boundary conditions which are much more common in practical applications.

To that end, we consider a set of sample Helmholtz problems in a single spatial dimension with a constant (normalized) speed of sound, $c = 1$, in the domain $0 \leq x \leq 2$ where we impose the impedance boundary condition

$$w_t + \vec{n} \cdot w_x = 0,$$

which we note is equivalent to the Sommerfeld radiation condition. The Helmholtz problem under consideration has no forcing and so $f = 0$. In this case the solution is not unique (we have $\sin(\omega x)$ and $\cos(\omega x)$ as solutions), but we nevertheless may apply the WaveHoltz iteration. We formulate the wave equation in first order form and apply the extended iteration (3.7) since the boundary conditions do not conserve energy. The Laplacian is discretized with a standard three-point finite difference approximation, and a fourth order Taylor scheme is used for timestepping. We define the initial conditions as

$$v_0(x) = \sin(\omega x) - \frac{1}{2} (\sin((\omega + 2\pi)x) + \sin((\omega - 2\pi)x)), \quad v_1(x) = -\frac{d}{dx}v_0(x),$$

1409 which are shown in Figure 3.1.

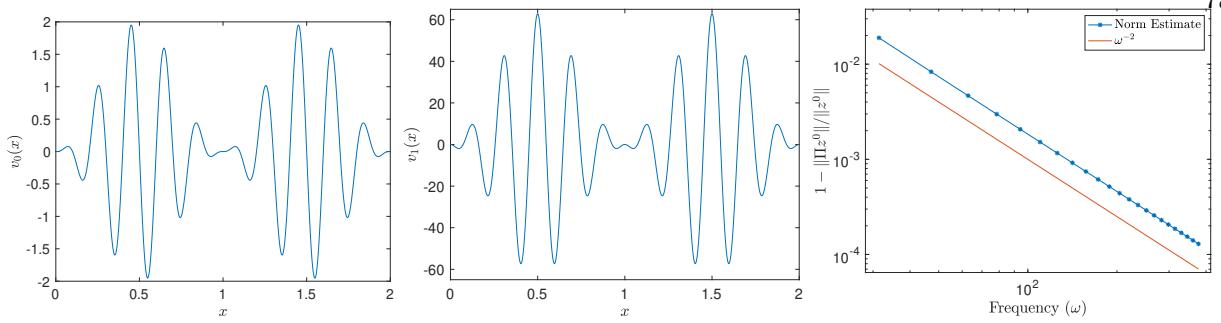


Figure 3.1: (Left, Middle) The initial conditions v_0 and v_1 for a Helmholtz frequency of $\omega = 10\pi$. (Right) The estimate of the quantity $1 - \|\mathcal{S}\|$ with increasing Helmholtz frequency ω .

1410 By definition, $\|\mathcal{S}\|_2 = \sup_{\|z\|_2 \neq 0} \|\mathcal{S}z\|_2 / \|z\|_2 \geq \|\mathcal{S}z^0\|_2 / \|z^0\|_2$ so that if $\|\mathcal{S}z^0\|_2 / \|z^0\| \approx 1 -$
 1411 $\mathcal{O}(\omega^{-2})$ is observed then the estimate of the spectral radius of the fixed point operator \mathcal{S} is tight
 1412 even for the problem with impedance boundary conditions. We consider a sweep of Helmholtz
 1413 frequencies $\omega = 10\pi, 15\pi, 20\pi, \dots, 120\pi$ with fifty points per wavelength and a CFL number of
 1414 1/10 for the solution of the wave equation. The results of this experiment are shown in Figure 3.1.

1415 On the left of Figure 3.1 we see the first part of the initial condition v^0 for a frequency of
 1416 $\omega = 10\pi$. We note that this specific initial condition is constructed such that it is close to a resonant
 1417 mode - which the filter-transfer function β weakly damps - as well as being close to zero at the
 1418 boundary so that a negligible amount of energy exits the system due to the impedance boundary
 1419 conditions in a single iteration. These two defining characteristics of the initial condition lead to
 1420 the norm estimate of the fixed-point iteration operator \mathcal{S} on the right of Figure 3.1. We observe
 1421 that the norm of \mathcal{S} does indeed approach unity at a rate of ω^{-2} , as predicted by theory. Thus, while
 1422 the preceding analysis “artificially” leveraged energy-conserving boundary conditions to obtain an
 1423 estimate of the convergence rate for open problems, it is possible to realize the ‘worst-case’ rate
 1424 implied by the energy-conserving regime.

1425 **Remark 3.5.1.** *We note that the estimate for the convergence rate is a **pointwise** estimate. Re-*
 1426 *peated application of the fixed-point iteration will (eventually) remove the modes close to resonance*
 1427 *and a faster convergence rate is observed. In Figure 3.2 we repeat the above experiment for the*

1428 frequencies $\omega = 10\pi, 40\pi$, and 70π but continue the iteration until the iterates converge to the zero
 1429 solution.

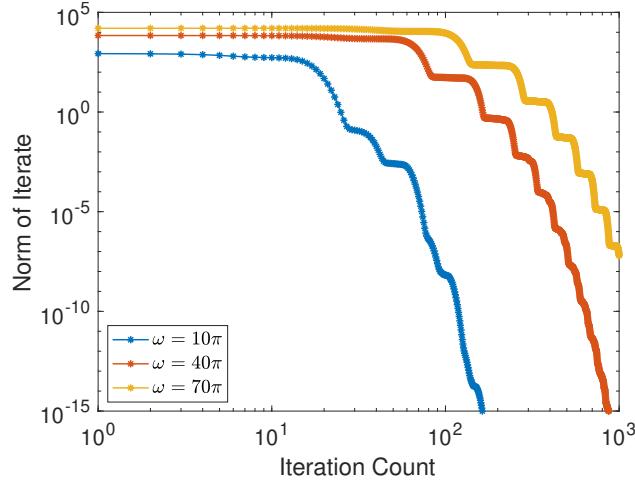


Figure 3.2: The norm of WaveHoltz iterates for increasing Helmholtz frequencies of $\omega = 10\pi, 40\pi$, and 70π for the adversarial example of Figure 3.1.

1430 We observe that after an initial phase the rate of convergence of the iterates to the solution increases
 1431 significantly since the data has propagated and exited the domain. We believe that the average
 1432 behavior over many fixed-point iterations leads to the much more attractive rates seen in the Krylov-
 1433 accelerated numerical experiments of [14]. Moreover, this example was pathologically constructed
 1434 and we note that so far we have been unable to construct initial conditions to realize the worst-case
 1435 rate in higher than one dimension.

1436 Assuming radially symmetric solutions to the Helmholtz equation, it is possible to cast higher
 1437 dimensional problems as 1D problems. Specifically, the wave equation in cylindrical/spherical
 1438 coordinates is

$$\begin{aligned} \frac{\partial^2 w}{\partial t^2} &= \frac{\partial^2 w}{\partial r^2} + \frac{\alpha}{r} \frac{\partial w}{\partial r}, \quad r \in \Omega, \quad 0 \leq t \leq T, \\ \frac{\partial w}{\partial r}(t, 0) &= 0, \quad 0 \leq t \leq T, \\ \frac{\partial w}{\partial n}(t, r) + \frac{\partial w}{\partial t}(t, r) &= 0, \quad x \in \partial\Omega, \end{aligned}$$

1439 where $\alpha = 1$ corresponds to cylindrical coordinates and $\alpha = 2$ corresponds to spherical. We use
 1440 a second order finite difference discretization (see [89] for details) with $\Omega = [0, 1]$. The initial
 1441 condition is analogous to the previous example,

$$v_0(r) = \sin(\omega(r+1)) - \frac{1}{2} (\sin((\omega + 2\pi)(r+1)) + \sin((\omega - 2\pi)(r+1))), \quad v_1(r) = -\frac{d}{dr}v_0(r).$$

1442 We consider a set of frequencies $10\pi, 11\pi, \dots, 30\pi$ and use fifty points per wavelength in the com-
 1443 putation with a CFL of 1/100. Below we show the results of the experiment in Figure 3.3.

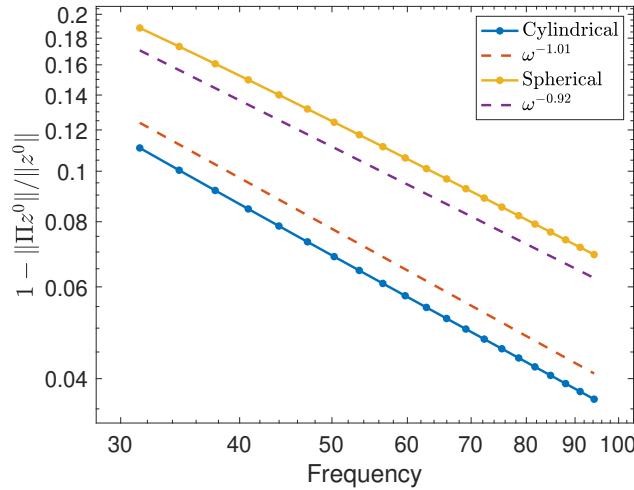


Figure 3.3: The estimate of the quantity $1 - \|\mathcal{S}\|$ with increasing Helmholtz frequency ω for a radially symmetric initial condition.

1444 From Figure 3.3 we observe that the norm of \mathcal{S} approaches unity at a nearly linear rate in
 1445 the frequency ω in 2D and a sublinear rate for the 3D problem, both of which are more favorable
 1446 than the quadratic rate in a single spatial dimension.

1447 **Remark 3.5.2.** *From Figure 3.3 it is clear that with a fixed discretization and initial condition,
 1448 the convergence rate improves with increasing dimension. This is perhaps unsurprising given an
 1449 increase in the local energy decay rate for the wave equation from two to three dimensions, along
 1450 with a richer set of directions in which waves may propagate and leave the domain.*

1451 **3.5.1.2 Time Discretization**

We consider solving the Helmholtz equation with $c = 1$ and constant exact solution

$$u(x) = 1, \quad 0 \leq x \leq 1.$$

1452 We take the frequency to be $\omega = 1$ and consider Dirichlet boundary conditions. We discretize the
1453 Laplacian with the standard three-point finite difference stencil and note that there is no error (aside
1454 from truncation errors) in the solution by a direct solution of the discrete Helmholtz equation. We
1455 use a centered modified equation timestepping scheme of both second and fourth order, with both
1456 the original frequency and a modified frequency $\tilde{\omega}$ with corresponding quadrature to remove time
1457 discretization errors. We use the WaveHoltz iteration as a fixed-point iteration with a convergence
1458 criterion that the relative L_2 norm between successive iterations is smaller than 10^{-13} .

1459 Using the original frequency in the calculation, we see from Figure 3.4 that the WaveHoltz
1460 solution converges to the discrete Helmholtz solution with the same order as that of the timestep
1461 scheme used.

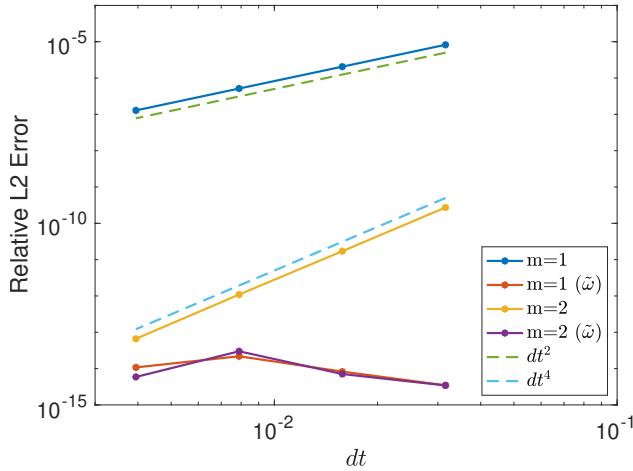


Figure 3.4: Convergence of the discrete WaveHoltz solution to the true solution of the discrete Helmholtz problem with fixed spatial discretization. Solid lines indicate relative errors between discrete solutions. The blue and yellow solid lines indicate relative errors between the usual discrete WaveHoltz solution and the true solution, and the red and purple solid lines indicate relative errors for the frequency corrected solution.

¹⁴⁶² With the modified frequency and quadrature, however, we see that the WaveHoltz iteration con-
¹⁴⁶³ verges to the discrete Helmholtz solution up to roundoff errors.

¹⁴⁶⁴ **Remark 3.5.3.** *While only centered timestepping schemes are presented here, this approach can*
¹⁴⁶⁵ *be extended to arbitrary timesteppers. A careful discrete analysis of the iteration isolated to a*
¹⁴⁶⁶ *single eigenmode of the wave solution reveals what the modified frequency should be, and a modified*
¹⁴⁶⁷ *quadrature as outlined above removes the time discretization error from the converged WaveHoltz*
¹⁴⁶⁸ *solution. Thus, the choice of a timestepper need not need be restricted to have the same order as*
¹⁴⁶⁹ *the spatial discretization. With a corrected scheme it may be more advantageous to take as large a*
¹⁴⁷⁰ *timestep as possible with a low order timestepper.*

1471 **3.5.1.3 Convergence Rate for Damped Helmholtz Equations**

To study how the number of iterations scale with the Helmholtz frequency ω we solve the wave equation on the domain $x \in [-6, 6]$ with constant wave speed $c^2(x) = 1$ and with a forcing

$$f(x) = \omega^2 e^{-(\omega x)^2},$$

1472 that results in the solution being $\mathcal{O}(1)$ for all ω . We discretize using the energy based DG method
1473 discussed above and use upwind fluxes which adds a small amount of dissipation. We keep the
1474 number of degrees of freedom per wave length fixed by letting the number of elements be $5[\omega]$. We
1475 always take the polynomial degree to be 7, the number of Taylor series terms in the timestepping
1476 to be 6, and use WHI accelerated by GMRES without restarts.

1477 We report the number of iterations it takes to reach a GMRES residual smaller than 10^{-10} for
1478 the six possible combinations of Dirichlet, Neumann and impedance boundary conditions for 200
1479 frequencies distributed evenly from 1 to 100. The results for three levels of damping are displayed
1480 in Figure 3.5.

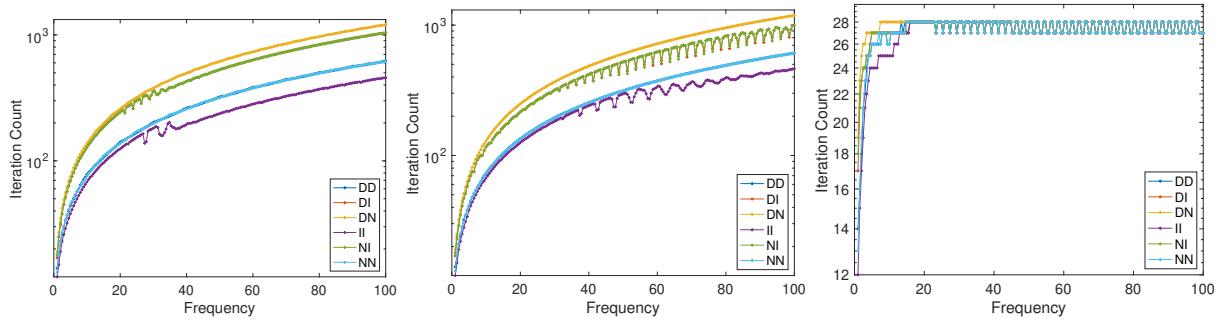


Figure 3.5: Number of iterations as a function of ω for different boundary conditions and damping parameters. Left: $\eta = 1/2\omega$, Middle: $\eta = 1/2$, Right: $\eta = \omega/2$. In the above legends each entry is made up of a two letter string where the first letter indicates the boundary condition on the left at $x = -6$, and the second letter indicates the boundary condition on the right at $x = 6$. Here D indicates Dirichlet, N indicates Neumann, and I indicates impedance/Sommerfeld conditions.

1481 On the left and middle of Figure 3.5 are damping parameters of $1/2\omega$ and $1/2$ respectively, from

which it is clear that the scaling is sub-linear with increasing frequency. On the right in Figure 3.5 are results from a damping parameter that grows with frequency, $\omega/2$, which demonstrates a number of iterations that is both frequency independent and modest for a given GMRES tolerance. Interestingly, in this case the curve for each set of boundary conditions collapses to the same curve so that the iteration is insensitive to boundary conditions for a sufficiently large damping parameter.

Remark 3.5.4. *As seen in the previous chapter, the impedance-impedance conditions take the fewest iterations to reach convergence for lower levels of damping. We point out the preceding analysis assumes energy-conserving boundary conditions to obtain estimates on the convergence rate of WaveHoltz as a fixed-point iteration. A different approach without the need for a Laplacian with a point-spectrum is needed to obtain rates depending on the specific boundary conditions.*

3.5.2 Examples in Two Dimensions

In this section we present experiments in two space dimensions. For the following examples, we consider solving the Helmholtz equation for the wedge model which we adapt from [45, 95]. The domain is the rectangle $[0, 600] \times [0, 1000]$ with the (discontinuous) speed of sound

$$c(x) = \begin{cases} c_1 = 2100, & y \leq x/6 + 400, \\ c_2 = 1000, & x/6 + 400 \leq y \leq 800 - x/3, \\ c_3 = 2900, & \text{else.} \end{cases}$$

The domain and mesh used for the examples is shown in Figure 3.6, where the blue region corresponds to c_1 , the green region with c_2 , and the magenta region with c_3 .

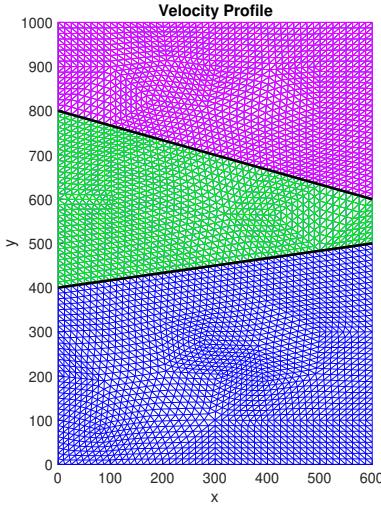


Figure 3.6: Computational domain where the mesh in blue corresponds to a wavespeed of $c = 2100$, the mesh in green corresponds to a speed of $c = 1000$, and the magenta mesh with $c = 2900$. The solid black line is not physical and is meant to more easily distinguish between regions.

On the boundary of the rectangle we impose the impedance boundary condition $w_t + c\nabla w \cdot \vec{n} = 0$. For the spatial discretization we use the SIPDG method with a penalty parameter choice of $\gamma = (p+1)^2$, where $p = 4$ is the polynomial order used in each element which results in a fifth order method. In time we use a fourth order Taylor method for timestepping. For each example, we use the point-source

$$f(x, y) = \omega^2 \delta(|x - x_0|) \delta(|y - y_0|),$$

where $x_0 = 300$, $y_0 = 0$, ω is the Helmholtz frequency, and $\delta(z)$ is the usual Dirac delta function.

These examples were implemented in the MFEM finite element discretization library [4].

3.5.2.1 Convergence for Damped Helmholtz Equations

We again study how the number of GMRES accelerated WHI iterations scale with the Helmholtz frequency ω for the exemplary wedge problem.

We report the number of iterations it takes to reach a GMRES residual smaller than 10^{-10} for the frequencies $1, 2, \dots, 100$, with damping $\eta = \omega/2$ with either impedance or Neumann conditions

on all sides of the rectangular domain. The results for this experiment are shown in Figure 3.7, from which it is clear that the number of iterations is essentially independent of frequency for larger frequencies as was the case in a single spatial dimension.

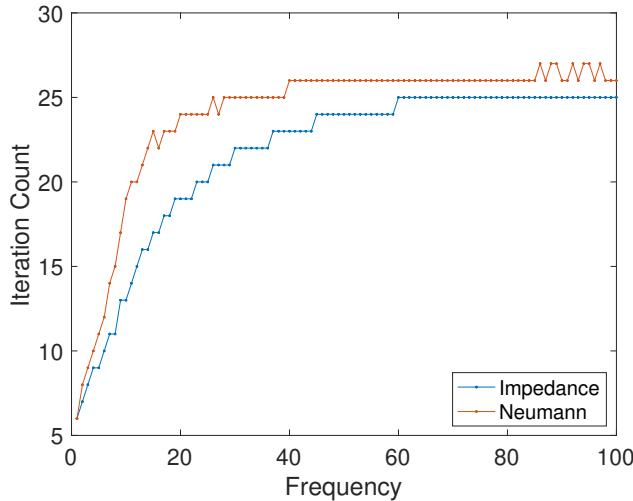


Figure 3.7: Number of iterations to reach a GMRES tolerance of 10^{-10} for the wedge problem in 2D with all Neumann or all impedance boundary conditions.

We again note that energy-conserving boundary conditions require more iterations than the impedance case even in the presence of damping.

For a final example, in Figure 3.8 we display the solution of the damped (and undamped) Helmholtz equation using the GMRES accelerated WHI for a frequency of $\omega = 40\pi$ with damping $\eta = \omega/2$ and 0, respectively.

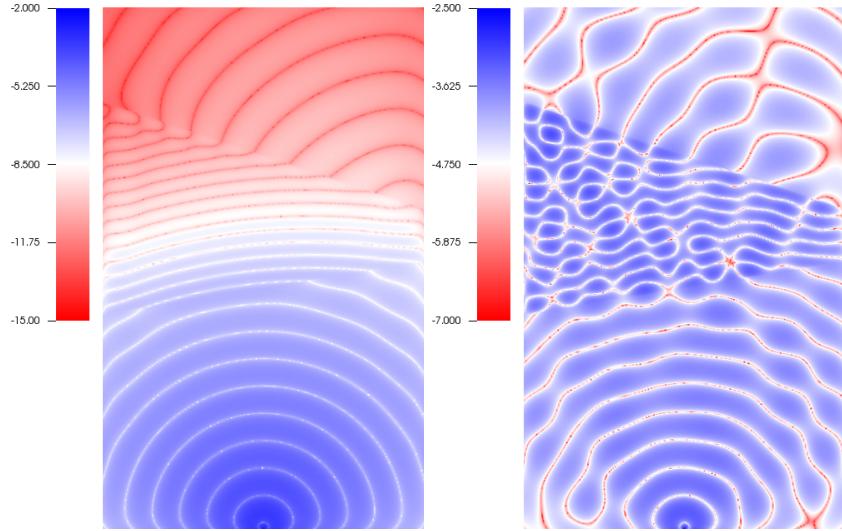


Figure 3.8: In the above we plot the \log_{10} of the absolute value of the real part of the Helmholtz solution with frequency $\omega = 40\pi$ for (Left) damping parameter $\eta = 20\pi$ and (Right) no damping.

1510 **3.6 Summary and Future Work**

1511 We have presented and extended analysis of the WaveHoltz iteration, an iterative method for
1512 solving the Helmholtz equation, applied to wave equations with and without damping. The general
1513 iteration has the same rate of convergence as the energy-conserving case presented in [14], but is
1514 a more general and appropriate formulation for considering problems with impedance/Sommerfeld
1515 boundary conditions. For problems with damping, the WaveHoltz iteration always converges and
1516 numerical experiments verify the frequency independent convergence of problems with sufficient
1517 levels of damping.

1518 We have provided analysis of the interior impedance problem in a single dimension and
1519 constructed an example in which the worst-case convergence rate is realized, despite the numerical
1520 results of the previous chapter indicating much more favorable scaling for non-energy conserving
1521 boundary conditions.

1522 Finally, here we have only considered acoustic wave propagation. In the following chapter we
1523 will apply the WaveHoltz iteration to the elastic Helmholtz equation. Moreover, we have not yet
1524 tried to leverage sweeping/domain decomposition ideas here and hope to study the numerical and

₁₅₂₅ theoretical properties of these in the future.

Chapter 4

El WaveHoltz Method

1528 Time harmonic wave propagation problems are notoriously difficult to solve by direct or
 1529 iterative methods due to the resolution requirements and the indefinite nature of the differential
 1530 operator, especially at high frequencies. In acoustic media, a prototypical model of time harmonic
 1531 wave propagation is given by the (heterogeneous) Helmholtz equation

$$\nabla \cdot (c^2(\mathbf{x}) \nabla u) + \omega^2 u = f(\mathbf{x}), \quad \mathbf{x} \in \Omega, \quad (4.1)$$

1532 for a domain Ω , frequency ω , and sound speed $c^2(\mathbf{x})$. The efficient solution of the Helmholtz
 1533 equation (4.1) via iterative methods is an active area of research with a variety of methods in both
 1534 the frequency and time domain. We refer to Chapter 2 for a more in-depth overview of the literature
 1535 on techniques for solving the Helmholtz equation, as well as the review articles [47, 50, 44].

For applications in solid mechanics, seismology and geophysics, however, it is more appropriate to consider the elastic wave equation instead of the acoustic wave equation. In contrast to the literature for the acoustic case, fewer effective solvers and preconditioners are available for time harmonic elastic waves governed by the elastic “Helmholtz” equation (also known as the Navier equation)

$$\rho \omega^2 \mathbf{v} + \nabla \cdot \mathcal{T}(\mathbf{v}) = \mathbf{f}(\mathbf{x}), \quad \mathbf{x} \in \Omega.$$

1536 Here $\mathbf{v} \in \mathbb{R}^d$ is the displacement vector and d the spatial dimension. The elastic wave equation
 1537 models both pressure and shear waves and, as is the case for the Helmholtz equation (4.1), the
 1538 above system of equations results in a discretization that is highly indefinite for large frequencies.

1539 As for any wave propagation problem the resolution must increase with the frequency, and here
 1540 the most stringent resolution constraint comes from the (shorter) shear wave wavelength. This, in
 1541 tandem with d times the number of unknowns leading to larger storage requirements, necessitates
 1542 parallel, memory lean, and scalable solvers that must be high order accurate to mitigate dispersive
 1543 errors, [75], causing the so-called *pollution effect* [16].

1544 While most methods have traditionally focused on solving the Helmholtz equation in the
 1545 frequency domain (we provide a review of some of these below), an alternative approach is to instead
 1546 construct iterative solvers in the time domain. One such method, the so-called Controllability
 1547 Method (CM), was first proposed by Bristeau et al. [29] and has recently received renewed interest
 1548 in a series of papers by Grote [63, 85, 33]. The CM was extended to elastic media in [87]. The
 1549 unknown in the CM is the initial data to the wave equation. In the CM this initial data is adjusted
 1550 so that it produces an approximation to the Helmholtz equation by solving a constrained least-
 1551 squares minimization where the objective function measures the deviation from time-periodicity.
 1552 The minimization can be efficiently implemented using the conjugate gradient method, where the
 1553 gradient is computed by solving the adjoint wave equation backwards in time.

1554 The method we present here is an extension of the WaveHoltz method introduced in [14]
 1555 for the scalar wave equation. As in the CM, the WaveHoltz iteration (and the Elastic version we
 1556 denote El WaveHoltz) iteratively updates the initial data to the wave equation but it does so by
 1557 filtering the wave equation solution over one period (or an integer number of periods). The filtered
 1558 solution is then used as the next initial data and thus the WaveHoltz method only requires one
 1559 wave solve per iteration while the CM requires two.

1560 In [14] we show that the (linear) iteration is convergent in both the continuous and discretized
 1561 setting and that, if formulated as a linear system of equations, the underlying matrix is positive
 1562 definite. We also showed that for closed waveguides with energy conserving boundary conditions
 1563 (Dirichlet or Neumann) the matrix is also symmetric as long as the numerical method is symmetric.

1564 We emphasize that the filter used in the WaveHoltz method is a bounded operator and
 1565 therefore the number of iterations (and the condition number of the problem) does **not** depend on

1566 the gridsize h . This is in contrast to methods that discretize and solve the PDE directly. Such
 1567 methods typically have a condition number that scales as h^{-2} which makes it increasingly difficult
 1568 to solve the problem as the solution becomes more accurate.

1569 The analysis in [14] predicted that the WaveHoltz method in d dimensions converges to a fixed
 1570 tolerance in $\mathcal{O}(\omega^d)$ iterations for energy conserving problems and numerical experiments indicated
 1571 that it converges in $\mathcal{O}(\omega)$ iterations for open problems. Numerical experiments also indicated that
 1572 some energy conserving examples may exhibit complexity closer to $\mathcal{O}(\omega^{d-0.5})$ for $d = 2$ and 3. The
 1573 analytical predictions from [14] are expected to hold here as well and in the experiments we carry
 1574 out below we observe $\mathcal{O}(\omega^d)$. All these results and observations are independent of grid resolution
 1575 indicating that our method can be particularly suitable when accurate solutions are required.

1576 In this chapter we focus solely on energy conserving boundary conditions (Dirichlet or normal
 1577 stress) and leave the cases of impedance and non-reflecting boundary conditions to future work. In
 1578 addition to introducing El WaveHoltz, we present several new results that are also retroactively ap-
 1579 plicable to our earlier work on WaveHoltz for the scalar wave equation [14] and Maxwell's equations
 1580 [93].

1581 First, for the energy conserving boundary conditions the continuous WaveHoltz operator
 1582 is symmetric positive definite. However, unless the semidiscretized wave equation has the form
 1583 $\mathbf{u}_{tt} = \mathbf{L}_h \mathbf{u}$ with \mathbf{L}_h SPD the discrete WaveHoltz iteration will not result in a symmetric matrix.
 1584 We show that for schemes that are symmetric in a weighted inner product there is a simple scaling
 1585 that can be applied to make the discrete WaveHoltz method symmetric. This then allows the
 1586 conjugate gradient or the conjugate residual method to be used.

1587 Following the ideas of Stolk [101] we introduce two new two-level time-stepping schemes – one
 1588 explicit and one implicit – that remove the time-stepping error from the WaveHoltz solution. When
 1589 either of these time-stepping methods are used the solution to the discrete WaveHoltz method is
 1590 identical to the solution obtained by directly discretizing the frequency domain equation.

1591 For high frequency, large scale problems, parallel solution of the time harmonic elastic equa-
 1592 tions is the only feasible option. For a parallel solver to scale well the ratio of communication to

1593 computation should be small. In general, there are two types of communication: a) the local com-
 1594 munications between processors to exchange local degrees of freedom needed for stencil operations
 1595 in the discretization of spatial derivatives, and b) global all-to-all operations such as computing
 1596 the inner product between two global vectors. The WaveHoltz method has an intrinsic advantage
 1597 compared to methods that work directly with the frequency domain equation in that the all-to-all
 1598 communication that is required to update search directions in CG, GMRES etc. only needs to be
 1599 computed once per $T = 2\pi/\omega$ -period. Here we explore the effect of filtering over an additional
 1600 number of periods to further reduce the number of all-to-all communications.

1601 We believe that the method we propose here is an alternative to previously proposed methods.
 1602 In particular, El WaveHoltz is easily implemented if an elastic wave equation solver is already
 1603 available. As we show in the numerical experiments section, El WaveHoltz can be one to two
 1604 orders of magnitude faster compared to an algebraic multigrid (AMG) preconditioned GMRES
 1605 solver for the frequency domain equation when using the symmetric interior penalty discontinuous
 1606 Galerkin implementation available in MFEM [4]. There are, of course, many other solvers available;
 1607 the question of which method will be most efficient will (most likely) depend on the details of the
 1608 problem to be solved. We now review some of the methods available in the literature.

1609 One of the most common preconditioners for acoustic problems is the shifted Laplacian pre-
 1610 conditioner (SLP), a more thorough review of which can be found in the review article by Erlangga
 1611 [44]. Perhaps one of the first extensions of the damping preconditioner to elastic media was intro-
 1612 duced by Airaksinen et al. [3], in which a finite element spatial discretization for the damped op-
 1613 erator is inverted by AMG. A more traditional finite-difference multigrid SLP with line-relaxations
 1614 was considered by Rizzuti and Mulder [96]. For both of these previous approaches, the effectiveness
 1615 of a straightforward SLP is degraded for nearly incompressible media as the prolongation operators
 1616 struggle to approximate the nullspace of the grad-div operator. To address this, a more recent ex-
 1617 tension was done by Treister [102] in which a mixed-formulation of the elastic Helmholtz equation
 1618 is considered. While nearly incompressible media could be handled by the methods of [102], this
 1619 comes at the cost of doubling the number of unknowns as well as additional storage requirements

1620 for precomputing the inverse of relaxation operators.

1621 Another important class of methods for the solution of the Helmholtz equation are domain
 1622 decomposition (DD) methods, for which we refer the reader to [50] for a review. In the short
 1623 article [30], it was shown that a classic Schwarz DD with overlap for elastic problems converges for
 1624 high frequencies, diverges for medium frequencies, and stagnates for small frequencies. Moreover,
 1625 overlapping DD as a preconditioner for a GMRES accelerated solver exhibits convergence behavior
 1626 that depends strongly on the frequency ω with degrading performance for increasing frequency.
 1627 To remedy this, Brunet et al. introduced more general transmission conditions at the boundaries
 1628 of overlapping domains in [31]. These transmission conditions, together with a sufficiently large
 1629 enough overlap, yield convergence of the DD method for all frequencies with the exception of
 1630 $\{\omega/C_s, \omega/C_p\}$, where C_s and C_p are the shear and pressure wave speeds, respectively.

1631 For unbounded problems one of the most promising classes of preconditioners for the Helmholtz
 1632 equation are the so-called sweeping preconditioners by Engquist and Ying [40, 41]. These precondi-
 1633 tioners construct an LDL^T decomposition by sweeping through the domain layer-by-layer, with the
 1634 key observation that the application of the Schur complement matrices found in the block diagonal
 1635 matrix D is equivalent to solving a quasi-1D(2D) problem in 2D(3D). In contrast to the acous-
 1636 tic case, however, the sweeping preconditioner for time harmonic elastic waves, [103], exhibited
 1637 an increase in the number of iterations with frequency for a heterogeneous media as the moving
 1638 perfectly matched layer (PML) does not approximate Green's function as well. We note that the
 1639 stable construction of PML for many elastic problems is still considered an open question [21, 11].
 1640 Similar to the sweeping preconditioner, Belonosov et al. [22] construct a preconditioner in 3D with
 1641 damping that sweeps through the domain along a coordinate axis while additionally homogenizing
 1642 the medium in each layer. The preconditioner of [22] is inverted using FFT's and is accelerated
 1643 with BiCGSTAB in the outer loop. As with the sweeping preconditioner, the choice of sweeping
 1644 direction is important. Thus for problems where heterogeneity is present in all directions this
 1645 preconditioner is less effective. Yet another solver with a sweeping nature is an extension of the
 1646 Gordon and Gordon [60] CARP-CG method for Helmholtz problems to elastic media [80]. Despite

1647 its simplicity this method requires a large number of iterations, especially for heterogeneous media
 1648 or problems with higher Poisson ratios. It should be emphasized that, although successful for un-
 1649 bounded problems, the efficiency of sweeping methods for energy conserving boundary conditions
 1650 has largely not been demonstrated and their parallel implementation remains cumbersome.

1651 Instead of the LDL^T decomposition used by the sweeping preconditioner, other approaches
 1652 constructing LU/ILU factorizations and preconditioners are available. In [37] an ILU precondi-
 1653 tioner based on wavelet transforms with Gibbs reordering is used in a GMRES accelerated solver
 1654 (with restarts) for time harmonic elastic waves. Wang et al. introduced a structured multifrontal
 1655 algorithm using nested dissection based domain decomposition, together with hierarchical semi-
 1656 separable (HSS) compression for frontal matrices with low off-diagonal ranks in [108]. The use
 1657 of multilevel sequentially semi-separable (MSSS) matrix structure of the discretized elastic wave
 1658 equation on Cartesian grids was leveraged in [19] inside of an induced dimension reduction (IDR)
 1659 accelerated ILU preconditioner. The drawback of LU/ILU methods for elastic Helmholtz problems
 1660 is the growth in memory and storage requirements.

1661 The rest of this chapter is organized as follows. In Section 4.1 we present the elastic Helmholtz
 1662 and wave equations. In Section 4.2 we introduce the WaveHoltz iteration applied to elastic problems
 1663 with Dirichlet and/or free surface boundary conditions. In Section 4.3 we outline the numerical
 1664 methods used to solve the elastic wave equation and present new results on time-stepping and
 1665 Krylov acceleration. Numerical examples are presented in Section 4.4. Finally, we summarize and
 1666 conclude in Section 4.5.

1667 4.1 Governing Equations

1668 4.1.1 The Time Harmonic Elastic Wave Equation

1669 For a linear isentropic elastic media the frequency domain equation is

$$\rho\omega^2\mathbf{v} + \nabla \cdot \mathcal{T}(\mathbf{v}) = \mathbf{f}(\mathbf{x}), \quad \mathbf{x} \in \Omega. \quad (4.2)$$

1670 For notational convenience we will refer to this as the elastic Helmholtz or, when there is no
 1671 ambiguity, simply the Helmholtz equation. This equation can be obtained by making the ansatz
 1672 $\mathbf{u}(\mathbf{x}, t) = e^{i\omega t}\mathbf{v}(\mathbf{x})$ and inserting it into the elastic wave equation (discussed below). We note that,
 1673 in general, the Helmholtz solution \mathbf{v} is complex-valued. However, for boundary conditions that
 1674 conserve the energy (such as Dirichlet and conditions on the normal stress) the corresponding
 1675 elastic wave equation solution \mathbf{v} becomes real-valued. For real-valued solutions, the ansatz then
 1676 simplifies to $\mathbf{u}(\mathbf{x}, t) = \cos(\omega t)\mathbf{v}(\mathbf{x})$. The El WaveHoltz method can be used to find the solution \mathbf{v}
 1677 in both cases, but as we exclusively consider the energy conserving case here we primarily describe
 1678 the method for that case.

1679 **4.1.2 The Elastic Wave Equation**

1680 The linear elastic wave equation in an isentropic material described by the density $\rho(\mathbf{x}, t)$,
 1681 the Lamé parameters $\mu(\mathbf{x}) > 0$ and $\lambda(\mathbf{x}) > 0$, and with a time harmonic forcing takes the form

$$\rho\mathbf{u}_{tt} = \nabla \cdot \mathcal{T}(\mathbf{u}) - \cos(\omega t)\mathbf{f}(\mathbf{x}), \quad \mathbf{x} \in \Omega, \quad 0 \leq t \leq T. \quad (4.3)$$

1682 Here $\mathbf{u} = (u(\mathbf{x}, t), v(\mathbf{x}, t), w(\mathbf{x}, t))$ is the displacement vector, $\mathbf{x} = (x, y, z)^T$ is the Cartesian coordi-
 1683 nate and t is time. The stress tensor $\mathcal{T}(\mathbf{u})$ can be decomposed into

$$\mathcal{T}(\mathbf{u}) = \lambda(\nabla \cdot \mathbf{u})I + 2\mu\mathcal{D}(\mathbf{u}), \quad (4.4)$$

1684 where $\mathcal{D}(\mathbf{u})$ is the symmetric part of the displacement gradient

$$\mathcal{D}(\mathbf{u}) = \frac{1}{2} \begin{pmatrix} 2u_x & u_y + v_x & u_z + w_x \\ u_y + v_x & 2v_y & v_z + w_y \\ u_z + w_x & v_z + w_y & 2w_z \end{pmatrix}. \quad (4.5)$$

1685 The equation (4.3) is closed by boundary conditions specifying the displacement

$$\mathbf{u}(\mathbf{x}, t) = \cos(\omega t)\mathbf{g}(\mathbf{x}), \quad x \in \partial\Omega_D, \quad (4.6)$$

1686 or the normal stress

$$\mathcal{T}(\mathbf{u})\mathbf{n} = \cos(\omega t)\mathbf{h}(\mathbf{x}), \quad x \in \partial\Omega_S, \quad (4.7)$$

1687 along with initial conditions

$$\mathbf{u}(\mathbf{x}, 0) = \mathbf{u}_0(\mathbf{x}), \quad \frac{\partial \mathbf{u}(\mathbf{x}, 0)}{\partial t} = \mathbf{u}_1(\mathbf{x}). \quad (4.8)$$

1688 Multiplying (4.3) by \mathbf{u}^T , integrating over Ω and invoking the divergence theorem yields the

1689 energy estimate

$$\frac{1}{2} \frac{d}{dt} \left(\|\sqrt{\rho} \mathbf{u}_t\|^2 + \int_{\Omega} \lambda (\nabla \cdot \mathbf{u}) I + 2\mu (\mathcal{D} : \mathcal{D}) d\mathbf{x} \right) = - \int_{\Omega} \cos(\omega t) \mathbf{u}^T \mathbf{f}(\mathbf{x}) d\mathbf{x} + \int_{\partial\Omega} \mathbf{u}_t^T \mathcal{T}(\mathbf{u}) \mathbf{n} dS. \quad (4.9)$$

1690 Here \mathbf{n} is the outward unit normal and the notation $(\mathcal{A} : \mathcal{B}) = \sum_{i=1}^d \sum_{j=1}^d a_{i,j} b_{i,j}$ is the standard
1691 tensor contraction over two indices.

1692 Thus, when there is no forcing, $\mathbf{f}(\mathbf{x}) = 0$, the energy is conserved in time as long as $\mathbf{u}_t^T \mathcal{T}(\mathbf{u}) \mathbf{n} = 0$
1693 on the boundary $\partial\Omega$. The condition $\mathcal{T}(\mathbf{u}) \mathbf{n} = 0$ indicates that the boundary is stress free or free
1694 of traction. The Dirichlet condition on the velocity $\mathbf{u}_t = 0$ also holds if the displacement vanishes
1695 for all time on the boundary, i.e. $\mathbf{u} = 0$.

1696 Note that if the initial data, $\mathbf{u}_0(\mathbf{x})$, gives rise to a solution of the form $\mathbf{u}(\mathbf{x}, t) = \cos(\omega t) \mathbf{v}(\mathbf{x})$
1697 then that solution coincides with the elastic Helmholtz solution to (4.2).

1698 **Remark 1.** *In the rest of this chapter, unless otherwise noted, we will assume that the equations*
1699 *have been non-dimensionalized and that $\rho = 1$.*

1700 4.2 The El WaveHoltz Iteration

1701 The El WaveHoltz iteration is a direct generalization of the WaveHoltz iteration introduced
1702 and analyzed in [14]. Precisely, if we consider the energy conserving case, applying the Wave-
1703 Holtz operator component wise to the initial displacement vector \mathbf{u}_0 defines the El WaveHoltz
1704 operator

$$\Pi \mathbf{u}_0 = \frac{2}{T} \int_0^T \left(\cos(\omega t) - \frac{1}{4} \right) \mathbf{u}(x, t) dt. \quad (4.10)$$

1705 Here $T = \frac{2\pi}{\omega}$ and $\mathbf{u}(\mathbf{x}, t)$ is the solution to (4.3) with the initial data \mathbf{u}_0 (recall that for the energy
1706 conserving case we always have $\mathbf{u}_1 = \frac{\mathbf{u}(\mathbf{x}, 0)}{\partial t} = 0$).

1707 As the analysis of this operator is the same as that for the scalar operator analyzed in [14], we
 1708 will not repeat the analysis in detail here. Instead, we now highlight its most important properties.
 1709 The first thing to note is that if $\mathbf{u}(\mathbf{x}, t) = \cos(\omega t)\mathbf{v}(\mathbf{x})$ (and thus $\mathbf{u}_0(\mathbf{x}) = \mathbf{v}(\mathbf{x})$), then the integral
 1710 in (4.10) can trivially be evaluated

$$\Pi\mathbf{v}(x) = \frac{2}{T} \int_0^T \left(\cos(\omega t) - \frac{1}{4} \right) \cos(\omega t)\mathbf{v}(x) dt = \mathbf{v}(x), \quad (4.11)$$

1711 showing that the elastic Helmholtz solution is a fixed point of the operator. Further, if we let (λ_j^2, ϕ_j)
 1712 be the eigendecomposition satisfying $\lambda^2 \phi_j = \nabla \cdot \mathcal{T}(\phi_j)$, then for a general initial displacement the
 1713 solution will be on the form $\sum_{j=0}^{\infty} d_j \cos(\lambda_j t) \phi_j$. Let

$$\beta(\lambda) \equiv \frac{2}{T} \int_0^T \left(\cos(\omega t) - \frac{1}{4} \right) \cos(\lambda t) dt.$$

1714 Then as in [14] we can define the operator \mathcal{S} as

$$\mathcal{S} \sum_{j=0}^{\infty} d_j \phi_j \equiv \sum_{j=0}^{\infty} \beta(\lambda_j) d_j \phi_j,$$

1715 which gives the filtered solution to the elastic wave equation when $\mathbf{f} = 0$. If $\omega \neq \lambda$ then the spectral
 1716 radius $\max |\beta| < 1$ (see Lemma 2.1 in [14]) so the iteration will converge. Since the operator
 1717 is linear, we may find the fixed point (or equivalently the elastic Helmholtz solution) by solving
 1718 the equation $(\mathcal{I} - \mathcal{S})\mathbf{v} \equiv \mathcal{A}\mathbf{v} = \mathbf{b} \equiv \Pi\mathbf{0}$. As is the case for the scalar Helmholtz equation, the
 1719 eigenvalues of \mathcal{A} lie in $(0, 3/2)$ and the condition number scales with the frequency as $\text{cond}(\mathcal{A}) \sim \omega^{2d}$
 1720 in d dimensions.

1721 We emphasize that here \mathcal{A} is a self-adjoint, positive definite and bounded operator. Thus
 1722 once \mathcal{A} is discretized it will be possible to apply the conjugate gradient method. Moreover, as
 1723 the condition number *does not* depend on the discretization size, the number of iterations are not
 1724 expected to increase as the solution becomes more accurate due to grid refinement. We also note
 1725 that since $\text{cond}(\mathcal{A}) \sim \omega^{2d}$ the conjugate gradient method is expected to converge to a fixed tolerance
 1726 in ω^d iterations.

1727 Finally, as mentioned above it is possible to define the iteration as the integral over multiple
 1728 periods in order to reduce the number of all-to-all communication in the Krylov iteration. For
 1729 example, if the number of periods is K then we can define the filtering as

$$\Pi_K \mathbf{u}_0 = \frac{2}{KT} \int_0^{KT} \left(\cos(\omega t) - \frac{1}{4} \right) \mathbf{u} dt, \quad T = \frac{2\pi}{\omega}. \quad (4.12)$$

1730 **Remark 2.** For general boundary conditions (e.g. non-reflecting or impedance), $\frac{u(x,0)}{\partial t} = \mathbf{u}_1(\mathbf{x})$
 1731 will not be zero and we must seek the initial data \mathbf{u}_0 and \mathbf{u}_1 simultaneously. The El WaveHoltz
 1732 operator then is

$$\Pi \begin{bmatrix} \mathbf{u}_0 \\ \mathbf{u}_1 \end{bmatrix} = \frac{2}{T} \int_0^T \left(\cos(\omega t) - \frac{1}{4} \right) \begin{bmatrix} \mathbf{u} \\ \mathbf{u}_t \end{bmatrix} dt, \quad T = \frac{2\pi}{\omega}.$$

1733 This operator is more difficult to analyze [51] but in practice the iteration converges much
 1734 faster, typically in $\sim \omega$ iterations independent of dimension.

1735 4.3 Numerical Methods and Discrete Analysis

1736 An attractive feature of El WaveHoltz is that it can be used together with any convergent
 1737 discretization of the elastic wave equation. Here we consider the conservative curvilinear finite
 1738 difference method from [12] and the symmetric interior penalty discontinuous Galerkin method
 1739 [35, 62]. We give a very brief description of these below methods and refer the reader to [12, 35]
 1740 for details.

1741 Although highly non-intrusive, the one additional discretizational detail necessitated by El
 1742 WaveHoltz is how to discretize the integral in (4.10). As the integrand is periodic (once converged)
 1743 we always use the trapezoidal rule.

1744 4.3.1 El WaveHoltz by Finite Differences

To discretize the elastic wave equation (4.1.2) in a general non-Cartesian geometry we write
 (4.1.2) in a curvilinear coordinate system that conforms with the boundaries of the domain but
 that can be mapped back to the unit square (cube). Thus, we assume that there is a one-to-one

mapping

$$x = x(q, r), \quad y = y(q, r), \quad (q, r) \in [0, 1]^2,$$

from the unit square to the domain of interest. Then the two dimensional version of (4.1.2) becomes

$$\begin{aligned} J\rho \frac{\partial^2 u}{\partial t^2} &= \frac{\partial}{\partial q} \left[Jq_x \left[(2\mu + \lambda) (q_x \partial_q + r_x \partial_r) u + \lambda (q_y \partial_q + r_y \partial_r) v \right] \right. \\ &\quad \left. + Jq_y \left[\mu ((q_x \partial_q + r_x \partial_r) v + (q_y \partial_q + r_y \partial_r) u) \right] \right] \\ &+ \frac{\partial}{\partial r} \left[Jr_x \left[(2\mu + \lambda) (q_x \partial_q + r_x \partial_r) u + \lambda (q_y \partial_q + r_y \partial_r) v \right] + Jr_y \left[\mu ((q_x \partial_q + r_x \partial_r) v + (q_y \partial_q + r_y \partial_r) u) \right] \right], \\ J\rho \frac{\partial^2 v}{\partial t^2} &= \frac{\partial}{\partial q} \left[Jq_x \left[\mu ((q_x \partial_q + r_x \partial_r) v + (q_y \partial_q + r_y \partial_r) u) \right] \right. \\ &\quad \left. + Jq_y \left[(2\mu + \lambda) (q_y \partial_q + r_y \partial_r) v + \lambda (q_x \partial_q + r_x \partial_r) u \right] \right] \\ &+ \frac{\partial}{\partial r} \left[Jr_x \left[\mu ((q_x \partial_q + r_x \partial_r) v + (q_y \partial_q + r_y \partial_r) u) \right] + Jr_y \left[(2\mu + \lambda) (q_y \partial_q + r_y \partial_r) v + \lambda (q_x \partial_q + r_x \partial_r) u \right] \right]. \end{aligned}$$

1745 Here $J = x_q y_r - x_r y_q$ is the Jacobian of the mapping. Also note that we have considered the case
1746 without forcing for brevity.

1747 We discretize the unit square $(q, r) \in [0, 1]^2$ by a uniform grid on which we introduce real
1748 valued grid functions $[u_{i,j}(t), v_{i,j}(t)] = [u(q_i, r_j, t), v(q_i, r_j, t)]$. On this grid we apply the an energy
1749 stable discretization

$$\rho J \frac{\partial^2 u_h}{\partial t^2} = L^{(u)}(u_h, v_h), \quad \rho J \frac{\partial^2 v_h}{\partial t^2} = L^{(v)}(u_h, v_h). \quad (4.13)$$

1750 Here ρJ is a diagonal matrix containing the metric information and u_h, v_h are vectors containing all
1751 the grid function values. The (lengthy) exact definitions of $L^{(u)}(u_h, v_h), L^{(v)}(u_h, v_h)$ can be found
1752 in [12].

Suppose we are to impose a free surface boundary condition at $q = 0$. We then use a modified stencil for which the method is stable in a modified inner product. Let w_h and u_h be real valued grid functions and $(w_h, u_h)_h$ be the discrete inner product

$$(w_h, u_h)_h = h_q h_r \sum_{j=1}^{N_r} \left(\frac{1}{2} w_{1,j} u_{1,j} + \sum_{i=2}^{N_q} w_{i,j} u_{i,j} \right),$$

with corresponding norm $\|w_h\|_h^2 = (w_h, w_h)_h$. In this inner product, the discretization (of the PDE and boundary conditions) is self adjoint. That is, for all real-valued grid functions (u^*, v^*) , (u^\dagger, v^\dagger) satisfying the discrete boundary conditions, we have

$$(u^*, L^{(u)}(u^\dagger, v^\dagger))_h + (v^*, L^{(v)}(u^\dagger, v^\dagger))_h = (u^\dagger, L^{(u)}(u^*, v^*))_h + (v^\dagger, L^{(v)}(u^*, v^*))_h. \quad (4.14)$$

To discretize the equations in time we either use the standard second order accurate centered differences, or one of the time-corrected schemes discussed below. For the standard second order accurate centered difference approximation in time, the fully discrete equations take the form

$$\begin{aligned} (\rho J)(u_h^{n+1} - 2u_h^n + u_h^{n-1}) &= \Delta t^2 L^{(u)}(u_h^n, v_h^n), \\ (\rho J)(v_h^{n+1} - 2v_h^n + v_h^{n-1}) &= \Delta t^2 L^{(v)}(u_h^n, v_h^n). \end{aligned} \quad (4.15)$$

1753 Then, if $(u, v)_{\rho J}$ is the weighted inner product defined by $(f, (\rho J)^{-1} g)_{\rho J} = (f, g)_h$, and $C_e(t^{n+1})$ is
1754 the discrete energy

$$C_e(t^{n+1}) = \|D_+^t u^n\|_{\rho J}^2 + \|D_+^t v^n\|_{\rho J}^2 - (u^{n+1}, (\rho J)^{-1} L^{(u)}(u^n, v^n))_{\rho J} - (v^{n+1}, (\rho J)^{-1} L^{(v)}(u^n, v^n))_{\rho J}, \quad (4.16)$$

1755 one can show that this discrete energy is conserved [12].

1756 Note that (4.15) is slightly non-symmetric and needs to be diagonally scaled to become
1757 symmetric. Here we scale by 2 along sides with free surface boundary conditions, and by 4 in
1758 corners where free surfaces meet. Incorporating this scaling through the multiplication by a scaling
1759 matrix Λ , the method can be formally written as

$$M(\mathbf{u}_h^{n+1} - 2\mathbf{u}_h^n + \mathbf{u}_h^{n-1}) = \Delta t^2 L_h \mathbf{u}_h. \quad (4.17)$$

1760 Here $M = \text{diag}(\Lambda \rho J, \Lambda \rho J)$ and $L_h = \text{diag}(\Lambda L^{(u)}, \Lambda L^{(v)})$ are symmetric and M is diagonal. However,
1761 as $M^{-1} L_h$ is not in general symmetric, the iteration (4.10) will produce a symmetrizable but not
1762 symmetric operator. We will show below that this necessitates a minor modification of the conjugate
1763 gradient algorithm when used together with the iteration (4.10).

1764 **Remark 3.** *Here we only consider domains that can be discretized by a single logically Cartesian*

1765 grid, but note that by using the overset grid version of the method, [7], more complex geometry
1766 could be handled.

1767 **4.3.2 El WaveHoltz by Symmetric Interior Penalty Discontinuous Galerkin Method**

As an alternative to the finite difference method outlined above, we will also consider the Symmetric Interior Penalty Discontinuous Galerkin (SIPDG) method [35, 62]. Let Ω_h be a finite element partition of the computational domain Ω , with Γ_h the set of all faces. Then (4.3) can be reformulated into the interior-penalty weak formulation: Find $\mathbf{u}_h \in (0, T) \times V_h$ such that

$$\sum_{E \in \Omega_h} (\rho \frac{d^2 \mathbf{u}_h}{dt^2}, \mathbf{v})_E + \sum_{E \in \Omega_h} B_E(\mathbf{u}_h, \mathbf{v}) + \sum_{\gamma \in \Gamma_h} J_\gamma(\mathbf{u}_h, \mathbf{v}; S, R) = -\cos(\omega t) \sum_{E \in \Omega_h} (\mathbf{f}, \mathbf{v})_E, \quad (4.18)$$

for all $\mathbf{v} \in V_h$. Here

$$\begin{aligned} (\mathbf{u}, \mathbf{v})_E &= \int_E \mathbf{u} \cdot \mathbf{v} dE, \\ B_E(\mathbf{u}, \mathbf{v}) &= \int_E [\lambda(\nabla \cdot \mathbf{u})(\nabla \cdot \mathbf{v}) + \mu(\nabla \mathbf{u} + \nabla \mathbf{u}^T) : \nabla \mathbf{v}] dE, \\ J_\gamma(\mathbf{u}, \mathbf{v}; S, R) &= - \int_\gamma \{\mathcal{T}(\mathbf{u})\mathbf{n}\} \cdot [\mathbf{v}] d\gamma + S \int_\gamma \{\mathcal{T}(\mathbf{v})\mathbf{n}\} \cdot [\mathbf{u}] d\gamma + R \int_\gamma \{\lambda + 2\mu\} [\mathbf{u}] \cdot [\mathbf{v}] d\gamma, \end{aligned}$$

1768 where $\{\cdot\}$ and $[\cdot]$ denote the average and jump of a function, respectively. The parameter R is the
1769 penalty and S determines the particular flavor of IPDG. We thus set $S = -1$, corresponding to
1770 the Symmetric IPDG [62]. In this case, J_γ is symmetric with respect to \mathbf{u}_h and \mathbf{v} so that together
1771 with the symmetry of B_E we have that the stiffness matrix is symmetric. Thus SIPDG provides a
1772 symmetric discretization of the elastic wave equation, which will allow the use of conjugate gradient
1773 to accelerate convergence of the El WaveHoltz iteration.

1774 Our solver is implemented in MFEM¹ [4] and is essentially a direct extension of example 17
1775 to the time domain. Depending on the mesh, our choice of finite element space V_h is typically one
1776 of two broken spaces. We choose either $\mathcal{P}^p(E)$, the space of polynomials of total degree at most
1777 p on triangles, or $\mathcal{Q}^p(E)$, the space of polynomials of at most degree p on quadrilaterals. Unless
1778 otherwise noted, for the penalty parameter we make the choice $R = (p+1)(p+2)$.

¹ www.mfem.org

1779 With the standard second order explicit time discretization, the matrix form of (4.18) becomes

$$M_\rho(\mathbf{u}_h^{n+1} - 2\mathbf{u}_h^n + \mathbf{u}_h^{n-1}) = \Delta t^2 [L_h \mathbf{u}_h^n - \cos(\omega t^n) \mathbf{f}].$$

1780 As for the finite difference method, $M_\rho^{-1} L_h$ is not (in general) symmetric and this will necessitate
1781 a minor modification of the conjugate gradient algorithm when this scheme is used together with
1782 the iteration (4.10).

For this explicit time-stepping and the error corrected time-stepping discussed below, we use the CFL condition from [10]

$$\Delta t < \frac{\text{CFL} \cdot h_{\min}}{(p + \frac{3}{2})^2 \sqrt{\frac{2\mu+\lambda}{\rho}}}, \quad (4.19)$$

1783 where h_{\min} is the smallest diameter of the elements and CFL depends on the time-stepper. For the
1784 second order centered scheme, we typically choose $\text{CFL} \sim 0.4\text{--}0.8$.

1785 4.3.3 Explicit Time-Corrected Scheme

1786 If the elastic Helmholtz equation equation (4.2) is discretized directly, the solution satisfies
1787 the equation (in this section we take $\rho = 1$ and for notational clarity we suppress the subscript h)

$$\omega^2 \mathbf{v} + L_h \mathbf{v} = \mathbf{f}(\mathbf{x}). \quad (4.20)$$

1788 As we show in [14, 51], when the elastic wave equation is time marched with e.g. the second
1789 order method

$$\mathbf{u}^{n+1} - 2\mathbf{u}^n + \mathbf{u}^{n-1} = \Delta t^2 [L_h \mathbf{u}^n - \cos(\omega t_n) \mathbf{f}], \quad (4.21)$$

started with the initial data

$$\mathbf{u}^0 = \mathbf{u}_0, \quad \mathbf{u}^{-1} = \mathbf{u}_0 - \frac{\Delta t^2}{2} L_h (\mathbf{u}_0 + \mathbf{f}).$$

1790 Once El WaveHoltz has converged the initial data \mathbf{u}_0 satisfies the elastic Helmholtz equation with
1791 a modified frequency

$$\tilde{\omega}^2 \mathbf{u}_0(\mathbf{x}) + L_h \mathbf{u}_0(\mathbf{x}) = \mathbf{f}(\mathbf{x}), \quad \tilde{\omega} = \frac{2 \sin(\Delta t \omega / 2)}{\Delta t}. \quad (4.22)$$

1792 For this second order time discretization the difference between the final converged \mathbf{u}_0 and \mathbf{v} is
 1793 $\mathcal{O}(\Delta t^2)$. Thus if a high order accurate spatial discretization is used, time discretization errors will
 1794 limit the accuracy of the El WaveHoltz solution. To reduce this error, a time discretization which
 1795 is at least as accurate as the spatial discretization can be used. It is also possible, however, to
 1796 use the technique proposed by Stolk in [101] to modify the second order time-stepping method
 1797 and eliminate the error altogether. The corrected scheme in [101], introduced as a time domain
 1798 preconditioner, is the straightforward modification

$$\mathbf{u}^{n+1} - 2\mathbf{u}^n + \mathbf{u}^{n-1} = \frac{\tilde{\omega}^2}{\omega^2} \Delta t^2 [L_h \mathbf{u}^n - \cos(\omega t_n) \mathbf{f}]. \quad (4.23)$$

As [101] solves the equations in the frequency domain no initial data is needed. Here, as we work in the time domain, we must also modify the computation of \mathbf{u}^0 accordingly:

$$\mathbf{u}^0 = \mathbf{u}_0, \quad \mathbf{u}^{-1} = \mathbf{u}_0 - \frac{\tilde{\omega}^2}{\omega^2} \frac{\Delta t^2}{2} L_h (\mathbf{u}_0 + \mathbf{f}).$$

1799 4.3.4 Implicit Time-Corrected Scheme

1800 For a DG discretization, the use of an explicit time-stepping scheme for the elastic wave
 1801 equation requires a CFL condition that shrinks as $\mathcal{O}(p^{-2})$ where p is the polynomial order within
 1802 an element. For meshes with geometrical stiffness and DG discretizations of high order, it is then
 1803 particularly desirable to consider the use of an implicit scheme to circumvent a potentially restrictive
 1804 time-step size demanded by an explicit scheme.

1805 To that end, consider the semi-discrete system

$$\rho \mathbf{u}_{tt} = L_h \mathbf{u} - \cos(\omega t) \mathbf{f}(\mathbf{x}), \quad \mathbf{x} \in \Omega, t \geq 0, \quad (4.24)$$

1806 where L_h is a symmetric, positive definite approximation to the continuous operator $\nabla \cdot \mathcal{T}$ including
 1807 boundary conditions. The values $\nabla \cdot \mathcal{T}(\mathbf{u})$ are then approximated by $L_h \mathbf{u}$. We assume L_h has the
 1808 eigenmodes (λ_j^2, ϕ_j) , such that $L_h \phi_j = \lambda_j^2 \phi_j$ for $j = 1, \dots, N$, where all λ_j are strictly positive and
 1809 ordered as $0 \leq \lambda_1 \leq \dots \leq \lambda_N$.

We let the discrete Helmholtz solution \mathbf{v} be given by

$$-L_h \mathbf{v} + \omega^2 \mathbf{v} = \mathbf{f}.$$

The numerical approximation of the iteration operator is denoted Π_h , and it is implemented as follows. Given a grid function $\mathbf{u} \in \mathbb{R}^N$, we use the following implicit time-stepping scheme to solve the elastic wave equation as

$$\frac{\mathbf{u}^{n+1} - \alpha \mathbf{u}^n + \mathbf{u}^{n-1}}{\Delta t^2} = \frac{1}{2} L_h (\mathbf{u}^{n+1} + \mathbf{u}^{n-1}) - \mathbf{f} \cos(\omega t_n) \cos(\omega \Delta t), \quad (4.25)$$

where

$$\alpha = \cos(\omega \Delta t)(2 + \omega^2 \Delta t^2) \approx 2 - \frac{5(\omega \Delta t)^4}{12} + \mathcal{O}(\Delta t^6). \quad (4.26)$$

- ₁₈₁₀ For the stability of the method it is necessary to have $|\alpha| < 2$. This choice of the time-step corresponds to a (mild) requirement of at least five time-steps per iteration (See details in Appendix .10).

With a time-step $\Delta t = T/k$ for some integer k , the scheme (4.25) is completed by initial data

$$\mathbf{u}^0 = \mathbf{u}_0, \quad \mathbf{u}^{-1} = \left(I - \frac{\Delta t^2}{2} L_h \right)^{-1} \left(\frac{\alpha}{2} \mathbf{u}_0 - \frac{\Delta t^2}{2} \cos(\omega \Delta t) \mathbf{f} \right).$$

- ₁₈₁₂ The trapezoidal rule is then used to compute $\Pi_h \mathbf{u}$,

$$\Pi_h \mathbf{u} = \frac{2\Delta t}{T} \sum_{n=0}^M \eta_n \left(\cos(\omega t_n) - \frac{1}{4} \right) \mathbf{u}^n, \quad \eta_n = \begin{cases} \frac{1}{2}, & n = 0 \text{ or } n = M, \\ 1, & 0 < n < M. \end{cases} \quad (4.27)$$

Define the discrete filter transfer function by

$$\beta_h(\lambda) = \frac{2\Delta t}{T} \sum_{n=0}^M \eta_n \cos(\lambda t_n) \left(\cos(\omega t_n) - \frac{1}{4} \right),$$

- ₁₈₁₃ In Appendix .11 we motivate the following Conjecture (we believe this can be proved but at the time of writing we have not yet done so.)

Conjecture 1. *Let $\Delta t = T/k$ for some integer k with $T = 2\pi/\omega$. The discrete filter transfer function β_h satisfies*

$$\begin{cases} |\beta_h(\lambda)| \geq 1, & \lambda \in [\omega(1 - 0.022 \cdot \Delta t^2), \omega], \\ |\beta_h(\lambda)| < 1, & \text{otherwise.} \end{cases}$$

1815 Under the assumption that Conjecture 1 holds we may then prove the following theorem.

1816 **Theorem 1.** Suppose there are no resonances, such that $\delta_h = \min_j |\lambda_j - \omega|/\omega > 0$. Moreover,
1817 assume that Δt satisfies the stability and accuracy requirements

$$\frac{|\alpha|}{2} = |\cos(\omega\Delta t)(1 + \omega^2\Delta t^2/2)| < 1, \quad \Delta t \leq \frac{\cos(2\pi/5)\omega^2\delta_h}{0.044 \cdot (1 + 2(\pi\lambda_N/5)^2)}. \quad (4.28)$$

Further assume that the properties of the discrete filter transfer function in Conjecture 1 hold. Then the fixed point iteration $\mathbf{v}^{(k+1)} = \Pi_h \mathbf{v}^{(k)}$ with $\mathbf{v}^{(0)} = 0$ converges to \mathbf{v} which is a solution to the discretized Helmholtz equation

$$-L_h \mathbf{v} + \omega^2 \mathbf{v} = \mathbf{f}.$$

Proof. We expand all functions in eigenmodes of L_h ,

$$\mathbf{u}^n = \sum_{j=1}^N \hat{u}_j^n \phi_j, \quad \mathbf{f} = \sum_{j=1}^N \hat{f}_j \phi_j, \quad \mathbf{v} = \sum_{j=1}^N \hat{v}_j \phi_j.$$

Then the Helmholtz eigenmodes of \mathbf{v} satisfy

$$\hat{v}_j = \frac{\hat{f}_j}{\omega^2 - \lambda_j^2}.$$

The wave solution eigenmodes are given by the difference equation

$$\left(1 + \frac{\Delta t^2}{2} \lambda_j^2\right) \hat{u}_j^{n+1} - \alpha \hat{u}_j^n + \left(1 + \frac{\Delta t^2}{2} \lambda_j^2\right) \hat{u}_j^{n-1} = -\Delta t^2 \hat{f}_j \cos(\omega t_n) \cos(\omega \Delta t), \quad (4.29)$$

with initial data

$$\hat{u}_j^0 = \hat{u}_{0,j}, \quad \hat{u}_j^{-1} = \left(1 + \frac{\Delta t^2}{2} \lambda_j^2\right)^{-1} \left(\frac{\alpha}{2} \hat{u}_{1,j} - \frac{1}{2} \Delta t^2 \cos(\omega \Delta t) \hat{f}_j\right).$$

By (4.28),

$$\left|\alpha \left(1 + \frac{\Delta t^2}{2} \lambda_j^2\right)^{-1}\right| = \left|2 \cos(\omega \Delta t) \frac{2 + \Delta t^2 \omega^2}{2 + \Delta t^2 \lambda_j^2}\right| < 2,$$

so that the characteristic polynomial for the equation, $r^2 - \alpha(1 + \Delta t^2 \lambda_j^2/2)^{-1}r + 1$, has two roots on the boundary of the unit circle. The solution is therefore stable and is given by (with a verification in Appendix .9)

$$\hat{u}_j^n = (\hat{u}_{0,j} - \hat{v}_j) \cos(\tilde{\lambda}_j t_n) + \hat{v}_j \cos(\omega t_n), \quad (4.30)$$

where $\tilde{\lambda}_j$ is defined by the relation

$$\cos(\tilde{\lambda}_j \Delta t) = \frac{\alpha}{2} \left(1 + \frac{\Delta t^2}{2} \lambda_j^2\right)^{-1} = \cos(\omega \Delta t) \left(1 + \frac{\omega^2 \Delta t^2}{2}\right) \left(1 + \frac{\Delta t^2}{2} \lambda_j^2\right)^{-1}. \quad (4.31)$$

Now, let

$$\Pi_h \mathbf{u}_0 = \sum_{j=1}^{\infty} \bar{u}_j \phi_j.$$

Then the numerical integration gives

$$\begin{aligned} \bar{u}_j &= \frac{2\Delta t}{T} \sum_{n=0}^M \eta_n \left(\cos(\omega t_n) - \frac{1}{4} \right) \left((\hat{u}_{0,j} - \hat{v}_j) \cos(\tilde{\lambda}_j t_n) + \hat{v}_j \cos(\omega t_n) \right) = (\hat{u}_{0,j} - \hat{v}_j) \beta_h(\tilde{\lambda}_j) + \hat{v}_j \beta_h(\omega) \\ &= \hat{u}_{0,j} \beta_h(\tilde{\lambda}_j) + (1 - \beta_h(\tilde{\lambda}_j)) \hat{v}_j. \end{aligned}$$

1818 Here we used the fact that the trapezoidal rule is exact, and equal to one, when $\lambda = \omega$. (Recall
1819 that for periodic functions the trapezoidal rule is exact for all pure trigonometric functions of order
1820 less than the number of grid points.)

1821 For the time-step restriction (4.28), we have that $|\tilde{\lambda}_j - \omega| > 0.022 \cdot \Delta t^2$ so that the bound
1822 $|\beta_h(\tilde{\lambda}_j)| \leq \rho_h < 1$ in the conjecture is uniform for all j . It follows that $\mathbf{v}^{(k)} \rightarrow \mathbf{v}$. This concludes
1823 the proof of the theorem. \square

Remark 4. We remark that it is also possible to remove the time discretization error by modifying the weights in the trapezoidal rule as in [93]

$$\frac{2\Delta t}{T} \sum_{n=0}^{N_t} \frac{\cos(\omega t_n)}{\cos(\frac{2\sin(\Delta t \omega/2)}{\Delta t} t_n)} \left(\cos(\omega t_n) - \frac{1}{4} \right) \mathbf{u}^n. \quad (4.32)$$

1824 It should however be noted that there is a risk that the denominator in this expression can become
1825 arbitrarily close to zero unless care is taken.

1826 4.3.5 Krylov Solution of the El WaveHoltz Iteration

Let Π_h be the matrix corresponding to a discretization of the El WaveHoltz method using either the finite difference or the SIPDG method. Then the iteration is (in this section a superscript

i denotes iteration and a superscript n denotes time-step)

$$\mathbf{u}_h^0 = \Pi_h \mathbf{0},$$

$$\mathbf{u}_h^{i+1} = \Pi_h \mathbf{u}_h^i, \quad i = 0, 1, \dots$$

1827 The solution to this fixed point iteration can also be found by solving

$$(I - \Pi_h) \mathbf{u}_h = \mathbf{b} \equiv \Pi_h \mathbf{0}, \quad (4.33)$$

1828 where the action of the matrix $(I - \Pi_h)$ requires (4.3) with $\mathbf{f} = 0$ to be solved for one period,

1829 $T = 2\pi/\omega$, and the right hand side is pre computed by solving (4.3) with the \mathbf{f} at hand.

1830 Let $Q = I - \Pi_h$. We know from [14] that the eigenvalues of Q are in the interval $(0, 3/2)$ so

1831 that Q is positive definite. We note that the methods for the elastic wave equation we consider here

1832 produce solutions $\{\mathbf{u}^1, \mathbf{u}^2, \dots, \mathbf{u}^{N_t}\}$ at time instances $0, \Delta t, 2\Delta t, \dots$, according to the recursion

$$\mathbf{u}^0 = a_0 \mathbf{u} + a_1 M^{-1} S \mathbf{u}^n, \quad (4.34)$$

$$\mathbf{u}^1 = \mathbf{u} \quad (4.35)$$

$$\mathbf{u}^{n+1} = \kappa \mathbf{u}^n - \mathbf{u}^{n-1} + \gamma M^{-1} S \mathbf{u}^n. \quad (4.36)$$

1833 It follows that the matrix

$$Q \mathbf{u} = \sum_{n=1}^{N_t} \alpha_n \mathbf{u}^n, \quad (4.37)$$

1834 will not be symmetric even if M and S are.

However, as the operator Q can be expressed as a polynomial P_Q of degree $N_t - 1$ in $M^{-1}S$,

we have that MQ is symmetric. Thus rather than applying the conjugate gradient method to

(4.33), we instead solve

$$M(I - \Pi_h) \mathbf{u}_h = M \mathbf{b}. \quad (4.38)$$

1835 We note that the main cost of applying the matrix Q is in computing $\Pi_h \mathbf{u}_h$. Since the matrix M

1836 is diagonal for the finite difference method and block-diagonal for SIPDG, the difference in cost

1837 between applying (4.38) over (4.33) is negligible compared to the advantage of not having to store

1838 a Krylov subspace when using the conjugate gradient or conjugate residual method.

1839 **Remark 5.** In some of the experiments below we use conjugate residual rather than conjugate
 1840 gradient. The reason for this is that it has the property that the residual is non-increasing, which
 1841 we have found gives a predictable and robust iteration count when doing parameter sweeps over ω .
 1842 When conjugate gradient is used we sometimes observe that we get “lucky” and converge in very
 1843 few iterations for a few frequencies. When considering practical applications it is of course good to
 1844 have such luck, but as we are trying to present the average behavior of our method here we prefer
 1845 conjugate residual.

1846 **4.4 Numerical Experiments**

1847 In this section we present numerical experiments that demonstrate the properties of the
 1848 method. We start with numerical experiments that demonstrate the spatial accuracy with and with-
 1849 out the time-stepping correction for the finite difference and the discontinuous Galerkin method.
 1850

1851 **4.4.1 Accuracy of the Finite Difference Method**

We consider solving the elastic Helmholtz equation with Lamé parameters $\lambda = \mu = 1.0$, where the forcing function is chosen so that the displacements are given by

$$u = v = 16^2 x^2 (x - 1)^2 y^2 (y - 1)^2. \quad (4.39)$$

1852 We take the frequency to be $\omega = 1.0$ and enforce Dirichlet boundary conditions on the boundary
 1853 of the unit square $(x, y) \in [0, 1] \times [0, 1]$. To verify accuracy, we set the tolerance to 10^{-15} in the
 1854 conjugate residual method as the stopping criteria and compute the error in u to the exact solution.
 1855 We use the finite difference method together with the standard explicit second order time-stepping
 1856 scheme, and verify the convergence of the method by grid refinement. To that end, we choose the
 1857 coarsest grid to have $n = 20$ points along each direction and refine by a factor of two up to $n = 160$
 1858 points per direction. In Table 4.1 we estimate the rate of convergence and observe second order
 1859 convergence, as expected.

n	L_1 error	Convergence	L_2 error	Convergence	L_∞ error	Convergence
20	3.86(-3)	-	3.86(-3)	-	3.86(-3)	-
40	9.21(-4)	2.06	9.21(-4)	2.06	9.21(-4)	2.06
80	2.25(-4)	2.03	2.25(-4)	2.03	2.25(-4)	2.03
160	5.55(-5)	2.02	5.55e(-5)	2.02	5.55(-5)	2.02

Table 4.1: L_1 , L_2 and L_∞ errors of the computed solution with corresponding estimated rates of convergence.

1860 4.4.2 Verification of Corrected Time-Steppers

1861 We consider solving the elastic Helmholtz equation with $\lambda = \mu = 1.0$ and choose the forcing so
 1862 that the exact solution is the same as (4.39). We take the frequency $\omega = 1$ and enforce homogeneous
 1863 Dirichlet conditions on the boundary of the square $(x, y) \in [0, 1] \times [0, 1]$. As the solution is a fourth
 1864 order polynomial, choosing $p = 4$ should ensure that the solution to the discrete elastic Helmholtz
 1865 equation is precisely (4.39). We use the conjugate gradient accelerated version of El WaveHoltz
 1866 with the corrected second order centered time-stepping scheme presented in Section 4.3.3.

1867 We partition the domain into four quadrilaterals of equal side length $h = 0.5$, set the absolute
 1868 conjugate gradient residual tolerance to 10^{-15} , and consider the error as the time-step size is
 1869 decreased. We see from Figure 4.1 that the standard centered scheme leads to a discrete solution
 1870 that converges to the true solution to second order.

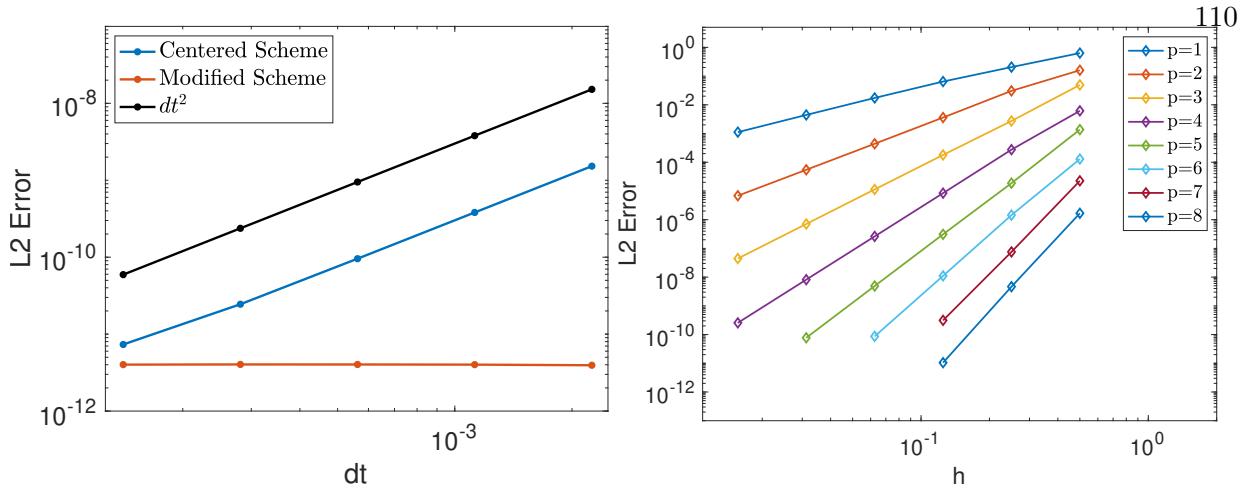


Figure 4.1: (Left) Convergence of the discrete WaveHoltz solution to the true solution of the discrete Helmholtz problem. (Right) Convergence of the discrete WaveHoltz solution to the true solution of the discrete Helmholtz problem for a manufactured solution.

1871 The modified scheme, however, maintains the same relative error to the true solution independent
 1872 of time-step size which indicates that (aside from roundoff errors) the time-stepping errors have
 1873 been removed. For the remaining numerical examples we use the modified time-stepping scheme
 1874 to remove time discretization errors.

1875 **4.4.3 Accuracy of the Symmetric Interior Penalty Discontinuous Galerkin Method**

Next we verify the rates of convergence for our symmetric interior penalty DG solver and for non-homogeneous problems using an example taken from [10]. We consider the unit square $S = [0, 1]^2$ and impose Dirichlet conditions on the boundary. The boundary conditions and forcing are chosen so that the Helmholtz solution is

$$u(x, y) = \sin(k_x x + x_0) \sin(k_y y + y_0),$$

$$v(x, y) = -\sin(k_x x + x_0) \sin(k_y y + y_0),$$

1876 where $k_x = 2.5\pi$, $k_y = 2\pi$, $x_0 = 5$, $y_0 = -10$. The mesh used is a uniform discretization of the unit
 1877 square split into smaller squares of side-length $h = 1/2^n$ for $n = 1, \dots, 6$.

1878 We set $\omega = 1$ and choose the material parameters to be the constants $\mu = 1$, $\lambda = 2$. Here we
 1879 use the modified time-stepping scheme of Section 4.3.3 with $CFL = 0.4$. The errors are plotted in
 1880 Figure 4.1 as a function of the grid size h . We additionally display estimated rates of convergence
 1881 calculated using linear least squares in Table 4.2 from which it is clear that the WaveHoltz method
 1882 converges with optimal rates with the error corrected time-stepper (which is formally only second
 1883 order accurate in Δt).

p	1	2	3	4	5	6	7	8
	1.84	2.94	4.00	4.94	6.01	6.86	8.07	8.64

Table 4.2: Estimated rates of convergence for the spatial discretization.

1884 4.4.4 Effects on Number of Iterations from Number of Periods and Accuracy

1885 In this section we investigate the efficiency of the filter (4.12), defined over K periods, for
 1886 various values of K . Let N_T be the number of time-steps for one period. Then we expect the
 1887 reduction in the number of all-to-all communications to be KN_T when compared to a direct dis-
 1888 cretization of (4.1.1). Here we consider energy conserving boundary conditions, for which we can
 1889 use the conjugate gradient method and avoid the need to store a Krylov subspace. We note that
 1890 for problems with impedance or non-reflecting boundary conditions (or with lower order damping
 1891 terms), El WaveHoltz will still result in a positive definite but non-symmetric system which can be
 1892 solved e.g. with GMRES. In that case, we also expect that the size of the GMRES Krylov subspace
 1893 will decrease by a factor KN_T compared to direct discretization, and by a factor K compared to
 1894 using a single period for the filter procedure.

1895 With these obvious advantages of filtering over K periods, it is natural to ask how the number
 1896 of iterations are affected by the increased filter time. In this experiment we numerically investigate
 1897 this. To do so, we use the corrected explicit version of the DG solver and consider the shaking of
 1898 a bar of (unitless) length 8 and height 1. We impose free surface boundary conditions on the top,

1899 bottom, and right of the domain, and on the left we set the boundary conditions to be

$$u(0, y, t) = v(0, y, t) = \cos(\omega t).$$

1900 The base computational mesh uses 8 square elements each with side length 1, which we uniformly
1901 refine by dividing each element in 4 parts for some number of refinements. We set $\lambda = 2$, $\mu = 1$,
1902 $\omega = 5.123$ and $\text{CFL} = 0.8$. We consider 2 cases: Case 1 uses $p = 5$ and refines the base grid 3 times,
1903 Case 2 uses $p = 15$ and refines one time. For both cases we use conjugate gradient and count the
1904 number of iterations it takes to reduce the relative residual by a factor 10^{-10} . The solution, along
1905 with the components of the stress tensor σ_{xx} , σ_{xy} and σ_{yy} , are displayed in Figure 4.2.

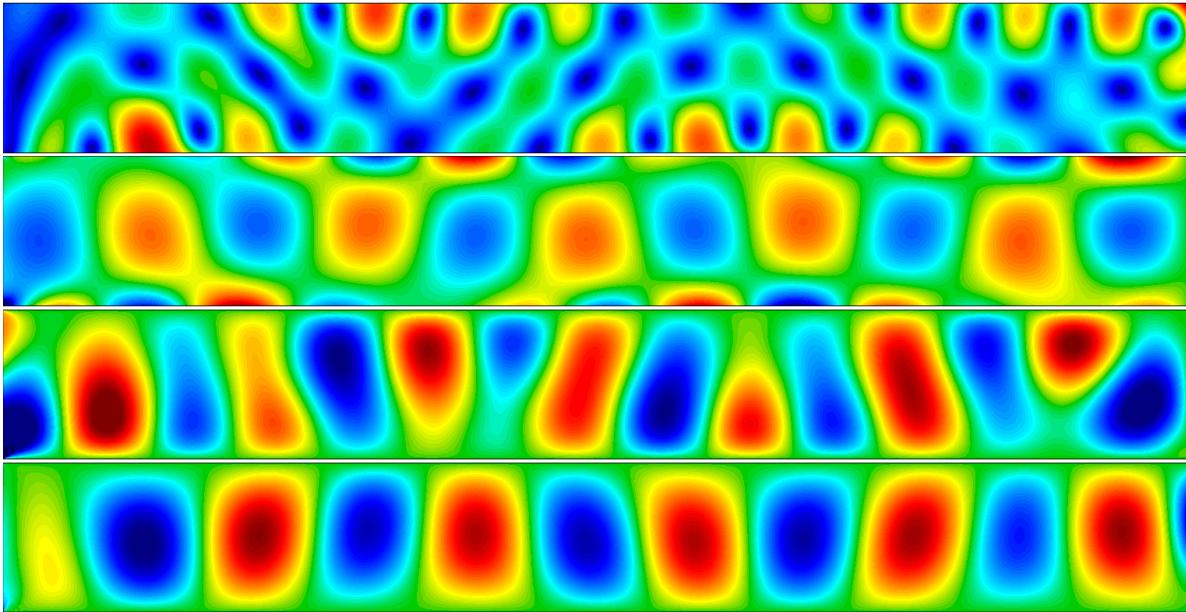


Figure 4.2: From top to bottom: displacement magnitude, σ_{xx} , σ_{xy} and σ_{yy} . The domain is $[0, 8] \times [0, 1]$ and the color scales are $[0, 8]$, $[-50, 50]$, $[-15, 15]$ and $[-40, 40]$ respectively.

1906 The number of iteration for the two cases and the relative efficiency are tabulated in Table 4.3.

Periods	1	2	3	4	5	10
Case 1 (#iter)	124	69	51	43	39	28
Efficiency	1	0.90	0.81	0.72	0.64	0.44
Case 2 (#iter)	151	96	78	68	62	53
Efficiency	1	0.79	0.65	0.56	0.49	0.29

Table 4.3: The table displays the number of iterations required and the efficiency of the longer times to reduce the relative residual by a factor 10^{-10} for the two cases (described in the text).

1907 As can be seen, the efficiency is relatively high when the number of periods are small and can thus
 1908 be deployed if the all-to-all communications (or the size of a GMRES Krylov space) becomes a
 1909 limiting factor.

1910 **4.4.5 Iteration Count as a Function of Frequency for Rectangles and Annular
 1911 Sectors**

1912 For energy conserving boundary conditions, the theoretical prediction (which is also observed
 1913 experimentally) is that the number of iterations scales as ω^d in d -dimensions. In this and the
 1914 next section, we study how the number of iterations depends on the frequency in two and three
 1915 dimensions. In this section we additionally investigate the dependence of the number of iterations on
 1916 frequency for different geometries. Here, we study these properties via three different computational
 1917 domains: a rectangle, a quarter annulus, and a half annulus (all with a characteristic length of 5).

1918 We use the finite difference method together with the standard explicit second order time-stepping
 1919 scheme. For each of the geometries we consider the set of frequencies $\omega = k + \sqrt{2}/10$, with
 1920 $k = 3, 4, \dots, 40$.

Let q and r be the coordinates in the (reference) unit square. We set n_q and n_r to be the number of cells in each coordinate direction. The (spatial) step size is given by $h_q = 1/n_q$ and $h_r = 1/n_r$, and our grid on the unit square is given by

$$q_i = i h_q, \quad i = 0, \dots, n_q, \quad r_j = j h_r, \quad j = 0, \dots, n_r.$$

We set the arc length of the outer arc at radii r_{out} of the annular sector to be the length, $L = 5$. Thus for the quarter annulus we have $r_{\text{out}} = \frac{2L}{\pi}$, and for the half-annulus we have $r_{\text{out}} = \frac{L}{\pi}$. For both cases, we take $r_{\text{in}} = r_{\text{out}} - 1$. Precisely, the coordinates of the two grids used for the quarter annular sector and the half annular sector are

$$x_{ij} = (r_{\text{in}} + (r_{\text{out}} - r_{\text{in}})q_i) \cos \left(n_{\text{an}} \frac{\pi}{2} r_j - \frac{\pi}{2} \right),$$

$$y_{ij} = (r_{\text{in}} + (r_{\text{out}} - r_{\text{in}})q_i) \sin \left(n_{\text{an}} \frac{\pi}{2} r_j - \frac{\pi}{2} \right).$$

1921 Here n_{an} is either 1 or 2 to indicate the quarter or half annulus, respectively. We set $n_r =$
 1922 $4L\omega + 1, n_q = 4\omega + 1$ so that the number of points per shear wavelength is around 20.

1923 For the forcing we use a discrete approximation of the delta function with amplitude $\omega^2 \cos(\omega t)$.
 1924 We locate this point source at $(x_{i^*j^*}, y_{i^*j^*})$ where $i^* = j^* = (n_q + 1)/2$ so that it is close to $(0.5, 0.5)$
 1925 in physical space. In Figure 4.3 we display the number of iterations required to reduce the relative
 1926 residual in the conjugate residual method by a factor 10^{-8} .

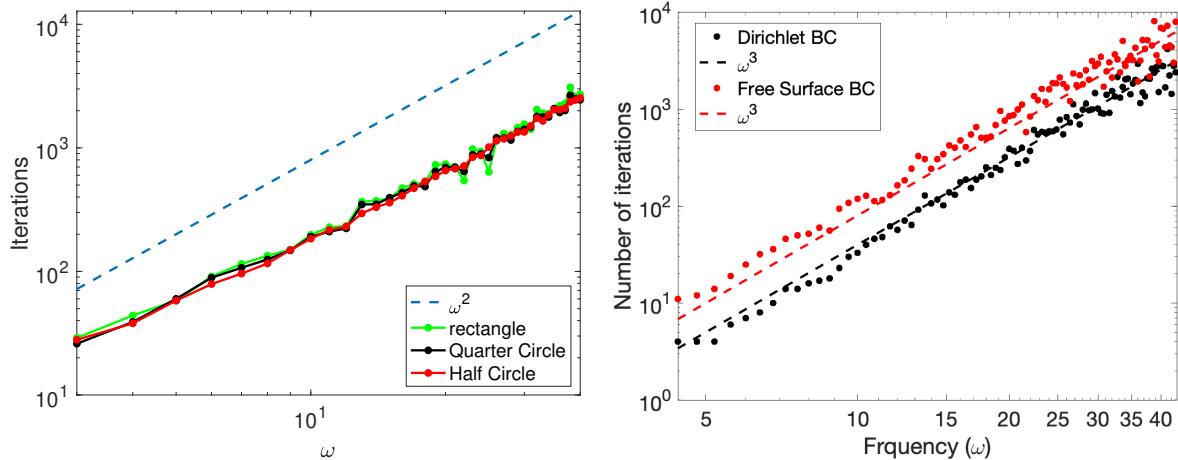


Figure 4.3: The number of iterations as a function of frequency to reach convergence for (Left) a rectangle, quarter circle and half circle, and (Right) the unit cube with Dirichlet or free surface conditions.

1927 From Figure 4.3 we see that the results for the three geometries are very similar, indicating that
 1928 (in this example at least) the geometry has little to no effect on the number of iterations needed.

1929 Moreover the number of iterations grow as ω^2 , as expected.

1930 **4.4.6 Effect of Boundary Conditions in A Cube**

1931 In this experiment we consider the unit cube with either Dirichlet boundary conditions on
1932 all sides, or with free surface boundary conditions on the top and bottom ($z = 0$ and $z = 1$) with
1933 Dirichlet boundary conditions on all other sides. We use the 3D version of the finite difference
1934 method described above with the standard explicit time-stepping method. Here we use the forcing

$$f_j(\mathbf{x}, t) = A_j e^{-\frac{\sigma}{2} \|\mathbf{x} - \mathbf{x}_j\|^2},$$

1935 with $A_j \sim \sqrt{\sigma^d}$ and with $\sigma \sim \omega$ so that each of the components of the forcing approaches a delta
1936 function as ω grows. We select \mathbf{x}_j slightly different for each j so that both $\nabla \times \mathbf{f} \neq 0$ and $\nabla \cdot \mathbf{f} \neq 0$,
1937 resulting in a solution with both shear and pressure waves.

1938 We use the conjugate residual method, keep the product $h\omega = 0.4$ fixed, and report the
1939 number of iterations required to reduce the initial residual (starting from zero initial data) by a
1940 factor 10^{-9} . The result, which can be found in Figure 4.3, confirm the prediction from [14] that
1941 the number of iterations scale as ω^3 .

1942 **4.4.7 Iteration Count as a Function of Wave Speed Ratio**

1943 The length of a domain, when measured in number of wavelengths, will increase both if the
1944 physical domain size is increased and if the wave speed is reduced. The compressional and shear
1945 wave speeds are $C_p = \sqrt{(2\mu + \lambda)/\rho}$ and $C_s = \sqrt{\mu/\rho}$, respectively. We expect that the number of
1946 iterations will depend on the smallest wave speed, but for El WaveHoltz there is no intuitive reason
1947 to think that a problem with $C_p \gg C_s$ should be more difficult than a problem with $C_p \approx C_s$. We
1948 note, however, that such behavior has been reported in the literature (see e.g. Table 3.1 on page
1949 11 of [102]) for other methods.

1950 To experimentally investigate how well El WaveHoltz works for different combinations of μ
1951 and λ , we use the SIPDG solver with the corrected explicit time-stepper for a geometry consisting

1952 of the unit square with a circular hole cut out (see Figure 4.4).

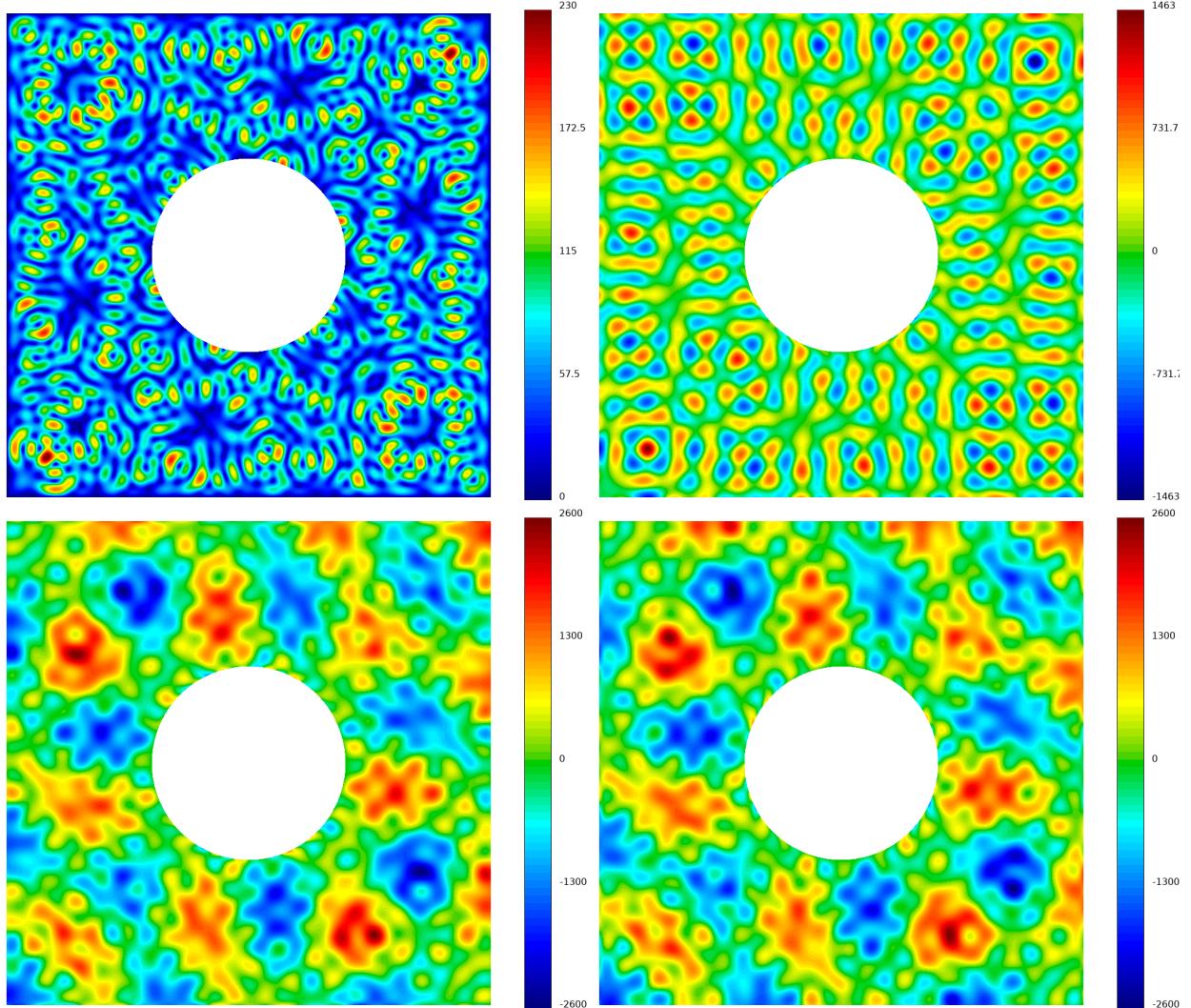


Figure 4.4: From top left to bottom right: displacement magnitude, σ_{xy} , σ_{xx} and σ_{yy} .

1953 This is the mesh `square-disc-nurbs.mesh`, which is part of the MFEM distribution. The Lamé
1954 parameters are constant in space and we choose the number of refinements so that the solution is
1955 well resolved (the largest and smallest element size is reported in Table 4.4).

λ	1	2	4	8	16	32	64	1	1
μ	1	1	1	1	1	1	1	1/4	1/16
$h_{\max} \times 10^2$	3.37	3.37	3.37	3.37	3.37	3.37	3.37	1.10	0.852
$h_{\min} \times 10^2$	2.25	2.25	2.25	2.25	2.25	2.25	2.25	1.70	0.547
#Iter.	45	47	35	38	38	56	44	126	247
#Iter. $\times \sqrt{\mu}$	45	47	35	38	38	56	44	63	62

Table 4.4: The effect on iteration count depending on different combinations of λ and μ .

1956 We impose the boundary conditions

$$u(0, y, t) = v(0, y, t) = \cos(\omega t),$$

1957 on the outer part of the domain, and let the circular hole be free of traction. For all experiments
1958 we set $\omega = 25.12$, $\text{CFL} = 0.8$ and we evolve the El WaveHoltz iteration over $K = 3$ periods. We
1959 stop the CG iteration when the relative residual falls below 10^{-6} . In Figure 4.4 we display the
1960 magnitude of the displacement and the components of the stress tensor σ_{xy} , σ_{xx} and σ_{yy} for the
1961 case when $\lambda = 1$ and $\mu = 1/16$.

1962 The results, displayed in Table 4.4, show that El WaveHoltz appears to be robust with respect
1963 to the ratio between λ and μ . Moreover, the number of iterations to reach the desired tolerance is
1964 primarily a function of the μ , or equivalently, the shear wave speed C_s .

1965 4.4.8 Comparison of Explicit and Implicit El WaveHoltz with Direct Discretization 1966 of Elastic Helmholtz

1967 In this example we compare the explicit error corrected SIPDG method, the implicit error
1968 corrected SIPDG method and the SIPDG method of `example17p` extended to the elastic Helmholtz
1969 problem (4.2). In this section we will refer to these solvers by the abbreviations WH, IWH and
1970 HH respectively. For the Helmholtz SIPDG solver we use GMRES preconditioned by the AMG
1971 solver provided by `HypreBoomerAMG`, with the elasticity specific options (see also [17]) provided by
1972 `SetElasticityOptions`. The GMRES solver is restarted every 100 iterations.

1973 For the implicit solver we must also invert the elasticity operator (but with a shift that

¹⁹⁷⁴ preserves its positive definiteness) and we do this by CG preconditioned by the same AMG setup
¹⁹⁷⁵ as for the Helmholtz SIPDG solver. We always use 10 time-steps for the implicit solver, and for the
¹⁹⁷⁶ explicit solver we use $CFL = 1.1$. For the WH and IWH solver we solve the El WaveHoltz problem
¹⁹⁷⁷ with conjugate gradient. For all three solvers the tolerance is set to be 10^{-10} .

¹⁹⁷⁸ We solve the equations on the unit square with a smooth (but narrow) forcing

$$\mathbf{f} = -\frac{200\omega^2}{\pi} e^{-1.2\omega^2[(x-0.25)^2 + (y-0.25)^2]} \begin{pmatrix} -y + 0.5 \\ x - 0.5 \end{pmatrix},$$

¹⁹⁷⁹ and with free surface boundary conditions on all sides. We consider different refinements and
¹⁹⁸⁰ polynomial degrees from 2 to 9 and estimate the error in the solution by computing a reference
¹⁹⁸¹ solution using degree 11 polynomials (note that the solutions are the same up to the tolerance of
¹⁹⁸² the iterative solvers since we have eliminated the time errors).

¹⁹⁸³ For all of the computations we record the number of iterations (a maximum of 500000) and
¹⁹⁸⁴ list them in Table 4.5.

		$p = 1$	2	3	4	5	6	7	8	9
h	WH	57	102	115	126	132	138	141	144	148
$h/2$	WH	68	108	114	126	130	133	138	143	146
$h/4$	WH	79	107	114	122	130	135	137	142	146
$h/8$	WH	83	108	114	125	130	133	137	142	146
h	IWH	81	182	160	173	213	237	192	235	201
$h/2$	IWH	97	151	157	168	220	263	188	276	196
$h/4$	IWH	112	147	153	163	171	178	264	188	275
$h/8$	IWH	117	147	154	165	171	213	217	N/A	N/A
h	HH	480	25084	18298	37926	80262	144863	204694	230688	500000
$h/2$	HH	15686	32063	57338	106688	256801	347279	500000	500000	500000
$h/4$	HH	46667	79865	184331	334665	500000	500000	500000	500000	500000
$h/8$	HH	500000	304561	500000	500000	500000	500000	500000	N/A	N/A

Table 4.5: Comparison of the number of iterations for the three different methods.

¹⁹⁸⁵ It can be seen that the fewest number of iterations are achieved with the WH method. We also
¹⁹⁸⁶ note that the number of iterations for WH is insensitive to the mesh resolution and has a weak
¹⁹⁸⁷ dependence on the polynomial degree. Again, the latter is due to the fact that the linear system

1988 we are solving comes from a bounded operator so that the condition number does not depend on h .
 1989 HH, which discretizes an unbounded and indefinite operator, behaves radically different with the
 1990 number of iterations increasing rapidly with decreasing mesh resolution. In addition, the number
 1991 of iterations for the HH method increases very quickly with the polynomial degree and, as a result,
 1992 many of the accurate test cases fail to converge. Similar to the WH method, the IWH method has
 1993 an iteration count that is relatively robust under grid refinement but with a slight increase with
 1994 order. We note that it does appear the higher order methods have more variation in iteration count
 1995 than the lower order methods, and in general the iteration count is larger than for the WH method.

1996 The number of iterations displayed in Table 4.5 are not useful for comparing the different
 1997 methods as each iteration comes with a different computational cost. In Table 4.6 we instead list
 1998 the number of right hand side (rhs) evaluations.

		$p = 1$	2	3	4	5	6	7	8	9
h	WH	1425	4998	9315	15120	22176	30774	40326	51552	64676
$h/2$	WH	3400	10476	18354	30240	43550	59318	78936	102245	127458
$h/4$	WH	7821	20758	36594	58438	86970	120285	156728	202918	254916
$h/8$	WH	16434	41904	73188	119750	173940	236873	313456	405836	509686
h	HH	1	5	2	3	4	5	5	4	∞
$h/2$	HH	5	3	3	4	6	6	∞	∞	∞
$h/4$	HH	6	4	5	6	∞	∞	∞	∞	∞
$h/8$	HH	∞	7	∞	∞	∞	∞	∞	N/A	N/A
h	IWH	81	96	67	63	72	72	53	60	48
$h/2$	IWH	72	67	59	53	69	70	46	59	40
$h/4$	IWH	59	59	51	48	46	41	57	36	49
$h/8$	IWH	50	53	47	43	42	46	42	N/A	N/A

Table 4.6: The number of right hand side evaluations (estimated) for the three different methods. The top four rows display the actual number of right hand side evaluations and the rows below indicate how many times more the HH and IWH method evaluates the right hand side. An infinity sign indicates that the computation did not converge.

1999 By a right hand side evaluation we mean a single application of the matrix corresponding to the
 2000 matrix discretizing the elastic operator. For the explicit method, the total number of rhs evaluations

2001 is $N_T N_{\text{iter}}$, where N_T is the number of time-steps needed to evolve the elastic wave equation one
 2002 period and N_{iter} is the number of iterations. For the IWH method we always take $N_T = 10$ so
 2003 that the number of rhs evaluations is $10N_{\text{inner}}N_{\text{iter}}$, where N_{inner} is the number of inner iterations
 2004 used by the AMG preconditioner (as reported in Table 4.5). For the HH method the number of
 2005 rhs evaluations is equal to the number of GMRES iterations.

2006 As can be seen in Table 4.6, the WH method is also more efficient with respect to the number
 2007 of rhs evaluations (note that we report total number of rhs for the WH method and multipliers for
 2008 the other methods). The advantage of the explicit method over the implicit method appears to be
 2009 decreasing with increased accuracy - both in terms of decreasing mesh size and increased polynomial
 2010 order. This is not unexpected as the number of time-steps needed for the explicit method grows
 2011 linearly with the reciprocal of the mesh size, and quadratically with the polynomial degree while
 2012 the implicit method maintains the number of time-steps constant. In terms of rhs evaluations, the
 2013 gap between the WH method and the HH method is smaller than between WH and IWH; though
 2014 the HH method degrades with increasing mesh refinement.

2015 Finally, in Table 4.7 we report the increase in compute time (as a multiplicative factor) for
 2016 the IWH and HH methods relative to the time required to solve the same problem with the explicit
 2017 WH method.

	p / meth	1	2	3	4	5	6	7	8	9
h	HH	15	164	56	56	60	64	62	59	∞
$h/2$	HH	265	157	109	96	117	113	∞	∞	∞
$h/4$	HH	587	311	91	260	∞	∞	∞	∞	∞
$h/8$	HH	∞	224	∞	∞	∞	∞	∞	N/A	N/A
h	IWH	192	197	137	113	103	90	59	73	44
$h/2$	IWH	169	185	125	112	107	121	55	42	29
$h/4$	IWH	261	295	61	152	62	68	71	36	42
$h/8$	IWH	28	30	100	69	62	46	34	N/A	N/A

Table 4.7: The table reports how many times longer a computation with the HH and IWH method takes compared to the explicit WH method.

2018 Throughout, the WH method is one to two orders of magnitude faster than the other methods.
 2019 The HH method becomes less attractive (especially when it stops converging) as the accuracy is
 2020 increased, while the IWH method improves with increased accuracy. It is of note that neither the
 2021 number of iterations nor the number of rhs is a good predictor of compute time. Possible causes
 2022 for this discrepancy are, a) that we only counted one right hand side evaluation per iteration and
 2023 neglected the cost of the AMG preconditioner, and b) that the GMRES solve for the HH method
 2024 actually has a significantly higher cost than conjugate gradient.

2025 **4.4.9 Materials with Spatially Varying Properties**

We next consider an example taken from [79] with elastic propagation in a heterogeneous medium. We define the domain as $\Omega = [0, 2] \times [0, 1]$ where there is an embedded inclusion, $\Omega_I = [0.5, 1.5] \times [0.4, 0.6]$, in the middle from a stiffer material. We let Ω_I be a material with $\lambda = 200, \mu = 100$, while the domain $\Omega \setminus \Omega_I$ has $\lambda = 2$ and $\mu = 1$. We impose traction-free boundary conditions at $y = 0, 1$ and at $x = 1$. We additionally have

$$\mathbf{u}(0, y, t) = \begin{pmatrix} 0 \\ \cos(\omega t) \end{pmatrix}, \quad \mathbf{f}(x, y, t) = \begin{pmatrix} 0 \\ \delta(|x - x_0| + |y - y_0|) \cos(\omega t) \end{pmatrix},$$

2026 where δ is a delta function centered at $x_0 = 0.1$ and $y_0 = 0.5$.

2027 We use the SIPDG method of Section 4.3.2 with a uniform quadrilateral mesh and poly-
 2028 mial degree $p = 6$. We consider two meshes with element widths of $h_1 = 1/20$ and $h_2 = 1/40$,
 2029 respectively. We use $CFL = 0.4$ with the (corrected) explicit leapfrog time-stepper of Section 4.3.3,
 2030 integrate over five periods, and accelerate convergence with conjugate gradient with a relative resid-
 2031 ual tolerance of 10^{-5} . We choose the frequency $\omega = 100$ so that we have at least one element per
 2032 wavelength when using element widths of h_1 , and at least two when the widths are h_2 . We plot
 2033 the \log_{10} of the magnitude of the displacement vector in Figure 4.5.

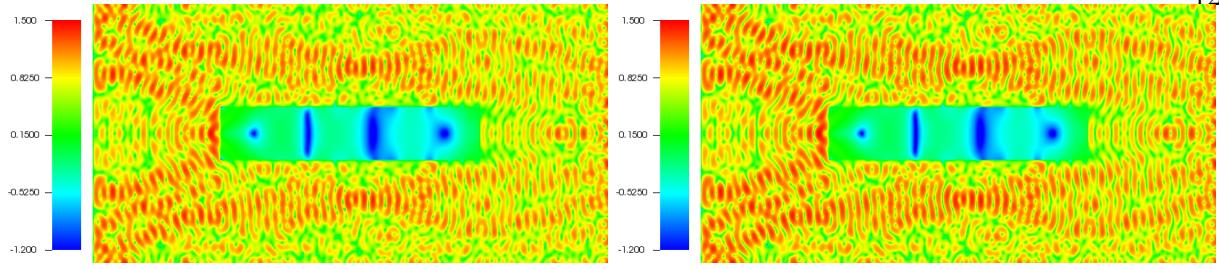


Figure 4.5: The \log_{10} of the magnitude of the displacements for the CG accelerated solution of WH for the inclusion problem using sixth order polynomials within each element. (Left) Solution using a grid resolution of at least one element per wavelength, and (Right) two elements per wavelength.

2034 From Figure 4.5 it is clear that the solution using one element per wavelength is visually quite
 2035 similar to that of the refined mesh with two elements per wavelength. Thus using one element per
 2036 wavelength with a higher order polynomial order is sufficient to produce reasonable results, as was
 2037 similarly seen in [103].

2038 **4.4.10 Vibrations of a Toroidal Shell**

2039 Finally, as a more realistic example in three dimensions we perform a simulation of a toroidal
 2040 shell parametrized by

$$x(\theta, \phi, r) = (R + r \cos(\theta)) \cos(\phi), \quad y(\theta, \phi, r) = (R + r \cos(\theta)) \sin(\phi), \quad z(\theta, \phi, r) = r \sin(\theta).$$

2041 Here we set $R = 4$, and let the partial toroidal shell occupy the volume $1 \leq r \leq 2$, and $0 \leq \phi, \theta \leq \pi$.
 2042 The surfaces at $r = 1$ and $r = 2$ are free, and we impose homogeneous Dirichlet conditions on all
 2043 other boundaries.

2044 We force the problem by

$$\mathbf{f} = \frac{\sqrt{\sigma^3}}{20} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} e^{-\zeta^2} \cos(\omega t), \quad \sigma = 100\omega, \quad \zeta^2 = 0.5\sigma((x - 4)^2 + (y - 0.5)^2 + (z - 1)^2)$$

2045 We consider two cases. In the first case the frequency is $\omega = 5.1234$ and we use a grid that consists
 2046 of $400 \times 120 \times 40$ points. In the second case the frequency is $\omega = 10.2468$ and we use a grid that
 2047 consists of $800 \times 240 \times 80$ points.

2048 The converged solution is displayed in Figure 4.6.

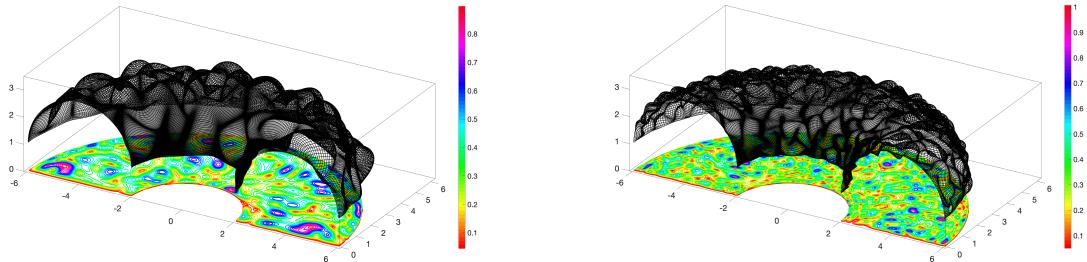


Figure 4.6: The solution in the toroidal shell for (Left) $\omega = 5.1234$, and (Right) $\omega = 10.2468$. The projection onto the xy-plane is the magnitude of the displacement on the outermost free surface $r = 2$. In black we display the (scaled) displaced mesh for $r = 2$.

2049 The projection onto the xy-plane is the magnitude of the displacement $\sqrt{u^2 + v^2 + w^2}$ on the
 2050 outermost free surface, $r = 2$. The mesh is the grid for that outermost surface with the (scaled)
 2051 displacements added to the grid coordinates.

2052 4.5 Conclusion

2053 In this chapter we applied the WaveHoltz iteration, a time-domain Krylov accelerated fixed-
 2054 point iteration, to the solution of the elastic Helmholtz equation for interior problems with Dirichlet
 2055 and/or free surface boundary conditions. With symmetric discretizations, the iteration results in a
 2056 positive definite and symmetric matrix, a notable advantage over direct discretizations of the elastic
 2057 Helmholtz equation which typically lead to highly indefinite systems. In this work we have also
 2058 introduced a new implicit time-stepping scheme and demonstrated that its use in the WaveHoltz
 2059 iteration completely removes time discretization errors.

2060 The implicit method did not offer any advantages for the one problem we considered here,

2061 but we believe it could be an advantageous approach for problems with anisotropic or refined
2062 non-conforming meshes. It could also prove to be a possible way to construct preconditioners.

2063 Finally, here we have only considered the energy conserving problem. In the future, we will
2064 revisit the elastic Helmholtz problem with impedance/absorbing boundary conditions which are a
2065 hallmark of scattering and seismic applications.

Chapter 5

2067 Optimal Control of Closed Quantum Systems via B-Splines with Carrier Waves

2068 In this chapter we consider the quantum optimal control problem of determining electromag-
 2069 netic pulses for implementing unitary gates in a quantum computer. A truncated modal expansion
 2070 of Schrödinger's equation is used to model the quantum system, in which the state of the quantum
 2071 system is described by the state vector¹ $\psi \in \mathbb{C}^N$. The elements of the state vector are complex
 2072 probability amplitudes, where the magnitude squared of each element represents the probability
 2073 that the quantum system occupies the corresponding energy level [90]. Because the probabilities
 2074 must sum to one, the state vector is normalized such that $\|\psi\|_2^2 = 1$. The evolution of the state
 2075 vector in the time interval $t \in [0, T]$ is governed by Schrödinger's equation:

$$\frac{d\psi}{dt} + iH(t; \alpha)\psi = 0, \quad 0 \leq t \leq T, \quad \psi(0) = g. \quad (5.1)$$

Here, $i = \sqrt{-1}$ is the imaginary unit and g is the initial state. The Hamiltonian matrix $H(t; \alpha) \in \mathbb{C}^{N \times N}$ (scaled such that Planck's constant becomes $\hbar = 1$) is Hermitian and is assumed to be of the form

$$H(t; \alpha) = H_s + H_c(t; \alpha), \quad (5.2)$$

2076 where H_s and H_c are the system and control Hamiltonians, respectively. The control Hamiltonian
 2077 models the action of external control fields on the quantum system. The time dependence in the
 2078 control Hamiltonian is parameterized by the control vector $\alpha \in \mathbb{R}^D$. As a result, the state vector
 2079 ψ depends implicitly on α through Schrödinger's equation.

¹ This chapter uses conventional matrix-vector notation. For finite-dimensional systems, it is equivalent to the bra-ket notation that often is used in quantum physics.

2080 To justify the truncation of the modal expansion of Schrödinger's equation, we divide the
 2081 state vector into $E > 0$ "essential" levels and $G \geq 0$ "guard" levels, such that $E + G = N$. The
 2082 population of the highest guard levels need to be small to minimize coupling to even higher energy
 2083 levels, which are excluded from the model.

2084 The goal of the quantum optimal control problem is to determine the parameter vector α such
 2085 that the time-dependence in the Hamiltonian matrix leads to a solution of Schrödinger's equation
 2086 such that $\psi(T) \approx V_{tg}g$, where V_{tg} is the target gate transformation. The gate transformation
 2087 should be satisfied for all initial conditions in the essential subspace of the state vector. A basis of
 2088 this subspace is provided by the matrix $U_0 \in \mathbb{R}^{N \times E}$. The definitions of U_0 and V_{tg} are described
 2089 in Appendix .12.

2090 To account for any initial condition in the essential subspace, we define the solution operator
 2091 matrix $U \in \mathbb{C}^{N \times E}$. Each column of this matrix satisfies (5.1), leading to Schrödinger's equation in
 2092 matrix form,

$$\frac{dU(t)}{dt} + iH(t; \alpha)U(t) = 0, \quad 0 \leq t \leq T, \quad U(0) = U_0. \quad (5.3)$$

The overlap between the target gate matrix and the solution operator matrix at the final time is defined by

$$O_{V_{tg}} := \langle U(T; \alpha), V_{tg} \rangle_F, \quad (5.4)$$

2093 where $\langle \cdot, \cdot \rangle_F$ denotes the Frobenius matrix scalar product. Because U_0 spans an E -dimensional
 2094 subspace of initial conditions, we have $|O_T| \leq E$. The difference between $U(T; \alpha)$ and V_{tg} can be
 2095 measured by the target gate infidelity [71, 78, 82, 83, 98],

$$\mathcal{J}_1(U_T(\alpha)) := 1 - \frac{1}{E^2} |\langle U_T(\alpha), V_{tg} \rangle_F|^2, \quad U_T(\alpha) := U(T; \alpha). \quad (5.5)$$

2096 Note that the target gate infidelity is invariant to global phase differences between U_T and V_{tg} . In
 2097 quantum physics, the global phase of a state is considered irrelevant because it can not be measured.

2098 The leakage of population to the guard states can be measured by the objective function

$$\mathcal{J}_2(U(\cdot; \alpha)) = \frac{1}{T} \int_0^T \langle U(t; \alpha), WU(t; \alpha) \rangle_F dt, \quad (5.6)$$

2099 where W is a diagonal $N \times N$ positive semi-definite weight matrix. The elements in W are zero
 2100 for all essential states and are positive for the guard states. The elements of W are typically larger
 2101 for higher energy levels in the model.

For the quantum control problem with guard states, we formulate the optimization problem
 as

$$\min_{\boldsymbol{\alpha}} \mathcal{G}(\boldsymbol{\alpha}) := \mathcal{J}_1(U_T(\boldsymbol{\alpha})) + \mathcal{J}_2(U(\cdot; \boldsymbol{\alpha})), \quad (5.7)$$

$$\frac{dU}{dt} + iH(t; \boldsymbol{\alpha})U = 0, \quad 0 \leq t \leq T, \quad U(0; \boldsymbol{\alpha}) = U_0, \quad (5.8)$$

$$\alpha_{min} \leq \alpha_q \leq \alpha_{max}, \quad q = 1, 2, \dots, D. \quad (5.9)$$

2102 For a discussion of the solvability of the quantum control problem, see for example Borzi et al. [27].

2103 In the quantum optimal control problem, the Schrödinger (state) equation is a time-dependent
 2104 Hamiltonian system. To ensure long-time numerical accuracy it is appropriate to discretize it using
 2105 a symplectic time-integration method [67]. For this purpose we use the Störmer-Verlet method,
 2106 which can be written as a partitioned Runge-Kutta scheme, based on the trapezoidal and implicit
 2107 midpoint rules. Our main theoretical contribution is the generalization of Ober-Blöbaum's [92]
 2108 work to the case of a time-dependent Hamiltonian system. We show that the compatible method
 2109 for the adjoint state equation resembles a partitioned Runge-Kutta scheme, except that the time-
 2110 dependent matrices must be evaluated at modified time levels.

2111 Our approach builds upon the works of Hager [65], Sanz-Serna [97] and Ober-Blöbaum [92].
 2112 Hager [65] first showed how the Hamiltonian structure in an optimization problem can be utilized
 2113 to calculate the gradient of the objective function. Hager considered the case in which the state
 2114 equation is discretized by one Runge-Kutta scheme, with the adjoint state equation discretized by
 2115 another Runge-Kutta scheme. It was found that the discrete gradient can be calculated exactly if
 2116 the pair of Runge-Kutta methods satisfy the requirements of a symplectic partitioned Runge-Kutta
 2117 method. Further details and generalizations are described in the review paper by Sanz-Serna [97].
 2118 Ober-Blöbaum [92] extended Hager's approach to the case where the state equation itself is a
 2119 Hamiltonian system that is discretized by a partitioned Runge-Kutta scheme. For autonomous

2120 state equations, it was shown that the compatible discretization of the adjoint state equation is
 2121 another partitioned Runge-Kutta scheme.

2122 Several numerical methods for the quantum control problem are based on the GRAPE al-
 2123 gorithm [73]. In this case, Schrödinger’s equation is discretized in time using the second order
 2124 accurate Magnus scheme [67], in which the Hamiltonian matrix is evaluated at the midpoint of
 2125 each time step. A stair-step approximation of the control functions is imposed such that each con-
 2126 trol function is constant within each time step. Thus, the time step determines both the numerical
 2127 accuracy of the dynamics of the quantum state **and** the number of control parameters. With Q
 2128 control functions, M time steps of size h , the control functions are thus described by M times Q
 2129 parameters $\alpha_{j,k}$. The propagator in the Magnus method during the j^{th} time step is of the form
 2130 $\exp(-ih(H_0 + \sum_k \alpha_{k,j} H_k))$. In general, the matrices H_0 and H_k do not commute, leading to an
 2131 integral expression for the derivative of the propagator with respect to the parameters, which is
 2132 needed for computing the gradient of the objective function. In the original GRAPE method, this
 2133 integral expression is approximated by the first term in its Taylor series expansion, leading to an
 2134 approximate gradient that is polluted by an $\mathcal{O}(h^2)$ error. As the gradient becomes smaller during
 2135 the optimization, the approximation error will eventually dominate the numerical gradient, which
 2136 may hamper the convergence of the optimization algorithm. A more accurate way of numerically
 2137 evaluating the derivative of the time-step propagator can be obtained by retaining more terms in
 2138 the Taylor series expansion, or by using a matrix commutator expansion [36]. More recently, the
 2139 GRAPE algorithm has been generalized to optimize objective functions that include a combination
 2140 of the target gate infidelity, integrals penalizing occupation of “forbidden states” and terms for
 2141 imposing smoothness and amplitude constraints on the control functions. Here, automatic differ-
 2142 entiation is used for computing the gradient of the objective function [78]. However, the number of
 2143 control parameters is still proportional to the number of time steps, which may become very large
 2144 when the duration of the gate is long, or the quantum state is highly oscillatory.

2145 As an alternative to calculating the gradient of the objective function by solving an adjoint
 2146 equation backwards in time, the gradient can be calculated by differentiating Schrödinger’s equa-

2147 tion with respect to each parameter in the control function, leading to a differential equation for
 2148 each component of the gradient of the state vector. This approach, implemented in the GOAT
 2149 algorithm [83], allows the gradient of the objective function to be calculated exactly, but requires
 2150 $(D + 1)$ Schrödinger systems to be solved when the control functions depend on D parameters.
 2151 This makes the method computationally expensive when the number of parameters is large.

2152 Using the stair-stepped approximation of the control functions often leads to a large number
 2153 of control parameters, which may hamper the convergence of the GRAPE algorithm. The total
 2154 number of parameters can be reduced by instead expanding the control functions in terms of basis
 2155 functions. By using the chain rule, the gradient from the GRAPE algorithm can then be used
 2156 to calculate the gradient with respect to the coefficients in the basis function expansion. This
 2157 approach is implemented in the GRAFS algorithm [82], where the control functions are expanded
 2158 in terms of Slepian sequences.

2159 Gradient-free optimization methods can also be applied to quantum optimal control problems.
 2160 These methods do not rely on the gradient to be evaluated and are therefore significantly easier to
 2161 implement. However, the convergence of these methods is usually much slower than for gradient-
 2162 based techniques, unless the number of control parameters is very small. One example of a gradient-
 2163 free methods for quantum optimal control is the CRAB algorithm [32].

2164 Many parameterizations of quantum control functions have been proposed in the literature,
 2165 for example cubic splines [48], Gaussian pulse cascades [38], Fourier expansions [110] and Slepian
 2166 sequences [82].

2167 This chapter presents a different approach, based on parameterizing the control functions by
 2168 B-spline basis functions with carrier waves. Our approach relies on the observation that transitions
 2169 between the energy levels in a quantum system are triggered by resonance, at frequencies that can
 2170 be determined by inspection of the system Hamiltonian. The carrier waves are used to specify the
 2171 frequency spectra of the control functions, while the B-spline functions specify their envelope and
 2172 phase. We find that this approach allows the number of control parameters to be independent of,
 2173 and significantly smaller than, the number of time steps for integrating Schrödinger's equation.

2174 The remainder of the chapter is organized as follows. In Section 5.1, we introduce a Hamiltonian
 2175 model and discuss the resonant frequencies needed to trigger transitions between the states in
 2176 the system. These resonant frequencies naturally motivate us to parameterize the control functions
 2177 using B-splines with carrier waves; details of this parameterization are presented in Section 5.2.
 2178 In Section 5.3, we introduce a real-valued formulation of Schrödinger’s equation and present the
 2179 symplectic Störmer-Verlet scheme that we use for its time-integration. To achieve an exact gradient
 2180 of the discrete objective function, we apply the “discretize-then-optimize” approach. Based on the
 2181 Störmer-Verlet scheme, in Section 5.4 we outline the construction of a discrete adjoint time integra-
 2182 tion method. Section 5.5 presents numerical examples. We illustrate how the proposed technique,
 2183 combined with the interior point L-BFGS algorithm [91] from the IPOPT package [106], is used
 2184 to optimize control functions for multi-level qudit gates. We additionally consider a simple noise
 2185 model and risk-neutral optimization to illustrate the construction of controls that are robust to
 2186 uncertainty in the Hamiltonian. The proposed scheme is implemented in the Julia [24] program-
 2187 ming language, in an open source package called Juqbox.jl [53]. In Section 5.6, we compare the
 2188 performance of Juqbox.jl and two implementations of the GRAPE algorithm. Concluding remarks
 2189 are given in Section 5.7.

2190 5.1 Hamiltonian model

Several Hamiltonian models exist for describing the quantum physics of super-conducting circuits [26, 84]. In this chapter, we consider a composite quantum system with $Q \geq 1$ subsystems (qubits/qudits) where the system Hamiltonian satisfies:

$$H_s = \sum_{q=1}^Q \left(\omega_q a_q^\dagger a_q - \frac{\xi_q}{2} a_q^\dagger a_q^\dagger a_q a_q - \sum_{p>q} \xi_{pq} a_p^\dagger a_p a_q^\dagger a_q \right). \quad (5.10)$$

In this model, ω_q is the ground state transition frequency and ξ_q is the self-Kerr coefficient of subsystem q ; the cross-Kerr coefficient between subsystems p and q is denoted ξ_{pq} . Furthermore, subsystem q is assumed to have $n_q \geq 2$ energy levels, with lowering operator a_q . The lowering

operator is constructed using Kronecker products,

$$a_q := I_{n_Q} \otimes \cdots \otimes I_{n_{q+1}} \otimes A_q \otimes I_{n_{q-1}} \otimes \cdots \otimes I_{n_1} \in \mathbb{R}^{N \times N}, \quad N = \prod_{q=1}^Q n_q, \quad (5.11)$$

where I_n denotes the $n \times n$ identity matrix and the single-system lowering matrix satisfies

$$A_q := \begin{pmatrix} 0 & \sqrt{1} & & & \\ & \ddots & \ddots & & \\ & & \ddots & \sqrt{n_q - 1} & \\ & & & & 0 \end{pmatrix} \in \mathbb{R}^{n_q \times n_q}. \quad (5.12)$$

We consider a control Hamiltonian with real-valued control functions that are parameterized by the control vector $\boldsymbol{\alpha}$,

$$H_c(t; \boldsymbol{\alpha}) = \sum_{q=1}^Q f_q(t; \boldsymbol{\alpha})(a_q + a_q^\dagger), \quad f_q(t; \boldsymbol{\alpha}) = 2 \operatorname{Re}(d_q(t; \boldsymbol{\alpha}) e^{i\omega_{r,q} t}). \quad (5.13)$$

where $\omega_{r,q}$ is the drive frequency in subsystem q .

5.1.1 Rotating wave approximation

To slow down the time scales in the state vector, we apply a rotating frame transformation in Schrödinger's equation through the unitary change of variables $\tilde{\psi}(t) = R(t)\psi(t)$, where

$$R(t) = \bigotimes_{q=Q}^1 \exp\left(i\omega_{r,q} t A_q^\dagger A_q\right), \quad (5.14)$$

and $\otimes_{q=Q}^1 C_q = C_Q \otimes C_{Q-1} \otimes \cdots \otimes C_1$. Note that we use $\omega_{r,q}$ as the frequency of rotation in subsystem q . The system Hamiltonian transforms into $H_s^{rw} = H_s - \sum \omega_{r,q} a_q^\dagger a_q$. Then, the rotating wave approximation is applied to transform the control Hamiltonian. Here, we substitute the laboratory frame control function $f_q(t; \boldsymbol{\alpha})$ from (5.13) and neglect terms oscillating with frequencies $\pm 2\omega_{r,q}$. As a result, the Hamiltonians (5.10) and (5.13) transform into (see Appendix .13 for details)

$$H_s^{rw} = \sum_{q=1}^Q \left(\Delta_q a_q^\dagger a_q - \frac{\xi_q}{2} a_q^\dagger a_q^\dagger a_q a_q - \sum_{p>q} \xi_{qp} a_q^\dagger a_q a_p^\dagger a_p \right), \quad (5.15)$$

$$H_c^{rw}(t; \boldsymbol{\alpha}) = \sum_{q=1}^Q \left(d_q(t; \boldsymbol{\alpha}) a_q + \bar{d}_q(t; \boldsymbol{\alpha}) a_q^\dagger \right), \quad (5.16)$$

where $\Delta_q = \omega_q - \omega_{r,q}$ is called the detuning frequency. The main advantages of the rotating frame approximation are the reduction of the spectral radius in the system Hamiltonian (5.15), and the absence of the highly oscillatory factor $\exp(i\omega_{r,q}t)$ in the control Hamiltonian (5.16). In the following we assume that the rotating wave approximation has already been performed, and the tilde on the state vector will be suppressed. We additionally note that the target unitary V_{tg} is similarly transformed into the rotating frame via $V_{tg}^{rw} = R(T)V_{tg}$.

5.1.2 Resonant frequencies

To simplify the presentation we restrict our analysis to a bipartite quantum system, i.e., $Q = 2$. The system Hamiltonian (5.15) is diagonal and we denote its elements by

$$\{H_s^{rw}\}_{j,k} = \begin{cases} \kappa_j, & j = k, \\ 0, & \text{otherwise,} \end{cases} \quad \kappa_j = \sum_{q=1}^2 \left(\Delta_q j_q - \frac{\xi_q}{2} j_q(j_q - 1) \right) - \xi_{12} j_1 j_2, \quad (5.17)$$

for $j_q \in [0, n_q - 1]$ and where $j = (j_2, j_1)$ is a multi-index. Let us consider the case when the control functions $d_k(t)$ oscillate with carrier wave frequencies $\{\Omega_1, \Omega_2\}$, and amplitude ϵ , where $0 < \epsilon \ll 1$. These assumptions give

$$H_c^{rw}(t) = \epsilon H_1(t), \quad H_1(t) = \sum_{k=1}^2 \left(e^{i\Omega_k t} a_k + e^{-i\Omega_k t} a_k^\dagger \right). \quad (5.18)$$

We make an asymptotic expansion of the solution of Schrödinger's equation (5.1), $\psi = \psi^{(0)} + \epsilon \psi^{(1)} + \mathcal{O}(\epsilon^2)$. The zero'th and first order terms satisfy

$$\frac{d\psi^{(0)}}{dt} + i H_s^{rw} \psi^{(0)} = 0, \quad \psi^{(0)}(0) = g, \quad (5.19)$$

$$\frac{d\psi^{(1)}}{dt} + i H_s^{rw} \psi^{(1)} = f(t), \quad \psi^{(1)}(0) = 0. \quad (5.20)$$

Because the system Hamiltonian is diagonal, (5.19) is a decoupled system of ordinary differential equation that is solved by $\psi_j^{(0)}(t) = g_j e^{-i\kappa_j t}$. The right hand side of (5.20) satisfies $f(t) := -i H_1(t) \psi^{(0)}(t)$, which can be written

$$f(t) = \sum_{k=1}^Q f^{(k)}(t), \quad f^{(k)}(t) = -i \left(e^{i\Omega_k t} a_k + e^{-i\Omega_k t} a_k^\dagger \right) \psi^{(0)}(t). \quad (5.21)$$

Because the matrix H_s^{rw} is diagonal, the system for the first order perturbation, (5.20), is also decoupled. We are interested in cases when $\psi_j^{(1)}(t)$ grows in time, corresponding to resonance. Let e_k denote the k^{th} unit vector and denote a shifted multi-index by $j \pm e_1 = (j_2, j_1 \pm 1)$ and $j \pm e_2 = (j_2 \pm 1, j_1)$.

Lemma 5.1.1. *The perturbation of the state vector, $\psi_j^{(1)}(t)$, grows linearly in time when the carrier wave frequencies and the initial condition satisfy:*

$$\Omega_k = \kappa_{j+e_k} - \kappa_j, \quad g_{j+e_k} \neq 0, \quad j_k \in [0, n_k - 2], \quad (5.22)$$

$$\Omega_k = \kappa_j - \kappa_{j-e_k}, \quad g_{j-e_k} \neq 0, \quad j_k \in [1, n_k - 1], \quad (5.23)$$

for $k = \{1, 2\}$.

Proof. See Appendix .14. \square

We can evaluate the conditions for resonance by inserting the Hamiltonian elements from (5.17) into (5.22) and (5.23). For $k = 1$ and $j_2 \in [0, n_2 - 1]$, resonance occurs in $\psi_j^{(1)}(t)$ when

$$\Omega_1 = \begin{cases} \Delta_1 - \xi_1 j_1 - \xi_{12} j_2, & g_{j+e_1} \neq 0, \quad j_1 \in [0, n_1 - 2], \\ \Delta_1 - \xi_1(j_1 - 1) - \xi_{12} j_2, & g_{j-e_1} \neq 0, \quad j_1 \in [1, n_1 - 1]. \end{cases} \quad (5.24)$$

For $k = 2$ and $j_1 \in [0, n_1 - 1]$, the resonant cases are

$$\Omega_2 = \begin{cases} \Delta_2 - \xi_2 j_2 - \xi_{12} j_1, & g_{j+e_2} \neq 0, \quad j_2 \in [0, n_2 - 2], \\ \Delta_2 - \xi_2(j_2 - 1) - \xi_{12} j_1, & g_{j-e_2} \neq 0, \quad j_2 \in [1, n_2 - 1]. \end{cases} \quad (5.25)$$

For example, when $n_1 = 3$, $n_2 = 3$ and $g_j \neq 0 \forall j$, the carrier wave frequencies:

$$\Omega_1 = \left[\Delta_1, \quad \Delta_1 - \xi_{12}, \quad \Delta_1 - 2\xi_{12}, \quad \Delta_1 - \xi_1, \quad \Delta_1 - \xi_1 - \xi_{12}, \quad \Delta_1 - \xi_1 - 2\xi_{12} \right],$$

$$\Omega_2 = \left[\Delta_2, \quad \Delta_2 - \xi_{12}, \quad \Delta_2 - 2\xi_{12}, \quad \Delta_2 - \xi_2, \quad \Delta_2 - \xi_2 - \xi_{12}, \quad \Delta_2 - \xi_2 - 2\xi_{12} \right],$$

lead to resonance.

Since Schrödinger's equation conserves total probability, the linear growth in time only occurs for short times. Thus, each resonant frequency corresponds to the initiation of a transition between two energy levels in the quantum system.

2210 **5.2 Quadratic B-splines with carrier waves**

Motivated by the results from the previous section, we parameterize the rotating frame control functions using basis functions that act as envelopes for carrier waves with fixed frequencies:

$$d_k(t; \boldsymbol{\alpha}) = \sum_{n=1}^{N_f} d_{k,n}(t; \boldsymbol{\alpha}), \quad d_{k,n}(t; \boldsymbol{\alpha}) = \sum_{b=1}^{N_b} \hat{S}_b(t) \alpha_{b,n}^k e^{it\Omega_{k,n}}, \quad k \in [1, Q]. \quad (5.26)$$

Here, $\Omega_{k,n} \in \mathbb{R}$ is the n^{th} carrier wave frequency for system k . These frequencies are chosen to match the resonant frequencies in the system Hamiltonian (5.15), as outlined above. The complex coefficients $\alpha_{b,n}^k = \alpha_{b,n}^{k(r)} + i\alpha_{b,n}^{k(i)}$ are control parameters that are to be determined through optimization, corresponding to a total of $D = 2QN_bN_f$ real-valued parameters. It is convenient to also define the real-valued functions

$$p_{k,n}(t; \boldsymbol{\alpha}) = \sum_{b=1}^{N_b} \hat{S}_b(t) \alpha_{b,n}^{k(r)}, \quad q_{k,n}(t; \boldsymbol{\alpha}) = \sum_{b=1}^{N_b} \hat{S}_b(t) \alpha_{b,n}^{k(i)}, \quad (5.27)$$

2211 such that $d_{k,n}(t; \boldsymbol{\alpha}) = (p_{k,n}(t; \boldsymbol{\alpha}) + iq_{k,n}(t; \boldsymbol{\alpha})) \exp(it\Omega_{k,n})$.

2212 The basis functions $\hat{S}_b(t)$ are chosen to be piece-wise quadratic B-spline wavelets (see Fig-
2213 ure 5.1), centered on a uniform grid in time,

$$t_m = (m - 1.5)\delta, \quad m = 1, \dots, D_1, \quad \delta = \frac{T}{D_1 - 2}. \quad (5.28)$$

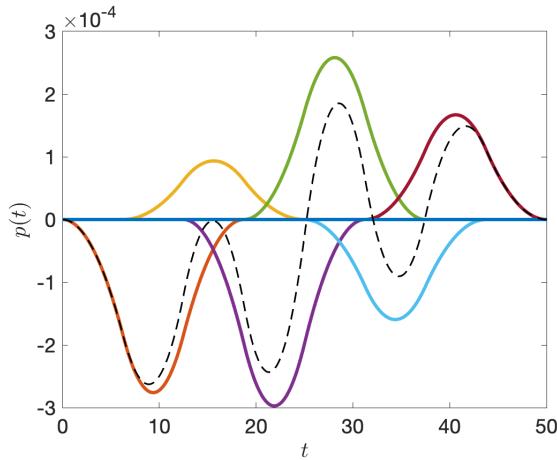


Figure 5.1: The real part of a quadratic B-spline control function, with zero carrier frequency (dashed black). The solid colored lines are the individual B-spline wavelets.

Each basis function $\hat{S}_b(t)$ is centered around $t = t_m$ and is easily expressed in terms of the scaled time parameter $\tau_m(t) = (t - t_m)/3\delta$,

$$\hat{S}_b(t) = \tilde{S}(\tau_m(t)), \quad \tilde{S}(\tau) = \begin{cases} \frac{9}{8} + \frac{9}{2}\tau + \frac{9}{2}\tau^2, & -\frac{1}{2} \leq \tau < -\frac{1}{6}, \\ \frac{3}{4} - 9\tau^2, & -\frac{1}{6} \leq \tau < \frac{1}{6}, \\ \frac{9}{8} - \frac{9}{2}\tau + \frac{9}{2}\tau^2, & \frac{1}{6} \leq \tau < \frac{1}{2}, \\ 0, & \text{otherwise.} \end{cases} \quad (5.29)$$

2214 The basis function $\hat{S}_b(t)$ has local support for $t \in [t_m - 1.5\delta, t_m + 1.5\delta]$. Thus, for any fixed time t
 2215 a control function will get contributions from at most three B-spline wavelets, allowing the control
 2216 functions to be evaluated very efficiently. We also remark that the control function (5.26) can be
 2217 evaluated at any time $t \in [0, T]$. Importantly, this allows the time-integration scheme to be chosen
 2218 independently of the parameterization of the control function, and allows the number of control
 2219 parameters to be chosen independently of the number of time steps for integrating Schrödinger's
 2220 equation.

2221 5.3 Real-valued formulation

2222 A real-valued formulation of Schrödinger's equation (5.1) is given by

$$\begin{bmatrix} \dot{\mathbf{u}} \\ \dot{\mathbf{v}} \end{bmatrix} = \begin{bmatrix} S(t) & -K(t) \\ K(t) & S(t) \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix} =: \begin{bmatrix} f^u(\mathbf{u}, \mathbf{v}, t) \\ f^v(\mathbf{u}, \mathbf{v}, t) \end{bmatrix}, \quad \begin{bmatrix} \mathbf{u}(0) \\ \mathbf{v}(0) \end{bmatrix} = \begin{bmatrix} \mathbf{g}^u \\ \mathbf{g}^v \end{bmatrix}, \quad (5.30)$$

2223 where,

$$\mathbf{u} = \operatorname{Re}(\psi), \quad \mathbf{v} = -\operatorname{Im}(\psi), \quad K = \operatorname{Re}(H), \quad S = \operatorname{Im}(H),$$

2224 Because the matrix H is Hermitian, $K^T = K$ and $S^T = -S$. The real-valued formulation of
 2225 Schrödinger's equation is a time-dependent Hamiltonian system corresponding to the Hamiltonian
 2226 functional,

$$\mathcal{H}(\mathbf{u}, \mathbf{v}, t) = \mathbf{u}^T S(t) \mathbf{v} + \frac{1}{2} \mathbf{u}^T K(t) \mathbf{u} + \frac{1}{2} \mathbf{v}^T K(t) \mathbf{v}. \quad (5.31)$$

2227 In general, $S(t) \neq 0$, which makes the Hamiltonian system non-separable.

2228 In terms of the real-valued formulation, let the columns of the solution operator matrix in
 2229 (5.3) satisfy $U = [\mathbf{u}_1 - i\mathbf{v}_1, \mathbf{u}_2 - i\mathbf{v}_2, \dots, \mathbf{u}_E - i\mathbf{v}_E]$. Here, $(\mathbf{u}_j, \mathbf{v}_j)$ satisfy (5.30) subject to the
 2230 initial conditions $\mathbf{g}_j^v = \mathbf{0}$ and $\mathbf{g}_j^u = \mathbf{e}_j$, where $j = (j_Q, j_{Q-1}, \dots, j_1)$ is a multi-index such that
 2231 $j_q \in \{0, 1, \dots, m_q - 1\}$ and m_q is the number of essential levels of subsystem q . The columns in the
 2232 target gate matrix $V_{tg} = [\mathbf{d}_1, \dots, \mathbf{d}_E]$ correspond to

$$V_{tg} = [\mathbf{d}_1^u - i\mathbf{d}_1^v, \mathbf{d}_2^u - i\mathbf{d}_2^v, \dots, \mathbf{d}_E^u - i\mathbf{d}_E^v], \quad \mathbf{d}_j^u = \text{Re}(\mathbf{d}_j), \quad \mathbf{d}_j^v = -\text{Im}(\mathbf{d}_j).$$

2233 Using the real-valued notation, the two parts of the objective function (5.7) can be written

$$\mathcal{J}_1(U_T(\boldsymbol{\alpha})) = \left(1 - \frac{1}{E^2} |O_V(U_T(\boldsymbol{\alpha}))|^2\right), \quad (5.32)$$

$$\mathcal{J}_2(U(\cdot, \boldsymbol{\alpha})) = \frac{1}{T} \sum_{j=0}^{E-1} \int_0^T \langle \mathbf{u}_j(t, \boldsymbol{\alpha}) - i\mathbf{v}_j(t, \boldsymbol{\alpha}), W(\mathbf{u}_j(t, \boldsymbol{\alpha}) - i\mathbf{v}_j(t, \boldsymbol{\alpha})) \rangle_2 dt, \quad (5.33)$$

where

$$O_V(U_T) = \sum_{j=0}^{E-1} \langle \mathbf{u}_j(T, \boldsymbol{\alpha}) - i\mathbf{v}_j(T, \boldsymbol{\alpha}), \mathbf{d}_j^u - i\mathbf{d}_j^v \rangle_2. \quad (5.34)$$

2234

2235 5.3.1 Time integration

Let $t_n = nh$, for $n = 0, 1, \dots, M$, be a uniform grid in time where $h = T/M$ is the time step. Also let $\mathbf{u}^n \approx \mathbf{u}(t_n)$ and $\mathbf{v}^n \approx \mathbf{v}(t_n)$ denote the numerical solution on the grid. We use a partitioned Runge-Kutta (PRK) scheme [67] to discretize the real-valued formulation of Schrödinger's equation,

$$\mathbf{u}^0 = \mathbf{g}^u, \quad \mathbf{v}^0 = \mathbf{g}^v, \quad (5.35)$$

$$\mathbf{u}^{n+1} = \mathbf{u}^n + h \sum_{i=1}^s b_i^u \boldsymbol{\kappa}^{n,i}, \quad \mathbf{v}^{n+1} = \mathbf{v}^n + h \sum_{i=1}^s b_i^v \boldsymbol{\ell}^{n,i}, \quad (5.36)$$

$$\boldsymbol{\kappa}^{n,i} = f^u(\mathbf{U}^{n,i}, \mathbf{V}^{n,i}, t_n + c_i^u h), \quad \boldsymbol{\ell}^{n,i} = f^v(\mathbf{U}^{n,i}, \mathbf{V}^{n,i}, t_n + c_i^v h), \quad (5.37)$$

$$\mathbf{U}^{n,i} = \mathbf{u}^n + h \sum_{j=1}^s a_{ij}^u \boldsymbol{\kappa}^{n,j}, \quad \mathbf{V}^{n,i} = \mathbf{v}^n + h \sum_{j=1}^s a_{ij}^v \boldsymbol{\ell}^{n,j}. \quad (5.38)$$

2236 Here, $s \geq 1$ is the number of stages. The stage variables $\mathbf{U}^{n,i}$ and $\mathbf{V}^{n,i}$ are set in a bold font to
2237 indicate that they are unrelated to the solution operator matrix $U(t, \boldsymbol{\alpha})$ and the target gate matrix
2238 V_{tg} .

The Störmer-Verlet scheme is a two-stage PRK method ($s = 2$) that is symplectic, time-reversible and second order accurate [67]. It combines the trapezoidal and the implicit midpoint rules, with Butcher coefficients:

$$a_{11}^u = a_{12}^u = 0, \quad a_{21}^u = a_{22}^u = \frac{1}{2}, \quad a_{11}^v = a_{21}^v = \frac{1}{2}, \quad a_{12}^v = a_{22}^v = 0, \quad (5.39)$$

$$b_1^u = b_2^u = \frac{1}{2}, \quad c_1^u = 0, \quad c_2^u = 1, \quad b_1^v = b_2^v = \frac{1}{2}, \quad c_1^v = c_2^v = \frac{1}{2}. \quad (5.40)$$

2239

2240 5.3.2 Time step restrictions for accuracy and stability

2241 The accuracy in the numerical solution of Schrödinger's equation is essentially determined
2242 by how well the fastest time scale in the state vector is resolved on the grid in time. The analysis
2243 of the time scales in the solution of Schrödinger's equation is most straightforward to perform in
2244 the complex-valued formulation (5.1).

There are two fundamental time scales that must be resolved in the solution of Schrödinger's equation. The first corresponds to how quickly the control functions must vary in time to trigger the desired transitions between the energy levels in the quantum system. This time scale is determined by the transition frequencies in the system Hamiltonian, which follow as the difference between its consecutive eigenvalues. In the Hamiltonian model (5.15) and (5.16), the angular transition frequencies between the essential energy levels (with detuning frequency Δ_1) are

$$\Omega_{1,n} = \Delta_1 - n\xi_1, \quad n = 0, \dots, N_f - 1.$$

2245 The second time scale is due to the harmonic oscillation of the phase in the state vector. It can
2246 be estimated by freezing the time-dependent coefficients in the Hamiltonian matrix at some time
2247 $t = t_*$ and considering Schrödinger's equation with the time-independent Hamiltonian matrix $H_* =$

2248 $H(t_*)$. The $N \times N$ matrix H_* is Hermitian and can be diagonalized by a unitary transformation,

$$H_* X = X \Gamma, \quad X^\dagger X = I_N, \quad \Gamma = \text{diag}(\gamma_1, \gamma_2, \dots, \gamma_N),$$

2249 where the eigenvalues γ_k are real. By the change of variables $\tilde{\psi} = X^\dagger \psi$, the solution of the
2250 diagonalized system follows as

$$\tilde{\psi}_k(t) = e^{-i\gamma_k t} \tilde{\psi}_k(0),$$

2251 corresponding to the period $\tau_k = 2\pi/|\gamma_k|$. The shortest period thus follows from the spectral radius
2252 of H_* , $\rho(H_*) = \max_k |\gamma_k|$.

To estimate the time step for the Störmer-Verlet method, we require that the shortest period
in the solution of Schrödinger's equation must be resolved by at least C_P time steps. Taking both
time scales into account leads to the time step restriction

$$h \leq \frac{2\pi}{C_P \max\{\rho(H_*), \max_n(|\Omega_{1,n}|)\}}. \quad (5.41)$$

2253 The value of C_P that is needed to obtain a given accuracy in the numerical solution depends on the
2254 order of accuracy, the duration of the time integration, as well as the details of the time-stepping
2255 scheme. For second order accurate methods such as the Störmer-Verlet method, acceptable accuracy
2256 for engineering applications can often be achieved with $C_P \approx 40$. With the Störmer-Verlet method,
2257 we note that the time-stepping can become unstable if $C_P \leq 2$, corresponding to a sampling rate
2258 below the Nyquist limit.

After freezing the coefficients, the Hamiltonian (5.15) and (5.16) becomes

$$H_* = -\frac{\xi_a}{2} a^\dagger a^\dagger aa + p_*(a + a^\dagger) + iq_*(a - a^\dagger), \quad p_* = p(t_*, \boldsymbol{\alpha}), \quad q_* = q(t_*, \boldsymbol{\alpha}).$$

We can estimate the spectral radius of $H_* \in \mathbb{C}^{N \times N}$ using the Gershgorin circle theorem [59].
Because H_* is Hermitian, all its eigenvalues are real. As a result, its spectral radius can be bounded
by

$$\rho(H_*) \leq \frac{|\xi_a|}{2}(N-1)(N-2) + 2d_\infty \sqrt{N-1}.$$

2259 Here we have used that the control function is bounded by $d_\infty = \max_t |d_1(t, \boldsymbol{\alpha})|$ for a given pa-
 2260 rameter vector $\boldsymbol{\alpha}$, in the interval $0 \leq t \leq T$. With this estimate in (5.41) we guarantee that the
 2261 time-dependent phase in the state vector is resolved by at least C_P time steps per shortest
 2262 period.

2263 If the optimization imposes amplitude constraints on the parameter vector, $|\boldsymbol{\alpha}|_\infty \leq \alpha_{max}$,
 2264 those constraints can be used to estimate the time step before the optimization starts. This allows
 2265 the same time step to be used throughout the iteration and eliminates the need to recalculate the
 2266 spectral radius of H_* when $\boldsymbol{\alpha}$ changes.

2267 Our implementation of the Störmer-Verlet scheme was verified to be second order accurate. It
 2268 was also found to give approximately the same accuracy as the second order Magnus integrator [67]
 2269 when the same time step was used in both methods.

2270 5.4 Discretizing the objective function and its gradient

2271 In this section, we develop a “discretize-then-optimize” approach in which we first discretize
 2272 the objective function and then derive a compatible scheme for discretizing the adjoint state equa-
 2273 tion, which is used for computing the gradient of the objective function. As was outlined in
 2274 the introduction, our approach builds upon the works of Hager [65], Sanz-Serna [97] and Ober-
 2275 Blöbaum [92].

2276 5.4.1 Discretizing the objective function

The Störmer-Verlet scheme can be written in terms of the stage variables $(\mathbf{U}^{n,i}, \mathbf{V}^{n,i})$ by substituting $(\boldsymbol{\kappa}^{n,i}, \boldsymbol{\ell}^{n,i})$ from (5.37) into (5.36),

$$\mathbf{u}^0 = \mathbf{g}^u, \quad \mathbf{v}^0 = \mathbf{g}^v, \tag{5.42}$$

$$\mathbf{u}^{n+1} = \mathbf{u}^n + \frac{h}{2} (S_n \mathbf{U}^{n,1} + S_{n+1} \mathbf{U}^{n,2} - K_n \mathbf{V}^{n,1} - K_{n+1} \mathbf{V}^{n,2}), \tag{5.43}$$

$$\mathbf{v}^{n+1} = \mathbf{v}^n + \frac{h}{2} (K_{n+1/2} (\mathbf{U}^{n,1} + \mathbf{U}^{n,2}) + S_{n+1/2} (\mathbf{V}^{n,1} + \mathbf{V}^{n,2})), \tag{5.44}$$

and into (5.38),

$$\mathbf{U}^{n,1} = \mathbf{u}^n, \quad (5.45)$$

$$\mathbf{U}^{n,2} = \mathbf{u}^n + \frac{h}{2} (S_n \mathbf{U}^{n,1} + S_{n+1} \mathbf{U}^{n,2} - K_n \mathbf{V}^{n,1} - K_{n+1} \mathbf{V}^{n,2}), \quad (5.46)$$

$$\mathbf{V}^{n,1} = \mathbf{v}^n + \frac{h}{2} (K_{n+1/2} \mathbf{U}^{n,1} + S_{n+1/2} \mathbf{V}^{n,1}), \quad (5.47)$$

$$\mathbf{V}^{n,2} = \mathbf{v}^n + \frac{h}{2} (K_{n+1/2} \mathbf{U}^{n,1} + S_{n+1/2} \mathbf{V}^{n,1}). \quad (5.48)$$

2277 Here, $S_n = S(t_n)$, $S_{n+1/2} = S(t_n + 0.5h)$, etc. Because $S(t) \neq 0$, the scheme is block implicit. Note
2278 that $\mathbf{u}^{n+1} = \mathbf{U}^{n,2}$ and $\mathbf{V}^{n,1} = \mathbf{V}^{n,2} = \mathbf{v}(t_{n+1/2}) + \mathcal{O}(h^2)$.

2279 The numerical solution at the final time step provides a second order accurate approximation
2280 of the continuous solution operator matrix U_T , which we denote U_{Th} . It is used to approximate
2281 the matrix overlap function O_T in (5.4),

$$O_{Vh}(U_{Th}) = \sum_{j=0}^{E-1} (\langle \mathbf{u}_j^M, \mathbf{d}_j^u \rangle_2 + \langle \mathbf{v}_j^M, \mathbf{d}_j^v \rangle_2) + i \sum_{j=0}^{E-1} (\langle \mathbf{v}_j^M, \mathbf{d}_j^u \rangle_2 - \langle \mathbf{u}_j^M, \mathbf{d}_j^v \rangle_2), \quad (5.49)$$

2282 which is then used as the first part of the discrete objective function,

$$\mathcal{J}_1^h(U_{Th}) = \left(1 - \frac{1}{E^2} |O_{Vh}(U_{Th})|^2 \right). \quad (5.50)$$

The integral in the objective function (5.6) can be discretized to second order accuracy by using the Runge-Kutta stage variables,

$$\mathcal{J}_2^h(\mathbf{U}, \mathbf{V}) = \frac{h}{T} \sum_{j=0}^{E-1} \sum_{n=0}^{M-1} \left(\frac{1}{2} \langle \mathbf{U}_j^{n,1}, W \mathbf{U}_j^{n,1} \rangle_2 + \frac{1}{2} \langle \mathbf{U}_j^{n,2}, W \mathbf{U}_j^{n,2} \rangle_2 + \langle \mathbf{V}_j^{n,1}, W \mathbf{V}_j^{n,1} \rangle_2 \right). \quad (5.51)$$

2283 Based on the above formulas we discretize the objective function (5.7) according to

$$\mathcal{G}^h(\boldsymbol{\alpha}) = \mathcal{J}^h(U_{Th}^\alpha, \mathbf{U}^\alpha, \mathbf{V}^\alpha), \quad \mathcal{J}^h(U_{Th}, \mathbf{U}, \mathbf{V}) := \mathcal{J}_1^h(U_{Th}) + \mathcal{J}_2^h(\mathbf{U}, \mathbf{V}). \quad (5.52)$$

2284 Here, U_{Th}^α , \mathbf{U}^α and \mathbf{V}^α represent the time-discrete solution of the Störmer-Verlet scheme for a
2285 given parameter vector $\boldsymbol{\alpha}$. We note that $\mathcal{G}^h(\boldsymbol{\alpha})$ can be evaluated by accumulation during the
2286 time-stepping of the Störmer-Verlet scheme.

2287 **5.4.2 The discrete adjoint approach**

The gradient of the discretized objective function can be derived from first order optimality conditions of the corresponding discrete Lagrangian. In this approach, let $(\boldsymbol{\mu}_j^n, \boldsymbol{\nu}_j^n)$ be the adjoint variables and let $(\mathbf{M}_j^{n,i}, \mathbf{N}_j^{n,i})$ be Lagrange multipliers. We define the discrete Lagrangian by

$$\begin{aligned} \mathcal{L}^h(\mathbf{u}, \mathbf{v}, \mathbf{U}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\nu}, \mathbf{M}, \mathbf{N}, \boldsymbol{\alpha}) = \\ \mathcal{J}^h(U_{Th}, \mathbf{U}, \mathbf{V}) - \sum_{j=0}^{E-1} \left(\langle \mathbf{u}_j^0 - \mathbf{g}_j^u, \boldsymbol{\mu}_j^0 \rangle_2 + \langle \mathbf{v}_j^0 - \mathbf{g}_j^v, \boldsymbol{\nu}_j^0 \rangle_2 + \sum_{k=1}^6 T_j^k \right). \quad (5.53) \end{aligned}$$

The first two terms in the sum enforce the initial conditions (5.42). The terms T_j^1 and T_j^2 enforce the time-stepping update formulas (5.43)-(5.44) in the Störmer-Verlet scheme,

$$T_j^1 = \sum_{n=0}^{M-1} \left\langle \mathbf{u}_j^{n+1} - \mathbf{u}_j^n - \frac{h}{2} \left(S_n \mathbf{U}_j^{n,1} + S_{n+1} \mathbf{U}_j^{n,2} - K_n \mathbf{V}_j^{n,1} - K_{n+1} \mathbf{V}_j^{n,2} \right), \boldsymbol{\mu}_j^{n+1} \right\rangle_2, \quad (5.54)$$

$$T_j^2 = \sum_{n=0}^{M-1} \left\langle \mathbf{v}_j^{n+1} - \mathbf{v}_j^n - \frac{h}{2} \left(K_{n+1/2} (\mathbf{U}_j^{n,1} + \mathbf{U}_j^{n,2}) + S_{n+1/2} (\mathbf{V}_j^{n,1} + \mathbf{V}_j^{n,2}) \right), \boldsymbol{\nu}_j^{n+1} \right\rangle_2. \quad (5.55)$$

2288 The terms T_j^3 to T_j^6 enforce the relations between the stage variables (5.45)-(5.48) using the La-
2289 grange multipliers $(\mathbf{M}_j^{n,i}$ and $\mathbf{N}_j^{n,i})$, see Appendix .15 for details.

To derive the discrete adjoint scheme, we note that the discrete Lagrangian (5.53) has a saddle point if

$$\frac{\partial \mathcal{L}^h}{\partial \boldsymbol{\mu}_j^n} = \frac{\partial \mathcal{L}^h}{\partial \boldsymbol{\nu}_j^n} = \frac{\partial \mathcal{L}^h}{\partial \mathbf{N}_j^{n,i}} = \frac{\partial \mathcal{L}^h}{\partial \mathbf{M}_j^{n,i}} = 0, \quad (5.56)$$

$$\frac{\partial \mathcal{L}^h}{\partial \mathbf{u}_j^n} = \frac{\partial \mathcal{L}^h}{\partial \mathbf{v}_j^n} = \frac{\partial \mathcal{L}^h}{\partial \mathbf{U}_j^{n,i}} = \frac{\partial \mathcal{L}^h}{\partial \mathbf{V}_j^{n,i}} = 0, \quad (5.57)$$

2290 for $n = 0, 1, \dots, M$, $i = 1, 2$ and $j = 0, 1, \dots, E - 1$. Here, the set of conditions in (5.56) result in
2291 the Störmer-Verlet scheme (5.42)-(5.48) for evolving $(\mathbf{u}_j^n, \mathbf{v}_j^n, \mathbf{U}_j^{n,i}, \mathbf{V}_j^{n,i})$ forwards in time. The set
2292 of conditions in (5.57) result in a time-stepping scheme for evolving the adjoint variables $(\boldsymbol{\mu}_j^n, \boldsymbol{\nu}_j^n)$
2293 backwards in time, as is made precise in the following lemma.

Lemma 5.4.1. *Let \mathcal{L}^h be the discrete Lagrangian defined by (5.53). Furthermore, let $(\mathbf{u}_j^n, \mathbf{v}_j^n, \mathbf{U}_j^{n,i}, \mathbf{V}_j^{n,i})$ satisfy the Störmer-Verlet scheme (5.42)-(5.48) for a given parameter vector $\boldsymbol{\alpha}$. Then, the set of*

saddle-point conditions (5.57) are satisfied if the Lagrange multipliers $(\boldsymbol{\mu}_j^n, \boldsymbol{\nu}_j^n)$ are calculated according to the reversed time-stepping scheme,

$$\boldsymbol{\mu}_j^M = \frac{\partial \mathcal{J}^h}{\partial \mathbf{u}_j^M}, \quad \boldsymbol{\nu}_j^M = \frac{\partial \mathcal{J}^h}{\partial \mathbf{v}_j^M}, \quad (5.58)$$

$$\boldsymbol{\mu}_j^n = \boldsymbol{\mu}_j^{n+1} - \frac{h}{2} (\boldsymbol{\kappa}_j^{n,1} + \boldsymbol{\kappa}_j^{n,2}), \quad (5.59)$$

$$\boldsymbol{\nu}_j^n = \boldsymbol{\nu}_j^{n+1} - \frac{h}{2} (\boldsymbol{\ell}_j^{n,1} + \boldsymbol{\ell}_j^{n,2}), \quad (5.60)$$

for $n = M-1, M-2, \dots, 0$. Because $S^T = -S$ and $K^T = K$, the slopes satisfy

$$\boldsymbol{\kappa}_j^{n,1} = S_n \mathbf{X}_j^n - K_{n+1/2} \mathbf{Y}_j^{n,1} - \frac{2}{h} \frac{\partial \mathcal{J}^h}{\partial \mathbf{U}_j^{n,1}}, \quad (5.61)$$

$$\boldsymbol{\kappa}_j^{n,2} = S_{n+1} \mathbf{X}_j^n - K_{n+1/2} \mathbf{Y}_j^{n,2} - \frac{2}{h} \frac{\partial \mathcal{J}^h}{\partial \mathbf{U}_j^{n,2}}, \quad (5.62)$$

$$\boldsymbol{\ell}_j^{n,1} = K_n \mathbf{X}_j^n + S_{n+1/2} \mathbf{Y}_j^{n,1} - \frac{2}{h} \frac{\partial \mathcal{J}^h}{\partial \mathbf{V}_j^{n,1}}, \quad (5.63)$$

$$\boldsymbol{\ell}_j^{n,2} = K_{n+1} \mathbf{X}_j^n + S_{n+1/2} \mathbf{Y}_j^{n,2} - \frac{2}{h} \frac{\partial \mathcal{J}^h}{\partial \mathbf{V}_j^{n,2}}, \quad (5.64)$$

where the stage variables are given by

$$\mathbf{X}_j^n = \boldsymbol{\mu}_j^{n+1} - \frac{h}{2} \boldsymbol{\kappa}_j^{n,2}, \quad (5.65)$$

$$\mathbf{Y}_j^{n,2} = \boldsymbol{\nu}_j^{n+1}, \quad (5.66)$$

$$\mathbf{Y}_j^{n,1} = \boldsymbol{\nu}_j^{n+1} - \frac{h}{2} (\boldsymbol{\ell}_j^{n,1} + \boldsymbol{\ell}_j^{n,2}). \quad (5.67)$$

²²⁹⁴ *Proof.* The lemma follows after a somewhat tedious but straightforward calculation shown in detail

²²⁹⁵ in Appendix .15. \square

²²⁹⁶ Corresponding to the continuous Schrödinger equation (5.30), the adjoint state equation

²²⁹⁷ (without forcing) is

$$\begin{bmatrix} \dot{\boldsymbol{\mu}} \\ \dot{\boldsymbol{\nu}} \end{bmatrix} = \begin{bmatrix} S(t) & -K(t) \\ K(t) & S(t) \end{bmatrix} \begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\nu} \end{bmatrix} =: \begin{bmatrix} \mathbf{f}^\mu(\boldsymbol{\mu}, \boldsymbol{\nu}, t) \\ \mathbf{f}^\nu(\boldsymbol{\mu}, \boldsymbol{\nu}, t) \end{bmatrix}, \quad (5.68)$$

²²⁹⁸ where we used that $S^T = -S$ and $K^T = K$.

Corollary 5.4.1.1. *The time-stepping scheme (5.59)-(5.67) (without forcing) is a consistent approximation of the continuous adjoint state equation (5.68). It can be written as a modified partitioned Runge-Kutta method, where the Butcher coefficients are*

$$a_{11}^\mu = a_{21}^\mu = 1/2, \quad a_{12}^\mu = a_{22}^\mu = 0, \quad a_{11}^\nu = a_{12}^\nu = 0, \quad a_{21}^\nu = a_{22}^\nu = 1/2, \quad (5.69)$$

$$b_1^\mu = b_2^\mu = \frac{1}{2}, \quad b_1^\nu = b_2^\nu = \frac{1}{2}, \quad (5.70)$$

corresponding to the implicit midpoint rule for the μ -equation and the trapezoidal rule for the ν -equation in (5.68). The modifications to the partitioned Runge-Kutta scheme concerns the formulae for the slopes, (5.61)-(5.64). Because of the time-levels at which the matrices K and S are evaluated, it is **not** possible to define Butcher coefficients c_i^μ and c_i^ν such that

$$\kappa_j^{n,i} = f^\mu(X_j^{n,i}, Y_j^{n,i}, t_n + c_i^\mu h),$$

$$\ell_j^{n,i} = f^\nu(X_j^{n,i}, Y_j^{n,i}, t_n + c_i^\nu h).$$

2299 *Proof.* See Appendix .16. □

2300 Only the matrices K and S depend explicitly on α in the discrete Lagrangian. When the
 2301 saddle point conditions (5.56) and (5.57) are satisfied, we can therefore calculate the gradient of
 2302 \mathcal{G}^h by differentiating (5.53),

$$\frac{\partial \mathcal{G}^h}{\partial \alpha_r} = \frac{\partial \mathcal{L}^h}{\partial \alpha_r}, \quad r = 0, 1, \dots, E - 1.$$

2303 This relation leads to the following lemma.

2304 **Lemma 5.4.2.** *Let \mathcal{L}^h be the discrete Lagrangian defined by (5.53). Assume that $(\mathbf{u}_j^n, \mathbf{v}_j^n, \mathbf{U}_j^{n,i}, \mathbf{V}_j^{n,i})$
 2305 are calculated according to the Störmer-Verlet scheme for a given parameter vector α . Furthermore,
 2306 assume that $(\boldsymbol{\mu}_j^n, \boldsymbol{\nu}_j^n, \mathbf{X}_j^n, \mathbf{Y}_j^{n,i})$ satisfy the adjoint time-stepping scheme in Lemma 5.4.1, subject to
 2307 the terminal conditions*

$$\boldsymbol{\mu}_j^M = -\frac{2}{E^2} (Re(O_{Vh})\mathbf{d}_j^u - Im(O_{Vh})\mathbf{d}_j^v), \quad \boldsymbol{\nu}_j^M = -\frac{2}{E^2} (Re(O_{Vh})\mathbf{d}_j^v + Im(O_{Vh})\mathbf{d}_j^u),$$

and the forcing functions

$$\begin{aligned}\frac{\partial \mathcal{J}^h}{\partial \mathbf{U}_j^{n,1}} &= \frac{h}{T} W \mathbf{U}_j^{n,1}, & \frac{\partial \mathcal{J}^h}{\partial \mathbf{U}_j^{n,2}} &= \frac{h}{T} W \mathbf{U}_j^{n,2} \\ \frac{\partial \mathcal{J}^h}{\partial \mathbf{V}_j^{n,1}} &= \frac{h}{T} W \mathbf{V}_j^{n,1}, & \frac{\partial \mathcal{J}^h}{\partial \mathbf{V}_j^{n,2}} &= 0.\end{aligned}$$

Then, the saddle-point conditions (5.56) and (5.57) are satisfied and the gradient of the objective function (5.52) is given by

$$\begin{aligned}\frac{\partial \mathcal{G}^h}{\partial \alpha_r} &= \frac{h}{2} \sum_{j=0}^{E-1} \sum_{n=0}^{M-1} \left\{ \left\langle S'_n \mathbf{U}_j^{n,1} + S'_{n+1} \mathbf{U}_j^{n,2} - (K'_n + K'_{n+1}) \mathbf{V}_j^{n,1}, \mathbf{X}_j^n \right\rangle_2 \right. \\ &\quad \left. + \left\langle K'_{n+1/2} \mathbf{U}_j^{n,1} + S'_{n+1/2} \mathbf{V}_j^{n,1}, \mathbf{Y}_j^{n,1} \right\rangle_2 + \left\langle K'_{n+1/2} \mathbf{U}_j^{n,2} + S'_{n+1/2} \mathbf{V}_j^{n,2}, \mathbf{Y}_j^{n,2} \right\rangle_2 \right\}, \quad (5.71)\end{aligned}$$

where $S'_n = \partial S / \partial \alpha_r(t_n)$, $K'_{n+1/2} = \partial K / \partial \alpha_r(t_{n+1/2})$, etc.

Proof. See Appendix .17. \square

As a result of Lemma 5.4.2, all components of the gradient can be calculated from $(\mathbf{u}_j^n, \mathbf{v}_j^n, \mathbf{U}_j^{n,i}, \mathbf{V}_j^{n,1})$ and the adjoint variables $(\boldsymbol{\mu}_j^n, \boldsymbol{\nu}_j^n, \mathbf{X}_j^n, \mathbf{Y}_j^{n,i})$. The first set of variables are obtained from time-stepping the Störmer-Verlet scheme forward in time, while the second set of variables follow from time-stepping the adjoint scheme backward in time.

We can avoid storing the time-history of $(\mathbf{u}_j^n, \mathbf{v}_j^n, \mathbf{U}_j^{n,i}, \mathbf{V}_j^{n,1})$ by using the time-reversibility of the Störmer-Verlet scheme. However, in order to do so, we must first calculate the terminal conditions $(\mathbf{u}_j^M, \mathbf{v}_j^M)$ by evolving (5.42)-(5.48) forwards in time. The time-stepping can then be reversed and the gradient of the objective function (5.71) can be accumulated by simultaneously time-stepping the adjoint system (5.59)-(5.67) backwards in time.

5.5 Numerical optimization

Our numerical solution of the optimal control problem is based on the general purpose interior-point optimization package IPOPT [106]. This open-source library implements a primal-dual barrier approach for solving large-scale nonlinear programming problems, i.e., it minimizes an objective function subject to inequality (barrier) constraints on the parameter vector. Because

2324 the Hessian of the objective function is costly to calculate, we use the L-BFGS algorithm [91] in
 2325 IPOPT, which only relies on the objective function and its gradient to be evaluated. Inequality
 2326 constraints that limit the amplitude of the parameter vector α are enforced internally by IPOPT.

2327 The routines for evaluating the objective function and its gradient are implemented in the Ju-
 2328 lia programming language [24], which provides a convenient interface to IPOPT. Given a parameter
 2329 vector α , the routine for evaluating the objective function solves the Schrödinger equation with the
 2330 Störmer-Verlet scheme and evaluates $\mathcal{G}^h(\alpha)$ by accumulation. The routine for evaluating the gradi-
 2331 ent first applies the Störmer-Verlet scheme to calculate terminal conditions for the state variables.
 2332 It then proceeds by accumulating the gradient $\nabla_\alpha \mathcal{G}^h$ by simultaneous reversed time-stepping of the
 2333 discrete adjoint scheme and the Störmer-Verlet scheme. These two fundamental routines, together
 2334 with functions for setting up the Hamiltonians, estimating the time step, setting up constraints
 2335 on the parameter vector, post-processing and plotting of the results have been implemented in the
 2336 software package Juqbox, which was used to generate the numerical results below.

2337 The adjoint gradient implementation has been verified against a centered finite difference
 2338 approximation of the discrete objective function by perturbing each component of the parameter
 2339 vector. To further verify our implementation, we also calculated the discrete gradient by differen-
 2340 tiating the Störmer-Verlet scheme with respect to each component of the parameter vector. This
 2341 gradient agreed with the adjoint gradient to within 11-12 digits.

2342 5.5.1 A CNOT gate on a single qudit with guard levels

2343 To test our methods on a quantum optimal control problem, we consider realizing a CNOT
 2344 gate on a single qudit with four essential energy levels and two guard levels. The qudit is modeled
 2345 in the rotating frame of reference (with detuning frequencies $\Delta_1 = \Delta_2 = 0$) using the system and
 2346 control Hamiltonians (5.15) and (5.16), respectively. We set the fundamental frequency $\omega_1/2\pi =$
 2347 4.10336 GHz and self-Kerr coefficient $\xi_1/2\pi = 0.2198$ GHz. We parameterize the two control
 2348 functions using B-splines with carrier waves and choose the frequencies to be $\Omega_1 = 0$, $\Omega_2 = -\xi_1$
 2349 and $\Omega_3 = -2\xi_1$. In the rotating frame, these frequencies correspond to transitions between the

2350 ground state and the first excited state, the first and second excited states and the second and third
 2351 excited states. We discourage population of the fourth and fifth excited states using the weight
 2352 matrix $W = \text{diag}[0, 0, 0, 0, 0.1, 1.0]$ in \mathcal{J}_2^h , see (5.6). We use $D_1 = 10$ basis functions per frequency
 2353 and control function, resulting in a total of $D = 60$ parameters. The amplitudes of the control
 2354 functions are limited by the constraint

$$\|\boldsymbol{\alpha}\|_\infty := \max_{1 \leq r \leq D} |\alpha_r| \leq \alpha_{\max}. \quad (5.72)$$

2355 We set the gate duration to $T = 100$ ns and estimate the time step using the technique in Sec-
 2356 tion 5.3.2. To guarantee at least $C_P = 40$ time steps per period, we use $M = 8,796$ time steps,
 2357 corresponding to $h \approx 1.136 \cdot 10^{-2}$ ns.

2358 As initial guess for the elements of the parameter vector, we use a random number generator
 2359 with a uniform distribution in $[-0.01, 0.01]$. In Figure 5.2 we present the convergence history with
 2360 the two parameter thresholds $\alpha_{\max}/2\pi = 4$ MHz and 3 MHz, respectively.

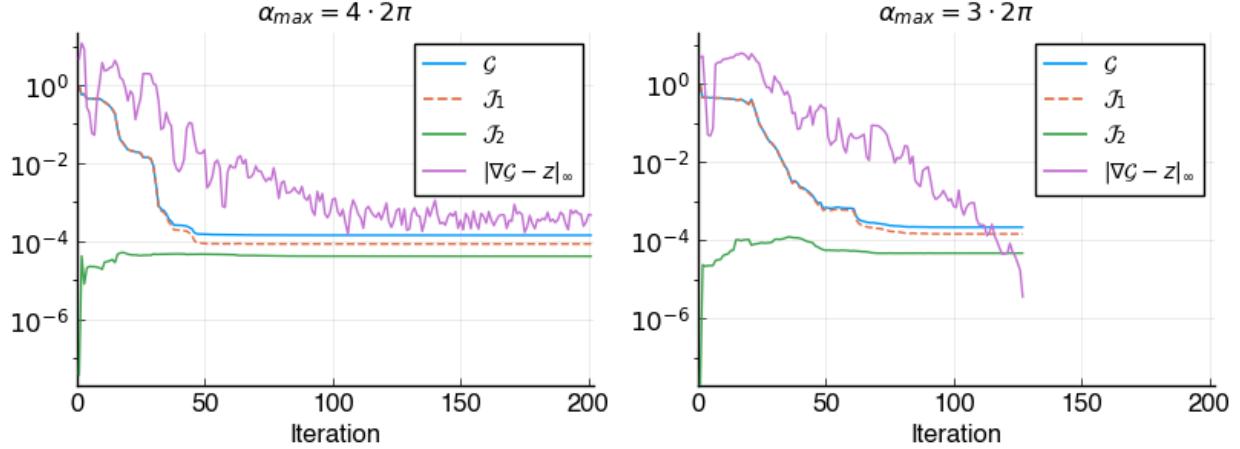


Figure 5.2: Convergence of the IPOPT iteration for the CNOT gate with the parameter constraint
 2361 $\|\boldsymbol{\alpha}\|_\infty \leq \alpha_{\max}$. Here, $\alpha_{\max}/2\pi = 4$ MHz (left) and $\alpha_{\max}/2\pi = 3$ MHz (right).

2361 We show the objective function \mathcal{G} , decomposed into \mathcal{J}_1^h and \mathcal{J}_2^h , together with the norm of the
 2362 dual infeasibility, $\|\nabla_{\boldsymbol{\alpha}} \mathcal{G} - z\|_\infty$, that IPOPT uses to monitor convergence, see [106] for details. For
 2363 the case with $\alpha_{\max}/2\pi = 3$ MHz, IPOPT converges well and needs 126 iteration to reduce the

2364 dual infeasibility to 10^{-5} , which was used as convergence criteria. However, when the parameter
2365 constraint is relaxed to $\alpha_{max}/2\pi = 4$ MHz, the convergence of IPOPT stalls after about 100
2366 iterations and is terminated after 200 iterations.

2367 For the converged solution with parameter constraint $\alpha_{max}/2\pi = 3$ MHz, the two parts of
2368 the objective function are $\mathcal{J}_1^h \approx 1.47 \cdot 10^{-4}$ and $\mathcal{J}_2^h \approx 4.72 \cdot 10^{-5}$, corresponding to a trace fidelity
2369 greater than 0.9998. The population of the guard states remains small for all times and initial
2370 conditions. In particular, the “forbidden” state $|5\rangle$ has a population that remains below $4.04 \cdot 10^{-7}$,
2371 see Figure 5.3.

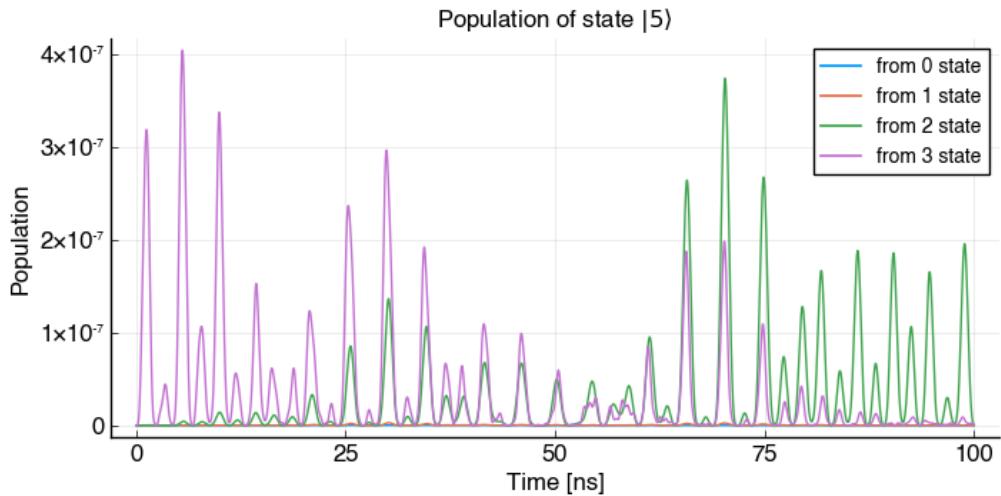


Figure 5.3: The population of the “forbidden” state $|5\rangle$ as function of time for the four initial conditions of the CNOT gate. Here, $\alpha_{max}/2\pi = 3$ MHz.

2372 The optimized control functions are shown in Figure 5.4 and the population of the essential
2373 states, corresponding to the four initial conditions of the CNOT gate, are presented in Figure 5.5.

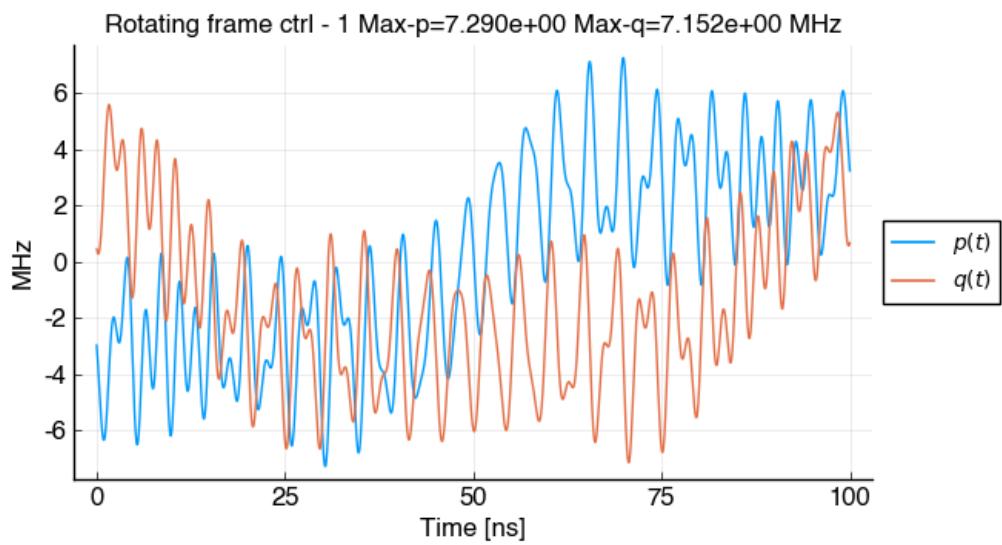


Figure 5.4: The rotating frame control functions $p(t)$ (blue) and $q(t)$ (orange) for realizing a CNOT gate with $D_1 = 10$ basis function per carrier wave and three carrier wave frequencies. Here, $\alpha_{max}/2\pi = 3$ MHz.

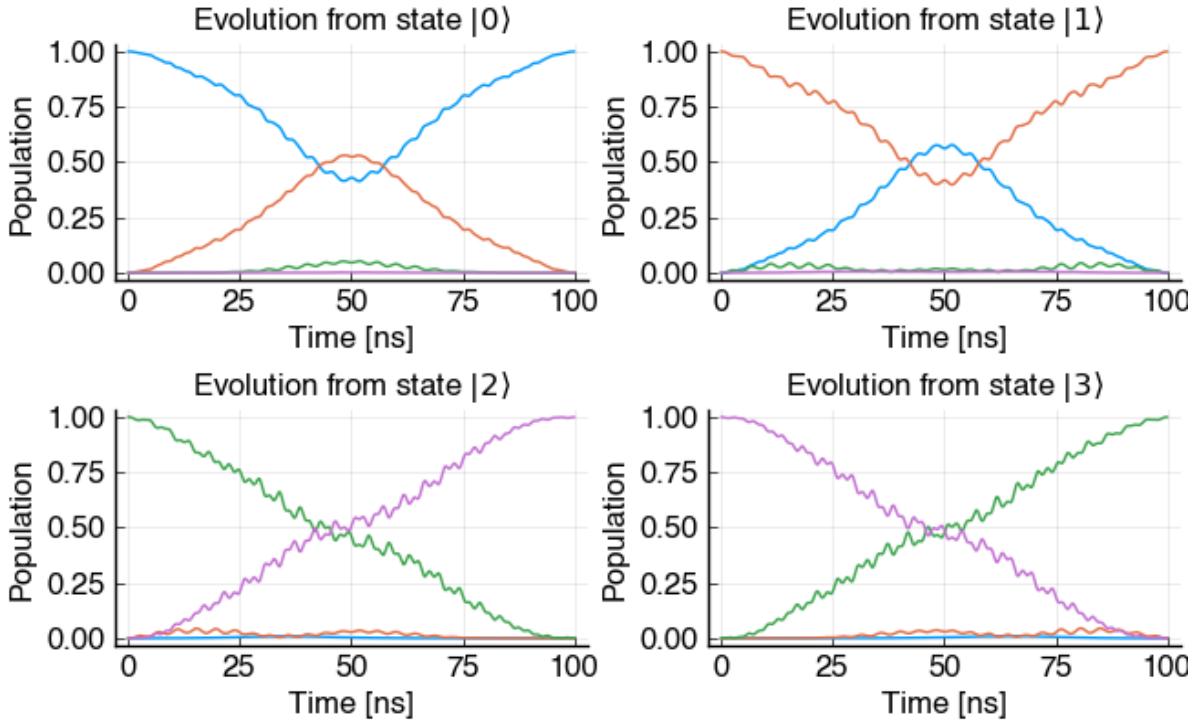


Figure 5.5: The population of the states $|0\rangle$ (blue), $|1\rangle$ (orange), $|2\rangle$ (green) and $|3\rangle$ (purple), as function of time, for each initial condition of the CNOT gate. Here, $\alpha_{max}/2\pi = 3$ MHz.

2374 Even though the dual infidelity does not reach the convergence criteria with the parameter
 2375 threshold $\alpha_{max}/2\pi = 4$ MHz, the resulting control functions give a very small objective function.
 2376 Here, $\mathcal{J}_1^h \approx 8.56 \cdot 10^{-5}$ and $\mathcal{J}_2^h \approx 4.15 \cdot 10^{-5}$, corresponding to a trace fidelity greater than 0.9999.
 2377 The population of the “forbidden” state $|5\rangle$ has a population that remains below $3.39 \cdot 10^{-7}$.

2378 **5.5.2 The Hessian of the objective function**

2379 The numerical results shown in Figure 5.2 illustrate that the convergence properties of the
 2380 optimization algorithm depend on the parameter constraints. To gain clarity into the local land-
 2381 scape of the optima we study the Hessian of the objective function. Let the optima correspond
 2382 to the parameter vector $\boldsymbol{\alpha}^*$. Based on the adjoint scheme for calculating the gradient, we can
 2383 approximate the elements of the Hessian matrix using a centered finite difference approximation,

$$\frac{\partial^2 \mathcal{G}^h(\boldsymbol{\alpha}^*)}{\partial \alpha_j \partial \alpha_k} \approx \frac{1}{2\varepsilon} \left(\frac{\partial \mathcal{G}^h}{\partial \alpha_j}(\boldsymbol{\alpha}^* + \varepsilon \mathbf{e}_k) - \frac{\partial \mathcal{G}^h}{\partial \alpha_j}(\boldsymbol{\alpha}^* - \varepsilon \mathbf{e}_k) \right) := L_{j,k}, \quad (5.73)$$

for $j, k = 1, 2, \dots, D$. To perform this calculation, the gradient must be evaluated for the $2D$ parameter vectors $(\boldsymbol{\alpha}^* \pm \varepsilon \mathbf{e}_k)$. Because the objective function and the parameter vector are real-valued, the gradient and the Hessian are also real-valued. Due to the finite difference approximation, the matrix L is only approximately equal to the Hessian. The accuracy in L is estimated in Table 5.1 by studying the norm of its asymmetric part, which is zero for the Hessian.

ε	$\ 0.5(L + L^T)\ _F$	$\ 0.5(L - L^T)\ _F$
10^{-4}	$4.95 \cdot 10^3$	$1.99 \cdot 10^{-4}$
10^{-5}	$4.95 \cdot 10^3$	$2.01 \cdot 10^{-6}$
10^{-6}	$4.95 \cdot 10^3$	$1.46 \cdot 10^{-6}$
10^{-7}	$4.95 \cdot 10^3$	$1.47 \cdot 10^{-5}$

Table 5.1: The Frobenius norm of the symmetric and asymmetric parts of the approximate Hessian, L , for the case $\alpha_{max}/2\pi = 3.0$ MHz.

Based on this experiment we infer that $\varepsilon = 10^{-6}$ is appropriate to use for approximating the Hessian in (5.73). To eliminate spurious effects from the asymmetry in the L matrix, we study the spectrum of its symmetric part, $L_s = 0.5(L + L^T)$. Because it is real and symmetric, it has a complete set of eigenvectors and all eigenvalues are real.

The eigenvalues of the Hessian are shown in Figure 5.6 for both values of the parameter threshold, α_{max} .

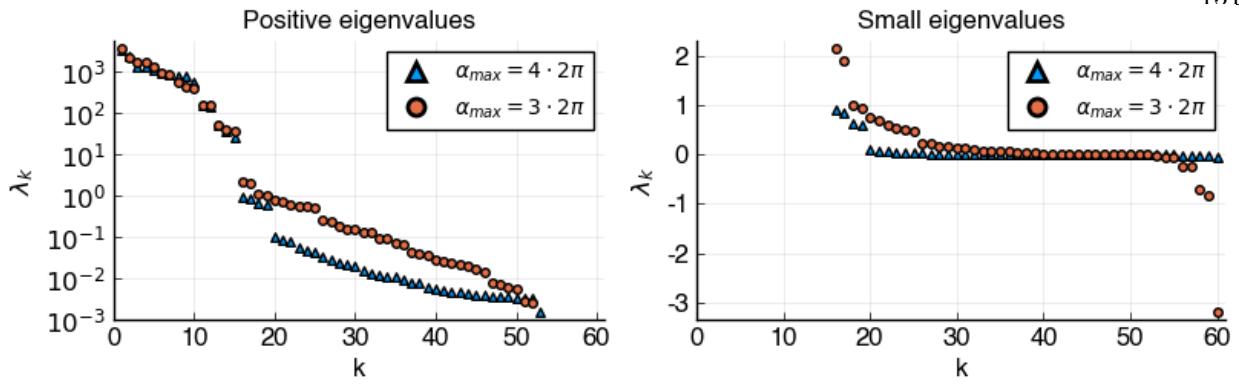


Figure 5.6: The eigenvalues of the symmetric part of the approximate Hessian, $0.5(L + L^T)$, evaluated at the optima for the parameter thresholds $\alpha_{max}/2\pi = 4$ MHz (blue triangles) and $\alpha_{max}/2\pi = 3$ MHz (orange circles). The positive eigenvalues are shown on a log-scale on the left and the small eigenvalues are shown on a linear scale on the right.

2395 Two properties of the spectra are noteworthy. First, a few eigenvalues are negative. This
 2396 may be an artifact related to the elements of the parameter vector that are close to their bounds.
 2397 As a result the landscape of the objective function may not be accurately represented by the
 2398 corresponding components of the Hessian. The second interesting property is that the 15 largest
 2399 eigenvalues are significantly larger than the rest. This indicates that the control functions are
 2400 essentially described by the 15 eigenvectors associated with those eigenvalues. As a result, the
 2401 objective function varies much faster in those directions than in the directions of the remaining 45
 2402 eigenvectors and this may hamper the convergence of the optimization algorithm in that subspace.
 2403 However, most of those 45 eigenvalues become larger when the parameter threshold is reduced from
 2404 $\alpha_{max}/2\pi = 4$ MHz to $\alpha_{max}/2\pi = 3$ MHz. This indicates that the constraints on the parameter
 2405 vector have a regularizing effect on the optimization problem and may explain why the latter case
 2406 converges better (see Figure 5.2).

2407 **5.5.3 Risk-neutral controls**

2408 In practice the entries of the Hamiltonian may have some uncertainty, especially for higher
 2409 energy levels, and it is desirable to design control pulses that are more robust to noise. There

2410 are several ways to design noise resilient controls, including robust optimization methods, in which
 2411 a min-max problem is solved [54], or risk-neutral/averse optimization approaches that minimize
 2412 the expectation of a utility function based on the original objective function subject to uncertain
 2413 parameters [55].

In this section we consider a risk-neutral strategy to design a $|0\rangle \leftrightarrow |2\rangle$ SWAP gate on a single qubit ($Q = 1$), with three essential levels and one guard level. Let $\epsilon \sim \text{Unif}(-\epsilon_{\max}, \epsilon_{\max})$ be a uniform random variable for some $\epsilon_{\max} > 0$. As a simple example, we consider the uncertain system Hamiltonian $H_s^u(\epsilon) = H_s^{rw} + H'(\epsilon)$ where H_s^{rw} is given by (5.15), and $H'(\epsilon)$ is a diagonal perturbation:

$$\frac{H'(\epsilon)}{2\pi} = \begin{pmatrix} 0 & & & \\ & \epsilon/100 & & \\ & & \epsilon/10 & \\ & & & \epsilon \end{pmatrix}.$$

Here, no perturbation is imposed on the control Hamiltonian (5.16). From these assumptions follow that the uncertain system Hamiltonian has expectation $\mathbb{E}[H_s^u(\epsilon)] = H_s^{rw}$. We may correspondingly update the original objective function, $\mathcal{G}(\boldsymbol{\alpha}, H_s^{rw})$, to the risk-neutral utility function $\tilde{\mathcal{G}}(\boldsymbol{\alpha}) = \mathbb{E}[\mathcal{G}(\boldsymbol{\alpha}, H_s^u(\epsilon))]$. Given the simple form of the random variable ϵ , we may compute $\tilde{\mathcal{G}}$ by quadrature:

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\alpha}, H_s^u(\epsilon))] = \int_{-\epsilon_{\max}}^{\epsilon_{\max}} \mathcal{G}(\boldsymbol{\alpha}, H_s^u(\epsilon)) d\epsilon \approx \sum_{k=1}^M w_k \mathcal{G}(\boldsymbol{\alpha}, H_s^u(\epsilon_k)), \quad (5.74)$$

2414 where w_k and ϵ_k are the weights and collocation points of a quadrature rule.

2415 For the following example, we compare the optimal control obtained using the standard
 2416 optimization procedure (no noise) and a risk-neutral control, in which the utility function (5.74)
 2417 is computed using the Gauss-Legendre quadrature with $N = 9$ collocation points and $\epsilon_{\max} = 10$
 2418 MHz. We set the gate duration to $T = 300$ ns, the maximum allowable amplitude to $\alpha_{\max}/2\pi = 12$
 2419 MHz, the fundamental frequency to $\omega_1/2\pi = 4.10336$ GHz, with detuning frequency $\Delta_1 = 0$, and
 2420 the self-Kerr coefficient to $\xi_1/2\pi = 0.2198$ GHz.

2421 The control functions are constructed using two carrier waves with frequencies $\Omega_{1,1} = 0$

and $\Omega_{1,2} = -\xi_1$ for both the “noise-free” (NF) and “risk-neutral” (RN) cases. In each case we use $D_1 = 12$ splines per control and carrier wave frequency for a total of $D = 48$ splines. We additionally constrain the controls to start and end at zero. We set the tolerance for L-BFGS to 10^{-5} , the maximum iteration count to 150, and use a maximum of five previous iterates to approximate the Hessian at each iteration. For the noise-free and risk-neutral optimized control functions, we use the perturbed Hamiltonian $H_s^u(\epsilon)$ to evaluate the objective function \mathcal{G} , for evenly spaced ϵ in the range $[-30, 30]$ MHz. The results are shown in Figure 5.7.

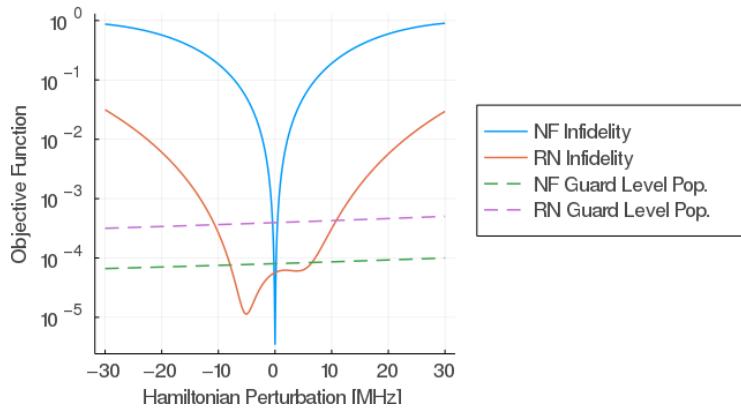


Figure 5.7: Infidelity objective (\mathcal{J}_1) and guard level objective (\mathcal{J}_2) as function of ϵ in $H_s^u(\epsilon)$. Here ‘NF’ and ‘RN’ correspond to the “Noise-Free” and “Risk-Neutral” cases.

From Figure 5.7 we note that the optimal control corresponding to the noise-free approach obtains the smallest infidelity for $\epsilon = 0$, but it grows rapidly for $|\epsilon| > 0$. By comparison, the optimal control found with the risk-neutral approach is much less sensitive to noise. We plot the control functions for both cases in Figure 5.8.

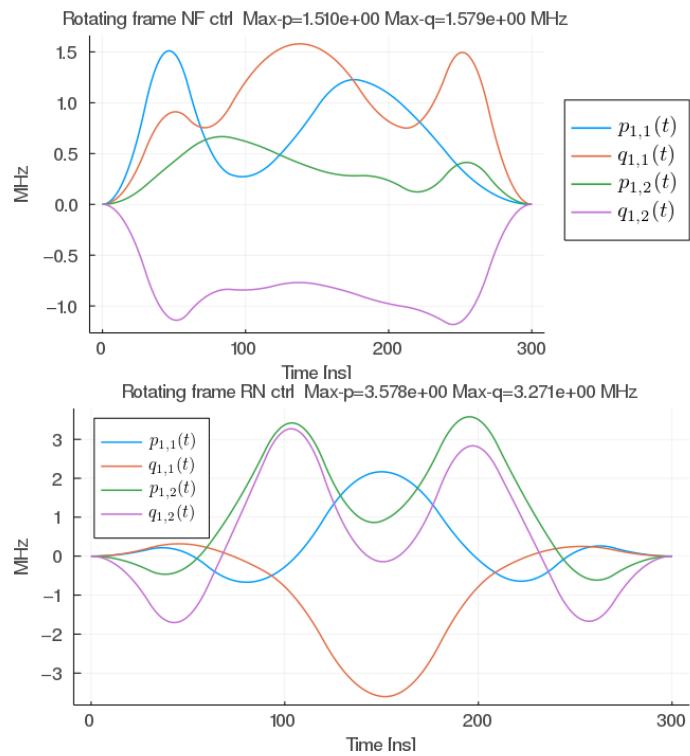


Figure 5.8: Control functions (without carrier waves) for the cases: “noise-free” (top), and “risk-neutral” (bottom). Here, $p_{k,n}(t)$ and $q_{k,n}(t)$ are defined in (5.27).

2433 Note that the risk-neutral controls (bottom panel) have larger amplitudes compared to the noise-
 2434 free controls (top panel), indicating a potential drawback of the risk-neutral approach. However, a
 2435 more systematic study of this issue is needed and left for future work.

2436 **5.6 Comparing Juqbox with QuTiP/pulse_optim and Grape-TF**

2437 The QuTiP/pulse_optim package is part of the QuTiP [71] framework and implements the
 2438 GRAPE algorithm in the Python language. The Grape-TF code (TF is short for TensorFlow [2]) is
 2439 also implemented in Python and provides an enhanced implementation of the GRAPE algorithm,
 2440 as described by Leung et al. [78]. It is callable from QuTiP and shares a similar problem setup
 2441 with the pulse_optim function.

To compare the Juqbox code with pulse_optim and Grape-TF, we consider a set of SWAP gates. These gates transform the ground state $|0\rangle$ to excited state $|d\rangle$, and vice versa. The trans-

formation can be described by the unitary matrix

$$V_g = \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & 0 \\ 1 & 0 & \cdots & 0 & 0 \end{bmatrix} \in \mathbb{C}^{(d+1) \times (d+1)}, \quad (5.75)$$

which involves $E = d + 1$ essential states. To evaluate how much leakage occurs to higher energy levels, we add one guard (forbidden) level ($G = 1$) and evolve a total of $N = d + 2$ states in Schrödinger's equation. As before, the guard level is left unspecified in the target gate transformation. We consider implementing the SWAP gates on a multi-level qudit that can be described by the fundamental frequency $\omega_1/2\pi = 4.8$ GHz and the self-Kerr coefficient $\xi_1/2\pi = 0.22$ GHz. We apply the rotating wave approximation, where the angular frequency of the rotation is ω_1 , resulting in the Hamiltonian model (5.15) and (5.16). As a realistic model for current superconducting quantum devices, we impose the control amplitude restrictions

$$\max_t |d(t; \boldsymbol{\alpha})| \leq c_\infty, \quad \frac{c_\infty}{2\pi} = 9 \text{ MHz}, \quad (5.76)$$

2442 in the rotating frame of reference.

2443 5.6.1 Setup of simulation codes

2444 QuTiP/pulse_optim can minimize the target gate fidelity, \mathcal{J}_1 , but does not suppress occupa-
2445 tion of higher energy states. Thus, it does **not** minimize terms of the type \mathcal{J}_2 . As a proxy for \mathcal{J}_2 ,
2446 we append one additional energy level to the simulation and measure its occupation as an estimate
2447 of leakage to higher energy states. In pulse_optim, the control functions are discretized on the same
2448 grid in time as Schrödinger's equation and no smoothness conditions are imposed. In our tests, we
2449 use a random initial guess for the parameter vector.

2450 Grape-TF discretizes the control functions on the same grid in time as Schrödinger's equation.
2451 It minimizes an objective function that consists of a number of user-configurable parts. In our test,

2452 we minimize the gate infidelity (\mathcal{J}_1) and the occupation of one guard (forbidden) energy level
 2453 (similar to \mathcal{J}_2). To smooth the control functions in time, the objective function also contains
 2454 additional terms to minimize their first and second time derivatives. The various parts of the
 2455 objective function are weighted together by user-specified coefficients. The gradient of the objective
 2456 function is calculated using the automatic differentiation (AD) technique, as implemented in the
 2457 TensorFlow package. In our tests, we use a random initial guess for the control vector.

In Juqbox, we trigger the first d transition frequencies in the system Hamiltonian by using d
 carrier waves in the control functions, with frequencies

$$\Omega_{1,k} = (k - 1)(-\xi_1), \quad k = 1, 2, \dots, N_f, \quad N_f = d.$$

2458 Similar to pulse_optim and Grape-TF, a pseudo-random number generator is used to construct the
 2459 initial guess for the parameter vector.

2460 The pulse_optim and Juqbox simulations were run on a Macbook Pro with a 2.6 GHz Intel
 2461 iCore-7 processor. To utilize the GPU acceleration in TensorFlow, the Grape-TF simulations were
 2462 run on one node of the Pascal machine at Livermore Computing, where each node has an Intel
 2463 XEON E5-2695 v4 processor with two NVIDIA P-100 GPUs.

2464 5.6.2 Numerical results

2465 A SWAP gate where the control functions meet the control amplitude bounds (5.76) can
 2466 only be realized if the gate duration is sufficiently long. Furthermore, the minimum gate duration
 2467 increases with d . For each value of d , we used numerical experiments to determine a duration
 2468 T_d such that at least two of the three simulation codes could find a solution with a small gate
 2469 infidelity. For Juqbox, we used the technique in Section 5.3.2 with $C_P = 80$ to obtain the number
 2470 of time steps. The number of control parameters follow from $D = 2N_f D_1$, where $N_f = d$ equals
 2471 the number of carrier wave frequencies and D_1 is the number of B-splines per control functions.
 2472 Here, $D_1 = 10$ for $d = 3, 4, 5$ and $D_1 = 20$ for $d = 6$. For pulse_optim and Grape-TF, we calculate
 2473 the number of time steps based on the shortest transition period, corresponding to the highest

transition frequency in the system. We then use 40 time steps per shortest transition period to resolve the control functions. For both GRAPE methods there are 2 control parameters per time step. The main simulation parameters are given in Table 5.2.

d	T_d [ns]	# time steps		# parameters	
		Juqbox	GRAPE	Juqbox	GRAPE
3	140	14,787	4,480	60	8,960
4	215	37,843	7,568	80	15,136
5	265	69,962	11,661	100	23,322
6	425	157,082	22,441	240	44,882

Table 5.2: Gate duration, number of time steps (M) and total number of control parameters (D) in the $|0\rangle \leftrightarrow |d\rangle$ SWAP gate simulations. The number of time steps and control parameters are the same for pulse_optim and Grape-TF.

2476

Optimization results for the pulse_optim, Grape-TF and Juqbox codes are presented in Tables 5.3, 5.4 and 5.5.

d	\mathcal{J}_1^*	$ \psi^{(d+1)} _\infty^2$	$ p _\infty$ [MHz]	$ q _\infty$ [MHz]	# iter	CPU [s]
3	4.35e-6	9.41e-3	9.00	9.00	38	30
4	3.91e-5	1.20e-2	9.00	9.00	93	108
5	1.57e-4	8.77e-3	9.00	9.00	215	385
6	1.76e-3	4.48e-2	9.00	9.00	246	894

Table 5.3: QuTiP/pulse_optim results for $|0\rangle \leftrightarrow |d\rangle$ SWAP gates. Note the larger infidelity and guard state population for $d = 6$.

d	\mathcal{J}_1	$ \psi^{(d+1)} _\infty^2$	$ p _\infty$ [MHz]	$ q _\infty$ [MHz]	# iter	CPU [s]
3	8.76e-6	4.03e-3	6.98	8.83	78	2,062
4	1.52e-5	3.39e-3	6.87	6.54	128	10,601
5	2.80e-5	1.78e-3	7.21	7.62	161	28,366
6	4.89e-1	2.33e-5	0.73	0.74	93	81,765

Table 5.4: Grape-TF results for $|0\rangle \leftrightarrow |d\rangle$ SWAP gates. Note the very large infidelity for $d = 6$. These simulations used two NVIDIA P-100 GPUs to accelerate TensorFlow.

d	\mathcal{J}_1	$ \psi^{(d+1)} _\infty^2$	$ p _\infty$ [MHz]	$ q _\infty$ [MHz]	# iter	CPU
3	2.71e-5	1.92e-3	7.59	8.99	177	55
4	4.91e-5	1.23e-3	7.78	5.33	166	151
5	4.95e-5	1.25e-3	7.42	7.24	173	291
6	7.41e-6	4.41e-3	4.55	5.39	229	1255

Table 5.5: Juqbox results for $|0\rangle \leftrightarrow |d\rangle$ SWAP gates.

2479 The pulse_optim code generates piecewise constant control functions that are very noisy and may
 2480 therefore be hard to realize experimentally. To obtain a realistic estimate of the resulting dynamics,
 2481 we interpolate the optimized control functions on a grid with 20 times smaller time step and use
 2482 the `mesolve()` function in QuTiP to calculate the evolution of the system from each initial state.
 2483 We then evaluate the gate infidelity using the evolved states at the final time, denoted by \mathcal{J}_1^* in
 2484 Table 5.3. Since the control functions from Grape-TF and Juqbox are significantly smoother, we
 2485 report the target gate fidelities as calculated by those codes. For the $|0\rangle \leftrightarrow |3\rangle$, $|0\rangle \leftrightarrow |4\rangle$ and
 2486 $|0\rangle \leftrightarrow |5\rangle$ SWAP gates, all three codes produce control functions with very small gate infidelities.
 2487 We note that the population of the guard level, $|\psi^{(d+1)}|^2$, is about an order of magnitude larger
 2488 with pulse_optim than with Juqbox; the guard level population from Grape-TF are somewhere in
 2489 between. The most significant difference between the results occur for the $d = 6$ SWAP gate. Here,
 2490 the Grape-TF code fails to produce a small gate infidelity after running for almost 23 hours and the
 2491 pulse_optim code results in a gate fidelity that is about 2 orders of magnitude larger than Juqbox.

2492 While pulse_optim and Juqbox require comparable amounts of CPU time to converge, the
 2493 Grape-TF code is between 50-100 times slower, despite the GPU acceleration.

We proceed by analyzing the optimized control functions and take the $|0\rangle \leftrightarrow |5\rangle$ SWAP gate as a representative example. In this case, the relevant transition frequencies in the laboratory frame of reference are

$$f_k = \frac{1}{2\pi} (\omega_1 - k\xi_1), \quad k = 0, 1, 2, 3, 4. \quad (5.77)$$

2494 To compare the smoothness of the optimized control functions, we study the Fourier spectra of the
 2495 laboratory frame control functions, see Figure 5.9.

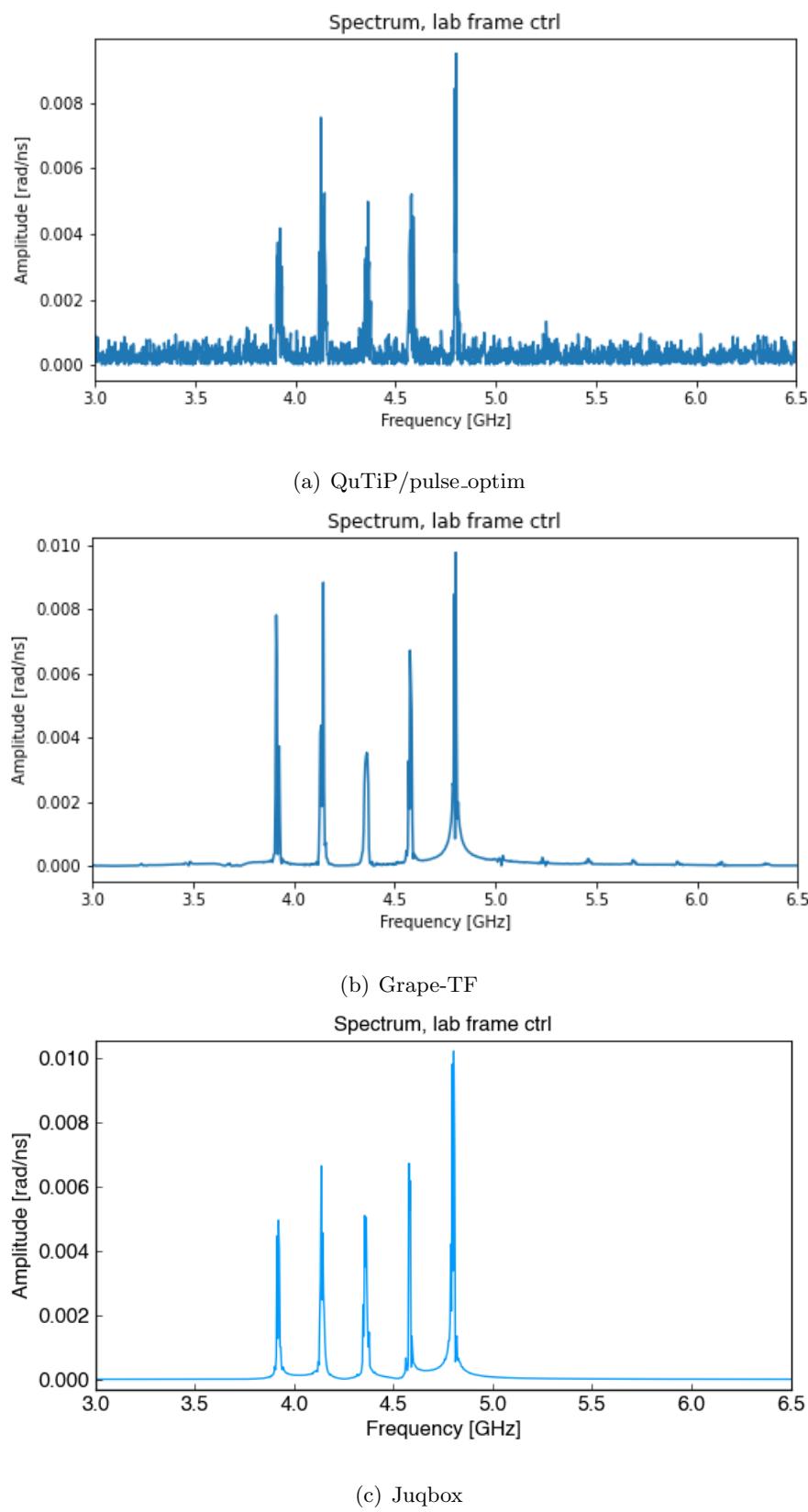


Figure 5.9: Magnitude of the Fourier spectrum of the laboratory frame control function for the $|0\rangle \leftrightarrow |5\rangle$ SWAP gate.

2496 We first note that pulse_optim produces a significantly noisier control function compared to the
 2497 other two codes. The control function from Grape-TF is significantly smoother, even though
 2498 its spectrum includes some noticeable peaks at frequencies that do not correspond to transition
 2499 frequencies in the system. The Juqbox simulation results in a laboratory frame control function
 2500 where each peak in the spectrum corresponds to a transition frequency in the Hamiltonian.

2501 **5.7 Conclusions**

2502 In this chapter we developed numerical methods for optimizing control functions for realizing
 2503 logical gates in closed quantum systems where the state is governed by Schrödinger's equation. By
 2504 asymptotic expansion, we calculated the resonant frequencies in the system Hamiltonian, corre-
 2505 sponding to transitions between energy levels in the state vector. We introduced a novel param-
 2506 eterization of the control functions using B-spline basis functions that act as envelopes for carrier
 2507 waves, with frequencies that match the transition frequencies. This approach allows the number of
 2508 control parameters to be independent of, and significantly smaller than, the number of time steps
 2509 for integrating Schrödinger's equation.

2510 The objective function in the optimal control problem consists of two parts: the infidelity
 2511 of the final gate transformation and a time-integral for evaluating leakage to higher energy levels.
 2512 We apply a “discretize-then-optimize” approach and outline the derivation of the discrete adjoint
 2513 equation that is solved to efficiently calculate the gradient of the objective function.

2514 To demonstrate our approach, we optimized the control functions for a CNOT gate with two
 2515 guard states, resulting in a gate fidelity exceeding 99.99%. Having a moderate number of control
 2516 parameters enabled us to study the spectrum of the Hessian of the objective function at an optima.
 2517 We found that imposing tighter bounds on the parameter vector results in a Hessian with larger
 2518 eigenvalues and thus improves the convergence of the optimization algorithm.

2519 Based on a simple noise model, we also generalized the proposed method to calculate risk-
 2520 neutral controls that are resilient to uncertainties in the Hamiltonian model. The results are
 2521 promising and indicate that a more systematic study of optimization under uncertainty can yield

2522 controls that are robust to noise in quantum systems. We finally compared the performance of
2523 the proposed method, implemented in the Juqbox package [53], and two implementations of the
2524 GRAPE algorithm: the pulse_optim method in QuTiP [71] and Grape-TensorFlow [78]. The codes
2525 were compared on a set of SWAP gates on a single qudit. Here, Juqbox was found to run 50-100
2526 times faster than Grape-TensorFlow and produce control functions that are significantly smoother
2527 than pulse_optim.

2528 In future work, we intend to generalize our approach to solve optimal control problems for
2529 larger quantum systems.

2530

Chapter 6

2531

IBM Open Science Prize – SWAP Gate Challenge

2532

On November 30, 2020, IBM announced an open science prize competition aimed at improving the fidelity of a SWAP gate between qubits 5 and 6 on their Quantum device named “Casablanca.” The details of the SWAP gate problem was outlined in a Jupyter notebook, which utilizes the IBM developed open source package Qiskit. The participants were allowed to work in teams with up to five people and restricted to only use open source software in solving the problem. The competition concluded on April 16, 2021, and on June 14, 2021, it was announced that there were no winners as no team achieved the desired 50% reduction in error of the SWAP gate.

2539

As a case study in the use of the optimal control techniques outlined in Chapter 5, in this (brief) chapter we outline an approach for the SWAP gate challenge. The approach in this chapter is based on the quantum optimal control techniques implemented in the open source packages Juqbox.jl (the methods for which were outlined in Chapter 5) and Quandary [64]. The results from the optimal control approach are only as good as the accuracy in the description of the quantum system dynamics, characterized by a Hamiltonian model that was provided by IBM. Based on the calibrated control pulses that IBM provide for their standard gate set, we describe a reverse engineering approach to calibrate our computational model, including effects of cross-talk. Techniques were developed to translate between Qiskit’s pulse representation and the B-spline formulation used in Juqbox and Quandary. The fidelity of the optimized pulse sequences were finally estimated using Qiskit’s randomized benchmarking (RB) techniques.

2550 **6.1 Hamiltonian model**

2551 The Hamiltonian model used in this study is based on work by Magesan and Gambetta [84].
 2552 In [84], the authors considered a system of two transmons coupled by a bus resonator. The bus
 2553 resonator is modeled as a harmonic oscillator with fundamental frequency ω^r and each transmon
 2554 is coupled to the bus resonator by a Jaynes-Cummings Hamiltonian with coupling strength g_j .
 2555 Let the $|01\rangle$ transition frequencies of the transmons be ω_j . It is assumed that the coupling is in
 2556 the dispersive regime, which means that resonator frequency is sufficiently detuned from the $|01\rangle$
 2557 transition frequencies to make $|\omega_j - \omega^r| \gg |g_j|$.

After transforming the Hamiltonian to a frame rotating with frequency ω^r in all three subsystems (two transmons and a bus resonator), the system and control Hamiltonians become (for notational convenience we set $\hbar = 1$)

$$H_{sys} = \sum_{j=1}^2 \left((\omega_j - \omega^r) b_j^\dagger b_j + \frac{\Delta_j}{2} b_j^\dagger b_j^\dagger b_j b_j \right) + g_j (b_j^\dagger c + b_j c^\dagger), \quad (6.1)$$

$$H_{ctrl}(t) = \sum_{j=1}^2 \text{Re}(e^{i\omega^r t} d_j(t)) (e^{-i\omega^r t} b_j + e^{i\omega^r t} b_j^\dagger), \quad (6.2)$$

2558 where b_j is the lowering operator for the j -th transmon and c is the lowering operator for the bus
 2559 resonator. The above model is an ideal starting point for an optimal control approach for designing
 2560 a SWAP gate. In practice, unfortunately, the resonator frequency ω^r and the coupling coefficients
 2561 g_j are not readily available from the IBM backend description of the Casablanca system.

Magesan and Gambetta [84] proceed by deriving a simplified Hamiltonian model using the following steps: 1) reorder the state vector into blocks of increasing transmon excitation number, 2) adiabatically eliminate the terms that couple the blocks, 3) project the Hamiltonian onto the zero-excitation subspace of the bus resonator. These steps results in an effective Hamiltonian for the two transmon system given in Equation (2.12) of [84]. This model is completely specified for the Casablanca system as all parameters can be accessed through the Qiskit interface. For these reasons, it is used as a starting point in our modeling effort. Based on this model, the lab frame

system Hamiltonian for qubits 5 and 6 of Casablanca is

$$H_{sys} = \sum_{j=5}^6 \left(\tilde{\omega}_j b_j^\dagger b_j + \frac{\Delta_j}{2} b_j^\dagger b_j^\dagger b_j b_j \right) + j_{56} \left(b_5^\dagger b_6 + b_5 b_6^\dagger \right). \quad (6.3)$$

Here $\tilde{\omega}_j$ is the dressed frequency of the j -th transmon, $b_5 = I \otimes a$, and $b_6 = a \otimes I$, where the lowering matrix for a single system is denoted by a . For simplicity, in the following we assume we use the dressed frequencies only and the tildes on the frequencies will be suppressed. Based on Equation (2.14) in [84] and the Hamiltonian entry in the backend of the Casablanca system, the lab frame control Hamiltonian is

$$H_{ctrl}(t) = \Omega_{d,5} \left(D_5(t) + U_{10}^{(5,6)}(t) \right) (b_5 + b_5^\dagger) + \Omega_{d,6} \left(D_6(t) + U_{11}^{(6,5)}(t) \right) (b_6 + b_6^\dagger), \quad (6.4)$$

where

$$D_5(t) = \text{Re} \left(e^{i\omega_5 t} d_5(t) \right), \quad U_{10}^{(5,6)}(t) = \text{Re} \left(e^{i\omega_6 t} u_{10}(t) \right), \quad (6.5)$$

$$D_6(t) = \text{Re} \left(e^{i\omega_6 t} d_6(t) \right), \quad U_{11}^{(5,6)}(t) = \text{Re} \left(e^{i\omega_5 t} u_{11}(t) \right). \quad (6.6)$$

For conciseness we have absorbed the phase factors $e^{i\phi}$ into the normalized control functions d_5 through u_{11} . Note that U_{10} is applied to qubit 5 but uses qubit 6's transition frequency. Correspondingly, U_{11} is applied to qubit 6, but uses qubit 5's transition frequency. Because $\text{Re}(z) = 0.5(z + \bar{z})$ for $z \in \mathbb{C}$, the control Hamiltonian can also be written

$$\begin{aligned} H_c(t) &= \frac{1}{2} \Omega_{d,5} \left(e^{i\omega_5 t} d_5(t) + e^{-i\omega_5 t} \bar{d}_5(t) + e^{i\omega_6 t} u_{10}(t) + e^{-i\omega_6 t} \bar{u}_{10}(t) \right) (b_5 + b_5^\dagger) \\ &\quad + \frac{1}{2} \Omega_{d,6} \left(e^{i\omega_6 t} d_6(t) + e^{-i\omega_6 t} \bar{d}_6(t) + e^{i\omega_5 t} u_{11}(t) + e^{-i\omega_5 t} \bar{u}_{11}(t) \right) (b_6 + b_6^\dagger). \end{aligned} \quad (6.7)$$

We apply a rotating frame transformation using the same frequency of rotation for both sub-systems (e.g. $\omega_{rot} = \omega_5$). In this frame, the system Hamiltonian becomes

$$\tilde{H}_{sys} = R H_{sys} R^\dagger + i R^\dagger \dot{R}, \quad (6.8)$$

$$= \sum_{j=5}^6 \left(\delta_j b_j^\dagger b_j + \frac{\Delta_j}{2} b_j^\dagger b_j^\dagger b_j b_j \right) + j_{56} \left(b_5^\dagger b_6 + b_5 b_6^\dagger \right), \quad (6.9)$$

where we have defined $\delta_j = \omega_j - \omega_{rot}$ for $j = 5, 6$. After applying the rotating wave approximation, the control Hamiltonian becomes

$$\tilde{H}_c(t) \approx \frac{\Omega_{d,5}}{2} \left(e^{i\delta_5 t} d_5(t) + e^{i\delta_6 t} u_{10}(t) \right) b_5 + \frac{\Omega_{d,6}}{2} \left(e^{i\delta_6 t} d_6(t) + e^{i\delta_5 t} u_{11}(t) \right) b_6 + \text{H.c.}, \quad (6.10)$$

2562 where H.c. stands for the Hermitian conjugate.

2563 6.2 Optimal control with Juqbox.jl and Quandary

We use numerical optimization in the open source packages Juqbox.jl [52] and Quandary [64] to determine the control functions $d_5(t)$, $d_6(t)$, $u_{10}(t)$, and $u_{11}(t)$ for realizing the SWAP gate transformation,

$$V_{SW} = \begin{bmatrix} 1 & & & \\ & 0 & 1 & \\ & 1 & 0 & \\ & & & 1 \end{bmatrix}. \quad (6.11)$$

Since higher energy levels play an important role in a cross-resonance gate, we model each transmon with 4 energy levels leading to a state vector with $N = 16$ elements. In the closed system setting, the time-evolution of the quantum system is unitary, yielding the transformation $\psi(t) = U(t, \alpha)\psi(0)$ for any initial quantum state $\psi(0)$. Here, $\alpha \in \mathbb{C}^D$ is the vector of control parameters and $U(t, \alpha) \in \mathcal{C}^{N \times N}$ is the unitary solution matrix, which solves Schrödinger's equation

$$\dot{U}(t, \alpha) = -i\tilde{H}(t, \alpha)U(t, \alpha) \quad 0 < t \leq T, \quad \text{with} \quad U(0) = I_N, \quad (6.12)$$

2564 where I_N is the $N \times N$ identity matrix and $\tilde{H}(t, \alpha) = \tilde{H}_{sys} + \tilde{H}_c(t, \alpha)$ denotes the Hamiltonian in 2565 the rotating frame.

The main target of the optimization is to find the vector of control parameters α that minimizes the difference between the target SWAP gate matrix, V_{SW} , and the final-time solution operator, $U(T, \alpha)$, projected onto the two lowest energy levels of each transmon. The difference

between the matrices is measured in terms of the trace infidelity

$$\mathcal{J}_1(\boldsymbol{\alpha}) = 1 - \frac{1}{E^2} \left| \text{Tr} \left(V_{SW}^\dagger \tilde{U}(T, \boldsymbol{\alpha}) \right) \right|^2, \quad \tilde{U} = PUP^\dagger. \quad (6.13)$$

2566 Here, $U(t, \boldsymbol{\alpha})$ solves Schrödinger's equation (6.12) and P extracts the first two energy levels of each
 2567 transmon from the full state vector. Thus, only $E = 4$ columns of the solution matrix U are used
 2568 for evaluating the trace infidelity. The time-averaged population of the highest energy level in each
 2569 transmon, $\mathcal{J}_2(\boldsymbol{\alpha})$, is used to discourage leakage to non-computational levels of the transmons. The
 2570 optimization then minimizes the total objective function $\mathcal{J}_1 + \mathcal{J}_2$, subject to constraints on the
 2571 amplitude of the control functions, as outlined in Chapter 5.

Both Juqbox and Quandary represent the control Hamiltonian in terms of their real and imaginary components,

$$\tilde{H}_{ctrl}(t) = p_5(t)(b_5 + b_5^\dagger) + iq_5(t)(b_5 - b_5^\dagger) + p_6(t)(b_6 + b_6^\dagger) + iq_6(t)(b_6 - b_6^\dagger). \quad (6.14)$$

To identify the relation between the real-valued control functions (p_k, q_k) and the complex-valued functions d_5, d_6, u_{10} and u_{11} in (6.10), it is convenient to first introduce the complex-valued functions $\zeta_k(t)$,

$$\zeta_k(t) = \sum_{\mathcal{L}=1}^{N_s} \alpha_{k,\mathcal{L}} \hat{B}_{\mathcal{L}}(t), \quad \alpha_{k,\mathcal{L}} \in \mathbb{C}. \quad (6.15)$$

In Juqbox and Quandary, $\hat{B}_{\mathcal{L}}(t)$, $\mathcal{L} = 1, 2, \dots, N_s$, are quadratic B-spline wavelets, uniformly spaced in time. We define,

$$p_5(t) + iq_5(t) := e^{i\delta_5 t} \zeta_5(t) + e^{i\delta_6 t} \zeta_{10}(t), \quad (6.16)$$

$$p_6(t) + iq_6(t) := e^{i\delta_6 t} \zeta_6(t) + e^{i\delta_5 t} \zeta_{11}(t). \quad (6.17)$$

2572 As there are four control functions, the total number of control parameters becomes $D = 4N_s$.
 2573 In the following, 30 B-spline coefficients were used to parameterize each control function. We
 2574 remark that the number of control parameters can be chosen independently of (and usually much
 2575 smaller than) the number of time steps for integrating Schrödinger's equation (In this study we
 2576 used $N_T = 105,625$ time steps to integrate Schrödinger's equation to time $T = 668.4$ ns).

The functions $\zeta_k(t)$ allow the control Hamiltonian (6.14) to be written as

$$\tilde{H}_{c,c}(t) = \left(e^{i\delta_5 t} \zeta_5(t) + e^{i\delta_6 t} \zeta_{10}(t) \right) b_5 + \left(e^{i\delta_6 t} \zeta_6(t) + e^{i\delta_5 t} \zeta_{11}(t) \right) b_6 + \text{H.c.} \quad (6.18)$$

The uncalibrated relation between the control functions in Qiskit and Juqbox is found by comparing (6.18) and (6.10),

$$d_5(t) = \frac{2}{\Omega_{d,5}} \zeta_5(t), \quad (6.19)$$

$$d_6(t) = \frac{2}{\Omega_{d,6}} \zeta_6(t), \quad (6.20)$$

$$u_{10}(t) = \frac{2}{\Omega_{d,5}} \zeta_{10}(t), \quad (6.21)$$

$$u_{11}(t) = \frac{2}{\Omega_{d,6}} \zeta_{11}(t). \quad (6.22)$$

2577 6.2.1 Open system optimal control

To account for system-environment interactions, the numerical optimization with Juqbox can be used as a starting point for optimal control with the Quandary code [64]. Quandary describes open quantum systems using a density matrix $\rho \in \mathbb{C}^{N \times N}$. The time evolution of $\rho(t)$ is modeled by Lindblad's master equation

$$\dot{\rho}(t) = -i(H(t)\rho(t) - \rho(t)H(t)) + L(\rho(t)). \quad (6.23)$$

Both decay and dephasing processes are modeled using the Lindblad terms:

$$L(\rho) = \sum_{k=5}^6 \sum_{l=1}^2 \mathcal{L}_{lk} \rho \mathcal{L}_{lk}^\dagger - \frac{1}{2} \left(\mathcal{L}_{lk}^\dagger \mathcal{L}_{lk} \rho + \rho \mathcal{L}_{lk}^\dagger \mathcal{L}_{lk} \right), \quad (6.24)$$

2578 where the collapse operators satisfy $\mathcal{L}_{1k} := \frac{1}{\sqrt{T_1^k}} a_k$ (decay) and $\mathcal{L}_{2k} := \frac{1}{\sqrt{T_2^k}} a_k^\dagger a_k$ (dephasing). Here,
2579 T_1^k and T_2^k correspond to the decay and dephasing times for system k .

2580 Quandary solves Lindblad's master equation numerically by applying the implicit midpoint
2581 time-stepping method, which is a symplectic, second-order time-integration scheme of Runge-Kutta
2582 type. In order to derive the discrete adjoint equations, techniques from Algorithmic Differentiation
2583 are applied to yield consistent and exact gradients of the objective function at costs that are inde-
2584 pendent of the number of control parameters. The optimization problem is then solved iteratively

2585 using gradient updates, preconditioned by the L-BFGS algorithm to incorporate Hessian information.
 2586 Constraints on the maximum amplitudes of the control parameters are incorporated using
 2587 a line-search procedure that projects the gradient onto the linear box-constraints of maximally
 2588 allowed control amplitudes.

2589 **6.3 Rabi pulse calibrations**

Qiskit support two types of channels that accept custom pulses: drive and control channels.

Pulses on the drive channels are specified in the following way:

$$D_i(t_j) = \operatorname{Re} \{ \exp(i2\pi f j dt + \phi) d_j \}. \quad (6.25)$$

Here f is a frequency that can be chosen by the user; it defaults to the qubit transition frequency for drive channels, and the frequency of the linked qubit for control channels. We additionally have that ϕ is a phase, the time step is set to $dt = 2/9$ ns for the Casablanca system, and d_j is the non-dimensional, complex-valued control amplitude at time $t_j = j dt$. For each channel, Qiskit specifies the maximum allowable amplitude signal, $\Omega_{d,i}$. Absorbing the phase into the dimensionless amplitude via $\tilde{d}_j = e^{i\phi} d_j$, then

$$\Omega_{d,i} D_i(t_j) = \Omega_{d,i} \left(\cos(2\pi f j dt) \operatorname{Re}\{\tilde{d}_j\} - \sin(2\pi f j dt) \operatorname{Im}\{\tilde{d}_j\} \right),$$

which gives the basic mapping between Qiskit and Juqbox:

$$p(t_j) = \frac{\operatorname{Re}\{\tilde{d}_j\}}{2\Omega_{d,i}}, \quad q(t_j) = \frac{\operatorname{Im}\{\tilde{d}_j\}}{2\Omega_{d,i}}. \quad (6.26)$$

To verify this mapping, we considered a Rabi pulse for a single qubit. With a single carrier-wave with zero frequency in the rotating frame and two energy levels, with constant $p = A_J, q = 0$, half a Rabi oscillation occurs in Juqbox simulations for the pulse duration

$$\tau_p = \frac{\pi}{|A_J|} \implies |A_J| = \frac{\pi}{\tau_p}.$$

Using Qiskit's amplitude convention, the corresponding relation is

$$A_Q = \operatorname{Re}\{\tilde{d}_j\} = \frac{2\pi}{\tau_p} \Omega_{d,i}.$$

The ratio between the amplitudes of the corresponding lab frame control signals in Qiskit and Juqbox becomes

$$\frac{|A_Q|}{2|A_J|} = \Omega_{d,i}.$$

2590 This implies that the amplitudes should be proportional to each other. Results from a Rabi
 2591 experiment on the Casablanca hardware are shown in Figure 6.1.

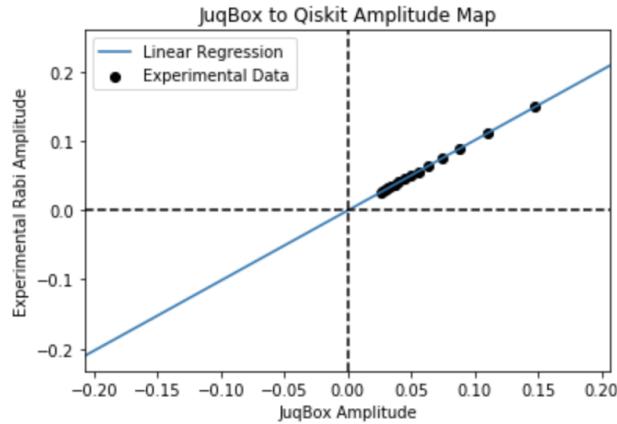


Figure 6.1: Results of the Rabi experiment on qubit 5 of the Casablanca hardware.

2592 For the Casablanca hardware, the fitted line $y = 1.014x - 1.9 \cdot 10^{-4}$ has a slope that is close
 2593 to $\Omega_{d,5} \approx 1.084$. This study confirms that the drive channel amplitude $\Omega_{d,5}$ is almost perfectly
 2594 calibrated.

2595 6.4 Gaussian square and DRAG pulses in Qiskit

We start by defining the zeroed Gaussian function, centered at time $T/2$:

$$g_z(t; A, T, \sigma) = \begin{cases} A \left(\exp\left(-\frac{(t-T/2)^2}{2\sigma^2}\right) - g_0(T, \sigma) \right), & t \in [0, T], \\ 0, & \text{otherwise.} \end{cases}$$

Here A is a complex-valued amplitude, T is the duration of the zeroed Gaussian, and σ is its standard deviation. Programmatically, however, the tails of the Gaussian are truncated by subtracting

out the constant g_0 , defined by

$$g_0(T, \sigma) = \exp\left(-\frac{(T/2 + dt)^2}{2\sigma^2}\right), \quad \text{where } dt = \frac{2}{9} \text{ ns.} \quad (6.27)$$

Based on the zeroed Gaussian pulse, we can now define the Derivative Removal by Adiabatic Gate (DRAG) pulse:

$$f(t; A, T, \sigma, \beta) = g_z(t; A, T, \sigma) + i\beta \underbrace{\left(-\frac{(t - T/2)}{\sigma^2}\right)}_{g'_z(t)} g(t; A, T, \sigma),$$

where β is a correction amplitude. Finally, the Gaussian square pulse is defined by its amplitude A , total duration T , and the duration of its constant part, w . Let the duration of the leading and trailing ramp be $r = (T - w)/2 > 0$. Then,

$$s(t; A, T, \sigma, w) = \begin{cases} \frac{g_z(t; A, 2r, \sigma)}{1 - g_0(2r, \sigma)}, & 0 \leq t \leq r, \\ A, & r \leq t \leq r + w, \\ \frac{g_z(T - t; A, 2r, \sigma)}{1 - g_0(2r, \sigma)}, & r + w \leq t \leq T. \end{cases}$$

2596 6.5 Converting Qiskit pulses to B-splines with carrier waves

For many basic gates, Qiskit provides parametric representations of the pulse schedule required to realize the chosen gate. These provided pulses take the form (6.25). For simplicity, suppose we have the signal D_0 and wish to represent it in Juqbox/Quandary via B-splines with carrier waves. Since each signal in Qiskit is associated with a single frequency we have

$$\Omega_{d,0} D_0(t_j) = \Omega_{d,0} \left(\cos(2\pi f_j dt) \operatorname{Re}\{\tilde{d}_j\} - \sin(2\pi f_j dt) \operatorname{Im}\{\tilde{d}_j\} \right) = 2p(t_j) \cos(\omega t_j) - 2q(t_j) \sin(\omega t_j).$$

The approximation of the real and imaginary parts of d_j can be done independently, so that in the following we focus on approximating the real part. This becomes a classical interpolation problem

$$\frac{\Omega_{d,0}}{2} \operatorname{Re}\{d_k\} = p(t_k) = \sum_{j=1}^{D_1} \alpha_j B_j(t_k), \quad \forall k = 1, 2, \dots, N,$$

where N is the number of samples d_k . This gives us the Vandermonde system

$$\underbrace{\begin{pmatrix} B_1(t_1) & B_2(t_1) & B_3(t_1) & \cdots & B_{D_1}(t_1) \\ B_1(t_2) & B_2(t_2) & B_3(t_2) & \cdots & B_{D_1}(t_2) \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ B_1(t_N) & B_2(t_N) & B_3(t_N) & \cdots & B_{D_1}(t_N) \end{pmatrix}}_{R^{N \times D_1}} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{D_1} \end{pmatrix} = \frac{\Omega_{d,0}}{2} \begin{pmatrix} \text{Re}\{d_1\} \\ \text{Re}\{d_2\} \\ \vdots \\ \text{Re}\{d_N\} \end{pmatrix},$$

where D_1 is the number of B-splines used in the approximation. Given that the time points t_k are distinct and $D_1 = N$ the above system is uniquely solvable. We also note that at a given time t_k only three B-splines are non-zero so that for a large number of splines, D_1 , the above system is sparse. Moreover, if we choose the interpolation points t_k to be the centers of each B-spline and pick $D_1 = N$ then the above is the tridiagonal system

$$\begin{pmatrix} 3/4 & 1/8 & & & \\ 1/8 & 3/4 & 1/8 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & 1/8 \\ & & & 1/8 & 3/4 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{D_1-1} \\ \alpha_{D_1} \end{pmatrix} = \frac{\Omega_{d,0}}{2} \begin{pmatrix} \text{Re}\{d_1\} \\ \text{Re}\{d_2\} \\ \vdots \\ \text{Re}\{d_{D_1-1}\} \\ \text{Re}\{d_{D_1}\} \end{pmatrix},$$

2597 which can be solved in $\mathcal{O}(D_1)$ time.

2598 6.6 Reverse model calibration using X- and Cx-gates

2599 The Casablanca system uses DRAG pulses to implement X-gates and a combination of Gaus-
2600 sian square and DRAG pulses to implement Cx gates. As the pulse coefficients are updated during
2601 each calibration of the system hardware, they can be used to calibrate the computational model
2602 for the Casablanca hardware. Given the high fidelity of the corresponding gates, the pulses can be
2603 used to engineer a mapping between the computational control functions we optimize with Juqbox
2604 / Quandary and the physical control functions that must be applied to the hardware.

2605 **6.6.1 X-gates**

The X-gate for qubit 5 is defined by the single DRAG pulse

$$d_5(t) = f(t; A_1, T, \sigma, \beta_1),$$

for $0 \leq t \leq 160 \cdot dt$ where $dt = 2/9$ ns. The X-gate for qubit 6 is also defined by the single DRAG pulse

$$d_6(t) = f(t; A_2, T, \sigma, \beta_2),$$

for $0 \leq t \leq 160 \cdot dt$. The parameters defining each DRAG pulse can be accessed using Qiskit by querying the backend for Casablanca. The coefficients change slightly after each system calibration, but the duration of the DRAG pulses appears to be fixed at $160 \cdot dt$. At some point during the spring of 2021, the coefficients were

$$A_1 = 0.17545065110530234,$$

$$A_2 = 0.20674287767710134,$$

$$T = 160,$$

$$\sigma = 40,$$

$$\beta_1 = 0.47609887200679674,$$

$$\beta_2 = 1.9314472856919194.$$

2606 The pulse schedules for the above X-gates are shown in Figure 6.2.

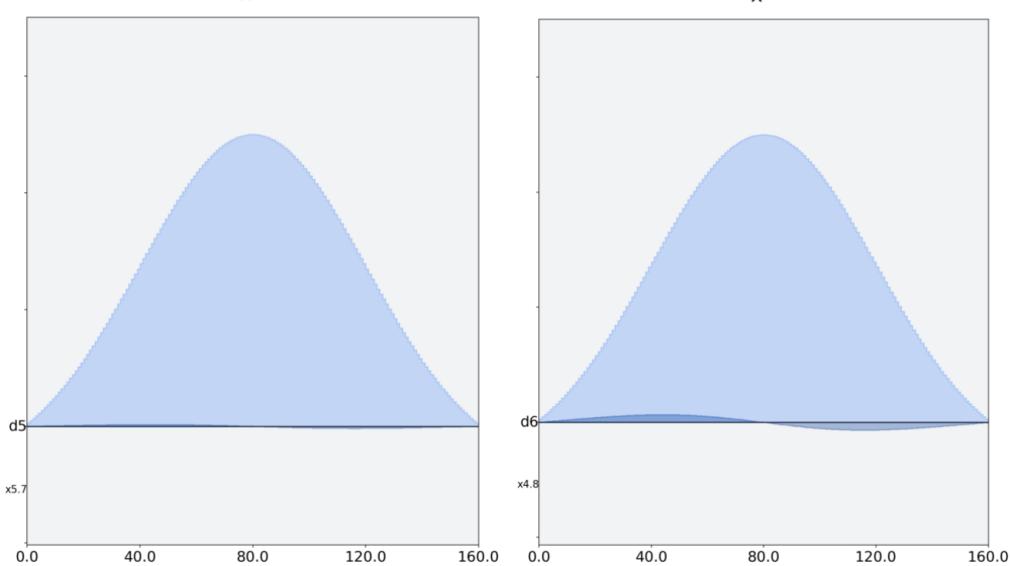


Figure 6.2: The pulse schedule for an X gate on Casablanca for qubit 5 (left), and qubit 6 (right).

Only the $d_5(t)$ function is active during the X5-gate and only the $d_6(t)$ function is active during the X6-gate, which allows the pulses to be calibrated independently. We want to improve the calculated average gate fidelity by modifying the control Hamiltonian (6.10) according to

$$d_5(t) \rightarrow \xi_5 d_5(t), \quad (6.28)$$

$$d_6(t) \rightarrow \xi_6 d_6(t), \quad (6.29)$$

where ξ_5 and ξ_6 are calibration factors. We were unable to achieve small trace gate infidelities based on (6.13). Upon closer examination, significant relative phase differences occurred between the target unitary and the simulated unitary evolution, which could not be explained by the rotating wave transformation. However, very good agreement in population was obtained. For this reason, we use the averaged gate fidelity, defined by

$$F_{avg} = \frac{1}{E} \sum_{j=1}^N \text{Tr}[V^\dagger \tilde{U}(T, \boldsymbol{\alpha})], \quad (6.30)$$

in this calibration. Here V is the target gate unitary and \tilde{U} is the projected solution matrix for Schrödinger's equation at final time T . The highest X-gate fidelities were obtained for the

coefficients

$$\xi_5 = 0.9927, \quad \xi_6 = 0.9909, \quad (6.31)$$

2607 indicating that the drive channels on the Casablanca system are very well calibrated. We remark
 2608 that $\xi_5 \Omega_{d,5} \approx 1.08$ is close to the slope 1.014 that was found in the Rabi calibration experiment in
 2609 Section 4.

2610 **6.6.2 Calibrating a cross-talk model using the Cx gates**

2611 The Casablanca backend holds calibrated pulse sequences for two CX-gates that involve
 2612 qubits 5 and 6: CX-56 (5 controls 6) and CX-65 (6 controls 5). These gates are implemented with
 2613 a combination of DRAG and Gaussian square pulses and use the control functions d_5 , d_6 and u_{11}
 2614 (but not u_{10}), see Figure 6.3.

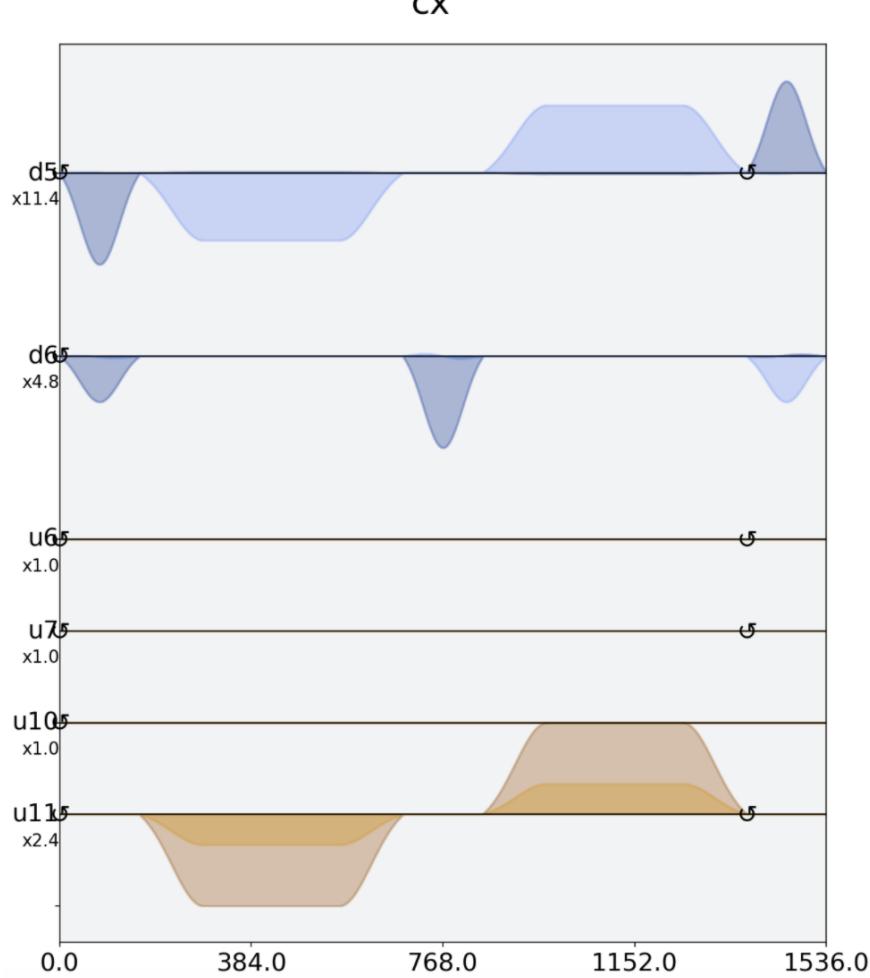


Figure 6.3: The pulse schedule for a CNOT gate on Casablanca where qubit 5 is the control qubit and qubit 6 is the target.

Since the control function u_{11} acts on qubit 6 but uses qubit 5's frequency, the corresponding signal may be subject to cross-talk. Following [84], we can account for this effect through the simple model

$$u_{11}(t)b_6 \rightarrow \xi_{11}u_{11}(t)b_6 + A_c e^{i\phi_c}u_{11}(t)b_5, \quad (6.32)$$

where ξ_{11} and A_c are cross-talk coefficients and ϕ_c is a phase shift that compensates for the physical distance between qubits 5 and 6 on the chip. After a parameter space sweep, we found that

$$\xi_{11} = 2.02, \quad A_c = 0.0583, \quad \phi_c = 1.2189, \quad (6.33)$$

²⁶¹⁵ result in gate fidelities of 0.994 for both the CX-56 gate and the CX-65 gate.

To complete our modeling we also need to handle potential cross-talk from control function u_{10} . In lieu of calibrated pulses, we use a symmetry argument to motivate the same cross-talk model for u_{10} as for u_{11} , i.e.,

$$u_{10}(t)b_5 \rightarrow \xi_{10}u_{10}(t)b_5 + A_c e^{i\phi_c} u_{10}(t)b_6, \quad \xi_{10} = \xi_{11}. \quad (6.34)$$

As a result of the above calibrations, the modified Hamiltonian model becomes $\tilde{H}_{comp}(t) = \tilde{H}_{sys} + \tilde{H}_{c,c}(t)$, where

$$\begin{aligned} \tilde{H}_{c,c}(t) &= \frac{\Omega_{d,5}}{2} \left(e^{i\delta_5 t} \xi_5 d_5(t) b_5 + e^{i\delta_6 t} u_{10}(t) (\xi_{10} b_5 + A_c e^{i\phi_c} b_6) \right) \\ &\quad + \frac{\Omega_{d,6}}{2} \left(e^{i\delta_6 t} \xi_6 d_6(t) b_6 + e^{i\delta_5 t} u_{11}(t) (\xi_{11} b_6 + A_c e^{i\phi_c} b_5) \right) + \text{H.c.} \end{aligned} \quad (6.35)$$

The Juqbox code represents the control Hamiltonian in terms of its real and imaginary components, as described above. By identifying the coefficients between (6.18) and (6.35), we arrive at the calibrated conversion

$$\zeta_5(t) = \frac{\Omega_{d,5}}{2} \xi_5 d_5(t) + \frac{\Omega_{d,6}}{2} A_c e^{i\phi_c} u_{11}(t), \quad (6.36)$$

$$\zeta_6(t) = \frac{\Omega_{d,6}}{2} \xi_6 d_6(t) + \frac{\Omega_{d,5}}{2} A_c e^{i\phi_c} u_{10}(t), \quad (6.37)$$

$$\zeta_{10}(t) = \frac{\Omega_{d,5}}{2} \xi_{10} u_{10}(t), \quad (6.38)$$

$$\zeta_{11}(t) = \frac{\Omega_{d,6}}{2} \xi_{11} u_{11}(t). \quad (6.39)$$

Thus, control functions $(\zeta_5, \zeta_6, \zeta_{10}, \zeta_{11})$ that are optimized with Juqbox should be converted to Qiskit according to

$$d_5(t) = \frac{2}{\Omega_{d,5} \xi_5} \left(\zeta_5(t) - A_c e^{i\phi_c} \frac{1}{\xi_{11}} \zeta_{11}(t) \right), \quad (6.40)$$

$$d_6(t) = \frac{2}{\Omega_{d,6} \xi_6} \left(\zeta_6(t) - A_c e^{i\phi_c} \frac{1}{\xi_{10}} \zeta_{10}(t) \right), \quad (6.41)$$

$$u_{10}(t) = \frac{2}{\Omega_{d,5} \xi_{10}} \zeta_{10}(t), \quad (6.42)$$

$$u_{11}(t) = \frac{2}{\Omega_{d,6} \xi_{11}} \zeta_{11}(t). \quad (6.43)$$

²⁶¹⁶ The control pulses used by Qiskit are defined by inserting the above functions into (6.5)-(6.6).

2617 **6.7 Implementation of custom pulses in Qiskit**

2618 In Juqbox/Quandary, each control signal is represented by a continuous approximation of
 2619 B-splines with carrier waves which we may simply evaluate at equispaced time points, $t_k = k dt$
 2620 where $dt = 2/9$ ns. As the Juqbox/Quandary samples have units of rad/ns, we use the mapping
 2621 (6.26) to obtain the dimensionless amplitude samples \tilde{d}_j that are required by Qiskit. Once we have
 2622 a set of dimensionless amplitude samples, we need to make Qiskit aware of our custom pulse. Our
 2623 chosen approach is to simply pass a non-basis element Clifford gate (such as the SWAP gate) to
 2624 `randomized_benchmarking_seq` as follows:

```
2625 1 # Use standard swap gate for interleaved circuit
2626 2 circ = QuantumCircuit(2)
2627 3 circ.swap(0,1)
2628 4 interleaved_elem = [circ]
2629 5
2630 6 # generate the RB circuit parameters
2631 7 length_vector = np.arange(1,200,20)
2632 8 nseeds = 5
2633 9
2634 0 rb_pattern = [[5,6]]
2635 1
2636 2 # Generate the RB circuits
2637 3 _,_cirlcs = randomized_benchmarking_seq(length_vector=length_vector,
2638 4                                     nseeds=nseeds,
2639 5                                     rb_pattern=rb_pattern,
2640 6                                     interleaved_elem=interleaved_elem)
```

2641 With the interleaved circuits built, we then manually add our custom pulse to each element as
 2642 follows:

```
2643 1 for circuits in cirlcs:
2644 2     for circuit in circuits:
2645 3         circuit.add_calibration("swap", qubits=[5, 6], schedule=sched)
```

2646 where `sched` is a `Schedule` object containing our custom pulse. This approach allows us to avoid
 2647 issues with `randomized_benchmarking_seq` decomposing a custom gate while providing a custom
 2648 pulse that the transpiler respects when executing the sequence on the actual device. When down-
 2649 loading the associated JSON files with each RB job on IBM Quantum using this approach, it can
 2650 be seen that each custom pulse is assigned to a 64 bit hexadecimal string indicating the signal is
 2651 being used on the actual device.

2652 A final practical matter with this approach is that directly adding the full custom waveform
 2653 as a single instruction when building a schedule can result in the error:

2654 `Waveform memory exceeds the maximum amount of memory currently available # [8018].`
 2655 To avoid this error, we have built the full schedule by splitting each custom waveform in smaller
 2656 chunks of (at most) 160 complex-valued amplitudes.

2657 6.8 Randomized benchmarking results

2658 To reduce influence of state preparation and measurement (SPAM) errors, the fidelity of the
 2659 SWAP gate is estimated using randomized benchmarking. This functionality is provided by the
 2660 routine `randomized_benchmarking_seq()` in Qiskit. This routine constructs two sets of random-
 2661 ized circuits for testing purposes. The first is composed of standard gates and is used as a reference.
 2662 In the second set, the circuits are augmented by one or more interleaved gates, which in this case
 2663 corresponds to a custom SWAP gate. As described above, the custom pulse schedule is explicitly
 2664 inserted by calling `add_calibration()`, before transpiling each interleaved circuit. An example of
 2665 a short interleaved pulse schedule is shown in Figure 6.4.

rb_interleaved_length_0_seed_1

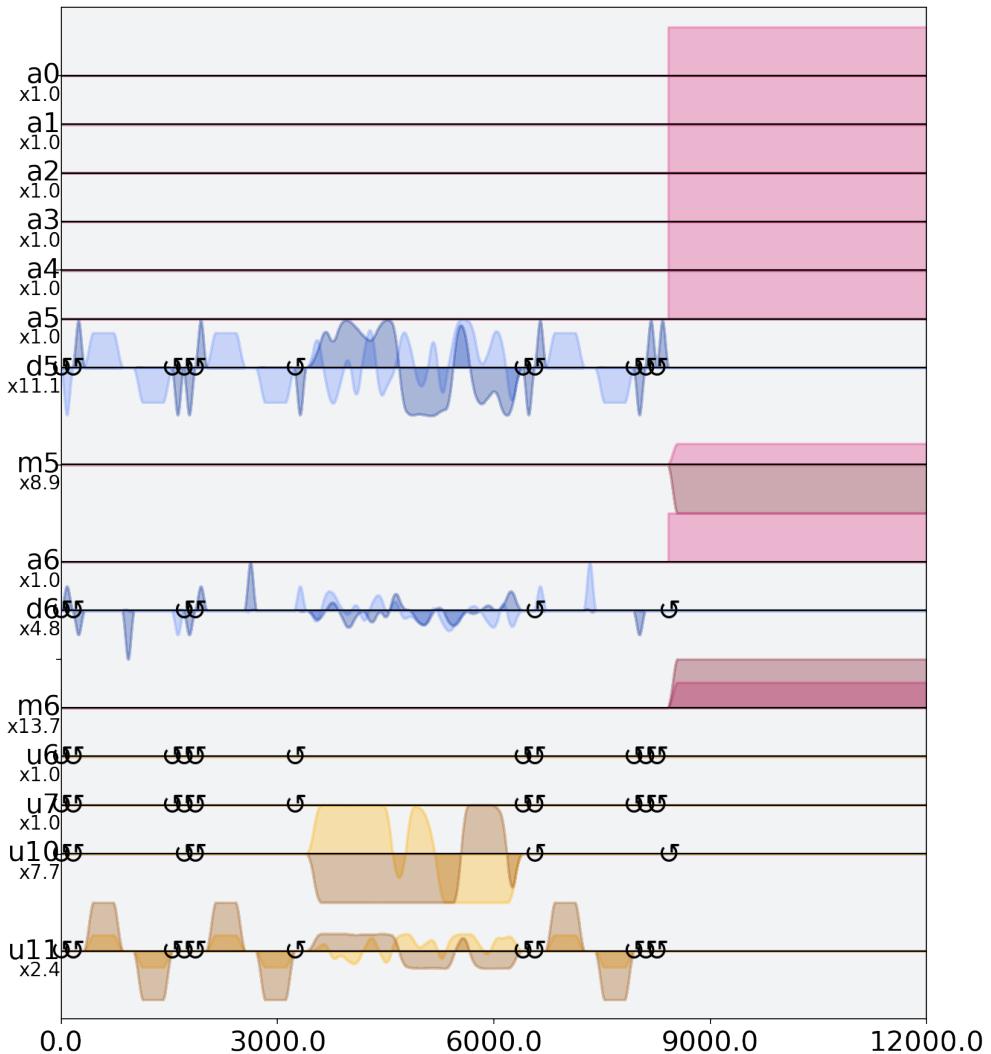


Figure 6.4: The pulse schedule for the first interleaved RB circuit.

2666 The results of the non-interleaved and interleaved randomized benchmarking are processed
 2667 by the Qiskit function `InterleavedRBfitter()`. It fits the data for each case to an exponential
 2668 function and estimates the SWAP gate error as the difference in exponential decay between the
 2669 fitted functions. With 5 randomized samples per case, 1000 shots per circuit, and considering
 2670 circuit lengths in the range of 1-181, results from the QASM simulator are shown in Figure 6.5.

epc_est: 0.023679604172967522 epc_err: 0.0014.

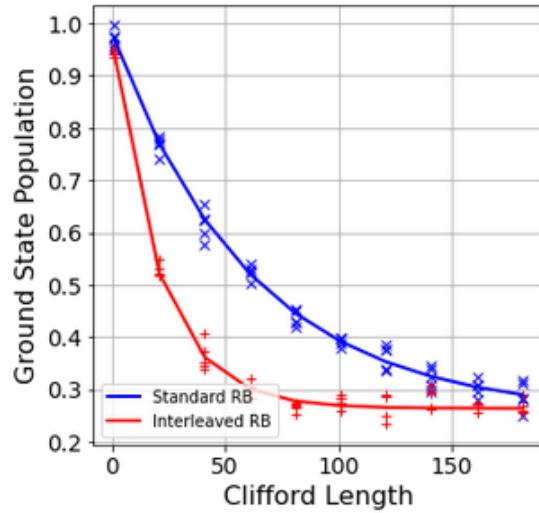


Figure 6.5: Simulated results. The observed ground state population as function of the length of the Clifford circuit. Here, the non-interleaved circuits are shown in blue and interleaved ones in red. The estimated error per Clifford (EPC) is 2.36%.

2671 In this case the estimated SWAP gate error was 2.36%.

2672 To test the custom SWAP gate pulse sequence on physical hardware, we consider a random-
 2673 ized benchmarking sequence consisting of 5 sets of interleaved and 5 sets of non-interleaved circuits.
 2674 For each circuit, we take the average of 1000 separate shots/experiments. By inspecting the result-
 2675 ing state output histograms for the shortest interleaved circuits, we observe that the custom gate
 2676 suffers from some severe accuracy problems, see Figure 6.6.

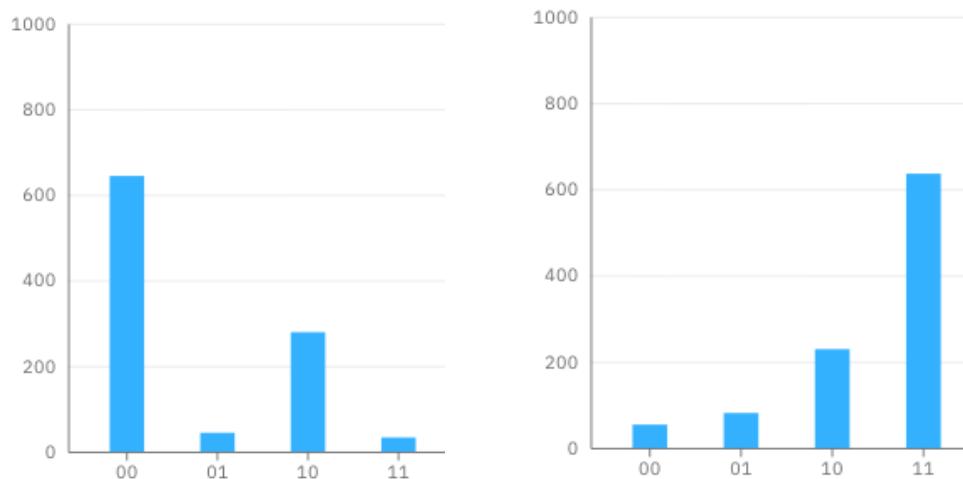


Figure 6.6: Classification results after 1000 shots in two of the randomized circuits with one interleaved SWAP gate. Ideally the $|00\rangle$ state should have 100% of the population.

2677 Due to the large spread in the ground state population, the `InterleavedRBfitter()` function
2678 had problems fitting the data to an exponential decay. The estimated error per Clifford (EPC) was
2679 $-1.32 \cdot 10^{-2}$, but with a very large standard deviation, see Figure 6.7.

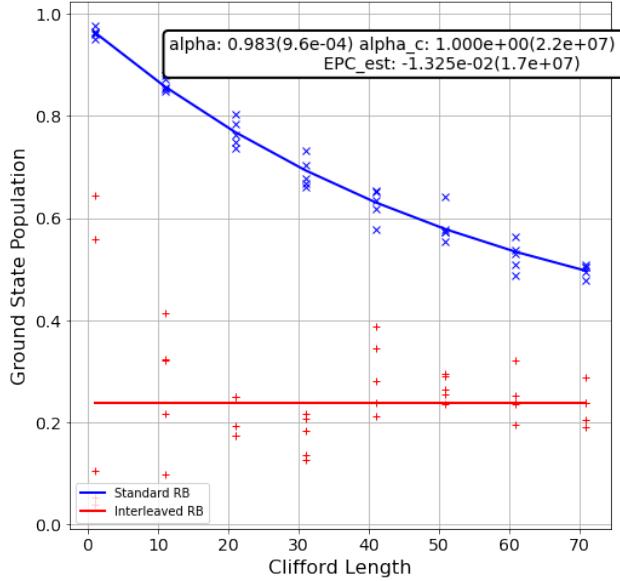


Figure 6.7: Randomized benchmarking results on the Casablanca hardware using the Interleave-dRBFitter.

2680 The large gate errors could well be due to some misinterpretation of the pulse schedules for
 2681 X and Cx gates. Another potential source of error is the Hamiltonian model.

2682 **6.9 Conclusion**

2683 As mentioned in the introduction of this chapter, no winners were selected for the IBM
 2684 SWAP gate challenge. It is clear that current NISQ-era machines are indeed noisy and difficult
 2685 to control in practice. This use case highlights many interesting areas of exploration to make
 2686 multi-qubit control possible in noisy-systems. Accurate system characterization, both of the bare
 2687 system Hamiltonian and for noise processes inherent to a specific system, are necessary for the
 2688 successful design of useful control signals. As seen in Chapter 5, even if armed with an (on average)
 2689 accurate Hamiltonian model it is possible for noise processes to severely degrade the performance
 2690 of predetermined optimal controls. It is clear that robust optimization techniques, or optimization
 2691 under uncertainty, could be leveraged to create control signals robust to noise.

2692 **Acknowledgment**

2693 This work was performed under the auspices of the U.S. Department of Energy by Lawrence
2694 Livermore National Laboratory under Contract DE-AC52-07NA27344. This report is contribution
2695 LLNL-TR-821599. We gratefully acknowledge financial support from Lawrence Livermore National
2696 Laboratory through the Laboratory Directed Research and Development (LDRD) program, grant
2697 # 20-ERD-028.

2698 **Disclaimer**

2699 This document was prepared as an account of work sponsored by an agency of the United
2700 States government. Neither the United States government nor Lawrence Livermore National Secu-
2701 rity, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any
2702 legal liability or responsibility for the accuracy, completeness, or usefulness of any information, ap-
2703 paratus, product, or process disclosed, or represents that its use would not infringe privately owned
2704 rights. Reference herein to any specific commercial product, process, or service by trade name,
2705 trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement,
2706 recommendation, or favoring by the United States government or Lawrence Livermore National Se-
2707 curity, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect
2708 those of the United States government or Lawrence Livermore National Security, LLC, and shall
2709 not be used for advertising or product endorsement purposes.

2710

Chapter 7

2711

Conclusion

2712 In this thesis, we investigated two distinct but related problems with regards to wave phe-
2713 nomena. We will now summarize the results presented and discuss further avenues of exploration
2714 for both topics.

2715 In Chapter 2 we presented and analyzed the WaveHoltz iteration, a new iterative method
2716 for solving the Helmholtz equation, for energy-conserving problems. We demonstrated that the
2717 iteration results in positive definite and sometimes symmetric matrices that are amenable to solution
2718 by iterative methods such as Krylov subspace methods. The numerical experiments indicated that
2719 the WaveHoltz iteration is a promising method with more favorable scaling for problems with
2720 outflow/impedance boundary conditions which are of much practical interest, e.g. for seismic
2721 applications.

2722 In Chapter 3 we extended the analysis of Chapter 2 to problems with damping and/or
2723 impedance conditions. We additionally showed that the WaveHoltz iteration converges to the
2724 discrete Helmholtz solution to an order matching the order of the timestepper for arbitrary order
2725 modified equation centered timestepping schemes. We then demonstrated that knowledge of how
2726 the timestepping modifies the discrete WaveHoltz iteration allows one to *completely* remove time
2727 discretization errors.

2728 In Chapter 4 we applied the WaveHoltz iteration to the “elastic” Helmholtz equation (also
2729 known as the Navier equation) for energy-conserving problems with Dirichlet and/or free surface
2730 boundary conditions. We additionally presented a second order implicit timestepping scheme with

2731 a modification to remove time discretization errors as was done for a family of explicit schemes
 2732 in Chapter 3. Numerical experiments indicate scaling similar to that of the acoustic Helmholtz
 2733 equation considered in Chapter 2.

2734 The WaveHoltz method has many avenues of exploration left. For the most part we have used
 2735 unconditioned Krylov solvers to accelerate the WaveHoltz iteration, but the spectral properties of
 2736 the WaveHoltz operator, $I - \mathcal{S}$, indicate that preconditioning should be possible. Given that the
 2737 spectral radius of \mathcal{S} is smaller than one for problems that are not in resonance, it may be possible
 2738 to construct polynomial preconditioners of $I - \mathcal{S}$ via a Neumann series or Padé approximations
 2739 of \mathcal{S} arising from slightly unstable WaveHoltz iterations with a small number of timesteps per
 2740 iteration. Further, we have not exploited adaptivity in space or time or any ideas from the sweeping
 2741 preconditioner class of methods.

2742 For the elastic WaveHoltz method, we have thus far only considered energy-conserving prob-
 2743 lems with Dirichlet and/or free surface boundary conditions. It is clear that further investigation
 2744 into problems with impedance/absorbing boundary conditions is needed. For DG discretizations,
 2745 the CFL condition of explicit timesteppers may be restrictive for high order spatial discretizations
 2746 on fine meshes. Improved methods for the inversion of the matrices for implicit time-corrected
 2747 schemes could drastically improve the performance and runtime of high order DG methods for
 2748 elastic problems.

2749 In this thesis, we presented some time-corrected centered schemes for the wave equation in
 2750 second order form. However, problems with damping or impedance/absorbing boundary conditions
 2751 require the use of the general WaveHoltz iteration in which the wave equation is solved as a first
 2752 order system in time. Another possible area of exploration is in devising modified timestepping
 2753 schemes for these systems. Related to this, we have not explored using spatial discretizations
 2754 specifically designed to reduce dispersion/pollution errors of the Helmholtz equation together with
 2755 the WaveHoltz iteration.

2756 Beginning in Chapter 5, we shifted focus from time-domain Helmholtz solvers to the optimal
 2757 control of quantum system. In this chapter, we considered closed quantum systems governed by

2758 the time-dependent Schrödinger equation where the controls are microwave pulses used to enact a
 2759 user-specified quantum logic gate. We defined a pair of objective functions measuring the infidelity
 2760 of the logic gate generated by a given control and a time-average of leakage into higher energy
 2761 level states of superconducting qubits to design high-fidelity gates. To decouple the timestep size
 2762 from the size of the parameter space, we introduced a novel basis of B-spline wavelets with carrier
 2763 waves designed specifically to drive transitions between energy levels in a quantum system. Using
 2764 a “discretize-then-optimize” approach, we devised a pair of partitioned Runge Kutta schemes to
 2765 compute *exact* discrete gradients. We demonstrated that this approach allows the construction of
 2766 high-fidelity gates using a small number of parameters for systems of superconducting qubits. We
 2767 additionally perform a brief study of risk-neutral controls to design controls that are more robust
 2768 to noise or uncertainty in the system Hamiltonian.

2769 In Chapter 6, we briefly outlined the approach for a submitted solution to the IBM SWAP
 2770 Gate Challenge as a practical application of the methods of Chapter 5. The goal of the IBM SWAP
 2771 Gate Challenge was to ask teams of researchers to attempt to reduce the errors of a standard
 2772 SWAP gate by 50% or more on IBM’s Casablanca system. Despite simulated results (both with
 2773 IBM’s simulator and the open-source quantum control toolbox Juqbox.jl) indicating the design of
 2774 a high-fidelity gate, the experimental results did not meet the desired target set by IBM.

2775 Despite the unsuccessful submission, Chapter 6 reveals many fruitful directions and research
 2776 questions to explore. One approach is in more thorough and advanced methods for system charac-
 2777 terization. As briefly elucidated in Chapter 5, controls that enact high-fidelity gates in a noise-free
 2778 optimization may quickly degrade in performance with the presence of unaccounted noise in the
 2779 Hamiltonian. Thus part of the system characterization procedure could benefit from characteri-
 2780 zation not only of a base Hamiltonian model, but also quantification and identification of noise
 2781 processes and their distributions. Armed with more accurate Hamiltonians, it would be advan-
 2782 tageous to explore robust and risk-neutral/averse optimization methods to extend the ideas of
 2783 Chapter 6. This would ultimately result in controls that need to (1) be calibrated less often, and
 2784 (2) are more resistant to noise.

Bibliography

- 2786 [1] Julia documentation. <https://docs.julialang.org/en/v1/>. Accessed: 2019-04-09.
- 2787 [2] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean,
2788 Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A
2789 system for large-scale machine learning. In 12th USENIX Symposium on Operating Systems
2790 Design and Implementation (OSDI 16), pages 265–283, 2016.
- 2791 [3] Tuomas Airaksinen, Anssi Pennanen, and Jari Toivanen. A damping preconditioner for time-
2792 harmonic wave equations in fluid and elastic material. Journal of computational physics,
2793 228(5):1466–1479, 2009.
- 2794 [4] R. Anderson, J. Andrej, A. Barker, J. Bramwell, J.-S. Camier, J. Cerveny V. Dobrev, Y. Du-
2795 douit, A. Fisher, Tz. Kolev, W. Pazner, M. Stowell, V. Tomov, I. Akkerman, J. Dahm,
2796 D. Medina, and S. Zampini. MFEM: A modular finite element library. Computers &
2797 Mathematics with Applications, 2020.
- 2798 [5] Laurent Anné, Patrick Joly, and Quang Huy Tran. Construction and analysis of higher order
2799 finite difference schemes for the 1D wave equation. Computational Geosciences, 4(3):207–249,
2800 2000.
- 2801 [6] D. Appelö. Absorbing Layers and Non-Reflecting Boundary Conditions for Wave Propagation
2802 Problems. PhD thesis, Royal Institute of Technology, October 2005.
- 2803 [7] D. Appelö, J. W. Banks, W. D. Henshaw, and D. W. Schwendeman. Numerical methods for
2804 solid mechanics on overlapping grids: Linear elasticity. Journal of Computational Physics,
2805 231(18):6012–6050, 2012.
- 2806 [8] D. Appelö and T. Colonius. A high-order super-grid-scale absorbing layer and its application
2807 to linear hyperbolic systems. Journal of Computational Physics, 228(11):4200–4217, 2009.
- 2808 [9] D. Appelö and T. Hagstrom. A new discontinuous Galerkin formulation for wave equations
2809 in second order form. SIAM Journal On Numerical Analysis, 53(6):2705–2726, 2015.
- 2810 [10] D. Appelö and T. Hagstrom. An energy-based discontinuous Galerkin discretization of the
2811 elastic wave equation in second order form. Comput. Meth. Appl. Mech. Engrg., 338:362–391,
2812 2018.
- 2813 [11] D. Appelö and G. Kreiss. A new absorbing layer for elastic waves. Journal of Computational
2814 Physics, 215(2):642–660, 2006.

- [12] D. Appelö and N. A. Petersson. A stable finite difference method for the elastic wave equation on complex geometries with free surfaces. *Communications in Computational Physics*, 5(1):84–107, 2009.
- [13] Daniel Appelö, Fortino Garcia, Allen Alvarez Loya, and Olof Runborg. El waveholtz method. *in preparation*.
- [14] Daniel Appelö, Fortino Garcia, and Olof Runborg. WaveHoltz: Iterative solution of the Helmholtz equation via the wave equation. *SIAM Journal on Scientific Computing*, 42(4):A1950–A1983, 2020.
- [15] Anton Arnold, Sjoerd Geevers, Ilaria Perugia, and Dmitry Ponomarev. An adaptive finite element method for high-frequency scattering problems with variable coefficients, 2021.
- [16] I. Babuška and S. Sauter. Is the pollution effect of the fem avoidable for the Helmholtz equation considering high wave numbers? *SIAM Journal on Numerical Analysis*, 34(6):2392–2423, 1997.
- [17] A. H. Baker, Tz. V. Kolev, and U. M. Yang. Improving algebraic multigrid interpolation operators for linear elasticity problems. *Numerical Linear Algebra with Applications*, 17(2-3):495–517, 2010.
- [18] Claude Bardos and Jeffrey Rauch. Variational algorithms for the Helmholtz equation using time evolution and artificial boundaries. *Asymptotic analysis*, 9(2):101–117, 1994.
- [19] Manuel M Baumann. *Fast Iterative Solution of the Time-Harmonic Elastic Wave Equation at Multiple Frequencies*. PhD thesis, PhD thesis, Delft University of Technology, 2018.
- [20] Alvin Bayliss, Charles I Goldstein, and Eli Turkel. An iterative method for the Helmholtz equation. *Journal of Computational Physics*, 49(3):443–457, 1983.
- [21] E. Bécache, S. Fauqueux, and P. Joly. Stability of perfectly matched layers, group velocities and anisotropic waves. *J. Comput. Phys.*, 188:399–433, 2003.
- [22] Mikhail Belonosov, Victor Kostin, Dmitry Neklyudov, and Vladimir Tcheverda. 3D numerical simulation of elastic waves with a frequency-domain iterative solver. *Geophysics*, 83(6):T333–T344, 10 2018.
- [23] Hadrien Bériot, Albert Prinn, and Gwénaël Gabard. Efficient implementation of high-order finite elements for helmholtz problems. *International Journal for Numerical Methods in Engineering*, 106(3):213–240, 2016.
- [24] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah. Julia: A fresh approach to numerical computing. *SIAM Review*, 59(1):65–89, 2017.
- [25] A. Björck. *Numerical methods in matrix computations*, volume 59. Springer, 2015.
- [26] Alexandre Blais, Arne L. Grimsmo, S. M. Girvin, and Andreas Wallraff. Circuit quantum electrodynamics. *Rev. Mod. Phys.*, 93:025005, May 2021.
- [27] A. Borzì, G. Ciarmella, and M. Sprengel. *Formulation and Numerical Solution of Quantum Control Problems*. Computational science and engineering. SIAM, 2017.

- [2852] [28] A. Brandt and I. Livshits. Wave-ray multigrid method for standing wave equations. *Electron. Trans. Numer. Anal.*, 6(162-181):91, 1997.
- [2853]
- [2854] [29] M.O. Bristeau, R. Glowinski, and J. P\'eriaux. Controllability methods for the computation of time-periodic solutions; application to scattering. *Journal of Computational Physics*, 147(2):265–292, 1998.
- [2855]
- [2856]
- [2857] [30] Romain Brunet, Victorita Dolean, and Martin J. Gander. Can classical Schwarz methods for time-harmonic elastic waves converge? In *Domain Decomposition Methods in Science and Engineering XXV*, pages 425–432, Cham, 2020. Springer International Publishing.
- [2858]
- [2859]
- [2860] [31] Romain Brunet, Victorita Dolean, and Martin J Gander. Natural domain decomposition algorithms for the solution of time-harmonic elastic waves. *SIAM Journal on Scientific Computing*, 42(5):A3313–A3339, 2020.
- [2861]
- [2862]
- [2863] [32] T. Caneva, T. Calarco, and S. Montangero. Chopped random-basis quantum optimization. *Physical Review A*, 84(2), Aug 2011.
- [2864]
- [2865] [33] T. Chaumont-Frelet, M. J. Grote, S. Lanteri, and J. H. Tang. A controllability method for Maxwell’s equations, 2021.
- [2866]
- [2867] [34] Z. Chen and X. Xiang. A source transfer domain decomposition method for Helmholtz equations in unbounded domain. *SIAM Journal on Numerical Analysis*, 51(4):2331–2356, 2013.
- [2868]
- [2869]
- [2870] [35] Jon\'as D De Basabe, Mrinal K. Sen, and Mary F Wheeler. The interior penalty discontinuous Galerkin method for elastic wave propagation: grid dispersion. *Geophysical Journal International*, 175(1):83–93, 10 2008.
- [2871]
- [2872]
- [2873] [36] P. de Fouquieres, S.G. Schirmer, S.J. Glaser, and Ilya Kuprov. Second order gradient ascent pulse engineering. *Journal of Magnetic Resonance*, 212(2):412–417, Oct 2011.
- [2874]
- [2875] [37] A El Kacimi and Omar Laghrouche. Wavelet based ILU preconditioners for the numerical solution by PUFEM of high frequency elastic wave scattering. *Journal of Computational Physics*, 230(8):3119–3134, 2011.
- [2876]
- [2877]
- [2878] [38] L. Emsley and G. Bodenhausen. Gaussian pulse cascades: New analytical functions for rectangular selective inversion and in-phase excitation in NMR. *Chem. Phys.*, 165(6):469–476, 1989.
- [2879]
- [2880]
- [2881] [39] B. Engquist and A. Majda. Absorbing boundary conditions for the numerical simulation of waves. *Math. Comp.*, 31:629, 1977.
- [2882]
- [2883] [40] B. Engquist and L. Ying. Sweeping preconditioner for the Helmholtz equation: hierarchical matrix representation. *Communications on pure and applied mathematics*, 64(5):697–735, 2011.
- [2884]
- [2885]
- [2886] [41] B. Engquist and L. Ying. Sweeping preconditioner for the Helmholtz equation: moving perfectly matched layers. *Multiscale Modeling & Simulation*, 9(2):686–710, 2011.
- [2887]
- [2888] [42] B. Engquist and H. Zhao. Approximate separability of the Green’s function of the Helmholtz equation in the high frequency limit. *Communications on Pure and Applied Mathematics*, 71(11):2220–2274, 2018.
- [2889]
- [2890]

- [43] Y. Erlangga, C. Oosterlee, and C. Vuik. A novel multigrid based preconditioner for heterogeneous Helmholtz problems. *SIAM Journal on Scientific Computing*, 27(4):1471–1492, 2006.
- [44] Y.A. Erlangga. Advances in iterative methods and preconditioners for the Helmholtz equation. *Archives of Computational Methods in Engineering*, 15(1):37–66, 2008.
- [45] Yogi A Erlangga, Cornelis Vuik, and Cornelis W Oosterlee. Comparison of multigrid and incomplete LU shifted-Laplace preconditioners for the inhomogeneous Helmholtz equation. *Applied numerical mathematics*, 56(5):648–666, 2006.
- [46] Yogi A Erlangga, Cornelis Vuik, and Cornelis Willebrordus Oosterlee. On a class of preconditioners for solving the Helmholtz equation. *Applied Numerical Mathematics*, 50(3-4):409–425, 2004.
- [47] O.G. Ernst and M.J. Gander. Why it is difficult to solve Helmholtz problems with classical iterative methods. In *Numerical analysis of multiscale problems*, pages 325–363. Springer, 2012.
- [48] B. Ewing, S. J. Glaser, and G. P. Drobny. Development and optimization of shaped NMR pulses for the study of coupled spin systems. *Chem. Phys.*, 147:121–129, 1990.
- [49] M. Gander and F. Nataf. AILU for Helmholtz problems: a new preconditioner based on an analytic factorization. *Comptes Rendus de l'Académie des Sciences - Series I - Mathematics*, 331(3):261–266, 2000.
- [50] M. Gander and H. Zhang. A class of iterative solvers for the Helmholtz equation: Factorizations, sweeping preconditioners, source transfer, single layer potentials, polarized traces, and optimized Schwarz methods. *SIAM Review*, 61(1):3–76, 2019.
- [51] Fortino Garcia, Daniel Appelö, and Olof Runborg. Analysis of an iterative solution of the Helmholtz equation via the wave equation for impedance boundary conditions. *in preparation*.
- [52] Fortino Garcia and N. Anders Petersson. Juqbox. Github, 2021.
- [53] Fortino Garcia and N. Anders Petersson. Juqbox.jl. Github, 2021.
- [54] Xiaozhen Ge, Haijin Ding, Herschel Rabitz, and Re-Bing Wu. Robust quantum control in games: An adversarial learning approach. *Physical Review A*, 101(5):052317, 2020.
- [55] Xiaozhen Ge and Re-Bing Wu. Risk-sensitive optimization for robust quantum controls. *arXiv preprint arXiv:2104.01323*, 2021.
- [56] J. Charles Gilbert and Patrick Joly. *Higher Order Time Stepping for Second Order Hyperbolic Problems and Optimal CFL Conditions*, pages 67–93. Springer Netherlands, Dordrecht, 2008.
- [57] A. Gillman, A.H. Barnett, and P-G. Martinsson. A spectrally accurate direct solution technique for frequency-domain scattering problems with variable media. *BIT Numerical Mathematics*, 55(1):141–170, Mar 2015.
- [58] R. Glowinski and T. Rossi. A mixed formulation and exact controllability approach for the computation of the periodic solutions of the scalar wave equation. (i): Controllability problem formulation and related iterative solution. *Comptes Rendus Math.*, 343(7):493–498, 2006.

- [59] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 1996.
- [60] Dan Gordon and Rachel Gordon. Robust and highly scalable parallel solution of the Helmholtz equation with large wave numbers. *Journal of Computational and Applied Mathematics*, 237(1):182–196, 2013.
- [61] M. J. Grote, F. Nataf, J. H. Tang, and P.-H. Tournier. Parallel Controllability Methods For the Helmholtz Equation. *arXiv e-prints*, page arXiv:1903.12522, Mar 2019.
- [62] M. J. Grote, A. Schneebeli, and D. Schötzau. Discontinuous Galerkin finite element method for the wave equation. *SIAM Journal on Numerical Analysis*, 44(6):2408–2431, 2006.
- [63] M.J. Grote and J.H. Tang. On controllability methods for the Helmholtz equation. *Journal of Computational and Applied Mathematics*, 358:306–326, 2019.
- [64] Stefanie Günther and N. Anders Petersson. Quandary: Optimal control for open quantum systems. <https://github.com/LLNL/quandary>, 2021.
- [65] William W. Hager. Runge-Kutta methods in optimal control and the transformed adjoint system. *Numerische Mathematik*, 87(2):247–282, Dec 2000.
- [66] T. Hagstrom. Radiation boundary conditions for the numerical simulation of waves. *Acta Numerica*, 8:47–106, 1999.
- [67] E. Hairer, C. Lubich, and G. Wanner. *Geometric Numerical Integration*. Number 31 in Springer series in computational mathematics. Springer-Verlag, Heidelberg, 2nd edition, 2006.
- [68] E. Heikkola, S. Mönkölä, A. Pennanen, and T. Rossi. Controllability method for acoustic scattering with spectral elements. *Journal of Computational and Applied Mathematics*, 204(2):344–355, 2007.
- [69] E. Heikkola, S. Mönkölä, A. Pennanen, and T. Rossi. Controllability method for the Helmholtz equation with higher-order discretizations. *Journal of Computational Physics*, 225(2):1553–1576, 2007.
- [70] C. V. Hile and G. A Kriegsmann. A hybrid numerical method for loaded highly resonant single mode cavities. *Journal of Computational Physics*, 142(2):506–520, 1998.
- [71] J.R. Johansson, P.D. Nation, and Franco Nori. Qutip 2: A python framework for the dynamics of open quantum systems. *Computer Physics Communications*, 184(4):1234 – 1240, 2013.
- [72] D. Ketcheson, L. Lóczsi, and T. Kocsis. On the absolute stability regions corresponding to partial sums of the exponential function. *IMA Journal of Numerical Analysis*, 35(3):1426–1455, 2015.
- [73] N. Khaneja, T. Reiss, C. Kehlet, T. Schulte-Herbruggen, and S. Glaser. Optimal control of coupled spin dynamics: design of NMR pulse sequences by gradient ascent algorithms. *J. Magnetic Resonance*, 172:296–305, 2005.
- [74] Q Kong and A Zettl. Eigenvalues of regular Sturm–Liouville problems. *Journal of differential equations*, 131(1):1–19, 1996.

- [75] H.-O. Kreiss and J. Oliger. Comparison of accurate methods for the integration of hyperbolic equations. *Tellus*, 24:199–215, 1972.
- [76] O. A. Ladyzhenskaya. On the limiting-amplitude principle. *Uspekhi Mat. Nauk*, 12(4):161–164, 1957.
- [77] Alistair L Laird and M Giles. Preconditioned iterative solution of the 2D Helmholtz equation. 2002.
- [78] N. Leung, M. Abdelhafez, Jens Koch, and D. Schuster. Speedup for quantum optimal control from automatic differentiation based on graphics processing units. *Phys. Rev. A*, 95:0432318, 2017.
- [79] Randall J. LeVeque. *Finite volume methods for hyperbolic problems*. Cambridge University Press, Cambridge, 2002.
- [80] Yang Li, Ludovic Métivier, Romain Brossier, Bo Han, and Jean Virieux. 2d and 3d frequency-domain elastic wave modeling in complex media with a parallel iterative solver. *Geophysics*, 80(3):T101–T118, 2015.
- [81] I. Livshits and A. Brandt. Accuracy properties of the Wave-ray multigrid algorithm for Helmholtz equations. *SIAM Journal on Scientific Computing*, 28(4):1228–24, 2006.
- [82] Dennis Lucarelli. Quantum optimal control via gradient ascent in function space and the time-bandwidth quantum speed limit. *Physical Review A*, 97(6), Jun 2018.
- [83] S. Machnes, E. Assémat, D. Tannor, and F. K. Wilhelm. Tunable, flexible, and efficient optimization of control pulses for practical qubits. *Physical Review Letters*, 120(15), Apr 2018.
- [84] E. Magesan and J. M. Gambetta. Effective Hamiltonian models of the cross-resonance gate. *Phys. Rev. A*, 101, 2020.
- [85] Marcus J. Grote, Frédéric Nataf, Jet Hoe Tang, and Pierre-Henri Tournier. Parallel controllability methods for the Helmholtz equation. *Computer Methods in Applied Mechanics and Engineering*, 362:112846, 2020.
- [86] K. Mattsson. Summation by parts operators for finite difference approximations of second-derivatives with variable coefficients. *Journal of Scientific Computing*, 51(3):650–682, Jun 2012.
- [87] Sanna Mönkölä, Erkki Heikkola, Anssi Pennanen, and Tuomo Rossi. Time-harmonic elasticity with controllability and higher-order discretization methods. *Journal of Computational Physics*, 227(11):5513–5534, 2008.
- [88] C.S. Morawetz. The limiting amplitude principle. *Communications on Pure and Applied Mathematics*, 15(3):349–361, 1962.
- [89] K. W. Morton and D. F. Mayers. *Numerical Solution of Partial Differential Equations: An Introduction*. Cambridge University Press, 2 edition, 2005.
- [90] M. Nielsen and I. Chuang. *Quantum computation and quantum information*. Cambridge University Press, 2000.

- 3004 [91] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, 2nd edition, 2006.
- 3005 [92] Sina Ober-Blöbaum. *Discrete mechanics and optimal control*. PhD thesis, University of
3006 Paderborn, 2008.
- 3007 [93] Zhichao Peng and Daniel Appelö. EM-Waveholtz: A flexible frequency-domain method built
3008 from time-domain solvers. *arXiv preprint arXiv:2103.14789*, 2021.
- 3009 [94] N. Anders Petersson and Fortino Garcia. Optimal control of closed quantum systems via
3010 B-splines with carrier waves. *arXiv preprint arXiv:2106.14310*, 2021.
- 3011 [95] R. E. Plessix and W. A. Mulder. Separation-of-variables as a preconditioner for an iterative
3012 Helmholtz solver. *Applied numerical mathematics*, 44(3):385–400, 2003.
- 3013 [96] G. Rizzuti and W.A. Mulder. Multigrid-based ‘shifted-Laplacian’ preconditioning for the
3014 time-harmonic elastic wave equation. *Journal of Computational Physics*, 317:47–65, 2016.
- 3015 [97] J. M. Sanz-Serna. Symplectic Runge–Kutta schemes for adjoint equations, automatic differ-
3016 entiation, optimal control, and more. *SIAM Review*, 58(1):3–33, 2016.
- 3017 [98] Yunong Shi, Nelson Leung, Pranav Gokhale, Zane Rossi, David I. Schuster, Henry Hoffmann,
3018 and Frederic T. Chong. Optimized compilation of aggregated instructions for realistic quan-
3019 tum computers. *Proceedings of the Twenty-Fourth International Conference on Architectural
3020 Support for Programming Languages and Operating Systems - ASPLOS ’19*, 2019.
- 3021 [99] Gregory R Shubin and John B Bell. A modified equation approach to constructing fourth
3022 order methods for acoustic wave propagation. *SIAM Journal on Scientific and Statistical
3023 Computing*, 8(2):135–151, 1987.
- 3024 [100] C.C. Stolk. A rapidly converging domain decomposition method for the Helmholtz equation.
3025 *Journal of Computational Physics*, 241:240–252, 2013.
- 3026 [101] Christiaan C. Stolk. A time-domain preconditioner for the Helmholtz equation, 2020.
- 3027 [102] Eran Treister. Shifted Laplacian multigrid for the elastic Helmholtz equation, 2018.
- 3028 [103] Paul Tsuji, Jack Poulson, Björn Engquist, and Lexing Ying. Sweeping preconditioners for
3029 elastic wave propagation with spectral element methods. *ESAIM: Mathematical Modelling
3030 and Numerical Analysis-Modélisation Mathématique et Analyse Numérique*, 48(2):433–447,
3031 2014.
- 3032 [104] B. R. Vainberg. On short-wave asymptotic behaviour of solutions to steady-state problems
3033 and the asymptotic behaviour as $t \rightarrow \infty$ of solutions of time-dependent problems. *Uspekhi
3034 Mat. Nauk*, 30(2):1–58, 1975.
- 3035 [105] A. Vion and C. Geuzaine. Double sweep preconditioner for optimized Schwarz methods
3036 applied to the Helmholtz problem. *Journal of Computational Physics*, 266:171–190, 2014.
- 3037 [106] A. Wächter and L. T. Biegler. On the implementation of an interior-point filter line-search
3038 algorithm for large-scale nonlinear programming. *Mathematical Programming*, 106(1):25–57,
3039 Mar 2006.

- 3040 [107] S. Wang, M.V. de Hoop, and J. Xia. On 3d modeling of seismic wave propagation via a
 3041 structured parallel multifrontal direct Helmholtz solver. *Geophysical Prospecting*, 59(5):857–
 3042 873, 2011.
- 3043 [108] Shen Wang, Maarten V de Hoop, Jianlin Xia, and Xiaoye S Li. Massively parallel structured
 3044 multifrontal solver for time-harmonic elastic waves in 3-D anisotropic media. *Geophysical*
 3045 *Journal International*, 191(1):346–366, 2012.
- 3046 [109] H. Weyl. Über die asymptotische verteilung der eigenwerte. *Nachr. Konigl. Ges. Wiss.*, pages
 3047 110–117, 1911.
- 3048 [110] D. B. Zax, G. Goelman, and S. Vega. Amplitude-modulated composite pulses. *J. Magn.*
 3049 *Reson.*, 80(2):375–382, 1988.
- 3050 [111] L. Zepeda-Núñez, A. Scheuer, R. J. Hewett, and L. Demanet. The method of polarized traces
 3051 for the 3D Helmholtz equation. *ArXiv e-prints*, January 2018.
- 3052 [112] L. Zepeda-Núñez and L. Demanet. The method of polarized traces for the 2d Helmholtz
 3053 equation. *Journal of Computational Physics*, 308:347–388, 2016.
- 3054 [113] L. Zepeda-Núñez and L. Demanet. Nested domain decomposition with polarized traces for
 3055 the 2d Helmholtz equation. *SIAM Journal on Scientific Computing*, 40(3):B942–B981, 2018.

3056 .1 Proof of Lemma 2.1.1

Proof. We show the results for the rescaled transfer function

$$\bar{\beta}(r) := \beta(r\omega) = \frac{2}{T} \int_0^T \left(\cos(\omega t) - \frac{1}{4} \right) \cos(r\omega t) dt = \frac{1}{\pi} \int_0^{2\pi} \left(\cos(t) - \frac{1}{4} \right) \cos(rt) dt.$$

By direct integration we get

$$\bar{\beta}(r) = \frac{1}{\pi} \int_0^{2\pi} \frac{1}{2} (\cos((r+1)t) + \cos((r-1)t)) - \frac{1}{4} \cos(rt) dt = \quad (1)$$

$$\frac{1}{2\pi} \left(\frac{\sin(2\pi(r+1))}{r+1} + \frac{\sin(2\pi(r-1))}{r-1} - \frac{1}{2} \frac{\sin(2\pi r)}{r} \right) = \text{sinc}(r+1) + \text{sinc}(r-1) - \frac{1}{2} \text{sinc}(r), \quad (2)$$

where

$$\text{sinc}(r) = \frac{\sin(2\pi r)}{2\pi r}.$$

- 3057 We use the fact that $\sin(x) \leq x - \tilde{\alpha}x^3$ in the interval $x \in [0, \pi]$ for any $\tilde{\alpha} \in [0, \pi^{-2}]$. This leads to
 3058 the following estimate for the sinc function

$$0 \leq \text{sinc}(r) \leq 1 - \alpha r^2, \quad r \in [-0.5, 0.5], \quad \alpha \in [0, 4]. \quad (3)$$

3059 We also note that $\text{sinc}(r+n) = \text{sinc}(r)r/(r+n)$ for all integer n .

We now first consider $0 \leq r \leq 0.5$ and use (3) with $\alpha = 4$ and $\alpha = 0$,

$$|\bar{\beta}(r)| = \text{sinc}(r) \left| \frac{r}{r+1} + \frac{r}{r-1} - \frac{1}{2} \right| = \frac{1}{2} \text{sinc}(r) \frac{1+3r^2}{1-r^2} \leq \frac{1}{2} \frac{1-4r^2+3r^2}{1-r^2} = \frac{1}{2}.$$

For $0.5 \leq r \leq 1.5$ we instead center around $r = 1$ and get for $|\delta| \leq 0.5$,

$$\bar{\beta}(1+\delta) = \text{sinc}(\delta) \frac{3(\delta+1)^2+1}{2(2+\delta)(1+\delta)} \geq 0,$$

since $\text{sinc}(\delta) \geq 0$. Moreover, using again (3) with $\alpha = 1$,

$$\bar{\beta}(1+\delta) \leq (1-\delta^2) \frac{3(\delta+1)^2+1}{2(2+\delta)(1+\delta)} = \frac{4+2\delta-3\delta^2-3\delta^3}{2(2+\delta)} \leq \frac{4+2\delta-2\delta^2-\delta^3}{2(2+\delta)} = 1 - \frac{\delta^2}{2}.$$

Finally, for $r > 1$ we have $1/(r+1) - 1/2r \geq 0$ and therefore

$$|\bar{\beta}(r)| = \frac{|\sin(2\pi r)|}{2\pi r} \left(\frac{r}{r+1} + \frac{r}{r-1} - \frac{1}{2} \right) \leq \frac{1}{2\pi} \left(\frac{1}{r+1} + \frac{1}{r-1} - \frac{1}{2r} \right) \leq \frac{3}{4\pi} \frac{1}{(r-1)}.$$

3060 We note also that for $r \geq 1.5$ this gives $|\bar{\beta}(r)| \leq 3/2\pi \leq 1/2$.

To prove (2.13) we use the Taylor expansion of $\bar{\beta}$ around $r = 1$,

$$\bar{\beta}(1+\delta) = 1 + \frac{\delta^2}{2} \bar{\beta}''(1) + \frac{\delta^3}{6} \bar{R}(\delta),$$

where $\bar{R}(\delta)$ is the remainder term, which can be bounded as

$$|\bar{R}(\delta)| \leq \sup_{r \geq 0} |\bar{\beta}^{(3)}(r)| \leq \frac{1}{\pi} \int_0^{2\pi} t^3 \left(1 + \frac{1}{4} \right) dt = 5\pi^3.$$

Hence, $|R(\delta)| \leq |\bar{R}(\delta)|/6 \leq 5\pi^3/6$. Finally,

$$\beta''(1) = \text{sinc}''(2) + \text{sinc}''(0) - \frac{1}{2} \text{sinc}''(1) = \frac{-1}{2} - \frac{(2\pi)^2}{3} + 1 = -2b_1.$$

3061 This shows (2.13) and concludes the proof of the lemma. \square

3062 .2 Verification of Discrete Solution

Here we verify that (2.19) is indeed a solution to the difference equation (2.18). Direct substitution yields

$$\begin{aligned}
 \hat{w}_j^{n+1} - 2\hat{w}_j^n + \hat{w}_j^{n-1} + \Delta t^2 \lambda_j^2 \hat{w}_j^n &= (\hat{v}_j - \hat{v}_j^\infty) \cos(\tilde{\lambda}_j t_n) \left(\cos(\tilde{\lambda}_j \Delta t) - 2 + \Delta t^2 \lambda_j^2 + \cos(\tilde{\lambda}_j \Delta t) \right) \\
 &\quad + \hat{v}_j^\infty \cos(\omega t_n) \left(\cos(\omega \Delta t) - 2 + \Delta t^2 \lambda_j^2 + \cos(\omega \Delta t) \right) \\
 &= (\hat{v}_j - \hat{v}_j^\infty) \cos(\tilde{\lambda}_j t_n) \left(-4 \sin^2(\tilde{\lambda}_j \Delta t / 2) + \Delta t^2 \lambda_j^2 \right) \\
 &\quad + \hat{v}_j^\infty \cos(\omega t_n) \left(-4 \sin^2(\omega \Delta t / 2) + \Delta t^2 \lambda_j^2 \right) \\
 &= \hat{v}_j^\infty \cos(\omega t_n) (-\tilde{\omega}^2 + \lambda_j^2) \Delta t^2 = -\Delta t^2 f_j \cos(\omega t_n).
 \end{aligned}$$

Second, the initial conditions are satisfied since

$$\begin{aligned}
 \hat{w}_j^0 &= (\hat{v}_j - \hat{v}_j^\infty) + \hat{v}_j^\infty = \hat{v}_j, \\
 \hat{w}_j^{-1} &= (\hat{v}_j - \hat{v}_j^\infty) \cos(\tilde{\lambda}_j \Delta t) + \hat{v}_j^\infty \cos(\omega \Delta t) = (\hat{v}_j - \hat{v}_j^\infty) \left(1 - \frac{1}{2} \Delta t^2 \lambda_j^2 \right) + \hat{v}_j^\infty \left(1 - \frac{1}{2} \Delta t^2 \tilde{\omega}^2 \right) \\
 &= \hat{v}_j \left(1 - \frac{1}{2} \Delta t^2 \lambda_j^2 \right) + \frac{1}{2} \Delta t^2 \hat{v}_j^\infty (\lambda_j^2 - \tilde{\omega}^2) = \hat{v}_j \left(1 - \frac{1}{2} \Delta t^2 \lambda_j^2 \right) - \frac{1}{2} \Delta t^2 \hat{f}_j.
 \end{aligned}$$

3063 This shows that (2.19) solves (2.18).

3064 .3 Proof of Lemma 2.1.5

Proof. In general, we introduce the trapezoidal rule applied to $\cos(\alpha t)$ in $[0, 1]$,

$$\mathcal{T}_h(\alpha) := h \sum_{n=0}^M \eta_n \cos(\alpha t_n) \approx \int_0^1 \cos(\alpha t) dt = \frac{\sin(\alpha)}{\alpha}, \quad h = 1/M,$$

3065 from which we attain the following lemma:

Lemma .3.1. *The error in $\mathcal{T}_h(\alpha)$ satisfies¹*

$$\left| \int_0^1 \cos(\alpha t) dt - \mathcal{T}_h(\alpha) \right| \leq \frac{h^2 |\alpha|}{\pi^2}, \quad \text{when } |h\alpha| \leq \pi.$$

¹ Note that this estimate is sharper than the standard error estimate for the trapezoidal rule, which would have the factor α^2 from the second derivative of the integrand, not just α .

Proof. A direct calculation shows that

$$\mathcal{T}_h(\alpha) = g(h\alpha) \int_0^1 \cos(\alpha t) dt, \quad g(x) = \frac{x}{2 \tan(x/2)}.$$

The function $g(x)$ can be bounded as $1 - x^2/\pi^2 \leq g(x) \leq 1$ for $|x| \leq \pi$. This gives

$$\left| \int_0^1 \cos(\alpha t) dt - \mathcal{T}_h(\alpha) \right| = |1 - g(h\alpha)| \left| \frac{\sin \alpha}{\alpha} \right| \leq \frac{(h\alpha)^2}{\pi^2} \left| \frac{\sin \alpha}{\alpha} \right| \leq \frac{h^2 |\alpha|}{\pi^2}.$$

3066

□

Since

$$\left(\cos(\omega t) - \frac{1}{4} \right) \cos(\lambda t) = \frac{1}{2} \left(\cos((\omega + \lambda)t) + \cos((\omega - \lambda)t) - \frac{1}{2} \cos(\lambda t) \right),$$

and $h = \Delta t/T$, we can write

$$\begin{aligned} \beta_h(\lambda) &= \frac{\Delta t}{T} \sum_{n=0}^M \eta_n \left(\cos((\omega + \lambda)t_n) + \cos((\omega - \lambda)t_n) - \frac{1}{2} \cos(\lambda t_n) \right) \\ &= \left[\mathcal{T}_h(T(\omega + \lambda)) + \mathcal{T}_h(T(\omega - \lambda)) - \frac{1}{2} \mathcal{T}_h(T\lambda) \right]. \end{aligned}$$

From Lemma .3.1 we then get that

$$|\beta(\tilde{\lambda}_j) - \beta_h(\tilde{\lambda}_j)| \leq \frac{h^2}{\pi^2} \left(|T(\omega + \tilde{\lambda}_j)| + |T(\omega - \tilde{\lambda}_j)| + \frac{1}{2} |T\tilde{\lambda}_j| \right) \leq \frac{5h^2 T}{2\pi^2} (\omega + \tilde{\lambda}_j) = \frac{5\Delta t^2}{2\pi^2 T} (\omega + \tilde{\lambda}_j),$$

when

$$\pi \geq hT(\omega + \tilde{\lambda}_j) = \Delta t(\omega + \tilde{\lambda}_j),$$

3067 which is true by (3.25) and the fact that $\arcsin(x) \leq \pi x/2$ for $x \in [0, 1]$:

$$\tilde{\lambda}_j = \frac{2}{\Delta t} \arcsin \left(\frac{\Delta t \lambda_j}{2} \right) \leq \frac{\pi}{2} \lambda_j. \quad (4)$$

3068 Next, we use the inequality $|x - \sin(x)| \leq x^3/6$ for $|x| \leq \pi/2$ to show that

$$\left| \frac{\sin(xh)}{h} - x \right| = \frac{1}{h} |\sin(xh) - xh| \leq \frac{(xh)^3}{6h} = \frac{h^2 x^3}{6}, \quad |hx| \leq \frac{\pi}{2}. \quad (5)$$

It gives us an estimate for $\tilde{\lambda}_j - \lambda_j$,

$$|\lambda_j - \tilde{\lambda}_j| = \left| \frac{\sin(\Delta t \tilde{\lambda}_j / 2)}{\Delta t / 2} - \tilde{\lambda}_j \right| \leq \frac{\Delta t^2}{24} \tilde{\lambda}_j^3,$$

3069 which is valid for all j since $\Delta t \tilde{\lambda}_j / 2 \leq \Delta t \pi \lambda_j / 4 \leq \Delta t \pi \lambda_N / 4 \leq \pi / 2$, by (3.25) and (4).

By Lemma 2.1.1

$$|\beta(\omega + r)| \leq \begin{cases} 1 - \frac{r^2}{2\omega^2}, & |r/\omega| \leq \frac{1}{2}, \\ \frac{1}{2}, & |r/\omega| \geq \frac{1}{2}. \end{cases}$$

We now claim that the statement (2.20) in the lemma holds for all j if $\Delta t \omega \leq \min(\delta_h, 1)$. On the one hand, if $|\omega - \tilde{\lambda}_j| \geq \omega / 2$, by (3.25) and (4)

$$|\beta_h(\tilde{\lambda}_j)| \leq |\beta(\tilde{\lambda}_j)| + \frac{5\Delta t^2}{2\pi^2 T} (\omega + \tilde{\lambda}_j) \leq \frac{1}{2} + \frac{5\Delta t^2}{2\pi^2 T} (\omega + \tilde{\lambda}_j) = \frac{1}{2} + \frac{5\Delta t \omega}{4\pi^3} \Delta t (\omega + \tilde{\lambda}_j) \leq \frac{1}{2} + \frac{5\Delta t \omega}{4\pi^2} \leq 0.63.$$

On the other hand, if $|\omega - \tilde{\lambda}_j| < \omega / 2$,

$$\frac{|\omega - \tilde{\lambda}_j|}{\omega} \geq \frac{|\omega - \lambda_j|}{\omega} - \frac{|\tilde{\lambda}_j - \lambda_j|}{\omega} \geq \delta_h - \frac{\Delta t^2}{24\omega} \tilde{\lambda}_j^3 \geq \delta_h - \frac{\Delta t^2}{24\omega} (\omega + |\tilde{\lambda}_j - \omega|)^3 \geq \delta_h - \frac{(3/2)^3}{24} \Delta t^2 \omega^2 \geq \frac{55}{64} \delta_h,$$

since $\min(\delta_h, 1)^2 \leq \delta_h$. Then

$$\begin{aligned} |\beta_h(\tilde{\lambda}_j)| &\leq |\beta(\tilde{\lambda}_j)| + \frac{5\Delta t^2}{48T} (\omega + \tilde{\lambda}_j) \leq 1 - \frac{1}{2} \left(\frac{|\omega - \tilde{\lambda}_j|}{\omega} \right)^2 + \frac{5\Delta t^2 \omega}{96\pi} (\omega + \omega + \omega/2) \\ &\leq 1 - \frac{55^2}{2 \cdot 64^2} \delta_h^2 + \Delta t^2 \omega^2 \frac{25}{192\pi} \leq 1 - \left(\frac{55^2}{2 \cdot 64^2} - \frac{25}{192\pi} \right) \delta_h^2 \leq 1 - 0.3 \delta_h^2. \end{aligned}$$

3070 This proves the lemma. □

3071

3072 .4 Proof of Lemma 3.1.1

We show the results for the rescaled function

$$\bar{\gamma}(r) := \gamma(r\omega) = \frac{2}{T} \int_0^T \left(\cos(\omega t) - \frac{1}{4} \right) \sin(r\omega t) dt = \frac{1}{\pi} \int_0^{2\pi} \left(\cos(t) - \frac{1}{4} \right) \sin(rt) dt.$$

By direct integration we get

$$\bar{\gamma}(r) = \frac{1}{\pi} \int_0^{2\pi} \frac{1}{2} (\sin((r+1)t) + \sin((r-1)t)) - \frac{1}{4} \sin(rt) dt = \frac{(1+3r^2) \sin^2(\pi r)}{2\pi r(r^2-1)} = \frac{\pi r(1+3r^2) \text{sinc}^2(r/2)}{2(r^2-1)},$$

where

$$\text{sinc}(r) = \frac{\sin(2\pi r)}{2\pi r}.$$

From [14] we have the following expression for β :

$$\bar{\beta}(r) = \frac{1}{\pi} \int_0^{2\pi} \frac{1}{2} (\cos((r+1)t) + \cos((r-1)t)) - \frac{1}{4} \cos(rt) dt = \frac{(1+3r^2)\sin(2\pi r)}{4\pi r(r^2-1)} = \frac{(1+3r^2)\text{sinc}(r)}{2(r^2-1)}.$$

Then the eigenvalues of the WaveHoltz operator applied to the first order system are

$$|\bar{\mu}(r)|^2 = \bar{\beta}^2(r) + \bar{\gamma}^2(r) = \frac{(1+3r^2)^2 \sin^2(\pi r)}{4\pi^2 r^2 (r^2-1)^2} = \frac{(1+3r^2)^2 \text{sinc}^2(r/2)}{4(r^2-1)^2}.$$

We now first consider $0 \leq r \leq 0.5$ and note that $|\mu(r)|^2$ is a positive, increasing function on this interval so that

$$|\bar{\mu}(r)|^2 \leq |\bar{\mu}(1/2)| = \frac{49}{9\pi^2} \leq 0.56.$$

For $1/2 \leq r \leq 3/2$ we instead center around $r = 1$ and get for $|\delta| \leq 1/2$,

$$|\bar{\mu}(1+\delta)|^2 = \frac{(3(\delta+1)^2+1)^2 \sin^2(\pi\delta)}{4\pi^2 \delta^2 (1+\delta)^2 (2+\delta)^2} = \frac{(3(\delta+1)^2+1)^2 \text{sinc}^2(\delta/2)}{4(1+\delta)^2 (2+\delta)^2}.$$

We use the fact that $\sin(x) \leq x - \tilde{\alpha}x^3$ in the interval $x \in [0, \pi]$ for any $\tilde{\alpha} \in [0, \pi^{-2}]$. This leads to the following estimate for the sinc function

$$0 \leq \text{sinc}(r/2) \leq 1 - \alpha r^2, \quad r \in [-0.5, 0.5], \quad \alpha \in [0, 1]. \quad (6)$$

Using (6) with $\alpha = 1$, gives

$$\begin{aligned} |\bar{\mu}(1+\delta)|^2 &\leq \frac{(3(\delta+1)^2+1)^2 (1-\delta^2)^2}{4(1+\delta)^2 (2+\delta)^2} = \frac{(4+2\delta-3\delta^2-3\delta^3)^2}{4(2+\delta)^2} \leq \frac{(4+2\delta-2\delta^2-\delta^3)^2}{4(2+\delta)^2} = \left(1 - \frac{\delta^2}{2}\right)^2 \\ &= 1 - \delta^2 + \frac{\delta^4}{4} \\ &\leq 1 - \frac{15}{16}\delta^2, \end{aligned}$$

since $|\delta| < 1/2$. A Taylor expansion around $\delta = 0$ for $|\delta| \leq 1/2$ immediately gives the bound

$$\sqrt{1-\delta^2} \leq 1 - \frac{\delta^2}{2} \implies |\bar{\mu}(1+\delta)| \leq \sqrt{1 - \frac{15}{16}\delta^2} \leq 1 - \frac{15\delta^2}{32}.$$

If we consider $r \geq 3/2$,

$$|\bar{\mu}(r)|^2 = \frac{(1+3r^2)^2 \text{sinc}^2(r/2)}{4(r^2-1)^2} \leq \frac{(1+3r^2)^2}{4(r^2-1)^2},$$

which is a positive and decreasing function. It follows that

$$|\bar{\mu}(r)|^2 \leq \frac{(1+3(3/2)^2)^2}{4((3/2)^2 - 1)^2} \leq 0.44,$$

for $r \geq 3/2$. Finally, for a more general bound for $r > 1$ we have $1/(r+1) - 1/2r \geq 0$ so that

$$|\bar{\mu}(r)|^2 = \frac{(1+3r^2)^2 \sin^2(\pi r)}{4\pi^2 r^2(r^2-1)^2} \leq \frac{(1+3r^2)^2}{4\pi^2 r^2(r^2-1)^2} = \frac{1}{\pi^2} \left(\frac{1}{r+1} + \frac{1}{r-1} - \frac{1}{2r} \right)^2 \leq \left(\frac{3}{2\pi(r-1)} \right)^2,$$

which gives

$$|\bar{\mu}(r)| \leq \frac{3}{2\pi(r-1)}.$$

To prove (3.13), we use a Taylor expansion of $\bar{\mu}(r)$ about $r = 1$ in the interval $r \in (1/2, 3/2)$,

$$|\bar{\mu}(1+\delta)| = 1 + \frac{\delta^2}{2} \frac{d^2}{dr^2} [|\bar{\mu}(r)|]_{r=1} + \frac{\delta^3}{6} \bar{R}(\delta),$$

where $\bar{R}(\delta)$ is the remainder term. We note that by product rule we have

$$\frac{d}{dr} |\bar{\mu}(r)| = \frac{1}{|\bar{\mu}|} (\bar{\beta}\bar{\beta}' + \bar{\gamma}\bar{\gamma}').$$

Since

$$\frac{d}{dr} |\bar{\mu}(r)|^{-s} = -s |\bar{\mu}(r)|^{-s-1} \frac{d}{dr} |\bar{\mu}(r)| = \frac{-s}{|\bar{\mu}(r)|^{s+2}} (\bar{\beta}\bar{\beta}' + \bar{\gamma}\bar{\gamma}'),$$

by repeated product rule we can then show that

$$\begin{aligned} \frac{d^3}{dr^3} |\bar{\mu}(r)| &= \frac{3}{|\bar{\mu}|^5} (\bar{\beta}\bar{\beta}' + \bar{\gamma}\bar{\gamma}')^2 - \frac{1}{|\bar{\mu}|^3} (\bar{\beta}\bar{\beta}'' + (\bar{\beta}')^2 + \bar{\gamma}\bar{\gamma}'' + (\bar{\gamma}')^2)(1 + \bar{\beta}\bar{\beta}' + \bar{\gamma}\bar{\gamma}') \\ &\quad + \frac{1}{|\bar{\mu}|} (\bar{\beta}\bar{\beta}''' + 3\bar{\beta}'\bar{\beta}'' + \bar{\gamma}\bar{\gamma}''' + 3\bar{\gamma}'\bar{\gamma}''). \end{aligned}$$

We note that $|\bar{\mu}(r)| \geq |\bar{\mu}(3/2)| \geq 1/\pi$ in the interval $1/2 \leq r \leq 3/2$, and that we have the following bound

$$\sup_{r \geq 0} \left| \bar{\beta}^{(s)}(r) \right| \leq \frac{1}{\pi} \int_0^{2\pi} t^s \left(1 + \frac{1}{4} \right) dt = 5 \frac{2^{s-1} \pi^s}{s+1},$$

which similarly holds for $\sup_{r \geq 0} |\bar{\gamma}^{(s)}(r)|$ for $s = 0, 1, 2, \dots$. Thus by Taylor's theorem we have

$$\begin{aligned} |\bar{R}(\delta)| &\leq \sup_{1/2 \leq r \leq 3/2} \left| \frac{d^3}{dr^3} |\bar{\mu}(r)| \right| \leq \frac{3}{|\bar{\mu}(3/2)|^5} \frac{25^2 \pi^2}{4} + \frac{3}{|\bar{\mu}(3/2)|^3} \left(\frac{50\pi^2}{3} + \frac{25\pi^2}{2} \right) \left(1 + \frac{25\pi}{2} \right) + \frac{3 \cdot 75\pi^3}{|\bar{\mu}(3/2)|} \\ &\leq \frac{3}{4} 25^2 \pi^7 + 3\pi^3 \left(\frac{50\pi^2}{3} + \frac{25\pi^2}{2} \right) \left(1 + \frac{25\pi}{2} \right) + 3 \cdot 75\pi^4 \\ &= \frac{25\pi^4}{4} (36 + 20\pi + 250\pi^2 + 75\pi^3). \end{aligned}$$

Then, $|R(\delta)| \leq |\bar{R}(\delta)|/6$. Finally,

$$\frac{d^2}{dr^2} [|\bar{\mu}(r)|]_{r=1} = \frac{1}{6} (3 - 2\pi^2) = -2b_1.$$

3073

3074 .5 Wave Equation Extension

3075 Let $\Omega = (-\infty, 0)$ and let $f \in L^2(\Omega)$ be compactly supported in Ω away from $x = 0$. Additionally, assume $1/c^2 \in L^1_{\text{loc}}(\Omega)$ with $c(0) = c_0$ on the interval $[-\delta, 0]$ for some $\delta > 0$. We consider 3076 the semi-infinite problem

$$\begin{aligned} w_{tt} &= \frac{\partial}{\partial x} \left[c^2(x) \frac{\partial}{\partial x} w \right] - \operatorname{Re}\{f(x)e^{i\omega t}\}, \quad x \leq 0, \quad 0 \leq t \leq T, \\ w(0, x) &= v_0(x), \quad w_t(0, x) = v_1(x), \\ \alpha w_t(t, 0) + \beta c_0 w_x(t, 0) &= 0, \end{aligned}$$

3078 Let \tilde{w} solve the extended wave equation

$$\begin{aligned} \tilde{w}_{tt} &= \frac{\partial}{\partial x} \left[\tilde{c}^2(x) \frac{\partial}{\partial x} \tilde{w} \right] - \operatorname{Re}\{\tilde{f}(x)e^{i\omega t}\}, \quad x \in \mathbb{R}, \quad 0 \leq t \leq T, \\ \tilde{w}(0, x) &= \tilde{v}_0(x), \quad \tilde{w}_t(0, x) = \tilde{v}_1(x), \\ \alpha \tilde{w}_t(t, 0) + \beta c_0 \tilde{w}_x(t, 0) &= 0. \end{aligned}$$

where \tilde{f} is a zero extension, \tilde{c} is the extended wavespeed

$$\tilde{c}(x) = \begin{cases} c_0, & -\delta < x \leq 0, \\ \tilde{c}_0, & x > 0, \end{cases}$$

and (I) $\tilde{v}_0 \in H^1$ and $\tilde{v}_1 \in L^2$. We then choose the extension of v_0 and v_1 such that

$$\tilde{v}_1(x) + \tilde{v}'_0(x) = 0, \quad x > 0, \quad (\text{II})$$

which will have purely right-going waves in the extended region $x > 0$. Moreover, since c is constant in $[-\delta, 0]$ the solution will then be of the form

$$\tilde{w}(t, x) = \begin{cases} w_L(x + c_0 t) + w_R(x - c_0 t), & -\delta \leq x \leq 0, \\ w_T(x - \tilde{c}_0 t), & x > 0, \end{cases}$$

for some functions w_L, w_R , and w_T . At $x = 0$ where \tilde{c} is (potentially) discontinuous, the weak solution satisfies the interface conditions that \tilde{w} and $\tilde{c}^2 \tilde{w}_x$ are both continuous. These requirements lead to the relations

$$w_L(c_0 t) + w_R(-c_0 t) = w_T(-\tilde{c}_0 t),$$

$$c_0^2(w'_L(c_0 t) + w'_R(-c_0 t)) = \tilde{c}_0^2 w'_T(-c_0 t).$$

It follows that

$$\tilde{w}_t(t, 0^-) = c_0(w'_L(c_0 t) - w_R(-c_0 t)) = -\tilde{c}_0 w'_T(-\tilde{c}_0 t), \quad c_0 \tilde{w}_x(t, 0^-) = \frac{\tilde{c}_0^2}{c_0} w'_T(-c_0 t),$$

so that the impedance condition

$$\alpha \tilde{w}_t(t, 0^-) + \beta c_0 \tilde{w}_x(t, 0^-) = \left(-\alpha \tilde{c}_0 + \beta \frac{\tilde{c}_0^2}{c_0} \right) w'_T(-c_0 t) = 0,$$

is satisfied if

$$\tilde{c}_0 = \frac{\alpha}{\beta} c_0. \quad (\text{III})$$

With this choice of the extended wavespeed \tilde{c}_0 , both \tilde{w} and w satisfy the same PDE and condition at $x = 0$ so that they must be equal for $x < 0$. In summary, if we have that conditions (I-III) are satisfied, we have that $\tilde{w}(t, x) = w(t, x)$ for $x < 0$. We note that a similar argument can be made for an interior impedance problem on a bounded domain, $a \leq x \leq b$, to a problem on \mathbb{R} . In

3083 this case, assuming $c(a) = c_a$, $c(b) = c_b$ where c is constant near the endpoints, then the following
 3084 problem has $\tilde{w}(t, x) = w(t, x)$ for $a \leq x \leq b$:

$$\tilde{w}_{tt} = \frac{\partial}{\partial x} \left[\tilde{c}^2(x) \frac{\partial}{\partial x} \tilde{w} \right] - \operatorname{Re}\{\tilde{f}(x)e^{-i\omega t}\}, \quad x \in \mathbb{R}, \quad 0 \leq t \leq T,$$

$$\tilde{w}(0, x) = \tilde{v}_0(x), \quad \tilde{w}_t(0, x) = \tilde{v}_1(x),$$

where \tilde{v}_0 and \tilde{c} are the constant extensions (with $\gamma = \alpha/\beta$)

$$\tilde{v}_0(x) = \begin{cases} v_0(a_0), & x < a, \\ v_0(x), & a \leq x \leq b, \\ v_0(b_0), & b < x, \end{cases} \quad \tilde{c}(x) = \begin{cases} \gamma c_a, & x < a, \\ c(x), & a \leq x \leq b, \\ \gamma c_b, & b < x, \end{cases}$$

3085 and \tilde{v}_1 , \tilde{f} are zero extensions of v_1 and f , respectively.

3086 Since the solutions to the wave equation have finite speed of propagation, we may replace the
 3087 domain \mathbb{R} for \tilde{w} by a large enough domain with any boundary condition given that any reflections
 3088 at the new boundary do not re-enter the region $a \leq x \leq b$. Let $\tilde{a} < a - c_a T/2$ and $\tilde{b} > b + c_b T/2$. We
 3089 define the extension operator E such that $[v_0, v_1]^T \rightarrow [\tilde{v}_0, \tilde{v}_1]^T$ where \tilde{v}_0 and \tilde{c} are the extensions as
 3090 above and \tilde{v}_1 , \tilde{f} are zero extensions of v_1 and f , respectively. We now consider the (finite interval)
 3091 extended problem with homogeneous Neumann conditions

$$\tilde{w}_{tt} = \frac{\partial}{\partial x} \left[\tilde{c}^2(x) \frac{\partial}{\partial x} \tilde{w} \right] - \operatorname{Re}\{\tilde{f}(x)e^{-i\omega t}\}, \quad \tilde{a} \leq x \leq \tilde{b}, \quad 0 \leq t \leq T,$$

$$\tilde{w}(0, x) = \tilde{v}_0(x), \quad \tilde{w}_t(0, x) = \tilde{v}_1(x),$$

$$\tilde{w}_x(t, \tilde{a}) = 0, \quad \tilde{w}_x(t, \tilde{b}) = 0.$$

3092 Defining the projection operator P as the restriction of \tilde{w} to $a \leq x \leq b$ then it follows that $P\tilde{w} = w$
 3093 where w is the original wave solution to the interior impedance problem.

3094 .6 Verification of Discrete Solution

Here we verify that (3.27) is a solution to the difference equation (3.26). Direct substitution yields

$$\begin{aligned}
& \hat{w}_j^{n+1} - 2\hat{w}_j^n + \hat{w}_j^{n-1} + 2 \left[\sum_{k=1}^m \frac{(-1)^{k+1} \Delta t^{2k} \lambda_j^{2k}}{(2k)!} \right] \hat{w}_j^n = \\
& (\hat{v}_j - \hat{v}_j^\infty) \cos(\tilde{\lambda}_j t_n) \left(\cos(\tilde{\lambda}_j \Delta t) - 2 + 2 \sum_{k=1}^m \frac{(-1)^{k+1} \Delta t^{2k} \lambda_j^{2k}}{(2k)!} + \cos(\tilde{\lambda}_j \Delta t) \right) \\
& + \hat{v}_j^\infty \cos(\omega t_n) \left(\cos(\omega \Delta t) - 2 + 2 \sum_{k=1}^m \frac{(-1)^{k+1} \Delta t^{2k} \lambda_j^{2k}}{(2k)!} + \cos(\omega \Delta t) \right) \\
& = (\hat{v}_j - \hat{v}_j^\infty) \cos(\tilde{\lambda}_j t_n) \left(-4 \sin^2(\tilde{\lambda}_j \Delta t / 2) + 2 \sum_{k=1}^m \frac{(-1)^{k+1} \Delta t^{2k} \lambda_j^{2k}}{(2k)!} \right) \\
& + \hat{v}_j^\infty \cos(\omega t_n) \left(-4 \sin^2(\omega \Delta t / 2) + 2 \sum_{k=1}^m \frac{(-1)^{k+1} \Delta t^{2k} \lambda_j^{2k}}{(2k)!} \right) \\
& = \hat{v}_j^\infty \cos(\omega t_n) \left(-4 \sin^2(\omega \Delta t / 2) + 2 \sum_{k=1}^m \frac{(-1)^{k+1} \Delta t^{2k} \lambda_j^{2k}}{(2k)!} \right).
\end{aligned}$$

Finally, if

$$\sin^2(\omega \Delta t / 2) = \sum_{k=1}^m \frac{(-1)^{k+1} (\Delta t \tilde{\omega})^{2k}}{2(2k)!},$$

then

$$\begin{aligned}
& \hat{w}_j^{n+1} - 2\hat{w}_j^n + \hat{w}_j^{n-1} + 2 \left[\sum_{k=1}^m \frac{(-1)^{k+1} \Delta t^{2k} \lambda_j^{2k}}{(2k)!} \right] \hat{w}_j^n \\
& = \hat{v}_j^\infty \cos(\omega t_n) \left(-4 \sin^2(\omega \Delta t / 2) + 2 \sum_{k=1}^m \frac{(-1)^{k+1} \Delta t^{2k} \lambda_j^{2k}}{(2k)!} \right) \\
& = \hat{v}_j^\infty \cos(\omega t_n) \left(2 \sum_{k=1}^m \frac{(-1)^k \Delta t^{2k} (\tilde{\omega}^{2k} - \lambda_j^{2k})}{(2k)!} \right) \\
& = \hat{v}_j^\infty \cos(\omega t_n) \left(2 \sum_{k=1}^m \frac{(-1)^k \Delta t^{2k}}{(2k)!} (\tilde{\omega}^2 - \lambda_j^2) \sum_{\ell=0}^{k-1} \tilde{\omega}^{2(k-\ell-1)} \lambda^{2\ell} \right) \\
& = \hat{f}_j \cos(\omega t_n) \left(2 \sum_{k=1}^m \frac{(-1)^k \Delta t^{2k}}{(2k)!} \sum_{\ell=0}^{k-1} \tilde{\omega}^{2(k-\ell-1)} \lambda^{2\ell} \right),
\end{aligned}$$

since

$$\tilde{\omega}^{2k} - \lambda^{2k} = (\tilde{\omega}^2 - \lambda^2) \sum_{\ell=0}^{k-1} \tilde{\omega}^{2(k-\ell-1)} \lambda^{2\ell}, \quad k = 1, 2, \dots,$$

as desired. Moreover, the initial conditions are satisfied as

$$\begin{aligned}
\hat{w}_j^0 &= (\hat{v}_j - \hat{v}_j^\infty) + \hat{v}_j^\infty = \hat{v}_j, \\
\hat{w}_j^{-1} &= (\hat{v}_j - \hat{v}_j^\infty) \cos(\tilde{\lambda}_j \Delta t) + \hat{v}_j^\infty \cos(\omega \Delta t) \\
&= (\hat{v}_j - \hat{v}_j^\infty) \left(1 - \sum_{k=1}^m \frac{(-1)^{k+1} (\Delta t \lambda_j)^{2k}}{2k!} \right) + \hat{v}_j^\infty \left(1 - \sum_{k=1}^m \frac{(-1)^{k+1} (\Delta t \tilde{\omega})^{2k}}{2k!} \right) \\
&= \hat{v}_j \left(1 - \sum_{k=1}^m \frac{(-1)^{k+1} (\Delta t \lambda_j)^{2k}}{2k!} \right) - \hat{v}_j^\infty \left(\sum_{k=1}^m \frac{(-1)^{k+1} \Delta t^{2k} (\tilde{\omega}^{2k} - \lambda^{2k})}{2k!} \right) \\
&= \hat{v}_j \left(1 + \sum_{k=1}^m \frac{(-1)^k (\Delta t \lambda_j)^{2k}}{2k!} \right) + \hat{f}_j \left(\sum_{k=1}^m \frac{(-1)^k \Delta t^{2k}}{2k!} \sum_{\ell=0}^{k-1} \tilde{\omega}^{2(k-\ell-1)} \lambda^{2\ell} \right).
\end{aligned}$$

3095 This shows that (3.27) solves (3.26).

3096 .7 Well-definedness of modified frequencies

For an order $2m$ ME scheme we have the relation

$$\sin^2(\omega \Delta t / 2) = \sum_{j=1}^m \frac{(-1)^{j+1} (\Delta t \tilde{\omega})^{2j}}{2(2j)!} = \sin^2(\tilde{\omega} \Delta t / 2) + \mathcal{O}(\Delta t^{2m+2}), \quad (7)$$

for which we note that, under the assumptions of Theorem 3.3.1, $\Delta t, \omega > 0$ so that $\Delta t \tilde{\omega} = 0$ does not satisfy the above relation. We seek to show that if $\Delta t \tilde{\omega} \leq 1$ then we have a well-defined (and unique) real-valued $\tilde{\omega}$ such that $0 \leq \Delta t \tilde{\omega} \leq 2$. To that end, we now define the polynomial

$$p(x) = \sum_{j=1}^m \frac{(-1)^{j+1} x^j}{2(2j)!} - \sin^2(\omega \Delta t / 2), \quad (8)$$

and note that $(\Delta t \tilde{\omega})^2$ is a root of $p(x)$. On the interval $[0, 1/2]$ we have that $\sin^2(x)$ is increasing so that

$$0 \leq \sin^2(\omega \Delta t / 2) \leq \sin^2(1/2) \leq 0.23 < 1,$$

immediately giving $p(0) < 0$. Moreover,

$$\begin{aligned} p(4) &= -\frac{1}{2} \sum_{j=1}^m \frac{(-1)^j 2^{2j}}{(2j)!} - \sin^2(\omega \Delta t / 2) = -\frac{1}{2} \sum_{j=1}^{\infty} \frac{(-1)^j 2^{2j}}{(2j)!} + \frac{1}{2} \sum_{j=m+1}^{\infty} \frac{(-1)^j 2^{2j}}{(2j)!} - \sin^2(\omega \Delta t / 2) \\ &= -\frac{1}{2}(\cos(2) - 1) + \frac{1}{2} \sum_{j=m+1}^{\infty} \frac{(-1)^j 2^{2j}}{(2j)!} - \sin^2(\omega \Delta t / 2) \\ &= \sin^2(1) + \frac{1}{2} \sum_{j=m+1}^{\infty} \frac{(-1)^j 2^{2j}}{(2j)!} - \sin^2(\omega \Delta t / 2). \end{aligned}$$

We note that

$$\sum_{j=2}^{\infty} \frac{2^{2j}}{(2j)!} = \sum_{j=0}^{\infty} \frac{2^{2j}}{(2j)!} - 3 = \cosh(2) - 3,$$

so that

$$\frac{1}{2} \sum_{j=m+1}^{\infty} \frac{(-1)^j 2^{2j}}{(2j)!} > -\frac{1}{2} \sum_{j=m+1}^{\infty} \frac{2^{2j}}{(2j)!} \geq -\frac{1}{2} \sum_{j=2}^{\infty} \frac{2^{2j}}{(2j)!} = -\frac{1}{2}(\cosh(2) - 3).$$

This gives that

$$\begin{aligned} p(4) &= \sin^2(1) + \frac{1}{2} \sum_{j=m+1}^{\infty} \frac{(-1)^j 2^{2j}}{(2j)!} - \sin^2(\omega \Delta t / 2) \geq \sin^2(1) - \frac{1}{2}(\cosh(2) - 3) - \sin^2(\omega \Delta t / 2) \\ &\geq \sin^2(1) - \frac{1}{2}(\cosh(2) - 3) - \sin^2(1) > 0. \end{aligned}$$

By the intermediate value theorem, it follows that $p(x)$ has a root in the interval $[0, 4]$. We next need to show that $p'(x) \neq 0$ on this interval to guarantee the root is unique. Taking a derivative,

$$\frac{d}{dx} p(x) = \sum_{j=1}^m \frac{(-1)^{j+1} j x^{j-1}}{(2j)!} = \frac{1}{2} \left[1 + \sum_{j=2}^m \frac{(-1)^{j+1} x^{j-1}}{(2j-1)!} \right].$$

We then have

$$\sum_{j=2}^m \frac{(-1)^{j+1} x^{j-1}}{(2j-1)!} \geq -\sum_{j=2}^m \frac{2^{j-1}}{(2j-1)!} = -\frac{1}{\sqrt{2}} \sum_{j=2}^m \frac{\sqrt{2}^{2j-1}}{(2j-1)!} \geq -\frac{1}{\sqrt{2}} \left[\sum_{j=1}^{\infty} \frac{\sqrt{2}^{2j-1}}{(2j-1)!} - \sqrt{2} \right] = -\frac{\sinh(\sqrt{2})}{\sqrt{2}} + 1,$$

so that

$$\frac{d}{dx} p(x) = \frac{1}{2} \left[1 + \sum_{j=2}^m \frac{(-1)^{j+1} x^{j-1}}{(2j-1)!} \right] \geq \frac{1}{2} - \frac{\sinh(\sqrt{2})}{\sqrt{2}} + 1 > \frac{1}{10} > 0. \quad (9)$$

3097 This gives that there is a unique, real-valued $\tilde{\omega}$ with $\Delta t \tilde{\omega} \leq 2$ that satisfies the relation (7), which
 3098 we choose as our modified frequency.

We may similarly prove that the relation (3.28) is well-defined by showing that the right hand side of (3.28) is smaller than (or equal) to one in magnitude. Under the assumptions of Theorem 3.3.1, we have $\Delta t \lambda_j \leq 1$ for each j so that we may bound the right hand side of (3.28) via

$$\left| \sum_{k=1}^m \frac{(-1)^{k+1} (\Delta t \lambda_j)^{2k}}{2(2k)!} \right| \leq \sum_{k=1}^m \frac{(\Delta t \lambda_j)^{2k}}{2(2k)!} \leq \frac{1}{2} \sum_{k=1}^m \frac{1}{(2k)!} = \frac{1}{2} (\cosh(1) - 1) \approx 0.2715 < 1. \quad (10)$$

3099 .8 Error in discrete Helmholtz frequency

We let $x = \Delta t \omega$, $\tilde{x} = \Delta t \tilde{\omega}$, and $R_m = p(x^2) - p(\tilde{x}^2)$ where $p(x)$ is defined as in (8). By the mean value theorem we have

$$|R_m| = |p(x^2) - p(\tilde{x}^2)| = |(x^2 - \tilde{x}^2)p'(\xi)| = |x - \tilde{x}| |x + \tilde{x}| |p'(\xi)|,$$

for some $\xi \in [0, 2]$, so that

$$|x - \tilde{x}| \leq \frac{|R_m|}{|x + \tilde{x}| |p'(\xi)|}.$$

A straightforward bound gives

$$\begin{aligned} |R_m| &= |p(x^2) - \sin^2(\omega \Delta t / 2) + \sin^2(\omega \Delta t / 2) - p(\tilde{x}^2)| = |p(x^2) - \sin^2(\omega \Delta t / 2)| = \left| \sum_{j=m+1}^{\infty} \frac{(-1)^{j+2} (\Delta t \omega)^{2j}}{2(2j)!} \right| \\ &\leq \frac{(\Delta t \omega)^{2m+2}}{2(2m+2)!}, \end{aligned}$$

which gives

$$|x - \tilde{x}| \leq \frac{|R_m|}{|x + \tilde{x}| |p'(\xi)|} \leq \frac{\Delta t^{2m+1} \omega^{2m+2}}{2(2m+2)! (\omega + \tilde{\omega}) |p'(\xi)|} \implies |\omega - \tilde{\omega}| \leq \frac{\Delta t^{2m} \omega^{2m+2}}{2(2m+2)! (\omega + \tilde{\omega}) |p'(\xi)|}.$$

By (9) we have that $|p'(x)| > 1/10$ in $[0, 4]$, which finally gives

$$|\omega - \tilde{\omega}| \leq \frac{10 \Delta t^{2m} \omega^{2m+2}}{2(2m+2)! (\omega + \tilde{\omega})} \leq \frac{5 \Delta t^{2m} \omega^{2m+1}}{(2m+2)! (1 + \tilde{\omega}/\omega)} \leq \frac{5 \Delta t^{2m} \omega^{2m+1}}{(2m+2)!} \approx \mathcal{O}(\Delta t^{2m}). \quad (11)$$

3100 .9 Verification of Discrete Solution

Here we verify that (4.30) is a solution to the difference equation (4.29). Direct substitution yields

$$\begin{aligned}
& \left(1 + \frac{\Delta t^2}{2} \lambda_j^2\right) \hat{u}_j^{n+1} - \alpha \hat{u}_j^n + \left(1 + \frac{\Delta t^2}{2} \lambda_j^2\right) \hat{u}_j^{n-1} = \\
& (\hat{u}_{0,j} - \hat{v}_j) \cos(\tilde{\lambda}_j \Delta t_n) \left(\left(1 + \frac{\Delta t^2}{2} \lambda_j^2\right) \cos(\tilde{\lambda}_j \Delta t) - \alpha + \left(1 + \frac{\Delta t^2}{2} \lambda_j^2\right) \cos(\tilde{\lambda}_j \Delta t) \right) \\
& + \hat{v}_j \cos(\omega \Delta t_n) \left(\left(1 + \frac{\Delta t^2}{2} \lambda_j^2\right) \cos(\omega \Delta t) - \alpha + \left(1 + \frac{\Delta t^2}{2} \lambda_j^2\right) \cos(\omega \Delta t) \right) \\
& = (\hat{u}_{0,j} - \hat{v}_j) \cos(\tilde{\lambda}_j \Delta t_n) \left(\frac{\alpha}{2} - \alpha + \frac{\alpha}{2} \right) \\
& + \hat{v}_j \cos(\omega \Delta t_n) \cos(\omega \Delta t) \left(\left(1 + \frac{\Delta t^2}{2} \lambda_j^2\right) - (2 + \Delta t^2 \omega^2) + \left(1 + \frac{\Delta t^2}{2} \lambda_j^2\right) \right) \\
& = \hat{v}_j \cos(\omega \Delta t_n) \cos(\omega \Delta t) \Delta t^2 (\lambda_j^2 - \omega^2) \\
& = -\Delta t^2 \hat{f}_j \cos(\omega \Delta t_n) \cos(\omega \Delta t),
\end{aligned}$$

as desired. Moreover, we can check that the initial condition is satisfied since

$$\hat{u}_j^0 = (\hat{u}_{0,j} - \hat{v}_j) + \hat{v}_j = \hat{u}_{0,j},$$

and

$$\begin{aligned}
\hat{u}_j^{-1} &= (\hat{u}_{0,j} - \hat{v}_j) \cos(\tilde{\lambda}_j \Delta t) + \hat{v}_j \cos(\omega \Delta t) = \hat{u}_{0,j} \cos(\tilde{\lambda}_j \Delta t) + \hat{v}_j \left(\cos(\omega \Delta t) - \cos(\tilde{\lambda}_j \Delta t) \right) \\
&= \hat{u}_{0,j} \cos(\tilde{\lambda}_j \Delta t) + \hat{v}_j \cos(\omega \Delta t) \left(1 - \left(1 + \frac{\omega^2 \Delta t^2}{2} \right) \left(1 + \frac{\Delta t^2}{2} \lambda_j^2 \right)^{-1} \right) \\
&= \hat{u}_{0,j} \cos(\tilde{\lambda}_j \Delta t) + \hat{v}_j \cos(\omega \Delta t) \left(1 + \frac{\Delta t^2}{2} \lambda_j^2 \right)^{-1} \frac{\Delta t^2}{2} (\lambda_j^2 - \omega^2) \\
&= \left(1 + \frac{\Delta t^2}{2} \lambda_j^2 \right)^{-1} \left(\frac{\alpha}{2} \hat{u}_{0,j} - \frac{\Delta t^2}{2} \hat{f}_j \cos(\omega \Delta t) \right).
\end{aligned}$$

3101 This shows that (4.30) is indeed a solution to the implicit scheme (4.25).

3102 .10 Time-step Restriction

3103 To understand how restrictive the requirement (4.26) is, we plot α for various values of $\omega \Delta t$
3104 below in Figure 1.

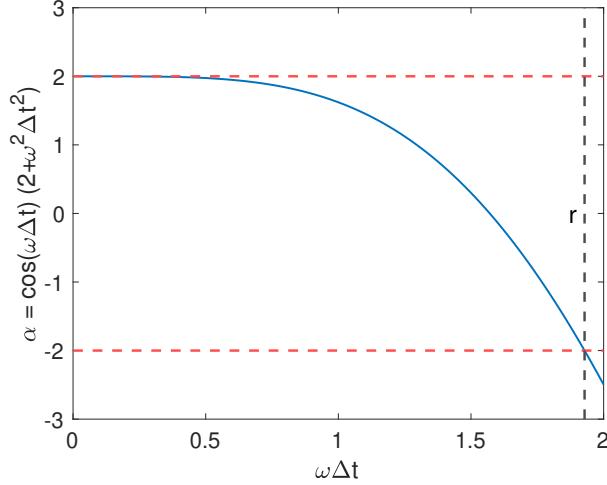


Figure 1: Values of $\alpha = \cos(\omega\Delta t)(2 + \omega^2\Delta t^2)$ for values of $\omega\Delta t$ in the interval $[0, 2]$. The red lines indicate the desired bound on α , and the black line indicates the maximum allowable value of $\omega\Delta t$ at $r \approx 1.93$.

From Figure 1 we see that $|\alpha| < 2$ for $\Delta t < r/\omega$ where $r \approx 1.93$. This choice of the time-step corresponds to a requirement of at least four time-steps per iteration. However, the WaveHoltz kernel is a constant with four time-steps so that at least five time-steps are needed for stability.

1.1 Motivation of Conjecture 1

For the stability requirement (4.28), which requires at least five time-steps per iteration, we note that from (4.31) we have that the right hand side is a decreasing function of λ so that

$$\cos(\tilde{\lambda}_j \Delta t) \leq \lim_{\lambda_j \rightarrow \infty} \cos(\omega \Delta t) \left(1 + \frac{\omega^2 \Delta t^2}{2}\right) \left(1 + \frac{\Delta t^2}{2} \lambda_j^2\right)^{-1} = 0,$$

implying that

$$0 \leq \tilde{\lambda}_j \leq \frac{\pi}{2\Delta t} = \frac{k}{4}\omega,$$

where $k \geq 5$ is the number of quadrature points used in the trapezoidal rule. We now consider the (continuous) rescaled filter transfer function,

$$\bar{\beta}(r) := \beta(r\omega) = \frac{2}{T} \int_0^T \left(\cos(\omega t) - \frac{1}{4} \right) \cos(r\omega t) dt = \frac{1}{\pi} \int_0^{2\pi} \left(\cos(t) - \frac{1}{4} \right) \cos(rt) dt,$$

with discrete analogue

$$\bar{\beta}_h(r) = \frac{\Delta t}{\pi} \sum_{n=0}^M \eta_n \cos(rt_n) \left(\cos(t_n) - \frac{1}{4} \right), \quad \eta_n = \begin{cases} \frac{1}{2}, & n = 0 \text{ or } n = M, \\ 1, & 0 < n < M. \end{cases}$$

Let us now take a look at the rescaled discrete filter function, $\bar{\beta}_h(r)$. It is sufficient to consider

only the range $r \in [0, k/4]$, which we plot in Figure 2.

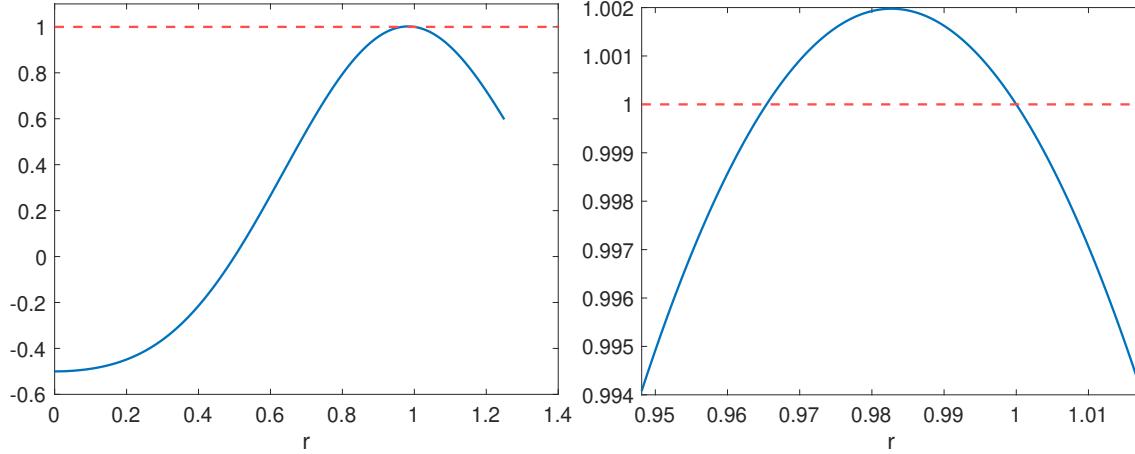


Figure 2: A plot of the discrete filter function using five time-steps $0 \leq r \leq 5/4$. On the left we plot the full range of values of r , and on the right we zoom in close to resonance, i.e. $r = 1$.

From Figure 2 we see that it is possible to integrate and get eigenvalues of the WHI operator to be larger than one for a small range near resonance, $r = 1$. To get a sense of the size of this gap, we perform a simple bisection where we find the leftmost point $r^* < 1$ such that $\beta_h(r^*) = 1$ for increasing number of quadrature points $k = 5, 6, \dots, 100$. We plot the result below in Figure 3.

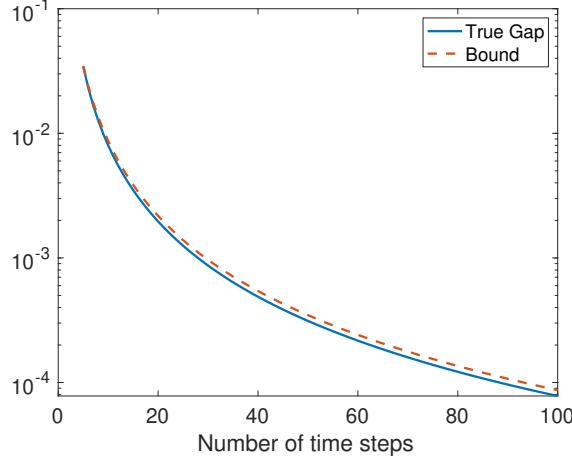


Figure 3: A bound on the gap from resonance that creates problematic modes. The blue curve is the true gap, $1 - r^*$, and the dotted red curve is a proposed bound.

From Figure 3 we see that, perhaps unsurprisingly, the gap shrinks with increasing number of quadrature points. The curve in red in Figure 3 indicates the bound

$$1 - r^* \leq 0.022 \cdot \Delta t^2,$$

so that we see that this gap shrinks as Δt^2 . Moreover, if $|1 - r| \geq 0.022 \cdot \Delta t^2$ then $|\beta_h(r)| < 1$. For $r = \tilde{\lambda}_j/\omega$, we may thus obtain convergence of the iteration if it can be guaranteed the time-step is chosen such that $\tilde{\lambda}_j \notin [\omega(1 - 0.022 \cdot \Delta t^2), \omega]$.

Assuming we have a bound similar to the above, let us now try to get a minimum on the distance $|\tilde{\lambda}_j - \omega|$. By the mean-value theorem we have the bound $|\cos(x) - \cos(y)| \leq |x - y|$, so that

$$\frac{|\cos(\tilde{\lambda}_j \Delta t) - \cos(\omega \Delta t)|}{\Delta t} \leq |\tilde{\lambda}_j - \omega|.$$

Some algebra shows that

$$\begin{aligned}
\cos(\omega\Delta t) - \cos(\tilde{\lambda}_j\Delta t) &= \cos(\omega\Delta t) - \cos(\omega\Delta t) \left(1 + \frac{\omega^2\Delta t^2}{2}\right) \left(1 + \frac{\Delta t^2}{2}\lambda_j^2\right)^{-1} \\
&= \cos(\omega\Delta t) \frac{\Delta t^2}{2} (\lambda_j^2 - \omega^2) \left(1 + \frac{\Delta t^2}{2}\lambda_j^2\right)^{-1} \\
&= \cos(\omega\Delta t) \frac{\Delta t^2}{2} (\lambda_j - \omega)(\lambda_j + \omega) \left(1 + \frac{\Delta t^2}{2}\lambda_j^2\right)^{-1} \\
&= \cos(\omega\Delta t) \frac{\Delta t^2}{2} \omega \delta_h (\lambda_j + \omega) \left(1 + \frac{\Delta t^2}{2}\lambda_j^2\right)^{-1},
\end{aligned}$$

where $\delta_h = \min_j |\lambda_j - \omega|/\omega > 0$ is the minimum gap to resonance. We then have

$$\begin{aligned}
|\tilde{\lambda}_j - \omega| &= \cos(\omega\Delta t) \frac{\Delta t}{2} \omega \delta_h (\lambda_j + \omega) \left(1 + \frac{\Delta t^2}{2}\lambda_j^2\right)^{-1} \geq \cos(\omega\Delta t) \frac{\Delta t}{2} \omega \delta_h (\lambda_1 + \omega) \left(1 + \frac{\Delta t^2}{2}\lambda_N^2\right)^{-1} \\
&\geq \cos(2\pi/5) \frac{\Delta t}{2} \omega \delta_h (\lambda_1 + \omega) \left(1 + \frac{\Delta t^2}{2}\lambda_N^2\right)^{-1} \\
&\geq \cos(2\pi/5) \frac{\Delta t}{2} \omega \delta_h (\lambda_1 + \omega) (1 + 2(\pi\lambda_N/5)^2)^{-1},
\end{aligned}$$

so that we need to choose Δt such that

$$\cos(\omega\Delta t) \frac{\Delta t}{2} \omega \delta_h (\lambda_1 + \omega) \left(1 + \frac{\Delta t^2}{2}\lambda_N^2\right)^{-1} \geq 0.022 \cdot \Delta t^2,$$

which gives the condition

$$\Delta t \leq \frac{\cos(2\pi/5)\omega^2\delta_h}{0.044 \cdot (1 + 2(\pi\lambda_N/5)^2)} \leq \frac{\cos(2\pi/5)\omega\delta_h(\lambda_1 + \omega)}{0.044 \cdot (1 + 2(\pi\lambda_N/5)^2)} \approx 7.02 \cdot \frac{\delta_h\omega(\lambda_1 + \omega)}{1 + 2(\pi\lambda_N/5)^2}.$$

3118

3119 .12 Composite quantum systems and essential states

To simplify the notation we assume a bipartite quantum system ($Q = 2$); the case $Q = 1$ is trivial and $Q > 2$ follows by straightforward generalizations. Let the number of energy levels in the subsystems be n_1 and n_2 , respectively, for a total of $N = n_1 \cdot n_2$ states in the coupled system. We use the canonical unit vectors $e_j^{(n_q)} \in \mathbb{R}^{n_q}$, for $j = 0, \dots, n_q - 1$, as a basis for subsystem q , where the superscript indicates its size. These basis vectors can be used to describe the total state of the

coupled system,

$$\psi = \sum_{j_2=0}^{n_2-1} \sum_{j_1=0}^{n_1-1} \psi_{j_2,j_1} \left(e_{j_2}^{(n_2)} \otimes e_{j_1}^{(n_1)} \right) = \sum_{k=0}^{N-1} \vec{\psi}_k e_k^{(N)}. \quad (12)$$

Here, $\vec{\psi} \in \mathbb{C}^N$ denotes the one-dimensional representation of the two-dimensional state vector ψ , using a natural ordering of the elements, i.e., $\vec{\psi}_k = \psi_{j_2,j_1}$ for $k = j_1 + n_1 j_2 =: k_{ind}(j_2, j_1)$. The mapping $k = k_{ind}(j_2, j_1)$ is invertible for $k \in [0, N - 1]$.

We classify the energy levels in the total state vector as either essential or guarded levels. The unitary gate transformation is only specified for the essential levels. The guard levels are retained to justify the truncation of the modal expansion of Schrödinger's equation, and to avoid leakage of probability to even higher energy levels.

Let the number of essential energy levels in the subsystems be m_1 and m_2 , respectively, where $0 < m_q \leq n_q$. Similar to the total state vector, we use the canonical unit vectors as a basis for the essential subspace of each subsystem. The total number of essential levels equals $E = m_1 \cdot m_2$. Let the essential energy levels in the total state vector be represented by the essential state vector ϕ . Similar to the full state vector, we flatten its two-dimensional indexing using a natural ordering,

$$\phi = \sum_{i_2=0}^{m_2-1} \sum_{i_1=0}^{m_1-1} \phi_{i_2,i_1} \left(e_{i_2}^{(m_2)} \otimes e_{i_1}^{(m_1)} \right) = \sum_{\ell=0}^{E-1} \vec{\phi}_{\ell} e_{\ell}^{(E)} \in \mathbb{C}^E, \quad (13)$$

where $\ell = i_1 + m_1 i_2 =: \ell_{ind}(i_2, i_1)$. The elements in the essential state vector are defined from the total state vector by $\phi_{i_2,i_1} = \psi_{i_2,i_1}$, for $i_1 \in [0, m_1 - 1]$ and $i_2 \in [0, m_2 - 1]$.

The initial condition for the solution operator matrix $U(t)$ in Schrödinger's equation (5.3) needs to span a basis for the E -dimensional essential state space. Here we use the canonical basis consisting of the unit vectors $e_{\ell}^{(E)}$. Let the columns of the initial condition matrix be $U_0 = [\mathbf{g}_0, \mathbf{g}_1, \dots, \mathbf{g}_{E-1}] \in \mathbb{R}^{N \times E}$. Because the total probabilities in each column vector \mathbf{g}_k must sum to one, the basis vectors in the total state space become

$$\mathbf{g}_{\ell} = U_0 e_{\ell}^{(E)}, \quad g_{k,\ell} = \begin{cases} 1, & k = k_{ind}(i_2(\ell), i_1(\ell)), \\ 0, & \text{otherwise,} \end{cases} \quad \text{for } \ell = 0, 1, \dots, E - 1. \quad (14)$$

³¹²⁹ Here, $i_2(\ell) = \lfloor \ell/m_1 \rfloor$ and $i_1(\ell) = \ell - m_1 \cdot i_2(\ell)$.

The target gate matrix $V_E \in \mathbb{C}^{E \times E}$ defines the unitary transformation between the essential levels in the initial and final states, $\phi_T = V_E \phi_0$, for all $\phi_0 \in \mathbb{C}^E$. Because V_E is unitary, each of its columns has norm one. To preserve total probabilities, we define the target gate transformation according to

$$V_{tg} = U_0 V_E \in \mathbb{C}^{N \times E}. \quad (15)$$

³¹³⁰ This implies that each column of V_{tg} also has norm one.

Example: A bipartite quantum system As a small example, consider a composite system considering of two subsystems, each with three energy levels, $n_1 = n_2 = 3$. In this case, the dimension of the full state vector is $N = 9$. It can be written as:

$$\psi = \sum_{j_2=0}^{n_2-1} \sum_{j_1=0}^{n_1-1} \psi_{j_2, j_1} \left(e_{j_2}^{(3)} \otimes e_{j_1}^{(3)} \right) = \begin{bmatrix} \psi_{0,0} \\ \psi_{0,1} \\ \psi_{0,2} \\ \psi_{1,0} \\ \psi_{1,1} \\ \psi_{1,2} \\ \psi_{2,0} \\ \psi_{2,1} \\ \psi_{2,2} \end{bmatrix}, \quad \vec{\psi} = \sum_{k=0}^8 \vec{\psi}_k e_k^{(9)} = \begin{bmatrix} \vec{\psi}_0 \\ \vec{\psi}_1 \\ \vec{\psi}_2 \\ \vec{\psi}_3 \\ \vec{\psi}_4 \\ \vec{\psi}_5 \\ \vec{\psi}_6 \\ \vec{\psi}_7 \\ \vec{\psi}_8 \end{bmatrix}. \quad (16)$$

If both systems have two essential levels, i.e. $m_1 = m_2 = 2$, there are $E = 4$ essential levels in the composite system. In this case the total and essential state vectors are related by

$$\vec{\phi} = \begin{bmatrix} \psi_{0,0} \\ \psi_{0,1} \\ \psi_{1,0} \\ \psi_{1,1} \end{bmatrix} = U_0^\dagger \vec{\psi}, \quad U_0^\dagger = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} \in \mathbb{R}^{4 \times 9}. \quad (17)$$

³¹³¹

3132 .13 The Hamiltonian in a rotating frame of reference

3133 The time-dependent and unitary change of variables $\tilde{\psi}(t) = R(t)\psi(t)$ where $R^\dagger R = I$, results
3134 in the transformed Schrödinger equation

$$\frac{d\tilde{\psi}}{dt} = -i\tilde{H}(t)\tilde{\psi}, \quad \tilde{H}(t) = R(t)H(t)R^\dagger(t) + i\dot{R}(t)R^\dagger(t). \quad (18)$$

The rotating frame of reference is introduced by taking the unitary transformation to be the matrix (5.14). Because both $R(t)$ and the system Hamiltonian (5.10) are diagonal, $RH_sR^\dagger = H_s$. The time derivative of the transformation can be written

$$\dot{R}(t) = \left(\bigoplus_{q=Q}^1 i\omega_{r,q} A_q^\dagger A_q \right) \left(\bigotimes_{q=Q}^1 \exp(i\omega_{r,q} t A_q^\dagger A_q) \right), \quad (19)$$

where \oplus denotes the Kronecker sum, $C \oplus D = C \otimes I_D + I_C \otimes D$. Therefore,

$$\dot{R}(t)R^\dagger(t) = \bigoplus_{q=Q}^1 i\omega_{r,q} A_q^\dagger A_q = \sum_{q=1}^Q i\omega_{r,q} a_q^\dagger a_q. \quad (20)$$

As a result, the first term in the Hamiltonian (5.10) is modified by the term $i\dot{R}(t)R^\dagger(t)$. After noting that $Ra_q R^\dagger = e^{-i\omega_{r,q} t} a_q$, the transformed Hamiltonian can be written as

$$H_s^{rw} = \sum_{q=1}^Q \left(\Delta_q a_q^\dagger a_q - \frac{\xi_q}{2} a_q^\dagger a_q^\dagger a_q a_q - \sum_{p>q} \xi_{qp} a_q^\dagger a_q a_p^\dagger a_p \right), \quad (21)$$

$$\tilde{H}_c(t) = \sum_{q=1}^Q f_q(t; \alpha) \left(e^{-i\omega_{r,q} t} a_q + e^{i\omega_{r,q} t} a_q^\dagger \right), \quad (22)$$

3135 where $\Delta_q = \omega_q - \omega_{r,q}$ is the detuning frequency. The above system Hamiltonian corresponds to
3136 (5.15).

To slow down the time scales in the control Hamiltonian, we want to absorb the highly oscillatory factors $\exp(\pm i\omega_{r,q} t)$ into $f_q(t)$. Because the control function $f_q(t)$ is real-valued, this can only be done in an approximate fashion. We make the ansatz,

$$f_q(t) := 2 \operatorname{Re} (d_q(t)e^{i\omega_{r,q} t}) = d_q(t)e^{i\omega_{r,q} t} + \bar{d}_q(t)e^{-i\omega_{r,q} t}, \quad (23)$$

where \bar{d}_q denotes the complex conjugate of d_q . By substituting this expression into the transformed control Hamiltonian (22), we get

$$\tilde{H}_c(t) = \sum_{q=1}^Q \left(d_q(t)a_q + \bar{d}_q(t)a_q^\dagger + \bar{d}_q(t)e^{-2i\omega_{r,q}t}a_q + d_q(t)e^{2i\omega_{r,q}t}a_q^\dagger \right).$$

3137 The rotating frame approximation follows by ignoring terms that oscillate with frequency, $\pm 2i\omega_{r,q}$,
3138 resulting in the approximate control Hamiltonian (5.16).

3139 .14 Conditions for resonance

Consider the scalar function $y(t) := \psi_j^{(1)}(t)$. It satisfies an ordinary differential equation of the form

$$\frac{dy(t)}{dt} + \kappa_j y(t) = \sum_{\ell} c_{\ell} e^{i\nu_{\ell} t}, \quad \nu_k \in \mathbb{R}. \quad (24)$$

3140 We are interested in cases when $y(t)$ grows in time, corresponding to resonance. Conditions for
3141 resonance are provided in the following lemma.

Lemma .14.1. *Let $\kappa \in \mathbb{R}$ and $\nu \in \mathbb{R}$ be constants. The solution of the scalar ordinary differential equation*

$$\frac{dy(t)}{dt} + i\kappa y(t) = ce^{i\nu t}, \quad y(0) = y_0, \quad (25)$$

is given by

$$y(t) = \begin{cases} y_0 e^{-i\kappa t} + cte^{-i\kappa t}, & \nu + \kappa = 0, \\ y_0 e^{-i\kappa t} - \frac{ic}{\nu + \kappa} (e^{i\nu t} - e^{-i\kappa t}), & \text{otherwise.} \end{cases} \quad (26)$$

3142 Corresponding to resonance, the function $y(t)$ grows linearly in time when $\nu + \kappa = 0$ and $c \neq 0$.

3143 *Proof.* Follows by direct evaluation. \square

We proceed by analyzing the right hand side of (5.20). It can be shown that the forcing function $\mathbf{f}^{(k)}(t)$ is of the form The forcing function $\mathbf{f}^{(k)}(t)$ contains the terms

$$\{a_k \psi^{(0)}\}_{\mathbf{j}} = \begin{cases} g_{\mathbf{j}+e_k} \sqrt{j_k + 1} e^{-(i\kappa_{\mathbf{j}+e_k} t)}, & j_k \in [0, n_1 - 2], \\ 0, & j_k = n_k - 1, \end{cases} \quad (27)$$

and

$$\{a_k^\dagger \psi^{(0)}\}_j = \begin{cases} 0, & j_k = 0, \\ g_{j-\mathbf{e}_k} \sqrt{j_k} e^{-(i\kappa_{j-\mathbf{e}_k} t)}, & j_k \in [1, n_k - 1]. \end{cases} \quad (28)$$

Therefore,

$$\mathbf{f}_j^{(k)}(t) = \begin{cases} -ig_{j+\mathbf{e}_k} \sqrt{j_k + 1} e^{i(\Omega_k - \kappa_{j+\mathbf{e}_k})t}, & j_k = 0, \\ \Theta_k(t), & j_k \in [1, n_k - 2], \\ -ig_{j-\mathbf{e}_k} \sqrt{j_k} e^{-i(\Omega_k + \kappa_{j-\mathbf{e}_k})t}, & j_k = n_k - 1, \end{cases} \quad (29)$$

3144 where $\Theta_k(t) = -ig_{j+\mathbf{e}_k} \sqrt{j_k + 1} e^{i(\Omega_k - \kappa_{j+\mathbf{e}_k})t} - ig_{j-\mathbf{e}_k} \sqrt{j_k} e^{-i(\Omega_k + \kappa_{j-\mathbf{e}_k})t}$.

The right hand side satisfies $\mathbf{f}(t) = \mathbf{f}^{(1)}(t) + \mathbf{f}^{(2)}(t)$. The first set of frequencies and coefficients on the right hand side of (24) satisfy

$$\nu_1 = \Omega_k - \kappa_{j+\mathbf{e}_k}, \quad c_1 = -ig_{j+\mathbf{e}_k} \sqrt{j_k + 1},$$

for $k = \{1, 2\}$ and $j_k \in [0, n_k - 2]$. The second set of frequencies and coefficients are

$$\nu_2 = -(\Omega_k + \kappa_{j-\mathbf{e}_k}), \quad c_2 = -ig_{j-\mathbf{e}_k} \sqrt{j_k}.$$

3145 for $k = \{1, 2\}$ and $j_k \in [1, n_k - 1]$. From Lemma .14.1, component $\psi_j^{(1)}(t)$ is in resonance if

3146 $(\kappa_j + \nu_1 = 0, c_1 \neq 0)$ or $(\kappa_j + \nu_2 = 0, c_2 \neq 0)$. These conditions are equivalent to (5.22) and (5.23),

3147 which

3148 .15 Derivation of the discrete adjoint scheme

3149 We seek to determine a scheme for evolving the Lagrange multiplier (adjoint) variables to
3150 satisfy the first order optimality conditions (5.57). In the following, let $\delta_{r,s}$ denote the usual
3151 Kronecker delta function.

The terms T_j^3 to T_j^6 in (5.53) enforce the relations between the stage variables (5.45)-(5.48)

according to

$$T_j^3 = \sum_{n=0}^{M-1} \left\langle \mathbf{U}_j^{n,1} - \mathbf{u}_j^n, \mathbf{M}_j^{n,1} \right\rangle_2, \quad (30)$$

$$T_j^4 = \sum_{n=0}^{M-1} \left\langle \mathbf{U}_j^{n,2} - \mathbf{u}_j^n - \frac{h}{2} \left(S_n \mathbf{U}_j^{n,1} + S_{n+1} \mathbf{U}_j^{n,2} - K_n \mathbf{V}_j^{n,1} - K_{n+1} \mathbf{V}_j^{n,2} \right), \mathbf{M}_j^{n,2} \right\rangle_2, \quad (31)$$

$$T_j^5 = \sum_{n=0}^{M-1} \left\langle \mathbf{V}_j^{n,1} - \mathbf{v}_j^n - \frac{h}{2} \left(K_{n+1/2} \mathbf{U}_j^{n,1} + S_{n+1/2} \mathbf{V}_j^{n,1} \right), \mathbf{N}_j^{n,1} \right\rangle_2, \quad (32)$$

$$T_j^6 = \sum_{n=0}^{M-1} \left\langle \mathbf{V}_j^{n,2} - \mathbf{v}_j^n - \frac{h}{2} \left(K_{n+1/2} \mathbf{U}_j^{n,1} + S_{n+1/2} \mathbf{V}_j^{n,1} \right), \mathbf{N}_j^{n,2} \right\rangle_2. \quad (33)$$

Taking the derivative of (5.53) with respect to \mathbf{u}_j^r

$$0 = \frac{\partial \mathcal{L}^h}{\partial \mathbf{u}_j^r} = \frac{\partial \mathcal{J}^h}{\partial \mathbf{u}_j^r} - \left[(\boldsymbol{\mu}_j^n - \boldsymbol{\mu}_j^{n+1}) \delta_{r,n} + \boldsymbol{\mu}_j^M \delta_{r,M} - (\mathbf{M}_j^{n,1} + \mathbf{M}_j^{n,2}) \delta_{r,n} \right],$$

which gives the conditions

$$\boldsymbol{\mu}_j^M = \frac{\partial \mathcal{J}^h}{\partial \mathbf{u}_j^M}, \quad \boldsymbol{\mu}_j^n - \boldsymbol{\mu}_j^{n+1} = \mathbf{M}_j^{n,1} + \mathbf{M}_j^{n,2}, \quad n = 0, 1, \dots, M-1.$$

Similarly, differentiating (5.53) with respect to \mathbf{v}_j^r gives

$$0 = \frac{\partial \mathcal{L}^h}{\partial \mathbf{v}_j^r} = \frac{\partial \mathcal{J}^h}{\partial \mathbf{v}_j^r} - \left[(\boldsymbol{\nu}_j^n - \boldsymbol{\nu}_j^{n+1}) \delta_{r,n} + \boldsymbol{\nu}_j^M \delta_{r,M} - (\mathbf{N}_j^{n,1} + \mathbf{N}_j^{n,2}) \delta_{r,n} \right],$$

which leads to the conditions

$$\boldsymbol{\nu}_j^n - \boldsymbol{\nu}_j^{n+1} = \mathbf{N}_j^{n,1} + \mathbf{N}_j^{n,2}, \quad \boldsymbol{\nu}_j^M = \frac{\partial \mathcal{J}^h}{\partial \mathbf{v}_j^M}.$$

Next we take the derivative of (5.53) with respect to $\mathbf{U}_j^{n,1}$,

$$\begin{aligned}\frac{\partial \mathcal{L}^h}{\partial \mathbf{U}_j^{n,1}} &= \frac{\partial \mathcal{J}^h}{\partial \mathbf{U}_j^{n,1}} - \sum_{i=1}^6 \frac{\partial T_j^i}{\partial \mathbf{U}_j^{n,1}} = 0, \\ \frac{\partial T_j^1}{\partial \mathbf{U}_j^{n,1}} &= -\frac{h}{2} S_n^T \boldsymbol{\mu}_j^{n+1}, \\ \frac{\partial T_j^2}{\partial \mathbf{U}_j^{n,1}} &= -\frac{h}{2} K_{n+1/2}^T \boldsymbol{\nu}_j^{n+1}, \\ \frac{\partial T_j^3}{\partial \mathbf{U}_j^{n,1}} &= \mathbf{M}_j^{n,1}, \\ \frac{\partial T_j^4}{\partial \mathbf{U}_j^{n,1}} &= -\frac{h}{2} S_n^T \mathbf{M}_j^{n,2}, \\ \frac{\partial T_j^5}{\partial \mathbf{U}_j^{n,1}} &= -\frac{h}{2} K_{n+1/2}^T \mathbf{N}_j^{n,1}, \\ \frac{\partial T_j^6}{\partial \mathbf{U}_j^{n,1}} &= -\frac{h}{2} K_{n+1/2}^T \mathbf{N}_j^{n,2},\end{aligned}$$

which, using the fact that $S_n^T = -S_n$ and $K_n^T = K_n$, we may write as

$$\mathbf{M}_j^{n,1} + \frac{h}{2} S_n (\boldsymbol{\mu}_j^{n+1} + \mathbf{M}_j^{n,2}) - \frac{h}{2} K_{n+1/2} (\boldsymbol{\nu}_j^{n+1} + \mathbf{N}_j^{n,1} + \mathbf{N}_j^{n,2}) = \frac{\partial \mathcal{J}^h}{\partial \mathbf{U}_j^{n,1}}.$$

Repeating this procedure for the derivative with respect to $\mathbf{U}_j^{n,2}$ gives

$$\begin{aligned}\frac{\partial \mathcal{L}^h}{\partial \mathbf{U}_j^{n,2}} &= \frac{\partial \mathcal{J}^h}{\partial \mathbf{U}_j^{n,2}} - \sum_{i=1}^6 \frac{\partial T_j^i}{\partial \mathbf{U}_j^{n,2}} = 0, \\ \frac{\partial T_j^1}{\partial \mathbf{U}_j^{n,2}} &= -\frac{h}{2} S_{n+1}^T \boldsymbol{\mu}_j^{n+1}, \\ \frac{\partial T_j^2}{\partial \mathbf{U}_j^{n,2}} &= -\frac{h}{2} K_{n+1/2}^T \boldsymbol{\nu}_j^{n+1}, \\ \frac{\partial T_j^4}{\partial \mathbf{U}_j^{n,2}} &= \mathbf{M}_j^{n,2} - \frac{h}{2} S_{n+1}^T \mathbf{M}_j^{n,2}, \\ \frac{\partial T_j^3}{\partial \mathbf{U}_j^{n,2}} &= \frac{\partial T_j^5}{\partial \mathbf{U}_j^{n,2}} = \frac{\partial T_j^6}{\partial \mathbf{U}_j^{n,2}} = 0,\end{aligned}$$

which we may write compactly as

$$\mathbf{M}_j^{n,2} + \frac{h}{2} S_{n+1} (\boldsymbol{\mu}_j^{n+1} + \mathbf{M}_j^{n,2}) - \frac{h}{2} K_{n+1/2} \boldsymbol{\nu}_j^{n+1} = \frac{\partial \mathcal{J}^h}{\partial \mathbf{U}_j^{n,2}}.$$

Taking the derivative of (5.53) with respect to $\mathbf{V}_j^{n,1}$ gives the set of equations

$$\begin{aligned}\frac{\partial \mathcal{L}^h}{\partial \mathbf{V}_j^{n,1}} &= \frac{\partial \mathcal{J}^h}{\partial \mathbf{V}_j^{n,1}} - \sum_{i=1}^6 \frac{\partial T_j^i}{\partial \mathbf{V}_j^{n,1}} = 0, \\ \frac{\partial T_j^1}{\partial \mathbf{V}_j^{n,1}} &= \frac{h}{2} K_n^T \boldsymbol{\mu}_j^{n+1}, \\ \frac{\partial T_j^2}{\partial \mathbf{V}_j^{n,1}} &= -\frac{h}{2} S_{n+1/2}^T \boldsymbol{\nu}_j^{n+1}, \\ \frac{\partial T_j^3}{\partial \mathbf{V}_j^{n,1}} &= 0, \\ \frac{\partial T_j^4}{\partial \mathbf{V}_j^{n,1}} &= \frac{h}{2} K_n^T \mathbf{M}_j^{n,2}, \\ \frac{\partial T_j^5}{\partial \mathbf{V}_j^{n,1}} &= \mathbf{N}_j^{n,1} - \frac{h}{2} S_{n+1/2}^T \mathbf{N}_j^{n,1}, \\ \frac{\partial T_j^6}{\partial \mathbf{V}_j^{n,1}} &= -\frac{h}{2} S_{n+1/2}^T \mathbf{N}_j^{n,2},\end{aligned}$$

which gives the condition

$$\mathbf{N}_j^{n,1} + \frac{h}{2} S_{n+1/2} \left(\boldsymbol{\nu}_j^{n+1} + \mathbf{N}_j^{n,1} + \mathbf{N}_j^{n,2} \right) + \frac{h}{2} K_n \left(\boldsymbol{\mu}_j^{n+1} + \mathbf{M}_j^{n,2} \right) = \frac{\partial \mathcal{J}^h}{\partial \mathbf{V}_j^{n,1}}.$$

Similarly, taking the derivative with respect to $\mathbf{V}_j^{n,2}$ gives

$$\begin{aligned}\frac{\partial \mathcal{L}^h}{\partial \mathbf{V}_j^{n,2}} &= \frac{\partial \mathcal{J}^h}{\partial \mathbf{V}_j^{n,2}} - \sum_{i=1}^6 \frac{\partial T_j^i}{\partial \mathbf{V}_j^{n,2}} = 0, \\ \frac{\partial T_j^1}{\partial \mathbf{V}_j^{n,2}} &= \frac{h}{2} K_{n+1}^T \boldsymbol{\mu}_j^{n+1}, \\ \frac{\partial T_j^2}{\partial \mathbf{V}_j^{n,2}} &= -\frac{h}{2} S_{n+1/2}^T \boldsymbol{\nu}_j^{n+1}, \\ \frac{\partial T_j^4}{\partial \mathbf{V}_j^{n,2}} &= \frac{h}{2} K_{n+1}^T \mathbf{M}_j^{n,2}, \\ \frac{\partial T_j^6}{\partial \mathbf{V}_j^{n,2}} &= \mathbf{N}_j^{n,2}, \\ \frac{\partial T_j^3}{\partial \mathbf{V}_j^{n,2}} &= \frac{\partial T_j^5}{\partial \mathbf{V}_j^{n,2}} = 0,\end{aligned}$$

giving

$$\mathbf{N}_j^{n,2} + \frac{h}{2} S_{n+1/2} \boldsymbol{\nu}_j^{n+1} + \frac{h}{2} K_{n+1} \left(\boldsymbol{\mu}_j^{n+1} + \mathbf{M}_j^{n,2} \right) = \frac{\partial \mathcal{J}^h}{\partial \mathbf{V}_j^{n,2}}.$$

In summary, the first order optimality conditions (5.57) are satisfied if the following equations hold:

$$\boldsymbol{\mu}_j^n - \boldsymbol{\mu}_j^{n+1} = \mathbf{M}_j^{n,1} + \mathbf{M}_j^{n,2}, \quad \boldsymbol{\mu}_j^M = \frac{\partial \mathcal{J}^h}{\partial \mathbf{u}_j^M}, \quad (34)$$

$$\boldsymbol{\nu}_j^n - \boldsymbol{\nu}_j^{n+1} = \mathbf{N}_j^{n,1} + \mathbf{N}_j^{n,2}, \quad \boldsymbol{\nu}_j^M = \frac{\partial \mathcal{J}^h}{\partial \mathbf{v}_j^M}, \quad (35)$$

$$\mathbf{M}_j^{n,1} + \frac{h}{2} S_n (\boldsymbol{\mu}_j^{n+1} + \mathbf{M}_j^{n,2}) - \frac{h}{2} K_{n+1/2} (\boldsymbol{\nu}_j^{n+1} + \mathbf{N}_j^{n,1} + \mathbf{N}_j^{n,2}) = \frac{\partial \mathcal{J}^h}{\partial \mathbf{U}_j^{n,1}}, \quad (36)$$

$$\mathbf{M}_j^{n,2} + \frac{h}{2} S_{n+1} (\boldsymbol{\mu}_j^{n+1} + \mathbf{M}_j^{n,2}) - \frac{h}{2} K_{n+1/2} \boldsymbol{\nu}_j^{n+1} = \frac{\partial \mathcal{J}^h}{\partial \mathbf{U}_j^{n,2}}, \quad (37)$$

$$\mathbf{N}_j^{n,1} + \frac{h}{2} S_{n+1/2} (\boldsymbol{\nu}_j^{n+1} + \mathbf{N}_j^{n,1} + \mathbf{N}_j^{n,2}) + \frac{h}{2} K_n (\boldsymbol{\mu}_j^{n+1} + \mathbf{M}_j^{n,2}) = \frac{\partial \mathcal{J}^h}{\partial \mathbf{V}_j^{n,1}}, \quad (38)$$

$$\mathbf{N}_j^{n,2} + \frac{h}{2} S_{n+1/2} \boldsymbol{\nu}_j^{n+1} + \frac{h}{2} K_{n+1} (\boldsymbol{\mu}_j^{n+1} + \mathbf{M}_j^{n,2}) = \frac{\partial \mathcal{J}^h}{\partial \mathbf{V}_j^{n,2}}. \quad (39)$$

We now consider the following change of variables

$$\mathbf{X}_j^n = \boldsymbol{\mu}_j^{n+1} + \mathbf{M}_j^{n,2}, \quad (40)$$

$$\mathbf{Y}_j^{n,1} = \boldsymbol{\nu}_j^{n+1} + \mathbf{N}_j^{n,1} + \mathbf{N}_j^{n,2}, \quad (41)$$

$$\mathbf{Y}_j^{n,2} = \boldsymbol{\nu}_j^{n+1}, \quad (42)$$

which, upon substitution into (36)-(39), gives the set of equations

$$\mathbf{M}_j^{n,1} + \frac{h}{2} S_n \mathbf{X}_j^n - \frac{h}{2} K_{n+1/2} \mathbf{Y}_j^{n,1} = \frac{\partial \mathcal{J}^h}{\partial \mathbf{U}_j^{n,1}}, \quad (43)$$

$$\mathbf{M}_j^{n,2} + \frac{h}{2} S_{n+1} \mathbf{X}_j^n - \frac{h}{2} K_{n+1/2} \mathbf{Y}_j^{n,2} = \frac{\partial \mathcal{J}^h}{\partial \mathbf{U}_j^{n,2}}, \quad (44)$$

$$\mathbf{N}_j^{n,1} + \frac{h}{2} S_{n+1/2} \mathbf{Y}_j^{n,1} + \frac{h}{2} K_n \mathbf{X}_j^n = \frac{\partial \mathcal{J}^h}{\partial \mathbf{V}_j^{n,1}}, \quad (45)$$

$$\mathbf{N}_j^{n,2} + \frac{h}{2} S_{n+1/2} \mathbf{Y}_j^{n,2} + \frac{h}{2} K_{n+1} \mathbf{X}_j^n = \frac{\partial \mathcal{J}^h}{\partial \mathbf{V}_j^{n,2}}. \quad (46)$$

By adding (43)-(44),

$$\mathbf{M}_j^{n,1} + \mathbf{M}_j^{n,2} = -\frac{h}{2} [(S_n + S_{n+1}) \mathbf{X}_j^n - K_{n+1/2} (\mathbf{Y}_j^{n,1} + \mathbf{Y}_j^{n,2})] + \frac{\partial \mathcal{J}^h}{\partial \mathbf{U}_j^{n,1}} + \frac{\partial \mathcal{J}^h}{\partial \mathbf{U}_j^{n,2}}. \quad (47)$$

Similarly, by adding (45)-(46),

$$\mathbf{N}_j^{n,1} + \mathbf{N}_j^{n,2} = -\frac{h}{2} [S_{n+1/2} (\mathbf{Y}_j^{n,1} + \mathbf{Y}_j^{n,2}) + (K_n + K_{n+1}) \mathbf{X}_j^n] + \frac{\partial \mathcal{J}^h}{\partial \mathbf{V}_j^{n,1}} + \frac{\partial \mathcal{J}^h}{\partial \mathbf{V}_j^{n,2}}. \quad (48)$$

Thus, (34)-(35) can be rewritten as

$$\boldsymbol{\mu}_j^n - \boldsymbol{\mu}_j^{n+1} = -\frac{h}{2} \left[(S_n + S_{n+1}) \mathbf{X}_j^n - K_{n+1/2} (\mathbf{Y}_j^{n,1} + \mathbf{Y}_j^{n,2}) \right] + \frac{\partial \mathcal{J}^h}{\partial \mathbf{U}_j^{n,1}} + \frac{\partial \mathcal{J}^h}{\partial \mathbf{U}_j^{n,2}} \quad (49)$$

$$\boldsymbol{\nu}_j^n - \boldsymbol{\nu}_j^{n+1} = -\frac{h}{2} \left[S_{n+1/2} (\mathbf{Y}_j^{n,1} + \mathbf{Y}_j^{n,2}) + (K_n + K_{n+1}) \mathbf{X}_j^n \right] + \frac{\partial \mathcal{J}^h}{\partial \mathbf{V}_j^{n,1}} + \frac{\partial \mathcal{J}^h}{\partial \mathbf{V}_j^{n,2}} \quad (50)$$

By combining $\mathbf{X}_j^n = \boldsymbol{\mu}_j^{n+1} + \mathbf{M}_j^{n,2}$ and (44),

$$\mathbf{X}_j^n = \boldsymbol{\mu}_j^{n+1} - \frac{h}{2} S_{n+1} \mathbf{X}_j^n + \frac{h}{2} K_{n+1/2} \mathbf{Y}_j^{n,2} + \frac{\partial \mathcal{J}^h}{\partial \mathbf{U}_j^{n,2}}. \quad (51)$$

Similarly, by combining $\mathbf{Y}_j^{n,1} = \boldsymbol{\nu}_j^{n+1} + \mathbf{N}_j^{n,1} + \mathbf{N}_j^{n,2}$ and (48),

$$\mathbf{Y}_j^{n,1} = \boldsymbol{\nu}_j^{n+1} - \frac{h}{2} \left[S_{n+1/2} (\mathbf{Y}_j^{n,1} + \mathbf{Y}_j^{n,2}) + (K_n + K_{n+1}) \mathbf{X}_j^n \right] + \frac{\partial \mathcal{J}^h}{\partial \mathbf{V}_j^{n,1}} + \frac{\partial \mathcal{J}^h}{\partial \mathbf{V}_j^{n,2}}. \quad (52)$$

3152 The time-stepping scheme is completed by the relation

$$\mathbf{Y}_j^{n,2} = \boldsymbol{\nu}_j^{n+1}. \quad (53)$$

3153 The scheme (49)-(53) may be written in the form of Lemma 5.4.1 by defining the slopes

3154 according to (5.61)-(5.64). This completes the proof of the lemma.

3155 **.16 Proof of Corollary 1**

By rearranging (5.59) and (5.60),

$$\boldsymbol{\mu}_j^{n+1} = \boldsymbol{\mu}_j^n + \frac{h}{2} (\boldsymbol{\kappa}_j^{n,1} + \boldsymbol{\kappa}_j^{n,2}), \quad (54)$$

$$\boldsymbol{\nu}_j^{n+1} = \boldsymbol{\nu}_j^n + \frac{h}{2} (\boldsymbol{\ell}_j^{n,1} + \boldsymbol{\ell}_j^{n,2}). \quad (55)$$

3156 Hence, $b_1^\mu = b_2^\mu = 1/2$ and $b_1^\nu = b_2^\nu = 1/2$.

To express the stage variables in standard form we substitute (54) into (5.65) and define

$\mathbf{X}_j^{n,1} = \mathbf{X}_j^{n,2} = \mathbf{X}_j^n$. Similarly, we substitute (55) into (5.66) and (5.67), resulting in

$$\mathbf{X}_j^{n,1} = \boldsymbol{\mu}_j^n + \frac{h}{2} \boldsymbol{\kappa}_j^{n,1},$$

$$\mathbf{X}_j^{n,2} = \boldsymbol{\mu}_j^n + \frac{h}{2} \boldsymbol{\kappa}_j^{n,1},$$

$$\mathbf{Y}_j^{n,1} = \boldsymbol{\nu}_j^n,$$

$$\mathbf{Y}_j^{n,2} = \boldsymbol{\nu}_j^n + \frac{h}{2} (\boldsymbol{\ell}_j^{n,1} + \boldsymbol{\ell}_j^{n,2}).$$

3157 From these relations we can identify $a_{11}^\mu = a_{21}^\mu = 1/2$ and $a_{12}^\mu = a_{22}^\mu = 0$. Furthermore, $a_{11}^\nu = a_{12}^\nu = 0$
3158 and $a_{21}^\nu = a_{22}^\nu = 1/2$.

For the case without forcing, the formulae for the slopes, (5.65)-(5.67), become

$$\kappa_j^{n,1} = S_n \mathbf{X}_j^{n,1} - K_{n+1/2} \mathbf{Y}_j^{n,1}, \quad (56)$$

$$\kappa_j^{n,2} = S_{n+1} \mathbf{X}_j^{n,2} - K_{n+1/2} \mathbf{Y}_j^{n,2}, \quad (57)$$

$$\ell_j^{n,1} = K_n \mathbf{X}_j^{n,1} + S_{n+1/2} \mathbf{Y}_j^{n,1}, \quad (58)$$

$$\ell_j^{n,2} = K_{n+1} \mathbf{X}_j^{n,2} + S_{n+1/2} \mathbf{Y}_j^{n,2}. \quad (59)$$

3159 They are consistent approximations of the time derivatives $\dot{\mu}(t_n)$ and $\dot{\nu}(t_n)$, respectively. The
3160 scheme is therefore a consistent approximation of the continuous adjoint system.

3161 .17 Computing the gradient of the discrete objective function

3162 Given a solution that satisfies the saddle point conditions of (5.56) and (5.57), the gradient
3163 of $\mathcal{L}_h(\boldsymbol{\alpha})$ satisfies

$$\frac{d\mathcal{L}_h}{d\alpha_r} = \frac{\partial \mathcal{J}_{1h}}{\partial \alpha_r}(\mathbf{u}, \mathbf{v}) + \frac{\partial \mathcal{J}_{2h}}{\partial \alpha_r}(\mathbf{U}, \mathbf{V}), \quad r = 1, 2, \dots, D.$$

The gradient of \mathcal{L}_h with respect to $\boldsymbol{\alpha}$ only gets a contribution from the terms in T_j^q that involve the matrices K and S . Let $S'_n = \partial S / \partial \alpha_r(t_n)$ and $K'_n = \partial K / \partial \alpha_r(t_n)$. We have,

$$\begin{aligned} \frac{\partial T_j^1}{\partial \alpha_r} &= -\frac{h}{2} \sum_{n=0}^{M-1} \left\langle S'_n \mathbf{U}_j^{n,1} - K'_n \mathbf{V}_j^{n,1} + S'_{n+1} \mathbf{U}_j^{n,2} - K'_{n+1} \mathbf{V}_j^{n,2}, \boldsymbol{\mu}_j^{n+1} \right\rangle_2, \\ \frac{\partial T_j^2}{\partial \alpha_r} &= -\frac{h}{2} \sum_{n=0}^{M-1} \left\langle K'_{n+1/2} (\mathbf{U}_j^{n,1} + \mathbf{U}_j^{n,2}) + S'_{n+1/2} (\mathbf{V}_j^{n,1} + \mathbf{V}_j^{n,2}), \boldsymbol{\nu}_j^{n+1} \right\rangle_2, \\ \frac{\partial T_j^3}{\partial \alpha_r} &= 0, \\ \frac{\partial T_j^4}{\partial \alpha_r} &= -\frac{h}{2} \sum_{n=0}^{M-1} \left\langle S'_n \mathbf{U}_j^{n,1} - K'_n \mathbf{V}_j^{n,1} + S'_{n+1} \mathbf{U}_j^{n,2} - K'_{n+1} \mathbf{V}_j^{n,2}, \mathbf{M}_j^{n,2} \right\rangle_2, \\ \frac{\partial T_j^5}{\partial \alpha_r} &= -\frac{h}{2} \sum_{n=0}^{M-1} \left\langle K'_{n+1/2} \mathbf{U}_j^{n,1} + S'_{n+1/2} \mathbf{V}_j^{n,1}, \mathbf{N}_j^{n,1} \right\rangle_2, \\ \frac{\partial T_j^6}{\partial \alpha_r} &= -\frac{h}{2} \sum_{n=0}^{M-1} \left\langle K'_{n+1/2} \mathbf{U}_j^{n,1} + S'_{n+1/2} \mathbf{V}_j^{n,1}, \mathbf{N}_j^{n,2} \right\rangle_2. \end{aligned}$$

³¹⁶⁴ We note that

$$\frac{\partial(T_j^5 + T_j^6)}{\partial \alpha_r} = -\frac{h}{2} \sum_{n=0}^{M-1} \left\langle K'_{n+1/2} \mathbf{U}_j^{n,1} + S'_{n+1/2} \mathbf{V}_j^{n,1}, \mathbf{N}_j^{n,1} + \mathbf{N}_j^{n,2} \right\rangle_2.$$

Let \mathbf{X}_j^n and $\mathbf{Y}_j^{n,i}$ be defined by (40)-(42). We have,

$$\begin{aligned} \frac{\partial T_j^4}{\partial \alpha_r} &= -\frac{h}{2} \sum_{n=0}^{M-1} \left\langle S'_n \mathbf{U}_j^{n,1} - K'_n \mathbf{V}_j^{n,1} + S'_{n+1} \mathbf{U}_j^{n,2} - K'_{n+1} \mathbf{V}_j^{n,2}, \mathbf{X}_j^n - \boldsymbol{\mu}_j^{n+1} \right\rangle_2, \\ \frac{\partial(T_j^5 + T_j^6)}{\partial \alpha_r} &= -\frac{h}{2} \sum_{n=0}^{M-1} \left\langle K'_{n+1/2} \mathbf{U}_j^{n,1} + S'_{n+1/2} \mathbf{V}_j^{n,1}, \mathbf{Y}_j^{n,1} - \boldsymbol{\nu}_j^{n+1} \right\rangle_2. \end{aligned}$$

³¹⁶⁵ Thus,

$$\frac{\partial(T_j^1 + T_j^4)}{\partial \alpha_r} = -\frac{h}{2} \sum_{n=0}^{M-1} \left\langle S'_n \mathbf{U}_j^{n,1} - K'_n \mathbf{V}_j^{n,1} + S'_{n+1} \mathbf{U}_j^{n,2} - K'_{n+1} \mathbf{V}_j^{n,2}, \mathbf{X}_j^n \right\rangle_2.$$

Furthermore, from the relation (42),

$$\begin{aligned} \frac{\partial(T_j^2 + T_j^5 + T_j^6)}{\partial \alpha_r} &= -\frac{h}{2} \sum_{n=0}^{M-1} \left\langle K'_{n+1/2} \mathbf{U}_j^{n,1} + S'_{n+1/2} \mathbf{V}_j^{n,1}, \mathbf{Y}_j^{n,1} \right\rangle_2 \\ &\quad - \frac{h}{2} \sum_{n=0}^{M-1} \left\langle K'_{n+1/2} \mathbf{U}_j^{n,2} + S'_{n+1/2} \mathbf{V}_j^{n,2}, \mathbf{Y}_j^{n,2} \right\rangle_2, \end{aligned}$$

We can further simplify the expressions by recognizing that $\mathbf{V}^{n,1} = \mathbf{V}^{n,2}$. By collecting the terms,

$$\begin{aligned} \frac{\partial \mathcal{L}_h}{\partial \alpha_r} &= \frac{h}{2} \sum_{j=0}^{E-1} \sum_{n=0}^{M-1} \left(\left\langle S'_n \mathbf{U}_j^{n,1} + S'_{n+1} \mathbf{U}_j^{n,2} - (K'_n + K'_{n+1}) \mathbf{V}_j^{n,1}, \mathbf{X}_j^n \right\rangle_2 \right. \\ &\quad + \left\langle K'_{n+1/2} \mathbf{U}_j^{n,1} + S'_{n+1/2} \mathbf{V}_j^{n,1}, \mathbf{Y}_j^{n,1} \right\rangle_2 \\ &\quad \left. + \left\langle K'_{n+1/2} \mathbf{U}_j^{n,2} + S'_{n+1/2} \mathbf{V}_j^{n,2}, \mathbf{Y}_j^{n,2} \right\rangle_2 \right). \end{aligned}$$

³¹⁶⁶ This completes the proof of the lemma.